High-Confidence Off-Policy (or Counterfactual) Variance Estimation

Yash Chandak, Shiv Shankar, Philip S. Thomas
University of Massachusetts
{ychandak, sshankar, pthomas}@cs.umass.edu

Abstract

Many sequential decision-making systems leverage data collected using prior policies to propose a new policy. For critical applications, it is important that high-confidence guarantees on the new policy's behavior are provided before deployment, to ensure that the policy will behave as desired. Prior works have studied high-confidence off-policy estimation of the *expected* return, however, high-confidence off-policy estimation of the *variance* of returns can be equally critical for high-risk applications. In this paper we tackle the previously open problem of estimating and bounding, with high confidence, the variance of returns from off-policy data.

Introduction

Reinforcement learning (RL) has emerged as a promising method for solving sequential decision-making problems (Sutton and Barto 2018). Deploying RL to real-world applications, however, requires additional consideration of reliability, which has been relatively understudied. Specifically, it is often desirable to provide high-confidence guarantees on the behavior of a given policy, *before* deployment, to ensure that the policy will behave as desired.

Prior works in RL have studied the problem of providing high-confidence guarantees on the *expected* return of an evaluation policy, π, using only data collected from a currently deployed policy called the *behavior policy*, β (Thomas, Theocharous, and Ghavamzadeh 2015; Hanna, Stone, and Niekum 2017; Kuzborskij et al. 2020). Analogously, researchers have also studied the problem of *counter-factually* estimating and bounding the average treatment effect, with high confidence, using data from past treatments (Bottou et al. 2013). While these methods present important contributions towards developing practical algorithms, real-world problems may require additional consideration of the *variance* of returns (effect) under any new policy (treatment) before it can be deployed responsibly.

For applications that have high stakes in the terms of financial cost or public well-being, only providing guarantees on the mean outcome might *not* be sufficient. *Analysis of variance* (ANOVA) has therefore become a *de-facto* standard for many industrial and medical applications (Tabachnick and Fidell 2007). Similarly, analysis of variance can inform numerous real-world applications of RL. For example, (a) analysing the variance of outcomes in a robotics application (Kuindersma, Grupen, and Barto 2013), (b) ensuring that the variance of outcomes for a medical treatment is not high, (c) characterizing the variance of customer experiences for a recommendation system (Teevan et al. 2009), or (d) limiting the variability of the performance of an autonomous driving system (Montgomery 2007).

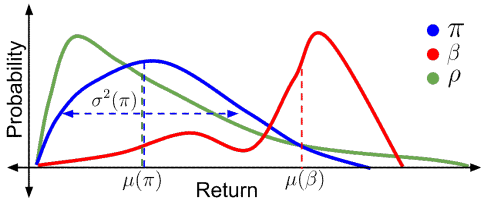

Figure 1: Illustrative example of the distributions of returns from a behavior policy β, and evaluation policy π, along with the *importance weighted returns* ρ, discussed later. Given trajectories from the behavior policy β, we aim to estimate and bound the variance, $\sigma^2(\pi)$, of returns under an evaluation policy π, with high confidence. Note that the distribution of *importance-weighted* returns ρ has the mean value $\mu(\pi)$, but might have variance not equal to $\sigma^2(\pi)$.

More generally, variance estimation can be used to account for risk in decision-making by designing objectives that maximize the mean of returns but minimize the variance of returns (Sato, Kimura, and Kobayashi 2001; Di Castro, Tamar, and Mannor 2012; La and Ghavamzadeh 2013). Variance estimates have also been shown to be useful for automatically adapting hyper-parameters, like the exploration rate (Sakaguchi and Takano 2004) or λ for eligibility-traces (White and White 2016), and might also inform other methods that depend on the entire distribution of returns (Bellemare, Dabney, and Munos 2017; Dabney et al. 2017).

Despite the wide applicability of variance analysis, estimating and bounding the variance of returns with high confidence, using only off-policy data, has remained an understudied problem. In this paper, we first formalize the problem statement; an illustration of which is provided in Figure 1. We show that the typical use of *importance sampling* (IS) in RL only corrects for the mean, and so

it does not directly provide unbiased off-policy estimates of variance. We then present an off-policy estimator of the variance of returns that uses IS twice, together with a simple double-sampling technique. To reduce the variance of the estimator, we extend the per-decision IS technique (Precup 2000) to off-policy variance estimation. Building upon this estimator, we provide confidence intervals for the variance using (a) concentration inequalities, and (b) statistical bootstrapping.

Advantages: The proposed variance estimator has several advantages: (a) it is a model-free estimator and can thus be used irrespective of the environment complexity, (b) it requires only off-policy data and can therefore be used *before* actual policy deployment, (c) it is unbiased and consistent. For high-confidence guarantees, (d) we provide both upper and lower confidence intervals for the variance that have guaranteed coverage (that is, they hold with any desired confidence level and without requiring false assumptions), and (e) we also provide bootstrap confidence intervals, which are approximate but often more practical.

Limitations: The proposed off-policy estimator of the variance relies upon IS and thus inherits its limitations. Namely, (a) it requires knowledge of the action probabilities from the behavior policy β, (b) it requires that the support of the trajectories under the evaluation policy π is a subset of the support under the behavior policy β, and (c) the variance of the estimator scales exponentially with the length of the trajectory (Guo, Thomas, and Brunskill 2017; Liu et al. 2018).

Background and Problem Statement

A *Markov decision process* (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, d_0)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{P}$ is the transition function, $\mathcal{R}$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and d_0 is the starting state distribution.[1] A policy π is a distribution over the actions conditioned on the state, i.e., $\pi(a|s)$ represents the probability of taking action a in state s. We assume that the MDP has finite horizon T, after which any action leads to an absorbing state $S_{(\infty)}$. In general, we will use subscripts with parentheses for the timestep and subscript without parentheses to indicate the episode number. Let $R_{i(j)} \in [R_{\texttt{min}}, R_{\texttt{max}}]$ represent the reward observed at timestep j of the episode i. Let the random variable $G_i \coloneqq \sum_{j=0}^{T} \gamma^j R_{i(j)}$ be the *return* for episode i. Let $c \coloneqq (1 - \gamma^T)/(1 - \gamma)$ so that the minimum and the maximum returns possible are $G_{\texttt{min}} \coloneqq cR_{\texttt{min}}$ and $G_{\texttt{max}} \coloneqq cR_{\texttt{max}}$, respectively. Let $\mu(\pi) \coloneqq \mathbb{E}_\pi[G]$ be the expected return, and $\sigma^2(\pi) \coloneqq \mathbb{V}_\pi[G]$ be the variance of returns, where the subscript π denotes that the trajectories are generated using policy π.

