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, \( \pi \), using only data collected from a currently deployed policy called the behavior policy, \( \beta \) (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).

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 \( \lambda \) 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
the maximum returns possible are 
variable observed at timestep represents the probability of taking action the starting state distribution. 
jectories under the evaluation policy behavior policy reliance upon IS and thus inherits its limitations. Namely, The proposed off-policy estimator of the var-
ance relies upon IS and thus inherits its limitations. Namely, (a) it requires knowledge of the action probabilities from the behavior policy \( \beta \), (b) it requires that the support of the trajectories under the evaluation policy \( \pi \) is a subset of the support under the behavior policy \( \beta \), and (c) the variance of the estimator scales exponentially with the length of the trajectory. 

**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 \( \beta \), (b) it requires that the support of the trajectories under the evaluation policy \( \pi \) is a subset of the support under the behavior policy \( \beta \), and (c) the variance of the estimator scales exponentially with the length of the trajectory. 

**Background and Problem Statement**

A Markov decision process (MDP) is a tuple \((S, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, d_0)\), where \( S \) 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. 

A policy \( \pi \) 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 \( R_{i(j)} \in [R_{\min}, R_{\max}] \) represent the reward observed at timestep \( j \) of the episode \( i \). Let the random variable \( G_i := \sum_{j=0}^{T} \gamma^j R_{i(j)} \) be the return for episode \( i \). Let \( c := (1 - \gamma)/\gamma \) so that the minimum and the maximum returns possible are \( G_{\min} := c R_{\min} \) and \( G_{\max} := c R_{\max} \), respectively. Let \( \mu(\pi) := \mathbb{E}_\pi[G] \) be the expected return, and \( \sigma^2(\pi) := \mathbb{V}_\pi[G] \) be the variance of returns, where the subscript \( \pi \) denotes that the trajectories are generated using policy \( \pi \).

1We 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 \( (T := \infty) \).

Let \( \mathcal{H}_{\pi(j)}^T \) be the set of all possible trajectories for a policy \( \pi \), from timestep \( i \) to timestep \( j \). Let \( H \) denote a complete trajectory: \((S_0, A_0, \Pr(A_0|S_0), R_0, S_1, \ldots, S(\infty))\), 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 \( \rho_i(0, T) := \prod_{j=0}^{T} \Pr(A_{i(j)}|S_{i(j)}) \) denote the product of importance ratios from timestep \( 0 \) to \( T \). For brevity, when the range of timesteps is not necessary, we write \( \rho_i := \rho_i(0, T) \). Similarly, when referring to \( \rho_i \) for an arbitrary \( i \in \{1, \ldots, n\} \), we often write \( \rho \). 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 \( \pi \), we aim to find an estimator \( \hat{\sigma}^2_n \) that is both an unbiased and consistent estimator of \( \sigma^2(\pi) \), i.e.,

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

**HCOVE Problem:** Given a set of trajectories \( \mathcal{D} \), an evaluation policy \( \pi \), and a confidence level \( 1 - \delta \), we aim to find a confidence interval \( C := [v^{1/2}, v^{1/2}] \), such that

\[
\Pr(\sigma^2(\pi) \in C) \geq 1 - \delta.
\]

**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 \( \pi \). For example, if \( \mathcal{D} \) has no trajectory that could be observed if policy \( \pi \) were to be executed, then \( \mathcal{D} \) provides little or no information about the possible outcomes under \( \pi \). To avoid this case, we make the following common assumption (Precup 2000), which is satisfied if \( \beta_i(a|s) = 0 \Rightarrow \pi(a|s) = 0 \) for all \( s \in S, a \in \mathcal{A} \), 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

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

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 \( \beta \), 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 \( \pi \) 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 \( \hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} \rho_i G_i \) of the importance weighted returns \( \{\rho_i G_i\}_{i=1}^{n} \) is an unbiased estimator of \( \mu(\pi) \) \cite{Precup2000}, i.e., \( \mathbb{E}_\pi[\hat{\mu}] = \mu(\pi) \). 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 \( \{\rho_i G_i\}_{i=1}^{n} \).

\[
\hat{\sigma}^{2n}_{\pi} := \frac{1}{n-1} \sum_{i=1}^{n} \left( \rho_i G_i - \frac{1}{n} \sum_{j=1}^{n} \rho_j G_j \right)^2. \tag{1}
\]

Unfortunately, \( \hat{\sigma}^{2n}_{\pi} \) 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 \[1\], \( \hat{\sigma}^{2n}_{\pi} \) may be a biased estimator of \( \sigma^2(\pi) \). That is, it is possible that \( \mathbb{E}_\pi[\hat{\sigma}^{2n}_{\pi}] \neq \sigma^2(\pi) \).

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

Since the on-policy variance is \( \mathbb{V}_\pi[(G - \mathbb{E}_\pi[G])^2] \), 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,

\[
\hat{\sigma}^2_n := \frac{1}{n-1} \sum_{i=1}^{n} \left( \rho_i G_i - \frac{1}{n} \sum_{j=1}^{n} \rho_j G_j \right)^2. \tag{2}
\]

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

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

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 \( \hat{\sigma}^2_n \) and \( \hat{\sigma}^{2n}_{\pi} \). Notice that an expansion of \[1\] and \[2\] produces self-coupled importance ratio terms. That is, terms consisting of \( \rho_i^2 \) and \( \rho_i \rho_j \). While \( \rho_i \) helps in correcting the distribution, its higher powers, \( \rho_i^2 \) and \( \rho_i^3 \), 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 \( \rho_i \) and \( \rho_j \) 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 \( \rho_i^2 \), but which may have \( \rho_i \rho_j \) terms, where \( i \neq j \). To do so, we consider the alternate formulation of variance,

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

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 \( \rho_i^2 \) terms. To remedy this problem, we explicitly split the set of sampled trajectories into two mutually exclusive sets, \( D_1 \) and \( D_2 \), 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 \( D_1 \) and the second expectation is estimated using samples from \( D_2 \). Based on this double sampling approach, we propose the following off-policy variance estimator,

\[
\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^{n} \rho_i G_i^2 - \frac{1}{|D_1|} \sum_{i=1}^{|D_1|} \rho_i G_i - \frac{1}{|D_2|} \sum_{i=1}^{|D_2|} \rho_i G_i. \tag{4}
\]

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, \( \hat{\sigma}_n^2 \xrightarrow{\text{a.s.}} \sigma^2(\pi) \).

**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 \( \hat{\sigma}_n^{2+} := \text{clip}((\hat{\sigma}_n^2, \min = 0, \max = \infty) \). 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 occurrences 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 \( \rho \) may become unstable when its denominator, \( \prod_{i=0}^{n} \beta(A(i)|S(i)) \), is small.
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 [Prelec, 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 corresponds to \( \mathbb{E}_\beta[\rho G] = \mu(\pi) \), a term for which we can directly leverage the existing PDIS estimator, \( \mathbb{E}_\beta[\rho G] = \mathbb{E}_\beta \left[ \sum_{i=0}^{T} \rho(i, i) \gamma^{i} R(i) \right] \). 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 \( \gamma^2 = (\sum_{i=0}^{T} \gamma^i R(i))^2 \). Expanding this expression, we obtain self-coupled and cross-coupled reward terms, \( R(i) \) and \( R(i) R(j) \), which makes PDIS not directly applicable. In the following theorem we present a new estimator, \textit{coupled-decision importance sampling} (CDIS), which we can directly leverage the existing PDIS estimation to use instead of the full-trajectory IS, to reduce variance without policy variance estimator \( \sigma^2(\pi) \).

