COLEP: Certifiably Robust Learning-Reasoning Conformal Prediction via Probabilistic Circuits

ICLR 2024

Mintong Kang · Nezihe Merve Gürel · Linyi Li · Bo Li

Vulnerability of Data-driven ML Models

- Vulnerability to adversarial perturbations

Adversarial Image by PGD

Classifier → "Fish" (Adversarial Target)

Solution: Certified Robustness

- Overconfidence:

Classifier → "River"

Solution: Conformal Prediction

Certified Robustness & Conformal Prediction

Adversarial Image with bounded perturbations

Classifier

Prediction always be “Dog”

Certified Robustness

(Test sample X_{n+1})

Conformal Model $\hat{\pi}$

Quantile estimation $\hat{q}_\alpha$ *on calibration set*

Prediction set $C_{n,\alpha}(X_{n+1})$: {Dog, River, Tree}

Conformal guarantee:

$$\mathbb{P}\left[Y_{n+1} \in C_{n,\alpha}(X_{n+1})\right] \geq 1 - \alpha$$

Conformal Prediction

Conformal Prediction (CP)

- n calibration samples $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \mathcal{X} = \mathbb{R}^d$, $Y_i \in \mathcal{Y} = \{1,2,\dots,C\}$, pretrained model $\hat{\pi}\colon \mathbb{R}^d \mapsto \Delta^C$, desired coverage level $1-\alpha \in [0,1]$, prediction set of test sample: $C_{n,\alpha}(X_{n+1})$
- Non-conformity score of sample: $S_{\hat{\pi}}(X_i, Y_i) \in [0,1]$
  - Measures how much non-conformity each sample has regarding the ground truth label
  - E.g., $S_{\hat{\pi}}(x, y) = 1 - \hat{\pi}_y(x)$
- Conformal prediction guarantee:
  - $\mathbb{P}\left[Y_{n+1} \in C_{n,\alpha}(X_{n+1})\right] \geq 1-\alpha$, where $C_{n,\alpha}(X_{n+1}) = \{y \in \mathcal{Y}\colon S_{\hat{\pi}}(X_{n+1}, y) \leq Q_{1-\alpha}(\{S_{\hat{\pi}}(X_i, Y_i)\}_{i=1}^n)\}$, where $Q_{1-\alpha}(\cdot)$ computes the $1-\alpha$ empirical quantile value

Conformal Prediction (CP)

(Test sample X_{n+1})

Conformal Model
$\hat{\pi}$

Prediction set
$C_{n,\alpha}(X_{n+1})$:
{Dog, River, Tree}

Quantile estimation
$\hat{q}_{\alpha}$
on calibration set

Conformal guarantee:

$$\mathbb{P}\left[Y_{n+1} \in C_{n,\alpha}(X_{n+1})\right] \geq 1 - \alpha$$

- Requirement: test distribution is identical to the calibration distribution
- Conformal guarantee is broken in the adversary setting
  - The adversary can add imperceptible noises to the test sample during inference time

