FIRST-ORDER CHARACTERIZATION OF PROBABILISTIC ROBUSTNESS FOR CLASSIFIERS VIA THE COAREA FORMULA

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training (AT) is a standard approach for improving worst-case robustness, but its worst-case formulation can be overly conservative when perturbations are predominantly stochastic. This has motivated PR-targeted training, where existing methods typically rely either on surrogate risk objectives or on adaptations of adversarial-search procedures, leaving the first-order behavior of PR itself largely unexplored. In this paper, we study the first-order optimization structure of PR and ask when its variation can be represented by only a few perturbations. Under mild regularity conditions, we express the PR gradient via the co-area formula as a distribution-weighted integral over the decision boundary, revealing an inherently aggregate boundary effect. To study this quantity in high dimensions, we use a Gaussian kernel as a computational probe, yielding a simple Monte Carlo estimator of the PR first variation. Low- and high-dimensional experiments show that the extent to which a small number of perturbations represent the aggregate PR variation depends strongly on the underlying geometry and dataset. We further study how the estimated PR variation depends on the chosen margin level and smoothing bandwidth, and explore its use in PR-oriented training.

1 INTRODUCTION

Deep learning (DL) models have achieved strong empirical performance across many tasks, yet remain vulnerable to adversarial examples (AEs) Szegedy et al. (2014), where small input perturbations can alter model predictions. Most work on adversarial robustness (AR) focuses on point-wise worst-case perturbations that maximize loss within a prescribed budget Huang et al. (2020). Probabilistic robustness (PR) Webb et al. (2019); Zhao (2025) instead adopts a distributional perspective, measuring the probability that a model remains correct under random perturbations. Such perturbations may arise from stochastic environmental effects, sensor noise, or simple random-noise attacks that are not explicitly optimized against the model.

Adversarial training (AT) is a standard approach to improving adversarial robustness, but its worst-case formulation can be overly conservative when perturbations are predominantly stochastic. This has motivated PR-oriented methods based either on surrogate risk measures of the perturbation-induced loss distribution Wang et al. (2021); Robey et al. (2022); Zhang et al. (2024a) or on adversarial-search-based proxies Zhang et al. (2025b;a). Yet the first-order optimization structure of PR remains poorly understood, since its indicator-based definition prevents direct gradient computation by standard backpropagation. This motivates the following central research question:

How does probabilistic robustness change under small model updates, and when can this change be captured by a small number of perturbations?

To address this question, we characterize the first-order behavior of PR under model updates. Under mild regularity conditions, we express the PR gradient as a distribution-weighted integral over the decision boundary via the co-area formula Federer (1969); Evans & Gariepy (1992). This reveals an aggregate boundary effect, while allowing representative perturbations to emerge under favorable geometry. We illustrate this behavior on a two-dimensional Palmer Penguins example with smooth MLPs. In high dimensions, directly integrating over the decision boundary is infeasible. We therefore use a Gaussian kernel centered at zero margin as a computational probe, which replaces

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.

the boundary integral by a weighted expectation over perturbations near the boundary. This yields a simple Monte Carlo estimator of the PR first variation and allows us to study how well a small number of perturbations can represent the underlying aggregate effect.

Our main contributions are summarized as follows:

- We derive the first variation of indicator-defined PR via the co-area formula, revealing a distribution-weighted aggregate effect over the decision boundary.
- We use a Gaussian kernel as a computational probe of this boundary integral, yielding a simple Monte Carlo estimator that makes the PR first variation accessible in high-dimensional models.
- We provide low- and high-dimensional empirical analyses showing that the extent to which a small number of perturbations can represent the aggregate PR variation is strongly geometry- and dataset-dependent. We further examine the roles of the target margin level and kernel bandwidth, and use the estimator in a simple training proof of concept.

2 Related Work

2.1 Adversarial vs. Probabilistic Robustness

AR has become a central topic in deep learning since the seminal observations that neural networks can be misled by small, visually imperceptible input perturbations (Szegedy et al., 2014; Goodfellow, 2015). In contrast, PR evaluates robustness from a distributional perspective, measuring the likelihood that a model remains correct when perturbations are drawn from a prescribed reference distribution (Webb et al., 2019; Weng et al., 2019; Zhang et al., 2023; 2024b; Zhao, 2025). While AR is determined by the existence of a worst-case perturbation, PR depends on the probability mass and aggregate geometry of the adversarial region under a reference perturbation distribution.

Early studies of PR primarily focused on estimating such probabilistic robustness from sampled perturbations. For example, Webb et al. (2019) formulated robustness evaluation as a statistical estimation problem, while subsequent work developed more efficient estimation and certification techniques (Weng et al., 2019; TIT et al., 2023; Wang et al., 2026). PR has also been extended beyond standard norm-bounded perturbations to functional transformations and other stochastic perturbation settings (Zhang et al., 2023; 2024b). Recent surveys and benchmarks provide broader comparisons of probabilistic robustness measures and estimators (Zhao, 2025; Zhang et al., 2025b).

2.2 Training for Probabilistic Robustness

Compared with the relatively extensive literature on PR evaluation, substantially less attention has been devoted to training models directly for probabilistic robustness. A first line of work formulates PR-oriented learning by optimizing statistics of the perturbation-induced surrogate-loss distribution. Representative approaches minimize the mean perturbation loss (Wang et al., 2021), conditional value-at-risk (CVaR) (Robey et al., 2022), or entropic value-at-risk (EVaR) (Zhang et al., 2024a), thereby placing different degrees of emphasis on high-loss perturbations. While these objectives provide tractable surrogates for PR-oriented training, they optimize statistics of a surrogate loss distribution rather than the indicator-defined probabilistic robustness itself.

Interestingly, AT, although originally developed for worst-case robustness, has also been observed to improve PR as a by-product (Zhang et al., 2025b). Building on this connection, AT-PR (Zhang et al., 2025a) explicitly adapts the AT framework toward PR by redesigning the inner search to identify adversarial examples associated with larger adversarial regions, using multi-start PGD together with decision-boundary exploration. While this AT-based route can achieve strong PR, the improvement is either indirect, as in conventional AT, or relies on a substantially more involved inner optimization, as in AT-PR, and may still inherit the clean-accuracy and generalization trade-offs of adversarial training.

Existing PR-targeted methods either optimize surrogate risk measures or rely on adversarial-search-based proxies, leaving the first-order response of the indicator-defined PR objective to model updates largely unexplored. We address this gap by deriving its first variation and expressing it, via the co-area formula, as a distribution-weighted integral over the decision boundary.

2.3 Margin and Decision-Boundary Methods

A substantial line of work has studied adversarial robustness through the geometry of the decision boundary. Early approaches such as DeepFool characterize vulnerability through the distance to the nearest decision boundary (Moosavi-Dezfooli et al., 2016), while MMA explicitly improves robustness by maximizing this input-space margin during training (Ding et al., 2020). Beyond boundary distance, subsequent studies have shown that local curvature, flatness, and boundary thickness also play important roles in adversarial robustness and generalization (Fawzi et al., 2016; 2018; Moosavi-Dezfooli et al., 2019; Yang et al., 2020). Together, these works establish decision-boundary geometry as a fundamental perspective for understanding and improving adversarial robustness.

Beyond geometric characterization, the decision boundary has also been used explicitly as a target for adversarial search and sample construction, for example in Boundary Attack, GeoDA, and FAB (Brendel et al., 2017; Rahmati et al., 2020; Croce & Hein, 2020). However, these methods primarily use boundary geometry to locate informative or minimal adversarial perturbations, rather than to characterize how boundary regions collectively determine a distribution-defined notion of robustness. In contrast, we show that the first-order variation of PR can be expressed as a distribution-weighted integral over the decision boundary, providing a direct link between probabilistic robustness and boundary geometry.

3 Preliminary

Consider a classification problem over $\mathcal{X} \times \mathcal{Y}$, where $(\boldsymbol{x}, y) \sim P$ denotes an input–label pair drawn from the data distribution P. Let $h_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^C$ be a classifier parameterized by $\boldsymbol{\theta} \in \Theta$, where C is the number of classes.

For each input $\boldsymbol{x}$, let $\Delta(\boldsymbol{x}) \subseteq \mathbb{R}^d$ denote the admissible perturbation region, consisting of perturbations $\boldsymbol{\delta}$ such that $\boldsymbol{x} + \boldsymbol{\delta}$ remains a valid input. A common example is an ℓ_p-bounded region,

$$\Delta(\boldsymbol{x}) = \{\boldsymbol{\delta} \mid \|\boldsymbol{\delta}\|_p \le \epsilon,\ \boldsymbol{x} + \boldsymbol{\delta} \in \mathcal{X}\}. \tag{1}$$

We further consider a reference perturbation distribution $q(\boldsymbol{\delta} \mid \boldsymbol{x})$ supported on $\Delta(\boldsymbol{x})$, which determines how perturbations within this region are weighted when evaluating probabilistic robustness.

Definition 3.1 (Margin)**.** Given a parameterized classifier $h_{\boldsymbol{\theta}}$, a labeled example $(\boldsymbol{x}, y)$, and a perturbation $\boldsymbol{\delta} \in \Delta(\boldsymbol{x})$, we define the margin vector as

$$\boldsymbol{m}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \triangleq \left(m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\right)_{j \neq y} \in \mathbb{R}^{C-1}, \qquad m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \triangleq h_{\boldsymbol{\theta},y}(\boldsymbol{x}+\boldsymbol{\delta}) - h_{\boldsymbol{\theta},j}(\boldsymbol{x}+\boldsymbol{\delta}). \tag{2}$$

When $(\boldsymbol{x}, y)$ is clear from context, we simply write $\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{\delta})$ and $m^{(j)}_{\boldsymbol{\theta}}(\boldsymbol{\delta})$. We further denote the minimum component of the margin vector by $m^{\min}_{\boldsymbol{\theta}}(\boldsymbol{\delta})$.

Definition 3.2 (Adversarial Region and Decision Boundary)**.** Under the setting of Def. 3.1, let $m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})$ denote the minimum margin. The adversarial region is defined as

$$\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \triangleq \left\{\boldsymbol{\delta} \mid m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) < 0\right\}. \tag{3}$$

and we denote the decision boundary as $\partial\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y}$. For each $j \neq y$, the corresponding class-wise decision face is given by

$$\partial_j\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \triangleq \left\{\boldsymbol{\delta} \in \partial\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \mid m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) = 0\right\}. \tag{4}$$

Definition 3.3 (Non-Robustness Probability (NPR))**.** Under Def. 3.2, the local NPR for an input–label pair $(\boldsymbol{x}, y)$ under a perturbation distribution $q(\boldsymbol{\delta} \mid \boldsymbol{x})$ supported on $\Delta(\boldsymbol{x})$ is defined as

$$\mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) \triangleq \int_{\Delta(\boldsymbol{x})} \mathbf{1}_{\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y}}(\boldsymbol{\delta}) q(\boldsymbol{\delta} \mid \boldsymbol{x})\, d\boldsymbol{\delta}. \tag{5}$$

At the population level, we define

$$\mathrm{NPR}_q(\boldsymbol{\theta}) \triangleq \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta})\right]. \tag{6}$$

Remark 3.4. PR and non-robustness probability (NPR) are complementary quantities, satisfying $\mathrm{PR}_q = 1 - \mathrm{NPR}_q$. Since our analysis is naturally expressed in terms of the probability mass of the adversarial region, we work with NPR throughout the paper.