[1] We formulate the problem in terms of MDPs, but it can analogously be formulated in terms of *structural causal models.* (Pearl 2009). For simplicity, we consider finite states and actions, but our results extend to POMDPs (by replacing states with *observations*) and to continuous states and actions (by appropriately replacing summations with integrals), and to infinite horizons ($T \coloneqq \infty$).

Let $\mathcal{H}^{\pi}_{(i):(j)}$ be the set of all possible trajectories for a policy π, from timestep i to timestep j. Let H denote a complete trajectory: $(S_{(0)}, A_{(0)}, \Pr(A_{(0)}|S_{(0)}), R_{(0)}, S_{(1)}, ..., S_{(\infty)})$, where T is the horizon length, and $S_{(0)}$ is sampled from d_0. Let $\mathcal{D}$ be a set of n trajectories $\{H_i\}_{i=1}^n$ generated using *behavior* policies $\{\beta_i\}_{i=1}^n$, respectively. Let $\rho_i(0, T) \coloneqq \prod_{j=0}^{T} \frac{\pi(A_{i(j)}|S_{i(j)})}{\beta_i(A_{i(j)}|S_{i(j)})}$ denote the product of *importance ratios* from timestep 0 to T. For brevity, when the range of timesteps is not necessary, we write $\rho_i \coloneqq \rho_i(0, T)$. Similarly, when referring to ρ_i for an arbitrary $i \in \{1, \ldots, n\}$, we often write ρ. With this notation, we now formalize the *off-policy variance estimation* (OVE) and the *high-confidence off-policy variance estimation* (HCOVE) problems.

OVE Problem: Given a set of trajectories $\mathcal{D}$ and an *evaluation* policy π, we aim to find an estimator $\hat{\sigma}^2_n$ that is both an unbiased and consistent estimator of $\sigma^2(\pi)$, i.e.,

$$\mathbb{E}[\hat{\sigma}^2_n] = \sigma^2(\pi), \qquad \hat{\sigma}^2_n \xrightarrow{\text{a.s.}} \sigma^2(\pi).$$

HCOVE Problem: Given a set of trajectories $\mathcal{D}$, an *evaluation* policy π, and a confidence level $1 - \delta$, we aim to find a confidence interval $\mathcal{C} \coloneqq [v^{\texttt{lb}}, v^{\texttt{ub}}]$, such that

$$\Pr\left(\sigma^2(\pi) \in \mathcal{C}\right) \geq 1 - \delta.$$

Remark 1. *It is worth emphasizing that the OVE problem is about estimating the variance of returns, and not the variance of the estimator of the mean of returns.*

These problems would not be possible to solve if the trajectories in $\mathcal{D}$ are not informative about the trajectories that are possible under π. For example, if $\mathcal{D}$ has no trajectory that could be observed if policy π were to be executed, then $\mathcal{D}$ provides little or no information about the possible outcomes under π. To avoid this case, we make the following common assumption (Precup 2000), which is satisfied if $(\beta_i(a|s) = 0) \implies (\pi(a|s) = 0)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, and $i \in \{1, \ldots, n\}$.

Assumption 1. *The set $\mathcal{D}$ contains independent trajectories generated using behavior policies $\{\beta_i\}_{i=1}^n$, such that*

$$\forall i,\ \mathcal{H}^{\pi}_{(0):(T)} \subseteq \mathcal{H}^{\beta_i}_{(0):(T)}.$$

The methods that we derive, and IS methods in general, do not require complete knowledge of $\{\beta_i\}_{i=1}^n$ (which might be parameterized using deep neural networks and might be hard to store). Only the probabilities, $\beta_i(a|s)$, for states s and actions a present in $\mathcal{D}$ are required. For simplicity, we restrict our notation to a single behavior policy β, such that $\forall i,\ \beta_i = \beta$.

Naïve Methods

In the *on-policy* setting, computing an estimate of $\mu(\pi)$ or $\sigma^2(\pi)$ is trivial—sample n trajectories using π and compute the sample mean or variance of the observed returns, $\{G_i\}_{i=1}^n$. In the *off-policy* setting, under Assumption 1,

the sample mean $\hat{\mu} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \rho_i G_i$ of the *importance weighted returns* $\{\rho_i G_i\}_{i=1}^{n}$, is an unbiased estimator of $\mu(\pi)$ (Precup 2000), i.e., $\mathbb{E}_\beta[\hat{\mu}] = \mu(\pi)$. Similarly, one natural way to estimate $\sigma^2(\pi)$ in the off-policy setting might be to compute the sample variance (with Bessel's correction) of the importance sampled returns $\{\rho_i G_i\}_{i=1}^{n}$,

$$\hat{\sigma}_n^{2!!} \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}\left(\rho_i G_i - \frac{1}{n}\sum_{j=1}^{n}\rho_j G_j\right)^2. \quad (1)$$

Unfortunately, $\hat{\sigma}_n^{2!!}$ is neither an unbiased nor consistent estimator of $\sigma^2(\pi)$, in general, as shown in the following properties. These properties also reveal that $\rho_i G_i$ only corrects the distribution for the mean and not for the variance, as depicted in Figure 1. Also, note that all proofs are deferred to the appendix.

Property 1. *Under Assumption* 1, $\hat{\sigma}_n^{2!!}$ *may be a biased estimator of* $\sigma^2(\pi)$. *That is, it is possible that* $\mathbb{E}_\beta[\hat{\sigma}_n^{2!!}] \neq \sigma^2(\pi)$.

Property 2. *Under Assumption* 1, $\hat{\sigma}_n^{2!!}$ *may not be a consistent estimator of* $\sigma^2(\pi)$. *That is, it is not always the case that* $\hat{\sigma}_n^{2!!} \xrightarrow{a.s.} \sigma^2(\pi)$.

Since the on-policy variance is $\mathbb{E}_\pi[(G - \mathbb{E}_\pi[G])^2]$, a natural alternative might be to construct an estimator that corrects the off-policy distribution for both the mean and the variance. That is, using the equivalence

$$\mathbb{V}_\pi(G) = \mathbb{E}_\pi[(G - \mathbb{E}_\pi[G])^2] = \mathbb{E}_\beta[\rho(G - \mathbb{E}_\beta[\rho G])^2],$$

an alternative might be to use a plug-in estimator for $\mathbb{E}_\beta[\rho(G - \mathbb{E}_\beta[\rho G])^2]$ (with Bessel's correction) as,

$$\hat{\sigma}_n^{2!} \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}\rho_i\left(G_i - \frac{1}{n}\sum_{j=1}^{n}\rho_j G_j\right)^2. \quad (2)$$

While $\hat{\sigma}_n^{2!}$ turns out to be a consistent estimator, it is still not an unbiased estimator of $\sigma^2(\pi)$. We formalize this in the following properties.