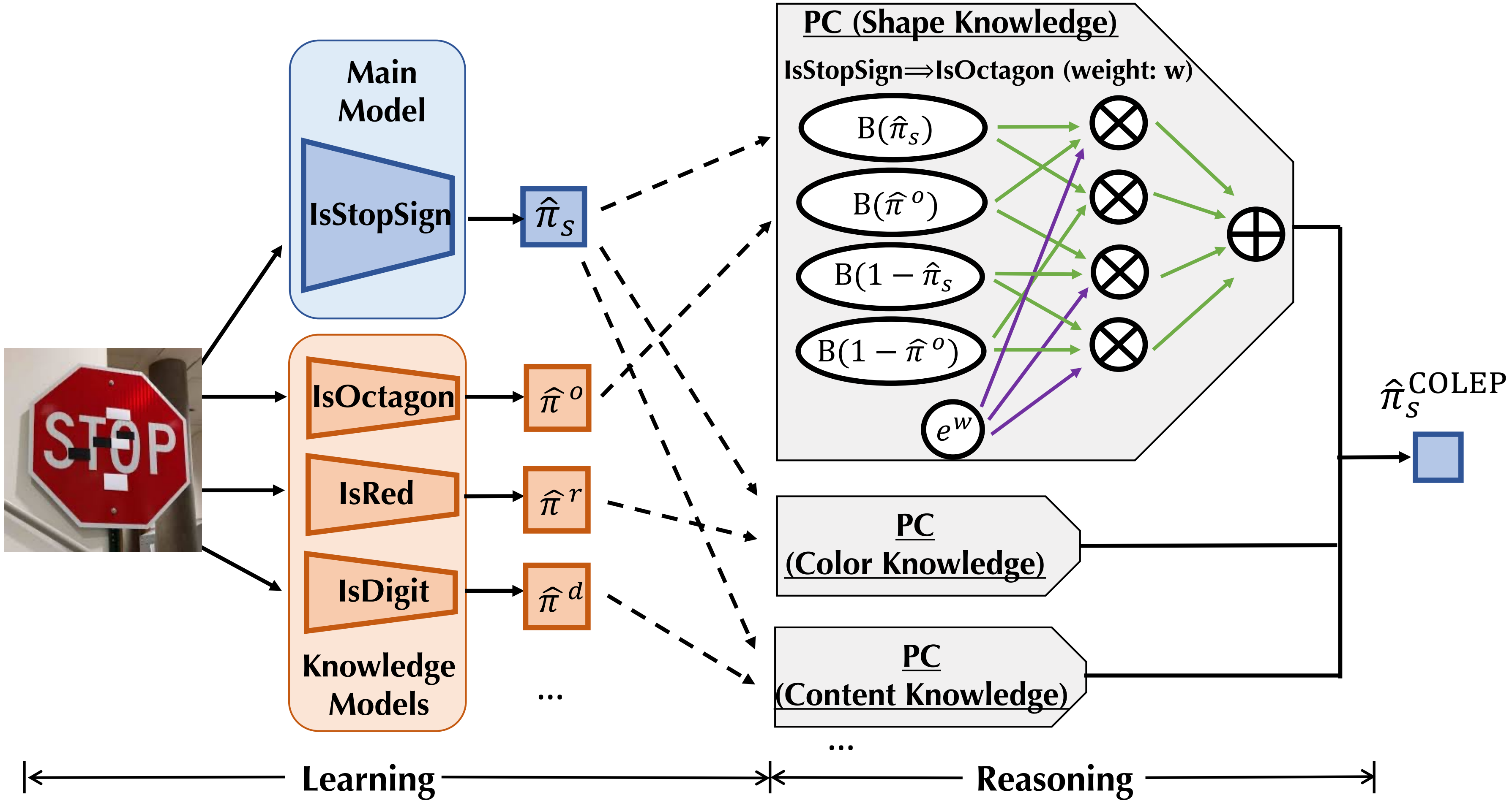
Standard CP is broken with adversary

Table I: Marginal coverage under ℓ_2 attack on GTSRB. The benign coverage is 0.9.

	$\boldsymbol{\delta = 0.125}$	$\boldsymbol{\delta = 0.25}$	$\boldsymbol{\delta = 0.5}$
Standard CP	0.3118	0.0484	0.0028
Smoothing CP	0.8306	0.7504	0.5478
COLEP (ours)	**0.9508**	**0.9324**	**0.8804**

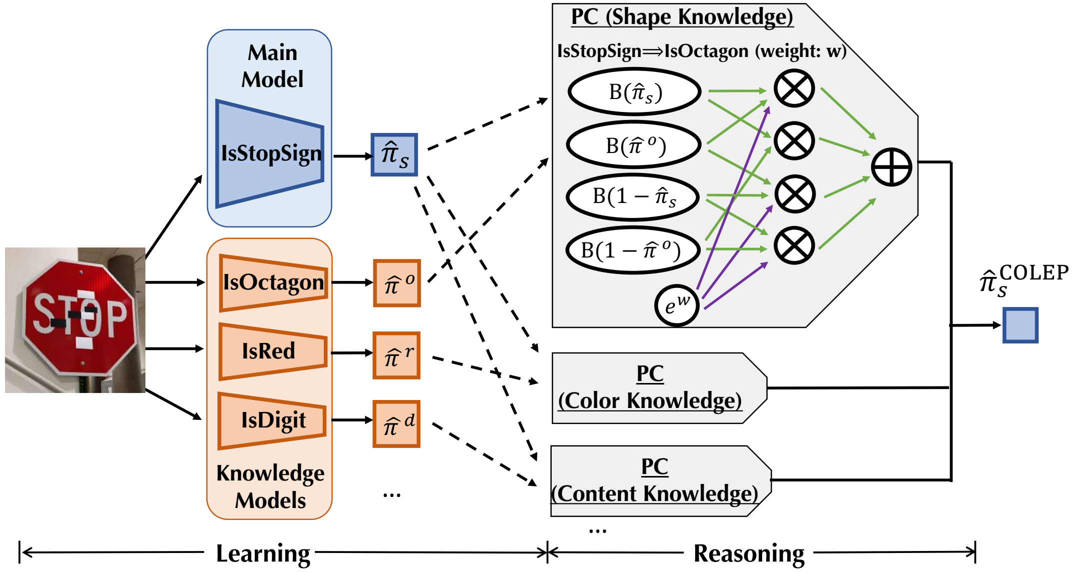
- Data-driven conformal model is vulnerable
- COLEP (ours):
  - Certifiably Robust Learning-Reasoning Conformal Prediction via Probabilistic Circuits
  - Integrate domain knowledge into the conformal prediction framework

Learning-reasoning CP framework via probabilistic circuit (PC)



Learning-reasoning CP framework

Stop sign is octagon: IsStopSign⟹IsOctagon (with weight w)
IsStopSign: Bernoulli random variable with success rate $\hat{\pi}_s$
IsOctagon: Bernoulli random variable with success rate $\hat{\pi}_o$



IsStopSign	IsOctagon	Likelihood
0	0	$(1-\hat{\pi}_s)(1-\hat{\pi}_o)e^w$
0	1	$\hat{\pi}_s(1-\hat{\pi}_o)e^w$
1	0	$(1-\hat{\pi}_s)\hat{\pi}_o$
1	1	$\hat{\pi}_s\hat{\pi}_o e^w$

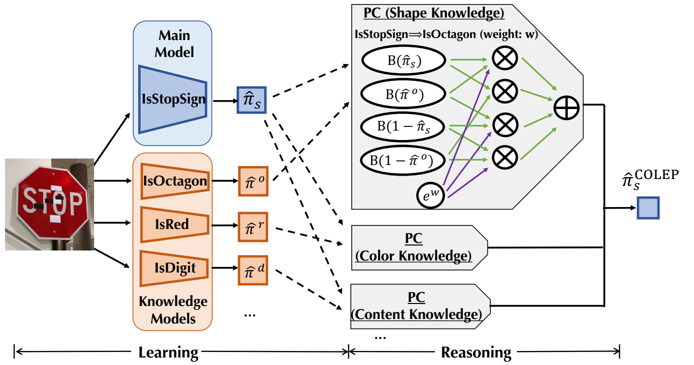
The likehood $p(\text{IsStopSign}=1,\text{IsOctagon}=0)$ is down-weighted by the correction of the knowledge rule.

