# FDR-CONTROLLED SELECTION IN REINFORCEMENT LEARNING VIA DISTRIBUTIONAL LIKELIHOOD RATIOS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In high-stakes reinforcement learning, reliable deployment requires identifying instances where a target policy is safe or effective, rather than relying solely on average performance. We study the problem of selecting a subset of initial states whose returns exceed a given threshold, while controlling the false discovery rate (FDR) of erroneous selections. This problem is challenging due to unobservable long-horizon returns and distribution shift in off-policy settings. To address this, we propose a unified framework for FDR-controlled selection in RL based on distributional modeling and likelihood-ratio principles. Our approach estimates the return distribution under the target policy and constructs selection rules that directly quantify evidence for desirable outcomes, avoiding the need for explicit p-value construction. We further incorporate a temporal decomposition strategy and an offline calibration procedure to enable reliable FDR control under partial observability and policy mismatch. We provide theoretical guarantees and demonstrate empirically that our method achieves accurate FDR control with improved selection power across a range of environments.

## 1 INTRODUCTION

In many high-stakes applications of reinforcement learning (RL), such as healthcare and autonomous systems, deploying a new policy requires careful evaluation before real-world use. While off-policy evaluation methods aim to estimate the expected return of a target policy from historical data (Uehara et al., 2022; Kallus & Uehara, 2020; Uehara et al., 2020), accurate value estimation alone is often insufficient in practice. A policy that performs well on average may still lead to undesirable outcomes for certain instances or subpopulations, which is unacceptable in safety-critical settings. Consequently, practitioners are often interested in **selective deployment**: identifying subsets of instances for which a target policy is reliably safe or effective. This leads to a fundamental decision problem:

*Given a policy of interest, can we select a subset of instances whose associated returns exceed a desired threshold, while controlling the proportion of incorrect selections?*

Such scenarios arise naturally across domains. In personalized healthcare, a clinician may wish to deploy a treatment policy only for patients whose initial conditions are likely to yield long-term health benefits, while avoiding those at risk of adverse outcomes (Chakraborty & Murphy, 2014; Gottesman et al., 2020). In digital platforms such as e-commerce, a new intervention policy (e.g., coupon allocation or recommendation strategy) may improve average performance but harm a subset of users, necessitating selective deployment to a high-confidence user group (Zhao et al., 2017; Chen et al., 2019). In autonomous driving, one must identify initial conditions under which a policy achieves sufficiently high cumulative reward or safety guarantees, since exhaustively simulating all possible scenarios is computationally infeasible (Garcıa & Fernández, 2015; Kiran et al., 2021).

We formalize this problem in the standard RL setting. Suppose there are $m$ instances or initial conditions awaiting selection. Given a target policy $\pi$, our goal is to identify a subset of instances whose associated returns exceed a predefined threshold $c$. To ensure reliability, we aim to control the proportion of erroneous selections among all selected instances. This corresponds to controlling the FDR, defined as

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\sum_{i=1}^{m} \mathbb{I}\{i \text{ is selected and } G_i \le c\}}{1 \vee \sum_{i=1}^{m} \mathbb{I}\{i \text{ is selected}\}}\right] \le \alpha, \tag{1}$$



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

where $G_i$ denotes the return associated with instance $i$, $\alpha \in (0, 1)$ is a user-specified error level, and the expectation is taken over the randomness in the data and learning procedure. This formulation connects selective deployment in RL with large-scale hypothesis testing, where each instance corresponds to a hypothesis about whether its return exceeds the target threshold.

However, solving this problem in RL is fundamentally challenging: First, the return is not directly observable at the state level and must be inferred from trajectory data. Most existing works focus on estimating the expected return, whereas our goal requires reasoning about the return itself at the instance level, which is significantly more demanding. Second, in off-policy settings, such estimation is further complicated by distribution shift between the behavior policy and the target policy, as well as temporal dependence along trajectories. These factors make accurate uncertainty quantification particularly difficult (Gan et al., 2025). Third, existing approaches for FDR control, such as conformal selection, rely on well-calibrated p-values (Jin & Candès, 2023b). Constructing such p-values in RL is nontrivial due to the lack of directly observed outcomes and the complex dependence structure of the data.

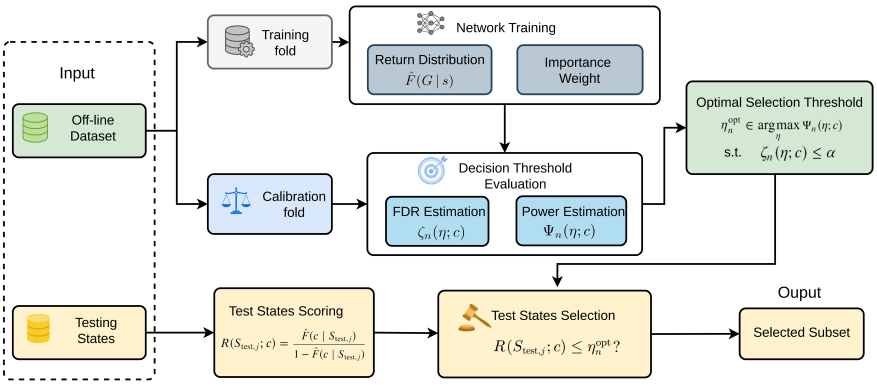


Figure 1: Pipeline of the proposed likelihood-ratio-based selection framework.

To address these challenges, we propose a selection framework tailored to RL that directly targets FDR control without relying on exact p-values. Our approach is built on four key ideas:

(i) We reformulate the selection problem in terms of the *conditional return distribution* under the target policy. This enables direct reasoning about whether the return exceeds a threshold at the instance level, which is essential for selective deployment. We leverage distributional RL as a practical tool to estimate this distribution, but emphasize that such model-based tail probabilities alone do not guarantee valid FDR control.

(ii) Bypassing the need for conformal p-values, we construct likelihood-ratio-based conformity scores inspired by the Neyman–Pearson principle. These scores directly quantify the evidence for $G > c$ using the estimated return distribution, and enable consistent estimation of the false discovery proportion, leading to more powerful selection procedures.

(iii) To enable valid FDR estimation in RL, we develop a calibration strategy tailored to RL. We introduce a temporal decomposition to handle unobservable long-horizon returns, use experience replay for subsampling, and incorporate importance weighting to correct for distribution shift between behavior and target policies. This unified framework can be viewed as a weighted conformal-style procedure adapted to RL, providing stable and accurate estimation of the false discovery proportion.

(iv) To construct practically useful FDR and power functions, we replace an indicator function with the conditional return distribution, thereby avoiding reliance on unobservable returns. This marks a significant departure from Qin et al. (2025)'s method, which relies on indicator functions and requires observable outcomes.

Together, these components enable principled and reliable FDR-controlled selection in RL, while achieving improved statistical power in long-horizon and off-policy settings. The overall framework is summarized in Figure 1

### 1.1 Related Works

Selective inference is well-established in statistical learning, especially in supervised settings. However, extending these ideas to RL remains largely unexplored due to the challenges posed by unobservable and long-horizon returns. Our work bridges RL and large-scale hypothesis testing by introducing FDR-controlled selection under such settings. We relate our work to two main strands of literature: (i) selective inference and conformal selection with FDR control, and (ii) distributional RL for policy evaluation.

**Conformal selection for FDR control.** Conformal prediction provides a distribution-free framework for constructing prediction sets with guaranteed coverage for a *single* test instance. In contrast, our work considers simultaneous inference over *multiple* units and is closely related to the emerging literature on conformalized selection (Jin & Candès, 2023b;a; Qin et al., 2025; Du et al., 2025). A representative approach is proposed by Jin & Candès (2023b), which constructs conformal p-values for each candidate unit and applies the Benjamini–Hochberg procedure to control the false discovery rate (FDR). Subsequent work extends this framework along several directions, including applications to new domains such as alignment in large language models (Gui et al., 2024), adaptive and interactive data analysis (Gui et al., 2025), selection under multiple conditions (Hao et al., 2025), and multivariate outcomes (Bai et al., 2025). Parallel developments also study conformalized multiple testing and selective inference under FDR or false coverage rate (FCR) control, focusing on data-dependent selection and improved calibration under complex sampling schemes (Wang et al., 2024; Bao et al., 2025; Huo et al., 2025).

Despite these advances, existing approaches predominantly rely on constructing well-calibrated conformal p-values followed by multiple testing procedures. Such requirements are difficult to satisfy in RL. In contrast, our approach bypasses p-values and directly estimates the false discovery proportion via likelihood-ratio-based conformity scores. A closely related idea has been explored in supervised learning (Qin et al., 2025), whereas our work extends this perspective to the significantly more challenging RL setting. This extension is nontrivial, as Qin et al. (2025)'s method relies on the availability of outcomes. In our setting, however, the outcomes (i.e., the long-horizon returns) are unobservable.

**Distributional RL for policy evaluation.** Distributional RL shifts the paradigm from estimating the scalar expected return to modeling the full return distribution (Bellemare et al., 2017a). Early work in this area primarily focuses on *online learning*, where the agent interacts with the environment and learns the return distribution via the distributional Bellman equation. Representative methods include Categorical DQN (C51) (Bellemare et al., 2017a), which discretizes the support of the return distribution, and Quantile Regression DQN (QR-DQN) (Dabney et al., 2018) and related approaches (Rowland et al., 2024), which directly estimate conditional quantiles. These methods have demonstrated strong empirical performance in control tasks and provide a richer characterization of uncertainty compared to value-based approaches (Bellemare et al., 2020; Fawzi et al., 2022).

More recent work extends distributional RL to *off-policy evaluation*, where the goal is to estimate the return distribution of a target policy from offline data. This setting is substantially more challenging due to distribution shift and the lack of direct interaction. Existing approaches include quantile-based methods adapted to the offline setting (Kuang et al., 2026) and generative modeling approaches that aim to learn the full return distribution, such as GAN-based methods (Qi et al., 2025; Hong et al., 2024; 2025). Our work falls into this line of research but differs in both objective and methodology: Rather than solely estimating the return distribution, we leverage it to construct likelihood-ratio-based selection rules with FDR control, tailored to the challenges of off-policy RL.

## 2 Problem Setup

Suppose we are given an offline dataset of $N$ trajectories $\mathcal{D} = \{\zeta_i\}_{i=1}^N$, each consisting of $T$ time steps. For simplicity, we assume trajectories have a uniform length, although our framework naturally extends to variable-length settings. Each trajectory $\zeta_i = \{(S_{it}, A_{it}, R_{it})\}_{t=0}^{T-1}$ consists of the state $S_{it}$, the action $A_{it}$, and the immediate reward $R_{it}$. The data are generated by a **behavior policy** $\pi_b$, such that $A_{it} \sim \pi_b(\cdot \mid S_{it})$, and evolve according to a transition kernel $\mathcal{P}$ with $(R_{it}, S_{i,t+1}) \sim \mathcal{P}(\cdot \mid S_{it}, A_{it})$. We consider a standard infinite-horizon RL setting, where the environment is modeled

as a time-homogeneous Markov Decision Process (MDP). In healthcare applications, for instance, each trajectory corresponds to a patient, where $S_{it}$ represents clinical features, $A_{it}$ denotes the administered treatment, and $R_{it}$ reflects the observed clinical response.

Let $\pi$ denote the **target policy** of interest. For an initial state $s$, the discounted return under $\pi$ is defined as $G^\pi(s) = \sum_{t=0}^{\infty} \gamma^t R_t$, where $\gamma \in (0,1)$ is the discount factor. The random variable $G^\pi(s)$ captures the long-term outcome induced by policy $\pi$ starting from state $s$. Given a collection of test states $\{S_{\text{test},j}\}_{j=1}^m$, let $G_j^\pi := G^\pi(S_{\text{test},j}), \quad 1 \le j \le m$, denote the corresponding returns. Our goal is to identify those states for which the return exceeds a pre-specified threshold $c$. This can be formulated as the following multiple testing problem:

$$H_{0,j} : G_j^\pi \le c \quad \text{versus} \quad H_{1,j} : G_j^\pi > c, \qquad 1 \le j \le m. \tag{2}$$

We aim to select a subset $\mathcal{S} \subseteq \{1, \dots, m\}$ while controlling the FDR at a target level $\alpha \in (0,1)$:

$$\text{FDR} = \mathbb{E}\left[ \frac{\sum_{j=1}^m \mathbb{I}\{j \in \mathcal{S},\, G_j^\pi \le c\}}{1 \vee |\mathcal{S}|} \right], \tag{3}$$

where $|\mathcal{S}|$ denotes the cardinality of the set $\mathcal{S}$ and $1 \vee |\mathcal{S}| = \max(1, |\mathcal{S}|)$. The FDR measures the expected proportion of selected states whose returns actually fall below the desired threshold $c$. In risk-sensitive applications, such errors correspond to deploying a policy in unfavorable or unsafe conditions (e.g., treating patients who may not benefit), and thus must be strictly controlled.

In addition to controlling FDR, it is desirable to maximize the number of truly beneficial selections. This is captured by the statistical power:

$$\text{Power} = \mathbb{E}\left[ \frac{\sum_{j=1}^m \mathbb{I}\{j \in \mathcal{S},\, G_j^\pi > c\}}{1 \vee \sum_{j=1}^m \mathbb{I}\{G_j^\pi > c\}} \right]. \tag{4}$$

Power measures the proportion of truly favorable states that are correctly selected. In risk-sensitive settings, this corresponds to identifying as many safe or beneficial deployment conditions as possible, without violating safety guarantees. Our objective is to ensure strict control of FDR while maximizing power subject to this constraint. This reflects a fundamental trade-off in high-stakes decision-making: avoiding harmful deployments is the primary priority, while improving coverage of beneficial cases is secondary.

Achieving FDR-controlled selection in RL is challenging due to unobserved outcomes and temporal dependence. A natural approach is to adapt conformal selection methods from supervised learning. Recent work (Gan et al., 2025) combines distributional RL with pseudo-return construction and experience replay to enable conformal prediction beyond finite horizons, which could in principle be used to construct p-values for our problem. However, extending such approaches to FDR-controlled selection requires additional calibration under dependence and multiple testing procedures (e.g., Benjamini–Hochberg), which can be conservative and difficult to implement in RL settings. Instead, we take a different route: leveraging the estimated return distribution, we directly construct likelihood-ratio-based selection rules that bypass p-values, resulting in a simpler and more powerful framework for FDR control.

## 3 Likelihood-Ratio-Based Selection with FDR Control

In this section, we develop a likelihood-ratio-based framework for FDR-controlled selection in reinforcement learning. We begin by estimating the conditional distribution of the return under the target policy. Specifically, we split the dataset into training and validation sets, and use the training set to learn the conditional cumulative distribution function (CDF) of $G^\pi$ given an initial state $S = s$, denoted by $F(g \mid s)$. Our selection framework can use different return distribution estimators. We introduce an energy-based generative implementation below.

### 3.1 Likelihood-Ratio-Based Selection from Return Distribution

Let $g(s)$ denote the marginal density of the state $S$, and $F_G(g)$ the marginal CDF of $G^\pi$. Recall that our goal is to test the hypotheses $H_{0,j} : G_j^\pi \le c$ versus $H_{1,j} : G_j^\pi > c$. Since the return $G_j^\pi$ is

not directly observable for test states, we instead consider the induced distribution of the observable variable $S_{\text{test},j}$ under the two hypotheses. By the law of total probability, the conditional densities of $S$ under the null and alternative hypotheses, denoted by $g_0(s;c)$ and $g_1(s;c)$, can be written as

$$g_0(s;c) := p(s \mid G^\pi \le c) = \frac{F(c \mid s)\, g(s)}{F_G(c)}, \quad g_1(s;c) := p(s \mid G^\pi > c) = \frac{(1 - F(c \mid s))\, g(s)}{1 - F_G(c)}.$$

*This reformulation shifts the problem from testing an unobserved outcome to distinguishing between two distributions over observable states.* Motivated by the Neyman–Pearson lemma, the most powerful test for discriminating between $H_0$ and $H_1$ is based on the likelihood ratio $R^*(s;c) := g_0(s;c)/g_1(s;c)$. Rejecting the null hypothesis when $R^*(S_{\text{test},j};c)$ is small yields the most powerful test. Plugging in the expressions for $g_0$ and $g_1$, we obtain

$$R^*(s;c) = \frac{F(c \mid s)}{1 - F(c \mid s)} \cdot \frac{1 - F_G(c)}{F_G(c)}.$$

Since the second factor does not depend on $s$, the rejection region is equivalently characterized by $\{R(s;c) \le \eta\}$, where we define the core conformity score as $R(s;c) := F(c \mid s)/\{1 - F(c \mid s)\}$.

This score has a natural interpretation: smaller values of $R(s;c)$ indicate stronger evidence against $H_0 : G^\pi(s) \le c$. Note that $R(s;c)$ is a strictly increasing function of $F(c \mid s)$, so smaller values of $R(s;c)$ correspond to smaller values of $F(c \mid s) = \mathbb{P}(G^\pi \le c \mid s)$. Therefore, rejecting $H_0$ for states with small $R(s;c)$ is equivalent to selecting states for which the probability of $G^\pi \le c$ is small, or equivalently, those likely to satisfy $G^\pi(s) > c$. Moreover, the likelihood-ratio-based score is straightforward to compute once the conditional return distribution is available, and avoids the need to construct p-values.

Under a homoscedastic location model, residual-based conformal scores and the LR score can induce the same ranking. The advantage of LR ranking is more pronounced when conditional returns are heteroskedastic, skewed, or multimodal, since states with similar means can have different failure probabilities. We compare these approaches empirically below.

### 3.2 FDR Control and Power Evaluation via Calibration Data and Temporal Decomposition

The likelihood-ratio-based score provides an optimal ranking of states for selection. However, in practice, the selection threshold cannot be chosen directly, as both the FDR in (3) and the statistical power in (4) depend on the unobserved returns $G^\pi$. To address this challenge, we introduce a calibration strategy that enables estimation of FDR and power using data, which in turn allows us to select the threshold in a data-driven manner.

Suppose we have a calibration dataset $\{(S_{i0}, A_{i0}, R_{i0}, S_{i1}, \ldots, S_{ik})\}_{i=1}^n$, consisting of $k$-step trajectories generated under the target policy $\pi$, where the initial states $S_{i0}$ share the same distribution as the test states. This dataset is introduced for conceptual clarity; we describe how to construct such data in the subsequent subsections. If the full return $G_i^\pi$ were observable, we could estimate FDR via empirical FDP and power at threshold $\eta$ as

$$\mathrm{FDR}(\eta;c) = \frac{\sum_{i=1}^n \mathbb{I}\{R(S_{i0};c) \le \eta\} \cdot \mathbb{I}\{G_i^\pi \le c\}}{1 \vee \sum_{i=1}^n \mathbb{I}\{R(S_{i0};c) \le \eta\}},$$
$$\mathrm{Power}(\eta;c) = \frac{\sum_{i=1}^n \mathbb{I}\{R(S_{i0};c) \le \eta\} \cdot \mathbb{I}\{G_i^\pi > c\}}{1 \vee \sum_{i=1}^n \mathbb{I}\{G_i^\pi > c\}}.$$

However, since the full return $G_i^\pi$ is not observable in practice, these formulae are of little practical use. To overcome this, we replace the indicator $\mathbb{I}\{G_i^\pi \le c\}$ with its conditional expectation given the observed partial trajectory, yielding the estimable surrogate $\mathbb{P}(G_i^\pi \le c \mid S_{i0}, A_{i0}, R_{i0}, \ldots, S_{ik})$.

