Probabilistically Robust Conformal Prediction

Abstract

Conformal prediction (CP) is a framework to quantify uncertainty of machine learning classifiers including deep neural networks. Given a testing example and a trained classifier, CP produces a prediction set of candidate labels with a user-specified coverage (i.e., true class label is contained with high probability). Almost all the existing work on CP assumes clean testing data and there is not much known about the robustness of CP algorithms w.r.t natural/adversarial perturbations to testing examples. This paper studies the problem of probabilistically robust conformal prediction (PRCP) which ensures robustness to most perturbations around clean input examples. PRCP generalizes the standard CP (cannot handle perturbations) and adversarially robust CP (ensures robustness w.r.t worst-case perturbations) to achieve better trade-offs between nominal performance and robustness. We propose a novel adaptive PRCP (aPRCP) algorithm to achieve probabilistically robust coverage. The key idea behind aPRCP is to determine two parallel thresholds, one for data samples and another one for the perturbations on data (aka "*quantile-of-quantile*" design). We provide theoretical analysis to show that aPRCP algorithm achieves robust coverage. Our experiments on CIFAR-10, CIFAR-100, and ImageNet datasets using deep neural networks demonstrate that aPRCP achieves better trade-offs than state-of-the-art CP and adversarially robust CP algorithms.

1 INTRODUCTION

Deep learning has shown significant success in diverse real-world applications. However, to deploy these deep models in safety-critical applications (e.g, autonomous driving and medical diagnosis), we need uncertainty quantification (UQ) tools to capture the deviation of the prediction from the ground-truth output. For example, producing a subset of candidate labels referred to as *prediction set* for classification tasks. Conformal prediction (CP) [Vovk et al., 1999, 2005, Shafer and Vovk, 2008] is a framework for UQ that provides formal guarantees for a user-specified *coverage*: ground-truth output is contained in the prediction set with a high probability $1 - \alpha$ (e.g., 90%). There are two key steps in CP. First, in the prediction step, we use a black-box classifier (e.g., deep neural network) to compute *(non-)conformity* scores which measure similarity between calibration examples and a testing input. Second, in the calibration step, we use the conformity scores on a set of calibration examples to find a threshold to construct prediction set which meets the coverage constraint (e.g., $1 - \alpha$=90%). The *efficiency* of CP [Sadinle et al., 2019] is measured in terms of size of the prediction set (smaller the better) which is important for human-ML collaborative systems [Rastogi et al., 2022].

In spite of the recent successes of CP [Vovk et al., 2005], there is little known about the robustness of CP to adversarial perturbations of clean inputs. Most CP methods [Cauchois et al., 2020, Gibbs and Candes, 2021, Tibshirani et al., 2019, Podkopaev and Ramdas, 2021, Guan and Tibshirani, 2022] are brittle as they assume clean input examples and cannot handle *any* perturbations. The recent work on adversarially robust CP [Gendler et al., 2022] ensures robustness to *all* perturbations bounded by a norm ball with radius r. However, this conservative approach of dealing with *worst-case* perturbations can degrade the nominal performance (evaluation on only clean inputs) of the CP method. For example, the prediction set size can be large even for clean and easy-to-classify inputs, which increases the burden of human expert in human-ML collaborative systems [Cai et al., 2019, Rastogi et al., 2022]. The main research question of this paper is: *how can we develop probably correct CP algorithms for ensuring robustness to most perturbations for (pre-trained) deep classifiers?*

To answer this question, we present a general notion for

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

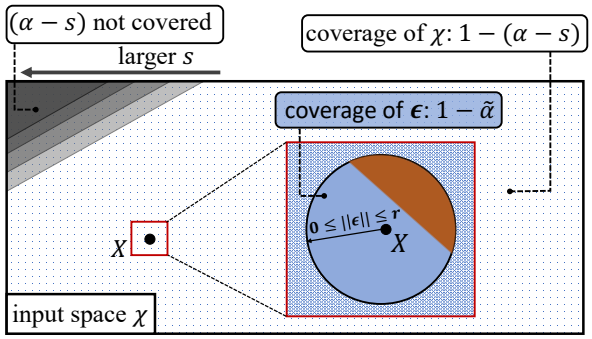

Figure 1: Conceptual illustration of the adaptive PRCP setting. The goal is to improve the robustness of the CP framework to handle perturbations ϵ bounded by r for every input $X \in \mathcal{X}$. The robust quantile corresponding to 1-$\tilde{\alpha}$ region (blue circle around X) is computed by accounting for most of the perturbed data $X + \epsilon$ (see (8)). s is a conservativeness parameter for the robust quantile that can be varied to achieve the target marginal coverage $1 - \alpha + s$ (see (9)). Adaptive PRCP can find a trade-off between the marginal coverage on feature space (X, Y) and the robustness for perturbation ϵ by changing the value of $\tilde{\alpha}$ and s to achieve probabilistically robust coverage (See Definition 3).

probabilistically robust coverage that balances the standard conformal coverage and the adversarial (worst-case) coverage as the fundamental setting. To address this challenge, we develop the adaptive PRCP algorithm (aPRCP) which is based on the principle of "*quantile-of-quantile*" design: consists of two parallel quantiles as illustrated in Figure 1: one defined in the perturbed noise space (see (8)), the other one in the data space (9). Our analysis fixes one quantile probability as a given hyper-parameter, and finds the other one to achieve the target probabilistically robust coverage. We provide theoretical analysis for probabilistic correctness of aPRCP at the population level and the approximation error of empirical quantiles as a function of the number of samples. As a result, aPRCP achieves improved trade-offs between nominal performance (evaluation on clean inputs) and robust performance (evaluation on perturbation inputs) for both probabilistic and worst-case settings as illustrated in Figure 2, which is analogous to the recent work on probabilistically robust learning Robey et al. [2022].

