# Probabilistically Robust Conformal Prediction (Supplementary material)

**Subhankar Ghosh**[1=] **Yuanjie Shi**[1=] **Taha Belkhouja**[1] **Yan Yan**[1] **Janardhan Rao Doppa**[1] **Brian Jones**[2]

[1]School of Electrical Engineering and Computer Science, Washington State University
[2]Proofpoint Inc.

## 1 TECHNICAL PROOFS

In this section, we prove the theoretical results in the main paper. To make it complete and self-contained, we also include the proof of Proposition 1, i.e., Theorem 1 in [Gendler et al., 2022], with the framework and notations used in our paper.

**Proposition 1.** *(Proposition 1 restated, 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.*

*Proof.* (of Proposition 1)

After reviewing the inflated quantile in the adversarial sense, we extend it to the following probabilistic sense.

$$\begin{aligned}\mathbb{P}_Z\{S(X+\epsilon,Y) \leq \tau^{\text{AR}}(\alpha)\} \geq& \mathbb{P}_Z\{S(X,Y) + M_r \leq \tau^{\text{AR}}(\alpha)\}\\ =& \mathbb{P}_Z\{S(X,Y) + M_r \leq Q(\alpha) + M_r\}\\ =& \mathbb{P}_Z\{S(X,Y) \leq Q(\alpha)\}\\ =& \mathbb{P}_{X,Y}\{S(X,Y) \leq Q(\alpha)\} \geq 1-\alpha,\end{aligned}$$

where the first inequality is due to the condition of $M_r$-adversarially inflated conformity score function (Definition 2), the first equality is due to the setting of the inflated threshold $\tau^{\text{AR}}(\alpha) = Q(\alpha) + M_r$, and the last inequality is due to the definition of quantile $Q(\alpha)$. □

**Proposition 2.** *(Proposition 2 restated, probabilistically robust coverage of iPRCP) Assume the score function $S$ is $M_{r,\eta}$-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.*

*Proof.* (of Proposition 2)

Denote $A_{r,\eta} = \{Z \in \mathcal{X} \times \mathcal{Y} \times \mathcal{E}_r : S(X+\epsilon,Y) \leq S(X,Y) + M_{r,\eta}\}$, which implies $\mathbb{P}_Z\{Z \in A_{r,\eta}\} \geq 1-\eta$. Recall $\tau^{\text{iPR}}(\alpha';\eta) = Q(\alpha') + M_{r,\eta}$ for $\alpha'$ and $\eta$.

$$\begin{aligned}&\mathbb{P}_Z\{S(X+\epsilon,Y) \leq \tau^{\text{iPR}}(\alpha';\eta)\}\\ =&\mathbb{P}\{Z \in A_{r,\eta}\} \cdot \mathbb{P}_Z\{S(X+\epsilon,Y) \leq \tau^{\text{iPR}}(\alpha';\eta) | Z \in A_{r,\eta}\}\\ &+ \mathbb{P}\{Z \notin A_{r,\eta}\} \cdot \mathbb{P}_Z\{S(X+\epsilon,Y) \leq \tau^{\text{iPR}}(\alpha';\eta) | Z \notin A_{r,\eta}\}\\ \geq&(1-\eta) \cdot \mathbb{P}_Z\{S(X+\epsilon,Y) \leq \tau^{\text{iPR}}(\alpha';\eta) | Z \in A_{r,\eta}\}\\ \geq&(1-\eta) \cdot \mathbb{P}_Z\{S(X,Y) + M_{r,\eta} \leq Q(\alpha') + M_{r,\eta} | Z \in A_{r,\eta}\}\\ =&(1-\eta) \cdot \mathbb{P}_{X,Y}\{S(X,Y) \leq Q(\alpha')\}\\ \geq&(1-\eta)(1-\alpha'),\end{aligned}$$

*Accepted for the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

where the first inequality is due to the non-negativity of probability and the definition of $A_{r,\eta}$, and the second inequality is due to $M_{r,\eta}$-probabilistically inflated score function (7).

In this case, define $\alpha^*_{\text{iPR}}(\alpha;\eta) := \max\{\alpha' : (1-\eta)(1-\alpha') \geq 1-\alpha\}$, and we can use $\tau^{\text{iPR}}(\alpha^*_{\text{iPR}}(\alpha;\eta);\eta)$ as the threshold to derive $(1-\alpha)$-probabilistically robust coverage. However, we have to know the conformity score function very well, so that we access the value of $M_{r,\eta}$ given $\eta$ to determine $\tau^*_{\text{iPR}}(\alpha;\eta)$, which is not always possible in practice. □

**Theorem 1.** *(Theorem 1 restated, 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.*

*Proof.* (of Theorem 1)

Denote $B = \{(X,Y) \in \mathcal{X}\times\mathcal{Y} : Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}}) \leq \tau^{\text{aPR}}(\alpha;s)\}$, which implies that

$$\mathbb{P}_{X,Y}\{(X,Y)\in B\} \geq 1-\alpha+s \tag{1}$$

due to the definition of $\tau^{\text{aPR}}(\alpha;s)$ in (9). We simply check whether $\tau^{\text{aPR}}(\alpha;s)$ can give us probabilistically robust coverage as follows:

$$\begin{aligned}
&\mathbb{P}_Z\{S(X+\epsilon,Y) \leq \tau^{\text{aPR}}(\alpha;s)\}\\
=&\mathbb{P}_{X,Y}\{X,Y : Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}}) \leq \tau^{\text{aPR}}(\alpha;s)\}\cdot\mathbb{P}_{\epsilon|X,Y}\{S(X+\epsilon,Y)\leq\tau^{\text{aPR}}(\alpha;s)\}\\
&+\mathbb{P}_{X,Y}\{X : Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}}) > \tau^{\text{aPR}}(\alpha;s)\}\cdot\mathbb{P}_{\epsilon|X,Y}\{S(X+\epsilon,Y)\leq\tau^{\text{aPR}}(\alpha;s)\}\\
\geq&\mathbb{P}_{X,Y}\{X,Y : Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}}) \leq \tau^{\text{aPR}}(\alpha;s)\}\cdot\mathbb{P}_{\epsilon|(X,Y)\in B}\{S(X+\epsilon,Y)\leq\tau^{\text{aPR}}(\alpha;s)\}\\
\geq&\mathbb{P}_{X,Y}\{(X,Y)\in B\}\cdot\mathbb{P}_{\epsilon|(X,Y)\in B}\{S(X+\epsilon,Y)\leq Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}})\}\\
\geq&(1-\alpha+s)\cdot\mathbb{P}_{\epsilon|(X,Y)\in B}\{S(X+\epsilon,Y)\leq Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}})\}\\
\geq&(1-\alpha+s)(1-\alpha^*_{\text{aPR}}),
\end{aligned} \tag{2}$$

where the first inequality is due to the non-negativity of probability, the second inequality is due to $Q^{\text{rob}}(X,Y;\alpha^{\text{aPR}}(\alpha)) \leq \tau^{\text{aPR}}(\alpha;s)$ for $(X,Y)\in B$, the third inequality is due to (1), and the last inequality is due to the definition of robust quantile $Q^{\text{rob}}(X,Y;\tilde{\alpha})$ in (8).

Recall $\alpha^*_{\text{aPR}} = 1-(1-\alpha)/(1-\alpha+s)$, so $(1-\alpha+s)(1-\alpha^*_{\text{aPR}}) = 1-\alpha$, which shows

$$\mathbb{P}_Z\{S(X+\epsilon,Y)\leq\tau^{\text{aPR}}(\alpha;s)\}\geq 1-\alpha.$$

□

**Lemma 1.** *(Inflated probability for cross domain noise) Assume $\mathbb{P}_{\epsilon\sim\mathcal{P}^{cal}_\epsilon}\{\epsilon\} - \mathbb{P}_{\epsilon\sim\mathcal{P}^{test}_\epsilon}\{\epsilon\} \leq d$ for all $\|\epsilon\| \leq r$. Then, for any threshold $\tau$, the following inequality holds:*

$$\mathbb{P}_{\epsilon\sim\mathcal{P}^{cal}_\epsilon|X,Y}\{S(X+\epsilon,Y)\leq\tau\} - \mathbb{P}_{\epsilon\sim\mathcal{P}^{test}_\epsilon|X,Y}\{S(X+\epsilon,Y)\leq\tau\} \leq d. \tag{3}$$

*Proof.* (of Lemma 1)

$$\begin{aligned}
&\mathbb{P}_{\epsilon\sim\mathcal{P}^{cal}_\epsilon}\{S(X+\epsilon,Y)\leq\tau\} - \mathbb{P}_{\epsilon\sim\mathcal{P}^{test}_\epsilon}\{S(X+\epsilon,Y)\leq\tau\}\\
=&\mathbb{E}_{\epsilon\sim\mathcal{P}^{cal}_\epsilon}[\mathbb{I}[S(X+\epsilon,Y)\leq\tau]] - \mathbb{E}_{\epsilon\sim\mathcal{P}^{test}_\epsilon}[\mathbb{I}[S(X+\epsilon,Y)\leq\tau]]\\
=&\int_\epsilon \mathbb{P}_{\epsilon\sim\mathcal{P}^{cal}_\epsilon}\{\epsilon\}\cdot\mathbb{I}[S(X+\epsilon,Y)\leq\tau]d\epsilon - \int_\epsilon \mathbb{P}_{\epsilon\sim\mathcal{P}^{test}_\epsilon}\{\epsilon\}\cdot\mathbb{I}[S(X+\epsilon,Y)\leq\tau]d\epsilon\\
=&\int_\epsilon \left(\mathbb{P}_{\epsilon\sim\mathcal{P}^{cal}_\epsilon}\{\epsilon\} - \mathbb{P}_{\epsilon\sim\mathcal{P}^{test}_\epsilon}\{\epsilon\}\right)\cdot\mathbb{I}[S(X+\epsilon,Y)\leq\tau]d\epsilon\\
\leq&\int_\epsilon (d\cdot 1)d\epsilon = d.
\end{aligned}$$

□

**Theorem 2.** *(Theorem 2 restated, probabilistically robust coverage of aPRCP for cross domain noise) Let $\mathcal{P}_\epsilon^{test}$ and $\mathcal{P}_\epsilon^{cal}$ denote different distributions of $\epsilon$ during the testing and calibration phase, 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 under $\mathcal{P}_\epsilon^{test}$.*

*Proof.* (of Theorem 2) We start with (2) in the proof of Theorem 2 which only considers the noise $\epsilon$ drawn from the same distribution during calibration and testing as follows.

$$\begin{aligned}
&\mathbb{P}_{X,Y,\epsilon\sim\mathcal{P}_\epsilon^{test}}\{S(X+\epsilon,Y) \leq \tau^{\text{aPR}}(\alpha;s)\}\\
\geq&(1-\alpha+s)\cdot\mathbb{P}_{\epsilon\sim\mathcal{P}_\epsilon^{test}|(X,Y)\in B}\{S(X+\epsilon,Y) \leq Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}})\}\\
\geq&(1-\alpha+s)\cdot\left(\mathbb{P}_{\epsilon\sim\mathcal{P}_\epsilon^{cal}|(X,Y)\in B}\{S(X+\epsilon,Y) \leq Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}})\}-d\right)\\
\geq&(1-\alpha+s)\cdot\left(1-\left(1-d-\frac{1-\alpha}{1-\alpha+s}\right)-d\right)\\
=&(1-\alpha+s)\cdot\frac{1-\alpha}{1-\alpha+s}=1-\alpha,
\end{aligned}$$

where the first inequality follows (2), the second inequality is due to inequality 3 in Lemma 1, and the third inequality is due to the definition $Q^{\text{rob}}(X,Y;\alpha^*_{\text{aPR}})$ in (8) with $\alpha^*_{\text{aPR}} = 1-d-(1-\alpha)/(1-\alpha+s)$. □

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

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

*Proof.* (of Corollary 3) For adaptive PRCP, if $s=0$, to achieve $(1-\alpha)$-probabilistically robust coverage over $Z$, we must have $\alpha^*_{\text{aPR}}=0$. Since $\alpha^*_{\text{aPR}}$ controls how aggressively we derive the robust quantile for $(X,Y)$, it indicates that we have to consider 1-robust quantile. This is equivalent to deriving the adversarial $S(X+\epsilon,Y)$ for all $(X,Y)$.