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 \( CI^t(\mathbb{E}[X], \delta) \), \( CL^t(\mathbb{E}[X], \delta) \), and \( CI^+_{\mathbb{E}[X], \delta} \) represent only upper, only lower, and both upper and lower \((1 - \delta)\)-confidence bounds for \( \mathbb{E}[X] \), respectively. That is, \( \Pr(\mathbb{E}[X, \delta] \geq \mathbb{E}[X]) \geq 1 - \delta, \Pr(\mathbb{E}[X, \delta] \leq \mathbb{E}[X]) \geq 1 - \delta \), etc. For brevity, we will sometimes suppress CI’s dependency on \( \delta \).

With the above notation, we now establish a high-confidence bound on (5). 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 (5) 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 \( \delta_4 \) be some constants in \([0, 0.5]\) such that \( \delta/2 = \delta_1 + \delta_2 = \delta_3 + \delta_4 \). The lower bound \( u^{lb} \) and the upper bound \( u^{ub} \) can be expressed as,

\[
u^{lb} := CI^t_{-\mathbb{E}[\rho G^2], \delta_1} - CI^+_{\mathbb{E}[\rho G^2], \delta_2}, \quad (5)
\]

\[
u^{ub} := CI^+_{\mathbb{E}[\rho G^2], \delta_3} - CL^-_{\mathbb{E}[\rho G^2], \delta_4}. \quad (6)
\]

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, \( \mathbb{E}[\rho G] \), can also be used to bound \( \mathbb{E}[\rho G^2] \), where \( G' := G^2 \).

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 \( CI^t(\mathbb{E}[\rho G^2], \delta) = \max(\mathbb{E}[\mathbb{E}[\rho G^2]], \mathbb{E}[\mathbb{E}[\rho G]]) \), and the resulting high-confidence lower bound is \( CL^-_{\mathbb{E}[\rho G^2], \delta} = 0 \) if both \( \mathbb{E}[\mathbb{E}[\rho G]] \leq 0 \) and \( CI^+_{\mathbb{E}[\rho G^2]} \geq 0 \), and \( CL^-_{\mathbb{E}[\rho G^2]} = \min(\mathbb{E}[\mathbb{E}[\rho G]], \mathbb{E}[\mathbb{E}[\rho G]]) \) otherwise.

Notice that these upper and lower high-confidence bounds on \( \mathbb{E}[\rho G^2] \) can always be reduced to \( \max(\mathbb{E}[\rho G^2], \mathbb{E}[\rho G^2]) \) the maximum squared return under any policy when they are larger.

In the following theorem, we prove that the resulting confidence interval, \( C \), 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 \( C := [u^{lb}, u^{ub}] \),

\[
\Pr(\sigma^2(\pi) \in C) \geq 1 - \delta.
\]
section, we aim to establish a uninformative CIs, as we discuss below. Therefore, in this
ness of these existing off-policy policy evaluation methods
When returning to an earlier point in the trajectory, we can use the weights of the state visited to
produce wide and uninformative confidence intervals, especially for short trajectories. This issue
is particularly acute with the weighted return of a ten timestep long trajectory can be on the order of $10^{10}$.

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. C can always be clipped via taking the intersection with the interval $[0, (G_{\text{max}} - G_{\text{min}})^2/4]$, 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, C, for $\sigma^2(\pi)$ using any concentration inequality that can be used to get CIs $\text{EI} \pm (E[|G|] + \epsilon)$ for $\mu(\pi)$. Hence, the tightness of $C$s scales directly with the tightness of these existing off-policy policy evaluation methods for the expected discounted return. However, naively 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 $\eta$ which 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 $E[X]$ obtained using common concentration inequalities, such as Hoeffding’s inequality (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 $10^8$. For example, even if $\forall \alpha \in A$ and $\forall S \in S$, if $\beta(\alpha|\gamma) > 0.1$, 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 $E[X]$, 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 $X' := \min\{X, c\}$, $E[X'] \leq E[X]$. Therefore, $X_{\text{ub}} := \text{CI}_{\max}(E[X], \delta)$ is a valid upper bound for $E[X]$ and $\Pr(X_{\text{ub}} \leq E[X]) \geq 1 - \delta$. 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 Theorem 4 and 5, it cannot be naively used to get the upper bounds. As $E[X'] \leq E[X]$, the upper bound $X_{\text{ub}} := \text{CI}_{\text{min}}(E[X'], \delta)$, may not be a valid upper bound for $E[X]$ and $\Pr(X_{\text{ub}} \geq E[X]) \geq 1 - \delta$. 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 $\eta$ 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 $\xi$, $\text{CI}^\pm(\mathbb{E}_\eta[\rho X], \delta) = \text{CI}^\pm(\mathbb{E}_\eta[\rho(X - \xi)], \delta) + \xi$.

Remark 5. When $\xi$ 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 $\xi$ 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 $\xi$ discussed above.

Theorem 6 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 6 still makes use of the full trajectory importance ratio $\rho$, 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 4 and the control over the tail behavior offered by Theorem 6 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 $\xi_R := \max(R_{\text{min}}^2, R_{\text{max}}^2)$ and $\xi_G := \max(G_{\text{min}}^2, G_{\text{max}}^2)$ then

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

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

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

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

$$\text{CI}^+ (\mathbb{E}_\eta[\rho G], \delta) = \text{CI}^+ (\mathbb{E}_\eta[Y], \delta) + G_{\text{max}}.$$
Remark 7. If some trajectories have horizon length \( t < T \), then they must be appropriately padded to ensure that \( \forall i \in [t+1, T] \), \( r(0, i) = r(0, t) \) and \( R_{i-1} = 0 \), 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 \cite{efron1994bootstrap}. The core idea of bootstrap is to re-sample the observed data \( D \) and construct pseudo-datasets \( \{D_i^{*}\}_{i=1}^B \) in a way such that each \( D_i^{*} \) resembles a draw from the true underlying data generating process. With each pseudo-data \( D_i \), an unbiased pseudo-estimate of a desired sample statistic can be created. For this problem, this statistic corresponds to \( \hat{\sigma}^2_n \), the estimate of \( \sigma^2(\pi) \) obtained using \( (4) \). Thereby, leveraging the entire set of pseudo-data \( \{D_i^{*}\}_{i=1}^B \), an empirical distribution for the estimates of the variance \( \{\hat{\sigma}^2_n\}_{i=1}^B \) can be obtained. This empirical distribution approximates the true distribution of \( \hat{\sigma}^2_n \) and can thus be leveraged to obtain CIs for \( \sigma^2(\pi) \) using the percentile method, the bias-corrected and accelerated (BCa) method, etc. \cite{di1996estimation}.

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

**Assumption 2.** The third moment of \( \hat{\sigma}^2_n \) is bounded. That is, \( \exists c_3 < \infty \) such that \( E_{\hat{\sigma}^2_n}[(\hat{\sigma}^2_n - E_{\hat{\sigma}^2_n}[\hat{\sigma}^2_n])^3] < c_1 \).

Assumption 2 is a typical requirement for bootstrap methods \cite{efron1994bootstrap}. Assumption 2 can easily be satisfied by commonly used entropy regularized behavior policies that ensure that \( \exists c_2 > 0 \) such that \( \forall a \in A, \forall s \in S, \beta(a|s) \geq c_2 \). This would ensure that the importance ratio \( \rho \leq 1/(c_2^2) \), and because \( G \leq G_{\text{max}}, \rho G \) and \( \rho G^2 \) would also be bounded. This ensures that \( \hat{\sigma}^2_n \) 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 \( \tilde{C} \) has a finite sample error of \( O(n^{-1/2}) \). That is,

\[
Pr \left( \sigma^2(\pi) \in \tilde{C} \right) \geq 1 - \delta - O \left( n^{-\frac{1}{2}} \right).
\]

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 \( O(n^{-1}) \) \cite{di1996estimation}.

