Off-Policy Risk Assessment for Markov Decision Processes

Audrey Huang
UIUC

Liu Leqi
CMU

Zachary C. Lipton
CMU

Kamyar Azizzadenesheli
Purdue University

Abstract

Addressing such diverse ends as mitigating safety risks, aligning agent behavior with human preferences, and improving the efficiency of learning, an emerging line of reinforcement learning research addresses the entire distribution of returns and various risk functionals that depend upon it. In the contextual bandit setting, recently work on *off-policy risk assessment* estimates the target policy's CDF of returns, providing finite sample guarantees that extend to (and hold simultaneously over) plugin estimates of an arbitrarily large set of risk functionals. In this paper, we lift OPRA to Markov decision processes (MDPs), where importance sampling (IS) CDF estimators suffer high variance on longer trajectories due to vanishing (and exploding) importance weights. To mitigate these problems, we incorporate model-based estimation to develop the first doubly robust (DR) estimator for the CDF of returns in MDPs. The DR estimator enjoys significantly less variance and, when the model is well specified, achieves the Cramer-Rao variance lower bound. Moreover, for many risk functionals, the downstream estimates enjoy both lower bias and lower variance. Additionally, we derive the first minimax lower bounds for off-policy CDF and risk estimation, which match our error bounds up to a constant. Finally, we demonstrate the efficacy of our DR CDF estimates experimentally on several different environments.

1 Introduction

In off-policy evaluation (OPE), we aim to estimate the risk associated with a *target* policy given only offline datasets collected from a deployed *behavior* policy. While OPE research has historically focused on estimating the expected return, a wave of recent works, addressing such desiderata as risk aversion (and seeking) and alignment with human preferences, has expanded OPE research to broader families of risk functionals. Each risk functional maps from *the distribution of returns* to a real value. Common risk functionals include the expectation, variance, value at risk, conditional value at risk, and broader families such as distortion risk functionals, coherent risk functionals, and the Lipschitz risk functionals.

Recent works on general *off-policy risk assessment* have recognized that we can simultaneously estimate multiple risk functionals, providing guarantees that hold simultaneously over an entire collection of risks [Huang et al., 2021b, Chandak et al., 2021a]. These results are achieved through a two-step process: (i) estimate the cumulative distribution function (CDF) of returns under the target policy; and (ii) calculate the risk functionals of interest on the estimated CDF. This approach has two advantages: (a) guarantees on the CDF estimation error yield corresponding guarantees on the risk estimates; and (b) because the risk estimates all share the same underlying CDF, the corresponding guarantees on any number of downstream risk estimates hold simultaneously without loss in power.

To our knowledge, in MDPs, importance sampling (IS) is the only established method for estimating CDFs [Chandak et al., 2021a]. However, IS estimators are known to suffer from high variance exponential in the horizon [Liu et al., 2018] that can render them impractical over longer trajectories. Further, practitioners often evaluate a deterministic target policy given a dataset collected by a stochastic behavioral policy. In such cases, the IS weight is zero for all trajectories where the target policy places zero probability on an action that the behavioral policy takes. This renders many samples unusable and can significantly reduce the effective size of an offline dataset [Sussex et al.].

In off-policy evaluation addressing the standard *expected return* risk functional, doubly robust estimators and their variants have been shown to mitigate these problems [Dudik et al., 2011, Dudík et al., 2014, Jiang and Li, 2016, Thomas and Brunskill, 2016, Wang et al.,

2017, Su et al., 2020]. In the contextual bandit setting, the paper introducing OPRA also proposed and analyzed a doubly robust estimator for the CDF [Huang et al., 2021b], but model-based and doubly robust risk estimation in MDPs, where it may be most impactful, remains an open problem. One primary challenge is that, in the expected return setting, extending the model-based and doubly robust estimators from contextual bandits to MDPs [Jiang and Li, 2016] relied on stepwise or recursive formulations of the importance sampling estimator, which lack clear analogs in CDF estimation.

In this paper, we define and analyze the first doubly robust CDF estimator for off-policy risk assessment in MDPs. Specifically, we contribute the following:

1. A recursive formulation of the CDF of returns in MDPs (§ 3.2), which enables us to derive importance sampling (§ 4) and model-based CDF estimators (§ 5).

2. Rate-matching upper and minimax lower bounds (§ 4.2) for the IS CDF estimator.

3. The first doubly robust off-policy estimator for the CDF of returns that subsumes the previously introduced doubly robust estimator for contextual bandits (§ 6). Analysis demonstrating its potential for significant variance reduction.

4. The first Cramer-Rao lower bound for off-policy CDF estimation and analysis demonstrating that the doubly robust estimator achieves this lower bound in certain cases (§ 6.3).

5. A demonstration that the risk estimates inherit the reduced variance in CDF estimation (§ 7).

6. Experimental evidence that the doubly robust CDF and risk estimates exhibit significantly lower error and variance than importance sampling estimators on benchmark problems, with applications to safe policy evaluation and improvement (§ 8).

2 Related Work

Several estimators have been studied for off-policy estimation of the expected return. A seminal work by Dudik et al. [2011] proposed the first doubly robust estimator for the expected return in contextual bandits, which was extended to the MDP setting by Jiang and Li [2016]. The doubly robust estimator leverages both importance sampling weights and models to reduce variance while retaining unbiasedness. Researchers have proposed several variants of the doubly robust estimator, aiming to improve its performance, with Thomas and Brunskill [2016], Wang et al. [2017] optimizing the balance between importance sampling and model-based estimators and Su et al. [2020] optimizing

the weights used in the estimator. A Cramer-Rao variance lower bound [Jiang and Li, 2016] and minimax lower bounds [Li et al., 2015, Wang et al., 2017, Ma et al., 2021] for off-policy mean estimation have been derived.

While the expected return has been the focus of most previous off-policy evaluation research, other risk functionals are increasingly of interest. In the reinforcement learning (RL) literature, the variance [Sani et al., 2013, Tamar et al., 2016], variance-constrained mean Mannor and Tsitsiklis [2011], and conditional value-at-risk [Rockafellar et al., 2000, Keramati et al., 2020, Huang et al., 2021a] risk functionals are often employed. Other risk functionals, including exponential utility [Denardo et al., 2007], distortion risk functionals [Dabney et al., 2018], and cumulative prospect weighting [Gopalan et al., 2017, Prashanth et al., 2016, 2020] have also been investigated.

As a result, a number of recent works have aimed to estimate risk functionals from off-policy datasets. Of these, [Chandak et al., 2021b] estimates the variance, while more recent works [Huang et al., 2021b, Chandak et al., 2021a] tackle the estimation of more general risks and are the closest works of comparison. Both [Huang et al., 2021b, Chandak et al., 2021a] take a two-step approach of first estimating the off-policy CDF of returns, then estimating their risks via a plug-in approach. Chandak et al. [2021a] proposed an IS CDF estimator in both stationary and nonstationary MDPs and derives confidence intervals for a number of risk estimates on a per-risk basis. Huang et al. [2021b] proposed both IS and DR estimators for the CDF of contextual bandits, establishing the first finite sample guarantees for off-policy CDF estimation, showing that the estimated CDFs converge uniformly achieving $O(1/\sqrt{n})$ rates. Additionally, they introduced the broad class of Lipschitz risk functionals, whose Lipschitz constants bound the errors in risk estimation by the errors in CDF estimation.

While Chandak et al. [2021a] proposes *only* IS CDF estimates for MDPs, Huang et al. [2021b] proposes variance-reduced DR CDF estimates in *only* contextual bandits. Thus *variance-reduced risk estimation in MDPs*, where high variance from IS is most problematic [Liu et al., 2018], is an open problem. Towards this, we (1) develop provably variance-reduced CDF/risk estimators in MDPs, and, demonstrating their optimality, (2) derive the first matching upper and lower bounds.

3 Preliminaries

3.1 Problem Setting

An MDP is defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \mu)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition function, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in [0, 1]$ is the discount factor, and μ is the fixed starting state distribution. We assume that rewards are bounded WLOG on the support $[0, D]$. A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps a state to a probability distribution over actions.

In the off-policy evaluation problem, we are concerned with estimating risk functionals of a policy π given a dataset of trajectories $\mathcal{D} = \{\tau^i\}_{i=1}^n$ collected by the behavioral policy β interacting with $\mathcal{M}$, where τ is an H-step trajectory $\tau = (s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1})$. Let $w(a, s) := \frac{\pi(a|s)}{\beta(a|s)}$ denote the importance weight, and let $w_{max} = \max_h \max_{a_h, s_h} w(a_h, s_h)$ be the maximum importance weight over all horizons, states, and actions. For convenience, by τ_j we denote a sub-trajectory of τ from horizon 1 to j, and by $w_j = \prod_{h=1}^{j} w(a_h, s_h)$ the importance weight of τ_j.

We denote by $\mathbb{P}$ the distribution of trajectories induced by π on $\mathcal{M}$, and by $\mathbb{P}_\beta$ the distribution induced by β. The random variable of returns of an H-step trajectory induced by π in $\mathcal{M}$ is given by $Z^{H,\pi} = \left(\sum_{h=1}^{H} \gamma^h R_h \Big| \pi \right)$ and its cumulative distribution function (CDF) is $F^{H,\pi}(t) = \mathbb{E}_\mathbb{P}\left[\mathbb{1}\left\{ Z^{H,\pi} \le t \right\} \right]$, where $\mathbb{1}\{\cdot\}$ is the indicator function. When clear from the context we drop the dependency on π. The CDF of returns conditioned on a state s_h at horizon h is defined as $F_{s_h}(t) = \mathbb{E}_\mathbb{P}\left[\mathbb{1}\left\{ Z^{H-h+1} \le t \right\} | s_h \right]$, and the state-action dependent distribution $F_{s_h, a_h}(t)$ is defined similarly. We also define $S^{H,\pi}(t) = 1 - F^{H,\pi}(t) = \mathbb{E}_\mathbb{P}\left[\mathbb{1}\left\{ Z^{H,\pi} > t \right\} \right]$, and S analogues of all previous definitions. For short, we use $\mathbb{E}_h^\beta[\cdot] = \mathbb{E}_{\mathbb{P}_\beta}[\cdot|\tau_h]$, and $\mathbb{E}_h[\cdot]$ is defined similarly for $\mathbb{P}$. We define conditional variances $\mathbb{V}_h^\beta, \mathbb{V}_h$ similarly.

3.2 CDF Bellman Equation

Our off-policy CDF estimators rely on a novel *CDF Bellman equation* that provides a recursive formulation of the CDF of returns F. In the off-policy setting, for state s_h at horizon h, we have:

$$F_{s_h}^{H+1-h}(t) := \mathbb{E}_{\mathbb{P}_\beta}\left[w(A_h, s_h) F_{S_{h+1}}^{H-h}\left(\frac{t - R_h}{\gamma} \right) \right] \quad (1)$$

where $F_{s_{H+1}}^0 = \mathbb{1}\{0 \le t\}$ for all s_{H+1} at the end of the horizon. This recursion was first introduced in Sobel [1982] for deterministic rewards in the value iteration setting, which we have extended to the more general stochastic setting. Using the definition of equivalence in distribution ($\overset{D}{=}$), it can be seen that the CDF Bellman operator is equivalent to the distributional Bellman operator [Bellemare et al., 2017, Dabney et al., 2018], $Z_{s_h}^{H-h+1} \overset{D}{=} R(s_h, A_h) + \gamma Z_{S_{h+1}}^{H-h}$.

3.3 Risk Functionals

Let $Z \in \mathcal{L}_\infty(\Omega, \mathcal{F}_Z, \mathbb{P}_Z)$ denote a real-valued random variable that admits a CDF $F_Z \in \mathcal{L}_\infty(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A *risk functional* $\rho : \mathcal{L}_\infty(\Omega, \mathcal{F}_Z, \mathbb{P}_Z) \to \mathbb{R}$ is said to be law-invariant if $\rho(Z)$ depends only on the distribution of Z [Kusuoka, 2001], i.e., if for any pair of random variables Z and Z', $F_Z = F_{Z'} \implies \rho(Z) = \rho(Z')$. When clear from the context, we write $\rho(F_Z)$ in place of $\rho(Z)$. In this paper, we restrict our focus to law invariant risk functionals, as it may not be practical to estimate risk functionals that are not law invariant from data [Balbás et al., 2009].

Examples of popularly studied law-invariant risk functionals include the expected value, variance, mean-variance, coherent risks (including conditional value-at-risk (CVaR) and proportional hazard), distortion risks, and cumulative prospect theory risks, which we discuss more extensively in Appendix A due to space constraints. We give particular attention to distortion risk functionals due to their widespread usage and general form. Distortion risks apply a distortion function g to the CDF F and have the formula

$$\rho(Z) = \int_0^\infty g(1 - F_Z(t))dt = \int_0^\infty g(S(t))dt.$$

Notably, expected value is a special case of distortion risk with $g(x) = x$, while CVaR_α has distortion $g(x) = \min\{\frac{x}{1-\alpha}, 1\}$. A large number of the aforementioned risk functionals satisfy a notion of smoothness in the CDF, which Huang et al. [2021b] define as the class of *Lipschitz risk functionals*:

Definition 3.1 (Lipschitz Risk Functionals). *A law invariant risk functional ρ is L-Lipschitz if for any pair of CDFs F_Z and $F_{Z'}$ and some $L \in (0, \infty)$, it satisfies*

$$|\rho(F_Z) - \rho(F_{Z'})| \le L\|F_Z - F_{Z'}\|_\infty. \quad (2)$$

Lipschitz risk functionals include distortion risk functionals with Lipschitz distortion functions g. As special cases, expected value is D-Lipschitz and CVaR at level α is $D/(1-\alpha)$-Lipschitz. Additionally, variance is $3D^2$-Lipschitz and so is mean-variance [Huang et al., 2021b].

As we analyze the theoretical properties of CDF and risk estimators later in our paper, we show that estimates of Lipschitz risk functionals retain many desirable properties, including characterizable finite sample error and consistency.

3.4 Off-Policy Risk Assessment

In the remainder of this paper, we will take the general approach of Algorithm 1 (OPRA) for off-policy risk assessment. OPRA takes as input a set of risk functionals of interest, and outputs estimates of their values under the target policy π. It first estimates the CDF of returns under π from off-policy data, then generates plug-in risk estimates on the CDF. In Sections 4, 5 and 6 we define and analyze CDF estimators. In Section 7, we turn to plug-in risk estimation.

Algorithm 1: Off-Policy Risk Assessment (OPRA)

Input: Dataset $\mathcal{D}$, policy π, probability δ, risk functionals $\{\rho_p\}_{p=1}^{P}$.

1 Estimate the CDF $\widehat{F}$ with error ϵ;
2 For $p = 1...P$, estimate risk $\widehat{\rho}_p = \rho_p(\widehat{F})$;

Output: Estimates with errors $\{\widehat{\rho}_p \pm f_p(\epsilon)\}_{p=1}^{P}$.

