Off-Policy Risk Assessment in Contextual Bandits

Audrey Huang
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, IL 61801
audreyh5@illinois.edu

Liu Leqi
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
leqil@cs.cmu.edu

Zachary C. Lipton
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
zlipton@cmu.edu

Kamyar Azizzadenesheli
Department of Computer Science
Purdue University
West Lafayette, IN 47907
kamyar@purdue.edu

Abstract

Even when unable to run experiments, practitioners can evaluate prospective policies, using previously logged data. However, while the bandits literature has adopted a diverse set of objectives, most research on *off-policy evaluation* to date focuses on the expected reward. In this paper, we introduce Lipschitz risk functionals, a broad class of objectives that subsumes conditional value-at-risk (CVaR), variance, mean-variance, many distorted risks, and CPT risks, among others. We propose *Off-Policy Risk Assessment* (OPRA), a framework that first estimates a target policy's CDF and then generates plugin estimates for any collection of Lipschitz risks, providing finite sample guarantees that hold simultaneously over the entire class. We instantiate OPRA with both importance sampling and doubly robust estimators. Our primary theoretical contributions are (i) the first uniform concentration inequalities for both CDF estimators in contextual bandits and (ii) error bounds on our Lipschitz risk estimates, which all converge at a rate of $O(1/\sqrt{n})$.

1 Introduction

Many practical tasks, including medical treatment [50] and content recommendation [31] are commonly modeled within the contextual bandits framework. In the online setting, an agent observes a context at each step and chooses among the available actions. The agent then receives a context-dependent reward corresponding to the action taken, but cannot observe the rewards corresponding to alternative actions. In a healthcare setting, the observed context might be a vector capturing vital signs, lab tests, and other available data, while the action space might consist of the available treatments. The reward to optimize could be a measure of patient health or treatment response.

While contextual bandits research has traditionally focused on the expected reward, stakeholders often care about other risk functionals (parameters of the reward distribution) that express real-world desiderata or have desirable statistical properties. For example, investors assess mutual funds via the Sharpe ratio, which normalizes returns by their variance [44]. Related works in reinforcement learning (RL) have sought to estimate the variance of returns [41, 48] and to optimize the mean return under variance constraints [33]. In safety-critical and financial applications, researchers often measure the conditional value-at-risk (CVaR), which captures the expected return among the lower α quantile of outcomes [40, 28]. In an emerging line of RL works, researchers have explored other risk functionals, including cumulative prospect weighting [22], distortion risk measures [13], and exponential utility functions [16].

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

In many real-world problems otherwise suited to the contextual bandits framework, experimentation turns out to be prohibitively expensive or unethical. In such settings, we might hope to evaluate prospective policies using the data collected under a previous policy. Formally, this problem is called *off-policy evaluation*, and our goal is to evaluate the performance of a target policy π using data collected under a behavior policy β. While most existing research focuses on estimating the expected value of the returns [20, 19], one recent paper evaluates the variance of returns [10].

In this paper, we propose practical methods and the first sample complexity guarantees for *off-policy risk evaluation*, addressing a diverse set of objectives of interest to researchers and practitioners. Towards this end, we introduce *Lipschitz risk functionals* which encompass all objectives for which the risk (i) depends only on the CDF of rewards; and (ii) is Lipschitz with respect to changes in the CDF (as assessed via the sup norm). We prove that for bounded rewards, this class subsumes many risk functionals of practical interest, including variance, mean-variance, conditional value-at-risk, and cumulative prospect weighting, among others.

Thus, given accurate estimates of the CDF of rewards under π, we can accurately estimate Lipschitz risks. Moreover, (sup norm) bounds on our CDF estimates imply bounds on the corresponding plugin estimates for any Lipschitz risks. The key remaining step is to establish finite sample guarantees on the error in estimating the target policy's CDF of rewards. Our analysis centers on an importance sampling estimator (Section 5.1), and a variance-reduced doubly robust estimator (Section 5.3). We derive finite sample concentrations for both CDF estimators, showing that they achieve the desired $O(1/\sqrt{n})$ rates, where n is the sample size. Moreover, the estimation error for any Lipschitz risk is scales with its Lipschitz constant, and similarly converges as $O(1/\sqrt{n})$.

We assemble these results into an algorithm called OPRA (Algorithm 1) that outputs a comprehensive risk assessment for a target policy π, using any set of Lipschitz risk functionals. Notably, because all risk estimates share the same underlying CDF estimate, our error guarantees hold simultaneously for all estimated risk functionals in the set, regardless of the cardinality (Section 6). Finally, we present experiments that demonstrate the practical applicability our estimators.

2 Related Work

The study of risk functionals and risk-aware algorithms is core to the decision making literature [3, 40, 29, 43, 1, 37, 25]. In the bandit literature, many works address regret minimization problems using risk functionals; popular examples include the CVaR, value-at-risk, and mean-variance [8, 41, 54, 60]. [49] studies optimistic UCB exploration for optimizing CVaR while [11, 6] study Thompson sampling, and [27, 7] study regret minimization for linear combinations of the mean and CVaR. Using the CPT risk functional, [22] considers regret minimization in both K-armed bandits and linear contextual bandits. [53, 35] tackle the problem of black-box function optimization under different risk functionals.

In off-policy evaluation, we face an additional challenge due to the discrepancy between the data distribution and that induced by the target policy. Importance sampling (IS) estimators are among the most prominent methods for dealing with distribution shift [2, 23, 45]. Doubly robust (DR) estimators [39, 5] leverage (possibly misspecified) models to achieve lower variance without sacrificing consistency. These estimators have been adapted for off-policy evaluation in multi-armed bandits [32, 52, 10], contextual bandits [20, 19, 57], and Markov decision processes [24, 51]

After deriving our key results, we learned of an independent (also unpublished) work [9] that also employs importance sampling to estimate CDFs for the purpose of providing off-policy estimates for parameters of the reward distribution. However, they do not establish uniform concentration of their CDF estimates or formally relate the errors in CDF and (downstream) parameter estimation leaving open questions concerning the convergence (both asymptotically and in finite samples) of the parameter estimates.

Our work formulates both importance sampling and variance-reduced doubly robust estimators and provides the first uniform finite sample concentration bounds for both types of CDF and risk estimates. For empirical CDF estimation, we build on seminal work by [21], which provided an approximation-theoretic concentration bound that was later tightened by [34].

3 Problem Setting

We denote contexts by X and the corresponding context space by $\mathcal{X}$. Similarly, we denote actions by A and the corresponding action space by $\mathcal{A}$. We study the contextual bandit problem characterized by a fixed probability measure over context space $\mathcal{X}$, and a conditional reward distribution $\mathcal{R}(\cdot|X, A)$. In the off-policy setting, we have access to a dataset $\mathcal{D}$ generated using a behavior policy β that interacts with the environment for n rounds as follows: at each round, a new context X is drawn and then the policy β chooses an action $A \sim \beta(\cdot|X)$. The environment then reveals the reward $R \sim \mathcal{R}(\cdot|X, A)$ for only the chosen action A. Running this process for n stepsgenerates a dataset $\mathcal{D} := \{x_i, a_i, r_i\}_{i=1}^n$. In the off-policy evaluation setting, our goal is to evaluate the performance of a target policy π, using only a dataset $\mathcal{D}$.

Next, we can express our sample space in terms of the contexts, actions, and rewards: $\Omega = (\mathcal{X} \times \mathcal{A} \times \mathbb{R})$. Let $(\Omega, \mathcal{F}, \mathbb{P}_\beta)$ be the probability space induced by the behavior policy β, and $(\Omega, \mathcal{F}, \mathbb{P})$ the probability space induced by the target policy π. We assume that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{P}_\beta$. For any context x and action a, the importance weight expresses the ratio between the two densities $w(\omega) = w(a, x) = \frac{\beta(a|x)}{\pi(a|x)}$, and the maximum weight $w_{\max} = \sup_{a,x} w(a, x)$ is simply the supremum taken over all contexts and actions. Further, let $w_2 = \mathbb{E}_{\mathbb{P}_\beta}\left[w(A, X)^2\right]$ denote the exponential of the second order Rényi divergence. Note that by definition, $w_2 \le w_{\max}$, and in practice, we often have $w_2 \ll w_{\max}$.

Finally, we introduce some notation for describing CDFs: For any $t \in \mathbb{R}$, let $F(t) = \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{\{R \le t\}}]$ denote the CDF under the target policy; further, let $G(t; X, A) = \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{\{R \le t\}}|X, A] = \mathbb{E}_{\mathbb{P}_\beta}[\mathbb{1}_{\{R \le t\}}|X, A]$ denote the CDF of rewards conditioned on a context X and action A, which is independent of the policy. Lastly, for any $t \in \mathbb{R}$, we denote the variance by $\sigma^2(t; X, A) = \mathbb{V}_{\mathbb{P}}\left[\mathbb{1}_{\{R \le t\}}|X, A\right] = \mathbb{V}_{\mathbb{P}_\beta}\left[\mathbb{1}_{\{R \le t\}}|X, A\right]$.

4 Lipschitz Risk Functionals

We now introduce Lipschitz risk functionals, a novel class of objectives for which absolute differences in the risk are bounded by sup norm differences in the CDF of rewards. After formally defining the class, we provide an in-depth review of common risk functionals and their relationship to the CDF of rewards. When possible, we derive the associated Lipschitz constants, when rewards are bounded on support $[0, D]$, relegating all proofs to Appendix A.2.

4.1 Defining the Lipschitz Risk Functionals

The Lipschitz risk functionals are a subset of the broader family of *law-invariant risk functionals*. Formally, let $Z \in \mathcal{L}_\infty(\Omega, \mathcal{F}_Z, \mathbb{P}_Z)$ denote a real-valued random variable that admits a CDF $F_Z \in \mathcal{L}_\infty(\mathbb{R}, \mathbb{B}(\mathbb{R}))$. A *risk functional* ρ is a mapping from a space of random variables to the space of real numbers $\rho : \mathcal{L}_\infty(\Omega, \mathcal{F}_Z, \mathbb{P}_Z) \to \mathbb{R}$. Any risk functional ρ is said to be law-invariant if $\rho(Z)$ depends only on the distribution of Z [30].

Definition 4.1 (Law-Invariant Risk Functional)**.** *A risk functional $\rho : \mathcal{L}_\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, is law-invariant 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 sometimes abuse notation by writing $\rho(F_Z)$ in place of $\rho(Z)$. In general, it may not be practical to estimate risk functionals that are not law invariant from data [4]. Thus focusing on law-invariant risks is only mildly restrictive.

We can now formally define the Lipschitz risk functionals:

Definition 4.2 (Lipschitz Risk Functional)**.** *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*

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

A risk functional is L-Lipschitz if, for any two random variables Z, Z', its value is upper bounded by the sup-norm of the difference between their corresponding CDFs. The significance of this Lipschitzness property in the contextual bandit setting is that, given a high confidence bound on the error of the estimated CDF of rewards for a policy π, we can obtain a high confidence bound on its evaluation under any L-Lipschitz law-invariant risk functional on the distribution of rewards.

4.2 Overview of Common Risk Functionals (and their Lipschitzness)

We now briefly describe some popular classes of risk functionals and, when possible, derive their associated Lipschitz constants.

Coherent Risk Functionals The set of risk functionals that satisfy properties called monotonicity, subadditivity, translation invariance, and positive homogeneity (see Appendix A.2), constitute the *coherent risk functionals* [3, 15]. While not all coherent risk functionals are law-invariant, nearly among those commonly addressed in the literature are. Examples include expected value, conditional value-at-risk (CVaR), entropic value-at-risk, and mean semideviation [11, 49, 47, 43]. Others include the Wang transform function [55] and the proportional hazard (PH) risk functional [58].

Distorted Risk Functionals When the random variable Z is required to be non-negative, law-invariant coherent risk functionals are examples of the more general class of law-invariant *distorted risk functionals* [17, 55, 56, 4]. For $Z \geq 0$, a distorted risk functional has the following form

$$\rho(F_Z) = \int_0^{\infty} g(1 - F_Z(t))dt,$$

where the distortion function $g : [0,1] \to [0,1]$ is an increasing function with $g(0) = 0$ and $g(1) = 1$. Distorted risk functionals are coherent if and only if g is concave [58]. For example, when $g(s) = \min\{\frac{s}{1-\alpha}, 1\}$ for $s \in [0,1]$ and $\alpha \in (0,1)$, CVaR at level α is recovered. When g is the identity map, the distorted risk functional is the expected value. The Wang risk functional at level α [55] is recovered when $g(s) = F(F^{-1}(s) - F^{-1}(\alpha))$, and the proportional hazard risk functional can by obtained by setting $g(s) = s^{\alpha}$ for $\alpha < 1$. 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.

Lemma 4.1 (Lipschitzness of Coherent and Distorted Risk Functionals)**.** *On the space of random variables with support in $[0, D]$, the distorted risk functional of any $\frac{L}{D}$-Lipschitz distortion function $g : [0,1] \to [0,1]$, i.e., $|g(t) - g(t')| \leq \frac{L}{D}|t - t'|$, is a L-Lipschitz risk functional.*

Remark 4.1 (Expected Value and CVaR)**.** *Both expected value and CVaR are examples of distorted risk functionals. Then using Lemma 4.1, on the space of random variables with support in $[0, D]$, the expected value risk functional is D-Lipschitz because g is the identity and thus 1-Lipschitz. On the same space, the risk functional CVaR$_{\alpha}$ is $\frac{D}{\alpha}$-Lipschitz because g is $\frac{1}{\alpha}$-Lipschitz.*

Cumulative Prospect Theory (CPT) Risk Functionals CPT risks [37] take the form:

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

where $g^{+}, g^{-}\colon [0,1] \to [0,1]$ are continuous functions 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 $c \in \mathbb{R}$. Importantly, the CPT functional handles gains and losses separately. The functions u^{+}, u^{-} compare the random variable Z to a baseline c, and the distortion g^{+} is applied to "gains" (when $Z \geq c$), while g^{-} is applied to "losses" (when $Z < c$).

Lemma 4.2 (Lipschitzness of CPT Functional)**.** *On the space of random variables with support in $[0, D]$, if the CPT distortion functions g^{+} and g^{-} are both $\frac{L}{D}$-Lipschitz, then the CPT risk functionals is L-Lipschitz.*

Other Risk Functionals The variance, mean-variance, and many other popular risks do not fit easily into the aforementioned classes, but are nevertheless law-invariant. For example, for a nonnegative random variable Z, the variance is defined as $\rho(F_Z) = 2\int_0^{\infty} t(1 - F_Z(t))dt - \left(\int_0^{\infty}(1 - F_Z(t))dt\right)^2$. Moreover, the variance and mean-variance are both L-Lipschitz.

Lemma 4.3 (Lipschitzness of Variance)**.** *On the space of random variables with support in $[0, D]$, variance is a $3D^2$-Lipschitz risk functional.*

A number of recent papers have addressed risk functionals expressed as weighted combinations of others, e.g., mean-variance [41]. Other papers have optimized constrained objectives, such as expected reward constrained by variance or CVaR below a certain threshold [12, 38]. When expressed as Lagrangians, these objectives can also be expressed as weighted combinations of the risk functionals involved. We extend the Lipschitzness property to risk functionals of this form:

Lemma 4.4 (Lipschitzness of Weighted Sum of Risk Functionals)**.** *Let ρ be a weighted sum of risk functionals $\rho_1, ..., \rho_K$ that are $L_1, ..., L_K$-Lipschitz, respectively, with weights $\lambda_1, ..., \lambda_K > 0$, i.e., $\rho(Z) = \sum_{k=1}^{K} \lambda_k \rho_k(Z)$. Then ρ is $\sum_k \lambda_k L_k$-Lipschitz.*

Remark 4.2. *Note that mean-variance is given by $\rho(Z) = \mathbb{E}[Z] + \lambda \mathbb{V}(Z)$ for some $\lambda > 0$. Then, using Lemma 4.4, we immediately obtain that mean-variance is $(1 + 3\lambda D^2)$-Lipschitz for bounded random variables.*

Though we have provided many examples of Lipschitz risk functionals in this section, it is worth noting that there are a number of risk functionals that do not satisfy the Lipschitzness property, such as the value-at-risk (VaR). For the sake of brevity, we omit consideration of such risk functionals in this paper, and outline future avenues of research on this topic in the discussion.

5 Off-Policy CDF Estimation

This section describes our method for high-confidence off-policy estimation of F, the CDF of returns under the policy π. The key challenge in estimating F is that the reward samples are observed only for actions taken by the behavior policy β. To overcome this limitation, one intuitive solution is to reweight the observed samples according to their importance sampling (IS) weight (Section 5.1). However, IS estimators are known to suffer from high variance. To mitigate this, we define the first doubly robust CDF estimator (Section 5.3).

5.1 CDF Estimation with Importance Sampling (IS)

Given an off-policy dataset $\mathcal{D} = \{x_i, a_i, r_i\}_{i=1}^{n}$, we define the following nonparametric IS-based estimator for the empirical CDF,

$$\widehat{F}_{\text{IS}}(t) := \frac{1}{n} \sum_{i=1}^{n} w\left(a_i, x_i\right) \mathbb{1}_{\{r_i \leq t\}}, \tag{1}$$

where $w(a, x) = \frac{\pi(a|x)}{\beta(a|x)}$ are the importance weights. The IS estimator is pointwise-unbiased, with variance given below (proof in Appendix B.1):

Lemma 5.1. *The IS estimator* (1) *is unbiased and its variance is*

$$\mathbb{V}_{\mathbb{E}_\beta}\left[\widehat{F}_{IS}(t)\right] = \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[w(A, X)^2\sigma^2(t; X, A)\right] + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[\mathbb{E}_{\mathbb{P}_\beta}\left[w(A, X)G(t; X, A)|X\right]\right]$$
$$+ \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[\mathbb{V}_{\mathbb{P}_\beta}\left[w(A, X)G(t; X, A)|X\right]\right].$$

The expression for variance is broken down into three terms. The first term represents randomness in the rewards. The second term represents variance due to the randomness over contexts X. The final term is the penalty arising from using importance sampling, and is proportional to the importance sampling weights w and the true CDF of conditional rewards G. The variance contributed by the third term can be large when the weights w have a wide range, which occurs when β assigns extremely small probabilities to actions where π assigns high probability.

