Understanding the Curse of Horizon in Off-Policy Evaluation via Conditional Importance Sampling

By: Yao Liu, Pierre-Luc Bacon, Emma Brunskill

Presenters: Jonah Phillion, Sana Tonekaboni

Liu, Y., Bacon, P.L. and Brunskill, E., 2020, November. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *International Conference on Machine Learning* (pp. 6184-6193). PMLR.

On-policy RL

$$\max_{\theta} \quad \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

$\longleftrightarrow$

Off-policy RL

$$\max_{\theta} \quad \mathbb{E}_{\tau \sim \mu}\left[\frac{\pi_\theta(\tau)}{\mu(\tau)} \sum_{t=0}^{T} \gamma^t r_t\right]$$

Target policy: $\pi(a|s)$

Behaviour policy: $\mu(a|s)$

Off-policy and Important sampling

- How to estimate value of a functions under a certain policy distribution, using samples from another distribution.

- **Importance sampling**, is a statistical technique for estimating expected values under one distributions, given samples from another.

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

+ Unbiased estimator
- High variance

Standard importance sampling techniques

1. Crude Importance Sampling (IS)

[Precup et al, "Eligibility Traces for Off-Policy Policy Evaluation", 2000.]

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

2. Per-Decision Importance Sampling (PDIS)

[Precup et al, "Eligibility Traces for Off-Policy Policy Evaluation", 2000.]

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

3. Stationary Importance Sampling (SIS)

[Hallak et al "Consistent On-Line Off-Policy Evaluation", 2017.
Liu et al, "Breaking the curse of horizon: Infinite-horizon off-policy estimation", 2018.]

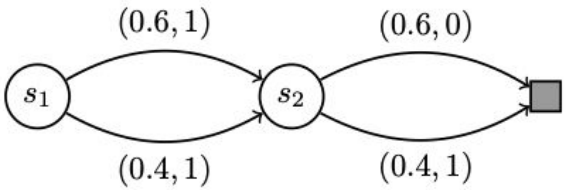
$$\hat{v}_{SIS} = \sum_{t=1}^{T} \frac{d_t^{\pi}(s_t, a_t)}{d_t^{\mu}(s_t, a_t)} \gamma^{t-1} r_t$$

$$d_t^{\mu}(s, a) = Pr(s_t = s, a_t = a | s_1 \sim p_1, a_i \sim \mu(a_i|s_i))$$

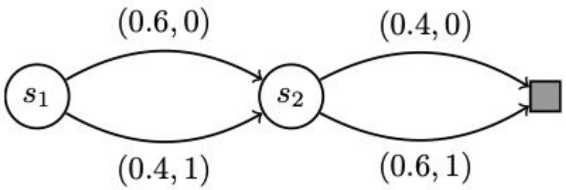
Intuitively, PDIS should be "better" than IS, and SIS should be "better" than PDIS.

Can we prove this?

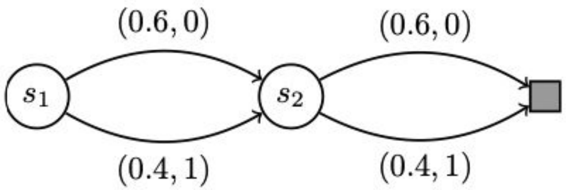
Counterexamples



(a) $\mathrm{Var}(\hat{v}_{\mathrm{IS}}) < \mathrm{Var}(\hat{v}_{\mathrm{SIS}}) < \mathrm{Var}(\hat{v}_{\mathrm{PDIS}})$



(b) $\mathrm{Var}(\hat{v}_{\mathrm{PDIS}}) < \mathrm{Var}(\hat{v}_{\mathrm{SIS}}) < \mathrm{Var}(\hat{v}_{\mathrm{IS}})$



(c) $\mathrm{Var}(\hat{v}_{\mathrm{IS}}) < \mathrm{Var}(\hat{v}_{\mathrm{PDIS}}) < \mathrm{Var}(\hat{v}_{\mathrm{SIS}})$

	IS	**PDIS**	**SIS**
(a)	1.4 +- 0.119	1.4 +- 0.244	1.4 +- 0.1999
(b)	1.0 +- 0.542	1.0 +- 0.452	1.0 +- 0.52
(c)	0.8 +- 0.230	0.8 +- 0.268	0.8 +- 0.32

We can better understand this observation when we note that all estimators are instances of **conditional expectation**.