Contributions. The key contribution of this paper is the development, theoretical analysis, and empirical evaluation of the aPRCP algorithm. Our specific contributions include:

- A general notion of probabilistically robust coverage for conformal prediction against perturbations of clean input examples.
- Development of the adaptive PRCP algorithm based on the principle of "*quantile-of-quantile*" design.
- Theory to show that aPRCP algorithm achieves probabilistically robust coverage for adversarial examples.
- Experimental evaluation of aPRCP method on classification benchmarks using deep models to demonstrate its efficacy over prior CP methods on CIFAR-10, CIFAR-100, and ImageNet. Our anonymized code is in the Appendix for review purposes.

2 BACKGROUND AND PROBLEM SETUP

We consider the problem of uncertainty quantification (UQ) of pre-trained deep models for classification tasks in the presence of adversarial perturbations. Suppose (X, Y) is a data sample where X is an input from the space $\mathcal{X}$ and $Y \in \mathcal{Y}$ is the corresponding ground-truth output. For classification tasks, $\mathcal{Y}$ is a set of C discrete class-labels $\{1, 2, \cdots, C\}$. Let ϵ denote the l_2-norm bounded noise, i,e,. $\mathcal{E}_r = \{\epsilon \in \mathcal{X} : \|\epsilon\|_2 \leq r\}$ that is independent from data sample (X, Y). Let $\mathcal{P}_{X,Y}$ and $\mathcal{P}_\epsilon$ denote the underlying distribution of (X, Y) and ϵ, respectively. We also define $Z = (X, Y, \epsilon)$ as the joint random variable and the perturbed input example $\widetilde{X} = X + \epsilon$ for notational simplicity.

Uncertainty Quantification. Let $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{cal}}$ correspond to sets of training and calibration examples drawn from a target distribution $\mathcal{P}_{X,Y}$. We assume the availability of a pre-trained deep model $F_\theta : \mathcal{X} \mapsto \mathcal{Y}$, where θ stands for the parameters of the deep model. For a given testing input $\widetilde{X}$, we want to compute UQ of the deep model F_θ in the form of a prediction set $\mathcal{C}(\widetilde{X})$, a subset of candidate class-labels $\{1, 2, \cdots, C\}$. The performance of UQ for clean data samples (i.e., ϵ=0) is measured using two metrics. First, the (marginal) *coverage* is defined as the probability that the ground-truth output Y is contained in $\mathcal{C}(X)$ for a testing example (X, Y) from the same data distribution $\mathcal{P}_{X,Y}$, i.e., $\mathbb{P}(Y \in \mathcal{C}(X))$. The empirical coverage `Cov` is measured over a given set of testing examples $\mathcal{D}_{\text{test}}$. Second, *efficiency*, denoted by `Eff`, measures the cardinality of the prediction set $\mathcal{C}(X)$. Smaller prediction set means higher efficiency. It is easy to achieve the desired coverage (say 90%) by always outputting $\mathcal{C}(X)$=$\mathcal{Y}$ at the expense of poor efficiency.

Conformal Prediction (CP). CP is a framework that allows us to compute UQ for any given predictor through a conformalization step. The key element of CP is a score function

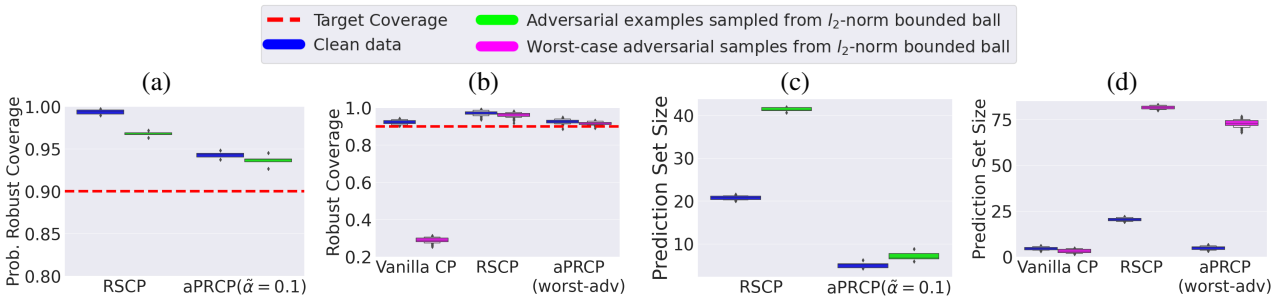

Figure 2: Results on CIFAR100 dataset using a ResNet model to illustrate the trade-offs between nominal performance (evaluation on clean data) and robust performance (evaluation on adversarial examples) for Vanilla CP, RSCP, and variants of the aPRCP algorithm. (a) and (c) show the evaluation against clean examples and their corresponding noisy samples (i.e., $\tilde{X} = X + \epsilon; ||\epsilon||_2 \leq r$) w.r.t probabilistic robustness. (b) and (d) show the evaluation against clean examples and their corresponding bounded adversarial examples. aPRCP(worst-adv) is the variant of aPRCP that works for worst adversarial data. Vanilla CP fails to achieve coverage for worst-case adversarial data. RSCP achieves a robust coverage much higher than the target (nominal) coverage, resulting in large prediction sets. aPRCP achieves better results (tighter coverage and smaller prediction set size) than vanilla CP and RSCP in terms of the joint performance on clean, noisy, and worst-adversarial data.

S that computes the *conformity* (or *non-conformity*) score, measures similarity between labeled examples, which is used to compare a given testing input to the calibration set $\mathcal{D}_{\text{cal}}$. Since any non-conformity score can be intuitively converted to a conformity measure [Vovk et al., 2005], we use non-conformity measure for ease of technical exposition. Let $S(X, Y)$ denote the non-conformity score function of data sample (X, Y). For a sample (X_i, Y_i) from the calibration set $\mathcal{D}_{\text{cal}}$, we use $S_i = S(X_i, Y_i)$ as a shorthand notation of its non-conformity score.

