PAC Off-Policy Prediction in Contextual Bandits

Yilong Wan, Yuqiang Li, Xianyi Wu

February 17, 2025

Abstract

This paper concentrates on the off-policy evaluation task in contextual bandits, aiming to rigorously quantify the performance of a target policy using data collected under a potentially different and unknown behavior policy. Recent research has shifted focus from estimating expectations to constructing reliable prediction intervals for the reward under a target policy. Based on conformal prediction, these methods control marginal coverage with finite-sample theoretical guarantees, making them particularly suited for safety-critical applications. In this paper, we further investigate how to achieve coverage conditional on a pre-collected offline dataset, introducing a novel algorithm that constructs probably approximately correct prediction intervals. Our approach primarily relies on rejection sampling and split conformal prediction. Theoretical results on the finite-sample properties and asymptotic behavior of our method are established, and simulation experiments are conducted to validate its effectiveness.

1 Introduction

In many fields such as healthcare, marketing, and content recommendation, understanding the potential impact of a decision-making policy prior to deployment is essential. Directly testing a new policy in the real world, however, is often impractical due to ethical considerations, resource constraints, or associated risks. Therefore, we may seek to evaluate the target policy's performance using offline data previously collected under a different behavior policy. This process is known as off-policy evaluation (OPE)

Problems in the aforementioned fields are commonly modeled within the contextual bandits framework. At each time step, the agent observes a context, selects an action according to a given policy, and then receives a random reward from the environment that depends on the context-action pair. While most OPE methods for contextual bandits have traditionally focused on the expectation of reward under the target policy, they may not be suitable in safety-critical settings due to their inability to capture the variability of the reward. Consequently, recent literature has turned to considering alternative measures of performance, including variance, quantiles, and conditional values at risk, among others; see e.g., [Keramati et al., 2020]; [Chandak et al., 2021]; [Huang et al., 2021].

A promising approach for uncertainty quantification is through prediction intervals (PIs). Unlike confidence intervals, which give a range for a population parameter, like a mean, PIs directly cover the reward itself with a specified confidence level. In [Taufiq et al., 2022], an algorithm for constructing finite-sample valid PIs was first proposed within the contextual bandits framework, incorporating stochastic policies and continuous action spaces. Notably, their method accounts for individual effects rather than evaluating the average impact of the target policy across all contexts, meaning that the constructed PIs are adaptive to test contexts, which is of significant interest in fields such as precision medicine [Lei and Candès, 2021]. However, their approach requires estimating the probability densities of rewards conditional on context-action pairs, which can be challenging when the model is unknown. To address this limitation, [Zhang et al., 2023] introduced a sub-sampling-based method and extended the framework to both contextual bandits and sequential decision-making scenarios. Nevertheless, their approach is restricted to discrete action spaces, motivating us to explore methods applicable to more general action spaces.

In both [Taufiq et al., 2022] and [Zhang et al., 2023], the authors employ conformal prediction (CP) [Vovk et al., 2005, Shafer and Vovk, 2008, Balasubramanian et al., 2014], a well-established and effective method for uncertainty quantification. CP constructs reliable PIs with finite-sample theoretical guarantees, relying solely on the exchangeability of the calibration dataset and the test point, irrespective of the underlying data distribution. However, a key limitation of CP is that its validity is inherently unconditional (or marginal), meaning the nominal coverage it provides is valid only under the randomness of both the calibration and test data. This marginal validity can be problematic in the OPE setting, where the calibration data is pre-collected and fixed. The PIs constructed by CP in this scenario are no longer guaranteed to achieve the required coverage and may result in undercoverage without proper control. Therefore, we aim to establish a stronger form of validity, referred to as training conditional validity in [Vovk, 2013], which ensures that the PIs attain the desired coverage, conditional on the given dataset.

In this paper, we propose PAC Off-Policy Prediction (PACOPP), a novel algorithm that applies a modified split CP method to construct PIs for the rewards under target policy in contextual bandits using an offline observational dataset. PACOPP enjoys both finite-sample theoretical guarantees and adaptivity with respect to the test context, without relying on any distributional or space assumptions. Our approach, to the best of our knowledge, is the first method with these properties that achieves probably approximately correct (PAC) validity, a type of training conditional validity, in the OPE task, guaranteeing that the constructed PIs attain the required coverage with a specified high probability.

In summary, our contributions are as follows: (i) Methodologically, we develop a novel procedure to construct off-policy PAC prediction intervals for a target policy's reward at any test context in bandits. our method achieves stronger validity and is more general than previous works, as it does not require model estimation and can be applied to continuous actions. (ii) Theoretically, we provide finite-sample theoretical guarantees for the validity of PACOPP and prove that it is asymptotically efficient when the estimators are consistent. Additionally, we extend the theoretical results of split CP with PAC validity, offering a two-sided bound in a PAC sense for the first time.

2 Preliminaries

We begin with some necessary preliminaries, including the formulation of our research problem, as well as the definition of PAC prediction intervals. For convenience, we define $\Delta(\mathcal{X})$ as the set of all probability distributions over the space $\mathcal{X}$, and $[n]$ as the index set $\{1, 2, \ldots, n\}$ for an integer $n > 0$ throughout this paper.

2.1 Problem formulation

We denote contexts, actions and rewards by S, A, and R, respectively, with their corresponding spaces given by $\mathcal{S}, \mathcal{A}$, and $\mathcal{R} \subset \mathbb{R}$. These spaces can be either discrete or continuous. In the OPE setting, we assume access to an observational dataset $\mathcal{D} = \{S_i, A_i, R_i\}_{i=1}^n$, collected over n rounds through interactions between a behavior policy π_b and the environment. In each round $i \in [n]$, a context $S_i = s_i$ is independently drawn from the context distribution $P_S \in \Delta(\mathcal{S})$. An action $A_i = a_i$ is then selected according to the behavior policy $\pi_b(\,\cdot\,|s_i)$, where a policy is defined as a mapping from $\mathcal{S}$ to $\Delta(\mathcal{A})$. The environment subsequently reveals a reward $R_i \sim P_R(\,\cdot\,|s_i, a_i)$, with $P_R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{R})$ mapping context-action pairs to distributions over the reward space $\mathcal{R}$. Given a target policy π_e that may differ from π_b, the goal of OPE is to predict the potential reward that would be obtained from the environment if π_e were implemented instead of π_b. Specifically, OPE aims to quantify the target reward R_{n+1}, which, together with a test context S_{n+1}, follows the target distribution given by

$$P^{\pi_e}(\mathrm{d}s, \mathrm{d}r) \coloneqq P_S(\mathrm{d}s) \int_{\mathcal{A}} P_R(\mathrm{d}r|s, a)\pi_e(\mathrm{d}a|s). \tag{1}$$

Using the dataset $\mathcal{D}$, methods based on CP construct distribution-free PIs $C_{\mathcal{D}}(S_{n+1})$ such that

$$\mathbb{P}\left[R_{n+1} \in C_{\mathcal{D}}(S_{n+1})\right] \geq 1 - \epsilon, \tag{2}$$

where ϵ is a pre-specified failure probability. These methods are often preferred over traditional approaches that estimate $\mathbb{E}[R_{n+1}]$, as the PIs better capture the variability in target rewards while provide finite-sample guarantees. However, these intervals are marginal valid because all variables in (2) are treated as random, including the test point (S_{n+1}, R_{n+1}) and the dataset $\mathcal{D}$. As a result, $C_{\mathcal{D}}(S_{n+1})$ does not guarantee $1-\epsilon$ coverage conditional on a specific S_{n+1} (object conditional validity) or on a fixed observational dataset $\mathcal{D}$ (training conditional validity).

In fact, it is inherently impossible to design non-trivial algorithms that output PIs with object conditional validity without making modeling assumptions ([Foygel Barber et al., 2021]), although training conditional validity is achievable. In this paper, we focus on achieving training conditional validity and aim to devise an algorithm that outputs PIs that are probably approximately correct, as defined by (3). To this end, we impose the following standard assumption on the weight function $w(s, a)$, which is defined as $w(s, a) \coloneqq \frac{\pi_e}{\pi_b}(\mathrm{d}a|s)$:

Assumption 1. *The weight function is uniformly upper bounded, i.e.,* $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} w(s, a) = b < \infty$. *Additionally, we set* $w(s, a) = 0$ *if* $\pi_b(\mathrm{d}a|s) = \pi_e(\mathrm{d}a|s) = 0$.

2.2 PAC Prediction Intervals

We restate the definition of PAC prediction intervals (sets) from [Park et al., 2020] as follows. Let $X \in \mathcal{X}$ be inputs, $Y \in \mathcal{Y}$ be outcomes and P be the joint distribution of (X, Y). A set-value function C, which takes an input $x \in \mathcal{X}$ and outputs a subset of $\mathcal{Y}$, is called *approximately correct* if the probability that $C(X)$ fails to contain Y is bounded by a given $\epsilon \in (0, 1)$, i.e.,

$$L_P(C) \coloneqq \mathbb{P}_{(X,Y)\sim P}\left[Y \notin C(X)\right] \leq \epsilon.$$

Given a dataset $\mathcal{D}$ treated as random, a set-valued function $C_{\mathcal{D}}$ constructed from $\mathcal{D}$ is said to be *probably approximately correct* if it is approximately correct with high probability. That is, for a given $\delta \in (0, 1)$,

$$\mathbb{P}\left[L_P(C_{\mathcal{D}}) \leq \epsilon\right] \geq 1 - \delta.$$

Finally, we refer to $C_{\mathcal{D}}(X)$ as a (ϵ, δ)-PAC prediction interval.

To our setting. For the OPE problem formulated in the previous subsection and any pre-specified $(\epsilon, \delta) \in (0, 1)^2$, the desired (ϵ, δ)-PAC prediction interval for the target reward R_{n+1} is the value of a set-valued function $\hat{C}$ at the test context S_{n+1}, which satisfies

$$\mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}) \leq \epsilon\right] \geq 1 - \delta, \tag{3}$$

where the coverage error is defined as

$$L_{P^{\pi_e}}(\hat{C}) = \mathbb{P}_{(S_{n+1},R_{n+1})\sim P^{\pi_e}}\left[R_{n+1} \notin \hat{C}(S_{n+1})\right],$$

and P^{π_e} is the target distribution defined in (1).

