Robust Conformal Prediction Using Privileged Information

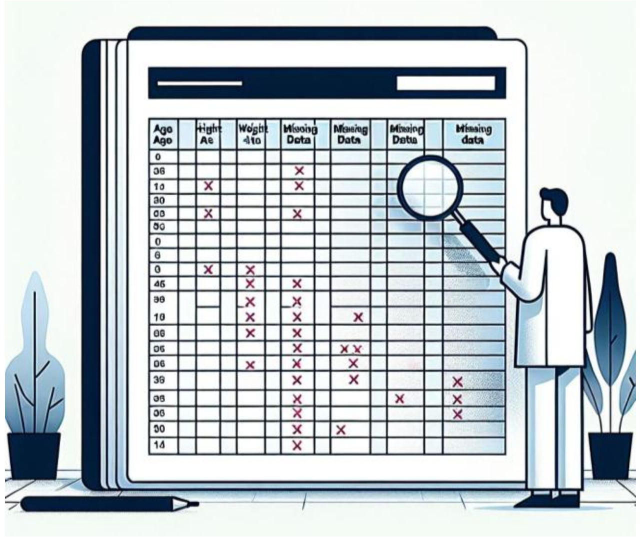
Shai Feldman, Yaniv Romano

Technion – Israel Institute of Technology

Various forms of corrections

- Noisy labels
- Missing values
- Low-quality data, uncertainty
- Sensor noise
- Failing measuring equipment

1. Dough (ImageNet label)
2. Pizza
3. Soup bowl
4. …



No 100% accurate data

→ corrupted samples

Uncertainty is inevitable!

Setup

- **Input:** n training points $\{(X_i, Y_i^{\text{obs}}, Z_i, M_i)\}_{i=1}^{n}$ and a test point $(X_{\text{test}}, ?)$

  → exchangeable (e.g., i.i.d.) samples from unknown joint dist.
- $X \in \mathcal{X}$: features

- $Y^{\text{obs}} \in \mathcal{Y}$: observed label/response

- $Y \in \mathcal{Y}$: ground truth label
- $Z \in \mathcal{Z}$: privileged information (PI) - available only during training time
  - E.g., The annotator's level of expertise
- $M \in \{0,1\}$: noise indicator $M = 1 \Leftrightarrow Y^{\text{obs}}$ is noisy
- Assumption: the PI Z explains the corruption appearances $(X, Y) \perp M \mid Z$

* See paper for a more general framework covering missing or noisy features and labels.

Ultimate goal: reliable UQ under corruptions

- **Input:** n training points $\{(X_i, Y_i^{\text{obs}}, Z_i, M_i)\}_{i=1}^{n}$ and a test point $(X_{\text{test}}, ?)$
  → exchangeable (e.g., i.i.d.) samples from unknown joint dist.
- $X_{\text{test}} = X_{n+1} \in \mathcal{X}$: clean test features
- $Y_{\text{test}} = Y_{n+1} \in \mathcal{Y}$: clean, unknown, test response

Wish to use any ML algorithm to construct a marginal **distribution-free prediction set**

$$\mathbb{P}[Y_{\text{test}} \in C(X_{\text{test}})] \geq 1 - \alpha \quad \text{(e.g., 90\%)}$$

$\alpha \in (0,1)$ is a user-specified miscoverage rate

Ultimate goal: reliable UQ under corruptions

- **Input:** n training points $\{(X_i, Y_i^{\text{obs}}, Z_i, M_i)\}_{i=1}^{n}$ and a test point $(X_{\text{test}}, ?)$
  → exchangeable (e.g., i.i.d.) samples from unknown joint dist.
- $X_{\text{test}} = X_{n+1} \in \mathcal{X}$: clean test features
- $Y_{\text{test}} = Y_{n+1} \in \mathcal{Y}$: clean, unknown, test response

Wish to use any ML algorithm to construct a marginal **distribution-free prediction set**

$$\mathbb{P}[Y_{\text{test}} \in C(X_{\text{test}})] \geq 1 - \alpha \quad \text{(e.g., 90\%)}$$

$\alpha \in (0,1)$ is a user-specified miscoverage rate

- Construct $C(X_{\text{test}})$ using the *observed* corrupted data
- Guarantee that clean Y_{test} is covered in $C(X_{\text{test}})$

how and under what conditions is it possible?