Due to the use of importance sampling weights, the estimated CDF $\widehat{F}_{\text{IS}}(t)$ may be greater than 1 for some t, even though a valid CDF must be in the interval $[0, 1]$ for all t. To mitigate this problem, a weighted importance sampling (WIS) estimator can be used, which normalizes each importance weight by the sum of importance weights:

$$\widehat{F}_{\text{WIS}}(t) = \frac{1}{\sum_{j=1}^{n} w(a_j, x_j)} \sum_{i=1}^{n} w(a_i, x_i) \mathbb{1}_{\{r_i \leq t\}},$$

which [9] shows is a biased but uniformly consistent estimator. Another option is the clipped estimator IS-Clip (2), which simply limits the estimator to the unit interval:

$$\widehat{F}_{\text{IS-CLIP}}(t) := \min\{\widehat{F}_{\text{IS}}(t), 1\}. \tag{2}$$

Although $\widehat{F}_{\text{IS-CLIP}}$ has lower variance than the IS estimator, it is potentially biased. However, given finite samples, we can bound with high confidence the sup-norm error between $\widehat{F}_{\text{IS-CLIP}}$ and F, in Theorem 5.1 below (proof in Appendix B.2):

Theorem 5.1. *Given n samples drawn from $\mathbb{P}_\beta$, for the* IS *estimator $\widehat{F}_{\text{IS}}$, we have*

$$\mathbb{P}_\beta\left(\|\widehat{F}_{\text{IS}} - F\|_\infty \leq \varepsilon_{\text{IS}_1} := \sqrt{\frac{8w_{\max}^2}{n}\log(4/\delta)}\right) \geq 1-\delta. \tag{3}$$

or, based on w_2, we obtain a Bernstein-style bound,

$$\mathbb{P}_\beta\left(\left\|\widehat{F}_{\text{IS}} - F\right\|_\infty \leq \varepsilon_{IS_2} := \frac{4w_{\max}\log(4/\delta)}{n} + 2\sqrt{\frac{2w_2\log(4/\delta)}{n}}\right) \geq 1-\delta. \tag{4}$$

The same bounds hold for $\|\widehat{F}_{\text{IS-CLIP}} - F\|_{\mathcal{L}_\infty}$.

When $w_2 \ll w_{max}$, we observe that inequality (4) is more favorable than inequality (3). Theorem 5.1 demonstrates that the $\widehat{F}_{\text{IS-CLIP}}$ uniformly converges to the true CDF at a rate of $O(1/\sqrt{n})$, with the uniform consistency of $\widehat{F}_{\text{IS-CLIP}}$ as an immediate consequence. To the best of our knowledge, it is the first DKW-style concentration inequality on the importance sampling estimator CDF estimator in off-policy evaluation. The bound has explicit constants and subsumes the classical DKW inequality.

5.2 Model-Based CDF Estimation

As we have shown previously, IS estimators can suffer from high variance, which can be limiting in practice. However, in many practical applications, we may have access to a model $\overline{G}(t; X, A)$ of the conditional distribution $G(t; X, A)$, which can be used in estimation with very low variance. In many cases, practitioners may have a model of $\overline{G}$ from expert studies or from a simulator, or can form a regression estimate of $\overline{G}$ from logged data. One simple model-based estimator can then be obtained using the *direct method*, which simply employs the model $\overline{G}$ for each observed context:

$$\widehat{F}_{\text{DM}}(t) = \frac{1}{n}\sum_{i=1}^{n}\overline{G}(t; x_i, \pi), \quad \text{where } \overline{G}(t; x_i, \pi) = \sum_a \pi(a|x_i)\overline{G}(t; x_i, a). \tag{5}$$

Because the DM estimator $\widehat{F}_{\text{DM}}$ does not use importance weights, it can have significantly lower variance than the IS and DR estimators. In general, however, the DI estimator is biased, and its error flows directly from error in the model $\overline{G}$ (full derivations of the bias and variance are given in Lemma E.3 of Appendix E). The magnitude and distribution of bias over the context and action space is difficult to characterize. In practice, $\overline{G}$ is often estimated or modeled agnostic to the target policy, and hence may not be well-approximated in areas that are important for π. If $\overline{G}$ is an accurate model of the conditional reward distribution, however, then $\widehat{F}_{\text{DM}}$ is a good approximation of F.

5.3 Doubly Robust (DR) CDF Estimation

We now define a doubly robust (DR) CDF estimator that takes advantage of both importance sampling and models $\overline{G}$ to obtain the best characteristics of both types of estimation. In particular, the DR estimator is unbiased, but has potentially significant reduction in variance. The DR estimator for the empirical CDF is defined to be

$$\widehat{F}_{\text{DR}}(t) := \frac{1}{n}\sum_{i=1}^{n} w(a_i, x_i)\Big(\mathbb{1}_{\{r_i \leq t\}} - \overline{G}(t; x_i, a_i)\Big) + \overline{G}(t; x_i, \pi), \tag{6}$$

where $\overline{G}(t; x, \pi) = \mathbb{E}_{\mathbb{P}_\beta}\left[\overline{G}(t; x, A)|x\right]$. Informally, the DR estimator takes the model $\overline{G}$ as a baseline, using the available data to apply a correction. While $\overline{G}$ alone may be biased, the DR estimator is an unbiased estimator of F, and can have reduced variance compared to the IS estimator (proof in Appendix C.1):

Lemma 5.2. *The DR estimator* (6) *is unbiased and its variance is*

$$\begin{aligned}\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{F}_{\text{DR}}(t)\right] &= \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[w(A,X)^2\sigma^2(t; X, A)\right] + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[\mathbb{E}_{\mathbb{P}_\beta}\left[w(A,X)G(t; X, a)|X\right]\right] \\ &\quad + \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[\mathbb{V}_{\mathbb{P}_\beta}\left[w(A,X)\left(G(t; X, A) - \overline{G}(t; X, A)\right)|X\right]\right].\end{aligned}$$

The variance reduction advantage of the DR estimator becomes apparent from a direct comparison of the three terms in the IS estimator variance (Lemma 5.1) and the DR estimator variance (Lemma 5.2). The first and second terms, which capture the variance in rewards and contexts, are identical. The third term, which represents the importance sampling penalty, is proportional to $G - \overline{G}$ in the DR estimator, but proportional to G in the IS estimator. When this difference $G - \overline{G}$ is smaller than G, which is often the case in practice, the third term has reduced variance in the DR estimator. The magnitude of variance reduction is greater when the weights w have a large range, which is precisely when large variance can become problematic in importance sampling.

Remark 5.1 (Double Robustness). *Although we consider the setting where the behavior policy β is known, when the behavior policy is unknown and needs to be estimated, the estimator $\widehat{F}_{\mathrm{DR}}$ is consistent when either $\overline{G}$ is consistent or the policy estimator is consistent. This is where the name "doubly robust" comes from. We demonstrate and discuss this fact further in Appendix E.*

Although the DR estimator $\widehat{F}_{\mathrm{DR}}$ has desirable reductions in variance, given finite samples, it is not guaranteed to be a valid CDF. Like the IS estimator, the DR estimator may be greater than 1 for some t due to the use of importance weighting. However, it may also be negative at some t as a consequence of the subtracted term in (6). As an additional consequence of this term, the DR estimator is not guaranteed to be a monotone function. As a result, in order to use the DR CDF estimate for risk estimation, we must transform $\widehat{F}_{\mathrm{DR}}$ into a monotone function bounded in $[0, 1]$. Examples of such transformations include isotonic approximation [46] and monotone $\mathcal{L}_p$ approximation [14].

For our analysis, however, we consider a simple monotone transformation that involves an accumulation function, which does not allow the CDF to decrease, followed by a clipping to $[0, 1]$:

$$\widehat{F}_{\text{M-DR}}(t) = \text{Clip}\left\{\max_{t' \leq t} \widehat{F}_{\mathrm{DR}}(t'), 0, 1\right\}, \tag{7}$$

which is a uniformly consistent estimator, as the following concentration guarantee shows:

Theorem 5.2. *The monotone transformation of the DR estimator $\widehat{F}_{\text{M-DR}}(t)$ satisfies*

$$\mathbb{P}_\beta\left(\left\|\widehat{F}_{\text{M-DR}} - F\right\|_\infty \leq \epsilon_{\mathrm{DR}} := \sqrt{\frac{72 w_{\max}^2}{n} \log\left(\frac{8n^{1/2}}{\delta}\right)}\right) \geq 1 - \delta. \tag{8}$$

The proof is in Appendix C.2. The purpose of Theorem 5.2 is to show the dependence of the error on the importance weights w_{max} and on the finite sample size n. Using the M-DR estimator, we again recover a sample complexity of $\widetilde{O}\left(1/\sqrt{n}\right)$. The proof is given in Appendix C.2. Note that (8) does not depend on the error $\overline{G} - G$, which is the term responsible for variance reduction, as given in Lemma E.5. A tighter bound that incorporates the error $G - \overline{G}$ remains an open problem.

6 Off-Policy Risk Assessment

Given any law-invariant risk functional ρ and CDF estimator $\widehat{F}$, we can estimate the value of the risk functional as $\widehat{\rho} := \rho(\widehat{F})$. However, the estimator $\widehat{\rho}$ may be biased even if $\widehat{F}$ is unbiased. For Lipschitz risk functionals introduced in Section 4, we can obtain their finite sample error bounds, using the error bound of the CDF estimator. Further, a set of risk functionals of interest can be evaluated using the same estimated CDF, which suggests that the error bound of the CDF gives error bounds on the risk estimators that hold simultaneously.

Theorem 6.1 utilizes our error bound of the estimated CDF to derive error bounds for estimators of a set of Lipschitz risk functionals. As we showed in Section 4, most if not all commonly studied risk functionals satisfy the property of Lipschitzness, showing our result's wide applicability.

Theorem 6.1. *Given a set of Lipschitz risk functionals $\{\rho_p\}_{p=1}^P$ with Lipschitz constants $\{L_p\}_{p=1}^P$, and a CDF estimator $\widehat{F}$, such that $\|\widehat{F} - F\|_\infty \leq \epsilon$ with probability at least $1 - \delta$, we have with probability at least $1 - \delta$ that for all $p \in \{1, \ldots P\}$,*

$$\left|\rho_p(\widehat{F}) - \rho_p(F)\right| \leq L_p \epsilon.$$

Thus, one powerful property of risk estimation using the estimated CDF approach is that, given a high-probability error bound on the CDF estimator, the corresponding error bounds on estimates of *all* Lipschitz risk functionals of interest hold *simultaneously* with the same probability. Further, because the error of the IS CDF estimator ϵ_{IS} (Theorem 5.1) and DR CDF estimator ϵ_{DR} (Theorem 5.2) converge at a rate of $O(1/\sqrt{n})$, Theorem 6.1 shows that the error of all Lipschitz risk functional estimators shrink at a rate of $O(1/\sqrt{n})$. Thus, $\rho_p(\widehat{F})$ are consistent risk functional estimators.

Putting these results together, we now provide an algorithm, called OPRA (Algorithm 1), which given an off-policy contextual bandit dataset and a set of Lipschitz risk functionals of interest, outputs for each risk functional an estimate of its value and a confidence bound. The algorithm first uses a valid CDF estimator, e.g., the clipped IS estimator (2) or monotonized DR estimator (7), to form $\widehat{F}$ with sup-norm error ϵ. OPRA then evaluates each L_p-Lipschitz risk functional ρ_p on $\widehat{F}$ to obtain $\widehat{\rho}_p$, along with its upper and lower confidence bound $\widehat{\rho}_p \pm L_p\epsilon$.

Algorithm 1: Off-Policy Risk Assessment (OPRA)

Input: Dataset $\mathcal{D}$, policy π, probability δ, models $\overline{G}$, Lipschitz risk functionals $\{\rho_p\}_{p=1}^P$ with Lipschitz constants $\{L_p\}_{p=1}^P$.

Estimate the CDF using a valid CDF estimator $\widehat{F}$;

Compute the corresponding CDF estimation error ϵ such that $\mathbb{P}(\|F - \widehat{F}\|_\infty < \epsilon) \geq 1 - \delta$;

for *p = 1 . . . P* **do**

  Estimate $\widehat{\rho}_p = \rho_p(\widehat{F})$;

end

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

OPRA can be used to obtain a full risk assessment of any given policy, using the input Lipschitz risk functionals of interest, which can include the popularly used mean, variance, and CVaR. As demonstrated in Theorem 6.1, the error guarantee on the risk estimators holds simultaneously for all P risk functionals with probability at least $1 - \delta$. Importantly, OPRA also demonstrates the computational efficiency of the distribution-centric risk estimation approach proposed in this paper. For a given π, the CDF only needs to be estimated once, and can be used repetitively to estimate the value of the risk functionals. Further, the error of the risk estimators are determined by the known error of the CDF estimator, multiplied by the known Lipschitz constants.

7 Empirical Studies

In this section, we give empirical evidence for the effectiveness of the doubly robust (DR) CDF and risk estimates, in comparison to the importance sampling (IS), weighted importance sampling (WIS), and direct method (DM) estimates. Further, we demonstrate the convergence of the CDF and risk estimation error in terms of the number of samples. In our experiments, the relevant baselines are the IS and WIS CDF estimators with plugin risk estimates, both of which were contemporaneously proposed by [9], which are the only existing off-policy estimators for general risks. In most of our experiments, the DR CDF and risk estimates exhibit reduced variance compared to the baselines, a possibility suggested by our theoretical results.

7.1 UCI Datasets

Setup. Following [20, 19, 57], we obtain our off-policy contextual bandit datasets by transforming classification datasets. The contexts are the provided features, and the actions correspond to the possible class labels. To obtain the evaluation policy π, we use the output probabilities of a trained logistic regression classifier. The behavior policy is defined as $\beta = \alpha\pi + (1-\alpha)\pi_{\text{UNIF}}$, where π_{UNIF} is a uniform policy over the actions, for some $\alpha \in (0, 1]$. We apply this process to the PageBlocks and OptDigits datasets [18], which have dimensions d and actions k using $\alpha = 0.1$ (Figure 1). When models $\overline{G}$ are used (for DM, DR estimators), as in [20], the dataset is divided into two splits, with each of the two splits used to calculate $\overline{G}$ via regression, which is then used with the other split to calculate the estimator. The two results are averaged to produce the final estimators. We provide further details and extensive evaluations in Appendix F.

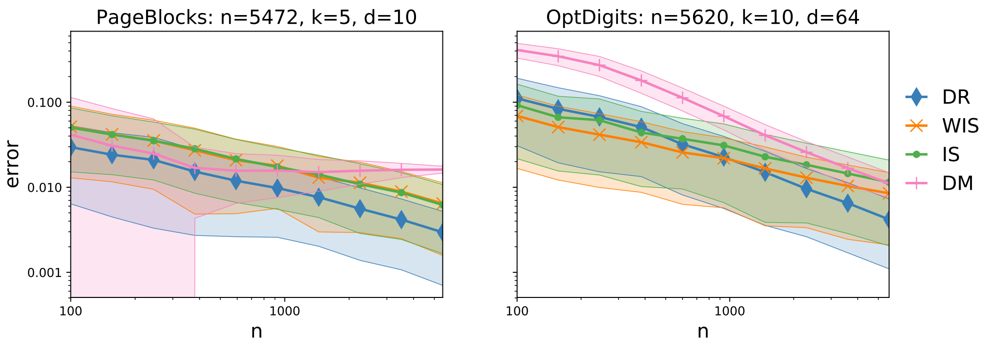


Figure 1: The error of the CDF estimators as a function of sample size n, for **(left)** the PageBlocks dataset and **(right)** the OptDigits dataset. Shaded area is the 95% quantile over 500 runs.

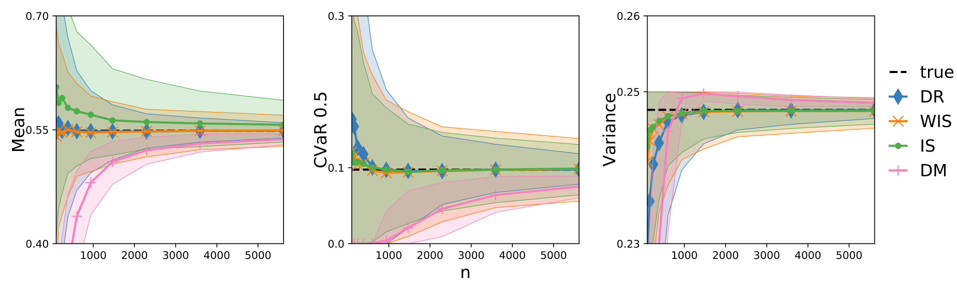


Figure 2: Estimated mean, $\mathrm{CVaR}_{0.5}$, and variance for the OptDigits dataset, compared to their true values (black). Shaded area is the standard deviation over 500 runs.

CDF Estimation. We evaluate the error $\|F - \widehat{F}\|_\infty$ of our CDF estimators against sample size for two UCI datasets (Figure 1). The IS and DR exhibit the expected $O\left(1/\sqrt{n}\right)$ rate of convergence in error previously derived in Theorems 5.1 and 5.2, respectively. We note that the WIS estimator, while biased, performs as well as the IS estimator if not better. In the PageBlocks dataset (Figure 1, left), the regression model for $\overline{G}$ is relatively well-specified as exemplified by the relatively low error of the DM estimator, though it has high variance for low samples sizes. The DR estimator leverages this model to outperform all other estimators for all sample sizes, without suffering the drawbacks of the DM estimator. It takes an order of magnitude less data to reach the same error compared to the IS and WIS estimators. In contrast, the regression model is less well-specified in the OptDigits dataset for lower sample sizes (Figure 1, right), and consequently, the DR estimator cannot perform as well as the IS and WIS estimators for small n. However, this trend reverses as data increases and the model improves, with the DR estimator outperforming the importance sampling estimators.

