CONFORMAL OFF-POLICY PREDICTION

Yingying Zhang
Academy of Statistics and Interdisciplinary Sciences, KLATASDS-MOE
East China Normal University
Shanghai, China

Chengchun Shi
Department of Statistics
London School of Economics and Political Science
London, United Kingdom

Shikai Luo
Bytedance
Beijing, China

ABSTRACT

Off-policy evaluation is critical in a number of applications where new policies need to be evaluated offline before online deployment. Most existing methods focus on the expected return, define the target parameter through averaging and provide a point estimator only. In this paper, we develop a novel procedure to produce reliable interval estimators for a target policy's return starting from any initial state. Our proposal accounts for the variability of the return around its expectation, focuses on the individual effect and offers valid uncertainty quantification. Our main idea lies in designing a pseudo policy that generates subsamples as if they were sampled from the target policy so that existing conformal prediction algorithms are applicable to prediction interval construction. Our methods are justified by theories, synthetic data and real data from short-video platforms.

1 Introduction

Policy evaluation plays a crucial role in many real-world applications including healthcare, marketing, social sciences, among many others. Before deploying any new policy, it is crucial to know the impact of this policy. However, in the aforementioned applications, it is often impractical to evaluate a new policy by directly running this policy. As a result, the new policy needs to be evaluated offline based on an observational dataset generated by a possibly different behavior policy. This formulates the off-policy evaluation (OPE) problem.

Most works in the literature focus on evaluating the *average* value of a target policy aggregated over different initial states. In many applications such as healthcare and technology industries, in addition to the average effect, it is crucial to learn the value under a given initial condition (e.g., the individual effect) as well. For instance, in precision medicine, it allows us to estimate the outcome of each individual patient following a given treatment regime. In online recommendation, it allows us to evaluate the effect of a new strategy for each individual visitor. Moreover, real world datasets often follow asymmetric or heavy-tailed distributions. An example is given in our real dataset collected from a world-leading short-video platform where the outcome distribution is highly heavy-tailed (see Figure 6). In these applications, in additional to a target policy's *mean* return, it is equally important to infer its outcome distribution. This motivates us to construct a prediction interval for the target policy's outcome.

1.1 Related Work

Off-policy evaluation. There is a huge literature on OPE. Existing methods can be divided into three categories, corresponding to the value-based method (see e.g., Le et al., 2019; Luckett et al., 2020; Liao et al., 2021; Chen and Qi, 2022; Shi et al., 2022), importance sampling (IS) or resampling-based method (see e.g., Precup, 2000; Li et al., 2011; Liu et al., 2018; Nachum et al., 2019; Schlegel et al., 2019) and doubly robust method (see e.g., Farajtabar et al., 2018; Kallus and Zhou, 2018; Tang et al., 2019; Uehara et al., 2020; Kallus and Uehara, 2020; Liao et al., 2022).

In addition, several papers have studied interval estimation of the policy's value for uncertainty quantification (Thomas et al., 2015; Jiang and Li, 2016; Hanna et al., 2017; Dai et al., 2020; Feng et al., 2020; Jiang and Huang, 2020; Chandak et al., 2021; Hao et al., 2021; Shi et al., 2021; Wang et al., 2021). These confidence intervals are typically derived based on concentration inequalities, normal approximations, bootstrap (Efron and Tibshirani, 1994) or the empirical likelihood method (Owen, 2001). However, all the aforementioned methods focused on the average effect of the target policy. To our knowledge, interval estimation of the individual effect has been less explored in the literature.

Conformal prediction. Our proposal is closely related to a line of research on conformal prediction (CP), which was originally introduced by Vovk et al. (2005) to construct valid model-free *prediction* intervals (PIs) for the response; see also, some follow-up works by Vovk et al. (2009); Vovk (2012); Lei and Wasserman (2014); Lei et al. (2015, 2018); Sesia and Candès (2020); Cauchois et al. (2021). Both PI and confidence interval (CI) express uncertainty in statistical estimates. Nonetheless, a CI gives a range for the conditional mean function of the response whereas a PI aims to cover the response itself. A key strength of CP lies in its generality and finite-sample guarantees. Specifically, it can accommodate *any* prediction model under minimal assumptions on the data and achieve nominal coverage even for small samples.

Recently, Tibshirani et al. (2019) developed a weighted CP method to handle settings under covariate shift. The weighted CP method was further extended and applied to a number of applications, including individual treatment effects estimation (Kivaranovic et al., 2020; Jin et al., 2021; Lei and Candès, 2021; Yin et al., 2022), survival analysis Candès et al. (2021), classification under covariate shift (Podkopaev and Ramdas, 2021), to name a few. These methods considered either a standard supervised learning setting, or a contextual bandit setting with “state-agnostic" target policies. These settings differ from ours that involve sequential decision making and a general target policy.

Finally, we notice that there is a closely related concurrent work by Taufiq et al. (2022) that studied conformal off-policy prediction in contextual bandits. However, they did not consider sequential decision making, which is more challenging. In addition, even when specialized to contextual bandits, the proposed methodology differs largely from theirs. See Section 3.2 for more details.

Distributional reinforcement learning. Recently, there is an emerging line of research on distributional reinforcement learning that estimates the entire distribution of the return under the optimal policy (see e.g., Bellemare et al., 2017; Dabney et al., 2018; Mavrin et al., 2019; Zhou et al., 2020). Our proposal shares similar spirits with these works in that it not only considers the expected return, but takes the variability of the return around its expectation into account as well.

1.2 Contribution

Methodologically, we develop a novel procedure to construct off-policy PIs for a target policy's return starting from any initial state in sequential decision making. It is ultimately different from many existing OPE methods that consider the average effect aggregated over different initial states, construct CIs for the expected return and ignore the variance of the return around its expectation. A key ingredient of our proposal lies in constructing a pseudo variable whose distribution depends on both the target and behavior policy. We next sample a subset of observations based on this pseudo variable and apply weighted conformal prediction method on the selected subsamples. Finally, we develop an importance-sampling-based method and a multi-sampling-based method to further improve efficiency.

Theoretically, we prove that the proposed PI achieves valid coverage asymptotically. In addition, when the behavior policy is known to us (e.g., as in randomized studies), it achieves exact coverage in finite samples. Such a property is particularly appealing as the sample size is usually limited in offline domains. Finally, our PI is asymptotically efficient when the regression estimator is consistent.

2 Preliminaries: Conformal Prediction

We begin with a brief overview for the CP algorithm in supervised learning. Given i.i.d. predictor-response pairs $\{Z_i = (X_i, Y_i)\}_{i=1}^n$, it is concerned with producing a prediction band $\widehat{C}(\bullet)$ (as a function of the predictor X_i) such that for an identically distributed test data pair $Z_{n+1} = (X_{n+1}, Y_{n+1})$,

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1})) \geq 1 - \alpha, \tag{1}$$

for a given desired coverage rate $1 - \alpha \in (0, 1)$. One example of $\widehat{C}(x)$ is given by $[q_{\alpha_L}(x), q_{\alpha_U}(x)]$ where $q_{\alpha_L}(x)$ and $q_{\alpha_U}(x)$ correspond to the α_Lth (lower) and α_Uth (upper) conditional quantiles of Y given $X = x$ such that $\alpha_U - \alpha_L = 1 - \alpha$. Given the observed data, we can employ state-of-the-art machine learning algorithms to learn these conditional quantiles. This yields the following $\widehat{C}(x)$,

$$\widehat{C}(x) = [\widehat{q}_{\alpha_L}(x), \widehat{q}_{\alpha_U}(x)]. \tag{2}$$

However, one disadvantage of the aforementioned PI is that the inequality (1) is not guaranteed to hold in finite samples, due to the estimation errors of the the conditional quantiles.

The CP algorithm is developed to address this challenge. At a high-level, CP allows us to calibrate PIs (such as (2)) computed by general black box machine learning algorithms with finite-sample guarantees such that (1) holds for any n. Specifically, CP first splits the data into training and calibration data-subsets. On the training dataset, it learns the conditional mean or quantile of Y given X using any black box machine learning algorithm. For instance, let $\widehat{\mu}$ denote the estimated conditional mean function. On the calibration dataset $\{(X_i^{cal}, Y_i^{cal})\}_i$, it calculates a nonconformity score (e.g., $|Y_i^{cal} - \widehat{\mu}(X_i^{cal})|$) that measures how each observation "conforms" to the training dataset. The resulting PI is constructed based on the empirical quantiles of these nonconformity scores and attains valid coverage as long as the data observations are exchangeable. There are many choices of the score function available and we refer readers to Gupta et al. (2021) for details. Another widely-used score function is given by

$$\max\{\widehat{q}_{\alpha_L}(X_i^{cal}) - Y_i^{cal}, Y_i^{cal} - \widehat{q}_{\alpha_U}(X_i^{cal})\} \tag{3}$$

where $\widehat{q}_{\alpha_L}$ and $\widehat{q}_{\alpha_U}$ are the conditional quantile estimators given in (2). The resulting algorithm is referred to as the conformal quantile regression (Romano et al., 2019).

We next review the weighted CP algorithm developed by Tibshirani et al. (2019). As commented earlier, the aforementioned CP algorithm relies on exchangeability — a key assumption that requires the joint distribution of calibration and testing samples to be invariant to the order of samples. This assumption is clearly violated under distributional shift where the calibration and testing samples follow different distribution functions. To address this concern, Tibshirani et al. (2019) introduced the so-called "weighted exchangeability" that relaxes the classical i.i.d. assumption and is automatically satisfied for independent samples.

Definition 1. *(Weighted Exchangeability) $Z_1, \ldots, Z_n$ and the testing sample Z_{n+1} are said to be weighted exchangeable, if the density f of their joint distribution can be factorized as*

$$f(z_1, \ldots, z_{n+1}) = \prod_{i=1}^{n+1} w_i(z_i) g(z_1, \ldots, z_{n+1}),$$

for certain weight functions $\{w_i\}_{i=1}^{n+1}$, and a permutation-invariance function g such that $g(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)}) = g(z_1, \ldots, z_{n+1})$ for any permutation σ of $\{1, \ldots, n+1\}$.

According to the definition, independent data are always "weighted exchangeable" with weight function corresponding to the likelihood ratios.

Lemma 1. *Let $Z_i \sim P_i$, $i = 1, \ldots, n+1$ be independent draws, where each P_i is absolutely continuous with respect to P_1 for $i \geq 2$. Then $Z_1, \ldots, Z_{n+1}$ are weighted exchangeable with weight functions $w_1 = 1$ and $w_i = dP_i/dP_1$, $i \geq 2$.*

Let $S_i = \mathcal{S}(Z_i^{cal}, \mathcal{Z}^{tr})$ denote the nonconformity score for the ith observation in the calibration data based on certain machine learning algorithm trained on the training dataset $\mathcal{Z}^{tr}$, and $S_{(x,y)} = \mathcal{S}((x, y), \mathcal{Z}^{tr})$ denote the one for an arbitrary predictor-response pair (x, y). Instead of relying on the empirical quantiles of these nonconformity scores, the weighted CP algorithm considers a weighted version and constructs the PI for Y given $X = x$ as

$$\{y : S_{(x,y)} \leq (1-\alpha)\text{th quantile of the mixture distribution} \sum_{i=1}^{n} p_i^w \delta_{S_i} + p_{n+1}^w \delta_\infty\}$$

where $1 - \alpha$ denotes the desired coverage rate, δ_a denotes the Dirac delta distribution that places all mass at the value a, and the mixing probabilities $\{p_i^w\}_{i=1}^{n+1}$ are functions of weights $\{w_i\}_{i=1}^{n+1}$ whose explicit expressions are given in Tibshirani et al. (2019).

