# Simultaneous Conformal Prediction of Missing Outcomes with Propensity Score $\varepsilon$-Discretization

Edgar Dobriban

Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

August 8, 2024

# Collaborators

Yonghoon Lee

Eric Tchetgen Tchetgen

# Table of Contents

# Introduction: Conformal prediction

- Major developing area in statistics: distribution-free predictive inference (a.k.a. conformal prediction)

# Introduction: Conformal prediction

- Major developing area in statistics: distribution-free predictive inference (a.k.a. conformal prediction)
- Goal, given $(X_1, Y_1), \ldots, (X_n, Y_n)$, find a prediction set $C$ such that for new $X_{n+1}$, $\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$ under *minimal assumptions*

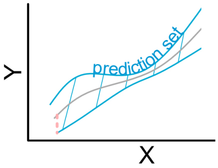



**Figure:** Towards DS

# Introduction: Conformal prediction

- Major developing area in statistics: distribution-free predictive inference (a.k.a. conformal prediction)
- Goal, given $(X_1, Y_1), \ldots, (X_n, Y_n)$, find a prediction set $C$ such that for new $X_{n+1}$, $\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$ under *minimal assumptions*

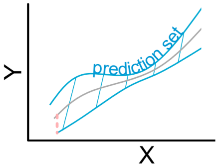
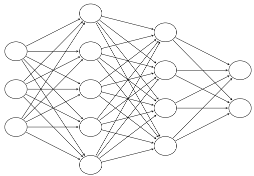



**Figure:** Towards DS



- Motivated by complex applications, e.g., where a machine learning model $\hat{\mu}$ is used to predict $Y_{n+1}$ based on $X_{n+1}$ (not known how to find distribution of $Y_{n+1} - \hat{\mu}(X_{n+1})$)

# Introduction: Conformal prediction

- It is known how to achieve this in many settings, due to extensive work by many, starting in the 90s (Vovk, Wasserman, J. Lei, R. J. Tibshirani, Barber, Candes, ... )
- Ideas date back to work on tolerance regions by Wilks, Wald, Tukey ... starting in the 1940s

Samuel S. Wilks

Abraham Wald

Vladimir Vovk

# Conformal prediction ctd.

- Typical setting: *exchangeable datapoints.*
    - For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, e.g., $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \ldots, s(X_{n+1}, Y_{n+1})$ are exchangeable (if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)

# Conformal prediction ctd.

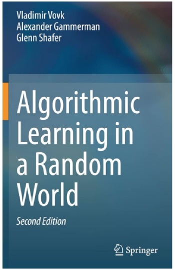
- Typical setting: *exchangeable datapoints.*
  - For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, e.g., $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \dots, s(X_{n+1}, Y_{n+1})$ are exchangeable (if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)
  - Hence, the rank of $s(X_{n+1}, Y_{n+1})$ is uniform over $1, \dots, n+1$ (if no ties)

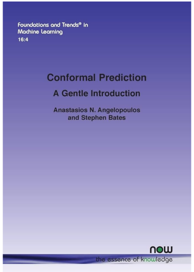
# Conformal prediction ctd.

- Typical setting: *exchangeable datapoints.*
  - For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, e.g., $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \ldots, s(X_{n+1}, Y_{n+1})$ are exchangeable (if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)
  - Hence, the rank of $s(X_{n+1}, Y_{n+1})$ is uniform over $1, \ldots, n+1$ (if no ties)
  - So $x \mapsto C(x) = \{y : \mathrm{rank}\{s(x, y) : s_1, \ldots, s_n\} \leq \lceil (1-\alpha)(n+1) \rceil\}$ satisfies $\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$

# Conformal prediction ctd.

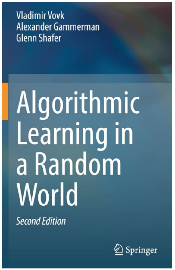
- Typical setting: *exchangeable datapoints.*
  - For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, e.g., $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \dots, s(X_{n+1}, Y_{n+1})$ are exchangeable (if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)
  - Hence, the rank of $s(X_{n+1}, Y_{n+1})$ is uniform over $1, \dots, n+1$ (if no ties)
  - So $x \mapsto C(x) = \{y : \mathrm{rank}\{s(x, y) : s_1, \dots, s_n\} \leq \lceil (1-\alpha)(n+1) \rceil\}$ satisfies $\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$

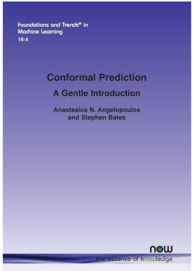


# Conformal prediction ctd.

- Typical setting: *exchangeable datapoints*.
  - For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, e.g., $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \ldots, s(X_{n+1}, Y_{n+1})$ are exchangeable (if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)
  - Hence, the rank of $s(X_{n+1}, Y_{n+1})$ is uniform over $1, \ldots, n+1$ (if no ties)
  - So $x \mapsto C(x) = \{y : \mathrm{rank}\{s(x, y) : s_1, \ldots, s_n\} \leq \lceil (1-\alpha)(n+1) \rceil\}$ satisfies $\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$

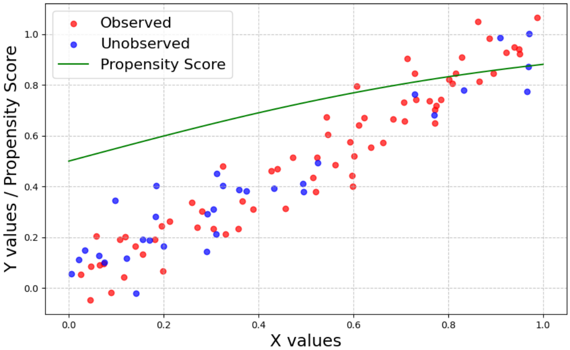



- However, there are scenarios that existing methods do not resolve, e.g., missing data

# Our problem setting

- Given data
$$(X_1, A_1, Y_1A_1), \dots, (X_n, A_n, Y_nA_n) \overset{\text{iid}}{\sim} P_X \times P_{A|X} \times P_{Y|X},$$
with outcomes *missing at random* (MAR). Thus,
$$A_i = 1 : Y_i \text{ is observed}, \qquad A_i = 0 : Y_i \text{ is unobserved}.$$

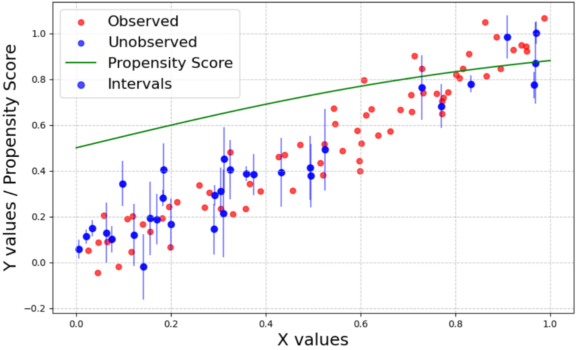


# Our problem setting: Missing At Random

- **Goal**: Simultaneous inference on the missing outcomes $\{Y_i : A_i = 0\}$.