Related Work

When samples are from the distribution whose variance needs to be estimated, then under the assumption that the distribution is normal, the \( \chi^2 \) 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 \cite{pearson1929introduction} and has led to a large body of literature on variance tests \cite{pearson1931introduction, box1953distribution, levend1960distribution}. Various modifications to \( \chi^2 \) tests have also been proposed to be robust against samples from non-normal distributions \cite{subrahmanian1966large, garcia2006approximate, pan1999correction, lim1996bootstrap}. The statistical bootstrap approach used in this paper to obtain bounds on the variance is closest to the bootstrap test developed by Shao \cite{shao1990bootstrap}. However, all of these methods are analogous to on-policy variance analysis.

In the context of RL, Sobel \cite{sobel1982unbiased} 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 \cite{tamar2016variance, la2013reinforcement, white2016variance, sherstan2018variance}. However, such TD methods might suffer from potential instabilities when used with function approximators and off-policy data \cite{sutton2018reinforcement}. Policy gradient style algorithms have also been developed for finding policies that optimize variance related objectives \cite{di1996estimation, tamar2013variance}. 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 \cite{horvitz1952estimator}) is common in econometrics \cite{hoover2011measurement, stock2015econometric} and causal inference \cite{pearl2009causality}. While these works have mostly focused on mean estimation, counterfactual probability density or quantiles of potential outcomes can also be estimated \cite{nardo2013statistical, melly2008estimating, chernozhukov2013inference, donald2014distribution}. 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 $\hat{\sigma}^2_n$ 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 $\hat{\sigma}^2_n$ can be visualized by comparing it with $\sigma^2(\pi)$. Notice that as the bootstrap confidence interval $\hat{\sigma}^2_n$ is only approximate, it can fail with more that $\delta$ probability. In comparison, confidence interval $\sigma^2(\pi)$ obtained using concentration inequalities provide guaranteed coverage. However, as clear from the plots, $C$ can be conservative, while $\hat{\sigma}^2_n$ provides a tighter interval.

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 $4 \times 4$ 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 $\beta$, in Figure 3 we provide the trend of our estimator $\hat{\sigma}^2_n$ for an evaluation policy $\pi$, and the confidence intervals $C$ and $\hat{C}$ as the number of trajectories increase (more details on how $\pi$ and $\beta$ were constructed can be found in Appendix G). As established in Theorem 1 and Theorem 2, $\hat{\sigma}^2_n$ 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 $C$ provides guaranteed coverage. In comparison, as established in Theorem 7, bootstrap bounds are approximate and can fail more than $\delta$ 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.
Broader Impact

Note to a wider audience: Methods developed in this work can be beneficial for researchers and practitioners working with applications that require reliability guarantees, especially before the proposed system/policy is even deployed. It is worth noting that if the failure rate $\delta$ is set to be too low, then our bounds can result in overly conservative intervals. Further, for settings where the probabilities from behavior policies are not available, and are instead estimated, $\hat{\sigma}_n$ might be biased. Consequently, for applications that require hard-guarantees, or have a batch of sampled actions without their sampling probabilities, our methods are not applicable.

Future research directions: While we took some measures to mitigate the variance of $\hat{\sigma}_n$, IS can still result in high variance. Recent off-policy mean estimation methods show how the Markov structure of an MDP (Liu et al. 2018; Xie, Ma, and Wang 2019; Rowland et al. 2020) or an estimate of the model of an MDP (Jiang and Li 2016; Thomas andBrunskill 2016) can be further leveraged for variance reduction. Alternatively, if the entire behavior policy is available, and not just the probabilities for the sampled actions, then multi-importance sampling (Sbert, Havran, and Szirmay-Kalos 2018) can be leveraged to relax Assumption 1, and also get tighter bounds on the mean return (Papini et al. 2019; Metelli et al. 2020). Extending these methods for OVE and HCOVE remains an interesting future direction.

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.; Quinonero-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.

Jiang, N.; and Li, L. 2016. Doubly robust off-policy value evaluation for reinforcement learning. In International Conference on Machine Learning, 652–661. PMLR.

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. 1953. Nonnormality and tests on variances. The Collected Works of Wassily Hoeffding, 409–426. Springer.

Lim, T.-S.; and Loh, W.-Y. 1996. A comparison of tests of equality of variances. Computational Statistics & Data Analysis 22(3): 287–301.
High-Confidence Off-Policy (or Counterfactual) Variance Estimation  
(Supplementary Material)

A: Proofs for the Naïve Estimator \( \hat{\sigma}^2_n \)

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

Proof. We prove this using a counter-example. Consider the MDP shown in Figure 4.

![MDP Diagram](https://example.com/mdp_diagram.png)

Figure 4: In this MDP, \( S_0 \) is the starting state and \( S_3 \) is the terminal/absorbing state. From \( S_0 \) there are two available actions: \( a \) and \( b \), which yield a reward of 1 and 0 respectively. All transitions are deterministic and \( \gamma = 1 \).

For the purpose of a counter-example, we now describe the evaluation policy \( \pi \) and behavior policy \( \beta \). Let \( \pi \) be a policy that always selects action \( a \), i.e., \( \pi(a|S_0) = 1 \) and \( \pi(b|S_0) = 0 \). Since \( \pi \) is deterministic and action \( a \) yields a reward of 1 always, the variance of returns observed under \( \pi \) is 0. Let \( \beta \) be the behavior policy which selects both actions with equal probability, i.e., \( \beta(a|S_0) = \beta(b|S_0) = 0.5 \).

Now, to show that \( \hat{\sigma}^2_n \) can be a biased estimator, we explicitly compute \( \mathbb{E}_\beta[\hat{\sigma}^2_n] \) for the above setting, when \( n = 2 \). The set of possible actions chosen by \( \beta \) when \( n = 2 \) can be \( \{(a, a), (a, b), (b, a), (b, b)\} \), each of which is equally likely and occurs with probability 1/4. Before computing variance using each of these possible outcomes, recall from (1),

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

For the case where sampled actions are \((a, a)\), the importance ratio are \((2, 2)\) and

\[
\hat{\sigma}^2_n = \frac{2}{2-1} \left( 2 \times 1 - \frac{1}{2} (2 \times 1 + 2 \times 1) \right)^2 = 0.
\]

Similarly, for the cases where sampled actions are \((a, b)\) or \((b, a)\), the importance ratios are \((2, 0)\) or \((0, 2)\) respectively, and

\[
\hat{\sigma}^2_n = \frac{2 \times 1 - \frac{1}{2} (2 \times 1 + 0))^2 + (0 - \frac{1}{2} (0 + 2 \times 1))^2}{2 - 1} = 1 + 1 = 2.
\]

For the cases where sampled actions are \((b, b)\) the importance ratios are \((0, 0)\) respectively, and \( \hat{\sigma}^2_n = 0 \). Therefore, the expected value of \( \hat{\sigma}^2_n = \frac{1}{4} (0 + 2 + 2 + 0) = 1 \neq 0 \), and it is a biased estimator.