Definition 3.5 (Margin Fiber)**.** Under the setting of Def. 3.1, for any $\boldsymbol{z} \in \mathbb{R}^{C-1}$, the margin fiber associated with $\boldsymbol{z}$ is defined as

$$\mathcal{M}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{z}) \triangleq \{\boldsymbol{\delta} \mid \boldsymbol{m}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) = \boldsymbol{z}\}. \tag{7}$$

To integrate quantities over the margin fibers introduced above, we recall the co-area formula for Lipschitz maps (Nicolaescu, 2021; Evans & Gariepy, 1992).

Theorem 3.6 (Co-area Formula for Lipschitz Maps)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be open and $F : \Omega \to \mathbb{R}^k$, $1 \le k \le d$, be Lipschitz. By Rademacher's theorem, F is differentiable for Lebesgue a.e. $\boldsymbol{\delta} \in \Omega$. At every point of differentiability, define the k-dimensional Jacobian of F by*

$$J_F(\boldsymbol{\delta}) \triangleq \sqrt{\det(DF(\boldsymbol{\delta})DF(\boldsymbol{\delta})^\top)}. \tag{8}$$

On the Lebesgue-null set where F is not differentiable, J_F may be defined arbitrarily. Then, for every nonnegative measurable function $\varphi : \Omega \to [0, \infty]$,

$$\int_\Omega \varphi(\boldsymbol{\delta}) J_F(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \int_{\mathbb{R}^k} \left(\int_{\Omega \cap F^{-1}(\boldsymbol{z})} \varphi(\boldsymbol{\delta})\, d\mathcal{H}^{d-k}(\boldsymbol{\delta}) \right) d\boldsymbol{z}, \tag{9}$$

where $\mathcal{H}^{d-k}$ denotes the $(d-k)$-dimensional Hausdorff measure.

4 First-Order Characterization of PR

Since PR and NPR are complementary, it suffices to study the first-order variation of NPR with respect to the model parameters. Although NPR is defined as the probability mass of the adversarial region in perturbation space, the co-area formula allows us to decompose this quantity along the fibers of the margin map. This decomposition connects the volumetric definition of NPR to the boundary-supported first variation derived below.

Proposition 4.1 (Co-area Reformulation of Local NPR)**.** *Fix an input–label pair $(\boldsymbol{x}, y)$ and $\boldsymbol{\theta} \in \Theta$. Let $k = C - 1 \le d$, and suppose that the margin map $\boldsymbol{m}_{\boldsymbol{\theta},\boldsymbol{x},y}$ is Lipschitz on an open neighborhood of $\Delta(\boldsymbol{x})$. Let $q(\boldsymbol{\delta} \mid \boldsymbol{x})$ be a measurable perturbation density on $\Delta(\boldsymbol{x})$, and assume that it assigns zero mass to the rank-deficient set of the margin map, i.e.,*

$$\int_{\left\{\boldsymbol{\delta} \in \Delta(\boldsymbol{x}) \mid J_{\boldsymbol{m}_{\boldsymbol{\theta},\boldsymbol{x},y}}(\boldsymbol{\delta}) = 0\right\}} q(\boldsymbol{\delta} \mid \boldsymbol{x})\, d\boldsymbol{\delta} = 0. \tag{10}$$

Then the local NPR admits the decomposition

$$\mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) = \int_{\{\boldsymbol{z} \in \mathbb{R}^{C-1} \mid \min_{j \neq y} z_j < 0\}} \left(\int_{\mathcal{M}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{z}) \cap \Delta(\boldsymbol{x})} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{J_{\boldsymbol{m}_{\boldsymbol{\theta},\boldsymbol{x},y}}(\boldsymbol{\delta})}\, d\mathcal{H}^{d-k}(\boldsymbol{\delta}) \right) d\boldsymbol{z}, \tag{11}$$

with $q/J_{\boldsymbol{m}} = 0$ on $\{J_{\boldsymbol{m}} = 0\}$.

By the co-area formula, the integral over perturbation space is decomposed into an outer integral over margin values $\boldsymbol{z}$ and an inner integral over the corresponding fibers $\mathcal{M}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{z})$. We assume that the margin map has full local rank for q-a.e. perturbation. This is a natural non-degeneracy condition when the input dimension is much larger than the number of margin coordinates. A more general treatment that relaxes this assumption and the proof of Prop. 4.1 are provided in Appendix A.1. Before analyzing the variation of NPR, we first state the assumptions used in our analysis.

Assumption 4.2 (Smoothness)**.** The classifier $h_{\boldsymbol{\theta}}$ is continuously differentiable jointly with respect to its parameters and inputs in a neighborhood of $\{\boldsymbol{\theta}\} \times (\boldsymbol{x} + \Delta(\boldsymbol{x}))$. The perturbation density $q(\boldsymbol{\delta} \mid \boldsymbol{x})$ is continuous in a neighborhood of the relevant class-wise decision faces.

Remark 4.3 (Piecewise-smooth classifiers). Assumption 4.2 is stated for continuously differentiable classifiers for clarity. For piecewise-smooth models such as ReLU networks, the analysis can instead be carried out stratum-wise, with additional conditions on the intersections between decision faces and nondifferentiable strata. We discuss this extension in Appx. B.

Assumption 4.4 (Regularity and transversality)**.** For each $j \neq y$, zero is a regular value of $m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}$ on $\Delta(\boldsymbol{x})$, i.e.,

$$\nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \neq \mathbf{0} \tag{12}$$

whenever $m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) = 0$. Moreover, distinct class-wise zero-margin hypersurfaces intersect transversally with one another and with each smooth stratum of $\partial\Delta(\boldsymbol{x})$ whenever such intersections occur.

Intuitively, regularity ensures that each class-wise zero-margin set is locally a smooth hypersurface with a well-defined normal. Transversality ensures that intersections between distinct faces, and between these faces and $\partial\Delta(\boldsymbol{x})$, are lower-dimensional and hence negligible under $\mathcal{H}^{d-1}$.

Theorem 4.5 (First Variation of NPR for Smooth Classifiers)**.** *Under the notation of Prop. 4.1, suppose Assumptions 4.2–4.4 hold. Then the local NPR is differentiable at $\boldsymbol{\theta}$, with*

$$\nabla_{\boldsymbol{\theta}} \mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\sum_{j \neq y} \int_{\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\left\| \nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \right\|_2} \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\, d\mathcal{H}^{d-1}(\boldsymbol{\delta}). \tag{13}$$

If the assumptions hold for P-almost every $(\boldsymbol{x}, y)$ and the local gradients are integrably dominated, then

$$\nabla_{\boldsymbol{\theta}} \mathrm{NPR}_q(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim P} \left[\nabla_{\boldsymbol{\theta}} \mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) \right]. \tag{14}$$

For geometric interpretation and subsequent estimation, it is convenient to collect the class-wise contributions into the minimum scalar margin.

Corollary 4.6 (Scalar-Margin Form of the NPR First Variation)**.** *Under the assumptions of Thm. 4.5, the first variation can equivalently be expressed in terms of the minimum scalar margin $m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}$ as*

$$\nabla_{\boldsymbol{\theta}} \mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\int_{\partial \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\left\| \nabla_{\boldsymbol{\delta}} m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \right\|_2} \nabla_{\boldsymbol{\theta}} m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\, d\mathcal{H}^{d-1}(\boldsymbol{\delta}). \tag{15}$$

The derivatives of $m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}$ are understood $\mathcal{H}^{d-1}$-almost everywhere on the decision boundary. Indeed, except on intersections between distinct class-wise faces, which are $\mathcal{H}^{d-1}$-null by transversality, the minimum margin locally coincides with a unique active class-wise margin.

The proof is provided in Appendix A.3. Cor. 4.6 provides a simple geometric picture of PR optimization. Improving PR can be viewed geometrically as pushing the decision boundary so that the adversarial region occupies progressively less probability mass inside $\Delta(\boldsymbol{x})$. The boundary weighting identifies where such motion matters most. Moving some parts of the boundary removes substantially more adversarial probability mass than moving others. Thus, PR optimization is not simply about moving the boundary, but about moving the most consequential parts of it.

To answer our central research question, Eq. 15 shows that the first-order evolution of PR is governed by the aggregate contribution of the decision boundary. Accordingly, a small number of perturbations can describe this evolution only to the extent that their local contributions are representative of the full boundary integral.

An additional observation is that the local contribution to PR variation is determined jointly by input-space geometry and parameter-space sensitivity. The factor $\|\nabla_{\boldsymbol{\delta}} m_{\boldsymbol{\theta}}(\boldsymbol{\delta})\|_2^{-1}$ controls how a margin change translates into boundary motion, while $\nabla_{\boldsymbol{\theta}} m_{\boldsymbol{\theta}}(\boldsymbol{\delta})$ determines how the model update changes that margin. Thus, input-space geometry alone is insufficient to determine the PR optimization direction.

To further examine these observations empirically, we introduce a simple estimator of the NPR descent direction and use it in the experiments that follow.

5 Estimating the NPR Descent Direction

Direct evaluation of the NPR boundary integral is generally intractable in high dimensions. We therefore seek a tractable estimator based on samples from the ambient perturbation distribution.

Following Cor. 4.6, we write

$$m(\boldsymbol{\delta}) \triangleq m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}), \qquad m^{-1}(0) = \partial\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x}), \tag{16}$$

where m is understood to be restricted to $\Delta(\boldsymbol{x})$. The surface integral in Eq. 15 therefore induces a boundary distribution on $m^{-1}(0)$ with surface density

$$\pi(d\boldsymbol{\delta}) \propto \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\|\nabla_{\boldsymbol{\delta}} m(\boldsymbol{\delta})\|_2}\, d\mathcal{H}^{d-1}(\boldsymbol{\delta}), \qquad \boldsymbol{\delta} \in m^{-1}(0). \tag{17}$$

Sampling from π is generally difficult in high dimensions because it is supported on the codimension-one level set $m^{-1}(0)$ and follows the non-uniform surface density in Eq. 17. In contrast, perturbations drawn from $q(\cdot \mid \boldsymbol{x})$ lie in the ambient space and hit $m^{-1}(0)$ with probability zero. We therefore replace exact sampling on the zero-margin boundary with a Gaussian kernel that localizes ambient perturbations near $m(\boldsymbol{\delta}) = 0$, i.e.,

$$K_\tau(t) = \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{t^2}{2\tau^2}\right), \tag{18}$$

where $\tau > 0$ controls the degree of localization. Define

$$g_\tau \triangleq \mathbb{E}_{\boldsymbol{\delta}\sim q(\cdot|\boldsymbol{x})}\left[K_\tau(m(\boldsymbol{\delta}))\nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta})\right]. \tag{19}$$

Applying the co-area formula yields

$$g_\tau = \int_{\mathbb{R}} K_\tau(t) \left[\int_{m^{-1}(t)} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\|\nabla_{\boldsymbol{\delta}} m(\boldsymbol{\delta})\|_2} \nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta})\, d\mathcal{H}^{d-1}(\boldsymbol{\delta})\right] dt. \tag{20}$$

Under the regularity conditions of Assumptions 4.2–4.4, together with suitable local domination, the level-set integral in brackets is continuous at $t = 0$. Since K_τ converges to the Dirac mass at zero as $\tau \to 0$,

$$\lim_{\tau\to 0} g_\tau = \int_{m^{-1}(0)} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\|\nabla_{\boldsymbol{\delta}} m(\boldsymbol{\delta})\|_2} \nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta})\, d\mathcal{H}^{d-1}(\boldsymbol{\delta}) = -\nabla_{\boldsymbol{\theta}} \mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}), \tag{21}$$

We discuss connections to smooth-loss formulations and other robustness surrogates in Appendix C. We now turn to a finite-sample Monte Carlo approximation of g_τ.