3 PAC Off-Policy Prediction in Contextual Bandits

The mismatch between π_e and π_b induces a discrepancy between the target distribution P^{π_e} and the joint context-reward distributions of the observational data, commonly referred to as a distribution shift. This shift violates the assumptions underlying standard CP-based method, which assume exchangeable or i.i.d. distributions between the test and observational data. While the "weighted CP" approach [Tibshirani et al., 2019] extend CP by relaxing the assumption to weighted exchangeability or independence, its application in OPE requires estimating the weights

$$\frac{P^{\pi_e}}{P^{\pi_b}}(\mathrm{d}s, \mathrm{d}r) = \frac{\int_{\mathcal{A}} P_R(\mathrm{d}r|s, a)\pi_e(\mathrm{d}a|s)}{\int_{\mathcal{A}} P_R(\mathrm{d}r|s, a)\pi_b(\mathrm{d}a|s)} \tag{4}$$

for all (s, r) pairs in the observational and test datasets, which can be challenging as the model of reward distribution P_R is unknown.

Another natural approach to address the distribution shift is to construct a subset from the observational dataset whose distribution resembles the target distribution. In this work, we adopt a rejection sampling strategy [Neumann, 1951, Owen, 2013, Park et al., 2022] to generate such sub-samples.

Rejection sampling. Rejection sampling (RS), also known as the acceptance-rejection method, is a fundamental technique for generating samples from a target distribution using samples from a proposal distribution, along with auxiliary variables and a scaling constant. Supposing the Assumption 1 holds and the behavior policy π_b is also known, this procedure can generate samples from the target distribution using $\mathcal{D}$. Specifically, for each (S_i, A_i, R_i) in the dataset $\mathcal{D}$, we independently sample a auxiliary uniform random variable $V_i \sim U([0,1])$ and retain (S_i, R_i) if $V_i \leq \frac{1}{b} w(S_i, A_i)$, where w is the weight function and b is the uniform upper bound. This process constructs a set $\mathcal{D}^{\text{rs}}$ with a random size $N_{\text{rs}} \leq n$, denoted by

$$\mathcal{D}^{\text{rs}} \coloneqq \left\{ (S_i, R_i) \mid (S_i, A_i, R_i) \in \mathcal{D} \text{ and } V_i \leq \frac{1}{b} w(S_i, A_i) \right\}. \tag{5}$$

We rigorously establish in Proposition 1 (with the proof provided in Appendix A.1) that each sample in $\mathcal{D}^{\text{rs}}$ independently follows the target distribution P^{π_e}.

Proposition 1. *Let $i_1, i_2, \ldots, i_{N_{\text{rs}}}$ be elements of the set $\{i \mid (S_i, R_i) \in \mathcal{D}^{\text{rs}}\}$, arranged in ascending order. Define $Z_j = (S_{i_j}, R_{i_j})$ for $j \in [N_{\text{rs}}]$. Under the randomness of $\mathcal{D}$ and $V \coloneqq (V_1, \ldots, V_n)$, conditional on the event $N_{\text{rs}} = n_{\text{rs}} \in [n]$, we have $Z_j \overset{i.i.d.}{\sim} P^{\pi_e}$ for $j \in [n_{\text{rs}}]$, denoted by $\mathcal{D}^{\text{rs}} \sim (P^{\pi_e})^{n_{\text{rs}}}$.*

The RS procedure circumvents the estimation of the conditional reward distribution P_R by introducing exogenous randomness, a strategy also employed in [Zhang et al., 2023]. However, unlike Zhang's approach, which is restricted to discrete action spaces, the RS method extends to continuous action spaces as well. Building on Proposition 1, we propose an algorithm, inspired by [Vovk, 2013] and [Park et al., 2020], to construct the PAC prediction interval as specified in (3).

Algorithm. After the RS procedure, we adhere to the structure of split CP approach [Romano et al., 2019] for constructing PIs. At first, the dataset $\mathcal{D}^{\text{rs}}$ is partitioned into two disjoint subsets: the training set $\mathcal{D}_1^{\text{rs}}$ of size $L = N_{\text{rs}} - M$ and the calibration set $\mathcal{D}_2^{\text{rs}}$ of size $M = \lfloor \gamma N_{\text{rs}} \rfloor$, where $\gamma \in (0,1)$ is a fixed proportion. Subsequently, the ϵ_{lo} (lower) and ϵ_{up} (upper) conditional quantile functions of R given S, denoted by $\hat{q}_{\epsilon_{\text{lo}}}$ and $\hat{q}_{\epsilon_{\text{up}}}$, are estimated using the training set $\mathcal{D}_1^{\text{rs}}$, with $\epsilon_{\text{up}} - \epsilon_{\text{lo}} = 1 - \epsilon$. A variety of algorithms, including linear regression [Koenker and Bassett Jr, 1978], neural networks [Taylor, 2000], and random forests [Meinshausen and Ridgeway, 2006], can be employed for this purpose.[1] For a given context s, these estimated quantile functions enable the parameterization of PIs in the following widely-used form[2]:

$$\hat{C}_\tau(s) \coloneqq [\hat{q}_{\epsilon_{\text{lo}}}(s) - \tau, \hat{q}_{\epsilon_{\text{up}}}(s) + \tau] \cap \mathcal{R}, \tag{6}$$

where the scalar τ is used to calibrate the black-box PI $[\hat{q}_{\epsilon_{\text{lo}}}(s), \hat{q}_{\epsilon_{\text{up}}}(s)]$.

After training $\hat{q}_{\epsilon_{\text{lo}}}$ and $\hat{q}_{\epsilon_{\text{up}}}$, for each observation (S_i, R_i) in the calibration set $\mathcal{D}_2^{\text{rs}}$, we define

$$\tau_i \coloneqq \max\{\hat{q}_{\epsilon_{\text{lo}}}(S_i) - R_i, R_i - \hat{q}_{\epsilon_{\text{up}}}(S_i)\} \tag{7}$$

[1] For the sake of simplicity in the discussion, we assume no additional exogenous randomness is introduced by the quantile prediction algorithm.

[2] This framework is flexible and can be adapted to other interval forms. For instance, if the target distribution P^{π_e} is characterized by a joint probability density function $f(s,r)$, the intervals can be parameterized in a highest-density form: $\hat{C}_\tau(s) = \{r \in \mathcal{R} \mid \hat{f}(s,r) \geq \tau\}$, where $\tau \geq 0$ and $\hat{f}$ is the estimated density function.

which represents the minimal τ such that $R_i \in \hat{C}_\tau(S_i)$. Note that these τ_i values are referred to as non-conformity scores within the framework of CP, where they quantify the extents to which the observations in the calibration set "conform" to the training set. In the split CP method, the empirical $1-\epsilon$ quantile of these scores is typically used to construct a $1-\epsilon$ marginally valid PI, which has been shown to also achieve (ε, δ)-PAC validity, with $\varepsilon \geq \epsilon + \sqrt{\frac{-\log \delta}{2M}}$ ([Vovk, 2013]). While one could employ split CP with the $1 - \epsilon + \sqrt{\frac{-\log \delta}{2M}}$ quantile for constructing a (ϵ, δ)-PAC valid PI, this requires that M exceeds $(-\log \delta)/2\epsilon^2$, a condition that may not be guaranteed in our setting.

To address this issue, we introduce a critical constant $k(M, \epsilon, \delta)$, defined as

$$k(M, \epsilon, \delta) \coloneqq \underset{k \in \{-1, 0, \ldots, M-1\}}{\arg\max} \; k \text{ s.t. } F_{\mathrm{B}(M,\epsilon)}(k) \leq \delta, \tag{8}$$

where $F_{\mathrm{B}(M,\epsilon)}(\cdot)$ is the cumulative distribution function of a binomial distribution $\mathrm{B}(M, \epsilon)$ with M trials and success probability ϵ. Rather than using a specific empirical quantile of the non-comformity scores, we define

$$\tilde{\tau} \coloneqq \tau_{(M-k(M,\epsilon,\delta))}, \tag{9}$$

where $\tau_{(k)}$ denotes the k-th smallest τ_i, with $\tau_{(M+1)} = \infty$. In Theorem 1 (with proof provided in Appendix A.2), we demonstrate that $\hat{C}_{\tilde{\tau}}(S_{n+1})$ forms a valid (ϵ, δ)-PAC prediction interval.

Theorem 1. *Suppose the Assumption 1 holds and the behavior policy π_b is known. For any $\tau \geq \tilde{\tau}$, where $\tilde{\tau}$ is defined in (9), it holds that*

$$\mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_\tau) \leq \epsilon\right] \geq 1 - \delta. \tag{10}$$

Using the RS method (5) with a known behavior policy, the above procedure generates PIs $\{\hat{C}_\tau(S_{n+1})\}_{\tau \geq \tilde{\tau}}$ that are approximately $1 - \epsilon$ correct with probabilities at least $1 - \delta$, where the probabilities are taken over both the observation dataset $\mathcal{D}$ and the auxiliary variables V. Clearly, typical measures, such as the cardinality or Lebesgue measure, of the interval size are non-decreasing as τ increases. Therefore, $\hat{C}_{\tilde{\tau}}(S_{n+1})$ is a better choice compared to any $\hat{C}_\tau(S_{n+1})$ with $\tau > \tilde{\tau}$ from the perspective of informativeness. To further illustrate the efficiency of $\hat{C}_{\tilde{\tau}}$, we establish in Theorem 2 two-side bounds on the probability that the coverage error of $\hat{C}_{\tilde{\tau}}$ is exact ϵ.