Marginal Probability:

$$p(\text{IsStopSign}=1) = \frac{p(\text{IsStopSign=1,IsOctagon=0})+p(\text{IsStopSign=1,IsOctagon=1})}{p(\text{IsStopSign=0,IsOctagon=0})+p(\text{IsStopSign=0,IsOctagon=1})+p(\text{IsStopSign=1,IsOctagon=0})+p(\text{IsStopSign=1,IsOctagon=1})}$$

Time complexity linear to the size of PC graph

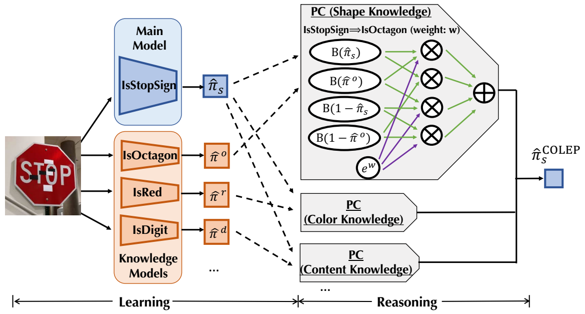
Learning-reasoning CP framework



- Formally:
  consider N_c class labels (main model) and L knowledge labels (knowledge models)
  $\mu \epsilon M = \{0,1\}^{N_c+L}$: a possible assignment
  $O_{\mathrm{root}}(\mu)$: output of PC given μ, indicating likelihood of assignment μ
  $F(\mu) = \exp\left\{\sum_{h=1}^{H} w_h \mathbb{I}[\mu \sim K_h]\right\}$: factor function, where $K_h$is the h-th rule with weight w_h
  $\mathbb{I}[\mu \sim K_h] = 1$ if assignment μ satisfies knowledge rule K_h
  let $T(a,b) = \log\left(ab + (1-a)(1-b)\right)$

$$\hat{\pi}_j^{\mathrm{COLEP}}(x) = \frac{\sum_{\mu \epsilon M, \mu_j=1} O_{\mathrm{root}}(\mu)}{\sum_{\mu \epsilon M} O_{\mathrm{root}}(\mu)} = \frac{\sum_{\mu \epsilon M, \mu_j=1} \exp\left\{\sum_{j'=1}^{N_c+L} T\left(\hat{\pi}_{j'}(x), \mu_{j'}\right)\right\} F(\mu)}{\sum_{\mu \epsilon M} \exp\left\{\sum_{j'=1}^{N_c+L} T\left(\hat{\pi}_{j'}(x), \mu_{j'}\right)\right\} F(\mu)}$$

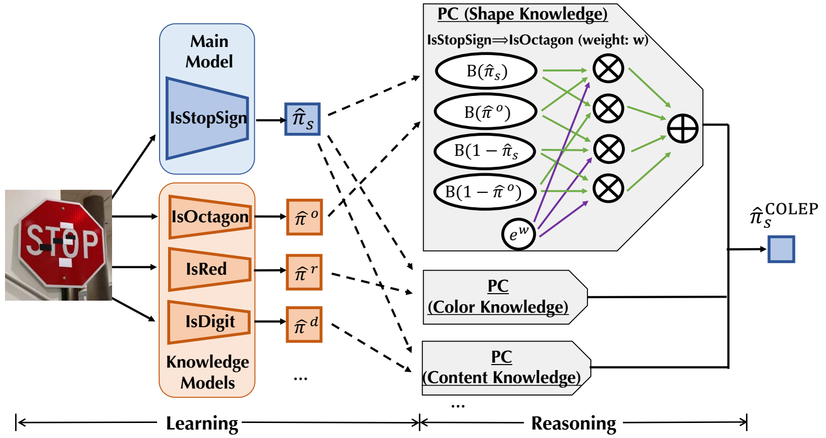
Learning-reasoning CP framework



- Conformal prediction with probability estimator $\hat{\pi}_j^{\text{COLEP}}$
  - Step 1: class-wise conformal prediction
    $\hat{C}_{n,\alpha_j}^{\text{COLEP}_j}(X_{n+1}) = \left\{ q^y \in \{0,1\}: S_{\hat{\pi}_j^{\text{COLEP}}}(X_{n+1}, q^y) \le Q_{1-\alpha_j}\left(\left\{ S_{\hat{\pi}_j^{\text{COLEP}}}(X_i, \mathbb{I}[Y_i = j]) \right\}_{i \in \mathcal{I}_{cal}} \right) \right\}$
  - Step 2: Final prediction set
    $$\hat{C}_{n,\alpha}^{\text{COLEP}}(X_{n+1}) = \left\{ j \in [N_c]: 1 \in \hat{C}_{n,\alpha_j}^{\text{COLEP}_j}(X_{n+1}) \right\}$$
  - Recall the conformal guarantee:
    $\mathbb{P}\left[Y_{n+1} \in \hat{C}_{n,\alpha}^{\text{COLEP}}(X_{n+1}) \right] \ge 1 - \alpha$
- Problem:
  - Conformal guarantee is broken with adversary $\tilde{X}_{n+1} = X_{n+1} + \varepsilon$
- Question:
  - What is the valid conformal guarantee with perturbation $\|\varepsilon\| < \delta$?