This substitution is natural, as the conditional expectation provides the optimal $L^2$ approximation of the unobserved indicator given the available information. To evaluate this quantity, we exploit a temporal decomposition inspired by temporal-difference learning. Specifically, we write $G_i^\pi = \sum_{j=0}^{k-1} \gamma^j R_{ij} + \gamma^k G^\pi(S_{ik})$, where $G^\pi(S_{ik})$ denotes the remaining return starting from state $S_{ik}$.

Consequently, $\{G_i^\pi \le c\}$ is equivalent to $\left\{G^\pi(S_{ik}) \le (c - \sum_{j=0}^{k-1} \gamma^j R_{ij})/\gamma^k\right\}$. Using the estimated conditional return distribution $\widehat{F}(u \mid s)$, we replace $\mathbb{I}\{G_i^\pi \le c\}$ in $\mathrm{FDR}(\eta; c)$ and $\mathrm{Power}(\eta; c)$ by

$$\widehat{\mathbb{P}}(G_i^\pi \le c \mid S_{i0}, \ldots, S_{ik}) = \widehat{F}\left(\frac{c - \sum_{j=0}^{k-1} \gamma^j R_{ij}}{\gamma^k} \,\middle|\, S_{ik}\right). \tag{5}$$

This construction combines observed short-term rewards with a model-based estimate of the remaining return. Compared with using $F(c \mid S_{i0})$ alone, it incorporates the observed partial trajectory when estimating the failure indicator. Increasing $k$ reduces the contribution of residual distribution error, but can also require more difficult tail estimation and increase off-policy weight variability. Thus, moderate horizons may provide a better finite-sample trade-off.

### 3.3 Constructing Calibration Data via Experience Replay and Weighted Adjustment

In the previous subsection, we assumed the existence of a calibration dataset consisting of i.i.d. samples drawn from the target policy and sharing the same distribution as the test data. In practice, however, we only have access to a calibration dataset of size $N_{\mathrm{cal}}$, extracted from the full offline data, $\{(S_{it}, A_{it}, R_{it}, \ldots, S_{i,t+k})\}_{i=1}^{N_{\mathrm{cal}}}$, generated under a behavior policy $\pi_b$. The $k$-step tuples extracted from this dataset are temporally correlated and, moreover, their distribution differs from that of the target test data due to both policy mismatch and state distribution shift.

We adopt an experience replay strategy by uniformly sampling $n$ tuples from the offline dataset. Replay is a computational subsampling step and does not itself ensure independence. Denote the resulting subset by $\mathcal{I}_{\mathrm{cal}} = \{(S_{it}, A_{it}, R_{it}, \ldots, S_{i,t+k})\}_{i=1}^{n}$. After subsampling, the sampled tuples may still deviate from both the target policy distribution and the test-state distribution. To correct for this discrepancy, we employ importance weighting. Let $\mathcal{P}_0$ denote the target distribution of rollout segments starting from test states under policy $\pi$, and $\mathcal{P}_{\mathrm{cal}}$ the distribution induced by the behavior policy $\pi_b$ on the calibration data. Our goal is to reweight samples drawn from $\mathcal{P}_{\mathrm{cal}}$ so as to approximate expectations under $\mathcal{P}_0$. The resulting density ratio can be decomposed as

$$\widehat{w}(s_0, a_0, \ldots, s_k) \propto \frac{d\mathcal{P}_0(s_0)}{d\mathcal{P}_{\mathrm{cal}}(s_0)} \prod_{h=0}^{k-1} \frac{\pi(a_h \mid s_h)}{\pi_b(a_h \mid s_h)}. \tag{6}$$

The first term corrects for the mismatch in initial state distributions (a form of covariate shift), while the second term accounts for the difference between the target policy $\pi$ and the behavior policy $\pi_b$. The state density ratio $d\mathcal{P}_0(s_0)/d\mathcal{P}_{\mathrm{cal}}(s_0)$ can be estimated using standard propensity score or density ratio estimation techniques, which can be formulated as a **binary classification problem** distinguishing samples from the two distributions. The behavior policy $\pi_b$ can be estimated from the training data. In the special case where $\pi = \pi_b$ (the on-policy setting), the action-dependent terms cancel, and the weight reduces to the state density ratio.

Let $\hat{w}_{itk} = \widehat{w}(S_{it}, \ldots, S_{i,t+k})$. By combining the probability surrogate introduced earlier with the importance weights, we obtain practically useful estimates of the FDR and power, as follows:

$$\zeta_n(\eta; c) = \frac{\sum_{(i,t)\in\mathcal{I}_{\mathrm{cal}}} \hat{w}_{itk} \mathbb{I}\{R(S_{it}; c) \le \eta\} \widehat{F}\left(\frac{c - \sum_{j=0}^{k-1} \gamma^j R_{i,t+j}}{\gamma^k} \,\middle|\, S_{i,t+k}\right)}{1 \vee \sum_{(i,t)\in\mathcal{I}_{\mathrm{cal}}} \hat{w}_{itk} \mathbb{I}\{R(S_{it}; c) \le \eta\}}, \tag{7}$$

$$\Psi_n(\eta; c) = \frac{\sum_{(i,t)\in\mathcal{I}_{\mathrm{cal}}} \hat{w}_{itk} \mathbb{I}\{R(S_{it}; c) \le \eta\} \left[1 - \widehat{F}\left(\frac{c - \sum_{j=0}^{k-1} \gamma^j R_{i,t+j}}{\gamma^k} \,\middle|\, S_{i,t+k}\right)\right]}{1 \vee \sum_{(i,t)\in\mathcal{I}_{\mathrm{cal}}} \hat{w}_{itk} \left[1 - \widehat{F}\left(\frac{c - \sum_{j=0}^{k-1} \gamma^j R_{i,t+j}}{\gamma^k} \,\middle|\, S_{i,t+k}\right)\right]}. \tag{8}$$

Finally, the selection threshold $\eta$ is chosen by maximizing $\Psi_n(\eta; c)$ subject to the constraint $\zeta_n(\eta; c) \le \alpha$, yielding an FDR-controlled selection rule that accounts for both unobservable returns and distribution mismatch in offline data. To better utilize the available data and improve stability, we repeat this procedure $B$ times and take the median of the resulting thresholds as the final selection threshold. A complete description of this algorithm is provided in Appendix A.2.

### 3.4 DISTRIBUTIONAL RL WITH EMGN RETURN MODELING

Our framework requires an estimate of the conditional return distribution $F(g \mid s)$. This can be obtained using different distributional RL methods. We introduce an energy-based generative network (EMGN), trained by minimizing the energy distance between generated and target samples (Bellemare et al., 2017b). This avoids adversarial training and provides a flexible generative implementation. The main contribution remains the selection and calibration framework, rather than a universally best return estimator. We use EMGN, Nadaraya–Watson, and IQN estimators in the three examples. Details are provided in Appendix A.1.

For an action-conditional generator $H(\epsilon, s, a)$, we obtain the state-conditional CDF by marginalizing over the target policy, $\widehat{F}(g \mid s) = \mathbb{E}_{a\sim\pi(\cdot|s)} \Pr\{H(\epsilon, s, a) \le g\}$. We sum over discrete actions and use Monte Carlo integration for continuous actions.

## 4 THEORETICAL RESULT

In this section, we establish theoretical guarantees of our proposed selection framework. In particular, we show that our likelihood-ratio-based procedure (i) asymptotically controls the FDR, and (ii) achieves optimal statistical power among all valid FDR-controlling procedures. Our analysis is conducted under regularity conditions on the estimated return distribution and the importance weights. Let $F(g \mid s; \boldsymbol{\theta})$ denote a parameterized approximation to the conditional CDF, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ represents the neural network parameters.

ASSUMPTION **1** (**Model Regularity**). *(i) $F(g \mid s; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ over a compact set $\boldsymbol{\Theta}$; (ii) $\widehat{\boldsymbol{\theta}}_n$ converges in probability to some $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$; (iii) for any $\tau \in [0, 1]$, threshold $c$, and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{P}\{F(c \mid S_{it}; \boldsymbol{\theta}) = \tau\} = 0$, where the probability is taken with respect to the distribution of $S_{it}$; and (iv) the oracle return distribution denoted by $F^\pi$ admits a Lebesgue density uniformly bounded by $L$.*

Assumption 1(i) is a standard continuity assumption and Assumption 1(ii) requires the estimated parameter to converge to a limiting value, allowing us to analyze the asymptotic behavior of the selection rule. Assumption 1(iii) excludes ties in the conformity score by ruling out point masses in $F(c \mid S_{it}; \boldsymbol{\theta})$; this ensures that thresholding the score yields a stable ranking of candidate states. Assumption 1(iv) imposes a mild smoothness Assumption on the oracle return distribution.

ASSUMPTION **2** (**Weight Stability**). *Let $P_{\mathrm{cal}}$ denote the probability distribution of the calibration data $\mathcal{D}_{\mathrm{cal}}$. Assume that: (i) $\mathbb{E}_{P_{\mathrm{cal}}}\big[\big|\widehat{w}(S_{it}, \cdots, S_{i,t+k}) - w(S_{it}, \cdots, S_{i,t+k})\big|\big] = o_p(1)$; (ii) $\widehat{w} \in \mathscr{W}$, where $\mathscr{W}$ is a Glivenko–Cantelli class of functions; (iii) the true importance weight is uniformly bounded, i.e., there exists a constant $C_w$ such that $w(s_{it}, \cdots, s_{i,t+k}) \le C_w$ for all $s_{it} \in \mathcal{S}$.*

Assumption 2(i) ensures the consistency of the estimated importance weights, so that the weighted calibration distribution converges to the target distribution. Assumption 2(ii) controls the complexity of the weight estimators through a Glivenko-Cantelli class, enabling uniform convergence of empirical weighted quantities. Assumption 2(iii) rules out extreme importance weights, which is needed to avoid variance explosion under distribution shift. Under these assumptions, we establish the following asymptotic FDR control guarantee.

The first bound allows the limiting return model to be misspecified. Correct specification is required for the exact limit and power result. The i.i.d. test-state assumption concerns independent deployment instances. Further discussion of sampling and weight stability is given in Appendix C.7.

THEOREM **1** (**Asymptotic FDR Control**). *Suppose the test data $\mathcal{D}_{\mathrm{test}}$ are i.i.d., and that Assumptions 1 and 2 hold. We further assume that $\mathbb{P}\{F(c \mid S_0; \boldsymbol{\theta}^*) \le \eta_n^{\mathrm{opt}}\}$ is bounded below by u as $n$ tends to infinity. Let $\alpha \in (0, 1)$ be a target level, and let $\mathcal{S}$ denote the selection set returned by Algorithm 1 (see Appendix A.2). Then,with probability tending to 1*

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{\sum_{j=1}^{m} I\big(j \in \mathcal{S},\, G_j^\pi \le c\big)}{1 \vee |\mathcal{S}|}\right] \le \alpha + \sqrt{2L\gamma^k\, \bar{W}_1(F_{\boldsymbol{\theta}^*}, F^\pi)}$$

*where $\bar{W}_1(F_{\theta^*}, F^\pi)$ denotes the worst-case 1-Wasserstein distance between $F^\pi$ and its estimate $F_{\boldsymbol{\theta}^*}$ (see Appendix B.1). Furthermore, if $F(g \mid s; \boldsymbol{\theta})$ is correctly specified with parameter $\boldsymbol{\theta}^*$, and*

$\eta_n^{\mathrm{opt}} \xrightarrow{p} \eta^*$, *then*

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{\sum_{j=1}^{m} I\big(j \in \mathcal{S},\, G_j^{\pi} \le c\big)}{1 \vee |\mathcal{S}|}\right] = \alpha\left\{1-(1-b)^m\right\}.$$

Theorem 1 bounds the effect of return-distribution error, with temporal decomposition attenuating its contribution through $\gamma^k$. The guarantee is asymptotic and is not distribution-free. Under correct specification, $b$ denotes the limiting selection probability, so $1-(1-b)^m$ is the probability of making at least one selection. This factor approaches one as $m$ increases and can otherwise make the procedure conservative. The following calibration recovers power. Additional interpretation is provided in Appendix C.7.

THEOREM **2** (**Power Recovery**). *Suppose the test data $\mathcal{D}_{\mathrm{test}}$ are i.i.d., and that Assumption 1 holds. Let $\alpha \in (0,1)$ be a prespecified FDR level, and let $\mathcal{S}$ be the output of Algorithm 1 using the calibrated level $\widetilde{\alpha} = \alpha\{1-(1-\widehat{b})^m\}^{-1}$, where $\widehat{b} = n^{-1}\sum_{i=1}^{n} I\{R(S_{it};\widehat{\boldsymbol{\theta}}_n, c) \le \eta_n^{\mathrm{opt}}\}$. Assume $F(g \mid s;\boldsymbol{\theta})$ is correctly specified and the power function $\Psi_n(\eta;\boldsymbol{\theta}, c)$ strictly increases with $\eta$. Then, as $n \to \infty$, this selection method asymptotically maximizes the proportion of valid units selected among all procedures that control the FDR at or below $\alpha$.*

Theorem 2 shows that, after calibration, our procedure fully utilizes the FDR budget and achieves maximal statistical power. This establishes a Neyman–Pearson-type optimality for our likelihood-ratio-based selection rule.

## 5 NUMERICAL EXPERIMENTS

We evaluate our LR-based selection method (Ours) against three baselines: Plug-in BH, which applies BH to estimated failure probabilities, C-BH, which uses conformal p-values based on pseudo-return residuals, and SC-BH, which uses studentized residuals. All methods share the same fitted return distribution within each environment. Baseline definitions are given in Appendix C.4. We consider the following three environments.

**Example 1 (Linear Dynamical Environment).** The state space is two-dimensional and continuous, with a binary action space and linear transition dynamics. In the on-policy setting ($\pi = \pi_b$), the behavior and target policies are identical sigmoid functions. In the off-policy setting ($\pi \neq \pi_b$), we introduce a policy shift by modifying the coefficients of the sigmoid policy, thereby evaluating the robustness of the method under distributional shift. This setup is motivated by the linear-state setting in Scenario B of (Shi et al., 2022).

**Example 2 (Pendulum-v1: Nonlinear Control).** To evaluate the adaptability of our framework in nonlinear dynamical systems with continuous action spaces, we consider a variant of the classic Pendulum-v1 environment. The state space is three-dimensional, and actions correspond to continuous torques. We mix a pre-trained Soft Actor-Critic (SAC) policy with uniform exploration, using different mixture probabilities for the behavior and target policies. See Appendix C.2 for details.

**Example 3 (Synthetic OhioT1DM: Healthcare).** To demonstrate the utility of our method in high-stakes healthcare settings, we simulate a type 1 diabetes management scenario using a 4th-order MDP (Shi et al., 2020) based on the OhioT1DM dataset (Marling & Bunescu, 2020). The state space is 15-dimensional, capturing glucose levels, meal intake, and physical activity, while actions correspond to discretized insulin injections. The reward function penalizes both hyperglycemia and hypoglycemia. The target policy is constructed as a weighted combination of a DQN-derived policy and a uniform random policy, contrasting with a stochastic behavior policy.

**Implementation Details** We use EMGN in Example 1, a Nadaraya–Watson estimator in Example 2, and an Implicit Quantile Network (IQN) in Example 3. These choices illustrate the compatibility of our selection framework with different return estimators, rather than a universally best estimator. We estimate state density ratios and, for discrete actions, behavior policies. Continuous-action importance ratios use the specified policy densities. We set $\alpha = 0.1$ and repeat each experiment over 50 independent runs. Figure 2 reports off-policy comparisons at $k = 2, 2, 4$ for Examples 1–3,

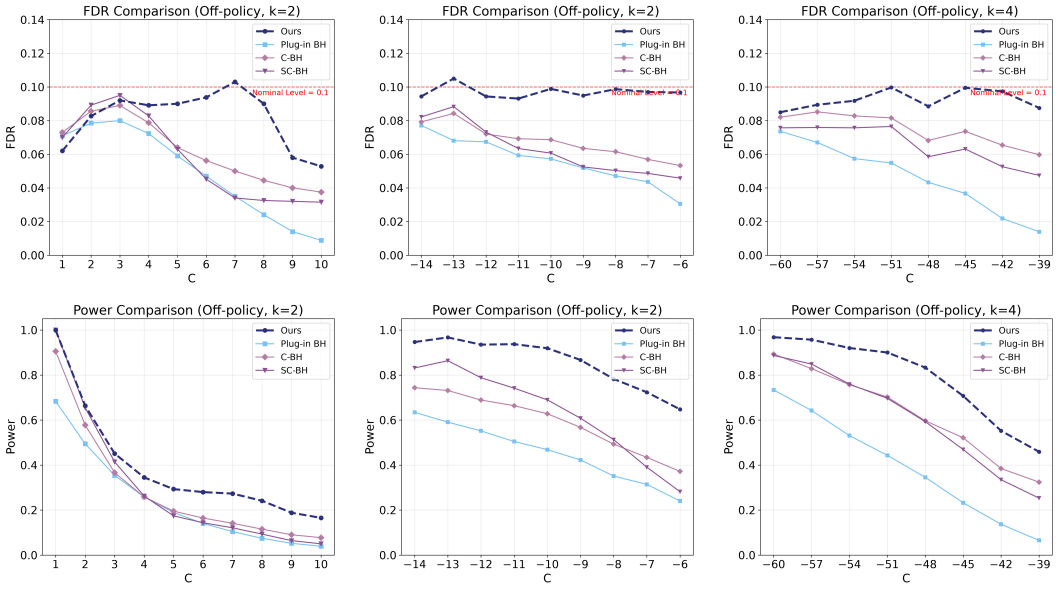


Figure 2: Off-policy comparison of Ours, Plug-in BH, C-BH, and SC-BH. Columns correspond to Example 1 ($k = 2$), Example 2 ($k = 2$), and Example 3 ($k = 4$). The top and bottom rows show empirical FDR and power, respectively, across thresholds $c$. The nominal FDR level is $\alpha = 0.1$.

respectively. Additional horizons, on-policy results, oracle comparisons, and uncertainty summaries are provided in Appendix C.

**Results** Figure 2 shows a consistent power advantage for our method in Examples 2 and 3, with FDR generally close to the nominal level. In Example 1, SC-BH is competitive at the lowest threshold, whereas our method retains more power as $c$ increases. Plug-in BH is particularly conservative at larger thresholds. Thus, the power curves should be read together with achieved FDR, rather than as comparisons at exactly matched error rates.

The additional comparisons in Appendix C.4 show that $k = 0$ can produce FDR inflation, while a longer rollout is not always better, notably at $k = 8$ in Example 3. The oracle and clipping analyses in Appendices C.5 and C.6 further examine return-model error and weight variability. These finite-sample effects are consistent with our asymptotic guarantees.

## 6 Conclusion

We propose a likelihood-ratio-based framework for selective deployment in RL with FDR control. By leveraging distributional RL to estimate the return distribution, our approach constructs selection rules that directly quantify the evidence for desirable outcomes, bypassing the need for p-values. We further develop a calibration procedure to enable data-driven threshold selection, achieving both reliable FDR control and high statistical power. Our guarantees are asymptotic and depend on return-distribution and importance-weight estimation. The experiments illustrate the power benefit of the proposed selection procedure and the finite-sample trade-off in the rollout horizon. Improving tail estimation and extending the dependence analysis are promising directions for future work.

