PAC Prediction Sets Under Covariate Shift

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

[1]University of Pennsylvania [2]Georgia Institute of Technology

Motivation

A standard prediction. A predictor $\hat{y} : \mathcal{X} \to \mathcal{Y}$ is a function from examples to a label. Uncertainty is implicit.

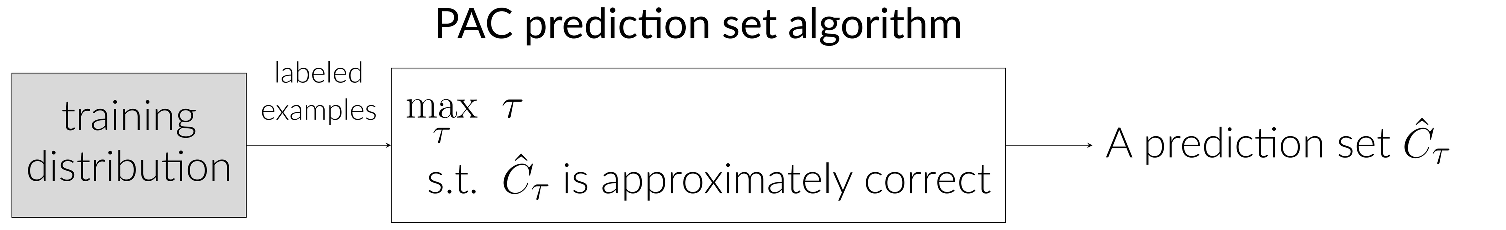
A prediction set. A prediction set $\hat{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ is a set-valued function from examples to a subset of labels (e.g., Wilks (1941); Vovk et al. (2005)). The size represents uncertainty.

PAC prediction sets. Consider a prediction set $\hat{C}$ that satisfies the PAC guarantee, i.e.,

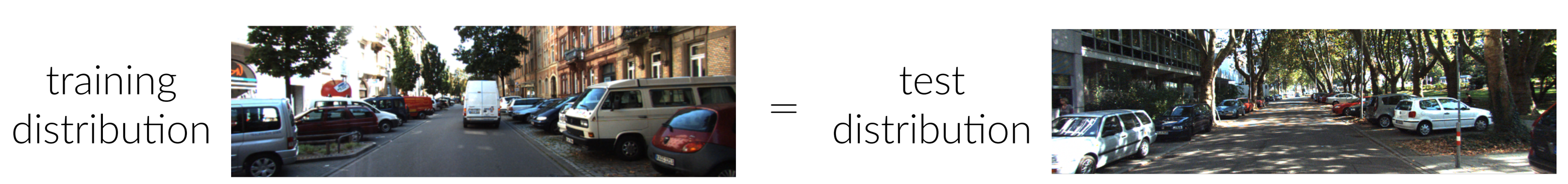
$$\mathbb{P}_{S \sim P^m}\left[\mathbb{P}_{(x,y)\sim P}\left[y \notin \hat{C}(x; S) \right] \leq \epsilon \right] \geq 1 - \delta.$$

PAC prediction sets algorithm (Park et al., 2020a). Consider a scalar-parameter prediction set $\hat{C}_\tau(x) \coloneqq \{y \mid \hat{f}(x,y) \geq \tau\}$, where a score function $\hat{f} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ is given. The following minimizes the expected prediction set size while satisfying the PAC guarantee.

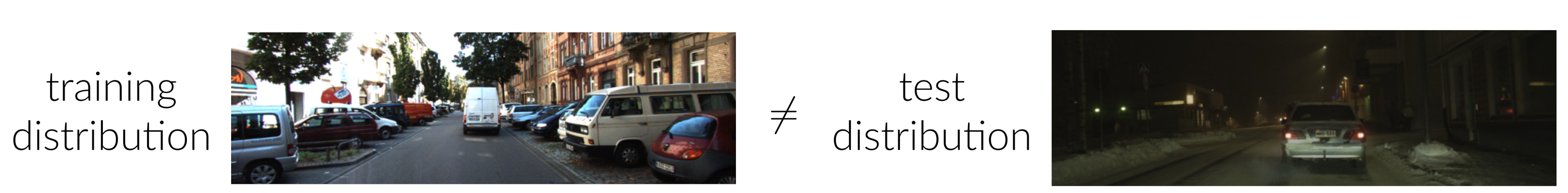
PAC prediction set algorithm



Assumption for PAC prediction sets. Assume training and test distributions are identical (i.e., the i.i.d. assumption)



Covariate shift. Consider training and test covariate distributions can be different.