A typical method based on split conformal prediction has a threshold τ to compute UQ in the form of prediction set for a given testing input X and deep model F_θ. A small set of calibration examples $\mathcal{D}_{\text{cal}}$ are used to select the threshold t for achieving the given coverage $1-\alpha$ (say 90%) empirically on $\mathcal{D}_{\text{cal}}$. Let $Q(\alpha) := \min\{t : \mathbb{P}_{X,Y}\{S(X, Y) \leq t\} \geq 1 - \alpha\}$ be the true quantile of the conformity score for (X, Y). Let $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ denote a calibration set with n exchangeably drawn random samples from the underlying distribution $\mathcal{P}_{X,Y}$. We denote the $(1 - \alpha)$-quantile derived from $\{S_i\}_{i=1}^n$ by $Q(\alpha; \{S_i\}_{i=1}^n) = S_{(\lceil (1-\alpha)n \rceil)}$. The prediction set for a new testing input X is given by $\mathcal{C}(X)$=$\{y : S(X, y) \leq \tau\}$ using a threshold τ. CP provides valid guarantees that $\mathcal{C}(X)$ has coverage $1 - \alpha$ on future examples drawn from the same distribution $\mathcal{P}_{X,Y}$.

For classification, several non-conformity scores can be employed. The homogeneous prediction sets (HPS) score is defined [Vovk et al., 2005, Lei et al., 2013] as follows:

$$S^{\text{HPS}}(X, y) = 1 - F_\theta(X)_y, \tag{1}$$

where $F_\theta(X)_y \in [0, 1]$ is the probability corresponding to the true class y using the deep model F_θ. Recent work has proposed the adaptive prediction sets (APS) [Romano et al., 2020b] score that is based on ordered probabilities. The score function of APS is defined as follows:

$$S^{\text{APS}}(X, y) = \sum_{y' \in \mathcal{Y}} F_\theta(X)_{y'} \mathbb{1}\left\{F_\theta(X)_{y'} > F_\theta(X)_y\right\} + u.F_\theta(X)_y, \tag{2}$$

where u is a random variable uniformly distributed over $[0, 1]$ and $\mathbb{1}$ is the indicator function.

Problem Definition. The high-level goal of this paper is to study methods to improve the robustness of the standard CP framework to adversarial/noisy examples of the form $\widetilde{X} = X + \epsilon$, where ϵ is the additive perturbation from $\mathcal{E}_r = \{\epsilon \in \mathbb{R}^d : \|\epsilon\|_2 \leq r\}$. Specifically, we propose a novel adaptive probabilistically robust conformal prediction (aPRCP) algorithm which accounts for $(1 - \tilde{\alpha})$ (see $\tilde{\alpha}$ for robust quantile in (8)) fraction of perturbations in $\mathcal{E}_r$ for each data (X, Y). Setting $\tilde{\alpha} = 0$ as an extreme case makes aPRCP handle all perturbations (i.e., worst-case), similar to RSCP [Gendler et al., 2022]. We theoretically and empirically analyze aPRCP to demonstrate improved trade-offs between nominal performance (evaluation on clean inputs) and robust performance (evaluation on perturbation inputs). Figure 1 conceptually illustrates the PRCP problem setting.

3 ROBUST CONFORMAL PREDICTION

This section describes our proposed adaptive probabilistically robust conformal prediction (aPRCP) algorithm. First, we introduce the notion of adversarially robust coverage and extend it to probabilistically robust coverage. Next, we motivate the significance of aPRCP algorithm and study the theoretical connection between aPRCP and adversarially robust CP setting [Gendler et al., 2022] in terms of probabilistically robust coverage and prediction set size. Finally, we analyze the gap between empirical and population level

quantiles in terms of the number of data samples.

3.1 PROBABILISTICALLY ROBUST COVERAGE

This section introduces the expanded notation of inflation condition on the conformity scoring function from the worst-case adversarial robustness setting to the more general probabilistic robustness setting. We start with the following definitions that are originally introduced for the ARCP setting [Gendler et al., 2022] and capture the inflation property of the score function for deriving adversarial robustness.

Definition 1. *(Adversarially robust coverage) A prediction set $\mathcal{C}(\widetilde{X})$ provides $(1-\alpha)$-adversarially robust coverage if for a desired coverage probability $1-\alpha \in (0,1)$:*

$$\mathbb{P}_{X,Y}\{Y \in \mathcal{C}(\widetilde{X} = X + \epsilon), \forall \epsilon \in \mathcal{E}_r\} \geq 1-\alpha. \quad (3)$$

Definition 2. *(M_r-adversarially inflated score function) $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is an M_r-adversarially inflated score function if the following inequality holds:*

$$S(X+\epsilon, Y) \leq S(X,Y) + M_r, \quad \forall X \in \mathcal{X}, Y \in \mathcal{Y} \text{ and } \epsilon \in \mathcal{E}_r. \quad (4)$$

The strategy of RSCP algorithm [Gendler et al., 2022] for the ARCP setting is to directly add an inflated quantity M_r to the quantile determined from the clean data (X,Y),

$$\tau^{\text{AR}}(\alpha) := Q(\alpha) + M_r, \quad (5)$$

and construct a prediction set with $\mathcal{C}^{\text{AR}}(X) = \{y \in \mathcal{Y} : S(X+\epsilon, y) \leq \tau^{\text{AR}}(\alpha)\}$. To this end, since $Q(\alpha)$ provides $(1-\alpha)$ marginal coverage on clean data (X,Y), $\tau^{\text{AR}}(\alpha)$ thus guarantees $(1-\alpha)$-adversarially robust coverage on adversarial data $(X+\epsilon, Y)$. This result is summarized in the following proposition.

Proposition 1. *(Adversarially robust coverage of RSCP, Theorem 1 in [Gendler et al., 2022]) Assume the score function S is M_r-adversarially inflated. Let $\mathcal{C}^{AR}(\widetilde{X}) = \{y \in \mathcal{Y} : S(\widetilde{X}, y) \leq \tau^{AR}(\alpha)\}$ be the prediction set for a testing sample $\widetilde{X}$. Then RSCP achieves $(1-\alpha)$-adversarially robust coverage.*

Now we extend the notion of adversarially robust coverage to the more general and relaxed condition, i.e., probabilistically robust coverage, by introducing the definition below.