Given i.i.d. perturbations $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_N \sim q(\cdot \mid \boldsymbol{x})$, we approximate g_τ by the Monte Carlo estimator

$$\widehat{g}_{\tau,N} \triangleq \frac{1}{N} \sum_{i=1}^{N} K_\tau(m(\boldsymbol{\delta}_i))\nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta}_i). \tag{22}$$

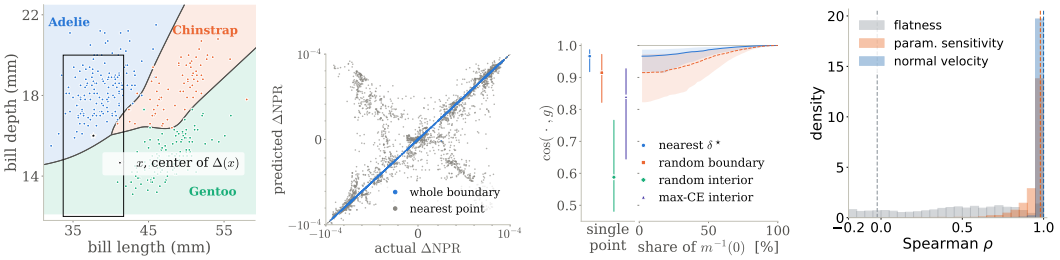
For fixed $\tau > 0$, $\widehat{g}_{\tau,N}$ is an unbiased estimator of g_τ and converges to g_τ as N increases. Together with Eq. 21, this provides a finite-sample approximation to the boundary-supported first-order variation of NPR.

6 Experiments

Our experiments are designed to examine the geometric implications of the NPR first variation derived above. We ask whether the boundary-based characterization predicts the observed change in NPR, and when a small number of perturbations can represent the aggregate boundary contribution. We study these questions first in a directly observable two-dimensional setting and then in high-dimensional image classification.

6.1 Experiment Setup

Low-dimensional setting. We use the Palmer Archipelago Penguins dataset and retain bill length and bill depth as the two input features, resulting in 342 examples after removing missing values. This two-dimensional setting allows the perturbation region and zero-margin decision boundary to be directly visualized and numerically tracked under model updates. We train smooth GELU MLPs across multiple widths and random seeds. Full training and numerical details are provided in Appendix D.



(a) (b) (c) (d)

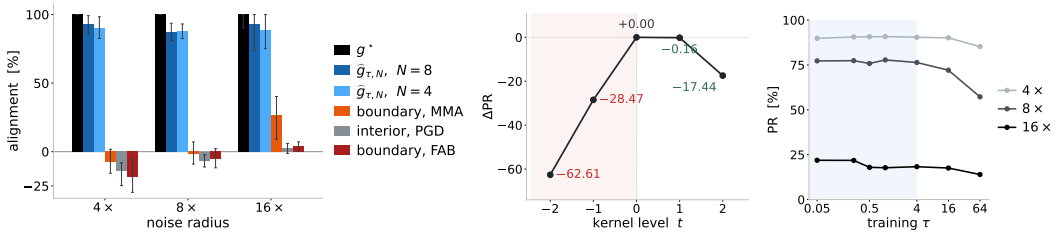
Figure 1: **Numerical assessment of the NPR first variation.** (a) The two-dimensional Palmer Penguins setting, showing the learned decision regions and an example perturbation region $\Delta(\boldsymbol{x})$. (b) Predicted versus actual NPR change under small parameter updates. The actual $\Delta\mathrm{NPR}$ is computed by centered finite differences, while the predicted $\Delta\mathrm{NPR}$ is obtained by numerically evaluating Eq. 15 and projecting the resulting gradient onto the same update direction. The full boundary prediction is compared with a nearest-boundary-point approximation (c.f. Appendix D.2). (c) Cosine alignment with the full boundary gradient for different single-point approximations and for progressively increasing fractions of $m^{-1}(0)$ starting from different boundary locations (c.f. Appendix D.3). (d) Local prediction of decision-boundary motion. We compare the observed normal displacement of boundary points with input-space flatness $1/\|\nabla_{\boldsymbol{\delta}} m\|_2$, parameter sensitivity $|D_{\boldsymbol{\theta}} m|$, and their combination given by the predicted normal velocity $|V|$. Spearman correlations quantify how well each quantity explains the observed motion (c.f. Appendix D.4).

High-dimensional setting. We use CIFAR-10 with a ResNet-18 and define $1\times = 8/255$. We evaluate the representativeness of different NPR directions and study how the resulting approximation depends on the target margin level and Gaussian-kernel bandwidth. Additional training experiments use the same estimator as a simple proof of concept. Full experimental protocols, hyperparameters, and evaluation details are provided in Appendix E.

6.2 Numerical Evaluation of the NPR First Variation

We first examine whether the boundary integral in Cor. 4.6 predicts the change in NPR along optimization directions encountered during training. For each recorded Adam update $\Delta\boldsymbol{\theta}_t$, we compute the centered finite-difference change in NPR and compare it with the first-order prediction obtained from the full decision-boundary integral. Fig. 1(b) shows that the full-boundary prediction closely follows the observed NPR change, whereas the nearest-boundary-point approximation is substantially more dispersed and can even predict the wrong sign. At $R = 4$, the nearest-point approximation gives the wrong sign in 28.8% of the evaluated cases. We further examine the representativeness of individual perturbations in Fig. 1(c) by comparing their parameter-space directions with the full boundary gradient and by progressively increasing the fraction of $m^{-1}(0)$ included in the integral. Although a single nearby boundary point can sometimes be strongly aligned with the aggregate direction, high-fidelity recovery can require contributions from a substantial portion of the boundary. These results support the interpretation of the NPR first variation as an aggregate boundary quantity whose representation by a small number of perturbations depends on the underlying geometry. Detailed procedures are provided in Appendix D.2 and Appendix D.3.

We next examine the local mechanism underlying this boundary variation. For each recorded parameter update, we measure the normal displacement of individual boundary elements and compare it with the input-space factor $1/\|\nabla_{\boldsymbol{\delta}} m\|_2$, the parameter sensitivity $|\langle \nabla_{\boldsymbol{\theta}} m, \Delta\boldsymbol{\theta}\rangle|$, and their combination $|\langle \nabla_{\boldsymbol{\theta}} m, \Delta\boldsymbol{\theta}\rangle|/\|\nabla_{\boldsymbol{\delta}} m\|_2$, corresponding to the predicted normal velocity. Fig. 1(d) shows that the normal-velocity prediction has the strongest correlation with the observed boundary motion, while the input-space factor alone is substantially weaker and more variable. This supports the view that boundary motion depends jointly on parameter-space sensitivity and input-space geometry, rather than on input-space flatness alone. Detailed procedures are provided in Appendix D.4.



(a) (b) (c)

Figure 2: **High-dimensional diagnostics and training ablations of NPR on CIFAR-10.** (a) Representativeness of candidate NPR directions across noise radii. The reference $g^\star$ is a high-sample Monte Carlo approximation of $g_\tau = \mathbb{E}_{\boldsymbol{\delta}\sim q}[K_\tau(m)\nabla_{\boldsymbol{\theta}} m]$ with $\tau = 0.25$, while $\widehat{g}_{\tau,N}$ uses only N perturbations per image. Alignment with NPR descent is evaluated by the kernel-free finite-difference change of the true NPR along each normalized direction and normalized so that $g^\star$ equals 100%. Error bars show 95% bootstrap intervals from 4000 resamples over images (c.f. Appendix E.1). (b) Effect of shifting the Gaussian kernel from the decision boundary to the level set $m = t$. Each value of t corresponds to a separately trained model, and $\Delta\mathrm{PR}$ denotes the change in PR relative to the corresponding $t = 0$ model (c.f. Appendix E.2). (c) Sensitivity to the Gaussian-kernel bandwidth used during training. Each value of τ corresponds to a separately trained model with all other settings fixed, and PR is evaluated under uniform noise at $4\times$, $8\times$, and $16\times$ (c.f. Appendix E.3).

6.3 NPR Descent Direction Estimation in High Dimensions

Figure 2 examines the NPR first variation in the high-dimensional ResNet-18/CIFAR-10 setting. Figure 2a directly revisits our central research question by asking whether a small number of perturbations can represent the aggregate NPR variation. Unlike the low-dimensional setting, where individual boundary points can sometimes be strongly representative, the single-point candidates on CIFAR-10 are substantially less aligned with NPR descent. In contrast, the high-sample reference $g^\star$ and its iid Monte Carlo approximations remain strongly aligned even with only $N = 4$ or 8 perturbations per image. The MMA anchor and FAB nearest-boundary point can be only weakly aligned or even point in the wrong direction. These results suggest that, in this setting, a few perturbations can represent NPR effectively when their contributions are aggregated, but no individual boundary point is reliably representative of the full boundary effect.

Figures 2b and 2c examine two additional properties of the boundary-localized construction. Figure 2b reveals a pronounced asymmetry across the decision boundary when PR is evaluated under $8\times$ uniform noise. Relative to the corresponding $t = 0$ model, shifting the kernel to $t = -1$ reduces PR by 28.47 points, whereas $t = +1$ reduces it by only 0.16 points. This asymmetry admits a natural geometric interpretation. For $t > 0$, the set $\{m \le t\}$ contains the adversarial region and also includes correctly classified points near the boundary, whereas $t < 0$ targets only a subset already inside the adversarial region. The observed behavior is therefore consistent with the different upper- and lower-level roles of the two sides of the boundary. Finally, Fig. 2c evaluates PR under uniform noise at $4\times$, $8\times$, and $16\times$ and shows relatively stable performance for $\tau \in [0.05, 4]$, with clearer degradation only for substantially broader kernels.

