On the Variance of Conditional Importance Sampling for Off-Policy Evaluation

Yao Liu

Computer Science Department, Stanford

September 30, 2020

(A joint work with Pierre-Luc Bacon and Emma Brunskill)
Sequential Decision Making

1-Step Decision Making

Sequential Decision Making

Figure: Gottesman et al. [2019]

- Key difficulty for sequential case: covariate shift
- Bottleneck of solving this: dependency on planning horizon $T$. 
States: context at each time

Action: treatment choice

- Policy: a map from context to decision, at each step.
- A common algorithm [Sutton and Barto, 2018]: iteration of policy evaluation and improvement.
Evaluating a counterfactual policy

Problem: evaluating an alternative policy from logged data, with Markov assumption. Three types of inverse probability weighting (IPW) estimator:

1. Purely based on propensity ratio [Rubinstein, 1981].
2. Leverages the temporal structure of the problem [Precup et al., 2000].

Which one is the best?

- Empirical results says \( \text{Var}(1) \geq \text{Var}(2) \geq \text{Var}(3) \).
- We will show an “surprising” example then an analysis of the variance.
Three types of inverse probability weighting (IPW) estimator:

1. Inverse probability weighting over **whole history**.
2. Inverse probability weighting over **history until now**.
3. Inverse probability weighting over **current context and decision**.
An “Surprising” Example

Three types of inverse probability weighting (IPW) estimators:

1. Inverse probability weighting over **whole history**.
2. Inverse probability weighting over **history until now**.
3. Inverse probability weighting over **current state and action**.

Figure: Example: \( \text{Var}(1) < \text{Var}(2) < \text{Var}(3) \)
State: \( s_t \in S \).
Action: \( a_t \in A \).
Initial state (pdf): \( p_1(s_1) \).
Transition of state (pdf): \( p(s_{t+1} | s_t, a_t) \).
Reward: \( r_t \in [0, 1] \) depending on \( s_t, a_t, \) and \( t \).
Horizon: \( T \).
Discounted factor: \( \gamma \in [0, 1] \).

Policy \( \pi \): conditional distribution of \( a_t \) given \( s_t \) with p.d.f.: \( \pi(a_t | s_t) \).
Value of policy:

\[
v^\pi = \mathbb{E} \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \mid a_t \sim \pi \right].
\]
Problem: Off-Policy Evaluation

Value of policy:

$$v^\pi = \mathbb{E} \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \mid a_t \sim \pi \right] .$$

Off-Policy Evaluation (OPE) Problem: Estimating $v^\pi$ with data collected from another policy $\mu$.

A T-step trajectory: $\tau_{1:T} = \{(s_t, a_t, r_t)\}_{t=1}^{T}$, where

$$s_1 \sim p_1, a_t \sim \mu(\cdot|s_t), r_t \sim p_r(\cdot|s_t, a_t, t), s_{t+1} \sim p(\cdot|s_t, a_t)$$
Define likelihood ratio terms as:

$$\rho_t = \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}, \quad \rho_{1:T} = \prod_{t=1}^{T} \rho_t.$$ 

The importance sampling estimator [Rubinstein, 1981], introduced to MDP case by Precup et al. [2000]

$$\hat{v}_{IS} = \rho_{1:T} \sum_{t=1}^{T} \gamma^{t-1} r_t$$

is an unbiased estimator when $\rho_t$’s are known.

$$\mathbb{E}_\mu]\rho_{1:T} \sum_{t=1}^{T} \gamma^{t-1} r_t] = \mathbb{E}_\pi \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \right] = v^\pi,$$
Precup et al. [2000]:

\[ \hat{v}_{\text{PDIS}} = \sum_{t=1}^{T} \gamma^{t-1} r_t \rho_{1:t} . \]

This is also an unbiased estimator:

\[ \mathbb{E} [\hat{v}_{\text{PDIS}}] = \sum_{t=1}^{T} \gamma^{t-1} \mathbb{E}_\pi [r_t] = v^\pi \]
Since samples are generated from a Markov chain, we define its marginal distribution and stationary distribution of state action pair as:

\[
d_{t}^{\mu}(s, a) = \int p_{1}(s_{1})\mu(a_{t}|s_{t})
\]

\[
\prod_{k=1}^{t-1} [\mu(a_{k}|s_{k})p(s_{k+1}|s_{k}, a_{k})] ds_{1}da_{1} \ldots ds_{t-1}da_{t-1}
\]

\[
d_{\gamma}^{\mu}(s, a) = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} \gamma^{t}d_{t}^{\mu}(s, a)}{\sum_{t=1}^{T} \gamma^{t}}
\]
If one know $\frac{d_t^\pi(s,a)}{d_t^\mu(s,a)}$ one could get the “stationary” IS estimators [Hallak and Mannor, 2017, Liu et al., 2018, Xie et al., 2019]:

$$\hat{v}_{\text{SIS}} = \sum_{t=1}^{T} \gamma^{t-1} r_t \frac{d_t^\pi(s_t, a_t)}{d_t^\mu(s_t, a_t)} .$$

