A New Doubly Robust Policy Estimator on Infinite Horizon Reinforcement Learning

Ziyang Tang [*1] **Yihao Feng** [*1] **Qiang Liu** [1]

Abstract

Recently infinite horizon off-policy evaluation method based on estimation of density ratio has been proposed (Liu et al., 2018). Before that doubly robust estimator is the strongest baseline in off-policy evaluation in finite horizon. A natural question is if we can apply doubly robust method in infinite horizon setting. This paper answer that question and reveal an interesting connection between density ratio function and value function. We provide both theoretical and empirical result to show that our method yields significantly advantage over previous method.

1. Introduction

Reinforcement learning (RL) (Sutton & Barto, 1998) is one of the most successful approaches to artificial intelligence, and has found successful applications in robotics, games, dialogue systems, and recommendation systems, among others. One of the key problems in RL is policy evaluation: given a fixed policy, estimate the average reward garnered by an agent that runs this policy in the environment. In this paper, we consider the off-policy estimation problem, in which we want to estimate the expected reward of a given target policy with samples collected by a different behavior policy. This problem is of great practical importance in many application domains where deploying a new policy can be costly or risky, such as medical treatments (Murphy et al., 2001), econometrics (Hirano et al., 2003), recommender systems (Li et al., 2011), education (Mandel et al., 2014), Web search (Li et al., 2015), advertising and marketing (Bottou et al., 2013; Chapelle et al., 2014; Tang et al., 2013; Thomas et al., 2017). It can also be used as a key component for developing efficient off-policy policy optimization algorithms (Dudík et al., 2011; Jiang & Li, 2016; Li et al., 2015; Thomas & Brunskill, 2016).

Most previous off-policy estimation methods are based on importance sampling(IS) for trajectory (e.g., Liu, 2001). A major limitation, however, is that this approach can become inaccurate due to high variance introduced by the long trajectory. Indeed, most existing IS-based estimators compute the weight as the product of the importance ratios of many steps in the trajectory. To alleviate the high variance problem, researcher introduce variance reduction technique for IS-based estimators. The most famous one is Doubly Robust estimator (Jiang & Li, 2015).

Recently, Liu et al. (2018) proposes an estimation problem for infinite horizon off-policy evaluation. The method is based on estimate the density ratio between stationary distribution, rather than trajectory, which avoid the cumulative product of across many steps in the trajectory, which substantially decrease its variance and eliminate the estimator's dependence on the horizon.

However, Liu et al. (2018) does not provide any variance reduction trick for its estimator. A natural question is to ask if there is a easy way to do variance reduction for the infinite horizon density ratio estimator.

In this paper, we develop a new doubly robust estimator based on the infinite horizon density ratio estimator. We find interesting connection between the density ratio function and value function, and use them to construct a doubly robust estimator to reduce the variance. Our method yields a significant advantage compare to the previous estimator.

2. Background

2.1. Off-Policy Evaluation

Consider a Markov decision process(MDP) $M = \langle \mathcal{S}, \mathcal{A}, r, \boldsymbol{T} \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, reward function r, and transition probability function $\boldsymbol{T}$. The average discounted reward for a target policy π is defined as:

$$R_\pi := \lim_{T\to\infty} \mathbb{E}_{\tau\sim\pi}[\frac{\sum_{t=0}^T \gamma^t r_t}{\sum_{t=0}^T \gamma^t}],$$

where $\tau = \{s_i, a_i, r_i\}_{1\le i\le n}$ is trajectory with state, action, reward simulated in MDP under policy π.

[*] Equal contribution [1] Department of Computer Science, UT Austin, TX, USA. Correspondence to: Ziyang Tang <ztang@cs.utexas.edu>, Yihao Feng <yihao@cs.utexas.edu>.

Real-world Sequential Decision Making workshop at ICML 2019.

Denoted $d_{\pi,t}(\cdot)$ as average visitation of s_t in time step t, we can define the stationary distribution, or sometimes we call it discounted average visitation, d_π as:

$$d_\pi(s) := \lim_{T\to\infty} \frac{\sum_{t=0}^{T} \gamma^t d_{\pi,t}(s)}{\sum_{t=0}^{T} \gamma^t} = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t d_{\pi,t}(s).$$

Here $(1-\gamma)$ is the normalization factor introduced by $\sum_{t=0}^{\infty} \gamma^t$.