Property 3. *Under Assumption* 1, $\hat{\sigma}_n^{2!}$ *may be a biased estimator of* $\sigma^2(\pi)$. *That is, it is possible that* $\mathbb{E}_\beta[\hat{\sigma}_n^{2!}] \neq \sigma^2(\pi)$.

Property 4. *Under Assumption* 1, $\hat{\sigma}_n^{2!}$ *is a consistent estimator of* $\sigma^2(\pi)$. *That is,* $\hat{\sigma}_n^{2!} \xrightarrow{a.s.} \sigma^2(\pi)$.

Before even considering confidence intervals for $\sigma^2(\pi)$, the lack of unbiased estimates from these naïve methods leads to a basic question: *How can we construct unbiased estimates of* $\sigma^2(\pi)$*?* We answer this question in the following section.

Off-Policy Variance Estimation

Before constructing an unbiased estimator for $\sigma^2(\pi)$, we first discuss one root cause for the bias of $\hat{\sigma}_n^{2!}$ and $\hat{\sigma}_n^{2!!}$. Notice that an expansion of (1) and (2) produces self-coupled importance ratio terms. That is, terms consisting of ρ_i^2 and ρ_i^3. While ρ_i helps in correcting the distribution, its higher powers, ρ_i^2 and ρ_i^3, do not.

Expansion of (1) and (2) also results in cross-coupled importance ratio terms, $\rho_i\rho_j$, where $i \neq j$. However, because $\mathbb{E}_\beta[\rho_i] = 1$ for all $i \in \{1, \ldots, n\}$ and because ρ_i and ρ_j are independent when $i \neq j$, these terms factor out in expectation. Hence, these terms do not create the troublesome higher powers of importance ratios.

Based on these observations, we create an estimator that does not have any self-coupled importance ratio terms like ρ_i^2, but which may have $\rho_i\rho_j$ terms, where $i \neq j$. To do so, we consider the alternate formulation of variance,

$$\mathbb{V}_\pi(G) = \mathbb{E}_\pi[G^2] - \mathbb{E}_\pi[G]^2 = \mathbb{E}_\beta[\rho G^2] - \mathbb{E}_\beta[\rho G]^2. \quad (3)$$

In (3), while a plug-in estimator of $\mathbb{E}_\beta[\rho G^2]$ would be unbiased and free of any self-coupled importance ratio terms, a plug-in estimator for $\mathbb{E}_\beta[\rho G]^2$ would neither be unbiased nor would it be free of ρ_i^2 terms. To remedy this problem, we explicitly split the set of sampled trajectories into two mutually exclusive sets, $\mathcal{D}_1$ and $\mathcal{D}_2$, of equal sizes, and re-express $\mathbb{E}_\beta[\rho G]^2$ as $\mathbb{E}_\beta[\rho G]\mathbb{E}_\beta[\rho G]$, where the first expectation is estimated using samples from $\mathcal{D}_1$ and the second expectation is estimated using samples from $\mathcal{D}_2$. Based on this *double sampling* approach, we propose the following off-policy variance estimator,

$$\hat{\sigma}_n^2 \coloneqq \frac{1}{n}\sum_{i=1}^{n}\rho_i G_i^2 - \left(\frac{1}{|\mathcal{D}_1|}\sum_{i=1}^{|\mathcal{D}_1|}\rho_i G_i\right)\left(\frac{1}{|\mathcal{D}_2|}\sum_{i=1}^{|\mathcal{D}_2|}\rho_i G_i\right). \quad (4)$$

This simple data-splitting trick suffices to create, $\hat{\sigma}_n^2$, an off-policy variance estimator that is both unbiased and consistent. We formalize this in the following theorems.

Theorem 1. *Under Assumption* 1, $\hat{\sigma}_n^2$ *is an unbiased estimator of* $\sigma^2(\pi)$. *That is,* $\mathbb{E}_\beta[\hat{\sigma}_n^2] = \sigma^2(\pi)$.

Theorem 2. *Under Assumption* 1, $\hat{\sigma}_n^2$ *is a consistent estimator of* $\sigma^2(\pi)$. *That is,* $\hat{\sigma}_n^2 \xrightarrow{a.s.} \sigma^2(\pi)$.

Remark 2. *It is possible that* $\hat{\sigma}_n^2$ *results in negative values (see Appendix C for an example). One practical solution to avoid negative values for variance can be to define* $\hat{\sigma}_n^{2+} \coloneqq \texttt{clip}(\hat{\sigma}_n^2, \texttt{min} = 0, \texttt{max} = \infty)$. *However, this may make* $\hat{\sigma}_n^{2+}$ *a biased estimator, i.e.,* $\mathbb{E}_\beta[\hat{\sigma}_n^{2+}] \neq \sigma^2(\pi)$. *Notice that this is the expected behavior of IS based estimators. For example, the IS estimates of expected return can be smaller or larger than the smallest and largest possible returns when* $\rho > 1$. *We refer the reader to the works by McHugh and Mielke (1968), Anderson (1965), and Nelder (1954) for other occurences of negative variance and its interpretations.*

Variance-Reduced Estimation of Variance

Despite $\hat{\sigma}_n^2$ being both an unbiased and a consistent estimator of variance, the use of IS can make its variance high. Specifically, the importance ratio ρ may become unstable when its denominator, $\prod_{i=0}^{T}\beta(A_{(i)}|S_{(i)})$, is small.

Algorithm 1: Variance-Reduced Off-Policy Variance Estimator

Input: Set of trajectories $\mathcal{D}$

$\mathcal{D}_1, \mathcal{D}_2 \leftarrow \texttt{equal\_split}(\mathcal{D})$

$X = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=0}^{T} \sum_{k=0}^{T} \rho_i(0, \max(j,k)) \gamma^{j+k} R_{i(j)} R_{i(k)}$

$Y = \frac{1}{|\mathcal{D}_1|} \sum_{i=1}^{|\mathcal{D}_1|} \sum_{j=0}^{T} \rho(0,j) \gamma^j R_{i(j)}$

$Y' = \frac{1}{|\mathcal{D}_2|} \sum_{i=1}^{|\mathcal{D}_2|} \sum_{j=0}^{T} \rho(0,j) \gamma^j R_{i(j)}$

Return $X - YY'$

To mitigate variance, it is common in off-policy mean estimation to use *per-decision importance sampling* (PDIS), instead of the full-trajectory IS, to reduce variance without incurring any bias (Precup 2000). It is therefore natural to ask: Is it also possible to have something like PDIS for off-policy variance estimation?

