Marginalized Off-Policy Evaluation for Reinforcement Learning

Tengyang Xie*[1] Yu-Xiang Wang*[2] Yifei Ma[3]

[1]UMass Amherst [2]UC Santa Barbara [3]Amazon AI (* Most of this work performed at Amazon AI)

Background

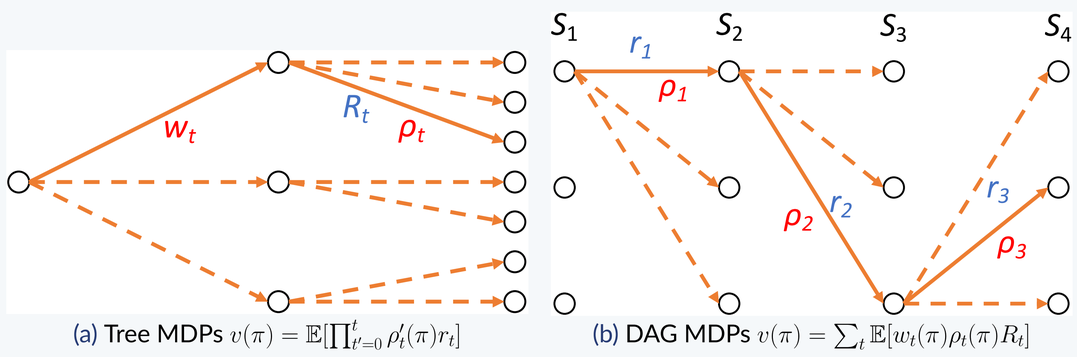
Off-policy evaluation: Evaluating the performance of *target policy* using data sampled by a *behavior policy*.

Importance: Crucial for using reinforcement learning (RL) algorithms responsibly in many real-world applications, e.g., medical treatment and digital marketing.

Challenge: The variance of importance sampling (IS)-based approaches tends to be too high to be useful for long-horizon problems because the variance of the *cumulative product of importance weights* is exploding exponentially.

Our solution:

- The cumulative product of importance weights is only necessary on the Markov decision process (MDP) model of tree MDPs.
- We can reduce the variance from the cumulative product of importance weights by designing estimators based on the DAG MDPs. Thus, what we need is only the marginalized state distribution at each time step.



(a) Tree MDPs $v(\pi) = \mathbb{E}[\prod_{t'=0}^{t} \rho'_t(\pi) r_t]$ (b) DAG MDPs $v(\pi) = \sum_t \mathbb{E}[w_t(\pi)\rho_t(\pi)R_t]$

Figure: Tree MDPs vs DAG MDPs. DAG MDPs avoid multiplicative factors.

The marginalized estimators work in the ***space of possible states***, instead of the ***space of trajectories***, resulting in a significant potential for variance reduction.

Review of Existing IS-based Methods

Discrete Tree MDPs: If an MDP satisfies:

- The state is represented by history, i.e., $s_t = h_t := o_1 a_1 \dots o_{t-1} a_{t-1} o_t$, where o_i is the observation at step $i (1 \le i \le t)$.
- The observations and actions are discrete.
- The initial state takes the form of $s_0 = o_0$. After taking action a at state $s = h$, the next can be only expressed in the form of $s' = hao$, with probability $\mathbb{P}(o|h,a)$.

Then, this MDP is a discrete tree MDP.

Generic framework of IS-based estimators:

$$\widehat{v}(\pi) = \frac{1}{n}\sum_{i=1}^{n} g(s_0^i) + \sum_{i=1}^{n}\sum_{t=0}^{H-1} \frac{\rho_{0:t}^i}{\phi_t(\rho_{0:t}^{1:n})} \gamma^t (r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)),$$

where $\rho_{0:t}^i := \prod_{t'=0}^{t} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}$ is the cumulative importance ratio for the i-th trajectory, $\phi_t : \mathbb{R}_+^n \to \mathbb{R}_+$ are the ``normalization'' functions for $\rho_{0:t}^i$, $g : \mathcal{S} \to \mathbb{R}$ and $f_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ are the ``value-related'' functions.