## Ethics Statement

FDR-controlled selective deployment may improve reliability in healthcare and autonomous systems. However, it does not guarantee a favorable outcome for every selected individual. Distribution shift and inaccurate return or weight estimates can affect reliability, so domain expertise, validation, and monitoring remain necessary. The healthcare experiment uses a synthetic simulator and does not establish clinical safety.

## AI Use Statement

Generative AI tools were used to assist with language editing, manuscript restructuring, and LaTeX formatting.

## Reproducibility Statement

Proofs, assumptions, and experimental details are provided in the appendix. The accompanying code specifies random seeds and experimental configurations, including shared calibration budgets and the common nominal level $\alpha$.

## REFERENCES


Tian Bai, Yue Zhao, Xiang Yu, and Archer Y Yang. Multivariate conformal selection. *arXiv preprint arXiv:2505.00917*, 2025.

Yajie Bao, Yuyang Huo, Haojie Ren, and Changliang Zou. Cap: A general algorithm for online selective conformal prediction with fcr control. *Journal of Machine Learning Research*, 26(287): 1–74, 2025.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017a.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. In *arXiv preprint arXiv:1705.10743*, 2017b.

Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1(1):447–464, 2014.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 456–464, 2019.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Wenqin Du, Wanteng Ma, Dong Xia, Yuan Zhang, and Wen Zhou. Conformal link prediction with false discovery rate control. *arXiv preprint arXiv:2507.07025*, 2025.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Feichen Gan, Lu Youcun, Yingying Zhang, and Yukun Liu. Conformal prediction beyond the horizon: Distribution-free inference for policy evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Celi, Emma Brunskill, and Finale Doshi-Velez. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *International Conference on Machine Learning*, pp. 3658–3667. PMLR, 2020.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 37:73884–73919, 2024.

Yu Gui, Ying Jin, Yash Nair, and Zhimei Ren. Acs: An interactive framework for conformal selection. *arXiv preprint arXiv:2507.15825*, 2025.

Qingyang Hao, Wenbo Liao, Bingyi Jing, and Hongxin Wei. Multi-condition conformal selection. *arXiv preprint arXiv:2510.08075*, 2025.

Sungee Hong, Zhengling Qi, and Raymond KW Wong. Distributional off-policy evaluation with bellman residual minimization. *arXiv preprint arXiv:2402.01900*, 2024.

Sungee Hong, Jiayi Wang, Zhengling Qi, and Raymond KW Wong. A principled path to fitted distributional evaluation. *arXiv preprint arXiv:2506.20048*, 2025.

Yuyang Huo, Xiaoyang Wu, Changliang Zou, and Haojie Ren. Unified conformalized multiple testing with full data efficiency. *arXiv preprint arXiv:2508.12085*, 2025.

Ying Jin and Emmanuel J Candès. Model-free selective inference under covariate shift via weighted conformal p-values. *arXiv preprint arXiv:2307.09291*, 2023a.

Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023b.

Olav Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, New York, 2nd edition, 2002.

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.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE transactions on intelligent transportation systems*, 23(6):4909–4926, 2021.

Qi Kuang, Chao Wang, Yuling Jiao, and Fan Zhou. Distributional off-policy evaluation with deep quantile process regression. *arXiv preprint arXiv:2604.18143*, 2026.

Cindy Marling and Razvan Bunescu. The ohiot1dm dataset for blood glucose level prediction: Update 2020. In *CEUR workshop proceedings*, volume 2675, pp. 71, 2020.

Zhengling Qi, Chenjia Bai, Zhaoran Wang, and Lan Wang. Distributional off-policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 120(551):1517–1530, 2025.

Jing Qin, Yukun Liu, Moming Li, and Chiung-Yu Huang. Revamping conformal selection with optimal power: A neyman–pearson perspective. *arXiv preprint arXiv:2502.16513*, 2025.

David Rodbard. Interpretation of continuous glucose monitoring data: glycemic variability and quality of glycemic control. *Diabetes technology & therapeutics*, 11(1_suppl):S–55, 2009.

Nathan Ross. Fundamentals of stein's method. 2011.

Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

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.

Bodhisattva Sen. A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 11:28–29, 2018.

Chengchun Shi, Runzhe Wan, Rui Song, Wenbin Lu, and Ling Leng. Does the markov decision process fit the data: Testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pp. 8807–8817. PMLR, 2020.

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.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 6414–6423. PMLR, 2020.

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

Xiaoning Wang, Yuyang Huo, Liuhua Peng, and Changliang Zou. Conformalized multiple testing after data-dependent selection. *Advances in Neural Information Processing Systems*, 37:58574–58609, 2024.

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.

Xiangyu Zhao, Liang Zhang, Long Xia, Zhuoye Ding, Dawei Yin, and Jiliang Tang. Deep reinforcement learning for list-wise recommendations. *arXiv preprint arXiv:1801.00209*, 2017.

# A Additional Methodological Details

## A.1 Distributional RL for Return Modeling

Under a time-homogeneous MDP and a stationary target policy, the conditional return law given a fixed state-action pair does not depend on time. It may differ across state-action pairs. We train an energy-based generative network (EMGN) using the distributional Bellman equation.

We model this conditional distribution using a generative network

$$H(\epsilon, S_t, A_t),$$

where $\epsilon \in \mathbb{R}^p$ is a $p$-dimensional noise vector. The network takes $(\epsilon, S_t, A_t)$ as input and outputs a sample intended to match the distribution of $G_t^\pi$ conditional on $(S_t, A_t)$. The existence of such a generator is guaranteed by the noise-outsourcing lemma, which underlies GAN-based models (Theorem 5.10 in Kallenberg (2002)).

To train $H$, we leverage the distributional Bellman operator (Bellemare et al., 2017a; 2023; Rowland et al., 2018), which implies

$$G_t^\pi(S_t, A_t) \stackrel{d}{=} R_t + \gamma G_{t+1}^\pi(S_{t+1}, A_{t+1}),$$

where $A_{t+1} \sim \pi(\cdot \mid S_{t+1})$. Accordingly, we define two random variables

$$U = \big(H(\epsilon_t, S_t, A_t), S_t, A_t\big), \quad V = \big(R_t + \gamma H(\epsilon_{t+1}, S_{t+1}, A_{t+1}), S_t, A_t\big),$$

and train $H$ by minimizing the energy distance between $U$ and $V$ (Székely & Rizzo, 2013). For a minibatch of $M$ transitions, let

$$U_i = (H(\epsilon_i, S_i, A_i), S_i, A_i), \qquad V_i = (R_i + \gamma H(\epsilon_i', S_i', A_i'), S_i, A_i),$$

where $A_i' \sim \pi(\cdot \mid S_i')$ and the latent variables are independent standard Gaussian draws. The minibatch loss is

$$L_M = \frac{2}{M^2} \sum_{i,j=1}^{M} \|U_i - V_j\| - \frac{1}{M(M-1)} \sum_{i \neq j} \|U_i - U_j\| - \frac{1}{M(M-1)} \sum_{i \neq j} \|V_i - V_j\|.$$

Bellman target samples $V_i$ are held fixed during each gradient update. In Example 1, distances use the raw two-dimensional states, binary actions, and returns, without coordinate standardization. The energy objective therefore depends on their relative scales.

After training, the learned generator $\widehat{H}$ provides an approximation to the conditional return distribution, so that for a given $(S_t, A_t)$,

$$\Pr\big(\widehat{H}(\epsilon, S_t, A_t) \leq u\big) \approx \Pr\big(G_t^\pi \leq u \mid S_t, A_t\big).$$

For Example 1, we use the proposed EMGN, where the noise variable $\epsilon$ is one-dimensional. For Example 2, we adopt a Nadaraya–Watson (NW) estimator, trained using Monte Carlo returns to obtain a nonparametric approximation of the conditional return distribution. For Example 3, we use an implicit quantile network trained with a quantile Huber loss.

## A.2 Algorithms

We summarize the procedure as follows: Algorithm 1 produces a selection subset based on an estimated threshold. Since this procedure relies on subsampling the calibration data to evaluate FDR and power, it can exhibit substantial variability. To improve stability, we propose Algorithm 2, which repeats this procedure multiple times and takes the median of the resulting thresholds to obtain the final selection.

**Algorithm 1:** Likelihood-Ratio-Based Selection with FDR Control

**Data:** $\mathcal{D} = \{(S_{it}, A_{it}, R_{it}, S_{i,t+1}) : 1 \le i \le N, 1 \le t \le T\}$ and test states $S_{\text{test},j}$, $1 \le j \le m$.

**Input:** $\alpha \in (0,1)$, target FDR; $c$, threshold for hypotheses (2); $\pi$, target policy; $\mathcal{A}$, a conditional return distribution estimator; $\mathcal{B}$, a propensity score estimation algorithm (e.g., neural network with softmax layer); $\mathcal{W}$, a density ratio estimation algorithm (e.g., logistic regression); $k$, rollout horizon; $n$: subsample size.

**Output:** Selection set $\mathcal{S}$

Split the data: $\mathcal{D} = \mathcal{D}_{\text{tr}} \bigcup \mathcal{D}_{\text{cal}}$ where $\mathcal{D}_{\text{tr}}$ is used for training and $\mathcal{D}_{\text{cal}} = \{(S_{it}, A_{it}, R_{it}, \ldots, S_{i,t+k}) : (i,t) \in \widetilde{\mathcal{I}}_{\text{cal}}\}$ is the calibration replay buffer;

Fit $\widehat{F}(y \mid s)$ using $\mathcal{A}$ and the training data, marginalizing over target-policy actions when needed;

Estimate $\widehat{\pi}_b$ on the training data using $\mathcal{B}$, unless its probabilities or densities are known;

Fit the state density ratio using training replay states and an independent reference sample from the deployment initial-state distribution. Multiply by the action ratios in Eq. (6) and normalize the buffer weights as described in Appendix C.7;

Sample $n$ data tuples $\{(S_{it}, A_{i,t}, R_{i,t}, \ldots, S_{i,t+k}) : (i,t) \in \mathcal{I}_{\text{cal}}\}$ from $\mathcal{D}_{\text{cal}}$.

**for** *each* $(i,t) \in \mathcal{I}_{\text{cal}}$ **do**

  Compute the conformity score: $\eta_{\text{calc}}^{(i,t)} = R(S_{it}; c) := \frac{\widehat{F}(c|S_{it})}{1-\widehat{F}(c|S_{it})}$;

Obtain a set of candidate critical values $E_n = \left\{\eta_{\text{calc}}^{(i,t)} : \zeta_n\left(\eta_{\text{calc}}^{(i,t)}; c\right) \le \alpha, (i,t) \in \mathcal{I}_{\text{cal}}\right\}$ using the weighted FDR estimation (7) with rollout horizon $k$;

Determine the optimal critical value:

**if** $E_n \neq \varnothing$ **then**

  Set $\eta_n^{\text{opt}} = \arg\max_{\eta \in E_n} \Psi_n(\eta; c)$ where $\Psi_n(\eta; c)$ is the power estimation (8) with rollout horizon $k$;

**else**

  Set $\eta_n^{\text{opt}} = -\infty$;

**Result:** $\mathcal{S} = \{j : R(S_{\text{test},j}; c) \le \eta_n^{\text{opt}}, 1 \le j \le m\}$

**Algorithm 2:** Aggregated Likelihood-Ratio-based Selection with FDR Control

**Data:** $\mathcal{D} = \{(S_{it}, A_{it}, R_{it}, S_{i,t+1}) : 1 \le i \le N, 1 \le t \le T\}$ and test states $S_{\text{test},j}$, $1 \le j \le m$.

**Input:** Inputs from Algorithm 1 $(\alpha, c, \pi, \mathcal{A}, \mathcal{B}, \mathcal{W}, k, n)$; $B$, the number of subsampling repetitions.

**Output:** Selection set $\mathcal{S}$.

Split the data into $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{cal}}$;

Fit the return and weight models as in Algorithm 1, using the same training data and independent initial-state reference sample. Keep these models and the normalized buffer weights fixed across repetitions;

**for** $b = 1, \ldots, B$ **do**

  Construct calibration set $\mathcal{D}_{\text{cal}}^{(b)}$ by uniformly subsampling indices $\mathcal{I}_{\text{cal}}^{(b)}$ from $\widetilde{\mathcal{I}}_{\text{cal}}$;

  Compute conformity scores $\eta_{\text{calc}}^{(i,t)} = R(S_{it}; c)$ for each $(i,t) \in \mathcal{I}_{\text{cal}}^{(b)}$ as in Algorithm 1;

  Calculate the optimal critical value $\eta_{n,b}^{\text{opt}}$ using $\mathcal{I}_{\text{cal}}^{(b)}$ following the weighted FDR calculation and power maximization procedure in Algorithm 1;

Aggregate the optimal critical values using the median: $\widetilde{\eta}_n^{\text{opt}} = \text{Median}\left(\eta_{n,1}^{\text{opt}}, \eta_{n,2}^{\text{opt}}, \ldots, \eta_{n,B}^{\text{opt}}\right)$;

**Result:** $\mathcal{S} = \{j : R(S_{\text{test},j}; c) \le \widetilde{\eta}_n^{\text{opt}}, 1 \le j \le m\}$

# B TECHNICAL DETAILS

## B.1 PROOF OF THEOREM 1

In this section, we provide the proof of Theorem 1. We denote by $\mathbb{E}_{\mathrm{te}}$ and $\mathbb{E}_{\mathrm{ca}}$ the expectations taken with respect to the distributions of the test data and calibration data, respectively.

*Proof of Theorem 1.* Recall that we are given a collection of test states $\{S_{\mathrm{test},j}\}_{j=1}^{m}$. Our procedure selects states whose likelihood-ratio score $R(S_{\mathrm{test},j}; c)$ is smaller than the threshold $\eta_n^{\mathrm{opt}}$. Since

$$R(S; c) = \frac{F(c \mid S; \boldsymbol{\theta})}{1 - F(c \mid S; \boldsymbol{\theta})}$$

is a strictly increasing function of $F(c \mid S; \boldsymbol{\theta})$, replacing $R(S; c)$ by $F(c \mid S; \boldsymbol{\theta})$ does not alter the resulting selection set. For notational simplicity, we therefore work directly with $F(c \mid S; \boldsymbol{\theta})$ in the proofs.

Define

$$\zeta_{\mathrm{te},m}(\eta; \boldsymbol{\theta}, c) = \frac{\sum_{j=1}^{m} I\{F(c \mid S_{\mathrm{test},j}; \boldsymbol{\theta}) \le \eta,\, G_j^{\pi} \le c\}}{1 \vee \sum_{j=1}^{m} I\{F(c \mid S_{\mathrm{test},j}; \boldsymbol{\theta}) \le \eta\}}.$$

Since the selection set is given by

$$\mathcal{S} = \left\{ j : F(c \mid S_{\mathrm{test},j}; \widehat{\boldsymbol{\theta}}_n) \le \eta_n^{\mathrm{opt}},\, 1 \le j \le m \right\},$$

we can rewrite the false discovery proportion as

$$\frac{\sum_{j=1}^{m} I\{j \in \mathcal{S},\, G_j^{\pi} \le c\}}{1 \vee |\mathcal{S}|} = \zeta_{\mathrm{te},m}(\eta_n^{\mathrm{opt}}; \widehat{\boldsymbol{\theta}}_n, c).$$

Therefore, it suffices to show that

$$\limsup_{n \to \infty} \mathbb{E}\big[\zeta_{\mathrm{te},m}(\eta_n^{\mathrm{opt}}; \widehat{\boldsymbol{\theta}}_n, c)\big] \le \alpha + \Lambda(\boldsymbol{\theta}^*, c),$$

where $\Lambda(\boldsymbol{\theta}^*, c) = \sqrt{2L\gamma^k\, \mathbb{E}\left[\bar{W}_1(F_{\boldsymbol{\theta}^*}, F^{\pi})\right]}$ and $\bar{W}_1(F_{\boldsymbol{\theta}^*}, F^{\pi})$ is defined in (12).

**(1)** To this end, we first define the set of candidate thresholds satisfying the estimated FDR constraint:

$$E_n = \left\{ \eta_i : \zeta_{\mathrm{ca},n}(\eta_i; \widehat{\boldsymbol{\theta}}_n, c) \le \alpha,\ i = 1, \ldots, n \right\},$$

where

$$\zeta_{\mathrm{ca},n}(\eta; \boldsymbol{\theta}, c) = \frac{\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\left[ \widehat{w}(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} F\{(c - \sum_{t=0}^{k-1} \gamma^t R_t)/\gamma^k \mid S_k; \boldsymbol{\theta}\} \right]}{\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\left[ \widehat{w}(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} \right]}.$$

Here, $\mathbb{P}_{\mathrm{ca},n}$ denotes the probability measure induced by uniformly sampling $n$ observations from the calibration dataset, denoted by $\mathcal{I}_{\mathrm{cal}}$ in the main text, and $\mathcal{H}_{0:k} = (S_0, A_0, R_0, \cdots, S_k)$.

**(2)** Second, we define the bridge statistic

$$\zeta(\eta; \boldsymbol{\theta}, c) = \frac{\mathbb{E}_{\mathrm{te}}\big[ I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta,\, G^{\pi} \le c\} \big]}{\mathbb{E}_{\mathrm{te}}\big[ I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} \big]},$$

where $S_0$ has the same distribution as the test states $S_{\mathrm{test},j}$. We further define the auxiliary threshold set

$$E = \left\{ \eta : \zeta(\eta; \boldsymbol{\theta}^*, c) \le \alpha + \Lambda(\boldsymbol{\theta}^*, c) \right\} \cap \left\{ F(c \mid S; \boldsymbol{\theta}^*) : S \in \mathcal{S} \right\}.$$

**(3)** By Fact 1 (proved at the end of this proof), the difference between the empirical and population functions is uniformly bounded:

$$\sup_{\eta \in [\underline{\eta}, 1]} \left| \zeta_{\mathrm{ca},n}(\eta; \widehat{\boldsymbol{\theta}}_n, c) - \zeta(\eta; \boldsymbol{\theta}^*, c) \right| \le \Lambda(\boldsymbol{\theta}^*, c) + o_p(1). \tag{9}$$

where $\underline{\eta} := \inf_{\eta'}\{\eta' : \mathbb{P}_{\text{te}}\{F(c|S_0;\boldsymbol{\theta}^*) \le \eta'\} \ge u\} > 0$ and $u$ is defined in the assumption of Theorem 1. By definition, $\eta_n^{\text{opt}} \ge \underline{\eta}$. Moreover, by Assumption 1 and the Dominated Convergence Theorem, $\zeta(\eta;\boldsymbol{\theta},c)$ is continuous in $(\eta,\boldsymbol{\theta})$. Because the functions converge uniformly up to the boundary $\Lambda(\boldsymbol{\theta}^*,c)$, as $n\to\infty$, every element of $E_n$ converges in probability to some element in $E$.