Conditional Monte-Carlo

According to the **law of total expectation**:

$$\begin{aligned}\mathbb{E}\left[\rho_{1:T}G_T\right] &= \mathbb{E}\left[\mathbb{E}\left[\rho_{1:T}G_T|\phi_T, G_T\right]\right]\\ &= \mathbb{E}\left[G_T\mathbb{E}\left[\rho_{1:T}|\phi_T, G_T\right]\right]\\ &= \mathbb{E}\left[G_T\mathbb{E}\left[\rho_{1:T}|\phi_T\right]\right].\end{aligned}$$

According to the **law of total variance**:

$$\begin{aligned}&\mathrm{Var}\left(G_T\mathbb{E}\left[\rho_{1:T}|\phi_T\right]\right)\\ &= \mathrm{Var}\left(G_T\rho_{1:T}\right) - \mathbb{E}\left[\mathrm{Var}(G_T\phi_{1:T}|\phi_T, G_T)\right]\\ &= \mathrm{Var}\left(G_T\rho_{1:T}\right) - \underbrace{\mathbb{E}\left[G_T^2\mathrm{Var}(\rho_{1:T}|\phi_T)\right]}_{+}\quad.\end{aligned}$$

This is the basis of **conditional Monte-Carlo** as a variance reduction method

Extended Conditional Importance Sampling

Conditioning in a stage-dependant manner rather than with a fixed statistics results in an estimator belonging to **extended conditional monte-carlo** estimator.

$$v^\pi = \mathbb{E}\left[G_T \rho_{1:T}\right] = \sum_{t=1}^{T} \gamma^{t-1} \mathbb{E}\left[\mathbb{E}\left[r_t \rho_{1:t} | \phi_t, r_t\right]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t \mathbb{E}\left[\rho_{1:t} | \phi_t\right]\right]$$

Conditioning history up to time t

$$\mathbb{E}[\rho_{1:T} | \tau_{1:t}] = \rho_{1:t}$$

$$v^\pi = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t \rho_{1:t}\right]$$

PDIS

Conditioning on state and action at time t

$$v^\pi = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t \mathbb{E}\left[\rho_{1:t} | s_t, a_t\right]\right]$$

SIS

Extended Conditional Importance Sampling

Law of total variance no longer implies a variance reduction because the variance is now over a sum of random variables and depends on **interaction of covariance terms across time steps**

$$\mathrm{Var}\left(\sum_{t=1}^{T} r_t w_t\right) = \sum_{t=1}^{T} \mathrm{Var}(r_t w_t) + \sum_{k \neq t} \mathrm{Cov}(r_k w_k, r_t w_t)$$

When would conditioning reduce variance?

Section 5: Finite-Horizon Analysis

IS vs. PDIS

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,* $Var(\hat{v}_{PDIS}) \leq Var(\hat{v}_{IS})$.

=> *PDIS is better if the target policy is more likely to take a trajectory with a higher reward*

Section 5: Finite-Horizon Analysis

PDIS vs. SIS

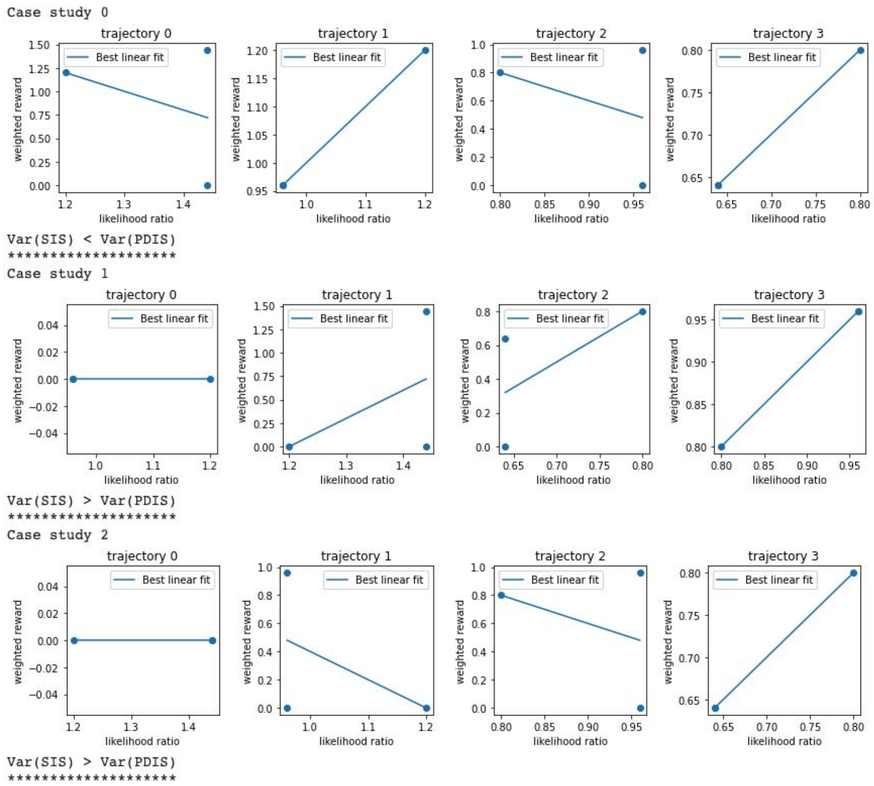
Theorem 2 (Variance reduction of SIS). *If for any fixed* $0 \leq t \leq k < T$,