---

4 Importance Sampling CDF Estimation

Off-policy evaluation faces the unique challenge that only rewards for actions taken by the behavioral policy are observed. Importance sampling (IS), which reweights samples according to the ratio between target and behavioral policy probabilities, can be used to overcome this distribution shift. Though the IS estimator for off-policy estimation has previously been introduced in Huang et al. [2021b], Chandak et al. [2021a], theoretical understanding of finite sample convergence remains an open question. In this section, we derive the first minimax lower bound for off-policy CDF estimation, which we demonstrate the convergence rate of the IS estimator matches. Further, we leverage unique properties of the CDF estimate to develop a novel variance-reduced IS estimator.

4.1 Importance Sampling Estimator

Using the recursion in (1), the importance sampling CDF estimator is:

$$\widehat{F}_{\text{IS}}(t) = \frac{1}{n} \sum_{i=1}^{n} \widehat{F}_{s_1^i}^{H}(t), \tag{3}$$

$$\text{where} \quad \widehat{F}_{s_h^i}^{H+1-h}(t) = w(a_h^i, s_h^i)\widehat{F}_{s_{h+1}^i}^{H-h}\left(\frac{t - r_h^i}{\gamma}\right)$$

Unrolling the recursion recovers the more intuitive form of the IS estimator previously introduced in Chandak et al. [2021a], that reweights each sample by the importance weight of the entire trajectory:

$$\widehat{F}_{\text{IS}}(t) = \frac{1}{n} \sum_{i=1}^{n} w_H^i \mathbb{1}\left\{\sum_{h=1}^{H} \gamma^h r_h^i \le t\right\}. \tag{4}$$

Lemma 4.1. *The* IS CDF *estimator* (3) *is unbiased and its variance is recursively given by*

$$\mathbb{V}_h^{\beta}\left[\widehat{F}_{S_h}(t)\right] = \mathbb{E}_h^{\beta}\left[\mathbb{V}_h^{\beta}\left[w(A_h, S_h)F_{S_h, A_h}(t)\Big|S_h\right]\right]$$
$$+ \mathbb{E}_h^{\beta}\left[w(A_h, S_h)^2\mathbb{V}_{h+1}\left[\widehat{F}_{S_{h+1}}\left(\frac{t - R_h}{\gamma}\right)\Big|S_h, A_h\right]\right]$$
$$+ \mathbb{V}_h\left[F_{S_h}(t)\right] \tag{5}$$

The first term expresses variance from importance-weighted actions, while the third term expresses variance from transitions. The second term encompasses variance from random rewards and includes the recursion. At horizon h, each term in the variance is multiplied by the product of importance weights from the trajectory thus far, w_{h-1}.

4.2 Finite Sample Error Bound

While the IS estimator is known to be consistent, obtaining finite sample error bounds has remained an important open problem. Such bounds characterize the convergence rate of the estimate towards the true F, and provide confidence intervals around estimates in the finite sample regimes of practical scenarios.

First, however, it is important to note that, given finite samples, the IS estimator may not be a valid CDF due to the use of importance weighting, e.g., it can be greater than 1. To address this issue, we can use a weighted importance sampling (WIS) estimator instead: $\widehat{F}_{\text{WIS}}(t) = \frac{n}{\sum_{j=1}^{n} w_H^j}\widehat{F}_{\text{IS}}(t)$, which normalizes the estimate using the sum of importance weights instead of the sample size, guaranteeing that $\widehat{F}_{\text{WIS}} \in [0, 1]$. As Chandak et al. [2021a] demonstrates, the WIS estimator is biased but uniformly consistent.

Alternatively, the IS estimator can simply be clipped to the unit range, which was first introduced by Huang et al. [2021b] for off-policy CDF estimation in contextual bandits.

$$\widehat{F}_{\text{IS-CLIP}}(t) = \min\left\{\widehat{F}_{\text{IS}}(t), 1\right\},$$

The IS-CLIP estimator, too, is biased but uniformly consistent, and is uniformly bounded as:

Lemma 4.2. *With probability at least* $1 - \delta$,

$$\|\widehat{F}_{\text{IS-CLIP}} - F\|_{\infty} \le c_1\sqrt{\frac{\sum_{\tau}\mathbb{P}_{\beta}(\tau)w^2(\tau)}{n}} + c_2\frac{w_{max}}{n}. \tag{6}$$

The IS estimator converges at a rate of $O(\mathbb{E}_{\mathbb{P}_{\beta}}[w(\tau)^2]/\sqrt{n})$, subsuming previous results on convergence in contextual bandits [Huang et al., 2021b]. For readability throughout the paper, we leave constants implicit and give their exact form in the proof (Appendix B).

How close to optimal is this convergence rate? For large enough sample size n, the IS estimator is in fact minimax rate-optimal, as the following lower bound demonstrates:

Theorem 4.1. *For a universal constant c, when $n \geq c \frac{w_{max}^2}{\sum_\tau \mathbb{P}_\beta(\tau) w^2(\tau)}$, we have*

$$\inf_{\widehat{F}} \sup_{F \in \mathcal{F}} \mathbb{E}\left[\|\widehat{F} - F^\pi\|_\infty\right] \geq c\sqrt{\frac{\sum_\tau \mathbb{P}_\beta(\tau) w^2(\tau)}{n}} \quad (7)$$

4.3 Variance-Reduced IS Estimation

Unlike the IS estimator of the expected return [Dudik et al., 2011, Jiang and Li, 2016], the IS estimator of the CDF has the unique property that its variance differs across its support as a function of t. Examining the variance for $\widehat{F}_{\mathrm{IS}}$ (5) at a fixed t, the first term is small if $F_{S_h, A_h}(t)$ is close to 0, which can occur at small t. However if $F_{S_h, A_h}(t)$ is close to its maximal value of 1, which occurs when t close to the maximum return D, the first term is close to its largest possible value.

4.3.1 Estimating $S = 1 - F$

Rather than estimating the CDF F, another possibility is to estimate $S = 1 - F$, which is directly used in the expression for many risk expressions (e.g. distortion risk functionals, see § 3). Importance sampling and direct method estimators for S take a similar form as those for F, but with the indicator function inequality swapped to $\mathbb{1}\{ \cdot > t\}$. For example,

$$\widehat{S}_{\mathrm{IS}}(t) = \frac{1}{n}\sum_{i=1}^n w_H^i \mathbb{1}\left\{\sum_{h=1}^H \gamma^h r_h^i > t\right\}, \quad (8)$$

Perhaps non-intuitively, $\widehat{S}_{\mathrm{IS}}$ and $\widehat{F}_{\mathrm{IS}}$ have different theoretical properties, such as variance:

$$\mathbb{V}_h^\beta\left[\widehat{S}_{S_h}(t)\right] = \mathbb{E}_h^\beta\left[\mathbb{V}_h^\beta\left[w(A_h, S_h)(1 - F_{S_h, A_h}(t))\big|S_h\right]\right]$$
$$+ \mathbb{E}_h^\beta\left[w(A_h, S_h)^2 \mathbb{V}_{h+1}\left[\widehat{S}_{S_{h+1}}\left(\frac{t - R_h}{\gamma}\right)\big|S_h, A_h\right]\right]$$
$$+ \mathbb{V}_h\left[S_{S_h}(t)\right]$$

The first term of the variance expression for $\widehat{S}_{\mathrm{IS}}$ takes the exact opposite trend as that of $\widehat{F}_{\mathrm{IS}}$: it is low where $F_{S_h, A_h}(t)$ is high, and vice versa. This implies that while $\widehat{F}_{\mathrm{IS}}$ has low variance in the lower tails but high variance in estimating the upper tails, $\widehat{S}_{\mathrm{IS}}$ has low variance in the upper tails but high variance in the lower tails. Lower variance leads to lower error of the CDF estimate. To formalize this argument, we present the following bounds on the error of the CDF estimate in terms of its variance at each t. Going

forward, we refer to the estimator in (4) as $\widehat{F}_{\mathrm{F\text{-}IS}}$, and $\widehat{F}_{\mathrm{S\text{-}IS}} = 1 - \widehat{S}_{\mathrm{IS}}$ from (8) to avoid confusion.

Lemma 4.3. *If $\mathbb{V}[\widehat{F}]$, $\mathbb{V}[\widehat{S}]$ have bounded variation (see Assumption B.1), there exists universal constants c_1, c_2 such that $w.p \geq 1 - \delta$, $\forall t$,*

$$|F(t) - \widehat{F}_{\mathrm{F\text{-}IS}}(t)| \leq \frac{c_1 \log(\sqrt{n}/\delta)}{n}$$
$$+ \sqrt{\frac{c_2 \mathbb{V}[\widehat{F}_{\mathrm{IS}}(t)] \log(\sqrt{n}/\delta)}{n}},$$

$$|F(t) - \widehat{F}_{\mathrm{S\text{-}IS}}(t)| \leq \frac{c_1 \log(\sqrt{n}/\delta)}{n}$$
$$+ \sqrt{\frac{c_2 \mathbb{V}[\widehat{S}_{\mathrm{IS}}(t)] \log(\sqrt{n}/\delta)}{n}}$$

This difference has important implications for downstream risk estimation. The choice of whether to estimate S or F can reduce the variance of the plug-in risk estimate, especially for risk functionals such as CVaR that heavily up-weight one tail of the distribution. In general, under reward distributions, $\widehat{F}_{\mathrm{S\text{-}IS}}$ will lead to lower variance estimates for risk-averse risk functionals that upweight the lower tails of the distribution. $\widehat{S}_{\mathrm{F\text{-}IS}}$ will lead to lower-variance risk estimates for risk-seeking risk functionals that upweight the higher tails.

The RHS of the error bound in Lemma 4.3 gives a uniform confidence band on the estimated CDF that contains F with high probability. However, the variance of $\widehat{F}_{\mathrm{F\text{-}IS}}$ or $\widehat{F}_{\mathrm{S\text{-}IS}}$ cannot be computed without knowledge of the underlying MDP. Instead, we can use the empirical variance of the estimate $\widehat{F}$ or $\widehat{S}$:

Lemma 4.4 (Empirical Bernstein Bound). *Let $\mathbb{V}_n$ denote the sample variance. For any choice of M points $\{t_j\}_{j=1}^M \in [0, D]$, $w.p. \geq 1 - \delta$ we have $\forall j$ that*

$$|F(t) - \widehat{F}_{\mathrm{F\text{-}IS}}(t)| \leq \frac{c_1' \log(\sqrt{n}/\delta)}{n}$$
$$+ \sqrt{\frac{c_2' \mathbb{V}_n[\widehat{F}_{\mathrm{IS}}(t)] \log(\sqrt{n}/\delta)}{n}},$$

$$|F(t) - \widehat{F}_{\mathrm{S\text{-}IS}}(t)| \leq \frac{c_1' \log(\sqrt{n}/\delta)}{n}$$
$$+ \sqrt{\frac{c_2' \mathbb{V}_n[\widehat{S}_{\mathrm{IS}}(t)] \log(\sqrt{n}/\delta)}{n}}$$

Note that Lemma 4.4 is a bound for a finite number of points, while Lemma 4.3 is a uniform bound over all t. In practice, the sample variance of only a finite number of points can be calculated.

4.4 Combining S and F

Lemma 4.3 also indicates that the variance of the resulting estimate can be further reduced by pointwise com-

bining the F and S estimators, such that the estimator with the lower variance is used at each t. Such a combined estimator may be especially effective for reducing the variance of plug-in estimates for risk functionals that apply a distortion to both upper and lower tails of the distribution, such as the expected return.

How should the choice between $\widehat{F}_{\text{F-IS}}(t)$ or $\widehat{F}_{\text{S-IS}}(t)$ be made for each t? Lemma 4.4 suggests that one effective method is to estimate the empirical variance of $\widehat{F}_{\text{IS}}(t)$ and $\widehat{S}_{\text{IS}}(t)$ at each t, and to simply use the estimate with lower empirical variance:

$$\widehat{F}_{\text{C-IS}}(t) = \begin{cases} \widehat{F}_{\text{F-IS}}(t), & \text{if } \mathbb{V}_n[\widehat{F}_{\text{IS}}(t)] < \mathbb{V}_n[\widehat{S}_{\text{IS}}(t)], \\ \widehat{F}_{\text{S-IS}}(t), & \text{otherwise.} \end{cases}$$
(9)

The error bound of such a combined estimator is

Proposition 4.1. *Let* $\mathbb{V}_n^{min}(t) = \min\{\mathbb{V}_n[\widehat{F}(t)], \mathbb{V}_n[\widehat{S}(t)]\}$, *and* $\Delta(\mathbb{V}_n) = \sqrt{\frac{2\log(2/\delta)}{n-1}} + \max_j |\mathbb{V}_n[\widehat{F}](t_j) - \mathbb{V}_n[\widehat{S}](t_j)|$. *Then for* $n \geq \frac{8\log(2n/\delta)}{\Delta(\mathbb{V}_n)^2}$ *and discrete points* $\{t_j\}_{j=1}^M$, *we have w.p.* $\geq 1 - \delta$ *we have* $\forall j$ *that*

$$|F(t_j) - \widehat{F}_{\text{C-IS}}(t_j)| \leq \frac{c_1'' \log(\sqrt{n}/\delta)}{n} + \sqrt{\frac{c_2'' \mathbb{V}_n^{min}(t_j) \log(\sqrt{n}/\delta)}{n}}.$$

5 Model-Based CDF Estimation

In the expected value OPE literature, model-based estimation can avoid the high variance associated with importance sampling by directly learning the expected return under π using a model of the underlying MDP learned from data [Jiang and Li, 2016]. Taking inspiration from such approaches, we now develop the first model-based off-policy CDF estimator. First, a model of the MDP $\overline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \overline{P}, \overline{R}, \gamma, \overline{\mu})$ is formed from data, upon which a model $\overline{F}_{s_h, a_h}^{H-h+1}$ of the return distribution under π can be directly computed using the recursion in (1) for all horizons h, states s_h, and a_h. Formally, for horizons $h = H, \ldots, 1$, $\overline{F}$ can be computed recursively:

$$\overline{F}_{s_h, a_h}^{H+1-h}(t) = \mathbb{E}_{\overline{P}, \overline{R}} \left[\overline{F}_{S_{h+1}}^{H-h} \left(\frac{t - R_h}{\gamma} \right) \right]$$

$$\overline{F}_{s_h}^{H+1-h}(t) = \mathbb{E}_\pi \left[\overline{F}_{s_h, A_h}^{H+1-h}(t) \right]$$

with $\overline{F}_{s_{H+1}}^0(t) = \mathbb{1}\{0 \leq t\}$ for all s_{H+1} at the end of the horizon. The models $\overline{F}$ can then be used for the *direct method* (DM) of estimation, which simply averages $\overline{F}$ for the starting states observed in the data:

$$\widehat{F}_{\text{DM}}(t) = \frac{1}{n} \sum_{i=1}^n \overline{F}_{s_1^i}^H(t)$$
(10)

Because this direct method of estimation directly deploys the target policy π in the estimated model, the error of the estimates $\widehat{F}$ and $\overline{F}$ are directly related to the misspecification of the MDP model $\overline{M}$. The model $\overline{M}$ and thus the direct estimator is frequently prone to uncharacterizable and possibly high bias, e.g. when dataset coverage is inadequate or the state space is high in dimension.

