PROVABLY ROBUST CONFORMAL PREDICTION WITH IMPROVED EFFICIENCY

Ge Yan
CSE, UCSD
`geyan@ucsd.edu`

Yaniv Romano
ECE, Technion
`yromano@technion.ac.il`

Tsui-Wei Weng
HDSI, UCSD
`lweng@ucsd.edu`

ABSTRACT

Conformal prediction is a powerful tool to generate uncertainty sets with guaranteed coverage using any predictive model, under the assumption that the training and test data are i.i.d.. Recently, it has been shown that adversarial examples are able to manipulate conformal methods to construct prediction sets with invalid coverage rates, as the i.i.d. assumption is violated. To address this issue, a recent work, Randomized Smoothed Conformal Prediction (RSCP), was first proposed to certify the robustness of conformal prediction methods to adversarial noise. However, RSCP has two major limitations: (i) its robustness guarantee is flawed when used in practice and (ii) it tends to produce large uncertainty sets. To address these limitations, we first propose a novel framework called `RSCP+` to provide provable robustness guarantee in evaluation, which fixes the issues in the original RSCP method. Next, we propose two novel methods, Post-Training Transformation (PTT) and Robust Conformal Training (RCT), to effectively reduce prediction set size with little computation overhead. Experimental results in CIFAR10, CIFAR100, and ImageNet suggest the baseline method only yields trivial predictions including full label set, while our methods could boost the efficiency by up to $4.36\times$, $5.46\times$, and $16.9\times$ respectively and provide practical robustness guarantee.

1 INTRODUCTION

Conformal prediction (Lei & Wasserman, 2014; Papadopoulos et al., 2002; Vovk et al., 2005) has been a powerful tool to quantify prediction uncertainties of modern machine learning models. For classification tasks, given a test input x_{n+1}, it could generate a prediction set $C(x_{n+1})$ with coverage guarantee:

$$\mathbb{P}[y_{n+1} \in C(x_{n+1})] \geq 1 - \alpha, \tag{1}$$

where y_{n+1} is the ground truth label and $1 - \alpha$ is user-specified target coverage. This property is desirable in safety-critical applications like autonomous vehicles and clinical applications. In general, it is common to set the coverage probability $1 - \alpha$ to be high, e.g. 90% or 95%, as we would like the ground truth label to be contained in the prediction set with high probability. It is also desired to have the smallest possible prediction sets $C(x_{n+1})$ as they are more informative. In this paper, we use the term "efficiency" to compare conformal prediction methods: we say a conformal prediction method is more efficient if the size of the prediction set is smaller.

Despite the power of conformal prediction, recent work (Gendler et al., 2021) showed that conformal prediction is unfortunately prone to adversarial examples – that is, the coverage guarantee in Eq. (1) may not hold anymore because adversarial perturbation on test data breaks the i.i.d. assumption and thus the prediction set constructed by vanilla conformal prediction becomes invalid. To solve this problem, Gendler et al. (2021) proposes a new technique, named Randomized Smoothed Conformal Prediction (RSCP), which is able to construct new prediction sets $C_\epsilon(\tilde{x}_{n+1})$ that is robust to adversarial examples:

$$\mathbb{P}[y_{n+1} \in C_\epsilon(\tilde{x}_{n+1})] \geq 1 - \alpha, \tag{2}$$

where $\tilde{x}_{n+1}$ denotes a perturbed example that satisfies $\|\tilde{x}_{n+1} - x_{n+1}\|_2 \leq \epsilon$ and $\epsilon > 0$ is the perturbation magnitude. The key idea of RSCP is to modify the vanilla conformal prediction procedure with randomized smoothing (Cohen et al., 2019; Duchi et al., 2012; Salman et al., 2019) so that the impact of adversarial perturbation could be bounded and compensated.

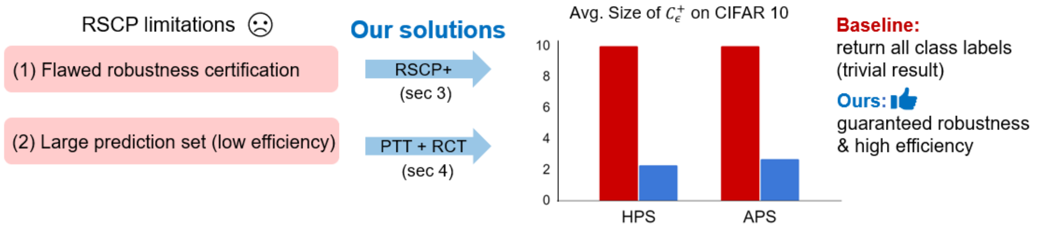


Figure 1: An overview of this work: We address two limitations of RSCP (Gendler et al., 2021) by proposing `RSCP+` (Sec. 3) & PTT + RCT (Sec. 4), which enables the first *provable* and *efficient* robust conformal prediction. As we show in the experiments in Sec. 5, our proposed method could provide useful robust prediction sets information while the baseline failed.

However, RSCP has two major limitations: (1) *the robustness guarantee of RSCP is flawed*: RSCP introduces randomized smoothing to provide robustness guarantee. Unfortunately, the derived guarantee is invalid when Monte Carlo sampling is used for randomized smoothing, which is how randomized smoothing is implemented in practice (Cohen et al., 2019). Therefore, their robustness certification is invalid, despite empirically working well. (2) *RSCP has low efficiency*: The average size of prediction sets of RSCP is much larger than the vanilla conformal prediction, as shown in our experiments (Fig. D.1).