Using stationary distribution we can rewrite our average discounted reward as:

$$R_\pi = \mathbb{E}_{s\sim d_\pi, a\sim\pi(\cdot|s)}[r(s,a)] \tag{1}$$

The Off-Policy evaluation problem gives you data $\{s_i, a_i, s_i', r_i\}$ collected under a behavior policy $\pi_0(a|s)$, and we want to estimate the average discounted reward for another target policy $\pi(a|s)$.

2.2. Value function Estimator

The value function for policy π is defined as the expected discounted reward sum start from a certain state:

$$V^\pi(s) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t r_t | s = s_0], \tag{2}$$

where $\tau = \{(s_0, a_0, r_0)\}_{i=0}^{\infty}$ is a trajectories draw from policy π.

Using the definition we can derive a Bellman equation for V^π:

$$V^\pi(s) = \mathbb{E}_{a,s'|s\sim\pi}[r(s,a) + \gamma V(s')], \forall s \in \mathcal{S} \tag{3}$$

If we can evaluate the value function accurately, we can use it to estimate the average discounted reward:

$$\begin{aligned} R_\pi =& (1-\gamma)\mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t r_t] \\ =& (1-\gamma)\mathbb{E}_{s\sim d_0}[\mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s]] \\ =& (1-\gamma)\mathbb{E}_{s\sim d_0}[V^\pi(s)]. \end{aligned}$$

Solving V^π If we define an Temporal Difference(TD)(Sutton, 1988) target over function f as:

$$\mathcal{B}f(s) = f(s) - \mathbb{E}_{a,s'|s\sim\pi}[r(s,a) + \gamma f(s')]. \tag{4}$$

We often use the TD method to solve V^π using the following fixed point recursion:

$$V(s) \leftarrow V(s) + \alpha(r + \gamma V(s') - V(s)). \tag{5}$$

Actually this can be viewed as a stochastic gradient descent under the following objective functional:

$$\min_V D[V] := \mathbb{E}_{s\sim d_{\pi_0}}[(\mathcal{B}V(s))^2]. \tag{6}$$

In this objective, the gradient with respect to $V(s)$ is $2(V(s) - (r + \gamma V(s')))$ which leads to equation (5).

Another way is to introduce a test function to the objective and let the functional G as:

$$G[f, V] := \mathbb{E}_{s\sim d_{\pi_0}}[f(s)\mathcal{B}V(s)]. \tag{7}$$

We know that if $V = V^\pi$, $G[f, V] = 0$ for all test function f.

Now we can redefine the objective for solving V as,

$$\tilde{D}[V] := \max_{f\in\mathcal{F}} G[f, V]. \tag{8}$$

When the function space $\mathcal{F} = \{\mathbb{E}_{s\sim d_{\pi_0}}[f(s)^2] \leq 1\}$ is in l_2, the new objective function $\widetilde{D}$ is equivalent to the previous objective function D.

Using V^π **to estimate** R_π Suppose now we have already get an estimation of V, we could estimate R_π as a functional:

$$R_{val}[V] = (1-\gamma)\mathbb{E}_{s\sim d_0}[V(s)]. \tag{9}$$

If $V = V^\pi$, the estimator is unbiased.

2.3. Density Ratio Estimation

The density ratio is defined as the ratio between stationary distribution of π and π_0:

$$w_{\pi/\pi_0}(s) = \frac{d_\pi(s)}{d_{\pi_0}(s)}. \tag{10}$$

Notice that if we get the true density ratio, we can estimate R_π using π_0 samples with importance sampling method:

$$\begin{aligned} R_\pi =& \mathbb{E}_{s\sim d_\pi, a\sim\pi(\cdot|s)}[r(s,a)] \\ =& \mathbb{E}_{s\sim d_{\pi_0}, a\sim\pi_0(\cdot|s)}[w_{\pi/\pi_0}(s)\frac{\pi(a|s)}{\pi_0(a|s)} r(s,a)]. \end{aligned} \tag{11}$$

