Robust Yet Efficient Conformal Prediction Sets

Soroush H. Zargarbashi, M. Sadegh Akhondzadeh, Aleksander Bojchevski

CISPA HELMHOLTZ-ZENTRUM FÜR INFORMATIONSSICHERHEIT

University of Cologne

TLDR

What is it about?

Conformal Prediction: Distribution free prediction sets with guaranteed probability to cover the true label.

$$\Pr[\text{airplane} \in \mathcal{C}(\ \ )] \geq 1-\alpha$$

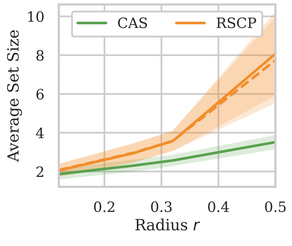
This guarantee can break for worst case noise.

$$\Pr[\text{airplane} \in \mathcal{C}(\ \ + 0.001\ \ )] \ngeq 1-\alpha$$

Robust CP: Recover the same guarantee for the worst case perturbed input.

Our Method

CAS: Returns smaller prediction sets with the same worst-case guarantee.



CAS in calibration-time robustness is also significantly faster (20x faster on ImageNet).

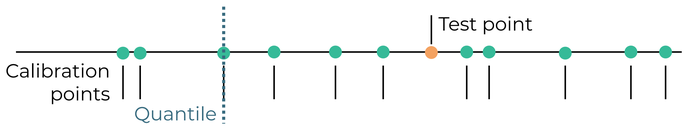
	Time (seconds)		No. Data
	Baseline	CAS (Ours)	
Calib.	0.15	0.79	204
Testing	2.93	0.15	100
Total	3.08	0.94	

Scan to view our code, and our paper.

Paper Code

Vanilla Conformal Prediction

A score function $s : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ e.g. model softmax, a holdout set of calibration datapoints $\mathcal{D}_{\text{cal}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with unseen labels.



Calibration. Define the threshold $q = \text{Quant}(\alpha; \{s(\boldsymbol{x}_i, y_i\}_{i=1}^n)$.

Prediction. Define the set $\mathcal{C}(\boldsymbol{x}_{n+1}) = \{y : s(\boldsymbol{x}_{n+1}, y) \geq q\}$.

Guarantee. If test and calibration are exchangeable then we have:

$$\Pr[y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{n+1})] \geq 1-\alpha$$

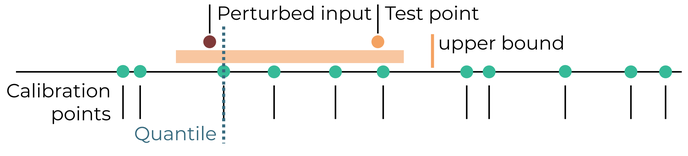
Robust Conformal Prediction

Bounded Perturbation. The set of all possible perturbations are given as $\mathcal{B}(\boldsymbol{x})$. Compute upper and lower bounds for the perturbed scores:

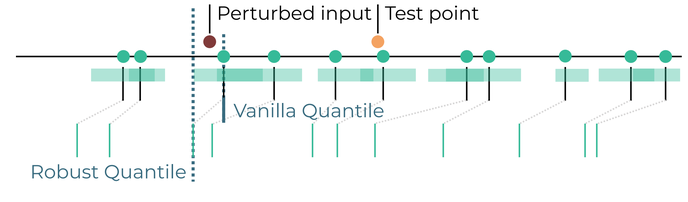
$$\forall \boldsymbol{x'} \in \mathcal{B}(\boldsymbol{x}) : \underline{s}(\boldsymbol{x}, y) \leq s(\boldsymbol{x'}, y) \leq \overline{s}(\boldsymbol{x}, y)$$

Test-time Robustness. We compare upper bound of the test score to the quantile of clean calibration set.

$$\overline{\mathcal{C}}(\tilde{\boldsymbol{x}}_{n+1}) = \{y : \overline{s}(\tilde{\boldsymbol{x}}_{n+1}, y) \geq q\} \quad q = \text{Quant}(\alpha; \{s(\boldsymbol{x}_i, y_i)\}_{i=1}^n)$$