6 Doubly Robust CDF Estimation

We now define a new *doubly robust* (DR) estimator for the distribution of returns in MDPs (Section 6.1), which takes advantage of both importance sampling (§ 4) and model-based (§ 5) estimators to retain the unbiasedness of the IS CDF estimator, with potentially significant reduction in variance. In § 6.3, we derive the first Cramer-Rao lower bound on off-policy CDF estimation, and show that the DR estimator achieves this lower bound when the regression model perfectly recovers the true distribution.

6.1 Doubly Robust (DR) Estimator

The recursive formulation in (3) is key for defining the doubly robust CDF estimator. At each horizon h in the recursive formulation, we use the model $\overline{F}$ as a baseline and apply an IS-weighted data-dependent correction. Specifically, this has the recursive form

$$\widehat{F}_{\text{DR}}(t) = \frac{1}{n} \sum_{i=1}^n \widehat{F}_{s_1^i}^H(t),$$

$$\text{where } \widehat{F}_{s_h}^{H+1-h}(t) = \overline{F}_{s_h}^{H+1-h}(t)$$
$$+ w(a_h, s_h) \left(\widehat{F}_{s_{h+1}}^{H-h} \left(\frac{t - r_h}{\gamma} \right) - \overline{F}_{s_h, a_h}^{H+1-h}(t) \right)$$
(11)

The model $\overline{F}$ can be obtained using the methods of § 5, and as before, $\widehat{F}_{s_{H+1}}^0(t) = \mathbb{1}\{0 \leq t\}$ for all s_{H+1}. Letting $z_h^i = \sum_{k=1}^h \gamma^k r_k^i$ be the return of trajectory i up until step h, the recursion in (11) can be unrolled over the H timesteps of the trajectory to obtain

$$\widehat{F}_{\text{DR}}(t) = \frac{1}{n} \sum_{i=1}^n \left(w_H^i \mathbb{1}\{z_H^i \leq t\} \right.$$
$$+ \sum_{h=1}^H w_{h-1}^i \overline{F}_{s_h^i} \left(\frac{t - z_h^i}{\gamma^h} \right)$$
$$\left. - \sum_{h=1}^H w_{h-1}^i w(a_h^i, s_h^i) \overline{F}_{s_h^i, a_h^i} \left(\frac{t - z_h^i}{\gamma^h} \right) \right).$$

Note that when $H = 1$, (11) recovers the DR estimator for contextual bandits previously derived in Huang et al. [2021b].

Lemma 6.1. *The doubly robust estimator* (11) *is unbiased, and its variance is given recursively as:*

$$
\mathbb{V}_h^\beta \left[\widehat{F}_{S_h}(t) \right] = \mathbb{E}_h^\beta \left[\mathbb{V}_h^\beta \left[w(A_h, S_h) \Delta_{S_h, A_h}(t) \Big| S_h \right] \right]
$$
$$
+ \mathbb{E}_h^\beta \left[w(A_h, S_h)^2 \mathbb{V}_{h+1} \left[\widehat{F}_{S_{h+1}} \left(\frac{t - R_h}{\gamma} \right) \Big| S_h, A_h \right] \right]
$$
$$
+ \mathbb{V}_h \left[F_{S_h}(t) \right]
$$

where $\Delta_{s,a}(t) = \overline{F}_{s,a}(t) - F_{s,a}(t)$.

Note that the IS estimator (3) is a special case of the DR estimator when $\overline{F} = 0$ for all t, for which $\Delta_{s,a}(t)$ reduces to $-F_{s,a}(t)$ in the first term to match the variance given for the IS estimator in Lemma 4.1. When $\Delta_{s,a} < F_{s,a}$, which is often the case in practice, the variance contributed by the first term can be significantly reduced in the DR estimator. This also shows that the DR estimator can only reduce variance contributed by importance sampling for stochastic actions; it cannot reduce variance from rewards or transitions.

6.2 Error Bound and Consistency

The DR estimator is not guaranteed to be a monotone function due to the subtracted terms (particularly at lower sample sizes). In practice, we transform the DR estimator to a valid CDF using:

$$
\widehat{F}_{\text{M-DR}}(t) = \text{Clip}\left(\max_{t' \leq t} \widehat{F}_{\text{DR}}(t'), 0, 1 \right), \tag{12}
$$

which simply does not allow the CDF to decrease, and clips the value to the unit interval. Taking inspiration from Thomas et al. [2015], one potentially variance-reduced alternative to clipping the estimator is to instead use weighted importance sampling after expanding the recursion in (11). Letting $\widetilde{w}_h = \sum_i w(\tau_h^i)$ and $z_h^i = \sum_{k=1}^h \gamma^k r_k^i$,

$$
\widehat{F}_{\text{WDR}}(t) = \sum_{i=1}^n \left(\frac{w_H^i}{\widetilde{w}_H} \mathbb{1}\left\{ z_H^i \leq t \right\} \right.
$$
$$
+ \sum_{h=1}^H \frac{w_{h-1}^i}{\widetilde{w}_{h-1}} \overline{F}_{s_h^i} \left(\frac{t - z_h^i}{\gamma^h} \right)
$$
$$
\left. - \sum_{h=1}^H \frac{w_{h-1}^i}{\widetilde{w}_{h-1}} w(a_h^i, s_h^i) \overline{F}_{s_h^i, a_h^i} \left(\frac{t - z_h^i}{\gamma^h} \right) \right).
$$

The WDR estimator is difficult to analyze, and we empirically demonstrate its efficacy in Section 8. The advantage of the M-DR estimator in (12), however, is that it obeys the following finite sample error bound:

Lemma 6.2. *With probability at least $1 - \delta$,*

$$
\left\| \widehat{F}_{\text{M-DR}} - F \right\|_\infty \leq w_{max}^H \sqrt{\frac{72}{n} \log \frac{8n^{1/2}}{\delta}},
$$

and thus $\widehat{F}_{\text{M-DR}}$ is uniformly consistent.

We achieve a convergence rate of $O(w_{max}^H / \sqrt{n})$. However, we note that the above bound is conservative in the sense that it assumes worst-case model misspecification, under which the DR estimator can perform worse than IS. This contributes looseness to the result, and a tighter bound remains an open problem.

6.3 Cramer-Rao Lower Bound

Does there exists an unbiased estimator of the CDF that can even further reduce variance, e.g. in either transitions or rewards? Inspired by Jiang and Li [2016], Huang and Jiang [2020], we derive the first Cramer-Rao lower bound on the variance of off-policy CDF estimation in MDPs, and show that the DR estimator can achieve this lower bound.

Theorem 6.1. *For unique directed acyclic graphs (Definition C.1), the variance of any unbiased off-policy CDF estimator is pointwise lower bounded by*

$$
\sum_{h=1}^H \mathbb{E}_{\mathbb{P}_\beta} \left[w_{h-1}^2 \mathbb{V}_h \left[F_{S_h} \left(t - \sum_{k-1}^h R_k \right) \right] \right]
$$

This lower bound on the variance shows that the intrinsic state transition stochasticity of the underlying MDP cannot be eliminated. In fact, the variance of the DR estimator in DAGs is equal to the Cramer Rao lower bound when $\overline{F}$ is a perfect model, that is, $\overline{F}_{s_h, a_h} = F_{s_h, a_h}$ for all s_h, a_h. Further, as the error between $\overline{F}$ and F decreases, the variance approaches the Cramer-Rao lower bound.

7 Risk Estimation

We have developed several methods for CDF estimation and their error bounds, which we now translate to our goal of off-policy risk assessment. Returning to Algorithm 1, we now have $\widehat{F}$ and ϵ (line 2). We can now form a plug-in estimate of any risk functional ρ:

$$
\widehat{\rho} = \rho\left(\widehat{F} \right).
$$

In this section, we characterize the theoretical properties of the risk estimate $\widehat{\rho}$ and show that the properties of the CDF estimators translate directly to the risk functionals estimated upon them. Notably, the reduced variance of the DR CDF estimator results in risk estimates with reduced bias and variance, and for Lipschitz risk functionals (Def. 3.1), the error of $\widehat{\rho}$ is a function of ϵ which implies consistent estimation.

7.1 Bias and Variance

In general, the risk estimate $\widehat{\rho}$ may not be biased even if the CDF estimator is unbiased. However, Proposition 7.1 allows us to approximately quantify the bias and variance of the risk estimate in terms of the variance of the CDF estimator $\widehat{F}$:

Proposition 7.1. *For any thrice Fréchet differentiable risk functional ρ in metric d, let $\nabla_F \rho = \frac{\partial \rho(x)}{dx}|_{x=F}$ and $\nabla_F^2 \rho = \frac{\partial^2 \rho(x)}{dx^2}|_{x=F}$. Then the bias of $\widehat{\rho}$ is*

$$\mathbb{E}[\rho(\widehat{F})] \lessapprox \rho(F) + \frac{1}{2}||\nabla_F^2 \rho||_{op}\mathbb{V}[\widehat{F}] \qquad (13)$$

and its variance is approximately

$$\mathbb{V}[\rho(\widehat{F})] \lessapprox ||\nabla_F \rho \otimes \nabla_F \rho||_{op}\mathbb{V}[\widehat{F}] \qquad (14)$$

Proposition 7.1 has two important implications. First, the estimate of $\rho(\widehat{F})$ is approximately unbiased when the variance of $\widehat{F}$ vanishes (e.g. with large samples) or when the derivative $\nabla_F \rho$ vanishes, e.g., when g is a constant. As $g(x) = 1$ for the expected return risk functional, the latter condition allows us to recover the unbiasedness of the plug-in expected return risk estimate. Further, since $\nabla_F \rho = 1$ when ρ is the expected value, using Thm. 6.1 it can be seen that $\mathbb{V}[\rho(\widehat{F}_{\mathrm{DR}})]$ matches the Cramer-Rao variance lower bound for off-policy expected return estimation first derived in Jiang and Li [2016].

Second, both the bias (13) and the variance (14) of the risk estimate derive directly from the variance of the CDF estimator $\mathbb{V}[\widehat{F}]$. Then because both the bias and variance of $\rho(\widehat{F})$ are proportional to $\mathbb{V}[\widehat{F}]$, a CDF estimator with lower variance will also result in a risk estimate lower bias and variance. We have previously shown that when the model of the conditional distribution $\overline{F}$ is good, the DR CDF estimator can have significantly reduced variance compared to the IS estimator. Proposition 7.1 extends this variance reduction property to risk estimates as well:

Proposition 7.2. *For any twice Frechet differentiable ρ, if $\mathbb{V}[\widehat{F}_{\mathrm{DR}}] < \mathbb{V}[\widehat{F}_{\mathrm{IS}}]$, then $\mathrm{BIAS}[\rho(\widehat{F}_{\mathrm{DR}})] < \mathrm{BIAS}[\rho(\widehat{F}_{\mathrm{IS}})]$, and $Var[\rho(\widehat{F}_{\mathrm{DR}})] < \mathbb{V}[\rho(\widehat{F}_{\mathrm{IS}})]$.*

7.2 Error Bounds and Consistency

For general risk functionals, the plug-in risk estimate is also not guaranteed to be consistent even if the CDF estimator is consistent. However, CDF consistency does indeed imply risk consistency for the class of Lipschitz risk functionals (Definition 3.1) which we have previously shown to contain many popular risk functionals. Further, the error of the risk estimate is the Lipschitz constant of the risk functional times the CDF estimation error. The following error bound formalizes these results:

Corollary 7.1. *For all Lipschitz risk functionals simultaneously, given a CDF estimate $\widehat{F}$ with error ϵ, we have w.p. $\geq 1 - \delta$ that*

$$|\rho(F) - \rho_p(\widehat{F})| \leq L\epsilon, \qquad (15)$$

where L is the Lipschitz constant of ρ. Further, if $\widehat{F}$ is uniformly consistent, then $\rho(\widehat{F}) \xrightarrow{a.s.} \rho(F)$.

As a result, the downstream risk estimates from the IS or DR CDF estimators, which have error that scale with $O(1/\sqrt{n})$, are guaranteed to be consistent.

We have previously demonstrated that the IS and DRCDF estimators are minimax rate-optimal, but are their downstream risk estimates also rate-optimal? For general risk functionals this remains an open problem, but we show that for the Lipschitz risk functionals of CVaR_α and expected return, their off-policy estimation has the following lower bound:

Theorem 7.1. *For $\alpha \in [0, 1)$ and $n \geq c\frac{w_{max}^2}{\sum_\tau \mathbb{P}_\beta(\tau)w^2(\tau)}$,*

$$\inf_{\widehat{\rho}} \sup_{F \in \mathcal{F}} \mathbb{E}\left[|\widehat{\rho} - \rho(F^\pi)|\right] \geq \frac{cD}{1-\alpha}\sqrt{\frac{\sum_\tau \mathbb{P}_\beta(\tau)w^2(\tau)}{n}}$$

which, for large samples sizes, also matches existing results on off-policy mean estimation [Li et al., 2015, Wang et al., 2017, Ma et al., 2021]. As the Lipschitz constants for CVaR_α and expected return are $\frac{D}{1-\alpha}$ and D, respectively, the convergence rate of Corollary 7.1 is indeed minimax rate-optimal.

8 Experiments

We now provide empirical support for our theoretical results, and compare the performance of the CDF and plug-in risk estimators in Figure 1. We compare the doubly robust DR and WDR estimators against the importance sampling estimators — WIS, F-IS (4), S-IS (8), C-IS (§ 4.3) — and model-based DM estimator. Our primary baselines of comparison are the F-IS and WIS estimators, which are the only two off-policy CDF/risk estimators in existing work [Chandak et al., 2021a]. As we will see, the DR and WDR estimators outperform the IS-based estimators in all studied cases.

Environment We utilize two experimental domains: a tabular Cliffwalk [Sutton and Barto, 2018], and continuous diabetes treatment simulator [Xie, 2018]. The horizons of both are $H = 200$, and the behavioral policy β is a mixture between the target π and a uniform policy UNIF, i.e. $\beta = \lambda\pi + (1-\lambda)\mathrm{UNIF}$. Appendix E includes the full setup and additional experiments. Both environments have risk-related interpretations; in Cliffwalk the agent must take the shortest path while avoiding a slippery, high-cost cliff, while in Simglucose the agent must maintain a patient's glucose levels within a healthy range.