**(4)** Now we are ready to prove the first result. By (3), each $\eta_{ni}$ in the set $E_n$ converges to some element $\eta^*$ in $E$ in probability. Because $|\zeta_{\text{te},m}(\eta;\boldsymbol{\theta},c)| \le 1$ and $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}^* + o_p(1)$, by the Dominated Convergence Theorem, we have

$$\lim_{n\to\infty} \mathbb{E}\big[\zeta_{\text{te},m}(\eta_{ni};\widehat{\boldsymbol{\theta}}_n,c)\big] = \mathbb{E}\big[\zeta_{\text{te},m}(\eta^*;\boldsymbol{\theta}^*,c)\big].$$

Because $\cup_{n=1}^{\infty} E_n \subset [0,1]$ and $[0,1]$ is a compact set, $\eta_n^{\text{opt}}$ must have a convergent subsequence, say $\eta_{n_k}^{\text{opt}}$. Its limit, say $\eta_k^{**}$, must belong to $E\cap[\underline{\eta},1]$. By the above result, we have

$$\lim_{k\to\infty} \mathbb{E}\big[\zeta_{\text{te},m}(\eta_{n_k}^{\text{opt}};\widehat{\boldsymbol{\theta}}_n,c)\big] = \mathbb{E}\big[\zeta_{\text{te},m}(\eta_k^{**};\boldsymbol{\theta}^*,c)\big].$$

Taking into consideration all the convergent subsequences of $\eta_n^{\text{opt}}$, we have

$$\limsup_{n\to\infty} \mathbb{E}\big[\zeta_{\text{te},m}(\eta_n^{\text{opt}};\widehat{\boldsymbol{\theta}}_n,c)\big] \le \sup_{\eta\in E\cap[\underline{\eta},1]} \mathbb{E}\big[\zeta_{\text{te},m}(\eta;\boldsymbol{\theta}^*,c)\big].$$

Finally, by Corollary 1 in Qin et al. (2025), we have

$$\mathbb{E}\big[\zeta_{\text{te},m}(\eta;\boldsymbol{\theta}^*,c)\big] \le \zeta(\eta;\boldsymbol{\theta}^*,c),$$

and thus, by the definition of $E$:

$$\sup_{\eta\in E\cap[\underline{\eta},1]} \mathbb{E}\big[\zeta_{\text{te},m}(\eta;\boldsymbol{\theta}^*,c)\big] \le \alpha + \Lambda(\boldsymbol{\theta}^*,c).$$

Therefore we have with probability tending to 1

$$\limsup_{n\to\infty} \mathbb{E}\big[\zeta_{\text{te},m}(\eta_n^{\text{opt}};\widehat{\boldsymbol{\theta}}_n,c)\big] \le \sup_{\eta\in E\cap[\underline{\eta},1]} \mathbb{E}\big[\zeta_{\text{te},m}(\eta;\boldsymbol{\theta}^*,c)\big] \le \alpha + \Lambda(\boldsymbol{\theta}^*,c)$$

This proves the first result of Theorem 1.

**(5)** Note that $\zeta(\eta;\boldsymbol{\theta}^*,c)$ is continuous in $\eta$. For the second result, if $\eta_n^{\text{opt}}$ converges in probability to a constant $\eta^{\text{opt}} \in E\cap[\underline{\eta},1]$ and $\boldsymbol{\theta}^*$ is the true value, then $\eta^{\text{opt}}$ must be the solution to $\zeta(\eta;\boldsymbol{\theta}^*,c)=\alpha$ since $F^* = F^{\pi}$ implies $\Lambda(\boldsymbol{\theta}^*,c)=0$. Therefore, by the Dominated Convergence Theorem, we have

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}_{\text{te}}\big[\zeta_{\text{te},m}(\eta_n^{\text{opt}};\widehat{\boldsymbol{\theta}}_n,c)\big] = \mathbb{E}_{\text{te}}\big[\zeta_{\text{te},m}(\eta^{\text{opt}};\boldsymbol{\theta}^*,c)\big] \\
&= \frac{\mathbb{P}_{\text{te}}\{F(c\mid S_0;\boldsymbol{\theta}^*) \le \eta^{\text{opt}}, G^{\pi} \le c\}}{\mathbb{P}_{\text{te}}\{F(c\mid S_0;\boldsymbol{\theta}^*) \le \eta^{\text{opt}}\}} \times \big[1-(1-\mathbb{P}_{\text{te}}\{F(c\mid S_0;\boldsymbol{\theta}^*) \le \eta^{\text{opt}}\})^m\big] \\
&= a \times \big[1-(1-\mathbb{P}_{\text{te}}\{F(c\mid S_0;\boldsymbol{\theta}^*) \le \eta^{\text{opt}}\})^m\big].
\end{aligned}$$

**Fact 1.** To prove (9), we define two functions

$$\bar{\zeta}_{\text{ca},n}(\eta;\boldsymbol{\theta},c) = \frac{\mathbb{E}_{\mathbb{P}_{\text{ca},n}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta}) \le \eta\}F\{(c-\sum_{t=0}^{k-1}\gamma^t R_t)/\gamma^k \mid S_k;\boldsymbol{\theta}\}\Big]}{\mathbb{E}_{\mathbb{P}_{\text{ca},n}}\big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta}) \le \eta\}\big]},$$

$$\widetilde{\zeta}^k(\eta;\boldsymbol{\theta},c) = \frac{\mathbb{E}_{\text{te}}\big[I\{F(c\mid S_0;\boldsymbol{\theta}) \le \eta\}F\{(c-\sum_{t=0}^{k-1}\gamma^t R_t)/\gamma^k \mid S_k;\boldsymbol{\theta}\}\big]}{\mathbb{E}_{\text{te}}\big[I\{F(c\mid S_0;\boldsymbol{\theta}) \le \eta\}\big]}.$$

Note that

$$\begin{aligned}
|\zeta_{\text{ca},n}(\eta;\widehat{\boldsymbol{\theta}}_n,c) - \zeta(\eta;\boldsymbol{\theta}^*,c)| \le& |\zeta_{\text{ca},n}(\eta;\widehat{\boldsymbol{\theta}}_n,c) - \bar{\zeta}_{\text{ca},n}(\eta;\widehat{\boldsymbol{\theta}}_n,c)| + \Big|\bar{\zeta}_{\text{ca},n}(\eta;\widehat{\boldsymbol{\theta}}_n,c) - \widetilde{\zeta}^k(\eta;\boldsymbol{\theta}^*,c)\Big| \\
&+ \Big|\widetilde{\zeta}^k(\eta;\boldsymbol{\theta}^*,c) - \zeta(\eta;\boldsymbol{\theta}^*,c)\Big| \\
:=& M_1(\eta) + M_2(\eta) + M_3(\eta).
\end{aligned}$$

We prove (9) by bounding $M_1(\eta)$, $M_2(\eta)$, and $M_3(\eta)$ uniformly over $\eta \in [\underline{\eta}, 1]$:

**(1)** By Lemma 7, we bound $M_1(\eta)$:

$$M_1(\eta) \leq \frac{2 \cdot \mathbb{E}_{\mathrm{ca}}\big[|w(\mathcal{H}_{0:k}) - \widehat{w}(\mathcal{H}_{0:k})|\big]}{\Pr\{F(c \mid S; \widehat{\boldsymbol{\theta}}_n) \leq \eta\} - \mathbb{E}_{\mathrm{ca}}\big[|w(\mathcal{H}_{0:k}) - \widehat{w}(\mathcal{H}_{0:k})|\big]} = o_p(1),$$

by the consistency of $\widehat{w}$.

**(2)** By Lemma 6, we bound $M_2(\eta)$:

$$\sup_{\eta \in [0,1]} \Bigg| \mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}} \Bigg[ w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \leq \eta\} F\{(c - \sum_{t=0}^{k-1} \gamma^t R_t)/\gamma^k \mid S_k; \widehat{\boldsymbol{\theta}}_n\} \Bigg] \\ - \mathbb{E}_{\mathrm{te}} \Big[ I\{F(c \mid S_0; \boldsymbol{\theta}^*) \leq \eta\} F(\widetilde{G}(c,k) | S_k; \boldsymbol{\theta}^*) \Big] \Bigg| = o_p(1),$$

$$\sup_{\eta \in [0,1]} \Big| \mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}} \big[ w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \leq \eta\} \big] - \mathbb{E}_{\mathrm{te}} \big[ I\{F(c \mid S_0; \boldsymbol{\theta}^*) \leq \eta\} \big] \Big| = o_p(1).$$

Therefore, $\sup_{\eta \in [0,1]} M_2(\eta) = o_p(1)$.

**(3)** By Lemma 8, we bound $M_3(\eta)$:

$$\sup_{\eta \in [\underline{\eta}, 1]} M_3(\eta) \leq \sqrt{2L\gamma^k \bar{W}_1(F_{\boldsymbol{\theta}^*}, F^\pi)}. \tag{10}$$

By combining the uniform bounds on $M_1(\eta)$, $M_2(\eta)$, and $M_3(\eta)$, we have

$$\sup_{\eta \in [\underline{\eta}, 1]} \big| \zeta_{\mathrm{ca},n}(\eta; \widehat{\boldsymbol{\theta}}_n, c) - \zeta(\eta; \boldsymbol{\theta}^*, c) \big| \leq \Lambda(\boldsymbol{\theta}^*, c) + o_p(1).$$

This completes the proof of (9). □

### B.2 Auxiliary Lemmas

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

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

Next, we define the distributional Bellman operator $\mathcal{T}^\pi$ as the following transformation:

$$(\mathcal{T}^\pi \eta^\pi)(s) = \mathcal{P}^\pi(R + \gamma G^\pi(S') \mid s), \tag{11}$$

where the transition $(s, R, S')$ is generated by sampling an action according to policy $\pi$, observing the reward $R$, and subsequently transitioning to the next state $S'$, with $G^\pi(S') \sim \eta^\pi(S')$. Under the time-homogeneity assumption, $\eta^\pi$ satisfies the following fixed-point condition:

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

A core property of the distributional Bellman operator $\mathcal{T}^\pi$ is that it is a $\gamma$-contraction with respect to the Wasserstein distance (detailed in Proposition 1). The $p$-Wasserstein distance between two measures $\mu$ and $\nu$ 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 $\Gamma(\mu, \nu)$ denotes the set of all couplings with marginal distributions $\mu$ and $\nu$. To introduce this property, we first review some fundamental results regarding the Wasserstein distance.

Next, we state the contraction property of the distributional Bellman operator $\mathcal{T}^\pi$. For any $\mu, \mu' \in \mathscr{P}^{\mathcal{S}}$, we denote

$$\bar{W}_p(\mu, \mu') := \sup_{s \in \mathcal{S}} W_p(\mu(s), \mu'(s)). \tag{12}$$

Let $\mathscr{P}$ represent the set of all probability distributions over $\mathbb{R}$.

PROPOSITION **1** (Bellemare et al., 2023, Proposition 4.15)**.** *The distributional Bellman operator is a $\gamma$-contraction on $\mathscr{P}^{\mathcal{S}}$ with respect to the supremum $p$-Wasserstein metric for $p \in [1, \infty)$. That is, for any $\mu, \mu' \in \mathscr{P}^{\mathcal{S}}$, we have $\bar{W}_p(\mathcal{T}^\pi \mu, \mathcal{T}^\pi \mu') \leq \gamma \bar{W}_p(\mu, \mu')$.*

Inspired by the insights in Xu et al. (2025), and unlike traditional approaches that rely on the total variation distance, we adopt the Wasserstein distance. The Kolmogorov distance serves as a crucial intermediary linking the coverage error and the Wasserstein distance, and is defined as follows:

DEFINITION **1** (Kolmogorov Distance)**.** *Let $F_\mu$ and $F_\nu$ be the cumulative distribution functions of probability measures $\mu$ and $\nu$ on $\mathbb{R}$, respectively. The Kolmogorov distance between $\mu$ and $\nu$ is defined as:*

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

LEMMA **1** ((Ross, 2011))**.** *If a probability measure $\mu$ on $\mathbb{R}$ has a Lebesgue density bounded by $L$, then for any probability measure $\nu$, it holds that $K(\mu, \nu) \leq \sqrt{2LW_1(\mu, \nu)}$.*

LEMMA **2** ((Qin et al., 2025), Lemma 1)**.** *Let $\{(A_i, B_i)\}_{i=1}^n$ be $n$ independent bivariate random vectors and $A_i, B_i \in \{0, 1\}$. Then for any positive integer $m$, it holds that*

$$\mathbb{E}\left(\frac{\sum_{j=1}^m A_j B_j}{1 \vee \sum_{i=1}^m A_i}\right) \leq \max_{1 \leq k \leq m} \frac{\mathbb{E}(A_k B_k)}{\mathbb{E}(A_k)}.$$

*If $\mathbb{E}(A_k B_k)/\mathbb{E}(A_k) = a$ for all $1 \leq k \leq m$, then*

$$\mathbb{E}\left(\frac{\sum_{j=1}^m A_j B_j}{1 \vee \sum_{i=1}^m A_i}\right) = a \times \left[1 - \prod_{i=1}^m \{1 - \mathbb{E}(A_i)\}\right].$$

Denote the local trajectory segment following $\pi_b$ as $\mathcal{H}_{t:t+k} := (S_t, A_t, \ldots, S_{t+k})$. Let $\mathbb{P}_n$ denote the empirical probability measure of $\{\mathcal{H}_{i,t:t+k}\}_{(i,t) \in \mathcal{I}_{cal}}$ and $P_{\mathrm{ca}}$ denote the mixture of probability measure of the selected tuple in different times $P_{\mathrm{ca}} = \frac{t}{T-k} \sum_{t=0}^{T-k} P_t(\mathcal{H}_{t:t+k})$, respectively, with some abuse of notation, and let $P_{\mathrm{te}}$ denote the test distribution following $\pi$. Further denote

$$\widetilde{G}(c, k) := (c - \sum_{t=0}^{k-1} \gamma^t R_t)/\gamma^k.$$

LEMMA **3** ((Qin et al., 2025), Lemma 2)**.** *Under Assumption 1, the function class*

$$\mathcal{G} = \{I(\eta - F(c \mid S; \boldsymbol{\theta}) \geq 0) : \ (\eta, \boldsymbol{\theta}) \in [0, 1] \times \Theta\}$$

*is Glivenko-Cantelli (GC) for any $P_t(S)$, $t = 1, \ldots, T$.*

LEMMA **4.** *Recall that $k$ denotes the step. For any $w \in \mathscr{W}$,under Assumption 1 and 2, we have*

$$\sup_{\eta \in [0,1]} \left| \mathbb{E}_{\mathbb{P}_{cal}}\left[w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \leq \eta\} F(\widetilde{G}(c, k) | S_k; \widehat{\boldsymbol{\theta}}_n)\right] \right.$$
$$\left. - \ \mathbb{E}_{\mathrm{te}}\left[I\{F(c \mid S_0; \boldsymbol{\theta}^*) \leq \eta\} F(\widetilde{G}(c, k) | S_k; \boldsymbol{\theta}^*)\right] \right| = o_p(1),$$
$$\sup_{\eta \in [0,1]} \left| \mathbb{E}_{\mathbb{P}_{cal}}\left[w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \leq \eta\}\right] - P_{\mathrm{te}}\{F(c \mid S_0; \boldsymbol{\theta}^*) \leq \eta\} \right| = o_p(1).$$

*Proof.* Note that

$$\sup_{\eta \in [0,1]} \left| \mathbb{E}_{\mathbb{P}_{cal}}\left[w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \leq \eta\} F(\widetilde{G}(c, k) | S_k; \widehat{\boldsymbol{\theta}}_n)\right] \right.$$
$$\left. - \ \mathbb{E}_{\mathrm{ca}}\left[w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \boldsymbol{\theta}^*) \leq \eta\} F(\widetilde{G}(c, k) | S_k; \boldsymbol{\theta}^*)\right] \right| = \ D_{n1} + D_{n2},$$

where

$$D_{n1} = \sup_{\eta\in[0,1]} \left| (\mathbb{E}_{\mathbb{P}_{cal}} - \mathbb{E}_{\mathrm{ca}}) \left[ w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \le \eta\} F(\widetilde{G}(c,k) | S_k; \widehat{\boldsymbol{\theta}}_n) \right] \right|,$$

$$D_{n2} = \sup_{\eta\in[0,1]} \left| \mathbb{E}_{\mathrm{ca}} \left[ w(\mathcal{H}_{0:k}) \left\{ I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \le \eta\} F(\widetilde{G}(c,k) | S_k; \widehat{\boldsymbol{\theta}}_n) \right.\right.\right.$$
$$\left.\left.\left. - I\{F(c \mid S_0; \boldsymbol{\theta}^*) \le \eta\} F(\widetilde{G}(c,k) | S_k; \boldsymbol{\theta}^*) \right\} \right] \right|.$$

Note that $\mathbb{P}_n$ is not an empirical probability measure of $P_{\mathrm{ca}}$, while after splitting the probability measure of different time, we can reconstruct $\mathbb{P}_n$ as $\mathbb{P}_{cal} = \frac{1}{T-k} \sum_{t=0}^{T-k} \mathbb{P}_t(S_t, A_t, R_t, ..., S_{t+k})$ where $\mathbb{P}_t$ is an empirical probability measure of $P_t$, thus

$$D_{n1} \le \sum_{t=0}^{T-k} \frac{1}{T-k} \sup_{\eta\in[0,1]} \left| (\mathbb{E}_{\mathbb{P}_t} - \mathbb{E}_{P_t}) \left[ w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \le \eta\} F(\widetilde{G}(c,k) | S_k; \widehat{\boldsymbol{\theta}}_n) \right] \right|.$$

Let $\mathscr{H} = \{f = w \cdot F \cdot I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} : w \in \mathscr{W}, F \in \mathscr{G}\}$. $I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\}$ is GC by Lemma 3 and $\mathscr{W}$ is also GC by Assumption 2. As $\mathscr{G}$ are pointwise compact classes with integrable envelope (Sen, 2018), $\mathscr{H}$ is GC for any $P_t(S)$. So $D_{n1} = o_p(1)$.

By Assumption 1(i), the Dominated Convergence Theorem and Assumption 2(ii),

$$\mathbb{E}_{\mathrm{ca}} \left[ w(\mathcal{H}_{0:k}) \cdot \left\{ I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \le \eta\} F(\widetilde{G}(c,k) | S_k; \widehat{\boldsymbol{\theta}}_n) \right.\right.$$
$$\left.\left. - I\{F(c \mid S_0; \boldsymbol{\theta}^*) \le \eta\} F(\widetilde{G}(c,k) | S_k; \boldsymbol{\theta}^*) \right\} \right]$$

is uniformly continuous in $(\eta, \boldsymbol{\theta})$. This together with the continuous mapping theorem and the consistency of $\widehat{\boldsymbol{\theta}}_n$ implies $D_{n2} = o_p(1)$. Therefore, $D_{n1} + D_{n2} = o_p(1)$.

The first result is proved by noting that

The second result can be proved in the same way; the details are omitted. □

With the same notation as defined above and given calibration data $\mathcal{D}_{\mathrm{cal}}$ of size $N_{cal}$, let $\mathbb{P}_{cal}$ denote the corresponding empirical probability measure induced by $\mathcal{D}_{\mathrm{cal}}$ and $\mathbb{P}_{ca,n}$ denote the probability measure induced by bootstrap sample from $\mathbb{P}_{cal}$ of size $n$.

