Supplementary Materials for Conformal Prediction Beyond the Horizon: Distribution-Free Inference for Policy Evaluation

Feichen Gan, Youcun Lu, Yingying Zhang* and Yukun Liu

KLATASDS - MOE, School of Statistics, East China Normal University

A Preliminaries

We impose the following standard assumptions in RL. In our notation, $\mathcal{P}$ denotes a probability distribution.

ASSUMPTION **1** (Markov Property)**.** *The decision process satisfies the Markov property: the next state and reward depend only on the current state and action. Formally, for all t,*

$$\mathcal{P}(S_{t+1}, R_t \mid A_t, S_t, R_{t-1}, A_{t-1}, S_{t-1}, \ldots, S_0) = \mathcal{P}(S_{t+1}, R_t \mid S_t, A_t).$$

ASSUMPTION **2** (Time-homogeneity)**.** *The distribution of the transition and reward remains stationary over time. Specifically, for all t, the joint distribution of the next state and reward given the current state and action satisfies*

$$\mathcal{P}(S_{t+1}, R_t \mid S_t, A_t) = \mathcal{P}(S_t, R_{t-1} \mid S_{t-1}, A_{t-1}).$$

ASSUMPTION **3** (Stationary Policy)**.** *The policy is stationary and Markovian: the action at each time step depends only on the current state and not on the full history. Formally, for all t,*

$$\pi_t(A_t \mid S_t, R_{t-1}, A_{t-1}, S_{t-1}, \ldots, S_0) = \pi(A_t \mid S_t).$$

Before proceeding with theoretical analysis, we introduce the distributional Bellman operator and several related results. We use $\eta^\pi(s)$ to denote the distribution of the return starting from the initial state s following policy π, that is,

$$\eta^\pi(s) := \mathcal{P}^\pi(G \,|\, S_0 = s) := \mathcal{P}^\pi(\sum_{t=0}^{\infty} \gamma^t R_t \,|\, S_0 = s).$$

We define the **distributional Bellman operator** $\mathcal{T}^\pi$ as the following transformation:

$$(\mathcal{T}^\pi \eta^\pi)(s) = \mathcal{P}^\pi(R + \gamma G^\pi(S') \,|\, s)$$

where the transition (s, R, S') is generated by sampling an action from π, observing the reward R, and transitioning to the next state S', and $G^\pi(S') \sim \eta^\pi(S')$.

Under the time-homogeneity assumption, η^π satisfies the fixed-point condition:

$$\eta^\pi(s) = (\mathcal{T}^\pi \eta^\pi)(s), \quad \forall s \in \mathcal{S}.$$

A key property of the distributional Bellman operator $\mathcal{T}^\pi$ is that it is a γ-contraction w.r.t. the Wasserstein distance, stated in Proposition 3. The p-Wasserstein distance between two measures μ and ν on the real space $\mathbb{R}$ is defined as

$$W_p(\mu, \nu) := \inf_{\kappa \in \Gamma(\mu,\nu)} \left(\int_{\mathbb{R}\times\mathbb{R}} |x - y|^p \kappa(dx, dy) \right)^{1/p}$$

where $\Gamma(\mu, \nu)$ is the set of all couplings with marginals μ and ν. We begin by presenting some fundamental results on the Wasserstein distance.

[1*]Corresponding author: yyzhang@fem.ecnu.edu.cn.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

PROPOSITION **1** (Duality Formula for 1-Wasserstein Distance [10]). *For any measures μ and ν,*

$$W_1(\mu,\nu) = \sup_{\psi:\|\psi\|_{\mathrm{Lip}}\le 1}\left\{\int \psi\, d\mu - \int \psi\, d\nu\right\},$$

where "$\|\psi\|_{\mathrm{Lip}} \le 1$" means that ψ is a 1-Lipschitz function.

PROPOSITION **2.** *Suppose $\|f\|_{\mathrm{Lip}} \le 1$ and b_f is an operator on measure such that $b_f(\mu)(A) = \mu(f^{-1}(A))$ for any measure μ and Borel set A. Then b_f is a contraction under 1-Wasserstein distance, i.e., $W_1(b_f(\mu), b_f(\nu)) \le W_1(\mu,\nu)$ for all measures μ, ν.*

Proof. For any 1-Lipschitz function ψ, the composition $\psi \circ f$ is also 1-Lipschitz, since the composition of 1-Lipschitz functions preserves the Lipschitz constant. By Proposition 1, we have, for any measures μ and ν,

$$\begin{aligned} W_1\left(b_f(\mu), b_f(\nu)\right) &= \sup_{\psi:\|\psi\|_{\mathrm{Lip}}\le 1}\left\{\int \psi\circ f\, d\mu - \int \psi\circ f\, d\nu\right\} \\ &\le \sup_{\psi:\|\tilde{\psi}\|_{\mathrm{Lip}}\le 1}\left\{\int \tilde{\psi}\, d\mu - \int \tilde{\psi}\, d\nu\right\} \\ &= W_1(\mu,\nu), \end{aligned}$$

where the first equation follows from the change-of-variables formula for measures. □

Applying Proposition 2, for any real random variables X and Y with laws $\mathcal{P}_X$ and $\mathcal{P}_Y$, since $f(x) = |x-a|$ for any $a \in \mathbb{R}$ is 1-Lipschitz continuous, we have $W_1(\mathcal{P}_{|X-a|}, \mathcal{P}_{|Y-a|}) \le W_1(\mathcal{P}_X, \mathcal{P}_Y)$.

We now state the contraction property of the distributional Bellman operator $\mathcal{T}^\pi$. Let $\mathscr{P}$ be the set of all probability distributions over $\mathbb{R}$. Please note that the conditional return distribution given a state ($s \in \mathcal{S}$) is a distribution that is indexed by the state s. That is, $\eta^\pi(\cdot) \in \mathscr{P}^{\mathcal{S}}$, and $\mathscr{P}^{\mathcal{S}}$ contains all possible conditional return distributions. We define the Wasserstein distance of two conditional distributions $\mu(\cdot), \nu(\cdot) \in \mathscr{P}^{\mathcal{S}}$ as $\bar{W}_p(\mu(\cdot),\nu(\cdot)) := \sup_{s\in\mathcal{S}} W_p(\mu(s),\nu(s))$.

PROPOSITION **3.** *[[1], Proposition 4.15] The distributional Bellman operator is a γ-contraction on $\mathscr{P}^{\mathcal{S}}$ w.r.t. the supreme p-Wasserstein metric for $p \in [1,\infty)$. That is, for any $\eta, \eta' \in \mathscr{P}^{\mathcal{S}}$, we have $\bar{W}_p(\mathcal{T}^\pi\eta, \mathcal{T}^\pi\eta') \le \gamma \bar{W}_p(\eta,\eta')$.*

We denote the learned DRL model in the proposed prediction procedure by $\widehat{\eta}^\pi(s)$. It is clear that given S_t, the one-step pseudo-return $\widetilde{G}^{(1)}(S_t) = R_t + \gamma\widetilde{G}^\pi(S_{t+1})$ with $\widetilde{G}^\pi(S_{t+1}) \sim \widehat{\eta}^\pi(S_{t+1})$, follows the distribution $(\mathcal{T}^\pi\widehat{\eta}^\pi)(S_t)$. The following proposition shows that a similar conclusion also holds when the step width is k. That is, the k-step pseudo-return starting from S_t follows $((\mathcal{T}^\pi)^k\widehat{\eta}^\pi)(S_t)$.

PROPOSITION **4** ([1], Lemma 4.33). *Let $\eta \in \mathscr{P}^{\mathcal{S}}$, and let G be an instantiation of η. For $s \in \mathcal{S}$, if $(S_t, A_t, R_t)_{t\ge 0}$ is a random trajectory with initial state $S_0 = s$ and generated by following π, independent of G, then $\sum_{t=0}^{k-1}\gamma_t R_t + \gamma^k G(S_k)$ is an instantiation of $((\mathcal{T}^\pi)^k\eta)(s)$.*

Proposition 4 allows us to investigate the k-step pseudo-return. As discussed in the main paper, we measure the coverage gap using the distributional distance between the estimated return distribution and the true return distribution. Unlike traditional approaches that rely on total variation distance, we adopt the Wasserstein distance, motivated by the insights in [11]. A key intermediary that links the coverage error and the Wasserstein distance is the Kolmogorov distance, which is defined as follows.