Risk Functionals As such, we estimate one each of risk-neutral, risk-averse, and risk-seeking risk functionals using OPRA: the expected return, CVaR_α, and
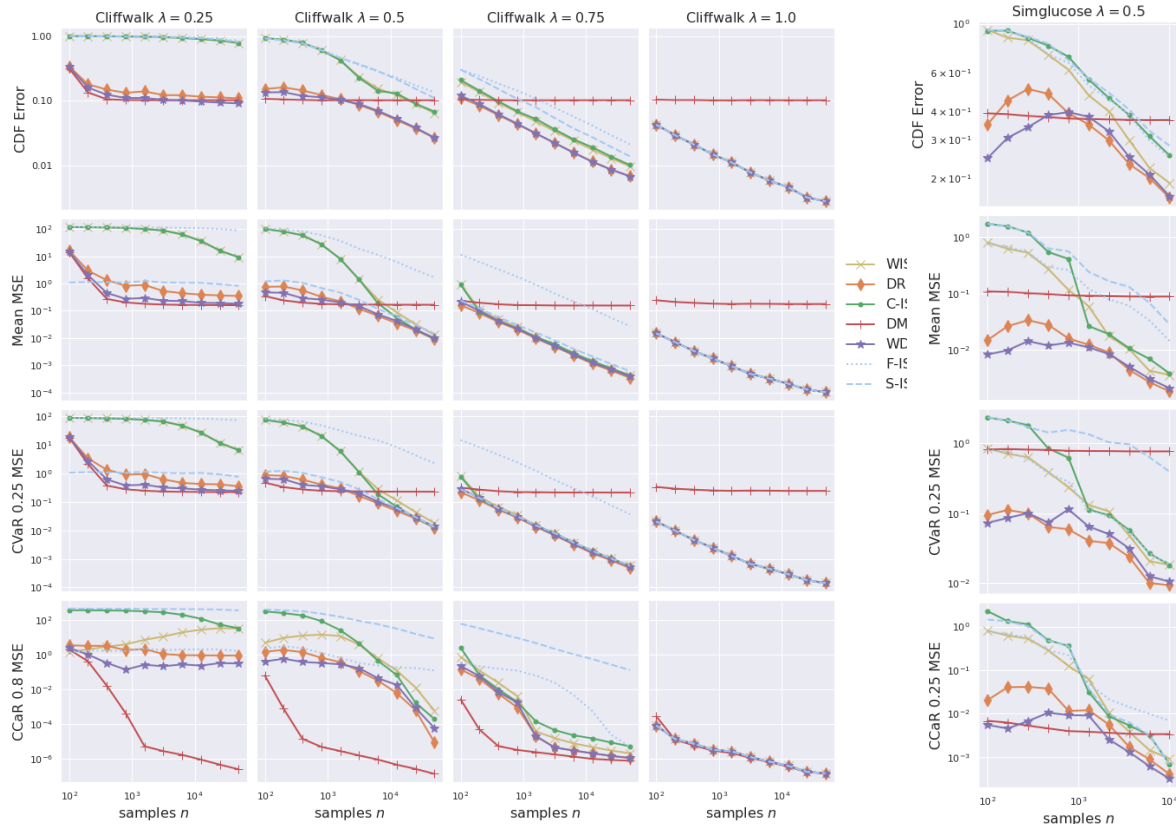
Figure 1: Sup-norm error of the CDF estimate **(row 1)**, and mean squared error of the expected return **(row 2)** and CVaR 0.25 **(row 3)** plug-in estimates for different λ **(column)** in Cliffwalk and Simglucose.

a risk functional we call CCaR_α (conditional cost-at-risk), respectively. While CVaR_α is the expectation of the worst-case α tail of the distribution, CCaR is the expectation of the best-case α tail.

Results In both environments (Fig 1), the DR and WDR estimators outperform the IS estimators and their variants in CDF and risk estimation error. These environments present challenges for IS estimators because of their long horizons, but DR and WDR effectively overcome them by incorporating model information. At very low sample sizes, however, the DM estimator can outperform the other estimators.

Our results also demonstrate the variability of CDF estimator performance for different risk functionals; the gap between F-IS and S-IS is especially obvious. Under the Cliffwalk cost distribution, S-IS far outperforms F-IS for both the risk-neutral expected return and risk-sensitive CVaR_α, For the risk-seeking CCaR_α, however, F-IS outperforms S-IS. The combined estimator C-IS never does worse than either F-IS or S-IS, and as n increases can do better than both, as Prop. 4.1 earlier implied. The DM estimator has low error under CCaR_α because the model estimates the lower tail well, but the upper tail poorly. The DR and WDR

estimator straddle best of both worlds; they perform well under all evaluated risk functionals.

9   Discussion

In this work, we have introduced and analyzed variance-reduced off-policy CDF and risk estimators for MDPs, including the first doubly robust estimator. Several future research directions are of interest. First, it is possible that estimators and bounds can be tailored and tightened for individual risk functionals, and a thorough comparison of such methods is an important avenue of future work. Second, our results have also highlighted the performance disparity between different types of estimators, which is more severe for risk-averse or risk-seeking risk functionals. As such, adaptive estimator selection (based on risk functionals of interest) may prove to be of both theoretical and practical interest. Lastly, in off-policy evaluation of the mean, marginalized importance sampling [Xie et al., 2019] has been shown to achieve the Cramer-Rao lower bound, and model-free regression [Duan et al., 2020] has been proven to be minimax optimal. We plan to investigate whether these methods can be extended to off-policy CDF or risk estimation, which, unlike the mean, are nonlinear functions of the return.

References

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Alejandro Balbás, José Garrido, and Silvia Mayoral. Properties of distortion risk measures. *Methodology and Computing in Applied Probability*, 11(3):385–399, 2009.

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

Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *arXiv preprint arXiv:2104.12820*, 2021a.

Yash Chandak, Shiv Shankar, and Philip S. Thomas. High-confidence off-policy (or counterfactual) variance estimation, 2021b.

Joel QL Chang, Qiuyu Zhu, and Vincent YF Tan. Risk-constrained thompson sampling for cvar bandits. *arXiv preprint arXiv:2011.08046*, 2020.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

Freddy Delbaen. Coherent risk measures on general probability spaces. In *Advances in finance and stochastics*, pages 1–37. Springer, 2002.

Eric V Denardo, Haechurl Park, and Uriel G Rothblum. Risk-sensitive and risk-neutral multiarmed bandits. *Mathematics of Operations Research*, 32(2):374–394, 2007.

Dieter Denneberg. Distorted probabilities and insurance premiums. *Methods of Operations Research*, 63 (3):3–5, 1990.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning, 2011.

Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Aditya Gopalan, LA Prashanth, Michael Fu, and Steve Marcus. Weighted bandits or: How bandits learn distorted values that are not expected. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. On the convergence and optimality of policy gradient for markov coherent risk, 2021a.

Audrey Huang, Liu Leqi, Zachary C Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. *arXiv preprint arXiv:2104.08977*, 2021b.

Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. In *International Conference on Machine Learning*, pages 4434–4443. PMLR, 2020.

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

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

Shigeo Kusuoka. On law invariant coherent risk measures. In *Advances in mathematical economics*, pages 83–95. Springer, 2001.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint arXiv:1810.12429*, 2018.

Cong Ma, Banghua Zhu, Jiantao Jiao, and Martin J Wainwright. Minimax off-policy evaluation for multi-armed bandits. *arXiv preprint arXiv:2101.07781*, 2021.

Shie Mannor and John Tsitsiklis. Mean-variance optimization in markov decision processes. *arXiv preprint arXiv:1104.5601*, 2011.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

LA Prashanth, Cheng Jie, Michael Fu, Steve Marcus, and Csaba Szepesvári. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, pages 1406–1415. PMLR, 2016.

LA Prashanth, Krishna Jagannathan, and Ravi Kolla. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *International Conference on Machine Learning*, pages 5577–5586. PMLR, 2020.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. *arXiv preprint arXiv:1301.1936*, 2013.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Matthew J Sobel. The variance of discounted markov decision processes. *Journal of Applied Probability*, pages 794–802, 1982.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

Scott Sussex, Omer Gottesman, Yao Liu, Susan Murphy, Emma Brunskill, and Finale Doshi-Velez. Stitched trajectories for off-policy learning.

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

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1468–1476, 2015.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research*, 17(1):361–396, 2016.

Alex Tamkin, Ramtin Keramati, Christoph Dann, and Emma Brunskill. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.

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

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

Shaun Wang. Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, 26(1):71–92, 1996.

Shaun S Wang, Virginia R Young, and Harry H Panjer. Axiomatic characterization of insurance prices. *Insurance: Mathematics and economics*, 21(2):173–183, 1997.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

Julia L Wirch and Mary R Hardy. Distortion risk measures: Coherence and stochastic dominance. In *International congress on insurance: Mathematics and economics*, pages 15–17, 2001.

Jinyu Xie. Simglucose v0. 2.1 (2018). *URL https://github. com/jxx123/simglucose*, 2018.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

A Preliminaries: Risk Overview

Let Z be a random variable. A *risk functional* ρ is a mapping from a space of random variables to the space of real numbers $\rho : \infty(\Omega, \mathcal{F}_Z, \mathbb{P}_Z) \to \mathbb{R}$. Risk functionals are law-invariant if they depend only on the distribution of Z [Kusuoka, 2001]. Formally $\rho(Z)$ is law invariant if for any pair of random variables Z and Z', $F_Z = F_{Z'} \implies \rho(Z) = \rho(Z')$.

Lipschitz risk functionals are a subset of law-invariatn risk functionals, and we now delineate popular classes of risks and their associated Lipschitz constants, where possible, adapted from Huang et al. [2021a].

Distorted Risk Functionals For $Z \geq 0$, distortion risk functionals are defined to be Denneberg [1990], Wang [1996], Wang et al. [1997], Balbás et al. [2009]

$$\rho(F_Z) = \int_0^\infty g(1 - F_Z(t))dt = \int_0^\infty g(S_Z(t))dt$$

where the distortion $g : [0, 1] \to [0, 1]$ has $g(0) = 0$ and $g(1) = 1$, and is increasing. When $g(x) = x$, the expected value is recovered. When $g(x) = \min\{\frac{x}{1-\alpha}, 1\}$ for $\alpha \in (0, 1)$, CVaR at level α is recovered.

When $g(x) = F(F^{-1}(x) - F^{-1}(\alpha))$, the Wang risk functional at level α [Wang, 1996] is recovered, and the proportional hazard risk functional can by obtained by setting $g(x) = x^\alpha$ for $\alpha < 1$. Distorted risk functionals are coherent if and only if g is concave [Wirch and Hardy, 2001]. Not all distorted risk functionals are coherent. For example, setting $g(s) = \mathbb{1}_{\{s \geq 1-\alpha\}}$ recovers the value-at-risk (VaR), which is not coherent.

Coherent Risk Functionals Coherent risk functionals satisfy properties called monotonicity, subadditivity, translation invariance, and positive homogeneity Artzner et al. [1999], Delbaen [2002]. Examples include expected value, conditional value-at-risk (CVaR), entropic value-at-risk, and mean semideviation, proportional hazard, and Wang transform [Chang et al., 2020, Tamkin et al., 2019, Tamar et al., 2015, Shapiro et al., 2014, Wirch and Hardy, 2001, Wang, 1996].

Lemma A.1 (Lipschitzness of Coherent and Distorted Risk Functionals Huang et al. [2021b]). *On the space of random variables with support in $[0, D]$, the distorted risk functional of any $\frac{L}{D}$-Lipschitz distortion function is a L-Lipschitz risk functional.*

The expected value risk functional is D-Lipschitz and $CVaR_\alpha$ is $\frac{D}{\alpha}$-Lipschitz.

Cumulative Prospect Theory (CPT) Risk Functionals CPT risks Prashanth et al. [2016] are defined as:

$$\rho(F_Z) = \int_0^{+\infty} g^+ \left(1 - F_{u^+(Z)}(t)\right) dt - \int_0^{+\infty} g^- \left(1 - F_{u^-(Z)}(t)\right) dt,$$

where $g^+, g^- : [0, 1] \to [0, 1]$ are continuous with $g^{+/-}(0) = 0$ and $g^{+/-}(1) = 1$. The functions $u^+, u^- : \mathbb{R} \to \mathbb{R}_+$ are continuous, with $u^+(z) = 0$ when $z \geq c$ and $u^-(z) = 0$ when $z < c$ for some constant $c \in \mathbb{R}$. The CPT functional handles gains and losses separately, applying distortion g^+ to the part of the distribution corresponding to gains (controlled by u^+), and g^- to the part corresponding to losses (controlled by u^-).

Lemma A.2 (Lipschitzness of CPT Functional). *If the CPT distortion functions g^+ and g^- are both $\frac{L}{D}$-Lipschitz, then the CPT risk functional is L-Lipschitz.*

Other Risk Functionals Many other risk functionals, though not members of a specific class, can be shown to be Lipschitz. This includes the variance and mean-variance. The variance is defined as $\rho(F_Z) = 2\int_0^\infty t(1 - F_Z(t))dt - \left(\int_0^\infty (1 - F_Z(t))dt\right)^2$.

Lemma A.3 (Lipschitzness of Variance). *On the space of random variables with support in $[0, D]$, variance is a $3D^2$-Lipschitz risk functional. For mean-variance given by $\rho(Z) = \mathbb{E}[Z] + \lambda \mathbb{V}(Z)$ for some $\lambda > 0$, it is $(1 + 3\lambda D^2)$-Lipschitz.*

B   Proofs for Importance Sampling CDF Estimation (Section 4)

B.1   Bias and Variance (Proof of Lemma 4.1)

As the IS estimator can be seen as a special case of the DR estimator with $\overline{F} = 0$, we defer the proof of bias and variance to Appendix C.1.

B.2   Error Upper Bound (Proof of Lemma 4.2)

The proof of the upper bound closely follows the proof of Theorem 5.1 from Huang et al. [2021b].

For any $\lambda > 0$, we have

$$
\begin{aligned}
\lambda \mathbb{E}\left[\|\widehat{F} - F\|_\infty\right] &= \mathbb{E}\left[\log \exp\left(\lambda\|\widehat{F} - F\|_\infty\right)\right] \\
&\leq \log \mathbb{E}\left[\exp\left(\lambda\|\widehat{F} - F\|_\infty\right)\right]
\end{aligned}
$$

To upper bound the RHS, define the following function class:

$$
\mathbb{F}(n) := \left\{ f(g) := \varrho \frac{1}{n} \mathbb{1}_{\{g \leq t\}} : \forall t \in \mathbb{R}; \forall g \in \mathbb{Q}, \varrho \in \{-1, +1\} \right\}
$$

Note that this is a countable set. Denote $w(\tau) = \prod_{h=1}^{H} w(a_h, s_h)$, and $g^i = \sum_{h=1}^{H} \gamma^h r_h^i$ the return of trajectory i for short. Using this definition, we have

$$
\sup_{t \in \mathbb{R}} \left|\widehat{F}_{\mathrm{IS}}(t) - F(t)\right| = \sup_{f \in \mathbb{F}(n)} \left| \frac{1}{n} \sum_{i=1}^{n} w(\tau^i) f(g^i) - \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{1}{n} \sum_{i=1}^{n} w(\tau^i) f(g^i)\right] \right|
$$

Then adapting the proof of Theorem 5.1 in Huang et al. [2021b], we have for any $\lambda \in (0, \frac{n}{2w_{max}})$ that,

$$
\mathbb{E}\left[\exp\left(\lambda\|\widehat{F} - F\|_\infty\right)\right] \leq 4 \exp\left(\frac{n\lambda^2 \frac{4\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2])}{n^2}}{2\left(1 - \lambda \frac{2}{n} w_{max}\right)}\right)
$$

Then

$$
\begin{aligned}
\mathbb{E}\left[\|\widehat{F} - F\|_\infty\right] &\leq \frac{1}{\lambda} \log\left(4 \exp\left(\frac{n\lambda^2 \frac{4\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2])}{n^2}}{2\left(1 - \lambda \frac{2}{n} w_{max}\right)}\right)\right) \\
&\leq \frac{\log 4}{\lambda} + \frac{2\lambda \mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]}{n(1 - \lambda \frac{2}{n} w_{max})}
\end{aligned}
\tag{16}
$$

Since this holds for any $\lambda \in (0, \frac{n}{2w_{max}})$, it also holds for $\min_\lambda$ of the RHS. To find the minimizer, we solve

$$
\frac{d}{d\lambda}\left(\frac{\log 4}{\lambda} + \frac{2\lambda \mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]}{n - 2\lambda w_{max}}\right) = 0
$$

which WolframAlpha tells us is solved by

$$
\lambda^* = \frac{n}{\sqrt{\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]n \log 2} + 2w_{max}}
$$

if $\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]n \neq w_{max}^2 \log 16$. We observe that $\lambda^* \in (0, \frac{n}{2w_{max}})$ satisfies the constraints on λ. Plugging λ^* back

into (16) we have,

$$
\mathbb{E}\left[\|\widehat{F} - F\|_\infty\right] \le \frac{\log 4}{\frac{n}{\sqrt{\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]n\log 2} + 2w_{max}}} + \frac{\frac{2n\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]}{\sqrt{\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]n\log 2} + 2w_{max}}}{n - \frac{2w_{max}n}{\sqrt{\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]n\log 2} + 2w_{max}}}
$$

$$
= \log 4 \sqrt{\log 2} \sqrt{\frac{\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]}{n}} + \frac{2\log 4 w_{max}}{n} + 2\sqrt{\frac{\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]}{n\log 2}}
$$

$$
\le c_1 \sqrt{\frac{\mathbb{E}_{\mathbb{P}_\beta}[w(\tau)^2]}{n}} + c_2 \frac{w_{max}}{n}
$$

B.3   CDF Estimation Lower Bound (Proof of Theorem 4.1)

We begin by deriving a CDF estimation lower bound in multi-armed bandits (MABs). We can reduce the estimation problem in the MDP case to the MAB case by treating entire trajectories τ as "actions".