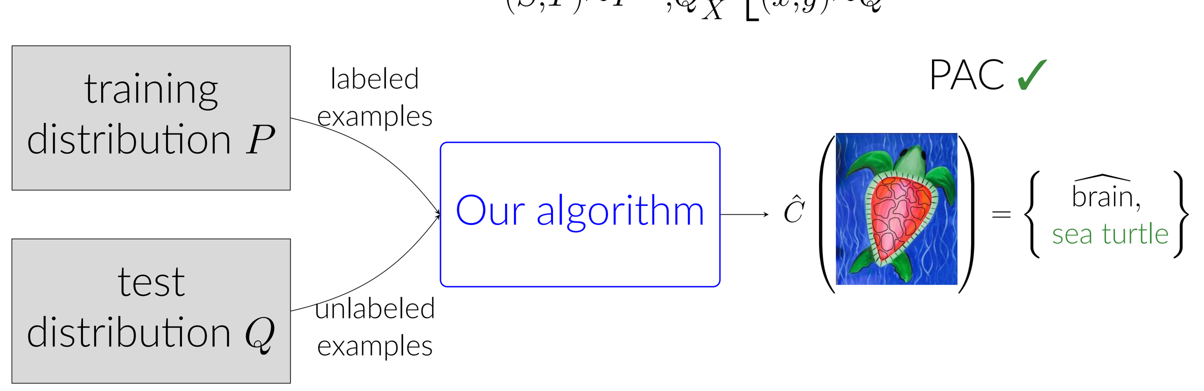


How to achieve the PAC guarantee under covariate shift?

Problem: PAC Prediction Sets under Covariate Shift

Find a probably approximately correct (PAC) prediction set $\hat{C}$, while ensuring its size is small—i.e.,

$$\min_{\hat{C}} \ \mathbb{E}_{x \sim Q_X}\left[\text{size}(\hat{C}(x))\right] \quad \text{subj. to} \quad \mathbb{P}_{(S,T)\sim P^m, Q_X^n}\left[\mathbb{P}_{(x,y)\sim Q}\left[y \notin \hat{C}(x; S, T) \right] \leq \epsilon \right] \geq 1 - \delta.$$

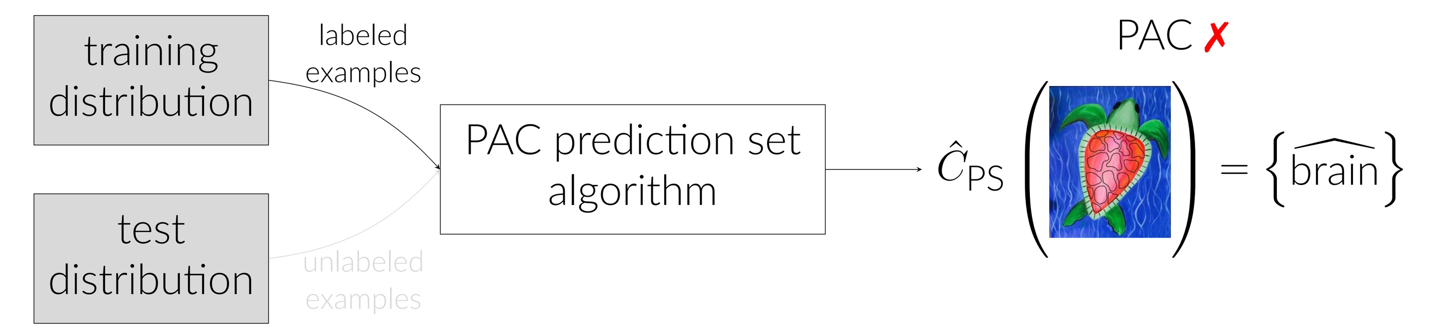


Closely Related Work: Conformal Prediction under Covariate Shift (Tibshirani et al., 2019) considers *fully-unconditional* validity, while we consider *training-conditional* validity.

Existing Approach: Use the Standard PAC Prediction Sets

Main idea. Ignore unlabeled examples from a test distribution, and simply run the standard PAC prediction set algorithm.