Finally, we remark that the (weighted) CP method possesses several appealing statistical properties. First, it does not depend on any specific model assumption in the conditional distribution of the outcome given the covariates; as such, it is applicable to complex nonlinear and high-dimensional settings. Second, it achieves exact coverage in the sense that $\mathbb{P}\{Y \in \widehat{C}_n(X)\} \geq 1-\alpha$ for any n. To the contrary, most interval estimation procedures are only *asymptotically* valid. Nonetheless, it is not straightforward to extend these methods to the OPE problem. See Section 3.2 for details.

3 Conformal Off-Policy Prediction in Contextual Bandits

3.1 Problem Formulation

To better illustrate the idea, in this section, we focus on a contextual bandit setting (i.e., single stage decision making) where the observed data consist of n i.i.d. samples $\{(X_i, T_i, Y_i)\}_{i=1}^n$ where X_i collects the contextual information of the ith instance, $T_i \in \{0, 1, \cdots, m-1\}$ denotes the treatment (e.g., action) that the ith instance receives where m denotes the number of treatment options, and Y_i is the corresponding response (e.g., reward). We adopt a counterfactual/potential outcome framework (Rubin, 2005) to formulate the OPE problem. Specifically, for any $0 \leq t \leq m-1$, let Y_i^t denote the reward that the ith instance would have been observed were they to receive action t.

A policy π is a (stochastic) decision rule that maps the contextual space to a distribution function over the action space. We use $\pi(t|x)$ to denote the probability that the agent selects treatment t given $X = x$. For a given target/evaluation policy π_e, we are interested in inferring the conditional distribution of the potential outcome $Y_{n+1}^{\pi_e}$ that would be observed were the instance to follow π_e given X_{n+1}. Specifically, given X_{n+1}, we aim to produce a PI for $Y_{n+1}^{\pi_e}$ with valid coverage guarantees. Notice that our objective differs from the standard OPE problem in which one aims to derive a CI for the expected value $\mathbb{E}Y_{n+1}^{\pi_e}$.

Finally, we impose standard assumptions in the causal inference literature (see e.g., Zhang et al., 2012; Zhu et al., 2017; Chen et al., 2022), including (1) $Y_i^{T_i} = Y_i$ almost surely for any i (i.e., consistency); (2) $(Y_i^0, \cdots, Y_i^{m-1}) \perp\!\!\!\perp T_i | X_i$ for any i (i.e., no unmeasured confounders); (3) The behavior policy $\pi_b(t|x) = \mathbb{P}(T_i = t | X_i = x)$ is uniformly bounded away from zero for any t, x (i.e., positivity).

3.2 Conformal Prediction for Off-Policy Evaluation

To motivate our proposed approach, we first outline two potential extensions of CP to the OPE problem in this section, corresponding to the direct method and the subsampling-based method, and discuss their limitations. We next illustrate the main idea of our proposal.

Direct method. OPE is essentially a policy evaluation problem under distribution shift where the target policy π_e differs from the behavior policy π_b that generates the offline data. By Lemma 1, the calibration dataset $\{(X_i^{cal}, Y_i^{cal})\}_i$ and the predictor-potential outcome pair $(X_{n+1}, Y_{n+1}^{\pi_e})$ in the target population satisfy weighed exchangeability with weights 1 for samples in the calibration dataset and $w_{n+1}(x,y)$ (given below) for the testing data

$$w_{n+1}(x,y) = \frac{dP_{Y^{\pi_e}|X}(y|x)}{dP_{Y|X}(y|x)}, \tag{4}$$

where $P_{Y|X}$ and $P_{Y^{\pi_e}|X}$ denote the conditional distributions of Y and Y^{π_e} given X, respectively. As a result, a direct application of the weighted CP method is valid for OPE given the weights $\{w_i\}_i$. We refer to the resulting algorithm as the direct method and notice that the concurrent work by Taufiq et al. (2022) adopted a similar idea. One key step in their proposal is to use the estimated conditional density function and the Monte Carlo method to learn the weight function (see equation 7 in Taufiq et al. (2022) for details). As such, their method can be sensitive to the specification of the conditional density model. On the contrary, our proposal below is robust to the model misspecification. We also conduct simulation studies in Section 5 to empirically verify the robustness property of our proposal.

To apply weighted CP, it remains to specify the weight w_{n+1}. Notice that both Y and Y^{π_e} correspond to a mixture of $\{Y^t : 1 \leq t \leq m\}$ with different weight vectors. Estimating w_{n+1} essentially requires to learn the conditional densities of Y^t given X — an extremely challenging task in complicated high-dimensional nonlinear systems. As will show later, this approach would fail to cover Y^{π_e} when the conditional density model is misspecified.

Subsampling-based method. Another approach to handle distributional shift is to take a data subset whose distribution is similar to the “target distribution" and apply standard CP to this sub-dataset. In particular, for each observation (X_i, T_i, Y_i) in the calibration dataset, we sample a pseudo action E_i following the evaluation policy π_e, select subsamples whose pseudo action matches the observed action, and apply CP to these subsamples. We refer to the resulting algorithm as the subsampling-based method.

However, this approach is not valid and is likely to produce PIs that undercover the target outcome $Y_{n+1}^{\pi_e}$ in general. This is because the distribution of the selected subsamples $\{(X_i, Y_i) : T_i = E_i\}$ generally differs from that of $(X_{n+1}, Y_{n+1}^{\pi_e})$. The two distributions coincide only when π_e is deterministic or π_b is uniformly random, as shown below.

Proposition 1. *Let E denote a pseudo action generated according to the target policy π_e. Then the conditional distribution of Y given $E = T$ and X follows a mixture distribution given as follows*

$$P_{Y|E=T,X} = \sum_{t=0}^{m-1} \frac{\pi_e(t|X)\pi_b(t|X)}{\sum_{t'} \pi_e(t'|X)\pi_b(t'|X)} P_{Y^t|X}.$$

The above mixture distribution equals $P_{Y^{\pi_e}|X} = \sum_t \pi_e(t|X)P_{Y^t|X}$ if and only if π_e is a deterministic policy or $\pi_b(0|X) = \pi_b(1|X) = \cdots = \pi_b(m-1|X)$.

Our proposal. The subsampling-based method fails because the distribution of the selected response differs from that of the potential outcome. To address this issue, instead of sampling according to the target policy π_e, we carefully design a pseudo/auxiliary policy π_a whose distribution depends on both π_e and π_b such that the resulting subsamples' distribution matches that of the potential outcome. More specifically, for any $0 \le t < m-1$ and x, π_a shall satisfy the following,

$$\frac{\pi_a(t|x)}{\pi_a(0|x)} = \frac{\pi_e(t|x)}{\pi_e(0|x)} \left[\frac{\pi_b(t|x)}{\pi_b(0|x)}\right]^{-1}. \tag{5}$$

In other words, $\pi_a(\bullet|x)$ shall be proportional to the ratio $\pi_e(\bullet|x)/\pi_b(\bullet|x)$ for any x. Similar to Proposition 1, let A denote the pseudo action generated according to π_a, we can show that subsamples with $A = T$ follow the following distribution,

$$\begin{aligned} P_{Y|A=T,X} &= \sum_{t=0}^{m-1} \frac{\pi_a(t|X)\pi_b(t|X)}{\sum_{t'} \pi_a(t'|X)\pi_b(t'|X)} P_{Y^t|X} \\ &= \sum_{t=0}^{m-1} \pi_e(t|X)P_{Y^t|X} = P_{Y^{\pi_e}|X}. \end{aligned}$$

This implies that subsampling according to the pseudo policy π_a yields the same conditional distribution as $P_{Y^{\pi_e}|X}$ in the target population. Nonetheless, the selected subsamples and the target possess different covariate distributions. Such a “covariate shift" problem can be naturally handled by the weighted CP algorithm. Using Lemma 1 again, the subsamples and the target population are weighted exchangeable with weights $w_i = 1$ for any i such that $A_i = T_i$ and

$$\begin{aligned} w_{n+1}(x, y) &= \frac{P_{X,Y^{\pi_e}}(x, y)}{P_{X,Y|A=T}(x, y)} = \frac{P_X(x)}{P_{X|A=T}(x)} \\ &= \frac{\mathbb{P}(A = T)}{\mathbb{P}(A = T|X = x)} \propto \frac{1}{\mathbb{P}(A = T|X = x)}, \end{aligned}$$

Compared to the direct method (see 7), the weight in the above expression depends only on the behavior policy which is known in randomized studies. and is independent of y. Consequently, our proposal is robust to the model misspecification of the conditional distribution $P_{Y^t|X}$, as shown later. When the behavior policy is unknown, it can be estimated based on existing supervised learning algorithms. We summarize our proposal in Algorithm 1, and call our method COPP, short for conformal off-policy prediction. Finally, we remark that by (5), $\pi_e = \pi_a$ only when π_e is deterministic or π_b is uniformly random. Consequently, the subsampling-based method is valid in these two special cases.

A numerical example. We conduct a simulation study to further demonstrate the sub-optimality of the direct and subsampling-based methods. We generate 500 data points from Example 1 of Section 5 for calibration and 10000 test data points. We consider a random target policy and a deterministic target policy. We further consider two conditional distribution models for $Y^t|X$, corresponding to a correctly specified model (denoted by “true"), and a misspecified model (denoted by “false") generated by injecting uniformly random noises on $(0, 1)$ to the oracle distribution function. It can be seen from Figure 1 that the direct method fails to cover the response when the conditional distribution model is misspecified whereas the subsampling-based method fails when the target policy is random. To the contrary, our proposal achieves valid coverage in all settings.

Statistical properties. Let $\widehat{\pi}_b(t|x)$ denote the estimated behavior policy, $w_{n+1}(x) = 1/\mathbb{P}(A = T|X = x)$ denote the oracle normalized weight function and $\widehat{w}_{n+1}$ denote the estimated weight function in Step 7 of Algorithm 1. We first show that COPP achieves valid coverage *asymptotically* when the behavior policy is consistently estimated. Notice that we do not require consistency of the estimated conditional outcome distribution.

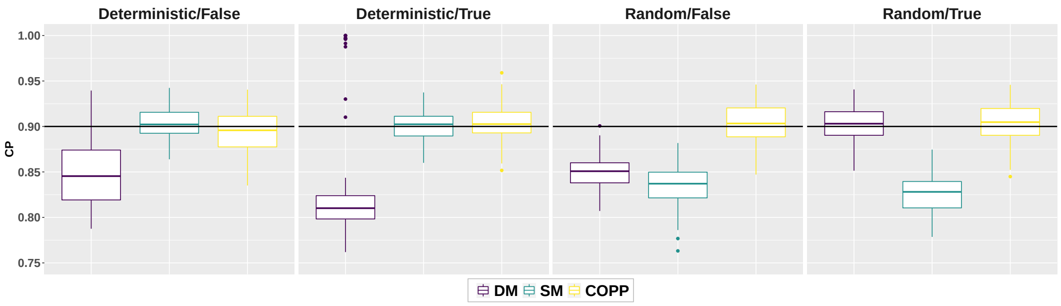

Figure 1: Empirical coverage probabilities (CPs) of PIs based on the Direct method (DM), Subsampling-based method (SM) and our proposal (COPP) in single-stage studies. The stochastic target policy is given by $\pi_e(1|X) = 1 - \pi_e(0|X) = \text{sigmoid}(-0.5 + X^{(1)} + X^{(2)} - X^{(3)} - X^{(4)})$ and the deterministic target policy is given by $\mathbb{I}(X^{(3)} + X^{(4)} > X^{(1)} + X^{(2)})$. The nominal level $1 - \alpha$ is 90%.