In this paper, we will address these two limitations of RSCP to allow *efficient* and *provably robust* conformal prediction by proposing a new theoretical framework `RSCP+` in Sec. 3 to guarantee robustness, along with two new methods (PTT & RCT) in Sec. 4 to effectively decrease the prediction set size. We summarize our contributions below:

1. We first identify the major issue of RSCP in robustness certification and address this issue by proposing a new theoretical framework called `RSCP+`. The main difference between `RSCP+` and RSCP is that our `RSCP+` uses the Monte Carlo estimator directly as the base score for RSCP, and amends the flaw of RSCP with simple modification on the original pipeline. To our best knowledge, `RSCP+` is the first method to provide *practical certified robustness* for conformal prediction.
2. We further propose two methods to improve the efficiency of `RSCP+`: a scalable, training-free method called PTT and a general robust conformal training framework called RCT. Empirical results suggest PTT and RCT are necessary for providing guaranteed robust prediction sets.
3. We conduct extensive experiments on CIFAR10, CIFAR100 and ImageNet with `RSCP+`, PTT and RCT. Results show that without our method the baseline only gives trivial predictions, which are uninformative and useless. In contrast, our methods provide practical robustness certification and boost the efficiency of the baseline by up to $4.36\times$ on CIFAR10, $5.46\times$ on CIFAR100, and $16.9\times$ on ImageNet.

2 BACKGROUND AND RELATED WORKS

2.1 CONFORMAL PREDICTION

Suppose $D = \{(x_i, y_i)\}_{i=1}^n$ is an i.i.d. dataset, where $x_i \in \mathbb{R}^p$ denotes the features of ith sample and $y_i \in [K] := \{1, \ldots, K\}$ denotes its label. Conformal prediction method divides D into two parts: a training set $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^m$ and a calibration set $D_{\text{cal}} = D \setminus D_{\text{train}}$. The training set D_{train} is utilized to train a classifier function $\hat{\pi}(x) : \mathbb{R}^p \to [0, 1]^K$. Given classifier $\hat{\pi}$, a non-conformity score function $S(x, y) : \mathbb{R}^p \times [K] \to \mathbb{R}$ is defined for each class y based on classifier's prediction $\hat{\pi}(x)$. Next, the calibration set D_{cal} is utilized to calculate threshold τ, which is the $(1-\alpha)(1 + 1/|D_{\text{cal}}|)$ empirical quantile of calibration scores $\{S(x, y)\}_{(x,y)\in D_{\text{cal}}}$. Given a test sample x_{n+1}, conformal prediction construct a prediction set $C(x_{n+1}; \tau)$ as:

$$C(x_{n+1}; \tau) = \{k \in [K] \mid S(x_{n+1}, k) \leq \tau\}, \tag{3}$$

where

$$\tau = Q_{1-\alpha}(\{S(x,y)\}_{(x,y)\in D_{\text{cal}}}) \tag{4}$$

and $Q_p(D_{\text{cal}})$ denote the $p(1+1/|D_{\text{cal}}|)$-th empirical quantile of the calibration scores. In the remainder of the paper, we may omit the parameter τ and write the prediction set simply as $C(x)$ when the context is clear. Conformal prediction ensures the coverage guarantee in Eq. (1) by showing that the score corresponding to the ground truth label is bounded by τ with probability $1-\alpha$, i.e. $\mathbb{P}(S(x_{n+1}, y_{n+1}) \leq \tau) \geq 1-\alpha$.

Note that the above conformal prediction pipeline works for any non-conformity score $S(x,y)$, but the statistical efficiency of conformal prediction is affected by the choice of non-conformity score. Common non-conformity scores include HPS (Lei et al., 2013; Sadinle et al., 2019) and APS (Romano et al., 2020):

$$S_{\text{HPS}}(x,y) = 1 - \hat{\pi}_y(x),\ S_{\text{APS}}(x,y) = \sum_{y'\in[K]} \hat{\pi}_{y'}(x)\mathbb{1}_{\{\hat{\pi}_{y'}(x) > \hat{\pi}_y(x)\}} + \hat{\pi}_y(x)\cdot u, \tag{5}$$

where u is a random variable sampled from a uniform distribution over $[0,1]$.

2.2 RANDOMIZED SMOOTHED CONFORMAL PREDICTION

To ensure the coverage guarantee still holds under adversarial perturbation, Gendler et al. (2021) proposed *Randomized Smoothed Conformal Prediction (RSCP)*, which defines a new non-conformity score $\tilde{S}$ that can construct new prediction sets that are robust against adversarial attacks. The key idea of RSCP is to consider the worst-case scenario that $\tilde{S}$ may be affected by adversarial perturbations:

$$\tilde{S}(\tilde{x}_{n+1}, y) \leq \tilde{S}(x_{n+1}, y) + M_\epsilon, \forall y \in [K], \tag{6}$$

where x_{n+1} denotes the clean example, $\tilde{x}_{n+1}$ denotes the perturbed example that satisfies $\|\tilde{x}_{n+1} - x_{n+1}\|_2 \leq \epsilon$ and M_ϵ is a non-negative constant. Eq. (6) indicates that the new non-conformity score $\tilde{S}$ on adversarial examples may be inflated, but fortunately the inflation can be bounded. Therefore, to ensure the guarantee in Eq. (2) is satisfied, the threshold τ in the new prediction set needs to be adjusted to τ_{adj} defined as $\tau_{\text{adj}} = \tau + M_\epsilon$ to compensate for potential adversarial perturbations, and then C_ϵ can be constructed as follows:

$$C_\epsilon(x; \tau_{\text{adj}}) = \{k \in [K] \mid \tilde{S}(x,k) \leq \tau_{\text{adj}}\}, \tag{7}$$

where x is any test example. From Eq. (6), the validity of C_ϵ could be verified by following derivation:

$$y_{n+1} \in C(x_{n+1}) \Rightarrow \tilde{S}(x_{n+1}, y_{n+1}) \leq \tau \Rightarrow \tilde{S}(\tilde{x}_{n+1}, y_{n+1}) \leq \tau_{\text{adj}} \Rightarrow y_{n+1} \in C_\epsilon(\tilde{x}_{n+1}). \tag{8}$$

