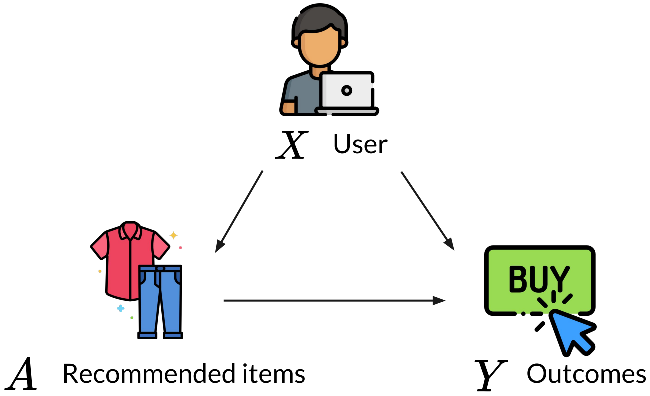
Conformal Off-Policy Prediction in Contextual Bandits

Muhammad Faaiz Taufiq*, Jean-Francois Ton*, Rob Cornish, Yee Whye Teh, Arnaud Doucet

Set up



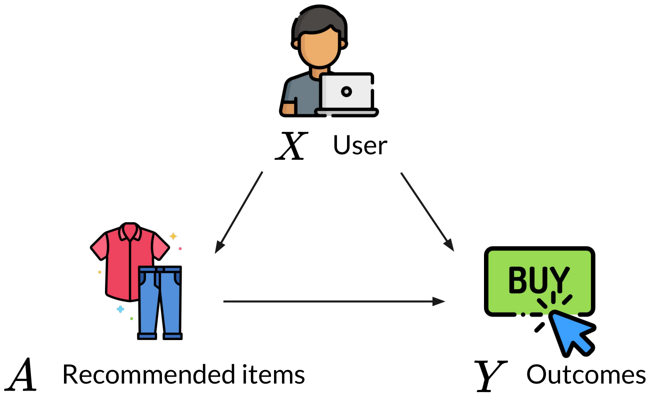
We are given logged data $\mathcal{D}_{obs} = \{x_i, a_i, y_i\}_{i=1}^{n_{obs}}$

Where, actions are sampled from behavioural policy π^b

$$A_i \mid X_i = x_i \sim \pi^b(\cdot \mid x_i)$$

Goal: Given a new target policy π^* and a user X, what are the probable outcomes for X if actions are chosen from π^*

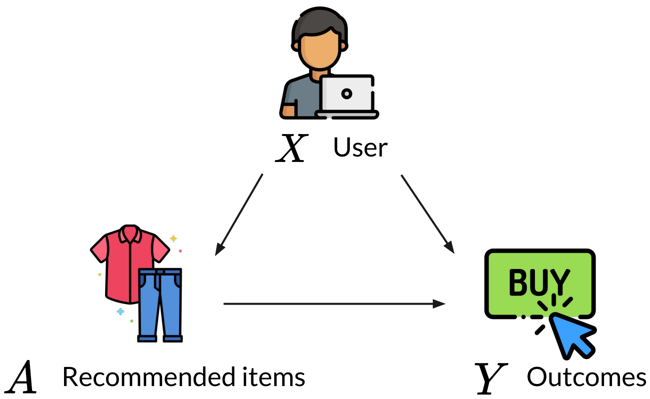
Set up



We achieve this by:
Constructing sets $\hat{C}(x)$ on the outcomes which are

1) Adaptive w.r.t. X
2) Capture variability in the outcome Y
3) Provide finite-sample guarantees.

Set up

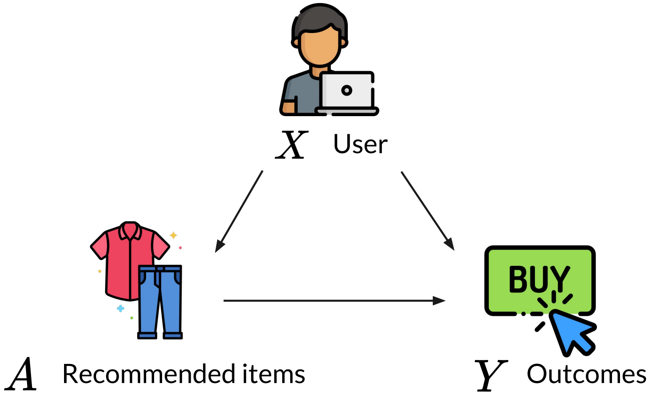


We achieve this by:
Constructing sets $\hat{C}(x)$ on the outcomes which are

1) Adaptive w.r.t. X
2) Capture variability in the outcome Y
3) Provide finite-sample guarantees.