\( \square \)

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

Proof. We begin by expanding (1),

\[
\begin{align*}
\lim_{n \to \infty} \hat{\sigma}^2_n &= \lim_{n \to \infty} \frac{1}{n-1} \sum_{i=1}^{n} \left( \rho_i G_i - \frac{1}{n} \sum_{j=1}^{n} \rho_j G_j \right)^2 \\
&= \lim_{n \to \infty} \frac{1}{n-1} \sum_{i=1}^{n} \left( \rho_i^2 G_i^2 - \frac{2}{n} (\rho_i G_i) \sum_{j=1}^{n} \rho_j G_j + \left( \frac{1}{n} \sum_{j=1}^{n} \rho_j G_j \right)^2 \right) \\
&= \lim_{n \to \infty} \left( \frac{1}{n-1} \sum_{i=1}^{n} \rho_i^2 G_i^2 - 2 \left( \frac{1}{n-1} \sum_{i=1}^{n} \rho_i G_i \right) \left( \frac{1}{n} \sum_{j=1}^{n} \rho_j G_j \right) + \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{1}{n} \sum_{j=1}^{n} \rho_j G_j \right)^2 \right) .
\end{align*}
\]
Notice, (a) $1/(n - 1) = (1/n)(n/(n - 1))$ and in the limit $n/(n - 1) \to 1$, therefore $1/(n - 1) \to 1/n$, and (b) using Kolmogorov’s strong law of large numbers, $\frac{1}{n} \sum_{j=1}^{n} \rho_{j} G_{j} \overset{a.s.}{\to} \mathbb{E}[\rho G] = \mathbb{E}[G]$, and $\frac{1}{n} \sum_{i=1}^{n} \rho_{i}^{2} G_{i} \overset{a.s.}{\to} \mathbb{E}[\rho^{2} G^{2}] = \mathbb{E}[\rho G^{2}]$. Therefore using Slutsky’s theorem, (4) can be simplified to,

$$
\lim_{n \to \infty} \hat{\sigma}^{2l}_{n} \xrightarrow{a.s.} \mathbb{E}[\rho G^{2}] - \mathbb{E}[G]^{2}.
$$

(8)

Since $\sigma^{2}(\pi) = \mathbb{V}_{\pi}(G) = \mathbb{E}_{\pi}[G^{2}] - \mathbb{E}_{\pi}[G]^{2}$, (8) may not be equal to $\sigma^{2}(\pi)$.

□

**B: Proofs for the Naïve Estimator $\hat{\sigma}^{2l}_{n}$**

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

**Proof.** This proof uses the same counter-example presented in the proof of Property 1. Recall from (3) that,

$$
\hat{\sigma}^{2l}_{n} = \frac{1}{n - 1} \sum_{i=1}^{n} \rho_{i} \left( G_{i} - \frac{1}{n} \sum_{j=1}^{n} \rho_{j} G_{j} \right)^{2}.
$$

For the case where sampled actions are $(a, a)$, the importance ratios are $(2, 2)$ and

$$
\hat{\sigma}^{2l}_{n} = \frac{2}{2 - 1} \left( 2 \times \left( 1 - \frac{1}{2} (2 \times 1 + 2 \times 1) \right) \right)^{2} = 4.
$$

Similarly, for the cases where sampled actions are $(a, b)$ or $(b, a)$, the importance ratios are $(2, 0)$ or $(0, 2)$ respectively, and

$$
\hat{\sigma}^{2l}_{n} = \frac{(2 \times (1 - \frac{1}{2} (2 \times 1 + 0) + 0) + 0 \times (0 - \frac{1}{2} (2 \times 1 + 0)))}{2 - 1} = 0 + 0 = 0.
$$

For the cases where sampled actions are $(b, b)$ the importance ratios are $(0, 0)$ respectively, and $\hat{\sigma}^{2l}_{n} = 0$. Therefore, the expected value of $\hat{\sigma}^{2l}_{n} = \frac{1}{4} (4 + 0 + 0 + 0) = 1 \neq 0$, and it is a biased estimator.

□

**Property 4.** Under Assumption 1 $\hat{\sigma}^{2l}_{n}$ is a consistent estimator of $\sigma^{2}(\pi)$. That is, $\hat{\sigma}^{2l}_{n} \overset{a.s.}{\to} \sigma^{2}(\pi)$.

**Proof.** We begin by expanding (1).

$$
\lim_{n \to \infty} \hat{\sigma}^{2l}_{n} = \lim_{n \to \infty} \frac{1}{n - 1} \sum_{i=1}^{n} \rho_{i} \left( G_{i} - \frac{1}{n} \sum_{j=1}^{n} \rho_{j} G_{j} \right)^{2}
\begin{align*}
&= \lim_{n \to \infty} \frac{1}{n - 1} \sum_{i=1}^{n} \left( \rho_{i} G_{i} - \frac{2}{n} (\rho_{i} G_{i}) \sum_{j=1}^{n} \rho_{j} G_{j} + \frac{1}{n} \sum_{j=1}^{n} \rho_{j} G_{j} \right)^{2} \\
&= \lim_{n \to \infty} \left( \frac{1}{n - 1} \sum_{i=1}^{n} \rho_{i} G_{i}^{2} - \frac{2}{n - 1} \sum_{i=1}^{n} \rho_{i} G_{i} \right) \left( \frac{1}{n - 1} \sum_{j=1}^{n} \rho_{j} G_{j} \right) + \left( \frac{1}{n - 1} \sum_{i=1}^{n} \rho_{i} G_{i} \right) \left( \frac{1}{n - 1} \sum_{j=1}^{n} \rho_{j} G_{j} \right)^{2}. \tag{9}
\end{align*}
$$

Notice, (a) $1/(n - 1) = (1/n)(n/(n - 1))$ and in the limit $n/(n - 1) \to 1$, therefore $1/(n - 1) \to 1/n$, and (b) using Kolmogorov’s strong law of large numbers, $\frac{1}{n} \sum_{j=1}^{n} \rho_{j} G_{j} \overset{a.s.}{\to} \mathbb{E}[\rho G] = \mathbb{E}[G]$, further $\frac{1}{n} \sum_{j=1}^{n} \rho_{j} G_{j}^{2} \overset{a.s.}{\to} \mathbb{E}[\rho^{2} G^{2}] = \mathbb{E}[\rho G^{2}]$, and $\frac{1}{n} \sum_{i=1}^{n} \rho_{i} \overset{a.s.}{\to} \mathbb{E}_{\beta}[\rho] = 1$. Therefore using Slutsky’s theorem, (9) can be simplified to,

$$
\lim_{n \to \infty} \hat{\sigma}^{2l}_{n} \overset{a.s.}{\to} \mathbb{E}_{\pi}[G^{2}] - \mathbb{E}_{\pi}[G]^{2}
= \mathbb{V}_{\pi}(G)
= \sigma^{2}(\pi).
$$

□
C: Proofs for the Proposed Estimator $\hat{\sigma}_n^2$

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

**Proof.**

\[
\mathbb{E}_\beta[\hat{\sigma}_n^2] = \mathbb{E}_\beta \left[ \frac{1}{n} \sum_{i=1}^{n} \rho_i G_i^2 - \left( \frac{1}{|D_1|} \sum_{i=1}^{|D_1|} \rho_i G_i \right) \left( \frac{1}{|D_2|} \sum_{i=1}^{|D_2|} \rho_i G_i \right) \right]
= \mathbb{E}_\beta \left[ \frac{1}{n} \sum_{i=1}^{n} \rho_i G_i^2 \right] - \mathbb{E}_\beta \left[ \left( \frac{1}{|D_1|} \sum_{i=1}^{|D_1|} \rho_i G_i \right) \right] \mathbb{E}_\beta \left[ \left( \frac{1}{|D_2|} \sum_{i=1}^{|D_2|} \rho_i G_i \right) \right]
= \mathbb{E}_\beta \left[ \rho G^2 \right] - \mathbb{E}_\beta \left[ \rho G \right] \mathbb{E}_\beta \left[ \rho G \right]
= \mathbb{E}_\pi \left[ G^2 \right] - \mathbb{E}_\pi \left[ G \right] \mathbb{E}_\pi \left[ G \right]
= \sigma^2(\pi).
\]