Theorem 2. *Assume the conditions in Theorem 1 hold. Additionally, suppose that all τ_i defined in (7) have no ties almost surely. Then, for any fixed $\Delta_\epsilon \in (0, \epsilon)$, there exists positive constants C_1 and C_2 depending only on ϵ, δ and Δ_ϵ, such that*

$$1 - \delta - \frac{C_1}{\sqrt{n}} < \mathbb{P}\left[\epsilon - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon\right] < 1 - \delta + \frac{C_2}{\sqrt{n}}. \tag{11}$$

Theorem 2 demonstrates that $\hat{C}_{\tilde{\tau}}$ achieves exact $1 - \epsilon$ coverage with probability $1 - \delta$ in an asymptotic sense (see Appendix A.3 for the proof). Simulations were conducted to verify this result and provide intuitive insights, as shown in Figure 1. Although equation (11) indicates

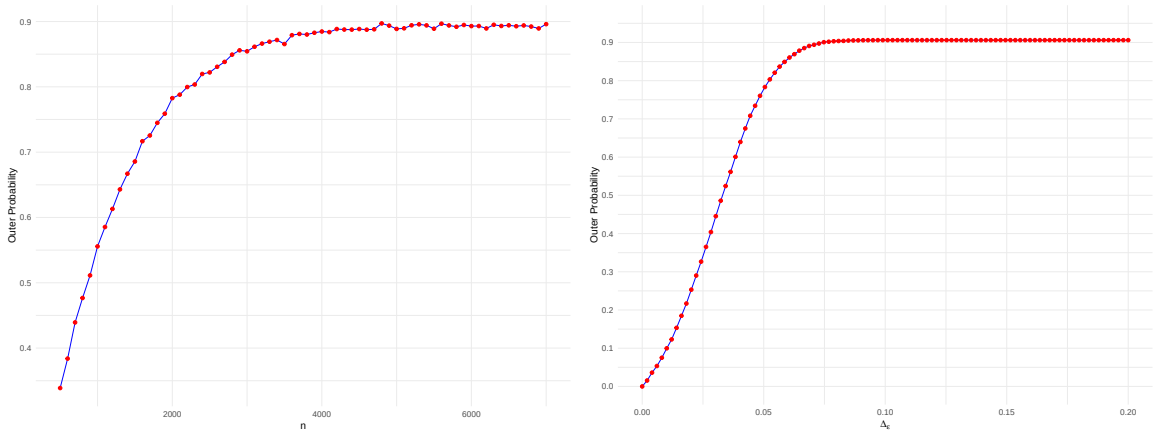

Figure 1: Empirical probabilities of $\mathbb{P}[\epsilon - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon]$ for varying sample sizes n (with $\Delta_\epsilon = 0.05$) and varying Δ_ϵ (with $n = 2000$), $\epsilon = 0.2$, $\delta = 0.1$.

that there is always an asymptotic probability δ such that the coverage of $\hat{C}_{\tilde{\tau}}(S_{n+1})$ is strictly less than $1 - \epsilon$, it can be similarly derived that

$$\delta - \frac{C}{\sqrt{n}} < \mathbb{P}\left[\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon + \Delta_\epsilon\right] \leq \delta,$$

for any fixed $\Delta_\epsilon \in (0, 1 - \epsilon]$ and a constant C. This implies $\hat{C}_{\tilde{\tau}}(S_{n+1})$ achieves exact $1 - \epsilon$ coverage over the joint distribution of (S_{n+1}, R_{n+1}) with a probability tending to 1. To further investigate the coverage conditional on the testing context S_{n+1}, let $q_\alpha(s)$ denote the α-th quantile of the conditional distribution of R_{n+1} given $S_{n+1} = s$, where $\alpha \in \{\epsilon_{\text{lo}}, \epsilon_{\text{up}}\}$. Then, a desirable oracle PI would be

$$C^{\text{oracle}}(S_{n+1}) = [q_{\epsilon_{\text{lo}}}(S_{n+1}), q_{\epsilon_{\text{up}}}(S_{n+1})], \tag{12}$$

which ensures exact $1 - \epsilon$ coverage conditional on any value of S_{n+1}. In Theorem 3, we show that, assuming the consistency of the quantile estimators (see Assumption 3 for details), along with a regularity condition, the symmetric difference between $\hat{C}_{\tilde{\tau}}(S_{n+1})$ and $C^{\text{oracle}}(S_{n+1})$ converges to zero in probability.

Theorem 3. *Assume the conditions in Theorem 1 hold. Under Assumptions 3 and 4, as* $n \to \infty$,

$$\mathcal{L}\left(\hat{C}_{\tilde{\tau}}(S_{n+1}) \Delta C^{\text{oracle}}(S_{n+1})\right) = o_{\mathbb{P}}(1).$$

Here, $\mathcal{L}(\cdot)$ *denotes the Lebesgue measure, and* Δ *is the symmetric difference operator.*

Thus far, we have assumed that the behavior policy π_b is known, thereby granting access to the oracle weight function w. However, in most real-world applications, π_b is typically unknown but can be estimated from observational data. We summarize our aforementioned approach in Algorithm 1, wherein the behavior policy is replaced with an estimator $\hat{\pi}_b$.

We proceed to establish a theoretical guarantee for Algorithm 1 (see Appendix A.5 for the proof), assuming the estimated behavior policy is sufficiently accurate. Formally, we make the following assumption:

Algorithm 1 PAC Off-Policy Prediction

1: **Input** Observational data $\mathcal{D} = \{S_i, A_i, R_i\}_{i=1}^n$; PAC parameters $(\epsilon, \delta) \in (0,1)^2$; a target policy π_e; a behavior policy estimation algorithm AL_p; a quantile prediction algorithm AL_q; quantile levels $\epsilon_{\text{lo}}, \epsilon_{\text{up}}$ with $\epsilon_{\text{up}} - \epsilon_{\text{lo}} = 1 - \epsilon$; and a test context S_{n+1}.
2: Split $\mathcal{D}$ into two disjoint subsets $\mathcal{D}_1 \cup \mathcal{D}_2$ of sizes n_1 and n_2, respectively.
3: Apply AL_p to $\mathcal{D}_1$ to estimate the behavior policy $\hat{\pi}_b$.
4: Set $\hat{w}(s,a) = \frac{\pi_e}{\hat{\pi}_b}(\mathrm{d}a|s)$ and calculate $\hat{b} = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \hat{w}(s,a)$. Independently sample $V_i \sim U([0,1])$ for $i = 1, \ldots, n$.
5: Extract (S_i, R_i) from (S_i, A_i, R_i) in $\mathcal{D}_1$ and $\mathcal{D}_2$ for which $V_i \leq \frac{1}{\hat{b}}\hat{w}(S_i, A_i)$. Denote these subsets by $\mathcal{D}_1^{\text{rs}}$ and $\mathcal{D}_2^{\text{rs}}$, respectively.
6: Use AL_q on $\mathcal{D}_1^{\text{rs}}$ to train conditional quantile functions $\hat{q}_{\epsilon_{\text{lo}}}$ and $\hat{q}_{\epsilon_{\text{up}}}$. Compute τ_i for each data in $\mathcal{D}_2^{\text{rs}}$ according to (7).
7: Set $\tilde{\tau}$ as the $(M - k(M,\epsilon,\delta))$-th smallest τ_i, where $k(M,\epsilon,\delta)$ is defined in (8) and $M = |\mathcal{D}_2^{\text{rs}}|$.
8: **Output** the Prediction interval: $\hat{C}_{\tilde{\tau}}(S_{n+1}) = [\hat{q}_{\epsilon_{\text{lo}}}(S_{n+1}) - \tilde{\tau}, \hat{q}_{\epsilon_{\text{up}}}(S_{n+1}) + \tilde{\tau}] \cap \mathcal{R}$.

Assumption 2. *There exist $(\epsilon', \delta') \in (0,1)^2$ such that the estimated weight function $\hat{w}(s,a) = \frac{\pi_e}{\hat{\pi}_b}(\mathrm{d}a|s)$ and $\hat{b} = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \hat{w}(s,a)$ in step 4 of Algorithm 1 satisfy*

$$\mathbb{P}\left[\mathbb{E}_{P^{\pi_b}}\left|\hat{w}(S,A)/\mathbb{E}_{P^{\pi_b}}[\hat{w}(S,A)] - w(S,A)\right| \leq 2\epsilon'\right] \geq 1 - \delta', \tag{13}$$

and $\hat{b} < \infty$ a.s., where, with a slight abuse of notation, $\mathbb{E}_{P^{\pi_b}}(\cdot)$ denotes the expectation over the joint distribution $P_S \times \pi_b$ conditional on the training set $\mathcal{D}_1$, and the probabilities are under the randomness of $\mathcal{D}_1$.

Theorem 4. *Suppose that the Assumption 2 holds. Then, the output $\hat{C}_{\tilde{\tau}}$ from Algorithm 1 satisfies*

$$\mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon + \epsilon'\right] \geq (1 - \delta')(1 - \delta). \tag{14}$$

In addition, if all τ_i have no ties almost surely. Then, for any fixed $\Delta_\epsilon \in (0, \epsilon)$, we also have

$$\mathbb{P}\left[\epsilon - \epsilon' - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon + \epsilon'\right] \geq (1 - \delta')(1 - \delta - \frac{C}{\sqrt{n}}), \tag{15}$$

for some positive constants C depending only on ϵ, δ and Δ_ϵ.

In general, π_b can be estimated using existing supervised learning algorithms. However, these methods typically fail to provide theoretical guarantees for satisfying condition (13). In Appendix B, we introduce a Maximum Likelihood Estimation (MLE) approach for estimating π_b and demonstrate that, under the realizable and uniformly bounded assumptions, condition (12) is satisfied when n_1, the size of $\mathcal{D}_1$, is sufficiently large. Furthermore, we can see that, as $n_1, n_2 \to \infty$, the output $\hat{C}_{\tilde{\tau}}$ of Algorithm 1 using the MLE (24) recovers exact $1 - \epsilon$ coverage asymptotically, as guaranteed by (15).

4 Synthetic Data Experiments

In the absence of established baselines for our problem, we compare our proposed method, PACOPP, with the following competing approaches, both of which are distribution-free off-

policy prediction algorithms capable of handling continuous action spaces.

Conformal Off-Policy Prediction (COPP). COPP was introduced in Taufiq et al. [2022] and has been demonstrated to achieve marginal valid coverage. Using a behavior policy estimator $\hat{\pi}_b$ and a reward distribution estimator $\hat{P}_R$, both trained on $\mathcal{D}_1$, COPP estimates the weight (4) via a Monte Carlo approach: $\hat{w}(s,r) = \sum_{i=1}^{h} \hat{P}_R(r|s,a_i^e) / \sum_{i=1}^{h} \hat{P}_R(r|s,a_i)$, where $a_i \sim \hat{\pi}_b(\cdot|s), a_i^e \sim \pi_e(\cdot|s)$ and h is the number of Monte Carlo samples. COPP constructs PIs based on the weighted CP framework. That is, it outputs (s,r) pairs whose non-conformity scores lie below the $1-\epsilon$ quantile of the weighted empirical distribution: $\hat{F}_{n_2}^{s,r} \coloneqq \sum_{\mathcal{D}_2} p_i^{\hat{w}}(s,r)\delta_{\tau_i} + p_{n_2+1}^{\hat{w}}(s,r)\delta_{\infty}$, where $p_i^{\hat{w}}(s,r) \coloneqq \frac{\hat{w}(S_i,R_i)}{\sum_{\mathcal{D}_2} \hat{w}(S_i,R_i)+\hat{w}(s,r)}$, and $p_{n_2+1}^{\hat{w}}(s,r) \coloneqq \frac{\hat{w}(s,r)}{\sum_{\mathcal{D}_2} \hat{w}(S_i,R_i)+\hat{w}(s,r)}$.