Algorithm 1 COPP for single-stage decision making

```
Input: Data $\{(X_i, T_i, Y_i)\}_{i=1}^n$; a test point $X_{n+1}$; a target policy $\pi_e$; number of treatment options $m$; propensity score training algorithm $\mathcal{P}$; quantile prediction algorithm $\mathcal{Q}$; quantile levels $\alpha_U$, $\alpha_L$ with $\alpha_U - \alpha_L = 1 - \alpha$.
    1: Split the data into two disjoint subsets $\mathcal{Z}^{tr} \cup \mathcal{Z}^{cal}$.
    2: Estimate $\pi_b(t|x)$ via $\mathcal{P}$ using samples from $\mathcal{Z}^{tr}$.
    3: Draw $\{A_i\}_{i=1}^n$ by plugging the propensity score estimator $\widehat{\pi}_b(t|x)$ in (5).
    4: Select subsamples satisfying $A_i = T_i$ in both data subsets. Denote them by $\mathcal{Z}^{tr,s}$ and $\mathcal{Z}^{cal,s}$.
    5: Apply quantile regressions via $\mathcal{Q}$ to selected subsamples from $\mathcal{Z}^{tr,s}$ to obtain the conditional quantile functions $\widehat{q}_{\alpha_L}$ and $\widehat{q}_{\alpha_U}$.
    6: Compute the conformity scores $\{S_i\}_i$ for all selected subsamples $i \in \mathcal{Z}^{cal,s}$ according to (3).
    7: For any $i \in \mathcal{Z}^{cal,s} \cup \{n+1\}$, estimate the weight $\widehat{w}_{n+1}(X_i) = \sum_{t=0}^{m-1} \pi_e(t|X_i)/\widehat{\pi}_b(t|X_i)$.
    8: For any $i$, compute the mixing probability $p_i^w = [\sum_{j \in \mathcal{Z}^{cal,s} \cup \{n+1\}} \widehat{w}_{n+1}(X_j)]^{-1} \widehat{w}_{n+1}(X_i)$.
    9: Compute $Q_{1-\alpha}(X_{n+1})$ as the $(1-\alpha)$th quantile of $\sum_{i \in \mathcal{Z}^{cal,s}} p_i^w \delta_{S_i} + p_{n+1}^w \delta_\infty$.
Output: the PI $\widehat{C}(X_{n+1}) = [\widehat{q}_{\alpha_L}(X_{n+1}) - Q_{1-\alpha}(X_{n+1}), \widehat{q}_{\alpha_U}(X_{n+1}) + Q_{1-\alpha}(X_{n+1})]$
```

Theorem 1 (Asymptotic coverage)**.** *Let $n_1 = |\mathcal{Z}^{tr}|$. Further, suppose that $\mathbb{E}[\widehat{w}_{n+1}(X)|\mathcal{Z}^{tr}] < \infty$, $\mathbb{E}[w_{n+1}(X)] < \infty$ and the consistency of behavior policy estimates (see the detailed requirements in Appendix A), then the output $\widehat{C}(x)$ from Algorithm 1 satisfies*

$$\lim_{n_1 \to \infty} \mathbb{P}(Y_{n+1}^{\pi_e} \in \widehat{C}(X_{n+1})) \geq 1 - \alpha.$$

Next, we show that if the propensity scores are known in advance, the proposed PI achieves exact coverage in finite samples.

Theorem 2 (Exact coverage)**.** *Suppose that $\mathbb{E}[w_{n+1}(X)] < \infty$, then the output $\widehat{C}(x)$ from Algorithm 1 with correctly specified propensity scores satisfies, for any sample size n,*

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

Finally, we show that the proposed PI is asymptotically efficient when the quantile regression estimator in Step 5 of Algorithm 1 is consistent.

Theorem 3 (Asymptotic efficiency)**.** *Suppose the behavior policy is known and the quantile regresssion estimates are consistent (see the detailed requirements in Appendix A), the output $\widehat{C}(x)$ from Algorihtm 1 satisfies*

$$L(\widehat{C}(X_{n+1}) \triangle C_\alpha^{oracle}(X_{n+1})) = o_p(1),$$

as $|\mathcal{Z}^{tr}|, |\mathcal{Z}^{cal}| \to \infty$. Here $L(A)$ indicates the Lebesgue measure of the set A, and $\triangle$ is the symmetric difference operator, i.e., $A \triangle B = (A \backslash B) \cup (B \backslash A)$, $C_\alpha^{oracle}(x)$ is the oracle interval defined as $[q_{\alpha_L}(x), q_{\alpha_U}(x)]$.

3.3 Extensions

In this section, we discuss two extensions of COPP, based on importance sampling and multi-sampling, respectively.

Extension 1. One limitation of COPP lies in that the PIs are constructed based only on observations in the subsamples. Nonetheless, when the target policy is stochastic, each observation has certain chance of being selected. To make full use of data, we adopt the importance sampling trick to compute the normalized weights and quantiles in Steps 8 and 9 of Algorithm 1, respectively. Specifically, in Step 7, we set the weight $\widehat{w}_{n+1}(X_i)$ for each of the sample in $\mathcal{Z}^{cal}$ to $\widehat{\pi}_a(T_i|X_i)\widehat{w}_{n+1}(X_i)$. These weights are then passed to Step 8 to compute $\widehat{p}_i$, and subsequently to Step 9 to calculate $Q_{1-\alpha}(X_{n+1})$ by replacing $\mathcal{Z}^{cal,s}$ with the whole calibration set $\mathcal{Z}^{cal}$. As we will show in Section 5, this procedure is much more efficient than COPP when the selected subsamples contains only a few observations. We next prove that such an extension achieves valid coverage as well.

Theorem 4. *Under the conditions of Theorem 1, we have*

$$\lim_{n_1\to\infty} \mathbb{P}(Y_{n+1}^{\pi_e} \in \widehat{C}(X_{n+1})) \geq 1-\alpha.$$

Extension 2. The second extension integrates COPP with the multi-sampling method. Notice that Algorithm 1 only implements subsampling once. The result can be very sensitive to the selected subsamples. To mitigate the randomness the single-sampling procedure introduces, we propose to repeat COPP multiple times and then aggregate all these PIs to gain efficiency. To combine multiple PIs, we adopt the idea proposed by Solari and Djordjilović (2022) for multi-split conformal prediction. A key observation is that, the PI in Algorithm 1 is equivalent to $\widehat{C}(X_{n+1}) = \{y : p(X_{n+1}, y) \geq \alpha\}$ where $p(X_{n+1}, y)$ is given by

$$\sum_{i\in\mathcal{Z}^{cal,s}} p_i^w \mathbb{I}[\max\{\widehat{q}_{\alpha_L}(X_{n+1}) - y, y - \widehat{q}_{\alpha_U}(X_{n+1})\} \leq S_i] + p_\infty^w,$$

serving as a p-value for the testing hypotheses $H_0 : Y_{n+1}^{\pi_e} = y$ against $H_1 : Y_{n+1}^{\pi_e} \neq y$ given X_{n+1}. This allows us to follow the idea of Meinshausen et al. (2009) for p-value aggregation. Let $p^b(x, y)$ for $1 \leq b \leq B$ be the p-values for B constructed PIs with significance level $\alpha\gamma$ for certain tuning parameter $0 < \gamma < 1$. We aggregate these p-values by setting $\bar{p}(X_{n+1}, y)$ to their empirical γ-quantile. The final PI is given by $\widehat{C}_{B,\gamma}(X_{n+1}) = \{y : \bar{p}(X_{n+1}, y) \geq \alpha\}$.

Theorem 5. *Under the conditions of Theorem 1, we have for any $B > 0$ and $0 < \gamma < 1$,*

$$\lim_{n_1\to\infty} \mathbb{P}(Y_{n+1}^{\pi_e} \in \widehat{C}_{B,\gamma}(X_{n+1})) \geq 1-\alpha.$$

Finally, we remark that we only derive the asymptotic coverage of the two extensions in Theorems 4 and 5. Nonetheless, when the behavior policy is known, these methods also achieve exact coverage.

4 Conformal Off-Policy Prediction in Sequential Decision Making

Problem formulation. In this section, we consider sequential desicion making where the observed data consist of n i.i.d samples $\{(X_{1i}, T_{1i}, X_{2i}, T_{2i}, \ldots, X_{Ki}, T_{Ki}, Y_i)\}_{i=1}^n$ where for the ith instance, X_{ki} collects the state information at the kth stage, $T_{ki} \in \{0, \ldots, m-1\}$ denotes the action at the kth stage , Y_i is the corresponding reward at the final stage. Such a sparse reward setting is frequently considered for precision medicine type applications (Murphy, 2003). Meanwhile, our method is equally applicable to settings with immediate rewards at each decision point (see Appendix B).

Let $H_k = \{X_1, T_1, \ldots, X_k\}$ denote the history up to the kth stage. We define a (history-dependent) policy $\Pi = (\pi_1(t_1|h_1), \pi_2(t_2|h_2), \ldots, \pi_K(t_K|h_K))$ as a sequence of (stochastic) decision rules where each $\pi_k(t_k|h_k)$ determines the probability that an agent selects action t_k at the kth stage given that $H_k = h_k$. For a given target policy π_e, we are interested in constructing PIs for the potential outcome Y^{π_e} that would be observed were the instance to follow π_e for any initial state X_1. To save space, we impose the consistency, sequential ignorability and positivity assumption in Appendix B.

COPP. We generalize our proposal in Section 3.2 to sequential making decision. We design a pseudo policy $\pi_a = \{\pi_{a,k}\}_k$ which relies on both $\pi_b = \{\pi_{b,k}\}_k$ and $\pi_e = \{\pi_{e,k}\}_k$, to generate subsamples whose outcome distribution conditional on the state-action history matches that of the potential outcome. Specifically, for any $1 \leq k \leq K$ and h_k, the pseudo policy $\pi_{a,k}(\bullet|h_k)$ shall be proportional to the ratio $\pi_{e,k}(\bullet|h_k)/\pi_{b,k}(\bullet|h_k)$. Similar to Proposition 1, we can show that the conditional density of $Y|A_K = T_K, H_K$ equals that of $Y^{\pi_e}|H_K$.

More importantly, by iteratively integrating over the space of $\{T_k, X_{k+1}, \cdots, X_K\}$, we can show that the conditional density of $Y|A_k = T_k, \cdots, A_K = T_K, H_k$ also equals that of $Y^{\pi_e}|H_k$ for each k. Using Lemma 1 again, the subsamples $\{(H_{1i}, Y_{1i}) : A_{ki} = T_{ki}, 1 \le k \le K, 1 \le i \le n\}$ and the target population $(H_{1,n+1}, Y_{n+1})$ are weighted exchangeable with weights $w_i = 1$ for any i and

$$w_{n+1}(h) \propto \mathbb{P}^{-1}(A_1 = T_1, \cdots, A_K = T_K|H_1 = h).$$

Based on these weights, the PIs can be similarly derived as in Algorithm 1. We defer the pseudocode and the statistical properties of the constructed PIs to Appendix B.

Extensions. Our proposal suffers from the “curse of horizon" (Liu et al., 2018) in that the number of selected subsamples decreases exponentially fast with respect to the number of decision stages. While this phenomenon is unavoidable without further model assumptions (Jiang and Li, 2016), the importance-sampling-based and multi-sampling-based approach alleviate this issue to some extent, as shown in our simulations. Since these extensions are very similar to those presented in Section 3.3, we omit them for brevity.

5 Synthetic Data Analysis

In this section, we conduct simulation studies to investigate the empirical performance of our proposed methods. In particular, we focus on the following three examples: two examples considered in Wang et al. (2018):

Example 1 (Single-Stage Decision Making):

- The baseline covariates $X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}$ are independently uniformly generated from $(0, 1)$.
- The action is binary and satisfies $\mathbb{P}(T = 1|X) = \text{sigmoid}(-0.5 - 0.5\sum_{j=1}^4 X^{(j)})$ where $\text{sigmoid}(t) = \exp(t)/[1 + \exp(t)]$.
- The return is given by $Y = 1 + X^{(1)} - X^{(2)} + (X^{(3)})^3 + \exp(X^{(4)}) + T(3 - 5X^{(1)} + 2X^{(2)} - 3X^{(3)} + X^{(4)}) + (1 + T)(1 + \sum_{j=1}^4 X^{(j)})\epsilon$ where ϵ is a standard normal variable independent of X and T.
- The target policy π_e satisfies $\pi_e(1|X) = \text{sigmoid}(-0.5 + X^{(1)} + X^{(2)} - X^{(3)} - X^{(4)}))$.

