Behavior Policy Search for Risk Estimators in RL

Elita Lobo
University of New Hampshire
eal1063@usnh.edu

Yash Chandak
University of Massachusetts Amherst
ychandak@cs.umass.edu

Dharmashankar Subramanian
IBM Research
dharmash@us.ibm.com

Josiah Hannah
University of Wisconsin Madison
jphanna@cs.wisc.edu

Marek Petrik
University of New Hampshire
mpetrik@cs.unh.edu

Abstract

In real-world sequential decision problems, exploration is expensive, and the risk of expert decision policies must be evaluated from limited data. In this setting, Monte Carlo (MC) risk estimators are typically used to estimate the risk of decision policies. Unfortunately, while these estimators have the desired low bias property, they often suffer from large variance. In this paper, we consider the problem of minimizing the asymptotic mean squared error and hence variance of MC risk estimators. We show that by carefully choosing the data sampling policy (*behavior policy*), we can obtain low variance estimates of the risk of any given decision policy.

1 Introduction

Reinforcement Learning (RL) aims to find optimal decision policies for sequential decision problems like portfolio management [6], marketing [9] and dynamic treatment regimes [8]. A crucial component of these algorithms is estimating the value of a given policy which is termed policy evaluation. Several methods [3, 5, 15, 16] have been proposed to evaluate the value of a given policy in online and offline settings. In online policy evaluation [19], the value of a policy is obtained by simulating the policy several times and computing the average returns of the policy. Whereas, in an offline setting [5, 17], the value of a policy must be estimated from logged data sampled using a different policy. In this case, the sampling policy is known as the *behavior policy*, and the policy to be evaluated is known as the *evaluation policy*. To account for the difference between the behavior and evaluation policies, importance sampling weights are used to reweigh each trajectory sampled from the behavior policy by its likelihood of being observed under the evaluation policy [5, 17]. It is also essential to estimate the risks of a given policy in many high stake domains. We can compute these risks from the Monte-Carlo simulations by applying a risk-metric to the distribution of observed returns [14]. However, in such domains, exploration is often expensive and limited. Monte-Carlo (MC) estimates of risks can have a high variance when high-risk events are insufficiently sampled. The high variance in risk estimates can result in severely underestimating or overestimating risks and deploying bad policies.

This paper aims to minimize the variance of online Monte-Carlo-based risk-estimators using a hybrid of online and offline policy evaluation methods. To achieve our goal, we ask: *Is there a behavior policy better than the evaluation policy such that it results in minimum variance risk-estimates of the given policy?* To answer this question, we must first understand that when the set of trajectories

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

used for evaluation is finite and limited, the evaluation policy may not be a good representative of the high-risk trajectories that occur with very small probabilities. In such cases, the optimal behavior policy is the one that samples high-risk trajectories with greater likelihood.

Several prior works [3, 5, 7, 15, 17] have used *importance sampling* as a variance reduction tool in policy evaluation. Hanna et al. [4] propose a framework for computing the optimal behavior policy that minimizes the mean square error and hence variance of an unbiased MC value estimate of a policy. However, minimizing mean square error (MSE) via behavior policy search for risk-estimators is difficult because most MC risk-estimators are usually biased. Additionally, MC estimators of some of the popular risk-measures like CVaR and VaR are highly sample-inefficient as they are computed from only a fraction of the sampled trajectories [11]. Hence, they suffer from very large variance. This work focuses on minimizing the asymptotic MSE and hence asymptotic variance of two popular Monte-Carlo-based risk-estimators, Conditional Value at Risk and Value at Risk estimator.

As the paper's main contribution, we formulate the problem of optimal behavior policy search for minimization of asymptotic mean square error of risk estimators in RL. Building on prior work [4], we derive policy gradient theorems that optimize the variance of CVaR and VaR estimators. Finally, we demonstrate the effectiveness of our methods on several discrete and continuous domains.

