PAC Prediction Sets under Label Shift

Wenwen Si[1], Sangdon Park[2], Insup Lee[1], Edgar Dobriban[1], Osbert Bastani[1]

[1]University of Pennsylvania

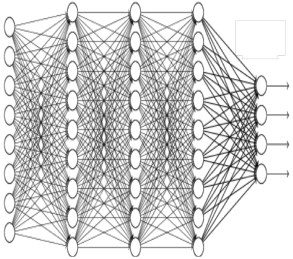
[2]POSTECH

Prediction Sets

No Predictor achieves 100% accuracy.

Incorrect!
(Ground truth label: y= “effusion”)

Image x



Predictor f (NN)

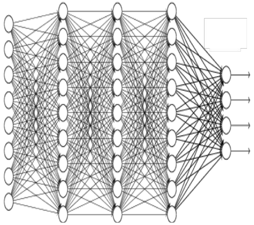
infiltration

Output
$\hat{y} = \arg\max_{y \in \mathcal{Y}} f(x)$

Idea: Modify predictor f to predict **sets of labels**

Now, we have $y \in C(x)$

Image x



Prediction Set C

$\left\{\begin{matrix}\text{effusion} \\ \text{infiltration}\end{matrix}\right\}$

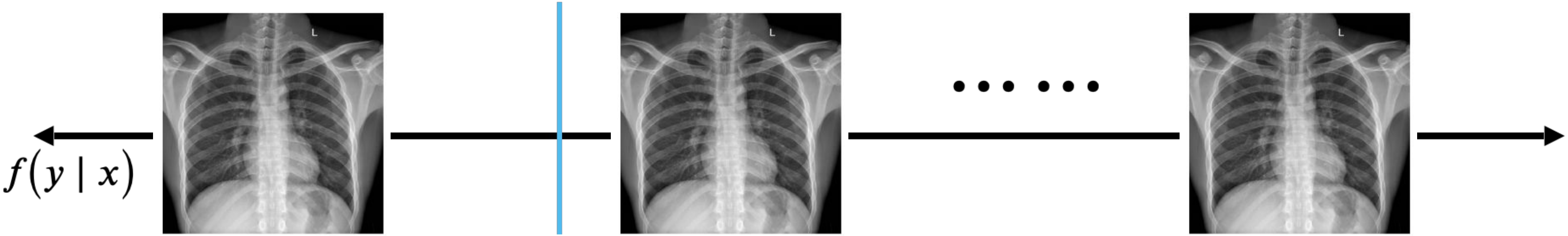
Output $\hat{y} = C(x)$

Prediction Sets

- Construct prediction sets based on an existing predictor $(x, y) \mapsto f(y \mid x)$:

$$\boxed{C_\tau(x) = \left\{ y \mid f(y \mid x) \geq \tau \right\}}$$

$$\tau = 0.35$$



$f(y \mid x)$

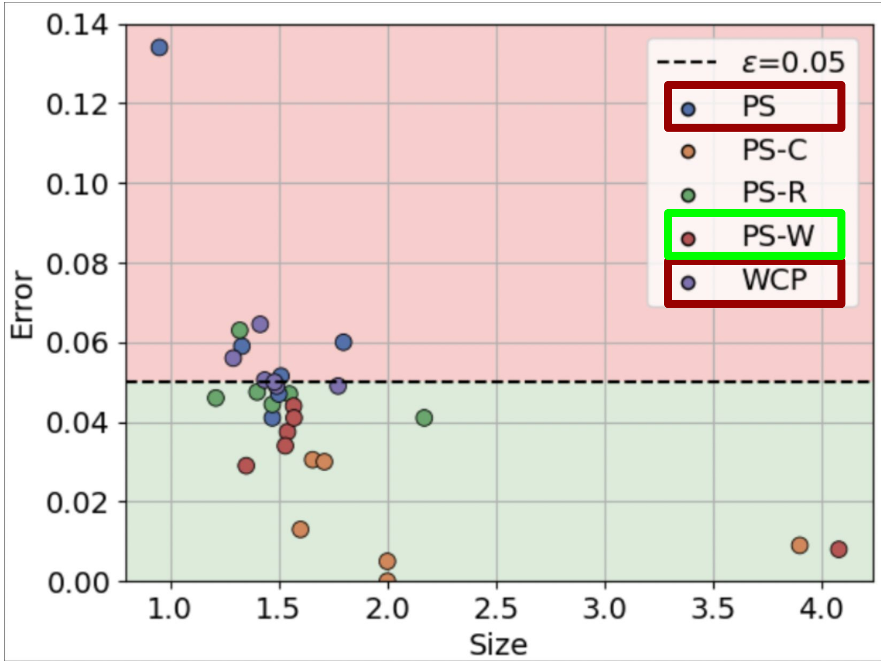
nodule (0.1)