We assume an oracle of $\frac{d_t^\pi(s,a)}{d_t^\mu(s,a)}$ here, then $\hat{v}_{\text{SIS}}$ is also unbiased. In practice, existing work proposed different estimators of this ratio.
Conditional Monte Carlo: \( \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y] \), \( \text{Var}(\mathbb{E}[Y|X]) \leq \text{Var}(Y) \).

Extended Conditional Monte Carlo: \( Y = \sum_t Y_t \).

\[
\text{Var} \left( \sum_t Y_t \right) - \text{Var} \left( \sum_t \mathbb{E}[Y_t|X_t] \right) \\
= \sum_t \mathbb{E}[Y_t^2] - \mathbb{E} \left[ (\mathbb{E}[Y_t|X_t])^2 \right] \\
\geq 0 \\
+ 2 \sum_{t<k} \mathbb{E}[Y_t Y_k] - \mathbb{E}[\mathbb{E}[Y_t|X_t]\mathbb{E}[Y_k|X_k]] .
\]

No guarantee in general.
By leveraging these observations, we could have unbiased conditional importance estimators:

- If \( \forall t, r_t \) is conditionally independent with \( \rho_{1:t} \) given \( \phi_t \), then:

\[
\nu^\pi = \mathbb{E} \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \rho_{1:T} \right] = \mathbb{E} \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \mathbb{E} \left[ \rho_{1:T} | \phi_t \right] \right].
\]
We show that PDIS and SIS are both conditional importance sampling estimator, with different choices of $\phi_t$.

- **PDIS:** $\phi_t$ is the first $t$ step trajectory $\tau_{1:t}$.
  \[
  \mathbb{E}[\rho_{1:T} | \phi_t] = \mathbb{E}[\rho_{1:T} | \tau_{1:t}] = \rho_{1:t} .
  \]

- **SIS:** $\phi_t$ is state action pair $s_t, a_t$. We show that
  \[
  \mathbb{E}[\rho_{1:T} | \phi_t] = \mathbb{E}[\rho_{1:T} | s_t, a_t] = \mathbb{E}[\rho_{1:t} | s_t, a_t] = \frac{d_\pi^t(s_t, a_t)}{d_\mu^t(s_t, a_t)} .
  \]
Though there is no consistent order in general between the variance of IS, PDIS and SIS, we are still interested in characterizing when a variance reduction can occur. For PDIS v.s. IS, we have the following theorem:

**Theorem 1 (Variance reduction of PDIS)**

If for any $1 \leq t \leq k \leq T$ and initial state $s$, $\rho_{0:k}(\tau)$ and $r_t(\tau)\rho_{0:k}(\tau)$ are positively correlated, $\text{Var}(\hat{v}_{PDIS}) \leq \text{Var}(\hat{v}_{IS})$. 
For SIS v.s. PDIS, we have the following theorem:

**Theorem 2 (Variance reduction of SIS)**

If for any fixed $0 \leq t \leq k < T$,

$$\text{Cov}(\rho_{1:t} r_t, \rho_{0:k} r_k) \geq \text{Cov}\left(\frac{d_{t}^{\pi}(s, a)}{d_{t}^{\mu}(s, a)} r_t, \frac{d_{k}^{\pi}(s, a)}{d_{k}^{\mu}(s, a)} r_k\right)$$

then $\text{Var}(\hat{v}_{\text{SIS}}) \leq \text{Var}(\hat{v}_{\text{PDIS}})$
When $T \to \infty$

How fast does the variance increase with respect to horizon $T$?
### Assumptions Needed for Infinite Horizon Analysis

<table>
<thead>
<tr>
<th>Assumption 1 (Harris ergodic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Markov chain of ( {s_t, a_t} ) under ( \mu ) is Harris ergodic. That is: the chain is aperiodic, ( \psi )-irreducible, and positive Harris recurrent [Meyn and Tweedie, 2012].</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assumption 2 (Drift condition)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(This is equivalent with geometric ergodicity.) There exist an everywhere-finite function ( B : S \times A \mapsto [1, \infty) ), a constant ( \lambda \in (0, 1) ), ( b &lt; \infty ) and a petite ( K \subset S \times A ) such that:</td>
</tr>
</tbody>
</table>

\[
\mathbb{E}_{s', a'|s, a} B(s', a') \leq \lambda B(s, a) + b \mathbbm{1}((s, a) \in K) .
\]