Thus, the coverage guarantee in Eq. (2) is satisfied. To obtain a valid M_ϵ, Gendler et al. (2021) proposed to leverage randomized smoothing (Cohen et al., 2019; Duchi et al., 2012) to construct $\tilde{S}$. Specifically, define

$$\tilde{S}(x,y) = \Phi^{-1}\left[S_{\text{RS}}(x,y)\right] \text{ and } S_{\text{RS}}(x,y) = \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I_p)} S(x+\delta, y), \tag{9}$$

where δ is a Gaussian random variable, σ is the standard deviation of δ which controls the strength of smoothing, and $\Phi^{-1}(\cdot)$ is Gaussian inverse cdf. We call $S_{\text{RS}}(x,y)$ the randomized smoothed score from a base score $S(x,y)$, as $S_{\text{RS}}(x,y)$ is the smoothed version of $S(x,y)$ using Gaussian noise on the input x. Since Φ^{-1} is defined on the interval $[0,1]$, the base score S must satisfy $S(x,y) \in [0,1]$. One nice property from randomized smoothing (Cohen et al., 2019) is that it guarantees that $\tilde{S}$ is Lipschitz continuous with Lipschitz constant $\frac{1}{\sigma}$, i.e. $\frac{|\tilde{S}(\tilde{x}_{n+1}, y_{n+1}) - \tilde{S}(x_{n+1}, y_{n+1})|}{\|\tilde{x}_{n+1} - x_{n+1}\|_2} \leq \frac{1}{\sigma}$. Hence, we have

$$\|\tilde{x}_{n+1} - x_{n+1}\|_2 \leq \epsilon \Longrightarrow \tilde{S}(\tilde{x}_{n+1}, y_{n+1}) \leq \tilde{S}(x_{n+1}, y_{n+1}) + \frac{\epsilon}{\sigma}, \tag{10}$$

which is exactly Eq. (6) with $M_\epsilon = \frac{\epsilon}{\sigma}$. Therefore, when using $\tilde{S}$ in conformal prediction, the threshold should be adjusted by:

$$\tau_{\text{adj}} = \tau + \frac{\epsilon}{\sigma}. \tag{11}$$

3 CHALLENGE 1: ROBUSTNESS GUARANTEE

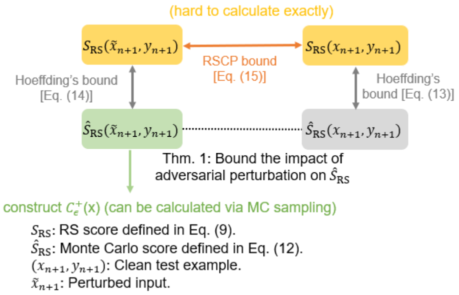
In this section, we point out a flaw in the robustness certification of RSCP (Gendler et al., 2021) and propose a new scheme called `RSCP+` to provide provable robustness guarantee in practice. As we discuss in Sec. 2.2, the key idea of RSCP is introducing a new conformity score $\tilde{S}$ that satisfies Eq. (10), which gives an upper bound to the impact of adversarial perturbation. However, in practice, $\tilde{S}$ is intractable due to expectation calculation in S_{RS}. A common practice in randomized smoothing literature is:

- **Step 1:** Approximate S_{RS} by Monte Carlo estimator:
$$\hat{S}_{\text{RS}}(x, y) = \frac{1}{N_{\text{MC}}} \sum_{i=1}^{N_{\text{MC}}} S(x + \delta_i, y), \delta_i \sim \mathcal{N}(0, \sigma^2 I_p). \tag{12}$$
- **Step 2:** Bound the estimation error via some concentration inequality.

In RSCP, however, **Step 2** is missing, because bounding the error simultaneously on the calibration set is difficult, as discussed in Appendix A.1. We argue that the missing error bound makes the robustness guarantee of RSCP invalid in practice.

To address this issue, we propose an elegant and effective approach, `RSCP+`, to fill in the gap and provide the guarantee. In particular, the intrinsic difficulty in bounding Monte Carlo error inspires us to avoid the estimation. Thus, in `RSCP+` we propose a new approach to incorporate the Monte Carlo estimator $\hat{S}_{\text{RS}}$ directly as the (non-)conformity score, which could be directly calculated, unlike S_{RS}. Here, one question may arise is: Can a randomized score (e.g. $\hat{S}_{\text{RS}}$) be applied in conformal prediction and maintain the coverage guarantee? The answer is yes: as we discuss in Appendix A.2, many classical (non-)conformity scores (e.g. APS (Romano et al., 2020)) are randomized scores, and the proofs for them are similar to the deterministic scores, as long as the i.i.d. property between calibration and test scores is preserved. Therefore, our $\hat{S}_{\text{RS}}$ is a legit (non-)conformity score.

The challenge of using $\hat{S}_{\text{RS}}$ is to derive an inequality similar to Eq. (10), i.e. connect $\hat{S}_{\text{RS}}(\tilde{x}_{n+1}, y)$ and $\hat{S}_{\text{RS}}(x_{n+1}, y)$ (the grey dotted line in Fig. 2), so that we can bound the impact from adversarial noises and compensate for it accordingly. To achieve this, we use S_{RS} as a bridge (as shown in Fig. 2), and present the result in Theorem 1.


(hard to calculate exactly)
$S_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1})$
$S_{\text{RS}}(x_{n+1}, y_{n+1})$
RSCP bound [Eq. (15)]
Hoeffding's bound [Eq. (14)]
Hoeffding's bound [Eq. (13)]
$\hat{S}_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1})$
$\hat{S}_{\text{RS}}(x_{n+1}, y_{n+1})$
Thm. 1: Bound the impact of adversarial perturbation on $\hat{S}_{\text{RS}}$
construct $C_\epsilon^+(x)$ (can be calculated via MC sampling)
S_{RS}: RS score defined in Eq. (9).
$\hat{S}_{\text{RS}}$: Monte Carlo score defined in Eq. (12).
(x_{n+1}, y_{n+1}): Clean test example.
$\tilde{x}_{n+1}$: Perturbed input.