effusion (0.4)

infiltration (0.9)

$$C_\tau(x) = \{\text{effusion, infiltration, ...} \quad \}$$

Post-processing: It gives a recipe for distribution-free finite sample prediction intervals, starting from an arbitrary score function S.



Conformal prediction

$$P\{Y_{n+1} \in \hat{C}_n(X_{n+1})\} \geq 1 - \alpha.$$

PAC/Conditional Valid Guarantee

$$\mathbb{P}_{S_m \sim P^m}[\mathbb{P}_{(X,Y)\sim P}[Y \in C_{\hat{\tau}(S_m)}(X)] \geq 1 - \varepsilon] \geq 1 - \delta.$$

Clopper-Pearson Interval

Label Shift

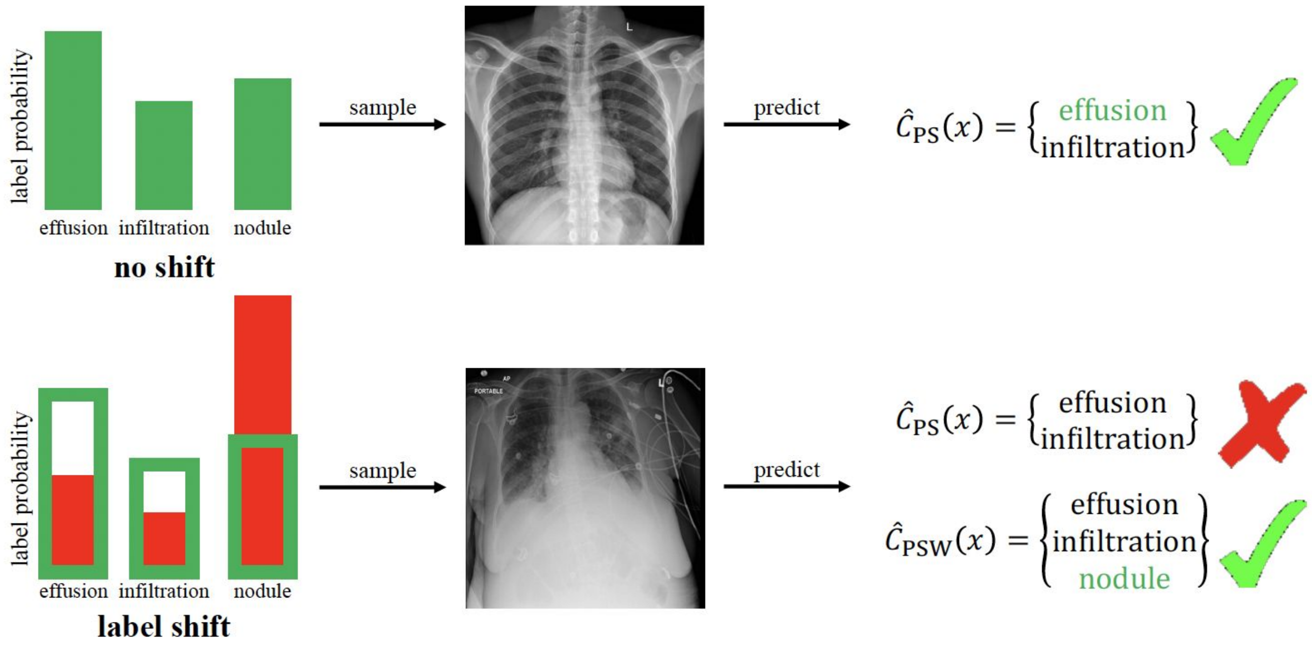
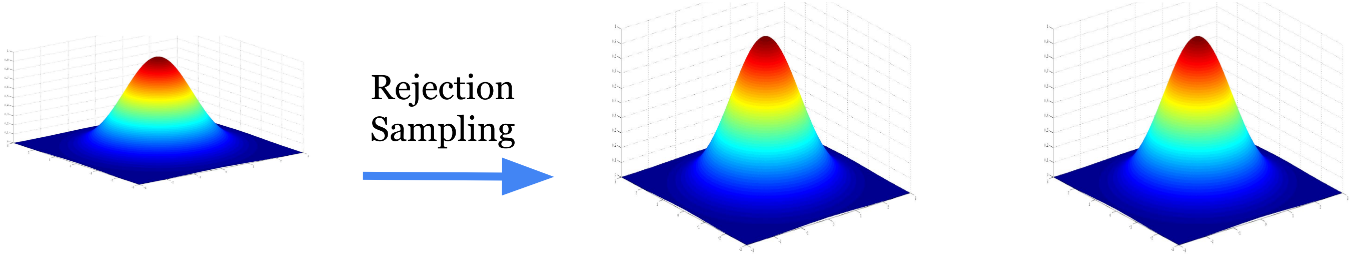