2 Preliminaries

As the fundamental model, we assume a finite-horizon discounted Markov Decision Process (MDP) defined as tuple $(\mathcal{S}, \mathcal{A}, P, r, p_0, \gamma)$ comprising a set of states $\mathcal{S} = \{1, 2, \ldots, S\}$, a set of actions $\mathcal{A} = \{1, 2, \ldots, A\}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta^S$, an initial state distribution $p_0 \in \Delta^S$, and a discount rate $\gamma \in (0, 1)$. A general solution to an MDP is a *randomized* stationary policy $\pi : \mathcal{S} \to \Delta^A$, which prescribes the probability of taking each action $a \in \mathcal{A}$ in each state $s \in \mathcal{S}$. We denote by $\Pi = (\Delta^A)^{\mathcal{S}}$ and $\Pi_D = \mathcal{A}^{\mathcal{S}}$, the sets of all randomized and deterministic policies, respectively.
In the remainder of the paper, we denote the evaluation policy by π and the behavior policy by π_b. A trajectory $H = \{s_0, a_0, r_0, \ldots s_{T-1}, a_{T-1}, r_{T-1}\}$ is a tuple of states, actions and rewards observed on simulating policy π over T time steps, where T is the maximum length of an episode. The return corresponding to the trajectory H is given by the sum of discounted rewards $G^\pi(H) = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$. We use $G^\pi(H)$ to denote a random variable that represents the returns of a trajectory H obtained by simulating policy π. Finally, we denote by F_π the cumulative density function (CDF) of the distribution of returns of policy π.

Risk Measures. Risk measures are often used in RL to measure risk associated with a given decision policy [2]. Let X be a real-valued random variable with cumulative density function F. Let VaR and CVaR denote the Value at Risk measure (VaR) and Conditional Value at Risk (CVaR) measure of X respectively. Then, for all $0 \leq \alpha \leq 1$, VaR and CVaR are defined as

$$\mathrm{VaR}(F) = F^{-1}(\alpha) = \inf\{x \mid F(x) \geq \alpha\} \tag{1}$$

$$\mathrm{CVaR}(F) = \frac{1}{\alpha} \int_{-\infty}^{\mathrm{VaR}} x \, dF(x) = \mathrm{VaR}(F) - \frac{1}{\alpha} \mathbb{E} \left[\mathrm{VaR} - X\right]_+ \tag{2}$$

where $[x]_+ = \max(x, 0)$. Intuitively, VaR equals the α-quantile of X and CVaR equals the mean of the values of X that are smaller than α-quantile. We measure the risk of a given policy π by considering the returns of the policy $G(H)$, with $H \sim \pi$, as our random variable of interest.

Monte-Carlo Estimation of Risk Estimators. To estimate the MC estimates of CVaR and VaR of returns of a policy π, we need to first estimate the MC estimate of the cumulative density function (CDF) of returns of policy π, denoted by F^π. Let $H_1, \ldots H_N \sim \pi_b$ represent trajectories obtained by simulating behavior policy π_b. Then, an unbiased and consistent estimator for F^π is given by [1]

$$\hat{F}_n(\nu) = \frac{1}{N} \sum_{i=1}^{N} \frac{\rho^\pi(H_i)}{\rho^{\pi_b}(H_i)} \mathbf{1}\{G(H_i) \leq \nu\}, \qquad \forall \nu \in \mathbb{R}, \tag{3}$$

where $\rho^\pi(H_i) = \prod_{t=0}^{T} \pi(s_t^i, a_t^i)$ represents the importance sampling weight for the i^{th} trajectory.

Let $v^{\pi,\alpha}$ and $c^{\pi,\alpha}$ represent the $\mathrm{VaR}(F^{\pi})$ and $\mathrm{CVaR}(F^{\pi})$ of returns of policy π. Then, one can compute the MC estimates of $v^{\pi,\alpha}$ and $c^{\pi,\alpha}$ using the estimators $\hat{v}_n^{\pi,\alpha}$ and $\hat{c}_n^{\pi,\alpha}$, which are defined as, as [13]

$$\begin{aligned}\hat{v}_n^{\pi,\alpha} &= \hat{F}_n^{-1}(\alpha) := \inf\left\{g \in (G(H_i))_{i=1}^{N} \mid \hat{F}_n(G(H_i)) \geq \alpha\right\} \\ \hat{c}_n^{\pi,\alpha} &= \hat{v}_n^{\pi,\alpha} - \frac{1}{n\alpha}\sum_{i=1}^{n}[\hat{v}_n^{\pi,\alpha} - G(H_i)]_+\end{aligned} \tag{4}$$

Unfortunately, the MC estimates $\hat{c}_n^{\pi,\alpha}$ and $\hat{v}_n^{\pi,\alpha}$ are sample-inefficient and suffer from large variance because they are based on order-statistics [10]. In the next section, we improve the sample efficiency of the risk estimators by adapting the behavior policy π_b to minimize the variance of $\hat{c}_n^{\pi,\alpha}$ and $\hat{v}_n^{\pi,\alpha}$.