DEFINITION **1** (Kolmogorov Distance). *F_μ and F_ν are the CDFs of probability measures μ and ν on $\mathbb{R}$, respectively. Kolmogorov distance between μ and ν is given by*

$$K(\mu,\nu) = \sup_{x\in\mathbb{R}} |F_\mu(x) - F_\nu(x)|.$$

LEMMA 1 ([7]). *If a probability measure μ in space $\mathbb{R}$ has Lebesgue density bounded by L, then for any probability measure ν, $K(\mu,\nu) \le \sqrt{2LW_1(\mu,\nu)}$.*

B Proof of Theorem 1

We now present the proof of the main theorem for the proposed PIs in the on-policy evaluation setting.

Proof of Theorem 1. Since $\widehat{C}^{\mathrm{on}}_{N,\alpha}(S_{\mathrm{test}})$ combines B intervals following [12, 9], it suffices to prove the validity of each single CP interval. With some abuse of notation, we denote the single CP interval at target coverage level $1-\alpha$ as $\widehat{C}^{\mathrm{on}}_{N,\alpha}(S_{\mathrm{test}})$.

We first consider the case where data splitting is performed in a trajectory-wise manner, and let n denote the number of trajectories in the calibration set $\mathcal{D}_{\mathrm{cal}}$. We index the trajectories in the calibration dataset $\mathcal{D}_{\mathrm{cal}}$ as $\{1, 2, \ldots, n\}$. Please note that, with a slight abuse of notation, n here denotes the number of trajectories, which differs from its definition in the main paper. In the main paper, n refers to the cardinality of the calibration set $\mathcal{D}_{\mathrm{cal}}$, where data are stored as tuples rather than trajectories.

Note that the step-width in constructing the pseudo-return is k. For a state variable S_{it} in the data, the corresponding pseudo-return is constructed as

$$\widetilde{G}^{(k)}(S_{it}) := \sum_{h=0}^{k-1} \gamma^h R_{i,t+h} + \gamma^k \widetilde{G}^{\pi}(S_{i,t+k}), \quad \widetilde{G}^{\pi}(S_{i,t+k}) \sim \widehat{\eta}^{\pi}(S_{i,t+k}).$$

Hereafter, for notational simplicity, we denote $\widetilde{G}^{(k)}_{it} := \widetilde{G}^{(k)}(S_{it})$. By Proposition 4,

$$\widetilde{G}^{(k)}_{it} \sim ((\mathcal{T}^{\pi})^k \widehat{\eta}^{\pi})(S_{it}).$$

Given all the data $\mathcal{D}$, the calibration set $\widetilde{\mathcal{D}}_{\mathrm{cal}}$, using experience replay and weighted subsampling, is a set of samples drawn from the distribution:

$$\widehat{F}_n(s, g) := \sum_{t=0}^{T-k} \sum_{i=1}^{n} \frac{\widehat{w}_{\mathrm{on}}(S_{it})}{\sum_{t=0}^{T-k} \sum_{j=1}^{n} \widehat{w}_{\mathrm{on}}(S_{jt})} I\{S_{it} \le s, \widetilde{G}^{(k)}_{it} \le g\}.$$

Main idea. The proof proceeds by successively isolating the effects of the two estimation errors: the approximation of $\eta^{\pi}(s)$ and the estimation of the weighting function. For notational simplicity, we abbreviate the return $G^{\pi}(S_{\mathrm{test}})$ on the test data as G_{test}.

We begin by noting that the true test point is drawn from

$$(S_{\mathrm{test}}, G_{\mathrm{test}}) \sim \mathcal{P}_{S_0} \times ((\mathcal{T}^{\pi})^k \eta^{\pi})(S_0),$$

where S_0 is the random initial state with marginal distribution $\mathcal{P}_{S_0}$. To quantify the error induced by approximating $\eta^{\pi}(s)$, we introduce an intermediate test point

$$(S_{\mathrm{test}}, \widetilde{G}_{\mathrm{test}}) \sim \mathcal{P}_{S_0} \times ((\mathcal{T}^{\pi})^k \widehat{\eta}^{\pi})(S_0),$$

which shares the same state distribution as the true test point but replaces η^{π} with its estimator $\widehat{\eta}^{\pi}$ (see (2) of this proof for details).

Next, to analyze the additional error due to weight estimation, we define another artificial test point

$$(\widehat{S}_{\mathrm{test}}, \widehat{G}_{\mathrm{test}}) \sim \widehat{F}_n(s, g),$$

which differs from $(S_{\mathrm{test}}, \widetilde{G}_{\mathrm{test}})$ only in the state distribution (see (3) of this proof for details).

Finally, conditional on $\mathcal{D}$, $(\widehat{S}_{\mathrm{test}}, \widehat{G}_{\mathrm{test}})$ is exchangeable with $\widetilde{\mathcal{D}}_{\mathrm{cal}}$. Hence, the standard conformal prediction argument applies, establishing the conditional coverage property in Eq. (S.1).

Given these new test points, we can bound the coverage probability of $G_{\text{test}} := G^{\pi}(S_{\text{test}})$ as

$$\begin{aligned}
&\Pr\left(G_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(S_{\text{test}})\right) \geq \Pr\left(\widehat{G}_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(\widehat{S}_{\text{test}})\right) \\
&\quad - \left|\Pr\left(\widehat{G}_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(\widehat{S}_{\text{test}})\right) - \Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(S_{\text{test}})\right)\right| \\
&\quad - \left|\Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(S_{\text{test}})\right) - \Pr\left(G_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(S_{\text{test}})\right)\right| \\
&:= M_1 - M_2 - M_3.
\end{aligned}$$

We now analyze M_1, M_2 and M_3 individually.

(1) Given $\mathcal{D}$, $(\widehat{S}_{\text{test}}, \widehat{G}_{\text{test}})$ is exchangeable with $\widetilde{\mathcal{D}}_{cal}$. Then, existing conclusions about coverage rate in SCP [4] gives

$$\Pr\left(\widehat{G}_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(\widehat{S}_{\text{test}}) \,|\, \mathcal{D}\right) \geq 1 - \alpha. \tag{S.1}$$

Taking expectation for the above inequality gives

$$M_1 \geq 1 - \alpha. \tag{S.2}$$

(2) Recall that $\widehat{C}^{\text{on}}_{N,\alpha}(S_{\text{test}}) = \widehat{v}^{\pi}(S_{\text{test}}) \pm \widehat{q}_{1-\alpha}$. By Propositions 2-4 and Lemma 1,

$$\begin{aligned}
&\left|\Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}_{N,\alpha}(S_{\text{test}}) \,|\, \mathcal{D}_{\text{tr}}, \widetilde{\mathcal{D}}_{\text{cal}}, S_{\text{test}}\right) - \Pr\left(G_{\text{test}} \in \widehat{C}_{N,\alpha}(S_{\text{test}}) \,|\, \mathcal{D}_{\text{tr}}, \widetilde{\mathcal{D}}_{\text{cal}}, S_{\text{test}}\right)\right| \\
= &\left|F_{|\widetilde{G}_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})|}(\widehat{q}_{1-\alpha}) - F_{|G_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})|}(\widehat{q}_{1-\alpha})\right| \\
\leq &K\left(\mathcal{P}_{|\widetilde{G}_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})|}, \mathcal{P}_{|G_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})|}\right) \quad \text{by Definition 1,} \\
\leq &\sqrt{2LW_1\left(\mathcal{P}_{|\widetilde{G}_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})|}, \mathcal{P}_{|G_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})|}\right)} \quad \text{by Lemma 1,} \\
\leq &\sqrt{2LW_1\left(\mathcal{P}_{\widetilde{G}_{\text{test}}}, \mathcal{P}_{G_{\text{test}}}\right)} \quad \text{by Proposition 2,} \\
\leq &\sqrt{2L\bar{W}_1\left((\mathcal{T}^{\pi})^k \widehat{\eta}^{\pi}, (\mathcal{T}^{\pi})^k \eta^{\pi}\right)} \quad \text{by Proposition 4,} \\
\leq &\sqrt{2L\gamma^k \bar{W}_1(\widehat{\eta}^{\pi}, \eta^{\pi})} \quad \text{by Proposition 3.}
\end{aligned}$$