For inflated PRCP, if $\eta=0$, to achieve $(1-\alpha)$-probabilistically robust coverage, we have $M_{\delta,\eta}=M_\delta$ and $\alpha^*_{\text{iPR}}=\alpha$, recovering ARCP (adversarially robust conformal prediction). This case is exactly the same with adpative PRCP with $s=0$. Therefore, $\tau^{\text{AR}}(\alpha)=\tau^{\text{iPR}}(\alpha;0)=\tau^{\text{aPR}}(\alpha;0)$.

Note that $\min_{s\in[0,\alpha]}\tau^{\text{aPR}}(\alpha;s)\leq\tau^{\text{aPR}}(\alpha;0)$ and $\min_{\eta\in[0,\alpha]}\tau^{\text{iPR}}(\alpha;\eta)\leq\tau^{\text{iPR}}(\alpha;0)$, so by tuning the value of $s$ for aPRCP and the value of $\eta$ for iPRCP, to achieve the same probabilistically robust coverage $1-\alpha$, we can have a more efficient threshold than ARCP. □

**Proposition 3.** *(Proposition 3 restated, 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 $\alpha$, and $\widehat{Q}_n(\alpha)=V_{(\lceil(n+1)(1-\alpha)\rceil)}$ 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.*

*Proof.* (of Proposition 3)

Define $Z_i=\mathbb{I}[V_i\leq Q(\alpha)]$ where $1\leq i\leq n$ and $\mathbb{I}[\cdot]$ is an indicator function. Then $Z_i$ is a Bernoulli random variable with $\mathbb{P}\{Z_i=1\}=1-\alpha$ and $\mathbb{P}\{Z_i=0\}=\alpha$ from the definition of $Q(\alpha)$. Let $\widehat{Z}=\frac{1}{n}\sum_{i=1}^n Z_i$ and $\mathbb{E}[\widehat{Z}]=1-\alpha$.

According to Chernoff bound, we know

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n Z_i-\mathbb{E}[\widehat{Z}]\right|\geq\varepsilon\mathbb{E}[\widehat{Z}]\right\}\leq 2\exp\left(-\mathbb{E}[\widehat{Z}]\varepsilon^2/3\right)=2\exp\left(-n(1-\alpha)\varepsilon^2/3\right).$$

By setting $\delta=2\exp(-n(1-\alpha)\varepsilon^2/3)$, i.e., $\varepsilon=\sqrt{(3\log(2/\delta))/((1-\alpha)n)}$, we have with probability at least $1-\delta$:

$$\left|\frac{1}{n}\sum_{i=1}^n\mathbb{I}[V_i\leq Q(\alpha)]-(1-\alpha)\right|\leq\varepsilon(1-\alpha)=\sqrt{(3(1-\alpha)\log(2/\delta))/n}=\tilde{O}(1/\sqrt{n}). \tag{4}$$

Recall the definition of the empirical quantile $\widehat{Q}_n(\alpha)$ given $\alpha$:

$$\widehat{Q}_n(\alpha) = \max\left\{t : \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[V_i \le t] \ge 1-\alpha\right\}.$$

Then we know the following upper bound and lower bound for $1-\alpha$:

$$(1-\alpha) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[V_i \le \widehat{Q}_n(\alpha)], \quad (1-\alpha) \ge \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[V_i \le \widehat{Q}_n(\alpha+1/n)].$$

Re-arranging (4) and using the above upper/lower bounds, with probability at least $1-\delta$, we have

$$\begin{aligned}
&(1-\alpha)(1-\varepsilon) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[V_i \le Q(\alpha)] \le (1-\alpha)(1+\varepsilon) \\
\Leftrightarrow\ & 1-(\underbrace{1-(1-\alpha)(1-\varepsilon)}_{=\alpha'}) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[V_i \le Q(\alpha)] \le 1-(\underbrace{1-(1-\alpha)(1+\varepsilon)}_{=\alpha''}) \\
\Rightarrow\ & \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[V_i \le \widehat{Q}_n(\alpha'+1/n)] \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[V_i \le Q(\alpha)] \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[V_i \le \widehat{Q}_n(\alpha'')] \\
\Leftrightarrow\ & \widehat{Q}_n(\alpha'+1/n) \le Q(\alpha) \le \widehat{Q}_n(\alpha'').
\end{aligned}$$

Finally, we analyze $\alpha'$ and $\alpha''$ as follows

$$\begin{aligned}
\alpha' &= 1-(1-\alpha)(1-\varepsilon) = \alpha+\varepsilon(1-\alpha) = \alpha+\sqrt{3(1-\alpha)\log(2/\delta)/n} = \alpha+\tilde{O}(1/\sqrt{n}), \\
\alpha'' &= 1-(1-\alpha)(1+\varepsilon) = \alpha-\varepsilon(1-\alpha) = \alpha-\sqrt{3(1-\alpha)\log(2/\delta)/n} = \alpha-\tilde{O}(1/\sqrt{n}).
\end{aligned}$$

Therefore, we have

$$\widehat{Q}_n(\alpha+\tilde{O}(1/\sqrt{n})) \le Q(\alpha) \le \widehat{Q}_n(\alpha-\tilde{O}(1/\sqrt{n})).$$

□

## 2 ADDITIONAL EXPERIMENTS AND IMPLEMENTATION DETAILS

**Implementation details.** Table 1 shows the testing accuracy of the different deep models using both standard training ($\sigma = 0$) and Gaussian augmented training ($\sigma > 0$).

| Architecture | Training | CIFAR10 | | CIFAR100 | | ImageNet | |
|---|---|---|---|---|---|---|---|
| | | Clean(%) | Adv(%) | Clean(%) | Adv(%) | Clean(%) | Adv(%) |
| ResNet-110 | $\sigma = 0.0$ | 89.99 | 26.71 | 71.12 | 12.20 | - | - |
| | $\sigma = 0.125$ | 81.70 | 67.80 | 58.11 | 42.01 | - | - |
| VGG-19 | $\sigma = 0.0$ | 93.10 | 54.96 | 72.22 | 23.10 | - | - |
| | $\sigma = 0.125$ | 86.50 | 72.10 | 55.12 | 40.85 | - | - |
| DenseNet-161 | $\sigma = 0.0$ | 95.42 | 23.28 | 77.10 | 04.30 | - | - |
| | $\sigma = 0.125$ | 88.17 | 73.15 | 60.32 | 46.91 | - | - |
| ResNet-50 | $\sigma = 0.0$ | - | - | - | - | 75.69 | 19.56 |
| | $\sigma = 0.250$ | - | - | - | - | 68.62 | 56.15 |

Table 1: Testing accuracy of different deep models on clean and adversarial test examples (generated using the PGD attack algorithm) for all three data sets.

## 2.1 CASE OF SIMILAR NOISE DISTRIBUTION FOR BOTH CALIBRATION AND TESTING

**Performance evaluation with a fixed $s$ hyper-parameter and varying $\tilde{\alpha}$.** We present in Figures 1 and 2 the probabilistic robust coverage and prediction set size performance of aPRCP using *the Uniform distribution as a noise distribution for both calibration and testing purposes* respectively for the CIFAR100 and CIFAR10 datasets with the three different models that are trained with clean data. Similarly, we present in Figures 3 and 4 the probabilistic robust coverage and prediction set size performance of aPRCP using *the Gaussian distribution as a noise distribution for both calibration and testing purposes*. For calibration, we sample $m_s = 128$ noisy data points from the surrounding of each data point ($||\epsilon||_2 \leq 0.125$). For testing, we sample $n_s = 128$ data points from the surrounding of each testing point ($||\epsilon||_2 \leq 0.125$). We observe that the probabilistic robust coverage for noisy data increases monotonically as we increase the quantile robust coverage for each ball from $1 - \tilde{\alpha} = 0.90$ to $1 - \tilde{\alpha} = 1.0$. These observations hold for both conformal scores (HPS and APS) and using different deep neural network models.

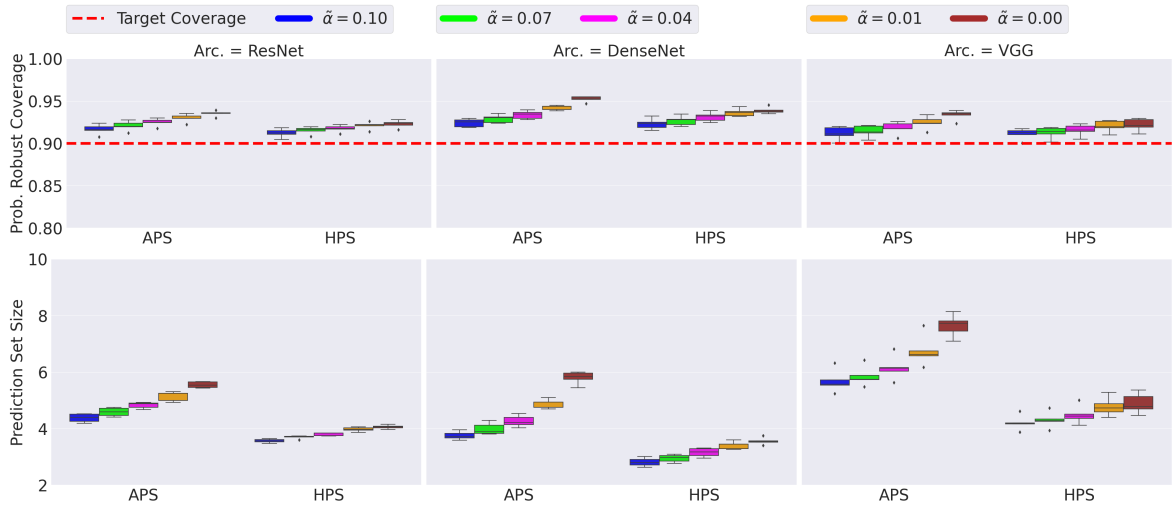


Figure 1: Probabilistic robust coverage (top) and prediction set size (bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$), aPRCP($\tilde{\alpha} = 0.03$), PRCP($\tilde{\alpha} = 0.06$), aPRCP($\tilde{\alpha} = 0.09$), and aPRCP($\tilde{\alpha} = 0.00$), evaluated on CIFAR100 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

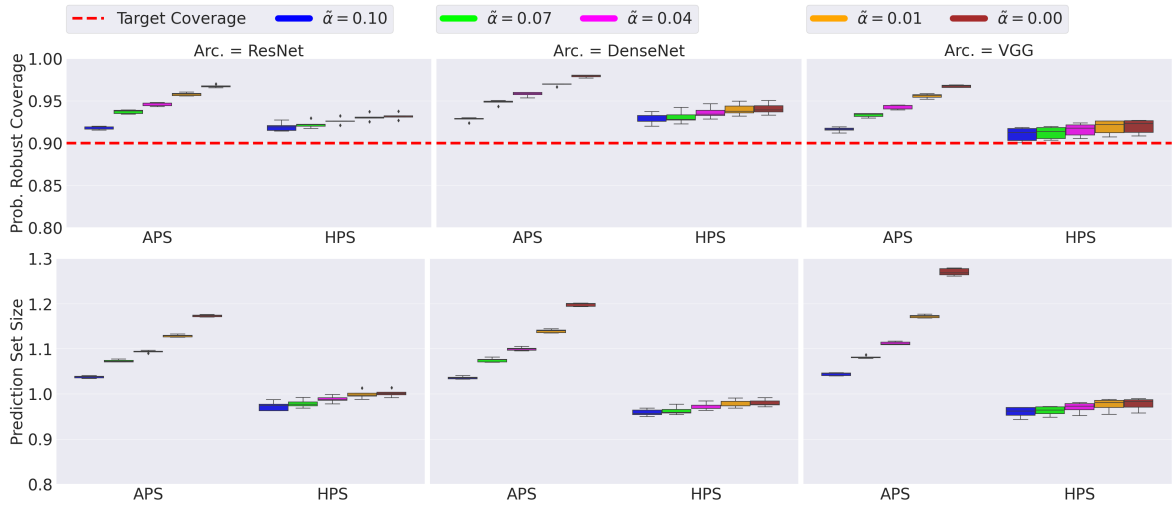


Figure 2: Probabilistic robust coverage (top) and prediction set size (bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$), aPRCP($\tilde{\alpha} = 0.03$), PRCP($\tilde{\alpha} = 0.06$), aPRCP($\tilde{\alpha} = 0.09$), and aPRCP($\tilde{\alpha} = 0.00$), evaluated on CIFAR10 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

