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)

Off-policy Evaluation

Evaluating the performance of ***target policy*** using data sampled by a ***behavior policy***.

Verify the performance of target policy before deploying it in the real system.

Crucial for using reinforcement learning (RL) algorithms responsibly in sensitive real-world applications, e.g.,

- medical treatment
- digital marketing & recommendation

Challenges

Importance Sampling (IS) [Precup et al., 2000; Sutton and Barto, 2018]:

$$\widehat{v}(\pi) = \sum_{i=1}^{n} \sum_{t=0}^{H-1} \prod_{t'=0}^{t} \frac{\pi(a_{t'}^{i} | s_{t'}^{i})}{\mu(a_{t'}^{i} | s_{t'}^{i})} r_t^i,$$

Challenges

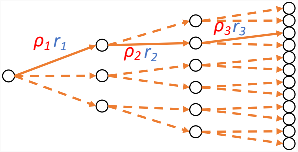
Importance Sampling (IS) [Precup et al., 2000; Sutton and Barto, 2018]:

$$\widehat{v}(\pi) = \sum_{i=1}^{n} \sum_{t=0}^{H-1} \prod_{t'=0}^{t} \frac{\pi(a_{t'}^{i}|s_{t'}^{i})}{\mu(a_{t'}^{i}|s_{t'}^{i})} r_{t}^{i},$$

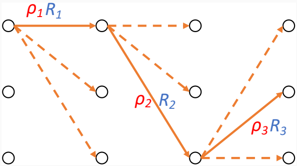
Challenges:

- $\prod_{t'=0}^{t} \frac{\pi(a_{t'}^{i}|s_{t'}^{i})}{\mu(a_{t'}^{i}|s_{t'}^{i})}$ exploding ***exponentially*** with Horizon H.
- The *variance* of IS-based approaches tends to be too high to be useful for long-horizon problems.

Idea: If we have observable states, we can use discrete directed acyclic graph (DAG) MDP model instead of discrete tree MDP model.



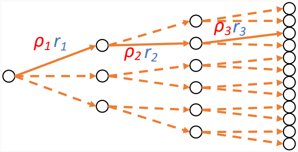
Tree MDPs



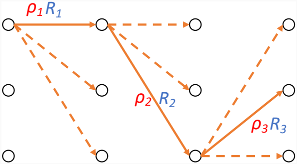
DAG MDPs

Marginalized Methods

Idea: If we have observable states, we can use discrete directed acyclic graph (DAG) MDP model instead of discrete tree MDP model.



Tree MDPs



DAG MDPs

Working space:

space of trajectories $\Rightarrow$ *space of possible states*

Marginalized Off-policy Evaluation

Marginalized IS estimators:

Tree MDPs $\Rightarrow$ DAG MDPs

$$w_t(s_t) := \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)}$$

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

Marginalized Off-policy Evaluation

Marginalized IS estimators:

Tree MDPs $\Rightarrow$ DAG MDPs

$$w_t(s_t) := \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)}$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1}\prod_{t'=0}^{t}\frac{\pi(a_{t'}^i|s_{t'}^i)}{\mu(a_{t'}^i|s_{t'}^i)} r_t^i, \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H-1}\widehat{w}_t^n(s_t^i)\frac{\pi(a_t^i|s_{t'}^i)}{\mu(a_t^i|s_{t'}^i)} r_t^i$$

The dependency of our recursive marginalized is **polynomial** on the horizon.

Estimating $w_t(s)$ recursively

Idea:

$$\begin{aligned} d_t^\pi(s_{t+1}) =& \mathbb{E}_{s_t \sim d_t^\pi}\left[T_t^\pi(s_{t+1}|s_t)\right] \\ =& \mathbb{E}_{(s_t,a_t)\sim d_t^\pi}\left[T_t(s_{t+1}|s_t,a_t)\right] \\ =& \mathbb{E}_{(s_t,a_t)\sim d_t^\mu}\left[\frac{d_t^\pi(s_t)}{d_t^\mu(s_t)}\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}T_t(s_{t+1}|s_t,a_t)\right]. \end{aligned}$$