**Example of negative variance estimate:** For completeness, we also work out the expected estimate of $\hat{\sigma}_n^2$ for the counter-example used to show $\hat{\sigma}_n^2$ and $\bar{\sigma}_n^2$ are biased. This example also shows that the $\bar{\sigma}_n$ can be negative. For the case where sampled actions are $(a, a)$, the importance ratio are $(2, 2)$ and $\bar{\sigma}_n$ is negative,

\[
\hat{\sigma}_n^2 = \frac{1}{2}(2 \times 1 + 2 \times 1) - (2 \times 1)(2 \times 1) = -2.
\]

Similarly, for the cases where sampled actions are $(a, b)$ or $(b, a)$, the importance ratios are $(2, 0)$ or $(0, 2)$ respectively, and

\[
\hat{\sigma}_n^2 = \frac{1}{2}(2 \times 1 + 0) - (2 \times 1) \times (0) = 1.
\]

For the cases where sampled actions are $(b, b)$, the importance ratios are $(0, 0)$ respectively, and $\hat{\sigma}_n^2 = 0$. Therefore, the expected value of $\hat{\sigma}_n^2 = \frac{1}{2}(-2 + 1 + 1 + 0) = 0$, as required.

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

**Proof.**

\[
\lim_{n \to \infty} \hat{\sigma}_n^2 = \lim_{n \to \infty} \left[ \frac{1}{n} \sum_{i=1}^{n} \rho_i G_i^2 - \left( \frac{1}{|D_1|} \sum_{i=1}^{|D_1|} \rho_i G_i \right) \left( \frac{1}{|D_2|} \sum_{i=1}^{|D_2|} \rho_i G_i \right) \right]. \tag{10}
\]

Now using Kolmogorov’s strong law of large numbers and Slutsky’s theorem, (10) can be simplified to

\[
\lim_{n \to \infty} = \left[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \rho_i G_i^2 \right] - \left[ \lim_{n \to \infty} \left( \frac{1}{|D_1|} \sum_{i=1}^{|D_1|} \rho_i G_i \right) \right] \left[ \lim_{n \to \infty} \left( \frac{1}{|D_2|} \sum_{i=1}^{|D_2|} \rho_i G_i \right) \right]
\xrightarrow{a.s.} \mathbb{E}_\beta \left[ \rho G^2 \right] - \mathbb{E}_\beta \left[ \rho G \right] \mathbb{E}_\beta \left[ \rho G \right]
= \mathbb{E}_\pi \left[ G^2 \right] - \mathbb{E}_\pi \left[ G \right] \mathbb{E}_\pi \left[ G \right]
= \mathbb{V}_\pi \left[ G \right]
= \sigma^2(\pi).
\]

**Theorem 3.** Under Assumption 1,

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

\[\begin{align*}
E_\beta[\rho G^2] &= E_\beta \left[ \rho(0, T) \left( \sum_{i=0}^{T} \gamma^i R_{(i)} \right) \right]^2 \\
&= E_\beta \left[ \rho(0, T) \left( \sum_{i=0}^{T} \gamma^{2i} R_{(i)}^2 + 2 \sum_{i=0}^{T} \sum_{j=i+1}^{T} \gamma^{i+j} R_{(i)} R_{(j)} \right) \right] \\
&= \left( \sum_{i=0}^{T} E_\beta \left[ \rho(0, T) \gamma^{2i} R_{(i)}^2 \right] \right) + \left( 2 \sum_{i=0}^{T} \sum_{j=i+1}^{T} E_\beta \left[ \rho(0, T) \gamma^{i+j} R_{(i)} R_{(j)} \right] \right) \\
&= \sum_{i=0}^{T} E_\beta \left[ \rho(0, T) \gamma^{2i} R_{(i)}^2 \right] + 2 \sum_{i=0}^{T} \sum_{j=i+1}^{T} E_\beta \left[ \rho(0, T) \gamma^{i+j} R_{(i)} R_{(j)} \right] \\
&= \left( \sum_{i=0}^{T} \sum_{h \in H_0^\beta} \Pr(h|\beta) \rho(0, T) \gamma^{2i} R_{(i)}^2 \right) + \left( 2 \sum_{i=0}^{T} \sum_{j=i+1}^{T} \sum_{h \in H_0^\beta} \Pr(h|\beta) \rho(0, T) \gamma^{i+j} R_{(i)} R_{(j)} \right) \\
&= \left( \sum_{i=0}^{T} \sum_{h' \in H_{i+1}^\beta} \Pr(h'|h, \pi) \gamma^{i} R_{(i)}^2 \right) + \left( 2 \sum_{i=0}^{T} \sum_{j=i+1}^{T} \sum_{h' \in H_{j+1}^\beta} \Pr(h'|h, \pi) \gamma^{i+j} R_{(i)} R_{(j)} \right) \\
&= \sum_{i=0}^{T} E_\beta \left[ \rho(0, i) \gamma^{2i} R_{(i)}^2 \right] + \left( 2 \sum_{i=0}^{T} \sum_{j=i+1}^{T} E_\beta \left[ \rho(0, j) \gamma^{i+j} R_{(i)} R_{(j)} \right] \right) \\
&= \sum_{i=0}^{T} \sum_{j=0}^{T} \rho(0, i) \gamma^{2i} R_{(i)}^2 + 2 \sum_{i=0}^{T} \sum_{j=i+1}^{T} \rho(0, j) \gamma^{i+j} R_{(i)} R_{(j)} \\
&= \sum_{i=0}^{T} \sum_{j=0}^{T} \rho(0, \max(i, j)) \gamma^{i+j} R_{(i)} R_{(j)} ,
\end{align*}\]

where steps (a) and (b) follow due to Assumption 1.

\[\square\]

Proofs for HCOVE using Concentration Inequalities

Theorem 4. Under Assumption 1 if \((\delta_1 + \delta_2 + \delta_3 + \delta_4) \leq \delta\), then for the confidence interval \(C := [v^{ib}, v^{ub}]\),

\[\Pr(\sigma^2(\pi) \in C) \geq 1 - \delta.\]

Proof. For brevity, let \(X := E_\beta[\rho G^2]\), and \(Y := E_\beta[\rho G^2]^2\).
\[
\Pr(v^{1b} \leq \sigma^2(\pi)) = \Pr(C_1^-(X, \delta_1) - C_1^+(Y^2, \delta_2) \leq X - Y) \\
\geq \Pr((C_1^-(X, \delta_1) \leq X) \cap (Y \leq C_1^+(Y^2, \delta_2))) \\
= 1 - \Pr(((C_1^-(X, \delta_1) \leq X) \cap (Y \leq C_1^+(Y^2, \delta_2))^c) \\
= 1 - \Pr(((C_1^-(X, \delta_1) \leq X)^c \cup (Y > C_1^+(Y^2, \delta_2))) \\
= 1 - \Pr((C_1^-(X, \delta_1) > X) \cup (Y > C_1^+(Y^2, \delta_2))) \\
= 1 - (\delta_1 + \delta_2),
\]

where superscript of \( c \) represents complement. Similarly, it can be shown that \( \Pr(v^{1b} \geq \sigma^2(\pi)) \geq 1 - (\delta_3 + \delta_4) \). Therefore, the maximum probability of failure, i.e., either \( v^{1b} < \sigma^2(\pi) \) or \( v^{1b} > \sigma^2(\pi) \), is less than \( (\delta_1 + \delta_2 + \delta_3 + \delta_4) \), which is not greater than \( \delta \). \( \square \)