Risk Functional Estimation. Figure 2 shows the mean, variance, and $\mathrm{CVaR}_{0.5}$ estimates, which are obtained by evaluating each risk functional on the CDF estimators for the OptDigits dataset. Here, the estimates are plotted against the true value (dashed line) to make the variance reduction effect of the DR estimators more apparent. The DM estimator, which appeared to have competitive performance in the CDF error plot, has relatively high risk estimate error, which occurs because the DM CDF may be poorly approximated in areas that are important for risk functional estimation. The IS, WIS, and DR risk estimates converge quickly to the true value as n increases, and as expected, their relative behavior echoes the trends in Figure 1 as a consequence of our distributional approach. The DR estimator has slightly worse performance for small samples sizes due to the poor specification of the model, but soon exhibits the desired variance reduction for $n > 1000$.

Comparison to Other Risk Estimators. One natural question to ask is how OPRA risk estimates, which can be generated for many risk functionals simultaneously, compare to existing estimators for individual risk functionals. To the best of our knowledge, of the risk functionals described in this paper, previous work has only derived off-policy estimators for the mean and variance.

[20] establishes IS and DR estimators for the mean, and [10] proposes an IS estimator for the variance. We compare these risk estimates to OPRA in Table 1, using the Pageblocks dataset of Figure 1 at select sample sizes n. We note that while a number of enhancements have been proposed

for off-policy mean estimator, e.g., clipping importance weights [57], we compare only to analogous estimators. In general, the OPRA DR estimates outperform all other estimators. The OPRA DR estimate of the mean have the same error as the DR mean estimator from [20], and OPRA IS estimates of both mean and variance have lower MSE than existing IS mean and variance estimators.

	Mean				Variance			
	IS [20]	OPRA IS	DR [20]	OPRA DR	IS [10]	OPRA IS	DR	OPRA DR
$n = 100$	7.5e-3	4.4e-3	1.6e-3	**1.6e-3**	5.0e-5	0.7e-5	N/A	**0.6e-5**
$n = 924$	7.5e-4	4.7e-4	2.3e-4	**2.3e-4**	4.0e-6	0.7e-6	N/A	**0.7e-6**
$n = 5472$	12.5e-5	7.1e-5	3.4e-5	**3.4e-5**	5.4e-7	**0.9e-7**	N/A	1.1e-7

Table 1: Comparison of mean squared error (MSE) for OPRA versus existing off-policy estimators of the mean and variance on the Pageblocks dataset (averaged over 1000 repetitions).

7.2 Application: Diabetes Treatment

To demonstrate the efficacy of our method in more complex applications, we use OPRA to evaluate the risks of a target policy for diabetes treatment in the Simglucose simulator [59]. In Simglucose, the agent must control insulin bolus injections given to a type 1 diabetes patient, and the state is a continuous vector consisting of the patient's blood glucose levels and the carbohydrate intake from the last meal. 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 (indicating hyperglycemia), the agent receives a reward of -1. If it falls under 70 (indicating hypoglycemia), the agent receives a reward of -2. Otherwise, the agent receives a reward of $+1$.

Correspondingly, we define the target policy π to be the built-in controller for the patient, and estimate the CDF, mean reward, and $\text{CVaR}_{0.05}$ of π, with results shown in Figure 3. The DR estimator has lower error for CDF and risk estimation compared to all baseline estimators, reinforcing our findings on the UCI data.

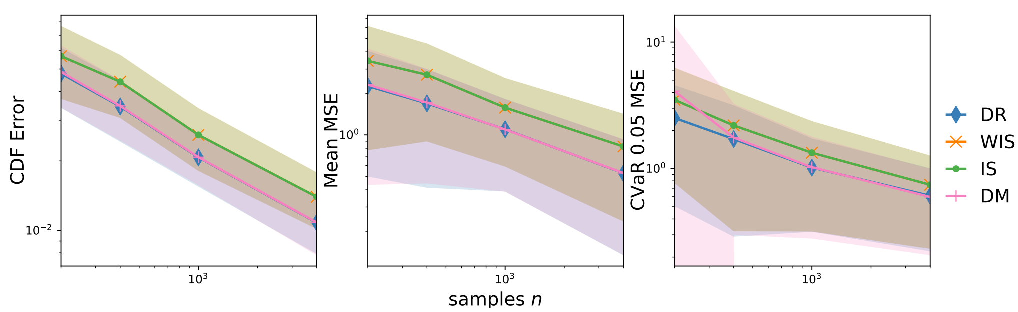


Figure 3: CDF and risk estimation error of OPRA in Simglucose (averaged over 100 repetitions).

8 Conclusion

This paper introduces a CDF-centric method for high confidence off-policy estimation of risk functionals, and the first doubly robust CDF estimator. From a theoretical point of view, we provide the first finite-sample concentration inequalities and confidence intervals for a variety of CDF and risk estimators, which are widely applicable in distributional RL settings [13, 28]. From a practical standpoint, our method can be used to comprehensively evaluate the behavior of a target policy before deployment under a wide range of risk functionals. Our work suggests several directions for future work. First, our uniform concentration bound for the doubly robust CDF estimator might be improve by incorporating its variance. Second, our heuristics for transforming CDF estimators into valid cumulative distribution functions (through a monotone transformation and clipping), could likely be improved. Finally, our work suggests a natural question: can concentration inequalities for risk functional estimators be derived if and only if they are Lipschitz, in some general sense?

References

[1] Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

[2] V. M. Aleksandrov, V. I. Sysoyev, and V. V. Shemeneva. Stochastic optimization. *Engineering Cybernetics*, 5(11-16):229–256, 1968.

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

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

[5] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[6] Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric-Ambryn Maillard. Thompson sampling for cvar bandits. *arXiv preprint arXiv:2012.05754*, 2020.

[7] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

[8] Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general framework for bandit problems beyond cumulative objectives.

[9] Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S. Thomas. Universal off-policy evaluation, 2021.

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

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

[12] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[13] 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.

[14] Richard B Darst and Robert Huotari. Best l1-approximation of bounded, approximately continuous functions on [0, 1] by nondecreasing functions. *Journal of approximation theory*, 43(2):178–189, 1985.

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

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

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

[18] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

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

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

[21] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[22] 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.

[23] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[24] 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.

[25] Cheng Jie, LA Prashanth, Michael Fu, Steve Marcus, and Csaba Szepesvári. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63(9):2867–2882, 2018.

[26] Elyès Jouini, Walter Schachermayer, and Nizar Touzi. Law invariant risk measures have the fatou property. In *Advances in mathematical economics*, pages 49–71. Springer, 2006.

[27] Anmol Kagrecha, Jayakrishnan Nair, and Krishna Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. *arXiv preprint arXiv:1906.00569*, 2019.

[28] 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.

[29] Pavlo A Krokhmal. Higher moment coherent risk measures. 2007.

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

[31] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

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

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

[34] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[35] Rémi Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. 2014.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[37] 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.

[38] A Prashanth L and Michael Fu. Risk-sensitive reinforcement learning: A constrained optimization viewpoint. *arXiv e-prints*, pages arXiv–1810, 2018.

[39] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

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

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

[42] Ekaterina N Sereda, Efim M Bronshtein, Svetozar T Rachev, Frank J Fabozzi, Wei Sun, and Stoyan V Stoyanov. Distortion risk measures in portfolio optimization. In *Handbook of portfolio construction*, pages 649–673. Springer, 2010.

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

[44] William F Sharpe. Mutual fund performance. *The Journal of business*, 39(1):119–138, 1966.

[45] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[46] PW Smith and JJ Swetits. Best approximation by monotone functions. *Journal of approximation theory*, 49(4):398–403, 1987.

[47] 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.

[48] 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.

[49] 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.

[50] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

[51] 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.

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

[53] Léonard Torossian, Aurélien Garivier, and Victor Picheny. $\mathcal{X}$ -armed bandits: Optimizing quantiles, cvar and other risks. In *Asian Conference on Machine Learning*, pages 252–267. PMLR, 2019.

[54] Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1330–1335. IEEE, 2015.

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

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

[57] 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.

[58] 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.

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

[60] Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.

Contents (Appendix)

A Appendix for Risk Functionals (Section 4)

A.1 Review of Risk Functionals

In Section 4.1 we introduced several classes of risk functionals and popular examples of each class. We now provide a formal axiomatic definition for classes of risk functionals, and begin by enumerating a set of prominent axioms used to describe risk functionals in the current literature [3, 42]. Classes of risk functionals are defined by the axioms that they satisfy, and one can define a class by choosing the subset best suited to the problem at hand.

Definition A.1 (Axioms of Risk Functionals)**.** *Consider a pair of random variables Z and Z' and a risk functional ρ. We have the following axioms:*

1. *Monotonicity: $\rho(Z) \leq \rho(Z')$ whenever $Z \leq Z'$.*
2. *Subadditivity: $\rho(Z + Z') \leq \rho(Z) + \rho(Z')$.*
3. *Additivity: $\rho(Z + Z') = \rho(Z) + \rho(Z')$ if Z and Z' are co-monotonic random variables (i.e., there exists a random variable Y and weakly increasing functions f, g such that $Z = f(Y)$ and $Z' = g(Y)$).*
4. *Translation invariance: $\rho(Z + c) = \rho(Z) + c, \forall c \in \mathbb{R}$.*
5. *Positive homogeneity: $\rho(tZ) = t\rho(Z)$ for $t > 0$.*
6. *Bounded above by the maximum cost, i.e., $\rho(Z) \leq \max(Z)$.*
7. *Bounded below by the mean cost, i.e., $\rho(Z) \geq \mathbb{E}[Z]$.*

Coherent Risk Functionals. The set of risk functionals that satisfy monotonicity (Axiom 1), subadditivity (Axiom 2), translation invariance (Axiom 4), and positive homogeneity (Axiom 5), constitute the class of *coherent risk functionals* [3, 15]. Further, if a law-invariant coherent risk functional additionally satisfies Additivity (Axiom 3), it is said to be a *spectral risk functional* [26, 1].

Distorted Risk Functionals. Distorted risk functionals have many desirable theoretical properties. They are translation invariant (Axiom 4) and positive homogeneous (Axiom 5), and are defined utilizing (Axiom 6) and (Axiom 7) [58]. They satisfy Axiom 7 if and only if $g(s) \geq s \; \forall s \in [0, 1]$ [58], and are subadditive (Axiom 2) if and only if g is concave, which preserves second order stochastic dominance [55]. In addition, all distorted risk functionals preserve stochastic first order dominance [58].

CPT-Inspired Risk Functionals. In general, due to the separate consideration of losses and gains, the CPT-inspired risk functional may not satisfy any of the above axioms. However, additional assumptions on the distortions g^+ and g^- may allow certain axioms to be satisfied. For example, if the random variable has nonnegative support and the threshold c is set to be 0 so that only gains are observed, and g^+ is additionally increasing, we recover the distorted risk functionals with axioms specified above. If g^+ is additionally concave, then we recover the coherent risk functionals.

A.2 Proofs for Risk Functionals

Proof of Lemma 4.1.

$$
\begin{aligned}
\left|\rho\left(F_Z\right) - \rho\left(F_{Z'}\right)\right| &= \left|\int_0^D g\left(1 - F_Z(t)\right) - g\left(1 - F_{Z'}(t)\right) dt\right| \\
&\leq \int_0^D \left|g\left(1 - F_Z(t)\right) - g\left(1 - F_{Z'}(t)\right)\right| dt \\
&\leq \int_0^D \frac{L}{D} \left|F_{Z'}(t) - F_Z(t)\right| dt \\
&\leq L \max_t \left|F_Z(t) - F_{Z'}(t)\right|,
\end{aligned}
$$

where the second to last step uses the L/D-Lipschitzness of ρ. $\square$

Proof of Lemma 4.2. Using the definition of the CDF, note that on the bounded support of $[0, D]$ the CPT functional can be rewritten as

$$\rho(F_Z) = \int_0^D g^+ \left(\mathbb{P}_Z \left(u^+(Z) > t\right)\right) dt - \int_0^D g^- \left(\mathbb{P}_Z \left(u^-(Z) > t\right)\right) dt.$$

Then,

$$\begin{aligned}
|\rho(Z) - \rho(Z')| &= \Big| \int_0^D g^+ \left(\mathbb{P}_Z \left(u^+(Z) > t\right)\right) dt - \int_0^D g^- \left(\mathbb{P}_Z \left(u^-(Z) > t\right)\right) dt \\
&\quad - \int_0^D g^+ \left(\mathbb{P}_{Z'} \left(u^+(Z') > t\right)\right) dt - \int_0^D g^- \left(\mathbb{P}_{Z'} \left(u^-(Z') > t\right)\right) dt \Big| \\
&\leq \left| \int_0^D g^+ \left(\mathbb{P}_Z \left(u^+(Z) > t\right)\right) dt - \int_0^D g^+ \left(\mathbb{P}_{Z'} \left(u^+(Z') > t\right)\right) dt \right| \\
&\quad + \left| \int_0^D g^- \left(\mathbb{P}_Z \left(u^-(Z) > t\right)\right) dt - \int_0^D g^- \left(\mathbb{P}_{Z'} \left(u^-(Z') > t\right)\right) dt \right| \\
&\leq \frac{L}{D} \int_0^D \left| \mathbb{P}_Z \left(u^+(Z) > t\right) - \mathbb{P}_{Z'} \left(u^+(Z') > t\right) \right| dt \\
&\quad + \frac{L}{D} \int_0^D \left| \mathbb{P}_Z \left(u^-(Z) > t\right) - \mathbb{P}_{Z'} \left(u^-(Z') > t\right) \right| dt \\
&\leq \frac{L}{D} \int_0^D \left| \mathbb{P}_Z (Z > t) - \mathbb{P}_{Z'} (Z' > t) \right| dt \\
&\quad + \frac{L}{D} \int_0^D \left| \mathbb{P}_Z (Z > t) - \mathbb{P}_{Z'} (Z' > t) \right| dt \\
&= 2\frac{L}{D} \int_0^D |F_{Z'}(t) - F_Z(t)| \, dt \\
&\leq 2L \max_t |F_Z(t) - F_{Z'}(t)|
\end{aligned}$$

□

Proof of Lemma 4.3. For the variance of any random variable Z with bounded support $[0, D]$, we have

$$\mathbb{V}(Z) = \mathbb{E}(Z^2) - \mathbb{E}(Z)^2.$$

Note that by the definition of expectation,

$$\mathbb{E}(Z^2) = \int_{t^2=0}^{D^2} 1 - F_{Z^2}(t^2) dt^2$$

Then using $dt^2 = 2tdt$ and the fact that $\mathbb{P}(Z^2 \geq t^2) = \mathbb{P}(Z \geq t)$ since t is nonnegative, with this change of variables we have

$$\mathbb{E}(Z^2) = 2 \int_{t=0}^{D} t \left(1 - F_Z(t)\right) dt.$$

This gives us the following expression for variance:

$$\mathbb{V}(Z) = 2 \int_0^D t(1 - F_Z(t)) dt - \left(\int_0^D (1 - F_Z(t)) dt \right)^2$$

Next, consider a pair of random variables Z and Z' with F_Z and $F_{Z'}$ as their CDF respectively. Therefore,

$$\begin{aligned}
|\mathbb{V}(Z) - \mathbb{V}(Z')| &\leq \left|2\int_0^D t(F_Z(t) - F_{Z'}(t))dt\right| + \left|\left(\int_0^D (1-F_Z(t))dt\right)^2 - \left(\int_0^D (1-F_{Z'}(t))dt\right)^2\right| \\
&\leq D^2\|F_Z(t) - F_{Z'}\|_\infty + \left|\int_0^D (F_Z(t) - F_{Z'}(t))dt\right| \left|\int_0^D (1-F_Z(t))dt + \int_0^D (1-F_{Z'}(t))dt\right| \\
&\leq D^2\|F_Z(t) - F_{Z'}\|_\infty + 2D\left|\int_0^D (F_Z(t) - F_{Z'}(t))dt\right| \\
&\leq D^2\|F_Z(t) - F_{Z'}\|_\infty + 2D^2\|F_Z(t) - F_{Z'}\|_\infty \\
&= 3D^2\|F_Z - F_{Z'}\|_\infty
\end{aligned}$$

□

Proof of Lemma 4.4. The proof of this lemma follows directly from the definition of Lipschitzness:

$$\begin{aligned}
\left|\sum_{k=1}^K \lambda_k\rho_k(Z) - \sum_{k=1}^K \lambda_k\rho_k(Z')\right| &\leq \sum_{k=1}^K \lambda_k \left|\rho_k(Z) - \rho_k(Z')\right| \\
&\leq \|F_Z - F_{Z'}\|_\infty \sum_{k=1}^K \lambda_k L_k.
\end{aligned}$$

□

B Proofs for Importance Sampling Estimators (Section 5.1)

B.1 Proof: Bias and Variance of IS CDF Estimate

Proof of Lemma 5.1. We take the expectation of the IS estimator (1) with respect to $\mathbb{P}_\beta$. Then for any $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\beta}[\widehat{F}_{\text{IS}}(t)] &= \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{1}{n}\sum_{i=1}^{n} w(A_i, X_i)\mathbb{1}_{\{R_i \le t\}}\right] \\
&= \mathbb{E}_{\mathbb{P}_\beta}\left[\mathbb{E}_{\mathbb{P}_\beta}\left[\frac{\pi(A|X)}{\beta(A|X)}\mathbb{E}_{\mathbb{P}_\beta}\left[\mathbb{1}_{\{R\le t\}}|X, A\right]\right]\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[w(A, X)\mathbb{1}_{\{R\le t\}}\right] \\
&= F(t).
\end{aligned}$$

Recall that $G(t; X, A) = \mathbb{E}[\mathbb{1}_{\{R\le t\}}|X, A\}]$. The variance of the IS estimator is derived using:

$$\begin{aligned}
\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{F}_{\text{IS}}(t)\right] &= \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[w(A, X)\mathbb{1}_{\{R\le t\}}\right] \\
&= \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[w(A, X)^2\mathbb{V}_{\mathbb{P}_\beta}\left[\mathbb{1}_{\{R\le t\}}|A, X\right]\right] + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[w(A, X)\mathbb{E}_{\mathbb{P}_\beta}\left[\mathbb{1}_{\{R\le t\}}|A, X\right]\right] \\
&= \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[w(A, X)^2\sigma^2(t; X, A)\right] + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[w(A, X)G(t; X, A)\right] \\
&= \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[w(A, X)^2\sigma^2(t; X, A)\right] + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[\mathbb{E}_{\mathbb{P}_\beta}\left[w(A, X)G(t; X, A)|X\right]\right] \\
&\quad + \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[\mathbb{V}_{\mathbb{P}_\beta}\left[w(A, X)G(t; X, A)|X\right]\right]
\end{aligned}$$

where the second equality uses the law of total variance conditioned on actions A and contexts X, and the third equality uses the definitions of σ^2 and G. The last equality is another application of the law of total variance conditioning on the context X.