Table 1 provides a downstream examination of the boundary-based picture suggested by our first-order analysis. The compared methods obtain robustness signals from rather different regions of the perturbation space. PGD-AT relies on high-loss adversarial perturbations in the interior of the feasible region, while AT-PR remains adversarial-search based and selects, among multiple PGD-generated adversarial candidates, the one associated with a larger local violation region. In contrast, MMA explicitly searches for the shortest successful perturbation and therefore places its training signal close to the decision boundary. At the larger evaluation radius $R = 8\times$, MMA achieves 66.39%, 61.14%, 61.51%, and 51.77% PR under uniform, Laplace, Gaussian, and sign noise, respectively, and is substantially more competitive than several interior adversarial-search baselines. This behavior is consistent with the theoretical role of the decision boundary in the NPR first variation, although differences between training objectives prevent this comparison from isolating boundary location as

Table 1: **Probabilistic and adversarial robustness on CIFAR-10 with ResNet-18.** All models are trained for 100 epochs under the same common training protocol. NPR-Aug. (Ours) is an MMA-based proof-of-concept that adds the NPR regularization term when the perturbation region contains a decision boundary, with regularization weight $\lambda = 1/2$. The EVaR baseline is trained at the $1\times$ perturbation scale. TRADES uses $\beta = 6$ and a learning rate of 0.01, while the remaining common training settings are kept the same across methods. Full training details are provided in Appendix E. Clean and PGD-10 denote clean accuracy and ℓ_∞ adversarial accuracy at $\epsilon = 8/255$. PR denotes classification accuracy under stochastic perturbations. Unif., Lap., Gauss., and Sign denote uniform, truncated Laplace, truncated Gaussian, and sign noise, respectively. R denotes the ℓ_∞ noise radius, with $1\times = 8/255$. Representative perturbations at different radii are shown in Fig. 3. Time reports the average wall-clock seconds per training epoch. Best accuracy and PR, together with the lowest training time, are shown in bold.

				$R = 4\times$				$R = 8\times$			
Method	Clean	PGD-10	Time (s/ep.)	Unif.	Lap.	Gauss.	Sign	Unif.	Lap.	Gauss.	Sign
Standard	**95.17**	0.17	**12.3**	47.36	31.31	32.03	16.78	13.74	11.34	11.33	10.38
EVaR	94.76	6.71	160.8	77.29	60.97	61.97	30.75	23.11	15.56	15.86	11.88
PGD-AT	85.36	47.88	95.5	81.99	79.07	79.35	71.75	63.72	50.46	51.25	31.31
AT-PR	84.42	47.43	608.3	81.67	78.94	79.07	72.25	64.87	52.59	53.41	31.98
TRADES†	77.09	46.12	107.0	74.92	72.75	72.77	67.23	63.58	56.61	56.90	41.90
MART	75.39	49.84	106.7	72.29	70.09	70.30	64.73	61.08	52.91	53.62	36.80
MMA	85.40	**51.72**	131.0	76.78	74.36	74.54	70.33	66.39	61.14	61.51	51.77
NPR-Aug. (Ours)	88.93	45.55	165.6	**83.01**	**82.03**	**82.14**	**79.54**	**76.02**	**69.01**	**69.61**	**54.45**

the sole cause. The pseudocode and a more complete set of experimental details are provided in Appendix F.

The high-dimensional diagnostic in Fig. 2a further shows that being on the decision boundary is not by itself sufficient. Individual boundary points can still be poor representatives of the aggregate NPR variation. We therefore use NPR-Aug. as a simple proof of concept to test whether aggregating near-boundary information can improve a boundary-based training procedure. NPR-Aug. retains MMA and, whenever the local perturbation region contains a decision boundary, adds the Gaussian-kernel NPR term using $4\times$ uniform perturbations with weight $\lambda = 1/2$. Relative to MMA, NPR-Aug. improves PR across all four stochastic perturbations reported in the main table. At $R = 8\times$, PR increases from 66.39% to 76.02% under uniform noise, from 61.14% to 69.01% under Laplace noise, from 61.51% to 69.61% under Gaussian noise, and from 51.77% to 54.45% under sign noise. These results support a more specific interpretation of the theory. Boundary information is important, but when a single boundary point is not sufficiently representative, aggregating contributions from multiple near-boundary perturbations can provide a more effective PR optimization signal. The gain in PR comes with lower PGD-10 accuracy (45.55% versus 51.72% for MMA) and higher training cost (165.6 versus 131.0 seconds per epoch), indicating that improved probabilistic robustness need not coincide with improved worst-case robustness.

7 Conclusion

In this paper, we investigated how probabilistic robustness changes under small model updates and when a small number of perturbations can represent this change. By applying the co-area formula, we showed that the first-order variation of NPR is characterized by a distribution-weighted integral over the decision boundary. This reveals that PR optimization is boundary-supported but generally aggregate, so the usefulness of individual perturbations depends on how well they represent the full boundary contribution. The same analysis also shows that local boundary motion is determined jointly by input-space geometry and parameter-space sensitivity.

Our experiments support this geometric picture in both low- and high-dimensional settings. While representative boundary points can emerge under favorable geometry, aggregation becomes substantially more important on CIFAR-10. A Gaussian-kernel Monte Carlo approximation makes this first variation accessible in high dimensions, and a simple MMA-based proof of concept further shows that aggregating near-boundary information can improve probabilistic robustness.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, figure preparation, literature search, and checking the presentation and consistency of mathematical proofs.

We did not use generative AI tools to originate the research ideas or methodology, independently derive or prove mathematical claims, execute experiments, generate synthetic datasets, or make final scientific interpretations or conclusions.

All technical arguments, mathematical derivations, experimental designs, implementations, analyses, and conclusions were developed and verified by the authors. All AI-assisted material was reviewed and, where appropriate, substantially revised by the authors, and references identified with AI assistance were checked against the original sources. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed assumptions and complete proofs of the theoretical results in the Appendix. The experimental setup, data preprocessing, model architectures, training hyperparameters, numerical procedures, and evaluation protocols for both the low- and high-dimensional experiments are also documented in the Appendix. Pseudocode for NPR-Aug. is provided together with its implementation details. We will release the code used to reproduce the experiments and figures.

REFERENCES

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International conference on machine learning*, pp. 2196–2205. PMLR, 2020.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020.

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, 1992.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. *Advances in neural information processing systems*, 29, 2016.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3762–3770. IEEE, 2018.

Herbert Federer. *Geometric Measure Theory*. Springer, New York, 1969.

Ian J Goodfellow. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations ICLR 2015*. International Conference on Learning Representations, 2015.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, and et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020. ISSN 1574-0137.

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, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9070–9078. IEEE, 2019.

Liviu I. Nicolaescu. The co-area formula, 2021. Notes for the "Blue Collar Seminar on Geometric Integration Theory", last revised January 22, 2021.

Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: a geometric framework for black-box adversarial attacks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8443–8452. IEEE, 2020.

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *In Proc. of 2nd Int. Conf. on Learning Representations*, 2014.

Karim TIT, Teddy Furon, and Mathias Rousset. Gradient-informed neural network statistical robustness estimation. In *Proc. of The 26th Int. Conf. on Artificial Intelligence and Statistics*, volume 206, pp. 323–334. PMLR, 2023.

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

Zheng Wang, Yi Zhang, Siddartha Khastgir, Carsten Maple, and Xingyu Zhao. Non-parametric probabilistic robustness: A conservative risk estimator under unknown perturbation distributions. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Stefan Webb, Tom Rainforth, Yee Whye Teh, and M Pawan Kumar. A statistical approach to assessing neural network robustness. In *Int. Conf. on Learning Representations*, 2019.

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 *Int. Conf. on Machine Learning*, pp. 6727–6736. PMLR, 2019.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Tianle Zhang, Wenjie Ruan, and Jonathan E. Fieldsend. Proa: A probabilistic robustness assessment against functional perturbations. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III*, pp. 154–170. Springer-Verlag, 2023. ISBN 978-3-031-26408-5.

Tianle Zhang, Yanghao Zhang, Ronghui Mu, Jiaxu Liu, Jonathan Fieldsend, and Wenjie Ruan. Prass: probabilistic risk-averse robust learning with stochastic search. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 559–567, 2024a.

Yi Zhang, Yun Tang, Wenjie Ruan, Xiaowei Huang, Siddartha Khastgir, Paul Jennings, and Xingyu Zhao. Protip: Probabilistic robustness verification on text-to-image diffusion models against stochastic perturbation. In *European Conference on Computer Vision*, 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, 2025a.

Yi Zhang, Zheng Wang, Zhen Chen, Wenjie Ruan, Qing Guo, Siddartha Khastgir, Carsten Maple, and Xingyu Zhao. Probabilistic robustness for free? revisiting training via a benchmark. *arXiv preprint arXiv:2511.01724*, 2025b.

Xingyu Zhao. Probabilistic robustness in deep learning: A concise yet comprehensive guide. *Adversarial Example Detection and Mitigation Using Machine Learning*, pp. 1–13, 2025.

A Omitted Proofs and Additional Discussions

A.1 Proof of Prop. 4.1

To prove Prop. 4.1, we first restate the co-area formula in an equivalent weighted form that is more convenient for our analysis.

Corollary A.1. *Let $F : \Omega \to \mathbb{R}^k$ be the Lipschitz map in Thm. 3.6, and let $q : \Omega \to [0, \infty]$ be measurable. Assume that the rank-deficient set of F carries zero q-mass,*

$$\int_{\{\boldsymbol{\delta} \in \Omega : J_F(\boldsymbol{\delta}) = 0\}} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = 0. \tag{23}$$

Then,

$$\int_{\Omega} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \int_{\mathbb{R}^k} \left(\int_{\Omega \cap F^{-1}(\boldsymbol{z})} \frac{q(\boldsymbol{\delta})}{J_F(\boldsymbol{\delta})} d\mathcal{H}^{d-k}(\boldsymbol{\delta}) \right) d\boldsymbol{z}. \tag{24}$$

Proof. Let

$$N \triangleq \{\boldsymbol{\delta} \in \Omega \mid J_F(\boldsymbol{\delta}) = 0\}. \tag{25}$$

By assumption,

$$\int_N q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = 0. \tag{26}$$

Define

$$\varphi(\boldsymbol{\delta}) \triangleq \begin{cases} \dfrac{q(\boldsymbol{\delta})}{J_F(\boldsymbol{\delta})}, & \boldsymbol{\delta} \in \Omega \setminus N, \\ 0, & \boldsymbol{\delta} \in N. \end{cases} \tag{27}$$

Then φ is nonnegative and measurable. Applying Theorem 3.6 yields

$$\int_{\Omega} \varphi(\boldsymbol{\delta}) J_F(\boldsymbol{\delta}) d\boldsymbol{\delta} = \int_{\mathbb{R}^k} \left(\int_{\Omega \cap F^{-1}(\boldsymbol{z})} \varphi(\boldsymbol{\delta}) d\mathcal{H}^{d-k}(\boldsymbol{\delta}) \right) d\boldsymbol{z}. \tag{28}$$

Since

$$\varphi(\boldsymbol{\delta}) J_F(\boldsymbol{\delta}) = q(\boldsymbol{\delta}) \tag{29}$$

on $\Omega \setminus N$, and $q = 0$ for Lebesgue-a.e. $\boldsymbol{\delta} \in N$ by the assumption. Therefore,

$$\int_{\Omega} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \int_{\mathbb{R}^k} \left(\int_{\Omega \cap F^{-1}(\boldsymbol{z})} \frac{q(\boldsymbol{\delta})}{J_F(\boldsymbol{\delta})} d\mathcal{H}^{d-k}(\boldsymbol{\delta}) \right) d\boldsymbol{z}, \tag{30}$$

where q/J_F is defined to be zero on N. □

Remark A.2. Cor. A.1 assumes that the rank-deficient set $\{\boldsymbol{\delta} : J_F(\boldsymbol{\delta}) = 0\}$ carries zero q-mass. If this assumption is relaxed, the regular co-area term remains valid, but an additional contribution supported on the rank-deficient set must be included.