label probability
effusion
infiltration
nodule
no shift
sample
predict
$\hat{C}_{\text{PS}}(x) = \begin{Bmatrix} \text{effusion} \\ \text{infiltration} \end{Bmatrix}$
L
label probability
effusion
infiltration
nodule
label shift
sample
AP
PORTABLE
predict
$\hat{C}_{\text{PS}}(x) = \begin{Bmatrix} \text{effusion} \\ \text{infiltration} \end{Bmatrix}$
$\hat{C}_{\text{PSW}}(x) = \begin{Bmatrix} \text{effusion} \\ \text{infiltration} \\ \text{nodule} \end{Bmatrix}$

Figure 1: An example of our approach on the ChestX-ray dataset. In the unshifted setting, standard PAC prediction sets guarantee high-probability coverage, but this guarantee fails under label shift. Our approach addresses this challenge and continues to work in the shifted environment.

Unsupervised Label Shift Adaptation

- Oracle - Importance weights aware rejection sampling adaptation



Rejection Sampling

Source Distribution

Target Distribution

- BBSE estimation

$$p(x \mid y) = q(x \mid y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}.$$

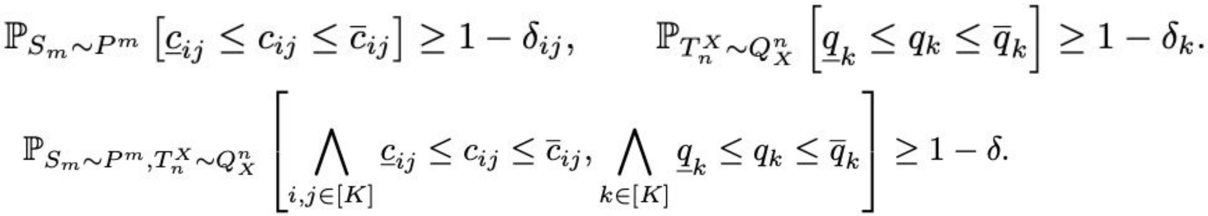
$$q_{\hat{y}} = \sum_{y \in \mathcal{Y}} q(\hat{y} \mid y) q(y) = \sum_{y \in \mathcal{Y}} p(\hat{y} \mid y) q(y) = \sum_{y \in \mathcal{Y}} p(\hat{y}, y) \frac{q(y)}{p(y)}$$

$$w^* = \mathbf{C}_P^{-1} q^*.$$

Lipton, Zachary, Yu-Xiang Wang, and Alexander Smola. "Detecting and correcting for label shift with black box predictors." *International conference on machine learning*. PMLR, 2018.