# Our problem setting: Missing At Random

- **Goal**: Simultaneous inference on the missing outcomes $\{Y_i : A_i = 0\}$.
- Specifically, construct prediction sets $\{\widehat{C}(X_i) : A_i = 0\}$ for $\{Y_i : A_i = 0\}$ with coverage guarantees



# Inferential target

- With i.i.d./exchangeable data $(X_1, Y_1), \cdots, (X_n, Y_n)$ and test input $X_{n+1}$, standard conformal prediction gives a prediction set $\widehat{C}_n(X_{n+1})$ with *marginal coverage*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha.$$

# Inferential target

- With i.i.d./exchangeable data $(X_1, Y_1), \cdots, (X_n, Y_n)$ and test input $X_{n+1}$, standard conformal prediction gives a prediction set $\widehat{C}_n(X_{n+1})$ with *marginal coverage*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha.$$

- **Question:** Under MAR:
  - In what sense can we do useful distribution-free inference for multiple unobserved outcomes?

# Inferential target

- With i.i.d./exchangeable data $(X_1, Y_1), \cdots, (X_n, Y_n)$ and test input $X_{n+1}$, standard conformal prediction gives a prediction set $\widehat{C}_n(X_{n+1})$ with *marginal coverage*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha.$$

- **Question:** Under MAR:
    - In what sense can we do useful distribution-free inference for multiple unobserved outcomes?
    - Is it possible to go beyond marginal coverage? E.g., have coverage conditional on the test inputs/feature observations with missing outcomes?

# Overview of results

- We consider coverage guarantees of the form

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}(X_i)\right\}\,\middle|\, X_{1:n}, A_{1:n}\right] \geq 1-\alpha, \tag{1}$$

  where $N^{(0)}$ is the number of unobserved labels, and $0/0 := 1$.
- The proportion of covered missing outcomes is on average at least $1-\alpha$, conditional on $X_{1:n}$ and the missingness pattern $A_{1:n}$.

# Overview of results

- We consider coverage guarantees of the form

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i\in\widehat{C}(X_i)\right\}\,\middle|\,X_{1:n},A_{1:n}\right]\geq 1-\alpha, \tag{1}$$

  where $N^{(0)}$ is the number of unobserved labels, and $0/0 := 1$.
- The proportion of covered missing outcomes is on average at least $1-\alpha$, conditional on $X_{1:n}$ and the missingness pattern $A_{1:n}$.
  - For discrete features $X$, we construct a procedure that achieves (1).

# Overview of results

- We consider coverage guarantees of the form

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i\in\widehat{C}(X_i)\right\}\,\middle|\,X_{1:n},A_{1:n}\right]\geq 1-\alpha, \tag{1}$$

  where $N^{(0)}$ is the number of unobserved labels, and $0/0 := 1$.
- The proportion of covered missing outcomes is on average at least $1-\alpha$, conditional on $X_{1:n}$ and the missingness pattern $A_{1:n}$.
  - For discrete features $X$, we construct a procedure that achieves (1).
  - For general features $X$, we prove an impossibility result for (1); and then relax it.

## Overview of results - continued

- As a relaxation, we consider

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}(X_i)\right\}\,\middle|\, B_{1:n}, A_{1:n}\right] \geq 1-\alpha, \tag{2}$$

  where $B_i = B_i(X_i)$ is a discretization of $X_i$ (defined soon).

# Overview of results - continued

- As a relaxation, we consider

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i\in\widehat{C}(X_i)\right\}\,\Big|\,B_{1:n},A_{1:n}\right]\geq 1-\alpha, \tag{2}$$

  where $B_i = B_i(X_i)$ is a discretization of $X_i$ (defined soon).
- **Challenge:** Even though we have MAR ($Y \perp\!\!\!\perp A \mid X$), this does not need to be preserved after discretization (may have $Y \not\perp\!\!\!\perp A \mid B$ for $B = B(X)$).

# Overview of results - continued

- As a relaxation, we consider

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i\in\widehat{C}(X_i)\right\}\,\middle|\,B_{1:n},A_{1:n}\right]\geq 1-\alpha, \tag{2}$$

  where $B_i = B_i(X_i)$ is a discretization of $X_i$ (defined soon).

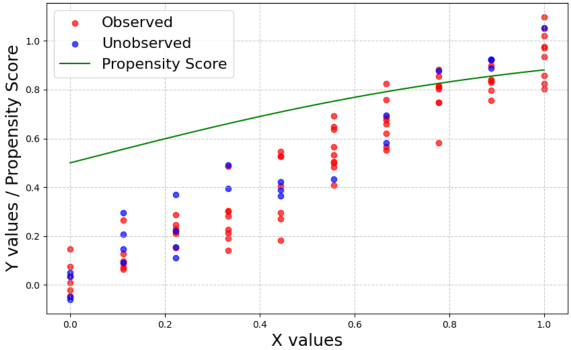
- **Challenge:** Even though we have MAR ($Y \perp\!\!\!\perp A \mid X$), this does not need to be preserved after discretization (may have $Y \not\!\perp\!\!\!\perp A \mid B$ for $B = B(X)$).

- We introduce a carefully designed propensity score partitioning scheme, and show how it can be used to obtain (2) in a distribution-free sense (for any dist. of $(X, Y)$).

# Table of Contents

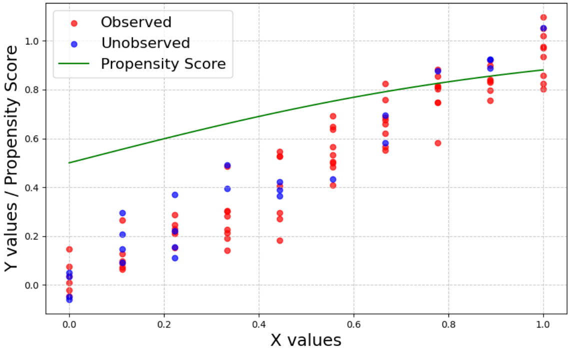
# First case: Discrete features

- Discrete features naturally form groups of outcomes $\{Y_i : X_i = x\}$, $x \in \mathcal{X}$.



