Non-Parametric Probabilistic Robustness

A Conservative Risk Estimator under Unknown Perturbation Distributions

Zheng Wang Yi Zhang Siddartha Khastgir Carsten Maple Xingyu Zhao

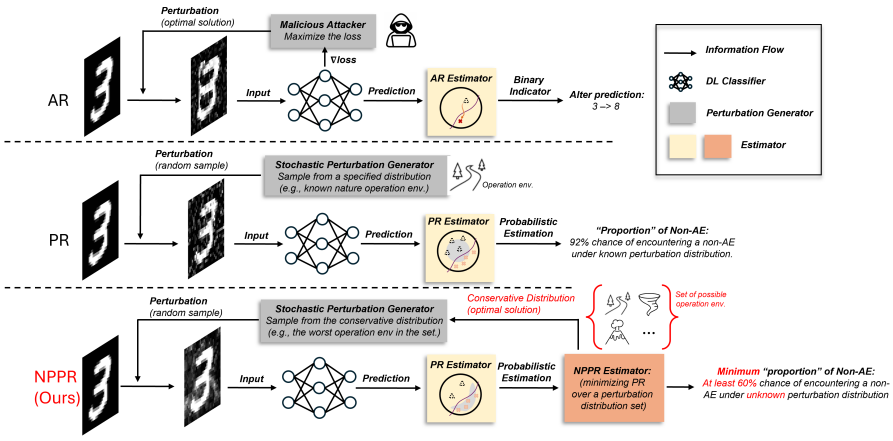
WMG, University of Warwick

ICML 2026

AR assumes a dedicated attacker — too pessimistic for random-perturbation settings

- Deep learning models fail under small input perturbations — adversarial examples reveal a fundamental vulnerability
- **Adversarial robustness (AR)** measures *worst-case* risk: can a carefully optimised attack fool the model?
- AR is appropriate for security settings, but real-world perturbations are often *random* — sensor noise, lighting shifts, transmission artefacts
- **Probabilistic robustness (PR) complements AR: it measures how likely a *random* perturbation is to cause failure**

Every existing PR method assumes the perturbation distribution is known — an assumption rarely satisfied in practice



Every existing PR method assumes the perturbation distribution is known — an assumption rarely satisfied in practice

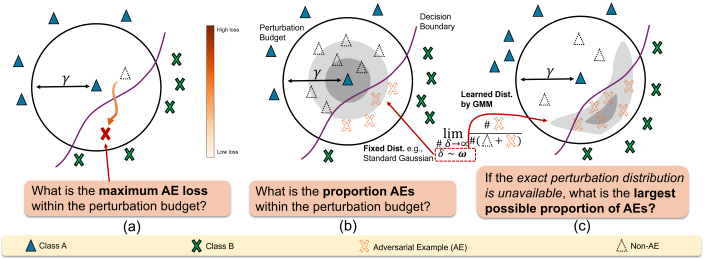
- PR estimates the fraction of random perturbations $\varepsilon \sim \omega$ that cause a model error — useful, but requires a fixed known ω
- All prior PR work (assessment *and* training) presupposes ω is given (e.g. Gaussian or Uniform) with no principled justification
- **In practice ω is almost never known; a misspecified distribution yields overconfident (too-high) robustness estimates**

We ask:

Can we evaluate PR conservatively — without assuming any fixed perturbation distribution — by learning the worst-case distribution directly from data?

Method

NPPR yields the most conservative robustness estimate by taking the infimum over all admissible perturbation distributions

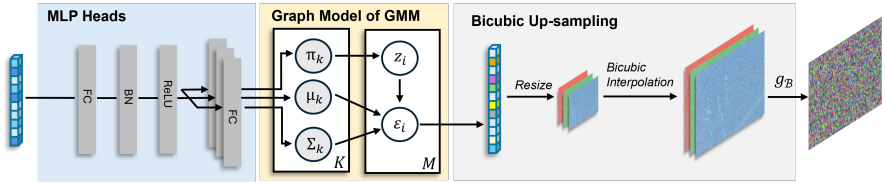


Definition (Non-parametric PR)

$$\mathfrak{S}_{\mathrm{NPPR}}(\boldsymbol{x}, y) \triangleq \inf_{\omega \in P_{\boldsymbol{\varepsilon}}} \mathbb{E}_{\boldsymbol{\varepsilon} \sim \omega(\cdot|\boldsymbol{x})} \left[\mathbf{1}_{h(\boldsymbol{x}+\boldsymbol{\varepsilon})=y} \right]$$

- **AR**: worst-case over one adversarial ε — extreme lower bound
- **PR**: expectation under a fixed ω — assumes distribution knowledge
- **NPPR**: minimises over *all* admissible ω — most conservative PR estimate

A GMM with MLP heads and bicubic up-sampling learns the conservative distribution — chosen for simplicity and interpretability



Component	Role
MLP heads	Output GMM parameters (π, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) conditioned on input features
GMM + Gumbel–Softmax	K-component mixture; differentiable relaxation enables end-to-end training
Bicubic up-sampling	Maps low-dim. perturbations to image space; scaled tanh enforces L_p budget
4 dependency modes	Independent · Label-dep. · Input-dep. · Joint (input + label)

Results

Under standard training, NPPR provides substantially more conservative estimates than fixed-distribution PR across all models

Clean-accuracy-normalised (%). **Bold** = NPPR (ours). L_∞ budget 16/255, joint dependency, K=7. AA = AutoAttack; S&P = Salt-and-Pepper.