Recall from (4) that the expectation of the terms inside the parentheses correspond to $\mathbb{E}_\beta[\rho G] = \mu(\pi)$, a term for which we can directly leverage the existing PDIS estimator, $\mathbb{E}_\beta[\rho G] = \mathbb{E}_\beta\left[\sum_{i=0}^{T} \rho(0,i)\gamma^i R_{(i)}\right]$. Intuitively, PDIS leverages the fact that the probability of observing an individual reward at timestep i only depends upon the probability of the trajectory up to timestep i.

However, the first term in the *right hand side* (RHS) of (4) contains $G^2 = (\sum_{i=o}^{T} \gamma^i R_{(i)})^2$. Expanding this expression, we obtain self-coupled and cross-coupled *reward* terms, $R_{(i)}^2$ and $R_{(i)}R_{(j)}$, which makes PDIS not directly applicable. In the following theorem we present a new estimator, *coupled-decision importance sampling* (CDIS), which extends PDIS to handle these coupled rewards.

Theorem 3. *Under Assumption* 1*,*

$$\mathbb{E}_\beta\left[\rho G^2\right] = \mathbb{E}_\beta\left[\sum_{i=0}^{T}\sum_{j=0}^{T} \rho\left(0, \max(i,j)\right) \gamma^{i+j} R_{(i)} R_{(j)}\right].$$

Borrowing intuition from PDIS, CDIS leverages the fact that the probability of observing a coupled-reward, $R_{(i)}R_{(j)}$, only depends on the probability of the trajectory up to the i or jth timestep, whichever is larger. Importance ratios beyond that timestep can therefore be discarded without incurring bias. In Algorithm 1, we combine both per-decision and coupled-decision IS to construct a variance-reduced estimator of $\sigma^2(\pi)$.

HCOVE using Concentration Inequalities

In the previous section we found that the reformulation presented in (3) was helpful for creating a variance reduced off-policy variance estimator $\hat{\sigma}_n^2$. In this section, we will again build upon (3) to obtain a *confidence interval* (CI) for $\sigma^2(\pi)$. One specific advantage of (3) is that it allows us to build upon existing concentration inequalities, which were developed for obtaining CIs for $\mu(\pi)$, to obtain a CI for $\sigma^2(\pi)$.

Before moving further, we define some additional notation. For any random variable X, let $\text{CI}^+(\mathbb{E}[X], \delta)$, $\text{CI}_-(\mathbb{E}[X], \delta)$, and $\text{CI}_-^+(\mathbb{E}[X], \delta)$ represent only upper, only lower, and both upper and lower $(1-\delta)$-confidence bounds for $\mathbb{E}[X]$, respectively. That is, $\Pr(\text{CI}^+(\mathbb{E}[X], \delta) \geq \mathbb{E}[X]) \geq 1-\delta$, $\Pr(\text{CI}_-(\mathbb{E}[X], \delta) \leq \mathbb{E}[X]) \geq 1-\delta$, etc. For brevity, We will sometimes suppress CI's dependency on δ.

With the above notation, we now establish a high-confidence bound on (3). Recall that (3) consists of one positive term $\mathbb{E}[\rho G^2]$ and a negative term $-\mathbb{E}[\rho G]^2$. Therefore, given a confidence interval for both of these terms, the high-confidence upper bound for (3) would be the high-confidence upper bound of $\mathbb{E}[\rho G^2]$ minus the high-confidence lower bound of $\mathbb{E}[\rho G]^2$, and vice-versa to obtain a high-confidence lower bound on (3). That is, let $\delta_1, \delta_2, \delta_3$ and δ_4 be some constants in $(0, 0.5]$ such that $\delta/2 = \delta_1 + \delta_2 = \delta_3 + \delta_4$. The lower bound $v^{\texttt{lb}}$ and the upper bound $v^{\texttt{ub}}$ can be expressed as,

$$v^{\texttt{lb}} \coloneqq \text{CI}_-\left(\mathbb{E}_\beta[\rho G^2], \delta_1\right) - \text{CI}^+\left(\mathbb{E}_\beta[\rho G]^2, \delta_2\right), \quad (5)$$

$$v^{\texttt{ub}} \coloneqq \text{CI}^+\left(\mathbb{E}_\beta[\rho G^2], \delta_3\right) - \text{CI}_-\left(\mathbb{E}_\beta[\rho G]^2, \delta_4\right). \quad (6)$$

For getting the desired CIs for the first terms in the RHS of (5) and (6), notice that any method for obtaining a CI on the expected return, $\mathbb{E}_\beta[\rho G]$, can also be used to bound $\mathbb{E}_\beta[\rho G']$, where $G' \coloneqq G^2$.

For getting the desired CIs in the second term in the RHS of (5) and (6), we perform *interval propagation* (Jaulin, Braems, and Walter 2002). That is, given a high confidence interval for $\mathbb{E}[\rho G]$, since $\mathbb{E}[\rho G]^2$ is a quadratic function of $\mathbb{E}[\rho G]$, the upper bound for the value of $\mathbb{E}[\rho G]^2$ would be the *maximum* of the squared values of the end-points of the interval for $\mathbb{E}[\rho G]$. Similarly, the lower bound on $\mathbb{E}[\rho G]^2$ would be 0 if the signs of upper and lower bounds for $\mathbb{E}[\rho G]$ are different, otherwise it would be the *minimum* of the squared value of the end-points of the interval for $\mathbb{E}[\rho G]$. An illustration of this concept is presented in Figure 2.

Using interval propagation, the resulting upper bound is

$$\text{CI}^+(\mathbb{E}_\beta[\rho G]^2) = \texttt{max}(\text{CI}_-(\mathbb{E}_\beta[\rho G])^2, \text{CI}^+(\mathbb{E}_\beta[\rho G])^2),$$

and the resulting high-confidence lower bound is $\text{CI}_-(\mathbb{E}_\beta[\rho G]^2) = 0$ if both $\text{CI}_-(\mathbb{E}_\beta[\rho G]) \leq 0$ and $\text{CI}^+(\mathbb{E}_\beta[\rho G]) \geq 0$, and $\text{CI}_-(\mathbb{E}_\beta[\rho G]^2) = \texttt{min}(\text{CI}^+(\mathbb{E}_\beta[\rho G])^2, \text{CI}_-(\mathbb{E}_\beta[\rho G])^2)$ otherwise. Notice that these upper and lower high-confidence bounds on $\mathbb{E}_\beta[\rho G]^2$ can always be reduced to $\max(G_{\texttt{min}}^2, G_{\texttt{max}}^2)$ (the maximum squared return under any policy) when they are larger.

In the following theorem, we prove that the resulting confidence interval, $\mathcal{C}$, has *guaranteed coverage*, i.e., that it holds with probability $1-\delta$.