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

\[
C_1^+(E_\beta[\rho X], \delta) = C_1^+(E_\beta[\rho(X - \xi)], \delta) + \xi.
\]

**Proof.**

\[
C_1^+(E_\beta[\rho X], \delta) = C_1^+(E_\beta[\rho X - \rho \xi + \rho \xi], \delta) \\
= (a) C_1^+(E_\beta[\rho(X - \xi)] + E_\beta[\rho \xi], \delta) \\
= (b) C_1^+(E_\beta[\rho(X - \xi)] + \xi, \delta) \\
= (c) C_1^+(E_\beta[\rho(X - \xi)], \delta) + \xi,
\]

where (a) and (c) follow because \( \xi \) is a fixed constant, and (b) follows because \( E_\beta[\rho] = 1 \). \( \square \)

**Theorem 6.** Under Assumption \( \mathbf{1} \), for any \( \delta \in [0, 0.5] \), let \( \xi_R := \max(R_{\text{max}, i}, R_{\text{max}, j}) \) and \( \xi_G := \max(G_{\text{max}, i}, G_{\text{max}, j}) \) then

\[
X := \sum_{i=0}^{T} \sum_{j=0}^{T} \rho(0, \text{max}(i, j)) \gamma^{i+j} (R_{i,j} - \xi_R), \\
Y := \sum_{i=0}^{T} \rho(0, i) \gamma^i (R_{i} - R_{\text{max}}),
\]

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

\[
C_1^+(E_\beta[\rho G^2], \delta) = C_1^+(E_\beta[X], \delta) + \xi_G, \\
C_1^+(E_\beta[\rho G^2], \delta) = C_1^+(E_\beta[Y], \delta) + G_{\text{max}}.
\]

**Proof.** Let \( c := (1 - \gamma^T)/(1 - \gamma) \). Then \( E_\beta[x] \) is,

\[
E_\beta[X] = E_\beta \left[ \sum_{i=0}^{T} \sum_{j=0}^{T} \rho(0, \text{max}(i, j)) \gamma^{i+j} (R_{i,j} - (\xi_R)) \right] \\
= E_\beta \left[ \sum_{i=0}^{T} \sum_{j=0}^{T} \rho(0, \text{max}(i, j)) \gamma^{i+j} (\xi_R) \right] \\
\leq (a) E_\beta [\rho G^2] - \xi_R \sum_{i=0}^{T} \sum_{j=0}^{T} \gamma^{i+j} E_\beta [\rho (0, \text{max}(i, j))] \\
\leq (b) E_\beta [\rho G^2] - \xi_R \sum_{i=0}^{T} \gamma^i \sum_{j=0}^{T} \gamma^j \\
= E_\beta [\rho G^2] - c^2 \xi_R, \\
\]

where (a) follows from Theorem \( \mathbf{3} \) and (b) follows because \( E_\beta [\rho (0, \text{max}(i, j))] = 1 \). Notice that as \( c^2 \xi_R \) is equivalent to \( \xi_G \), therefore substituting it into (11) gives,

\[
C_1^+(E_\beta[X], \delta) + \xi_G = C_1^+(E_\beta[\rho G^2] - \xi_G, \delta) + \xi_G. \\
= C_1^+(E_\beta[\rho G^2], \delta). 
\]
Similarly,
\[
E_\beta[Y] = E_\beta \left[ \sum_{i=0}^{T} \rho(0, i) \gamma^i R(i) \right] - E_\beta \left[ \sum_{i=0}^{T} \rho(0, i) \gamma^i R_{\text{max}} \right]
\]
\[
= E_\beta \left[ \rho G \right] - R_{\text{max}} \sum_{i=0}^{T} \gamma^i E_\beta \left[ \rho(0, i) \right]
\]
\[
(c) = E_\beta \left[ \rho G \right] - cR_{\text{max}},
\]
where (c) follows because \( E_\beta \left[ \rho(0, i) \right] = 1 \). Further, as \( cR_{\text{max}} = G_{\text{max}} \),
\[
\mathrm{CI}^+ (E_\beta[Y], \delta) + G_{\text{max}} = \mathrm{CI}^+ (E_\beta \left[ \rho G \right] - G_{\text{max}}, \delta) + G_{\text{max}}.
\]
\[
= \mathrm{CI}^+ (E_\beta \left[ \rho G \right], \delta).
\]

\[\square\]

### E. Proofs for HCOVE using Statistical Bootstrapping

**Theorem 7.** Under Assumptions 1 and 2, the confidence interval \( \hat{C} \) has a finite sample error of \( O(n^{-1/2}) \). That is,
\[
\mathrm{Pr} \left( \sigma^2(\pi) \in \hat{C} \right) \geq 1 - \delta - O \left( n^{-1/2} \right).
\]

**Proof.** This proof directly leverages the finite-sample coverage error result by [Efron and Tibshirani (1994)](https://www.jstor.org/stable/2533046). A similar technique has been used by [Kostrikov and Nachum (2020)](https://www.jmlr.org/papers/v21/20-001.html) for establishing the finite-sample coverage error of the CIs for the the mean return \( \mu(\pi) \) using off-policy data. Our result is inspired by theirs and establishes finite-sample coverage error of the CIs for the variance of returns, \( \sigma^2(\pi) \). Before proceeding, we first define some additional notation and then review Hadamard differentiability ([Wasserman (2006)](https://www.springer.com/gp/book/9780387492432), which is a key property for establishing the validity of bootstrap.

For brevity, let \( \mathcal{H} := \mathcal{H}_{(0)}(T) \). For a trajectory \( x \in \mathcal{H} \), let \( \rho(x) \) be the importance ratio of the entire trajectory, \( g_1(x) := G \) be the return, and \( g_2(x) := G^2 \) be the return squared. Considering a finite set of possible trajectories \( \mathcal{H} \), for a given set of trajectories \( \mathcal{D} \), let the empirical distribution over the trajectories be,
\[
d^D(x) := \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \delta_{\{x \}}.
\]

**Hadamard Differentiability:** Suppose \( F \) is a functional mapping distributions \( \mathcal{P} \) over trajectories to \( \mathbb{R} \). Denote \( \mathcal{P}_L \) as the linear space generated by \( \mathcal{P} \). The functional \( F \) is said to be Hadamard differentiable at \( d^D \in \mathcal{P} \) if there exists a linear functional \( L_D \) on \( \mathcal{P}_L \) such that for any \( \epsilon_n \to 0 \) and \( \{P, P_1, P_2, P_3, \ldots \} \subset \mathcal{P}_L \) such that \( \|P_n - P\|_\infty \to 0 \) and \( d^D + \epsilon_n P_n \in \mathcal{P} \),
\[
\lim_{n \to \infty} \left| \frac{F(d^D + \epsilon_n P_n) - F(d^D)}{\epsilon_n} - L_D(P) \right| = 0.
\]