Definition 3. *(Probabilistically robust coverage) A prediction set $\mathcal{C}(\widetilde{X})$ provides $(1-\alpha)$-probabilistically robust coverage if for a desired coverage probability $1-\alpha \in (0,1)$:*

$$\mathbb{P}_{X,Y,\epsilon}\{Y \in \mathcal{C}(\widetilde{X} = X + \epsilon)\} \geq 1-\alpha. \quad (6)$$

We highlight that the key difference between adversarially robust coverage (Definition 1) and probabilistically robust coverage (Definition 3) is whether the distribution of the perturbation ϵ is involved in the comparison with the target probability $1-\alpha$: probabilistically robust coverage goes though the joint distribution involving ϵ, i.e., $\mathbb{P}_{X,Y,\epsilon}\{\cdot\}$ in (6) instead of $\mathbb{P}_{X,Y}\{\cdot, \forall \epsilon \in \mathcal{E}_r\}$ in (3). Based on this understanding, we can see that a conformal prediction method can achieve $(1-\alpha)$-probabilistically robust coverage if it can satisfy $(1-\alpha)$-adversarially robust coverage. For the same target probability $(1-\alpha)$, adversarially robust coverage is more difficult to achieve than probabilistically robust coverage. Hence, the notion of probabilistic robustness for CP is more general and relaxed.

Naturally, we now extend the definition of the uniform inflated score function (Definition 2) to the following one.

Definition 4. *($M_{r,\eta}$-probabilistically inflated score function) $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is an $M_{r,\eta}$-probabilistically inflated score function if the following inequality holds for $\eta \in [0, \alpha]$:*

$$\mathbb{P}_Z\big\{S(X+\epsilon, Y) \leq S(X,Y) + M_{r,\eta}\big\} \geq 1-\eta. \quad (7)$$

The above definition regarding the inflation of the score function is general and includes (4) given in Definition 2 as a special case: By simply setting $\eta = 0$, we get $\mathbb{P}_Z\{S(X+\epsilon, Y) \leq S(X,Y) + M_{r,0}\} \geq 1$, i.e., $M_{r,0} = M_r$. Again, we highlight that the above condition involves the joint distribution on Z, as in Definition 3.

Based on the extension from adversarial to probabilistic robustness setting, it is easy to develop a similar principle on the *inflated* score function to derive probabilistically robust coverage, which we refer to as inflated probabilistically robust conformal prediction (`iPRCP`). To this end, let

$$\tau^{\text{iPR}}(\alpha; \eta) := Q(\alpha^*_{\text{iPR}}) + M_{r,\eta},$$

where $\alpha^*_{\text{iPR}} = 1-(1-\alpha)/(1-\eta)$. $\tau^{\text{iPR}}(\alpha;\eta)$ is the threshold determined by iPRCP that treats η from probabilistically inflated score function as a hyper-parameter. We use α^*_{iPR} as the probability for deriving the quantile on clean data, as (5) in ARCP.

Proposition 2. *(Probabilistically robust coverage of iPRCP) Assume the score function S is an $M_{r,\eta}$-probabilistically inflated. Let $\mathcal{C}^{iPR}(\widetilde{X}) = \{y \in \mathcal{Y} : S(\widetilde{X}, y) \leq \tau^{iPR}(\alpha;\eta)\}$ be the prediction set for a testing sample $\widetilde{X} = X + \epsilon$. Then iPRCP achieves $(1-\alpha)$-probabilistically robust coverage.*

This result shows that we can guarantee the $(1-\alpha)$-probabilistically robust coverage if we use $\tau^{\text{iPR}}(\alpha;\eta)$ to construct the prediction set $\mathcal{C}^{\text{iPR}}$. While the idea is simple and follows the inflation quantile used in the ARCP setting, it implies that we *have to know* $M_{r,\eta}$, the inflated quantity on the clean quantile. This requires us to know the score

Algorithm 1 adaptive PRCP (`aPRCP`)

1: **Input**: target probability $\alpha \in (0,1)$; the hyper-parameter s; set $\tilde{\alpha} = 1 - \frac{1-\alpha}{1-\alpha+s}$; split data into disjoint training set $\mathcal{D}_{\text{tr}}$ and calibration set $\mathcal{D}_{\text{cal}}$ with $|\mathcal{D}_{\text{cal}}| = n$.
2: Train a classifier F_θ on $\mathcal{D}_{\text{tr}}$.
3: Draw $\epsilon_{ij} \sim \mathcal{P}_\epsilon$ where $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, m\}$ denote the indices of data (X_i, Y_i) and its m perturbations.
4: Compute scores: $S_{ij} = S(X_i + \epsilon_{ij}, Y_i), \forall i, j$.
5: Compute empirical robust quantiles: $\widehat{Q}_i^{\text{rob}} = \widehat{Q}^{\text{rob}}(X_i, Y_i; \tilde{\alpha}) = Q(\tilde{\alpha}, \{S_{ij}\}_{j=1}^m)$ via (8), $\forall i$.
6: Determine threshold $\tau^{\text{aPR}}(\alpha; s) = \widehat{Q}^{\text{rob}}_{(\lceil n(1-\alpha+s)\rceil)}$ from empirical robust quantiles according to (9).
7: Receive $\widetilde{X}_{n+1}$ and construct prediction set: $\mathcal{C}(\widetilde{X}_{n+1}) = \{y \in \mathcal{Y} : S(\widetilde{X}_{n+1}, y) \leq \tau^{\text{aPR}}(\alpha; s)\}$.

function very well. Otherwise, we have to design a score function that satisfies the desired condition, similar to how the randomly smoothed score function was designed by RSCP algorithm to work for the ARCP setting [Gendler et al., 2022]. It was carefully designed to offer a uniform Lipschitz continuity with the requirement of an additional set of Gaussian random samples. This design may introduce additional restrictions, since extra samples are required every time the score function is applied, including each calibration and testing sample. Therefore, we would like to address the following question: *Can we design an adaptive algorithm to fit the underlying distribution without any prior knowledge or special design of the score function?*

3.2 ADAPTIVE PRCP ALGORITHM

This section presents our adaptive algorithm for achieving probabilistically robust coverage (aPRCP). We summarize it in Algorithm 1 and elaborate it below. First, we define the $(1-\tilde{\alpha})$-*robust quantile* for a given X as follows

$$Q^{\text{rob}}(X, Y; \tilde{\alpha}) := \min\{t : \mathbb{P}_\epsilon\{S(\widetilde{X}, Y) \leq t\} \geq 1 - \tilde{\alpha}\}. \quad (8)$$

Given (X, Y) and $\tilde{\alpha}$, $Q^{\text{rob}}(X, Y; \tilde{\alpha})$ returns the quantile from all randomly perturbed $\widetilde{X} = X + \epsilon$ over $\epsilon \in \mathcal{E}_r$. It acquires the inflated quantity from a local region of X as $\tilde{\alpha}$ indicates how conservative this inflation can be. We denote the empirical robust quantile (in Line 5 of Algorithm 1) by $\widehat{Q}^{\text{rob}}$.