$$1 - \alpha \leq \mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}(X)) \leq 1 - \alpha + o_{n_{obs}}(1)$$

Set up



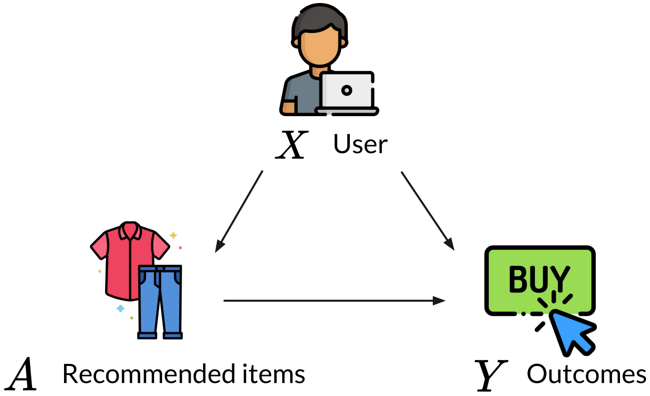
We achieve this by:
Constructing sets $\hat{C}(x)$ on the outcomes which are

1) Adaptive w.r.t. X
2) Capture variability in the outcome Y
3) Provide finite-sample guarantees.

$$1-\alpha \leq \mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}(X)) \leq 1-\alpha + o_{n_{obs}}(1)$$

Joint distribution of (X, Y) under target policy

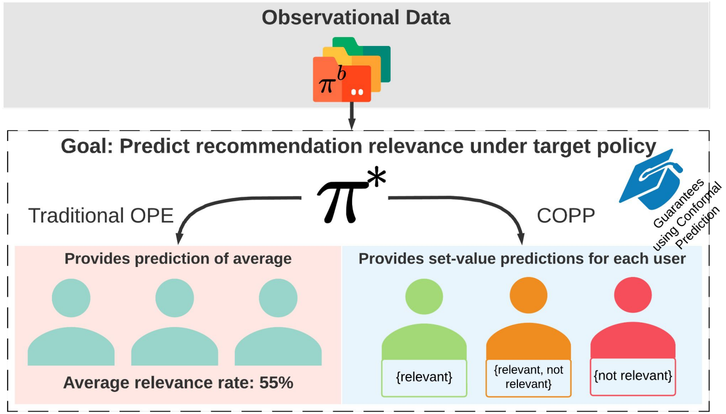
Comparison with Traditional Off-Policy Evaluation



Traditional OPE Methods focus on estimating average outcomes under a target policy.

1. This does not account for the variability in the outcomes
2. The resulting policy value is not adaptive w.r.t. X

In risk-sensitive settings, this measure may not be informative of the uncertainty.



Background

- In standard conformal prediction we require the calibration and test data to be **exchangeable**.
- If this assumption is fulfilled we are able to construct sets with the following guarantee:

$$1 - \alpha \leq \mathbb{P}_{(X,Y)\sim P_{X,Y}}(Y \in \hat{C}_n(X)) \leq 1 - \alpha + \frac{1}{n+1}.$$

Background

- In standard conformal prediction we require the calibration and test data to be **exchangeable**.
- If this assumption is fulfilled we are able to construct sets with the following guarantee:

$$1 - \alpha \leq \mathbb{P}_{(X,Y)\sim P_{X,Y}}(Y \in \hat{C}_n(X)) \leq 1 - \alpha + \frac{1}{n+1}.$$

- However this assumption can be easily violated in cases where **distribution shift is present**.
- For the case of covariate shift Tibshirani et al 2018 to use the idea weighted exchangeability:
  - As for most covariate shift problem, estimation of $w(x) := \mathrm{d}\tilde{P}_X / \mathrm{d}P_X(x)$ is crucial.
  - Tibshirani et al. show that if we are able to estimate the ratio well, CP is still applicable.