Examples:

$$\text{(IS[2])}: \quad g(s_0^i) = 0;\ \phi_t(\rho_{0:t}^{1:n}) = n;\ f_t(s_t^i, a_t^i, s_{t+1}^i) = 0$$

$$\text{(DR[1])}: \quad g(s_0^i) = \widehat{V}^\pi(s_0);\ \phi_t(\rho_{0:t}^{1:n}) = n;\ f_t(s_t^i, a_t^i, s_{t+1}^i) = -\widehat{Q}^\pi(s_t^i, a_t^i) + \gamma \widehat{V}^\pi(s_{t+1}^i)$$

$$\text{(WIS[3])}: \quad g(s_0^i) = 0;\ \phi_t(\rho_{0:t}^{1:n}) = \sum_{j=1}^{n} \rho_{0:t}^j;\ f_t(s_t^i, a_t^i, s_{t+1}^i) = 0$$

$$\text{(WDR[3])}: \quad g(s_0^i) = \widehat{V}^\pi(s_0);\ \phi_t(\rho_{0:t}^{1:n}) = \sum_{j=1}^{n} \rho_{0:t}^j;\ f_t(s_t^i, a_t^i, s_{t+1}^i) = -\widehat{Q}^\pi(s_t^i, a_t^i) + \gamma \widehat{V}^\pi(s_{t+1}^i).$$

Marginalized IS-based Estimators

Discrete Directed Acyclic Graph (DAG) MDPs: If an MDP satisfies:

- The state space and action space are finite.
- Each state can only occur at a particular time step.

Then, this MDP is a discrete DAG MDP.

$$w_t(s_t) := \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)} \quad \Rightarrow \quad v(\pi) = \sum_{t=0}^{H-1} \mathbb{E}_{\tau\sim\pi}[r_t] = \sum_{t=0}^{H-1} \mathbb{E}_{s_t\sim\pi}[r_t] = \sum_{t=0}^{H-1} \mathbb{E}_{s_t\sim\mu}[w_t(s_t) r_t]$$

Marginalized IS-based estimators:

$$\widehat{v}_M(\pi) = \frac{1}{n}\sum_{i=1}^{n} g(s_0^i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1} \widehat{w}_t^n(s_t^i)\rho_t^i \gamma^t (r_t^i + f_t(s_t^i, a_t^i, s_{t+1}^i)).$$

Note that the ``normalization'' function ϕ has not appeared in the framework above is because it can be a part in $\widehat{w}_t(s)$.

Properties

Theorem. Let $\phi_t(\rho_{0:t}^{1:n}) = n$ in the normal framework, then the marginalized framework could keep the unbiasedness and consistency same as in the normal framework if $\widehat{w}_t^n(s)$ is a unbiased or consistent estimator for marginalized ratio $w_t(s)$ for all t:

- If an unbiased estimator falls in the normal framework, then its marginalized estimator is also a unbiased estimator of $v(\pi)$ given unbiased estimator $\widehat{w}_t^n(s)$ for all t.
- If a consistent estimator falls in the normal framework, then its marginalized estimator is also a consistent estimator of $v(\pi)$ given consistent estimator $\widehat{w}_t^n(s)$ for all t.

Estimating $w_t(s)$:

- Unbiased and consistent Monte-Carlo based estimation

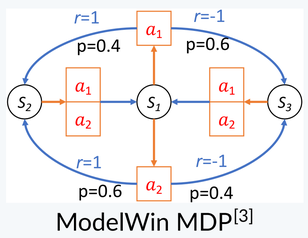
$$\widehat{w}_t^n(s) = \frac{\sum_{i=1}^{n} \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \mathbf{1}(s_t^i = s)}{\sum_{i=1}^{n} \mathbf{1}(s_t^i = s)}.$$