LEMMA **5.** *Under Assumption 1 & 2, for any $\delta > 0$, there exist $r_\delta$, s.t. with probability at least $1 - \delta$*

$$\limsup_{n\to\infty} \sup_{\eta\in[0,1], \boldsymbol{\theta}\in B_{r_\delta}(\boldsymbol{\theta}^*)} \left| (\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}} - \mathbb{E}_{\mathbb{P}_{cal}}) \left[ w(\mathcal{H}_{0:k}) \cdot I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} F(\widetilde{G}(c,k) | S_k; \boldsymbol{\theta}) \right] \right|$$
$$\le \frac{C_w}{N_{cal}}$$

$$\limsup_{n\to\infty} \sup_{\eta\in[0,1], \boldsymbol{\theta}\in B_{r_\delta}(\boldsymbol{\theta}^*)} \left| (\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}} - \mathbb{E}_{\mathbb{P}_{cal}}) \left[ w(\mathcal{H}_{0:k}) \cdot I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} \right] \right|$$
$$\le \frac{C_w}{N_{cal}}$$

*Proof.* Note that $B_{r_\delta}(\boldsymbol{\theta}^*)$ stands for a ball of radius $r_\delta$ centered at $\boldsymbol{\theta}^*$. We denote the initial state of samples in $\mathcal{D}_{\mathrm{cal}}$ as $S_i, i \in [N_{\mathrm{cal}}]$, and we denote the corresponding conditional cumulative probability function $F(c \mid S_i; \boldsymbol{\theta})$ as $\eta_i$ with distinct values a.s., since $\sup_{(\tau,\boldsymbol{\theta})\in[0,1]\times\Theta} P\big(F(c \mid S; \boldsymbol{\theta}) = \tau\big) = 0$

and $F(g \mid S; \boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$. So we can always find $r$ s.t. sets of $F(c \mid S_i; \boldsymbol{\theta} \in B_r(\boldsymbol{\theta}^*))$ are separated in $\mathbb{R}$. Thus as a direct corollary, for any $\delta > 0$, there exists $r_\delta$, s.t. with probability (w.r.t. $\mathcal{D}_{\text{cal}}$) at least $1-\delta$, $\sup_{\boldsymbol{\theta} \in B_{r_\delta}(\boldsymbol{\theta}^*), \eta \in [0,1]} P_{N_{\text{cal}}}\big(F(c \mid S; \boldsymbol{\theta}) = \tau\big) = \frac{1}{N_{\text{cal}}}$.

Define $m(S, c \mid \boldsymbol{\theta}, \eta) = \widehat{w}(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \boldsymbol{\theta}) \leq \eta\} F(\widetilde{G}(c, k) \mid S_k; \boldsymbol{\theta})$. For an open set $B \subset \mathbb{R}^{q+1}$, define

$$
\begin{aligned}
m_B(S, c) &= \inf\{m(S, c \mid \eta, \boldsymbol{\theta}) : (\eta, \boldsymbol{\theta}) \in B\}, \\
m^B(S, c) &= \sup\{m(S, c \mid \eta, \boldsymbol{\theta}) : (\eta, \boldsymbol{\theta}) \in B\}.
\end{aligned}
$$

Then $[m_B, m^B]$ constitutes a bracket. Given a sequence of balls $B_k$ with center a given $\eta$ and radii decreasing to 0, we must have

$$
\lim_{k \to \infty} (m^{B_k} - m_{B_k}) \subset w \cdot \{I(\eta - F(c \mid S; \boldsymbol{\theta}) = 0)\} \leq C_w \cdot \{I(\eta - F(c \mid S; \boldsymbol{\theta}) = 0)\}.
$$

Note that $0 \leq m_{B_k} \leq m^{B_k} \leq C_w$ is uniformly bounded. Taking expectation w.r.t. $P_{N_{\text{cal}}}$, by the Dominated Convergence Theorem, we have

$$
\begin{aligned}
\lim_{k \to \infty} \mathbb{P}_{\text{cal}}\big[|m^{B_k} - m_{B_k}|\big] &= \mathbb{P}_{\text{cal}}\left[\lim_{k \to \infty} |m^{B_k} - m_{B_k}|\right] \\
&\leq C_w \cdot \mathbb{P}_{\text{cal}}\big[I(\eta - F(c \mid S; \boldsymbol{\theta}) = 0)\big] \\
&= C_w \cdot P_{N_{\text{cal}}}\big(F(c \mid S; \boldsymbol{\theta}) = \eta\big) \\
&\leq \frac{C_w}{N_{\text{cal}}}.
\end{aligned}
$$

Therefore, given $\epsilon > 0$, for every $(\eta, \boldsymbol{\theta}) \in [0,1] \times B_{r_\delta}(\boldsymbol{\theta}^*)$, there exists a ball $B(\eta, \boldsymbol{\theta})$ around this $(\eta, \boldsymbol{\theta})$ such that the bracket $[m_{B(\eta,\boldsymbol{\theta})}, m^{B(\eta,\boldsymbol{\theta})}]$ has size at most $\frac{C_w}{N_{\text{cal}}} + \epsilon$. By the compactness of $[0,1] \times B_{r_\delta}(\boldsymbol{\theta}^*)$, the collection of balls constructed in this way has a finite cover, that is, there exist $K > 0$ and $(\eta_i, \boldsymbol{\theta}_i), 1 \leq i \leq K$, such that $\mathscr{G} \subset \cup_{i=1}^K [m_{B(\eta_i,\boldsymbol{\theta}_i)}, m^{B(\eta_i,\boldsymbol{\theta}_i)}]$.

In summary, for each $\epsilon > 0$ there exists a bracket of size at most $\frac{C_w}{N_{\text{cal}}} + \epsilon$ that covers $\mathscr{G}$. Following a similar derivation as in Theorem 2.4.1 of van de Vaart and Wellner (1996), we assert with probability (w.r.t. $\mathcal{D}_{\text{cal}}$) at least $1-\delta$,

$$
\begin{aligned}
\limsup_{r \to \infty} \sup_{\eta \in [0,1], \boldsymbol{\theta} \in B_{r_\delta}(\boldsymbol{\theta}^*)} \left| (\mathbb{E}_{\mathbb{P}_{\text{ca},n}} - \mathbb{E}_{\mathbb{P}_{\text{ca}}}) \left[ \widehat{w}(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \boldsymbol{\theta}) \leq \eta\} F(\widetilde{G}(c, k) \mid S_k; \boldsymbol{\theta}) \right] \right| \\
\leq \frac{C_w}{N_{\text{cal}}}.
\end{aligned}
$$

The second result can be proved in the same way; the details are omitted. □

LEMMA **6.** *With the same notation as defined above, as $n \to \infty$,*

$$
\begin{aligned}
\sup_{\eta \in [0,1]} \Bigg| \mathbb{E}_{\mathbb{P}_{\text{ca},n}} \Big[ w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \leq \eta\} F(\widetilde{G}(c, k) \mid S_k; \widehat{\boldsymbol{\theta}}_n) \Big] \\
- \mathbb{E}_{\text{te}} \Big[ I\{F(c \mid S_0; \boldsymbol{\theta}^*) \leq \eta\} F(\widetilde{G}(c, k) | S_k; \boldsymbol{\theta}^*) \Big] \Bigg| = o_p(1),
\end{aligned}
$$

$$
\sup_{\eta \in [0,1]} \Big| \mathbb{E}_{\mathbb{P}_{\text{ca},n}} \big[ w(\mathcal{H}_{0:k}) I\{F(c \mid S_0; \widehat{\boldsymbol{\theta}}_n) \leq \eta\} \big] - \mathbb{E}_{\text{te}} \big[ I\{F(c \mid S_0; \boldsymbol{\theta}^*) \leq \eta\} \big] \Big| = o_p(1).
$$

*Proof.* It can be seen that

$$
\begin{aligned}
&\sup_{\eta\in[0,1]}\Bigg|\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\\
&\qquad-\mathbb{E}_{\mathrm{te}}\Big[I\{F(c\mid S_0;\boldsymbol{\theta}^*)\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta}^*)\Big]\Bigg|\\
\le&\sup_{\eta\in[0,1]}\Bigg|(\mathbb{P}_{\mathrm{cal}}-\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}})\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\Bigg|\\
&+\sup_{\eta\in[0,1]}\Bigg|\mathbb{P}_{\mathrm{cal}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\\
&\qquad-\mathbb{E}_{\mathrm{te}}\Big[I\{F(c\mid S_0;\boldsymbol{\theta}^*)\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta}^*)\Big]\Bigg|\\
\le&\sup_{\eta\in[0,1]}\Bigg|(\mathbb{P}_{\mathrm{cal}}-\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}})\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\Bigg|\\
&+\sup_{\eta\in[0,1]}\Bigg|\mathbb{E}_{\mathbb{P}_{cal}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\\
&\qquad-\mathbb{E}_{\mathrm{te}}\Big[I\{F(c\mid S_0;\boldsymbol{\theta}^*)\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta}^*)\Big]\Bigg|\\
=&\sup_{\eta\in[0,1]}\Bigg|(\mathbb{P}_{\mathrm{cal}}-\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}})\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\Bigg|+o_p(1).
\end{aligned}
$$

Let $\Omega:=\{\omega$ s.t. the 1st inequality in Lemma 5 holds true$\}$. The former term above is bounded by:

$$
\begin{aligned}
&\le I\{\omega\in\Omega\}\cdot\sup_{\eta\in[0,1]}\Bigg|(\mathbb{P}_{\mathrm{cal}}-\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}})\Big[w(\mathcal{H}_{0:k})I\{d(\boldsymbol{\theta}^*,\widehat{\boldsymbol{\theta}}_n)\le r_\delta\}\\
&\qquad\cdot I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\Bigg|\\
&+\sup_{\eta\in[0,1]}\Bigg|(\mathbb{P}_{\mathrm{cal}}-\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}})\Big[w(\mathcal{H}_{0:k})I\{d(\boldsymbol{\theta}^*,\widehat{\boldsymbol{\theta}}_n)> r_\delta\}\\
&\qquad\cdot I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\Bigg|\\
&+I\{\omega\notin\Omega\}\cdot\sup_{\eta\in[0,1]}\Bigg|(\mathbb{P}_{\mathrm{cal}}-\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}})\Big[w(\mathcal{H}_{0:k})I\{d(\boldsymbol{\theta}^*,\widehat{\boldsymbol{\theta}}_n)\le r_\delta\}\\
&\qquad\cdot I\{F(c\mid S_0;\widehat{\boldsymbol{\theta}}_n)\le\eta\}F(\widetilde{G}(c,k)\mid S_k;\widehat{\boldsymbol{\theta}}_n)\Big]\Bigg|.
\end{aligned}
$$

The first term $\Delta_1$ is $o_p(1)$ by Lemma 2. The second term $\Delta_2\le 2C_w\cdot P(d(\boldsymbol{\theta}^*,\widehat{\boldsymbol{\theta}}_n)>r_\delta)$ is $o_p(1)$. For the third term $\Delta_3\le 2C_w\cdot I\{\omega\notin\Omega\}$, letting $\delta\to 0$ gives $P(\omega\notin\Omega)\to 0$, so it is also $o_p(1)$.

The second result can be proved in the same way; the details are omitted. □

LEMMA **7.** *With the same notation as defined above, we further define*

$$
\zeta_{\mathrm{ca},n}(\eta;\boldsymbol{\theta},c)=\frac{\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[\widehat{w}(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta})\Big]}{\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\big[\widehat{w}(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\big]}
$$

*and*

$$\bar{\zeta}_{\mathrm{ca},n}(\eta;\boldsymbol{\theta},c)=\frac{\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta})\Big]}{\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\big]}.$$

*With probability tending to 1 as n goes to infinity, we have*

$$|\zeta_{\mathrm{ca},n}(\eta;\boldsymbol{\theta},c)-\bar{\zeta}_{\mathrm{ca},n}(\eta;\boldsymbol{\theta},c)|\le\frac{2\cdot\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]}{\Pr\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}-\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]}:=\Lambda_{\widehat{w}}(\eta;\boldsymbol{\theta},c)$$

*Proof.* Regarding the numerator,

$$\begin{aligned}
&\Big|\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[\widehat{w}(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta})\Big]\\
&\qquad-\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta})\Big]\Big|\\
&\le\Big|(\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}-\mathbb{E}_{\mathrm{ca}})\Big[\widehat{w}(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta})\Big]\Big|\\
&\quad+\Big|(\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}-\mathbb{E}_{\mathrm{ca}})\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta})\Big]\Big|\\
&\quad+\Big|\mathbb{E}_{\mathrm{ca}}\Big[(w-\widehat{w})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}F(\widetilde{G}(c,k)|S_k;\boldsymbol{\theta})\Big]\Big|.
\end{aligned}$$

The 3rd term can be bounded by $\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]$. Analogous to Lemma 3, the first term and second term is $o_p(1)$. Similarly, for the deniminator,

$$\begin{aligned}
&\Big|\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[\widehat{w}(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\Big]-\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\Big]\Big|\\
&\quad\le\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]+o_p(1).
\end{aligned}$$

As a result, $\mathbb{E}I\{F(c\mid S;\boldsymbol{\theta})\le\eta,\widetilde{G}^{(k)}\le c\}$

$$\begin{aligned}
&\min\{\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[\widehat{w}(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\Big],\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\Big]\}\\
&\ge\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\Big]-\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]\\
&\ge\mathbb{E}_{\mathrm{te}}\big[I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\big]-o_p(1)-\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]\quad\text{(similar to Lemma 6).}
\end{aligned}$$

Since $|\frac{a}{b}-\frac{c}{d}|\le\frac{2\cdot\max\{|a-c|,|b-d|\}}{\min\{b,d\}}$ for $a,b,c,d>0$, as $n\to\infty$, with probability tending to 1 we have

$$\begin{aligned}
&|\zeta_{\mathrm{ca},n}(\eta;\boldsymbol{\theta},c)-\bar{\zeta}_{\mathrm{ca},n}(\eta;\boldsymbol{\theta},c)|\\
&\le\frac{2\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]+o_p(1)}{\min\{\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[\widehat{w}(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\Big],\mathbb{E}_{\mathbb{P}_{\mathrm{ca},n}}\Big[w(\mathcal{H}_{0:k})I\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}\Big]\}}\\
&\le\frac{4\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]}{\Pr\{F(c\mid S_0;\boldsymbol{\theta})\le\eta\}-\mathbb{E}_{\mathrm{ca}}\big[|w-\widehat{w}|\big]}
\end{aligned}$$

□

LEMMA **8.** *Recall that we denote the learned DRL model in the proposed prediction procedure by $\eta^{\pi}_{\boldsymbol{\theta}}(s)$, whose conditional cumulative distribution function is parameterized as $F(\cdot \mid S_0; \boldsymbol{\theta})$. With the same notation as defined above, we further define*

$$\widetilde{\zeta}^k(\eta; \boldsymbol{\theta}, c) = \frac{\mathbb{E}_{\mathrm{te}}[I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} F(\widetilde{G}(c, k) | S_k; \boldsymbol{\theta})\}]}{\mathbb{E}_{\mathrm{te}}[I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\}]},$$
$$\zeta(\eta; \boldsymbol{\theta}, c) = \frac{\mathbb{E}_{\mathrm{te}} I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta, G \le c\}}{\mathbb{E}_{\mathrm{te}} I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\}}.$$

*Then,*

$$\left|\widetilde{\zeta}^k(\eta; \boldsymbol{\theta}, c) - \zeta(\eta; \boldsymbol{\theta}, c)\right| \le \sqrt{2L\gamma^k \bar{W}_1(\eta^{\pi}_{\boldsymbol{\theta}}, \eta^{\pi})} := \Lambda_{\widehat{F}}(\eta; \boldsymbol{\theta}, c)$$

*Proof.* Let $((\mathcal{T}^{\pi})^k F_{\boldsymbol{\theta}})(\cdot \mid S_0)$ and $((\mathcal{T}^{\pi})^k F_{\boldsymbol{\theta}}$ denote the conditional and marginal probability measure, respectively, obtained by applying the distributional Bellman operator $k$ times to the learned model $F_{\boldsymbol{\theta}}$ at the initial state $S_0$. We define its cumulative distribution function (CDF) as $F^{(k)}(\cdot \mid S_0; \boldsymbol{\theta})$. By the definition of the distributional Bellman equations (Bellemare et al., 2023), evaluating this CDF at $c$ is equivalent to taking the expectation of the model's CDF after $k$ steps of environment transitions:

$$F^{(k)}(c \mid S_0; \boldsymbol{\theta}) = \mathbb{E}_{\mathrm{te}}\left[F(\widetilde{G}(c, k) \mid S_k; \boldsymbol{\theta}) \mid S_0\right].$$

Recall that $F^{\pi}$ is the true return distribution. Since $F^{\pi}$ is the fixed point of the Bellman operator, we have the identity $F^{\pi} = (\mathcal{T}^{\pi})^k F^{\pi}$.