# First case: Discrete features

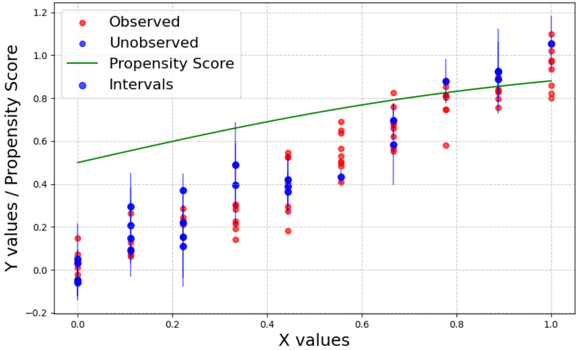
- Discrete features naturally form groups of outcomes $\{Y_i : X_i = x\}$, $x \in \mathcal{X}$.




- Within each group, the outcomes are *exchangeable* conditional on $X_i = x$.

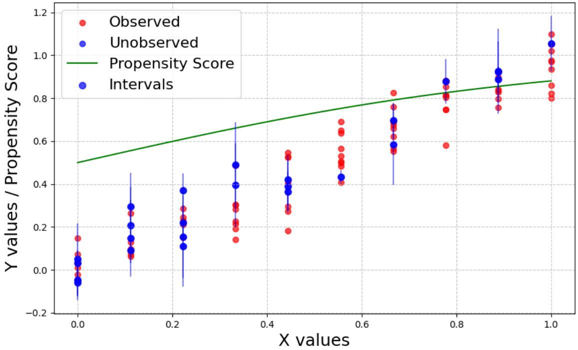
# Procedure for discrete features: Naive approach

- Direct method: run split conformal prediction separately for each $x$.



# Procedure for discrete features: Naive approach

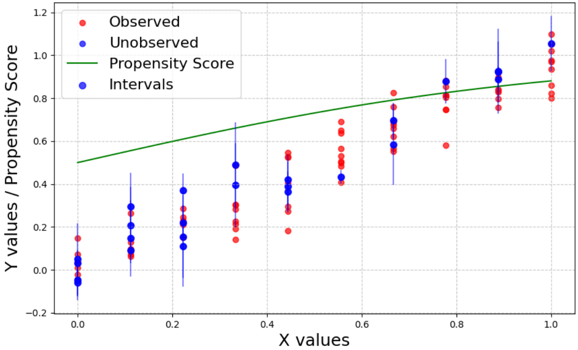
- Direct method: run split conformal prediction separately for each $x$.




- This method attains $\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}(X_i)\right\}\,\middle|\, X_{1:n}, A_{1:n}\right] \geq 1-\alpha$.

# Procedure for discrete features: Naive approach

- Direct method: run split conformal prediction separately for each $x$.




- This method attains $\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i\in\widehat{C}(X_i)\right\}\,\middle|\,X_{1:n},A_{1:n}\right]\geq 1-\alpha$.
- However, it can produce infinite-width prediction sets in small groups with $\geq \alpha$ missingness.

# Procedure for discrete features: our method

- Alternative method: simultaneous inference across multiple feature values.

# Procedure for discrete features: our method

- Alternative method: simultaneous inference across multiple feature values.
- Let
  1. Nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and $S_i = s(X_i, Y_i)$ if $A_i = 1$
  2. Distinct $X$ values observed: $X'_1, \cdots, X'_M$
  3. Indices of datapoints with features equal to $X'_k$: $I_k = \{i \in [n] : X_i = X'_k\}$,
  4. Indices partitioned according to unobserved and observed outcomes, resp.: $I_k^0 = \{i \in [n] : X_i = X'_k, A_i = 0\}$, $I_k^1 = \{i \in [n] : X_i = X'_k, A_i = 1\}$.
  5. Sample sizes $N_k = |I_k|$, $N_k^0 = |I_k^0|$

# Procedure for discrete features: our method

- Alternative method: simultaneous inference across multiple feature values.
- Let
  1. Nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and $S_i = s(X_i, Y_i)$ if $A_i = 1$
  2. Distinct $X$ values observed: $X'_1, \cdots, X'_M$
  3. Indices of datapoints with features equal to $X'_k$: $I_k = \{i \in [n] : X_i = X'_k\}$,
  4. Indices partitioned according to unobserved and observed outcomes, resp.: $I_k^0 = \{i \in [n] : X_i = X'_k, A_i = 0\}$, $I_k^1 = \{i \in [n] : X_i = X'_k, A_i = 1\}$.
  5. Sample sizes $N_k = |I_k|$, $N_k^0 = |I_k^0|$
- Our prediction set:

$$\widehat{C}(x) = \left\{ y \in \mathcal{Y} : s(x, y) \leq Q_{1-\alpha}\left( \sum_{k=1}^{M} \sum_{i \in I_k^1} \frac{N_k^0}{N_k N^{(0)}} \delta_{S_i} + \sum_{k=1}^{M} \frac{(N_k^0)^2}{N_k N^{(0)}} \delta_{+\infty} \right) \right\}. \quad (3)$$

# Procedure for discrete features: our method

- Alternative method: simultaneous inference across multiple feature values.
- Let
  1. Nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and $S_i = s(X_i, Y_i)$ if $A_i = 1$
  2. Distinct $X$ values observed: $X'_1, \cdots, X'_M$
  3. Indices of datapoints with features equal to $X'_k$: $I_k = \{i \in [n] : X_i = X'_k\}$,
  4. Indices partitioned according to unobserved and observed outcomes, resp.: $I_k^0 = \{i \in [n] : X_i = X'_k, A_i = 0\}$, $I_k^1 = \{i \in [n] : X_i = X'_k, A_i = 1\}$.
  5. Sample sizes $N_k = |I_k|$, $N_k^0 = |I_k^0|$
- Our prediction set:

$$\widehat{C}(x) = \left\{ y \in \mathcal{Y} : s(x, y) \le Q_{1-\alpha} \left( \sum_{k=1}^{M} \sum_{i \in I_k^1} \frac{N_k^0}{N_k N^{(0)}} \delta_{S_i} + \sum_{k=1}^{M} \frac{(N_k^0)^2}{N_k N^{(0)}} \delta_{+\infty} \right) \right\}. \tag{3}$$

- Idea: symmetry of data distribution; see also *SymmPI* (D. & Yu, 2023)
- Provides uniform-width prediction sets for all $x$ values.

# Procedure for discrete features: guarantee

**Theorem 1**

The prediction set (3) satisfies *feature- and missingness-conditional coverage*

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i\in\widehat{C}(X_i)\right\}\,\middle|\,X_{1:n},A_{1:n}\right]\geq 1-\alpha.$$

# Discrete features: improvement via partitioning

- If missingness proportion is high, this can still be conservative.

# Discrete features: improvement via partitioning

- If missingness proportion is high, this can still be conservative.
- Idea: *Partition* datapoints. For each partition, use $\widehat{C}$ on *all datapoints with observed labels* to predict outcomes missing in that partition.