In the following, we directly leverage the finite sample coverage error rate established for bootstrap [Efron and Tibshirani (1994)](https://www.jstor.org/stable/2533046) by considering the functional \( F \) to be our estimator \( \hat{\sigma}^2_n \) and showing that \( F \) is Hadamard differentiable for all \( d^D \). To make the dependence of \( \hat{\sigma}^2_n(d^D) \) explicit, we write \( \hat{\sigma}^2_n(d^D) \) instead of \( \hat{\sigma}^2_n \). Now using (4),
\[
\hat{\sigma}^2_n(d^D + \epsilon_n P_n) = \sum_{x \in \mathcal{H}} \left( d^D + \epsilon_n P_n \right)(x) \rho(x) g_2(x) - \left( \sum_{x \in \mathcal{H}} \left( d^{D_1} + \epsilon_n P_n \right)(x) \rho(x) g_1(x) \right) \left( \sum_{y \in \mathcal{H}} \left( d^{D_2} + \epsilon_n P_n \right)(y) \rho(y) g_1(y) \right),
\]
\[
(12)
\]
\[
\hat{\sigma}^2_n(d^D) = \sum_{x \in \mathcal{H}} d^D(x) \rho(x) g_2(x) - \left( \sum_{x \in \mathcal{H}} d^{D_1}(x) \rho(x) g_1(x) \right) \left( \sum_{y \in \mathcal{H}} d^{D_2}(y) \rho(y) g_1(y) \right).
\]
(13)

Using (12) and (13), as \( n \to \infty \) then \( \epsilon_n \to 0 \),
\[
\lim_{\epsilon_n \to 0} \frac{\hat{\sigma}^2_n(d^D + \epsilon_n P_n) - \hat{\sigma}^2_n(d^D)}{\epsilon_n} = \sum_{x \in \mathcal{H}} P_n(x) \rho(x) g_2(x)
\]
\[
- \sum_{x \in \mathcal{H}} \sum_{y \in \mathcal{H}} \rho(x) g_1(x) \rho(y) g_1(y) \left[ d^{D_1}(x) P_n(y) + d^{D_2}(y) P_n(x) + \epsilon_n P_n(x) P_n(y) - 0 \text{ as } \epsilon_n \to 0 \right].
\]
(14)

It can be seen that (14) is linear in \( P_n \), so there exists a linear functional \( L_D \) on \( P_L \) such that \( \hat{\sigma}^2_n(d^D) \) is Hadamard differentiable.
Algorithm 2: HCOVE - upper bound (Concentration Inequality)

1 Input: Dataset $D$, Confidence level $1 - \delta$

# Upper Bound on $E[|\rho G|^2]$
2 $\xi_R = \max(R_{\min}^2, R_{\max}^2)$ and $\xi_G = \max(G_{\min}^2, G_{\max}^2)$
3 $X_i = \sum_{j=0}^{T} \sum_{k=0}^{T} \rho_i(0, \max(j, k)) \gamma^{j+k} (R_{i(j)} R_{i(k)} - \xi_R)$
4 $X^{\text{ub}} = \text{CI}^+ \left( \{X_i\}_{i=1}^n, \delta/2 \right) + \xi_G$

# Upper and Lower Bounds on $E[\rho G]$
5 $Y_i^- = \sum_{j=0}^{T} \rho_i(0, j) \gamma^{j} (R_{i(j)} - R_{\min})$,
6 $Y_i^+ = \sum_{j=0}^{T} \rho_i(0, j) \gamma^{j} (R_{i(j)} - R_{\max})$,
7 $Y^{\text{lb}} = \text{CI}^- \left( \{Y_i^\text{ }_i\}_{i=1}^n, \delta/4 \right) + G_{\min}$
8 $Y^{\text{ub}} = \text{CI}^+ \left( \{Y_i^\text{ }_i\}_{i=1}^n, \delta/4 \right) + G_{\max}$

# Lower Bound on $E[\rho G]^2$ using bound propagation
9 if $Y^{\text{lb}} \leq 0 \leq Y^{\text{ub}}$ then
10 $Z^{\text{lb}} = 0$
11 else
12 $Z^{\text{lb}} = \min(\{Y^{\text{lb}}, Y^{\text{ub}}\})$
13 Return $X^{\text{ub}} - Z^{\text{lb}}$

F. Algorithms

In this section we present the algorithms to obtain high-confidence bounds for $\sigma^2(\pi)$. Algorithms 2, 3 provide lower and upper bounds using concentration inequalities. Algorithm 4 provides lower and upper bounds using statistical bootstrapping. In the following, we briefly review the concentration inequality established by Thomas, Theocharous, and Ghavamzadeh (2015), which we also use in Algorithms 2, 3.

Theorem 8 (Thomas, Theocharous, and Ghavamzadeh [2015]). Let $\{X_i\}_{i=1}^n$ be $n$ independent real-valued bounded random variables such that for each $i \in \{1, \ldots, n\}$, we have $\Pr(0 \leq X_i) = 1$, $E[X_i] \leq \mu$, and the fixed real-valued threshold $c_i > 0$. Let $\delta > 0$ and $Y_i := \min(X_i, c_i)$.

$$\text{CI}^- \left( \{X_i\}_{i=1}^n, \delta \right) := \left( \sum_{i=1}^{\infty} \frac{1}{c_i} \right)^{-1} \sum_{i=1}^{\infty} \frac{Y_i}{c_i} - \left( \sum_{i=1}^{\infty} \frac{1}{c_i} \right)^{-1} \frac{7n \ln(2/\delta)}{3(n-1)} - \left( \sum_{i=1}^{\infty} \frac{1}{c_i} \right)^{-1} \sqrt{\frac{\ln(2/\delta)}{n-1} \sum_{i,j=1}^{\infty} \left( \frac{Y_i}{c_i} - \frac{Y_j}{c_j} \right)^2} \quad (15)$$

Then with probability at least $1 - \delta$, we have $\mu \geq \text{CI}^- \left( \{X_i\}_{i=1}^n, 1 \right)$.

Similarly, let $\{A_i\}_{i=1}^n$ be $n$ independent real-valued bounded random variables where $\Pr(0 \geq A_i) = 1$ and $E[A_i] \geq \mu$, then for a fixed real-valued threshold $c_i < 0$ and $B_i := \max(A_i, c_i)$, the expected value $E[B_i] \geq E[A_i]$. Therefore, an upper bound on $E[B_i]$ is also an upper bound on $E[A_i]$. Consequently, to get an upper bound on $E[B_i]$ we flip the bound in Theorem 8 (i.e., let $Y_i := -B_i$ in (15) and then negate the resulting bound, since $\forall \nu$, $\Pr(\mathbb{E}[Y_i] \geq \nu) = \Pr(-\mathbb{E}[B_i] \geq \nu) = \Pr(\mathbb{E}[B_i] \leq -\nu)$).

$$\text{CI}^+ \left( \{A_i\}_{i=1}^n, \delta \right) := \left( \sum_{i=1}^{\infty} \frac{1}{c_i} \right)^{-1} \sum_{i=1}^{\infty} \frac{B_i}{c_i} + \left( \sum_{i=1}^{\infty} \frac{1}{c_i} \right)^{-1} \frac{7n \ln(2/\delta)}{3(n-1)} + \left( \sum_{i=1}^{\infty} \frac{1}{c_i} \right)^{-1} \sqrt{\frac{\ln(2/\delta)}{n-1} \sum_{i,j=1}^{\infty} \left( \frac{B_i}{c_i} - \frac{B_j}{c_j} \right)^2} \quad (16)$$

Then with probability at least $1 - \delta$, we have $\mu \leq \text{CI}^+ \left( \{A_i\}_{i=1}^n, 1 \right)$.