Using the law of total expectation, we can express the numerators of $\widetilde{\zeta}^k$ and $\zeta$ conditioned on the initial state $S_0$:

$$\mathbb{E}_{\mathrm{te}}\left[I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} F(\widetilde{G}(c, k) \mid S_k; \boldsymbol{\theta})\right] = \mathbb{E}_{\mathrm{te}}\left[I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} F^{(k)}(c \mid S_0; \boldsymbol{\theta})\right],$$
$$\mathbb{E}_{\mathrm{te}}\left[I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta, G \le c\}\right] = \mathbb{E}_{\mathrm{te}}\left[I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} F_G(c \mid S_0)\right].$$

The absolute difference between the empirical conditional coverage $\widetilde{\zeta}^k$ and the true conditional coverage $\zeta$ can thus be bounded by the expected difference between these two CDFs:

$$\left|\widetilde{\zeta}^k(\eta; \boldsymbol{\theta}, c) - \zeta(\eta; \boldsymbol{\theta}, c)\right|$$
$$\le \frac{1}{\mathbb{E}_{\mathrm{te}}\left[I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\}\right]} \mathbb{E}_{\mathrm{te}}\left[I\{F(c \mid S_0; \boldsymbol{\theta}) \le \eta\} \left|F^{(k)}(c \mid S_0; \boldsymbol{\theta}) - F_G(c \mid S_0)\right|\right].$$

Here, the term $\Delta$ is exactly bounded by the Kolmogorov distance between the distributions $((\mathcal{T}^{\pi})^k \eta^{\pi}_{\boldsymbol{\theta}})(S_0)$ and $\eta^{\pi}(S_0)$. Applying Lemma 1 and the contraction property from Proposition 1, we have:

$$\begin{aligned}
\Delta &= \left|F^{(k)}(c \mid S_0; \boldsymbol{\theta}) - F^{\pi}(c \mid S_0)\right| \\
&\le K\left(((\mathcal{T}^{\pi})^k F_{\boldsymbol{\theta}})(\cdot \mid S_0), F^{\pi}(\cdot \mid S_0)\right) && \text{(by Definition 1)} \\
&\le \sqrt{2LW_1\left(((\mathcal{T}^{\pi})^k F^{\pi}_{\boldsymbol{\theta}})(\cdot \mid S_0), F^{\pi}(\cdot \mid S_0)\right)} && \text{(by Lemma 1)} \\
&\le \sqrt{2L\bar{W}_1\left((\mathcal{T}^{\pi})^k F_{\boldsymbol{\theta}}, (\mathcal{T}^{\pi})^k F^{\pi}\right)} && \text{(by (12))} \\
&\le \sqrt{2L\gamma^k \bar{W}_1(F_{\boldsymbol{\theta}}, F^{\pi})} && \text{(by Proposition 1).}
\end{aligned}$$

Therefore

$$\left|\widetilde{\zeta}^k(\eta; \boldsymbol{\theta}, c) - \zeta(\eta; \boldsymbol{\theta}, c)\right| \le \sqrt{2L\gamma^k \bar{W}_1(F_{\boldsymbol{\theta}}, F^{\pi})}.$$

□

### B.3 PROOF OF THEOREM 2

*Proof of Theorem 2.* We show in the proof of Theorem 1 that $\zeta_{\mathrm{ca},n}(\eta;\widehat{\boldsymbol{\theta}}_n,c)$ converges to $\widetilde{\zeta}^k(\eta;\boldsymbol{\theta}^*,c) = \mathbb{E}_{\mathrm{te}}\big[I\{F(c \mid S_0;\boldsymbol{\theta}^*) \le \eta, G^\pi \le c\}\big]/\mathbb{E}_{\mathrm{te}}\big[I\{F(c \mid S_0;\boldsymbol{\theta}^*) \le \eta\}\big]$. Since we assume $F(g \mid s;\boldsymbol{\theta})$ is correctly specified, thus $\widetilde{\zeta}^k(\eta;\boldsymbol{\theta}^*,c) = \zeta(\eta;\boldsymbol{\theta}^*,c)$. This, along with $E_n = \{\eta : \zeta_{\mathrm{ca},n}(\eta;\widehat{\boldsymbol{\theta}}_n,c) \le \alpha\} \cap \{\eta_i : 1 \le i \le n\}$, implies that the selection set $E_n$ converges in probability to $E = \{\eta : \zeta(\eta;\boldsymbol{\theta}^*,c) \le \alpha\} \cap \{F(c \mid S_0;\boldsymbol{\theta}^*) : S_0 \in \mathcal{S}\}$ as $n \to \infty$. When the power function $\Psi_n(\eta;\boldsymbol{\theta},c)$ is increasing in $\eta$ for each $\boldsymbol{\theta}$ and $c$, one can further show that $\eta_n^{\mathrm{opt}} = \sup\{\eta : \eta \in E_n\}$ converges in probability to $\eta^{\mathrm{opt}} = \sup\{\eta : \eta \in E\}$. Thus Theorem 1 implies that the asymptotic FDR of our method with nominal FDR level $\alpha$ is given by

$$h(\alpha) = \alpha\,\{1-(1-b)^m\} \quad \text{with } b = \Pr\{F(c \mid S_0;\boldsymbol{\theta}^*) \le \eta^{\mathrm{opt}}\}.$$

When the nominal FDR level is set to $\widetilde{\alpha} = \alpha\left\{1-(1-\widehat{b})^m\right\}^{-1}$ with $\widehat{b} = (1/n)\sum_{i=1}^n I\{F(c \mid S_{i0};\widehat{\boldsymbol{\theta}}_n) \le \eta_n^{\mathrm{opt}}\}$, the asymptotic FDR becomes $h(\alpha \times \{1-(1-b)^m\}^{-1}) = \alpha$, thereby fully utilizing the target FDR level $\alpha$. Therefore the limit power of our method is

$$\mathrm{Power}(c) = \mathbb{E}_{\mathrm{te}}\left[\frac{\sum_{j=1}^m I\{F(c \mid S_{\mathrm{test},j};\boldsymbol{\theta}^*) \le \eta^{\mathrm{opt}}, G_j^\pi > c\}}{1 \vee \sum_{j=1}^m I\{G_j^\pi > c\}}\right],$$

where $\eta^{\mathrm{opt}}$ is the maximum point satisfying

$$\frac{\Pr\{F(c \mid S_0;\boldsymbol{\theta}^*) \le \eta^{\mathrm{opt}}, G^\pi \le c\}}{\Pr\{F(c \mid S_0;\boldsymbol{\theta}^*) \le \eta^{\mathrm{opt}}\}} = \alpha. \tag{13}$$

Let the selected set by a selection rule be $\mathcal{T} = \{j : S_{\mathrm{test},j} \in T_n\}$. Suppose that $T_n$ converges in probability to $T^*$. If this is not the case, we replace $T_n$ with one of its convergent subsequences. Suppose that this selection rule has asymptotic FDR control, i.e.,

$$\alpha \ge \limsup_{n\to\infty} \mathbb{E}_{\mathrm{te}}\left[\frac{\sum_{j=1}^m I\{j \in \mathcal{T}, G_j^\pi \le c\}}{1 \vee |\mathcal{T}|}\right] = \frac{\Pr(S_0 \in T^*, G^\pi \le c)}{\Pr(S_0 \in T^*)}. \tag{14}$$

The power of this selection method is

$$\mathrm{Power}_{n1}(c) = \mathbb{E}_{\mathrm{te}}\left[\frac{\sum_{j=1}^m I\{j \in \mathcal{T}, G_j^\pi > c\}}{1 \vee \sum_{j=1}^m I\{G_j^\pi > c\}}\right] = \mathbb{E}_{\mathrm{te}}\left[\frac{\sum_{j=1}^m I\{S_{\mathrm{test},j} \in T_n, G_j^\pi > c\}}{1 \vee \sum_{j=1}^m I\{G_j^\pi > c\}}\right],$$

which, as $n \to \infty$, converges to

$$\mathrm{Power}_1(c) = \mathbb{E}_{\mathrm{te}}\left[\frac{\sum_{j=1}^m I\{S_{\mathrm{test},j} \in T^*, G_j^\pi > c\}}{1 \vee \sum_{j=1}^m I\{G_j^\pi > c\}}\right].$$

Direct calculation gives

$$\mathrm{Power}(c) - \mathrm{Power}_1(c) = \sum_{k=1}^m \mathbb{E}_{\mathrm{te}}\left[\frac{I\{F(c \mid S_{\mathrm{test},k};\boldsymbol{\theta}^*) \le \eta^{\mathrm{opt}}\} - I\{S_{\mathrm{test},k} \in T^*\}}{1 \vee \sum_{j=1}^m I\{G_j^\pi > c\}} \cdot I\{G_k^\pi > c\}\right].$$

Because all the test data are i.i.d., by symmetry, we have

$$\mathrm{Power}(c) - \mathrm{Power}_1(c) = m\mathbb{E}_{\mathrm{te}}\left[\frac{I\{F(c \mid S_{\mathrm{test},1};\boldsymbol{\theta}^*) \le \eta^{\mathrm{opt}}\} - I\{S_{\mathrm{test},1} \in T^*\}}{1 \vee \sum_{j=1}^m I\{G_j^\pi > c\}} \cdot I\{G_1^\pi > c\}\right].$$

Because the term in the square brackets in the above equation is nonzero only if $G_1^\pi > c$, it follows that

$$\mathrm{Power}(c) - \mathrm{Power}_1(c) = m\mathbb{E}_{\mathrm{te}}\left[\frac{1}{1 + \sum_{j\ne 1}^m I\{G_j^\pi > c\}}\right] \times \Pr(G^\pi > c) \times \Delta,$$

where

$$\Delta = \mathbb{E}_{\mathrm{te}}\big[I\{F(c \mid S_0;\boldsymbol{\theta}^*) \le \eta^{\mathrm{opt}}\} - I\{S_0 \in T^*\} \mid G^\pi > c\big].$$

To prove this theorem, it suffices to show $\Delta \geq 0$ for any $\mathcal{T}$.

As the conditional density function of $S_0$ given $G^\pi > c$ is $g_1(s;c)$, we have

$$\Delta = \int_{S^*} g_1(s;c)\,ds - \int_{T^*} g_1(s;c)\,ds,$$

where $S^* = \{s : F(c \mid s;\boldsymbol{\theta}^*) \leq \eta^{\text{opt}}\}$. Because $g_0(s;c) = F(c \mid s)g(s)/F_G(c)$ and $g_1(s;c) = (1 - F(c \mid s))g(s)/(1 - F_G(c))$, the ratio $g_0(s;c)/g_1(s;c)$ is an increasing function of $F(c \mid s;\boldsymbol{\theta}^*)$. We may write

$$S^* = \{s : F(c \mid s;\boldsymbol{\theta}^*) \leq \eta^{\text{opt}}\} = \{s : g_0(s;c)/g_1(s;c) \leq r^*\}$$

for some $r^* > 0$. It is clear that

$$0 \leq \{I\{s \in S^*\} - I\{s \in T^*\}\}\{r^* g_1(s;c) - g_0(s;c)\}, \quad \forall s.$$

Taking integration on both sides of the above inequality, we have

$$\begin{aligned} 0 &\leq \int \{I\{s \in S^*\} - I\{s \in T^*\}\}\{r^* g_1(s;c) - g_0(s;c)\}\,ds \\ &= r^* \left\{\int_{S^*} g_1(s;c)\,ds - \int_{T^*} g_1(s;c)\,ds\right\} - \left\{\int_{S^*} g_0(s;c)\,ds - \int_{T^*} g_0(s;c)\,ds\right\}, \end{aligned}$$

which implies

$$\Delta = \int_{S^*} g_1(s;c)\,ds - \int_{T^*} g_1(s;c)\,ds \geq \frac{1}{r^*}\left\{\int_{S^*} g_0(s;c)\,ds - \int_{T^*} g_0(s;c)\,ds\right\}.$$

In the above inequality, if the equality holds, then

$$0 = \{I\{s \in S^*\} - I\{s \in T^*\}\}\{r^* g_1(s;c) - g_0(s;c)\}, \quad \forall s,$$

or equivalently $S^* = T^*$ almost surely. Therefore the selection rule determined by $S^*$ is the same powerful as that determined by $T^*$, and the result to be proved holds to be true.

Next we consider the other case, where

$$\Delta = \int_{S^*} g_1(s;c)\,ds - \int_{T^*} g_1(s;c)\,ds > \frac{1}{r^*}\left\{\int_{S^*} g_0(s;c)\,ds - \int_{T^*} g_0(s;c)\,ds\right\}, \tag{15}$$

i.e., the strict inequality holds.

Using the definitions of $g_0(s;c)$ and $g_1(s;c)$, from (13) and (14), we have

$$\begin{aligned} \frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \leq c)} \int_{S^*} g_1(s;c)\,ds &= \int_{S^*} g_0(s;c)\,ds, \\ -\frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \leq c)} \int_{T^*} g_1(s;c)\,ds &\leq -\int_{T^*} g_0(s;c)\,ds. \end{aligned} \tag{16}$$

Putting them into (15), we have

$$\begin{aligned} \Delta &> \frac{1}{r^*}\frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \leq c)}\left\{\int_{S^*} g_1(s;c)\,ds - \int_{T^*} g_1(s;c)\,ds\right\} \\ &= \frac{1}{r^*}\frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \leq c)}\Delta, \end{aligned}$$

which is equivalent to

$$\left\{1 - \frac{1}{r^*}\frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \leq c)}\right\}\Delta > 0. \tag{17}$$

To prove $\Delta > 0$, as (17) is a strict inequality, we need only prove

$$1 - \frac{1}{r^*}\frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \leq c)} \geq 0,$$

which is equivalent to

$$\frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \le c)} \le r^*.$$

By the definition of $S^* = \{s : g_0(s;c)/g_1(s;c) \le r^*\}$, we have

$$\int_{S^*} g_0(s;c)\, ds \le r^* \int_{S^*} g_1(s;c)\, ds,$$

which is equivalent to

$$0 \le \int \{r^* g_1(s;c) - g_0(s;c)\} \times I\{r^* g_1(s;c) - g_0(s;c) \ge 0\}\, ds.$$

This together with equality (16) implies that

$$\frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \le c)} \int_{S^*} g_1(s;c)\, ds \le r^* \int_{S^*} g_1(s;c)\, ds.$$

We must have $\int_{S^*} g_1(s;c)\, ds > 0$. (If $\int_{S^*} g_1(s;c)\, ds = 0$, then $\int_{S^*} g_0(s;c)\, ds = 0$ by (16). It follows that $\int I\{r^* g_1(s;c) - g_0(s;c) \ge 0\}\{r^* g_1(s;c) - g_0(s;c)\}\, ds = 0$, which implies that $r^* = 1$ and $g_1(s;c) = g_0(s;c)$. This contradicts $\int_{S^*} g_0(s;c)\, ds = 0$.) Therefore

$$\frac{\alpha \Pr(G^\pi > c)}{(1-\alpha)\Pr(G^\pi \le c)} \le r^*,$$

which together with inequality (17) implies $\Delta > 0$ and $\mathrm{Power}(c) - \mathrm{Power}_1(c) > 0$. This proves Theorem 2. □

# C Experimental Details

In this section, we provide implementation details and supplementary experiments, including baseline comparisons, oracle diagnostics, and importance-weight sensitivity.

The completed experiments, including baseline and clipping comparisons, use 50 independent runs. We summarize FDR and power by their mean and standard deviation across runs. The $B = 20$ replay repetitions are performed within each run to obtain the median selection threshold and are distinct from independent experimental runs. Within each run, all methods share the return estimator and test instances. Test returns are used only for evaluation. Calibration subsamples are drawn without replacement, with tied scores processed as a single group and an empty selection returned if no threshold is feasible.

## C.1 Details for Example 1

**Dynamics and reward.** We use the legacy branch of `Shi.transition`. Let $S_t = (S_{t,1}, S_{t,2})$, $A_t \in \{0, 1\}$, $d_t = 2A_t - 1$, and $r_t = \|S_t\|_2$. The initial state is $S_0 \sim \mathcal{N}(0_2, I_2)$, and

$$S_{t+1,1} = 0.9 d_t S_{t,1} + z_{t,1}, \qquad S_{t+1,2} = -0.9 d_t S_{t,2} + z_{t,2}, \qquad z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_2, 0.2 I_2).$$

The reward is evaluated at the current state, before the transition:

$$R_t = \begin{cases} 3 + 0.5 \exp(r_t), & r_t > 1.5,\ S_{t,1} \geq 0, \\ -3 - 0.8 r_t, & r_t > 1.5,\ S_{t,1} < 0, \\ 1.5 r_t - 0.5 d_t, & 0.5 < r_t \leq 1.5, \\ 0.5 r_t - 0.25 d_t, & r_t \leq 0.5. \end{cases}$$

Transitions are linear conditional on the action, whereas rewards depend asymmetrically and nonlinearly on the state. The identifier `Shi` in code and figure filenames denotes this adapted environment, not an exact reproduction of Scenario B in Shi et al. (2022). The separate non-legacy branch is not used for this setting.

**Policy formulation.** Write $\sigma(x) = (1 + e^{-x})^{-1}$ and

$$\pi_\lambda(1 \mid s) = \lambda \sigma(s_1) + (1 - \lambda)\sigma(s_2).$$

The behavior policy has $\lambda_b = 0.5$. On-policy sampling uses $\lambda = 0.5$ while $\lambda = 0.7$ is set for its off-policy target.. They share the legacy dynamics above but represent different policy shifts. A matched comparison requires the same target-policy coefficient for every method. The modular runner requires `shi_variant=legacy` or `-legacy-shi`, since its stored full-run preset defaults to the non-legacy branch.

**Configuration.** We consider a dataset of $N = 1200$ trajectories, each of length $T = 30$, with discount factor $\gamma = 0.8$. The data are randomly split into a training set and a calibration set with $N_{\text{train}} = N_{\text{cal}} = 600$. The training data are used to estimate the return distribution $\widehat{F}$ and the weight function, while the calibration data are used to evaluate FDR and power and to select the optimal threshold. In each of $B = 20$ repetitions, we subsample 4000 calibration segments and maximize estimated power subject to the estimated FDR constraint. The median threshold is evaluated on 10,000 independent test instances. Examples 2 and 3 use the same calibration and evaluation protocol, with 200 training and 200 calibration trajectories each.

(1) **Estimation of the return distribution.** We train a generator $H$ to approximate the conditional return distribution. The generator is parameterized as a deep MLP that takes as input the state $s \in \mathbb{R}^2$, action $a \in \{0, 1\}$, and a latent noise variable $\epsilon \sim \mathcal{N}(0, 1)$, forming a 4-dimensional input. The network architecture is $4 \to 256 \to 256 \to 32 \to 1$, with LeakyReLU activations (slope 0.1) applied after each hidden layer. The generator is trained by minimizing the energy distance (Appendix A.1) using RMSprop with learning rate $1 \times 10^{-4}$, batch size 256, for 64 epochs. To improve generalization, we apply global unstructured $L_1$ pruning with a progressive schedule, increasing sparsity from 1% to 10% after a 60% warm-up phase. After training, we compute $\widehat{F}(g \mid s)$ using 500 generator samples and marginalize over actions under the target policy.

(2) **Estimation of the weight function.** We estimate the state density ratio with a binary classifier having hidden layers of sizes 64 and 32, using balanced samples from the deployment initial-state distribution and training replay states. A separate classifier estimates the behavior action probabilities. Both estimators use training data only.

### C.2 Details for Example 2

**Dynamics:** We consider the standard Pendulum environment. The 3-dimensional state is given by $(\cos\theta, \sin\theta, \dot{\theta})$, and the action is a continuous torque in $[-2, 2]$. The transition dynamics follow Pendulum-v1. We use the reward $R_t = -0.1(\theta_t^2 + 0.1\dot{\theta}_t^2 + 0.001A_t^2)$, with $\theta_t \in [-\pi, \pi]$, so return thresholds are expressed on this scaled reward axis.

**Policy Formulation:** We mix the stochastic SAC policy $\pi_{\mathrm{SAC}}$ with the uniform density on $[-2, 2]$:

$$\pi_\epsilon(a \mid s) = (1-\epsilon)\pi_{\mathrm{SAC}}(a \mid s) + \epsilon/4.$$

Off-policy data use $\epsilon_b = 0.4$, while the target policy uses $\epsilon = 0.2$. In the on-policy setting, both policies use $\epsilon = 0.2$. Action importance ratios are computed from these mixture densities, including the change-of-variables factor for the squashed Gaussian SAC policy.

**Configuration:** The sample size is $N = 400$ with horizon $T = 100$, and the discount factor is $\gamma = 0.99$. We approximate the conditional return distribution $\widehat{F}(g \mid s)$ using a non-parametric Nadaraya–Watson (NW) estimator based on Monte Carlo rollouts.

(1) **Estimation of the return distribution.** The NW estimator uses target-policy Monte Carlo returns generated from reference states in the training trajectories. Because these rollouts follow the target policy, their returns do not require behavior-to-target importance weighting. Calibration segments still use importance weights. This implementation requires simulator access for return estimation. We standardize reference states, use a Gaussian kernel, and choose the bandwidth by cross-validation grouped by training trajectory. A small uniform contribution to the kernel weights shrinks the estimate toward the empirical reference-return distribution when local kernel mass is small.

(2) **Estimation of the weight function.** A binary classifier with hidden layers of sizes 64 and 32 estimates the state density ratio. In the off-policy setting, we multiply this ratio by the product of target-to-behavior action density ratios over the observed segment.

### C.3 Details for Example 3

**Dynamics:** The environment is modeled as a fourth-order MDP following Shi et al. (2020):

$$GL_{0,t} = \alpha + \sum_{i=1}^{4}(\beta_i^T S_{0,t-i} + c_i A_{0,t-i}) + \epsilon_{0,t}$$

where the simulator uses historical states and actions with coefficients based on OhioT1DM, augmented by Gaussian noise $\epsilon_t \sim \mathcal{N}(0, 9)$. Patient behaviors (meal consumption and exercise) occur with specific probabilities and magnitudes drawn from normal distributions. The 15-dimensional state tracks glucose, carbs, exercise, and past actions. The reward $R_t$ uses a modified Index of Glycemic Control (Rodbard, 2009), penalizing states where $GL_t < 80$ or $GL_t > 140$, and returning 0 within the safe target range.

