

Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling



Tengyang Xie¹ Yifei Ma² Yu-Xiang Wang³

¹University of Illinois at Urbana-Champaign

²Amazon AI

³University of California, Santa Barbara



Challenges

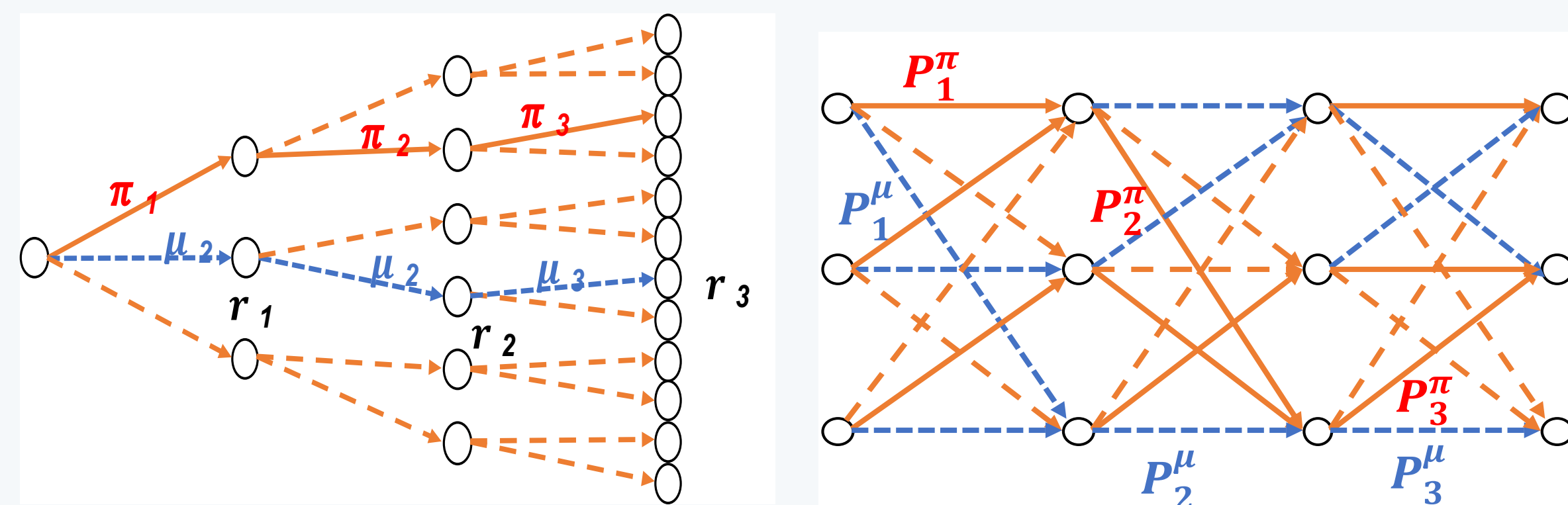
Off-policy evaluation: Evaluating the performance of *target policy* using data sampled by a *behavior policy*.

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

- Cumulative product of importance weights is only necessary without state observability.
- Reduce variance by marginalizing the actions to get the state distribution at every step.



(a) Vanilla IS

$$\hat{v}_{IS}^{\pi} = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^H \left[\prod_{t=1}^h \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})} \right] r_h^{(i)}$$

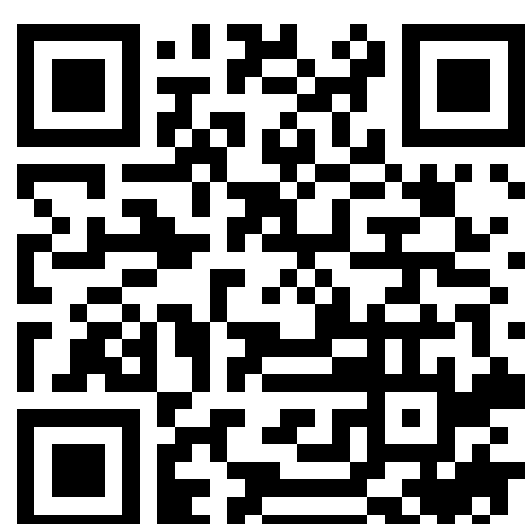
(b) Our MIS

$$\hat{v}_{MIS}^{\pi} = \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^H \frac{\hat{d}_t^{\pi}(s_t^{(i)})}{\hat{d}_t^{\mu}(s_t^{(i)})} \hat{r}_t^{\pi}(s_t^{(i)})$$

Figure: Vanilla IS versus our MIS. MIS changes mean-of-products to product-of-means.