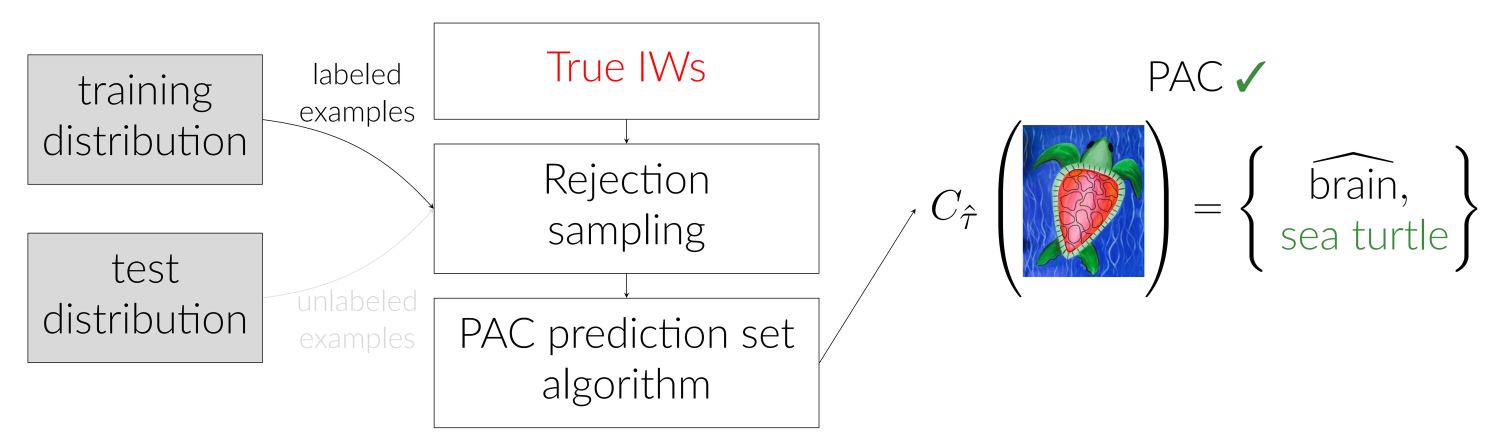
Limitation. The standard algorithm assumes i.i.d. labeled examples, thus the algorithm is not PAC for the target distribution.



Ours Part 1: Use Rejection Sampling + Exact IWs

Main idea. Use rejection sampling (von Neumann, 1951) to generate target labeled examples from source labeled examples by leveraging true IWs.

Limitation. The true IWs are not known in general.



Theorem

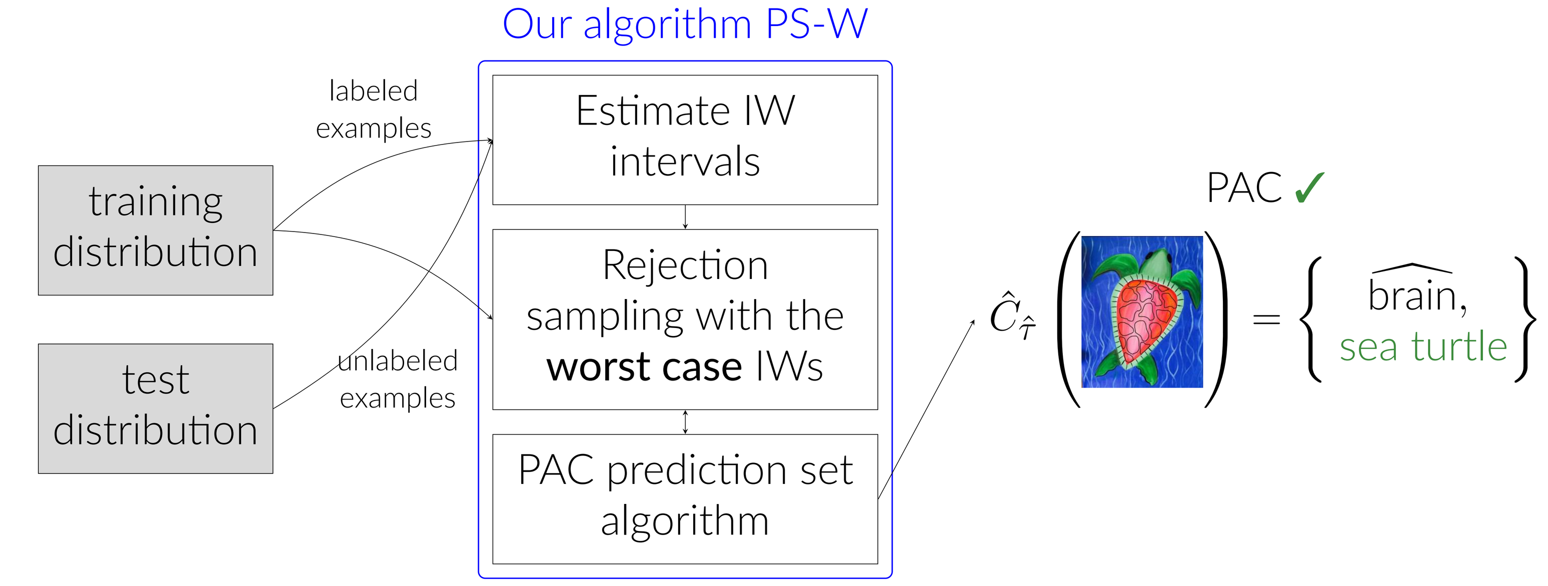
We have $\mathbb{P}\left[L_Q(C_\tau) \leq \varepsilon\right] \geq 1 - \delta$ for any $\tau \leq \hat{\tau}$.