3 Method

In this section, we propose a new method to minimize the *asymptotic* MSE of the risk estimators in (4). We minimize the asymptotic MSE because the asymptotic risk estimators are unbiased and, therefore, their MSE can be minimized efficiently. In contrast, the true MSE is difficult to optimize because finite-sample risk estimators are biased.

Let $\pi_{b,\theta}$ denote a behavior policy parameterized by $\theta \in \Theta$ where $\Theta \in \mathbb{R}^p$ is a class of parameters that can sufficiently represent any policy $\pi \in \Pi$. We use $\hat{\Psi}(\pi_{b,\theta})$ and $\Psi(\pi_{b,\theta})$ to represent the Monte-Carlo risk estimator and the target risk estimand as a function of the behavior policy $\pi_{b,\theta}$. In this setting, $\Psi(\pi_{b,\theta})$ is either $v^{\pi,\alpha}(\theta)$ or $c^{\pi,\alpha}(\theta)$. Our goal is to find the optimal behavior policy π_{b,θ^*} that minimizes the asymptotic MSE of $\hat{\Psi}(\pi_{b,\theta})$, that is,

$$\theta^* = \underset{\theta\in\Theta}{\operatorname{argmin}}\, MSE[\hat{\Psi}(\pi_{b,\theta})] = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathrm{Var}[\hat{\Psi}(\pi_{b,\theta})] + Bias[\hat{\Psi}(\pi_{b,\theta})]^2 \tag{5}$$

where $\mathrm{Var}[\hat{\Psi}(\pi_{b,\theta})]$ represents the variance of $\hat{\Psi}(\pi_{b,\theta})$. It is known that $v^{\pi,\alpha}$ and $c^{\pi,\alpha}$ are asymptotically unbiased [13]. Hence, we can write the objective in (5) as

$$\theta^* \in \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathrm{Var}[\hat{\Psi}(\pi_{b,\theta})] \tag{6}$$

Next, we require the following assumptions to derive the asymptotic variance of $\hat{c}_n^{\pi,\alpha}$ and $\hat{v}_n^{\pi,\alpha}$.

Assumption 1 *There exists an ϵ such that $\forall\, G(H) \in (v^{\pi,\alpha} - \epsilon, v^{\pi,\alpha} + \epsilon)$, $f^{\pi}(G(H)) \geq 0$ and $f^{\pi}(G(H))$ is differentiable. Further, there exist a constant $c \geq 0$ such that $\forall g \in \{G(H) : G(H) \leq v^{\pi,\alpha} + \epsilon\}$, $g \leq c$.*

Assumption 1 establishes that $G(H)$ has a non-zero density around $v^{\pi,\alpha}$ implying continuity around $v^{\pi,\alpha}$. It also requires the likelihood ratio to be bounded for all trajectories with returns in the α-tail of the distribution.

Proposition 1 (Vandervaart et al. [18]) *Given an estimator $\hat{\Psi}$ of a target estimand Ψ, the efficient influence function of the estimator given by $\hat{\Phi}$ is the L_2 norm of the gradient of Ψ with respect to the input instances. Then, the asymptotic form of Ψ is given by $\sqrt{N}(\hat{\Psi} - \Psi) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(\hat{\Phi}))$*

Proposition 1 shows that the asymptotic variance of an estimator can be easily computed using the efficient influence-function of the estimator.

Proposition 2 *Suppose that Assumption 1 is satisfied. Then,*

$$\begin{aligned}\sqrt{n}(\hat{v}_n^{\pi,\alpha} - v^{\pi,\alpha}) &\xrightarrow{d} \frac{1}{f^{\pi}(v^{\pi,\alpha})}\sqrt{\mathrm{Var}\left[\mathbf{1}\{\{G(H_i) \leq v^{\pi,\alpha}\}\}\frac{\rho^{\pi}(H_i)}{\rho^{\pi_b}(H_i)}\right]} \cdot N(0,1) \\ \sqrt{n}(\hat{c}_n^{\pi,\alpha} - c^{\pi,\alpha}) &\xrightarrow{d} \frac{1}{\alpha}\sqrt{\mathrm{Var}\left[[v_\alpha - G_i]_+\frac{\rho^{\pi}(H_i)}{\rho^{\pi_b}(H_i)}\right]} \cdot N(0,1)\,.\end{aligned}$$

where $\forall\alpha \in [0,1]$, $\hat{v}_n^{\pi,\alpha} \to v^{\pi,\alpha}$ and $\hat{c}_n^{\pi,\alpha} \to c^{\pi,\alpha}$ with probability 1.0 as $n \to \infty$. The proof for Proposition 2 follows directly from applying Proposition 1 to the risk estimators $\hat{v}_n^{\pi,\alpha}$ and $\hat{c}_n^{\pi,\alpha}$.