Estimating $w_t(s)$ recursively

Idea:

$$\begin{aligned}
d_t^\pi(s_{t+1}) =& \mathbb{E}_{s_t \sim d_t^\pi}\left[T_t^\pi(s_{t+1}|s_t)\right] \\
=& \mathbb{E}_{(s_t,a_t) \sim d_t^\pi}\left[T_t(s_{t+1}|s_t, a_t)\right] \\
=& \mathbb{E}_{(s_t,a_t) \sim d_t^\mu}\left[\frac{d_t^\pi(s_t)}{d_t^\mu(s_t)} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} T_t(s_{t+1}|s_t, a_t)\right].
\end{aligned}$$

Option 1:

$$\widehat{d}_{t+1}^\pi(s) = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{d}_t^\pi(s_t^i)}{\widehat{d}_t^\mu(s_t^i)} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)} \mathbf{1}(s_{t+1}^i = s).$$

Option 2 (with self-normalization):

$$\widetilde{d}_{t+1}^\pi(s) = \frac{\sum_{i=1}^{n} \frac{\widehat{d}_t^\pi(s_t^i)}{\widehat{d}_t^\mu(s_t^i)} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)} \mathbf{1}(s_{t+1}^i = s)}{\sum_{i=1}^{n} \frac{\widehat{d}_t^\pi(s_t^i)}{\widehat{d}_t^\mu(s_t^i)} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}}.$$

Error propagation:

Let $\widehat{\varepsilon}_t^\pi(s) \coloneqq \widehat{d}_t^\pi(s) - d_t^\pi(s)$ and $\dfrac{\pi(a|s)}{\mu(a|s)} \leq 1/\eta$, then

$$\sum_s |\widehat{\varepsilon}_{t+1}^\pi(s)| \leq \left(1 + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\eta^2}{n}}\right)\right) \sum_{s_t} |\varepsilon_t^\pi(s_t)| + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right).$$

Theoretical Analysis

Error propagation:

Let $\widehat{\varepsilon}_t^\pi(s) := \widehat{d}_t^\pi(s) - d_t^\pi(s)$ and $\dfrac{\pi(a|s)}{\mu(a|s)} \le 1/\eta$, then

$$\sum_s |\widehat{\varepsilon}_{t+1}^\pi(s)| \le \left(1 + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\eta^2}{n}}\right)\right) \sum_{s_t} |\varepsilon_t^\pi(s_t)| + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right).$$

If $n \gg \frac{|S|^2 H^2}{\eta^2}$, then with high probability,

$$\sum_s |\widehat{\varepsilon}_t^\pi(s)| = \widetilde{\mathcal{O}}\left(\frac{H|S|}{\eta\sqrt{n}}\right).$$

Theoretical Analysis

Error propagation:

Let $\widehat{\varepsilon}_t^\pi(s) := \widehat{d}_t^\pi(s) - d_t^\pi(s)$ and $\dfrac{\pi(a|s)}{\mu(a|s)} \leq 1/\eta$, then

$$\sum_s |\widehat{\varepsilon}_{t+1}^\pi(s)| \leq \left(1 + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\eta^2}{n}}\right)\right) \sum_{s_t} |\varepsilon_t^\pi(s_t)| + \widetilde{\mathcal{O}}\left(|\mathcal{S}|\sqrt{\frac{\max_s(w_t^2(s))}{n\eta^2}}\right).$$

If $n \gg \frac{|S|^2 H^2}{\eta^2}$, then with high probability,

$$\sum_s |\widehat{\varepsilon}_t^\pi(s)| = \widetilde{\mathcal{O}}\left(\frac{H|S|}{\eta\sqrt{n}}\right).$$

$\Rightarrow$ the absolute error bound of off-policy policy evaluation is $\widetilde{\mathcal{O}}\left(\frac{H^2|S|}{\eta\sqrt{n}}\right)$

Thanks!

References

Precup, D., Sutton, R. S., and Singh, S. P. (2000). Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766. Morgan Kaufmann Publishers Inc.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.