□

B.2 Proof: Error Bound of IS CDF Estimate

Proof Theorem 5.1. Define the following function class:

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

Note that this is a countable set. Using this definition, we have

$$\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\text{IS}}(t) - F(t)\right| = \sup_{f\in\mathbb{F}(n)}\left|\left(\sum_{i}^{n}\left(w(A_i, X_i)f(R_i) - \mathbb{E}_{\mathbb{P}_\beta}[w(A_i, X_i)f(R_i)]\right)\right)\right|$$

Using this equality, for $\lambda > 0$, we have:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\text{IS}}(t) - F(t)\right|\right)\right] \\
&\quad = \mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda\sup_{f\in\mathbb{F}(n)}\left|\left(\sum_{i}^{n}\left(w(A_i, X_i)f(R_i) - \mathbb{E}_{\mathbb{P}_\beta}[w(A_i, X_i)f(R_i)]\right)\right)\right|\right)\right] \\
&\quad = \mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda\sup_{f\in\mathbb{F}(n)}\left|\left(\mathbb{E}_{\mathbb{P}_\beta}\left[\sum_{i}^{n}\left(w(A_i, X_i)f(R_i) - w(X_i', A_i')f(R_i')\right)\middle|\{X_i, A_i, R_i\}_i^n\right]\right)\right|\right)\right] \\
&\quad \le \mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda\sup_{f\in\mathbb{F}(n)}\left|\mathbb{E}_{\mathbb{P}_\beta}\left[\left(\sum_{i}^{n}\left(w(A_i, X_i)f(R_i) - w(X_i', A_i')f(R_i')\right)\middle|\{X_i, A_i, R_i\}_i^n\right]\right)\right|\right)\right] \\
&\quad \le \mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda\mathbb{E}_{\mathbb{P}_\beta}\left[\sup_{f\in\mathbb{F}(n)}\left|\left(\sum_{i}^{n}\left(w(A_i, X_i)f(R_i) - w(X_i', A_i')f(R_i')\right)\middle|\{X_i, A_i, R_i\}_i^n\right]\right)\right|\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\le \mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda \sup_{f\in\mathbb{F}(n)}\left|\left(\sum_i^n (w(A_i,X_i)f(R_i) - w(X_i',A_i')f(R_i'))\right)\right|\right)\right] \\
&= \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\lambda \sup_{f\in\mathbb{F}(n)}\left|\left(\sum_i^n \xi_i(w(A_i,X_i)f(R_i) - w(X_i',A_i')f(R_i'))\right)\right|\right)\right] \\
&\le \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(2\lambda \sup_{f\in\mathbb{F}(n)}\left|\left(\sum_i^n \xi_i w(A_i,X_i)f(R_i)\right)\right|\right)\right] \\
&= \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\sup_{f\in\mathbb{F}(n)}\exp\left(2\lambda \left|\left(\sum_i^n \xi_i w(A_i,X_i)f(R_i)\right)\right|\right)\right]
\end{aligned}
$$

with $\mathfrak{R}$ a Rademacher measure on a set of Rademacher random variable $\{\xi_i\}$ a Rademacher random variable.

Next, permute the indices i such that $R_1 \le \ldots R_i \ldots \le R_n$. Consider a function $f(r) = \frac{1}{n}\varrho \mathbb{1}_{\{r\le t\}}$. For such a function, $\sum_i^n \xi_i w(A_i,X_i)f(R_i)$ is equal to

- 0 if $t < \min_i\{R_i\}_i^n$,
- $\frac{1}{n}\varrho\sum_i^j w(A_i,X_i)\xi_i$ when $R_j \le t < R_{j+1}$ for a $j\in\{1,\ldots,n-1\}$,
- $\frac{1}{n}\varrho\sum_i^n w(A_i,X_i)\xi_i$ otherwise.

Then,

$$
\begin{aligned}
&\sup_{f\in\mathbb{F}(n)}\exp\left(2\lambda\left|\left(\sum_i^n \xi_i w(A_i,X_i)f(R_i)\right)\right|\right) \\
&\qquad = \max_{\varrho,j}\exp\left(\frac{2\lambda}{n}\varrho\sum_i^j w(A_i,X_i)\xi_i\right) \\
&\qquad = \max_j\Bigg(\exp\left(\frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i^j w(A_i,X_i)\xi_i\ge 0\}} \\
&\qquad\qquad + \exp\left(-\frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i^j w(A_i,X_i)\xi_i< 0\}}\Bigg) \\
&\qquad = \max_j\left(\exp\left(\frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i^j w(A_i,X_i)\xi_i\ge 0\}}\right) \\
&\qquad\qquad + \max_j\left(\exp\left(-\frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i^j w(A_i,X_i)\xi_i< 0\}}\right)
\end{aligned}
$$

Which gives us the inequality

$$
\mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\mathrm{IS}}(t)-F(t)\right|\right)\right] \le 2\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\max_j\exp\left(\frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i^j w(A_i,X_i)\xi_i\ge 0\}}\right] \tag{9}
$$

Now we are left to bound the right hand side of (9). Using Lemma B.1, for the right hand side of the (9) we have,

$$
\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\max_j\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\max_j\sum_i^j w(A_i,X_i)\xi_i\ge 0\}}\right]
$$

$$
\begin{aligned}
&= \mathbb{P}_\beta\{\max_j \frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i \geq 0\} \\
&\qquad\qquad + \lambda \int_0^\infty \exp(\lambda t)\mathbb{P}\{\max_j \frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i \geq t\}dt \\
&\leq \mathbb{P}_\beta\{\max_j \frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i \geq 0\} \\
&\qquad\qquad + 2\lambda \int_0^\infty \exp(\lambda t)\mathbb{P}\{\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i \geq t\}dt
\end{aligned} \tag{10}
$$

Note that similarly we have,

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i w(A_i,X_i)\xi_i \geq 0\}}\right] \\
=&\mathbb{P}_\beta\{\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i \geq 0\} + \lambda\int_0^\infty \exp(\lambda t)\mathbb{P}\{\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i \geq t\}dt
\end{aligned} \tag{11}
$$

Putting these two statements, i.e., (10), and (11) together, and applying the result of Lemma B.2, we have,

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\max_j \frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i^j w(A_i,X_i)\xi_i \geq 0\}}\right] \\
&\qquad \leq \mathbb{P}_\beta\{\max_j \frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i \geq 0\} \\
&\qquad\qquad + 2\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i w(A_i,X_i)\xi_i \geq 0\}}\right] \\
&\qquad\qquad\qquad - 2\mathbb{P}_\beta\{\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i \geq 0\} \\
&\qquad \leq 2\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i w(A_i,X_i)\xi_i \geq 0\}}\right] \\
&\qquad \leq 2\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i\right)\right]
\end{aligned}
$$

Note that $\frac{2}{n}w(A_i,X_i)\xi_i$ is a mean zero random variable with values in $[-\frac{2}{n}w_{\max}, \frac{2}{n}w_{\max}]$. Therefore, it is a sub-Gaussian random variable with sub-Gaussian constant as $\left(\frac{2}{n}\right)^2 w_{\max}^2$. Using this, we have, $\frac{2}{n}\sum_i w(A_i,X_i)\xi_i$ is $\frac{4}{n}w_{\max}^2$ sub-Gaussian random variable. Therefore, we have,

$$
\begin{aligned}
\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\max_j \frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i^j w(A_i,X_i)\xi_i \geq 0\}}\right] &\leq 2\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i\right)\right] \\
&\leq 2\exp\left(\lambda^2\frac{2}{n}w_{\max}^2\right)
\end{aligned}
$$

Putting this with the (9), we have

$$
\begin{aligned}
\mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda \sup_{t\in\mathbb{R}}\left|\widehat{F}_{\mathrm{IS}}(t)-F(t)\right|\right)\right] &\leq 2\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\max_j \frac{2\lambda}{n}\sum_i^j w(A_i,X_i)\xi_i\right)\mathbb{1}_{\{\sum_i^j w(A_i,X_i)\xi_i\geq 0\}}\right]\\
&\leq 4\exp\left(\lambda^2\frac{2}{n}w_{\max}^2\right)
\end{aligned}
$$

Using Markov inequality we have

$$
\begin{aligned}
\mathbb{P}_\beta\left(\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\mathrm{IS}}(t)-F(t)\right|\geq\epsilon\right) &= \mathbb{P}_\beta\left(\exp\left(\lambda\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\mathrm{IS}}(t)-F(t)\right|\right)\geq\exp(\lambda\epsilon)\right)\\
&\leq 4\exp\left(\lambda^2\frac{2}{n}w_{\max}^2\right)\exp(-\lambda\epsilon)\\
&= 4\exp\left(\lambda^2\frac{2}{n}w_{\max}^2-\lambda\epsilon\right)
\end{aligned}
$$

This holds for any choice of $\lambda>0$, resulting in

$$
\mathbb{P}_\beta\left(\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\mathrm{IS}}(t)-F(t)\right|\geq\epsilon\right)\leq\inf_{\lambda>0}4\exp\left(\lambda^2\frac{2}{n}w_{\max}^2-\lambda\epsilon\right)=4\exp\left(\frac{-n\epsilon^2}{8w_{\max}^2}\right)
$$

Using this, we have

$$
\mathbb{P}_\beta\left(\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\mathrm{IS}}(t)-F(t)\right|\leq\sqrt{\frac{8w_{\max}^2}{n}\log\left(\frac{4}{\delta}\right)}\right)\geq 1-\delta.
$$

Bernstein style: To bound this $\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i\right)\right]$ now we use Bernstein's. As discussed, the random variable $w(A_i,X_i)\xi_i$ is in $[-w_{\max},w_{\max}]$. However, if we look at its variance, we have $\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[w(A_i,X_i)^2\xi_i^2\right]=\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[w(A_i,X_i)^2\right]$ which is the second order Rényi divergence w_2. Therefore, for $0<\lambda<\frac{n}{2w_{\max}}$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i w(A_i,X_i)\xi_i\right)\right] &= \prod_i \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}w(A_i,X_i)\xi_i\right)\right]\\
&\leq\prod_i\exp\left(\frac{\lambda^2\frac{4w_2}{n^2}}{2\left(1-\lambda\frac{2}{n}w_{\max}\right)}\right)\\
&=\exp\left(\frac{n\lambda^2\frac{4w_2}{n^2}}{2\left(1-\lambda\frac{2}{n}w_{\max}\right)}\right)
\end{aligned}
$$

Using the Markov inequality, we have,

$$
\mathbb{P}_\beta\left(\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\mathrm{IS}}(t)-F(t)\right|\geq\epsilon\right)=4\exp\left(\frac{n\lambda^2\frac{4w_2}{n^2}}{2\left(1-\lambda\frac{2}{n}w_{\max}\right)}-\lambda\epsilon\right)
$$

Setting $\lambda=\frac{\epsilon}{\frac{2w_{\max}\epsilon}{n}+n\frac{4w_2}{n^2}}$, we have,

$$
\mathbb{P}_\beta\left(\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\mathrm{IS}}(t)-F(t)\right|\geq\epsilon\right)\leq 4\exp\left(\frac{-\epsilon^2}{2\left(\frac{2}{n}w_{\max}\epsilon+n\frac{4w_2}{n^2}\right)}\right)
$$

$$= 4\exp\left(\frac{-n\epsilon^2}{4w_{\max}\epsilon + 8w_{\mathbb{2}}}\right)$$

which results in,

$$\mathbb{P}_\beta\left(\sup_{t\in\mathbb{R}}\left|\widehat{F}_{\text{IS}}(t) - F(t)\right| \leq \frac{4w_{\max}\log(\frac{4}{\delta})}{n} + 2\sqrt{\frac{2w_{\mathbb{2}}\log(\frac{4}{\delta})}{n}}\right) \geq 1-\delta.$$

Finally, we note that since $\sup_t|\widehat{F}_{\text{IS-clip}}(t) - F(t)| \leq \sup_t|\widehat{F}_{\text{IS}}(t) - F(t)|$, the above results for $\widehat{F}_{\text{IS}}$ also hold for $\widehat{F}_{\text{IS-clip}}$. □

Auxiliary Lemmas

Lemma B.1. *For any random variable X, with probability measure $\mathbb{P}$, we have*

$$\mathbb{E}\left[\exp(\lambda X)\mathbb{1}_{\{X\geq 0\}}\right] = \mathbb{P}\{X \geq 0\} + \lambda\int_0^\infty \exp(\lambda t)\mathbb{P}\{X\geq t\}dt.$$

Proof. for any random variable X, with probability measure $\mathbb{P}$, we have

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda X)\mathbb{1}_{\{X\geq 0\}}\right] &= \mathbb{E}\left[\left(\exp(0) + \int_0^X \lambda\exp(\lambda t)dt\right)\mathbb{1}_{\{X\geq 0\}}\right]\\
&= \mathbb{E}\left[\mathbb{1}_{\{X\geq 0\}}\exp(0)\right] + \mathbb{E}\left[\mathbb{1}_{\{X\geq 0\}}\lambda\int_0^X \exp(\lambda t)\mathbb{1}_{\{X\geq 0\}}dt\right]\\
&= \mathbb{P}\{X\geq 0\} + \mathbb{E}\left[\lambda\int_0^X \exp(\lambda t)\mathbb{1}_{\{X\geq 0\}}dt\right]\\
&= \mathbb{P}\{X\geq 0\} + \lambda\int_0^\infty \exp(\lambda t)\mathbb{P}\{X\geq t\}dt. \qquad (12)
\end{aligned}$$

□

Lemma B.2. *For $\gamma > 0$, we have,*

$$\mathbb{P}_\beta\left[\max_j\sum_i^j w(A_i, X_i)\xi_i \geq \gamma\right] \leq 2\mathbb{P}_\beta\left[\sum_i^n w(A_i, X_i)\xi_i \geq \gamma\right] \qquad (13)$$

Proof. Consider events $E_j := \{\sum_i^j w(A_i, X_i)\xi_i \geq \gamma, \sum_i^l w(A_i, X_i)\xi_i < \gamma, \forall l < j\}$ with $E_0 := \emptyset$. Using these definitions, we have,

$$\{\max_j\sum_i^j w(A_i, X_i)\xi_i \geq \gamma\} \subset \bigcup_j E_j$$

Also,

$$\bigcup_j\left(E_j\bigcap\{\sum_{i>j} w(A_i, X_i)\xi_i \geq 0\}\right) \subset \{\sum_i w(A_i, X_i)\xi_i \geq \gamma\}$$

Also note that

$$\mathbb{P}_\beta\left[\sum_{i>j} w(A_i, X_i)\xi_i \geq 0\right] \geq \frac{1}{2}$$

since this quantity is mean zero and symmetric. Also note that the event $\sum_{i>j} w(A_i, X_i)\xi_i$ is independent of E_j.

Using these, we have,

$$\mathbb{P}_\beta\left[E_j \bigcap \{\sum_{i>j} w(A_i, X_i)\xi_i \geq 0\}\right] = \mathbb{P}_\beta\left[E_j\right]\mathbb{P}_\beta\left[\{\sum_{i>j} w(A_i, X_i)\xi_i \geq 0\}\right] \geq \frac{\mathbb{P}_\beta\left[E_j\right]}{2}$$

As a result we have,

$$\begin{aligned}
\mathbb{P}_\beta\left[\sum_i w(A_i, X_i)\xi_i \geq \gamma\right] &\geq \mathbb{P}_\beta\left[\bigcup_j \left(E_j \bigcap \{\sum_{i>j} w(A_i, X_i)\xi_i \geq 0\}\right)\right] \\
&= \sum_j \mathbb{P}_\beta\left[E_j \bigcap \{\sum_{i>j} w(A_i, X_i)\xi_i \geq 0\}\right] \\
&\geq \sum_j \frac{\mathbb{P}_\beta\left[E_j\right]}{2} \\
&\geq \frac{\mathbb{P}_\beta\left[\{\max_j \sum_i^j w(A_i, X_i)\xi_i \geq \gamma\}\right]}{2}
\end{aligned}$$

which concludes the statement. $\square$

C Proofs for Doubly Robust Estimators (Section 5.3)

C.1 Proof: Bias and Variance of DR CDF Estimate

Proof of Lemma 5.2. The expectation of the DR estimator (22) is as follows:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\beta}\left[\widehat{F}_{\mathrm{DR}}(t)\right] &= \mathbb{E}_{\mathbb{P}_\beta}\left[w(A,X)\mathbb{1}_{\{R\leq t\}}\right] + \mathbb{E}_{\mathbb{P}_\beta}\left[\overline{G}(t;X,\pi) - w(A,X)\overline{G}(t;X,A)\right] \\
&= F(t) + \mathbb{E}_{\mathbb{P}_\beta}\left[\overline{G}(t;X,\pi) - \mathbb{E}_{\mathbb{P}_\beta}[w(A,X)\overline{G}(t;X,A)|X]\right] \\
&= F(t) + \mathbb{E}_{\mathbb{P}_\beta}\left[\overline{G}(t;X,\pi) - \overline{G}(t;X,\pi)\right] \\
&= F(t).
\end{aligned}$$

Next, we derive the variance.

$$\begin{aligned}
\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{F}_{\mathrm{DR}}(t)\right] &= \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[w(A,X)\left(\mathbb{1}_{\{R\leq t\}} - \overline{G}(t;X,A)\right) + \overline{G}(t;X,\pi)\right] \\
&= \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[w(A,X)^2\sigma^2(t;X,A)\right] \\
&\quad + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[w(A,X)\left(G(t;X,A) - \overline{G}(t;X,A)\right) + \overline{G}(t;X,\pi)\right] \\
&= \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[w(A,X)^2\sigma^2(t;X,A)\right] + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[\mathbb{E}_{\mathbb{P}_\beta}\left[w(A,X)G(t;X,A)|X\right]\right] \\
&\quad + \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[\mathbb{V}_{\mathbb{P}_\beta}\left[w(A,X)\left(G(t;X,A) - \overline{G}(t;X,A)\right)|X\right]\right]
\end{aligned}$$