# Discrete features: improvement via partitioning

- If missingness proportion is high, this can still be conservative.
- Idea: *Partition* datapoints. For each partition, use $\widehat{C}$ on *all datapoints with observed labels* to predict outcomes missing in that partition.
- Since guarantee is feature- and missingness-conditional, this is still valid!

# Discrete features: improvement via partitioning

- If missingness proportion is high, this can still be conservative.
- Idea: *Partition* datapoints. For each partition, use $\widehat{C}$ on *all datapoints with observed labels* to predict outcomes missing in that partition.
- Since guarantee is feature- and missingness-conditional, this is still valid!
- Previous methods are at two endpoints: partition is all singletons ("naive method") vs whole set ("our method").

# Discrete features: improvement via partitioning

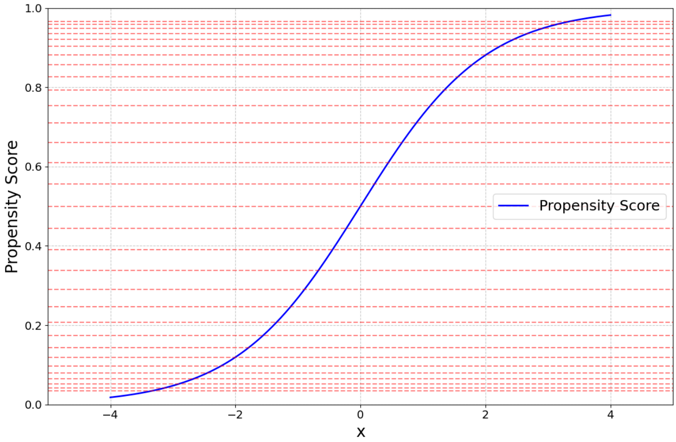
- If missingness proportion is high, this can still be conservative.
- Idea: *Partition* datapoints. For each partition, use $\widehat{C}$ on *all datapoints with observed labels* to predict outcomes missing in that partition.
- Since guarantee is feature- and missingness-conditional, this is still valid!
- Previous methods are at two endpoints: partition is all singletons ("naive method") vs whole set ("our method").
- Why practically useful? Partition can depend on $X_{1:n}, A_{1:n}$; can aim to ensure small missingness per group.

# Procedure for general feature distributions

- If the propensity score $x \mapsto p_{A|X}(x) = \mathbb{P}\{A = 1 \mid X = x\}$ is known, $\varepsilon$-**discretize** it
- Let $\varepsilon$ be a predefined discretization level, and $z_k = (1+\varepsilon)^k/[1+(1+\varepsilon)^k]$ for all integers $k$

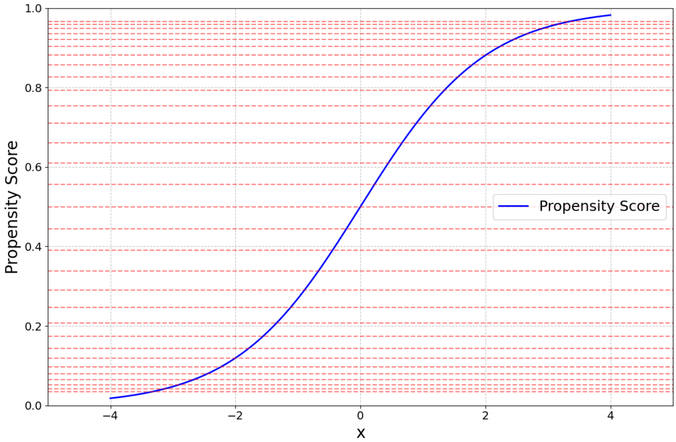
# Procedure for general feature distributions

- If the propensity score $x \mapsto p_{A|X}(x) = \mathbb{P}\{A = 1 \mid X = x\}$ is known, $\varepsilon$-**discretize** it
- Let $\varepsilon$ be a predefined discretization level, and $z_k = (1+\varepsilon)^k / [1 + (1+\varepsilon)^k]$ for all integers $k$
- Partition the feature space into $D_k = \left\{x : p_{A|X}(x) \in [z_k, z_{k+1})\right\}$, $\mathcal{B} = \{D_k : k \in \mathbb{Z}\}$.



# Procedure for general feature distributions

- If the propensity score $x \mapsto p_{A|X}(x) = \mathbb{P}\{A = 1 \mid X = x\}$ is known, $\varepsilon$-**discretize** it
- Let $\varepsilon$ be a predefined discretization level, and $z_k = (1+\varepsilon)^k/[1+(1+\varepsilon)^k]$ for all integers $k$
- Partition the feature space into $D_k = \left\{x : p_{A|X}(x) \in [z_k, z_{k+1})\right\}$, $\mathcal{B} = \{D_k : k \in \mathbb{Z}\}$.



# Pro-CP

- Show *approximate within-partition exchangeability* of the scores, enabling inference.

# Pro-CP

- Show *approximate within-partition exchangeability* of the scores, enabling inference.
- **Propensity score discretization-based conformal prediction (pro-CP)**: Procedure (4) applied to the discretized data $(B_i, A_i, A_i Y_i)_{i \in [n]}$, i.e.,

$$\widehat{C}^{\text{pro-CP}}(x) = \left\{ y \in \mathcal{Y}, :, s(x,y) \leq Q_{1-\alpha} \left( \sum_{k=1}^{M} \sum_{i \in I_k^{\mathcal{B},1}} \frac{N_k^{\mathcal{B},0}}{N^{(0)} N_k^{\mathcal{B}}} \cdot \delta_{S_i} + \frac{1}{N^{(0)}} \sum_{k=1}^{M} \frac{(N_k^{\mathcal{B},0})^2}{N_k^{\mathcal{B}}} \cdot \delta_{+\infty} \right) \right\}. \qquad (4)$$

# Pro-CP

- Show *approximate within-partition exchangeability* of the scores, enabling inference.
- **Propensity score discretization-based conformal prediction (pro-CP)**: Procedure (4) applied to the discretized data $(B_i, A_i, A_i Y_i)_{i \in [n]}$, i.e.,

$$\widehat{C}^{\text{pro-CP}}(x) = \left\{ y \in \mathcal{Y}, :, s(x,y) \leq Q_{1-\alpha} \left( \sum_{k=1}^{M} \sum_{i \in I_k^{\mathcal{B},1}} \frac{N_k^{\mathcal{B},0}}{N^{(0)} N_k^{\mathcal{B}}} \cdot \delta_{S_i} + \frac{1}{N^{(0)}} \sum_{k=1}^{M} \frac{(N_k^{\mathcal{B},0})^2}{N_k^{\mathcal{B}}} \cdot \delta_{+\infty} \right) \right\}. \tag{4}$$