Since $f(x) = \sqrt{x}$ is a concave function, taking expectations on both sides of the inequality and applying Jensen's inequality yields:

$$\begin{aligned}
M_3 &= \left|\mathbb{E}\left[\Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}_{N,\alpha}(S_{\text{test}}) \,\middle|\, \mathcal{D}_{\text{tr}}, \widetilde{\mathcal{D}}_{\text{cal}}, S_{\text{test}}\right) - \Pr\left(G_{\text{test}} \in \widehat{C}_{N,\alpha}(S_{\text{test}}) \,\middle|\, \mathcal{D}_{\text{tr}}, \widetilde{\mathcal{D}}_{\text{cal}}, S_{\text{test}}\right)\right]\right| \\
&\leq \mathbb{E}\left|\Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}_{N,\alpha}(S_{\text{test}}) \,\middle|\, \mathcal{D}_{\text{tr}}, \widetilde{\mathcal{D}}_{\text{cal}}, S_{\text{test}}\right) - \Pr\left(G_{\text{test}} \in \widehat{C}_{N,\alpha}(S_{\text{test}}) \,\middle|\, \mathcal{D}_{\text{tr}}, \widetilde{\mathcal{D}}_{\text{cal}}, S_{\text{test}}\right)\right| \\
&\leq \mathbb{E}\left[\sqrt{2L\gamma^k \bar{W}_1\left(\widehat{\eta}^{\pi}, \eta^{\pi}\right)}\right] \leq \sqrt{2L\gamma^k \,\mathbb{E}\left[\bar{W}_1\left(\widehat{\eta}^{\pi}, \eta^{\pi}\right)\right]} \quad \text{by Jensen's inequality.}
\end{aligned} \tag{S.3}$$

(3) Let $\mathcal{P}_t(s, g)$ denote the distribution of $(S_t, \widetilde{G}_t^{(k)})$ conditioned on $\mathcal{D}_{\text{tr}}$. While the marginal distribution of S_t may vary across time steps, the conditional distribution of $\widetilde{G}_t^{(k)} \mid S_t$ remains time-homogeneous. Now we analyze M_2 and first define $M_2(\mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}})$ as follows.

$$\begin{aligned}
M_2(\mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) &:= \left|\Pr\left(\widehat{G}_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(\widehat{S}_{\text{test}}) \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right) - \Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}^{\text{on}}_{N,\alpha}(S_{\text{test}}) \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\right| \\
&= \Bigg|\sum_{t=0}^{T-k}\sum_{i=1}^{n} \frac{\widehat{w}_{\text{on}}(S_{it})}{\sum_{t=0}^{T-k}\sum_{j=1}^{n}\widehat{w}_{\text{on}}(S_{jt})} I\left\{|\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| \leq \widehat{q}_{1-\alpha}\right\} \\
&\qquad - \Pr\left(|\widetilde{G}_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})| \leq \widehat{q}_{1-\alpha} \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\Bigg| \leq M_{21} + M_{22},
\end{aligned}$$

where

$$M_{21} := \sup_{x\in\mathbb{R}} \left| \sum_{t=0}^{T-k} \sum_{i=1}^{n} \frac{\widehat{w}_{\mathrm{on}}(S_{it})}{\sum_{t=0}^{T-k}\sum_{j=1}^{n} \widehat{w}_{\mathrm{on}}(S_{jt})} I\left\{ |\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| \le x \right\} - B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right|,$$

$$M_{22} := \left| B(\widehat{q}_{1-\alpha} \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) - \Pr\left(|\widetilde{G}_{\mathrm{test}} - \widehat{v}^{\pi}(S_{\mathrm{test}})| \le \widehat{q}_{1-\alpha} \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}} \right) \right|, \quad \text{where}$$

$$B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) := \frac{1}{T-k+1} \sum_{t=0}^{T-k} \int \widehat{w}_{\mathrm{on}}(s) I\{|g - \widehat{v}^{\pi}(s)| \le x\} \, d\mathcal{P}_t(s,g).$$

(3.1) To analyze M_{21}, we first define the normalization constant for weights as

$$W_n = \frac{1}{n(T-k+1)} \sum_{t=0}^{T-k} \sum_{i=1}^{n} \widehat{w}_{\mathrm{on}}(S_{it}).$$

Thus the first term in M_{21} becomes

$$\frac{1}{W_n} \frac{1}{n(T-k+1)} \sum_{t=0}^{T-k} \sum_{i=1}^{n} \widehat{w}_{\mathrm{on}}(S_{it}) I\left\{ |\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| \le x \right\} := \frac{1}{W_n} B_{\mathrm{emp}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}),$$

where $B_{\mathrm{emp}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}})$ is an empirical version of $B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}})$. By a simple algebraic calculation, we have

$$M_{21} \le \frac{1}{W_n} \sup_{x\in\mathbb{R}} \left| B_{\mathrm{emp}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) - B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right| + \left(\frac{1}{W_n} - 1 \right) \sup_{x\in\mathbb{R}} \left| B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right|.$$

Since $\frac{1}{T-k+1} \sum_{t=0}^{T-k} \mathbb{E}[\widehat{w}_{\mathrm{on}}(S_t) \mid \mathcal{D}_{\mathrm{tr}}] = 1$, by law of large numbers,

$$\lim_{n\to\infty} W_n = \frac{1}{T-k+1} \sum_{t=0}^{T-k} \mathbb{E}[\widehat{w}_{\mathrm{on}}(S_t) \mid \mathcal{D}_{\mathrm{tr}}] = 1. \tag{S.4}$$

Hence, for sufficiently large n, $W_n \ge 1/2$ and

$$M_{21} \le \underbrace{2 \sup_{x\in\mathbb{R}} \left| B_{\mathrm{emp}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) - B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right|}_{E} + \underbrace{\left| \frac{1}{W_n} - 1 \right| \sup_{x\in\mathbb{R}} \left| B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right|}_{F}.$$

(3.1.1) For E, since $\mathbb{E}[\widehat{w}_{\mathrm{on}}(S_{it}) \mid \mathcal{D}_{\mathrm{tr}}] < \infty$ for $0 \le t \le T-k$, the function class $\{\widehat{w}_{\mathrm{on}}(s) I\{|g - \widehat{v}^{\pi}(s)| \le x\} : x \in \mathbb{R}\}$ is $\{\mathcal{P}_t(s,g) : 0 \le t \le T-k\}$-Glivenko-Cantelli. Therefore, for all $0 \le t \le T-k$,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \widehat{w}_{\mathrm{on}}(S_{it}) I\left\{ |\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| \le x \right\} - \int \widehat{w}_{\mathrm{on}}(s) I\left\{ |g - \widehat{v}^{\pi}(s)| \le x \right\} \, d\mathcal{P}_t(s,g) \right| = 0.$$

Averaging over t gives

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| B_{\mathrm{emp}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) - B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right| = 0. \tag{S.5}$$

(3.1.2) For F, we have $\lim_{n\to\infty} (1/W_n - 1) = 0$ by Eq.(S.4) and

$$\sup_{x\in\mathbb{R}} \left| B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right| \le \frac{1}{T-k+1} \sum_{t=0}^{T-k} \int \widehat{w}_{\mathrm{on}}(s) \, d\mathcal{P}_t(s,g) = \frac{1}{T-k+1} \sum_{t=0}^{T-k} \mathbb{E}\left[\widehat{w}_{\mathrm{on}}(S_t) \mid \mathcal{D}_{\mathrm{tr}} \right] = 1,$$

by Eq.(S.4). Then combining (S.5), we conclude that

$$\lim_{n\to\infty} M_{21} = 0. \tag{S.6}$$

(3.2) Bound on M_{22}. Recall that

$$M_{22} := \left| B(\widehat{q}_{1-\alpha} \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) - \Pr\left(|\widetilde{G}_{\mathrm{test}} - \widehat{v}^{\pi}(S_{\mathrm{test}})| \le \widehat{q}_{1-\alpha} \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}} \right) \right|, \quad \text{where}$$

$$B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) := \frac{1}{T-k+1} \sum_{t=0}^{T-k} \int \widehat{w}_{\mathrm{on}}(s) I\{|g - \widehat{v}^{\pi}(s)| \le x\} \, d\mathcal{P}_t(s,g),$$

where $\mathcal{P}_t(s,g)$ denotes the conditional distribution of $(S_t, \widetilde{G}_t^{(k)})$ given the training data $\mathcal{D}_{\text{tr}}$.