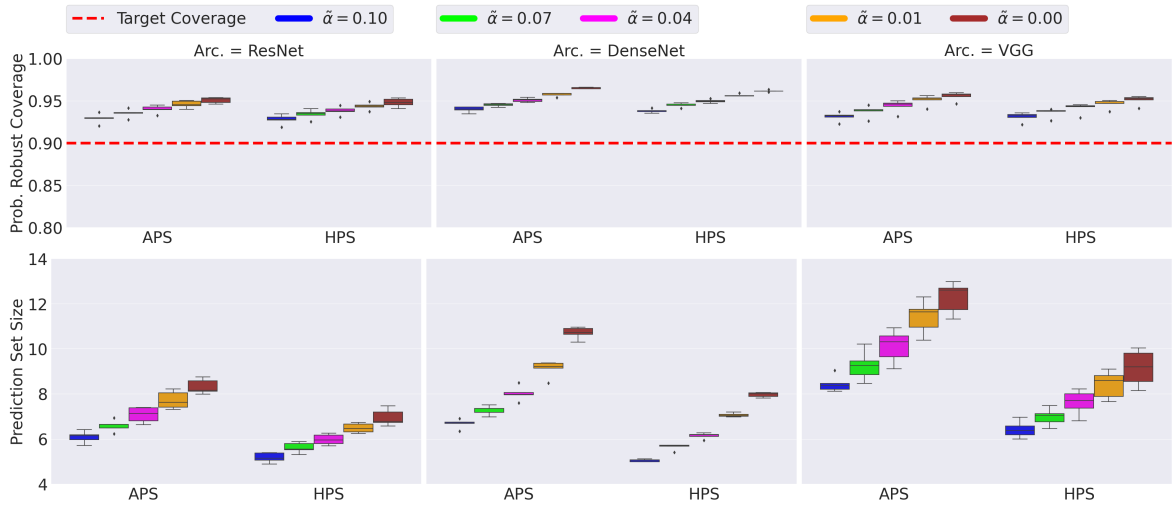


Figure 3: Probabilistic robust coverage(top) and Prediction set size(bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$), aPRCP($\tilde{\alpha} = 0.03$), PRCP($\tilde{\alpha} = 0.06$), aPRCP($\tilde{\alpha} = 0.09$), and aPRCP($\tilde{\alpha} = 0.00$), evaluated on CIFAR100 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

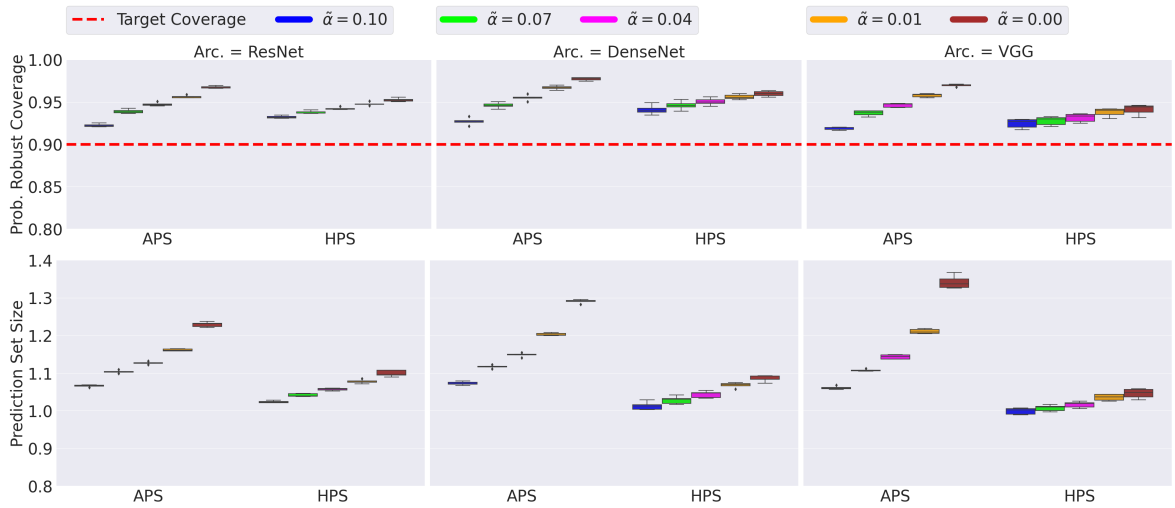


Figure 4: Probabilistic robust coverage(top) and Prediction set size(bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$), aPRCP($\tilde{\alpha} = 0.03$), PRCP($\tilde{\alpha} = 0.06$), aPRCP($\tilde{\alpha} = 0.09$), and aPRCP($\tilde{\alpha} = 0.00$), evaluated on CIFAR10 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

**Performance evaluation with a fixed $\tilde{\alpha}$ hyper-parameter and varying $s$.**

Figures 5 and 6 show the probabilistic robust coverage and prediction set size respectively for the CIFAR100 and CIFAR10 datasets with three different deep models that are trained using standard training. For calibration, we sample $m_s = 128$ noisy data points using the uniform sampling distribution from the surrounding of each data point ($||\epsilon||_2 \leq 0.125$). For testing, we sample $n_s = 128$ data points uniformly from the surrounding of each testing point ($||\epsilon||_2 \leq 0.125$). We observe that the probabilistic robust coverage for noisy data increases as we increase the $s$ parameter value from 0.0 to 0.09. This observation matches our proposition as a higher $s$ value produces higher coverage. The above observations hold for both conformal scores (APS and HPS) using different deep neural network models.

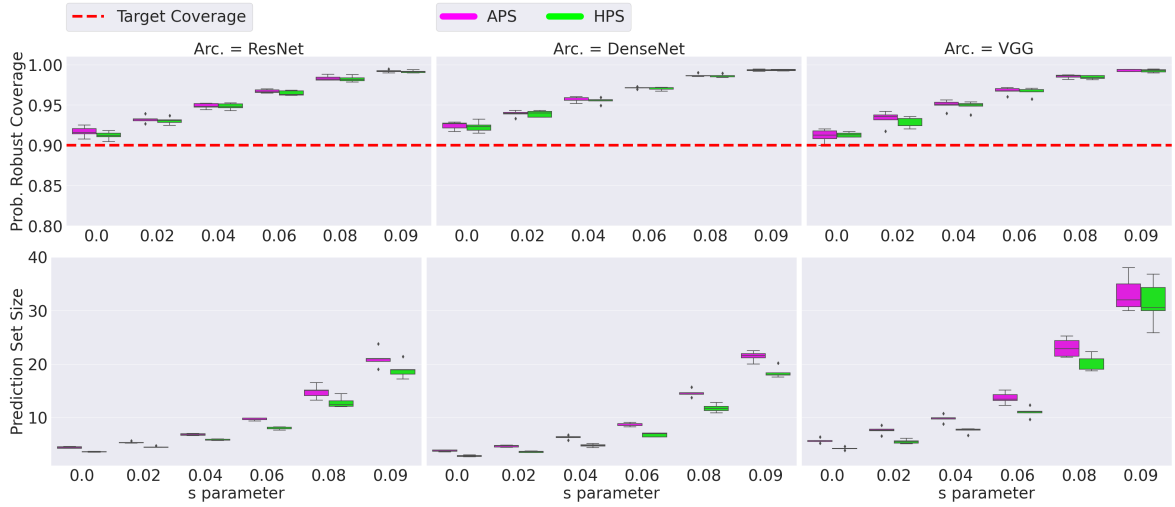


Figure 5: Probabilistic robust coverage(top) and Prediction set size(bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$) while varying the $s$ parameter, evaluated on CIFAR100 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

**Performance evaluation with fixed $s$ and $\tilde{\alpha}$ hyper-parameter and varying sampling radius ($||\epsilon||_2 \leq r$) around test samples.** Figures 7 and 8 present the probabilistic robust coverage and the prediction set size respectively for the CIFAR10 dataset. Similarly, figures 9 and 10 present probabilistic robust coverage and prediction set size for the CIFAR100 dataset. We employ three different deep models that are trained with clean data. For calibration, we sample $m_s = 128$ noisy data points using the uniform sampling distribution from the surrounding of each data point ($||\epsilon||_2 \leq 0.125$), where $\epsilon$ is sampled

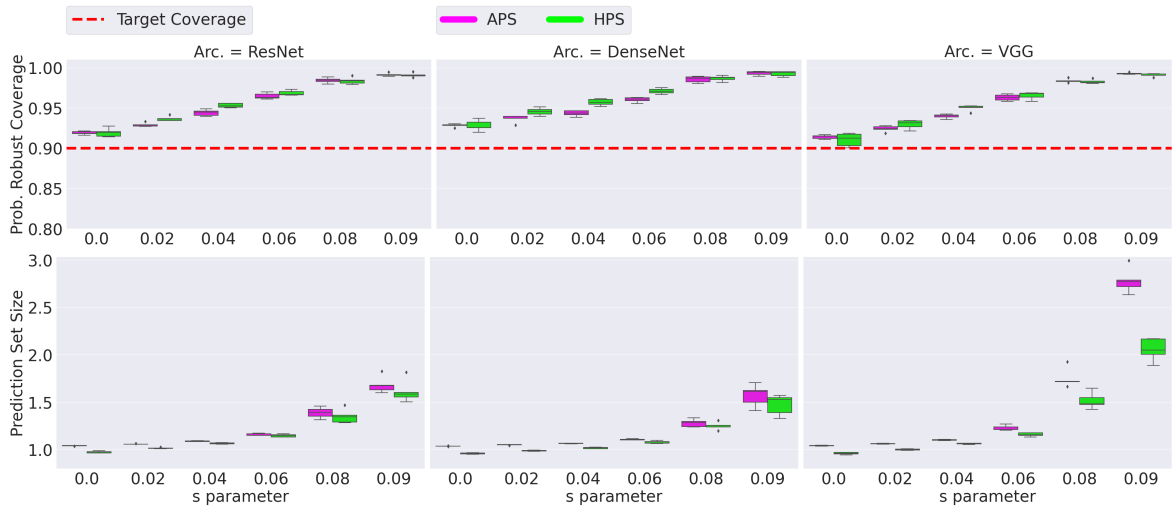


Figure 6: Probabilistic robust coverage(top) and Prediction set size(bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$) while varying the $s$ parameter, evaluated on CIFAR10 dataset for three different models. The target coverage is $90\%$. The results are shown over 50 runs forall three neural network models.

uniformly over the segment $[0, 0.125]$. For testing, we sample $n_s = 128$ data points uniformly from the surrounding of each testing point ($||\epsilon||_2 \leq \{1.0, 2.0, 3.0\}$), where $\epsilon$ is uniformly sampled over the segment $[0, 1], [0, 2], [0, 3]$ respectively. We observe that the probabilistic robust coverage for noisy data decays as we increase the sampling radius. Additionally, we note that when we set the $d$ parameter to $0.1$ (accounting for the change in noise distribution between calibration and testing as per Theorem 2), we guarantee achieving the target coverage. These observations hold for both conformal scores (APS and HPS) using different deep neural network models.