Note that the probability is taken over labeled examples and the randomness of rejection sampling.

Ours Part 2: Use Rejection Sampling + IW Intervals

Main idea. Use IW intervals instead of the true IWs.

Our algorithm PS-W



Theorem

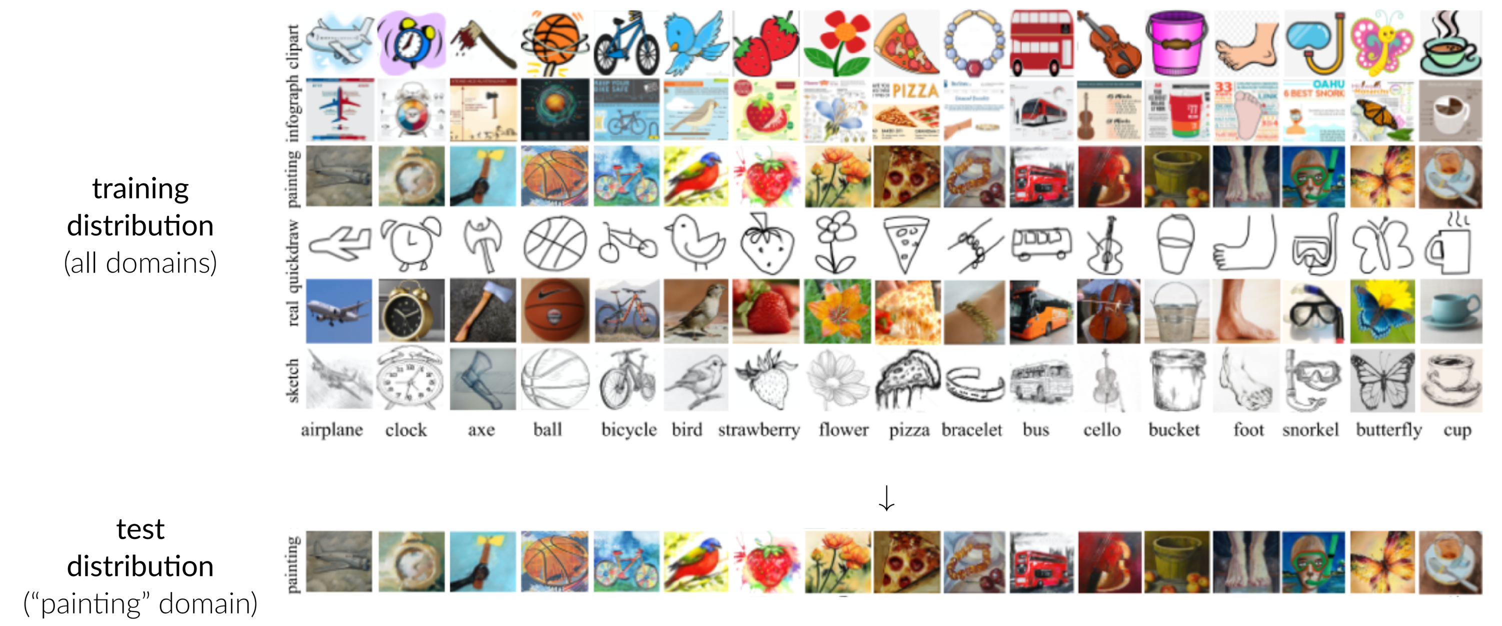
We have $\mathbb{P}[L_Q(C_\tau) \leq \varepsilon] \geq 1 - \delta_C - \delta_w$ for any $\tau \leq \hat{\tau}$.