Define a new probability measure

$$\frac{1}{T-k+1}\sum_{t=0}^{T-k} \widehat{w}_{\text{on}}(s)\, d\mathcal{P}_t(s,g),$$

and let $(\widetilde{S}, \widetilde{G})$ be drawn from this measure. Then M_{22} can be equivalently written as

$$M_{22} = \left| \Pr\left(|\widetilde{G} - \widehat{v}^{\pi}(\widetilde{S})| \le \widehat{q}_{1-\alpha} \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right) - \Pr\left(|\widetilde{G}_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})| \le \widehat{q}_{1-\alpha} \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\right|.$$

Since the conditional distributions $\widetilde{G} \mid \widetilde{S}$ and $\widetilde{G}_{\text{test}} \mid S_{\text{test}}$ are identical, by Eq. (A.9) in [5], we have

$$M_{22} \le d_{TV}(\mathcal{P}_{\widetilde{S}}, \mathcal{P}_{S_{\text{test}}}),$$

where d_{TV} denotes the total variation distance.

Denote the marginal distribution of $\mathcal{P}_t(s,g)$ as $\mathcal{P}_t(s)$ and define the calibration marginal $\mathcal{P}_{\text{cal}}(s) = \frac{1}{T-k+1}\sum_{t=0}^{T-k} \mathcal{P}_t(s)$. Then $\widetilde{S} \sim \widehat{w}_{\text{on}}(s)\mathcal{P}_{\text{cal}}(s)$ and $S_{\text{test}} \sim w_{\text{on}}(s)\mathcal{P}_{\text{cal}}(s)$. It follows that

$$\begin{aligned} M_{22} &\le \frac{1}{2}\int \left|\widehat{w}_{\text{on}}(s) - w_{\text{on}}(s)\right| d\mathcal{P}_{\text{cal}}(s) \\ &= \frac{1}{2(T-k+1)}\sum_{t=0}^{T-k} \mathbb{E}\Big[\left|\widehat{w}_{\text{on}}(S_t) - w_{\text{on}}(S_t)\right| \mid \mathcal{D}_{\text{tr}}\Big], \end{aligned} \tag{S.7}$$

where the last equality follows directly from the definition of $\mathcal{P}_{\text{cal}}(s)$.

The desired result in Theorem 1 follows from (S.2) - (S.7).

Extension. We now extend the above arguments to the setting where data splitting is performed at the tuple level—that is, on tuples of the form $(S_{it}, A_{it}, R_{it}, \ldots, S_{i,t+k})$, for $1 \le i \le N$ and $0 \le t \le T-k$. Let n denote the number of tuples in $\mathcal{D}_{\text{cal}}$, and let n_t be the number of t-stage tuples included. Then it holds that $\sum_{t=0}^{T-k} n_t = n$. We index the data points of the t-th stage separately as $\{1, 2, \ldots, n_t\}$ for notational simplicity. Given all data $\mathcal{D}$, $\widetilde{\mathcal{D}}_{\text{cal}}$ is a set of sample drawn from

$$\widehat{F}_n^*(s,g) := \sum_{t=0}^{T-k}\sum_{i=1}^{n_t} \frac{\widehat{w}_{\text{on}}(S_{it})}{\sum_{t=0}^{T-k}\sum_{j=1}^{n_t}\widehat{w}_{\text{on}}(S_{jt})} I\{S_{it} \le s, \widetilde{G}_{it}^{(k)} \le g\}.$$

Similarly we consider three new points

$$(\widehat{S}_{\text{test}}^*, \widehat{G}_{\text{test}}^*) \sim \widehat{F}_n^*(s,g), \quad (S_{\text{test}}, \widetilde{G}_{\text{test}}) \sim \mathcal{P}_{S_0} \times ((\mathcal{T}^{\pi})^k \widehat{\eta}^{\pi})(S_0), \quad (S_{\text{test}}, G_{\text{test}}) \sim \mathcal{P}_{S_0} \times \eta^{\pi}(S_0).$$

Then the coverage probability satisfies:

$$\begin{aligned} &\Pr\left(G_{\text{test}} \in \widehat{C}_{N,\alpha}^{\text{on}}(S_{\text{test}})\right) \ge \Pr\left(\widehat{G}_{\text{test}}^* \in \widehat{C}_{N,\alpha}^{\text{on}}(\widehat{S}_{\text{test}}^*)\right) \\ &\quad - \left|\Pr\left(\widehat{G}_{\text{test}}^* \in \widehat{C}_{N,\alpha}^{\text{on}}(\widehat{S}_{\text{test}}^*)\right) - \Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}_{N,\alpha}^{\text{on}}(S_{\text{test}})\right)\right| \\ &\quad - \left|\Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}_{N,\alpha}^{\text{on}}(S_{\text{test}})\right) - \Pr\left(G_{\text{test}} \in \widehat{C}_{N,\alpha}^{\text{on}}(S_{\text{test}})\right)\right| \\ &:= M_1^* - M_2^* - M_3. \end{aligned}$$

The analysis of M_1^* mirrors that of M_1, and the treatment of M_3 remains unchanged from the previous case. We now focus on the detailed analysis of M_2^*. Similarly we define $M_2^*(\mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}})$ as follows:

$$\begin{aligned} M_2^*(\mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) &:= \left|\Pr\left(\widehat{G}_{\text{test}}^* \in \widehat{C}_{N,\alpha}^{\text{on}}(\widehat{S}_{\text{test}}^*) \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right) - \Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}_{N,\alpha}^{\text{on}}(S_{\text{test}}) \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\right| \\ &= \left|\sum_{t=0}^{T-k}\sum_{i=1}^{n_t} \frac{\widehat{w}_{\text{on}}(S_{it})}{\sum_{t=0}^{T-k}\sum_{j=1}^{n_t}\widehat{w}_{\text{on}}(S_{jt})} I\left\{|\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| \le \hat{q}_{1-\alpha}\right\}\right. \\ &\qquad \left. - \Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}_{N,\alpha}^{\text{on}}(S_{\text{test}}) \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\right| \le M_{21}^* + M_{22} \end{aligned}$$

where

$$M_{21}^* := \sup_{x\in\mathbb{R}} \left| \sum_{t=0}^{T-k} \sum_{i=1}^{n_t} \frac{\widehat{w}_{\mathrm{on}}(S_{it})}{\sum_{t=0}^{T-k}\sum_{j=1}^{n_t} \widehat{w}_{\mathrm{on}}(S_{jt})} I\left\{ |\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| \le x \right\} - B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right|.$$

Then, we introduce an intermediate value for each time point t:

$$B_{\mathrm{emp}}(x \mid t, \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) := \frac{1}{n_t} \sum_{i=1}^{n_t} \widehat{w}_{\mathrm{on}}(S_{it}) I\left\{ |\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| \le x \right\},$$

which is an empirical version of $B(x \mid t, \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}})$ defined similarly:

$$B(x \mid t, \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) := \int \widehat{w}_{\mathrm{on}}(s) I\{|g - \widehat{v}^{\pi}(s)| \le x\} \, d\mathcal{P}_t(s, g).$$

Let n_t denote the number of tuples at time step t, for $0 \le t \le T-k$. The vector $(n_0, n_1, \ldots, n_{T-k})$ follows a multinomial distribution with total count n and uniform probabilities over the $T-k+1$ time steps:

$$(n_0, n_1, \ldots, n_{T-k}) \sim \text{Multinomial}\left(n; \left\{\tfrac{1}{T-k+1}, \ldots, \tfrac{1}{T-k+1}\right\}\right).$$

As $n \to \infty$, it follows that $n_t \to \infty$ for all t. Applying the same argument as in Equation (S.5), we obtain

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \frac{1}{T-k+1} \left\{ B_{\mathrm{emp}}(x \mid t, \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) - B(x \mid t, \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right\} \right| = 0. \tag{S.8}$$

Define the new normalization constant for weights as

$$W_n^* = \frac{1}{n} \sum_{t=0}^{T-k} \sum_{i=1}^{n_t} \widehat{w}_{\mathrm{on}}(S_{it}).$$

Since $\lim_{n\to\infty} n_t/n = 1/(T-k+1)$, it follows from law of large numbers that

$$\lim_{n\to\infty} W_n^* = \lim_{n\to\infty} \sum_{t=0}^{T-k} \frac{n_t}{n} \cdot \frac{1}{n_t} \sum_{i=1}^{n_t} \widehat{w}_{\mathrm{on}}(S_{it}) = \frac{1}{T-k+1} \sum_{t=0}^{T-k} \mathbb{E}[\widehat{w}_{\mathrm{on}}(S_t) \mid \mathcal{D}_{\mathrm{tr}}] = 1. \tag{S.9}$$

By simple algebra calculations and $\lim_{n\to\infty} n_t/n = 1/(T-k+1)$, we have

$$\begin{aligned} \lim_{n\to\infty} M_{21}^* \le & \lim_{n\to\infty} \frac{1}{W_n^*} \sup_{x\in\mathbb{R}} \left| \sum_{t=0}^{T-k} \frac{n_t}{n} \left\{ B_{\mathrm{emp}}(x \mid t, \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) - B(x \mid t, \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right\} \right| \\ & + \lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \frac{1}{W_n^*} \sum_{t=0}^{T-k} \frac{n_t}{n} B(x \mid t, \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) - B(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\mathrm{cal}}) \right| = 0 \quad \text{by (S.8) } \textit{and} \text{ (S.9)}. \end{aligned}$$

The desired result in Theorem 1 follows immediately.