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

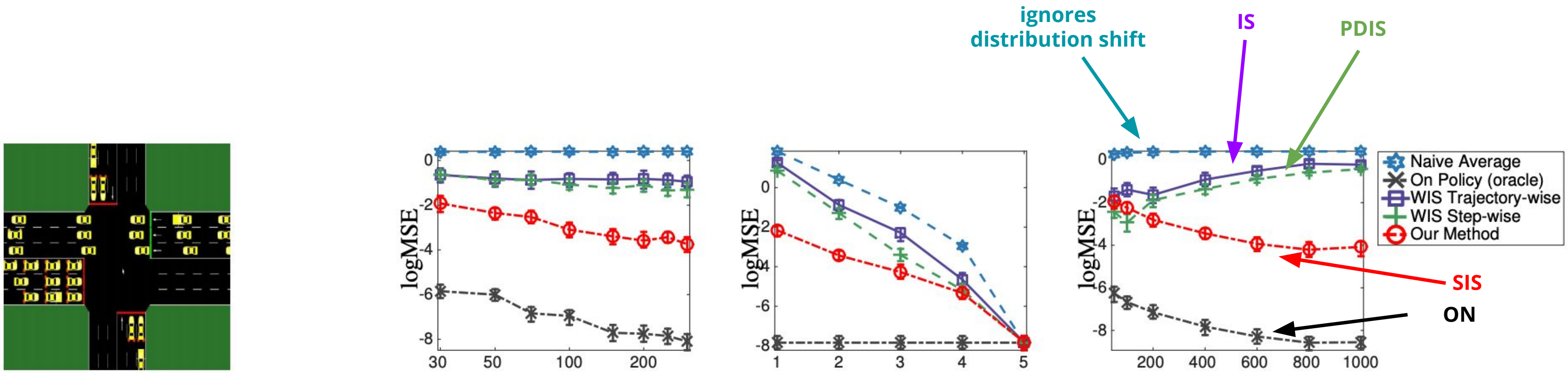
then $Var(\hat{v}_{SIS}) \leq Var(\hat{v}_{PDIS})$

=> *SIS is better for long time horizons in MDPs where high reward early in the MDP is correlated with reward later in the MDP*

Verified: Those bounds explain the results from the 2-state MDP



Section 6: Asymptotic Analysis



(a) Environment (b) # of Trajectories (n) (c) Different Behavior Policies (d) Truncated Length T

Maybe SIS is provably better for large T?

Liu et al, “Breaking the curse of horizon: Infinite-horizon off-policy estimation”, 2018.

Asymptotic Analysis

IS Variance of IS always scales exponentially with T

Theorem 4 (Variance of IS estimator). *Under Assumption 1, 2 and 3, there exist $T_0 > 0$ such that for all $T > T_0$,*

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

where B is defined in Assumption 2, c_1 is some constant defined in lemma 3, $\boxed{c = \mathbb{E}_{d^\mu}[D_{KL}(\mu||\pi)]}$. If $\mathbb{E}_{a\sim\mu}\left[\frac{\pi(a|s)^2}{\mu(a|s)^2}\right] \leq M_\rho^2$ for any s, then $Var(\hat{v}_{IS}) \leq \boxed{T^2 M^{2T} - (v^\pi)^2.}$

Asymptotic Analysis

PDIS Variance of PDIS can be better than quadratic (when the reward decreases fast enough)

Let $U_\rho = \sup_{s,a} \frac{\pi(a|s)}{\mu(a|s)} < \infty$, $Var(\hat{v}_{PDIS}) \le \boxed{T \sum_{t=1}^{T} U_\rho^{2t} \gamma^{2t-2} \mathbb{E}_\mu[r_t^2]} - (v^\pi)^2$.