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

link: <https://arxiv.org/abs/1906.03393>



Scan QR code to view the paper

Marginalized Importance Sampling (MIS)

Notations: behavior and target policy $\mu_t(a_t|s_t)$ and $\pi_t(a_t|s_t)$, resp.; transition function $T(s_{t+1}|s_t, a_t)$; state distribution $d_t^{\mu}(s_t)$ and $d_t^{\pi}(s_t)$.

Observation: Policy-induced state transitions are temporally independent

$$d_t^{\pi}(s_t) = \sum_{s_{t-1}} P_t^{\pi}(s_t|s_{t-1}) d_{t-1}^{\pi}(s_{t-1}),$$

where $P_t^{\pi}(s_t|s_{t-1}) = \sum_{a_{t-1}} T_t(s_t|s_{t-1}, a_{t-1}) \pi(a_{t-1}|s_{t-1})$.

Off-policy evaluation with MIS

$$\hat{v}_{MIS}^{\pi} = \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^H \frac{\hat{d}_t^{\pi}(s_t^{(i)})}{\hat{d}_t^{\mu}(s_t^{(i)})} \hat{r}_t^{\pi}(s_t^{(i)}),$$

where the functions of the states are estimated as

$$\hat{d}_t^{\mu}(s_t) = \frac{1}{n} \sum_i \mathbf{1}(s_t^{(i)} = s_t); \quad \hat{d}_t^{\pi}(s_t) = \sum_{s_{t-1}} \hat{P}_t^{\pi}(s_t|s_{t-1}) \hat{d}_{t-1}^{\pi}(s_{t-1}), \text{ where}$$

$$\hat{P}_t^{\pi}(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}} \sum_{i=1}^n \frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})} \mathbf{1}(s_{t-1} = s_{t-1}, s_t = s_t^{(i)});$$

$$\text{and } \hat{r}_t^{\pi}(s_t) = \frac{1}{n_{s_t}} \sum_{i=1}^n \frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)} r_t^{(i)} \mathbf{1}(s_t = s_t^{(i)}).$$

Theoretical Analysis -- Methodology

Why MIS breaks the exponential dependency on horizon:

- Ergodicity - all states are visited with probability at least $d_t^{\mu} > d_m > 0$.
- Sufficient data $n > O(d_m^{-1})$, so every state is empirically visited with high probability.
- Let $\tau_a \tau_s$ be the max importance weight; $n > O(\tau_a \tau_s)$ controls the variance.

Bellman equation for variance decomposition (Lemma B.3)

- Define $\tilde{d}(s)$, $\tilde{V}(s)$, and $\tilde{r}(s)$ to be "fictitious" tail-clipped estimators;
- Their vector forms include their values on all the states.

Bellman equation $V_t^{\pi}(s_t) = r_t^{\pi}(s_t) + \sum_{s_{t+1}} P_t^{\pi}(s_{t+1}|s_t) V_{t+1}^{\pi}(s_{t+1})$

$$\Rightarrow \text{Var}[\tilde{v}^{\pi}] = \frac{\text{Var}[V_1^{\pi}(s_1^{(1)})]}{n} + \sum_{h=1}^H \sum_{s_h} \mathbb{E} \left[\frac{\tilde{d}_h^{\pi}(s_h)^2}{n_{s_h}} \mathbf{1}(E_h) \right] \text{Var}_{\mu} \left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)} (V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)}) \middle| s_h^{(1)} = s_h \right]$$

Theoretical Analysis -- Optimality

MIS has optimal sample complexity upto a factor of H (Theorem 4.1)

