Understanding Probabilistic Robustness (PR): An Introductory yet Rigorous Guide

Kenneth

Abstract

Probabilistic Robustness (PR) measures the likelihood that a machine learning model's prediction remains correct under random, bounded perturbations of its inputs. Unlike adversarial robustness, which considers worst-case perturbations, PR adopts a statistical view, providing insights into a model's average-case behavior under realistic noise. This article offers beginners an intuitive and formal understanding of PR, covering definitions, estimation techniques, and practical considerations.

1 Introduction

Modern machine learning models, particularly deep neural networks, often face unpredictable variations in their inputs due to sensor noise, compression artifacts, or environmental changes. While adversarial robustness focuses on worst-case crafted attacks, *probabilistic robustness* (PR) quantifies the probability that a model's prediction remains stable under *stochastic* perturbations within a prescribed bound.

2 Intuitive Illustration

Imagine a handwritten digit classifier that labels the digit "3" correctly. Now suppose we add random pixel noise drawn uniformly in a small range to each image—most of these noisy variants will still look like a "3" to a human and should be classified as such. PR asks: *What fraction of these random perturbations does the model classify correctly?*

3 Formal Definition

Let

$$f : \mathbb{R}^d \to \{1, \ldots, K\}$$

be a classifier, and let (x, y) be a test example with true label $y = f(x)$. Fix a perturbation radius $\gamma > 0$ and a noise distribution $\mathcal{D}$ over $\mathbb{R}^d$. We define the *probabilistic robustness* of f at (x, y) by

$$\mathrm{PR}_{\mathcal{D}}(x, y; \gamma) \;=\; \Pr_{\delta \sim \mathcal{D}}\big[\|\delta\|_p \le \gamma \;\wedge\; f(x + \delta) = y\big].$$

Equivalently, one often conditions on the norm constraint:

$$\mathrm{PR}_{\mathcal{D}}(x, y; \gamma) \;=\; \Pr_{\substack{\delta \sim \mathcal{D} \\ \|\delta\|_p \le \gamma}} \big[f(x + \delta) = y\big].$$

Common choices for $\mathcal{D}$ include the uniform distribution on the L_p ball of radius γ, or an isotropic Gaussian $\mathcal{N}(0, \sigma^2 I)$ with σ tuned so that $\Pr(\|\delta\|_p \le \gamma)$ is high.

4 Estimating PR via Monte Carlo

Exact computation of PR is intractable for high-dimensional inputs. Instead, we approximate by drawing N i.i.d. samples $\{\delta_i\}_{i=1}^N$ from $\mathcal{D}$ (rejecting those with $\|\delta_i\|_p > \gamma$ if necessary) and compute

$$\widehat{\mathrm{PR}}(x, y; \gamma) \;=\; \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{f(x + \delta_i) = y\}.$$

By the law of large numbers, $\widehat{\mathrm{PR}} \to \mathrm{PR}$ as $N \to \infty$. A confidence interval can be constructed via the binomial distribution:

$$\widehat{\mathrm{PR}} \pm z_{\alpha/2} \sqrt{\frac{\widehat{\mathrm{PR}}(1 - \widehat{\mathrm{PR}})}{N}}.$$

5 Comparison with Adversarial Robustness

- **Adversarial Robustness (AR):** Guarantees $\min_{\|\delta\|_p \leq \gamma} f(x + \delta) = y$ (worst-case). Very stringent, often pessimistic.
- **Probabilistic Robustness (PR):** Measures $\Pr_{\|\delta\|_p \leq \gamma}[f(x + \delta) = y]$ (average-case). Reflects performance under typical noise.

PR is often easier to estimate at scale and more aligned with real-world scenarios where noise is random rather than adversarially chosen.

6 Practical Considerations

6.1 Choice of Distribution $\mathcal{D}$

- *Uniform noise* on the L_∞ ball models worst-case bounded random errors.
- *Gaussian noise* captures sensor noise or natural variations.

6.2 Computational Cost

Monte Carlo sampling can be expensive for large N or costly models. Strategies to reduce cost:

- *Variance reduction* (e.g. importance sampling).
- *Adaptive sampling* focusing on boundary regions.
- *Surrogate models* for faster approximate queries.

7 Applications and Insights

- **Model Comparison:** PR provides a scalar metric to compare models' resilience to random perturbations.
- **Diagnostics:** Identifies inputs (or classes) that a model handles poorly under noise.
- **Robust Training:** One can train to maximize PR by augmenting with random noise or by optimizing a risk-based objective:

$$\min_{\theta} \mathbb{E}_{(x,y),\, \delta \sim \mathcal{D}}[\mathcal{L}(f_\theta(x + \delta), y)].$$

8 Conclusion

Probabilistic robustness complements adversarial robustness by evaluating model stability under realistic random perturbations. Beginners can implement simple Monte Carlo estimators to gain intuition about their models' behavior in noisy environments and leverage PR as a practical metric for robustness benchmarking.

Further Reading:

- S. Kolter and E. Wong (2018), "Provable defenses against adversarial examples via the convex outer adversarial polytope."
- D. Hendrycks and T. Dietterich (2019), "Benchmarking neural network robustness to common corruptions and perturbations."
- A. Cohen, X. Rosenfeld, and Z. Kolter (2019), "Certified adversarial robustness via randomized smoothing."