Theorem 4 (Guaranteed coverage)**.** *Under Assumption* 1*, if* $(\delta_1 + \delta_2 + \delta_3 + \delta_4) \leq \delta$*, then for the confidence interval* $\mathcal{C} \coloneqq [v^{lb}, v^{ub}]$*,*

$$\Pr\left(\sigma^2(\pi) \in \mathcal{C}\right) \geq 1 - \delta.$$

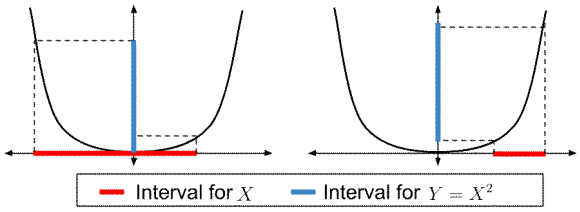

Figure 2: Two separate examples that show how interval propagation can be used to map confidence intervals for X (in red) to confidence intervals for $Y = X^2$ (in blue).

Remark 3. *Theorem 4 presents a two-sided interval. If only a lower bound or only an upper bound is required, then it suffices if only $(\delta_1 + \delta_2) \leq \delta$ or $(\delta_3 + \delta_4) \leq \delta$, respectively.*

Remark 4. *$\mathcal{C}$ can always be clipped via taking the intersection with the interval $[0, (G_{max} - G_{min})^2/4]$, since the variance will always be within this range (see Popoviciu's inequality for the deterministic upper bound on variance).*

A Tale of Long-Tails

One important advantage of Theorem 4 is that it constructs a CI, $\mathcal{C}$, for $\sigma^2(\pi)$ using *any* concentration inequality that can be used to get CIs $\text{CI}_+(\mathbb{E}_\beta[\rho G])$ and $\text{CI}_-(\mathbb{E}_\beta[\rho G])$) for $\mu(\pi)$. Hence, the tightness of $\mathcal{C}$ scales directly with the tightness of these existing off-policy policy evaluation methods for the expected discounted return. However, naïvely using common concentration inequalities can result in wide and uninformative CIs, as we discuss below. Therefore, in this section, we aim to establish a *control-variate* that is designed to produce tighter CIs for $\sigma^2(\pi)$.

Typically, for a random variable $X \in [a, b]$, the width of the confidence interval for $\mathbb{E}[X]$ obtained using common concentration inequalities, like Hoeffding's (Hoeffding 1994) or an empirical Bernstein inequality (Maurer and Pontil 2009), have a direct dependence on the range, $(b - a)$. Unfortunately, as shown by Thomas, Theocharous, and Ghavamzadeh (2015), IS based estimators may exhibit *extremely* long tail behavior and can have a range in the order of 10^{10}. For example, even if $\forall a \in \mathcal{A}$ and $\forall s \in \mathcal{S}$, if $\beta(a|s) > 0.1$, then the maximum possible importance weighted return of a ten timestep long trajectory can be on the order of $(1/0.1)^{10} = 10^{10}$ even when returns are normalized to the $[0, 1]$ interval. Such a large range causes Hoeffding's inequality and empirical Bernstein inequalities to produce wide and uninformative confidence intervals, especially when the number of samples is not enormous.

To construct a *lower bound* for $\mathbb{E}[X]$, while being robust to the long tail, Thomas, Theocharous, and Ghavamzadeh (2015) notice that truncating the upper-tail of X to a constant c can only lower the expected value of X, i.e., for $X' := \texttt{min}(X, c)$, $\mathbb{E}[X'] \leq \mathbb{E}[X]$. Therefore, $X'_{\text{lb}} := \text{CI}_-(\mathbb{E}[X'], \delta)$ is also a valid lower bound for $\mathbb{E}[X]$ and $\Pr(X'_{\text{lb}} \leq \mathbb{E}[X]) \geq 1 - \delta$. Additionally, truncating allows for significantly shrinking the range from $[a, b]$ to $[a, c]$, thereby effectively leading to a much tighter lower bound when c is chosen appropriately. For completeness, we review this bound in Appendix F.

While this bound was designed specifically for getting the *lower bounds* required in (5) and (6), it cannot be naïvely used to get the *upper bounds*. As $\mathbb{E}[X'] \leq \mathbb{E}[X]$, the upper bound $X'_{\text{ub}} := \text{CI}^+(\mathbb{E}[X'], \delta)$, may not be a valid upper bound for $\mathbb{E}[X]$ and $\Pr(X'_{\text{ub}} \geq \mathbb{E}[X]) \not\geq 1 - \delta$. A natural question is then: How can an upper bound be obtained that is robust to the long upper tail?

To answer this question, notice that if instead of the upper tail, the *lower* tail of the distribution was long, then the upper bound constructed after truncating the lower tail would still be valid. Therefore, we introduce a control-variate ξ which can be used to switch the tails of the distribution of an IS based estimator, such that both upper and lower valid bounds can be obtained using the resulting distribution. We formalize this in the following theorem.

Theorem 5. *Let X be either G or G^2, then for any $\delta \in (0, 0.5]$ and a fixed constant ξ,*

$$\text{CI}^+_-(\mathbb{E}_\beta[\rho X], \delta) = \text{CI}^+_-(\mathbb{E}_\beta[\rho(X - \xi)], \delta) + \xi.$$

Remark 5. *When ξ is set to be the maximum value that X can take, then the random variable $\rho(X - \xi)$ will have an upper bound of 0 and a long lower tail since $\rho \geq 0$ and $(X - \xi) \leq 0$. Similarly, when ξ is set to be the minimum value that X can take, then the random variable $\rho(X - \xi)$ will have a lower bound of 0 and a long upper tail. When a two-sided interval is required, two different estimators need to be constructed using the values for ξ discussed above.*

Theorem 5 allows us to control the tail-behavior such that the tight bounds presented by Thomas, Theocharous, and Ghavamzadeh (2015) can be leveraged to obtain *both* valid upper and valid lower high-confidence bound. However, Theorem 5 still makes use of the full trajectory importance ratio ρ, which can result in high-variance and inflate the confidence intervals.

To mitigate the above problem as well, we combine the variance reduction property of per-decision and coupled-decision IS offered by Theorem 3, and the control over the tail behavior offered by Theorem 5, and present the following theorem (see Appendix F for the complete algorithm).