- Consistent estimation following the insight of self-normalization

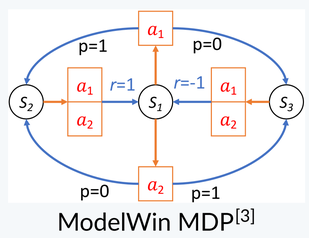
$$\widetilde{w}_t^n(s) = \frac{\sum_{i=1}^{n} \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \mathbf{1}(s_t^i = s)}{\frac{1}{n}\sum_{i=1}^{n} \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} \sum_{j=1}^{n} \mathbf{1}(s_t^j = s)} = \frac{n}{\sum_{i=1}^{n} \prod_{t'=0}^{t-1} \frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)}} \widehat{w}_t^n(s).$$

Experiments

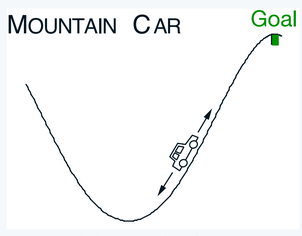
The estimators begin with "M" are the marginalized version with $\widehat{w}_t^n(s)$, and the estimators begin with "W-M" are the marginalized version with $\widetilde{w}_t^n(s)$.



ModelWin MDP[3]



ModelWin MDP[3]



Mountain Car

Horizon	IS	WIS	DR	WDR	MIS	W-MIS	MDR	W-MDR
8	0.317	0.310	0.885	0.885	0.164	**0.156**	0.822	0.821
16	0.532	0.433	1.138	1.026	0.148	**0.108**	0.838	0.836
32	1.768	0.618	2.655	1.125	0.808	**0.073**	1.182	0.857
64	3.89	0.707	2.417	1.157	1.360	**0.053**	1.159	0.862
128	2.559	0.706	4.481	1.071	0.856	**0.038**	0.778	0.855
256	3.167	0.718	1.471	1.012	1.310	**0.027**	2.268	0.861
512	1.519	0.731	1.266	0.953	1.101	**0.019**	0.670	0.863
1024	3.793	0.755	6.656	0.945	2.072	**0.013**	1.853	0.862

Table: Relative RMSE on the ModelWin Domain.

Horizon	IS	WIS	DR	WDR	MIS	W-MIS	MDR	W-MDR
8	0.079	0.035	0.863	0.861	0.056	**0.026**	0.861	0.860
16	0.158	0.055	0.883	0.864	0.104	**0.018**	0.870	0.864
32	0.586	0.106	1.078	0.864	0.64	**0.013**	1.036	0.863
64	2.844	0.191	1.872	0.861	1.604	**0.009**	1.960	0.861
128	2.220	0.332	1.321	0.859	1.436	**0.006**	1.591	0.859
256	1.293	0.478	1.203	0.859	1.000	**0.004**	0.780	0.858
512	2.459	0.596	0.539	0.858	1.885	**0.003**	1.821	0.857
1024	0.943	0.702	0.230	0.856	0.940	**0.002**	0.147	0.856

Table: Relative RMSE on the ModelFail Domain

Data Size	IS	WIS	DR	WDR	MIS	W-MIS	MDR	W-MDR
16	534.51	24.55	724.48	22.81	533.67	**5.39**	695.25	27.85
32	398.98	24.63	504.99	23.81	398.84	**4.12**	497.50	24.48
64	335.24	24.63	404.75	23.42	335.21	**3.06**	400.46	20.2
128	307.73	24.63	357.06	21.93	307.63	**2.40**	344.29	15.75
256	275.34	24.63	260.40	18.93	275.19	**2.04**	252.05	11.00
512	256.75	24.63	210.47	13.91	256.63	**1.80**	205.58	9.34
1024	249.97	24.63	166.67	6.71	249.94	**1.70**	163.30	11.62
2048	246.08	24.63	108.34	6.13	246.07	**1.64**	106.86	17.46

Table: RMSE on the Mountain Car Domain

References

[1] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652--661, 2016.

[2] Doina Precup, Richard S Sutton, and Satinder P Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759--766. Morgan Kaufmann Publishers Inc., 2000.

[3] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139--2148, 2016.