Algorithm 1: `RSCP+`: Generate provably robust prediction set.

```
Input  : Input image $x_{n+1}$, calibration set $D_{cal}$, target
         coverage level $1 - \alpha$, $\epsilon$, $\sigma$, $\beta$
Output : Guaranteed robust prediction set $C_\epsilon^+$
Calculate calibration scores with randomized
  smoothing $\{\hat{S}_{\text{RS}}(x, y)\}_{(x,y) \in D_{\text{cal}}}$;
Calculate threshold $\tau_{\text{MC}}$ as $\frac{1-\alpha-2\beta}{1+1/|D_{\text{cal}}|}$ empirical quantile of
  calibration scores;
Calculate test conformity score for each class with randomized
  smoothing $\hat{S}_{\text{RS}}(\tilde{x}_{n+1}, k)_{k=1}^{K}$;
Calculate prediction set $C_\epsilon^+$ with Eq. (17).
```

Figure 2: Diagram illustrating our `RSCP+`. (Left) (1) The dotted line shows our target: bound Monte-Carlo estimator score $\hat{S}_{\text{RS}}$ under perturbation; (2) The orange arrow denotes the bound of the randomized smoothed score S_{RS} under perturbation, given by (Gendler et al., 2021); (3) The grey arrows denote Hoeffding's inequality connecting randomized smoothed score S_{RS} and Monte Carlo estimator score $\hat{S}_{\text{RS}}$. The target (1) could be derived by (2) + (3). (Right) `RSCP+` algorithm.

Theorem 1. *Let* (x_{n+1}, y_{n+1}) *be the clean test sample and* $\tilde{x}_{n+1}$ *be perturbed input data that satisfies* $\|\tilde{x}_{n+1} - x_{n+1}\|_2 \leq \epsilon$. *Then, with probability* $1 - 2\beta$*:*

$$\hat{S}_{RS}(\tilde{x}_{n+1}, y_{n+1}) - b_{Hoef}(\beta) \leq \Phi\left[\Phi^{-1}[\hat{S}_{RS}(x_{n+1}, y_{n+1}) + b_{Hoef}(\beta)] + \frac{\epsilon}{\sigma}\right],$$

where $b_{Hoef}(\beta) = \sqrt{\frac{-\ln\beta}{2N_{MC}}}$, N_{MC} *is the number of Monte Carlo examples,* Φ *is standard Gaussian cdf,* σ *is smoothing strength and* $\hat{S}_{RS}$ *is the Monte Carlo score defined in Eq.* (12).

Proof of Theorem 1. The main idea of the proof is connecting $\hat{S}_{\text{RS}}(x_{n+1}, y_{n+1})$ and $\hat{S}_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1})$ via the corresponding S_{RS}, as shown in Fig. 2. By Hoeffding's inequality (See Appendix A.3 for further discussion), we have

$$S_{\text{RS}}(x_{n+1}, y_{n+1}) \leq \hat{S}_{\text{RS}}(x_{n+1}, y_{n+1}) + b_{\text{Hoef}}(\beta) \tag{13}$$

by Eq. (A.8) and

$$S_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1}) \geq \hat{S}_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1}) - b_{\text{Hoef}}(\beta) \tag{14}$$

by Eq. (A.9), both with probability $1 - \beta$. Meanwhile, by plugging in the definition of $\tilde{S}$, Eq. (10) is equivalent to

$$\Phi^{-1}[S_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1})] \leq \Phi^{-1}[S_{\text{RS}}(x_{n+1}, y_{n+1})] + \frac{\epsilon}{\sigma}. \tag{15}$$

Combining the three inequalities above and applying union bound gives:

$$\begin{aligned} & S_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1}) \leq \Phi\left[\Phi^{-1}[S_{\text{RS}}(x_{n+1}, y_{n+1})] + \frac{\epsilon}{\sigma}\right] \\ \xrightarrow[\text{with prob. } 1-\beta]{\text{Eq. (13)}} & S_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1}) \leq \Phi\left[\Phi^{-1}[\hat{S}_{\text{RS}}(x_{n+1}, y_{n+1}) + b_{\text{Hoef}}] + \frac{\epsilon}{\sigma}\right] \\ \xrightarrow[\text{with prob. } 1-2\beta]{\text{Eq. (14)}} & \hat{S}_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1}) - b_{\text{Hoef}}(\beta) \leq \Phi\left[\Phi^{-1}[\hat{S}_{\text{RS}}(x_{n+1}, y_{n+1}) + b_{\text{Hoef}}(\beta)] + \frac{\epsilon}{\sigma}\right], \end{aligned} \tag{16}$$

with probability $1 - 2\beta$, which proves Theorem 1. □

Remark. *The bound in Theorem 1 could be further improved using Empirical Bernstein's inequality (Maurer & Pontil, 2009). We found in our experiments that the improvement is light on CIFAR10 and CIFAR100, but could be significant on ImageNet. For more discussion see Appendix A.3.3.*

With Theorem 1, we could construct the prediction set accordingly and derive the robustness guarantee in Corollary 2 in the following.

Corollary 2. *(Robustness guarantee for `RSCP+`) The `RSCP+` prediction set*

$$C_\epsilon^+(\tilde{x}_{n+1}; \tau_{MC}) = \left\{k \in [K] \mid \hat{S}_{RS}(\tilde{x}_{n+1}, k) - b_{Hoef}(\beta) \leq \Phi\left[\Phi^{-1}[\tau_{MC} + b_{Hoef}(\beta)] + \frac{\epsilon}{\sigma}\right]\right\} \tag{17}$$

satisfies robust coverage guarantee in Eq. (2)*, i.e.* $\mathbb{P}(y_{n+1} \in C_\epsilon^+(\tilde{x}_{n+1}; \tau_{MC})) \geq 1 - \alpha$. *Here, the threshold* τ_{MC} *is calculated according to Eq.* (4) *with* $S = \hat{S}_{RS}$ *and* $1 - \alpha$ *replaced by* $1 - \alpha + 2\beta$, *i.e.* $\tau_{MC} = Q_{1-\alpha+2\beta}(\{\hat{S}_{RS}(x, y)\}_{(x,y)\in D_{cal}})$.