Theorem 6. *Under Assumption 1, for any $\delta \in [0.0.5]$, let $\xi_R := \texttt{max}(R^2_{min}, R^2_{max})$ and $\xi_G := \texttt{max}(G^2_{min}, G^2_{max})$ then*

$$X := \sum_{i=0}^{T} \sum_{j=0}^{T} \rho\left(0, \texttt{max}(i, j)\right) \gamma^{i+j} \left(R_{(i)} R_{(j)} - \xi_R\right),$$

$$Y := \sum_{i=0}^{T} \rho(0, i) \gamma^i \left(R_{(i)} - R_{max}\right),$$

then $\Pr(X \leq 0) = \Pr(Y \leq 0) = 1$, and

$$\text{CI}^+\left(\mathbb{E}_\beta[\rho G^2], \delta\right) = \text{CI}^+\left(\mathbb{E}_\beta[X], \delta\right) + \xi_G,$$

$$\text{CI}^+\left(\mathbb{E}_\beta[\rho G], \delta\right) = \text{CI}^+\left(\mathbb{E}_\beta[Y], \delta\right) + G_{max}.$$

Remark 6. *For a lower bound on $\mathbb{E}_\beta[\rho G^2]$, notice that $\rho G^2 \geq 0$ always and thus the lower bound on $\mathbb{E}_\beta[X]$, where ξ_R and ξ_G are set to 0, can be used. Lower bound on $\mathbb{E}_\beta[\rho G]$ can be constructed by using the lower bound on $\mathbb{E}_\beta[Y]$, when R_{max} and G_{max} are replaced by R_{min} and G_{min}.*

Remark 7. *If some trajectories have horizon length $t < T$, then they must be appropriately padded to ensure that $\forall i \in [t+1, T],\ \rho(0,i) = \rho(0,t)$ and $R_{(i)} = 0$, such that in expectation the total amount added/subtracted by the control variate is zero.*

HCOVE using Statistical Bootstrapping

Bootstrap is a popular non-parametric technique for finding *approximate* confidence intervals (Efron and Tibshirani 1994). The core idea of bootstrap is to re-sample the observed data $\mathcal{D}$ and construct B pseudo-datasets $\{\mathcal{D}_i^*\}_{i=1}^B$ in a way such that each $\mathcal{D}_i^*$ resembles a draw from the true underlying *data generating process*. With each pseudo-data $\mathcal{D}_i$, an unbiased pseudo-estimate of a desired sample statistic can be created. For our problem, this statistic corresponds to $\hat{\sigma}_n^{2*}$, the estimate of $\sigma^2(\pi)$ obtained using (4). Thereby, leveraging the entire set of pseudo-data $\{\mathcal{D}_i^*\}_{i=1}^B$, an empirical distribution for the estimates of the variance $\{\hat{\sigma}_{n,i}^{2*}\}_{i=1}^B$ can be obtained. This empirical distribution approximates the true distribution of $\hat{\sigma}_n^2$ and can thus be leveraged to obtain CIs for $\sigma^2(\pi)$ using the *percentile* method, the *bias-corrected and accelerated* (BCa) method, etc. (DiCiccio and Efron 1996).

A drawback of bootstrap is the increased computational cost required for re-sampling and analysing B pseudo datasets. Further, the CIs obtained from bootstrap are only approximate, meaning that they can fail with more than δ probability. However, the primary advantage of using bootstrap is that it provides much tighter CIs, as compared to the ones obtained using concentration inequalities, and hence can be more informative for certain applications in practice.

Let $\hat{\mathcal{C}}$ be the approximate interval for $\sigma^2(\pi)$, for a given confidence δ, obtained using bootstrap (see Appendix F for the complete algorithm). Then under the following assumption on the higher-moments of $\hat{\sigma}_n^2$, we directly leverage the results for bootstrap to obtain an error-rate for $\hat{\mathcal{C}}$.

Assumption 2. *The third moment of $\hat{\sigma}_n^2$ is bounded. That is, $\exists c_1 < \infty$ such that $\mathbb{E}_\beta[(\hat{\sigma}_n^2 - \mathbb{E}_\beta[\hat{\sigma}_n^2])^3] < c_1$.*

Assumption 2 is a typical requirement for bootstrap methods (Efron and Tibshirani 1994). Assumption 2 can easily be satisfied by commonly used entropy regularized behavior policies that ensure that $\exists c_2 > 0$ such that $\forall a \in \mathcal{A}, \forall s \in \mathcal{S},\ \beta(a|s) \geq c_2$. This would ensure that the importance ratio $\rho \leq 1/(c_2^T)$, and because $G \leq G_{\max}$, ρG and ρG^2 would also be bounded. This ensures that $\hat{\sigma}_n^2$ is bounded, and therefore all its moments are also bounded, as required by Assumption 2. We formalize the asymptotic correctness of bootstrap confidence intervals in the following theorem.

Theorem 7. *Under Assumptions 1 and 2, the confidence interval $\hat{\mathcal{C}}$ has a finite sample error of $O(n^{-1/2})$. That is,*

$$\Pr\left(\sigma^2(\pi) \in \hat{\mathcal{C}}\right) \geq 1 - \delta - O\left(n^{-\frac{1}{2}}\right).$$

Remark 8. *Other variants of bootstrap (Bootstrap-t, BCa, etc.) can also be used, which typically offer higher order refinement by reducing the finite sample error-rate to $O(n^{-1})$ (DiCiccio and Efron 1996).*

Related Work

When samples are from the distribution whose variance needs to be estimated, then under the assumption that the distribution is normal, the χ^2 distribution can be used for providing CIs for the variance. Effects of non-normality on tests of significance were first analyzed by Pearson and Adyanthaya (1929) and has led to a large body of literature on variance tests (Pearson 1931; Box 1953; Levene 1960). Various modifications to χ^2 tests have also been proposed to be robust against samples from non-normal distributions (Subrahmaniam 1966; García-Pérez 2006; Pan 1999; Lim and Loh 1996). The statistical bootstrap approach used in this paper to obtain bounds on the variance is closest to the bootstrap test developed by Shao (1990). However, all of these methods are analogous to on-policy variance analysis.

In the context of RL, Sobel (1982) first introduced Bellman operators for the second moment and combined it with the first moment to compute the variance. Temporal difference (TD) style algorithms have been subsequently developed for estimating the variance of returns (Tamar, Di Castro, and Mannor 2016; La and Ghavamzadeh 2013; White and White 2016; Sherstan et al. 2018). However, such TD methods might suffer from potential instabilities when used with function approximators and off-policy data (Sutton and Barto 2018). Policy gradient style algorithms have also been developed for finding policies that optimize variance related objectives (Di Castro, Tamar, and Mannor 2012; Tamar and Mannor 2013), however, these are limited to the on-policy setting. We are not aware of any work in the RL literature that provides unbiased and consistent off-policy variance estimators, nor high-confidence bounds for thereon.