□

C Proof of Theorem 2

This section proves Theorem 2, which analyzes the coverage probability of the proposed PIs in the context of off-policy evaluation. We focus on the case where data splitting is performed in a trajectory-wise manner, and let n denote the number of trajectories in $\mathcal{D}_{\mathrm{cal}}$. Please note that, with a slight abuse of notation, n here denotes the number of trajectories, which differs from its definition in the main paper. In the main paper, n refers to the cardinality of the calibration set $\mathcal{D}_{\mathrm{cal}}$, where data are stored as tuples rather than trajectories. The result can be readily extended to the tuple-data-splitting setting, as discussed in the proof of Theorem 1.

Proof of Theorem 2. Since $\widehat{C}^{\text{off}}_{N,\alpha}(S_{\text{test}})$ combines B intervals following [12, 9], it suffices to prove the validity of each CP interval. With some abuse of notation, we denote the single CP interval at target coverage level $1-\alpha$ as $\widehat{C}^{\text{off}}_{N,\alpha}(S_{\text{test}})$.

First, we index the data points in the calibration dataset $\mathcal{D}_{\text{cal}}$ as $\{1, 2, \ldots, n\}$. Given $\mathcal{D}$, $\widetilde{\mathcal{D}}_{\text{cal}}$ is a set of samples drawn from the distribution

$$\widehat{F}_n(s,g) = \sum_{t=0}^{T-k}\sum_{i=1}^{n} \frac{\widehat{w}_{\text{off}}(\mathcal{H}_{i,t:t+k})}{\sum_{t=0}^{T-k}\sum_{j=1}^{n}\widehat{w}_{\text{off}}(\mathcal{H}_{j,t:t+k})} I(S_{it} \le s, \widetilde{G}^{(k)}_{it} \le g),$$

where $\mathcal{H}_{i,t:t+k} = (S_{it}, A_{it}, \cdots, S_{i,t+k})$ denotes the local trajectory segment following the behavior policy. Following the main idea of the proof of Theorem 1, we consider two new test points:

$$(\widehat{S}_{\text{test}}, \widehat{G}_{\text{test}}) \sim \widehat{F}_n(s,g)$$

and

$$(S_{\text{test}}, \widetilde{G}_{\text{test}}) \sim \mathcal{P}_{S_0} \times \left((\mathcal{T}^\pi)^k \widehat{\eta}^\pi\right)(S_0)$$

which are drawn independently. Then for $G_{\text{test}} := G^\pi(S_{\text{test}})$, we have

$$\begin{aligned}
&\Pr\left(G_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(S_{\text{test}})\right) \ge \Pr\left(\widehat{G}_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(\widehat{S}_{\text{test}})\right) \\
&\quad - \left|\Pr\left(\widehat{G}_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(\widehat{S}_{\text{test}})\right) - \Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(S_{\text{test}})\right)\right| \\
&\quad - \left|\Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(S_{\text{test}})\right) - \Pr\left(G_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(S_{\text{test}})\right)\right| \\
&:= \widetilde{M}_1 - \widetilde{M}_2 - \widetilde{M}_3.
\end{aligned}$$

Note that the dataset $\mathcal{D}$ is sampled from the behavior policy π_b while $(S_{\text{test}}, G_{\text{test}})$ is generated by the target policy π. We now analyze $\widetilde{M}_1$, $\widetilde{M}_2$ and $\widetilde{M}_3$ separately.

(1) Given $\mathcal{D}$, $(\widehat{S}_{\text{test}}, \widehat{G}_{\text{test}})$ is exchangeable with $\widetilde{\mathcal{D}}_{\text{cal}}$. Existing result on coverage rate of SCP interval [4] gives

$$\widetilde{M}_1 = \mathbb{E}\left[\Pr\left(\widehat{G}_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(\widehat{S}_{\text{test}}) \,|\, \mathcal{D}\right)\right] \ge 1-\alpha. \tag{S.10}$$

(2) Similar to the treatment of M_3 in the proof of Theorem 1, we have

$$\widetilde{M}_3 \le \mathbb{E}\left[\sqrt{2L\bar{W}_1((\mathcal{T}^\pi)^k\widehat{\eta}^\pi, (\mathcal{T}^\pi)^k\eta^\pi)}\right] \le \sqrt{2L\gamma^k\mathbb{E}[\bar{W}_1(\widehat{\eta}^\pi, \eta^\pi)]}. \tag{S.11}$$

(3) Let $\mathcal{P}_t(s_0, a_0, \ldots, s_k, g)$ denote the joint probability distribution of $(\mathcal{H}_{t:t+k}, \widetilde{G}^{(k)}_t)$ given $\mathcal{D}_{tr}$ with some abuse of notation. Note that here $(\mathcal{H}_{t:t+k}, \widetilde{G}^{(k)}_t)$ is generated by π_b, consistent with the data. We further denote $h_{0:k} := (s_0, a_0, \ldots, s_k)$ for notational simplicity. Then

$$\begin{aligned}
\widetilde{M}_2(\mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) &:= \left|\Pr\left(\widehat{G}_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(\widehat{S}_{\text{test}}) \,|\, \mathcal{D}, \widehat{\mathcal{D}}_{\text{cal}}\right) - \Pr\left(\widetilde{G}_{\text{test}} \in \widehat{C}^{\text{off}}_{N,\alpha}(S_{\text{test}}) \,|\, \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\right| \\
&= \left|\Pr\left(|\widehat{G}_{\text{test}} - \widehat{v}^\pi(\widehat{S}_{\text{test}})| \le \widehat{q}_{1-\alpha} \,|\, \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\right. \\
&\qquad \left. - \Pr\left(|\widehat{G}_{\text{test}} - \widehat{v}^\pi(\widehat{S}_{\text{test}})| \le \widehat{q}_{1-a} \,|\, \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\right| \le \widetilde{M}_{21} + \widetilde{M}_{22},
\end{aligned}$$

where

$$\begin{aligned}
&\widetilde{M}_{21} := \sup_{x\in\mathbb{R}}\left|\sum_{t=0}^{T-k}\sum_{i=1}^{n}\frac{\widehat{w}_{\text{off}}(\mathcal{H}_{i,t:t+k})}{\sum_{t=0}^{T-k}\sum_{j=1}^{n}\widehat{w}_{\text{off}}(\mathcal{H}_{j,t:t+k})} I\left\{|\widetilde{G}^{(k)}_{it} - \widehat{v}^\pi(S_{it})| \le x\right\} - B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}})\right|, \\
&\widetilde{M}_{22} := \left|B^{\text{off}}(\hat{q}_{1-\alpha} \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) - \Pr\left(|\widetilde{G}_{\text{test}} - \widehat{v}^\pi(S_{\text{test}})| \le \widehat{q}_{1-\alpha} \,\middle|\, \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}\right)\right|, \\
&B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) := \frac{1}{T-k+1}\sum_{t=0}^{T-k}\int \widehat{w}_{\text{off}}(h_{0:k}) \cdot I\left\{|g - \widehat{v}^\pi(s_0)| \le x\right\} d\mathcal{P}_t(h_{0:k}, g).
\end{aligned}$$

(3.1) To analyze $\widetilde{M}_{21}$, we first define the normalization constant for weights as

$$W_n^{\text{off}} = \frac{1}{n(T-k+1)} \sum_{t=0}^{T-k} \sum_{i=1}^{n} \widehat{w}_{\text{off}}(\mathcal{H}_{i,t:t+k}).$$