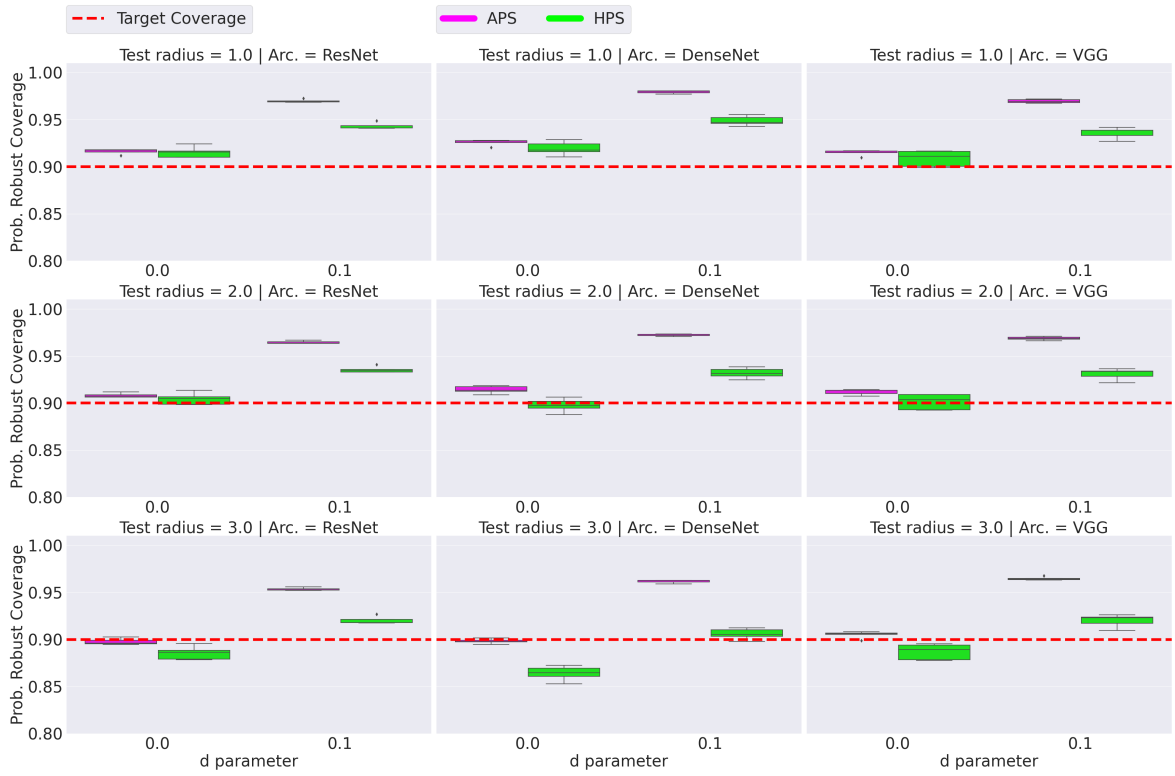


Figure 7: Probabilistic robust coverage evaluated on CIFAR10 dataset for three different models. The target coverage is $90\%$. The results are shown over 50 runs forall three neural network models.

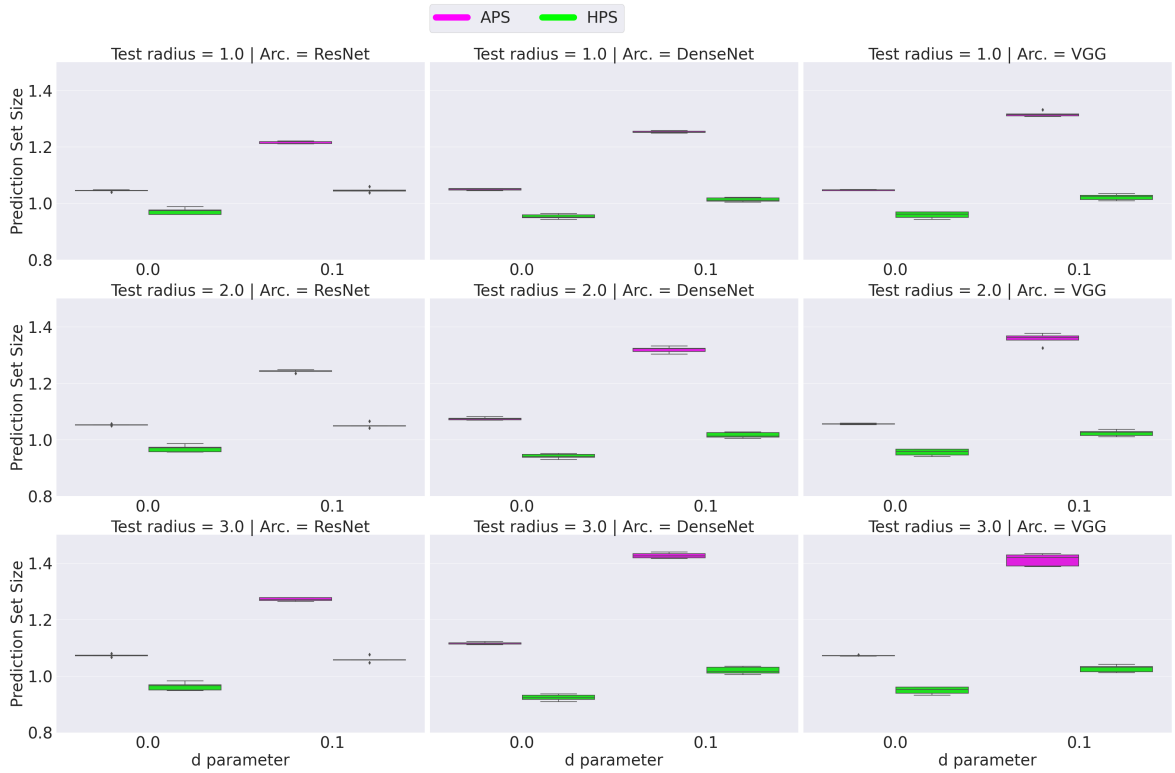


Figure 8: Prediction set size evaluated on CIFAR10 dataset for three different deep models. The results are shown over 50 different runs.

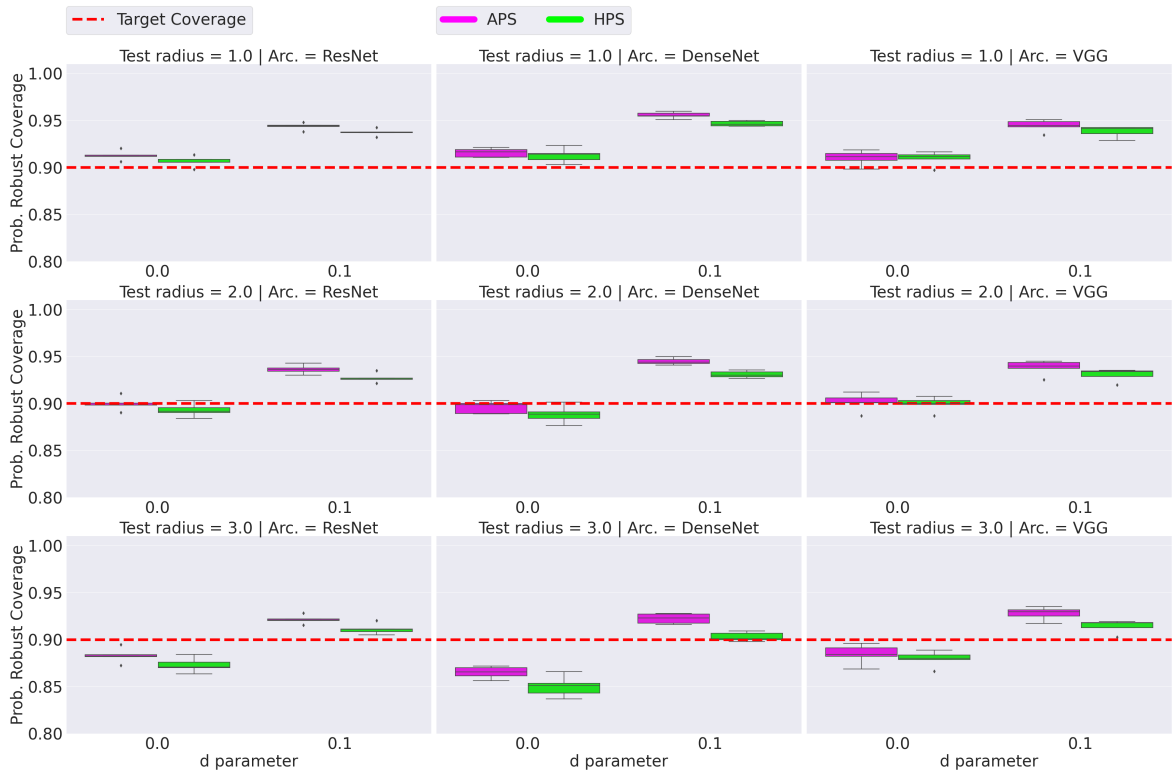


Figure 9: Probabilistic robust coverage evaluated on CIFAR100 dataset for three different deep models. The target coverage is 90%. The results are shown over 50 different runs.

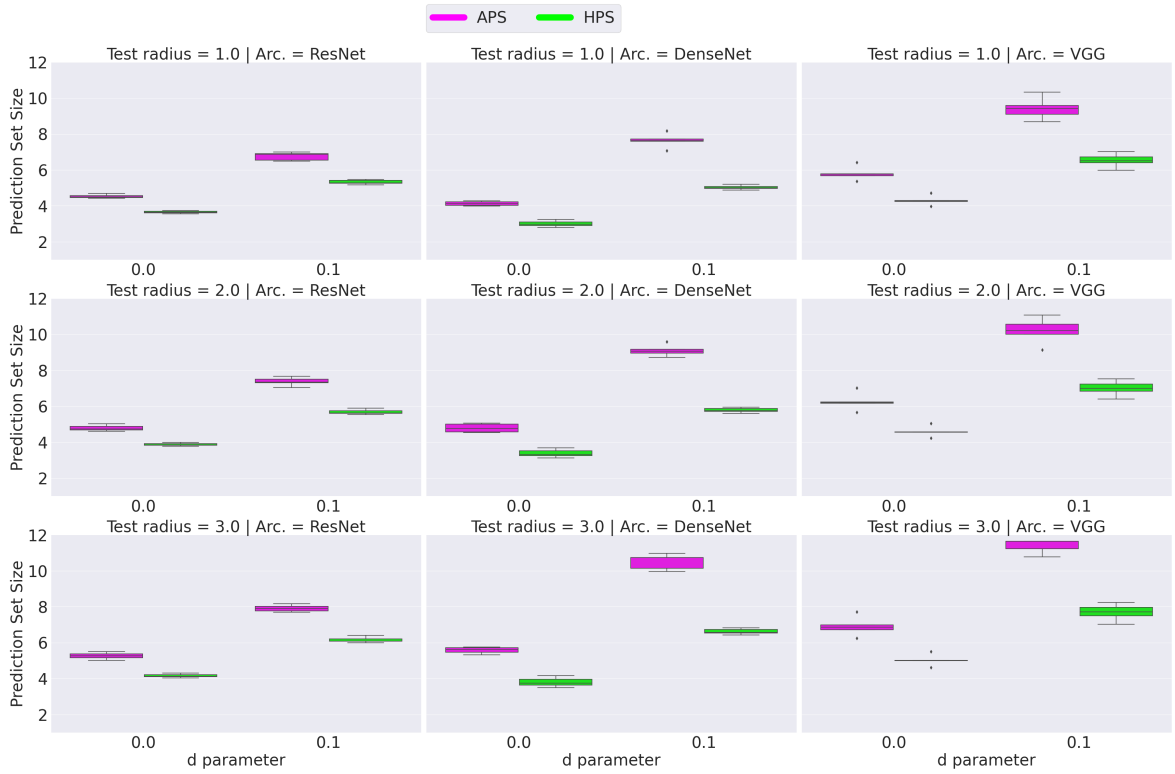


Figure 10: Prediction set size evaluated on CIFAR100 dataset for three different deep models. The results are shown over 50 different runs.

## 2.2 CASE OF DISSIMILAR NOISE DISTRIBUTIONS FOR CALIBRATION AND TESTING

**Gaussian distribution for Calibration and Uniform distribution for Testing with a fixed $s$ hyper-parameter and varying $\tilde{\alpha}$.** Figures 11 and 12 present probabilistic robust coverage and prediction set size respectively for the CIFAR100 and CIFAR10 datasets with three different deep models that are trained with clean data. For calibration, we sample $m_s = 128$ data points using the Gaussian sampling distribution from the surrounding of each data point($||\epsilon||_2 \leq 0.125$). For testing, we sample $n_s = 128$ data points uniformly from the surrounding of each testing point($||\epsilon||_2 \leq 0.125$). We observe that the probabilistic robust coverage increased over the case of using the same distribution for sampling during the testing and calibration phases.