Outside RL, variants of off-policy (or counterfactual) estimation using importance sampling (or *inverse propensity estimator* (Horvitz and Thompson 1952)) is common in econometrics (Hoover 2011; Stock and Watson 2015) and causal inference (Pearl 2009). While these works have mostly focused on mean estimation, counterfactual probability density or quantiles of potential outcomes can also be estimated (DiNardo, Fortin, and Lemieux 1995; Melly 2006; Chernozhukov, Fernández-Val, and Melly 2013; Donald and Hsu 2014). Such distribution estimation methods can possibly also be used to estimate off-policy variance; however it is unclear how to obtain unbiased estimates of the variance from an unbiased estimate of the distribution. Instead, focusing directly on the variance can be more data-efficient and can also lead to unbiased estimators. Further, these works neither leverage any MDP structure to reduce variance resulting from IS, nor do they provide any methods that provide high-confidence bounds on the variance. In the RL setting, the problem of high variance in IS is exacerbated as sequential interaction leads to multiplicative importance ratios, thereby requiring additional consideration for long tails to obtain tight bounds.

Experimental Study

Inspired by real-world applications where OVE and HCOVE can be useful, we validate our proposed estimators empirically on two domains motivated by real-world ap-

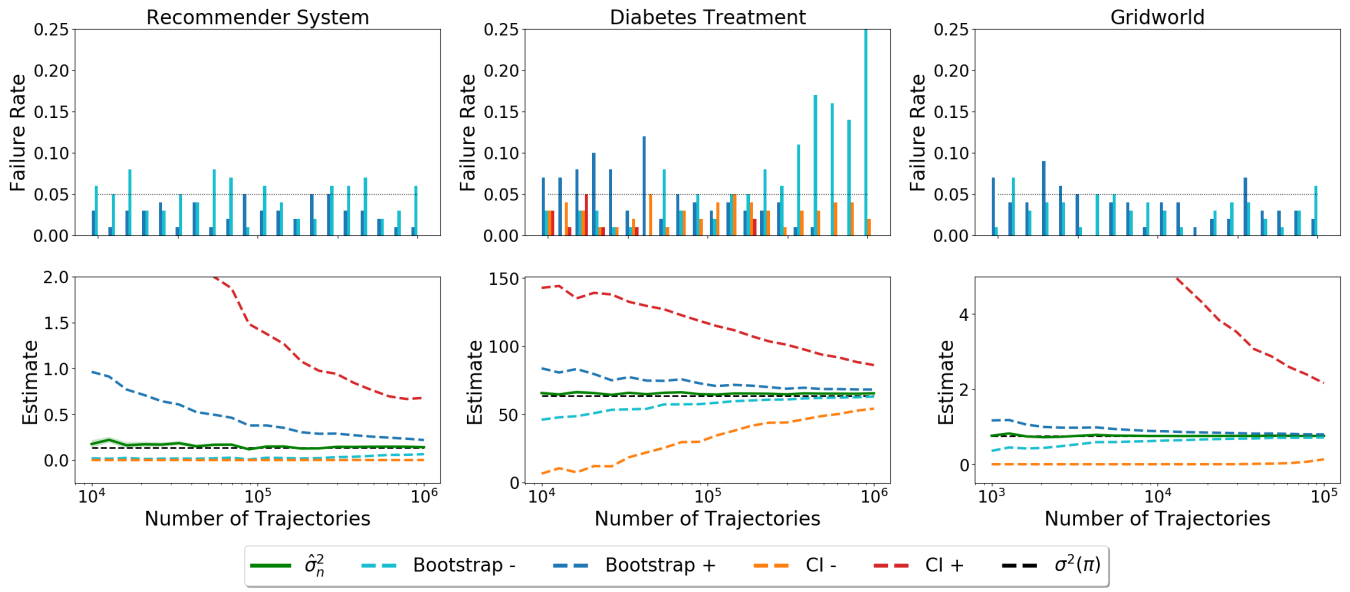

Figure 3: Experimental results using 100 trials. (Top) Empirically observed fraction (out of 100 trials) for which the computed confidence interval did not include the actual variance, for the given number of trajectories (plotted on the shared horizontal axis), for the proposed upper and lower high-confidence bounds that were constructed using concentration inequalities (labeled as: CI $+$, and CI $-$) and using bootstrap (labeled as: Bootstrap $+$, and Bootstrap $-$). The color of the bars refer to the legend, and these bars should ideally be below the line representing the confidence level $\delta = 0.05$. (Bottom) The dashed, colored, lines represent the value of the respective high-confidence bounds, constructed with confidence level $1 - \delta$ each. The green line represents the value of our proposed estimator $\hat{\sigma}_n^2$ and the shaded area around it (almost negligible) corresponds to the standard error. Black dashed line represents the true variance, $\sigma^2(\pi)$. The unbiased and consistent property of $\hat{\sigma}_n^2$ can be visualized by comparing it with $\sigma^2(\pi)$. Notice that as the bootstrap confidence interval $\hat{\mathcal{C}}$ is only approximate, it can fail with more that δ probability. In comparison, confidence interval $\mathcal{C}$ obtained using concentration inequalities provide guaranteed coverage. However, as clear from the plots, $\mathcal{C}$ can be conservative, while $\hat{\mathcal{C}}$ provides a tighter interval.

plications. Here, we only provide a brief description about the experimental setup and the main results. Appendix G contains additional experimental details.

Diabetes treatment: This domain is based on an open-source implementation (Xie 2019) of the FDA approved Type-1 Diabetes Mellitus simulator (T1DMS) (Man et al. 2014) for treatment of Type-1 Diabetes, where the objective is to control an insulin pump to regulate the blood-glucose level of a patient. High-confidence estimation of the variance of a controller's outcome, before deployment, can be informative when assessing potential harm to the patient that may be caused by the controller.

Recommender system: This domain simulates the problem of providing online recommendations based on customer interests, where it is often useful to obtain high-confidence estimates for the variance of customer's experience, before actually deploying the system, to limit financial loss.

Gridworld: We also consider a standard 4×4 Gridworld with stochastic transitions. There are eight discrete actions corresponding to up, down, left, right, and the four diagonal movements.

Given trajectories collected using a behavior policy β, in Figure 3 we provide the trend of our estimator $\hat{\sigma}_n^2$ for an evaluation policy π, and the confidence intervals $\mathcal{C}$ and $\hat{\mathcal{C}}$ as the number of trajectories increase (more details on how π and β were constructed can be found in Appendix G). As established in Theorem 1 and Theorem 2, $\hat{\sigma}_n^2$ can be seen to be both an unbiased and a consistent estimator of $\sigma^2(\pi)$. Similarly, as established in Theorem 4, the $(1 - \delta)$-confidence interval $\mathcal{C}$ provides guaranteed coverage. In comparison, as established in Theorem 7, bootstrap bounds are approximate and can fail more than δ fraction of the time. However, bootstrap bounds can still be useful in many applications as they provide tighter intervals.

Conclusion

In this work, we addressed an understudied problem of estimating and bounding $\sigma^2(\pi)$ using only off-policy data. We took the first steps towards developing a model-free, off-policy, unbiased, and consistent estimator of $\sigma^2(\pi)$ using a simple double-sampling trick. We then showed how bound propagation using concentration inequalities, or statistical bootstrap, can be used to obtain CIs for $\sigma^2(\pi)$. Finally, empirical results were provided to support the established theoretical results.