<table>
<thead>
<tr>
<th>Assumption 3 (Bounded likelihood ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
</tr>
</tbody>
</table>
Let \( c = \mathbb{E}_{d^\mu_{\gamma=1}} [D_{KL}(\mu||\pi)] \) be the expected KL distance between \( \pi \) and \( \mu \). We consider cases when \( c > 0 \). We can characterize asymptotic distribution of IS estimator with this constant.

**Theorem 3 (Almost sure convergence with exponential rate)**

Under Assumption 1, for \( \pi \neq \mu \), \( \lim_T (\rho_{1:T})^{1/T} = e^{-c} \), \( \lim_T |\hat{\nu}_{IS}|^{1/T} < e^{-c} \) a.s.
Asymptotic Variance of IS

Theorem 4 (Variance of IS estimator)

Under Assumption 1, 2 and 3, there exist $T_0 > 0$ such that for all $T > T_0$,

$$\text{Var}(\hat{v}_{IS}) \geq \frac{(v^\pi)^2}{4} \exp \left( \frac{Tc^2}{8c_1^2 \| B \|_\infty} \right) - (v^\pi)^2$$

where $B$ is defined in Assumption 2, $c_1$ is some constant. If $U_{\rho} := \sup_s, a \frac{\pi(a|s)}{\mu(a|s)} < \infty$, $\text{Var}(\hat{v}_{IS}) \leq T U_{\rho}^T - (v^\pi)^2$.

- This theorem indicate that variance of IS is exponential for any MDP satisfying assumptions, bounded from both sides.
- The key of LB proof is that $\log \rho_{1:T}$ is a martingale with bounded martingale differences, bounds achieved through the drift inequality.
Asymptotic Variance of PDIS

Theorem 5 (Variance of the PDIS estimator)

Under Assumption 1, 2 and 3, \( \exists T_0 > 0 \) s.t. \( \forall T > T_0 \),

\[
\text{Var}(\hat{v}_{\text{PDIS}}) \geq \sum_{t=T_0}^{T} \gamma^{2t-2}(\mathbb{E}_{\pi}(r_t))^2 \exp\left(\frac{tc^2}{8c_1\|B\|_{\infty}}\right) - (v^\pi)^2
\]

where \( B \), \( c_1 \) and \( c \) are same constants in theorem 4. If \( U_\rho < \infty \),

\[
\text{Var}(\hat{v}_{\text{PDIS}}) \leq T \sum_{t=1}^{T} U_\rho^{2t} \gamma^{2t-2}(\mathbb{E}_{\mu}[r_t])^2 - (v^\pi)^2
\]

- Under mild assumptions (\( \gamma \) close to 1, stationary reward), the LB part of this theorem indicate an exponential LB on variance of PDIS.
- Only if \( \gamma \) is small enough or rewards diminish exponentially fast, the variance of PDIS could by bounded by \( O(T) \).
Theorem 6 (Variance of the SIS estimator)

\[ \text{Var}(\hat{v}_{\text{SIS}}) \leq T \sum_{t=1}^{T} \gamma^{t-1} \left( \mathbb{E} \left[ \left( \frac{d_{t}^{\pi}(s_{t}, a_{t})}{d_{t}^{\mu}(s_{t}, a_{t})} \right)^{2} \right] - 1 \right) \]

- Under mild assumptions (sequence of \( d_{t}^{\pi} \) are continuous), this upper bound is \( O(T^{2}) \) as long as \( \frac{d_{t}^{\pi}}{d_{t}^{\mu}} \) is bounded.
Takeaways

To help shed the light into data efficient off-policy evaluation in Markov decision processes:

- We provide a unified view of recent proposed importance sampling estimators for off-policy evaluation.
- We formalize and clarify some intuitions on the variance reduction of those estimators.
  - $\text{Var}(\text{IS}) \geq \text{Var}(\text{PDIS}) \geq \text{Var}(\text{SIS})$ is not always true.
  - When $T \to \infty$, $\text{Var}(\text{IS})$ is always $\exp(T)$, $\text{Var}(\text{PDIS})$ is $\exp(T)$ in most cases, and $\text{Var}(\text{SIS})$ is $\text{poly}(T)$. 