**Uniform distribution for Calibration and Gaussian distribution for Testing with a fixed $s$ hyper-parameter and varying $\tilde{\alpha}$.** Figures 13 and 14 present probabilistic robust coverage and prediction size for CIFAR100 and CIFAR10 datasets respectively with three different deep models that are trained with clean data. For calibration, we sample $m_s = 128$ data points using the Uniform sampling distribution from the surrounding of each data point ($||\epsilon||_2 \leq 0.125$). For testing, we sample $n_s = 128$ data points using Gaussian distribution from the surrounding of each testing point ($||\epsilon||_2 \leq 0.125$). We observe a slightly different performance of aPRCP compared to the case of using the same distribution for noise during the testing and calibration phases. This observation corroborate the statement of Theorem 2 and Remark 2 explaining the relation between the gap of the density probability between the calibration and testing noise distributions with the probabilistic robust coverage for aPRCP.

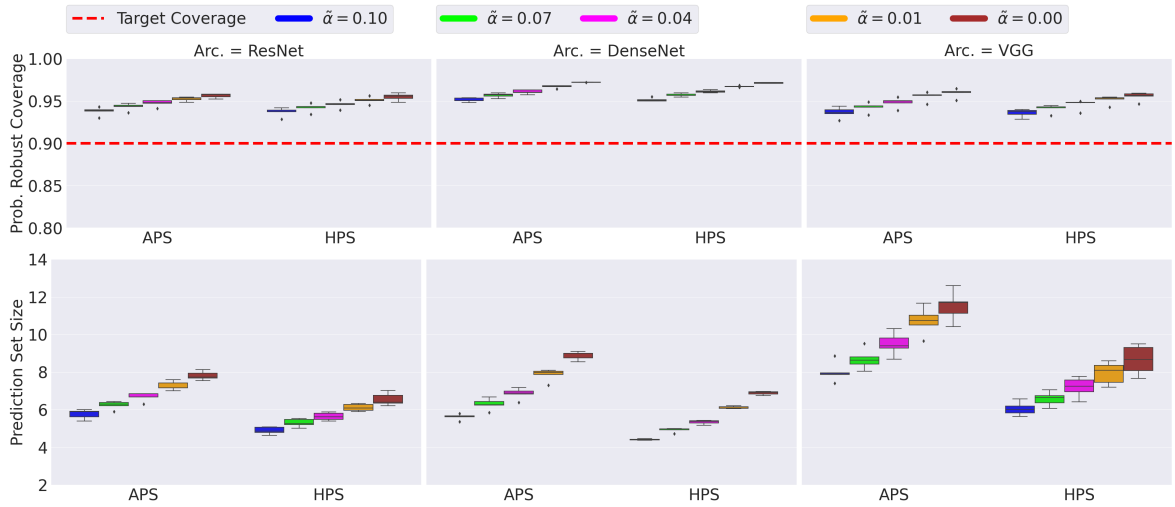


Figure 11: Probabilistic robust coverage(top) and Prediction set size(bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$), aPRCP($\tilde{\alpha} = 0.03$), PRCP($\tilde{\alpha} = 0.06$), aPRCP($\tilde{\alpha} = 0.09$), and aPRCP($\tilde{\alpha} = 0.00$), evaluated on CIFAR100 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

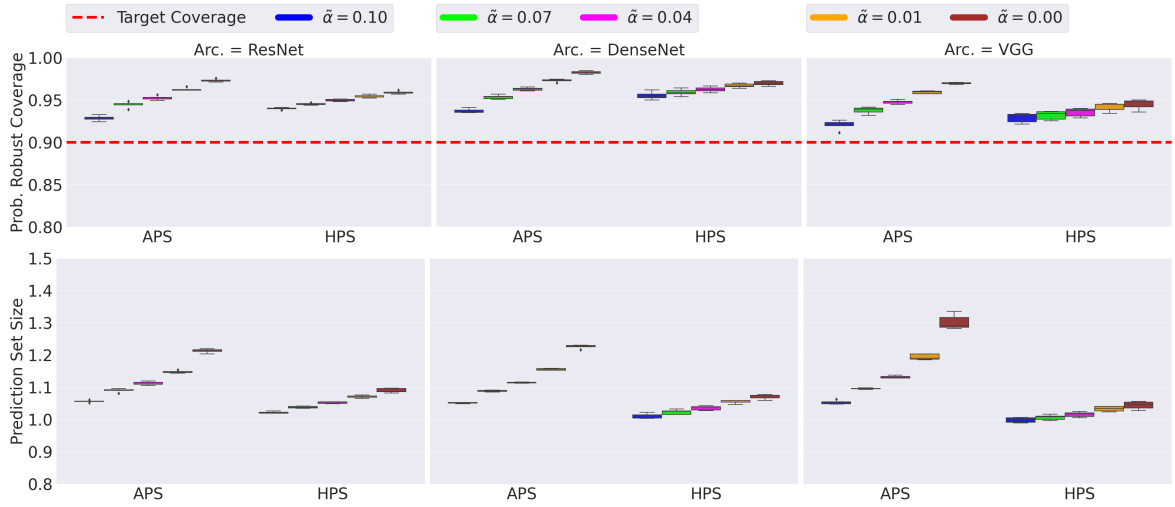


Figure 12: Probabilistic robust coverage(top) and Prediction set size(bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$), aPRCP($\tilde{\alpha} = 0.03$), PRCP($\tilde{\alpha} = 0.06$), aPRCP($\tilde{\alpha} = 0.09$), and aPRCP($\tilde{\alpha} = 0.00$), evaluated on CIFAR10 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

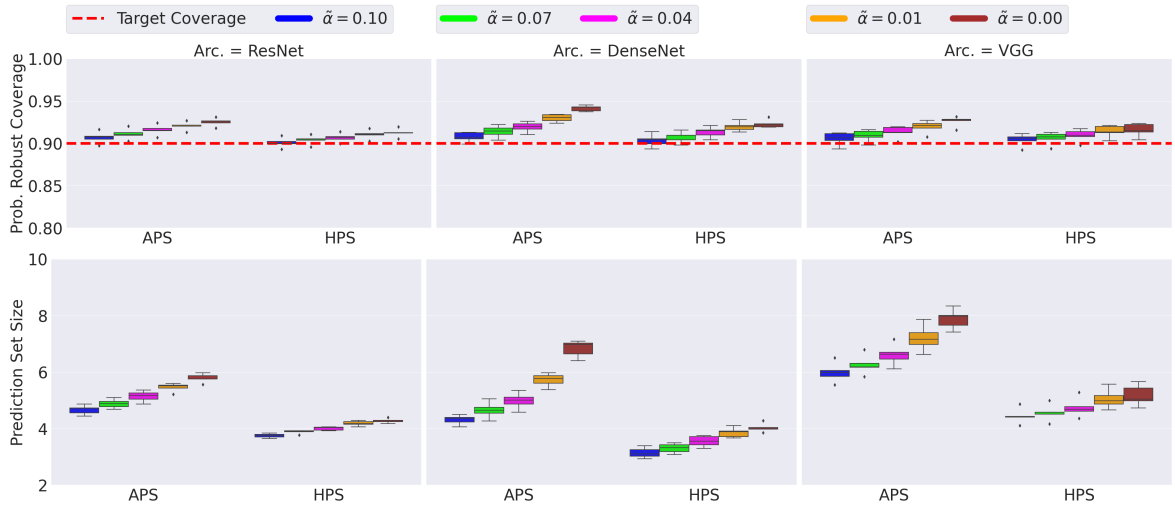


Figure 13: Probabilistic robust coverage(top) and Prediction set size(bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$), aPRCP($\tilde{\alpha} = 0.03$), PRCP($\tilde{\alpha} = 0.06$), aPRCP($\tilde{\alpha} = 0.09$), and aPRCP($\tilde{\alpha} = 0.00$), evaluated on CIFAR100 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

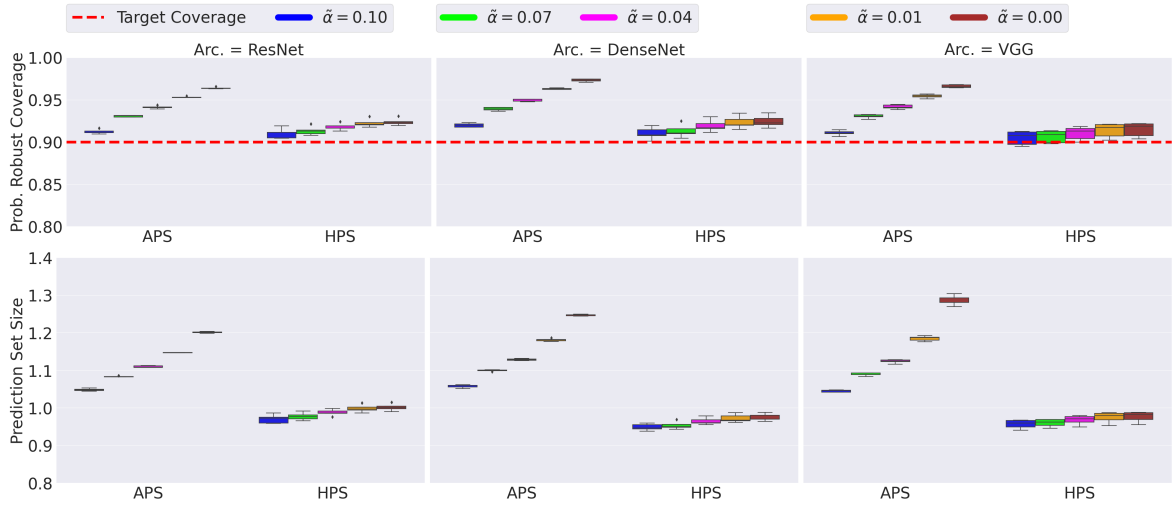


Figure 14: Probabilistic robust coverage(top) and Prediction set size(bottom) obtained by aPRCP($\tilde{\alpha} = 0.10$), aPRCP($\tilde{\alpha} = 0.03$), PRCP($\tilde{\alpha} = 0.06$), aPRCP($\tilde{\alpha} = 0.09$), and aPRCP($\tilde{\alpha} = 0.00$), evaluated on CIFAR10 dataset for three different deep models. The target coverage is $90\%$. The results are shown over 50 different runs.

### 2.3 PERFORMANCE OF APRCP (WORST-ADV) WITH VARYING $m_s$

Figures 15 and 16 show the performance of aPRCP with three different deep models when varying $m_s$ (number of noisy samples for calibration) for CIFAR10 and CIFAR100 datasets respectively. We show the robust coverage and prediction set size for both APS and HPS conformity scores. Both figures show that the aPRCP (worst-adv) reported performance is consistent for different values of $m_s$.

We show in Figure 17 the comparison of the prediction set size and the coverage between RSCP and aPRCP (worst-adv) using both APS and HPS. We employ ResNet110 model trained with Gaussian augmented data ($\sigma = 0.125$). We observe that RSCP is more conservative compared to our method aPRCP (worst-adv) for both APS and HPS conformity scores.

We show in Figure 18 and 19 the comparison of the prediction set size and coverage between RSCP and aPRCP (worst-adv) for two different deep models trained with Gaussian augmented data ($\sigma = 0.0625$ and $\sigma = 0.125$). We observe that aPRCP (worst-adv) produces smaller prediction sets than RSCP.