Let a multi-armed bandit be defined by K arms, and suppose the behavioral policy is given by $\beta(a)$ for $a \in [K]$, and the target policy is $\pi(a)$ for $a \in [K]$, with importance weight $w(a) := \beta(a)/\pi(a)$. Let $w_{max} = \max_a w(a)$.

Lemma B.1 (CDF Lower Bound in MABs). *When $n \ge c\frac{w_{max}^2}{\sum_a \beta(a)w^2(a)}$ for a universal constant c, the minimax lower bound on off-policy CDF estimation in MABs is*

$$
\inf_{\widehat{F}} \sup_{F \in \mathcal{F}} \mathbb{E}\left[\|\widehat{F} - F^\pi\|_\infty\right] \ge c\sqrt{\frac{\sum_{a \in [K]} \beta(a)w^2(a)}{n}}. \tag{17}
$$

We now lift this result to the MDP case. For a horizon of length H, the space of possible actions of size $|\tau| = |\mathcal{S}|^{H+1}|\mathcal{A}|^H$. Recall that in MDPs, $w_{max} = \max_\tau w(\tau)$, and $w(\tau) = \frac{\pi(\tau)}{\beta(\tau)} = \prod_{h=1}^H \frac{\pi(a_h|s_h)}{\beta(a_h|s_h)}$. Thus using Lemma B.1, when $n \ge c\frac{w_{max}^2}{\sum_\tau \mathbb{P}_\beta(\tau)w^2(\tau)}$, we have

$$
\inf_{\widehat{F}} \sup_{F \in \mathcal{F}} \mathbb{E}\left[\|\widehat{F} - F^\pi\|_\infty\right] \ge c\sqrt{\frac{\sum_\tau \mathbb{P}_\beta(\tau)w^2(\tau)}{n}}. \tag{18}
$$

This matches our Bernstein-style bound on the error of IS estimation up to a constant:

$$
\mathbb{E}\left[\|\widehat{F} - F\|_\infty\right] \le c_1 \sqrt{\frac{\sum_\tau \mathbb{P}_\beta(\tau)w^2(\tau)}{n}} + c_2 \frac{w_{max}}{n}. \tag{19}
$$

Proof of Lemma B.1. We use Le Cam's method to derive the lower bound. We first need to select two problem MAB problem instances, by defining their reward distributions R_1, R_2. Then define distributions $P_1 = \beta \circ R_1$, and $P_2 = \beta \circ R_2$.

Intuitively, we want to select them so that the difference $\|F_1^\pi - F_2^\pi\|_\infty$ is large enough that an estimator can distinguish between them, but with the distance $\mathrm{D}_{\mathrm{KL}}(P_1\|P_2)$ to be small enough that the testing problem is not too easy.

For each action $a \in [K]$, let $R_2(\cdot|a)$ be a Bernoulli distribution on the set $\{0,1\}$ with probability $p = 1/2$, and let $R_1(\cdot|a)$ be a Bernoulli distribution with $p = 1/2 + \delta(a)$ for some $\delta(a) \in [0, 1/2]$. Then

$$
\|F_1^\pi - F_2^\pi\|_\infty = \max_{t \in \{0,1\}} |F_1^\pi(t) - F_2^\pi(t)|
$$

$$
= |F_1^\pi(t = 0) - F_2^\pi(t = 0)|
$$

$$
= \left|\sum_a \pi(a)\delta(a)\right|
$$

At the same time,

$$D_{\mathrm{KL}}(P_1\|P_2) = \sum_a \beta(a) D_{\mathrm{KL}}(R_1\|R_2) \le 4 \sum_a \beta(a) \delta^2(a)$$

Since we want the probability of error in our test to be a constant, we want to upper bound the above by $\frac{1}{2n}$. Thus to obtain as tight a lower bound as possible, we want to solve the optimization problem

$$\max_{\delta \in \mathbb{R}^K} \sum_a \pi(a)\delta(a) \quad \text{s.t.} \quad 4\sum_a \beta(a)\delta^2(a) \le \frac{1}{2n}, \ \delta \in [0, 1/2] \tag{20}$$

Using Lemma 3 from Ma et al. [2021], the Fenchel dual of the optimization problem in (21) is

$$\min_{\nu \in \mathbb{R}^K} \sqrt{\frac{1}{8n} \sum_a \frac{(\pi(a) - \nu(a))^2}{\beta(a)}} + \frac{1}{2} \sum_a |\nu(a)| \tag{21}$$

Let ν^* be the solution to (21) and define the set of actions

$$S^* = \{a \in [K] | \nu^*(a) > 0\}$$

Then Lemma 2 from Ma et al. [2021] states that

$$\min_{\nu \in \mathbb{R}^K} \sqrt{\frac{1}{8n} \sum_a \frac{(\pi(a) - \nu(a))^2}{\beta(a)}} + \frac{1}{2} \sum_a |\nu(a)| \asymp \pi(S^*) + \sqrt{\frac{\sum_{a \notin S^*} \beta(a) w^2(a)}{n}}$$

Thus we have that for some universal constant c,

$$\inf_{\widehat{F}} \sup_{F \in \mathcal{F}} \mathbb{E}\left[\|\widehat{F} - F^\pi\|_\infty\right] \ge c \left\{ \pi(S^*) + \sqrt{\frac{\sum_{a \notin S^*} \beta(a) w^2(a)}{n}} \right\} \tag{22}$$

When we have sample size n such that $n \ge c \frac{w_{max}^2}{\sum_a \beta(a) w^2(a)}$, $S^* = \emptyset$ and the lemma statement follows. □

B.4 Proofs for Section 4.3

B.4.1 Bernstein Bounds for S and F (Proof of Lemma 4.3)

Definition B.1 (Bounded Variation). *A function f has bounded variation if there exists $C < \infty$ such that*

$$\sup_{N, \{t_i\}_{i=1}^N} \sum_{i=1}^N |f(t_{i+1}) - f(t_i)| \le C$$

Assumption B.1. *Suppose that the random variable of returns takes values on a compact set. Assume the variance of the IS estimator $\mathbb{V}^{\widehat{F}}(t)$ has bounded variation for all $t \in \mathbb{R}$.*

Lemma B.2. *If a function f has bounded variation, then there exists monotone functions f^+, f^- with bounded variation such that $\forall t$,*

$$f(t) = f^+(t) - f^-(t)$$

Proof. Using Lemma B.2, let $\mathbb{V}^{\widehat{F}} = \mathbb{V}^+ - \mathbb{V}^-$, and note that $\mathbb{V}^+, \mathbb{V}^- \in [0, d_\infty]$ and are monotone functions. As they are monotone functions, let $\{s_j^+\}_{j=1}^M$ and $\{s_j^-\}_{j=1}^M$ be the respective stepping points for their stepwise approximations. Further, let $\{s_k^F\}_{k=1}^{M'}$ be the stepping points for the CDF F. Finally, overriding notation slightly, let $\{s_j\}_{j=1}^{M'+2M}$ be the union of these three sets of stepping points. We know that between adjacent stepping

points, F, $\mathbb{V}^+$, and $\mathbb{V}^-$ do not change much. Let $\zeta^F, \zeta^+, \zeta^-$ be the step functions constructed using the stepping points for each function, respectively, and let $\zeta^{\mathbb{V}^{\widehat{F}}} = \zeta^+ - \zeta^-$. Formally, this means that

$$\sup_{t\in\mathbb{R}} |F(t) - \zeta^F(t)| \leq 1/2M'$$

$$\sup_{t\in\mathbb{R}} |\mathbb{V}^+(t) - \zeta^+(t)| \leq 1/2M$$

$$\sup_{t\in\mathbb{R}} |\mathbb{V}^-(t) - \zeta^-(t)| \leq 1/2M$$

$$\sup_{t\in\mathbb{R}} |\mathbb{V}^{\widehat{F}}(t) - \zeta^{\mathbb{V}^{\widehat{F}}}(t)| \leq 1/M$$

At a given s_j, Bernstein's inequality gives us

$$|F(s_j) - \widehat{F}(s_j)| \leq \frac{c_1 \log(1/\delta)}{n} + \sqrt{\frac{c_2 \mathbb{V}^{\widehat{F}}(s_j)\log(1/\delta)}{n}}$$

$$= \frac{c_1 \log(1/\delta)}{n} + \sqrt{\frac{c_2 \left(\mathbb{V}^+(s_j) - \mathbb{V}^-(s_j)\right)\log(1/\delta)}{n}}$$

Taking a union bound, this holds for all $\{s_j\}_{j=1}^{M'+2M}$ with probability at least $1 - \delta(M'+2M)$.

For each j and $t \in [s_j, s_{j+1}]$, if $F(t) \geq \widehat{F}(t)$,

$$F(t) - \widehat{F}(t) \leq F(s_{j+1}) - \widehat{F}(s_j)$$

$$\leq \frac{1}{M'} + F(s_j) - \widehat{F}(s_j)$$

$$\leq \frac{1}{M'} + \frac{c_1 \log(1/\delta)}{n} + \sqrt{\frac{c_2 \mathbb{V}^{\widehat{F}}(s_j)\log(1/\delta)}{n}}$$

$$\leq \frac{1}{M'} + \frac{c_1 \log(1/\delta)}{n} + \sqrt{\frac{c_2 \mathbb{V}^{\widehat{F}}(t)\log(1/\delta)}{n}} + \sqrt{\frac{c_2 \log(1/\delta)}{Mn}}$$

If $\widehat{F}(t) > F(t)$,

$$\widehat{F}(t) - F(t) \leq \widehat{F}(s_{j+1}) - F(s_j)$$

$$\leq \frac{1}{M'} + |\widehat{F}(s_{j+1}) - F(s_{j+1})|$$

$$\leq \frac{1}{M'} + \frac{c_1 \log(1/\delta)}{n} + \sqrt{\frac{c_2 \mathbb{V}^{\widehat{F}}(s_{j+1})\log(1/\delta)}{n}}$$

$$\leq \frac{1}{M'} + \frac{c_1 \log(1/\delta)}{n} + \sqrt{\frac{c_2 \mathbb{V}^{\widehat{F}}(t)\log(1/\delta)}{n}} + \sqrt{\frac{c_2 \log(1/\delta)}{Mn}}$$

Choosing $M, M' \propto \sqrt{n}$ gives the resulting bound.

Similarly, for $\widehat{F}_{\text{S-IS}}$, we have

$$F - \widehat{F}_{\text{S-IS}} = 1 - S - (1 - \widehat{S}_{\text{IS}}) = \widehat{S}_{\text{IS}} - S$$

The remainder of the proof occurs in the same manner, but with $\mathbb{V}[\widehat{S}_{\text{IS}}]$ on the RHS. $\qquad\square$

B.4.2 Empirical Bernstein Bound (Proof of Lemma 4.4)

Proof. We begin with the Bernstein bounds for $\widehat{F}_{\text{F-IS}}(t_j)$, $\widehat{F}_{\text{S-IS}}(t_j)$ at a given t_j. That is,

$$|F(t_j) - \widehat{F}_{\text{F-IS}}(t_j)| \leq \frac{c_1 \log(1/\delta)}{n} + \sqrt{\frac{c_2 \mathbb{V}[\widehat{F}_{\text{IS}}(t_j)] \log(1/\delta)}{n}}$$

and a similar statement holds for $\widehat{F}_{\text{S-IS}}$. From Maurer and Pontil [2009], we have the following bound on estimation of the variance:

Lemma B.3 (Theorem 10 from Maurer and Pontil [2009])**.** *With probability at least $1 - \delta$, for random variables $X \in [0, 1]$,*

$$\left| \sqrt{\mathbb{V}_n} - \sqrt{\mathbb{V}} \right| \leq \sqrt{\frac{2 \ln 1/\delta}{n - 1}} \tag{23}$$

Taking a union bound between Lemma B.3 and Lemma 4.3 at all M points, we have that with probability at least $1 - \delta$ that for all t,

$$\left| F(t) - \widehat{F}_{\text{F-IS}}(t) \right| \leq \frac{c_1 \log(2M/\delta)}{n} + \sqrt{\frac{c_2 \mathbb{V}_n[\widehat{F}_{\text{IS}}(t)] \log(2M/\delta)}{n}}$$
$$+ \sqrt{\frac{2c_2 \log^2(2M/\delta)}{n(n-1)}} + \sqrt{\frac{c_2 \log(2M/\delta)}{Mn}}$$

Similar steps give the bound for $\widehat{F}_{\text{S-IS}}$, with $\mathbb{V}_n[\widehat{S}_{\text{IS}}(t)]$ on the RHS.

$\square$

B.4.3 Empirical Bernstein Bound for Combined Estimator (Proof of Proposition 4.1)

In order to correctly choose either $\widehat{F}_{\text{F-IS}}$ or $\widehat{F}_{\text{S-IS}}$ for each t_j, we need that for all t_j, the empirical variance estimates $\mathbb{V}_n$ are separated by at least $\epsilon_{\mathbb{V}}(t) = \frac{1}{2}|\mathbb{V}[\widehat{F}_{\text{F-IS}}(t)] - \mathbb{V}[\widehat{F}_{\text{S-IS}}(t)]| \leq \frac{1}{2}|\mathbb{V}_n[\widehat{F}_{\text{F-IS}}(t)] - \mathbb{V}_n[\widehat{F}_{\text{S-IS}}(t)]| + \sqrt{\frac{\ln 1/\delta'}{2(n-1)}}$. For this to be the case, we set

$$\delta(t) = \frac{1}{2} \exp\left(-\frac{1}{2}(n-1)\epsilon_{\mathbb{V}}(t)^2 \right),$$

and $\delta = \max_j \delta(t_j)$. Thus if

$$n \geq \frac{8 \log(2n/\delta)}{\max_j \epsilon_{\mathbb{V}}(t_j)^2},$$

then we correctly choose $\widehat{F}_{\text{F-IS}}$ or $\widehat{F}_{\text{S-IS}}$ for each t_j, which gives the result.