- Define \mathcal{P} be the projection to feasible policy values.
- Let $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$ and $\tau_s := \max_{t,s_t} \frac{d_t^{\pi}(s_t)}{d_t^{\mu}(s_t)}$.
- H is total horizon, σ is observation noise, and R_{\max} is maximal immediate reward.

$$\text{then } \mathbb{E}[(\mathcal{P}\hat{v}_{MIS}^{\pi} - v^{\pi})^2] \leq \frac{1}{n} \sum_{t=1}^H \mathbb{E}_{\mu} \left[\frac{d_t^{\pi}(s_t)^2}{d_t^{\mu}(s_t)^2} \text{Var}_{\mu} \left[\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} (V_{t+1}^{\pi}(s_{t+1}) + r_t) \middle| s_t \right] \right] + \tilde{O}(n^{-1.5}).$$

Our paper shows the worst-case bound of our estimator is $\mathbb{E}[(\mathcal{P}\hat{v}_{MIS}^{\pi} - v^{\pi})^2] \leq \frac{4}{n} \tau_a \tau_s (H\sigma^2 + H^3 R_{\max}^2)$, which is optimal upto a factor of H compared with the CR lower bound [Jiang and Li, 2016].

Experimental Study

Tabular MDPs: (common) start with State s_1 , choose between Action a_1/a_2 , where $\mu(a_1) = 0.5; \pi(a_1) = 0.2$. Random transition to state s_2/s_3 . ModelWin and ModelFail MDPs are described as follows:

- ModelWin: rewards decided by the action on the observable state s_1 . Set $p = 0.4$.
- ModelFail: actions lead to unobservable states, where rewards are decided; set $p = 1$.

MountainCar: drive back and forth until at top of the hill. State = (position, speed); action=acceleration. Evaluate Q-learning policy π from soft-Q-policy for μ .

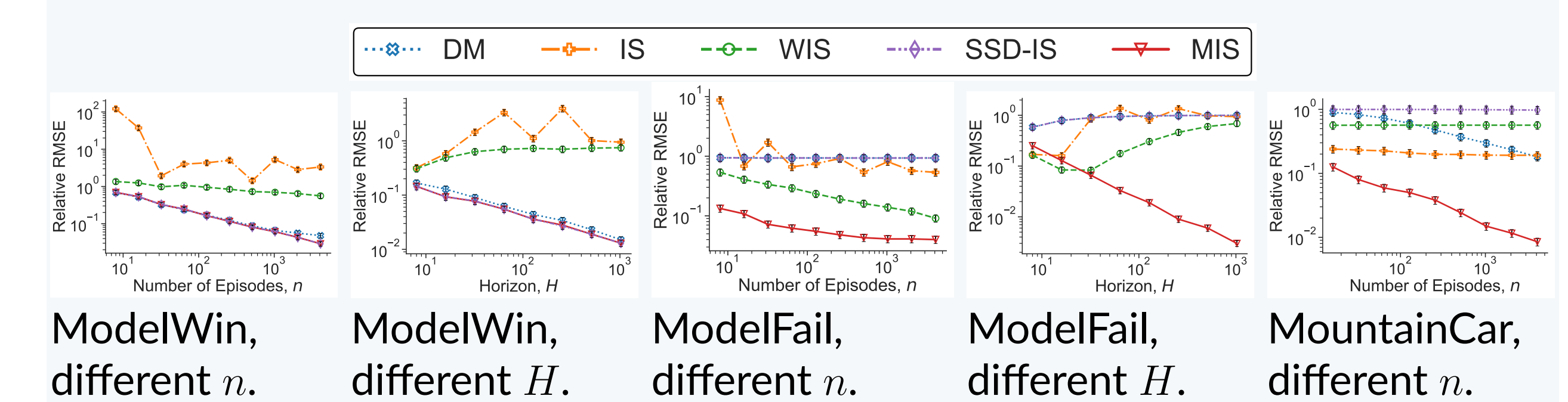


Figure: Relative RMSE error for policy evaluation. MIS matches DM (model-based method) on ModelWin and outperforms IS/WIS on ModelFail and MountainCar, both of which are the best existing methods on their respective domains.

References

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