**Policy Formulation:** The off-policy behavior policy samples the five insulin actions with probabilities $(0.799, 0.083, 0.065, 0.026, 0.027)$. The target policy selects the action maximizing a trained Deep Q-Network (DQN) with probability $1 - \epsilon$ and a uniformly sampled action with probability $\epsilon = 0.1$. On-policy data are generated by this target policy.

**Configuration:** We generate an offline dataset consisting of $N = 400$ trajectories, each of length $T = 168$ (one week of hourly observations). The discount factor is $\gamma = 0.9$.

(1) **Policy estimation and off-policy setup.** For off-policy calibration, we estimate $\widehat{\pi}_b$ from the training split by behavioral cloning. The policy is parameterized by an MLP with hidden layers of size 128 and 64, trained using cross-entropy loss with the Adam optimizer. The target policy $\pi$ is

derived from a Deep Q-Network (DQN) with architecture $15 \to 256 \to 128 \to 5$, and employs an $\epsilon$-greedy exploration strategy with $\epsilon = 0.1$.

(2) **Estimation of the return distribution.** To model the conditional return distribution, we employ an Implicit Quantile Network (IQN). Here, $H$ denotes the IQN-based return quantile network, which maps a state-action pair and a quantile fraction to a return quantile. By randomizing the quantile level $\tau$, $H$ can generate samples from the conditional return distribution. The generator processes a 16-dimensional state-action input through a fully connected layer of 512 units. A quantile fraction $\tau \sim U(0,1)$ is encoded via a 64-dimensional cosine embedding and projected to 512 units. Both representations are normalized (LayerNorm) and activated (SiLU), then fused via a Hadamard product. The fused features are passed through a projection head with dimensions $512 \to 256 \to 32 \to 1$, followed by a negative exponential output $-\exp(ax+b)$ with learnable parameters. This constrains generated returns to be negative.

The model minimizes a quantile Huber loss ($\kappa = 1$) with a dispersion regularizer

$$-\lambda_e \, \mathbb{E}\big[|H(S,A,\tau_1) - H(S,A,\tau_2)|\big], \qquad \tau_1, \tau_2 \overset{\text{i.i.d.}}{\sim} U(0,1),$$

where $\lambda_e = \max(0.01, 0.4 \times 0.98^e)$ at epoch $e$. This discourages collapse of the generated return distribution. Rewards are multiplied by $0.1$ during IQN training and predictions are rescaled before calibration. The implementation also retains a positive-output hinge term, which is identically zero under the negative exponential parameterization.

We optimize the model using AdamW with weight decay $1 \times 10^{-4}$ and a cosine annealing learning rate schedule starting at $1 \times 10^{-4}$ over 150 epochs. To stabilize training, we apply dynamic percentile gradient clipping at the 90th percentile.

(3) **Regularization and stabilization.** To mitigate over-parameterization, we adopt a delayed pruning strategy: the first state-action linear layer is frozen. Unstructured $L_1$ pruning removes 10% of weights in the remaining linear layers after 40% of training. Target-network parameters are updated by exponential averaging with coefficient $0.002$. This network regularization is separate from the importance-weight clipping experiment in Appendix C.6.

(4) **Evaluation.** Calibration and evaluation follow the protocol for Example 1.

### C.4 Baseline comparisons

**Shared return estimation and baseline definitions.** Ours, Plug-in BH, C-BH, and SC-BH use the same fitted return distribution and test instances. Plug-in BH applies the Benjamini–Hochberg (BH) procedure directly to $\widehat{F}(c \mid S_{\text{test},j})$ at the nominal level $\alpha$. C-BH and SC-BH use weighted calibration, repeated subsampling, randomized p-value construction, and arithmetic averaging, followed by BH at the same nominal level $\alpha$.

**Calibration returns and residual scores.** For each truncation length $k$, the calibration buffer contains overlapping trajectory segments starting at $t = 0, \ldots, T-k-1$. Each segment is completed using one draw from the fitted conditional return distribution:

$$\widetilde{G}_{it}^{(k)} = \sum_{h=0}^{k-1} \gamma^h R_{i,t+h} + \gamma^k \widetilde{Z}_{it}, \qquad \widetilde{Z}_{it} \sim \widehat{F}(\cdot \mid S_{i,t+k}).$$

Let $\widehat{\mu}(s)$ and $\widehat{\sigma}(s)$ denote the estimated conditional mean and standard deviation, and define $\widehat{\sigma}_\epsilon(s) = \max\{\widehat{\sigma}(s), 10^{-6}\}$. The calibration and test scores for C-BH are

$$V_{it}^{\mathrm{C}} = \widetilde{G}_{it}^{(k)} - \widehat{\mu}(S_{it}), \qquad v_j^{\mathrm{C}}(c) = c - \widehat{\mu}(S_{\text{test},j}).$$

SC-BH uses the studentized scores

$$V_{it}^{\mathrm{SC}} = \frac{\widetilde{G}_{it}^{(k)} - \widehat{\mu}(S_{it})}{\widehat{\sigma}_\epsilon(S_{it})}, \qquad v_j^{\mathrm{SC}}(c) = \frac{c - \widehat{\mu}(S_{\text{test},j})}{\widehat{\sigma}_\epsilon(S_{\text{test},j})}.$$

The completed returns and residual scores are constructed once for each $k$ and held fixed across subsampling repetitions.

**Shared calibration weights and subsampling configuration.** C-BH and SC-BH use the same segment importance weights as ours. These weights account for the initial-state distribution shift and the action-probability ratios along each segment. The subsample size and number of repetitions are also controlled by the same configuration entries as ours. Subsets are sampled uniformly within each repetition, independently across repetitions. For each $k$, Ours, C-BH, and SC-BH share the exact same collection $\{\mathcal{I}_b\}_{b=1}^{B}$, which is also reused across thresholds $c$.

**Randomized p-values, averaging, and BH selection and comparability..** For either residual baseline, let $V_{it}$ and $v_j(c)$ denote its corresponding calibration and test scores. For repetition $b$, the implemented randomized p-value is

$$p_j^{(b)}(c) = \frac{\sum_{(i,t)\in\mathcal{I}_b} w_{it} I\{V_{it} \le v_j(c)\} + U_j^{(b)}}{1 + \sum_{(i,t)\in\mathcal{I}_b} w_{it}}, \qquad U_j^{(b)} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1). \tag{18}$$

For each fixed $k$, the uniform draws are independent across test instances and repetitions, and are shared between C-BH and SC-BH and across thresholds to reduce comparison variability.

The $B$ p-values for each test instance are combined by their arithmetic mean:

$$\overline{p}_j(c) = \frac{1}{B}\sum_{b=1}^{B} p_j^{(b)}(c).$$

BH is then applied once to the combined values at level $\alpha$. Specifically, let $\overline{p}_{(1)}(c) \le \cdots \le \overline{p}_{(m)}(c)$ be their ordered values and define

$$\widehat{r} = \max\left\{ r \in \{1,\dots,m\} : \overline{p}_{(r)}(c) \le \frac{\alpha r}{m} \right\},$$

with $\widehat{r} = 0$ when no rank satisfies the inequality. The procedure selects the test instances corresponding to the first $\widehat{r}$ ordered values.

The weighted calibration baselines account for trajectory segments and policy shift, providing empirical comparators for Ours. All methods use the same environment, target policy, training data, fitted return distribution, test instances, and nominal level $\alpha$. Ours, C-BH, and SC-BH additionally share the calibration buffer, segment importance weights, and sampled subsets, so the performance of these baselines is assessed empirically.

**Additional rollout horizons and on-policy comparisons.** The main text compares the methods at $k = 2$ in Examples 1 and 2 and $k = 4$ in Example 3. Figure 3 gives the corresponding on-policy comparison with the same horizons. Figures 4 and 5 provide off-policy comparisons at $k = 0$ in all three examples and at $k = 4, 4, 8$ in Examples 1–3, respectively. Together, these comparisons assess whether the observed power advantage persists across policy settings and choices of $k$. Plug-in BH uses only the initial-state failure probability, so its definition does not depend on the rollout horizon when the fitted return model and test sample are fixed.

**Remarks on policy setting and rollout horizon.** In Figure 3, our method retains higher power over most of the displayed threshold range. The separation is especially clear in Examples 2 and 3, so the observed advantage is not confined to off-policy comparisons. On-policy sampling removes behavior–target action-policy mismatch, but does not eliminate return-model or finite-sample calibration error.

Figure 4 shows FDR inflation at several thresholds in all three examples when $k = 0$. Relative to these curves, the moderate horizons used in the main text give a more favorable FDR–power trade-off. Increasing $k$ further has different effects across environments: Example 2 remains close to the nominal level at $k = 4$, whereas Example 3 has repeated exceedances at $k = 8$ in Figure 5. Its higher power at some thresholds should therefore not be interpreted as an improvement under the same FDR constraint. The figures compare fixed horizons and do not establish a data-driven rule for choosing $k$.

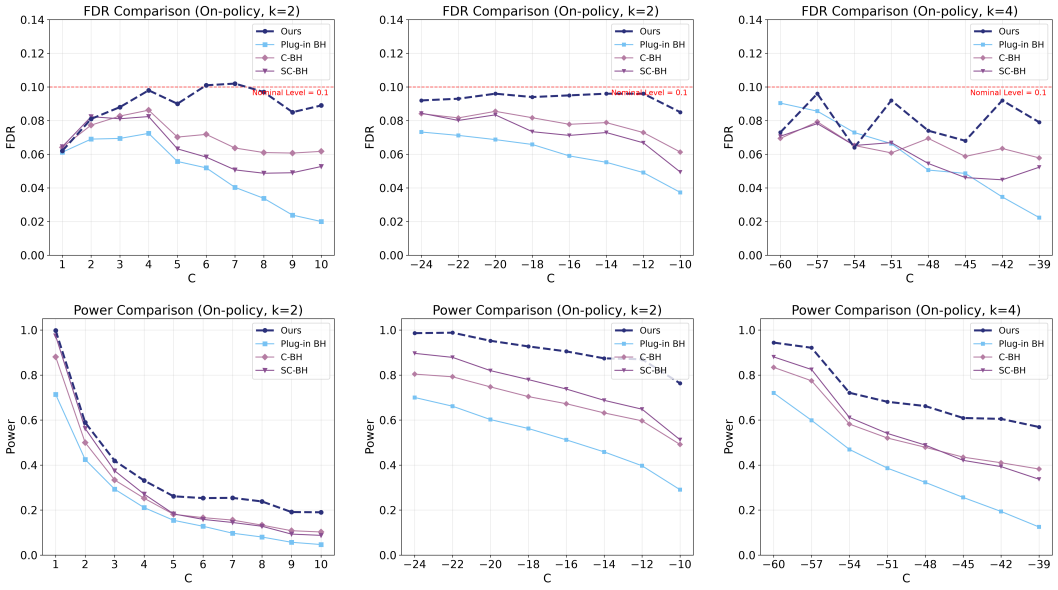


Figure 3: On-policy comparison of Ours, Plug-in BH, C-BH, and SC-BH. Columns correspond to Examples 1–3 at $k = 2, 2, 4$, respectively. The top row shows FDR and the bottom row shows power across thresholds $c$. The nominal FDR level is $\alpha = 0.1$.

**Power and variability in Example 1.** Tables 1 and 2 report mean and standard deviation summaries for the Shi off-policy setting. Our method achieves higher power in most reported configurations, whereas the BH-based baselines are generally more conservative. SC-BH is competitive at the lowest threshold, but studentization does not provide a uniform improvement over C-BH. Some empirical FDR values exceed the nominal level, especially at $k = 0$, so the power comparison should be interpreted together with achieved FDR. These finite-sample deviations do not contradict the asymptotic nature of our guarantee. For example, at $k = 2$ and $c = 5$, our method has FDR $0.090$ and power $0.293$, compared with powers $0.189$, $0.195$, and $0.174$ for Plug-in BH, C-BH, and SC-BH, respectively. Their FDR values are also lower, at $0.059$, $0.064$, and $0.063$. At $c = 7$, our empirical FDR is $0.103$, illustrating why the nominal level should not be described as satisfied in every configuration. The reported standard deviations quantify variation across runs, not standard errors or confidence intervals for the means.

The likelihood-ratio score uses the full conditional return distribution. This can be advantageous when states with similar predicted means have different failure probabilities, as can occur with heterogeneous return variability or tail behavior. Residual standardization accounts for scale heterogeneity but need not recover the ordering of conditional tail probabilities. This explains why no single residual score dominates uniformly.

### C.5 Oracle comparisons and rollout horizons

Figures 6 and 7 compare our method across rollout horizons with a Monte Carlo oracle under off-policy and on-policy sampling, respectively. The oracle replaces the fitted conditional return distribution with Monte Carlo estimates while retaining the same importance weights. It therefore serves as a diagnostic for return-distribution error, rather than an oracle for every component of the procedure.

The oracle power advantage is most pronounced at demanding thresholds in Examples 2 and 3, and is also visible under on-policy sampling. This indicates that return-distribution estimation remains an important limitation even without action-policy mismatch. In Example 1, the gap is smaller at many thresholds, although the estimated method does not uniformly match the oracle.

FDR is not monotone in the rollout horizon. Several estimated and oracle curves fluctuate around $0.1$, so the oracle should not be interpreted as a finite-sample guarantee of exact nominal control.

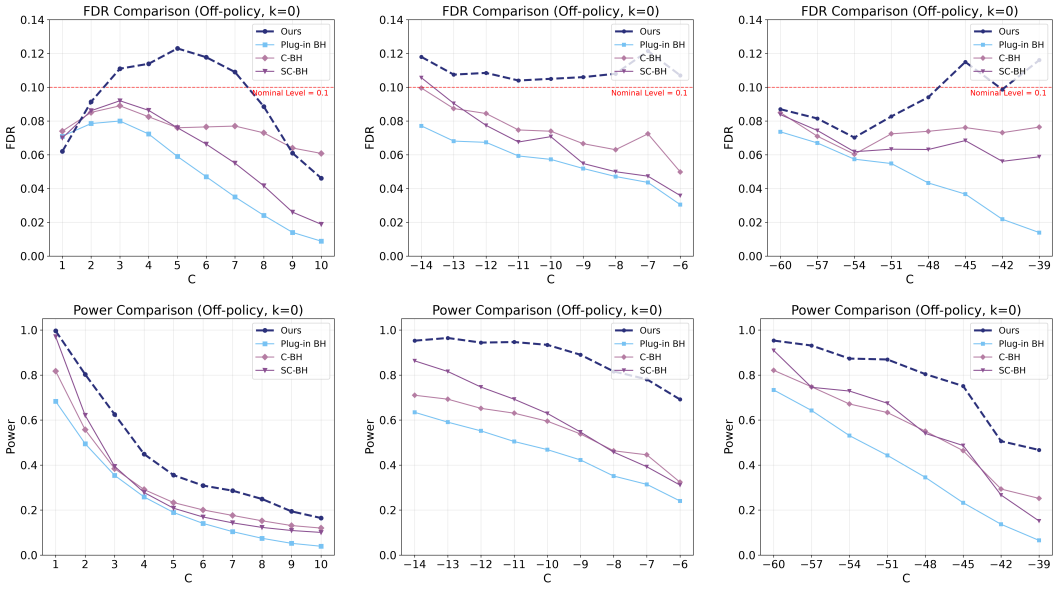


Figure 4: Additional off-policy baseline comparisons at $k = 0$. Columns correspond to Examples 1–3. The top row shows FDR and the bottom row shows power across thresholds $c$, for Ours, Plug-in BH, C-BH, and SC-BH. The nominal FDR level is $\alpha = 0.1$.

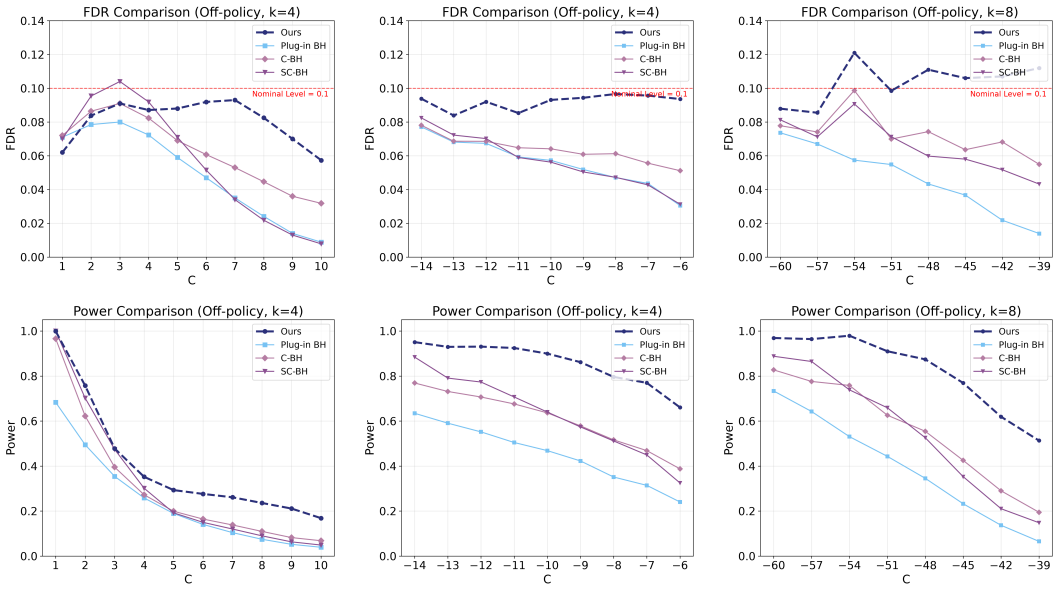


Figure 5: Additional off-policy baseline comparisons at longer rollout horizons. Columns correspond to Example 1 ($k = 4$), Example 2 ($k = 4$), and Example 3 ($k = 8$). The top row shows FDR and the bottom row shows power across thresholds $c$, for Ours, Plug-in BH, C-BH, and SC-BH. The nominal FDR level is $\alpha = 0.1$.