C Proofs for Doubly Robust CDF Estimation (Section 6)

C.1 Bias and Variance (Proof of Lemma 6.1)

We will prove the unbiasedness of the DR estimator by induction. As the base case, $F^0_{s_{H+1}}(t) = 0$ for all states s and for all t, which is unbiased by definition.

As the induction assumption, we assume that the DR estimator at the step $h + 1$ is unbiased, that is $\mathbb{E}_{h+1}[\widehat{F}_{S_{h+1}}(t)] = \mathbb{E}_{h+1}[F_{S_{h+1}}](t)$.

We need to prove that it is also unbiased for the hth step. Recall for now we assume that rewards are deterministic. Then for any t:

$$
\begin{aligned}
\mathbb{E}^{\beta}_h[\widehat{F}_{S_h}(t)] &= \mathbb{E}_{\mathbb{P}_\beta} \left[\overline{F}^{H+1-h}_{S_h}(t) + w(A_h, S_h) \left(\widehat{F}^{H-h}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) - \overline{F}^{H+1-h}_{S_h, A_h}(t) \right) \right] \\
&= \mathbb{E}^{\beta}_h \left[\overline{F}^{H+1-h}_{S_h}(t) - w(A_h, S_h) \overline{F}^{H+1-h}_{S_h, A_h}(t) \right] + \mathbb{E}^{\beta}_h \left[w(A_h, S_h) \widehat{F}^{H-h}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) \right] \\
&= \mathbb{E}_h \left[\overline{F}^{H+1-h}_{Z(S_h)}(t) - \mathbb{E}_h \left[\overline{F}^{H+1-h}_{S_h, A_h}(t) \Big| S_h \right] \right] \\
&\quad + \mathbb{E}_h \left[\widehat{F}^{H-h}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) \right] \\
&= \mathbb{E}_h \left[\overline{F}^{H+1-h}_{Z(s_h)}(t) - \overline{F}^{H+1-h}_{Z(s_h)}(t) \right] + \mathbb{E}_h \left[\mathbb{E}_{h+1} \left[\widehat{F}^{H-h+1}_{S_h+1} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) \right] \right] \\
&= \mathbb{E}_h \left[\mathbb{E}_{h+1} \left[F^{H-h+1}_{S_h+1} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) \right] \right] \\
&= \mathbb{E}_h \left[F_{S_h}(t) \right]
\end{aligned}
$$

where the third equality incorporates the importance sampling weight w into a change of measure from β to π, the foruth equality uses the conditional expectation. The second to last equality uses the induction assumption and the last uses the recursive identity.

Next, we derive the variance of the DR estimator.

$$
\begin{aligned}
\mathbb{V}^{\beta}_h \left[\widehat{F}_{S_h}(t) \right] &= \mathbb{E}^{\beta}_h \left[\left(\widehat{F}_{S_h}(t) \right)^2 \right] - \mathbb{E}^{\beta}_h \left[\widehat{F}_{S_h}(t) \right]^2 \\
&= \mathbb{E}^{\beta}_h \left[\left(\overline{F}^{H+1-h}_{S_h}(t) + w(A_h, S_h) \left(\widehat{F}^{H-h}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) - \overline{F}^{H+1-h}_{S_h, A_h}(t) \right) \right)^2 - F_{S_h}(t)^2 \right] \\
&\quad + \mathbb{E}^{\beta}_h \left[F_{S_h}(t)^2 \right] - \mathbb{E}^{\beta}_h \left[F_{S_h}(t) \right]^2 \\
&= \mathbb{E}^{\beta}_h \left[\left(\overline{F}^{H+1-h}_{S_h}(t) + w(A_h, S_h) \left(\widehat{F}^{H-h}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) - \overline{F}^{H+1-h}_{S_h, A_h}(t) \right) \right)^2 - F_{S_h}(t)^2 \right] \\
&\quad + \mathbb{V}_h \left[F_{S_h}(t) \right] \\
&= \mathbb{E}^{\beta}_h \Bigg[\Bigg(\overline{F}^{H+1-h}_{S_h}(t) + w(A_h, S_h) F_{S_h, A_h}(t) - w(A_h, S_h) \overline{F}^{H+1-h}_{S_h, A_h} \\
&\quad + w(A_h, S_h) \left(\widehat{F}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) - F_{S_h, A_h}(t) \right) \Bigg)^2 - F_{S_h}(t)^2 \Bigg] + \mathbb{V}_h \left[F_{S_h}(t) \right] \\
&= \mathbb{E}^{\beta}_h \Bigg[\Bigg(\overline{F}^{H+1-h}_{S_h}(t) - w(A_h, S_h) \Delta_{S_h, A_h}(t) \\
&\quad + w(A_h, S_h) \left(\widehat{F}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) - \mathbb{E}_{h+1} \left[\widehat{F}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) \right] \right) \Bigg)^2 - F_{S_h}(t)^2 \Bigg] \\
&\quad + \mathbb{V}_h \left[F_{S_h}(t) \right] \\
&= \mathbb{E}^{\beta}_h \left[\left(\overline{F}^{H+1-h}_{S_h}(t) - w(A_h, S_h) \Delta_{S_h, A_h}(t) \right)^2 - F_{S_h}(t)^2 \right]
\end{aligned}
$$

$$+ \mathbb{E}_h \left[\left(w(A_h, S_h) \left(\widehat{F}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) - \mathbb{E}_{h+1} \left[\widehat{F}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) \right] \right) \right)^2 \right]$$

$$+ \mathbb{V}_h \left[F_{S_h}(t) \right]$$

$$= \mathbb{E}_h^\beta \left[\mathbb{V}_h^\beta \left[w(A_h, S_h) \Delta_{S_h, A_h}(t) \Big| S_h \right] \right] + \mathbb{E}_h^\beta \left[w(A_h, S_h)^2 \mathbb{V}_{h+1} \left[\widehat{F}_{S_{h+1}} \left(\frac{t - R(S_h, A_h)}{\gamma} \right) \Big| S_h, A_h \right] \right]$$

$$+ \mathbb{V}_h \left[F_{S_h}(t) \right]$$

where $\Delta_{s,a}(t) = \overline{F}_{s,a}(t) - F_{s,a}(t)$.

C.2 Error Bound on Doubly Robust Estimator (Proof of Lemma 6.2)

We separate the bound into two parts using the definition of $\widehat{F}_{\mathrm{DR}}$:

$$
\left\| F - \widehat{F}_{\mathrm{DR}} \right\|_\infty \leq \left\| F - \frac{1}{n} \sum_{i=1}^n w(\tau^i) \mathbb{1} \left\{ \sum_{h=1}^H \gamma^h r_h \leq t \right\} \right\|_\infty
$$

$$
+ \left\| \frac{1}{n} \sum_{i=1}^n w(\tau_{1:h-1}^i) \left(w(s_h^i, a_h^i) \overline{F}_{S_h, A_h} \left(\frac{t - \sum_{k=1}^h R_k^i}{\gamma^h} \right) - \overline{F}_{S_h} \left(\frac{t - \sum_{k=1}^h R_k^i}{\gamma^h} \right) \right) \right\|_\infty \tag{24}
$$

Lemma 4.2 upper bounds the first term, and we are left to bound the second.

Let $\tau = (s^0, a^0, r^0, \ldots, s^h, a^h, r^h, \ldots, s^H, a^H, r^H, s^{H+1})$ be a trajectory, and define $w(\tau_{1:h}) = \prod_{k=1}^h w(s_h, a_h)$. Define the function class

$$
\mathbb{F}(m, w) = \left\{ f\left(\{\{x_j^h\}_{j=1}^M \}_{h=1}^H, \tau \right) := \frac{\varrho}{m} \sum_{h=1}^H w(\tau_{1:h-1}) \sum_{j=1}^m w(a^h, s^h) \mathbb{1}_{\{x_j^h \leq t\}} : \forall t \in \mathbb{R}, \varrho \in \{-1, +1\}, \{x_j^h\}_{j,h=1}^{M,H} \in \mathbb{Q}^{M \times H} \right\}
$$

Note that (24) is

$$
\leq \sup_{\zeta \in \mathbb{F}(m, w)} \left| \frac{1}{n} \sum_{i=1}^n \zeta \left(\{x_{S_i^h, A_i^h}^j\}, \tau_i \right) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbb{P}_\beta} \left[\zeta \left(\{x_{S_i^h, A_i^h}^j\}, \tau_i \right) \Big| \{S_i^h\}_{h=1}^H \right] \right| + \frac{1}{m}
$$

Going forward, we refer to $\zeta \left(\{x_{S_i^h, A_i^h}^j\}, \tau \right)$ as $\zeta(\tau_i)$ for short. For $\lambda > 0$ we have

$$
\mathbb{E}_{\mathbb{P}_\beta} \left[\exp \left(\lambda \sup_{\zeta \in \mathbb{F}(m, w)} \left(\frac{1}{n} \sum_{i=1}^n \zeta(\tau_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbb{P}_\beta} \left[\zeta(\tau_i) \Big| \{S_i^h\}_{h=1}^H \right] \right) \right) \right]
$$

$$
= \mathbb{E}_{\mathbb{P}_\beta} \left[\exp \left(\lambda \sup_{\zeta \in \mathbb{F}(m, w)} \left(\frac{1}{n} \sum_{i=1}^n \zeta(\tau_i) - \frac{1}{n} \sum_{i=1}^n \zeta(\tau_i') \right) \right) \right]
$$

where $\tau_i' = \{s_i^h, a_i'^h\}_{h=1}^H$. Using symmetrization, this is

$$
\leq \mathbb{E}_{\mathbb{P}_\beta, \Re} \left[\exp \left(2\lambda \sup_{\zeta \in \mathbb{F}(m, w)} \frac{1}{n} \sum_{i=1}^n \xi_i \zeta(\tau_i) \right) \right]
$$

$$
= \mathbb{E}_{\mathbb{P}_\beta, \Re} \left[\sup_{\zeta \in \mathbb{F}(m, w)} \exp \left(2\lambda \frac{1}{n} \sum_{i=1}^n \xi_i \zeta(\tau_i) \right) \right]
$$

$$
= \mathbb{E}_{\mathbb{P}_\beta, \Re} \left[\sup_{\varrho, t} \exp \left(2\lambda \frac{\varrho}{nm} \sum_{i=1}^n \xi_i \sum_{h=1}^H w(\tau_i^{1:h-1}) \sum_{j=1}^m w(A_i^h, S_i^h) \mathbb{1}_{\{x_{S_i^h, A_i^h}^j \leq t\}} \right) \right]
$$

Now for each h, j, permute the indices of i with a new indexing (i) such that

$$
x_{S_{(1)}^h, A_{(1)}^h}^j \leq \ldots \leq x_{S_{(i)}^h, A_{p(i)}^h}^j \leq \ldots \leq x_{S_{(n)}^h, A_{(n)}^h}^j.
$$

Then we have

$$\mathbb{E}_{\mathbb{P}_{\beta}, \Re} \left[\sup_{\varrho, t} \exp \left(2\lambda \frac{\varrho}{nm} \sum_{i=1}^{n} \xi_i \sum_{h=1}^{H} w(\tau_i^{1:h-1}) \sum_{j=1}^{m} w(A_i^h, S_i^h) \mathbb{1}_{\{x_{S_i^h, A_i^h}^j \leq t\}} \right) \right]$$

$$= \mathbb{E}_{\mathbb{P}_{\beta}, \Re} \left[\max_{\varrho, k} \exp \left(2\lambda \frac{\varrho}{nm} \sum_{h=1}^{H} \sum_{j=1}^{m} \sum_{i=1}^{k} \xi_{(i)} w(\tau_{(i)}^{1:h-1}) w(A_{(i)}^h, S_{(i)}^h) \right) \right]$$

$$\leq \mathbb{E}_{\mathbb{P}_{\beta}, \Re} \left[\max_{\varrho, j, k} \exp \left(2\lambda \frac{\varrho}{n} \sum_{i=1}^{k} \xi_{(i)} \sum_{h=1}^{H} w(\tau_{(i)}^{1:h-1}) w(A_{(i)}^h, S_{(i)}^h) \right) \right]$$

$$\leq 2 \sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_{\beta}, \Re} \left[\exp \left(\frac{2\lambda}{n} \sum_{i=1}^{n} \xi_{j(i)} \sum_{h=1}^{H} w(\tau_{j(i)}^{1:h}) \right) \mathbb{1}_{\{\sum_{i=1}^{n} \xi_{j(i)} \sum_{h=1}^{H} w(\tau_{j(i)}^{1:h}) \geq 0\}} \right]$$

$$\leq 2m \mathbb{E}_{\mathbb{P}_{\beta}, \Re} \left[\exp \left(\frac{2\lambda}{n} \sum_{i=1}^{n} \xi_{j(i)} \sum_{h=1}^{H} w(\tau_{j(i)}^{1:h}) \right) \right]$$

$$\leq 2m \exp \left(\frac{2\lambda^2 w_{max}(w_{max}^H - 1)}{n(w_{max} - 1)} \right)$$

And we have the result,

$$\mathbb{P}_{\beta} \left(\cdot \geq \epsilon + \frac{1}{m} \right) \leq 4m \exp \left(\frac{2\lambda^2 w_{max}(w_{max}^H - 1)}{n(w_{max} - 1)} - \lambda \epsilon \right)$$

$$\leq 4m \min_{\lambda > 0} \exp \left(\frac{2\lambda^2 w_{max}(w_{max}^H - 1)}{n(w_{max} - 1)} - \lambda \epsilon \right)$$

$$= 4m \exp \left(\frac{-n\epsilon(w_{max} - 1)^2}{w_{max}^2(w_{max}^H - 1)^2} \right)$$

where the second line results because the inequality holds for all $\lambda > 0$, so we can minimize over λ.

Then we have

$$\mathbb{P}_{\beta} \left(\cdot \geq \sqrt{\frac{w_{max}^2(w_{max}^H - 1)^2}{n(w_{max} - 1)^2} \log \frac{4m}{\delta}} + \frac{1}{m} \right) \leq \delta$$

Combining this with Lemma 4.2 to bound the first term gives the result and choice of $m \propto \sqrt{n}$ gives the result.

C.3 Cramer-Rao Variance Lower Bound of Off-Policy CDF Estimation (Proof of Theorem 6.1)

We prove a Cramer-Rao lower bound for UDAGs, defined below.

Definition C.1 (Unique Directed Acyclic Graph). *An* MDP *is unique directed acyclic graph (UDAG) if the state and action spaces are finite. A state only occurs at a particular horizon h, that is, for any $s \in \mathcal{S}$, there exists a unique h such that $\max_{\pi} P(s_h = s | \pi) > 0$. In addition, each state has a unique reward function. For simplicity, assume $\gamma = 1$.*

Proof. First, we transform the MDP of Definition C.1 into one with an augmented state space, including the history of rewards encountered. That is, a trajectory is now defined as

$$\tau = (s_0, g_0, a_0, ... s_H, g_H, a_H, s_{H+1}, g_{H+1}),$$

where $g_h = \sum_{k=0}^{h-1} r_h$. Note that $P(g_{h+1}|s_h, g_h, a_h) = P(r_h|s_h, a_h)$ and $P(s_{h+1}|s_h, g_h, a_h) = P(s_{h+1}|s_h, a_h)$.