Next, we define the threshold of the proposed adaptive PRCP (`aPRCP`) for a hyper-parameter $s \in [0, \alpha]$ as follows.

$$\tau^{\text{aPR}}(\alpha; s) = \min\{t : \mathbb{P}_{X,Y}\{Q^{\text{rob}}(X, Y; \alpha^*_{\text{aPR}}) \leq t\} \geq 1 - \alpha + s\}, \quad (9)$$

where $\alpha^*_{\text{aPR}} = 1 - (1-\alpha)/(1-\alpha+s)$ is a conservativeness parameter for the robust quantile in (8) that depends on the target probability α and the hyper-parameter s. In practice, the empirical threshold $\widehat{\tau}^{\text{aPR}} = \widehat{Q}^{\text{rob}}_{(\lceil n(1-\alpha+s)\rceil)}$ is selected from empirical robust quantiles $\{\widehat{Q}_i^{\text{rob}}\}_{i=1}^n$ (in Line 6 of Algorithm 1). Our aPRCP algorithm is adaptive since it finds α^*_{aPR} that is adaptive to the underlying distribution of (X, Y) as long as α and s are fixed apriori. The following formal result guarantees the probabilistically robust coverage for the aPRCP algorithm.

Theorem 1. *(Probabilistically robust coverage of aPRCP) Let $\mathcal{C}^{aPR}(\widetilde{X} = X + \epsilon) = \{y \in \mathcal{Y} : S(\widetilde{X}, y) \leq \tau^{aPR}(\alpha; s)\}$ be the prediction set for a testing sample $\widetilde{X}$. Then aPRCP achieves $(1-\alpha)$-probabilistically robust coverage.*

Remark 1. In fact, $\tau^{\text{aPR}}(\alpha; s)$ is the $(1-\alpha+s)$-th quantile (going through (X, Y)) of the $(1-\alpha^*_{\text{aPR}})$-robust quantiles (going through ϵ). One benefit of aPRCP is the transfer of the inflation from the score function to the specified probability (i.e., an s increase in probability). Therefore, it is not required to have a prior knowledge of either M_r as in ARCP or $M_{r,\eta}$ as in iPRCP. Instead, aPRCP requires finding a feasible and a good value for α^*_{aPR} by treating s as a hyper-parameter, though it inflates the specified probability, i.e., $1-\alpha+s \geq 1-\alpha$, and $1-\alpha^*_{\text{aPR}} \geq 1-\alpha$.

Theorem 2. *(Probabilistically robust coverage of aPRCP for cross-domain noise) Let $\mathcal{P}_\epsilon^{test}$ and $\mathcal{P}_\epsilon^{cal}$ denote different distributions of ϵ during the testing and calibration phases, respectively. Assume $\mathbb{P}_{\epsilon\sim\mathcal{P}_\epsilon^{cal}}\{\epsilon\} - \mathbb{P}_{\epsilon\sim\mathcal{P}_\epsilon^{test}}\{\epsilon\} \leq d$ for all $\|\epsilon\| \leq r$. Set $\alpha^*_{aPR} = 1 - d - (1-\alpha)/(1-\alpha+s)$ in (9). Let $\mathcal{C}^{aPR}(\widetilde{X} = X + \epsilon) = \{y \in \mathcal{Y} : S(\widetilde{X}, y) \leq \tau^{aPR}(\alpha; s)\}$ be the prediction set for a testing sample $\widetilde{X}$. Then aPRCP achieves $(1-\alpha)$-probabilistically robust coverage.*

Remark 2. The key assumption we make is $\mathbb{P}_{\epsilon\sim\mathcal{P}_\epsilon^{cal}}\{\epsilon\} - \mathbb{P}_{\epsilon\sim\mathcal{P}_\epsilon^{test}}\{\epsilon\} \leq d$, which is analogous to L^1-distance used in the domain adaptation literature [Redko et al., 2020, Ben-David et al., 2006]. One can interpret it as the maximal gap of the density probability between the calibration and testing distributions when fixing ϵ. As per our analysis, when this gap can be bounded by a sufficiently small constant d, with an inflated nominated coverage in the robust quantile (i.e., setting $\alpha^*_{\text{aPR}} = 1 - d - (1-\alpha)/(1-\alpha+s)$ in (9)), we can guarantee probabilistically robust coverage for aPRCP.

3.3 CONNECTION BETWEEN ARCP AND PRCP

Although ARCP algorithm can achieve adversarially robust coverage, we can still connect ARCP and PRCP in the sense of *probabilistically robust coverage* and understand their performance in terms of *efficiency*. Recall that efficiency of conformal prediction algorithms refers to the measured size of prediction sets for testing samples when some desired coverage is achieved. For example, for the same target probability $1-\alpha$, a smaller threshold indicates better efficiency. The following result shows the possibly improved

efficiency of iPRCP and aPRCP when compared to ARCP after that their hyper-parameters were tuned properly (i.e., η for iPRCP and s for aPRCP).

Corollary 3. *To achieve the same $(1-\alpha)$-probabilistically robust coverage on Z, the following inequalities hold:*

$$\min_{\eta\in[0,\alpha]} \tau^{iPR}(\alpha;\eta) \leq \tau^{AR}(\alpha), \quad \min_{s\in[0,\alpha]} \tau^{aPR}(\alpha;s) \leq \tau^{AR}(\alpha).$$

When all three algorithms achieve $(1-\alpha)$-probabilistically robust coverage, smaller thresholds yield better efficiency, i.e., iPRCP and aPRCP. The idea of the above result is to particularly set $\eta = 0$ and $s = 0$, which makes iPRCP and aPRCP degenerate to ARCP, resulting in the same threshold. For aPRCP with $s = 0$, we have $\alpha^*_{\text{aPR}} = 0$, i.e., 1-robust quantile for each (X, Y) used, which recovers ARCP.