We next remove the zero-q-mass assumption and derive the corresponding residual and rank-stratified decompositions.

Theorem A.3 (Weighted co-area decomposition with rank-deficient residual)**.** *Let $F : \Omega \to \mathbb{R}^k$ be the Lipschitz map in Thm. 3.6, and let $q : \Omega \to [0, \infty]$ be measurable. Define the rank-deficient set*

$$N \triangleq \{\boldsymbol{\delta} \in \Omega \mid J_F(\boldsymbol{\delta}) = 0\}. \tag{31}$$

Then

$$\int_{\Omega} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \int_N q(\boldsymbol{\delta})\, d\boldsymbol{\delta} + \int_{\mathbb{R}^k} \left(\int_{\Omega \cap F^{-1}(\boldsymbol{z})} \frac{q(\boldsymbol{\delta}) \mathbf{1}_{\{J_F > 0\}}(\boldsymbol{\delta})}{J_F(\boldsymbol{\delta})}\, d\mathcal{H}^{d-k}(\boldsymbol{\delta}) \right) d\boldsymbol{z}. \tag{32}$$

Proof. Define

$$\varphi(\boldsymbol{\delta}) \triangleq \begin{cases} \dfrac{q(\boldsymbol{\delta})}{J_F(\boldsymbol{\delta})}, & \boldsymbol{\delta} \in \Omega \setminus N, \\ 0, & \boldsymbol{\delta} \in N. \end{cases} \tag{33}$$

Applying Thm. 3.6 gives

$$\int_{\Omega \setminus N} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \int_{\mathbb{R}^k} \left(\int_{\Omega \cap F^{-1}(\boldsymbol{z})} \frac{q(\boldsymbol{\delta}) \mathbf{1}_{\{J_F > 0\}}(\boldsymbol{\delta})}{J_F(\boldsymbol{\delta})}\, d\mathcal{H}^{d-k}(\boldsymbol{\delta}) \right) d\boldsymbol{z}. \tag{34}$$

Since

$$\int_{\Omega} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \int_{N} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} + \int_{\Omega \setminus N} q(\boldsymbol{\delta})\, d\boldsymbol{\delta}, \tag{35}$$

the result follows. □

Theorem A.4 (Rank-stratified co-area decomposition)**.** *Let*

$$F = (F_1, \ldots, F_k) \mid \Omega \to \mathbb{R}^k \tag{36}$$

be the Lipschitz map in Thm. 3.6, and let $D_F \subseteq \Omega$ denote the set where F is differentiable. For $r = 0, \ldots, k$, define the rank-r stratum

$$S_r \triangleq \{\boldsymbol{\delta} \in D_F \mid \operatorname{rank} DF(\boldsymbol{\delta}) = r\}. \tag{37}$$

For $1 \le r \le k$, let

$$\mathfrak{I}_r \triangleq \{I \subseteq \{1, \ldots, k\} \mid |I| = r\}. \tag{38}$$

For each $I = \{i_1 < \cdots < i_r\} \in \mathfrak{I}_r$, define the component submap

$$F_I \triangleq (F_{i_1}, \ldots, F_{i_r}) \mid \Omega \to \mathbb{R}^r. \tag{39}$$

Fix any total order $\prec$ on $\mathfrak{I}_r$, and define

$$S_{r,I} \triangleq \left\{\boldsymbol{\delta} \in S_r \mid J_{F_I}(\boldsymbol{\delta}) > 0,\ J_{F_{I'}}(\boldsymbol{\delta}) = 0 \text{ for all } I' \prec I\right\}. \tag{40}$$

Then, for every $r \ge 1$,

$$S_r = \bigsqcup_{I \in \mathfrak{I}_r} S_{r,I}. \tag{41}$$

Moreover, for every measurable $q : \Omega \to [0, \infty]$,

$$\int_{\Omega} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \int_{S_0} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} + \sum_{r=1}^{k} \sum_{I \in \mathfrak{I}_r} \int_{\mathbb{R}^r} \left(\int_{S_{r,I} \cap F_I^{-1}(\boldsymbol{u})} \frac{q(\boldsymbol{\delta})}{J_{F_I}(\boldsymbol{\delta})}\, d\mathcal{H}^{d-r}(\boldsymbol{\delta}) \right) d\boldsymbol{u}. \tag{42}$$

Proof. By Rademacher's theorem, $\Omega \setminus D_F$ has Lebesgue measure zero. Hence

$$\int_{\Omega} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \sum_{r=0}^{k} \int_{S_r} q(\boldsymbol{\delta})\, d\boldsymbol{\delta}. \tag{43}$$

Fix $r \ge 1$. For every $\boldsymbol{\delta} \in S_r$, the matrix $DF(\boldsymbol{\delta})$ has rank r, so it contains r linearly independent rows. Therefore, there exists at least one $I \in \mathfrak{I}_r$ such that

$$J_{F_I}(\boldsymbol{\delta}) > 0. \tag{44}$$

The total order $\prec$ assigns each point in S_r to the first such index set, yielding the disjoint partition

$$S_r = \bigsqcup_{I \in \mathfrak{I}_r} S_{r,I}. \tag{45}$$

For fixed $I \in \mathfrak{I}_r$, define

$$\varphi_{r,I}(\boldsymbol{\delta}) \triangleq \begin{cases} \dfrac{q(\boldsymbol{\delta})}{J_{F_I}(\boldsymbol{\delta})}, & \boldsymbol{\delta} \in S_{r,I}, \\ 0, & \boldsymbol{\delta} \notin S_{r,I}. \end{cases} \tag{46}$$

Since $J_{F_I} > 0$ on $S_{r,I}$, applying Thm. 3.6 to $F_I \mid \Omega \to \mathbb{R}^r$ gives

$$\int_{S_{r,I}} q(\boldsymbol{\delta})\, d\boldsymbol{\delta} = \int_{\mathbb{R}^r} \left(\int_{S_{r,I} \cap F_I^{-1}(\boldsymbol{u})} \frac{q(\boldsymbol{\delta})}{J_{F_I}(\boldsymbol{\delta})}\, d\mathcal{H}^{d-r}(\boldsymbol{\delta}) \right) d\boldsymbol{u}. \tag{47}$$

Summing over $I \in \mathfrak{I}_r$ and $r = 1, \ldots, k$, and adding the rank-zero contribution, yields Eq. 42. □

A.2 Proofs of Thm. 4.5

Proof. Fix an arbitrary direction $\boldsymbol{v} \in \mathbb{R}^{\dim(\boldsymbol{\theta})}$ and define

$$\boldsymbol{\theta}_s \triangleq \boldsymbol{\theta} + s\boldsymbol{v}. \tag{48}$$

Let

$$E_s \triangleq \left\{ \boldsymbol{\delta} \in \Delta(\boldsymbol{x}) \mid m^{\min}_{\boldsymbol{\theta}_s,\boldsymbol{x},y}(\boldsymbol{\delta}) \le 0 \right\}. \tag{49}$$

Then

$$\mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}_s) = \int_{E_s} q(\boldsymbol{\delta} \mid \boldsymbol{x})\, d\boldsymbol{\delta}. \tag{50}$$

By Assumption 4.4, the moving part of ∂E_0 is, up to an $\mathcal{H}^{d-1}$-null set, the union of the class-wise decision faces

$$\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x}), \qquad j \neq y. \tag{51}$$

Intersections between distinct class-wise faces, and between these faces and $\partial\Delta(\boldsymbol{x})$, are lower-dimensional and hence negligible under $\mathcal{H}^{d-1}$. The remaining portion of ∂E_0 lies on the fixed boundary $\partial\Delta(\boldsymbol{x})$ and has zero normal velocity.

Consider a smooth boundary trajectory $\boldsymbol{\delta}_s$ on the j-th class-wise face, so that

$$m^{(j)}_{\boldsymbol{\theta}_s,\boldsymbol{x},y}(\boldsymbol{\delta}_s) = 0. \tag{52}$$

Differentiating at $s = 0$ gives

$$\left\langle \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}), \boldsymbol{v} \right\rangle + \left\langle \nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}), \dot{\boldsymbol{\delta}}_0 \right\rangle = 0. \tag{53}$$

On this face, the outward unit normal of the adversarial region is

$$\boldsymbol{n}_j(\boldsymbol{\delta}) = \frac{\nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})}{\left\| \nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \right\|_2}. \tag{54}$$

Hence, the corresponding normal velocity is

$$V_j(\boldsymbol{\delta}; \boldsymbol{v}) \triangleq \left\langle \dot{\boldsymbol{\delta}}_0, \boldsymbol{n}_j(\boldsymbol{\delta}) \right\rangle = -\frac{\left\langle \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}), \boldsymbol{v} \right\rangle}{\left\| \nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \right\|_2}. \tag{55}$$

Applying the standard first-variation formula for a moving domain, while noting that the fixed portion on $\partial\Delta(\boldsymbol{x})$ contributes zero, yields

$$D_{\boldsymbol{v}} \mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) = \sum_{j \neq y} \int_{\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})} q(\boldsymbol{\delta} \mid \boldsymbol{x}) V_j(\boldsymbol{\delta}; \boldsymbol{v})\, d\mathcal{H}^{d-1}(\boldsymbol{\delta}) \tag{56}$$

$$= -\sum_{j \neq y} \int_{\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\left\| \nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \right\|_2} \left\langle \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}), \boldsymbol{v} \right\rangle d\mathcal{H}^{d-1}(\boldsymbol{\delta}). \tag{57}$$

Since $\boldsymbol{v}$ is arbitrary, the directional derivative is represented by the linear functional

$$\nabla_{\boldsymbol{\theta}} \mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\sum_{j \neq y} \int_{\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\left\| \nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \right\|_2} \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\, d\mathcal{H}^{d-1}(\boldsymbol{\delta}), \tag{58}$$

which is Eq. 13. Finally, under the stated integrable domination condition, differentiation may be interchanged with expectation, giving

$$\nabla_{\boldsymbol{\theta}} \mathrm{NPR}_q(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim P} \left[\nabla_{\boldsymbol{\theta}} \mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) \right], \tag{59}$$

which yields Eq. 14. □

A.3 Proof of Cor. 4.6

Proof. Recall that

$$m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) = \min_{j\neq y} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}). \tag{60}$$

By Assumption 4.4, the decision boundary $\partial\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})$ is, up to an $\mathcal{H}^{d-1}$-null set, the union of the class-wise decision faces

$$\partial_j\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x}), \qquad j \neq y. \tag{61}$$

Indeed, intersections between distinct class-wise faces are lower-dimensional and therefore have zero $\mathcal{H}^{d-1}$ measure.