Background on conformal prediction

Conformal prediction [Vovk et al. ’99; Papadopoulos et al. ‘12, Lei et al. ’18; …]

- **Input:** pre-trained predictive model $\hat{f}$, and holdout calibration set $\{(X_i, Y_i)\}_{i=1}^{n}$
- **Process**
  - Compute non-conformity scores $s_i = S(X_i, Y_i)$ for all i

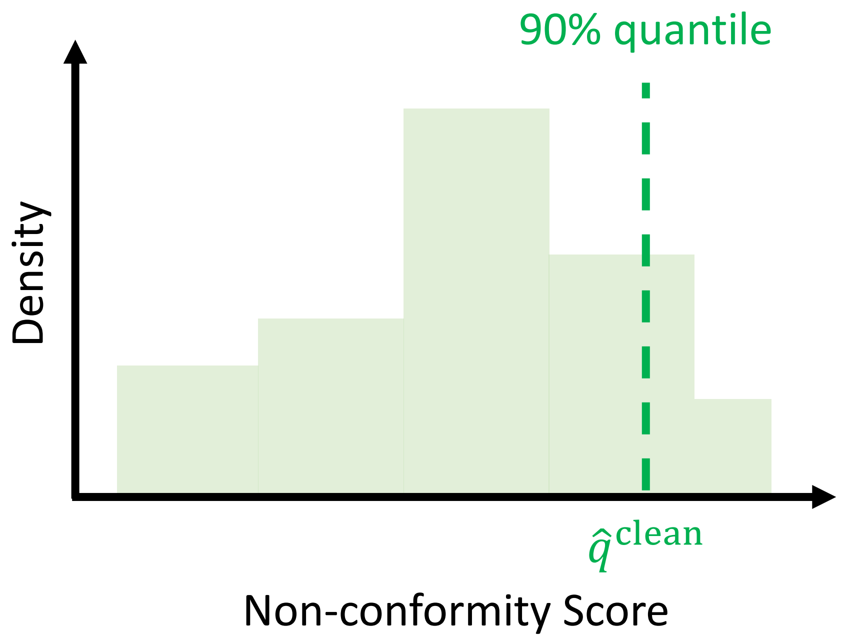
    a measure of goodness-of-fit (the lower the better), e.g., $s_i = \left|\hat{f}(X_i) - Y_i\right|$

Conformal prediction [Vovk et al. '99; Papadopoulos et al. '12, Lei et al. '18; …]

- **Input:** pre-trained predictive model $\hat{f}$, and holdout calibration set $\{(X_i, Y_i)\}_{i=1}^{n}$
- **Process**
  - Compute non-conformity scores $s_i = S(X_i, Y_i)$ for all i
  - Compute* $\hat{q}^{\text{clean}} =$ the $(1-\alpha)$-empirical quantile of $\{s_i\}_{i=1}^{n}$
- **Output:** prediction set

$$C\left(X_{\text{test}}, \hat{q}^{\text{clean}}\right) = \left\{y : s(X_{\text{test}}, y) \leq \hat{q}^{\text{clean}}\right\}$$

Sweep over all $y \in \mathcal{Y}$ and return the guessed y's whose score falls below $\hat{q}^{\text{clean}}$



*missing a small correction term

Conformal prediction is valid under exchangeability

<u>Theorem</u> (Vovk et al. '99; Papadopoulos et al. '12; Lei et al. '18; R., Patterson, Candes '19, …)

If $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X_{\text{test}}, Y_{\text{test}})$ are exch. Then,

$$\mathbb{P}\left[Y_{\text{test}} \in C\left(X_{\text{test}}, \hat{q}^{\text{clean}}\right)\right] \geq 1 - \alpha \quad \text{(e.g., 90\%)}$$

\+ Exchangeability is the only assumption

\- Assumes that the training data is clean

Weighted conformal prediction [Tibshirani et al. ’19]

- We consider only the scores of non-corrupted samples and **weight** their distribution by the ratio of likelihoods between the test and train data:

$$w(z) = \frac{\mathbb{P}(M = 0)}{\mathbb{P}(M = 0 \mid Z = z)} \quad \Rightarrow \text{accounts for distr. shift}$$

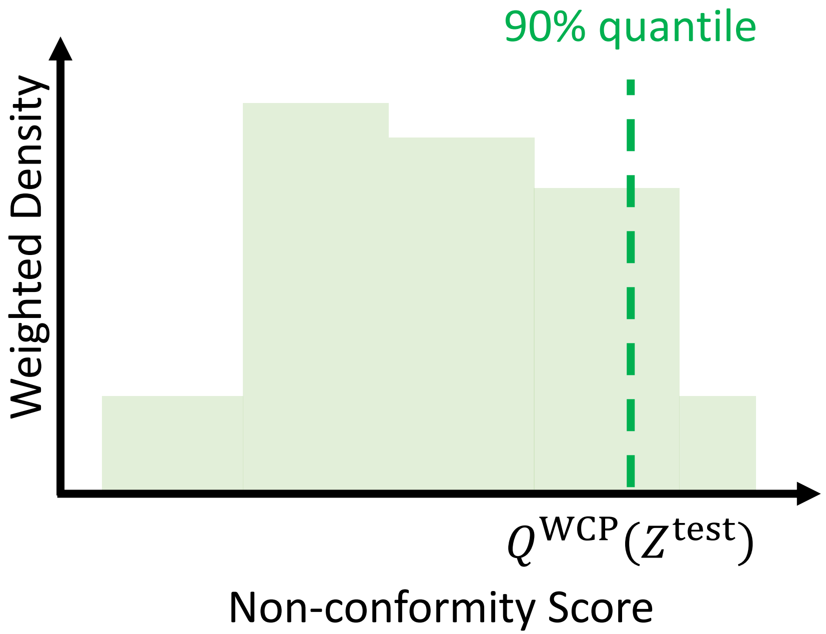
***Note**: Here, only uncorrupted data points are used, as they reflect the true distribution of the scores under covariate shift.