3.4 APPROXIMATION ERROR OF EMPIRICAL QUANTILES

In the above sections, we presented algorithms and their analysis directly in the population sense, including the true quantile $Q(\alpha)$ and $Q^{\text{rob}}(X;\alpha)$. However, when executing a given conformal prediction method on exchangeable samples $\mathcal{D}_{\text{cal}}$, we employ empirical quantiles in practice. To close this gap between theory and practice, we additionally discuss the concentration inequalities for empirical approximation to these quantities (i.e., the gap between empirical and true quantiles) as a function of the number of samples.

Proposition 3. *(Concentration inequality for quantiles) Let $Q(\alpha) = \max\{t : \mathbb{P}_V\{V \leq t\} \geq 1-\alpha\}$ be the true quantile of a random variable V given α, and $\widehat{Q}_n(\alpha) = V_{(\lceil n(1-\alpha)\rceil)}$ be the empirical quantile estimated by n randomly sampled set $\{V_1, ..., V_n\}_{i=1}^n$. Then with probability at least $1-\delta$, we have $\widehat{Q}_n(\alpha + \tilde{O}(1/\sqrt{n})) \leq Q(\alpha) \leq \widehat{Q}_n(\alpha - \tilde{O}(1/\sqrt{n}))$ where $\tilde{O}$ hides the logarithmic factor.*

The above result shows that more data samples from the underlying distribution for (X, Y) or ϵ will help in improving the approximation of empirical quantiles on score function S at a rate of $\tilde{O}(1/\sqrt{n})$, where n is number of samples. Note that we only use this proposition to fill the gap between empirical and true quantiles. Some prior work also studied similar concentration results [Vovk, 2012].

4 EXPERIMENTS AND RESULTS

In this section, we present the empirical evaluation of our proposed aPRCP algorithm along different dimensions.

4.1 EXPERIMENTAL SETUP

Classification Datasets. We consider three benchmark datasets for evaluation: CIFAR10 [Krizhevsky et al., 2009], CIFAR100 [Krizhevsky et al., 2009], and ImageNet [Deng et al., 2009] using the standard training and test split.

Deep Neural Network Models. We consider ResNet-110 [He et al., 2016] as the main model architecture for CIFAR10 and CIFAR100 and ResNet-50 for ImageNet in our experiments. We provide results on additional deep neural networks in the Appendix due to space constraints noting that we find similar patterns. We train each model using two different approaches : *1) Standard training:* The training is only performed using clean training examples; and *2) Gaussian augmented training:* The training procedure employs Gaussian augmented examples [Gendler et al., 2022] parameterized by a given standard deviation $\sigma = 0.125$.

Methods and Baselines. We consider two relevant state-of-the-art CP algorithms as our baselines. First, we employ `Vanilla CP` [Romano et al., 2020a] designed for clean input examples. Second, we use randomly smooth conformal prediction (`RSCP`) [Gendler et al., 2022] which is designed to handle worst-case adversarial examples. We employ the publicly available implementations of `Vanilla CP`[1] and `RSCP`[2] using the best settings suggested by their authors.

We consider different configurations of our proposed adaptive probabilistically robust CP (`aPRCP`) algorithm. `aPRCP`(worst-adv) refers to the configuration where the evaluation of `aPRCP` is performed over adversarial examples generated using an adversarial attack algorithm. `aPRCP`($\tilde{\alpha}$) refers to the configuration where the evaluation is performed over noisy examples with a bounded perturbation on the test data. We provide additional results using different values for $\tilde{\alpha}$ in the Appendix.

Adversarial Attack Algorithms. To generate adversarial examples, we employ the white-box `PGD` attack algorithm [Gendler et al., 2022] to evaluate `Vanilla CP` algorithm. For `RSCP` and `aPRCP (worst-adv)`, we employ an adapted `PGD` algorithm for smoothed classifiers as proposed in Salman et al. [2019]. We provide additional results using different adversarial algorithms in the Appendix.

Evaluation Methodology. We present all our experimental results for desired coverage as $(1-\alpha)$=90%. We report the average metrics (coverage and prediction set size) over 50 different runs for all datasets. We consider two different evaluation settings at the inference time as described below.

(a) **Probabilistic robustness evaluation**: We randomly sample $n_s = 128$ examples for each clean testing input: $X^j = X + \epsilon_j$ (j=1 to n_s), where $||\epsilon_j||_2 \leq r = 0.125$ for the CIFAR data and $||\epsilon_j||_2 \leq r = 0.25$ for the ImageNet data. For a better span during the sampling procedure for each clean testing input, we sample two perturbations ϵ_j for each $r^{(k)}$ in $0 < r^{(1)} < \cdots < r^{(k)} \leq r$ such that $\|\epsilon_j\|_2 = r^{(k)}$.