Certifiably robust learning-reasoning CP

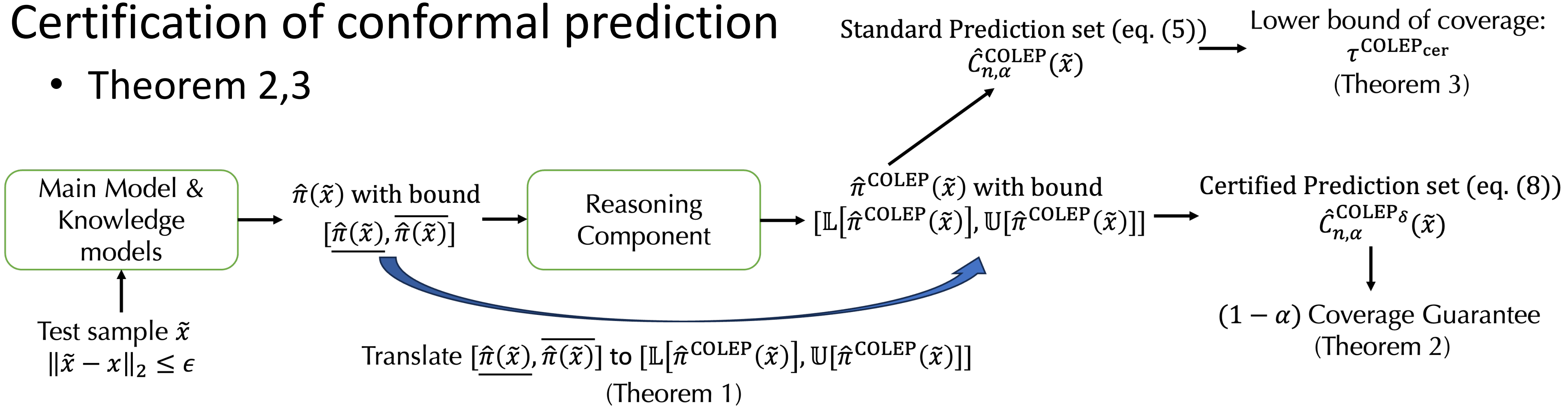
- Inference stage:
  - Learning component:
    - Compute class probability $\hat{\pi}$ for the main model and knowledge models
  - Reasoning component:
    - Compute corrected class probability $\hat{\pi}^{\text{COLEP}}$
  - Conformal prediction:
    - Compute final prediction set using $\hat{\pi}^{\text{COLEP}}$



- Certification Goal:
  - For adversary $\tilde{X}_{n+1} = X_{n+1} + \varepsilon$ with $\|\varepsilon\| < \delta$, construct and certify the prediction set with the desired coverage $1 - \alpha$.

Certification framework in COLEP

- End-to-end certification framework
  - Robustness certification of the learning component
    - Probabilistic certification: randomized smoothing
    - Deterministic certification: bound propagation approaches (e.g., CROWN-IBP)
  - Robustness certification framework of the reasoning component (PC)
    - Theorem 1
  - Certification of conformal prediction
    - Theorem 2,3



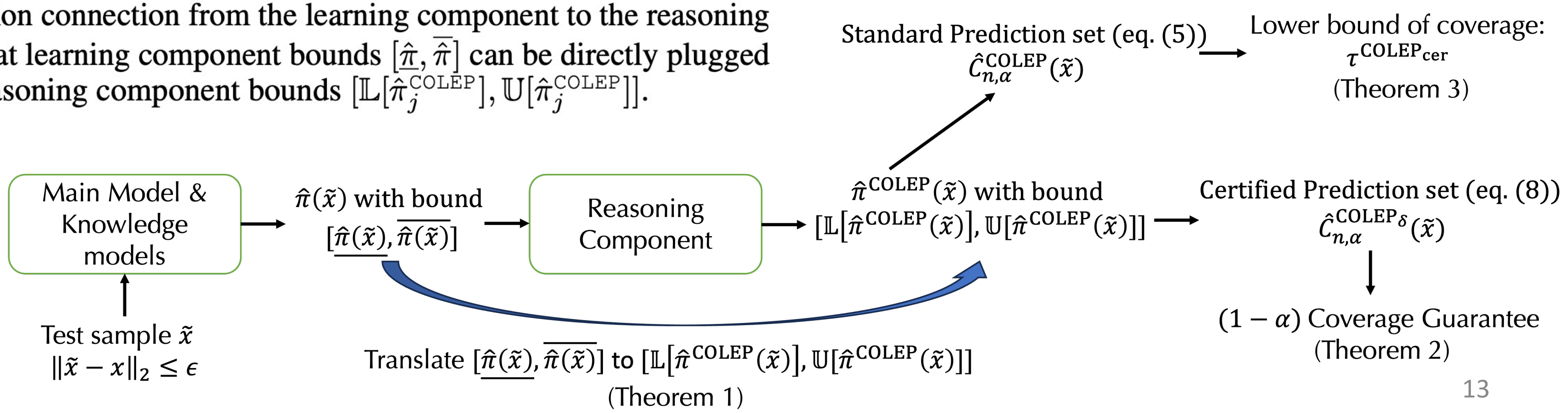
Certification of the reasoning component