Weighted conformal prediction [Tibshirani et al. ’19]

- We consider only the scores of non-corrupted samples and **weight** their distribution by the ratio of likelihoods between the test and train data:

$$w(z) = \frac{\mathbb{P}(M = 0)}{\mathbb{P}(M = 0 \mid Z = z)}$$

- The threshold $Q^{\mathrm{WCP}}(Z^{\mathrm{test}})$ is the $1 - \alpha$ empirical quantile of the **weighted distribution** of the uncorrupted samples’ scores



Weighted conformal prediction [Tibshirani et al. ’19]

- We consider only the scores of non-corrupted samples and **weight** their distribution by the ratio of likelihoods between the test and train data:

$$w(z) = \frac{\mathbb{P}(M = 0)}{\mathbb{P}(M = 0 \mid Z = z)}$$

- The threshold $Q^{\text{WCP}}(Z^{\text{test}})$ is the $1 - \alpha$ empirical quantile of the **weighted distribution** of the uncorrupted samples’ scores
- The prediction set is constructed as

$$C^{\text{WCP}}(X^{\text{test}}, Z^{\text{test}}) = \{y\colon S(X^{\text{test}}, y) \leq Q^{\text{WCP}}(Z^{\text{test}})\}$$

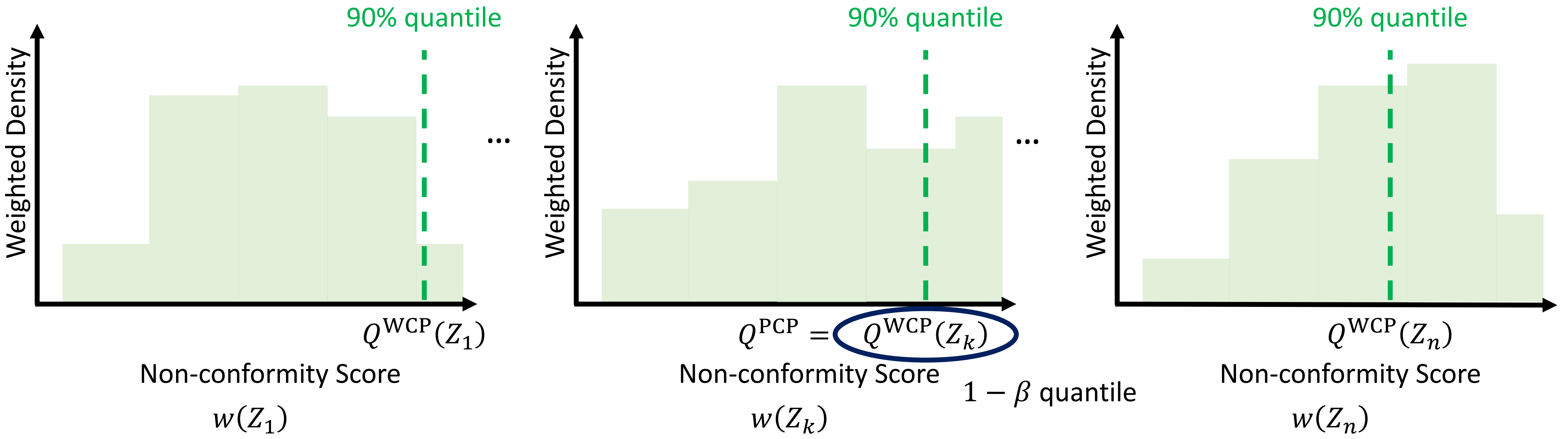
+ Achieves the desired coverage level even under presence of corrupted samples!