Solving density ratio w_{π/π_0} Liu et al. (2018) proposes a method to estimate the density ratio between distribution over stationary on state, using a Bellman like recursive equation

We skip the derivation for how we get those equation, but just state the result. The estimation of density ratio w can be written as a optimization problem on the following objective functional:

$$\begin{aligned} L[w, f] := \gamma\mathbb{E}_{(s,a,s')\sim d_{\pi_0}}[(w(s)\beta_{\pi/\pi_0}(a|s) - \\ w(s'))f(s')] + (1-\gamma)\mathbb{E}_{s\sim d_0}[(1-w(s))f(s)]. \end{aligned}$$

We further define $D[w]$ to be the the largest value of $L[w, f]$ among all the test function f, our goal is to find w minimize $D[w]$:

$$\min_{w} \left\{ D[w] := \max_{f \in \mathcal{F}} L[w, f]^2 \right\}, \tag{12}$$

where $\mathcal{F}$ is the space of test function f, typically taken to be Reproducing Kernel Hilbert Space(RKHS).

Estimate R_π using w Suppose now we have already get an estimation of w, we could estimate R_π as a functional:

$$R_{den}[w] = \mathbb{E}_{(s,a) \sim d_{\pi_0}} [w(s) \beta_{\pi/\pi_0}(a|s) r(s, a)] \tag{13}$$

where $\beta_{\pi/\pi_0}(a|s) = \frac{\pi(a|s)}{\pi_0(a|s)}$ for short.

2.4. Connection

Liu et al. (2018) also proposes an interesting connection for $L(w, f)$ and the value function V^π for the estimation:

Theorem 2.1. *Define $\mathcal{R}_{den}[w]$ to be the reward estimate using estimated density ratio $w(s)$ (which may not equal the true ratio w_{π/π_0}) and infinite number of trajectories from d_{π_0}, that is,*

$$\mathcal{R}_{den}[w] := \mathbb{E}_{(s,a,s') \sim d_{\pi_0}} [w(s) \beta_{\pi/\pi_0}(a|s) r(s, a)] \, .$$

Assume w is properly normalized such that $\mathbb{E}_{s \sim d_{\pi_0}}[w(s)] = 1$, we have $L[w, V^\pi] = \mathcal{R}_\pi - \mathcal{R}_{den}[w]$. Therefore, if $\pm V^\pi \in \mathcal{F}$, we have $|\mathcal{R}_{den}[w] - \mathcal{R}_\pi| \leq \max_{f \in \mathcal{F}} L[w, f]$.

Under this theorem, if $V^\pi \in \mathcal{F}$ we can safely use $R[w]$ as estimation of R_π.

3. Doubly Robust Estimator

Doubly robust estimator is first proposed to solve Causal effect (Funk et al., 2011) as an estimator combining Inverse Propensity Score(IPS) estimator and Regression modeling(Reg) estimator.

Jiang & Li (2015) introduce the idea of doubly robust estimator into off policy evaluation in Reinforcement Learning. It incorporates an approximate value function as a control variate to reduce the variance of importance sampling estimator.

Inspired by this method and Liu et al. (2018)'s theorem about the connection for density ratio w_{π/π_0} and value function V^π, we come up with an idea to use a new doubly robust estimator as our infinite horizon off policy estimator.

Suppose we have already get the our function approximation for density w and value function V, where w is close to w_{π/π_0} and V is close to value function for policy π V^π.

Algorithm 1 A New Doubly Robust Estimator

```
Input: History transition data $\mathcal{D} = \{s_i, a_i, s'_i, r_i\}_{1 \leq i \leq n}$ from policy $\pi_0$; a target policy $\pi$ for which we want to estimate the expected reward;
Use $D[V]$ in Equation (5) as objective to train a good value function $V$ under some parametric space.
Use $D[w]$ in Equation (12) as objective to train a good density ratio $w$ under some parametric space.
Use $R_{doubly}[w, V]$ in (16) to estimate $R_\pi$ using sample from $\mathcal{D}$.
```

Notice that if $V = V^\pi$, we can use $R[V] + G(V, f)$ to as an unbiased estimator, since $G(V, f) = 0$ for all test function f. Similarly, if $w = w_{\pi/\pi_0}$, we can use $R[w] + L(w, f)$ as an unbiased estimator.