Theorem 1 (Bounds for Conditional Class Probabilities $\hat{\pi}_j^{\texttt{COLEP}}(x)$ within the Reasoning Component)**.** *Given any input x and perturbation bound δ, we let $[\underline{\hat{\pi}_{j_\forall}}(x), \overline{\hat{\pi}_{j_\forall}}(x)]$ be bounds for the estimated conditional class and concept probabilities by all models with $j_\forall \in [N_c + L]$ (for example, achieved via randomized smoothing). Let V_d^j be the set of index of conditional variables in the PC except for $j \in [N_c]$ and V_s^j be that of consequence variables. Then the bound for* `COLEP`*-corrected estimate of the conditional class probability $\hat{\pi}_j^{\texttt{COLEP}}$ is given by:*

$$\mathbb{U}[\hat{\pi}_j^{\texttt{COLEP}}(x)] = \left\{ \frac{(1 - \overline{\hat{\pi}_j}(x)) \sum_{\mu_j = 0} \exp\left\{ \sum_{j_\forall \in V_d^j} T(\overline{\hat{\pi}_{j_\forall}}(x), \mu_{j_\forall}) + \sum_{j_\forall \in V_s^j} T(\underline{\hat{\pi}_{j_\forall}}(x), \mu_{j_\forall}) \right\} F(\mu)}{\overline{\hat{\pi}_j}(x) \sum_{\mu_j = 1} \exp\left\{ \sum_{j_\forall \in V_d^j} T(\underline{\hat{\pi}_{j_\forall}}(x), \mu_{j_\forall}) + \sum_{j_\forall \in V_s^j} T(\overline{\hat{\pi}_{j_\forall}}(x), \mu_{j_\forall}) \right\} F(\mu)} + 1 \right\}^{-1} \tag{6}$$

where $T(a, b) = \log(ab + (1-a)(1-b))$. We similarly give the lower bound $\mathbb{L}[\hat{\pi}_j^{\texttt{COLEP}}]$ in Appendix E.1.

Remarks. Thm. 1 establishes a certification connection from the learning component to the reasoning component. In other words, we show that learning component bounds $[\underline{\hat{\pi}}, \overline{\hat{\pi}}]$ can be directly plugged into a closed-form formula to obtain reasoning component bounds $[\mathbb{L}[\hat{\pi}_j^{\texttt{COLEP}}], \mathbb{U}[\hat{\pi}_j^{\texttt{COLEP}}]]$.



Certifiably robust conformal prediciton

Theorem 2 (Certifiably Robust Conformal Prediction of COLEP). *Consider a new test sample X_{n+1} drawn from P_{XY}. For any bounded perturbation $\|\epsilon\|_2 \leq \delta$ in the input space and the adversarial sample $\tilde{X}_{n+1} := X_{n+1} + \epsilon$, we have the following guaranteed marginal coverage:*

$$\mathbb{P}[Y_{n+1} \in \hat{C}_{n,\alpha}^{COLEP_\delta}(\tilde{X}_{n+1})] \geq 1 - \alpha \tag{7}$$

if we construct the certified prediction set of COLEP where

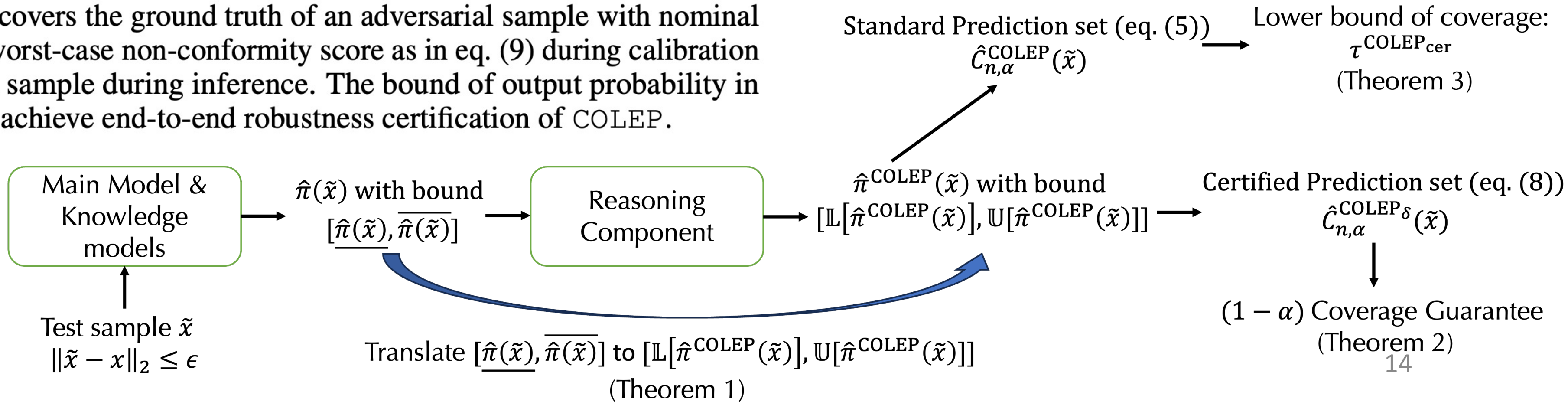
$$\hat{C}_{n,\alpha}^{COLEP_\delta}(\tilde{X}_{n+1}) = \left\{ j \in [N_c] : S_{\hat{\pi}_j^{COLEP}}(\tilde{X}_{n+1}, 1) \leq Q_{1-\alpha}(\{S_{\hat{\pi}_j^{COLEP_\delta}}(X_i, \mathbb{I}_{[Y_i=j]})\}_{i \in \mathcal{I}_{cal}}) \right\} \tag{8}$$

and $S_{\hat{\pi}_j^{COLEP_\delta}}(\cdot, \cdot)$ is a function of worst-case non-conformity score considering perturbation radius δ:

