The Ohio State University

A Generic Framework for Conformal Fairness

Aditya Vadlamani*, Anutam Srinivasan*, Pranav Maneriker, Ali Payani, Srinivasan Parthasarathy

Conformal Prediction

Goal: Given a coverage rate of $1-\alpha \in (0,1)$, we want to construct a prediction set $\mathcal{C}$ such that the true label for an unseen test point is in $\mathcal{C}$ with probability $1-\alpha$.

Theorem (Vovk et al, 2005)

Given a non-conformity score function $s: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and holdout out calibration set $\mathcal{D}_{calib} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_n, y_n)\}$, let $q = \text{Quantile}\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}, \{s(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}\right)$ and $\mathcal{C}_q(\boldsymbol{x}) = \{y \in \mathcal{Y}, s(\boldsymbol{x}, y) \leq q\}$. Then,

$$1-\alpha \leq \Pr\left[y_{n+1} \in \mathcal{C}_q(\boldsymbol{x}_{n+1})\right] \leq 1-\alpha+\frac{1}{n+1}$$

Group Fairness

- Intuitive notion of fairness which requires that different groups are treated equally
- Well-established metrics including:
  - Demographic (Statistical) Parity
  - Equal Opportunity
  - Predictive Equality
  - Equalized Odds

Conformal Group Fairness Metrics

- We define conformal variations of the metrics for multiclass classification
- Let $\mathcal{C}_\lambda(X)$ be the prediction set for X given threshold λ
  - We change $\hat{Y} = \tilde{y}$ to be $\tilde{y} \in \mathcal{C}_\lambda(X)$.

Metric	Definition
Demographic (or Statistical) Parity	$\Pr\left[\tilde{y} \in \mathcal{C}_\lambda(X) \,\middle\vert\, X \in g_a\right] = \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(X) \,\middle\vert\, X \in g_b\right], \forall g_a, g_b \in \mathcal{G}, \forall \tilde{y} \in \mathcal{Y}^+$
Equal Opportunity	$\Pr\left[\tilde{y} \in \mathcal{C}_\lambda(X) \,\middle\vert\, Y = \tilde{y}, X \in g_a\right] = \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(X) \,\middle\vert\, Y = \tilde{y}, X \in g_b\right], \forall g_a, g_b \in \mathcal{G}, \forall \tilde{y} \in \mathcal{Y}^+$
Predictive Equality	$\Pr\left[\tilde{y} \in \mathcal{C}_\lambda(X) \,\middle\vert\, Y \neq \tilde{y}, X \in g_a\right] = \Pr\left[\tilde{y} \in \mathcal{C}_\lambda(X) \,\middle\vert\, Y \neq \tilde{y}, X \in g_b\right], \forall g_a, g_b \in \mathcal{G}, \forall \tilde{y} \in \mathcal{Y}^+$
Equalized Odds	Equal Opp. and Pred. Equality

Motivation and Intuition

- In practice, achieving perfect fairness can be challenging or even impossible [Barocas, 2023].

- Instead, we control the ***fairness disparity*** between groups and labels to be within a closeness criterion, c.
  - Ex. For Demographic Parity, we want

$$|\Pr[y_{n+1} \in \mathcal{C}_\lambda(\boldsymbol{x}_{n+1}) \,|\boldsymbol{x}_{n+1} \in g_a] - \Pr[y_{n+1} \in \mathcal{C}_\lambda(\boldsymbol{x}_{n+1}) \,|\boldsymbol{x}_{n+1} \in g_b]| < c$$

Key Theoretical Insights

1. CP guarantee holds when applied to a **subset** of D_{calib}
   - **Intuition:** Fairness is evaluated on groups within the population

2. Using the **inverse quantile function**, we can recover the coverage level, for a **given** threshold λ
   - **Intuition:** We want to reverse the CP process to recover coverage

3. CP guarantee holds when considering **any fixed label**
   - **Intuition:** Essential to balance disparity between groups **for all** positive labels

Key Theoretical Insights

1. CP guarantee holds when applied to a **subset** of D_{calib}
   - **Intuition:** Fairness is evaluated on groups within the population

2. Using the **inverse quantile function**, we can recover the coverage level, for a **given** threshold λ
   - **Intuition:** We want to reverse the CP process to recover coverage

3. CP guarantee holds when considering **any fixed label**
   - **Intuition:** Essential to balance disparity between groups **for all** positive labels