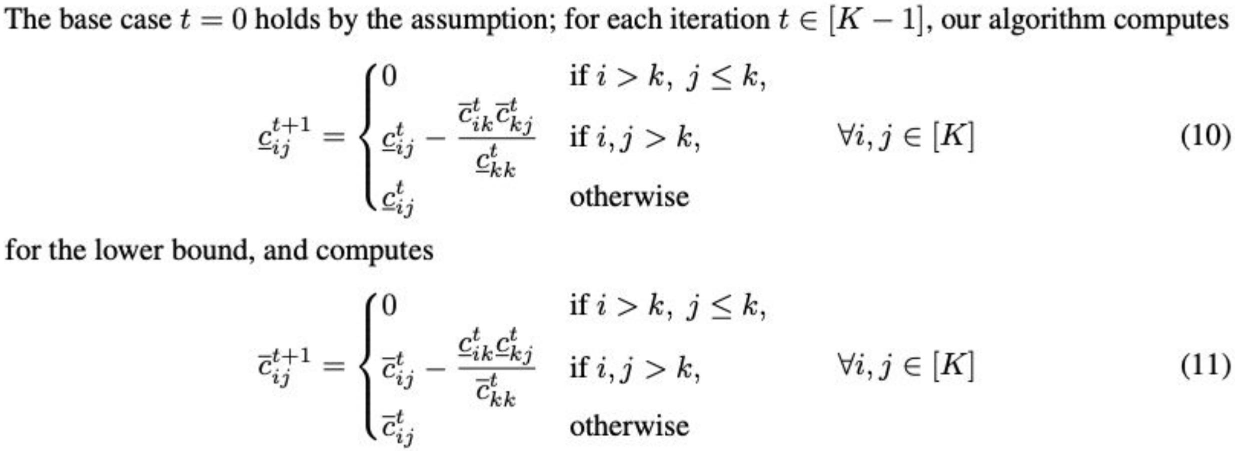
Greedy rejection sampling with confidence intervals

- Clopper-Pearson bound + Gaussian Elimination



$$\mathbb{P}_{S_m\sim P^m}\left[\underline{c}_{ij} \le c_{ij} \le \overline{c}_{ij}\right] \ge 1-\delta_{ij}, \qquad \mathbb{P}_{T_n^X\sim Q_X^n}\left[\underline{q}_k \le q_k \le \overline{q}_k\right] \ge 1-\delta_k.$$

$$\mathbb{P}_{S_m\sim P^m, T_n^X\sim Q_X^n}\left[\bigwedge_{i,j\in[K]} \underline{c}_{ij} \le c_{ij} \le \overline{c}_{ij}, \bigwedge_{k\in[K]} \underline{q}_k \le q_k \le \overline{q}_k\right] \ge 1-\delta.$$