$$S_{\hat{\pi}_j^{COLEP_\delta}}(X_i, \mathbb{I}_{[Y_i=j]}) = \begin{cases} \mathbb{U}_\delta[\hat{\pi}_j^{COLEP}(X_i)] + u(1 - \mathbb{U}_\delta[\hat{\pi}_j^{COLEP}(X_i)]), & Y_i \neq j \\ 1 - \mathbb{L}_\delta[\hat{\pi}_j^{COLEP}(X_i)] + u\mathbb{L}_\delta[\hat{\pi}_j^{COLEP}(X_i)], & Y_i = j \end{cases} \tag{9}$$

with $\mathbb{U}_\delta[\hat{\pi}_j^{COLEP}(x)] = \max_{|\eta|_2 \leq \delta} \hat{\pi}_j^{COLEP}(x + \eta)$ and $\mathbb{L}_\delta[\hat{\pi}_j^{COLEP}(x)] = \min_{|\eta|_2 \leq \delta} \hat{\pi}_j^{COLEP}(x + \eta)$.

Remarks. Thm. 2 shows that the coverage guarantee of COLEP in the adversary setting is still valid if we construct the prediction set by considering the worst-case perturbation as in eq. (8). That is, the prediction set of COLEP in eq. (8) covers the ground truth of an adversarial sample with nominal level $1 - \alpha$. To achieve that, we use a worst-case non-conformity score as in eq. (9) during calibration to counter the influence of adversarial sample during inference. The bound of output probability in eq. (9) can be computed by Thm. 1 to achieve end-to-end robustness certification of COLEP.



Worst-case coverage of COLEP

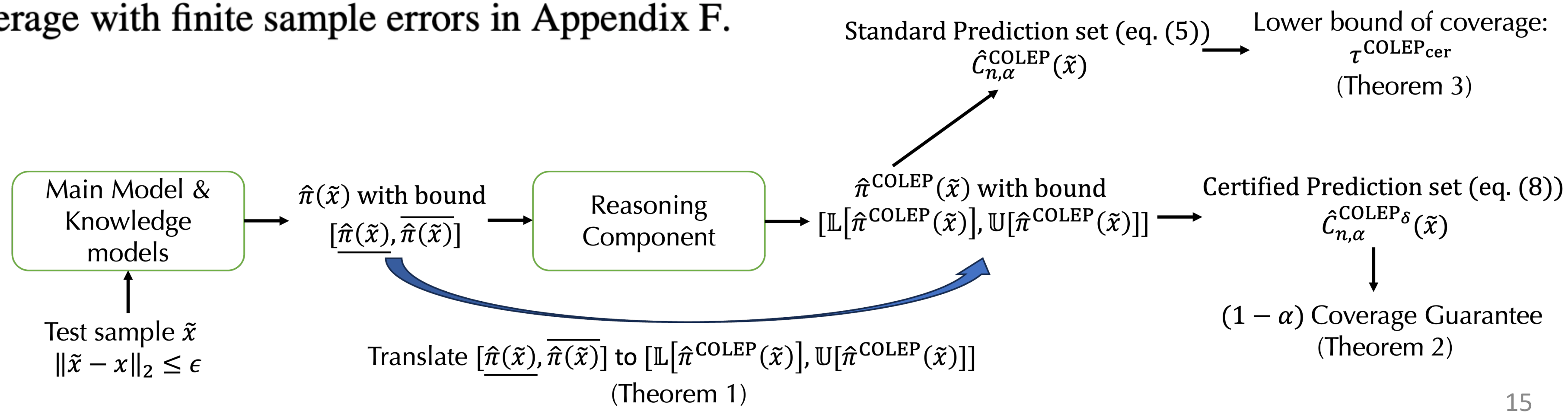
Theorem 3 (Certified (Worst-Case) Coverage of COLEP). *Consider the new sample X_{n+1} drawn from P_{XY} and adversarial sample $\tilde{X}_{n+1} := X_{n+1} + \epsilon$ with any perturbation $\|\epsilon\|_2 \leq \delta$ in the input space. We have:*

$$\mathbb{P}[Y_{n+1} \in \hat{C}_{n,\alpha}^{COLEP}(\tilde{X}_{n+1})] \geq \tau^{COLEP_{cer}} := \min_{j \in [N_c]} \left\{ \tau_j^{COLEP_{cer}} \right\}, \tag{10}$$