For brevity, denote $\widetilde{s} = (s, g)$ to be the augmented state space. We now derive a pointwise Constrained Cramer-Rao Bound (CCRB) for $F(t)$ in the augmented MDP,

$$KU(UIU)^{-1}U^\top K^\top$$

where K is the Jacobian of $F(t)$ and I is the Fisher information matrix. U is a matrix whose column vectors consist of an orthonormal basis for the null space of a matrix N, which is a block-diagonal matrix with $N_{(\widetilde{s},a),(\widetilde{s}',a')} = \mathbb{1}\{\widetilde{s}, a = \widetilde{s}', \widetilde{a}'\}$. Further, define $\eta_{\widetilde{s},a,\widetilde{s}',g'} = P(\widetilde{s}'|\widetilde{s},a)$, and note that $N\eta = \mathbf{1}$.

To compute the Fisher information matrix I,

$$I = \mathbb{E}\left[\left(\frac{\partial \log P_\beta(\tau)}{\partial \eta}\right)\left(\frac{\partial \log P_\beta(\tau)}{\partial \eta}\right)^\top\right]$$

where $P_\beta(\tau)$ is the probability of a trajectory τ under the behavioral policy β,

$$P_\beta(\tau) = \mu(s_1, g_1)\beta(a_1|s_1)P(s_2, |s_1, a_1)P(g_2|s_1, a_1, g_1)\ldots P(s_H, g_H|s_{H-1}, g_{H-1}a_{H-1})\beta(a_H|s_H)P(s_{H+1}, g_{H+1}|s_H, g_H, a_H)$$
$$= \mu(\widetilde{s}_1)\beta(a_1|\widetilde{s}_1)P(\widetilde{s}_2, |\widetilde{s}_1, a_1)P(\widetilde{s}_2|\widetilde{s}_1, a_1)\ldots P(\widetilde{s}_H|\widetilde{s}_{H-1}a_{H-1})\beta(a_H|\widetilde{s}_H)P(\widetilde{s}_{H+1}|\widetilde{s}_H, a_H)$$

Define a new indicator vector $g(\tau)$ with $g(\tau)_{\widetilde{s}_h,a_h,\widetilde{s}_{h+1}} = 1$ if $(\widetilde{s}_h, a_h, \widetilde{s}_{h+1}) \in \tau$. Then, letting $\circ$ denote an elementwise operation, we have

$$\frac{\partial \log P_\beta(\tau)}{\partial \eta} = \eta^{\circ - 1} \circ g(\tau)$$

which makes

$$I = \mathbb{E}\left[[1/\eta_i\eta_j]_{ij} \circ \left(g(\tau)g(\tau)^\top\right)\right] = [1/\eta_i\eta_j]_{ij} \circ \mathbb{E}\left[g(\tau)g(\tau)^\top\right]$$

As a result, the elements of I are as follows, where P_M refers to the marginal probability:

- diagonal: $\frac{P_M(\widetilde{s}_h, a_h)}{P(\widetilde{s}_{h+1}|\widetilde{s}_h, a_h)}$
- row $= (\widetilde{s}_h, a_h, \widetilde{s}_{h+1})$, column $= (\widetilde{s}'_h, a'_h, \widetilde{s}'_{h+1})$: $P_M(\widetilde{s}'_h, a'_h)P_M(\widetilde{s}_h, a_h|\widetilde{s}'_{h+1})$
- otherwise: 0

We know calculate the term $(U^\top IU)^{-1}$, and we use a similar strategy to Jiang and Li [2016], Huang and Jiang [2020] to avoid taking its inverse. Note that

$$U^\top IU = U^\top \left(N^\top X^\top + I + XN\right)U.$$

where X is arbitrary, N was defined previously. Let $D = N^\top X^\top + I + XN$. Our goal is to define X such that D is a diagonal matrix with diagonal identical to the diagonal of I. Note that XN and $N^\top X^\top$ are symmetric, so we can design the former to eliminate the upper triangle of I, and the latter to eliminate the lower triangle. To verify we can do this, set

- row $= (\widetilde{s}_h, a_h, \widetilde{s}_{h+1})$, column $= (\widetilde{s}_h, a_h, 2) : -P_M(\widetilde{s}_h, a_h)$
- row $= (\widetilde{s}_h, a_h, \widetilde{s}_{h+1})$, column $= (\widetilde{s}'_h, a'_h, \{1 \text{ or } 2\}) : -P_M(\widetilde{s}_h, a_h)P_M(\widetilde{s}'_h, a'_h, \widetilde{s}_{h+1})\mathbb{1}\{h < h'\}$
- otherwise: 0

Then with the proper choice of U,

$$U(U^\top IU)^{-1}U^\top = \mathrm{Diag}(\{B(\widetilde{s}_h, a_h)\}_0^H)$$

where $\mathrm{Diag}(\{\cdot\})$ is a block-diagonal matrix consisting of matrices in the set $\{\cdot\}$, and

$$B(\widetilde{s}_h, a_h) = \frac{\mathrm{Diag}(P(\cdot|\widetilde{s}_h, a_h)) - P(\cdot|\widetilde{s}_h, a_h)P(\cdot|\widetilde{s}_h, a_h)^\top}{P_M(\widetilde{s}_h, a_h)}$$

where $P(\cdot|\widetilde{s}_h, a_h)$ is the transition vector.

We now need to calculate the Jacobian K. Recall that the estimation objective is

$$F(t) = \sum_{\widetilde{s}_1} \mu(\widetilde{s}_1) \sum_{a_1} \pi(a_1|\widetilde{s}_1) \ldots \sum_{s_H} P(\widetilde{s}_H|\widetilde{s}_{H-1}, a_{H-1}) \mathbb{1}\left\{g_H \leq t\right\}$$

Then $K(\widetilde{s}_h, a_h, \widetilde{s}_{h+1})$ is

$$\frac{\partial F(t)}{\partial P(r_h, s_{h+1}|s_h, a_h)} = \frac{\partial F(t)}{\partial P(s_{h+1}, g_{h+1}|s_h, g_h, a_h)} = \mathbb{P}(s_h, g_h, a_h) P(r_h|s_h, a_h) P(s_{h+1}|s_h, a_h) F_{s_{h+1}}(t - g_h - r_h)$$

Denote by $K_{(s_h, a_h, \cdot)}$ the vector fragment of K whose index tuple starts with s_h, a_h. Then we have that

$$KU(UIU)^{-1}U^\top K^\top = \sum_{h=1}^{H+1} \sum_{\widetilde{s}_h, a_h} \left(K_{(\widetilde{s}_h, a_h, \cdot)}\right)^\top B(\widetilde{s}_h, a_h) K_{(\widetilde{s}_h, a_h, \cdot)} \tag{25}$$

We have

$$\left(K_{(s_h, g_h, a_h, \cdot)}\right)^\top B(s_h, g_h, a_h) K_{(s_h, g_h, a_h, \cdot)}$$

$$= \frac{\mathbb{P}(s_h, g_h, a_h)^2}{\mathbb{P}_\beta(s_h, g_h, a_h)} \left(\sum_{r_h} P(r_h|s_h, a_h) \sum_{s_{h+1}} P(s_{h+1}|s_h, a_h) F_{s_{h+1}}(t - g_h - r_h)^2 \right.$$

$$\left. - \left(\sum_{r_h} P(r_h|s_h, a_h) \sum_{s_{h+1}} P(s_{h+1}|s_h, a_h) F_{s_{h+1}}(t - g_h - r_h) \right)^2 \right)$$

$$= \frac{\mathbb{P}(s_h, g_h, a_h)^2}{\mathbb{P}_\beta(s_h, g_h, a_h)} \mathbb{V}_{r_h, s_{h+1}|s_h, a_h}\left[F_{s_{h+1}}(t - g_{h+1})\right]$$

Under Definition C.1, we have that $\mathbb{P}(g_h) = \mathbb{P}(\tau_h)$ so the lower bound is

$$\sum_{h=1}^{H} \mathbb{E}_{\mathbb{P}_\beta}\left[w_{1:h-1}^2 \mathbb{V}_{r_h, s_{h+1}|s_h, a_h}\left[F_{s_{h+1}}\left(t - \sum_{k=1}^{h} r_k\right)\right]\right].$$

$\square$

Comparison with DR Variance. When expanded over H horizons, the DR variance (6.1) in MDP of Definition C.1 is

$$\sum_{h=1}^{H} \mathbb{E}_{\mathbb{P}_\beta}\left[w_{1:h-1}^2 \mathbb{V}_h\left[w(A_h, S_h)\Delta_{S_h, A_h}\left(t - \sum_{k=1}^{h} R_k\right) \Big| S_h\right]\right] + \sum_{h=1}^{H} \mathbb{E}_{\mathbb{P}_\beta}\left[w_{1:h-1}^2 \mathbb{V}_h\left[F_{S_h}\left(t - \sum_{k=1}^{h} R_k\right)\right]\right]$$

where $\Delta_{s,a}(t) = \overline{F}_{s,a}(t) - F(t)$. Thus when $\overline{F} = F$, the first term of the above expression goes to 0 and the DR variance becomes

$$\sum_{h=1}^{H} \mathbb{E}_{\mathbb{P}_\beta}\left[w_{1:h-1}^2 \mathbb{V}_h\left[F_{S_h}\left(t - \sum_{k=1}^{h} R_k\right)\right]\right]$$

which attains the Cramer-Rao lower bound.

D Proofs for Risk Estimation (Section 7)

D.1 Risk Bias and Variance

Proof of Proposition 7.1. Note that the CDF $F \in \mathcal{L}_\infty(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ and risk functional $\rho : \mathcal{L}_\infty(\mathbb{R}, \mathbb{B}(\mathbb{R})) \to \mathbb{R}$. If the Fréchet derivative of ρ exists with respect to F in the $\mathcal{L}_\infty$ space, then using the Taylor expansion, we have,

$$
\begin{aligned}
\mathbb{E}\left[\rho(\widehat{F})\right] &= \mathbb{E}\left[\rho\left(F + (\widehat{F} - F)\right)\right] \\
&\approx \rho(F) + \mathbb{E}\left[\langle \nabla_F \rho, \widehat{h} \rangle + \frac{1}{2}\langle \widehat{h}\langle \nabla_F^2 \rho, \widehat{h}\rangle\rangle\right] \\
&= \rho(F) + \frac{1}{2}\mathbb{E}\left[\langle \widehat{h}\langle \nabla_F^2 \rho, \widehat{h}\rangle\rangle\right] \\
&= \rho(F) + \frac{\|\nabla_F^2 \rho\|}{2}\mathbb{E}\left[\langle \widehat{h}, \widehat{h}\rangle\right]
\end{aligned}
$$

where $\widehat{h} = \widehat{F} - F$, and the orders of Fréchet derivatives are maps to $\mathbb{R}$ with proper interproducts. Let $\otimes$ denote operator outer product, then we have,

$$
\begin{aligned}
\mathbb{V}\left[\rho(\widehat{F})\right] &= \mathbb{V}\left[\rho\left(F + (\widehat{F} - F)\right)\right] \\
&\approx \mathbb{V}\left[\nabla_F \rho \widehat{h} + \frac{1}{2}\langle \nabla_F^2 \rho, \widehat{h}\rangle\widehat{h}\right] \\
&\approx \mathbb{E}\left[\widehat{h}\nabla_F \rho \otimes \nabla_F \rho \ \widehat{h}\right]
\end{aligned}
$$

$\square$

Proof of Corollary 7.1. The bound is a direct application of the definition of Lipschitz risks (Definition 7.1) and the error bounds for off-policy CDF estimates (e.g. Lemma 4.2). $\square$

D.2 CVaR Risk Lower Bound (Proof of Theorem 7.1)

To prove the lower bound for CVaR estimation, we make an adaptation of the proof in B.3. From Le Cam's method, we have

$$
\inf_{\widehat{F}} \sup_{P_1, P_2 \in \mathcal{P}} \mathbb{E}_P\left[|\widehat{\rho} - \rho_P^\pi|\right] \geq \frac{1}{8}|\rho_1^\pi - \rho_2^\pi|e^{-n\mathrm{D}_{\mathrm{KL}}(P_1\|P_2)} \tag{26}
$$

For a distortion risk functional ρ,

$$
|\rho_1^\pi - \rho_2^\pi| = \left|\int_0^D g(1 - F_1^\pi(t)) - g(1 - F_2^\pi(t))dt\right|
$$

In the case of CVaR_α, with $\alpha \in [0, 1)$, we have $g(x) = \min\{\frac{x}{1-\alpha}, 1\}$ so

$$
|\mathrm{CVaR}_\alpha(F_1^\pi) - \mathrm{CVaR}_\alpha(F_2^\pi)| = \left|\int_0^D \min\left\{\frac{1 - F_1^\pi(t)}{1 - \alpha}, 1\right\} - \min\left\{\frac{1 - F_2^\pi(t)}{1 - \alpha}, 1\right\} dt\right|
$$

Note that, under this definition, the expected return is CVaR with level $\alpha = 0$.

Recall the Bernoulli problem instances we previously used to prove the lower bound for MABs. For the MAB P_2, we set the arm reward distribution to be $R_2(0|a) = c$ for all $a \in \mathcal{A}$. For the MAB P_1, we set the arm reward distribution to be $R_1(0|a) = c + \delta(a)$ for some $\delta(a) \in [0, 1 - c]$. Then

$$
|\mathrm{CVaR}_\alpha(F_1^\pi) - \mathrm{CVaR}_\alpha(F_2^\pi)| = \left|\int_0^1 \min\left\{\frac{1 - F_1^\pi(t)}{1 - \alpha}, 1\right\} - \min\left\{\frac{1 - F_2^\pi(t)}{1 - \alpha}, 1\right\} dt\right|
$$

$$= \left| \min\left\{ \frac{1 - F_1^\pi(t=0)}{1-\alpha}, 1 \right\} - \min\left\{ \frac{1 - F_2^\pi(t=0)}{1-\alpha}, 1 \right\} \right|$$

$$= \left| \min\left\{ \frac{1 - c - \sum_a \pi(a)\delta(a)}{1-\alpha}, 1 \right\} - \min\left\{ \frac{1-c}{1-\alpha}, 1 \right\} \right|$$

Then note that the difference above has three possibilities:

$$= \begin{cases} 0, & \text{if } 1 - \alpha \le 1 - c - \sum_a \pi(a)\delta(a) \\ \frac{|\alpha - c|}{1-\alpha}, & \text{if } 1 - c - \sum_a \pi(a)\delta(a) \le 1 - \alpha \le 1 - c \\ \frac{\sum_a \pi(a)\delta(a)}{1-\alpha}, & \text{if } 1 - c \le 1 - \alpha \end{cases}$$

By setting $c = \alpha$ and $\delta(a) \in [0, 1-\alpha]$, it can be seen that the first scenario (a vacuous lower bound) is impossible, and that the third scenario encompasses the second scenario. That is, if $\delta(a) = 0 \ \forall a$, then both the second and third cases become 0.