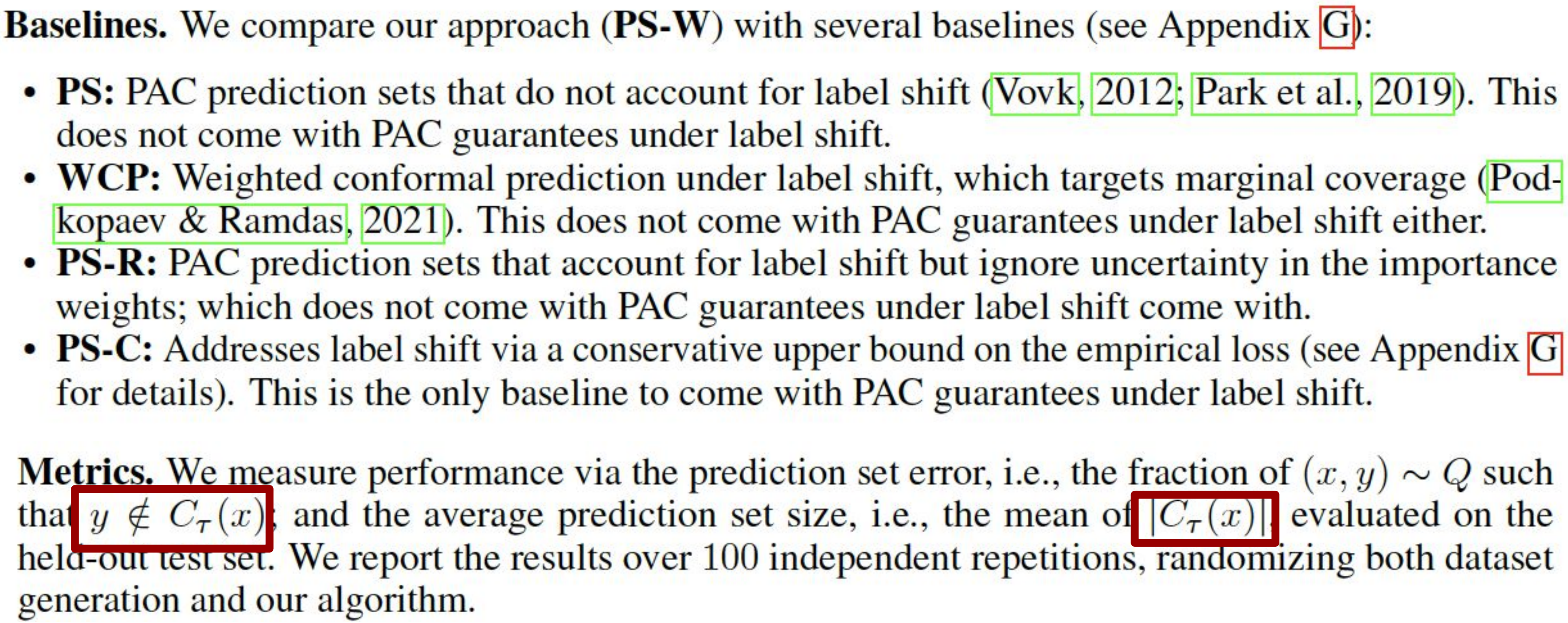
The base case $t = 0$ holds by the assumption; for each iteration $t \in [K-1]$, our algorithm computes

$$\underline{c}_{ij}^{t+1} = \begin{cases} 0 & \text{if } i > k,\ j \le k, \\ \underline{c}_{ij}^{t} - \dfrac{\overline{c}_{ik}^{t}\overline{c}_{kj}^{t}}{\underline{c}_{kk}^{t}} & \text{if } i, j > k, \\ \underline{c}_{ij}^{t} & \text{otherwise} \end{cases} \qquad \forall i, j \in [K] \tag{10}$$

for the lower bound, and computes

$$\overline{c}_{ij}^{t+1} = \begin{cases} 0 & \text{if } i > k,\ j \le k, \\ \overline{c}_{ij}^{t} - \dfrac{\underline{c}_{ik}^{t}\underline{c}_{kj}^{t}}{\overline{c}_{kk}^{t}} & \text{if } i, j > k, \\ \overline{c}_{ij}^{t} & \text{otherwise} \end{cases} \qquad \forall i, j \in [K] \tag{11}$$