Away from these intersections, each boundary point belongs locally to a unique active class-wise face. Hence, for $\mathcal{H}^{d-1}$-almost every $\boldsymbol{\delta} \in \partial_j\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})$, the minimum margin locally coincides with the active class-wise margin, so that

$$\nabla_{\boldsymbol{\delta}} m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) = \nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}), \qquad \nabla_{\boldsymbol{\theta}} m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) = \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}). \tag{62}$$

Substituting these identities into Thm. 4.5 and combining the class-wise integrals gives

$$\nabla_{\boldsymbol{\theta}}\mathrm{NPR}_q(\boldsymbol{x},y,\boldsymbol{\theta}) = -\int_{\partial\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y}\cap\Delta(\boldsymbol{x})} \frac{q(\boldsymbol{\delta}\mid\boldsymbol{x})}{\left\|\nabla_{\boldsymbol{\delta}} m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\right\|_2} \nabla_{\boldsymbol{\theta}} m^{\min}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\, d\mathcal{H}^{d-1}(\boldsymbol{\delta}), \tag{63}$$

which is Eq. 15. □

B Extension to Piecewise-Smooth Classifiers

The smoothness assumption in the main text is adopted to present the first-variation formula without introducing additional stratification notation. We now show that the same characterization extends to piecewise-smooth classifiers, including ReLU networks, under mild regularity conditions on the interaction between decision faces and nonsmooth strata. Fix an input–label pair $(\boldsymbol{x}, y)$ and define

$$\widetilde{h}(\boldsymbol{\vartheta},\boldsymbol{\delta}) \triangleq h_{\boldsymbol{\vartheta}}(\boldsymbol{x}+\boldsymbol{\delta}). \tag{64}$$

Let $\mathcal{U}$ be a neighborhood of $\{\boldsymbol{\theta}\} \times \Delta(\boldsymbol{x})$ in the joint parameter–perturbation space.

Assumption B.1 (Piecewise-smooth stratification)**.** There exists a locally finite smooth stratification $\{\mathcal{S}_r\}_{r\in\mathcal{R}}$ of $\mathcal{U}$ such that $\widetilde{h}$ is continuously differentiable on every top-dimensional stratum. Let

$$\mathcal{N} \triangleq \mathcal{U} \setminus \bigcup_{r\in\mathcal{R}_{\mathrm{top}}} \mathcal{S}_r \tag{65}$$

denote the union of the lower-dimensional strata, and define its slice at $\boldsymbol{\theta}$ by

$$\mathcal{N}_{\boldsymbol{\theta}} \triangleq \{\boldsymbol{\delta} \mid (\boldsymbol{\theta},\boldsymbol{\delta}) \in \mathcal{N}\}. \tag{66}$$

For every $j \neq y$, assume that

$$\mathcal{H}^{d-1}\left(\partial_j\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \mathcal{N}_{\boldsymbol{\theta}}\right) = 0, \tag{67}$$

and that

$$\nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \neq \mathbf{0} \tag{68}$$

for every

$$\boldsymbol{\delta} \in \partial_j\mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \setminus \mathcal{N}_{\boldsymbol{\theta}}. \tag{69}$$

We further assume that distinct smooth class-wise decision faces intersect transversally with one another and with the smooth strata of $\mathcal{N}_{\boldsymbol{\theta}}$ and $\partial\Delta(\boldsymbol{x})$ whenever such intersections occur.

Intuitively, Assumption B.1 separates each class-wise decision face into smooth pieces and a lower-dimensional nonsmooth remainder. The smooth pieces admit the same local differential analysis as in the main text, while the condition

$$\mathcal{H}^{d-1}\left(\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \mathcal{N}_{\boldsymbol{\theta}}\right) = 0 \tag{70}$$

ensures that the nonsmooth part does not contribute to the boundary integral. The transversality condition further guarantees that intersections between different pieces remain lower-dimensional. Hence, the smooth first-variation formula can be applied on each stratum and then summed over the resulting boundary pieces.

The argument is the same as in the smooth case, except that the first-variation formula is applied separately on each smooth stratum. The nonsmooth interfaces are lower-dimensional and therefore do not contribute to the boundary integral.

Theorem B.2 (First Variation of NPR for Piecewise-Smooth Classifiers)**.** *Suppose Assumption B.1 holds and $q(\boldsymbol{\delta} \mid \boldsymbol{x})$ is continuous in a neighborhood of the regular parts of the class-wise decision faces. Then, for any parameter direction $\boldsymbol{v}$,*

$$D_{\boldsymbol{v}}\mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\sum_{j \neq y} \int_{\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\left\|\nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\right\|_2} \left\langle \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}), \boldsymbol{v} \right\rangle \, d\mathcal{H}^{d-1}(\boldsymbol{\delta}), \tag{71}$$

where the derivatives are understood piecewise and $\mathcal{H}^{d-1}$-almost everywhere on the decision faces.

If, in addition, the stratum-wise first-order expansion is uniform in the parameter direction, then the local NPR is differentiable at $\boldsymbol{\theta}$ and

$$\nabla_{\boldsymbol{\theta}}\mathrm{NPR}_q(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\sum_{j \neq y} \int_{\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\left\|\nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\right\|_2} \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}) \, d\mathcal{H}^{d-1}(\boldsymbol{\delta}). \tag{72}$$

Proof. For each $j \neq y$, define the regular part of the class-wise decision face by

$$\Gamma^{(j)}_{\boldsymbol{\theta},\mathrm{reg}} \triangleq \left(\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})\right) \setminus \mathcal{N}_{\boldsymbol{\theta}}. \tag{73}$$

By Assumption B.1, $\Gamma^{(j)}_{\boldsymbol{\theta},\mathrm{reg}}$ is the locally finite union of smooth pieces lying in top-dimensional strata of $\widetilde{h}$.

On each such piece, $m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}$ is continuously differentiable and $\nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y} \neq \mathbf{0}$. Hence, the same implicit-function argument as in the smooth case applies. For the parameter path $\boldsymbol{\theta}(t) = \boldsymbol{\theta} + t\boldsymbol{v}$ and a boundary trajectory $\boldsymbol{\delta}(t)$ satisfying

$$m^{(j)}_{\boldsymbol{\theta}(t),\boldsymbol{x},y}(\boldsymbol{\delta}(t)) = 0, \tag{74}$$

differentiation at $t = 0$ yields the normal velocity

$$V^{(j)}_{\boldsymbol{v}}(\boldsymbol{\delta}) = -\frac{\left\langle \nabla_{\boldsymbol{\theta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta}), \boldsymbol{v} \right\rangle}{\left\|\nabla_{\boldsymbol{\delta}} m^{(j)}_{\boldsymbol{\theta},\boldsymbol{x},y}(\boldsymbol{\delta})\right\|_2}. \tag{75}$$

Applying the smooth moving-domain first-variation formula on each smooth piece and summing over the strata gives

$$D_{\boldsymbol{v}}\mathrm{NPR}_q = \sum_{j \neq y} \int_{\Gamma^{(j)}_{\boldsymbol{\theta},\mathrm{reg}}} q(\boldsymbol{\delta} \mid \boldsymbol{x}) V^{(j)}_{\boldsymbol{v}}(\boldsymbol{\delta}) \, d\mathcal{H}^{d-1}(\boldsymbol{\delta}). \tag{76}$$

The portion of $\partial\Delta(\boldsymbol{x})$ is fixed with respect to the model parameters and therefore has zero normal velocity.

By Assumption B.1,

$$\mathcal{H}^{d-1}\left[\left(\partial_j \mathcal{A}_{\boldsymbol{\theta},\boldsymbol{x},y} \cap \Delta(\boldsymbol{x})\right) \cap \mathcal{N}_{\boldsymbol{\theta}}\right] = 0. \tag{77}$$

The nonsmooth portions therefore contribute no $(d-1)$-dimensional boundary mass. Transversality likewise makes intersections between distinct class-wise faces and stratum interfaces $\mathcal{H}^{d-1}$-null. Thus the smooth-piece integrals recover the integral over the full class-wise decision faces, yielding Eq. 71.

Under the additional uniform first-order condition, the resulting directional derivative defines the differential at $\boldsymbol{\theta}$, which gives Eq. 72. □

C Connections to Smooth and Risk-Based Surrogates

Probit interpretation. The Gaussian-kernel construction in Sec. 5 admits an equivalent smooth-loss interpretation. Define

$$L_\tau(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\delta}\sim q(\cdot|\boldsymbol{x})}\left[\Phi\left(-\frac{m(\boldsymbol{\delta})}{\tau}\right)\right], \tag{78}$$

where Φ is the standard normal cumulative distribution function. Since

$$\frac{d}{dm}\Phi\left(-\frac{m}{\tau}\right) = -K_\tau(m), \tag{79}$$

we have

$$\nabla_{\boldsymbol{\theta}} L_\tau(\boldsymbol{\theta}) = -g_\tau. \tag{80}$$

Moreover, $\Phi(-m/\tau) \to \mathbf{1}_{\{m<0\}}$ as $\tau \to 0$. Thus, the Gaussian-kernel estimator can equivalently be viewed as a probit smoothing of the NPR indicator.

Relation to common robustness surrogates. The first-order characterization also provides a geometric distinction between NPR and commonly used continuous surrogates. NPR concentrates its first-order variation on the zero-margin boundary, whereas cross-entropy assigns weight over a broader range of margins, tail-risk objectives such as CVaR and EVaR emphasize high-loss regions, and worst-case adversarial training concentrates on loss maximizers. Thus, these objectives differ in which regions of the perturbation landscape contribute to their parameter updates.

Binary violations. The above distinction concerns risk measures applied to continuous margin- or loss-based random variables. If CVaR or EVaR is instead applied directly to

$$V_{\boldsymbol{\theta}}(\boldsymbol{\delta}) = \mathbf{1}_{\{m(\boldsymbol{\delta})<0\}}, \tag{81}$$

the resulting objective is a function of the Bernoulli parameter NPR_q. In its non-saturated regime, its gradient is therefore collinear with $\nabla_{\boldsymbol{\theta}}\mathrm{NPR}_q$. This does not remove the non-differentiability of the underlying indicator, which motivates the smooth boundary-localized construction used in the main text.

D Low-Dimensional Experiment Setup

Dataset and split. We use the Palmer Archipelago Penguins dataset, which contains 344 penguins from three species. We retain bill length and bill depth as the two input features, both measured in millimeters. After removing two samples with missing values, 342 examples remain. The data are split into stratified 70% training and 30% test sets using a fixed NumPy random seed of 0, and the same split is used for all models. Unless otherwise stated, we consider the ℓ_∞ perturbation region

$$\Delta_R(\boldsymbol{x}) = \left\{\boldsymbol{\delta} \in \mathbb{R}^2 \mid \|\boldsymbol{\delta}\|_\infty \le R\right\}, \tag{82}$$

with $R = 4$, and take $q(\boldsymbol{\delta} \mid \boldsymbol{x})$ to be uniform over $\Delta_R(\boldsymbol{x})$. Thus, in the original input coordinates, the corresponding perturbed inputs lie in the rectangle $\boldsymbol{x} + \Delta_R(\boldsymbol{x})$.

The two-dimensional input space is particularly convenient for our validation experiments. The perturbation region and the decision boundary can be directly visualized, while the zero-margin curves can be numerically traced before and after a parameter update. This allows us to measure the motion of the decision boundary and its induced change in NPR directly.

Model and preprocessing. We use a two-hidden-layer multilayer perceptron with architecture

$$2 \to W \to W \to 3, \qquad W \in \{64, 128\}, \tag{83}$$

and GELU activations. The smooth activation ensures that the network logits and class-wise margins are smooth functions of the input. Under the regularity conditions considered in the main text, their zero-margin sets are therefore locally smooth curves away from multiclass intersections.