Conformal Off-Policy Prediction with Rejection Sampling (COPP-RS) To better address the question of "why use PAC PIs instead of marginally valid PIs," we propose a new algorithm, COPP-RS, which, like PACOPP, employs rejection sampling to handle distribution shift. The only difference is that COPP-RS directly uses the $1-\epsilon$ empirical quantile of the non-conformity scores for construction. Specifically, denote by $\tau^{(1-\epsilon)}$ the $1-\epsilon$ quantile of the distribution $\sum_{(S_i,R_i)\in\mathcal{D}_2^{\text{rs}}} \frac{1}{M+1}\delta_{\tau_i} + \frac{1}{M+1}\delta_{\infty}$. COPP-RS then outputs the PI: $\hat{C}_{\tau^{(1-\epsilon)}}(S_{n+1}) = [\hat{q}_{\epsilon_{\text{lo}}}(S_{n+1}) - \tau^{(1-\epsilon)}, \hat{q}_{\epsilon_{\text{up}}}(S_{n+1}) + \tau^{(1-\epsilon)}] \cap \mathcal{R}$. It is straightforward to verify that $\hat{C}_{\tau^{(1-\epsilon)}}(S_{n+1})$ also attains marginal coverage at level $1-\epsilon$ under the conditions in Theorem 1.

Implementation details. We consider an experimental setup similar to the continuous action space scenario described in [Taufiq et al., 2022]. The observational data $\mathcal{D}$ is generated according to the following distributions:

$$S_i \overset{\text{i.i.d.}}{\sim} N(0,4);\ A_i|s_i \sim N(s_i/4,4);\ R_i|s_i,a_i \sim N(s_i+a_i,1) + U([-2,2]).$$

A total of $n = 2000$ samples are generated, with 75% allocated for training and the remaining for calibration. In the training set, neural networks (NNs) are used to estimate the behavior policy $\hat{\pi}_b$ and the quantiles $\hat{q}_{\epsilon/2}, \hat{q}_{1-\epsilon/2}$. For the COPP algorithm, in addition to these estimates, we also estimate the reward distribution P_R. Here, we assume the conditional density model is misspecified as $\hat{P}_R(r|s,a) = N(\mu(s,a), \sigma(s,a))$, where $\mu(s,a)$ and $\sigma(s,a)$ are both NNs. The number of Monte Carlo samples is $h = 100$. We define the target policy as

$$\pi_e(\cdot|s) = N(s/4, 1).$$

In each simulation, 10000 test data points are generated from the target distribution to evaluate the coverage probability. Finally, we set the nominal miscoverage level $\epsilon = 0.2$ and conduct 1000 simulations.

Results. Figure 2 presents the coverages and interval lengths of different methods. For the PACOPP algorithm, we set $\delta = 0.5$, 0.25, 0.1, and 0.01, denoted as PAC-0.5, PAC-0.25, PAC-0.1, and PAC-0.01, respectively. Our findings are summarized as follows. First, since COPP requires estimating the reward distribution P_R, it fails to achieve the nominal coverage when the conditional density model is misspecified. In contrast, methods based on rejection sampling (COPP-RS and PACOPP) achieve valid coverage. Second, compared to COPP-RS, PACOPP provides control over the probability of undercoverage, i.e., when the coverage falls below $1-\epsilon$. As observed in our experiments, when $\delta = 0.5$, the performance of PACOPP's PI

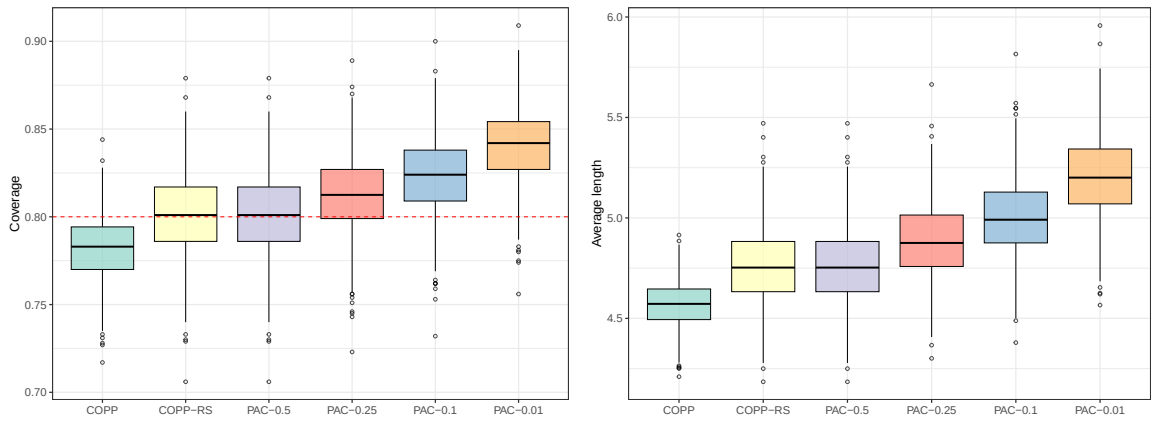

Figure 2: Empirical coverages and average lengths of prediction intervals based on COPP, COPP-RS and PACOPP with $\delta = 0.5, 0.15, 0.1$ and 0.01. The nominal level is 80%.

closely matches that of COPP-RS. As δ decreases, the probability of undercoverage decreases, while the interval length increases accordingly. Compared to marginally valid methods, PACOPP provides an additional mechanism to balance the trade-off between the probability of undercoverage and interval efficiency. Notably, even when $\delta = 0.01$, PACOPP's PI is not excessively conservative. Therefore, in scenarios requiring strict control over the risk of undercoverage, PACOPP is more suitable than marginally valid methods.

5 Related Work

Off-policy evaluation. OPE is one of the most fundamental topics in Reinforcement learning ([Sutton and Barto, 2018]) and has been extensively studied, resulting in a vast body of literature. The primary challenge in OPE arises from the distribution shift in rewards, induced by the discrepancy between the behavior and target policies. Current methods, which typically focus on estimating the expected reward (policy value), are broadly categorized into three main approaches: (i) importance sampling ([Precup, 2000]; [Liu et al., 2018]; [Schlegel et al., 2019]), known for its unbiased nature but susceptible to high variance; (ii) direct methods ([Thomas and Brunskill, 2016]; [Le et al., 2019]; [Shi et al., 2022]), which directly learn the model before policy evaluation, potentially introducing bias but offering lower variance; and (iii) doubly robust methods ([Dudík et al., 2011]; [Jiang and Li, 2016]; [Kallus and Uehara, 2020]), which combine the first two approaches to achieve more robust estimators. For an extensive review, we refer readers to [Uehara et al., 2022].

In addition to point estimates for the value of the target policy, less attention has been paid to interval estimates of the policy value for uncertainty quantification. To provide confidence regarding the accuracy of these estimates, [Thomas et al., 2015] proposed high confidence off-policy evaluation, which derives a lower confidence bound on the target policy's value by applying concentration inequalities to importance sampling estimates. Other approaches, such as bootstrap ([Hanna et al., 2017]), kernel Bellman loss ([Feng et al., 2020]) and empirical likelihood methods ([Dai et al., 2020]), have also been employed to derive confidence intervals.

Nevertheless, all of these methods focus on the average effect of the target policy and do not account for the variability in the reward itself.

Conformal prediction. CP has gained popularity in OPE due to its ability to construct distribution-free PIs with finite-sample guarantees and account for individual effects. The application of CP to OPE originated from [Tibshirani et al., 2019], who developed the "weighted conformal prediction" method that extends standard CP to covariate shift settings, in which the covariate distributions of test and training data differ, while the conditional distributions remain the same. This method offers an valuable approach to addressing distribution shift and was subsequently applied to OPE in contextual bandits [Taufiq et al., 2022] (COPP) and Markov decision processes [Foffano et al., 2023].

As discussed in Section 4, these direct application of weighted CP in OPE requires estimating the conditional probability densities of rewards, and may underperform if the model is misspecified. Additionally, COPP directly estimates the conditional quantiles from observational data. As a result, the algorithm essentially calibrates intervals constructed from the estimated conditional quantiles under the behavior policy. Even if the quantile estimation algorithm is consistent, the resulting PIs will not converge to the oracle intervals defined in (12). In contrast, PACOPP ensures this convergence property (Theorem 3). Furthermore, since the weight (4) depends on both s and r, COPP must use a grid of potential values for R_{n+1} corresponding to each S_{n+1} when generating the final PI, introducing additional computational overhead. By contrast, PACOPP explicitly outputs PIs without this extra burden.

In [Zhang et al., 2023], the authors avoid model estimation by selecting subsamples from observational data where the action matches the pseudo action generated by a designed auxiliary policy. These subsamples preserve the same conditional distribution as the target population, enabling the use of weighted CP for PI construction. However, their method is inherently limited, as it is only applicable when the action space is discrete. In contrast, our approach, PACOPP, is applicable to continuous action spaces, such as the doses of medication administered in precision medicine.

As discussed in Section 3, although split (inductive) CP has been shown in [Vovk, 2013] to automatically achieve training conditional validity in a PAC type, it is not applicable in our context due to its sample size requirement. In [Park et al., 2020], the authors proposed an adjusted version of split CP that constructs confidence sets for deep neural networks with finite-sample PAC validity. This method was further extended to handle covariate shift settings in [Park et al., 2022], where it was modified using Clopper-Pearson upper bounds to handle cases when the importance weights (the likelihood ratio of covariate distributions) are unknown, but confidence intervals for these weights are available. The distinction between our approach and theirs is that PACOPP explicitly provides the calibration parameter as specified in (9), while their approach requires solving an optimization problem. Additionally, we derive theoretical probability bounds for our approach achieving exact nominal coverage (Theorem 2) and establish its asymptotic efficiency. In contrast, their method lacks these theoretical foundations.

6 Conclusion

In this paper, we introduce PACOPP, a novel algorithm for constructing predictive intervals on off-policy rewards in contextual bandit settings. PACOPP enjoys distribution-free finite-sample guarantees and adaptivity to individuals by adapting conformal prediction to the OPE task. Unlike previous approaches, PACOPP is the first to achieve probably approximately correct validity, with the ability to control undercoverage probability through a separate confidence level.