## Theorem 2

Suppose $0 < p_{A|X}(X) < 1$ almost surely. Then $\widehat{C}^{\text{pro-CP}}$ from (4) satisfies *propensity score discretized feature- and missingness-conditional coverage*:

$$\mathbb{E}\left[ \frac{1}{N^{(0)}} \sum_{i:A_i=0} \mathbb{1}\left\{ Y_i \in \widehat{C}^{\text{pro-CP}}(X_i) \right\} \,\middle|\, B_{1:n}, A_{1:n} \right] \geq 1 - \alpha - \varepsilon.$$

# Pro-CP

- Show *approximate within-partition exchangeability* of the scores, enabling inference.
- **Propensity score discretization-based conformal prediction (pro-CP)**: Procedure (4) applied to the discretized data $(B_i, A_i, A_i Y_i)_{i \in [n]}$, i.e.,

$$\widehat{C}^{\text{pro-CP}}(x) = \left\{ y \in \mathcal{Y}, :, s(x,y) \le Q_{1-\alpha}\left( \sum_{k=1}^{M} \sum_{i \in I_k^{\mathcal{B},1}} \frac{N_k^{\mathcal{B},0}}{N^{(0)} N_k^{\mathcal{B}}} \cdot \delta_{S_i} + \frac{1}{N^{(0)}} \sum_{k=1}^{M} \frac{(N_k^{\mathcal{B},0})^2}{N_k^{\mathcal{B}}} \cdot \delta_{+\infty} \right) \right\}. \qquad (4)$$

### Theorem 2

Suppose $0 < p_{A|X}(X) < 1$ almost surely. Then $\widehat{C}^{\text{pro-CP}}$ from (4) satisfies *propensity score discretized feature- and missingness-conditional coverage*:

$$\mathbb{E}\left[ \frac{1}{N^{(0)}} \sum_{i: A_i = 0} \mathbb{1}\left\{ Y_i \in \widehat{C}^{\text{pro-CP}}(X_i) \right\} \,\middle|\, B_{1:n}, A_{1:n} \right] \ge 1 - \alpha - \varepsilon.$$

- The error from discretization is bounded by $\varepsilon$, for *any* $n$ and $\#$ of missing outcomes.

## Pro-CP with estimated propensity score

- If the propensity score is unknown, we may run pro-CP with an estimator $\hat{p}_{A|X}$ of $p_{A|X}$.

# Pro-CP with estimated propensity score

- If the propensity score is unknown, we may run pro-CP with an estimator $\hat{p}_{A|X}$ of $p_{A|X}$.

**Theorem 3**

Suppose $0 < p_{A|X}(X) < 1$ and $0 < \hat{p}_{A|X}(X) < 1$ almost surely. Then pro-CP run with $\hat{p}_{A|X}$ satisfies

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\right\}\,\middle|\, B_{1:n}, A_{1:n}\right] \geq 1-\alpha-(\varepsilon+\delta_{\hat{p}_{A|X}}+\varepsilon\delta_{\hat{p}_{A|X}}),$$

where

$$\delta_{\hat{p}_{A|X}} = e^{2\|\log f_{p,\hat{p}}\|_\infty} - 1, \qquad f_{p,\hat{p}}(x) = \frac{p_{A|X}(x)/(1-p_{A|X}(x))}{\hat{p}_{A|X}(x)/(1-\hat{p}_{A|X}(x))}.$$

# Pro-CP with estimated propensity score

- If the propensity score is unknown, we may run pro-CP with an estimator $\hat{p}_{A|X}$ of $p_{A|X}$.

**Theorem 3**

Suppose $0 < p_{A|X}(X) < 1$ and $0 < \hat{p}_{A|X}(X) < 1$ almost surely. Then pro-CP run with $\hat{p}_{A|X}$ satisfies

$$\mathbb{E}\left[\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\right\}\,\middle|\, B_{1:n}, A_{1:n}\right] \geq 1-\alpha-(\varepsilon+\delta_{\hat{p}_{A|X}}+\varepsilon\delta_{\hat{p}_{A|X}}),$$

where

$$\delta_{\hat{p}_{A|X}} = e^{2\|\log f_{p,\hat{p}}\|_\infty}-1, \qquad f_{p,\hat{p}}(x) = \frac{p_{A|X}(x)/(1-p_{A|X}(x))}{\hat{p}_{A|X}(x)/(1-\hat{p}_{A|X}(x))}.$$

- The error from estimation does not grow with the number of missing outcomes.

## New result underlying pro-CP guarantee

- Balancing property of the propensity score [Rosenbaum and Rubin (1983)]: the missingness is independent of the outcome conditional on the propensity: $A \perp\!\!\!\perp Y \mid p_{A|X}$.

# New result underlying pro-CP guarantee

- Balancing property of the propensity score [Rosenbaum and Rubin (1983)]: the missingness is independent of the outcome conditional on the propensity: $A \perp\!\!\!\perp Y \mid p_{A|X}$.
- We show *approximate version*: dist. of $s(X,Y)$ close for $A=0,1$ given small range of $p_{A|X}$

**Lemma (Bounded prop. score implies closeness of cond. distrib. for obs. and missing)**

*Suppose that* $(X,Y,A) \sim P_X \times P_{Y|X} \times \text{Bernoulli}(p_{A|X})$ *on* $\mathcal{X} \times \mathcal{Y} \times \{0,1\}$*, and that for a set* $B \subset \mathcal{X}$ *and* $t \in (0,1)$*,* $\varepsilon \geq 0$*,*

$$t \leq \frac{p_{A|X}(x)}{1-p_{A|X}(x)} \leq t(1+\varepsilon), \text{ for all } x \in B.$$

*Let* $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *be any measurable function and let* $S = s(X,Y)$*. Then*

$$d_{TV}(P_{S|A=1,X\in B}, P_{S|A=0,X\in B}) \leq \varepsilon.$$

# Related ideas in the literature

- Sub-classification based on propensity score [Rosenbaum and Rubin (1984)]: can reduce bias in causal effect estimation by partitioning based on estimated propensity
- Similar principle, but does not specify partitioning scheme, and for a different goal (bias reduction); w/o any technical overlap

# Table of Contents

# Application to simultaneous inference on ITEs

- Consider a potential outcomes model

$$(X_i, T_i, Y_i(0), Y_i(1))_{1\le i\le n} \overset{\text{iid}}{\sim} P_X \times P_{T|X} \times P_{Y(1)|X} \times P_{Y(0)|X},$$

  where we observe $(X_i, T_i, T_i Y_i(1) + (1 - T_i) Y_i(0))_{1\le i\le n}$.