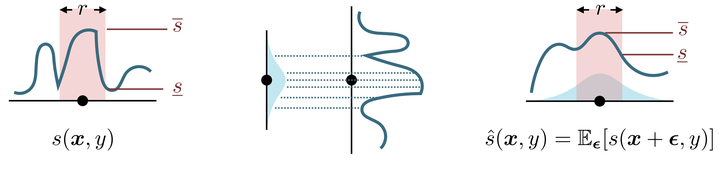


Calibration-time Robustness. We compare the test score to the quantile of lower bound calibration scores.



Bounds for **Randomized Smoothing**

Randomly Smoothed Scores. Blackbox bounds via smoothing.

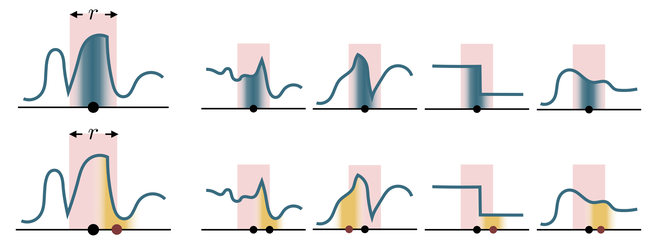


CDF Aware Sets (CAS)

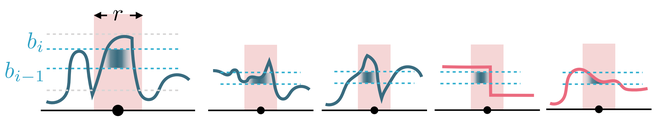
Bound Through Smoothing. We search over the space of all possible functions (due to black-box access).

$$\max_{\tilde{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{x})} \mathbb{E}[s(\tilde{\boldsymbol{x}}+\boldsymbol{\epsilon}, y)] \leq \max_{\tilde{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{x}), h \in \mathcal{H}} \mathbb{E}[h(\tilde{\boldsymbol{x}}+\boldsymbol{\epsilon}, y)]$$

Mean Constraint. We limit the search space to all functions with the same mean around the reference point. $\mathbb{E}[h(\boldsymbol{x}+\boldsymbol{\epsilon}, y)] = \mathbb{E}[s(\boldsymbol{x}+\boldsymbol{\epsilon}, y)]$

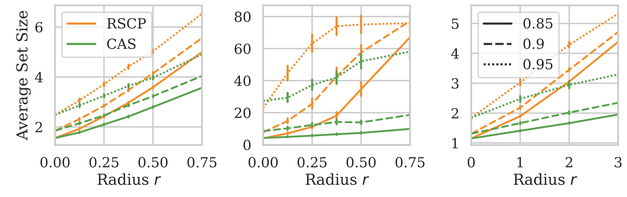


CDF Constraint. We limit the space to all functions with same CDF around the reference point. $\Pr[h(\boldsymbol{x}+\boldsymbol{\epsilon}, y) \leq b_i] = \Pr[s(\boldsymbol{x}+\boldsymbol{\epsilon}, y) \leq b_i]$

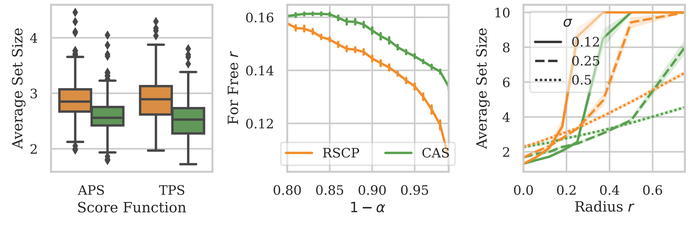


Results

CAS returns smaller prediction sets with the same guarantee.



This efficiency gain is across all radii, all smoothing magnitudes, and all score functions sets. It also results in larger "for free" radius — a radius for which the robust prediction set remains the same as vanilla.



We further propose robustness to poisoning on both feature and labels. These results can be combined with evasion.