The first equality follows from applying the law of total variance, noting that the variance $\mathbb{V}_{\mathbb{P}_\beta}\left[\overline{G}(t;X,A)|X,A\right] = 0$, and using the definitions of G and σ^2. The second equality again applies the law of total variance. □

C.2 Proof: Error Bound of DR CDF Estimate

Proof of Theorem 5.2. Recall that the DR estimator $\widehat{F}_{\mathrm{DR}}(t)$ is defined as

$$\widehat{F}_{\mathrm{DR}}(t) = \frac{1}{n}\sum_{i=1}^{n} w(a_i,x_i)\left(\mathbb{1}_{\{r_i\leq t\}} - \overline{G}(t;x_i,a_i)\right) + \overline{G}(t;x_i,\pi)$$

where $\overline{G}(t;x,\pi) = \mathbb{E}_{A\sim\pi}\left[\overline{G}(t;x,A)\right]$. We can decompose the error of the DR estimator as:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta}\left[\sup_t |\widehat{F}_{\mathrm{DR}}(t) - F(t)|\right] \\
&\quad = \mathbb{E}_{\mathbb{P}_\beta}\left[\sup_t\left|\left(\frac{1}{n}\sum_{i=1}^{n} w(A_i,X_i)\left(\mathbb{1}_{\{R_i\leq t\}} - \overline{G}(t;X_i,A_i)\right) + \overline{G}(t;X_i,\pi)\right) - F(t)\right|\right] \\
&\quad \leq \mathbb{E}_{\mathbb{P}_\beta}\left[\sup_t\left(\left|\frac{1}{n}\sum_{i=1}^{n} w(A_i,X_i)\mathbb{1}_{\{R_i\leq t\}} - F(t)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\overline{G}(t;X_i,\pi) - w(A_i,X_i)\overline{G}(t;X_i,A_i)\right|\right)\right] \\
&\quad \leq \mathbb{E}_{\mathbb{P}_\beta}\left[\sup_t\left|\frac{1}{n}\sum_{i=1}^{n} w(A_i,X_i)\mathbb{1}_{\{R_i\leq t\}} - F(t)\right| + \sup_t\left|\frac{1}{n}\sum_{i=1}^{n}\overline{G}(t;X_i,\pi) - w(A_i,X_i)\overline{G}(t;X_i,A_i)\right|\right].
\end{aligned}$$

We have already bounded the first term in Theorem 5.1, and Lemma C.1 bounds the second term. Then in total, we have

$$\mathbb{P}_\beta\left(\sup_t\left|\widehat{F}_{\mathrm{DR}}(t) - F(t)\right| \geq \sqrt{\frac{8w_{max}^2}{n}\log\left(\frac{4}{\delta}\right)} + \sqrt{\frac{32w_{max}^2}{n}\log\frac{(2n)^{1/2}}{w_{max}\delta}}\right) \leq 2\delta$$

Simplifying,

$$\mathbb{P}_\beta\left(\sup_t\left|\widehat{F}_{DR}(t) - F(t)\right| \geq \sqrt{\frac{72w_{max}^2}{n}\log\left(\frac{4n^{1/2}}{\delta}\right)}\right) \leq 2\delta \tag{14}$$

which gives us our error bound for the DR estimator $\widehat{F}_{\text{DR}}$.

As mentioned previously, however, $\widehat{F}_{\text{DR}}$ may not be monotone, and in practice we must use a monotone transformation of the estimator. Consider a monotone transformation $\mathcal{M}$ of $\widehat{F}_{\text{DR}}$ that is a simple accumulation function, e.g. $\forall t$,

$$\mathcal{M}\left(\widehat{F}_{DR}(t)\right) = \max_{t' \leq t} \widehat{F}_{DR}(t')$$

Now we want to bound the error between the monotonized estimate $\mathcal{M}\left(\widehat{F}_{DR}(t)\right)$ and F. Using our error bound in (14), let $\epsilon = \sqrt{\frac{72 w_{max}^2}{n} \log\left(\frac{8n^{1/2}}{\delta}\right)}$. Then with probability at least $1-\delta$, for all $t \in \mathbb{R}$,

$$\max_t |\widehat{F}_{\text{DR}}(t) - F(t)| \leq \epsilon.$$

On this event, $\forall t$ there exists some $t' \leq t$ for which

$$\max_{t' \leq t} \widehat{F}_{DR}(t') - F(t) = \widehat{F}_{DR}(t') - F(t)$$

Using the fact that F is monotone thus $F(t') \leq F(t)$, when $\widehat{F}_{DR}(t') \geq F(t)$ we have

$$\widehat{F}_{DR}(t') - F(t) \leq \widehat{F}_{DR}(t') - F(t') \leq \epsilon$$

Similarly, when $\widehat{F}_{DR}(t') \leq F(t)$,

$$F(t) - \widehat{F}_{DR}(t') \leq F(t) - \widehat{F}_{DR}(t) \leq \epsilon$$

Putting these two inequalities together, we have

$$\max_t \left|\mathcal{M}\left(\widehat{F}_{DR}\right)(t) - F(t)\right| \leq \epsilon.$$

The theorem statement, which applies to the clipped monotone transformation, follows from the fact that

$$\max_t \left|\min\left\{\mathcal{M}\left(\widehat{F}_{DR}\right)(t), 1\right\} - F(t)\right| \leq \max_t \left|\mathcal{M}\left(\widehat{F}_{DR}\right)(t) - F(t)\right|.$$

□

Lemma C.1. *Let $\overline{G}(t;x,a)$ be a valid conditional CDF for all $x \in \mathcal{X}, a \in \mathcal{A}$, and let $w : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ be the importance sampling weights. Then for $\delta \in (0,1]$,*

$$\mathbb{P}_\beta\left(\sup_t \left|\frac{1}{n}\sum_{i=1}^n \overline{G}(t; X_i, \pi) - \frac{1}{n}\sum_{i=1}^n w(A_i, X_i)\overline{G}(t; X_i, A_i)\right| \geq \sqrt{\frac{32 w_{max}^2}{n} \log \frac{(2n)^{1/2}}{w_{max}\delta}}\right) \leq \delta.$$

where $\overline{G}(t;x,\pi) = \mathbb{E}_{\mathbb{P}}[\overline{G}(t;x,A)|x]$.

Proof. Since $\overline{G}$ is a valid CDF, we apply Lemma C.2 to $\overline{G}$. Consider a function of the form

$$\overline{\zeta}(t; s^1, ..., s^m) = \frac{1}{m}\sum_{j=1}^m \mathbb{1}_{\{s^i \leq t\}}$$

The function $\overline{\zeta}$ can be seen as a stepwise CDF function, where each step is $1/m$ and occurs at points $\{s^j\}_{j=1}^m$.

Lemma C.2 approximates $\overline{G}$ using such $1/m$-stepwise CDFs. For each context x and action a, let $s_{x,a}^1, ..., s_{x,a}^m \in \mathbb{Q}^m$ be the points chosen according to the deterministic procedure in Lemma C.2, such that the following inequality holds:

$$\sup_t \left|\overline{G}(t;x,a) - \overline{\zeta}\left(t; \{s_{x,a}^j\}_{j=1}^m\right)\right| \leq \frac{1}{2m}. \tag{15}$$

Next, consider the class of functions

$$\mathcal{G}(m) := \left\{ \zeta(s^1, ..., s^m) := \frac{1}{m} \varrho \sum_{j=1}^{m} \mathbb{1}_{\{s^j \le t\}} : \forall t \in \mathbb{R}, \varrho \in \{-1, +1\}; \{s^j\}_{j=1}^m \in \mathbb{Q}^m \right\}$$

$$\mathcal{G}(m) := \left\{ \zeta(\cdot; s^1, ..., s^m) : \mathbb{R} \to [0, 1] := \frac{1}{m} \varrho \sum_{j=1}^{m} \mathbb{1}_{\{s^j \le t\}} : \varrho \in \{-1, +1\}; \{s^j\}_{j=1}^m \in \mathbb{Q}^m \right\}$$

Note that, $\overline{\zeta}$ is a subset of the function class $\mathcal{G}(m)$, e.g. $\overline{\zeta}\left(t; \{s^j_{x,a}\}_{j=1}^m\right) \in \mathcal{G}(m)$.

Then our problem becomes

$$\begin{aligned}
&\sup_t \left| \frac{1}{n} \sum_{i=1}^{n} w(A_i, X_i) \overline{G}(t; X_i, A_i) - \frac{1}{n} \sum_{i=1}^{n} \overline{G}(t; X_i, \pi) \right| \\
&\quad = \sup_t \left| \frac{1}{n} \sum_{i=1}^{n} w(A_i, X_i) \overline{G}(t; X_i, A_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}} \left[\overline{G}(t; X_i, A) | X_i \right] \right| \\
&\quad = \sup_t \left| \frac{1}{n} \sum_{i=1}^{n} w(A_i, X_i) \overline{G}(t; X_i, A_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}_\beta} \left[w(X_i, A) \overline{G}(t; X_i, A) | X_i \right] \right| \\
&\quad \le \sup_t \left| \frac{1}{n} \sum_{i=1}^{n} w(A_i, X_i) \overline{\zeta}\left(t; \{s^j_{X_i,A_i}\}_{j=1}^m\right) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}_\beta} \left[w(A, X_i) \overline{\zeta}\left(t; \{s^j_{X_i,A}\}_{j=1}^m\right) \Big| X_i \right] \right| + \frac{1}{m} \\
&\quad \le \sup_{\zeta \in \mathcal{G}(m)} \left| \frac{1}{n} \sum_{i=1}^{n} w(A_i, X_i) \zeta(\{s^j_{X_i,A_i}\}_{j=1}^m) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}_\beta} \left[w(A, X_i) \zeta(\{s^j_{X_i,A}\}_{j=1}^m) \Big| X_i \right] \right| + \frac{1}{m}
\end{aligned}$$

where the second line uses the definition of $\overline{G}(t; X_i, \pi)$, the third line uses a change of measure through the importance sampling weight w, the fourth line uses (C.2), and the last line uses the fact that, conditioned on $\{s^j_{x,a}\}_{j=1}^m$, the function $\overline{\zeta}$ is a member of $\mathcal{G}(m)$.

We can now upper bound the RHS. Going forward, we refer to $\zeta(\{s^j_{X,A}\}_{j=1}^m)$ as $\zeta(X, A)$ for short. Then for $\lambda > 0$ we have:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta} \left[\exp\left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \left(\frac{1}{n} \sum_{i=1}^{n} w(A_i, X_i) \zeta(X_i, A_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}_\beta} \left[w(A, X_i) \zeta(X_i, A) \Big| X_i \right] \right) \right) \right] \\
&\quad = \mathbb{E}_{\mathbb{P}_\beta} \left[\exp\left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}_\beta} \left[w(A_i, X_i) \zeta(X_i, A_i) - w(A'_i, X_i) \zeta(X_i, A'_i) \Big| \{X_i, A_i\}_{i=1}^n \right] \right) \right) \right] \\
&\quad \le \mathbb{E}_{\mathbb{P}_\beta} \left[\exp\left(\lambda \mathbb{E}_{\mathbb{P}_\beta} \left[\sup_{\zeta \in \mathcal{G}(m)} \frac{1}{n} \sum_{i=1}^{n} (w(A_i, X_i) \zeta(X_i, A_i) - w(A'_i, X_i) \zeta(X_i, A'_i)) \Big| \{X_i, A_i\}_{i=1}^n \right] \right) \right] \\
&\quad \le \mathbb{E}_{\mathbb{P}_\beta} \left[\exp\left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \frac{1}{n} \sum_{i=1}^{n} (w(A_i, X_i) \zeta(X_i, A_i) - w(A'_i, X_i) \zeta(X_i, A'_i)) \right) \right] \\
&\quad \le \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}} \left[\exp\left(2\lambda \sup_{\zeta \in \mathcal{G}(m)} \frac{1}{n} \sum_{i=1}^{n} \xi_i w(A_i, X_i) \zeta(X_i, A_i) \right) \right] \\
&\quad = \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}} \left[\sup_{\zeta \in \mathcal{G}(m)} \exp\left(2\lambda \frac{1}{n} \sum_{i=1}^{n} \xi_i w(A_i, X_i) \zeta(X_i, A_i) \right) \right] \\
&\quad = \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}} \left[\sup_{t, \varrho} \exp\left(2\lambda \frac{\varrho}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} \xi_i w(A_i, X_i) \mathbb{1}_{\{s^j_{X_i,A_i} \le t\}} \right) \right]
\end{aligned} \tag{16}$$

where $\{A'\}_i^n$ are the ghost variables, the second to last inequality uses symmetrization (Lemma C.3), and the last line uses the definition of $\zeta(X_i, A_i) = \zeta(s^1_{X_i,A_i}, ..., s^m_{X_i,A_i})$.

Now, for each j, permute the indices i such that $s^j_{X_{j(1)},A_{j(1)}} \le ... \le s^j_{X_{j(i)},A_{j(i)}} \le ... \le s^j_{X_{j(n)},A_{j(n)}}$. Then, for a given j, consider the function

$$\sum_{i=1}^{n} \xi_{j(i)} w(A_{j(i)}, X_{j(i)}) \mathbb{1}_{\{s^j_{X_{j(i)},A_{j(i)}} \le t\}},$$

which equals

1. 0 if $t < s^j_{X_{j(1)},A_{j(1)}}$,
2. $\varrho \sum_{i=1}^{k} w(A_{j(i)}, X_{j(i)})\xi_{j(i)}$ if there exists $k \in \{1, ..., n-1\}$ such that $s^j_{X_{j(k)},A_{j(k)}} \le t \le s^j_{X_{j(k)+1},A_{j(k)+1}}$,
3. $\varrho \sum_{i=1}^{n} w(A_{j(i)}, X_{j(i)})\xi_{j(i)}$ otherwise.

Then the RHS of (19) equals

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}} \left[\sup_{t,\varrho} \exp\left(2\lambda \frac{\varrho}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} \xi_i w(A_i, X_i) \mathbb{1}_{\{s^j_{X_i,A_i} \le t\}} \right) \right] \\
&= \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}} \left[\max_{k,\varrho} \exp\left(2\lambda \frac{\varrho}{nm} \sum_{j=1}^{m} \sum_{i=1}^{k} \xi_{j(i)} w(A_{j(i)}, X_{j(i)}) \right) \right] \\
&\le \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}} \left[\max_{j,k,\varrho} \exp\left(2\lambda \frac{\varrho}{n} \sum_{i=1}^{k} \xi_{j(i)} w(A_{j(i)}, X_{j(i)}) \right) \right].
\end{aligned}$$

Further, we have that

$$\begin{aligned}
&\max_{j,k,\varrho} \exp\left(2\lambda \frac{\varrho}{n} \sum_{i=1}^{k} \xi_{j(i)} w(A_{j(i)}, X_{j(i)}) \right) \\
&= \max_{j,k} \Bigg(\exp\left(\frac{2\lambda}{n} \sum_{i}^{k} w(A_{j(i)}, X_{j(i)})\xi_{j(i)} \right) \mathbb{1}_{\{\sum_i^k w(A_{j(i)},X_{j(i)})\xi_{j(i)} \ge 0\}} \\
&\qquad + \exp\left(-\frac{2\lambda}{n} \sum_{i}^{k} w(A_{j(i)}, X_{j(i)})\xi_{j(i)} \right) \mathbb{1}_{\{\sum_i^k w(a_{j(i)},x_{j(i)})\xi_{j(i)} < 0\}} \Bigg) \\
&\le 2 \max_{j,k} \exp\left(\frac{2\lambda}{n} \sum_{i}^{k} w(A_{j(i)}, X_{j(i)})\xi_{j(i)} \right) \mathbb{1}_{\{\sum_i^k w(A_{j(i)},X_{j(i)})\xi_{j(i)} \ge 0\}}.
\end{aligned}$$

Putting it together, we have that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta} \left[\exp\left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \left| \frac{1}{n} \sum_{i=1}^{n} w(A_i, X_i)\zeta(X_i, A_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{A\sim\pi(\cdot|X_i)} \left[\zeta(X_i, A) | X_i \right] \right| \right) \right] \\
&\le 2\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}} \left[\max_{j,k} \exp\left(\frac{2\lambda}{n} \sum_{i}^{k} w(A_{j(i)}, X_{j(i)})\xi_{j(i)} \right) \mathbb{1}_{\{\sum_i^k w(A_{j(i)},X_{j(i)})\xi_{j(i)} \ge 0\}} \right] \quad (17)
\end{aligned}$$

Now we are left to bound the RHS of (17). Using Lemma B.1,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}} \left[\max_{j,k} \exp\left(\frac{2\lambda}{n} \sum_{i}^{k} w(A_{j(i)}, X_{j(i)})\xi_{j(i)} \right) \mathbb{1}_{\{\max_k \sum_i^k w(A_{j(i)},X_{j(i)})\xi_{j(i)} \ge 0\}} \right] \\
&\le \mathbb{P}_\beta \left(\max_{k} \frac{2\lambda}{n} \sum_{i}^{k} w(A_{j(i)}, X_{j(i)})\xi_{j(i)} \ge 0 \right) + 2\lambda \sum_{j} \int_0^\infty \exp(\lambda t) \mathbb{P}\left(\frac{2\lambda}{n} \sum_{i=1}^{n} w(A_{j(i)}, X_{j(i)})\xi_{j(i)} \ge t \right) dt.
\end{aligned}$$

Similarly, for any j, we have

$$\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^k w(A_{j(i)},X_{j(i)})\xi_{j(i)}\geq 0\}}\right]$$
$$=\mathbb{P}_\beta\left(\frac{2\lambda}{n}\sum_i^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\geq 0\right)+\lambda\int_0^\infty \exp(\lambda t)\mathbb{P}\left(\frac{2\lambda}{n}\sum_{i=1}^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\geq t\right)dt$$