Proof of Corollary 2. Since we have $\tau_{\text{MC}} = Q_{1-\alpha+2\beta}(\{\hat{S}_{\text{RS}}(x, y)\}_{(x,y)\in D_{\text{cal}}})$, conformal prediction guarantees coverage on clean examples:

$$\mathbb{P}[\hat{S}_{\text{RS}}(x_{n+1}, y_{n+1}) \leq \tau_{\text{MC}}] \geq 1 - \alpha + 2\beta. \tag{18}$$

Plug Eq. (18) into Eq. (16) in Theorem 1 and apply union bound, we get

$$\mathbb{P}\left\{\hat{S}_{\text{RS}}(\tilde{x}_{n+1}, y_{n+1}) - b_{\text{Hoef}}(\beta) \leq \Phi\left[\Phi^{-1}[\tau_{\text{MC}} + b_{\text{Hoef}}(\beta)] + \frac{\epsilon}{\sigma}\right]\right\} \geq 1 - \alpha. \tag{19}$$

□

4 CHALLENGE 2: IMPROVING EFFICIENCY

So far, we have modified RSCP to `RSCP+` that can provide a certified guarantee in Sec. 3. However, there exists another challenge – directly applying `RSCP+` often leads to trivial prediction sets that give the entire label set, as shown in our experiment Tabs. 1 and 2. The reason is that RSCP is *conservative*: instead of giving an accurate coverage as vanilla CP, RSCP attains a higher coverage due to its threshold inflation (Eq. (11)), and thus gives a larger prediction set on both clean and perturbed data. We define *conservativeness* of RSCP as the increase in the average size of prediction sets after threshold inflation: see Appendix A.4 where we give a formal definition. Since `RSCP+` is modified from RSCP, it's expected to inherit the conservativeness, leading to trivial predictions. To address this challenge and make `RSCP+` useful, in this section, we propose to address this problem by modifying the base score S with two new methods: Post Training Transformation (PTT) and Robust Conformal Training (RCT).

4.1 POST-TRAINING TRANSFORMATION (PTT)

Intuition. We first start with a quantitative analysis of the conservativeness by threshold inflation. As an approximation to the conservativeness, we measure the coverage gap between inflated coverage $1-\alpha_{\text{adj}}$ and target coverage $1-\alpha$:

$$\alpha_{\text{gap}} = (1-\alpha_{\text{adj}}) - (1-\alpha) = \alpha - \alpha_{\text{adj}}. \tag{20}$$

Next, we conduct a theoretical analysis on α_{gap}. Let $\Phi_{\tilde{S}}(t)$ be the cdf of score $\tilde{S}(x,y)$, where $(x,y) \sim P_{xy}$. For simplicity, suppose $\Phi_{\tilde{S}}(t)$ is known. Recall that in conformal prediction, the threshold τ is the minimum value that satisfies the coverage condition:

$$\tau = \operatorname*{argmin}_{t\in\mathbb{R}} \left\{ \mathbb{P}_{(x,y)\sim P_{xy}}[\tilde{S}(x,y) \leq t] \geq (1-\alpha). \right\} \tag{21}$$

Notice that $\mathbb{P}_{(x,y)\sim P_{xy}}[\tilde{S}(x,y) \leq t]$ is exactly $\Phi_{\tilde{S}}(t)$, we have:

$$\Phi_{\tilde{S}}(\tau) = 1-\alpha. \tag{22}$$

Suppose the threshold is inflated as $\tau_{\text{adj}} = \tau + M_\epsilon$. Similarly, we could derive $1-\alpha_{\text{adj}} = \Phi_{\tilde{S}}(\tau_{\text{adj}}) = \Phi_{\tilde{S}}(\tau + M_\epsilon)$ by Eq. (11). Now the coverage gap α_{gap} can be computed as:

$$\alpha_{\text{gap}} = \alpha - \alpha_{\text{adj}} = \Phi_{\tilde{S}}(\tau + M_\epsilon) - \Phi_{\tilde{S}}(\tau) \approx \Phi'_{\tilde{S}}(\tau) \cdot M_\epsilon \tag{23}$$

The last step is carried out by the linear approximation of $\Phi_{\tilde{S}}$: $g(x+z) - g(x) \approx g'(x) \cdot z$.

Key idea. Eq. (23) suggests that we could reduce α_{gap} by **reducing the slope** of $\Phi_{\tilde{S}}$ near the original threshold τ, i.e. $\Phi'_{\tilde{S}}(\tau)$. This inspires us to the idea: can we perform a transformation on $\tilde{S}$ to reduce the slope while keeping the information in it? Directly applying transformation on $\tilde{S}$ is not a valid option because it would break the Lipschitz continuity of $\tilde{S}$ in Eq. (10): for example, applying a discontinuous function on $\tilde{S}$ may make it discontinuous. However, we could apply a transformation $\mathcal{Q}$ on the base score S, which modifies $\tilde{S}$ indirectly while preserving the continuity, as long as the transformed score, $\mathcal{Q} \circ S$, still lies in the interval $[0,1]$. The next question is: how shall we design this transformation $\mathcal{Q}$? Here, we propose that the desired transformation $\mathcal{Q}$ should satisfy the following two conditions:

1. **(Slope reduction)** By applying $\mathcal{Q}$, we should reduce the slope $\Phi'_{\tilde{S}}(\tau)$, thus decrease the coverage gap α_{gap}. Since we are operating on base score S, we approximate this condition by reducing the slope $\Phi'_{S}(\tau)$. We give a rigorous theoretical analysis of a synthetic dataset and an empirical study on real data to justify the effectiveness of this approximation in Appendices B.6 and B.7, respectively.
2. **(Monotonicity)** $\mathcal{Q}$ should be monotonically non-decreasing. It could be verified that under this condition, $(\mathcal{Q} \circ S)$ is equivalent to S in vanilla CP (See our proof in Appendix B.5). Hence, the information in S is kept after transformation $\mathcal{Q}$.