Example 2 (Two-Stage Decision Making):

- Observations and actions are generated as follows:

$$\begin{aligned} X_1 &\sim \text{Uniform}(0, 1), \\ T_1|X_1 &\sim \text{Bernoulli}(\text{sigmoid}(-0.5 + X_1)), \\ X_2|X_1, T_1 &\sim \text{Uniform}(X_1, X_1 + 1), \\ T_2|X_1, T_1, X_2 &\sim \text{Bernoulli}(\text{sigmoid}(-0.5 - X_2)). \end{aligned}$$

- The final return is given by $Y = 1 + X_1 + T_1[1 - 3(X_1 - 0.2)^2] + X_2 + T_2[1 - 5(X_2 - 0.4)^2] + (1 + 0.5T_1 - T_1X_1 + 0.5T_2 - T_2X_2)\epsilon$ for a standard normal variable ϵ independent of observations and actions.
- The target policy is defined as follows

$$\begin{aligned} E_1|X_1 &\sim \text{Bernoulli}(\text{sigmoid}(0.5X_1 - 0.5)), \\ E_2|X_1, E_1, X_2 &\sim \text{Bernoulli}(\text{sigmoid}(0.5X_2 - 1)). \end{aligned}$$

For each example, we further consider two settings. In the high-dimensional setting, we manually include $100 - p_0$ null variables that are uniformly distributed on $(0, 1)$ in the state with $p_0 = 4$ and 1 in Examples 1 and 2, respectively. In the low-dimensional setting, these null variables are not included. This yields a total of four different scenarios. The sample size is fixed to 2000.

Example 3 (Multi-Stage Decision Making): We design an additional simulation studies where the number of horizon (e.g., the decision stages) equals 3, 4 or 5, and investigate the performance of our methods under this setting.

- Observations and actions are generated as follows:

$$\begin{aligned} X_1 &= 0.5\epsilon_1, \epsilon_k \sim N(0, 1), 1 \le k \le m, \\ T_k|X_k &\sim \text{Bernoulli}(\text{sigmoid}(-0.5 + X_k)), \\ X_k &= 0.5X_{k-1} + 0.1T_{k-1} + 0.5\epsilon_k. \end{aligned}$$

- The final return is given by $Y_m = X_m$.
- The target policy is defined as follows

$$D_k|X_k \sim \text{Bernoulli}(\text{sigmoid}(-0.5 + 0.5X_k)).$$

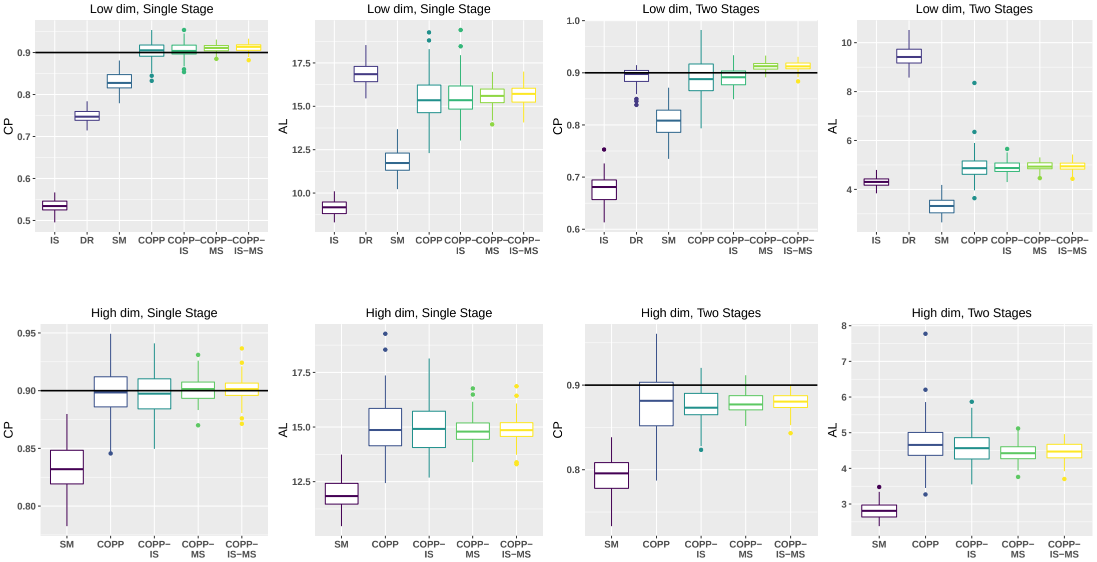

Figure 2: Empirical coverage probabilities (CPs) and average lengths (ALs) of intervals based on SM, IS, DR, and our proposed COPP, COPP-IS, COPP-MS, COPP-IS-MS in four settings. The nominal level is 90%.

The data consist of 2000 observations, in which three quarters are used for training and the rest for validation.

Implementation details. We estimate the behavior policy using logistic regression. In the high-dimensional setting, we apply penalized regression to improve the estimation efficiency. The conditional quantile functions are estimated based on quantile regression forest (Meinshausen and Ridgeway, 2006). Following Sesia and Candès (2020), we use 75% of the data for training and the rest for calibration. We fix $\alpha_L = \alpha/2$ and $\alpha_U = 1 - \alpha/2$ in all settings. In addition, to implement the multi-sampling-based method, we fix $\gamma = 1/2$ and set the significance level to α instead of $\alpha\gamma = \alpha/2$ to improve the precision (interval length). We find that the resulting PI achieves nominal coverage in practice. The number of intervals B is set to 100 in the low-dimensional setting, and 50 in high dimension to reduce computation time. Finally, in each simulation, we generate 10000 test data points in the target population to evaluate the converge probability. The R code is released here.

Benchmark specification. We compare our proposed methods against the subsampling-based method (SM) detailed in Section 3.2. In low-dimensional settings, we also compare with the standard importance sampling (IS) and doubly robust (DR) method (see e.g., Dudík et al., 2011; Zhang et al., 2012; Jiang and Li, 2016) designed for off-policy confidence interval estimation. These methods focus on the average effect. We couple them with kernel density estimation to infer the individual effect conditional on the initial state. Please refer to Appendix C for the detailed implementation.

Results. Figure 2 reports the coverage probability and average length of various interval estimators for Examples 1 and 2, aggregated over 100 simulations. We denote the extensions of our proposal based on importance-sampling, multi-sampling alone and a combination of the two are denoted by COPP-IS, COPP-MS and COPP-IS-MS, respectively. We summarize our findings below. First, intervals based on SM, IS and DR significantly undercover the potential outcome. As we have commented, these methods are not valid in general. SM requires either a uniformly random behavior policy, or a deterministic target policy. IS and DR focus on the expected return and ignore the variability of the return around its expectation. Second, all the proposed methods achieve nominal coverage in most cases. Among them, the multi-sampling-based methods (COPP-MS and COPP-IS-MS) achieve the best performance, with substantially reduced variability compared to the single-sampling-based methods. In addition, COPP-IS performs much better than COPP in two-stage settings where the number of subsamples is limited, as expected.

We report the results for Example 3 in Figure 3, where the proposed COPP, COPP-IS, COPP-MS, COPP-IS-MS with horizon m are labelled as m-1, m-2, m-3 and m-4, respectively. It can be seen that the proposed method is able to achieve nominal coverage in general. Nonetheless, as commented in our paper, it suffers from the curse of horizon and would be inefficient in long-horizon settings. It remains unclear how to break the curse of horizon and we leave it as future work.

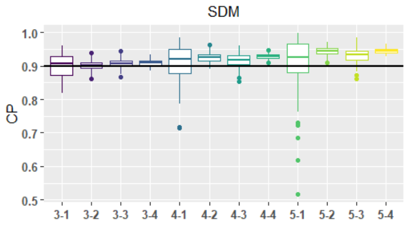

Figure 3: Coverage Probability for SDM for Horizons 3,4,5.

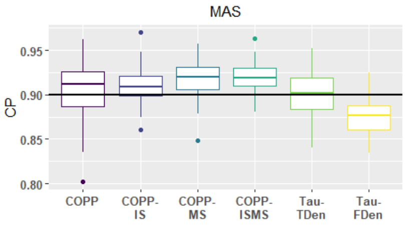

Figure 4: Coverage Probability for Multiple Action Space

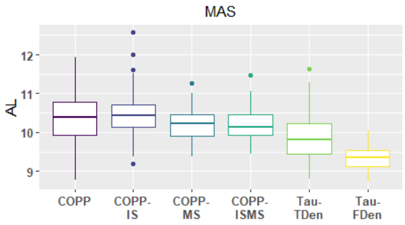

Figure 5: Average Length for Multiple Action Space

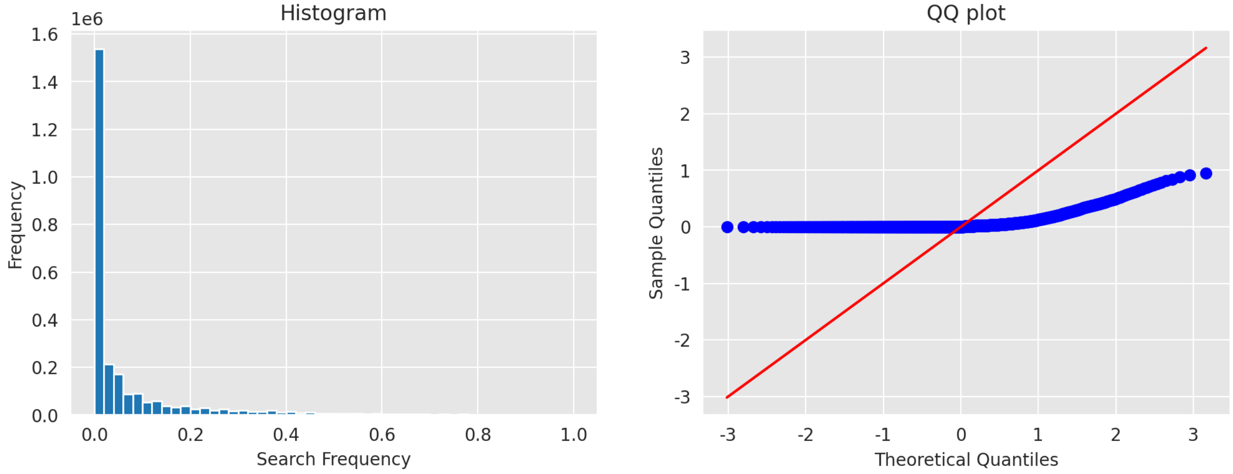

Figure 6: Histogram and QQ plot of users' search frequencies. In the QQ plot, the blue curve depicts the empirical quantiles of the search frequency against those from a standard normal distribution function.

Comparison with Taufiq et al. (2022). We consider the simulation setting described in Section 6.1.1 of Taufiq et al. (2022) with four actions, implement their methods with a correctly specified f (denoted by "Tau-TDen") and a misspecified model (denoted by"Tau-FDen"), and compare both methods against our proposal with a misspecified f. All these methods additionally require to specify the propensity score model and we use a correctly specified model. It is seen from Figures 4 and 5 below that unlike our proposal, Taufig et al. (2022)'s method is sensitive to the specification of the conditional density function.

6 Real Data Analysis

We illustrate the usefulness of our method based on a dataset collected from a world-leading technology company. This company has one of the largest mobile platforms for production, aggregation and distribution of short-form videos with extensive search functionality. It implements a strategy to encourage users to explore its search functionality. Specifically, when a user launches the app for the first time in a day, they will see a pop-up window that recommends them to use the search feature. However, pop-ups are annoying for some users. As such, the company's interested in 'pop-up' policies that implement this strategy to a subgroup of users to increase their search frequency.