Note that the probability is taken over labeled and unlabeled examples and the randomness of rejection sampling; the PAC guarantee **accounts for the uncertainty** of IW intervals.

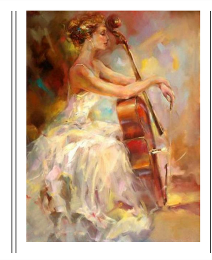
Estimate IW intervals. Please check out our paper!

Experiments on Natural Shift

Natural Shift via DomainNet. We evaluate our approach over the DomainNet dataset (Peng et al., 2019). Parameters are $m = 50,000$, $\varepsilon = 0.1$, and $\delta = 10^{-5}$.



Qualitative Results. The comparison between the standard PAC prediction set $\hat{C}_{\text{PS}}$ and the proposed approach $\hat{C}_{\text{PS-W}}$. The green label is the true label and the label with the hat is the predicted label. The proposed prediction sets include true labels.

Example x	$\hat{C}_{\text{PS}}(x)$	$\hat{C}_{\text{PS-W}}(x)$ (proposed)	Example x	$\hat{C}_{\text{PS}}(x)$	$\hat{C}_{\text{PS-W}}(x)$ (proposed)
	{$\widehat{\text{raccoon}}$}	{owl, $\widehat{\text{raccoon}}$}		{angel, $\widehat{\text{harp}}$}	{angel, cello, $\widehat{\text{harp}}$, microphone, piano, violin}

Experiments on 9 Shifts

PS-W satisfies the PAC guarantee, while producing the smallest prediction sets over 9 shifts.

Shift	Baselines						Ablations				Ours	
	PS		WSCI		PS-C		PS-R		PS-M		PS-W	
	error	size	error	size	error	size	error	size	error	size	error	size
All	✓ (0.094)	10.5	✓ (0.099)	9.5	✓ (0.093)	10.7	✓ (0.094)	10.6	✓ (0.094)	10.8	✓ (0.070)	17.0
Sketch	✗ (0.142)	13.1	✗ (0.116)	18.6	✓ (0.020)	141.7	✓ (0.097)	28.2	✗ (0.105)	26.1	✓ (0.078)	40.3
Painting	✗ (0.159)	15.4	✗ (0.113)	30.0	✓ (0.025)	125.4	✓ (0.096)	37.7	✗ (0.103)	34.5	✓ (0.076)	52.8
Quickdraw	✓ (0.069)	5.9	✓ (0.097)	3.8	✓ (0.021)	23.8	✓ (0.088)	4.3	✓ (0.087)	4.2	✓ (0.067)	6.1
Real	✓ (0.079)	8.7	✓ (0.087)	7.2	✓ (0.032)	47.8	✓ (0.080)	8.7	✓ (0.087)	7.1	✓ (0.068)	11.8
Clipart	✗ (0.105)	10.2	✗ (0.101)	10.9	✓ (0.000)	345.0	✓ (0.080)	19.4	✓ (0.086)	14.8	✓ (0.060)	25.7
Infograph	✗ (0.363)	36.4	✗ (0.114)	165.1	✓ (0.000)	345.0	✓ (0.085)	202.6	✗ (0.107)	177.4	✓ (0.078)	216.4
ImageNet-PGD	✓ (0.090)	5.5	✓ (0.096)	4.7	✓ (0.000)	1000.0	✓ (0.000)	1000.0	✓ (0.074)	7.8	✓ (0.049)	13.9
ImageNet-C13	✗ (0.127)	9.3	✗ (0.111)	67.0	✓ (0.000)	1000.0	✓ (0.000)	1000.0	✓ (0.095)	15.9	✓ (0.061)	35.8
mean normalized size	–		–		0.0338		0.0257		–		**0.0047**	

Acknowledgement

This work was supported in part by ARO W911NF-20-1-0080, AFRL and DARPA FA8750-18-C-0090, NSF award CCF 1910769, and NSF award 2031895 on the Mathematical and Scientific Foundations of Deep Learning (MoDL). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Air Force Research Laboratory (AFRL), the Army Research Office (ARO), the Defense Advanced Research Projects Agency (DARPA), or the Department of Defense, or the United States Government. The authors are grateful to Art Owen for helpful comments.