Notice that the asymptotic variance of $\hat{v}_n^{\pi,\alpha}$ and $\hat{c}_n^{\pi,\alpha}$ in Proposition 2 depends on two main factors: a) the likelihood ratio of risky trajectories, i.e., trajectories with returns that fall in the lower-tail of the returns distribution, and b) the magnitude of the returns of the risky trajectories. Thus, an optimal behavior policy will assign a higher probability to those actions, which increases the likelihood of observing risky trajectories.

Our approach to improving the behavioral policy is based on gradient descent. The following proposition derives the gradient of the asymptotic variance of the two risk estimators.

Proposition 3 (Policy Gradient) *The gradients of* $\mathrm{Var}(\hat{v}_n^{\pi,\alpha}(\theta))$, $\mathrm{Var}(\hat{c}_n^{\pi,\alpha}(\theta))$ *with respect to* θ *are*

$$\nabla_\theta \operatorname{Var}(\hat{v}_n^{\pi,\alpha}(\theta)) = \mathbb{E}_{H\sim\pi_{b,\theta}}\left[\frac{1}{f^\pi(v^{\pi,\alpha})^2}\left(\mathbf{1}\left\{\left\{(v^{\pi,\alpha}-G_i)_+\frac{\rho^\pi(H_i)}{\rho^{\pi_b}(H_i)}\right\}\right\}\right)^2\sum_{t=0}^{T}\nabla_\theta\log(\pi_{b,\theta}(a_t|s_t))\right],$$

$$\nabla_\theta \operatorname{Var}[\hat{c}_n^{\pi,\alpha}(\theta)] = \mathbb{E}_{H\sim\pi_{b,\theta}}\left[\frac{1}{\alpha^2}\left((v^{\pi,\alpha}-G_i)_+\frac{\rho^\pi(H_i)}{\rho^{\pi_b}(H_i)}\right)^2\sum_{t=0}^{T}\nabla_\theta\log(\pi_{b,\theta}(a_t|s_t))\right],$$

Note that optimizing the asymptotic variance of $\hat{c}_n^{\pi,\alpha}$ requires an estimate of $v^{\pi,\alpha}$. Therefore, we recommend using an upper bound on $v^{\pi,\alpha}$ that we can compute from trajectories sampled from π. For more details on how to obtain the upper bound on $v^{\pi,\alpha}$, please refer to Theorem 3 in [1].

We can now solve the optimization problem in (6) using stochastic gradient descent (SGD) method [12].

Algorithm 1: Behavior Policy Search for $\hat{c}_n^{\pi,\alpha}$

```
**Input:** Evaluation policy parameters $\theta_e$, batch size $n$, step size $\eta$, number of iterations K,
        number of evaluation trajectories $M$
**Output:** Final behavior policy $\pi_{b,\theta_k}$, MC estimate of $\psi$
*Initialize:* $\theta_0 \leftarrow \theta_e$, $D_{10} \leftarrow []$ ;
Sample $H_1, H_2 \dots H_M \sim \pi_{\theta_e}$ and estimate an upper-bound on $v^{\pi,\alpha}$ using Theorem 3 in [1];
**for** $i = 0, \dots, k-1$ **do**
    $B_i \leftarrow$ Set of $M$ trajectories $H_1, H_2 \dots H_n \sim \pi_{b,\theta_i}$;
    $D_{i+1} = D_i \cup \{B_i, \theta_i\}$
    $\theta_{i+1} = \theta_i + \eta\frac{d\Psi(\pi_{b,\theta_i}, B_i)}{d\theta}$ ;
**end**
Compute $\hat{c}_n^{\pi,\alpha}$ using $D_K$;
**return** $\hat{c}_n^{\pi,\alpha}$
```