Corollary 3. *Let* $U_\rho = \sup_{s,a} \frac{\pi(a|s)}{\mu(a|s)}$. *If* $U_\rho \gamma \le 1$ *or* $U_\rho \gamma \lim_T \left(\mathbb{E}_\pi[r_T]\right)^{1/T} < 1$, $\boxed{Var(\hat{v}_{PDIS}) = O(T^2)}$.

PDIS Variance of PDIS can also be worse than exponential (when reward doesn't drop fast enough)

Theorem 5 (Variance of the PDIS estimator)**.** *Under Assumption 1, 2 and 3,* $\exists T_0 > 0$ *s.t.* $\forall T > T_0$,

$$Var(\hat{v}_{PDIS}) \ge \sum_{t=T_0}^{T} \boxed{\frac{\gamma^{2t-2}(\mathbb{E}_\pi(r_t))^2}{4} \exp\left(\frac{tc^2}{8c_1^2\|B\|_\infty}\right)} - (v^\pi)^2$$

Corollary 2. *With theorem 5 holds,* $Var(\hat{v}_{PDIS}) = \boxed{\Omega(\exp(\epsilon T))}$ *if the following conditions hold: 1)* $\gamma \ge \exp\left(\frac{-c^2}{16c_1^2\|B\|_\infty}\right)$*; 2) There exist a* $\epsilon > 0$ *such that*

$$\mathbb{E}_\pi(r_t) = \Omega\left(\exp\left(-t\left(\frac{c^2}{16c_1^2\|B\|_\infty} + \log\gamma - \epsilon/2\right)\right)\right)$$

Asymptotic Analysis

SIS Variance of SIS scales quadratically in general

Theorem 6 (Variance of the SIS estimator).

$$Var\left(\hat{v}_{SIS}\right) \leq T \sum_{t=1}^{T} \gamma^{t-1} \left(\boxed{\mathbb{E}\left[\left(\frac{d_t^{\pi}(s_t, a_t)}{d_t^{\mu}(s_t, a_t)} \right)^2 \right]} - 1 \right)$$