Putting these two together, we have

$$\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\max_{j,k}\exp\left(\frac{2\lambda}{n}\sum_i^k w(A_{j(i)},X_{j(i)})\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^k w(A_{j(i)},X_{j(i)})\xi_{j(i)}\geq 0\}}\right]$$
$$\leq \sum_j \mathbb{P}_\beta\left(\max_k \frac{2\lambda}{n}\sum_i^k w(A_{j(i)},X_{j(i)})\xi_{j(i)}\geq 0\right)-2\sum_j \mathbb{P}_\beta\left(\frac{2\lambda}{n}\sum_i^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\geq 0\right)$$
$$+2\sum_j \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\geq 0\}}\right]$$
$$\leq 2\sum_j \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\geq 0\}}\right]$$
$$\leq 2m\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i^n w(A_{j(i)},X_{j(i)})\xi_{j(i)}\right)\right]$$
$$\leq 2m\exp\left(\frac{2\lambda^2 w_{max}^2}{n}\right)$$

where the last inequality uses the fact that ξ is a Rademacher random variable, and $w(A,X)\leq w_{max}$. Finally, using Markov's inequality,

$$\mathbb{P}_\beta\left(\sup_t\left|\frac{1}{n}\sum_{i=1}^n \overline{G}(t;X_i,\pi)-w(A_i,X_i)\overline{G}(t;X_i,A_i)\right|\geq \epsilon+\frac{1}{m}\right)$$
$$\leq \mathbb{P}_\beta\left(\exp\left(\lambda\sup_t\left|\frac{1}{n}\sum_{i=1}^n w(A_i,X_i)\zeta(X_i,A_i)-\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}_\beta}\left[\zeta(X_i,A)|X_i\right]\right|\right)\geq \exp(\lambda\epsilon)\right)$$
$$\leq 4m\exp\left(\frac{2\lambda^2 w_{max}^2}{n}-\lambda\epsilon\right)$$

Because this holds for any $\lambda>0$, we can minimize the RHS over λ:

$$\mathbb{P}_\beta\left(\sup_t\left|\frac{1}{n}\sum_{i=1}^n \overline{G}(t;X_i,\pi)-w(A_i,X_i)\overline{G}(t;X_i,A_i)\right|\geq \epsilon+\frac{1}{m}\right)\leq \inf_{\lambda>0} 4m\exp\left(\frac{2\lambda^2 w_{max}^2}{n}-\lambda\epsilon\right)$$
$$=4m\exp\left(\frac{-n\epsilon}{8w_{max}^2}\right).$$

Then we have

$$\mathbb{P}_\beta\left(\sup_t\left|\frac{1}{n}\sum_{i=1}^n \overline{G}(t;X_i,\pi)-w(A_i,X_i)\overline{G}(t;X_i,A_i)\right|\geq \sqrt{\frac{8w_{max}^2}{n}\log\frac{4m}{\delta}}+\frac{1}{m}\right)\leq \delta.$$

Setting $m=\sqrt{n/8w_{max}^2}$ gives the theorem statement:

$$\mathbb{P}_\beta\left(\sup_t\left|\frac{1}{n}\sum_{i=1}^n \overline{G}(t;X_i,\pi)-w(A_i,X_i)\overline{G}(t;X_i,A_i)\right|\geq \sqrt{\frac{32w_{max}^2}{n}\log\frac{(2n)^{1/2}}{w_{max}\delta}}\right)\leq \delta.$$

□

Auxiliary Lemmas

Lemma C.2. *For any ζ, a non-decreasing function with support $[0, D]$, there exists m points $s^1....s^m \in \mathbb{Q}^m$ such that for a function of the form,*

$$\overline{\zeta}(t; s^1, ..., s^m) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{\{s^j \leq t\}}, \quad \forall t \in \mathbb{R}$$

the following inequality holds:

$$\|\zeta - \overline{\zeta}\|_\infty \leq \frac{1}{2m}.$$

Proof of Lemma C.2. Uniformly partition the interval $[0, D]$ to m partitions, with partition points $\{\frac{j}{D}\}_{j=0}^m$. We construct the set $\{s_j\}_{j=1}^m$ using the following procedure. For any $j \in \{1, \ldots, m\}$ and the corresponding partition point $\frac{j-1}{D}$, let $s^j \in \mathbb{Q}$ be a point such that either $\lim_{t \to s^j_-} \zeta(t) = \frac{j-1}{m} + \frac{1}{2m}$ or $\lim_{t \to s^j_+} \zeta(t) = \frac{j-1}{m} + \frac{1}{2m}$ (e.g., as illustrated in Figure 4). Then for any t, $\overline{\zeta}(t)$ is $\frac{1}{2m}$-close to $\zeta(t)$. ☐

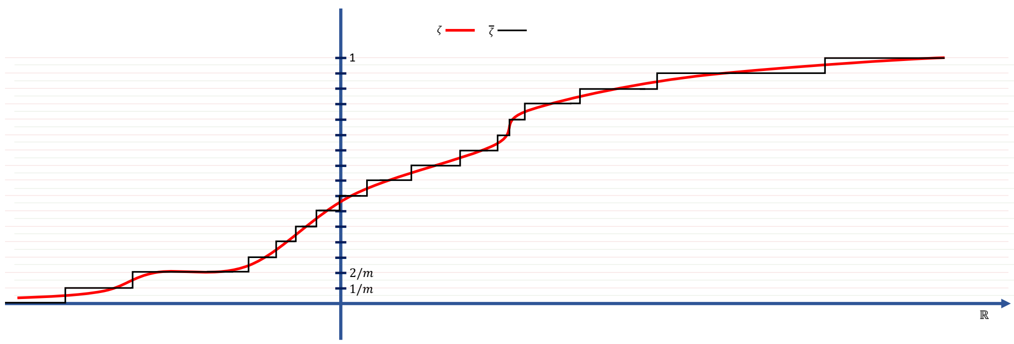


Figure 4: Approximating monotonic function ζ with $\overline{\zeta}$ **Kamyar:** is shown when the support is $[0, 1]$.

Lemma C.3. *For the function class $\mathcal{G}$ defined in Appendix C.2, we have for any $\lambda > 0$ that*

$$\mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \frac{1}{n}\sum_{i=1}^{n}(w(A_i, X_i)\zeta(X_i, A_i) - w(A'_i, X_i)\zeta(X_i, A'_i))\right)\right]$$
$$\leq \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}}\left[\sup_{\zeta \in \mathcal{G}(m)} \exp\left(2\lambda \frac{1}{n}\sum_{i=1}^{n} \xi_i w(A_i, X_i)\zeta(X_i, A_i)\right)\right]$$

where contexts and actions $X, A, A' \sim \mathbb{P}_\beta$, and Rademacher random variables $\xi_i \sim \mathfrak{R}$.

Proof. For each $i = 1, ..., n$, and let ξ_i be i.i.d. Rademacher random variables. Set

$$A_i^+ = \begin{cases} A_i, & \text{if } \xi_i = 1 \\ A'_i, & \text{if } \xi_i = -1 \end{cases}$$

$$A_i^- = \begin{cases} A'_i, & \text{if } \xi_i = 1 \\ A_i, & \text{if } \xi_i = -1 \end{cases}$$

We have that, conditioned on X_i, $(A_i^+, A_i^-) \overset{d}{=} (A_i, A_i')$. Then

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda \sup_{\zeta\in\mathcal{G}(m)} \frac{1}{n}\sum_{i=1}^n \left(w(A_i, X_i)\zeta(X_i, A_i) - w(A_i', X_i)\zeta(X_i, A_i')\right)\right)\right] \\
&\quad = \mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda \sup_{\zeta\in\mathcal{G}(m)} \frac{1}{n}\sum_{i=1}^n \left(w(A_i^+, X_i)\zeta(X_i, A_i^+) - w(A_i^-, X_i)\zeta(X_i, A_i^-)\right)\right)\right] \\
&\quad = \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}}\left[\exp\left(\lambda \sup_{\zeta\in\mathcal{G}(m)} \frac{1}{n}\sum_{i=1}^n \xi_i\left(w(A_i, X_i)\zeta(X_i, A_i) - w(A_i', X_i)\zeta(X_i, A_i')\right)\right)\right]
\end{aligned}$$

Our last step is to bound the last line of the above display.

$$\begin{aligned}
&= \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}}\left[\exp\left(\lambda \sup_{\zeta\in\mathcal{G}(m)} \frac{1}{2}\left(\frac{2}{n}\sum_{i=1}^n \xi_i w(A_i, X_i)\zeta(X_i, A_i) - \frac{2}{n}\sum_{i=1}^n \xi_i w(A_i', X_i)\zeta(X_i, A_i')\right)\right)\right] \\
&\leq \frac{1}{2}\mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}}\left[\exp\left(\lambda \sup_{\zeta\in\mathcal{G}(m)} \frac{2}{n}\sum_{i=1}^n \xi_i w(A_i, X_i)\zeta(X_i, A_i)\right)\right] \\
&\quad + \frac{1}{2}\mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}}\left[\exp\left(\lambda \sup_{\zeta\in\mathcal{G}(m)} \frac{2}{n}\sum_{i=1}^n (-\xi_i) w(A_i', X_i)\zeta(X_i, A_i')\right)\right] \\
&= \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}}\left[\exp\left(\lambda \sup_{\zeta\in\mathcal{G}(m)} \frac{2}{n}\sum_{i=1}^n \xi_i w(A_i, X_i)\zeta(X_i, A_i)\right)\right]
\end{aligned}$$

□

Lemma C.4. *Let $G(t; X, \pi) = \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{\{R\leq t\}}|X]$ be the conditional CDF of returns for all $x \in \mathcal{X}$. Then for $\delta \in (0, 1]$,*

$$\mathbb{P}_\beta\left(\sup_t \left|\frac{1}{n}\sum_{i=1}^n G(t; X_i, \pi) - F(t)\right| \geq \sqrt{\frac{32}{n}\log\frac{(2n)^{1/2}}{\delta}}\right) \leq \delta.$$

Proof. Since G is a valid CDF, we apply Lemma C.2 to $\overline{G}$. Consider a function of the form

$$\overline{\zeta}(t; s^1, ..., s^m) = \frac{1}{m}\sum_{j=1}^m \mathbb{1}_{\{s^i\leq t\}}$$

The function $\overline{\zeta}$ can be seen as a stepwise CDF function, where each step is $1/m$ and occurs at points $\{s^j\}_{j=1}^m$.

Lemma C.2 approximates $\overline{G}$ using such $1/m$-stepwise CDFs. For each context x, let $s_x^1, ..., s_x^m \in \mathbb{Q}^m$ be the points chosen according to the deterministic procedure in Lemma C.2, such that the following inequality holds:

$$\sup_t \left|G(t; x, \pi) - \overline{\zeta}\left(t; \{s_x^j\}_{j=1}^m\right)\right| \leq \frac{1}{2m}. \tag{18}$$

Next, consider the class of functions

$$\mathcal{G}(m) := \left\{\zeta(s^1, ..., s^m) := \frac{1}{m}\varrho\sum_{j=1}^m \mathbb{1}_{\{s^j\leq t\}} : \forall t\in\mathbb{R}, \varrho\in\{-1, +1\}; \{s^j\}_{j=1}^m \in \mathbb{Q}^m\right\}$$

Note that, $\overline{\zeta}$ is a subset of the function class $\mathcal{G}(m)$, e.g. $\overline{\zeta}\left(t; \{s_x^j\}_{j=1}^m\right) \in \mathcal{G}(m)$.

Then our problem becomes

$$
\begin{aligned}
&\sup_t \left| \frac{1}{n} \sum_{i=1}^n G(t; X_i, \pi) - F(t) \right| \\
&\quad = \sup_t \left| \frac{1}{n} \sum_{i=1}^n G(t; X_i, \pi) - \mathbb{E}_{\mathbb{P}_\beta} \left[\frac{1}{n} \sum_{i=1}^n G(t; X, \pi) \right] \right| \\
&\quad \leq \sup_t \left| \frac{1}{n} \sum_{i=1}^n \overline{\zeta} \left(t; \{s_{X_i}^j\}_{j=1}^m \right) - \mathbb{E}_{\mathbb{P}_\beta} \left[\frac{1}{n} \sum_{i=1}^n \overline{\zeta} \left(t; \{s_{X_i}^j\}_{j=1}^m \right) \right] \right| + \frac{1}{m} \\
&\quad \leq \sup_{\zeta \in \mathcal{G}(m)} \left| \frac{1}{n} \sum_{i=1}^n \zeta \left(\{s_{X_i}^j\}_{j=1}^m \right) - \mathbb{E}_{\mathbb{P}_\beta} \left[\frac{1}{n} \sum_{i=1}^n \zeta \left(\{s_{X_i}^j\}_{j=1}^m \right) \right] \right| + \frac{1}{m}
\end{aligned}
$$

We can now upper bound the RHS. Going forward, we refer to $\zeta(\{s_X^j\}_{j=1}^m)$ as $\zeta(X)$ for short. Then for $\lambda > 0$ we have:

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\beta} \left[\exp \left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \left(\frac{1}{n} \sum_{i=1}^n \zeta(X_i) - \mathbb{E}_{\mathbb{P}_\beta} \left[\frac{1}{n} \sum_{i=1}^n \zeta(X_i) \right] \right) \right) \right] \\
&\qquad = \mathbb{E}_{\mathbb{P}_\beta} \left[\exp \left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbb{P}_\beta} \left[\zeta(X_i) - \zeta(X_i') \middle| \{X_i\}_{i=1}^n \right] \right) \right) \right] \\
&\qquad \leq \mathbb{E}_{\mathbb{P}_\beta} \left[\exp \left(\lambda \mathbb{E}_{\mathbb{P}_\beta} \left[\sup_{\zeta \in \mathcal{G}(m)} \frac{1}{n} \sum_{i=1}^n \left(\zeta(X_i) - \zeta(X_i') \right) \middle| \{X_i\}_{i=1}^n \right] \right) \right] \\
&\qquad \leq \mathbb{E}_{\mathbb{P}_\beta} \left[\exp \left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \frac{1}{n} \sum_{i=1}^n \left(\zeta(X_i) - \zeta(X_i', A_i') \right) \right) \right] \\
&\qquad = \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}} \left[\exp \left(\lambda \sup_{\zeta \in \mathcal{G}(m)} \frac{1}{n} \sum_{i=1}^n \xi_i \left(\zeta(X_i) - \zeta(X_i') \right) \right) \right] \\
&\qquad \leq \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}} \left[\exp \left(2\lambda \sup_{\zeta \in \mathcal{G}(m)} \frac{1}{n} \sum_{i=1}^n \xi_i \zeta(X_i) \right) \right] \\
&\qquad = \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}} \left[\sup_{\zeta \in \mathcal{G}(m)} \exp \left(2\lambda \frac{1}{n} \sum_{i=1}^n \xi_i \zeta(X_i) \right) \right] \\
&\qquad = \mathbb{E}_{\mathbb{P}_\beta, \mathfrak{R}} \left[\sup_{t, \varrho} \exp \left(2\lambda \frac{\varrho}{nm} \sum_{j=1}^m \sum_{i=1}^n \xi_i \mathbb{1}_{\{s_{X_i}^j \leq t\}} \right) \right]
\end{aligned}
\tag{19}
$$

where $\{X'\}_i^n$ are the ghost variables, and the last line uses the definition of $\zeta(X_i) = \zeta(s_{X_i}^1, ..., s_{X_i}^m)$.

Now, for each j, permute the indices i such that $s_{X_{j(1)}}^j \leq ... \leq s_{X_{j(i)}}^j \leq ... \leq s_{X_{j(n)}}^j$. Then, for a given j, consider the function

$$
\sum_{i=1}^n \xi_{j(i)} \mathbb{1}_{\{s_{X_{j(i)}}^j \leq t\}},
$$

which equals

1. 0 if $t < s_{X_{j(1)}}^j$,
2. $\varrho \sum_{i=1}^k \xi_{j(i)}$ if there exists $k \in \{1, ..., n-1\}$ such that $s_{X_{j(k)}}^j \leq t \leq s_{X_{j(k)+1}}^j$,
3. $\varrho \sum_{i=1}^n \xi_{j(i)}$ otherwise.