References

Anderson, R. 1965. Negative variance estimates. *Technometrics* 7(1): 75–76.

Bastani, M. 2014. Model-free intelligent diabetes management using machine learning. *M.S. Thesis, University of Alberta* .

Bellemare, M. G.; Dabney, W.; and Munos, R. 2017. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887* .

Bottou, L.; Peters, J.; Quiñonero-Candela, J.; Charles, D. X.; Chickering, D. M.; Portugaly, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14(1): 3207–3260.

Box, G. E. 1953. Non-normality and tests on variances. *Biometrika* 40(3/4): 318–335.

Chernozhukov, V.; Fernández-Val, I.; and Melly, B. 2013. Inference on counterfactual distributions. *Econometrica* 81(6): 2205–2268.

Dabney, W.; Rowland, M.; Bellemare, M. G.; and Munos, R. 2017. Distributional reinforcement learning with quantile regression. *arXiv preprint arXiv:1710.10044* .

Di Castro, D.; Tamar, A.; and Mannor, S. 2012. Policy gradients with variance related risk criteria. *arXiv preprint arXiv:1206.6404* .

DiCiccio, T. J.; and Efron, B. 1996. Bootstrap confidence intervals. *Statistical Science* 189–212.

DiNardo, J.; Fortin, N. M.; and Lemieux, T. 1995. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Technical report, National Bureau of Economic Research.

Donald, S. G.; and Hsu, Y.-C. 2014. Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics* 178: 383–397.

Efron, B.; and Tibshirani, R. J. 1994. *An introduction to the Bootstrap*. CRC press.

García-Pérez, A. 2006. Chi-square tests under models close to the normal distribution. *Metrika* 63(3): 343–354.

Guo, Z.; Thomas, P. S.; and Brunskill, E. 2017. Using options and covariance testing for long horizon off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, 2492–2501.

Hanna, J. P.; Stone, P.; and Niekum, S. 2017. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*.

Hoeffding, W. 1994. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, 409–426. Springer.

Hoover, K. D. 2011. *Counterfactuals and causal structure*. Oxford University Press Oxford.

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

Jaulin, L.; Braems, I.; and Walter, E. 2002. Interval methods for nonlinear identification and robust control. In *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, volume 4, 4676–4681. IEEE.

Kostrikov, I.; and Nachum, O. 2020. Statistical bootstrapping for uncertainty estimation in off-policy Evaluation. *arXiv preprint arXiv:2007.13609* .

Kuindersma, S. R.; Grupen, R. A.; and Barto, A. G. 2013. Variable risk control via stochastic optimization. *The International Journal of Robotics Research* 32(7): 806–825.

Kuzborskij, I.; Vernade, C.; György, A.; and Szepesvári, C. 2020. Confident off-policy evaluation and selection through self-normalized importance weighting. *arXiv preprint arXiv:2006.10460* .

La, P.; and Ghavamzadeh, M. 2013. Actor-critic algorithms for risk-sensitive MDPs. *Advances in Neural Information Processing Systems* 26: 252–260.

Levene, H. 1960. Robust tests for equality of variances. Contributions to probability and statistics In Olkin I, Ed.

Lim, T.-S.; and Loh, W.-Y. 1996. A comparison of tests of equality of variances. *Computational Statistics & Data Analysis* 22(3): 287–301.

Liu, Q.; Li, L.; Tang, Z.; and Zhou, D. 2018. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, 5356–5366.

Man, C. D.; Micheletto, F.; Lv, D.; Breton, M.; Kovatchev, B.; and Cobelli, C. 2014. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology* 8(1): 26–34.

Maurer, A.; and Pontil, M. 2009. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740* .

McHugh, R. B.; and Mielke, P. W. 1968. Negative variance estimates and statistical dependence in nested sampling. *Journal of the American Statistical Association* 63(323): 1000–1003.

Melly, B. 2006. Estimation of counterfactual distributions using quantile regression. Technical report, Universität St. Gallen.

Montgomery, D. C. 2007. *Introduction to statistical quality control*. John Wiley & Sons.

Nelder, J. 1954. The interpretation of negative components of variance. *Biometrika* 41(3/4): 544–548.

Pan, G. 1999. On a Levene type test for equality of two variances. *Journal of Statistical Computation and Simulation* 63(1): 59–71.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Pearson, E. S. 1931. The analysis of variance in cases of non-normal variation. *Biometrika* 114–133.

Pearson, E. S.; and Adyanthaya, N. 1929. The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika* 21(1/4): 259–286.

Precup, D. 2000. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series* 80.

Sakaguchi, Y.; and Takano, M. 2004. Reliability of internal prediction/estimation and its application. I. Adaptive action selection reflecting reliability of value function. *Neural Networks* 17(7): 935–952.

Sato, M.; Kimura, H.; and Kobayashi, S. 2001. TD algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence* 16(3): 353–362.

Shao, J. 1990. Bootstrap estimation of the asymptotic variances of statistical functionals. *Annals of the Institute of Statistical Mathematics* 42(4): 737–752.

Sherstan, C.; Ashley, D. R.; Bennett, B.; Young, K.; White, A.; White, M.; and Sutton, R. S. 2018. Comparing direct and indirect temporal-difference methods for estimating the variance of the return. In *UAI*, 63–72.

Sobel, M. J. 1982. The variance of discounted Markov decision processes. *Journal of Applied Probability* 19(4): 794–802.

Stock, J. H.; and Watson, M. W. 2015. *Introduction to Econometrics*. Pearson Press.

Subrahmaniam, K. 1966. Some contributions to the theory of non-normality-I (univariate case). *Sankhyā: The Indian Journal of Statistics, Series A* 389–406.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press, 2 edition.

Tabachnick, B. G.; and Fidell, L. S. 2007. *Experimental Designs Using ANOVA*. Thomson/Brooks/Cole Belmont, CA.

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

Tamar, A.; and Mannor, S. 2013. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697* .

Teevan, J. B.; Dumais, S. T.; Liebling, D. J.; and Horvitz, E. J. 2009. Using variation in user interest to enhance the search experience. US Patent App. 12/163,561.

Thomas, P. S.; Theocharous, G.; and Ghavamzadeh, M. 2015. High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Wasserman, L. 2006. *All of nonparametric statistics*. Springer Science & Business Media.

White, M.; and White, A. 2016. A greedy approach to adapting the trace parameter for temporal difference learning. *arXiv preprint arXiv:1607.00446* .

Xie, J. 2019. *Simglucose v0.2.1 (2018)*. URL https://github.com/jxx123/simglucose.