Currently, we address distribution shift through rejection sampling, which, while feasible, reduces the available sample size, resulting in lower data utilization and ultimately leading to more conservative PIs. Improving the data utilization in this context presents a challenge that requires further exploration. Additionally, extending PACOPP to sequential decision-making scenarios would be an interesting direction for further research.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32, 2019.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:27475–27490, 2021.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33:9398–9411, 2020.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104, 2011.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Yihao Feng, Tongzheng Ren, Ziyang Tang, and Qiang Liu. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pages 3102–3111. PMLR, 2020.

Daniele Foffano, Alessio Russo, and Alexandre Proutiere. Conformal off-policy evaluation in markov decision processes. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 3087–3094. IEEE, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Josiah Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Audrey Huang, Liu Leqi, Zachary Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. *Advances in Neural Information Processing Systems*, 34: 23714–23726, 2021.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

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

Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4436–4443, 2020.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Von Neumann. Various techniques used in connection with random digits. *Notes by GE Forsythe*, pages 36–38, 1951.

Art B Owen. Monte carlo theory, methods and examples, 2013.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. Pac confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022.

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

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, and Martha White. Importance resampling for off-policy prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1), 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):765–793, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Muhammad Faaiz Taufiq, Jean-Francois Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. *Advances in Neural Information Processing Systems*, 35:31512–31524, 2022.

James W Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of forecasting*, 19(4):299–311, 2000.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92 (2):349–376, 2013.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 2751–2768. PMLR, 2023.

A Proofs

A.1 Proof of Proposition 1

Proof. Clearly, for each $i \in [n]$, the probability that $V_i \leq \frac{1}{b} w(S_i, A_i)$ equals to

$$\int_{\mathcal{S}} \int_{\mathcal{A}} \frac{1}{b} \frac{\pi_e}{\pi_b}(\mathrm{d}a|s) \pi_b(\mathrm{d}a|s) P_S(\mathrm{d}s) = \frac{1}{b}.$$

Therefore, the size N_{rs} follows a binomial distribution $B(n, \frac{1}{b})$ because of the independence. For any $\{(s_j, r_j)\}_{j=1}^{N_{\mathrm{rs}}} \in (\mathcal{S} \times \mathcal{R})^{N_{\mathrm{rs}}}$, conditional on the event that $N_{\mathrm{rs}} = n_{\mathrm{rs}} \in [n]$,

$$\begin{aligned}
&\mathbb{P}(Z_j \in (\mathrm{d}s_j, \mathrm{d}r_j), \forall j \in [N_{\mathrm{rs}}] \mid N_{\mathrm{rs}} = n_{\mathrm{rs}}) \\
=&\frac{1}{\mathbb{P}(N_{\mathrm{rs}} = n_{\mathrm{rs}})} \mathbb{P}\left(\exists\, \sigma_1 < \cdots < \sigma_{n_{\mathrm{rs}}} \in [n] \ s.t.\ (S_{\sigma_j}, R_{\sigma_j}) \in (\mathrm{d}s_j, \mathrm{d}r_j), V_{\sigma_j} \leq \frac{1}{b} w(S_{\sigma_j}, R_{\sigma_j}), \forall j \in [n_{\mathrm{rs}}] \right. \\
&\qquad \left. \text{and } V_i > \frac{1}{b} w(S_i, A_i), \forall i \in [n] \setminus \{\sigma_1, \ldots, \sigma_{n_{\mathrm{rs}}}\} \right) \\
=&\frac{1}{\binom{n}{n_{\mathrm{rs}}} b^{-n_{\mathrm{rs}}} (1 - \frac{1}{b})^{n - n_{\mathrm{rs}}}} \sum_{\sigma_1 < \cdots < \sigma_{n_{\mathrm{rs}}}} b^{-n_{\mathrm{rs}}} (1 - \frac{1}{b})^{n - n_{\mathrm{rs}}} \prod_{j \in [n_{\mathrm{rs}}]} \mathbb{P}\left((S_{\sigma_j}, R_{\sigma_j}) \in (\mathrm{d}s_j, \mathrm{d}r_j) \mid V_{\sigma_j} \leq \frac{1}{b} w(S_{\sigma_j}, R_{\sigma_j}) \right),
\end{aligned}$$

where the summation is taken over all possible choices of $\sigma_1, \ldots, \sigma_{n_{\mathrm{rs}}}$. Moreover, for each $i \in [n]$ and any $(s, r) \in \mathcal{S} \times \mathcal{R}$,

$$\begin{aligned}
\mathbb{P}\left((S_i, R_i) \in (\mathrm{d}s, \mathrm{d}r) \mid V_i \leq \frac{1}{b} w(S_i, R_i) \right) =& b \int_{\mathcal{A}} \frac{1}{b} \frac{\pi_e}{\pi_b}(\mathrm{d}a|s) P_S(\mathrm{d}s) \pi_b(\mathrm{d}a|s) P_R(\mathrm{d}r|s, a) \\
=& \int_{\mathcal{A}} P_S(\mathrm{d}s) \pi_e(da|s) P_R(\mathrm{d}r|s, a) \\
=& P^{\pi_e}(\mathrm{d}s, \mathrm{d}r).
\end{aligned}$$

Finally, we have

$$\mathbb{P}(Z_j \in (\mathrm{d}s_j, \mathrm{d}r_j), \forall j \in [N_{\mathrm{rs}}] \mid N_{\mathrm{rs}} = n_{\mathrm{rs}}) = \prod_{j \in [n_{\mathrm{rs}}]} P^{\pi_e}(\mathrm{d}s_j, \mathrm{d}r_j).$$

The proof is complete. □

A.2 Proof of Theorem 1

Proof. We first consider the probability of $\{L_{P^{\pi_e}}(\hat{C}_\tau) \leq \epsilon\}$ conditional on the event that the rejection sampling procedure generates $N_{\mathrm{rs}} = n_{\mathrm{rs}}$ samples. We omit the trivial cases of $n_{\mathrm{rs}} = 0$

or 1, where $\hat{C}_\tau(s)$ can be set to $\mathcal{R}$. For nontrivial cases, Proposition 1 enables us to interpret the conditional probability as

$$\mathbb{P}[L_{P^{\pi_e}}(\hat{C}_\tau) \le \epsilon \mid N_{\mathrm{rs}} = n_{\mathrm{rs}}] = \mathbb{P}_{\mathcal{D}^{\mathrm{rs}} \sim (P^{\pi_e})^{n_{\mathrm{rs}}}}[L_{P^{\pi_e}}(\hat{C}_\tau) \le \epsilon].$$

Now, for any partition of $\mathcal{D}^{\mathrm{rs}}$ and any realization of the set $\mathcal{D}_1^{\mathrm{rs}}$, the parameterized interval $\hat{C}_\tau(s)$ for a fixed s is nonrandom after training $\hat{q}_{\epsilon_{\mathrm{lo}}}$ and $\hat{q}_{\epsilon_{\mathrm{up}}}$, and depends only on τ. It can be easily verified from the definition (6) that the coverage error $L_{P^{\pi_e}}(\hat{C}_\tau) = \mathbb{P}_{(S_{n+1},R_{n+1})\sim P^{\pi_e}}(R_{n+1} \notin \hat{C}_\tau(S_{n+1}))$, as a function of τ, is monotonically decreasing and right-continuous. Next, we define

$$\tau^* := \inf\{\tau \in \mathbb{R} : L_{P^{\pi_e}}(\hat{C}_\tau) \le \epsilon\}, \tag{16}$$

and let $\{\alpha_j\}_{i=1}^{\infty}$ be a positive sequence such that $\alpha_j \downarrow 0$. Denote by $m = \lfloor \gamma n_{\mathrm{rs}} \rfloor$ the size of $\mathcal{D}_2^{\mathrm{rs}}$. For $\tilde{\tau} = \tau_{(m-k(m,\epsilon,\delta))}$ with $k(m,\epsilon,\delta) \neq -1$, the right-continuity implies that the event

$$\{L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) > \epsilon\} \Longleftrightarrow \{\tilde{\tau} < \tau^*\} \Longleftrightarrow \bigcup_{j=1}^{\infty} \{\tilde{\tau} < \tau^* - \alpha_j\}.$$

Since τ_i represents the minimal τ such that $R_i \in \hat{C}_\tau(S_i)$ in $\mathcal{D}_2^{\mathrm{rs}}$,

$$\begin{aligned}\{\tilde{\tau} < \tau^* - \alpha_j\} &\Longleftrightarrow \{\exists \text{ at most } k(m,\epsilon,\delta) \text{ indeices } i \ s.t.\ \tau_i \ge \tau^* - \alpha_j\} \\ &\Longrightarrow \{\exists \text{ at most } k(m,\epsilon,\delta) \text{ indeices } i \ s.t.\ \tau_i > \tau^* - \alpha_j\} \\ &\Longleftrightarrow \{\exists \text{ at most } k(m,\epsilon,\delta) \text{ indeices } i \ s.t.\ R_i \notin \hat{C}_{\tau^*-\alpha_j}(S_i)\}.\end{aligned}$$

As $\mathcal{D}_2^{\mathrm{rs}} \sim (P^{\pi_e})^m$, each sample in $\mathcal{D}_2^{\mathrm{rs}}$ independently satisfies $R_i \notin \hat{C}_{\tau^*-\alpha_j}(S_i)$ with probability $L_{P^{\pi_e}}(\hat{C}_{\tau^*-\alpha_j})$, and we have that $L_{P^{\pi_e}}(\hat{C}_{\tau^*-\alpha_j}) > \epsilon$ by the definition of τ^*. Then, it holds

$$\mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}}\sim(P^{\pi_e})^m}(\tilde{\tau} < \tau^* - \alpha_j) \le F_{\mathrm{B}(m,L_{P^{\pi_e}}(\hat{C}_{\tau^*-\alpha_j}))}(k(m,\epsilon,\delta)) \le F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)) \le \delta,$$

where the second inequality follows from the fact that, for a fixed k, the c.d.f. $F_{\mathrm{B}(m,\varepsilon)}(k) := \sum_{i=0}^{k} \binom{m}{i} \varepsilon^i (1-\varepsilon)^{m-i}$ is decreasing w.r.t. $\varepsilon \in [0,1]$, and the last inequality follows from the definition of $k(m,\epsilon,\delta)$. Together with the continuity of measures, we have

$$\mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}}\sim(P^{\pi_e})^m}[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) > \epsilon] = \lim_{j\to\infty} \mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}}\sim(P^{\pi_e})^m}(\tilde{\tau} < \tau^* - \alpha_j) \le \delta, \tag{17}$$

which also holds if $k(m,\epsilon,\delta) = -1$, in which case $L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) = 0$.