The dataset is collected from an online experiment which involves two millions daily active users and has been scaled due to privacy. The features available to us consist of each user's history information including the frequency they used the app and the search functionality prior to the experiment. The reward is the user's search frequency after treatment and is highly heavy-tailed as shown in Figure 6. As such, instead of focusing on a target policy's expected return, we are interested in its entire distribution. As commented earlier, most existing OPE methods are not directly applicable. In addition, since the behavior policy is known to us, the proposed method is robust to the model misspecification of the outcome distribution, and achieves exact coverage.

To investigate the validity of the proposed method, we equally split the dataset into two, one for learning an optimal policy and the other for policy evaluation. On the evaluation dataset, we further employ ten-fold cross-validation to test our method. Specifically, we randomly split the data into ten folds, use nine of them to train the proposed PIs and the remaining fold to estimate their coverage probabilities. We further aggregate these coverage probabilities over different folds to get the full efficiency. In our implementation, the number of intervals B is set to 100, and other configurations are consistent with those in Section 5. It turns out that the average coverage probabilities of COPP, COPP-MS, COPP-IS and COPP-IS-MS are all close to the nominal level 90%, and equal 91.0%, 90.5%, 89.6% and 90.4%, respectively. The lower and upper bounds of the proposed PIs offer a more accurate characterization of the target policy's return and give practitioners more information when conducting online A/B tests.

7 Conclusion and Discussion

To our knowledge, our proposal is the first to procedure statistically sound PIs for a target policy's return in sequential decision making. The proposed PIs focus on the individual effect, take the variability of the return around its mean into consideration, achieve finite-sample coverage guarantees and are robust to the misspecification of the conditional outcome model. Currently, we consider a discrete action space. It would be practically interesting to extend our proposal to the continuous action setting. However, this is beyond the scope of the current paper and we leave it for future research.

Acknowledgements

Zhang's research is supported by the National Natural Science Foundation of China (NSFC) grant nos. 12101241. Shi's research is partly support by the EPSRC grant EP/W014971/1. The authors wish to thank all the constructive comments from the referees, the area chair and the chair, which has lead to a significant improvement of the earlier version of this paper.

References

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, volume 70, pages 449–458. PMLR.

Candès, E. J., Lei, L., and Ren, Z. (2021). Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*.

Cauchois, M., Gupta, S., and Duchi, J. C. (2021). Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.*, 22:81–1.

Chandak, Y., Niekum, S., da Silva, B., Learned-Miller, E., Brunskill, E., and Thomas, P. S. (2021). Universal off-policy evaluation. In *Advances in Neural Information Processing Systems*, volume 34.

Chen, G., Li, X., and Yu, M. (2022). Policy learning for optimal individualized dose intervals. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pages 1671–1693. PMLR.

Chen, X. and Qi, Z. (2022). On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. *arXiv preprint arXiv:2201.06169*.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. (2018). Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2892–2901.

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvári, C., and Schuurmans, D. (2020). Coindice: Off-policy confidence interval estimation. In *Advances in neural information processing systems*, volume 33, pages 9398–9411.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, volume 80, pages 1447–1456. PMLR.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. (2020). Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, volume 119, pages 3102–3111. PMLR.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2021). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496.

Hanna, J. P., Stone, P., and Niekum, S. (2017). Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 4933–4934.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvari, C., and Wang, M. (2021). Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR.

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 2747–2758.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, volume 48, pages 652–661. PMLR.

Jin, Y., Ren, Z., and Candès, E. J. (2021). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *arXiv preprint arXiv:2111.12161*.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63.

Kallus, N. and Zhou, A. (2018). Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR.

Kivaranovic, D., Ristl, R., Posch, M., and Leeb, H. (2020). Conformal prediction intervals for the individual treatment effect. *arXiv preprint arXiv:2006.01474*.

Le, H., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning*, volume 97, pages 3703–3712. PMLR.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J., Rinaldo, A., and Wasserman, L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1):29–43.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Li, L., Chu, W., Langford, J., and Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306.

Liao, P., Klasnja, P., and Murphy, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391.

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. (2022). Batch policy learning in average reward markov decision processes. *Annals of Statistics*, accepted.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, volume 31.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706.

Mavrin, B., Yao, H., Kong, L., Wu, K., and Yu, Y. (2019). Distributional reinforcement learning for efficient exploration. In *International conference on machine learning*, volume 97, pages 4424–4434. PMLR.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(6):983–999.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32.

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, volume 161, pages 844–853. PMLR.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. In *Advances in neural information processing systems*, volume 32.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Schlegel, M., Chung, W., Graves, D., Qian, J., and White, M. (2019). Importance resampling for off-policy prediction. *Advances in Neural Information Processing Systems*, 32.

Sesia, M. and Candès, E. J. (2020). A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. (2021). Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, volume 139, pages 9580–9591. PMLR.

Shi, C., Zhang, S., Lu, W., and Song, R. (2022). Statistical inference of the value function for reinforcement learning in infinite horizon settings. *Journal of the Royal Statistical Society*, 84:765–793.

Solari, A. and Djordjilović, V. (2022). Multi split conformal prediction. *Statistics & Probability Letters*, 184:109395.

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. (2019). Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*.

Taufiq, M. F., Ton, J.-F., Cornish, R., Teh, Y. W., and Doucet, A. (2022). Conformal off-policy prediction in contextual bandits. *arXiv preprint arXiv:2206.04405*.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. (2015). High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pages 3000–3006.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in neural information processing systems*, volume 32.

Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, volume 119, pages 9659–9668. PMLR.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR.

Vovk, V., Gammerman, A., and Shafer, G. (2005). Conformal prediction. *Algorithmic learning in a random world*, pages 17–51.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2009). On-line predictive linear regression. *The Annals of Statistics*, pages 1566–1590.

Wang, J., Qi, Z., and Wong, R. K. (2021). Projected state-action balancing weights for offline reinforcement learning. *arXiv preprint arXiv:2109.04640*.

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018). Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254.

Yin, M., Shi, C., Wang, Y., and Blei, D. M. (2022). Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, accepted.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhou, F., Wang, J., and Feng, X. (2020). Non-crossing quantile regression for distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 15909–15919.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S., and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, 73(2):391–400.

A PROOF OF MAIN RESULT

This section provides the additional assumptions and proofs for Theorems 1-5 for COPP, COPP-IS, COPP-MS in contextual bandits. We first provide two assumptions concerning the consistency of behavior policy estimates $\widehat{\pi}_b(t|x)$. Assumption 1 is similar to A1 in Lei and Candès (2021). Let $\widetilde{A}$ and A denote the pseudo actions generated according to the estimated behavior policy $\widehat{\pi}_b$ obtained from the training dataset $\mathcal{Z}^{tr}$ and the oracle behavior policy π_b, respectively. Let $P_{X|\widetilde{A}=T}$ be the distribution function of the contextual information in selected samples $\mathcal{Z}^{cal,s}$ as an estimate of $P_{X|A=T}$. Since observations in the calibration dataset follow same distribution, it suffices to estimate the weight function w_{n+1}. To ease notation, we will omit the subscript $n+1$ in w_{n+1} and $\widehat{w}_{n+1}$ when there is no confusion.

Assumption 1. *For any* $0 \le t \le m-1$, $\lim\limits_{n_1\to\infty} \mathbb{E}_{X\sim P_X} |1/\widehat{\pi}_b(t|X) - 1/\pi_b(t|X)| \to 0$.

Assumption 2. $\mathbb{E}_{X\sim P_{X|\widetilde{A}=T}}[\widehat{w}(X)|\mathcal{Z}^{tr}] < \infty$.

Proof of Theorem 1 (Asymptotic coverage). The proof is similar to that of Theorem 3 in Lei and Candès (2021). First, we index the data points in the selected calibration dataset $\mathcal{Z}^{cal,s}$ by $\{1, 2, \dots, n_2'\}$. Notice that $P_{X|\widetilde{A}=T}$ is close to $P_{X|A=T}$ given a consistent behavior policy estimator. In addition, the two conditional distributions are the same if the propensity score is known in advance. The distribution of selected outcome is given by $P_{Y|\widetilde{A}=T,X}$, which aims to approximate $P_{Y|A=T,X}$. By definition, we have

$$P_{Y|\widetilde{A}=T,X} \quad = \quad \sum_{t=0}^{m-1} \frac{\pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x)}{\sum_{t'} \pi_e(t'|x)\pi_b(t'|x)/\widehat{\pi}_b(t'|x)} P_{Y^t|X}.$$

Using Bayesian rule,

$$\begin{aligned} dP_{X|\widetilde{A}=T}(x) &= \frac{P(\widetilde{A}=T|X=x)dP_X(x)}{\int_x P(\widetilde{A}=T|X=x)dP_X(x)} \\ &= \frac{(\sum_{t=0}^{m-1} \pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x))/(\sum_{t=0}^{m-1} \pi_e(t|x)/\widehat{\pi}_b(t|x))dP_X(x)}{\int_x (\sum_{t=0}^{m-1} \pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x))/(\sum_{t=0}^{m-1} \pi_e(t|x)/\widehat{\pi}_b(t|x))dP_X(x)}. \end{aligned} \tag{6}$$

Since the denominator is a constant and the weighted conformal inference is invariant to rescaling of the likelihood ratio, we can reset the estimated weights by multiplying the denominator in the above. In other words, instead of setting $\widehat{w}(x)$ to $\sum_{t=0}^{m-1} \pi_e(t|x)/\widehat{\pi}_b(t|x)$ as in the main paper, we define

$$\widehat{w}(x) = \frac{\int_x (\sum_{t=0}^{m-1} \pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x))/(\sum_{t=0}^{m-1} \pi_e(t|x)/\widehat{\pi}_b(t|x))dP_X(x)}{1/(\sum_{t=0}^{m-1} \pi_e(t|x)/\widehat{\pi}_b(t|x))}. \tag{7}$$

Let $Z_i = (X_i, Y_i)$ for $1 \le i \le n_2'$ where $(X_i, Y_i) \sim P_{X|\widetilde{A}=T} \times P_{Y|X,\widetilde{A}=T}$ and $Z_{n+1} = (X_{n+1}, Y_{n+1}) \sim P_X \times P_{Y^{\pi_e}|X}$. Recall that $P_{Y|X,A=T} = P_{Y^{\pi_e}|X}$ and $P_{Y|X,\widetilde{A}=T} = P_{Y^{\pi_e}|X}$ if and only if $\widehat{\pi}_b(t|x) = \pi_b(t|x)$ for any $0 \le t \le m-1$ and x. The true weight function w is set to

$$w(x) = \frac{dP_X(x)}{dP_{X|A=T}(x)},$$

accordingly. Under the assumption that $\mathbb{E}[w(X)] < \infty$, we have $\mathbb{P}_{X\sim P_X}(w(X) < \infty) = 1$. Similarly, under the assumption that $\mathbb{E}[\widehat{w}(X)|\mathcal{Z}^{tr}] < \infty$, we obtain

$$\mathbb{P}_{X\sim P_X}(\widehat{w}(X) < \infty) = 1.$$

Let $\widetilde{P}_X$ denote a context distribution function such that $d\widetilde{P}_X(x)$ is proportional to $\widehat{w}(x)dP_{X|\widetilde{A}=T}(x)$. Under Assumption 2, we have that $\mathbb{P}_{X\sim P_{X|\widetilde{A}=T}}(\widehat{w}(X) < \infty|\mathcal{Z}^{tr}) = 1$, which in turn implies that $\widetilde{P}_X$ is a probability measure. Consider now a new sample $\widetilde{Z}_{n+1} = (\widetilde{X}_{n+1}, \widetilde{Y}_{n+1}) \sim \widetilde{P}_X \times P_{Y|\widetilde{A}=T,X}$.