Then the RHS of (19) equals

$$\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\sup_{t,\varrho}\exp\left(2\lambda\frac{\varrho}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\xi_i\mathbb{1}_{\{s^j_{X_i}\le t\}}\right)\right]$$
$$=\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\max_{k,\varrho}\exp\left(2\lambda\frac{\varrho}{nm}\sum_{j=1}^{m}\sum_{i=1}^{k}\xi_{j(i)}\right)\right]$$
$$\le\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\max_{j,k,\varrho}\exp\left(2\lambda\frac{\varrho}{n}\sum_{i=1}^{k}\xi_{j(i)}\right)\right].$$

Further, we have that

$$\max_{j,k,\varrho}\exp\left(2\lambda\frac{\varrho}{n}\sum_{i=1}^{k}\xi_{j(i)}\right)$$
$$=\max_{j,k}\left(\exp\left(\frac{2\lambda}{n}\sum_{i}^{k}\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^k\xi_{j(i)}\ge 0\}}+\exp\left(-\frac{2\lambda}{n}\sum_{i}^{k}\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^k\xi_{j(i)}<0\}}\right)$$
$$\le 2\max_{j,k}\exp\left(\frac{2\lambda}{n}\sum_{i}^{k}\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^k\xi_{j(i)}\ge 0\}}.$$

Putting it together, we have that

$$\mathbb{E}_{\mathbb{P}_\beta}\left[\exp\left(\lambda\sup_{\zeta\in\mathcal{G}(m)}\left|\frac{1}{n}\sum_{i=1}^{n}\zeta(X_i)-\mathbb{E}_{\mathbb{P}_\beta}\left[\frac{1}{n}\sum_{i=1}^{n}\zeta(X_i)\right]\right|\right)\right]$$
$$\le 2\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\max_{j,k}\exp\left(\frac{2\lambda}{n}\sum_{i}^{k}\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^k\xi_{j(i)}\ge 0\}}\right] \tag{20}$$

Now we are left to bound the RHS of (20). Using Lemma B.1,

$$\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\max_{j,k}\exp\left(\frac{2\lambda}{n}\sum_{i}^{k}\xi_{j(i)}\right)\mathbb{1}_{\{\max_k\sum_i^k\xi_{j(i)}\ge 0\}}\right]$$
$$\le\mathbb{P}_\beta\left(\max_k\frac{2\lambda}{n}\sum_{i}^{k}\xi_{j(i)}\ge 0\right)+2\lambda\sum_j\int_0^\infty\exp(\lambda t)\mathbb{P}\left(\frac{2\lambda}{n}\sum_{i=1}^{n}\xi_{j(i)}\ge t\right)dt.$$

Similarly, for any j, we have

$$\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_{i}^{n}\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^k\xi_{j(i)}\ge 0\}}\right]$$
$$=\mathbb{P}_\beta\left(\frac{2\lambda}{n}\sum_{i}^{n}w(A_{j(i)},X_{j(i)})\xi_{j(i)}\ge 0\right)+\lambda\int_0^\infty\exp(\lambda t)\mathbb{P}\left(\frac{2\lambda}{n}\sum_{i=1}^{n}\xi_{j(i)}\ge t\right)dt$$

Putting these two together, we have

$$\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\max_{j,k}\exp\left(\frac{2\lambda}{n}\sum_{i}^{k}\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^k\xi_{j(i)}\ge 0\}}\right]$$
$$\le\sum_j\mathbb{P}_\beta\left(\max_k\frac{2\lambda}{n}\sum_{i}^{k}\xi_{j(i)}\ge 0\right)-2\sum_j\mathbb{P}_\beta\left(\frac{2\lambda}{n}\sum_{i}^{n}\xi_{j(i)}\ge 0\right)$$
$$+2\sum_j\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_{i}^{n}\xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^n\xi_{j(i)}\ge 0\}}\right]$$

$$
\begin{aligned}
&\leq 2\sum_j \mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i^n \xi_{j(i)}\right)\mathbb{1}_{\{\sum_i^n \xi_{j(i)}\geq 0\}}\right] \\
&\leq 2m\mathbb{E}_{\mathbb{P}_\beta,\mathfrak{R}}\left[\exp\left(\frac{2\lambda}{n}\sum_i^n \xi_{j(i)}\right)\right] \\
&\leq 2m\exp\left(\frac{2\lambda^2}{n}\right)
\end{aligned}
$$

where the last inequality uses the fact that ξ is a Rademacher random variable. Finally, using Markov's inequality,

$$
\begin{aligned}
&\mathbb{P}_\beta\left(\sup_t\left|\frac{1}{n}\sum_{i=1}^n G(t;X_i,\pi)-F(t)\right|\geq \epsilon+\frac{1}{m}\right) \\
&\quad\leq \mathbb{P}_\beta\left(\exp\left(\lambda\sup_t\left|\frac{1}{n}\sum_{i=1}^n \zeta(X_i)-\mathbb{E}_{\mathbb{P}_\beta}\left[\frac{1}{n}\sum_{i=1}^n\zeta(X_i)\right]\right|\right)\geq\exp(\lambda\epsilon)\right) \\
&\quad\leq 4m\exp\left(\frac{2\lambda^2}{n}-\lambda\epsilon\right)
\end{aligned}
$$

Because this holds for any $\lambda > 0$, we can minimize the RHS over λ:

$$
\begin{aligned}
\mathbb{P}_\beta\left(\sup_t\left|\frac{1}{n}\sum_{i=1}^n G(t;X_i,\pi)-F(t)\right|\geq \epsilon+\frac{1}{m}\right) &\leq \inf_{\lambda>0} 4m\exp\left(\frac{2\lambda^2}{n}-\lambda\epsilon\right) \\
&= 4m\exp\left(\frac{-n\epsilon}{8}\right).
\end{aligned}
$$

Then we have

$$
\mathbb{P}_\beta\left(\sup_t\left|\frac{1}{n}\sum_{i=1}^n G(t;X_i,\pi)-F(t)\right|\geq \sqrt{\frac{8}{n}\log\frac{4m}{\delta}}+\frac{1}{m}\right)\leq\delta.
$$

Setting $m=\sqrt{n/8}$ gives the theorem statement:

$$
\mathbb{P}_\beta\left(\sup_t\left|\frac{1}{n}\sum_{i=1}^n G(t;X_i,\pi)-F(t)\right|\geq \sqrt{\frac{32}{n}\log\frac{(2n)^{1/2}}{\delta}}\right)\leq\delta.
$$

□

D Proofs for Risk Functional Estimation (Section 6)

Proof of Theorem 6.1. By the definition of L-Lipschitz risk functionals, for the CDFs F and $\widehat{F}$,

$$
\begin{aligned}
|\rho(\widehat{F})-\rho(F)| &\leq L\|\widehat{F}-F\|_\infty \\
&\leq L\epsilon
\end{aligned}
$$

with probability at least $1-\delta$, where the last line uses the fact that $\widehat{F}$ is ϵ-close to F with probability at least $1-\delta$. □

E Risk Estimation with Unknown Behavior Policy

We begin this section with a consideration of estimators when the behavior policy is unknown, and must be modeled or estimated, which we call $\widehat{\beta}$. We first define the IS, DR, and DI estimators using $\widehat{\beta}$, then derive their bias and variance expressions. To differentiate between the estimator that use β and the estimators that use $\widehat{\beta}$, we call the latter $\widetilde{F}$ while continuing to call the former $\widehat{F}$.

The proofs of bias and variance begins with derivations for the DR estimator with estimated policy, from which the bias and variance of the remaining estimators can be derived as special cases.

Let $\widehat{\beta}$ be the estimated behavior policy, and let $\widehat{w}(a,x) := \frac{\pi(a|x)}{\widehat{\beta}(a|x)}$ be the importance weight with estimated policy. Then the importance sampling (IS) estimator is given by

$$\widetilde{F}_{\text{IS}}(t) := \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(a_i, x_i)\mathbb{1}_{\{r_i \le t\}} \tag{21}$$

Then doubly robust (DR) estimator is:

$$\widetilde{F}_{\text{DR}}(t) := \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(a_i, x_i)\left(\mathbb{1}_{\{r_i \le t\}} - \overline{G}(t; x_i, a_i)\right) + \overline{G}(t; x_i, \pi) \tag{22}$$

And the direct method (DI) estimator is still defined to be

$$\widehat{F}_{\text{DI}}(t) := \frac{1}{n}\sum_{i=1}^{n} \overline{G}(t; x_i, \pi) \tag{23}$$

Note that the direct estimator does not depend on the behavior policy, and thus we continue to call it $\widetilde{F}_{\text{DI}}$.

E.1 Bias and Variance

Next, we analyze the bias and variance of these estimators. Define $\Delta(a,x,t)$ to be the additive error between G and the model $\overline{G}$, and define $\delta(x,a)$ to be the multiplicative error of the estimate $\widehat{\beta}$, that is:

$$\Delta(t; x, a) := \overline{G}(t; x, a) - G(t; x, a),$$
$$\delta(x, a) := 1 - \beta(a|x)/\widehat{\beta}(a|x).$$

Note that when β is known or $\widehat{\beta} = \beta$ for all x, a, $\delta(x,a) = 0$, The bias of the IS estimator then given in Lemma E.1, in terms of δ and the conditional reward distribution G.

Lemma E.1 (Bias and Variance of IS Estimator with $\widehat{\beta}$.)**.** *The expectation of the IS estimator is*

$$\mathbb{E}_{\mathbb{P}_\beta}[\widetilde{F}_{IS}(t)] = F(t) + \mathbb{E}_{\mathbb{P}}[\delta(A,X)G(t;X,\pi)]$$

When $\widehat{\beta}(a|x) = \beta(a|x)$ for all a, x, the IS estimator is unbiased and $\mathbb{E}_{\mathbb{P}_\beta}[\widehat{F}_{IS}(t)] = F(t)$. Further, the variance is

$$\begin{aligned}\mathbb{V}_{\mathbb{P}_\beta}[\widetilde{F}_{IS}(t)] &= \frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[(1-\delta(A,X))^2\,\sigma^2(t;X,A)\right] + \frac{1}{n}\mathbb{V}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[(1-\delta(A,X))G(t;X,A)|X\right]\right] \\ &\quad + \frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{w}(A,X)G(t;X,A)|X\right]\right]\end{aligned} \tag{24}$$

The expression for variance is broken down into three terms. The first represents randomness in the rewards, and the second represents variance from the aleatoric uncertainty due to randomness over contexts X. The final term represents variance arising from using importance sampling, and is proportional to the true CDF of conditional rewards G.

The following lemma, similarly, derives the bias and variance for the DR estimator:

Lemma E.2 (Bias and Variance of DR Estimator with $\widehat{\beta}$.)**.** *The pointwise expectation of the DR estimator is*

$$\mathbb{E}_{\mathbb{P}_\beta}[\widetilde{F}_{DR}(t)] = F(t) + \mathbb{E}_{\mathbb{P}}[\delta(X, A)\Delta(t; X, A)]$$

Further, when there is perfect knowledge of the behavior policy β, e.g. $\hat{\beta}(a|x) = \beta(a|x)$ for all a, x, the DR estimator is unbiased and

$$\mathbb{E}_{\mathbb{P}_\beta}[\widetilde{F}_{DR}(t)] = F(t)$$

The variance of the doubly robust estimator is given by

$$\begin{aligned}\mathbb{V}_{\mathbb{P}_\beta}[\widetilde{F}_{DR}(t)] = \frac{1}{n}\mathbb{E}_{\mathbb{P}}\Big[(1-\delta(A,X))^2\,\sigma^2(t;X,A)\Big] + \frac{1}{n}\mathbb{V}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[\delta(A,X)\Delta(t;X,A) + G(t;X,A)|X\right]\right] \\ + \frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{w}(A,X)\Delta(t;X,A)|X\right]\right] \qquad (25)\end{aligned}$$

Because the DR estimator takes advantage of both policy and reward estimates, it is unbiased whenever either the estimated policy or estimated reward is unbiased. Further, when we have access to the true behavior policy β and $\widehat{w} = w$, it retains the unbiasedness of the IS estimator.

Compared to the IS estimator, the DR estimator may also have pointwise reduced variance. When the variances of the IS estimator (24) and the DR estimator (25) are compared, the first term is identical, and the middle term is of similar magnitude because the randomness in contexts X is endemic. The third term is the primary difference. For the IS estimator, it is proportional to G, but for the DR estimator, it is proportional to the error Δ between the estimated conditional CDF $\overline{G}$ and the true G. Thus, this term can be much larger in the IS estimator when $\widehat{w}$ is large and the error Δ is smaller than G. This demonstrates that the DR estimator retains the low bias of the IS estimator, but has the advantage of reduced variance.

Next, Lemma E.3 gives the bias and variance of the DI estimator, which is directly related to the bias and variance of the conditional distribution model $\overline{G}$.

Lemma E.3 (Bias of DI Estimator with $\widehat{\beta}$.)**.** *The bias is*

$$\mathbb{E}_{x,a\sim\beta,r}[\widetilde{F}_{DI}(t)] = F(t) + \mathbb{E}_{\mathbb{P}}[\Delta]$$

and the variance is

$$\mathbb{V}[\widetilde{F}_{DI}(t)] = \frac{1}{n}\mathbb{V}_{\mathbb{P}}\Big[G(t;X,\pi) + \Delta\Big].$$

While the DI estimator has lower variance than both the IS and DR estimators, it suffers from potentially high bias from $\overline{G}$. Unlike the other two estimators, it is biased even when $\widehat{\beta}$ is a perfect estimate of β, which in practice is undesirable. Though the DI estimator has low bias when $\overline{G}$ is a good model of the condition reward distribution, it is often much easier to form accurate models of β than of G.

Proofs: Bias and Variance

We begin by proving the bias and variance expressions of the DR estimator with $\widehat{\beta}$. The bias and variance of the other estimators can be derived as special cases, which we show later.

Proof of Lemma E.2. First, we take the expectation of the DR estimator (22) with respect to $\mathbb{P}_\beta$:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_\beta}\Big[\widetilde{F}(t)\Big] &= \mathbb{E}_{\mathbb{P}_\beta}\Big[\frac{\pi(A|X)}{\widehat{\beta}(A|X)}\mathbb{1}\{R \le t\}\Big] + \mathbb{E}_{\mathbb{P}_\beta}\Big[\Big(\frac{\pi(A|X)}{\widehat{\beta}(A|X)} - \sum_a \pi(A|X)\Big)\overline{G}(t;X,A)\Big] \\ &= \mathbb{E}_{\mathbb{P}}\Big[\frac{\beta(A|X)}{\widehat{\beta}(A|X)}\mathbb{1}\{R \le t\}\Big] + \mathbb{E}_{\mathbb{P}}\Big[\Big(\frac{\beta(A|X)}{\widehat{\beta}(A|X)} - 1\Big)\overline{G}(t;X,A)\Big] \\ &= F(t) + \mathbb{E}_{\mathbb{P}}\Big[\Big(\frac{\beta(A|X)}{\widehat{\beta}(A|X)} - 1\Big)\mathbb{1}\{R \le t\}\Big] + \mathbb{E}_{\mathbb{P}}\Big[\Big(\frac{\beta(A|X)}{\widehat{\beta}(A|X)} - 1\Big)\overline{G}(t;X,A)\Big] \\ &= F(t) + \mathbb{E}_{\mathbb{P}}\Big[\Big(\frac{\beta(A|X)}{\widehat{\beta}(A|X)} - 1\Big)\Big(G(t;X,A) - \overline{G}(t;X,A)\Big)\Big] \\ &= F(t) + \mathbb{E}_{\mathbb{P}}\Big[\delta(A,X)\Delta(t;X,A)\Big]\end{aligned}$$

When $\widehat{\beta} = \beta$ for all a, x, we have $\delta = 0$, giving the unbiasedness of the estimator.

Starting from the second line of the proof of variance for the DR estimator (Appendix C.1), we have

$$\begin{aligned}
\mathbb{V}_{\mathbb{P}_\beta}\left[\widetilde{F}_{\text{DR}}(t)\right] &= \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{w}(A,X)\left(\mathbb{1}_{\{R\le t\}} - \overline{G}(t;X,A)\right) + \overline{G}(t;X,\pi)\right] \\
&= \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[\widehat{w}(A,X)^2\sigma^2(t;X,A)\right] \\
&\quad + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{w}(A,X)\left(G(t;X,A) - \overline{G}(t;X,A)\right) + \overline{G}(t;X,\pi)\right] \\
&= \frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[\left(\frac{\beta(A,X)}{\widehat{\beta}(A,X)}\right)^2 \sigma^2(t;X,A)\right] \\
&\quad + \frac{1}{n}\mathbb{V}_{\mathbb{P}_\beta}\left[\mathbb{E}_{\mathbb{P}_\beta}\left[\widehat{w}(A,X)\left(G(t;X,A) - \overline{G}(t;X,A)\right) + \overline{G}(t;X,\pi)|X\right]\right] \\
&\quad + \frac{1}{n}\mathbb{E}_{\mathbb{P}_\beta}\left[\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{w}(A,X)\left(G(t;X,A) - \overline{G}(t;X,A)\right)|X\right]\right] \\
&= \frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[\left(1 - \delta(A,X)\right)^2 \sigma^2(t;X,A)\right] \\
&\quad + \frac{1}{n}\mathbb{V}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[\delta(A,X)\Delta(t;X,A) + G(t;X,A)|X\right]\right] \\
&\quad + \frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[\mathbb{V}_{\mathbb{P}_\beta}\left[\widehat{w}(A,X)\Delta(t;X,A)|X\right]\right]
\end{aligned}$$

the second line uses a change of measure in the first term, and the law of total variance conditioned on the context X. The third line follows again from change of measure and substituting in the definition of δ and Δ.

□

Lemma E.1 is derived from Lemma E.2 using the fact that the IS estimator is a special case of the DR estimator with $\overline{G} = 0$.

Lemma E.3 is derived from Lemma E.2 by using $\widehat{\beta} \to \infty$ which means $\widehat{w} = 0$, e.g. importance weighting is not used, and $\delta = 1$.

E.2 CDF and Risk Estimate Error Bounds

Theorem E.1 generalizes the CDF error bounds established for the IS and DR estimators with known behavior policy to the case where $\widehat{\beta}$ is estimated, given an additional high-probability guarantee on the quality of $\widehat{\beta}$.

Theorem E.1. *For the IS or DR CDF estimator $\widetilde{F}$ that uses estimated weights $\widehat{w}(a,x) = \pi(a|x)/\widehat{\beta}(a,x)$, given an estimate $\widehat{\beta}$ that is ϵ_β-close to the true behavior policy β, that is*

$$\sup_{a,x} |\beta(a|x) - \widehat{\beta}(a|x)| \le \epsilon_\beta,$$

we have with probability at least $1 - \delta$ that

$$\mathbb{P}_\beta\left(\sup_{t\in\mathbb{R}}\left|\widetilde{F}(t) - F(t)\right| \le \epsilon + c\epsilon_\beta\right) \ge 1 - \delta$$

where ϵ is either ϵ_{IS} or $\epsilon = \epsilon_{DR}$ depending the choice of $\widehat{F}$, and $c = w_{max}\left(\inf_{a,x}\widehat{\beta}(a|x)\right)^{-1}$.

Similarly, for L-Lipschitz risk functionals, the general error bound given in Theorem 6.1 can be extended to the case of $\widehat{\beta}$ by adding the additional error term from the policy estimation.

Corollary E.1. *For the IS or DR CDF estimator $\widetilde{F}$ that uses estimated weights $\widehat{w}(a,x) = \pi(a|x)/\widehat{\beta}(a,x)$, given an estimate $\widehat{\beta}$ that is ϵ_β-close to the true behavior policy β, we have with probability at least $1-\delta$ that*

$$\left|\rho(\widetilde{F}) - \rho(F)\right| \leq L\left(\epsilon + c\epsilon_\beta\right)$$

where $c = w_{max}\left(\inf_{a,x}\widehat{\beta}(a|x)\right)^{-1}$.

Note that the error contributed by policy estimation, $c\epsilon_\beta$, is primarily dependent upon two factors. First, the quality of $\widehat{\beta}$ estimation determines the magnitude of ϵ_β; a poor estimate naturally leads to a higher value of this constant. Second, c is a problem-dependent constant proportional to the maximum importance weight w_{max} and the minimum probability of the estimated behavior policy $\inf_{a,x}\widehat{\beta}(a|x)$. If $\inf_{a,x}\widehat{\beta}(a|x)$ is particularly small, the error bound is also large. This reflects the fact that CDF estimation can be difficult when the behavior policy places low probability in some area of the context and action space.