# Application to simultaneous inference on ITEs

- Consider a potential outcomes model

$$(X_i, T_i, Y_i(0), Y_i(1))_{1\le i\le n} \overset{\text{iid}}{\sim} P_X \times P_{T|X} \times P_{Y(1)|X} \times P_{Y(0)|X},$$

  where we observe $(X_i, T_i, T_i Y_i(1) + (1 - T_i) Y_i(0))_{1\le i\le n}$.

- Applying pro-CP, we can construct $\widehat{C}^{\text{counterfactual}}$ such that

$$\mathbb{E}\left[\tfrac{1}{N^{(0)}}\textstyle\sum_{i:T_i=0} \mathbb{1}\left\{Y_i(1) \in \widehat{C}^{\text{counterfactual}}(X_i)\right\} \,\middle|\, B_{1:n}, T_{1:n}\right] \ge 1 - \alpha.$$

# Application to simultaneous inference on ITEs

- Consider a potential outcomes model

$$(X_i, T_i, Y_i(0), Y_i(1))_{1\le i\le n} \overset{\text{iid}}{\sim} P_X \times P_{T|X} \times P_{Y(1)|X} \times P_{Y(0)|X},$$

  where we observe $(X_i, T_i, T_i Y_i(1) + (1 - T_i) Y_i(0))_{1\le i\le n}$.

- Applying pro-CP, we can construct $\widehat{C}^{\text{counterfactual}}$ such that

$$\mathbb{E}\left[\tfrac{1}{N^{(0)}} \textstyle\sum_{i:T_i=0} \mathbb{1}\left\{Y_i(1) \in \widehat{C}^{\text{counterfactual}}(X_i)\right\} \,\middle|\, B_{1:n}, T_{1:n}\right] \ge 1 - \alpha.$$

- By letting $\widehat{C}_i^{\text{ITE}} = \{y - Y_i(0) : y \in \widehat{C}^{\text{counterfactual}}(X_i)\}$, we obtain prediction sets for individual treatment effects

$$\mathbb{E}\left[\tfrac{1}{N^{(0)}} \textstyle\sum_{i\in I_{T=0}} \mathbb{1}\left\{(Y_i(1) - Y_i(0)) \in \widehat{C}_i^{\text{ITE}}\right\} \,\middle|\, B_{1:n}, T_{1:n}\right] \ge 1 - \alpha.$$

# Table of Contents

# Achieving a stronger guarantee on the coverage proportion

- Can we achieve a stronger guarantee beyond bounding the *mean coverage*?

# Achieving a stronger guarantee on the coverage proportion

- Can we achieve a stronger guarantee beyond bounding the *mean coverage*?
- Possible goal: PAC-type guarantee of the form

$$\mathbb{P}\left\{\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\right\} \geq 1-\alpha\right\} \geq 1-\delta.$$

# Achieving a stronger guarantee on the coverage proportion

- Can we achieve a stronger guarantee beyond bounding the *mean coverage*?
- Possible goal: PAC-type guarantee of the form

$$\mathbb{P}\left\{\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\right\} \geq 1-\alpha\right\} \geq 1-\delta.$$

- Turns out to be hard to achieve in the distribution-free sense

# Achieving a stronger guarantee on the coverage proportion

- Can we achieve a stronger guarantee beyond bounding the *mean coverage*?
- Possible goal: PAC-type guarantee of the form

$$\mathbb{P}\left\{\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\right\} \geq 1-\alpha\right\} \geq 1-\delta.$$

- Turns out to be hard to achieve in the distribution-free sense
- As a surrogate target, we consider bounding the *squared coverage*

$$\mathbb{E}\left[\left(\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \notin \widehat{C}^{\text{pro-CP}}(X_i)\right\}\right)^2\right] \leq \alpha^2.$$

# Achieving a stronger guarantee on the coverage proportion

- Can we achieve a stronger guarantee beyond bounding the *mean coverage*?
- Possible goal: PAC-type guarantee of the form

$$\mathbb{P}\left\{\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\right\} \geq 1-\alpha\right\} \geq 1-\delta.$$

- Turns out to be hard to achieve in the distribution-free sense
- As a surrogate target, we consider bounding the *squared coverage*

$$\mathbb{E}\left[\left(\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \notin \widehat{C}^{\text{pro-CP}}(X_i)\right\}\right)^2\right] \leq \alpha^2.$$

(motivated by Lee et. al. (2023): Hierarchical CP)

# Interpretation of the squared-coverage guarantee

- Let $\hat{m} = \frac{1}{N^{(0)}} \sum_{i:A_i=0} \mathbb{1}\left\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\right\}$ denote the *miscoverage proportion*.

# Interpretation of the squared-coverage guarantee

- Let $\hat{m} = \frac{1}{N^{(0)}} \sum_{i:A_i=0} \mathbb{1}\left\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\right\}$ denote the *miscoverage proportion*.
- Conditional on (discretized) features, pro-CP attains $\mathbb{E}\left[\hat{m}\right] \leq \alpha$.

# Interpretation of the squared-coverage guarantee

- Let $\hat{m} = \frac{1}{N^{(0)}} \sum_{i:A_i=0} \mathbb{1}\left\{ Y_i \in \widehat{C}^{\text{pro-CP}}(X_i) \right\}$ denote the *miscoverage proportion*.
- Conditional on (discretized) features, pro-CP attains $\mathbb{E}[\hat{m}] \leq \alpha$.
- The squared-coverage guarantee is $\mathbb{E}\left[\hat{m}^2\right] \leq \alpha^2$, and provides a stronger control over $\hat{m}$ being close to unity, preventing e.g., $\hat{m} = 0$ w.p. $1 - \alpha$ and 1 w.p. $\alpha$.

# Pro-CP2 procedure

- Define
  1. For all $i \in [n]$, $\bar{S}_i = S_i$ if $A_i = 1$ and $\bar{S}_i = +\infty$ if $A_i = 0$.
  2. Pairwise minima: $\bar{S}_{ij} := \min\{\bar{S}_i, \bar{S}_j\}$ for all $i, j$.