Let $E_{\widetilde{z}}$ denote the event that $\{Z_1, \dots, Z_{n_2'}, \widetilde{Z}_{n+1}\} = \{z_1, \dots, z_{n_2'}, \widetilde{z}_{n+1}\}$. The corresponding nonconformity scores are denoted as $\widetilde{S} = (S_1, \dots, S_{n_2'}, \widetilde{S}_{n+1})$, $s_i = \mathcal{S}(z_i, \mathcal{Z}^{tr})$ for $1 \le i \le n_2'$ and $\widetilde{s}_{n+1} = \mathcal{S}(\widetilde{z}_{n+1}, \mathcal{Z}^{tr})$. Without loss of generality, assume these scores are discrete-valued. For each $1 \le i \le n_2'$,

$$\mathbb{P}\{\widetilde{S}_{n+1} = s_i|E_{\widetilde{z}}\} = \mathbb{P}\{\widetilde{Z}_{n+1} = z_i|E_{\widetilde{z}}\} = \frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})}{\sum_\sigma f(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})}$$

where σ corresponds to the permutation of numbers $\{1, \dots, n_2', n+1\}$ and f is the joint density of $\{Z_1, \dots, Z_{n_2'}, \widetilde{Z}_{n+1}\}$. For instance, suppose $\sigma(i) = n+1$, then we have $z_{\sigma(i)} = \widetilde{z}_{n+1}$. Notice that the condition distribution of $\widetilde{Y}_{n+1}$ given $\widetilde{X}_{n+1}$ is the same as those in the calibration data. It follows from Lemma 1 that

$$\mathbb{P}\{\widetilde{S}_{n+1} = s_i | E_{\widetilde{z}}\} = \frac{\sum_{\sigma:\sigma(n+1)=i} \widehat{w}(x_{\sigma(n+1)})}{\sum_{\sigma} \widehat{w}(x_{\sigma(n+1)})} = \frac{\widehat{w}(x_i)}{\sum_{i\in\mathcal{Z}^{cal,s}} \widehat{w}(x_i) + \widehat{w}(\widetilde{x}_{n+1})} \equiv \widehat{p}_i(\widetilde{x}_{n+1}), \tag{8}$$

Similarly,

$$\mathbb{P}\{\widetilde{S}_{n+1} = \widetilde{s}_{n+1} | E_{\widetilde{z}}\} = \frac{\widehat{w}(x_i)}{\sum_{i\in\mathcal{Z}^{cal,s}} \widehat{w}(x_i) + \widehat{w}(\widetilde{x}_{n+1})} \equiv \widehat{p}_{n+1}(\widetilde{x}_{n+1}).$$

This yields that

$$\widetilde{S}_{n+1} | E_{\widetilde{z}} \sim \sum_{i=1}^{n_2'} \widehat{p}_i(\widetilde{x}_{n+1}) \delta_{s_i} + \widehat{p}_{n+1}(\widetilde{x}_{n+1}) \delta_{\widetilde{S}_{n+1}}.$$

By Lemma 1 in Tibshirani et al. (2019), it is equivalent to

$$\widetilde{S}_{n+1} | E_{\widetilde{z}} \sim \sum_{i=1}^{n_2'} \widehat{p}_i(\widetilde{x}_{n+1}) \delta_{v_i} + \widehat{p}_{n+1}(\widetilde{x}_{n+1}) \delta_{\infty}. \tag{9}$$

As a consequence,

$$\begin{aligned} &\mathbb{P}(\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{X}_{n+1}) | \mathcal{Z}^{tr}) = \mathbb{P}(\widetilde{S}_{n+1} \le \eta(\widetilde{X}_{n+1}) | \mathcal{Z}^{tr}) \\ &= \mathbb{P}(\widetilde{S}_{n+1} \le \text{Quantile}(1-\alpha; \sum_{i=1}^{n_2'} \widehat{p}_i(\widetilde{x}_{n+1}) \delta_{v_i} + \widehat{p}_{n+1}(\widetilde{x}_{n+1}) \delta_{\infty} | \mathcal{Z}^{tr})) \ge 1-\alpha, \end{aligned}$$

where the last inequality follows from (9). In addition, we have that

$$\begin{aligned} &\left|\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1}) | \mathcal{Z}^{tr}, \mathcal{Z}^{cal}) - \mathbb{P}(\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{X}_{n+1}) | \mathcal{Z}^{tr}, \mathcal{Z}^{cal})\right| \\ &\le d_{\text{TV}}(\widetilde{P}_X \times P_{Y|\widetilde{A}=T,X}, P_X \times P_{Y^{\pi^e}|X}). \end{aligned}$$

Taking expectation with respect to $\mathcal{Z}^{cal}$ on both sides, we obtain that

$$\left|\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1}) | \mathcal{Z}^{tr}) - \mathbb{P}(\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{X}_{n+1}) | \mathcal{Z}^{tr})\right| \le d_{\text{TV}}(\widetilde{P}_X \times P_{Y|\widetilde{A}=T,X}, P_X \times P_{Y^{\pi^e}|X}).$$

Recall that $\widetilde{P}_X$ is defined as the distribution function such that $d\widetilde{P}_X(x)$ is proportional to $\widehat{w}(x) dP_{X|\widetilde{A}=T}(x)$. It follows that

$$\begin{aligned} & d_{\text{TV}}(\widetilde{P}_X \times P_{Y|\widetilde{A}=T,X}, P_X \times P_{Y^{\pi^e}|X}) \\ \le\ & \frac{1}{2} \sum_{t=0}^{m-1} \int \left| \widehat{w}(x) \frac{\pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x)}{\sum_{t'} \pi_e(t'|x)\pi_b(t'|x)/\widehat{\pi}_b(t'|x)} dP_{X|\widetilde{A}=T}(x) - w(x)\pi_e(t|x) dP_{X|A=T}(x) \right| \\ +\ & \frac{1}{2} \sum_{t=0}^{m-1} \int \left| \widehat{w}(x) \frac{\pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x)}{\sum_{t'} \pi_e(t'|x)\pi_b(t'|x)/\widehat{\pi}_b(t'|x)} dP_{X|\widetilde{A}=T}(x) \left(1 - \frac{1}{\int \widehat{w}(x) dP_{X|A=T}(x)}\right) \right| \\ \overset{(1)}{=}\ & \frac{1}{2} \sum_{t=0}^{m-1} \int \left| \pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x) dP_X(x) - \pi_e(t|x) dP_X(x) \right| \\ +\ & \frac{1}{2} \sum_{t=0}^{m-1} \int \left| \pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x) dP_X(x) \left(1 - \frac{1}{\int \sum_{t=0}^{m-1} \pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x) dP_X(x)}\right) \right| \\ =\ & \frac{1}{2} \sum_{t=0}^{m-1} \int \pi_e(t|x)\pi_b(t|x) \left|1/\widehat{\pi}_b(t|x) - 1/\pi_b(t|x)\right| dP_X(x) + \frac{1}{2} \sum_{t=0}^{m-1} \int \pi_e(t|x)\pi_b(t|x)/\widehat{\pi}_b(t|x) dP_X(x) - \frac{1}{2} \\ \le\ & \sum_{t=0}^{m-1} \int \left|1/\widehat{\pi}_b(t|x) - 1/\pi_b(t|x)\right| dP_X(x) \overset{(2)}{\to} 0, \end{aligned}$$

where (1) follows from (6) and (7) and (2) follows from Assumption 1.

□

Proof of Theorem 2 (Exact coverage). The proof is very similar to that of Theorem 1 in Tibshirani et al. (2019) and is hence omitted. □

Assumption 3. *(consistency of quantile regression estimates). For n_1 large enough,*

$$\mathbb{P}\left[\mathbb{E}\left[(\widehat{q}_{\alpha_L}(X_{n+1}) - q_{\alpha_L}(X_{n+1}))^2 | \widehat{q}_{\alpha_L}, \widehat{q}_{\alpha_U}\right] \leq \eta_{n_1}\right] \geq 1 - \rho_{n_1},$$
$$\mathbb{P}\left[\mathbb{E}\left[(\widehat{q}_{\alpha_U}(X_{n+1}) - q_{\alpha_U}(X_{n+1}))^2 | \widehat{q}_{\alpha_L}, \widehat{q}_{\alpha_U}\right] \leq \eta_{n_1}\right] \geq 1 - \rho_{n_1},$$

for some sequences $\eta_{n_1} = o(1)$ and $\rho_{n_1} = o(1)$ as $n_1 \to \infty$.

Proof of Theorem 3 (Asymptotic efficiency). The proof is similar to that of Theorem 1 in Sesia and Candès (2020). First, we rewrite Assumption 3 as

$$\mathbb{P}\left[\mathbb{E}\left[(\widehat{q}_{\alpha_L}(X_{n+1}) - q_{\alpha_L}(X_{n+1}))^2 | \widehat{q}_{\alpha_L}, \widehat{q}_{\alpha_U}\right] \leq \eta_{n_1}/2\right] \geq 1 - \rho_{n_1}/2,$$
$$\mathbb{P}\left[\mathbb{E}\left[(\widehat{q}_{\alpha_U}(X_{n+1}) - q_{\alpha_U}(X_{n+1}))^2 | \widehat{q}_{\alpha_L}, \widehat{q}_{\alpha_U}\right] \leq \eta_{n_1}/2\right] \geq 1 - \rho_{n_1}/2,$$

for some sequences $\eta_{n_1} = o(1)$ and $\rho_{n_1} = o(1)$ as $n \to \infty$. Recall that our prediction band is defined as

$$\widehat{C}(X_{n+1}) = [\widehat{q}_{\alpha_L}(X_{n+1}) - Q_{1-\alpha}(X_{n+1}), \widehat{q}_{\alpha_U}(X_{n+1}) + Q_{1-\alpha}(X_{n+1})]$$

where $Q_{1-\alpha}(X_{n+1}) = (1-\alpha)$th quantile of $\sum_{i \in \mathcal{Z}^{cal,s}} p_i(X_{n+1})\delta_{V_i} + p_\infty(X_{n+1})\delta_\infty$. The oracle band is given by

$$C_\alpha^{oracle}(X_{n+1}) = [q_{\alpha_L}(X_{n+1}), q_{\alpha_U}(X_{n+1})].$$

It suffices to show

$$(i) \qquad |\widehat{q}_\beta(X_{n+1}) - q_\beta(X_{n+1})| = o_P(1) \text{ for } \beta = \alpha_L, \alpha_U. \tag{10}$$
$$(ii) \qquad |Q_{1-\alpha}(X_{n+1})| = o_P(1) \tag{11}$$

(i) Define random sets

$$B_{n,U} = \{x : |\widehat{q}_{\alpha_U}(x) - q_{\alpha_U}(x)| \geq \eta_{n_1}^{1/3}\}, B_{n,L} = \{x : |\widehat{q}_{\alpha_L}(x) - q_{\alpha_L}(x)| \geq \eta_{n_1}^{1/3}\}$$

and $B_n = B_{n,U} \cup B_{n,L}$. we have

$$\begin{aligned} & \mathbb{P}[X_{n+1} \in B_n | \widehat{q}_{\alpha_L}, \widehat{q}_{\alpha_U}] \\ \leq \; & \mathbb{P}[|\widehat{q}_{\alpha_U}(x) - q_{\alpha_U}(x)|^2 \geq \eta_{n_1}^{2/3} | \widehat{q}_{\alpha_U}] + \mathbb{P}[|\widehat{q}_{\alpha_L}(x) - q_{\alpha_L}(x)|^2 \geq \eta_{n_1}^{2/3} | \widehat{q}_{\alpha_L}] \\ \leq \; & \eta_{n_1}^{-2/3}\mathbb{E}[|\widehat{q}_{\alpha_U}(x) - q_{\alpha_U}(x)|^2] + \eta_{n_1}^{-2/3}\mathbb{E}[|\widehat{q}_{\alpha_L}(x) - q_{\alpha_L}(x)|^2] \leq \eta_{n_1}^{1/3} \end{aligned}$$

with probability at least $1 - \rho_{n_1}$ by Assumption 3. This implies (10).