Now we connect them together and define a new bivariate functional $C[w, V]$ as:

$$\begin{aligned} C[w, V] := \mathbb{E}_{s \sim d_{\pi_0}} [w(s)(V(s) - \\ \gamma \mathbb{E}_{s', a|s \sim \pi_0} [\beta_{\pi/\pi_0}(a|s) V(s')])]. \end{aligned}$$

Now we show our key observation that this connection functional can connect the density ratio and value function together.

Lemma 3.1. *Compare $C[w, V]$ with $G[w, V]$, we have:*

$$C[w, V] = G[w, V] + R_{den}[w] \tag{14}$$

Proof. Collect the term inside expectation we can easily get the equation. □

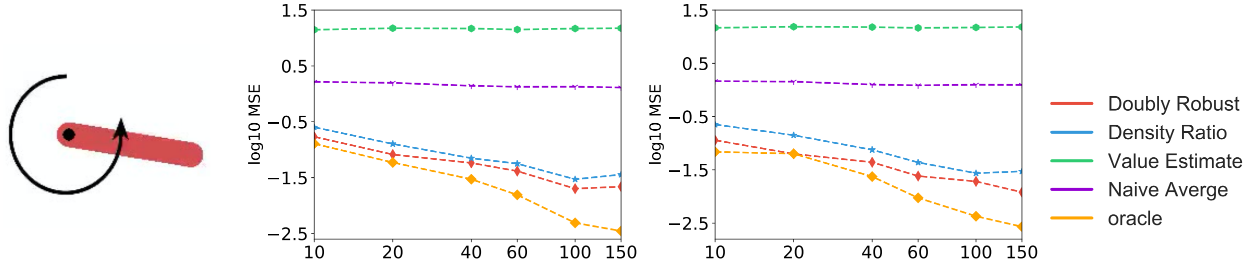
Lemma 3.2. *Compare $C[w, V]$ with $L[w, V]$, we have:*

$$C[w, V] = R_{val}[V] - L[w, V] \tag{15}$$

Proof. $R_{val}[V] = (1 - \gamma) \mathbb{E}_{s \sim d_0}[V(s)]$. Rewrite $L[w, V]$ using the lemma (5) and equation (25) in (Liu et al., 2018) we have (or simply use $d_{\pi_0} = (1 - \gamma) d_0 + \gamma T^{\pi_0} d_{\pi_0}$ trick.):

$$\begin{aligned} L[w, V] =& \gamma \mathbb{E}_{(s,a,s') \sim d_{\pi_0}} \left[\left(w(s) \beta_{\pi/\pi_0}(a|s) - w(s') \right) V(s') \right] + \\ & (1 - \gamma) \mathbb{E}_{s \sim d_0} [(1 - w(s)) V(s)] \\ =& \gamma \mathbb{E}_{(s,a,s') \sim d_{\pi_0}} \left[\left(w(s) \beta_{\pi/\pi_0}(a|s) \right) V(s') \right] - \\ & \gamma \mathbb{E}_{s' \sim T^{\pi_0} d_{\pi_0}} [w(s') V(s')] + (1 - \gamma) \mathbb{E}_{s \sim d_0} [(1 - w(s)) V(s)] \\ =& \gamma \mathbb{E}_{(s,a,s') \sim d_{\pi_0}} \left[\left(w(s) \beta_{\pi/\pi_0}(a|s) \right) V(s') \right] - \mathbb{E}_{s \sim d_{\pi_0}} [w(s) V(s)] \\ & + (1 - \gamma) \mathbb{E}_{s \sim d_0} [w(s) V(s)] + (1 - \gamma) \mathbb{E}_{s \sim d_0} [(1 - w(s)) V(s)] \\ =& - \mathbb{E}_{s \sim d_{\pi_0}} [w(s)(V(s) - \gamma \beta_{\pi/\pi_0}(a|s) V(s'))] \\ & + (1 - \gamma) \mathbb{E}_{s \sim d_0} [V(s)] \\ =& - C[w, V] + R_{val}[V]. \end{aligned}$$