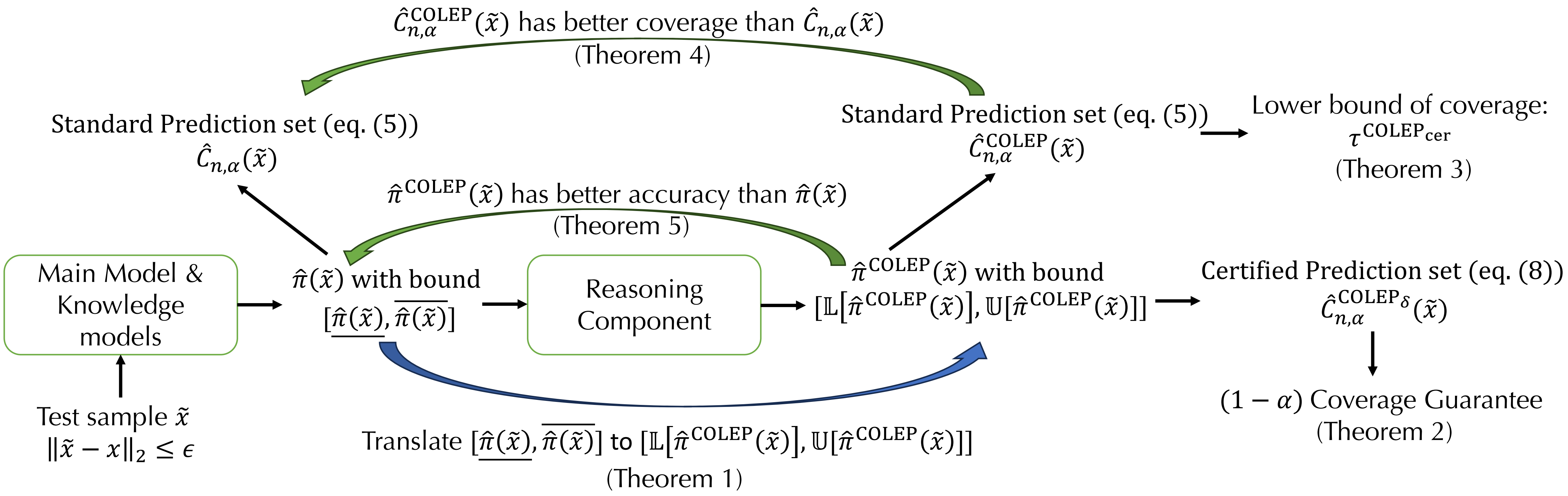
where the certified (worst-case) coverage of the j-th class label $\tau_j^{COLEP_{cer}}$ is formulated as:

$$\tau_j^{COLEP_{cer}} = \max \left\{ \tau : Q_\tau(\{S_{\hat{\pi}_j^{COLEP_\delta}}(X_i, \mathbb{I}_{[Y_i=j]})\}_{i \in \mathcal{I}_{cal}}) \leq Q_{1-\alpha}(\{S_{\hat{\pi}_j^{COLEP}}(X_i, \mathbb{I}_{[Y_i=j]})\}_{i \in \mathcal{I}_{cal}}) \right\}. \tag{11}$$

Remarks. Recall that thm. 2 constructs a certified prediction set as eq. (8) considering the worst-case perturbation and proves that the prediction set has $1-\alpha$ coverage guarantees. In contrast, thm. 3 provides a lower bound to the coverage of COLEP ($\tau_j^{\mathrm{COLEP_{cer}}}$) with the standard prediction set as eq. (5) in the adversary setting. In addition to the certified coverage, we consider finite calibration set size and certified coverage with finite sample errors in Appendix F.



How does the reasoning component benefit in COLEP?



COLEP achieves higher marginal coverage than a standard conformal model

Theorem 4 (Comparison of Marginal Coverage of COLEP and Main Model). *Consider the adversary setting that the calibration set $\mathcal{I}_{cal}$ consists of $n_{\mathcal{D}_b}$ samples drawn from the benign distribution $\mathcal{D}_b$, while the new sample (X_{n+1}, Y_{n+1}) is drawn $n_{\mathcal{D}_a}$ times from the adversarial distribution $\mathcal{D}_a$. Assume that $A(\hat{\pi}_j, \mathcal{D}_a) < 0.5 < A(\hat{\pi}_j, \mathcal{D}_b)$ for $j \in [N_c]$, where $A(\hat{\pi}_j, \mathcal{D})$ is the expectation of prediction accuracy of $\hat{\pi}_j$ on $\mathcal{D}$. Then we have:*

$$\mathbb{P}[Y_{n+1} \in \hat{C}^{COLEP}_{n,\alpha}(\tilde{X}_{n+1})] > \mathbb{P}[Y_{n+1} \in \hat{C}_{n,\alpha}(\tilde{X}_{n+1})], \quad w.p.$$

$$1 - \max_{j \in [N_c]} \{\exp\left\{-2n_{\mathcal{D}_a}(0.5 - A(\hat{\pi}_j, \mathcal{D}_a))^2 \boldsymbol{\epsilon}^2_{j,1,\mathcal{D}_a}\right\} + n_{\mathcal{D}_b} \exp\{-2n_{\mathcal{D}_b}\left((A(\hat{\pi}_j, \mathcal{D}_b) - 0.5)\sum_{c \in \{0,1\}} p_{jc} \boldsymbol{\epsilon}_{j,c,\mathcal{D}_b}\right)^2\}\} \tag{17}$$

where $p_{j0} = \mathbb{P}_{\mathcal{D}_b}[\mathbb{I}_{[Y \neq j]}]$ and $p_{j1} = \mathbb{P}_{\mathcal{D}_b}[\mathbb{I}_{[Y = j]}]$ are class probabilities on benign distribution.

Remarks. Thm. 4 shows that COLEP can achieve better marginal coverage than a single model with a high probability exponentially approaching 1. The probability increases in particular with a higher quality of models represented by $\boldsymbol{\epsilon}_{j,1,\mathcal{D}_a}$, $\boldsymbol{\epsilon}_{j,c,\mathcal{D}_b}$, $A(\hat{\pi}_j, \mathcal{D}_b)$. It also increases with lower $A(\hat{\pi}_j, \mathcal{D}_a)$, indicating COLEP improves marginal coverage more likely in a stronger adversary setting.