Our behavior policy search algorithm for risk measures (BPS-R) given in Algorithm 1 proceeds as follows. The algorithm takes as input the number of iterations K, the number of trajectories per iteration n, the number of trajectories to compute upper-bound on $v^{\pi,\alpha}$ given by M, and the evaluation policy parameters θ_e. We initialize the behavior policy with the policy to be evaluated i.e. $\pi_{b,\theta} = \pi_{\theta_e}$. In the first iteration of the algorithm, we sample M trajectories by simulation policy π and derive an upper-bound on $v^{\pi,\alpha}$ using Theorem 3 in [1]. In each of the subsequent iterations of the algorithm k, we collect n trajectories using the current behavior policy $\pi_{b,\theta}^k$ and use it to adapt the behavior policy using the gradient updates in Proposition 3. After repeating this step K times, we obtain the estimates of $\Psi(\pi_{b,\theta_k})$ using M trajectories sampled from $\pi_{b,\theta}^K$.

4 Experiments

To provide an intuition of how our algorithm works, we evaluate it on a single-state MDP. This example shows that our algorithm BPS-R can successfully find the optimal behavior policy that minimizes variance.

The single-state MDP consist of 1 state and four actions where each action yields some reward and transitions the agent to the terminal state. Thus, the length of the horizon in this setting is always be 1. The rewards corresponding to the four actions a_1, a_2, a_3, a_4 are $-50, -1, 0.05, 0.1$ respectively. To ensure that our evaluation policy results in rare risky events, we chose an evaluation policy that assigns a small probability to actions with small rewards. That is, our evaluation policy selects actions a_0, a_1, a_2, a_3 with probabilities $0.003, 0.097, 0.12, 0.80$ respectively.

In each of our experiment, we execute the BPS-R algorithm by setting the number of trajectories per iteration $M = 5$, number of iterations $K = 20$, and confidence level $\alpha = 0.9$. In each iteration of BPS-R, we sample 30 sets of trajectories (5 trajectories per set) using both the behavior policy and the evaluation policy and store them in memory. After every five iterations, we compute two sets of estimates of $\hat{c}_n^{\pi,\alpha}$ using the sets of trajectories sampled from the evaluation policy and the intermediate behavior policies respectively. Finally, we use these sets of $\hat{c}_n^{\pi,\alpha}$ estimates to obtain the variance in the Monte-Carlo and BPS-R estimates of $\hat{c}_n^{\pi,\alpha}$. We repeat this procedure 6 times and evaluate the standard error in the variance estimates.

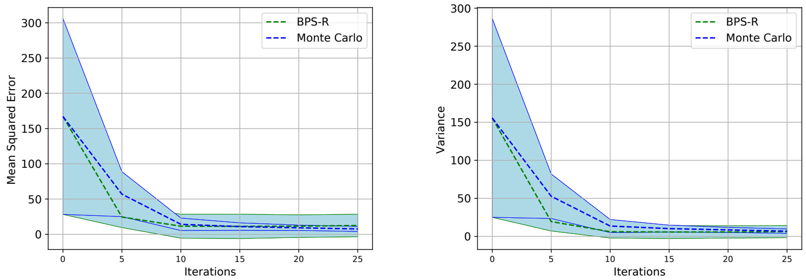
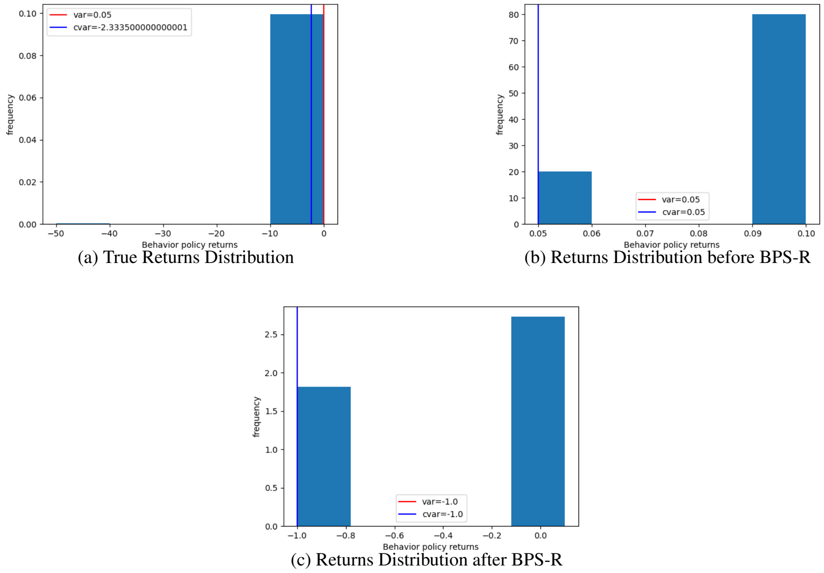


Figure 1: Mean Squared Error (left) and Variance (right) in estimates of $c_{0.9}^{\pi,\alpha}$ obtained using vanilla Monte-Carlo and BPS-R method.