In (15), $c_i$’s help in truncating the upper tail of the distribution, and in (16), $c_i$’s help in truncating the lower tail of the distribution. Further, note the use of absolute values for $c_i$’s in our presentation of the bounds by Thomas, Theocharous, and Ghavamzadeh (2015); while in (16) this is redundant as $c_i > 0$, in (15) this is important to prevent the change in sign of the random variable when normalized using $c_i$’s as in this equation $c_i < 0$. For simplicity, Thomas, Theocharous, and Ghavamzadeh (2015) suggest setting a common $c$ for all $c_i$’s. Further, since the value of $c$ should be chosen independent of that data being analyzed, they suggest partitioning the data into two sets $D_{\text{pre}}$ and $D_{\text{post}}$ in the ratio $1/20 : 19/20$ and searching the value of $c$ that optimizes the bound on the data from $D_{\text{pre}}$. The value of this $c$ is then used to get the desired bounds using data from $D_{\text{post}}$. We refer the readers to the work by Thomas, Theocharous, and Ghavamzadeh (2015) for more details.
Algorithm 3: HCOVE - lower bound (Concentration Inequality)

1. **Input:** Dataset $D$, Confidence level $1 - \delta$
   
   # Lower Bound on $\mathbb{E}_{\beta}[\rho G^2]$
   (Control variate is not needed as $\rho G^2$ is always positive.)

2. $X_i = \sum_{j=0}^{T} \sum_{k=0}^{T} \rho_i (0, \max(j, k)) \gamma^{j+k} (R_{i(j)} R_{i(k)})$

3. $X^{lb} = \mathbb{C}^{-} \left( \{X_i\}_{i=1}^{D}, \delta/2 \right)$
   
   # Upper and Lower Bounds on $\mathbb{E}_{\beta}[\rho G]$

4. $Y_i^- = \sum_{j=0}^{T} \rho_i (0, j) \gamma^{j} (R_{i(j)} - R_{\min})$

5. $Y_i^+ = \sum_{j=0}^{T} \rho_i (0, j) \gamma^{j} (R_{i(j)} - R_{\max})$

6. $Y^{lb} = \mathbb{C}^{-} \left( \{Y_i^+\}_{i=1}^{D}, \delta/4 \right) + G_{\min}$

7. $Y^{ub} = \mathbb{C}^{-} \left( \{Y_i^-\}_{i=1}^{D}, \delta/4 \right) + G_{\max}$

# Upper Bound on $\mathbb{E}_{\beta}[\rho G]^2$ using bound propagation

8. $Z^{ub} = \max \{(Y^{lb})^2, (Y^{ub})^2\}$

9. **Return** $X^{lb} - Z^{ub}$

Algorithm 4: HCOVE - upper and lower bounds (Bootstrap)

1. **Input:** Dataset $D$, Confidence level $1 - \delta$

2. Compute variance estimate $\hat{\sigma}_n^2 = \text{Algorithm 1}(D)$

3. Compute $B$ bootstrapped datasets $\{D_{i}^*\}_{i=1}^{B}$

4. Compute bootstrapped variance estimates $\{\hat{\sigma}_{n,i}^2\}_{i=1}^{B}$ using Algorithm 1 for $\{D_{i}^*\}_{i=1}^{B}$.

5. Compute $\delta/2$ and $1 - \delta/2$ quantiles $z_{\delta/2}, z_{1-\delta/2}$ of $\{\hat{\sigma}_{n,i}^2 - \hat{\sigma}_n^2\}_{i=1}^{B}$

6. **Return** $\hat{C} := [\hat{\sigma}_n^2 - z_{1-\delta/2}, \hat{\sigma}_n^2 - z_{\delta/2}]$

G. Empirical Details

In this section, we discuss domain details and how $\pi$ and $\beta$ were selected for both the domains.

**Recommender System:** Online recommendation systems are popular for tutorials, movies, advertisements, etc. In all these settings it may be beneficial to consider the user’s experience once the new system/policy is deployed. To abstract such settings, we create a simulated domain where the interest of the user for a finite set of items is represented using the reward for the corresponding item.

We use an actor-critic algorithm [Sutton and Barto 2018] to find a near optimal policy $\pi$, which we use as the evaluation policy. Let $\pi^{\text{rand}}$ be a random policy with uniform distribution over the actions (items). Then for an $\alpha = 0.5$, we define the behavior policy $\beta(a|s) := \alpha \pi(a|s) + (1 - \alpha)\pi^{\text{rand}}(a|s)$ for all states and actions.

**Gridworld:** We also consider a standard $4 \times 4$ Gridworld with stochastic transitions. There are eight discrete actions corresponding to up, down, left, right, and the four diagonal movements. Behavior and the evaluation policy for this domain were obtained in a similar way as discussed for the recommender system domain.

**Diabetes Treatment:** This domain is modeled using an open-source implementation [Xie 2019] of the U.S. Food and Drug Administration (FDA) approved Type-1 Diabetes Mellitus simulator (T1DMS) [Man et al. 2014] for the treatment of Type-1 diabetes. An episode corresponds to a day, and each step of an episode corresponds to a minute in an *in-silico* patient’s body and is governed by a continuous time non-linear ordinary differential equation (ODE) [Man et al. 2014].

To control the insulin injection, which is required for regulating the blood glucose level, we use a parameterized policy based on the amount of insulin that a person with diabetes is instructed to inject prior to eating a meal [Bastański 2014]:

$$\text{injection} = \frac{\text{current blood glucose} - \text{target blood glucose}}{CF} + \frac{\text{meal size}}{CR},$$
where ‘current blood glucose’ is the estimate of the person’s current blood glucose level, ‘target blood glucose’ is the desired blood glucose, ‘meal size’ is the estimate of the size of the meal the patient is about to eat, and $CR \in [CR_{\text{min}}, CR_{\text{max}}]$ and $CF \in [CF_{\text{min}}, CF_{\text{max}}]$ are two real-valued parameters that must be tuned based on the body parameters to make the treatment effective.

The action distribution for the policy is parameterized using a normal distribution $N(\mu, \sigma)$, whose mean $\mu$ is obtained using a sigmoid function (scaled for the desired range), and the standard deviation $\sigma$ is kept fixed. We use an actor-critic algorithm (Sutton and Barto 2018) to find a near optimal policy $\pi$ having normal distribution $N(\mu_\pi, \sigma)$, which we use as the evaluation policy. Let $\pi_{\text{rand}}$ be a random policy parameterized using a normal distribution $N(\mu_\text{rand}, \sigma)$, where $\mu_\text{rand} := [(CR_{\text{max}} - CR_{\text{min}})/2, (CF_{\text{max}} - CF_{\text{min}})/2]$. Then for an $\alpha = 0.5$, we parameterize the behavior policy $\beta$ using a normal distribution $N(\mu_\beta, \sigma)$, where $\mu_\beta := \alpha \mu_\pi + (1 - \alpha) \mu_\text{rand}$.

### Additional Experimental Results

In Figure 5 we present comparison of the two naive estimators, and the proposed estimators (with and without CDIS) on three domains. We see a similar behavior for the diabetes treatment, since the output of the policy corresponds to the parameters of another insulin controlling policy (Bastani 2014), which makes the horizon effectively of length one. The variance reduction benefit of CDIS can be observed in the Gridworld setting, which has a longer horizon.

The output of the naive estimator $\hat{\sigma}_n^{21}$ was outside the limits of the plotted $y$-axis for all the domains and hence it is not visible. The output of the other naive estimator $\hat{\sigma}_n^{21}$ is nearly unbiased for the recommender systems and the Gridworld domains, both of which have discrete actions. For the diabetes treatment domain, it can be observed that $\hat{\sigma}_n^{21}$ results in biased estimates when the number of samples are small.