□



(a) Pendulum-v0 (b) # of Trajectories (n), $T = 500$ (c) # of Trajectories (n), $T = 1000$

Figure 1. Result on Pendulum-v0 environment. (b)-(c) show the results in the discount case with $\gamma = 0.99$, (b)

3.1. Doubly Robust Estimator of R_π

Now we use the following estimator to estimate R_π:

$$R_{doubly}[w, V] = R_{den}[w] + R_{val}[V] - C[w, V] \quad (16)$$

Our main theorem tells us that if either V or w is estimated correctly, the doubly robust estimator is unbiased.

Theorem 3.3. *If $w = w_{\pi/\pi_0}$ or $V = V^\pi$, $R_{doubly}[w, V]$ is an unbiased estimator of R_π.*

Proof. If $w = w_{\pi/\pi_0}$, $R_{den}[w] = R_\pi$, $R_{val}[V] - C(w, V) = L[w, V] = 0$. If $V = V^\pi$, $R_{val}[V] = R_\pi$, $R_{den}[w] - C[w, V] = -G[w, V] = 0$. □

We can further analyze its variance in future.

3.2. Proposed algorithm

Solve the optimization for value function V and density ratio w respectively by minimizing $D[w]$ and $D[V]$, then use $R_{doubly}[w, V]$ to estimate R_π. A detail procedure is described in Algorithm 1.

4. Experiment

4.1. Evaluation on Pendulum

We test Pendulum, which has a continuous state space of $\mathbb{R}^3$ and action space of $[-2, 2]$. The initial state randomly start with angle from $-\pi$ to π and a random velocity from -1 to 1. In this environment, we want to control the pendulum to make it stand up as fast as possible. The policy is taken to be a truncated Gaussian whose mean is a neural network of the states and variance a constant. There is no ending state for this environment, and at each state the reward is calculated as how much you are close to the stand up status. We train a near-optimal policy π_* using REINFORCE and set it to be the target policy. The behavior policy is set to be $\pi = (1 - \alpha)\pi_* + \alpha\pi_+$, where α is a mixing ratio, and π_+ is another policy from REINFORCE when it has not converged.

Our results are shown in Figure 1, where we find that our proposed doubly robust policy evaluation method outperforms density ratio estimation (Liu et al., 2018) and off policy value estimation methods on two different trajectory length cases ($T = 500$ and $T = 1000$). As is shown in Figure 1, our proposed doubly robust can be accurate as long as either density ratio $R_{den}[w]$ or value function $R_{val}[V]$ estimator is accurate, which can enjoy the benefits of both estimators in different settings to make the reward estimator robust when the samples from target policy is unavailable.

5. Conclusion

In this paper, we develop a new doubly robust estimator based on the infinite horizon density ratio and off policy value estimation. Our new proposed doubly robust estimator can be accurate as long as one of the estimators are accurate, which yields a significant advantage comparing to previous estimators. As future work, we will validate our method on more complex environments and provide a theoretical variance and bias sample analysis.

References

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

Chapelle, O., Manavoglu, E., and Rosales, R. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology*, 5(4): 61:1–61:34, 2014.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

Hirano, K., Imbens, G. W., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

Jiang, N. and Li, L. Doubly robust off-policy evaluation for reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 652–661, 2016.

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM)*, pp. 297–306, 2011.

Li, L., Chen, S., Gupta, A., and Kleban, J. Counterfactual analysis of click metrics for search engine optimization: A case study. In *Proceedings of the 24th International World Wide Web Conference (WWW), Companion Volume*, pp. 929–934, 2015.

Liu, J. S. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer-Verlag, 2001. ISBN 0387763694.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018.

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 1077–1084, 2014.

Murphy, S. A., van der Laan, M., and Robins, J. M. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, March 1998. ISBN 0-262-19398-1.

Tang, L., Rosales, R., Singh, A., and Agarwal, D. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 1587–1594, 2013.

Thomas, P. S. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 2139–2148, 2016.

Thomas, P. S., Theocharous, G., Ghavamzadeh, M., Durugkar, I., and Brunskill, E. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4740–4745, 2017.