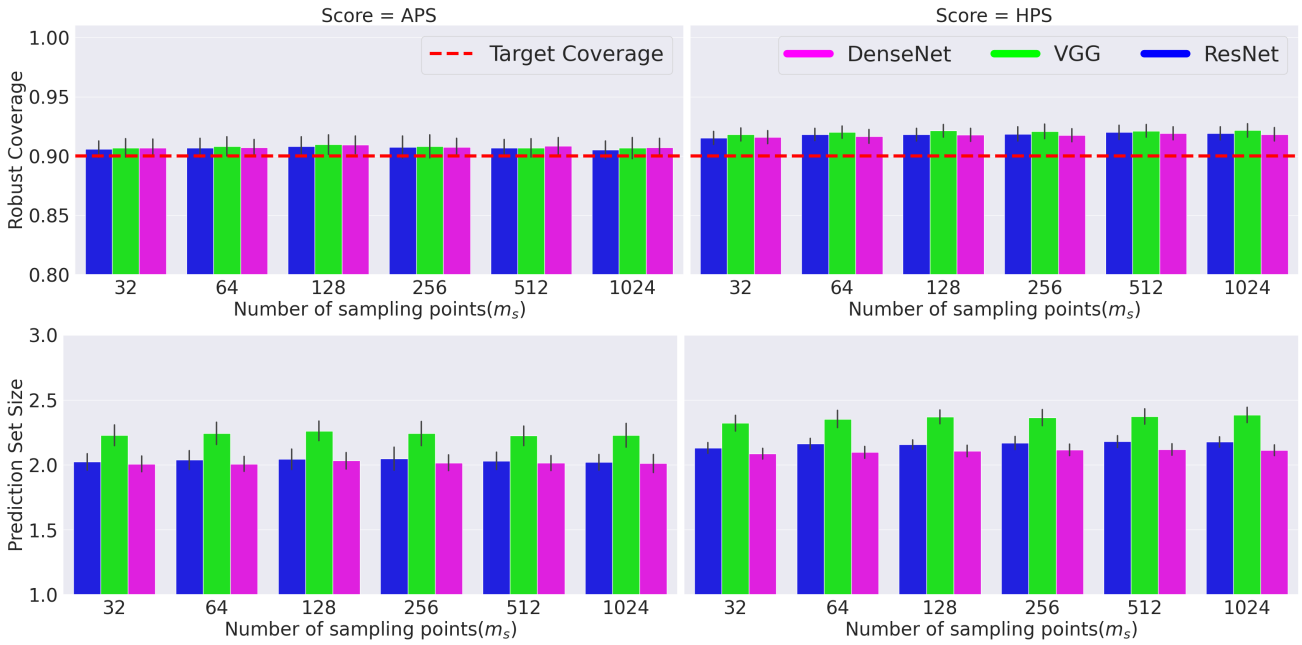


Figure 15: Robust coverage (top) and prediction set size (bottom) performance of two conformity scores (APS and HPS) for different deep models with varying $m_s$ samples on calibration data for CIFAR10 dataset. The results are reported over 50 different runs. We use all models trained with Gaussian augmented data using standard deviation $\sigma = 0.25$.

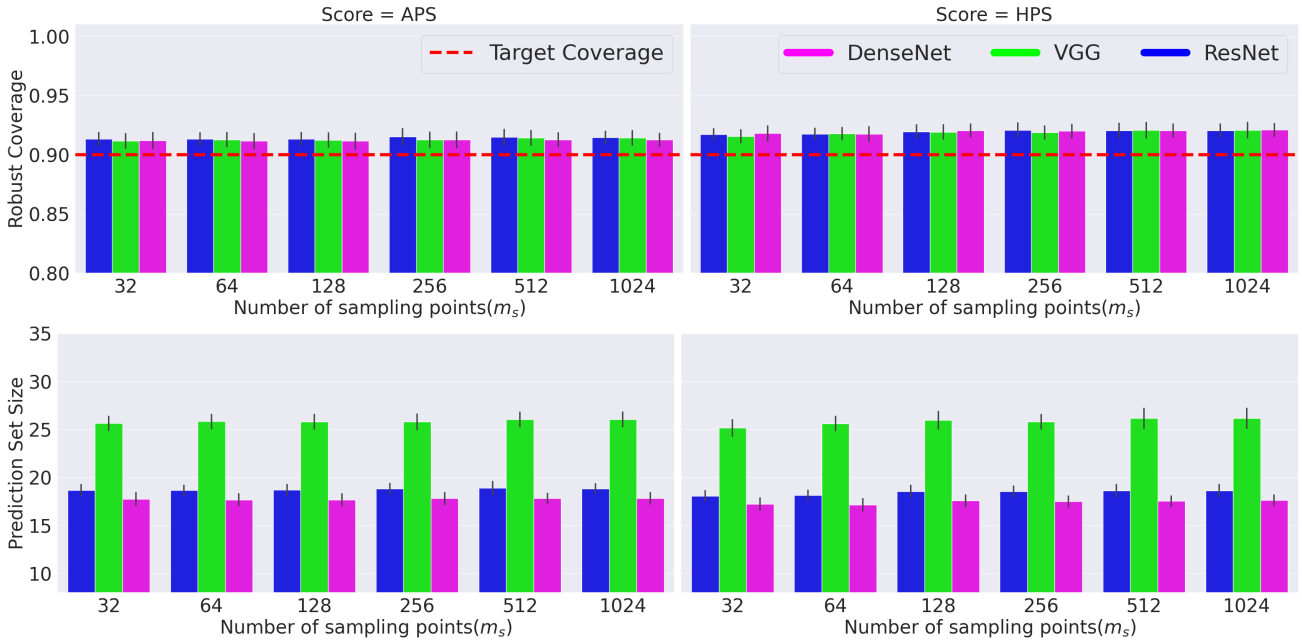


Figure 16: Robust coverage (top) and prediction set size (bottom) performance of two scores for different deep models with varying $m_s$ samples on calibration data for CIFAR100 dataset. The results are reported over 50 different runs. We use all models trained with Gaussian augmented data with standard deviation $\sigma = 0.25$.

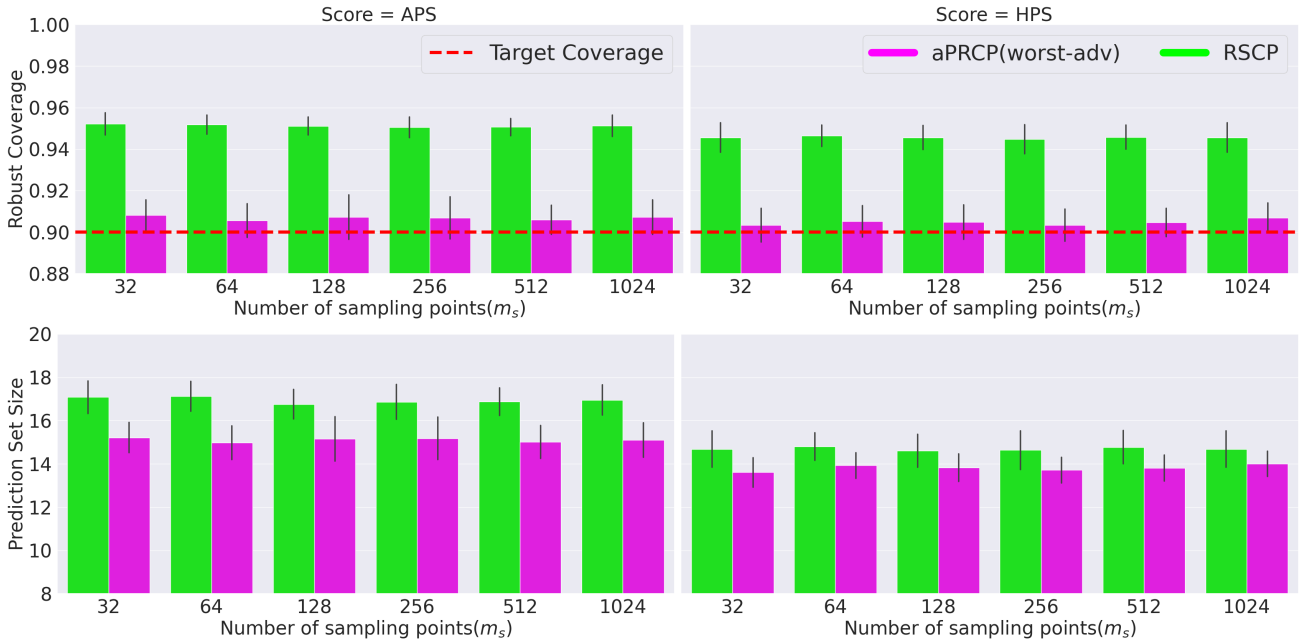


Figure 17: Robust coverage (top) and prediction set size (bottom) performance of two methods, namely, aPRCP(worst-adv) and RSCP, with varying $m_s$ samples on calibration data for CIFAR100 dataset. The results are reported over 50 different runs. We use all models trained with Gaussian augmented data of standard deviation $\sigma = 0.125$.

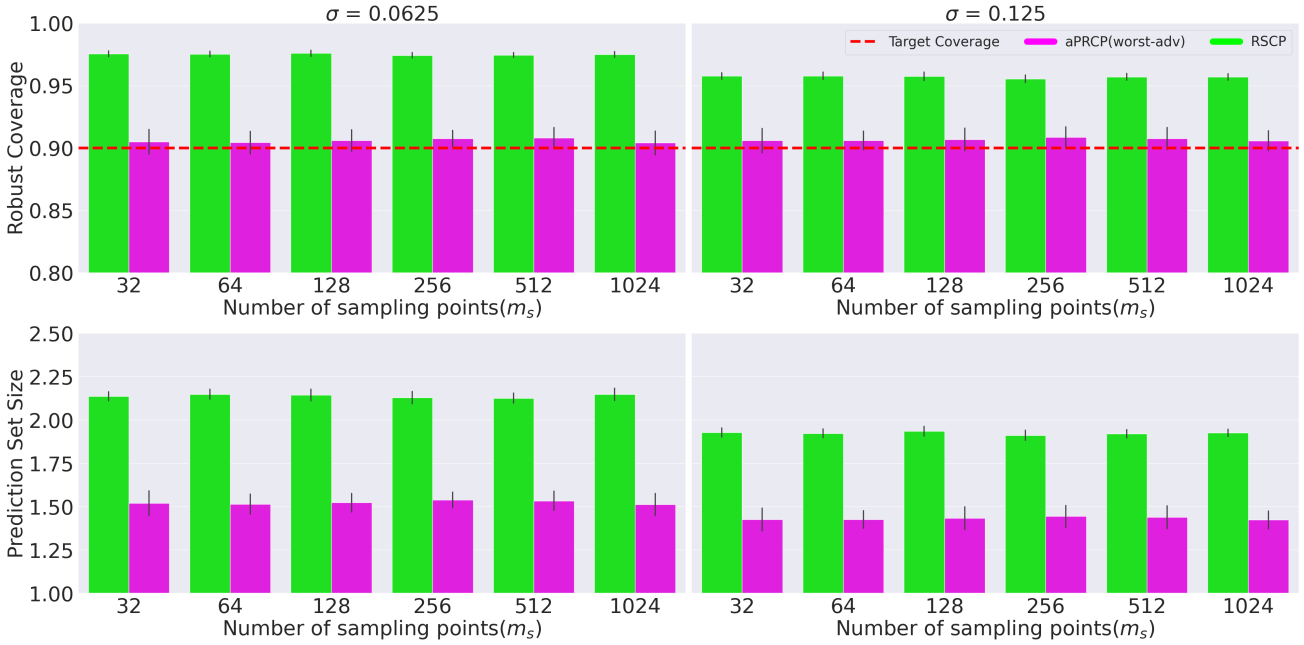


Figure 18: Robust coverage (top) and prediction set size (bottom) performance of two different models trained with Gaussian augmented data using standard deviation $\sigma = 0.0625$ and $\sigma = 0.125$ with varying $m_s$ samples on calibration data for CIFAR10 dataset. The results are reported over 50 different runs.