The two input features are standardized using the training-set mean and standard deviation before being passed to the network. These statistics remain fixed throughout training. The perturbation regions, boundary locations, and all geometric distances are nevertheless defined in the original millimeter coordinates.

Training. Each model is trained using full-batch cross-entropy and Adam for 3000 optimization steps. The learning rate follows cosine decay from 10^{-2} to 10^{-4}, with weight decay 10^{-3}. We train both widths $W \in \{64, 128\}$ using random seeds $\{0, 1, 2\}$, giving six model configurations in total. No mini-batching or data augmentation is used.

Recorded parameter updates. To evaluate the first-order characterization along directions that arise naturally during optimization, we record the model parameters $\boldsymbol{\theta}_t$ immediately before an optimizer update together with the realized parameter change

$$\Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \tag{84}$$

at training steps

$$t \in \{100, 300, 1000, 2000, 2999\}. \tag{85}$$

The first-variation predictions are therefore evaluated along actual optimization directions rather than artificially chosen parameter directions.

Numerical precision. All training and geometric calculations are performed in `float64`. This is important because several validation experiments involve NPR differences on the order of 10^{-7}.

D.1 (A) Two-dimensional Data and Decision Geometry

Panel (a) visualizes the final $W = 64$, seed-0 model. We evaluate the classifier on a dense grid in the two-dimensional input space and assign each grid point to $\arg\max_c h_{\boldsymbol{\theta},c}(\boldsymbol{x})$. The resulting prediction regions are shown by the background shading, while decision boundaries are obtained by extracting the zero contours of the corresponding active class-wise margins.

The black star marks a representative test point $\boldsymbol{x}$, and the black square shows the corresponding perturbed-input region $\boldsymbol{x} + \Delta_4(\boldsymbol{x})$. This visualization makes both the zero-margin geometry and its motion under parameter updates directly observable.

D.2 (B) Predicting NPR Changes along Optimization Directions

We next compare the first-order boundary prediction with the NPR change induced by optimization directions encountered during training. We use all six model configurations, $W \in \{64, 128\}$ with seeds $\{0, 1, 2\}$, the perturbation radius $R = 4$, and the recorded Adam updates $\Delta\boldsymbol{\theta}_t$ at $t \in \{100, 300, 1000, 2000, 2999\}$.

For each test point, we estimate the realized first-order change using the centered finite difference

$$\Delta\mathrm{NPR}_{\mathrm{FD}} = \frac{\mathrm{NPR}(\boldsymbol{\theta}_t + s\Delta\boldsymbol{\theta}_t) - \mathrm{NPR}(\boldsymbol{\theta}_t - s\Delta\boldsymbol{\theta}_t)}{2}, \qquad s = \frac{1}{16}. \tag{86}$$

Each NPR value is computed deterministically on an 801×801 grid over $\Delta_4(\boldsymbol{x})$, using linear interpolation of margin-zero crossings and trapezoidal integration of the misclassified area.

The full-boundary first-order prediction is

$$\Delta\mathrm{NPR}_{\mathrm{whole}} = s\,\langle \nabla_{\boldsymbol{\theta}}\mathrm{NPR}_q, \Delta\boldsymbol{\theta}_t \rangle\,, \tag{87}$$

where the NPR descent direction is evaluated numerically by extracting the zero-margin contours and integrating

$$-\frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\|\nabla_{\boldsymbol{\delta}} m(\boldsymbol{\delta})\|_2} \nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta}) \tag{88}$$

along their arc length.

As a single-point comparison, we locate the decision-boundary point $\boldsymbol{\delta}^\star$ nearest to the unperturbed input and approximate the entire boundary contribution using its local value,

$$g_{\mathrm{near}} = -L_q \frac{\nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta}^\star)}{\|\nabla_{\boldsymbol{\delta}} m(\boldsymbol{\delta}^\star)\|_2}, \tag{89}$$

where

$$L_q \triangleq \int_{m^{-1}(0)} q(\boldsymbol{\delta} \mid \boldsymbol{x})\, d\mathcal{H}^1(\boldsymbol{\delta}) \tag{90}$$

is the q-weighted boundary length. The corresponding prediction is $s\langle g_{\mathrm{near}}, \Delta\boldsymbol{\theta}_t \rangle$.

Panel (b) compares the predicted and centered finite-difference changes on symmetric logarithmic axes, with the diagonal indicating exact agreement.

D.3 (c) Single-point Representativeness and Boundary Coverage

We next examine when a single perturbation, or a subset of the decision boundary, can represent the full NPR first variation. Again using all six model configurations at $R = 4$, we first compute the full boundary direction

$$g = -\int_{m^{-1}(0)} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\|\nabla_{\boldsymbol{\delta}} m(\boldsymbol{\delta})\|_2} \nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta})\, d\mathcal{H}^1(\boldsymbol{\delta}). \tag{91}$$

The left part of the panel compares four single-point candidates. These are the nearest boundary point, a randomly sampled boundary point, a random interior point at least 0.5 mm from the decision boundary, and the interior point with the largest cross-entropy gradient norm. For each candidate $\boldsymbol{\delta}$, we use the parameter-space direction

$$c(\boldsymbol{\delta}) = -\nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta}) \tag{92}$$

and report $\cos(c(\boldsymbol{\delta}), g)$. For a boundary point, the omitted factor $q(\boldsymbol{\delta} \mid \boldsymbol{x})/\|\nabla_{\boldsymbol{\delta}} m(\boldsymbol{\delta})\|_2$ is positive and therefore does not affect its cosine direction. The interior candidates are included as controls and need not correspond to local terms in the exact boundary integral. Results are summarized across examples by the median and interquartile range.

The right part progressively replaces a single boundary point by an increasing subset of the boundary. Starting either from the nearest boundary point or from a randomly selected boundary point, boundary elements are included according to their distance along the boundary. For a partial boundary $m^{-1}(0)_r$, we compute

$$g_r = -\int_{m^{-1}(0)_r} \frac{q(\boldsymbol{\delta} \mid \boldsymbol{x})}{\|\nabla_{\boldsymbol{\delta}} m(\boldsymbol{\delta})\|_2} \nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta})\, d\mathcal{H}^1(\boldsymbol{\delta}), \tag{93}$$

and report $\cos(g_r, g)$ against the fraction of the total boundary arc length included. The solid and dashed curves correspond respectively to expansion from the nearest and random boundary points, while the shaded regions show the interquartile range across test points and model configurations.

D.4 (d) Local Predictors of Decision-boundary Motion

Finally, we examine the local normal motion underlying the NPR first variation. The analysis uses all six model configurations, the five recorded optimization steps, $R = 4$, and all test examples satisfying $0 < \mathrm{NPR}_q < 1$.

For each element of the decision boundary before an update, we identify its active class-wise margin and follow the old outward normal direction. Under the updated model, a one-dimensional

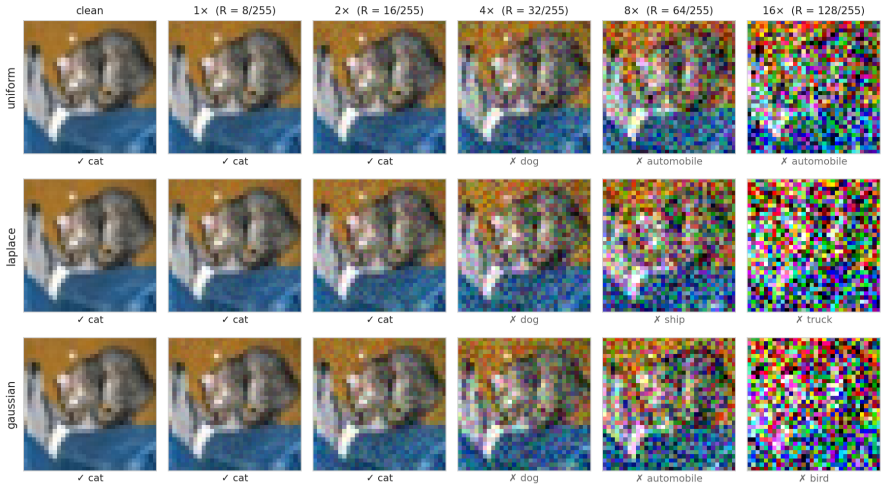


Figure 3: Representative uniform, Laplace, and Gaussian perturbations at increasing noise radii on CIFAR-10.

root search is then used to locate the zero of the corresponding class-wise margin. The signed displacement along this normal direction gives the observed normal displacement d_k of that boundary element.

At the same pre-update boundary location, we compute

$$S_{\text{flat}} = \frac{1}{\|\nabla_{\boldsymbol{\delta}} m\|_2}, \qquad S_{\text{param}} = |\langle \nabla_{\boldsymbol{\theta}} m, \Delta\boldsymbol{\theta} \rangle| \,, \tag{94}$$

together with their combination predicted by the implicit boundary-motion formula,

$$S_{\text{vel}} = \frac{|\langle \nabla_{\boldsymbol{\theta}} m, \Delta\boldsymbol{\theta} \rangle|}{\|\nabla_{\boldsymbol{\delta}} m\|_2}. \tag{95}$$

For each model–training-step–test-point combination, we compute the Spearman correlation between $|d_k|$ and each of S_{flat}, S_{param}, and S_{vel}. Panel (d) reports the resulting distributions of correlation coefficients as density histograms, with dashed vertical lines marking the median correlation for each quantity.

E High-Dimensional Experimental Setup

Common training setup. We use CIFAR-10 with a CIFAR-style ResNet-18, replacing the ImageNet stem by a 3×3 stride-one convolution and removing the initial max-pooling layer. Images are augmented by random cropping with four-pixel reflection padding and random horizontal flipping. Unless otherwise stated, models are optimized with SGD using momentum 0.9, Nesterov acceleration, weight decay 5×10^{-4}, batch size 128, and a cosine learning-rate schedule. The initial learning rate is 0.1 for all methods except TRADES, for which we use $\beta = 6$ and learning rate 0.01 for numerical stability. The models reported in the main robustness comparison are trained for 100 epochs. Throughout this section, $1\times = 8/255$ denotes the reference ℓ_∞ perturbation scale.

As a downstream proof of concept, NPR-Aug. incorporates the proposed NPR first-variation estimate into an MMA-based training procedure. MMA uses $d_{\max} = 32/255 = 4\times$. The NPR component uses iid perturbations drawn uniformly from the $4\times$ feasible box together with Gaussian localization around the zero-margin boundary. We use $\tau = 0.5$, four perturbation samples per input, and coefficient $\lambda = 1/2$. All remaining method-specific hyperparameters follow the corresponding baseline implementations.

Robustness evaluation. We report clean accuracy, PGD-10 adversarial accuracy, and PR. PGD-10 is evaluated under an ℓ_∞ constraint with $\epsilon = 8/255$. PR is evaluated as classification accuracy

under perturbations sampled from the specified stochastic distribution. The main comparison reports results at $R = 4\times$ and $R = 8\times$, with additional radii reported in the Appendix.