$$\mathrm{d}P_X \longrightarrow \mathrm{d}\tilde{P}_X$$

Proposed Method COPP

$$P^{\pi^b}(x, y) \longrightarrow P^{\pi^*}(x, y)$$

Proposed Method COPP

$$P^{\pi^b}(x,y) \longrightarrow P^{\pi^*}(x,y)$$

The key insight in COPP is to note is the following decomposition of the joint distribution of (X,Y)

$$P^{\pi^b}(x,y) = P(x)\int P(y|x,a)\pi^b(a|x)\mathrm{d}a = P(x)P^{\pi^b}(y|x)$$

$$P^{\pi^*}(x,y) = P(x)\int P(y|x,a)\pi^*(a|x)\mathrm{d}a = P(x)P^{\pi^*}(y|x)$$

Proposed Method COPP

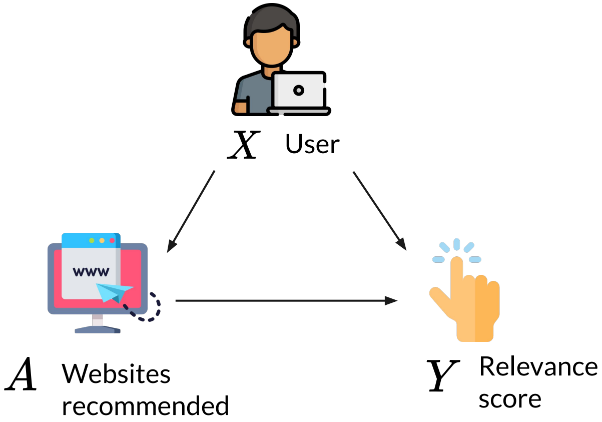
Adapting ideas from Tibshirani et al 2018, we show that for Off-Policy Prediction we only require estimation of the joint density ratio.

Following the previous decomposition we get the following weights.

$$w(x, y) = \mathrm{d}P^{\pi^*}_{X,Y} / \mathrm{d}P^{\pi^b}_{X,Y}(x, y) = \mathrm{d}P^{\pi^*}_{Y|X} / \mathrm{d}P^{\pi^b}_{Y|X}(x, y)$$

For exact details on how we construct the conformal intervals for Off-Policy Prediction see our paper

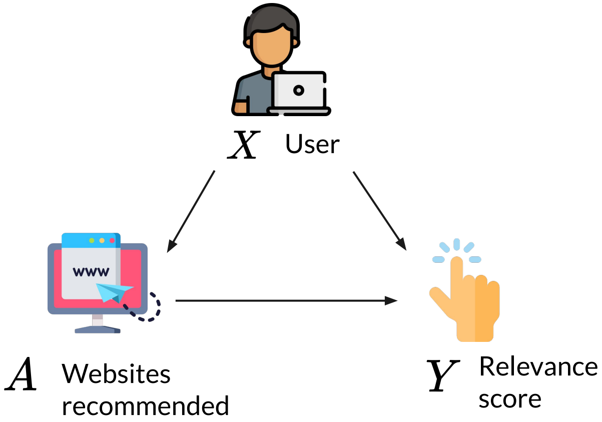
Application to Microsoft Ranking Dataset



- Data for 10,000 users.
- Relevance score is between 0 and 4.

Goal: Given a new target policy π^* and a user X, find the set of probable outcomes $\hat{C}(X)$

Application to Microsoft Ranking Dataset



Coverage of COPP vs other baselines with increasing policy shift.
Nominal coverage: 90%

	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$	$\Delta_\epsilon = 0.3$	$\Delta_\epsilon = 0.4$
COPP (Ours)	**0.90 ± 0.00**	**0.90 ± 0.02**	**0.90 ± 0.01**	**0.89 ± 0.01**	**0.91 ± 0.01**
WIS	1.00 ± 0.00	1.00 ± 0.00	0.92 ± 0.00	0.94 ± 0.00	0.91 ± 0.00
SBA	0.99 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.96 ± 0.00
CP (no policy shift)	**0.91 ± 0.02**	**0.92 ± 0.02**	0.93 ± 0.01	0.94 ± 0.01	0.96 ± 0.01

Interesting avenues for future work

- Conditional coverage guarantees rely on strong assumptions.
    - Interesting question for future work: Can these assumptions be weakened?
- Extending this to sequential decision making with evolving policies.
- Applying COPP to robust policy learning by optimising the worst case outcome.

Thanks for listening! Check out our paper at:

https://arxiv.org/abs/2206.04405