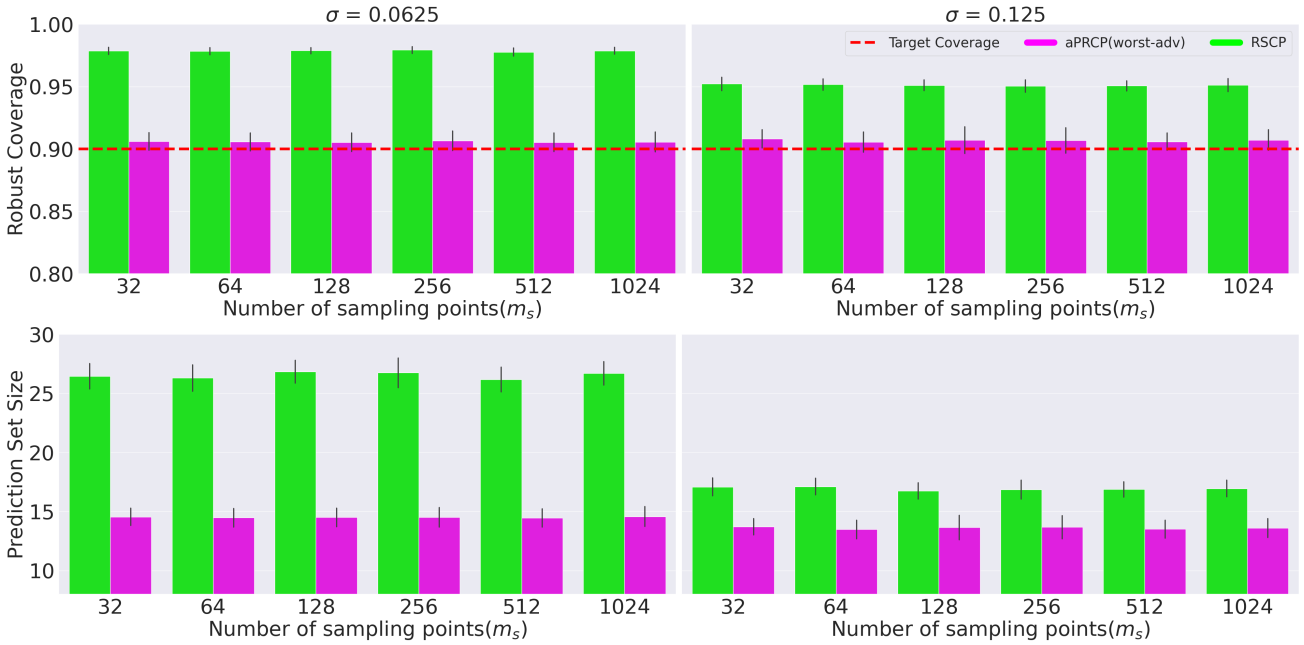


Figure 19: Robust coverage (top) and prediction set size (bottom) performance of two different models trained with Gaussian augmented data using standard deviation $\sigma = 0.0625$ and $\sigma = 0.125$ with varying $m_s$ samples on calibration data for CIFAR100 dataset. The results are reported over 50 different runs.

### 2.4 THE EFFECT OF VARYING $||\epsilon||_2 \leq r$ DURING CALIBRATION

We show in Figure 20 the robust coverage and the prediction set size achieved by aPRCP(worst-adv) on CIFAR100 with a ResNet model that is trained with Gaussian augmented data ($\sigma = 0.125$). For calibration, we sample $m_s = 128$ noisy data points using the uniform sampling distribution from the surrounding of each data point ($||\epsilon||_2 \leq r$), where $r = \{0.125, 0.250, 1.0\}$. For testing, we generate data using an adversarial attack algorithm of energy $0.125$. We observe that the effect of the small changes in the sampling radius is negligible.

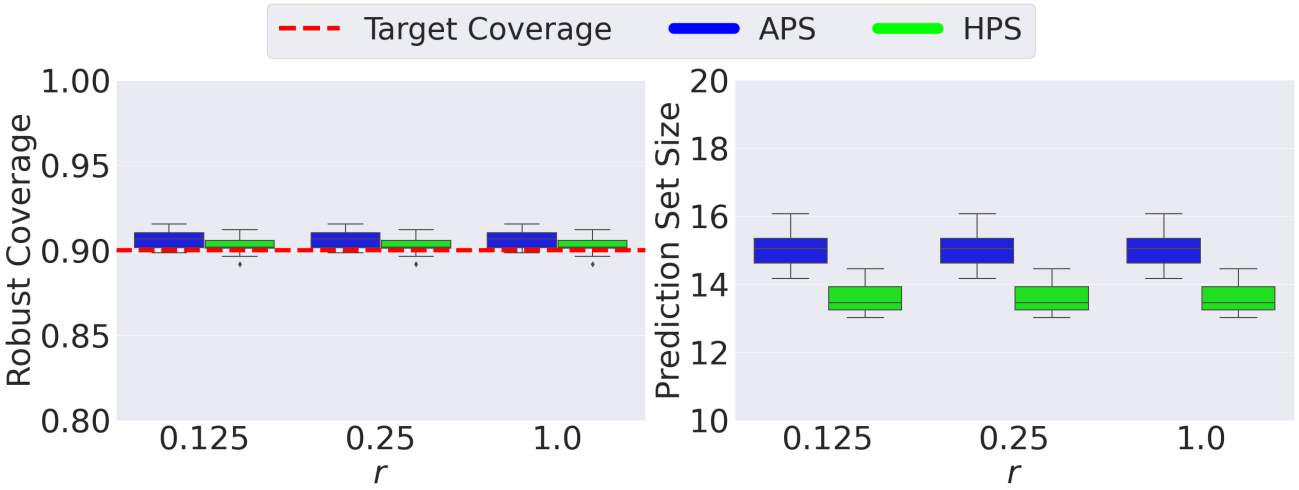


Figure 20: Robust coverage (left) and prediction set size (bottom) performance of the ResNet model trained with Gaussian augmented data using standard deviation $\sigma = 0.125$ with varying radius of robust quantile balls during calibration for CIFAR100 dataset. The results are reported over 50 different runs.

### 2.5 PERFORMANCE OF APRCP(WORST-ADV) WITH DIFFERENT DEEP MODELS

Figure 21 shows the performance of our `aPRCP`(worst-adv) using DenseNet[Iandola et al., 2014] and VGG[Simonyan and Zisserman, 2014] models on the CIFAR10 and CIFAR100 datasets. We use the same adversarial attack algorithm for test examples with a magnitude of $r = 0.125$. During calibration, we sample $m_s = 128$ noisy samples ($r = 0.125$) for each calibration example. We observe that the robust coverage is achieved on all three deep models with small prediction sets.

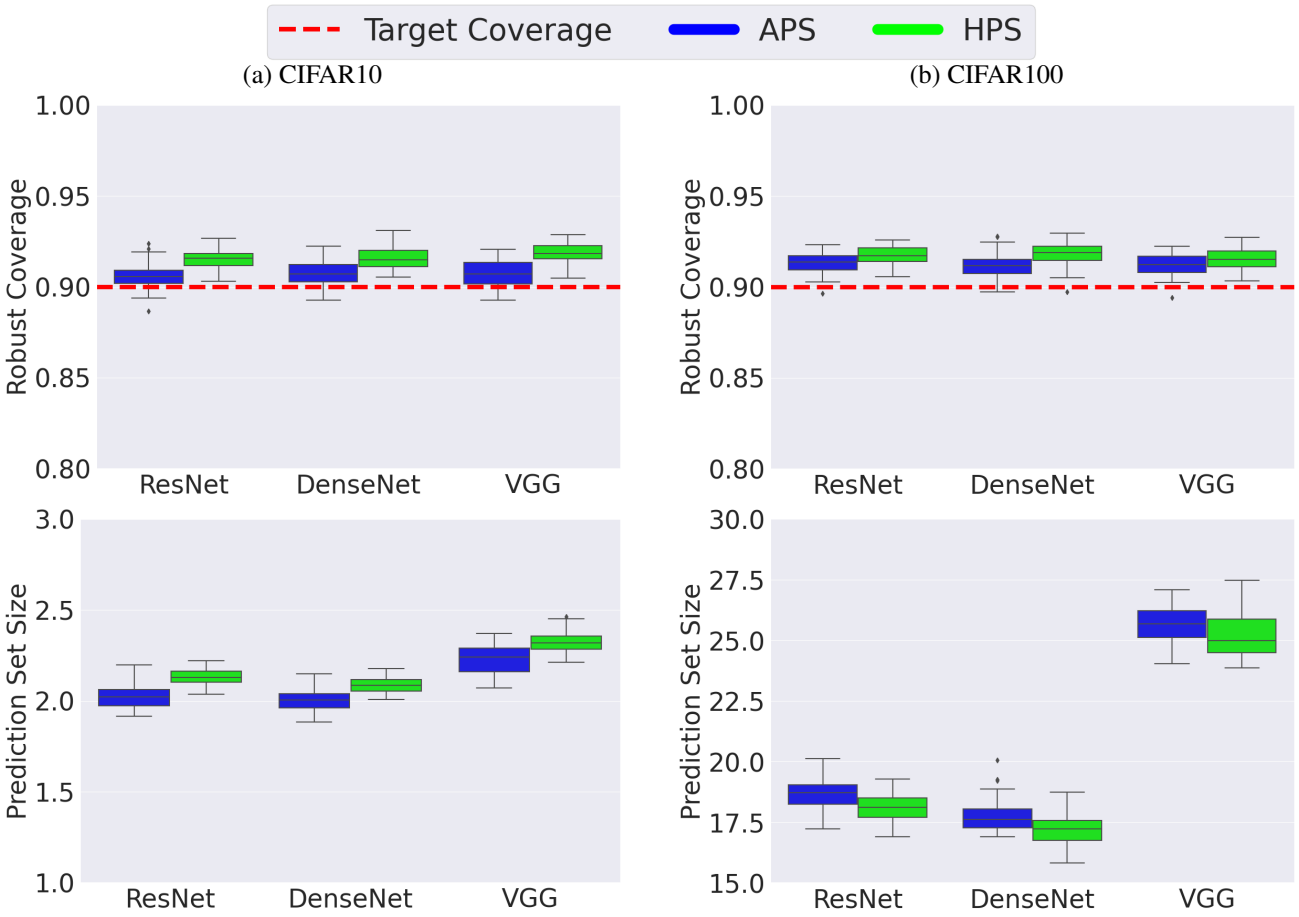


Figure 21: Robust coverage (top) and prediction set size (bottom) constructed by `aPRCP(worst-adv)` method for CIFAR10 (left) and CIFAR100 (right) datasets. The neural network models used are trained with Gaussian augmented data using standard deviation $\sigma = 0.25$. The results are reported over 50 different runs. As can be seen, all models guarantee target coverage and VGG produces larger prediction sizes compared to other models.

## 2.6 RESULTS ON ADVERSARIAL EXAMPLES GENERATED FROM A PROBABILITY DENSITY DISTRIBUTION

We evaluate the performance of aPRCP with a different adversarial attack algorithm, namely NATTACK [Li et al., 2019]. This attack algorithm generates a probability density distribution centered around an input from which adversarial examples can be sampled. We employ this algorithm using an adversarial magnitude $||\epsilon||_2 \leq r = 0.125$ to generate adversarial examples for the test data of CIFAR10 and CIFAR100 on three different deep models trained with Gaussian augmented data ($\sigma = 0.125$). In all our experiments, we set $T = 1000$ as the number of maximum iterations, and a learning rate $\eta = 0.008$.

Both Figures 22 and 23 show that aPRCP is the only algorithm that can guarantee the adversarial robust coverage. This can be explained by the fact that RSCP requires the design of a specialized scoring function to guarantee coverage while aPRCP uses a quantile-of-quantile design and can employ any existing score function.