Experiments

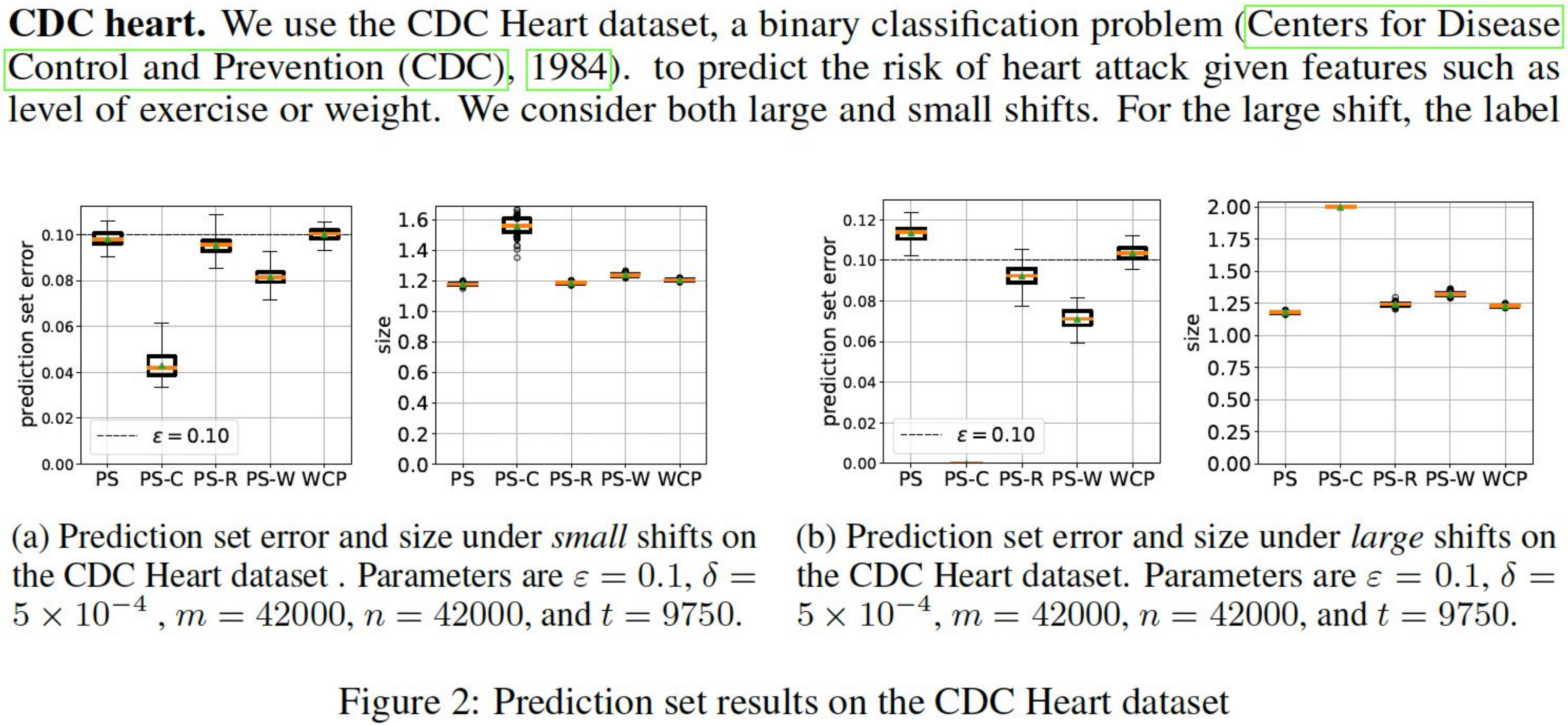


Baselines. We compare our approach (**PS-W**) with several baselines (see Appendix G):

- **PS:** PAC prediction sets that do not account for label shift (Vovk, 2012; Park et al., 2019). This does not come with PAC guarantees under label shift.
- **WCP:** Weighted conformal prediction under label shift, which targets marginal coverage (Podkopaev & Ramdas, 2021). This does not come with PAC guarantees under label shift either.
- **PS-R:** PAC prediction sets that account for label shift but ignore uncertainty in the importance weights; which does not come with PAC guarantees under label shift come with.
- **PS-C:** Addresses label shift via a conservative upper bound on the empirical loss (see Appendix G for details). This is the only baseline to come with PAC guarantees under label shift.