(ii) Consider the following partition of the data in $\mathcal{Z}^{cal,s}$, where $n_2' = |\mathcal{Z}^{cal,s}|$:

$$\mathcal{Z}_a^{cal,s} = \{i \in \mathcal{Z}^{cal,s} : Z_i \in B_n^c\}, \; \mathcal{Z}_b^{cal,s} = \{i \in \mathcal{Z}^{cal,s} : Z_i \in B_n\}.$$

First, by Hoeffding's inequality,

$$\mathbb{P}\left[|\mathcal{Z}_b^{cal,s}| \geq n_2'\eta_{n_1}^{1/3} + t\right] \leq \mathbb{P}\left[\frac{1}{n_2'}\sum_{i \in \mathcal{Z}^{cal,s}} \mathbb{I}[Z_i \in B_n] \geq \mathbb{P}[Z_i \in B_n] + \tfrac{t}{n_2'}\right] \leq \exp\left(-\frac{2t^2}{n_2'}\right).$$

Set $t = c\sqrt{n_2' \log n_2'}$, we obtain that $|\mathcal{Z}_b^{cal,s}| = o_P(n_2') = o_P(n)$.

Next, define $\widetilde{S}_i = \max\{q_{\alpha_L}(X_i) - Y_i, Y_i - q_{\alpha_U}(X_i)\}$ for any $i \in \mathcal{Z}^{cal,s}$. By definition, for all $i \in \mathcal{Z}_a^{cal,s}$,

$$\widetilde{S}_i - \eta_{n_1}^{1/3} \leq S_i \leq \widetilde{S}_i + \eta_{n_1}^{1/3}. \tag{12}$$

Let F_n and $\widetilde{F}_n$ denote the empirical distribution $\sum_{i \in \mathcal{Z}^{cal,s}} p_i(X_{n+1})\delta_{S_i} + p_\infty(X_{n+1})\delta_\infty$ and $\sum_{i \in \mathcal{Z}^{cal,s}} p_i(X_{n+1})\delta_{\widetilde{S}_i} + p_\infty(X_{n+1})\delta_\infty$ respectively. Define $F_{n,a}$ and $\widetilde{F}_{n,a}$ as versions of F_n and $\widetilde{F}_n$ when restricting attentions to observations that belong to $\mathcal{Z}_a^{cal,s}$ only. For sufficiently large n, we can show $|\mathcal{Z}_b^{cal,s}|/|\mathcal{Z}_a^{cal,s}| \leq \alpha$ by noting that $|\mathcal{Z}_b^{cal,s}| = o_P(n)$. We next show that

$$F_{n,a}^{-1}\left(1 - \frac{n_2'\alpha}{|\mathcal{Z}_a^{cal,s}|}\right) \leq F_n^{-1}(1-\alpha) \leq F_{n,a}^{-1}\left(1 - \frac{n_2'\alpha - |\mathcal{Z}_b^{cal,s}|}{n_2'|\mathcal{Z}_a^{cal,s}|}\right).$$

To prove the first inequality, notice that for those observations that belong to $\mathcal{Z}_b^{cal,s}$, if their scores are in the lower $1-\alpha$ quantile of F_n, $F_{n,a}^{-1}(1-n_2'\alpha/|\mathcal{Z}_a^{cal,s}|) = F_n^{-1}(1-\alpha)$. However, in general, the quantiles of $F_{n,a}$ will be smaller. The second inequality can be proven in a similar manner. Combining this together with (12) yields that

$$\widetilde{F}_{n,a}^{-1}\left(1-\frac{n_2'\alpha}{|\mathcal{Z}_a^{cal,s}|}\right) - \eta_{n_1}^{1/3} \le F_n^{-1}(1-\alpha) \le \widetilde{F}_{n,a}^{-1}\left(1-\frac{n_2'\alpha - |\mathcal{Z}_b^{cal,s}|}{n_2'|\mathcal{Z}_a^{cal,s}|}\right) + \eta_{n_1}^{1/3}.$$

It in turn yields that $|\widetilde{F}_{n,a}^{-1}(1-\alpha) - \widetilde{F}_n^{-1}(1-\alpha)| = o_P(1)$ and hence, $|\widetilde{F}_n^{-1}(1-\alpha) - F_n^{-1}(1-\alpha)| = o_P(1)$. By definition, $\widetilde{F}_n^{-1}(1-\alpha) = o_P(1)$. It follows that $Q_{1-\alpha}(X_{n+1}) := F_n^{-1}(1-\alpha) = o_P(1)$. This yields (11). □

Proof of Theorem 4 (COPP-IS). We index the data points in the calibration dataset $\mathcal{Z}^{cal}$ by $\{1, 2, \ldots, n_2\}$. Recall that $Z_i = (X_i, T_i, Y_i)$ for any $i \in \mathcal{Z}^{cal}$. Let $\widetilde{Z}_{n+1} = (\widetilde{X}_{n+1}, \widetilde{E}_{n+1}, \widetilde{Y}_{n+1})$ where $(\widetilde{X}_{n+1}, \widetilde{Y}_{n+1}) \sim \widetilde{P}_X \times P_{Y|\widetilde{A}=T,X}$ and $\widetilde{E}_{n+1}$ is the latent treatment variable. Let $E_{\widetilde{z}}$ denote the event that $\{Z_1, \ldots, Z_n, \widetilde{Z}_{n+1}\} = \{z_1, \ldots, z_n, \widetilde{z}_{n+1}\}$. Notice that each Z_i involves the binary treatment variable. The corresponding nonconformity scores are denoted by $\widetilde{S} = (S_1, \ldots, S_n, \widetilde{S}_{n+1})$ and $s_i = \mathcal{S}((x_i, y_i), \mathcal{Z}^{tr})$ for $1 \le i \le n_2$, $\widetilde{s}_{n+1} = \mathcal{S}((\widetilde{x}_{n+1}, \widetilde{y}_{n+1}), \mathcal{Z}^{tr})$. Notice that these scores are independent of the binary variable.

Similar to the proof of Theorem 1, for each $1 \le i \le n_2$,

$$\begin{aligned}
& \mathbb{P}\{\widetilde{S}_{n+1} = v_i | E_{\widetilde{z}}, A_1, \ldots, A_{n_2}\} \\
= \; & \mathbb{P}\{(\widetilde{X}_{n+1}, \widetilde{Y}_{n+1}) = (x_i, y_i) | E_{\widetilde{z}}, A_1, \ldots, A_{n_2}\} \\
= \; & \frac{\widehat{w}(z_i) I(A_i = t_i)}{\sum_{i \in \mathcal{Z}^{cal}} \widehat{w}(z_i) I(A_i = t_i) + \widehat{w}(\widetilde{z}_{n+1})} = \widehat{p}_i(\widetilde{x}_{n+1} | A_1, \ldots, A_{n_2}).
\end{aligned}$$

The quantile $Q_{1-\alpha}(\widetilde{x}_{n+1}|A_1, \ldots, A_{n_2})$ used to construct the interval is defined as

$$(1-\alpha)\text{th quantile of } \sum_{i \in \mathcal{Z}^{cal}} \widehat{p}(\widetilde{x}_{n+1}|A_1, \ldots, A_{n_2}) \delta_{V_i} + \widehat{p}_{n+1}(\widetilde{x}_{n+1}|A_1, \ldots, A_{n_2}) \delta_\infty.$$

In COPP-IS,

$$\mathbb{P}\{\widetilde{S}_{n+1} = v_i | E_{\widetilde{z}}\} = \frac{\widehat{w}(z_i) \widehat{\pi}_A(t_i|x_i)}{\sum_{i \in \mathcal{Z}^{cal}} \widehat{w}(z_i) \widehat{\pi}_A(t_i|x_i) + \widehat{w}(\widetilde{z}_{n+1})} = \widehat{p}_i(\widetilde{x}_{n+1}).$$

The quantile $Q_{1-\alpha}(\widetilde{x}_{n+1})$ used to construct the interval is defined as

$$(1-\alpha)\text{th quantile of } \sum_{i \in \mathcal{Z}^{cal}} \widehat{p}(\widetilde{x}_{n+1}) \delta_{V_i} + \widehat{p}_{n+1}(\widetilde{x}_{n+1}) \delta_\infty.$$

By law of large numbers, $\lim_{n_2 \to \infty} \{Q_{1-\alpha}(\widetilde{x}_{n+1}|A_1, \ldots, A_{n_2}) - Q_{1-\alpha}(\widetilde{x}_{n+1})\} = 0$. Other results can be similarly proven.

□

Proof of Theorem 5 (COPP-MS). By Markov inequality, we have

$$\begin{aligned}
& \mathbb{P}_{(X, Y^{\pi^e}) \sim P_X \times P_{Y^{\pi^e}|X}}(Y^{\pi^e} \notin \widehat{C}_{B,\gamma}(X)) \\
= \; & \mathbb{P}_{(X, Y^{\pi^e}) \sim P_X \times P_{Y^{\pi^e}|X}}(\sum_{b=1}^B \mathbb{I}(Y^{\pi^e} \notin \widehat{C}^b_{n_1,n_2}(X)) \ge \gamma B) \\
= \; & \mathbb{E}[\mathbb{I}(\sum_{b=1}^B \mathbb{I}(Y^{\pi^e} \notin \widehat{C}^b_{n_1,n_2}(X)) \ge \gamma B)] \\
\le \; & \frac{1}{\gamma B} \mathbb{E}[\sum_{b=1}^B \mathbb{I}(Y^{\pi^e} \notin \widehat{C}^b_{n_1,n_2}(X))] \\
= \; & \frac{1}{\gamma} \mathbb{P}_{(X, Y^{\pi^e}) \sim P_X \times P_{Y^{\pi^e}|X}}(Y^{\pi^e} \notin \widehat{C}^b_{n_1,n_2}(X)).
\end{aligned}$$

According to Theorem 1, we have

$$\lim_{n_1,n_1'\to\infty}\mathbb{P}_{(X,Y^{\pi^e})\sim P_X\times P_{Y^{\pi^e}|X}}(Y^{\pi^e}\notin \widehat{C}^b_{n_1,n_2}(X))\leq \alpha\gamma.$$

Combining the two results, we obtain that

$$\begin{aligned}
&\lim_{n_1,n_1'\to\infty}\mathbb{P}_{(X,Y^{\pi^e})\sim P_X\times P_{Y^{\pi^e}|X}}(Y^{\pi^e}\notin \widehat{C}_{B,\gamma}(X))\\
\leq\quad &\lim_{n_1,n_1'\to\infty}\frac{1}{\gamma}\mathbb{P}_{(X,Y^{\pi^e})\sim P_X\times P_{Y^{\pi^e}|X}}(Y^{\pi^e}\notin \widehat{C}^b_{n_1,n_2}(X))\leq \alpha.
\end{aligned}$$

This is equivalent to

$$\lim_{n_1,n_1'\to\infty}\mathbb{P}_{(X,Y^{\pi^e})\sim P_X\times P_{Y^{\pi^e}|X}}(Y^{\pi^e}\in \widehat{C}^b_{n_1,n_2}(X))\geq 1-\alpha.$$

The proof is hence completed. □

B SEQUENTIAL DECISION MAKING

This section provides assumptions, pseudo codes and theories in sequential decision making. Finally, we conclude this section with a discussion to extend our proposal to settings with immediate rewards at each decision point. We begin with the consistency, sequential ignorability and positive assumption in sequential decision making. Let $Y_i(t_1, t_2, \ldots, t_K)$ denote the reward that the ith instance would be observed were they to receive action $t_1, t_2, \ldots, t_K$ sequentially. The standard assumptions are (1) $Y_i(T_{1i}, T_{2i}, \ldots, T_{Ki}) = Y_i$ almost surely for any i (i.e., consistenct); (2) A policy π_b satisfies sequential ignorability, that is at any stage k, conditional on the history H_k generated by the policy, the action T_k generated by the policy is independent of the potential outcomes $\{X_{k+1}(t_1, \ldots, t_k), X_{k+2}(t_1, \ldots, t_{k+1}), \ldots X_K(t_1 \ldots, t_{K-1}), Y(t_1, t_2, \ldots, t_K)\}$ for all $t_k \in \{0, 1, \ldots, m-1\}$. (3) $\pi_{b_k}(t_k|h_k)$ is uniformly bounded away from zero for any t_k, h_k (i.e., positivity).