(a) True Returns Distribution

(b) Returns Distribution before BPS-R

(c) Returns Distribution after BPS-R

Figure 2: Figure 2a shows the true distribution of returns policy π. Figure 2b shows the empirical distribution of returns of policy π computed from a set of 5 trajectories sampled using $\pi_b = \pi$. Figure 2c shows the empirical distribution of returns of policy π computed from a set of 5 trajectories sampled using the new behavior policy $\pi_{b,\theta_{15}}$. This example demonstrates that BPS-R correctly increases the likelihood of observing rare events.

Figure 1 and Figure 2 shows the mean squared error and the variance in the $\hat{c}_n^{\pi,\alpha}$ estimates obtained using vanilla Monte-Carlo and BPS-R algorithms. As evident in Figure 2c, BPS-R algorithm successfully finds a behavior policy that assigns higher probability to rare events with low returns and thus results in low variance estimates of $\hat{c}_n^{\pi,\alpha}$.

5 Discussion and Future Work

In this paper, we proposed a framework for finding the optimal behavior policy that results in low-variance Monte-Carlo estimates of CVAR and VAR of returns of a given evaluation policy. Although we observe promising results on the single-state domain, the algorithm can be brittle to the gradient estimates. One possible way of minimizing the brittleness could be reusing all sampled trajectories for gradient estimation via Multiple Importance Sampling (MIS). We leave the incorporation of MIS in BPS-R for future work.

References

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

[2] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *International Conference on Neural Information Processing Systems*, pages 3509–3517. MIT Press, 2014.

[3] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation, 2018.

[4] Josiah P. Hanna, Philip S. Thomas, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search, 2017.

[5] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661. PMLR, 2016.

[6] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. *ArXiv*, abs/1706.10059, 2017.

[7] Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[8] Ning Liu, Ying Liu, Brent Logan, Zhiyuan Xu, Jian Tang, and Yanzhi Wang. Deep reinforcement learning for dynamic treatment regimes on medical registry data, 2018.

[9] Marek Petrik and Ronny Luss. Interpretable policies for dynamic product recommendations. In *UAI*, 2016.

[10] R. TYRRELL Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[11] R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000.

[12] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[13] Lihua Sun and L. Jeff Hong. A general framework of importance sampling for value-at-risk and conditional value-at-risk. In *Winter Simulation Conference*, WSC '09, page 415–422. Winter Simulation Conference, 2009.

[14] Yichuan Charlie Tang, Jian Zhang, and Ruslan Salakhutdinov. Worst cases policy gradients. *Preprint arXiv:1911.03618*, 2019.

[15] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2380–2388, Lille, France, 07–09 Jul 2015. PMLR.

[16] Philip S. Thomas, Georgios Teocharous, and Mohammad Ghavamzadeh. High Confidence Off-Policy Evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[17] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence off-policy evaluation. AAAI'15, page 3000–3006. AAAI Press, 2015.

[18] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[19] Martha White and Michael Bowling. Learning a value analysis tool for agent evaluation. pages 1976–1981, 01 2009.

[20] Shengyu Zhang, R. Martin, and Anthony Christidis. Influence functions for risk and performance estimators. *Journal of Mathematical Finance*, 11:15–47, 01 2021.

A Proofs

A.1 Influence Function and Asymptotic Variance

Consider a risk-estimator $\hat{\theta}_n = \hat{\theta}(r_1, r_2, \dots r_n)$ where r_1, r_2, r_3, r_n are i.i.d with distribution $F = F(\xi)$. Next, we assume a mixed distribution constructed by slighting perturbing the distribution f.

$$F_\xi(x) = (1-\xi)F(x) + \xi\delta_r(x) \tag{7}$$

where $\delta_r(x)$ is a discrete distribution with all the probability mass concentrated on the value r. Then, the influence function of the estimator $\hat{\theta}_n$is defined as

$$IF(r; \hat{\theta}, F) = \left.\frac{d\hat{\theta}(F_\xi)}{d\xi}\right|_{\xi=0} \tag{8}$$

We can now compute the asymptotic variance of the estimator $\hat{\theta}_n$ as

$$\mathrm{Var}(\hat{\theta}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} IF(r_i, \hat{\theta}, F)\right) \tag{9}$$

We will use 9 to compute the asymptotic variance of $\hat{c}_n^{\pi,\alpha}$ and $\hat{v}_n^{\pi,\alpha}$.