# Pro-CP2 procedure

- Define
  1. For all $i \in [n]$, $\bar{S}_i = S_i$ if $A_i = 1$ and $\bar{S}_i = +\infty$ if $A_i = 0$.
  2. Pairwise minima: $\bar{S}_{ij} := \min\{\bar{S}_i, \bar{S}_j\}$ for all $i, j$.
- Pro-CP2 prediction set:

$$\widehat{C}^{\text{pro-CP2}}(x) = \Bigg\{ y \in \mathcal{Y} : s(x,y) \leq Q_{1-\alpha^2}\Bigg( \sum_{k=1}^{M} \sum_{i \in I_k^{\mathcal{B}}} \frac{1}{(N^{(0)})^2} \cdot \frac{N_k^{\mathcal{B},0}}{N_k^{\mathcal{B}}} \cdot \delta_{\bar{S}_i}$$

$$+ \sum_{k=1}^{M} \sum_{\substack{i,j \in I_k^{\mathcal{B}} \\ i \neq j}} \frac{N_k^{\mathcal{B},0}(N_k^{\mathcal{B},0} - 1)}{(N^{(0)})^2 N_k^{\mathcal{B}}(N_k^{\mathcal{B}} - 1)} \delta_{\bar{S}_{ij}} + \sum_{1 \leq k \neq k' \leq M} \sum_{i \in I_k^{\mathcal{B}}} \sum_{j \in I_{k'}^{\mathcal{B}}} \frac{N_k^{\mathcal{B},0} N_{k'}^{\mathcal{B},0}}{(N^{(0)})^2 N_k^{\mathcal{B}} N_{k'}^{\mathcal{B}}} \delta_{\bar{S}_{ij}} \Bigg) \Bigg\}.$$

# Pro-CP2 procedure

- Define
  1. For all $i \in [n]$, $\bar{S}_i = S_i$ if $A_i = 1$ and $\bar{S}_i = +\infty$ if $A_i = 0$.
  2. Pairwise minima: $\bar{S}_{ij} := \min\{\bar{S}_i, \bar{S}_j\}$ for all $i, j$.
- Pro-CP2 prediction set:

$$\widehat{C}^{\text{pro-CP2}}(x) = \Bigg\{ y \in \mathcal{Y} : s(x,y) \le Q_{1-\alpha^2}\Bigg( \sum_{k=1}^{M} \sum_{i \in I_k^{\mathcal{B}}} \frac{1}{(N^{(0)})^2} \cdot \frac{N_k^{\mathcal{B},0}}{N_k^{\mathcal{B}}} \cdot \delta_{\bar{S}_i} + \sum_{k=1}^{M} \sum_{\substack{i,j \in I_k^{\mathcal{B}} \\ i \neq j}} \frac{N_k^{\mathcal{B},0}(N_k^{\mathcal{B},0}-1)}{(N^{(0)})^2 N_k^{\mathcal{B}}(N_k^{\mathcal{B}}-1)} \delta_{\bar{S}_{ij}} + \sum_{1 \le k \neq k' \le M} \sum_{i \in I_k^{\mathcal{B}}} \sum_{j \in I_{k'}^{\mathcal{B}}} \frac{N_k^{\mathcal{B},0} N_{k'}^{\mathcal{B},0}}{(N^{(0)})^2 N_k^{\mathcal{B}} N_{k'}^{\mathcal{B}}} \delta_{\bar{S}_{ij}} \Bigg) \Bigg\}.$$

- Similar intuition as before; but use invariance to find probability of $\mathbb{1}\left\{\min\{S_{i^*}, S_{j^*}\} \le q_{1-\alpha^2}(\tilde{S}_1, \ldots, \tilde{S}_M)\right\}$, where $i^*, j^*$ are random data indices with $A = 0$.

# Squared coverage error control of Pro-CP2

### Theorem 4

If $0 < p_{A|X}(X) < 1$ almost surely, then $\widehat{C}^{\text{pro-CP2}}$ satisfies

$$\mathbb{E}\left[\left(\frac{1}{N^{(0)}}\sum_{i:A_i=0}\mathbb{1}\left\{Y_i \notin \widehat{C}^{\text{pro-CP2}}(X_i)\right\}\right)^2 \,\middle|\, B_{1:n}, A_{1:n}\right] \leq \alpha^2 + 2\varepsilon.$$

# Table of Contents

# Simulation 1

**Weighted conformal (Tibshirani et al., 2019) vs pro-CP**: marginal vs conditional coverage

# Simulation 1

**Weighted conformal (Tibshirani et al., 2019) vs pro-CP**: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10],\ \ Y \mid X \sim N(X, (3+X)^2),\ \ A \mid X \sim \text{Bernoulli}(p_{A|X}(X))$

# Simulation 1

**Weighted conformal (Tibshirani et al., 2019) vs pro-CP**: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10],\ Y \mid X \sim N(X, (3 + X)^2),\ A \mid X \sim \text{Bernoulli}(p_{A|X}(X))$
2. $(1) : p_{A|X}(x) = 0.9 - 0.02x, (2) : p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$

# Simulation 1

**Weighted conformal (Tibshirani et al., 2019) vs pro-CP**: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10]$, $Y \mid X \sim N(X, (3 + X)^2)$, $A \mid X \sim \text{Bernoulli}(p_{A|X}(X))$
2. $(1) : p_{A|X}(x) = 0.9 - 0.02x, (2) : p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$
3. Fit OLS with $n_{\text{train}} = 500$, $s(x, y) = |y - \hat{\mu}(x)|$

# Simulation 1

**Weighted conformal (Tibshirani et al., 2019) vs pro-CP**: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10],\ \ Y \mid X \sim N(X, (3 + X)^2),\ \ A \mid X \sim \text{Bernoulli}(p_{A|X}(X))$
2. $(1) : p_{A|X}(x) = 0.9 - 0.02x, (2) : p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x)\sin 3x$
3. Fit OLS with $n_{\text{train}} = 500$, $s(x, y) = |y - \hat{\mu}(x)|$
4. 500 trials, $n = 500$, Pro-CP $\varepsilon = 0.1$, $\alpha = 0.2$, partition of size 10;

# Simulation 1

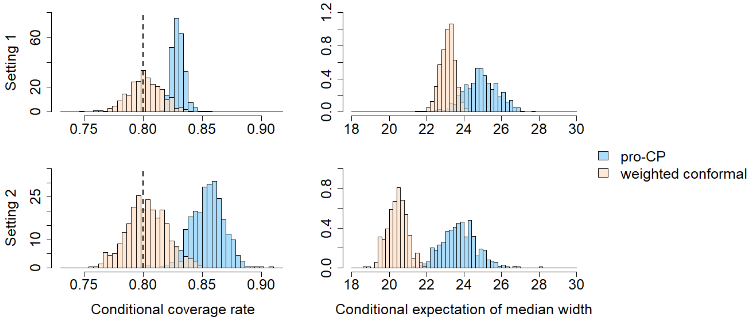
**Weighted conformal (Tibshirani et al., 2019) vs pro-CP**: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10]$, $Y \mid X \sim N(X, (3 + X)^2)$, $A \mid X \sim \text{Bernoulli}(p_{A|X}(X))$
2. $(1) : p_{A|X}(x) = 0.9 - 0.02x, (2) : p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$
3. Fit OLS with $n_{\text{train}} = 500$, $s(x, y) = |y - \hat{\mu}(x)|$
4. 500 trials, $n = 500$, Pro-CP $\varepsilon = 0.1$, $\alpha = 0.2$, partition of size 10;
5. Given $X_{1:n}, A_{1:n}$, 100x gen $(X_i', Y_i')_{1 \leq i \leq n} \mid B_i \sim P_{X|B} \times P_{Y|X}$, $n = 500$