We consider four representative perturbation distributions in the main comparison. Uniform noise samples each coordinate independently from $U(-R, R)$. Laplace noise is sampled from a zero-centered Laplace distribution and truncated to $[-R, R]$. Gaussian noise is sampled from $\mathcal{N}(0, R^2)$ and truncated to the same interval. Sign noise independently takes each coordinate to either $-R$ or $+R$. These distributions share the same ℓ_∞ radius but are not matched in ℓ_2 energy. Representative perturbations at different radii are visualized in Fig. 3. The complete evaluation additionally includes low-frequency and sparse noise. Wall-clock cost is reported as average training time per epoch and excludes evaluation.

E.1 (A) Kernel-free Evaluation of Candidate NPR Directions

In the low-dimensional experiment, the complete decision boundary can be numerically traced and integrated. This is infeasible for CIFAR-10, so we instead use a high-sample Monte Carlo approximation of the smoothed NPR descent direction as a reference. At a fixed checkpoint, we use 128 images with 256 iid perturbations per image to construct

$$g^\star \approx g_\tau = \mathbb{E}_{\boldsymbol{\delta}\sim q}\left[K_\tau(m(\boldsymbol{\delta}))\nabla_{\boldsymbol{\theta}} m(\boldsymbol{\delta})\right], \tag{96}$$

with $\tau = 0.25$. We compare this reference with Monte Carlo estimates using $N \in \{4, 8\}$ perturbations per image and with three single-point candidates shown in Fig. 2(a): the MMA boundary anchor, the FAB nearest-boundary point, and a PGD interior point. Each candidate is converted into its corresponding parameter-space direction and normalized before evaluation.

Importantly, the candidate directions are evaluated independently of the kernel construction. For each unit-norm direction $\boldsymbol{v}$, we directly perturb the model parameters and compute the centered finite difference of the indicator-defined NPR,

$$D(\boldsymbol{v}) = \frac{\mathrm{NPR}(\boldsymbol{\theta} + h\boldsymbol{v}) - \mathrm{NPR}(\boldsymbol{\theta} - h\boldsymbol{v})}{2h}, \qquad \mathrm{NPR}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\delta}\sim q}\mathbf{1}\{m_{\boldsymbol{\theta}}(\boldsymbol{\delta}) \le 0\}. \tag{97}$$

Common random numbers are used for the two NPR evaluations, so the evaluation itself contains neither a smoothing kernel nor a surrogate loss. For sufficiently small h,

$$-D(\boldsymbol{v}) \approx \|\nabla_{\boldsymbol{\theta}}\mathrm{NPR}\|_2 \cos\left(\boldsymbol{v}, -\nabla_{\boldsymbol{\theta}}\mathrm{NPR}\right). \tag{98}$$

Thus, $-D(\boldsymbol{v})$ provides a kernel-free measure of alignment with NPR descent. The values reported in Fig. 2(a) are normalized by the corresponding value of $g^\star$, which is set to 100%. Error bars show 95% bootstrap intervals obtained from 4000 resamples over images.

E.2 (B) Level-set ablation.

We next examine whether the location of the localized margin level matters.

Starting from the same NPR training construction, we replace $\Phi(-m/\tau)$ by $\Phi((t-m)/\tau)$, whose gradient is localized around $m = t$ through $K_\tau(m - t)$. So that the update is localized around the level set $m = t$ rather than the decision boundary $m = 0$. All other components of the loss and training procedure are unchanged. These experiments use a $4\times$ uniform perturbation distribution, $N = 12$ iid perturbations per input, $\tau = 0.5$, NPR coefficient $\lambda = 1$, clean CE coefficient 1, and 40 training epochs.

We evaluate

$$t \in \{-2, -1, 0, +1, +2\}. \tag{99}$$

The $t = \pm 1$ experiments use three training seeds and are compared with the $t = 0$ model from the corresponding seed. The $t = \pm 2$ experiments use seed 42 only. Figure 2(b) reports the change in PR relative to the corresponding same-seed $t = 0$ model, reducing between-seed offsets in the comparison.

The two sides of the decision boundary have different geometric roles. For $t > 0$,

$$\{m \le 0\} \subseteq \{m \le t\}, \tag{100}$$

and therefore

$$F_{\boldsymbol{\theta}}(t) = \Pr[m_{\boldsymbol{\theta}} \le t] \ge \mathrm{NPR}_q. \tag{101}$$

Thus, the positive level defines an upper level containing the adversarial region. By contrast, for $t < 0$, $\{m \le t\} \subseteq \{m \le 0\}$, so the corresponding level does not upper-control NPR.

E.3 (c) Bandwidth ablation.

Finally, we examine sensitivity to the Gaussian bandwidth used during training. We use the same NPR training configuration as above with $R = 4\times$, $N = 12$, $\lambda = 1$, clean CE coefficient 1, 40 training epochs, and seed 42, changing only τ. We evaluate

$$\tau \in \{0.05, 0.25, 0.5, 1, 4, 16, 64\}. \tag{102}$$

PR is evaluated under uniform perturbations at

$$R \in \{4\times, 8\times, 16\times\}. \tag{103}$$

The range $\tau \in [0.05, 4]$ examines sensitivity within a relatively localized regime, whereas $\tau = 16$ and 64 produce substantially broader weighting over the observed margin distribution. For the normalized Gaussian kernel, changing τ also changes the overall scale of the NPR term. Accordingly, this experiment is interpreted as a bandwidth sensitivity analysis rather than as an isolated test of localization alone.

F Additional Experiment and Pseudo-code

Algorithm 1 NPR-Aug. Training

Require: Batch $(\boldsymbol{x}, \boldsymbol{y})$, $d_{\max} = 32/255$, $R_{\mathrm{PR}} = 4\times$, $\tau = 0.5$, $\lambda = 0.5$, $c = 1/3$, $N = 4$

1: Compute $L_{\mathrm{clean}} = \mathrm{CE}(h_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{y})$.
2: Apply standard MMA to obtain boundary anchors $\boldsymbol{a}_i$ and their corresponding boundary radii.
3: Let $\mathcal{K} = \{i : \text{boundary radius}_i < d_{\max}\}$.
4: Compute $L_{\mathrm{anchor}} = \sum_{i \in \mathcal{K}} \mathrm{CE}(h_{\boldsymbol{\theta}}(\boldsymbol{a}_i), y_i)$ and the standard MMA loss L_{wrong} for misclassified samples.
5: **if** $\mathcal{K} \neq \varnothing$ **then**
6: **for** $i \in \mathcal{K}$ **do**
7: Sample N perturbations $\boldsymbol{\delta}_{ik}$ uniformly from the feasible R_{PR} perturbation box.
8: Compute $m_{ik} = h_{\boldsymbol{\theta},y_i}(\boldsymbol{x}_i + \boldsymbol{\delta}_{ik}) - \max_{j \neq y_i} h_{\boldsymbol{\theta},j}(\boldsymbol{x}_i + \boldsymbol{\delta}_{ik})$.
9: Compute the probit-smoothed NPR loss $\ell_i^{\mathrm{NPR}} = N^{-1} \sum_k \Phi(-m_{ik}/\tau)$.
10: **end for**
11: $L_{\mathrm{boundary}} = (1 - \lambda) L_{\mathrm{anchor}} + \lambda \sum_{i \in \mathcal{K}} \ell_i^{\mathrm{NPR}}$.
12: **else**
13: $L_{\mathrm{boundary}} = L_{\mathrm{anchor}}$.
14: **end if**
15: $L_{\mathrm{MMA}} = (L_{\mathrm{boundary}} + L_{\mathrm{wrong}})/B$.
16: $L = (1 - c) L_{\mathrm{MMA}} + c L_{\mathrm{clean}}$.
17: Update $\boldsymbol{\theta}$ by SGD.

Table 2: **CIFAR-10 results with ResNet-18.** We report clean accuracy, PGD-10 adversarial accuracy, training cost, and probabilistic robustness (PR) in percentage. PR is evaluated under six stochastic perturbation distributions at $R \in \{4\times, 8\times, 16\times\}$, where $1\times = 8/255$. All perturbations within each radius block share the same L_∞ radius but are not matched in L_2 energy. The EVaR baseline is trained at the $1\times$ perturbation scale. Bold denotes the highest PR among the methods shown for each $(R, \text{noise distribution})$ condition.

Method	Clean	PGD-10	Time / epoch (s)	Total time	Rel. MMA
Standard	95.17	0.17	12.3	21 min	0.09×
EVaR	94.76	6.71	160.8	268 min	1.23×
PGD-AT	85.36	47.88	95.5	159 min	0.73×
TRADES	77.09	46.12	107.0	178 min	0.82×
MART	75.39	49.84	106.7	178 min	0.81×
AT-PR	84.42	47.43	608.3	1014 min	4.64×
MMA	85.40	**51.72**	131.0	218 min	1.00×
NPR-Aug.	88.93	45.55	165.6	276 min	1.26×

Method	Uniform	Laplace	Gaussian	Low-freq	Sign	Sparse
$\mathbf{R} = 4\times$						
Standard	47.36	31.31	32.03	89.99	16.78	90.57
EVaR	77.29	60.97	61.97	**91.05**	30.75	**94.33**
PGD-AT	81.99	79.07	79.35	83.94	71.75	84.95
TRADES	74.92	72.75	72.77	75.89	67.23	77.22
MART	72.29	70.09	70.30	74.13	64.73	74.90
AT-PR	81.67	78.94	79.07	83.12	72.25	84.07
MMA	76.78	74.36	74.54	82.92	70.33	83.20
NPR-Aug.	**83.01**	**82.03**	**82.14**	87.54	**79.54**	88.29
$\mathbf{R} = 8\times$						
Standard	13.74	11.34	11.33	75.55	10.38	78.73
EVaR	23.11	15.56	15.86	75.04	11.88	**89.02**
PGD-AT	63.72	50.46	51.25	76.23	31.31	83.06
TRADES	63.58	56.61	56.90	69.84	41.90	76.41
MART	61.08	52.91	53.62	69.72	36.80	73.87
AT-PR	64.87	52.59	53.41	75.39	31.98	82.52
MMA	66.39	61.14	61.51	75.20	51.77	78.79
NPR-Aug.	**76.02**	**69.01**	**69.61**	**78.08**	**54.45**	84.67
$\mathbf{R} = 16\times$						
Standard	10.32	10.29	10.25	42.92	10.27	49.08
EVaR	11.16	10.70	10.76	38.20	10.63	68.87
PGD-AT	22.27	14.32	14.46	45.73	9.88	76.72
TRADES	34.56	21.30	21.62	47.25	10.86	72.40
MART	30.26	19.57	19.68	48.15	10.69	70.22
AT-PR	22.29	12.96	13.26	44.46	10.34	77.01
MMA	45.20	32.24	32.58	**49.85**	14.81	71.55
NPR-Aug.	**46.62**	**33.47**	**33.89**	47.81	**19.65**	**78.44**

Notes. $1\times = 8/255$. Uniform, Laplace, Gaussian, low-frequency, sign, and sparse perturbations share the same L_∞ radius within each block but have different L_2 energies. Training time is measured over the full 100-epoch run, and relative cost is normalized by MMA. Bold denotes the highest PR among all methods shown for each $(R, \text{noise distribution})$ condition. All reported results use a single training seed.