Thus the first term in $\widetilde{M}_{21}$ becomes

$$\frac{1}{W_n^{\text{off}}} \frac{1}{n(T-k+1)} \sum_{t=0}^{T-k} \sum_{i=1}^{n} \widehat{w}_{\text{off}}(\mathcal{H}_{i,t:t+k}) I\left\{|\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| \le x\right\} := \frac{1}{W_n^{\text{off}}} B_{\text{emp}}^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}),$$

where $B_{\text{emp}}^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}})$ is the empirical version of $B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}})$. By a simple algebraic calculation, we have

$$\begin{aligned}\widetilde{M}_{21} \le & \frac{1}{W_n^{\text{off}}} \sup_{x \in \mathbb{R}} \left| B_{\text{emp}}^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) - B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) \right| \\ & + \left(\frac{1}{W_n^{\text{off}}} - 1 \right) \sup_{x \in \mathbb{R}} \left| B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) \right|.\end{aligned}$$

As $\frac{1}{T-k+1} \sum_{t=0}^{T-k} \mathbb{E}[\widehat{w}_{\text{off}}(\mathcal{H}_{t:t+k}) \mid \mathcal{D}_{\text{tr}}] = 1$, by law of large numbers, $\lim_{n\to\infty} W_n^{\text{off}} = 1$. Hence, for sufficiently large n, $W_n^{\text{off}} \ge 1/2$ and

$$\widetilde{M}_{21} \le 2 \underbrace{\sup_{x \in \mathbb{R}} \left| B_{\text{emp}}^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) - B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) \right|}_{\widetilde{E}} + \underbrace{\left| \frac{1}{W_n^{\text{off}}} - 1 \right| \sup_{x \in \mathbb{R}} \left| B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) \right|}_{\widetilde{F}}.$$

(3.1.1) For $\widetilde{E}$, since $\mathbb{E}[\widehat{w}_{\text{off}}(\mathcal{H}_{t:t+k})] < \infty$ for $0 \le t \le T-k$, the function class $\{\widehat{w}_{\text{off}}(h_{0:k}, g) I\{|g - \widehat{v}^{\pi}(s_0)| \le x\} : x \in \mathcal{R}\}$ is $\{\mathcal{P}_t(h_{0:k}, g) : 0 \le t \le T-k\}$-Glivenko-Cantelli. Applying the same argument as in Equation (S.5), we obtain

$$\lim_{n\to\infty} \sup_{x \in \mathbb{R}} \left| B_{\text{emp}}^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) - B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) \right| = 0. \tag{S.12}$$

(3.1.2) For $\widetilde{F}$, we have $\lim_{n\to\infty} \left(1/W_n^{\text{off}} - 1\right) = 0$, and

$$\begin{aligned}\sup_{x \in \mathbb{R}} \left| B^{\text{off}}(x \mid \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}}) \right| &\le \frac{1}{T-k+1} \sum_{t=0}^{T-k} \int \widehat{w}_{\text{off}}(h_{0:k}) d\mathcal{P}_t(h_{0:k+1}, g) \\ &= \frac{1}{T-k+1} \sum_{t=0}^{T-k} \mathbb{E}\left[\widehat{w}_{\text{off}}(\mathcal{H}_{t:t+k}) \mid \mathcal{D}_{\text{tr}}\right] = 1.\end{aligned}$$

Combining these results with (S.12), we obtain

$$\lim_{n\to\infty} \widetilde{M}_{21} = 0. \tag{S.13}$$

(3.2) Bound on $\widetilde{M}_{22}$. Following the proof of Theorem 1, we define a new probability measure

$$\frac{1}{T-k+1} \sum_{t=0}^{T-k} \widehat{w}_{\text{off}}(h_{0:k}) d\mathcal{P}_t(h_{0:k}, g),$$

and let $(\widetilde{\mathcal{H}}_{0:k}, \widetilde{G})$ be drawn from this measure with $\widetilde{\mathcal{H}}_{0:k} = (\widetilde{S}_0, \widetilde{A}_0, \cdots, \widetilde{S}_k)$. Then $\widetilde{M}_{22}$ can be equivalently written as

$$\widetilde{M}_{22} := \left| \Pr\left(|\widetilde{G} - \widehat{v}^{\pi}(\widetilde{S}_0)| \le \widehat{q}_{1-\alpha} \,\middle|\, \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}} \right) - \Pr\left(|\widetilde{G}_{\text{test}} - \widehat{v}^{\pi}(S_{\text{test}})| \le \widehat{q}_{1-\alpha} \,\middle|\, \mathcal{D}, \widetilde{\mathcal{D}}_{\text{cal}} \right) \right|.$$

Denote the marginal distribution of $\mathcal{P}_t(h_{0:k}, g)$ as $\mathcal{P}_t(h_{0:k})$ and define the calibration marginal distribution as $\mathcal{P}_{\text{cal}}(h_{0:k}) = \frac{1}{T-k+1} \sum_{t=0}^{T-k} \mathcal{P}_t(h_{0:k})$. Then $\widetilde{\mathcal{H}}_{0:k} \sim \widehat{w}_{\text{off}}(h_{0:k}) \mathcal{P}_{\text{cal}}(h_{0:k})$, and the unobserved $\mathcal{H}_{\text{test},0:\text{k}} = (S_{\text{test},0}, A_{\text{test},0}, \cdots, S_{\text{test,k}}) \sim w_{\text{off}}(h_{0:k}) \mathcal{P}_{\text{cal}}(h_{0:k})$, where $S_{\text{test},0} = S_{\text{test}}$.

Since the conditional distributions $\widetilde{G} \mid \widetilde{\mathcal{H}}_{0:k}$ and $\widetilde{G}_{\text{test}} \mid \mathcal{H}_{\text{test},0:k}$ are the identical, by Eq. (A.9) in [5], we have

$$\begin{aligned}\widetilde{M}_{22} &\leq d_{TV}(\mathcal{P}_{\widetilde{\mathcal{H}}_{0:k}}, \mathcal{P}_{\mathcal{H}_{\text{test},0:k}}) \\ &\leq \frac{1}{2(T-k+1)} \sum_{t=0}^{k} \mathbb{E}\left[|\widehat{w}_{\text{off}}(\mathcal{H}_{t:t+k}) - w_{\text{off}}(\mathcal{H}_{t:t+k})| \mid \mathcal{D}_{\text{tr}}\right].\end{aligned}$$

The desired result follows by combining (S.10) - (S.14).

□

D Algorithm for Off-Policy Setting

Algorithm 1 presents the proposed algorithm for the off-policy setting, which closely parallels that in the on-policy case.

Algorithm S.1: *CP for Infinite Horizon Off-policy Evaluation*

Data: $\mathcal{D} = \{(S_{it}, A_{it}, R_{it}, S_{i,t+1}) : 1 \leq i \leq N, 1 \leq t \leq T\}$, a test initial state S_{test} and a target policy π.

Input: $1-\alpha$, target coverage level; $\widetilde{\mathcal{A}}$, an off-policy distributional RL algorithm; $\mathcal{B}$, a propensity score training algorithm; $\mathcal{W}$, a density ratio estimation algorithm; k, step width; B, resampling number; l, subsample size; ξ, multiple subsampling parameter

Output: Prediction interval for $G^{\pi}(S_{\text{test}})$

Split the data: $\mathcal{D} = \mathcal{D}_{\text{tr}} \bigcup \mathcal{D}_{\text{cal}}$ where $\mathcal{D}_{\text{tr}} = \{(S_{it}, A_{it}, R_{it}, S_{i,t+1}) : (i,t) \in \mathcal{I}_{\text{tr}}\}$ and $\mathcal{D}_{\text{cal}} = \{(S_{it}, A_{it}, R_{it}, \ldots, S_{i,t+k}) : (i,t) \in \mathcal{I}_{\text{cal}}\}$. Here, $\mathcal{I}_{\text{tr}}$ and $\mathcal{I}_{\text{cal}}$ denote the indices of transitions in the training and calibration datasets, respectively.

Train a conditional return model $\widehat{\eta}^{\pi}(s)$ using $\widetilde{\mathcal{A}}$ based on $\mathcal{D}_{\text{tr}}$.

Obtain the value function estimator $\widehat{v}^{\pi}(s)$, the expectation of $\widehat{\eta}^{\pi}(s)$.

Obtain $\widehat{w}_{\text{on}}(s)$ as an estimator of the density ratio (2) in the main paper based on $\{S_{i0} : (i,0) \in \mathcal{I}_{\text{tr}}\}$ and $\{S_{it} : (i,t) \in \mathcal{I}_{\text{tr}}\}$ using $\mathcal{W}$.