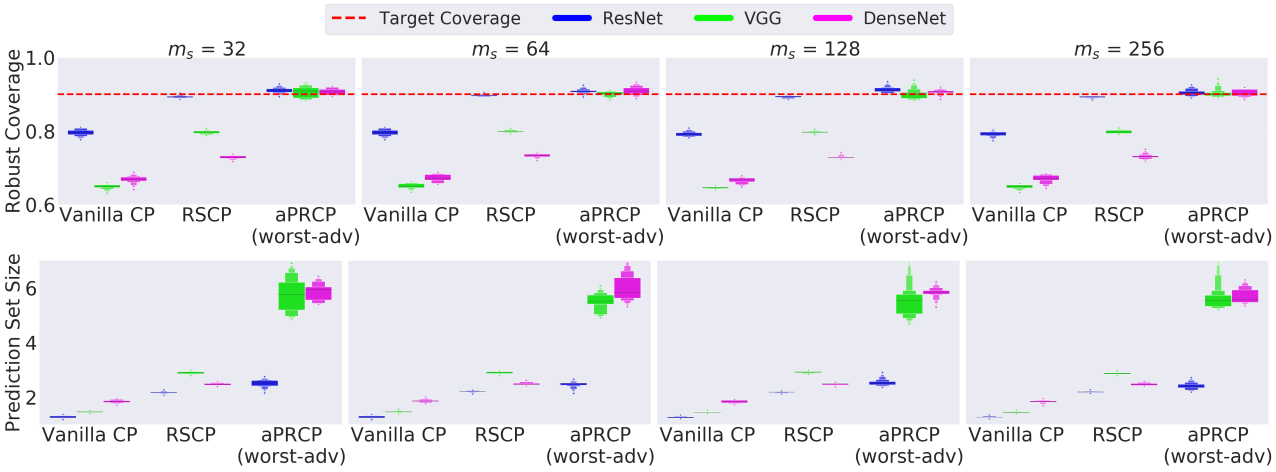


Figure 22: Robust coverage (top) and prediction set size (bottom) constructed by three different CP methods. The target coverage is 90%. The results are reported over 50 different runs for the CIFAR10 data set.

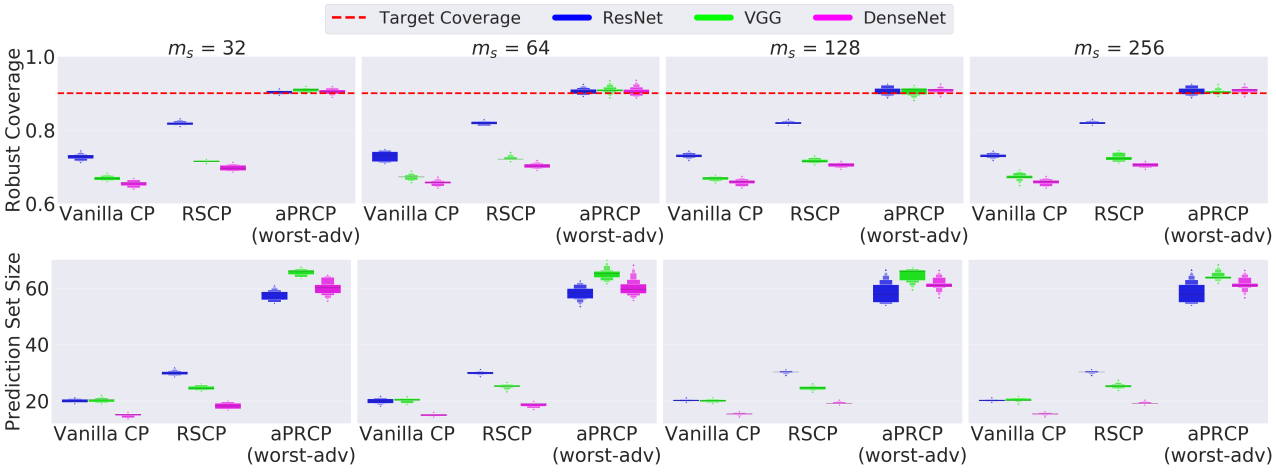


Figure 23: Robust coverage (top) and prediction set size (bottom) constructed by three different CP methods. The target coverage is 90%. The results are reported over 50 different runs for the CIFAR100 data set.

## 2.7 IMPORTANCE OF GAUSSIAN AUGMENTED TRAINING

While aPRCP can work without any assumption on the base classifier, Figure 24 shows the importance of the model robustness to produce smaller prediction sets. Both `RSCP` and `aPRCP`(worst-adv) construct prediction sets that are larger when the base model is not adversarially robust.

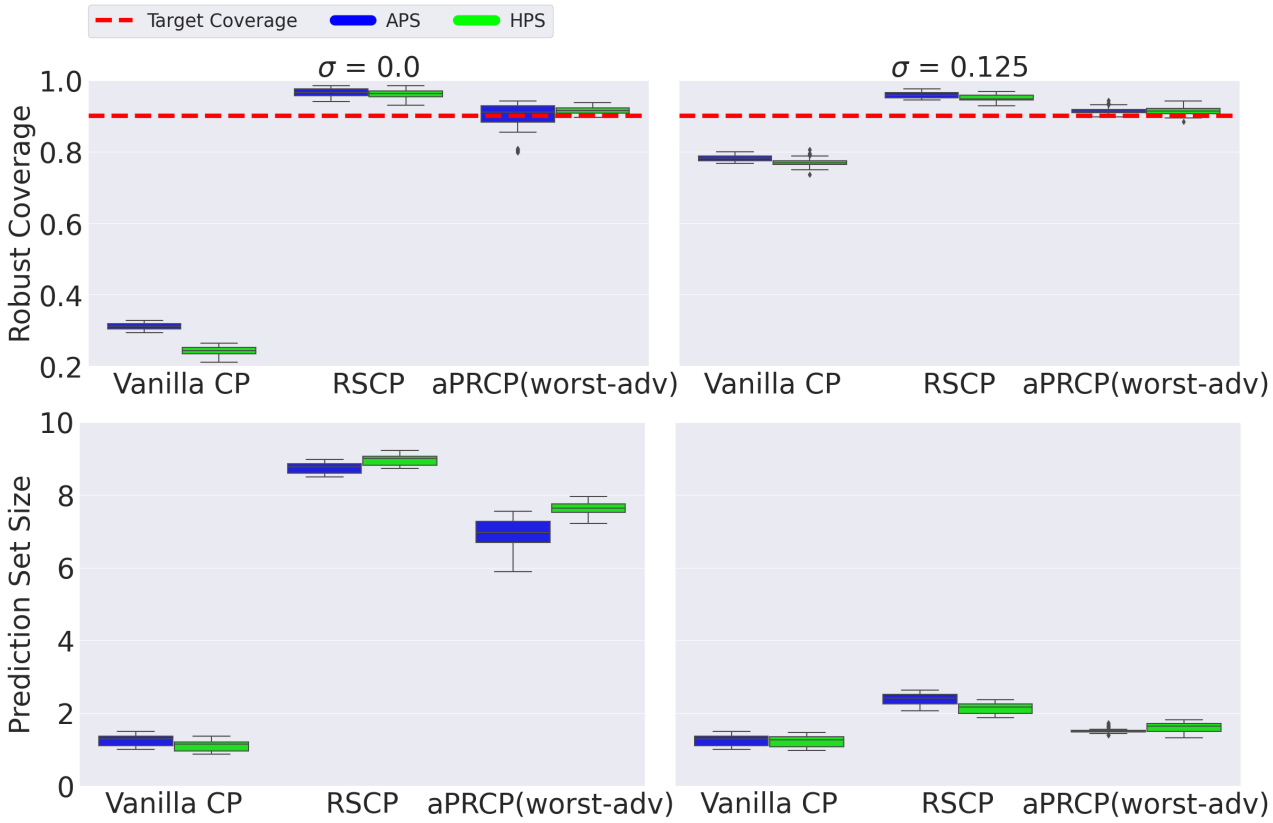


Figure 24: Robust coverage (top) and prediction set size (bottom) constructed by three different CP methods. The target coverage is $90\%$. The results are reported over 50 different runs for the CIFAR10 data set.

### 2.8 APRCP NOMINAL PERFORMANCE

Figure 25 shows a comparison of the nominal performance (evaluation on only clean inputs) on CIFAR10 and CIFAR100 datasets. We employ $m_s = 128$ for calibration and standard training to train the base model. We can observe that aPRCP achieves better trade-off between the nominal performance (evaluation on clean inputs) and the robust performance (evaluation on perturbed inputs). For both datasets, aPRCP achieves a tighter empirical coverage (closer to 90%) with smaller prediction sets than RSCP.

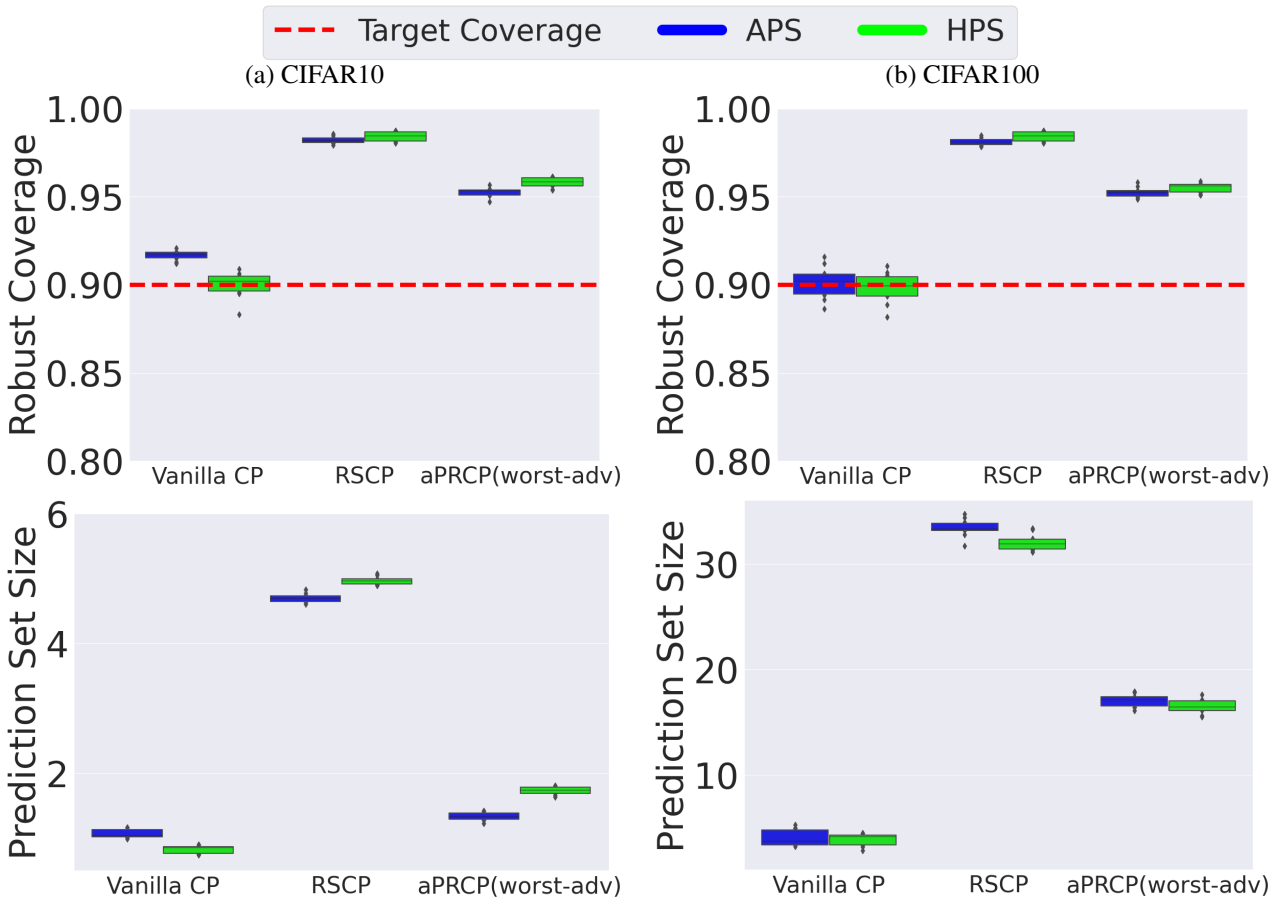


Figure 25: Robust coverage (top) and prediction set size (bottom) constructed by Vanilla CP, RSCP, and aPRCP(worst-adv) using HPS and APS conformity scoring functions (target coverage is $90\%$) for the CIFAR10 and CIFAR100 data sets. Results are averaged over 50 different runs.

## References


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

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.

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`.

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