$$\downarrow$$

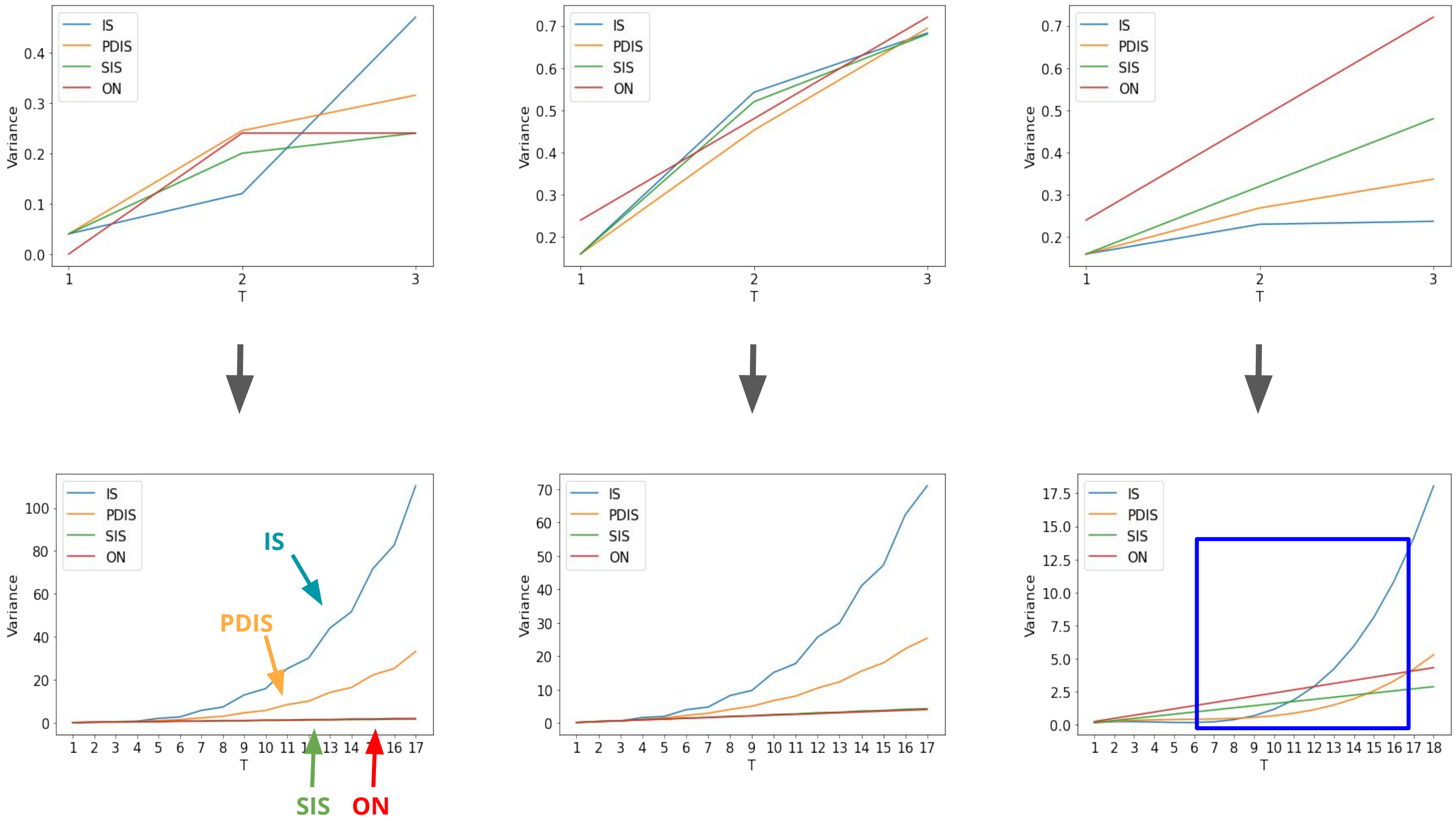
Corollary 4. *If $d_t^{\mu}(s_t)$ and $d_t^{\pi}(s_t)$ are asymptotically equi-continuous, $\frac{d^{\pi}(s)}{d^{\mu}(s)} \leq U_s$, and $\frac{\pi(a|s)}{\mu(a|s)} \leq U_{\rho}$, then* $\boxed{Var\left(\hat{v}_{SIS}\right) = O(T^2)}$

Warning: SIS does not give us reduced variance for free

In general, the density of the stationary distribution of a policy is something we need to fit

Corollary 5. *Under the same condition of Corollary 4, $\hat{v}_{ASIS}$ with w_t such that where* $\boxed{\mathbb{E}_\mu\left(w_t(s_t, a_t) - \frac{d_t^\pi(s_t,a_t)}{d_t^\mu(s_t,a_t)}\right)^2 \leq \epsilon_w}$ *has a MSE of* $\boxed{O(T^2(1+\epsilon_w))}$

Our experiment: Generalize Toy MDPs for T>2



Conclusion

$$\max_{\theta} \quad \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t\right] \longleftrightarrow \max_{\theta} \quad \mathbb{E}_{\tau \sim \mu}\left[\frac{\pi_\theta(\tau)}{\mu(\tau)} \sum_{t=0}^{T} \gamma^t r_t\right]$$

- It is not hard to get un-biased off-policy value estimators. The challenge is finding *low-variance* un-biased value estimators.
- PDIS is not always better than IS, SIS is not always better than PDIS, and ON is not always better than SIS.
- For large T, the ranking of the estimators provably aligns with the ranking generally found in empirical experiments
- IS scales exponentially, PDIS scales exponentially or polynomially, PDIS scales quadratically

Limitations

- Descriptive vs. prescriptive
- The restrictions required for finite time domains are very narrow
- Limited analysis of how weak/strong the bounds are

Questions

- Parameterize policies such that stationary distribution is known?
- Can world models be used to compute the stationary distribution?
- Use importance sampling to motivate better exploration strategies?

Thank you!

Jonah Phillion, Sana Tonekaboni