\- Infeasible! Requires access to the unknown Z^{test}

Proposed method: Privileged Conformal Prediction

Privileged conformal prediction

- Apply WCP on each calibration point to obtain a corresponding threshold $Q^{\mathrm{WCP}}(Z_i)$ for the i-th sample
- Take Q^{PCP} as the $(1-\beta)$-empirical quantile of $\{Q^{\mathrm{WCP}}(Z_i)\}_{i=1}^{n}$



Privileged conformal prediction

- Apply WCP on each calibration point to obtain a corresponding threshold $Q^{\mathrm{WCP}}(Z_i)$ for the i-th sample
- Take Q^{PCP} as the $(1-\beta)$-empirical quantile of $\left\{Q^{\mathrm{WCP}}(Z_i)\right\}_{i=1}^{n}$
- Construct the prediction set for Y_{test}

$$C^{\mathrm{PCP}}(X_{\mathrm{test}}) = \{y : S(X_{\mathrm{test}}, y) \leq Q^{\mathrm{PCP}}\}$$

Privileged conformal prediction

- Apply WCP on each calibration point to obtain a corresponding threshold $Q^{\text{WCP}}(Z_i)$ for the i-th sample
- Take Q^{PCP} as the $(1-\beta)$-empirical quantile of $\left\{Q^{\text{WCP}}(Z_i)\right\}_{i=1}^{n}$
- Construct the prediction set for Y_{test}

$$C^{\text{PCP}}(X_{\text{test}}) = \{y: S(X_{\text{test}}, y) \leq Q^{\text{PCP}}\}$$

$\{w(Z_i)\}_i$ are exch. + Q is an increasing function

$\Rightarrow Q^{\text{PCP}}$ is conservative $Q^{\text{WCP}}(Z^{\text{test}})$

$\Rightarrow$ PCP is valid

Privileged conformal prediction is valid

<u>Theorem</u>

If $\{(X_i, Y_i, Z_i, M_i)\}_{i=1}^{n+1}$ are exch., and P_Z is absolutely continuous with respect to $P_{Z|M=0}$, then,

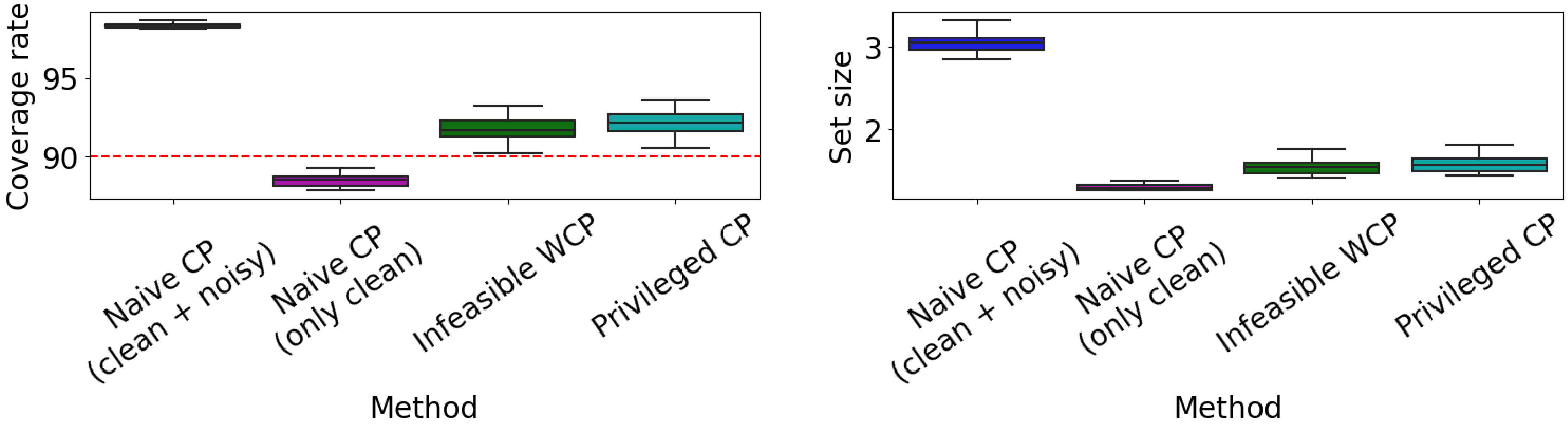
$$\mathbb{P}\left[Y_{\text{test}} \in C^{\text{PCP}}(X_{\text{test}})\right] \geq 1 - \alpha$$

+ Finite sample, dist. free guarantee!
+ Does not require Z^{test}!

Application: noisy labels

Experiment: CIFAR-10N – noisy labels

- Task: classify the object in an image ($K = 10$ classes)
- Clean Y: the correct object label
- Observed Y^{obs}: obtained by a single human annotator (incorrect for $M = 1$)
- PI Z = information about the annotator.



Conclusion and uncovered topics

Conclusion

- Proposed PCP to handle imperfect data using PI
- PCP achieves comparable performance to the infeasible WCP
- Coverage rate is supported by theoretical guarantees

Uncovered topics (ongoing work)

- Adaptation of PCP for scarce data
- Is PCP robust to inaccurate weights?
- Is PCP still valid if the PI Z does not satisfy the conditional independence assumption?
  - $(X, Y) \perp M \mid Z$

Thank you!