Train $\widehat{\pi}^{b}(a \mid s)$ based on $\{(S_{it}, A_{it}) : (i,t) \in \mathcal{I}_{\text{tr}}\}$ using $\mathcal{B}$.

Obtain $\widehat{w}_{\text{off}}(\cdot)$ by plugging in $\widehat{w}_{\text{on}}$ and $\widehat{\pi}_b$ in (3) of the main paper.

for $b = 1 : B$ **do**

- Sample l data tuples $\{(S_{it}, A_{i,t}, R_{i,t}, \ldots, S_{i,t+k}) : (i,t) \in \mathcal{I}_{\text{cal}}^{(b)}\}$ from $\mathcal{D}_{\text{cal}}$ accoring to the importance weight $\widehat{w}_{\text{off}}(S_{it}, A_{it}, \ldots, S_{i,t+k})$.
- Calculate pseudo-return (1) in the main paper and obtain $\widetilde{\mathcal{D}}_{\text{cal}}^{(b)} := \{(S_{it}, \widetilde{G}_{it}^{(k)}) : (i,t) \in \mathcal{I}_{\text{cal}}^{(b)}\}$.
- Calculate the nonconformity scores: $\{V_{it} := |\widetilde{G}_{it}^{(k)} - \widehat{v}^{\pi}(S_{it})| : (i,t) \in \mathcal{I}_{\text{cal}}^{(b)}\}\}$.
- Calculate $\widehat{q}_{1-\alpha\xi}^{(b)}$, the $\lceil l(1-\alpha\xi)\rceil$-th smallest value of $\{V_{it} : (i,t) \in \mathcal{I}_{\text{cal}}^{(b)}\}$.
- Obtain $\widehat{C}_{N,\alpha\xi}^{(b)}(S_{\text{test}}) = \widehat{v}^{\pi}(S_{\text{test}}) \pm \hat{q}_{1-\alpha\xi}^{(b)}$.

Result: A conformal predictive region for $G^{\pi}(S_{\text{test}})$ with a coverage rate of $1-\alpha$ is

$$\widehat{C}_{N,\alpha}^{\text{off}}(S_{\text{test}}) = \left\{G : \frac{1}{B}\sum_{b=1}^{B} I\left\{G \in \widehat{C}_{N,\alpha\xi}^{(b)}(S_{\text{test}})\right\} \geq 1-\xi\right\}. \quad \text{(S.14)}$$

E Implementation Details and Additional Results

We provide additional implementation details for the numerical experiments. The code is available at: https://github.com/yyzhangecnu/CPbeyonghorizon.

Example 1. We adopt the QTD algorithm (Algorithm 1 in [8]) to estimate the quantiles of the return distribution. The learning rate ρ is set to 0.1, and the discount factor γ is 0.8. We use 20 quantile levels in the estimation. The behavior policy is estimated based on the empirical frequency of (s, a) pairs in the training set, and the importance weights are computed similarly using frequency-based estimates. The hyperparameter ξ, which controls the aggregation of multiple prediction intervals, is selected via grid-based cross-fitting since simulations allow us to generate trajectories with sufficiently large T to get accurate return. We set the number of aggregated intervals to $B = 100$, with each interval constructed from a subsample of 400 tuples drawn from the calibration dataset. We repeat the experiment over 100 simulation runs and report the boxplots of the empirical coverage probabilities and the average lengths of PIs. The nominal coverage level is fixed at 90%.

Influence of k. Based on Example 1, we further investigate the effect of using larger k values, specifically for $k = 6, 7, 8$. Each experiment is repeated 100 times, and we report the mean and standard deviation of the empirical coverage probability (cov) and prediction interval length (len) under the nominal 90% coverage level.

As shown in Table 1, increasing k consistently results in overcoverage and, consequently, wider prediction intervals. This observation aligns with our theoretical results in Section 4 (Theorems 1 and 2), which reveal an inherent trade-off. A larger k reduces the approximation error in estimating $\widehat{\eta}^{\pi}$, but at the same time, it increases the difficulty of accurately estimating the off-policy weights and maintaining the approximate independence of calibration samples particularly under substantial distributional shifts. Empirically, we find that choosing $k = 2$ or 3 provides a good balance between these competing factors.

Table 1: Coverage (cov) and average length (len) for different k under on-policy and off-policy settings with $\xi = 0.8$. Standard errors are shown in parentheses.

on	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
cov	0.87(0.01)	0.90(0.01)	0.91(0.01)	0.92(0.01)	0.92(0.01)	0.93(0.01)	0.94(0.01)	0.94(0.01)
len	7.78(0.10)	8.24(0.10)	8.56(0.13)	8.78(0.14)	9.00(0.15)	9.15(0.19)	9.31(0.23)	9.50(0.22)
off	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
cov	0.87(0.01)	0.91(0.01)	0.92(0.01)	0.92(0.01)	0.93(0.01)	0.93(0.01)	0.93(0.01)	0.94(0.02)
len	7.57(0.10)	8.13(0.11)	8.47(0.14)	8.67(0.14)	8.90(0.17)	9.02(0.18)	9.20(0.18)	9.26(0.20)

Influence of ξ. We conduct experiments for Example 1 with ξ varying from 0.1 to 0.9 and $k = 2, 3, 4$. Each setting is repeated 100 times, and we report the mean and standard deviation of the coverage probability (cov) and interval length (len) at the nominal 90% coverage level, as shown in Table 2. The results show that smaller ξ and larger k tend to cause overcoverage, whereas settings with $\xi \geq 0.5$ and $k = 2, 3$ generally achieve satisfactory performance.

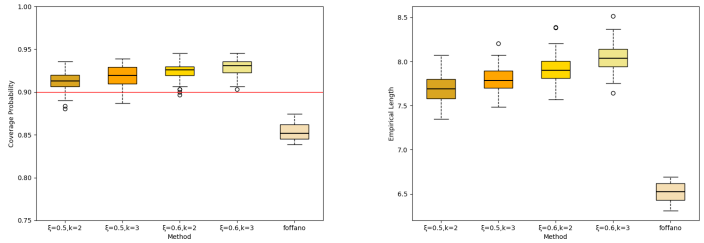
Comparison with [3]. We compare the performance of our method and that of [3] in the off-policy setting for Example 1 with a fixed horizon of 20. For Foffanos method, we follow their gradient-based approach to train the likelihood ratio model $w(x, y)$ via linear regression and apply WCP to construct prediction intervals. For our method, we replace the nonconformity score with the double-quantile (DQ) score from [3], setting ξ to 0.5 and 0.6, and k to 2 and 3. To better accommodate the DQ score, we employ the interval aggregation technique proposed by [6]. Each experiment is repeated 100 times, with the nominal coverage level fixed at 90%. The results, shown in Figure S.1, indicate that our method achieves superior performance in terms of both coverage probability and average interval length.

Example 2. The state space is continuous in this setting. To apply the QTD algorithm, we train a quantile network with 20 quantile levels. The input to the network is the state, and the architecture consists of three layers with 32 hidden neurons and 40 output units, each corresponding to a specific quantile level for a given state-action pair. The behavior policy is estimated using a separate neural network with architecture $2 \to 32 \to 32 \to 2$, where the outputs represent the action probabilities. Following the QR-DQN algorithm in [2], we replace the quantile regression loss with the Huber quantile loss to improve stability.