Then since

$$\mathrm{D}_{\mathrm{KL}}(P_1 \| P_2) = \sum_a \beta(a) \mathrm{D}_{\mathrm{KL}}(R_1(a) \| R_2(a)) \le 4 \sum_a \beta(a)\delta(a)^2,$$

Our optimization problem boils down to

$$\max_{\delta \in \mathbb{R}^{\mathcal{A}}} \frac{\sum_a \pi(a)\delta(a)}{1-\alpha} \quad \text{s.t.} \quad \sum_a \beta(a)\delta(a)^2 \le \frac{1}{8n}, \ 0 \le \delta \le 1 - \alpha$$

Thus we have the Lagrangian

$$L(\delta, \lambda, \nu) = -\frac{\sum_a \pi(a)\delta(a)}{1-\alpha} + \lambda \left(\sum_a \beta(a)\delta(a)^2 - \frac{1}{8n} \right) + \sum_a \nu(a)\left(\delta(a) - 1 + \alpha \right)$$

$$= -\frac{\lambda}{8n} + \sum_a \left\{ \lambda\beta(a)\delta^2(a) + \delta(a)\left(\nu(a) - \frac{\pi(a)}{1-\alpha} \right) - (1-\alpha)\nu(a) \right\}$$

The optimal solution is

$$\delta^*(a) = \frac{\frac{\pi(a)}{1-\alpha} - \nu(a)}{2\lambda\beta(a)}$$

When we plug this in, we have

$$-\frac{\sum_a \pi(a)\delta^*(a)}{1-\alpha} = \max_{\lambda \ge 0, \nu \ge 0} g(\lambda, \nu) = \max_{\nu \ge 0} \left\{ -\sqrt{\frac{1}{8n} \sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu(a) \right)^2}{\beta(a)}} - (1-\alpha)\sum_a \nu(a) \right\}$$

$$= -\min_{\nu \ge 0} \left\{ \sqrt{\frac{1}{8n} \sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu(a) \right)^2}{\beta(a)}} + (1-\alpha)\sum_a \nu(a) \right\} \tag{27}$$

Finally, we have from Lemma D.1 that

$$\min_{\nu \ge 0} \left\{ \sqrt{\frac{1}{8n} \sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu(a) \right)^2}{\beta(a)}} + (1-\alpha)\sum_a \nu(a) \right\} \asymp \frac{1}{1-\alpha} \left\{ \sum_{a \in S^*} \pi(a) + \sqrt{\frac{\sum_{a \notin S} \beta(a)w(a)^2}{n}} \right\}$$

Lemma D.1.

$$\min_{\nu \geq 0} \left\{ \sqrt{\frac{1}{8n} \sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu(a)\right)^2}{\beta(a)}} + (1-\alpha) \sum_a \nu(a) \right\} \asymp \sum_{a \in S^*} \pi(a) + \frac{1}{1-\alpha} \sqrt{\frac{\sum_{a \notin S} \beta(a) w(a)^2}{n}} \quad (28)$$

Proof. First, note that $0 \leq \nu^* \leq \pi(a)$ from the optimization problem (27). Further, WLOG we can assume that $\pi(a) > 0$ and $\beta(a) > 0$. This is because if $\pi(a) = 0$, then $w(a) = 0$, which contributes 0 to both sides of the equation. Similarly, if $\beta(a) = 0$, then $\nu^*(a) = \pi(a)$, which means $a \in S^*$, and $\pi(a)$ is contributed to both sides of the equation.

We separate the proof into three cases: $v^* = \frac{\pi}{1-\alpha}$, $v^* = 0$, $0 \neq v^* \neq \frac{\pi}{1-\alpha}$.

$\mathbf{v^* = \pi.}$ In this case, $S^* = [k]$, and the LHS of (28) is

$$(1-\alpha) \sum_a \frac{\pi(a)}{1-\alpha} = 1$$

while the RHS is

$$\sum_a \pi(a) = 1.$$

$\mathbf{v^* = 0.}$ In this case, $S^* = \{\}$, and the LHS of (28) is

$$\sqrt{\frac{1}{8n} \sum_a \frac{\frac{\pi(a)}{1-\alpha}^2}{\beta(a)}} = \frac{1}{1-\alpha} \sqrt{\frac{1}{8n} \sum_a \beta(a) w(a)^2},$$

while the RHS is

$$\frac{1}{1-\alpha} \sqrt{\frac{\sum_a \beta(a) w(a)^2}{n}}.$$

$\mathbf{0 \neq v^* \neq \pi.}$ First, we write the optimality conditions of (27):

$$\left(\frac{w(a)}{1-\alpha} - \frac{\nu^*(a)}{\beta(a)}\right)^2 = 8n(1-\alpha)^2 \sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu^*(a)\right)^2}{\beta(a)} \quad \text{for } a \in S^* \quad (29)$$

$$w(a)^2 \leq 8n(1-\alpha)^4 \sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu^*(a)\right)^2}{\beta(a)} \quad \text{for } a \notin S^* \quad (30)$$

Denote

$$T_1 := \sum_{a \in S^*} \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu^*(a)\right)^2}{\beta(a)}$$

$$T_2 := \sum_{a \notin S^*} \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu^*(a)\right)^2}{\beta(a)} = \frac{1}{(1-\alpha)^2} \sum_{a \notin S^*} \beta(a) w(a)^2$$

Next, we will show that $T_1 = \frac{1-\epsilon}{\epsilon} T_2$ for some $\epsilon \in (0,1)$ such that $\beta(S^*) = (1-\epsilon)/8(1-\alpha)^2 n$.

Summing (29) over $a \in S^*$, we have

$$T_1 = \sum_{a \in S^*} \beta(a) \left(\frac{w(a)}{1-\alpha} - \frac{\nu^*(a)}{\beta(a)}\right)^2$$

$$= 8n(1-\alpha)^2 \beta(S^*) \sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu^*(a)\right)^2}{\beta(a)}$$

$$= 8n(1-\alpha)^2 \beta(S^*)(T_1 + T_2)$$

which implies that $S^* \neq [k]$. Then since $w(a) > 0$ and $\nu^* \neq 0$, both $T_2 > 0$ and $\beta(S^*) > 0$. Then we have

$$T_1 > 8n(1-\alpha)^2 \beta(S^*) T_1$$

which implies that

$$\frac{1}{8n(1-\alpha)^2} > \beta(S^*) > 0.$$

This means that for some $\epsilon \in (0, 1)$,

$$\beta(S^*) = \frac{1-\epsilon}{8n(1-\alpha)^2}.$$

Then combining this with the previous relationship,

$$T_1 = 8n(1-\alpha)^2 \frac{1-\epsilon}{8n(1-\alpha)^2}(T_1 + T_2)$$

$$= \frac{1-\epsilon}{\epsilon} T_2.$$

Next, we show that (27) has optimal value

$$\pi(S^*) + \sqrt{\frac{T_2}{8n}}\epsilon.$$

From (29), for $a \in S^*$ we have that

$$\nu^*(a) = \frac{\pi(a)}{1-\alpha} - \beta(a)\sqrt{8(1-\alpha)^2 n(T_1 + T_2)}.$$

Then

$$\sqrt{\frac{1}{8n} \sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu^*(a)\right)^2}{\beta(a)}} + (1-\alpha)\sum_a |\nu^*(a)| = \sqrt{\frac{1}{8n}(T_1 + T_2)}$$

$$+ (1-\alpha) \sum_{a \in S^*} \left(\frac{\pi(a)}{1-\alpha} - \beta(a)\sqrt{8(1-\alpha)^2 n(T_1 + T_2)}\right)$$

$$= \pi(S^*) + \left(1 - \beta(S^*)8n(1-\alpha)^2\right)\sqrt{\frac{1}{8n}(T_1 + T_2)}$$

$$= \pi(S^*) + \epsilon\sqrt{\frac{1}{8n}(T_1 + T_2)}$$

$$= \pi(S^*) + \sqrt{\frac{\epsilon}{8n}T_2}$$

Going back to our original problem, note that for $\epsilon \in [1/2, 1]$ we achieve the desired equivalence. Thus we focus on the case where $\epsilon \in (0, 1/2)$, which means that $\frac{1}{16(1-\alpha)^2 n} < \beta(S^*) < \frac{1}{8(1-\alpha)^2 n}$. Assume WLOG that the actions are ordered according to likelihood ratios $w(1) \leq \ldots \leq w(k)$. From (29) and (30), $S^* = \{t+1, \ldots, k\}$ for some $t \in [k]$. Then applying (29), we have that for action $a = t+1$,

$$\frac{w(t+1)^2}{(1-\alpha)^2} \geq \left(\frac{w(t+1)}{1-\alpha} - \frac{\nu^*(t+1)}{\beta(t+1)}\right)^2 = 8n(1-\alpha)^2(T_1 + T_2) = \frac{8n(1-\alpha)^2 T_2}{\epsilon}$$

Using the fact that $\frac{\pi(S^*)}{\beta(S^*)} \geq w(t+1)$, we have that

$$\frac{\pi(S^*)}{\beta(S^*)} \geq (1-\alpha)^2 \sqrt{\frac{8nT_2}{\epsilon}}$$

Then since $\beta(S^*) > \frac{1}{16(1-\alpha)^2 n}$,

$$\pi(S^*) \geq \sqrt{\frac{T_2}{8n\epsilon}} \geq \sqrt{\frac{T_2}{4n}}\epsilon$$

Then

$$\sqrt{\frac{1}{8n}\sum_a \frac{\left(\frac{\pi(a)}{1-\alpha} - \nu^*(a)\right)^2}{\beta(a)}} + (1-\alpha)\sum_a |\nu^*(a)| = \pi(S^*) + \sqrt{\frac{\epsilon}{8n}T_2} \asymp \pi(S^*) \asymp \pi(S^*) + \sqrt{\frac{T_2}{n}}$$

which achieves the target equivalence using the definition of T_2.

Finally, examining the optimality condition for $a \notin S^*$ (30) shows that $\nu^* = 0$ when

$$n \geq \frac{w_{max}^2}{8(1-\alpha)^2 \sum_a \beta(a)w^2(a)},$$

which means that the lower bound is

$$\frac{cD}{1-\alpha}\sqrt{\frac{\sum_a \beta(a)w(a)^2}{n}}$$

for some universal constant c, giving the theorem statement. □

E   Experiments

E.1   Implementation

We describe each environment and their modeling methods below. For both environments we use a maximum horizon $H = 200$ and $\gamma = 1$.

Cliffwalk.   Cliffwalk Sutton and Barto [2018] is a 4×12 tabular environment where the agent must travel from the start state (lower left corner) to the goal state (lower right corner). The agent can take one of four cardinal actions (left, down up, right). The bottom row of cells represents a cliff, and moving into the cliff incurs a cost of 100. Each other state incurs a cost of 1, thus incentivizing the agent to take the shortest path between start a goal. As the original Cliffwalk is deterministic, we introduce a random transition towards the cliff with probability $p = 0.05$ in each state.

The target policy π is learned via Q-learning, and the behavioral policy is a mixture between the target and a uniform policy, that is for a constant $\lambda \in [0, 1]$, $\beta = \lambda \pi + (1 - \lambda)\text{UNIF}$. For estimators (DM, DR, WDR) involving models $\overline{F}$, half of the dataset is used to construct the model and the other half is used for estimation. This process is then repeated with the halves switched, and the resulting estimate is an average of the two. First, an estimate of the MDP $\overline{M}$ is constructed using empirical averages for the transitions and rewards. Following Thomas et al. [2015], we assume that the model is given imperfect information of the horizon as $H = 201$. $\overline{F}$ is then computed recursively via the relations from Section 5:

$$\overline{F}_{s_h, a_h}^{H+1-h}(t) = \mathbb{E}_{\overline{P}, \overline{R}}\left[\overline{F}_{S_{h+1}}^{H-h}\left(\frac{t - R_h}{\gamma}\right)\right]$$
$$\overline{F}_{s_h}^{H+1-h}(t) = \mathbb{E}_{\pi}\left[\overline{F}_{s_h, A_h}^{H+1-h}(t)\right]$$

The Cliffwalk results shown in Figure 1 were averaged over 1000 repetitions.

Simglucose.   In Simglucose, the agent must control insulin bolus injections to a patient with type 1 diabetes. The state is a continuous vector with the patient's blood glucose levels and the carbohydrate intake from the last meal. We discretize the space of possible bolus injections into 6 actions. The agent receives a reward according to whether the patient's blood glucose levels are within acceptable limits or not. If the patient's blood glucose exceeds 180, a condition called hyperglycemia, the agent receives a reward of -1. If it falls under 70, called hypoglycemia, the agent receives a reward of -2. Otherwise, the agent receives a reward of +1.

The target policy is the built-in controller for the patient, and the behavioral policy is defined as the same way for Cliffwalk. As the state space is continuous, the model was built by first discretizing the state space, then using empirical averages to estimate the MDP $\overline{M}$ in the discretized space, upon which $\overline{F}$ is calculated.

The Simglucose results shown in Figure 1 were averaged over 100 repetitions.

E.2   Mean vs Plug-in + CDF

As an additional experiment, we compare the plug-in mean estimate on the estimated CDFs with direct mean estimation for existing IS, WIS and DR Jiang and Li [2016] estimates in the Cliffwalk environment. For WIS, at low sample sizes the direct mean estimates can outperform the CDF estimates, but reach the same error as n increases. For the IS and DR estimates, however, the CDF version of the estimates perform just as well, if not better, than direct mean estimation.
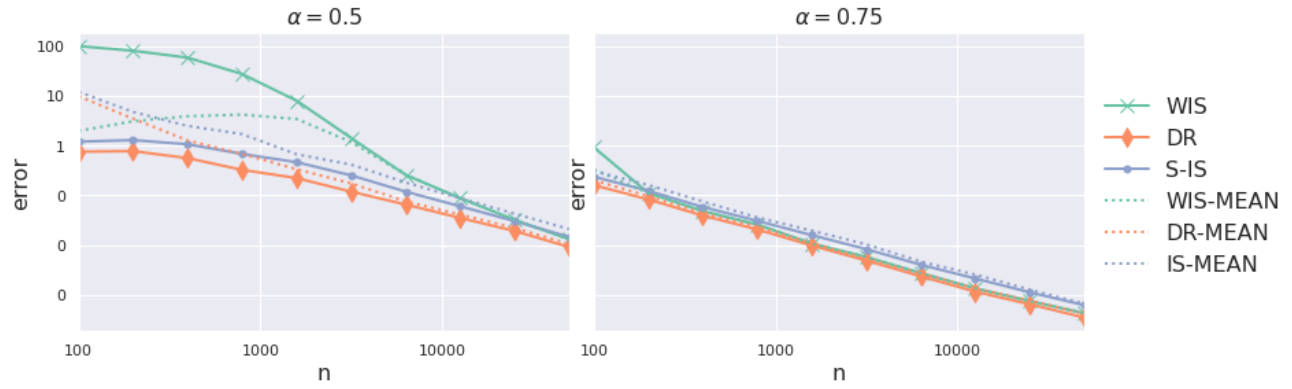
Figure 2: Normalized MSE for different λ in the Cliffwalk environment. Dashed lines show mean analogues of CDF estimates.