Key Theoretical Insights

1. CP guarantee holds when applied to a **subset** of D_{calib}
   - **Intuition:** Fairness is evaluated on groups within the population

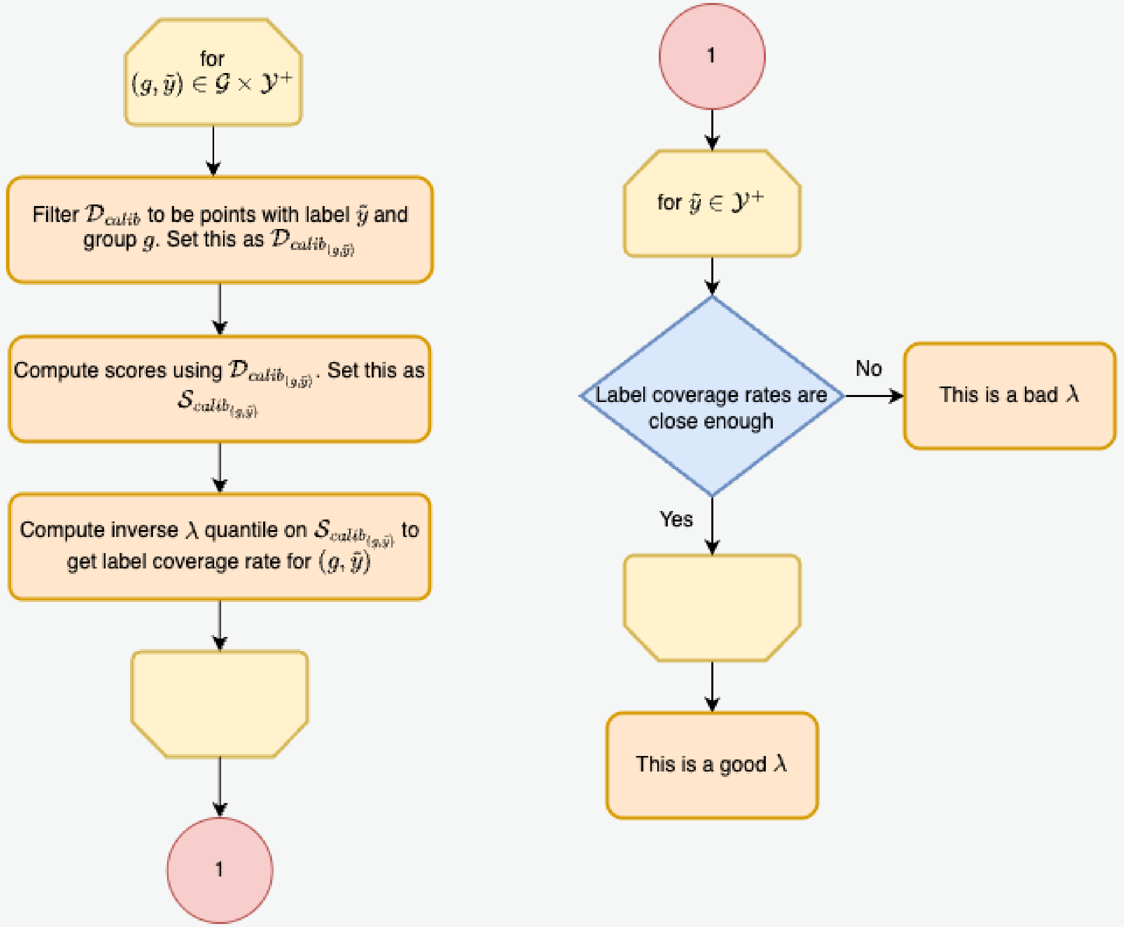
2. Using the **inverse quantile function**, we can recover the coverage level, for a **given** threshold λ
   - **Intuition:** We want to reverse the CP process to recover coverage

3. CP guarantee holds when considering **any fixed label**
   - **Intuition:** Essential to balance disparity between groups **for all** positive labels

Algorithm

- We evaluate if a threshold λ can control the fairness disparity for every group and positive label pair within some closeness criterion, c.