Metrics. We measure performance via the prediction set error, i.e., the fraction of $(x, y) \sim Q$ such that $\boxed{y \notin C_\tau(x)}$; and the average prediction set size, i.e., the mean of $\boxed{|C_\tau(x)|}$, evaluated on the held-out test set. We report the results over 100 independent repetitions, randomizing both dataset generation and our algorithm.

Experiments



CDC heart. We use the CDC Heart dataset, a binary classification problem (Centers for Disease Control and Prevention (CDC), 1984). to predict the risk of heart attack given features such as level of exercise or weight. We consider both large and small shifts. For the large shift, the label

(a) Prediction set error and size under *small* shifts on the CDC Heart dataset . Parameters are $\varepsilon = 0.1$, $\delta = 5 \times 10^{-4}$, $m = 42000$, $n = 42000$, and $t = 9750$.

(b) Prediction set error and size under *large* shifts on the CDC Heart dataset. Parameters are $\varepsilon = 0.1$, $\delta = 5 \times 10^{-4}$, $m = 42000$, $n = 42000$, and $t = 9750$.

Figure 2: Prediction set results on the CDC Heart dataset

CIFAR-100 Dataset

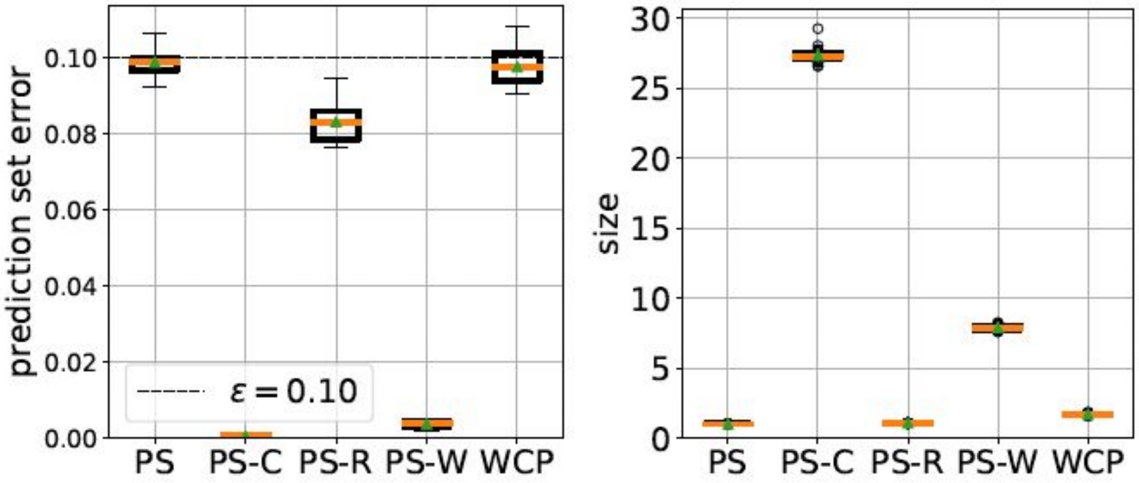


Figure 13: Prediction set error and size on the CIFAR-100 dataset. Parameters are $\varepsilon = 0.1$, $\delta = 5 \times 10^{-4}$, $m = 270k$, $n = 180k$, and $o = 5950$. Label distribution is $([1.01\%] \times 99 + [0.3\%])$ for source, and $([0.84\%] \times 99 + [16.8\%])$ for target.

Conclusions

- We have proposed a PAC prediction set algorithm for the label shift setting, and illustrated its effectiveness in experiments.
- Our approach is focused on problem settings when sufficient calibration data is available; and may produce conservative prediction sets otherwise.
- This reflects the intrinsic difficulty of the problem in these settings.
- Directions for future work include improving performance when the calibration dataset is small.

Thank you!