Denote $\mathbf{A} = (A_1, \ldots, A_K)^\top$ as the actions generated by the pseudo policy with $\pi_{a_k}(t_k|h_k) \propto \pi_{e_k}(t_k|h_k)/\pi_{b_k}(t_k|h_k)$ and $\mathbf{T} = (T_1, \ldots, T_K)^\top$ as the treatment generated by the behavior policy, and $\widetilde{\mathbf{A}}$ as the one generated by the estimated pseudo policy $\pi_{\widetilde{a}_k}(t_k|h_k) \propto \pi_{e_k}(t_k|h_k)/\widehat{\pi}_{b_k}(t_k|h_k)$. We summarize the pseudocode of our proposal COPP in Algorithm 2.

Let $n_1 = |\mathcal{Z}^{tr}|$. The following assumption provides the consistency of behavior policy estimates.

Assumption 4. *For the output in Step 2 of Algorithm 2 and any* $1 \leq k \leq K$, $\widehat{\pi}_{b_k}(t_k|h_k)$*s are uniformly bounded away from zero and*

$$\lim_{n_1\to\infty}\mathbb{E}\big|\widehat{\pi}_{b_k}(t_k|H_k) - \pi_{b_k}(t_k|H_k)\big| = 0.$$

Assumption 5. *Suppose that the output* $\widehat{\mathbb{P}}(\widetilde{\mathbf{A}} = \mathbf{T}|X_1, \mathcal{Z}^{tr})$ *in Step 4 of Algorithm 2 is uniformly bounded away from zero and*

$$\lim_{n_1\to\infty}\mathbb{E}_{X_1\sim P_{X_1}}\big|\widehat{\mathbb{P}}(\widetilde{\mathbf{A}} = \mathbf{T}|X_1, \mathcal{Z}^{tr}) - \mathbb{P}(\widetilde{\mathbf{A}} = \mathbf{T}|X_1, \mathcal{Z}^{tr})\big| = 0.$$

Let $P_{X_1|\widetilde{\mathbf{A}}=\mathbf{T}}$ be the probability measure of the initial state for selected samples in $\mathcal{Z}^{cal,s}$ as an estimate of $P_{X_1|\mathbf{A}=\mathbf{T}}$. Denote $w(X_1) = 1/\mathbb{P}(\mathbf{A} = \mathbf{T}|X_1)$ and $\widehat{w}(X_1) = 1/\widehat{\mathbb{P}}(\widetilde{\mathbf{A}} = \mathbf{T}|X_1, \mathcal{Z}^{tr})$ as the output in Step 4 of Algorithm 2. We summarize the theoretical results below.

Theorem 6 (Asymptotic coverage for SDM). *Let* $n_1' = |\mathcal{Z}^{tr,s}|$. *Suppose that Assumptions 4-5 hold and* $\mathbb{E}_{X_1\sim P_{X_1}}[w(X_1)] < \infty$, $\mathbb{E}_{X_1\sim P_{X_1}}[\widehat{w}(X_1)|\mathcal{Z}^{tr}] < \infty$, $\mathbb{E}_{X_1\sim P_{X_1|\widetilde{\mathbf{A}}=\mathbf{T}}}[\widehat{w}(X_1)|\mathcal{Z}^{tr}] < \infty$, *then the output* $\widehat{C}(x)$ *from Algorithm 2 satisfies*

$$\lim_{n_1,n_1'\to\infty}\mathbb{P}_{(X_1,Y^{\pi^e})\sim P_{X_1}\times P_{Y^{\pi^e}|X_1}}(Y^{\pi^e}\in \widehat{C}(X_1))\geq 1-\alpha, \tag{13}$$

Discussion. We conclude this section by extending our proposal to settings with immediate rewards at each decision point. Suppose the observed data can be summarized as $\{(X_{1i}, T_{1i}, Y_{1i}, X_{2i}, T_{2i}, Y_{2i}, \ldots, X_{Ki}, T_{Ki}, Y_{Ki})\}_{i=1}^n$ where Y_{ki} is the immediate reward at the kth stage. Similarly, we can show that the conditional distribution $Y_k|A_1 = T_1, \ldots, A_k = T_k, H_1$ is the same as that of $Y_k^{\pi^e}|H_1$ for each $1 \leq k \leq K$. As such, for each k, we can apply our proposal to construct a PI for $Y_k^{\pi^e}$. These PIs can be potentially further aggregated to cover the sum $\sum_k Y_k^{\pi^e}$. We leave it for future research.

Algorithm 2 COPP: Conformal off-policy prediction in multi-stage decision making

```
Input: Data $\{(X_{1i}, T_{1i}, \ldots, X_{Ki}, T_{Ki}, Y_i)\}_{i=1}^{n}$; a test point with initial state $X_{1,n+1}$; a sequence of target policies $\pi_e = \{\pi_{e_k}(t_k|h_k)\}_{k=1}^{K}$; propensity score training algorithm $\mathcal{P}$; quantile prediction algorithm $\mathcal{Q}$; conformity score $\mathcal{S}$; and coverage level $1-\alpha$ with $\alpha_U - \alpha_L = 1-\alpha$.
  Split the data into two disjoint subsets $\mathcal{Z}^{tr}$ and $\mathcal{Z}^{cal}$.
  Train $\{\widehat{\pi}_{b_k}(t_k|h_k)\}_{k=1}^{K}$ using $\mathcal{P}$ on all samples from $\mathcal{Z}^{tr}$, i.e.,
      $\widehat{\pi}_{b_k}(t_k|h_k) \leftarrow \mathcal{P}(\{(H_{ki}, T_{ki})\}_{i \in \mathcal{Z}^{tr}}), H_{ki} = \{(X_{1i}, T_{1i}, \ldots, X_{ki})\},$ for $1 \leq k \leq K$.
  Draw $\widetilde{\mathbf{A}}_i = (A_{1i}, \ldots, A_{Ki})$ for $i = 1, \ldots, n$ with plugging $\widehat{\pi}_{b_k}(t_k|h_k)$.
  Train $\widehat{w}(X_1)$ using $\mathcal{P}$ on all samples from the $\mathcal{Z}^{tr}$ augmented by $\{\widetilde{\mathbf{A}}_i\}_{i \in \mathcal{Z}^{tr}}$, i.e.,
      $\widehat{e}(X_1) = \widehat{\mathbb{P}}(\widetilde{\mathbf{A}} = \mathbf{T}|X_1, \mathcal{Z}^{tr}) \leftarrow \mathcal{P}(\{(\widetilde{\mathbf{A}}_i, \mathbf{T}_i, X_{1i})\}_{i \in \mathcal{Z}^{tr}}), \widehat{w}(X_1) = 1/\widehat{e}(X_1).$
  Select subsamples satisfying $\widetilde{\mathbf{A}}_i = \mathbf{T}_i$ in both subsets denoted as $\mathcal{Z}^{tr,s}$ and $\mathcal{Z}^{cal,s}$.
  Train quantile regressions using $\mathcal{Q}$ on selected subsamples from $\mathcal{Z}^{tr,s}$, i.e.,
      $\widehat{q}_{\alpha_L}(x; \mathcal{Z}^{tr,s}) \leftarrow \mathcal{Q}(\alpha_L, \{(X_{1i}, Y_i)\}_{i \in \mathcal{Z}^{tr,s}}), \widehat{q}_{\alpha_U}(x; \mathcal{Z}^{tr,s}) \leftarrow \mathcal{Q}(\alpha_U, \{(X_{1i}, Y_i)\}_{i \in \mathcal{Z}^{tr,s}}).$
  Compute the nonconformity scores for all selected subsamples $i \in \mathcal{Z}^{cal,s}$:
      $S_i = \max\{\widehat{q}_{\alpha_L}(X_{1i}; \mathcal{Z}^{tr,s}) - Y_i, Y_i - \widehat{q}_{\alpha_U}(X_{1i}; \mathcal{Z}^{tr,s})\}.$
  Compute the normalized weights for $i \in \mathcal{Z}^{cal,s}$ and the test point $X_{n+1}$
      $\widehat{p}_i(X_{1,n+1}) = \frac{\widehat{w}(X_{1i})}{\sum_{i \in \mathcal{Z}^{cal,s}} \widehat{w}(X_{1i}) + \widehat{w}(X_{1,n+1})}, \ \widehat{p}_\infty(X_{1,n+1}) = \frac{\widehat{w}(X_{1,n+1})}{\sum_{i \in \mathcal{Z}^{cal,s}} \widehat{w}(X_{1i}) + \widehat{w}(X_{1,n+1})}.$
  Compute the $(1-\alpha)$th quantile of $\sum_{i \in \mathcal{Z}^{cal,s}} \widehat{p}_i(X_{1,n+1})\delta_{S_i} + \widehat{p}_\infty(X_{1,n+1})\delta_\infty$ as $Q_{1-\alpha}(X_{1,n+1})$.
  Construct a prediction set for $X_{1,n+1}$:
      $\widehat{C}(X_{1,n+1}) = [\widehat{q}_{\alpha_L}(X_{1,n+1}; \mathcal{Z}^{tr,s}) - Q_{1-\alpha}(X_{1,n+1}), \widehat{q}_{\alpha_U}(X_{1,n+1}; \mathcal{Z}^{tr,s}) + Q_{1-\alpha}(X_{1,n+1})].$
Output: A prediction set $\widehat{C}(X_{1,n+1})$ for the outcome $Y_{n+1}^{\pi^e}$.
```

C ADDITIONAL IMPLEMENTATION DETAILS

This section provides additional implementation details for the competing methods in the simulation study. First, notice that the importance sampling (IS) method can be naturally coupled with the kernel method to evaluate the individual treatment effect (ITE). Specifically, consider the following estimator,

$$\widehat{\mathbb{E}}[Y_{n+1}^{\pi_e}|X_{n+1}] = \sum_{i=1}^{n} \frac{\pi_e(T_i|X_i)}{\widehat{\pi}_b(T_i|X_i)} Y_i \frac{K((X_i - X_{n+1})/h)}{\sum_{i=1}^{n} K((X_i - X_{n+1})/h)},$$

for certain kernel function K with bandwidth h, and the logistic regression estimator estimator $\widehat{\pi}_b$. Its standard deviation can be estimated based on the sampling variance estimator and the corresponding confidence interval (CI) can be derived. In our experiments, the bandwidth parameter h is manually selected so that the resulting CI achieves the best empirical coverage rate.

Second, the double robust (DR) estimator can be coupled with kernel method for ITE evaluation as well. Specifically, define

$$\widehat{\mathbb{E}}[Y_{n+1}^{\pi_e}|X_{n+1}] = \sum_{i=1}^{n} \left\{ \frac{\pi_e(T_i|X_i)}{\widehat{\pi}_b(T_i|X_i)} [Y_i - \widehat{\mu}(X_i)] + \widehat{\mu}(X_i) \right\} \frac{K((X_i - X_{n+1})/h)}{\sum_{i=1}^{n} K((X_i - X_{n+1})/h)},$$

where $\widehat{\mu}(x)$ denotes the estimated regression function obtained via random forest. The corresponding confidence interval can be similarly constructed.