These two conditions ensure that transformation $\mathcal{Q}$ could alleviate the conservativeness of RSCP without losing the information in the original base score. With the above conditions in mind, we design a two-step transformation $\mathcal{Q}$ by composing **(I)** ranking and **(II)** Sigmoid transformation on base score S, denoted as $\mathcal{Q} = \mathcal{Q}_{\text{sig}} \circ \mathcal{Q}_{\text{rank}}$. We describe each transformation below.

Transformation (I): ranking transformation $\mathcal{Q}_{\text{rank}}$. The first problem we encounter is that we have no knowledge about the score distribution Φ_S in practice, which makes designing transformation difficult. To address this problem, we propose a simple data-driven approach called ranking transformation to turn the unknown distribution Φ_S into a uniform distribution. With this, we could design the following transformations on it and get the analytical form of the final transformed score distribution $\Phi_{\mathcal{Q}\circ S}$. For ranking transformation, we sample an i.i.d. holdout set $D_{\text{holdout}} = \{(x_i, y_i)\}_{i=1}^{N_{\text{holdout}}}$ from P_{XY}, which is disjoint with the calibration set D_{cal}. Next, scores $\{S(x,y)\}_{(x,y)\in D_{\text{holdout}}}$ are calculated on the holdout set and the transformation $\mathcal{Q}_{\text{rank}}$ is defined as:

$$\mathcal{Q}_{\text{rank}}(s) = \frac{r\left[s; \{S(x,y)\}_{(x,y)\in D_{\text{holdout}}}\right]}{|D_{\text{holdout}}|}.$$

Here, $r(x; H)$ denotes the rank of x in set H, where ties are broken randomly. We want to emphasize that this rank is calculated on the holdout set D_{holdout} for both calibration samples and test samples. We argue that the new score $\mathcal{Q}_{\text{rank}} \circ S$ is uniformly distributed,which is a well-known result in statistics(Kuchibhotla, 2020). See more discussion in Appendix B.3.

Transformation (II): Sigmoid transformation $\mathcal{Q}_{\text{sig}}$. After ranking transformation, we get a uniformly distributed score. The next goal is reducing $\Phi'_S(\tau)$. For this, we introduce Sigmoid transformation $\mathcal{Q}_{\text{sig}}$. In this step, a sigmoid function ϕ is applied on S:

$$\mathcal{Q}_{\text{sig}}(s) = \phi\left[(s - b)/T\right],$$

where b, T are hyper-parameters controlling this transformation. Due to space constraint, we discuss more details of Sigmoid transformation in Appendix B.4, where we show that the distribution of transformed score $\Phi_{\mathcal{Q}_{\text{sig}} \circ \mathcal{Q}_{\text{rank}} \circ S}$ is the inverse of Sigmoid transformation $\mathcal{Q}_{\text{sig}}^{-1}$ (Eq. (B.2)), and by setting $b = 1 - \alpha$ and T properly small, the Sigmoid transformation could reduce $\Phi'_S(\tau)$.

Summary. Combining ranking transformation and sigmoid transformation, we derive a new (non-)conformity score S_{PTT}:

$$S_{\text{PTT}}(x, y) = (\mathcal{Q}_{\text{sig}} \circ \mathcal{Q}_{\text{rank}} \circ S)(x, y). \tag{24}$$

It could be verified that $S_{\text{PTT}}(x, y) \in [0, 1]$ for any S thanks to the sigmoid function, hence we could plug in $S \leftarrow S_{\text{PTT}}(x, y)$ into Eq. (9) as a base score. Additionally, $S_{\text{PTT}}(x, y)$ is monotonically non-decreasing, satisfying the monotonicity condition described at the beginning of this section. We provide a rigorous theoretical study on PTT over on a synthetic dataset in Appendix B.7. Additionally, we craft a case in Appendix B.8 where PTT may not improve the efficiency. Despite this theoretical possibility, we observe that PTT consistently improves over the baseline in experiments.

4.2 ROBUST CONFORMAL TRAINING (RCT)

While our proposed PTT provides a training-*free* approach to improve efficiency, there is another line of work (Einbinder et al., 2022b; Stutz et al., 2021) studying how to train a better base classifier for conformal prediction. However, these methods are designed for standard conformal prediction instead of *robust* conformal prediction considered in our paper. In this section, we introduce a training pipeline called RCT, which simulates the RSCP process in training to further improve the efficiency of robust conformal prediction.

Conformal training. Stutz et al. (2021) proposed a general framework to train a classifier for conformal prediction. It simulates conformal prediction in training by splitting the training batch B into a calibration set B_{cal} and a prediction set B_{pred}, then performing conformal prediction on them. The key idea is to use soft surrogate τ^{soft} and $c(x, y; \tau^{\text{soft}})$ to approximate the threshold τ and prediction set $C(x; \tau)$, making the pipeline differentiable: $\tau^{\text{soft}} = Q^{\text{soft}}_{1-\alpha}(\{S_\theta(x, y)\}_{(x,y)\in B_{\text{cal}}})$, where $Q^{\text{soft}}_q(H)$ denotes the $q(1 + \frac{1}{|H|})$-quantile of set H derived by smooth sorting (Blondel et al., 2020; Cuturi et al., 2019), and $c(x, y; \tau^{\text{soft}}) = \phi\left[\frac{\tau^{\text{soft}} - S_\theta(x,y)}{T_{\text{train}}}\right]$, where $\phi(z) = 1/(1 + e^{-z})$ is the sigmoid function and temperature T_{train} is a hyper-parameter. We introduce more details in Appendix C.1.