[1] `https://github.com/msesia/arc`
[2] `https://github.com/Asafgendler/RSCP`

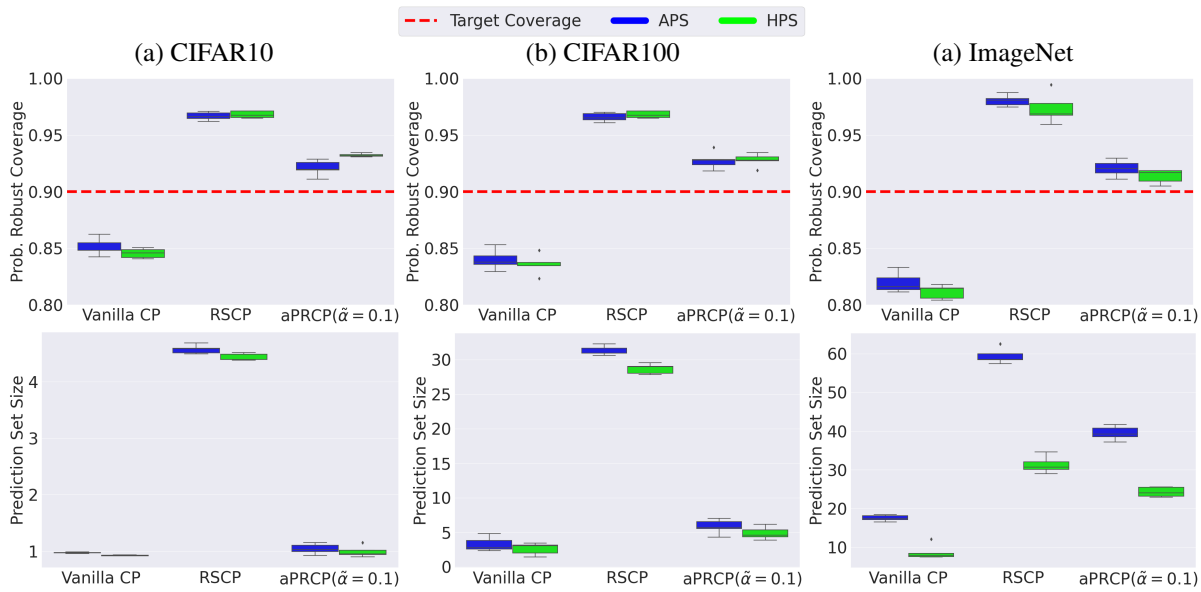

Figure 3: Probabilistic robust coverage (top) and prediction set size (bottom) constructed by `Vanilla CP`, `RSCP`, and `aPRCP`($\tilde{\alpha} = 0.1$) using HPS and APS scoreing functions (target coverage is 90%). Results are reported over 50 runs.

We define both coverage and prediction set size metrics to adapt to the probabilistic robustness setting as follows: *Coverage*: fraction of examples for which prediction set contains the ground-truth output.

$$\text{Coverage} = \frac{1}{n_s}\sum_{j=1}^{n_s} \mathbb{1}[Y_{n+1} \in \tilde{C}(X_{n+1} + \epsilon_j)]. \quad (10)$$

Efficiency: average prediction set size, small values mean high efficiency.

$$\text{Prediction Set Size} = \frac{1}{n_s}\sum_{j=1}^{n_s} |\tilde{C}(X_{n+1} + \epsilon_j)|, \quad (11)$$

where $||\epsilon_j||_2 \leq r = 0.125$ for CIFAR dataset, and $||\epsilon_j||_2 \leq r = 0.25$ for the ImageNet dataset. These re-defined metrics allow us to evaluate `aPRCP`($\tilde{\alpha}$) with different values of probability parameters $\tilde{\alpha}$ for probabilistic robustness. We provide additional results explaining the impact of the choice of the sampling distributions in the Appendix.

(b) **Worst-case evaluation:** We employ adversarial attack algorithms as mentioned above to create one worst-case adversarial example ($\tilde{X}$) for each clean testing input (X). We define both metrics for this setting as follows:

$$\text{Coverage} = \mathbb{1}[Y_{n+1} \in \tilde{C}(\tilde{X}_{n+1})]. \quad (12)$$

$$\text{Prediction Set Size} = |\tilde{C}(\tilde{X}_{n+1})|. \quad (13)$$

4.2 RESULTS AND DISCUSSION

Probabilistic Robust Coverage Performance. Figure 3 shows the probabilistic robustness performance (in terms of coverage and prediction set size) obtained by `Vanilla CP`, `RSCP`, and `aPRCP`($\tilde{\alpha} = 0.1$) for all three datasets using standard training. We make the following observations. 1) `Vanilla CP` algorithm fails in achieving the target probabilistic robust coverage. 2) `RSCP` algorithm achieves the desired probabilistic coverage, but has an empirical coverage significantly larger then 90%. This yields very large prediction sets. Using APS, `RSCP` yields on average a prediction set of 30 labels for CIFAR100 and 60 for ImageNet. 3) `aPRCP`($\tilde{\alpha} = 0.1$) produces smaller prediction sets by keeping the actual coverage close to the target coverage. `aPRCP`($\tilde{\alpha} = 0.1$) reduces the prediction set by an average of 20 labels for CIFAR100 and ImageNet compared to RSCP method using any of the two non-conformity scores.

Adversarial Robust Coverage Performance. Figure 4 shows the robust coverage and prediction set size obtained by `Vanilla CP`, `RSCP`, and `aPRCP`(worst-adv) achieved on the worst-case examples for three different datasets using Gaussian augmented training. We observe similar patterns as the probabilistic robust coverage results. 1) `Vanilla CP` fails to achieve the target coverage empirically. For all datasets, it achieves empirical coverage lower then 80%. 2) Similar to the probabilistic robustness results, `RSCP` method achieves an empirical coverage larger then 95% for all datasets, yielding significantly large prediction sets for all datasets. 3) `aPRCP`(worst-adv) produces smaller prediction sets by keeping the actual coverage close to the target coverage (by a margin of 2%) on worst-case adversarial examples. `aPRCP (worst-adv)` reduces the prediction set by more then 10 labels for CIFAR100 and ImageNet compared to RSCP method using any of the two non-conformity scores (HPS and APS).