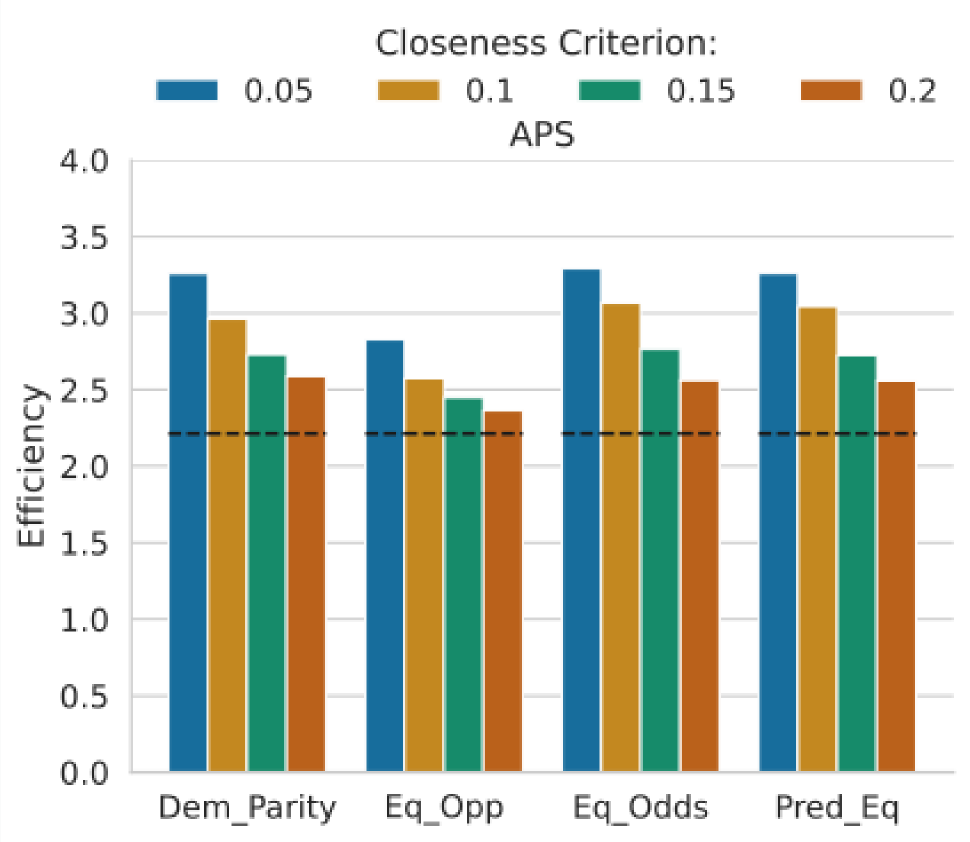
- For the best efficiency, we choose the smallest λ, which still gives the base CP guarantee