Table 1: FDR of Ours and the three baselines in the Shi off-policy setting. Entries are mean $\pm$ standard deviation. Both means and standard deviations are multiplied by 10. Selected thresholds and rollout horizons are shown.

| $k$ | $c$ | Ours | Plug-in BH | C-BH | SC-BH |
|---|---|---|---|---|---|
| 0 | 1 | $0.62 \pm 0.03$ | $0.71 \pm 0.06$ | $0.74 \pm 0.10$ | $0.70 \pm 0.07$ |
| 0 | 3 | $1.11 \pm 0.06$ | $0.80 \pm 0.22$ | $0.89 \pm 0.23$ | $0.92 \pm 0.13$ |
| 0 | 5 | $1.23 \pm 0.22$ | $0.59 \pm 0.16$ | $0.76 \pm 0.26$ | $0.76 \pm 0.20$ |
| 0 | 7 | $1.09 \pm 0.15$ | $0.35 \pm 0.10$ | $0.77 \pm 0.35$ | $0.55 \pm 0.17$ |
| 0 | 9 | $0.61 \pm 0.08$ | $0.14 \pm 0.09$ | $0.64 \pm 0.58$ | $0.26 \pm 0.22$ |
| 2 | 1 | $0.62 \pm 0.03$ | $0.71 \pm 0.06$ | $0.73 \pm 0.09$ | $0.70 \pm 0.07$ |
| 2 | 3 | $0.92 \pm 0.14$ | $0.80 \pm 0.22$ | $0.89 \pm 0.23$ | $0.95 \pm 0.07$ |
| 2 | 5 | $0.90 \pm 0.17$ | $0.59 \pm 0.16$ | $0.64 \pm 0.27$ | $0.63 \pm 0.21$ |
| 2 | 7 | $1.03 \pm 0.13$ | $0.35 \pm 0.10$ | $0.50 \pm 0.41$ | $0.34 \pm 0.08$ |
| 2 | 9 | $0.58 \pm 0.06$ | $0.14 \pm 0.09$ | $0.40 \pm 0.43$ | $0.32 \pm 0.41$ |
| 4 | 1 | $0.62 \pm 0.03$ | $0.71 \pm 0.06$ | $0.72 \pm 0.09$ | $0.70 \pm 0.07$ |
| 4 | 3 | $0.91 \pm 0.13$ | $0.80 \pm 0.22$ | $0.91 \pm 0.24$ | $1.04 \pm 0.09$ |
| 4 | 5 | $0.88 \pm 0.22$ | $0.59 \pm 0.16$ | $0.69 \pm 0.25$ | $0.71 \pm 0.15$ |
| 4 | 7 | $0.93 \pm 0.14$ | $0.35 \pm 0.10$ | $0.53 \pm 0.39$ | $0.34 \pm 0.08$ |
| 4 | 9 | $0.70 \pm 0.07$ | $0.14 \pm 0.09$ | $0.36 \pm 0.35$ | $0.13 \pm 0.26$ |

Table 2: Power of Ours and the three baselines in the Shi off-policy setting. Entries are mean $\pm$ standard deviation. Both means and standard deviations are multiplied by 10. Selected thresholds and rollout horizons are shown.

| $k$ | $c$ | Ours | Plug-in BH | C-BH | SC-BH |
|---|---|---|---|---|---|
| 0 | 1 | $\mathbf{9.96 \pm 0.16}$ | $6.83 \pm 0.54$ | $8.17 \pm 1.00$ | $9.70 \pm 0.34$ |
| 0 | 3 | $\mathbf{6.25 \pm 0.64}$ | $3.54 \pm 0.45$ | $3.84 \pm 0.57$ | $3.94 \pm 0.64$ |
| 0 | 5 | $\mathbf{3.55 \pm 0.40}$ | $1.89 \pm 0.44$ | $2.33 \pm 0.45$ | $2.08 \pm 0.26$ |
| 0 | 7 | $\mathbf{2.86 \pm 0.28}$ | $1.04 \pm 0.37$ | $1.76 \pm 0.44$ | $1.43 \pm 0.36$ |
| 0 | 9 | $\mathbf{1.94 \pm 0.11}$ | $0.52 \pm 0.23$ | $1.31 \pm 0.35$ | $1.09 \pm 0.31$ |
| 2 | 1 | $9.99 \pm 0.01$ | $6.83 \pm 0.54$ | $9.05 \pm 0.82$ | $\mathbf{10.00 \pm 0.00}$ |
| 2 | 3 | $\mathbf{4.51 \pm 0.47}$ | $3.54 \pm 0.45$ | $3.66 \pm 0.55$ | $4.14 \pm 0.53$ |
| 2 | 5 | $\mathbf{2.93 \pm 0.32}$ | $1.89 \pm 0.44$ | $1.95 \pm 0.46$ | $1.74 \pm 0.23$ |
| 2 | 7 | $\mathbf{2.73 \pm 0.24}$ | $1.04 \pm 0.37$ | $1.41 \pm 0.32$ | $1.21 \pm 0.19$ |
| 2 | 9 | $\mathbf{1.88 \pm 0.08}$ | $0.52 \pm 0.23$ | $0.90 \pm 0.24$ | $0.64 \pm 0.10$ |
| 4 | 1 | $9.99 \pm 0.01$ | $6.83 \pm 0.54$ | $9.66 \pm 0.42$ | $\mathbf{10.00 \pm 0.00}$ |
| 4 | 3 | $\mathbf{4.78 \pm 0.51}$ | $3.54 \pm 0.45$ | $3.95 \pm 0.53$ | $4.74 \pm 0.47$ |
| 4 | 5 | $\mathbf{2.93 \pm 0.40}$ | $1.89 \pm 0.44$ | $1.99 \pm 0.44$ | $1.92 \pm 0.19$ |
| 4 | 7 | $\mathbf{2.61 \pm 0.25}$ | $1.04 \pm 0.37$ | $1.38 \pm 0.32$ | $1.20 \pm 0.19$ |
| 4 | 9 | $\mathbf{2.11 \pm 0.09}$ | $0.52 \pm 0.23$ | $0.82 \pm 0.20$ | $0.63 \pm 0.14$ |

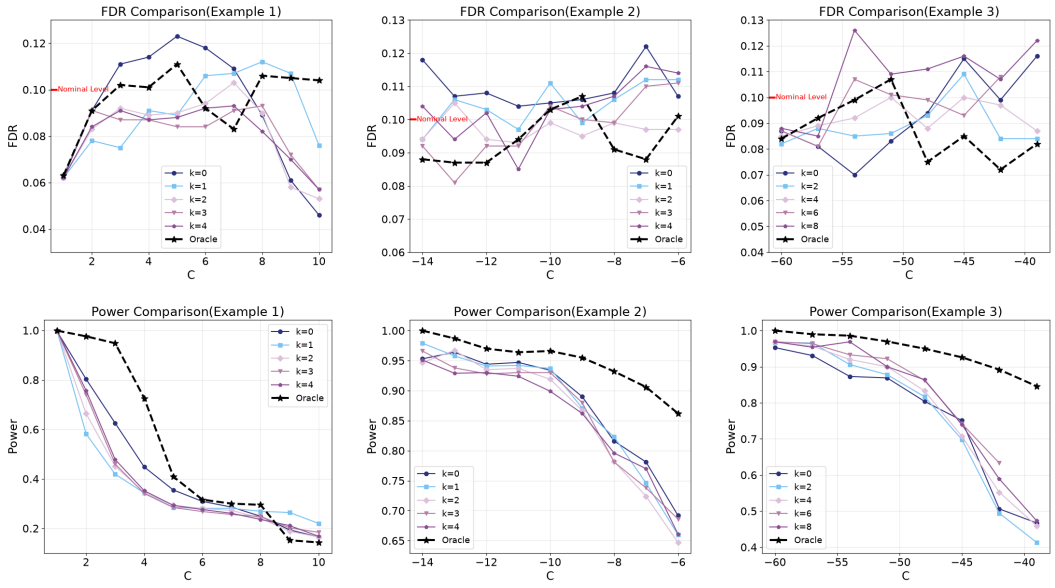


Figure 6: Off-policy comparison of our method at different rollout horizons $k$ with the Monte Carlo oracle. Columns correspond to Examples 1–3. The top row shows FDR and the bottom row shows power across thresholds $c$. The oracle retains the same importance weights and replaces only the return-distribution estimate.

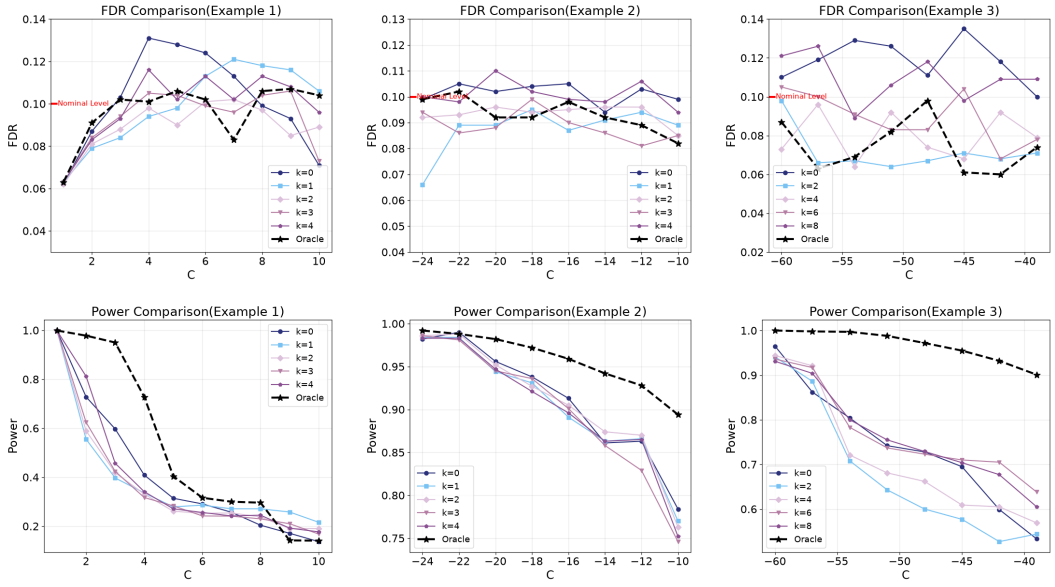


Figure 7: On-policy comparison of our method at different rollout horizons $k$ with the Monte Carlo oracle. Columns correspond to Examples 1–3. The top row shows FDR and the bottom row shows power across thresholds $c$. The oracle retains the same importance weights and replaces only the return-distribution estimate.

It approximates the return distribution by Monte Carlo while preserving the remaining estimation and calibration steps. The oracle gap is therefore a controlled diagnostic of replacing the return model, rather than a complete decomposition of all sources of error. Taken together with the baseline comparisons, these results support moderate horizons as a practical compromise in the displayed settings, not a universal optimum.

Table 3: Importance-weight clipping in the Shi off-policy setting (Example 1). Clip uses the empirical 5th and 95th percentiles of the weights as lower and upper cutoffs. FDR and power entries are mean $\pm$ standard deviation, with both multiplied by 10. Normalized ESS is unscaled.

| | $k=0$ | | $k=2$ | | $k=4$ | | $k=6$ | |
|---|---|---|---|---|---|---|---|---|
| $c$ | Clip | No clip | Clip | No clip | Clip | No clip | Clip | No clip |
| ESS/$n$ | 0.8904 | 0.8694 | 0.7869 | 0.7522 | 0.7016 | 0.6240 | 0.6381 | 0.4994 |
| | | | | **FDR** | | | | |
| 1 | $0.62 \pm 0.03$ | $0.62 \pm 0.03$ | $0.62 \pm 0.02$ | $0.63 \pm 0.04$ | $0.62 \pm 0.03$ | $0.67 \pm 0.08$ | $0.62 \pm 0.03$ | $0.65 \pm 0.03$ |
| 3 | $1.11 \pm 0.07$ | $1.09 \pm 0.09$ | $0.94 \pm 0.15$ | $0.95 \pm 0.12$ | $0.90 \pm 0.12$ | $1.20 \pm 0.30$ | $1.09 \pm 0.20$ | $1.79 \pm 0.42$ |
| 5 | $1.25 \pm 0.24$ | $1.34 \pm 0.25$ | $0.87 \pm 0.16$ | $0.91 \pm 0.21$ | $0.89 \pm 0.18$ | $0.86 \pm 0.22$ | $1.17 \pm 0.23$ | $2.08 \pm 0.54$ |
| 7 | $1.07 \pm 0.17$ | $1.18 \pm 0.14$ | $1.04 \pm 0.12$ | $1.05 \pm 0.11$ | $0.92 \pm 0.14$ | $0.90 \pm 0.12$ | $0.97 \pm 0.18$ | $2.15 \pm 0.54$ |
| 9 | $0.61 \pm 0.09$ | $0.60 \pm 0.09$ | $0.57 \pm 0.06$ | $0.61 \pm 0.08$ | $0.71 \pm 0.07$ | $0.89 \pm 0.16$ | $0.88 \pm 0.11$ | $1.12 \pm 0.18$ |
| | | | | **Power** | | | | |
| 1 | $9.95 \pm 0.18$ | $9.95 \pm 0.17$ | $9.98 \pm 0.01$ | $9.98 \pm 0.01$ | $9.99 \pm 0.01$ | $9.91 \pm 0.03$ | $9.99 \pm 0.01$ | $9.99 \pm 0.01$ |
| 3 | $6.31 \pm 0.69$ | $6.28 \pm 0.72$ | $4.46 \pm 0.51$ | $4.23 \pm 0.69$ | $4.82 \pm 0.49$ | $4.21 \pm 0.41$ | $4.54 \pm 0.54$ | $3.54 \pm 0.78$ |
| 5 | $3.51 \pm 0.42$ | $3.70 \pm 0.43$ | $2.96 \pm 0.31$ | $3.15 \pm 0.26$ | $2.89 \pm 0.43$ | $2.76 \pm 0.50$ | $3.06 \pm 0.48$ | $2.12 \pm 0.52$ |
| 7 | $2.90 \pm 0.27$ | $2.92 \pm 0.32$ | $2.71 \pm 0.27$ | $3.06 \pm 0.37$ | $2.64 \pm 0.23$ | $2.51 \pm 0.42$ | $2.40 \pm 0.30$ | $2.04 \pm 0.49$ |
| 9 | $1.93 \pm 0.12$ | $1.92 \pm 0.14$ | $1.90 \pm 0.07$ | $2.24 \pm 0.14$ | $2.10 \pm 0.10$ | $1.88 \pm 0.26$ | $1.23 \pm 0.14$ | $1.27 \pm 0.20$ |

### C.6 WEIGHT SENSITIVITY

We compare our method with and without importance-weight clipping in the Shi off-policy setting, keeping the return estimator and the remaining calibration procedure fixed. The completed sensitivity experiment uses 50 independent runs.

**Percentile-based weight clipping.** We use percentile-based clipping of the estimated importance weights. In Example 1, which corresponds to Table 3, the lower and upper cutoffs are the empirical 5th and 95th percentiles of the weights. Let $\widehat{q}_{0.05}$ and $\widehat{q}_{0.95}$ denote these quantiles of the unclipped weights $\{\widehat{w}_i\}_{i=1}^n$. The clipped weights are

$$\widehat{w}_i^{\text{clip}} = \min\{\widehat{q}_{0.95}, \max\{\widehat{w}_i, \widehat{q}_{0.05}\}\}.$$

Weights below the lower cutoff are raised to $\widehat{q}_{0.05}$, those above the upper cutoff are capped at $\widehat{q}_{0.95}$, and weights between the cutoffs remain unchanged. This operation retains all observations. For each horizon, these quantiles are computed from the complete segment weights in the calibration buffer before replay subsampling. Clipping is applied to the full state-density and action-ratio product, followed by mean normalization as described below.

**Weight concentration and selection performance.** Table 3 reports FDR, power, and normalized effective sample size,

$$\text{ESS}/n = \frac{(\sum_{i=1}^n v_i)^2}{n \sum_{i=1}^n v_i^2},$$

where $v_i = \widehat{w}_i^{\text{clip}}$ for the clipped procedure and $v_i = \widehat{w}_i$ for the unclipped procedure. The reported normalized ESS decreases with $k$ and is larger with clipping, indicating reduced weight concentration. At longer horizons, clipping reduces FDR inflation and variability in several configurations and can improve power. The effect is not uniform across thresholds, and clipping does not eliminate all finite-sample FDR deviations. ESS measures concentration rather than weight-estimation accuracy and does not verify the assumptions required by the theory. Clipping can also introduce bias. At $k = 6$ and $c = 7$, clipping reduces mean FDR from $0.215$ to $0.097$ and its standard deviation from $0.054$ to $0.018$, while power increases from $0.204$ to $0.240$. Normalized ESS increases from $0.4994$ to $0.6381$. In contrast, at $k = 2$ and $c = 9$, clipping lowers power from $0.224$ to $0.190$. Thus, its benefit depends on the threshold and horizon. Together with the oracle comparisons, this analysis distinguishes return-model sensitivity from sensitivity to extreme weights.

### C.7 CALIBRATION AND NUMERICAL DETAILS

**Sampling and weight stability.** Replay sampling is a computational subsampling strategy. The calibration proof requires convergence of the weighted empirical quantities under the stated sampling assumptions. Temporal dependence requires suitable mixing conditions, and the weight assumptions govern the correction for distribution shift. Independent test states represent separate deployment instances rather than consecutive states in a trajectory.

The state density ratio is estimated by a balanced classifier using training replay states and an independent reference sample from the deployment initial-state distribution. This does not require the evaluation test states or their realized returns. To fix the scale of Eq. (6) relative to the $1\vee$ denominators, let $v_{itk}$ be the estimated segment weight after any specified clipping and set

$$\hat{w}_{itk} = \frac{v_{itk}}{M_k^{-1}\sum_{(u,v)\in\widetilde{\mathcal{I}}_{\mathrm{cal}}} v_{uvk}}, \qquad M_k = |\widetilde{\mathcal{I}}_{\mathrm{cal}}|.$$

The buffer weights therefore have mean one before replay subsampling. Products of action ratios are computed in log space and exponentiated after subtracting a common maximum. Estimated class probabilities used for density ratios are floored at $10^{-10}$.

**Repeated calibration.** Taking the median over a fixed number $B$ of replay repetitions reduces sensitivity to subsampling. It does not create independent datasets or itself establish finite-sample FDR control. Its asymptotic interpretation requires the repeated threshold estimates to converge to the same population threshold. For numerical evaluation, $R = +\infty$ when $\widehat{F} = 1$, avoiding division by zero. Candidate thresholds include all tied scores, and $-\infty$ denotes an empty selection.

**Selection mass and the FDR limit.** The lower bound $u > 0$ keeps the population selection denominator away from zero. If a uniform conditional error bound $\Delta(S_0) \le \varepsilon_k$ holds, retaining the selection indicator gives

$$\frac{\mathbb{E}_{\mathrm{te}}[I\{S_0 \text{ selected}\}\Delta(S_0)]}{\Pr_{\mathrm{te}}(S_0 \text{ selected})} \le \varepsilon_k,$$

so this step does not require an additional $1/u$ factor. For a fixed limiting rule with selection probability $b$ and conditional failure probability $\rho$, independence gives $\mathrm{FDR} = \rho\{1-(1-b)^m\}$. The theorem's exact expression corresponds to an active calibration constraint, $\rho = \alpha$. The nominal-level adjustment should use the selection mass of the resulting rule. Under off-policy sampling, this mass is a target-state probability and requires the appropriate weighting.

**Action marginalization.** For discrete actions, we evaluate

$$\widehat{F}(g \mid s) = \sum_a \pi(a \mid s)\Pr\{H(\epsilon, s, a) \le g\}.$$

For continuous actions, the sum is replaced by an average over target-policy action samples. Increasing the number of action and generator samples reduces numerical error, which is distinct from return-model error.

**Rollout horizon and error diagnostics.** At $k = 0$, calibration relies entirely on the initial-state return model. Larger $k$ reduces the contribution of residual distribution error, but the transformed threshold $(c - \sum_{t<k}\gamma^t R_t)/\gamma^k$ can move into a difficult tail and the product of policy ratios can become more variable. The Monte Carlo oracle keeps the same estimated weights while replacing the return model, which helps assess the contribution of return-distribution error. Weight clipping and ESS provide complementary diagnostics for calibration stability. These diagnostics do not establish accuracy in poorly covered regions. The observed state is assumed to be Markov sufficient, so unobserved long-horizon returns should not be interpreted as a general latent-state POMDP setting.