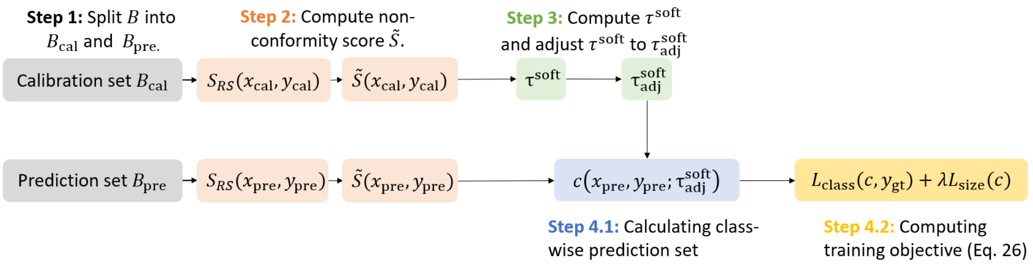

Step 1: Split B into B_{cal} and B_{pre}.
Step 2: Compute non-conformity score $\tilde{S}$.
Step 3: Compute τ^{soft} and adjust τ^{soft} to $\tau^{\text{soft}}_{\text{adj}}$
Calibration set B_{cal}
$S_{RS}(x_{\text{cal}}, y_{\text{cal}})$
$\tilde{S}(x_{\text{cal}}, y_{\text{cal}})$
τ^{soft}
$\tau^{\text{soft}}_{\text{adj}}$
Prediction set B_{pre}
$S_{RS}(x_{\text{pre}}, y_{\text{pre}})$
$\tilde{S}(x_{\text{pre}}, y_{\text{pre}})$
$c(x_{\text{pre}}, y_{\text{pre}}; \tau^{\text{soft}}_{\text{adj}})$
$L_{\text{class}}(c, y_{\text{gt}}) + \lambda L_{\text{size}}(c)$
Step 4.1: Calculating class-wise prediction set
Step 4.2: Computing training objective (Eq. 26)

Figure 3: Pipeline of our proposed Robust Conformal Training (RCT) method.

Incorporating RSCP into training. Inspired by Stutz et al. (2021), we propose to incorporate RSCP (Gendler et al., 2021) (and of course, `RSCP+` since the major steps are the same) into the training stage as shown in Fig. 3. We adopt soft threshold τ^{soft} and soft prediction $c(x, y; \tau^{\text{soft}})$ from Stutz et al. (2021), and add randomized smoothing $\tilde{S}$ and threshold adjustment $\tau^{\text{soft}}_{\text{adj}} = \tau^{\text{soft}} + \frac{\epsilon}{\sigma}$ to the pipeline as in RSCP. Next, we need to examine the differentiability of our pipeline. The threshold adjustment and Gaussian inverse cdf Φ^{-1} step in the calculation of $\tilde{S}$ is differentiable, but the gradient of $S_{\text{RS}} = \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I_p)} S(x+\delta, y)$ is difficult to evaluate, as the calculation of $S(x, y)$ involves a deep neural network and expectation. Luckily, several previous works (Salman et al., 2019; Zhai et al., 2020) have shown that the Monte-Carlo approximation works well in practice:

$$\nabla_\theta \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I_p)} S(x+\delta, y) \approx \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \nabla_\theta S(x+\delta_i, y). \tag{25}$$

With these approximations, the whole pipeline becomes differentiable and training could be performed by back-propagation. For the training objective, we can use the same loss function:

$$L(x, y_{\text{gt}}) = L_{\text{class}}(c(x, y; \tau^{\text{soft}}), y_{\text{gt}}) + \lambda L_{\text{size}}(c(x, y; \tau^{\text{soft}})), \tag{26}$$

where classification loss $L_{\text{class}}(c(x, y; \tau^{\text{soft}}), y_{\text{gt}}) = 1 - c(x, y_{\text{gt}}; \tau^{\text{soft}})$, size loss $L_{\text{size}}(c(x, y; \tau^{\text{soft}})) = \max(0, \sum_{y=1}^{K} c(x, y; \tau^{\text{soft}}) - \kappa)$, y_{gt} denotes the ground truth label, $c(x, y; \tau^{\text{soft}})$ denotes the soft prediction introduced in Stutz et al. (2021), κ is a hyper-parameter.

Remark. *Since the methods we proposed in Sec. 4 (PTT and RCT) are directly applied to base scores, they are orthogonal to the `RSCP+` we proposed in Sec. 3. That is to say, PTT and RCT not only work on `RSCP+` but also work on original RSCP as well. Nevertheless, we argue that `RSCP+` with PTT/RCT would be more desirable in practice since it provides* ***guaranteed robustness*** *which is the original purpose of provable robust conformal prediction. Hence, we will focus on this benchmark in the experiments section in the main text. However, we also provide experiment results on RSCP + PTT/RCT as an empirical robustness benchmark in Appendix D.2, which shows that our PTT and RCT are not limited to our `RSCP+` scheme.*

5 EXPERIMENTS

In this section, we evaluate our methods in Secs. 3 and 4. Experiments are conducted on CIFAR10, CIFAR100 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009) and target coverage is set to $1 - \alpha = 0.9$. We choose perturbation magnitude $\epsilon = 0.125$ on CIFAR10 and CIFAR100 and $\epsilon = 0.25$ on ImageNet.

Evaluation metrics and baseline. We present the average size of prediction sets $C^{+}_{\epsilon}(x)$ as a key metric, since the robustness is guaranteed by our theoretical results for `RSCP+`(Corollary 2). For the baseline, we choose the vanilla method from Gendler et al. (2021), where HPS and APS are directly applied as the base score without any modifications.