# Simulation 1

**Weighted conformal (Tibshirani et al., 2019) vs pro-CP**: marginal vs conditional coverage

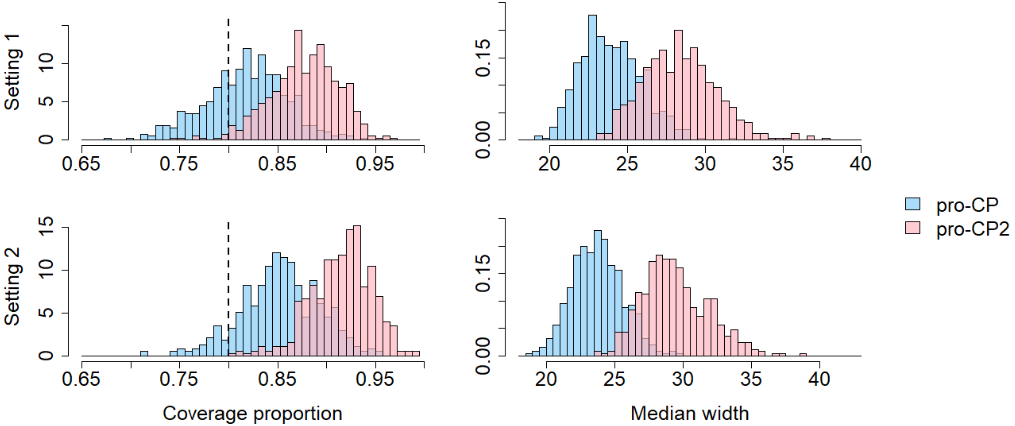
1. $X \sim \text{Unif}[0, 10]$, $Y \mid X \sim N(X, (3 + X)^2)$, $A \mid X \sim \text{Bernoulli}(p_{A|X}(X))$
2. $(1) : p_{A|X}(x) = 0.9 - 0.02x, (2) : p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$
3. Fit OLS with $n_{\text{train}} = 500$, $s(x, y) = |y - \hat{\mu}(x)|$
4. 500 trials, $n = 500$, Pro-CP $\varepsilon = 0.1$, $\alpha = 0.2$, partition of size 10;
5. Given $X_{1:n}, A_{1:n}$, 100x gen $(X_i', Y_i')_{1<i<n} \mid B_i \sim P_{X|B} \times P_{Y|X}$, $n = 500$



# Simulation 2

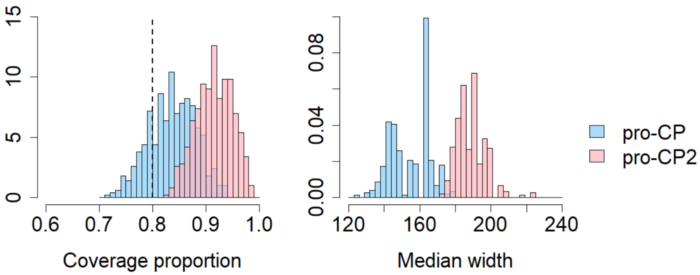
**pro-CP vs pro-CP2**: controlling mean vs squared miscoverage

- Same setting as Simulation 1, but evaluate marginal coverage & estimate propensity score with kernel regression on training data

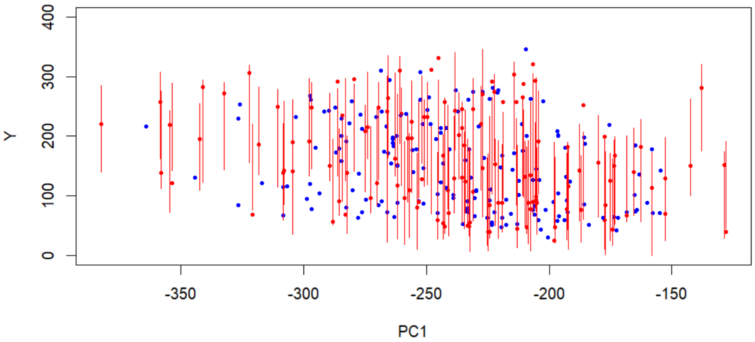


# Illustration on diabetes dataset (Efron et al., 2004)

- $X$: ten features (age, bmi, LDL/HDL, ...) of patients (sample sizes: train: 142; calibration+test: 300)
- $A$: missingness generated from a known logistic model
- $Y$: a measure of disease progression one year after baseline

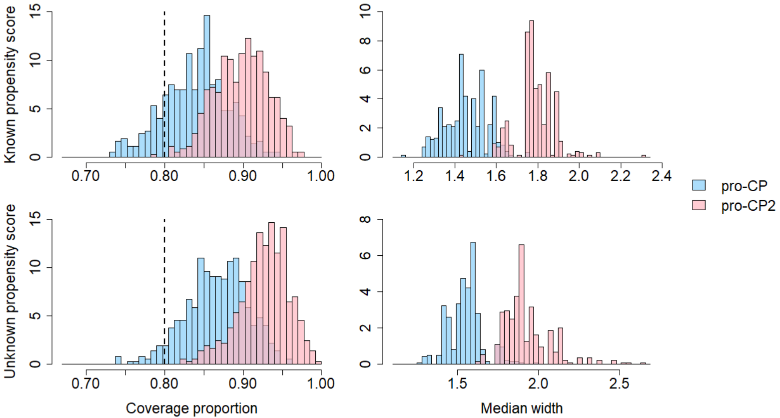


# Illustration on diabetes dataset (Efron et al., 2004): II



# Illustration on JOBS II dataset (Imai et al., 2010)

- $X$: job seekers: $n_{\text{train}} = 379$, $n = 500$; with 14 demographic features
- $A$: job skills workshop (to evaluate our methods, simulate via logistic model; estimate via RF)
- $Y(0), Y(1)$: pre- and post-treatment depression measure



# Discussion

- Introduced Pro-CP, a method for simultaneous prediction of multiple missing outcomes, and provided coverage guarantees
- Pro-CP2: stronger squared error miscoverage error control
- What applications might this have an impact on? Where could it be used?
- Preprint: `arxiv.org/abs/2403.04613`. Code: `github.com/yhoon31/pro-CP`
- Thanks!