Dataset	Model	Clean	NPPR↓	PR_{Gauss}	PR_{Unif}	AR_{AA}	S&P
CIFAR-10	ResNet18	87.72	**76.29**	95.28	97.21	0.00	83.63
CIFAR-10	VGG16	92.28	**74.53**	89.08	93.36	0.00	84.90
CIFAR-100	ResNet18	63.38	**58.65**	86.51	91.48	0.02	69.03
Tiny-IN	ResNet18	56.99	**53.20**	92.26	95.17	0.00	71.61

Full 12-model table: Appendix Table 1

Under adversarial training, fixed-distribution PR saturates near 100% and becomes uninformative — NPPR remains sensitive

Clean-accuracy-normalised (%). PGD-10 adversarial training. **Bold** = NPPR (ours).

Dataset	**Model**	**Clean**	**NPPR↓**	PR_{Gauss}	PR_{Unif}	AR_{PGD}	AR_{AA}
CIFAR-10	ResNet18	72.42	**95.59**	99.78	99.84	60.67	51.53
CIFAR-10	WRN50	78.83	**96.33**	99.86	99.97	58.33	50.35
CIFAR-100	ResNet18	49.39	**89.54**	99.96	99.95	43.88	34.64
Tiny-IN	ResNet18	21.34	**91.06**	99.70	99.78	44.05	27.65

Full 9-model table: Appendix Table 1

NPPR estimates are robust across different feature extractors and target classifiers, enabling flexible deployment

CIFAR-10, $L_\infty(16/255)$, joint dependency, K=7. Rows = feature extractor for GMM head; columns = target classifier. **Bold** = same-backbone baseline.

Feat. \ **Target**	ResNet18	ResNet50	VGG16	WRN50
ResNet18	**76.27**	80.12	74.99	77.26
ResNet50	86.41	**86.81**	86.03	86.09
VGG16	76.60	80.66	**74.47**	77.84
WRN50	82.40	84.21	79.79	**81.32**

NPPR adds only modest inference overhead over fixed-distribution PR — and is substantially faster than AutoAttack

Inference wall-clock time (seconds) on a single NVIDIA RTX 3090. NPPR training cost per epoch shown in parentheses.

Dataset	**Model**	**NPPR** (train/ep)	PR_{Gauss}	AR_{PGD}	AR_{AA}
CIFAR-10	ResNet18	**3.96** (69s)	2.55	2.96	30.20
CIFAR-10	WRN50	**16.15** (226s)	14.07	13.76	127.12
Tiny-IN	ResNet18	**8.36** (203s)	6.57	6.11	40.08
Tiny-IN	WRN50	**37.16** (908s)	34.29	30.28	244.73

NPPR is $\approx$1.5$\times$ PR inference cost; AutoAttack is **7–30**$\times$ **slower** than NPPR.

Conclusions

1. NPPR removes the need for a predefined perturbation distribution in PR evaluation
2. Provably: $\mathcal{G}_{\mathrm{AR}} \leq \mathcal{G}_{\mathrm{NPPR}} \leq \mathcal{G}_{\mathrm{PR}}$ — NPPR interpolates between worst-case and average-case robustness
3. Fixed-distribution PR saturates under adversarial training; NPPR remains informative
4. Lightweight GMM head is compatible with any frozen classifier and feature extractor

xingyu.zhao@warwick.ac.uk github.com/squarewang2077/nppr

References I


Webb, S., Rainforth, T., Teh, Y. W., & Kumar, M. P. (2018). A statistical approach to assessing neural network robustness. *ICLR*.


Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *ICLR*.


Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *IEEE S&P*.


Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *ICML*.


Rice, L., Garg, S., & Kolter, J. Z. (2021). Robustness between the worst and average case. *NeurIPS*.


Zhao, X., et al. (2025). Probabilistic robustness: a survey. *arXiv*.

Appendix A — Cross-radius analysis confirms NPPR degrades monotonically with larger perturbation budgets

NPPR (%) on ResNet18/CIFAR-10, trained at 16/255, evaluated across radii. Joint dependency, K=7.

Train \ **Eval**	4/255	8/255	16/255	32/255
4/255	96.56	91.58	79.63	59.79
8/255	96.77	91.31	77.66	55.60
16/255	96.75	90.99	**76.27**	52.49

NPPR decreases monotonically with larger evaluation radius; training at a larger radius gives more conservative estimates.

Appendix B — Ablation: joint dependency and more GMM components yield more conservative NPPR estimates

ResNet18 / CIFAR-10. ↓ lower NPPR = more conservative. **Bold** = best per dependency column.

Configuration	Independent	Label dep.	Input dep.	Joint dep.
Base (K=3)	74.99	75.79	85.36	79.20
+ Learnable decoder	81.27	69.08	84.76	78.59
+ K=7	73.25	64.88	84.17	76.27
+ K=12	74.43	**64.87**	**83.31**	**73.11**