Finally, since (17) is true for any partition and realization of $\mathcal{D}_1^{\mathrm{rs}}$, we can marginalize to obtain

$$\mathbb{P}_{\mathcal{D}^{\mathrm{rs}}\sim(P^{\pi_e})^{n_{\mathrm{rs}}}}[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \le \epsilon] \ge 1 - \delta.$$

Multiplying by the probability of $N_{\mathrm{rs}} = n_{\mathrm{rs}}$ and summing over n_{rs}, it then holds for any $\tau \ge \tilde{\tau}$ that

$$\mathbb{P}[L_{P^{\pi_e}}(\hat{C}_\tau) \le \epsilon] \ge 1 - \delta.$$

The proof is complete. □

A.3 Proof of Theorem 2

To facilitate the proof of Theorem 2, we first establish an upper bound on the probability that $\hat{C}_{\tilde{\tau}}$ is approximately correct, as detailed in Lemma 2. The proof of this lemma relies on the well-known Berry-Esseen inequality, presented in Lemma 1 (Theorem 3.4.17 in [Durrett, 2019]). Throughout this part, the letters C_1, C_2 and C denote positive constants that do not depend on m, l or n. Their values may vary at different places.

Lemma 1 (Berry-Esseen inequality)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma^2$ and $\mathbb{E}|X_i|^3 = \rho < \infty$. If $F_n(x)$ is the distribution function of $\sum_{i=1}^n X_i/\sigma\sqrt{n}$ and $\Phi(x)$ is the standard normal distribution function, then it holds for all n that*

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{3\rho}{\sigma^3\sqrt{n}}.$$

Lemma 2. *Assume the conditions in Theorem 2 hold. For $\tilde{\tau}$ defined in (9), we have*

$$\mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon\right] < 1 - \delta + \frac{C}{\sqrt{n}}. \tag{18}$$

Proof of Lemma 2
Similar to the proof of Theorem 1, we first show that

$$\mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon | N_{\mathrm{rs}} = n_{\mathrm{rs}}, \mathcal{D}_1^{\mathrm{rs}}\right] = \mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}}\sim(P^{\pi_e})^m}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon\right] < 1 - \delta + \frac{C}{\sqrt{m}}, \tag{19}$$

where $\tilde{\tau} = \tau_{(m-k(m,\epsilon,\delta))}$ and $m = \lfloor \gamma n_{\mathrm{rs}} \rfloor$. Suppose that $k(m,\epsilon,\delta) \neq -1$, which is always true if $m \geq m_0 \coloneqq \log_{1-\epsilon}\delta$. Then, for τ^* defined in (16),

$$\begin{aligned}
\{L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon\} &\Longleftrightarrow \{\tilde{\tau} \geq \tau^*\} \Longleftrightarrow \{\exists \text{ at least } k(m,\epsilon,\delta)+1 \text{ indeices } i \text{ s.t. } \tau_i \geq \tau^*\} \\
&\Longrightarrow \{\exists \text{ at least } k(m,\epsilon,\delta) \text{ indeices } i \text{ s.t. } \tau_i > \tau^*\} \\
&\Longleftrightarrow \{\exists \text{ at most } k(m,\epsilon,\delta) \text{ indeices } i \text{ s.t. } R_i \notin \hat{C}_{\tau^*-\alpha_j}(S_i)\}.
\end{aligned}$$

since there are almost surely no ties. Together with the fact that $L_{P^{\pi_e}}(\hat{C}_{\tau^*}) \leq \epsilon$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}}\sim(P^{\pi_e})^m}[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon] \\
\leq& 1 - F_{\mathrm{B}(m, L_{P^{\pi_e}}(\hat{C}_{\tau^*}))}(k(m,\epsilon,\delta)-1) \\
\leq& 1 - F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)-1) \\
=& 1 - F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)+1) + F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)+1) - F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)-1) \\
<& 1 - \delta + F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)+1) - F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)-1),
\end{aligned}$$

where the last inequality follows from the definition of $k(m,\epsilon,\delta)$. To bound the last two terms, it follows from Lemma 1 that

$$\begin{aligned}
&F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)+1) - F_{\mathrm{B}(m,\epsilon)}(k(m,\epsilon,\delta)-1) \\
\leq& 2\sup_{x\in\mathbb{R}}\left|F_{\mathrm{B}(m,\epsilon)}(\sigma_\epsilon\sqrt{m}x + m\epsilon) - \Phi(x)\right| + \Phi\left(\frac{k(m,\epsilon,\delta)+1-m\epsilon}{\sigma_\epsilon\sqrt{m}}\right) - \Phi\left(\frac{k(m,\epsilon,\delta)-1-m\epsilon}{\sigma_\epsilon\sqrt{m}}\right) \\
\leq& \frac{6(\epsilon(1-\epsilon)^3 + (1-\epsilon)\epsilon^3)}{\sigma_\epsilon^3}\frac{1}{\sqrt{m}} + \sqrt{\frac{2}{\pi}}\frac{1}{\sigma_\epsilon\sqrt{m}},
\end{aligned}$$

where $\sigma_\epsilon = \sqrt{\epsilon(1-\epsilon)}$. Therefore, (19) holds for $m \geq m_0$. After marginalizing, it holds for $n_{\mathrm{rs}} \geq \lfloor m_0/\gamma \rfloor + 1$ that

$$\mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon | N_{\mathrm{rs}} = n_{\mathrm{rs}}\right] < 1 - \delta + \frac{C_1}{\sqrt{n_{\mathrm{rs}}}}.$$

Finally, for $n \geq \lfloor m_0/\gamma \rfloor + 1$,

$$\begin{aligned}
\mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon\right] &\leq \sum_{n_{\mathrm{rs}}=0}^{\lfloor m_0/\gamma \rfloor} \mathbb{P}(N_{\mathrm{rs}} = n_{\mathrm{rs}}) + \sum_{n_{\mathrm{rs}}=\lfloor m_0/\gamma \rfloor+1}^{n} \mathbb{P}(N_{\mathrm{rs}} = n_{\mathrm{rs}}) \mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon | N_{\mathrm{rs}} = n_{\mathrm{rs}}\right] \\
&< F_{B(n,1/b)}(\lfloor m_0/\gamma \rfloor) + \sum_{n_{\mathrm{rs}}=\lfloor m_0/\gamma \rfloor+1}^{n} \mathbb{P}(N_{\mathrm{rs}} = n_{\mathrm{rs}})(1 - \delta + \frac{C_1}{\sqrt{n_{\mathrm{rs}}}}) \\
&\leq 1 - \delta + \frac{C_2}{\sqrt{n}} + C_1 \sum_{n_{\mathrm{rs}}=\lfloor m_0/\gamma \rfloor+1}^{n} \frac{1}{\sqrt{n_{\mathrm{rs}}}} \binom{n}{n_{\mathrm{rs}}} \frac{1}{b^{n_{\mathrm{rs}}}} (1 - \frac{1}{b})^{n - n_{\mathrm{rs}}} \\
&= 1 - \delta + \frac{C_2}{\sqrt{n}} + C_1 b \sum_{n_{\mathrm{rs}}=\lfloor m_0/\gamma \rfloor+1}^{n} \frac{1}{\sqrt{n_{\mathrm{rs}}}} \frac{n_{\mathrm{rs}}+1}{n+1} \binom{n+1}{n_{\mathrm{rs}}+1} \frac{1}{b^{n_{\mathrm{rs}}+1}} (1 - \frac{1}{b})^{n - n_{\mathrm{rs}}} \\
&\leq 1 - \delta + \frac{C_2}{\sqrt{n}} + \frac{C_1}{\sqrt{n}}.
\end{aligned}$$

As (18) always holds for suitable C if $n \leq \lfloor m_0/\gamma \rfloor$, we complete the proof.

Proof. Consider the probability

$$\mathbb{P}\left[\epsilon - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon | N_{\mathrm{rs}} = n_{\mathrm{rs}}, \mathcal{D}_1^{\mathrm{rs}}\right] = \mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}} \sim (P^{\pi_e})^m}\left[\epsilon - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon\right]$$

with $m = \lfloor \gamma n_{\mathrm{rs}} \rfloor$. Define $\delta_\Delta \coloneqq F_{\mathrm{B}(m,\epsilon-\Delta_\epsilon)}(k(m,\epsilon,\delta))$, then we have $k(m,\epsilon-\Delta_\epsilon,\delta_\Delta) = k(m,\epsilon,\delta)$ by the definition (8). Hence, for $\tilde{\tau} = \tau_{(m-k(m,\epsilon,\delta))}$, (19) yields

$$\mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}} \sim (P^{\pi_e})^m}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon - \Delta_\epsilon\right] < 1 - \delta_\Delta + \frac{C}{\sqrt{m}}.$$

Together with (17), it holds that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}} \sim (P^{\pi_e})^m}\left[\epsilon - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon\right] \\
=&1 - \mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}} \sim (P^{\pi_e})^m}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon - \Delta_\epsilon\right] - \mathbb{P}_{\mathcal{D}_2^{\mathrm{rs}} \sim (P^{\pi_e})^m}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) > \epsilon\right] \\
>&\delta_\Delta - \delta - \frac{C}{\sqrt{m}}.
\end{aligned}$$

We next show that $\delta_\Delta = F_{\mathrm{B}(m,\epsilon-\Delta_\epsilon)}(k(m,\epsilon,\delta))$ approaches 1 at an exponential rate. By treating $B(m,\epsilon)$ as the sum of m independent Bernoulli random variables with success probability ϵ, it follows from Hoeffding's inequality that

$$F_{\mathrm{B}(m,\epsilon)}(m\epsilon - mt) \leq \exp\{-2mt^2\}.$$

Choosing $t = \sqrt{\frac{\log \delta}{-2m}}$ leads to a lower bound for $k(m, \epsilon, \delta)$:

$$k(m, \epsilon, \delta) \geq (\epsilon - \sqrt{\frac{\log \delta}{-2m}})m. \tag{20}$$

Then, for sufficiently large m satisfying $\sqrt{\frac{\log(1/\delta)}{2m}} < \Delta_\epsilon$, we can obtain again from Hoeffding's inequality that

$$\delta_\Delta \geq 1 - \exp(-2m[\Delta_\epsilon - \sqrt{\frac{\log(1/\delta)}{2m}}]^2).$$

Therefore, after marginalizing,

$$\mathbb{P}\left[\epsilon - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon | N_{\mathrm{rs}} = n_{\mathrm{rs}}\right] > 1 - \delta - \frac{C}{\sqrt{n_{\mathrm{rs}}}}.$$

Finally, we obtain the lower bound of (11) in a manner similar to the proof of Lemma 2. The upper bound follows directly from Lemma 2. The proof is complete.