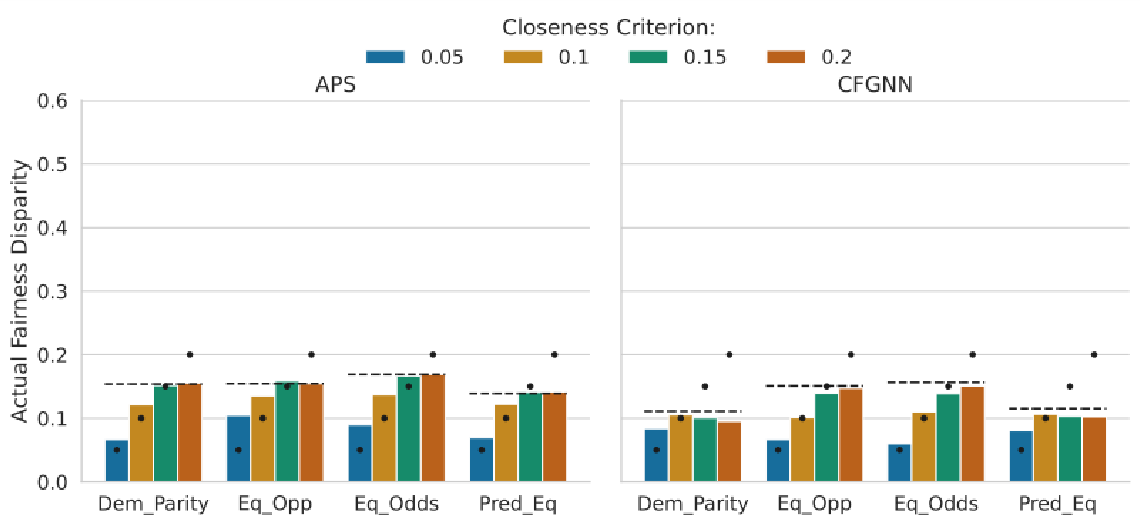
Experimental Setup

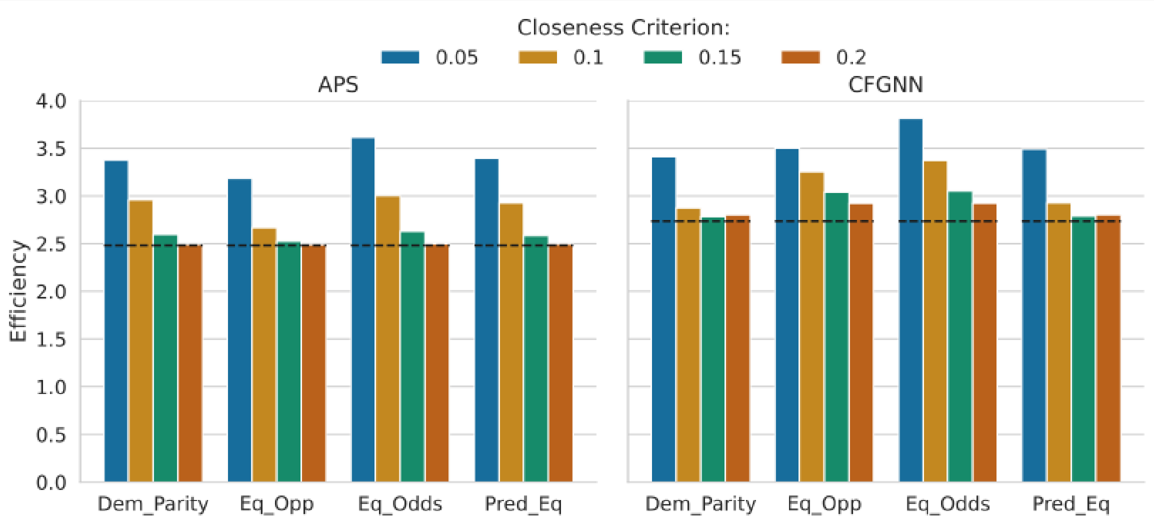
- Conduct experiments on five different datasets, including a mix of graph and tabular datasets
  - See the paper for the complete set of results
- **Baseline:** Conformal predictor using the quantile threshold
- **Evaluation Metrics:**
  - **Worst-case fairness disparity** (↓)**:** Greatest coverage difference between every pair of groups and positive labels
  - **Efficiency** (↓)**:** Prediction set size

Results (ACSIncome with APS)



Results (Pokec-n with APS): Intersectional Fairness





Auditing

- In addition to ***ensuring fairness*** of conformal predictors, our framework can also ***audit fairness*** for a user given closeness criterion

- Our framework can audit fairness for both known and black-box conformal predictors.
  - For black-box predictors, we can use a separate (exchangeable) $\mathcal{D}_{audit}$ set for evaluation

Framework Extensibility

The algorithm can be modified to:

1. Determine thresholds **for each class** independently to get better efficiency (prediction set sizes)
2. Accommodate other (user-defined) metrics with minor changes (e.g., Predictive Parity Proxy).

We don’t require group information at inference time as this isn’t necessarily available in online settings.

- A key differentiator of our work from existing works in group conditional CP [Gibbs 2023, Jung 2023, Lu 2022]

THE OHIO STATE UNIVERSITY

Thank you

CISCO

Aditya Vadlamani

Anutam Srinivasan

Pranav Maneriker

Ali Payani

Srinivasan Parthasarathy

Acknowledgements:
NSF AI-Edge grant (#2112471)
Cisco Research grant (US202581249)

Source Code

Contact:
vadlamani.12@osu.edu,
srinivasan.268@osu.edu
srini@cse.ohio-state.edu

References

[1] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. Algorithmic Learning in a Random World, Volume 29. Springer.
[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023
[3] Isaac Gibbs, John J Cherian, and Emmanuel J Candes. Conformal prediction with conditional guarantees. arXiv preprint arXiv:2305.12616, 2023.
[4] Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. 11th International Conference on Learning Representations (ICLR), 2023.
[5] Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11):12008–12016, Jun. 2022.