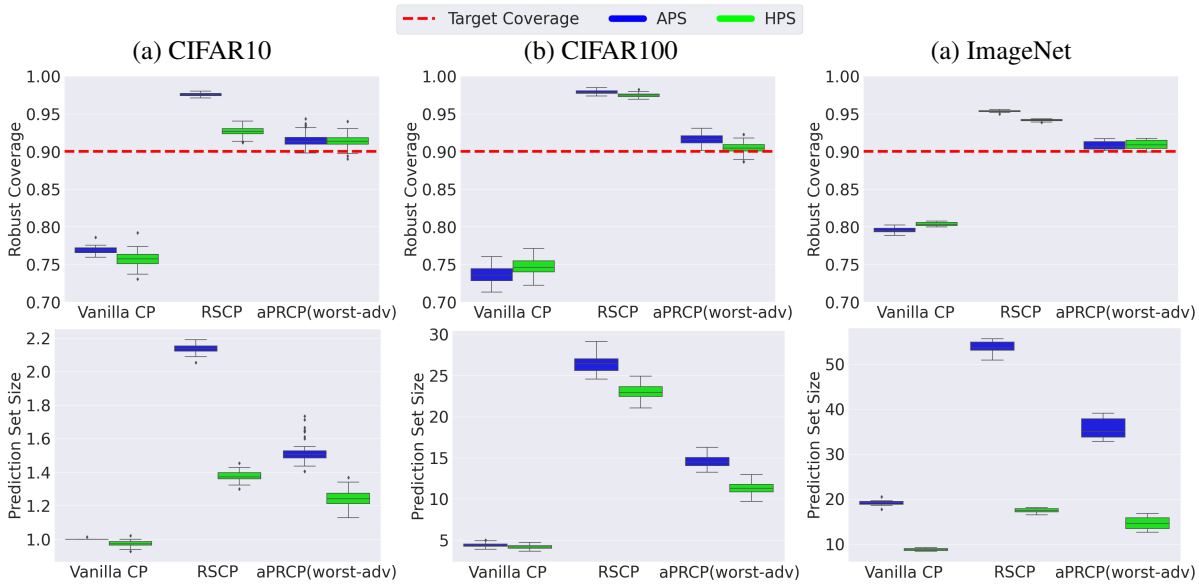

Figure 4: Adversarially robust coverage (top) and prediction set size (bottom) constructed by Vanilla CP, RSCP, and aPRCP(worst-adv) using HPS and APS scoring functions (target coverage is 90%). Results are reported over 50 runs.

5 RELATED WORK

Conformal Prediction. CP is a general framework for uncertainty quantification that provides marginal coverage guarantees without any assumptions on the underlying data distribution [Shafer and Vovk, 2008]. CP can be used for regression [Vovk et al., 2018, Lei et al., 2018, Romano et al., 2019, Izbicki et al., 2019, Guan, 2019, Gupta et al., 2022, Kivaranovic et al., 2020, Barber et al., 2021, Foygel Barber et al., 2021] to produce prediction intervals and for classification [Lei et al., 2013, Sadinle et al., 2019, Romano et al., 2020b, Angelopoulos et al., 2021] to produce prediction sets. Prior work has also considered instantiations of the CP framework to handle the differences between training and test distributions that is caused by long-term distribution shift [Gibbs and Candes, 2021], covariate shift[Tibshirani et al., 2019], and label-distribution shift [Podkopaev and Ramdas, 2021]. However, none of these existing works focus on the robustness setting where the distributional shift is caused by a bounded adversarial perturbation. While using adversarial training seems intuitive to mitigate this problem, it was shown that vanilla CP cannot achieve the target coverage on adversarial data [Gendler et al., 2022].

Robust Conformal Prediction. CP methods for robust coverage due to natural or adversarial perturbations is a new line of research that requires theoretical and empirical analysis. Very few works have proposed variants of CP to handle adversarial robust settings. The work on cautious deep learning [Hechtlinger et al., 2018] proposed a CP-based prediction set construction that accounts for adversarial examples. However, this method does not provide any theoretical guarantees. Recently, randomly smoothed conformal prediction (RSCP) [Gendler et al., 2022] was proposed as a generalization for adversarial examples using randomized smoothing. This generalization is achieved by introducing a constant inflation condition that adjusts the CP quantile to adversarial perturbations. This adjustment is proportional to the potential adversarial perturbations that can affect the test data. Hence, RSCP is prone to produce large prediction sets along with high marginal coverage to achieve robustness.

We study the general setting of probabilistically robust CP and develop probably correct algorithms to achieve improved trade-offs for nominal and robust performance over vanilla CP and RSCP. The key differences between our work (aPRCP) and RSCP are: 1) aPRCP uses a *quantile-of-quantile* design and does not require finding a score inflation constant like RSCP. 2) RSCP requires the design of a specialized scoring function while aPRCP can employ any existing score function. 3) aPRCP does not have test-time overhead unlike RSCP due to the generation of samples.

6 SUMMARY AND FUTURE WORK

This paper studied the novel problem of probabilistic robustness for conformal prediction (PRCP) based uncertainty quantification of deep classifiers. We developed the adaptive PRCP (aPRCP) algorithm based on the principle of quantile-of-quantile design and theoretically analyzed its effectiveness to achieve improved trade-offs between performance on clean data and robustness to adversarial examples. Our experiments on multiple image datasets using deep classifiers demonstrated the effectivess of aPRCP over vanilla CP methods and adversarially robust CP methods. Future work should study and analyze end-to-end PRCP algorithms.

References

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations(ICLR)*, 2021. URL `https://openreview.net/forum?id=eNdiU_DbM9`.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 2021.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Leying Guan. Conformal prediction with localization. *arXiv preprint arXiv:1908.08558*, 2019.

Leying Guan and Robert Tibshirani. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 84 (2):524, 2022.

Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Yotam Hechtlinger, Barnabás Póczos, and Larry Wasserman. Cautious deep learning. *arXiv preprint arXiv:1805.09460*, 2018.

Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

Rafael Izbicki, Gilson T Shimizu, and Rafael B Stern. Flexible distribution-free conditional predictive bands using density estimators. *arXiv preprint arXiv:1910.05575*, 2019.

Danijel Kivaranovic, Kory D Johnson, and Hannes Leeb. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics(AISTATS)*. PMLR, 2020.

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

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 2018.

Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3866–3876. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/li19g.html`.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. A unifying framework for combining complementary strengths of humans and ml toward better predictive decision-making. *arXiv preprint arXiv:2204.10806*, 2022.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

Alexander Robey, Luiz FO Chamon, George J Pappas, and Hamed Hassani. Probabilistically robust learning: Balancing average-and worst-case performance. *arXiv preprint arXiv:2202.01136*, 2022.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems(NeurIPS)*, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems(NeurIPS)*, volume 33, pages 3581–3591. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/244edd7e85dc81602b7615cd705545f5-Paper.pdf.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020b.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 2019.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*. PMLR, 2018.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.