□

A.4 Proof of Theorem 3

This proof is based on the proof of Theorem 1 in [Sesia and Candès, 2020]. Similar to Assumption A4 in [Lei et al., 2018], the following consistency assumption is weaker than requiring L_2-convergences and can be achieved by many consistent estimators.

Assumption 3. *Denote by l the size of the training set $\mathcal{D}_1^{\mathrm{rs}}$ used to fit the conditional quantile functions $\hat{q}_{\epsilon_{\mathrm{lo}}}$ and $\hat{q}_{\epsilon_{\mathrm{up}}}$. For sufficiently large l, the following conditions hold:*

$$\mathbb{P}\left[\mathbb{E}_S\left[\left(\hat{q}_{\epsilon_{\mathrm{lo}}}(S_{n+1}) - q_{\epsilon_{\mathrm{lo}}}(S_{n+1})\right)^2\right] \leq \eta_l/2\right] \geq 1 - \rho_l/2,$$
$$\mathbb{P}\left[\mathbb{E}_S\left[\left(\hat{q}_{\epsilon_{\mathrm{up}}}(S_{n+1}) - q_{\epsilon_{\mathrm{up}}}(S_{n+1})\right)^2\right] \leq \eta_l/2\right] \geq 1 - \rho_l/2,$$

for some sequences $\eta_l = o(1)$ and $\rho_l = o(1)$, as $l \to \infty$. Here, $\mathbb{E}_S(\cdot)$ denotes the expectation w.r.t S_{n+1}, and the probabilities are taken over $\mathcal{D}_1^{\mathrm{rs}}$.

In addition, a regularity assumption is needed.

Assumption 4. *The probability density of the random variable*

$$T \coloneqq \max\{q_{\epsilon_{\mathrm{lo}}}(S) - R, R - q_{\epsilon_{\mathrm{up}}}(S)\},$$

where $(S, R) \sim P^{\pi_e}$, is bounded away from zero in a neighborhood of zero.

Proof. It suffices to show that, as $|\mathcal{D}^{\mathrm{rs}}| = n^{\mathrm{rs}} \to \infty$,

(i) $|\hat{q}_{\epsilon_{\mathrm{lo}}}(S_{n+1}) - q_{\epsilon_{\mathrm{lo}}}(S_{n+1})| = o_{\mathbb{P}}(1)$ and $|\hat{q}_{\epsilon_{\mathrm{up}}}(S_{n+1}) - q_{\epsilon_{\mathrm{up}}}(S_{n+1})| = o_{\mathbb{P}}(1)$;

(ii) $\tilde{\tau} = o_{\mathbb{P}}(1)$.

Here and in the subsequent part of the proof, the probabilities are taken over $\mathcal{D}^{\text{rs}}$ and S_{n+1}; the sizes of the training set $\mathcal{D}_1^{\text{rs}}$ and the calibration set $\mathcal{D}_2^{\text{rs}}$ are denoted by l and m, respectively.

(i) Define the random set

$$B_{l,\text{lo}} = \{s : |\hat{q}_{\epsilon_{\text{lo}}}(s) - q_{\epsilon_{\text{lo}}}(s)| \geq \eta_l^{1/3}\}, B_{l,\text{up}} = \{s : |\hat{q}_{\epsilon_{\text{up}}}(s) - q_{\epsilon_{\text{up}}}(s)| \geq \eta_l^{1/3}\},$$

and $B_l = B_{l,\text{lo}} \cup B_{l,\text{up}}$. We have by Markov's inequality that

$$\begin{aligned}
&\mathbb{P}(S_{n+1} \in B_l \mid \mathcal{D}_1^{\text{rs}}) \\
\leq &\mathbb{P}(S_{n+1} \in B_{l,\text{lo}} \mid \mathcal{D}_1^{\text{rs}}) + \mathbb{P}(S_{n+1} \in B_{l,\text{lo}} \mid \mathcal{D}_1^{\text{rs}}) \\
= &\mathbb{P}(|\hat{q}_{\epsilon_{\text{lo}}}(S_{n+1}) - q_{\epsilon_{\text{lo}}}(S_{n+1})| \geq \eta_l^{1/3} \mid \mathcal{D}_1^{\text{rs}}) + \mathbb{P}(|\hat{q}_{\epsilon_{\text{up}}}(S_{n+1}) - q_{\epsilon_{\text{up}}}(S_{n+1})| \geq \eta_l^{1/3} \mid \mathcal{D}_1^{\text{rs}}) \\
\leq &\eta_l^{-2/3}\mathbb{E}_S\left[\left(\hat{q}_{\epsilon_{\text{lo}}}(S_{n+1}) - q_{\epsilon_{\text{lo}}}(S_{n+1})\right)^2\right] + \eta_l^{-2/3}\mathbb{E}_S\left[\left(\hat{q}_{\epsilon_{\text{up}}}(S_{n+1}) - q_{\epsilon_{\text{up}}}(S_{n+1})\right)^2\right]. \\
\leq &\eta_l^{1/3}
\end{aligned}$$

with probability at least $1-\rho_l$, by Assumption 3. This implies $|\hat{q}_{\epsilon_{\text{lo}}}(S_{n+1}) - q_{\epsilon_{\text{lo}}}(S_{n+1})| = o_{\mathbb{P}}(1)$ and $|\hat{q}_{\epsilon_{\text{up}}}(S_{n+1}) - q_{\epsilon_{\text{up}}}(S_{n+1})| = o_{\mathbb{P}}(1)$.

(ii) Consider the following partition of the calibration set $\mathcal{D}_2^{\text{rs}}$:

$$\mathcal{D}_{2,a}^{\text{rs}} = \{(S_i, R_i) \in \mathcal{D}_2^{\text{rs}} : S_i \in B_l^c\},\ \mathcal{D}_{2,b}^{\text{rs}} = \{(S_i, R_i) \in \mathcal{D}_2^{\text{rs}} : S_i \in B_l\}.$$

Since B_l only depends on $\mathcal{D}_1^{\text{rs}}$, the size of $\mathcal{D}_{2,b}^{\text{rs}}$ conditional on $\mathcal{D}_1^{\text{rs}}$ can be bounded using Hoeffding's inequality as

$$\mathbb{P}(|\mathcal{D}_{2,b}^{\text{rs}}| \geq m\eta_l^{1/3} + t) \leq \mathbb{P}(\sum_{(S_i,R_i)\in\mathcal{D}_2^{\text{rs}}} \mathbb{1}(\{S_i \in B_l\}) \geq m\mathbb{P}(S_i \in B_l) + t) \leq \exp\left(-\frac{2t^2}{m}\right).$$

Choosing $t = C\sqrt{m \log m}$ leads to $|\mathcal{D}_{2,b}^{\text{rs}}| = o_{\mathbb{P}}(m) = o_{\mathbb{P}}(n^{\text{rs}})$.

Now, for each $(S_i, R_i) \in \mathcal{D}_2^{\text{rs}}$, define $T_i = \max\{q_{\epsilon_{\text{lo}}}(S_i) - R_i, R_i - q_{\epsilon_{\text{up}}}(S_i)\}$. By the definition (7) of τ_i, it can be easily derived that

$$|T_i - \tau_i| \leq \eta_l^{1/3}, \text{ for } i \text{ s.t. } (S_i, R_i) \in \mathcal{D}_{2,a}^{\text{rs}}. \tag{21}$$

Recall that $\tilde{\tau} = \tau_{(m-k(m,\epsilon,\delta))}$, we also define the k-th smallest T_i as $T_{(k)}$. In addition, when restricted to the dataset $\mathcal{D}_{2,a}^{\text{rs}}$, define $\tau_{(k)}^a$ and $T_{(k)}^a$ as the k-th smallest τ_i and T_i for i s.t. $(S_i, R_i) \in \mathcal{D}_{2,a}^{\text{rs}}$. As demonstrated in (20), a similar bound for $k(m, \epsilon, \delta)$ can be established as

$$(\epsilon - \sqrt{\frac{\log\delta}{-2m}})m \leq k(m,\epsilon,\delta) \leq m(\epsilon + \sqrt{\frac{\log(1-\delta)}{-2m}}). \tag{22}$$

For n^{rs} large enough, without loss of generality, we assume $|\mathcal{D}_{2,b}^{\text{rs}}| < m - k(m, \epsilon, \delta)$ because $|\mathcal{D}_{2,b}^{\text{rs}}| = o_{\mathbb{P}}(m)$. Then, it is straightforward to verify that

$$\tau^a_{(m-k(m,\epsilon,\delta)-|\mathcal{D}_{2,b}^{\text{rs}}|)} \leq \tau_{(m-k(m,\epsilon,\delta))} \leq \tau^a_{(m-k(m,\epsilon,\delta))}.$$

Together with (21), we have

$$T^a_{(m-k(m,\epsilon,\delta)-|\mathcal{D}^{\mathrm{rs}}_{2,b}|)} - \eta_l^{1/3} \leq \tilde{\tau} \leq T^a_{(m-k(m,\epsilon,\delta))} + \eta_l^{1/3},$$

which in turn yields

$$T_{(m-k(m,\epsilon,\delta)-|\mathcal{D}^{\mathrm{rs}}_{2,b}|)} - \eta_l^{1/3} \leq \tilde{\tau} \leq T_{(m-k(m,\epsilon,\delta)+|\mathcal{D}^{\mathrm{rs}}_{2,b}|)} + \eta_l^{1/3}.$$

Therefore, it suffices to prove that $T_{(m-k(m,\epsilon,\delta)-|\mathcal{D}^{\mathrm{rs}}_{2,b}|)}$ and $T_{(m-k(m,\epsilon,\delta)+|\mathcal{D}^{\mathrm{rs}}_{2,b}|)}$ is $o_{\mathbb{P}}(1)$. In fact, by Assumption 4, for any sufficiently small $\alpha > 0$, there exists $\epsilon_\alpha > 0$ s.t. $\mathbb{P}(T_i > \alpha) = \epsilon_\alpha < \epsilon$ for each T_i. Then, by Hoeffding's inequality and (22), for sufficiently large m,

$$\begin{aligned}\mathbb{P}(T_{(m-k(m,\epsilon,\delta)+|\mathcal{D}^{\mathrm{rs}}_{2,b}|)} \leq \alpha) &= F_{B(m,\epsilon_\alpha)}(k(m,\epsilon,\delta) - |\mathcal{D}^{\mathrm{rs}}_{2,b}|) \\ &\geq 1 - \exp\left(-2m[\frac{k(m,\epsilon,\delta) - |\mathcal{D}^{\mathrm{rs}}_{2,b}|}{m} - \epsilon_\alpha]^2\right).\end{aligned}$$

One can prove in a similar manner that $\mathbb{P}(T_{(m-k(m,\epsilon,\delta)+|\mathcal{D}^{\mathrm{rs}}_{2,b}|)} \geq -\alpha) \to 1$ as $m \to \infty$. Thus, we obtain $T_{(m-k(m,\epsilon,\delta)+|\mathcal{D}^{\mathrm{rs}}_{2,b}|)} = o_{\mathbb{P}}(1)$, and $T_{(m-k(m,\epsilon,\delta)-|\mathcal{D}^{\mathrm{rs}}_{2,b}|)} = o_{\mathbb{P}}(1)$ analogously. Finally, this proof is complete.