A.2 Asymptotic Variance of Risk Estimators

With simple algebraic manipulations, we can show that the influence-function of $\mathrm{CVaR}(F)$ and $\mathrm{Var}(F)$ as

$$IF(r; \mathrm{CVaR}_\alpha, F) = -\mathrm{VaR}(F) - \mathrm{CVaR}(F) - \frac{(\mathrm{VaR}(F) - r)_+)}{\alpha} \tag{10}$$

$$IF(r; \mathrm{VaR}_\alpha, F) = \frac{\mathbf{1}_{\{r \le \mathrm{VaR}\}} - \alpha}{f(\mathrm{VaR})} \tag{11}$$

For detailed derivations of the influence functions of risk estimators, please refer [20].

A.3 Proof of Proposition 2

Next, we use 9 and 10, to derive the asymptotic variance of $\hat{c}_n^{\pi,\alpha}$. Substituting $r = G(H), H \sim \pi_b$ and $\hat{\theta}_n = \hat{c}_n^{\pi,\alpha}$ in 9, we get

$$\begin{aligned}\mathrm{Var}(\hat{c}_n^{\pi,\alpha}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{N}(-v^{\pi,\alpha} - c^{\pi,\alpha} - \frac{(v^{\pi,\alpha} - G(H)_i)_+}{\alpha}\frac{\rho^\pi(H_i)}{\rho^{\pi_b}(H_i)}\right) \\ &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{(v^{\pi,\alpha} - G(H)_i)_+}{\alpha}\frac{\rho^\pi(H_i)}{\rho^{\pi_b}(H_i)}\right)\end{aligned} \tag{12}$$

We can similarly derive the asymptotic variance of $\hat{v}_n^{\pi,\alpha}$ using 9 and 11. Substituting $r = G(H), H \sim \pi_b$ and $\hat{\theta}_n = \hat{v}_n^{\pi,\alpha}$ in 9, gives

$$\begin{aligned}\mathrm{Var}(\hat{v}_n^{\pi,\alpha}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{N}\frac{\mathbf{1}\{G(H_i) \le v^{\pi,\alpha}\} - \alpha}{f^\pi(v^{\pi,\alpha})}\right) \\ &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{1}\{v^{\pi,\alpha} - G(H_i)_+\}}{\alpha}\frac{\rho^\pi(H_i)}{\rho^{\pi_b}(H_i)}\right)\end{aligned} \tag{13}$$

Proposition 2 then directly follows from Central limit theorem, 12 and 13.

A.4 Proof of Proposition 3

Consider the variance of $\hat{c}_n^{\pi,\alpha}$.

$$\begin{align}
\mathrm{Var}(\hat{c}_n^{\pi,\alpha}) &= \mathrm{VaR}(\frac{1}{N}\sum_{i=1}^{N}\frac{(v^{\pi,\alpha}-G(H_i))_+}{\alpha}\frac{\rho^{\pi}(H_i)}{\rho^{\pi_b}(H_i)}) \tag{14}\\
&= \frac{1}{\alpha^2}\left(\mathbb{E}_{H\sim\pi_b}\left[(v^{\pi,\alpha}-G(H))_+^2\left(\frac{\rho^{\pi}(H)}{\rho^{\pi_b}(H)}\right)^2\right]-\mathbb{E}_{H\sim\pi}\left[(v^{\pi,\alpha}-G(H))_+^2\right]\right) \tag{15}\\
&= \left(\mathbb{E}_{H\sim\pi_b}\left[\frac{(v^{\pi,\alpha}-G(H))_+^2}{\alpha^2}\left(\frac{\rho^{\pi}(H)}{\rho^{\pi_b}(H)}\right)^2\right]\right)-(c^{\pi,\alpha})^2 \tag{16}
\end{align}$$