Model. We choose ResNet-110 (He et al., 2016) for CIFAR10 and CIFAR100 and ResNet-50 (He et al., 2016) for ImageNet. The pre-trained weights are from Cohen et al. (2019) for CIFAR10 and ImageNet and from Gendler et al. (2021) for CIFAR100.

Hyperparameters. In `RSCP+`, we choose $\beta = 0.001$ and the number of Monte Carlo examples $N_{\text{MC}} = 256$. For PTT, we choose $b = 0.9$ and $T = 1/400$ and we discuss this choice in Appendix B.4. The size of holdout set $|D_{\text{holdout}}| = 500$. We discuss more experimental details in Appendix D.

5.1 RESULTS AND DISCUSSION

Tab. 1 and Tab. 2 compare the average size of prediction sets on all three datasets with our `RSCP+` benchmark. Specifically, the first row shows the baseline method using base scores in Gendler et al. (2021) directly equipped with our `RSCP+`. Note that the baseline method gives trivial prediction sets (the prediction set size = total number of class, which is totally uninformative) due to its conservativeness. Our methods successfully address this problem and provide a meaningful prediction set with robustness guarantee.

Base score	HPS		APS	
Method / Dataset	CIFAR10	CIFAR100	CIFAR10	CIFAR100
Baseline (Gendler et al., 2021)	10	100	10	100
PTT **(Ours)**	**2.294**	26.06	**2.685**	21.96
PTT+RCT **(Ours)**	**2.294**	**18.30**	2.824	**20.01**
Improvement over baseline: PTT	**4.36**×	**3.84**×	**3.72**×	**4.55**×
Improvement over baseline: PTT + RCT	**4.36**×	**5.46**×	**3.54**×	**5.00**×

Table 1: **Average prediction set ($C_\epsilon^+(x)$) size of `RSCP+` on CIFAR10 and CIFAR100.** For CIFAR10 and CIFAR100, $\epsilon = 0.125$ and $\sigma = 0.25$. Following Gendler et al. (2021), we take $N_{\text{split}} = 50$ random splits between calibration set and test set and present the average results (Same for Tab. 2). We could see that the baseline method only gives trivial predictions containing the whole label set, while with PTT or PTT + RCT we can give informative and compact predictions.

Method / Base score	HPS	APS
Baseline (Gendler et al., 2021)	1000	1000
PTT **(Ours)**	1000	94.66
PTT + Bernstein **(Ours)**	**59.12**	**70.87**
Improvement over baseline: PTT	-	**10.6**×
Improvement over baseline: PTT + Bernstein	**16.9**×	**14.1**×

Table 2: **Average prediction set ($C_\epsilon^+(x)$) size of `RSCP+` on ImageNet.** For ImageNet, $\epsilon = 0.25$ and $\sigma = 0.5$. The ImageNet dataset is more challenging and our PTT only works for APS score, but we find by applying the improvement with Empirical Bernstein's bound (denoted as "PTT + Bernstein") we discussed in Appendix A.3.3, we could largely reduce the size of prediction sets.

N_{MC}	256	512	1024	2048	4096
Average size of prediction sets $C_\epsilon^+(x)$	2.294	2.094	1.954	1.867	1.816

Table 3: **Average size vs. Number of Monte Carlo samples N_{MC}.** The experiment is conducted on CIFAR10 dataset with PTT method. The base score is HPS. It could be seen that by increasing the number of Monte Carlo examples, we could further improve the efficiency of `RSCP+`, at the cost of higher computational expense.

Why the baseline gives trivial results under `RSCP+`? The key reason is conservativeness. RSCP is conservative compared to vanilla conformal prediction, and the challenging task of giving guaranteed robustness makes the situation worse. The result is that: without the boost of our PTT and RCT methods, the predictor is so conservative that it gives the whole label set to guarantee robustness, which is not the goal of users. This again justifies the necessity of our methods.

Impact of number of Monte Carlo samples N_{MC}. In Tab. 3, we study how the number of Monte Carlo samples (N_{MC}) influences the average size. It could be observed that the average size decreases as more Monte Carlo samples are taken. This is expected as more Monte Carlo samples reduce the error and provide a tighter bound in Eqs. (13) and (14). Therefore, a trade-off between prediction set size and computation cost needs to be considered in practice, since increasing N_{MC} also significantly boosts the computation requirement.

6 CONCLUSION

This paper studies how to generate prediction sets that are robust to adversarial attacks. We point out that the previous method RSCP (Gendler et al., 2021) has two major limitations: flawed robustness certification and low efficiency. We propose a new theoretically sound framework called `RSCP+` which resolves the flaw in RSCP and provides a provable guarantee. We also propose a training-free and scalable method (PTT) and robust conformal training method (RCT) to significantly boost the efficiency of RSCP. We have conducted extensive experiments and the empirical results support our theoretical analysis. Experiments show that the baseline gives trivial prediction sets (all class labels), while our methods are able to provide meaningful prediction sets that boost the efficiency of the baseline by up to $4.36\times$ on CIFAR10, $5.46\times$ on CIFAR100, and $16.9\times$ on ImageNet.

REPRODUCIBILITY STATEMENT

We provide the training details of RCT, hyperparameters, and other details of our experiments in Appendix D. The code of our experiments will be released to the public upon acceptance.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022.

Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pp. 950–959. PMLR, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 248–255. Ieee, 2009.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Bat-Sheva Einbinder, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to label noise. *arXiv preprint arXiv:2209.14295*, 2022a.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *arXiv preprint arXiv:2205.05878*, 2022b.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, pp. 681–690. PMLR, 2023.

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*, pp. 770–778, 2016.

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

Arun Kumar Kuchibhotla. Exchangeability, conformal prediction, and rank tests. *arXiv preprint arXiv:2005.06095*, 2020.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

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

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

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

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 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.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

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

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. *arXiv preprint arXiv:2001.02378*, 2020.