□

A.5 Proof of Theorem 4

Proof. For a realization of $\mathcal{D}_1$, the algorithm AL_p outputs a behavior policy estimator $\hat{\pi}_b$.[3] Then, it can be shown analogous to Proposition 1 that each sample in $\mathcal{D}^{\mathrm{rs}}_2$ independently follows the joint distribution

$$\hat{P}^{\pi_e}(\mathrm{d}s, \mathrm{d}r) \coloneqq \int_{\mathcal{A}} \frac{\hat{w}(s,a)}{\mathbb{E}_{p^{\pi_b}}[\hat{w}(S,A)]} P_S(\mathrm{d}s)\pi_b(\mathrm{d}a|s)P_R(\mathrm{d}r|s,a).$$

Therefore, $\hat{\pi}_b$ can be regarded as the true behavior policy for test data following $\hat{P}^{\pi_e}$. By using the same approach as in the proof of Theorem 1, we can obtain

$$\mathbb{P}[L_{\hat{P}^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \leq \epsilon \mid \mathcal{D}_1] \geq 1 - \delta. \tag{23}$$

Let $d_{\mathrm{TV}}(\hat{P}^{\pi_e}, P^{\pi_e})$ be the total variation distance between $\hat{P}^{\pi_e}$ and P^{π_e}. We have

$$\begin{aligned}d_{\mathrm{TV}}(\hat{P}^{\pi_e}, P^{\pi_e}) &= \frac{1}{2}\int_{\mathcal{S}}\int_{\mathcal{R}} \left|\hat{P}^{\pi_e}(\mathrm{d}s, \mathrm{d}r) - P^{\pi_e}(\mathrm{d}s, \mathrm{d}r)\right| \\ &\leq \frac{1}{2}\int_{\mathcal{S}}\int_{\mathcal{A}}\int_{\mathcal{R}} \left|\hat{w}(s,a)/\mathbb{E}_{p^{\pi_b}}[\hat{w}(S,A)] - w(s,a)\right| P_R(\mathrm{d}r|s,a)\pi_b(\mathrm{d}a|s)P_S(\mathrm{d}s) \\ &= \frac{1}{2}\mathbb{E}_{P^{\pi_b}}\left|\hat{w}(S,A)/\mathbb{E}_{p^{\pi_b}}[\hat{w}(S,A)] - w(S,A)\right|.\end{aligned}$$

[3]For simplicity, we also assume that AL_p introduces no exogenous randomness.

Denote by $\hat{E}$ the event $\left\{\frac{1}{2}\mathbb{E}_{P^{\pi_b}}|\hat{w}(S,A)/\mathbb{E}_{p^{\pi_b}}[\hat{w}(S,A)] - w(S,A)| \le \epsilon'\right\}$. By the fact that $\left|L_{\hat{P}^{\pi_e}}(\hat{C}_{\tilde{\tau}}) - L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}})\right| \le d_{\mathrm{TV}}(\hat{P}^{\pi_e}, P^{\pi_e})$, we then have by (13) and (23) that

$$\begin{aligned}\mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \le \epsilon + \epsilon'\right] &\ge \mathbb{P}\left[L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \le \epsilon + \epsilon', \hat{E}\right] \\ &\ge \mathbb{P}\left[L_{\hat{P}^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \le \epsilon, \hat{E}\right] \\ &\ge (1-\delta')(1-\delta).\end{aligned}$$

Thus, (14) is obtained. To prove (15), it follows similarly from Theorem 2 that

$$\mathbb{P}[\epsilon - \Delta_\epsilon < L_{\hat{P}^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \le \epsilon \mid \mathcal{D}_1] \ge 1 - \delta - \frac{C}{\sqrt{n}},$$

which, together with (13), yields

$$\begin{aligned}\mathbb{P}\left[\epsilon - \epsilon' - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \le \epsilon + \epsilon'\right] &\ge \mathbb{P}\left[\epsilon - \epsilon' - \Delta_\epsilon < L_{P^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \le \epsilon + \epsilon', E\right] \\ &\ge \mathbb{P}\left[\epsilon - \Delta_\epsilon < L_{\hat{P}^{\pi_e}}(\hat{C}_{\tilde{\tau}}) \le \epsilon, E\right] \\ &\ge (1-\delta')(1-\delta-\frac{C}{\sqrt{n}}).\end{aligned}$$

The proof is complete. □

B An MLE approach for estimating π_b

In this section, we establish the validity of Assumption 2 under an MLE approach and conduct a detailed analysis of the associated sample complexity. Specifically, we consider a policy class $\Pi = \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$ that satisfies the following assumptions:

Assumption 5.

- Π *is rich enough such that* $\pi_b \in \Pi$*;*
- *For all* $\pi \in \Pi$, $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\pi_e}{\pi}(\mathrm{d}a|s) \le B < \infty$.

We assume Π is discrete with size $|\Pi|$ for simplicity. Given the training set $\mathcal{D}_1$ with size n_1, our objective is to identify a policy within Π that maximizes the likelihood of fitting $\mathcal{D}_1$. This corresponds to solving the following MLE problem:

$$\hat{\pi}_b = \arg\max_{\pi\in\Pi} \sum_{(S_i,A_i,R_i)\in\mathcal{D}_1} \log p_\pi(A_i, S_i), \tag{24}$$

where $p_\pi(a,s)$ is the probability density or mass of π selecting action a at s. Leveraging an existing theoretical result for MLE, we derive the following sample complexity for the estimator (24) to satisfy condition (13).

Theorem 5. *Suppose the Assumption 5 holds. For the estimated weighted function* $\hat{w}(s,a) = \frac{\pi_e}{\hat{\pi}_b}(\mathrm{d}a|s)$ *with* $\hat{\pi}_b$ *given in (24), it holds that, with probability at least* $1-\delta'$*,*

$$\mathbb{E}_{P^{\pi_b}} |\hat{w}(S,A)/\mathbb{E}_{P^{\pi_b}}[\hat{w}(S,A)] - w(S,A)| \le 2\epsilon',$$

provided $n_1 \ge \frac{8B^2 \log(|\Pi|/\delta')}{\epsilon'^2}$*, for any* $(\epsilon', \delta') \in (0,1)^2$*.*

Proof. By Jensen's inequality and the fact that $\|p-q\|_{L_1} = 2d_{\mathrm{TV}}(p,q)$ for two distributions p and q, we have

$$\begin{aligned}
\mathbb{E}_{P^{\pi_b}} |\hat{w}(S,A) - w(S,A)| &= \int_{\mathcal{S}} \int_{\mathcal{A}} \left| \frac{\pi_e}{\hat{\pi}_b}(\mathrm{d}a|s) - \frac{\pi_e}{\pi_b}(\mathrm{d}a|s) \right| \pi_b(\mathrm{d}a|s) P_{\mathcal{S}}(\mathrm{d}s) \\
&= \int_{\mathcal{S}} \int_{\mathcal{A}} \frac{\pi_e}{\hat{\pi}_b}(\mathrm{d}a|s) \left| \pi_b(\mathrm{d}a|s) - \hat{\pi}_b(\mathrm{d}a|s) \right| P_{\mathcal{S}}(\mathrm{d}s) \\
&\le B \int_{\mathcal{S}} \int_{\mathcal{A}} \left| \pi_b(\mathrm{d}a|s) - \hat{\pi}_b(\mathrm{d}a|s) \right| P_{\mathcal{S}}(\mathrm{d}s) \\
&\le B \left(\int_{\mathcal{S}} \left| \int_{\mathcal{A}} |\pi_b(a|s) - \hat{\pi}_b(a|s)| \, \mathrm{d}a \right|^2 P_{\mathcal{S}}(\mathrm{d}s) \right)^{1/2} \\
&= 2B \left(\int_{\mathcal{S}} [d_{\mathrm{TV}}(\pi_b(\cdot|s), \hat{\pi}_b(\cdot|s))]^2 P_{\mathcal{S}}(\mathrm{d}s) \right)^{1/2}.
\end{aligned}$$

Together with Theorem 15.2 in [Agarwal et al., 2019], we obtain that, with probability at least $1-\delta'$,

$$\mathbb{E}_{P^{\pi_b}} |\hat{w}(S,A) - w(S,A)| \le 2B \sqrt{\frac{2\log(|\Pi|/\delta')}{n_1}} \le \epsilon',$$

which implies $|\mathbb{E}_{P^{\pi_b}}[\hat{w}(S,A)] - 1| \le \epsilon'$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{P^{\pi_b}} |\hat{w}(S,A)/\mathbb{E}_{P^{\pi_b}}[\hat{w}(S,A)] - w(S,A)| \\
\le &\mathbb{E}_{P^{\pi_b}} |\hat{w}(S,A) - w(S,A)| + \left| 1 - \frac{1}{\mathbb{E}_{P^{\pi_b}}[\hat{w}(S,A)]} \right| \mathbb{E}_{P^{\pi_b}}[\hat{w}(S,A)] \\
\le &\epsilon' + |\mathbb{E}_{P^{\pi_b}}[\hat{w}(S,A)] - 1| \\
\le &2\epsilon'.
\end{aligned}$$

we complete the proof. □