The importance weights are estimated using logistic regression. The hyperparameter ξ, which governs the aggregation of multiple PIs, is selected via grid-based cross-fitting since simulations allow

Table 2: Coverage probability (cov) and interval length (len) for different ξ under on-policy and off-policy settings. Standard errors are shown in parentheses.

on		cov			len	
ξ	$k=2$	$k=3$	$k=4$	$k=2$	$k=3$	$k=4$
0.1	0.95(0.01)	0.96(0.01)	0.96(0.01)	10.21(0.20)	10.71(0.21)	11.04(0.26)
0.2	0.95(0.01)	0.95(0.01)	0.95(0.01)	9.68(0.15)	10.10(0.16)	10.40(0.17)
0.3	0.94(0.01)	0.95(0.01)	0.95(0.01)	9.30(0.12)	9.67(0.14)	9.95(0.16)
0.4	0.92(0.01)	0.94(0.01)	0.95(0.01)	8.98(0.10)	9.34(0.13)	9.62(0.15)
0.5	0.92(0.01)	0.93(0.01)	0.94(0.01)	8.73(0.08)	9.07(0.13)	9.33(0.15)
0.6	0.91(0.01)	0.92(0.01)	0.93(0.01)	8.53(0.09)	8.87(0.13)	9.09(0.16)
0.7	0.91(0.01)	0.92(0.01)	0.92(0.01)	8.37(0.09)	8.69(0.12)	8.92(0.14)
0.8	0.90(0.01)	0.91(0.01)	0.92(0.01)	8.24(0.10)	8.56(0.13)	8.78(0.14)
0.9	0.90(0.01)	0.91(0.01)	0.92(0.01)	8.20(0.12)	8.51(0.14)	8.72(0.16)
off		cov			len	
ξ	$k=2$	$k=3$	$k=4$	$k=2$	$k=3$	$k=4$
0.1	0.96(0.01)	0.96(0.01)	0.97(0.01)	10.17(0.19)	10.68(0.22)	10.95(0.30)
0.2	0.95(0.01)	0.95(0.01)	0.96(0.01)	9.62(0.15)	10.08(0.17)	10.33(0.20)
0.3	0.94(0.01)	0.95(0.01)	0.96(0.01)	9.24(0.12)	9.64(0.15)	9.90(0.17)
0.4	0.93(0.01)	0.94(0.01)	0.95(0.02)	8.93(0.10)	9.30(0.13)	9.57(0.15)
0.5	0.92(0.01)	0.93(0.01)	0.94(0.01)	8.68(0.11)	9.03(0.14)	9.28(0.15)
0.6	0.92(0.01)	0.93(0.01)	0.93(0.01)	8.43(0.11)	8.78(0.14)	8.99(0.14)
0.7	0.91(0.01)	0.92(0.01)	0.93(0.01)	8.26(0.11)	8.61(0.13)	8.82(0.14)
0.8	0.91(0.01)	0.92(0.01)	0.92(0.01)	8.13(0.11)	8.47(0.14)	8.67(0.14)
0.9	0.91(0.01)	0.92(0.01)	0.92(0.01)	8.07(0.13)	8.41(0.16)	8.61(0.15)



(a) Coverage probability (b) Average length

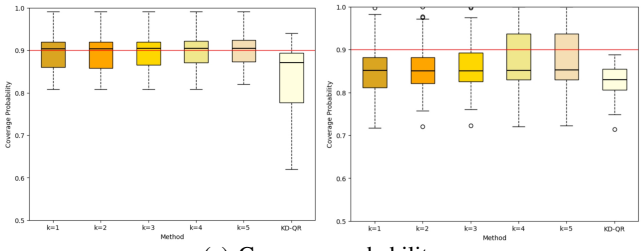
Figure S.1: Coverage probability and average interval length at the 90% level for the proposed method with $\xi = 0.5, 0.6$ and $k = 2, 3$ (from left to right) and Foffano's method (rightmost).

us to generate trajectories with sufficiently large T to get accurate return. We set the number of aggregated intervals to $B = 50$, with each interval constructed from a subsample of 200 tuples drawn from the calibration dataset. We repeat the experiment over 100 simulation runs and report boxplots of the empirical coverage probabilities and the average lengths of the resulting PIs. The nominal coverage level is fixed at 90%.

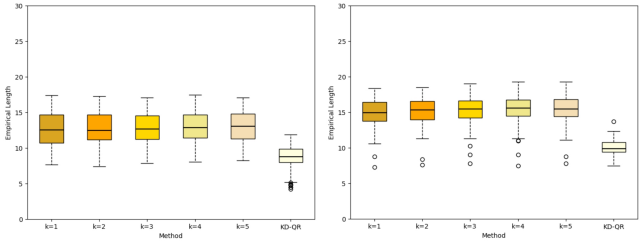
Example 3. Mountain car is a classic RL control problem. We first use RBF-based feature engineering to search for a suboptimal policy denoted by π_Q via Q-learning. To better illustrate that our proposal is a wrapper, we apply kernel density estimation to approximate the return distribution from Monte Carlo rollouts. The discount factor γ is set to 0.99. The remaining procedure of the experiment is the same as Example 2. We set the number of aggregated intervals to $B = 50$, with each interval constructed from a subsample of 200 tuples drawn from the calibration dataset. We repeat the experiment over 50 simulation runs and report boxplots of the empirical coverage probabilities and the average lengths of the resulting PIs. The nominal coverage level is fixed at 90%.

Figure S.2 presents the results for both on-policy and off-policy settings in Example 3. These experiments demonstrate that our proposed method consistently outperforms the kernel-density-based approach, even when the kernel density is estimated using Monte Carlo rollouts under the target

policy. Notably, all intervals exhibit greater variance compared to those in Examples 1 and 2. This increased variance arises from the challenging nature of the environment, where the agent receives a constant reward of -1 until reaching the goal (the flag). As a result, the immediate reward provides limited information, making learning and accurate value estimation more difficult.



(a) Coverage probability



(b) Average length

Figure S.2: Coverage probability and average interval length at the 90% level for the proposed method with k-step pseudo-returns ($k = 1, \ldots, 5$, from left to right) and KD-QR (rightmost), under on-policy (left) and off-policy (right) settings in Example 3.

Example 4. We extend Example 1 to a high-dimensional setting with 50 states, denoted by $\mathbf{S}_t = (S_{1t}, S_{2t}, \cdots, S_{50t})^\top$, where each feature S_{jt} for $1 \leq j \leq 50$ is binary, taking values x_1 or x_2. The action space is $\{0, 1\}$ and only affects transitions of the first state S_{1t}. The remaining states independently take values x_1 or x_2 with equal probability at each time step, thereby serving as confounders. The agent, however, does not know which state is directly influenced by the action. The reward follows the same distribution as in Example 1. The behavior policy specifies transition probabilities of 0.4 for $x_1 \to x_2$ and 0.8 for $x_2 \to x_1$, while the target policy remains the same as in Example 1 for the off-policy setting.

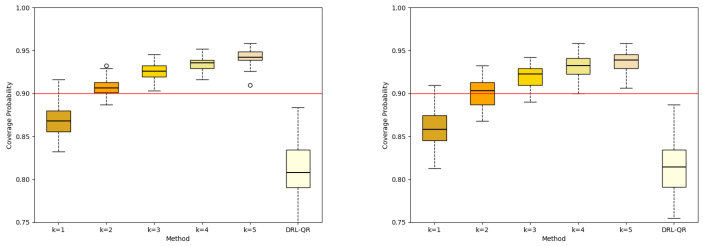
We employ quantile temporal difference (QTD) learning with linear regression and a ridge penalty to alleviate overfitting. The number of aggregated intervals is set to $B = 50$ and the hyperparameter is fixed at $\xi = 0.8$. Each interval is constructed from a subsample of 200 tuples drawn from 6000 calibration tuples. Experiments are conducted for $k = 1, \ldots, 5$, each repeated 50 times. We report boxplots of the empirical coverage probabilities and average interval lengths in Figure S.3, with the nominal coverage level fixed at 90%. The results show that our proposed method consistently outperforms the DRL-QR baseline in this high-dimensional setting.

References

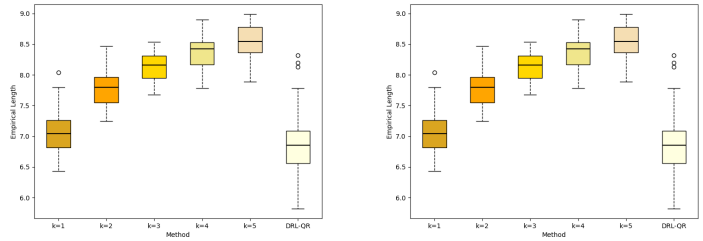
[1] Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023.

[2] Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, volume 32, 2018.

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



(a) Coverage probability



(b) Average length

Figure S.3: Coverage probability and average interval length at the 90% level for the proposed method with k-step pseudo-returns ($k = 1, \ldots, 5$, from left to right) and DRL-QR (rightmost), under on-policy (left) and off-policy (right) settings in Example 4.

[4] 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.

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

[6] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[7] Nathan Ross. Fundamentals of steins method. 2011.

[8] Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G. Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research*, 25(163):1–47, 2024.

[9] Aldo Solari and Vera Djordjilović. Multi split conformal prediction. *Statistics & Probability Letters*, 184:109395, 2022.

[10] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

[11] Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. Wasserstein-regularized conformal prediction under general distribution shift. *arXiv preprint arXiv:2501.13430*, 2025.

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