COLEP achieves higher prediction accuracy than a single standard ML model

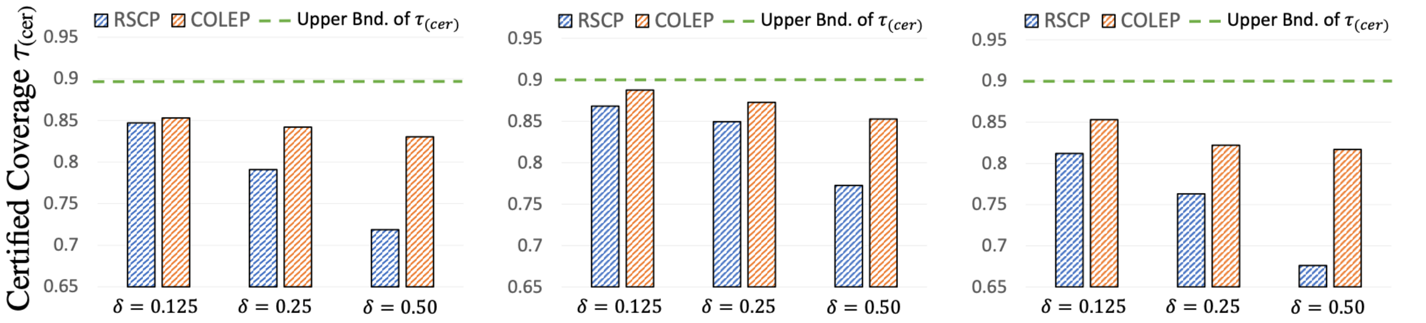
Theorem 5 (Comparison of Prediction Accuracy of `COLEP` and Main Model)**.** *Suppose that we evaluate the expected prediction accuracy of $\hat{\pi}_j^{COLEP}(\cdot)$ and $\hat{\pi}_j(\cdot)$ on n samples drawn from $\mathcal{D}_m$ and denote the prediction accuracy as $A(\hat{\pi}_j^{COLEP}(\cdot), \mathcal{D}_m)$ and $A(\hat{\pi}_j(\cdot), \mathcal{D}_m)$. Then we have:*

$$A(\hat{\pi}_j^{COLEP}(\cdot), \mathcal{D}_m) \geq A(\hat{\pi}_j(\cdot), \mathcal{D}_m), \quad \textit{w.p. } 1 - \sum_{\mathcal{D} \in \{\mathcal{D}_a, \mathcal{D}_b\}} p_{\mathcal{D}} \sum_{c \in \{0,1\}} \mathbb{P}_{\mathcal{D}}[Y = j] \exp\left\{-2n(\boldsymbol{\epsilon}_{j,c,\mathcal{D}})^2\right\}. \quad (18)$$

Remarks. Thm. 5 shows that `COLEP` achieves better prediction accuracy than the main model with a high probability exponentially approaching 1. The probability increases with a high utility of models and knowledge rules (i.e., a large $\boldsymbol{\epsilon}_{j,c,\mathcal{D}}$). In Appendix H, we further show that `COLEP` achieves higher prediction accuracy with more useful knowledge rules.

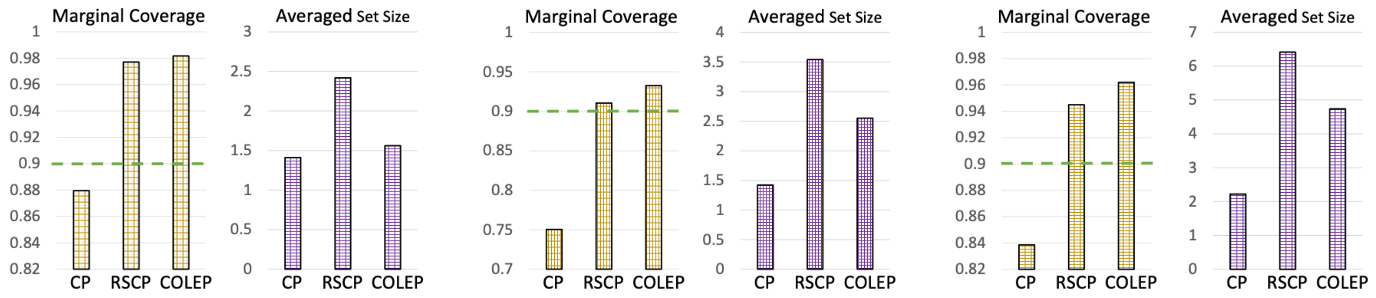
Evaluation

- Certified coverage
  - Baseline: RSCP[1] data-driven smoothed conformal model
  - COLEP achieves higher certified coverage than RSCP
- Marginal coverage under PGD
  - Metric: marginal coverage, average set size
  - Baseline: CP, RSCP
  - Coverage is broken with CP
  - COLEP achieves better tradeoff between coverage and efficiency than RSCP



(a) GTSRB (b) CIFAR-10 (c) AwA2

Figure 2: Comparison of certified coverage between COLEP ($\tau^{\mathrm{COLEP}_{\mathrm{cer}}}$) and RSCP under bounded perturbations $\delta = 0.125, 0.25, 0.50$ on GTSRB, CIFAR-10, and AwA2. The upper bound of certified coverage $\tau_{(\mathrm{cer})}$ is 0.9.



(a) GTSRB (b) CIFAR-10 (c) AwA2

Figure 3: Comparison of the marginal coverage and averaged set size for CP, RSCP, and COLEP under PGD attack ($\delta = 0.25$) on GTSRB, CIFAR-10, and AwA2. The nominal coverage level (green line) is 0.9.

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

Conclusion

- A certifiably robust conformal prediction framework via knowledge-enabled logical reasoning: COLEP
- Derive the conformal guarantee with COLEP
- Prove that with the reasoning component, COLEP achieves better coverage/prediction accuracy than a single standard ML model
- Empirically show the validity and effectiveness of COLEP on GTSRB, CIFAR-10, and AwA2