We can now compute the gradient $\nabla_\theta MSE(\hat{c}_n^{\pi,\alpha}(\theta))$ as

$$\begin{align}
\nabla_\theta MSE(\hat{c}_n^{\pi,\alpha}(\theta)) &= \nabla_\theta\left(\mathrm{Var}(\hat{c}_n^{\pi,\alpha}(\theta))+Bias(\hat{c}_n^{\pi,\alpha}(\theta))\right) \tag{17}\\
&= \nabla_\theta\,\mathrm{Var}(\hat{c}_n^{\pi,\alpha}(\theta))+0 \tag{18}\\
&= \nabla_\theta\left(\mathbb{E}_{H\sim\pi_{b,\theta}}\left[\frac{(v^{\pi,\alpha}-G(H))_+^2}{\alpha^2}\left(\frac{\rho^{\pi}(H)}{\rho^{\pi_{b,\theta}}(H)}\right)^2\right]\right) \tag{19}\\
& \tag{20}
\end{align}$$

Let $RISK(H,\theta)=\frac{(v^{\pi,\alpha}-G(H))_+}{\alpha}\left(\frac{\rho^{\pi}(H)}{\rho^{\pi_{b,\theta}}(H)}\right)$. Then,

$$\begin{align}
\nabla_\theta\mathbb{E}_{H\sim\pi_{b,\theta}}\left[RISK(H,\theta)^2\right] &= \nabla_\theta\sum_H Pr(H|\theta)RISK(H,\theta)^2 \tag{21}\\
&= \sum_H Pr(H|\theta)\nabla_\theta RISK(H,\theta)^2+RISK(H,\theta)^2\nabla_\theta Pr(H|\theta) \tag{22}
\end{align}$$

Substituting $Pr(H|\theta)=Pr(H)\rho^{\pi_{b,\theta}}(H)$, we get

$$\nabla_\theta\mathbb{E}_{H\sim\pi_{b,\theta}}\left[RISK(H,\theta)^2\right]=\sum_H\left(Pr(H|\theta)\nabla_\theta RISK(H,\theta)^2+RISK(H,\theta)^2Pr(H)\nabla_\theta\rho^{\pi_{b,\theta}}(H)\right) \tag{23}$$

Expanding $\nabla_\theta\rho^{\pi_{b,\theta}}(H)$, we get

$$\begin{aligned}
\nabla_\theta\rho^{\pi_{b,\theta}}(H) &= \nabla_\theta\prod_{t=0}^{T}\pi_{b,\theta}(a_t|s_t)\\
&= \left(\prod_{t=0}^{T}\pi_{b,\theta}(a_t|s_t)\right)\left(\sum_{t=0}^{T}\nabla_\theta\log(\pi_{b,\theta}(a_t|s_t))\right)
\end{aligned} \tag{24}$$

Combining 23 and 24, we get

$$\nabla_\theta MSE(\hat{c}_n^{\pi,\alpha}(\theta))=\mathbb{E}_{H\sim\pi_{b,\theta}}\left[RISK(H,\theta)^2\sum_{t=0}^{T}\nabla_\theta\log\pi_{b,\theta}(a_t|s_t)+\nabla_\theta RISK(H,\theta)^2\right] \tag{25}$$

Consider the gradient $\nabla_\theta RISK(H,\theta)^2$.

$$\begin{align}
\nabla_\theta RISK(H,\theta)^2 &= \nabla_\theta\left(\frac{(v^{\pi,\alpha}-G(H))_+}{\alpha}\left(\frac{\rho^{\pi}(H)}{\rho^{\pi_{b,\theta}}(H)}\right)\right)^2 \notag\\
&= 2\left(\frac{(v^{\pi,\alpha}-G(H))_+}{\alpha}\left(\frac{\rho^{\pi}(H)}{\rho^{\pi_{b,\theta}}(H)}\right)\right)\nabla_\theta\left(\frac{(v^{\pi,\alpha}-G(H))_+}{\alpha}\left(\frac{\rho^{\pi}(H)}{\rho^{\pi_{b,\theta}}(H)}\right)\right) \notag\\
&= -2\left(\frac{(v^{\pi,\alpha}-G(H))_+}{\alpha}\left(\frac{\rho^{\pi}(H)}{\rho^{\pi_{b,\theta}}(H)}\right)\right)^2\sum_{t=0}^{T}\nabla_\theta\log(\pi_{b,\theta}(a_t|s_t)) \tag{26}
\end{align}$$

Plugging 26 in 25 gives Proposition 3. The gradient $\nabla_\theta MSE(\hat{v}_n^{\pi,\alpha})$ can be similarly derived.