Remark E.1. *When actions and contexts are discrete, and $\widehat{\beta}$ is estimated using empirical averages, standard concentrations for the mean of a random variable can be used to determine ϵ_β. If $\widehat{\beta}$ is estimated using regression, depending on the estimator ϵ_β can also be determined from concentration inequalities.*

v2

Proofs: Error Bounds

The proof of these results is given below.

Proof of Theorem E.1. We can decompose the error $\widehat{F} - F$ as:

$$\begin{aligned}\sup_t |\widetilde{F}(t) - F(t)| &\leq \sup_t \left(|\widehat{F}(t) - F(t)| + |\widetilde{F}(t) - \widehat{F}(t)|\right) \\ &\leq \sup_t |\widehat{F}(t) - F(t)| + \sup_t |\widetilde{F}(t) - \widehat{F}(t)|\end{aligned}$$

Theorem 5.1 gives a bound for the first term, and the bound for the second term bound is given in Lemma E.4 for the IS estimator, and in Lemma E.5 for the DR estimator. ☐

Proof of Corollary E.1. This result follows directly from applying the general risk estimation error bound in Theorem 6.1 to the error from Theorem E.1. ☐

The intermediary lemmas are defined and proved below:

Lemma E.4. *Suppose that $|\widehat{\beta}(a|x) - \beta(a|x)| \leq \epsilon_\beta$ for all a, x with probability at least $1-\delta$. Then with probability at least $1-\delta$,*

$$\sup_t |\widetilde{F}_{IS}(t) - \widehat{F}_{IS}(t)| \leq c\epsilon_\beta$$

where $c = w_{max}\left(\inf_{a,x}\widehat{\beta}(a|x)\right)^{-1}$.

Proof. We can bound the LHS of the lemma statement as follows.

$$\begin{aligned}
\sup_t |\widetilde{F}_{\text{IS}}(t) - \widehat{F}_{\text{IS}}(t)| &= \sup_t \Big| \frac{1}{n} \sum_{i=1}^{n} (w(a_i, x_i) - \widehat{w}(a_i, x_i)) \mathbb{1}_{\{r_i \le t\}} \Big| \\
&= \sup_t \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{\pi(a_i|x_i)}{\beta(a_i|x_i)} - \frac{\pi(a_i|x_i)}{\widehat{\beta}(a_i|x_i)} \Big) \mathbb{1}_{\{r_i \le t\}} \Big| \\
&\le \frac{1}{n} \sum_{i=1}^{n} \Big| \frac{\pi(a_i|x_i)}{\beta(a_i|x_i)} - \frac{\pi(a_i|x_i)}{\widehat{\beta}(a_i|x_i)} \Big| \\
&\le w_{max} \frac{1}{n} \sum_{i=1}^{n} \left| 1 - \frac{\beta(a_i|x_i)}{\widehat{\beta}(a_i|x_i)} \right| \\
&= w_{max} \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\widehat{\beta}(a_i|x_i) - \beta(a_i|x_i)}{\widehat{\beta}(a_i|x_i)} \right| \\
&\le w_{max} \left(\inf_{a,x} \widehat{\beta}(a|x) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \left| \widehat{\beta}(a_i|x_i) - \beta(a_i|x_i) \right| \\
&\le w_{max} \left(\inf_{a,x} \widehat{\beta}(a|x) \right)^{-1} \epsilon_\beta
\end{aligned}$$

where the last line follows from using the assumption that $|\widehat{\beta}(a|x) - \beta(a|x)| \le \epsilon_\beta$ for all a, x. □

Lemma E.5. *Suppose that $|\widehat{\beta}(a|x) - \beta(a|x)| \le \epsilon_\beta$ for all a, x with probability at least $1 - \delta$. Then with probability at least $1 - \delta$,*

$$\sup_t |\widetilde{F}_{DR}(t) - \widehat{F}_{DR}(t)| \le c\epsilon_\beta$$

where $c = w_{max} \left(\inf_{a,x} \widehat{\beta}(a|x) \right)^{-1}$.

Proof. We can bound the LHS of the lemma statement as follows. Using the definitions of the DR estimators,

$$\begin{aligned}
\sup_t |\widetilde{F}_{\text{DR}}(t) - \widehat{F}_{\text{DR}}(t)| &= \sup_t \Big| \frac{1}{n} \sum_{i=1}^{n} (w(a_i, x_i) - \widehat{w}(a_i, x_i)) \left(\mathbb{1}_{\{r_i \le t\}} - \overline{G}(t; x_i, a_i) \right) \Big| \\
&= \sup_t \Big| \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{\pi(a_i|x_i)}{\beta(a_i|x_i)} - \frac{\pi(a_i|x_i)}{\widehat{\beta}(a_i|x_i)} \Big) \left(\mathbb{1}_{\{r_i \le t\}} - \overline{G}(t; x_i, a_i) \right) \Big| \\
&\le \frac{1}{n} \sum_{i=1}^{n} \Big| \frac{\pi(a_i|x_i)}{\beta(a_i|x_i)} - \frac{\pi(a_i|x_i)}{\widehat{\beta}(a_i|x_i)} \Big| \\
&\le w_{max} \frac{1}{n} \sum_{i=1}^{n} \left| 1 - \frac{\beta(a_i|x_i)}{\widehat{\beta}(a_i|x_i)} \right| \\
&= w_{max} \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\widehat{\beta}(a_i|x_i) - \beta(a_i|x_i)}{\widehat{\beta}(a_i|x_i)} \right| \\
&\le w_{max} \left(\inf_{a,x} \widehat{\beta}(a|x) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \left| \widehat{\beta}(a_i|x_i) - \beta(a_i|x_i) \right| \\
&\le w_{max} \left(\inf_{a,x} \widehat{\beta}(a|x) \right)^{-1} \epsilon_\beta
\end{aligned}$$

where the last line uses the assumption that $|\widehat{\beta}(a|x) - \beta(a|x)| \le \epsilon_\beta$ for all a, x. □

F Additional Experiments

Implementation Details. Following [20, 19, 57], we obtain our off-policy contextual bandit datasets by transforming classification datasets. The contexts are the provided features, and the actions correspond to the possible class labels. To obtain the evaluation policy π, we use the output probabilities of a trained logistic regression classifier [36]. The behavior policy is defined as $\beta = \alpha\pi + (1-\alpha)\pi_{\text{UNIF}}$, where π_{UNIF} is a uniform policy over the actions, for some $\alpha \in (0, 1]$. Each dataset is generated by drawing actions for each context according to the probabilities of β, and the deterministic reward is 1 if the action matches the ground truth label, and 0 otherwise.

We apply this process to the set of 9 UCI datasets [18] used in [20, 19, 57], which each have differing dimensions d, actions k, and sample size n. Models $\overline{G}$ must be constructed for the DM and DR estimators. As in [20], the dataset is divided into two splits, with each of the two splits used to estimate $\overline{G}$, which is then used with the other split to calculate the estimator. The two results are averaged to produce the final estimators. In order to estimate $\overline{G}$, we discretize the reward support into $t \in [0, 1]$, and train a logistic regression classifier [36] for each action a and each t, with regularization parameter $C = 1$ and tolerance 0.0001. The code to reproduce these experiments is provided in the supplementary. On a CPU, they take roughly half a day of compute in total.

Relationship With α. We plot the error over the range of α, which controls the mismatch between the behavioral policy β and the target policy π and is thus proportional to w_{max}, for the PageBlocks dataset (also in Figure 1). The CDF error is shown in Figure 5 and the mean squared error (MSE) for the mean, CVaR 0.5, and variance risk functionals are shown in Figure 6.

The DR estimator exhibits lower error than any other estimator, and significantly lower variance than the IS and WIS estimators, across the range of α. This is particularly obvious in the region where α is small, which is where importance weights can become larger and the IS-based estimators are prone to higher variance. Note that the $\text{CVaR}_{0.5}$ MSE is close to 0 for all estimators.

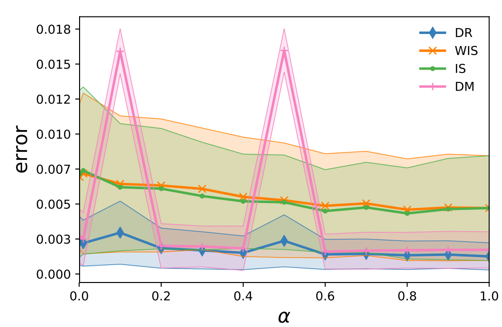


Figure 5: Sup-norm CDF error over α for PageBlocks. Shaded region shows one empirical standard deviation.

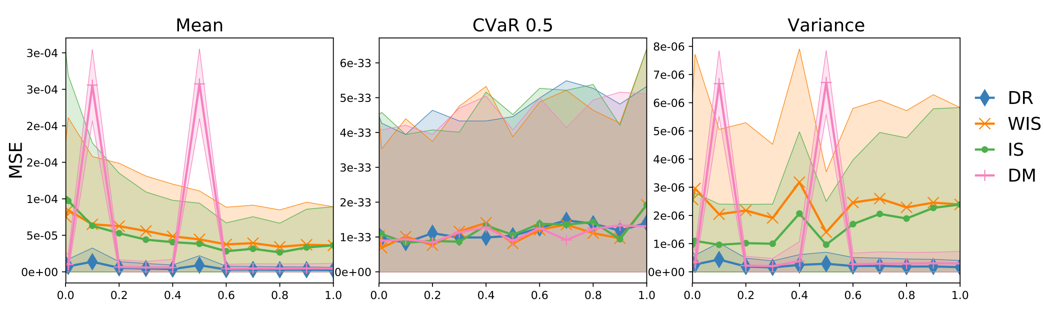
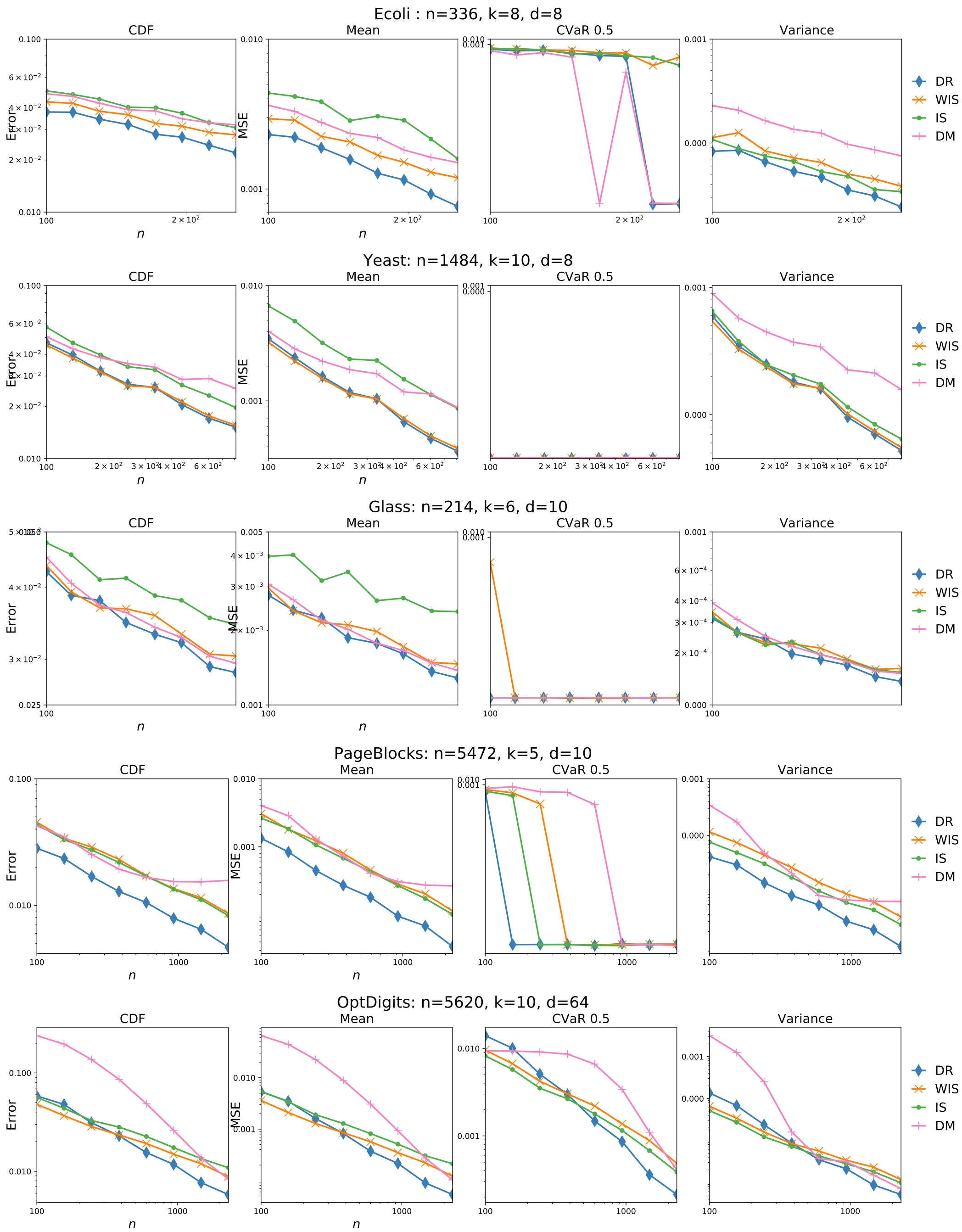
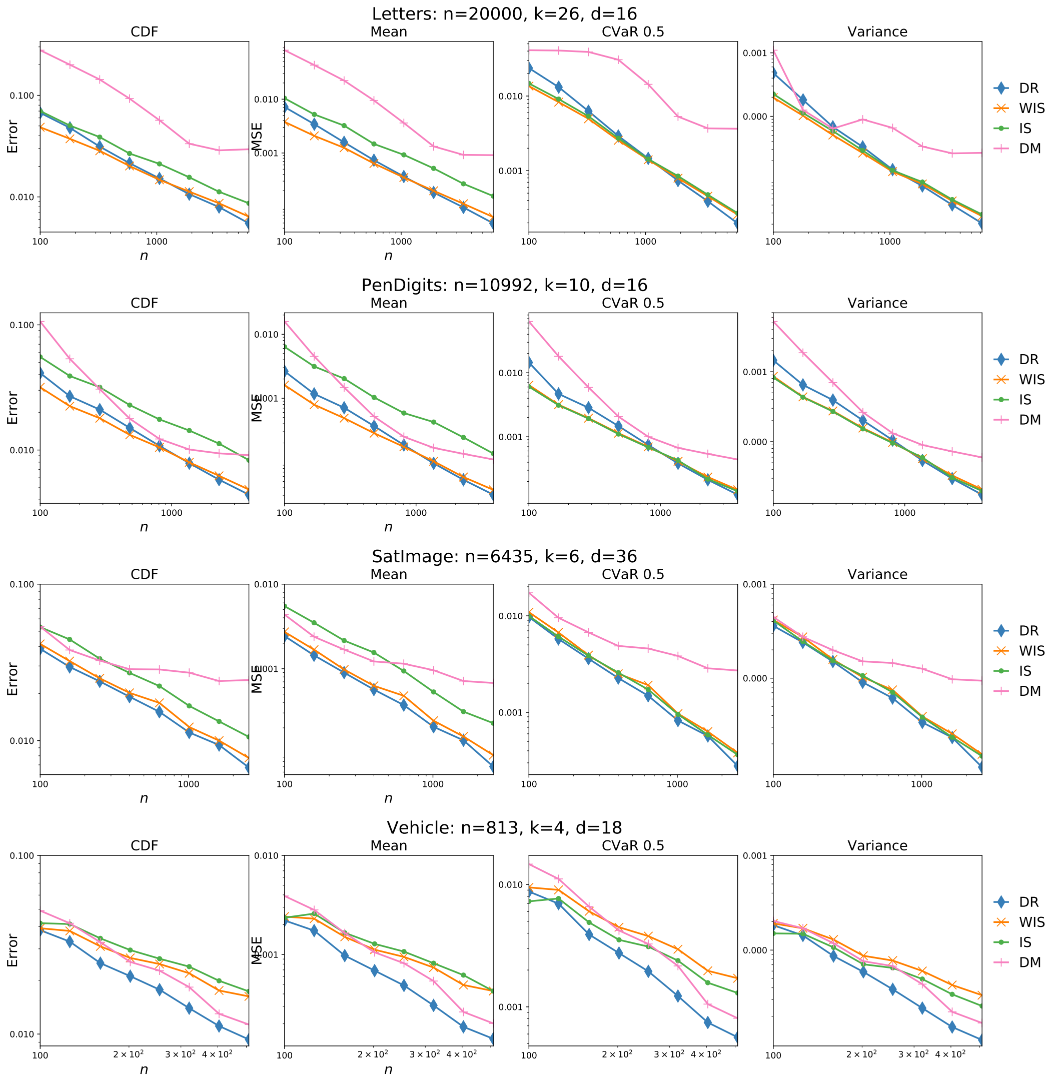


Figure 6: Mean squared error (MSE) over α for different risk functionals evaluated in the PageBlocks dataset. Shaded region shows one empirical standard deviation.

Evaluation Over UCI Datasets. We display the sup-norm error of the estimated CDF and the mean-squared error (MSE) of estimated risk functionals (mean, $\text{CVaR}_{0.5}$, and variance) for the 9 UCI datasets below. Here, $\alpha = 0.5$ is fixed. All plots are shown over 500 repetitions, with error bars omitted for readability but similar to those shown in Figure 1.

The general trends reflect analysis presented in Section 7. As expected of our distribution-based approach, trends in CDF estimation performance are reflected in risk estimation performance. Both the DR and IS estimators exhibit the expected $O(1/\sqrt{n})$ error convergence across the estimation tasks. Generally, the DR estimator does as well as if not better than the other estimators; where the model is difficult to specify well, the DR estimator may suffer slightly in performance in the low sample regime, but always outperforms the other estimators as the number of samples n increases.




Letters: n=20000, k=26, d=16
PenDigits: n=10992, k=10, d=16
SatImage: n=6435, k=6, d=36
Vehicle: n=813, k=4, d=18
CDF
Mean
CVaR 0.5
Variance
Error
MSE
n
DR
WIS
IS
DM