Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation

Qiang Liu
The University of Texas at Austin
Austin, TX, 78712
lqiang@cs.utexas.edu

Lihong Li
Google Brain
Kirkland, WA, 98033
lihong@google.com

Ziyang Tang
The University of Texas at Austin
Austin, TX, 78712
ztang@cs.utexas.edu

Dengyong Zhou
Google Brain
Kirkland, WA, 98033
dennyzhou@google.com

Abstract

We consider off-policy estimation of the expected reward of a target policy using samples collected by a different behavior policy. Importance sampling (IS) has been a key technique for deriving (nearly) unbiased estimators, but is known to suffer from an excessively high variance in long-horizon problems. In the extreme case of *infinite*-horizon problems, the variance of an IS-based estimator may even be unbounded. In this paper, we propose a new off-policy estimator that applies IS *directly* on the stationary state-visitation distributions to avoid the exploding variance faced by existing methods. Our key contribution is a novel approach to estimating the density ratio of two *stationary state distributions*, with *trajectories* sampled from only the behavior distribution. We develop a mini-max loss function for the estimation problem, and derive a closed-form solution for the case of RKHS. We support our method with both theoretical and empirical analyses.

1 Introduction

Reinforcement learning (RL) [36] 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 [26], econometrics [13], recommender systems [19], education [23], Web search [18], advertising and marketing [4, 5, 38, 40]. It can also be used as a key component for developing efficient off-policy policy optimization algorithms [7, 14, 18, 39].

Most state-of-the-art off-policy estimation methods are based on importance sampling (IS) [e.g., 22]. A major limitation, however, is that this approach can become inaccurate due to the high variance introduced by the importance weights, especially when the trajectory is long. Indeed, most existing IS-based estimators compute the weight as the product of the importance ratios of many steps in the trajectory. Variances in individual steps accumulate *multiplicatively*, so that the overall IS weight of a random trajectory can have an exponentially high variance to result in an unreliable estimator. In the extreme case when the trajectory length is infinite, as in infinite-horizon average-reward problems, some of these estimators are not even well-defined. Ad hoc approaches can be used, such as truncating

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

the trajectories, but often lead to a hard-to-control bias in the final estimation. Analogous to the well-known "curse of dimensionality" in dynamic programming [2], we call this problem the "curse of horizon" in off-policy learning.

In this work, we develop a new approach that tackles the curse of horizon. The key idea is to apply importance sampling on the *average visitation distribution* of single steps of state-action pairs, instead of the much higher dimensional distribution of whole trajectories. This avoids the cumulative product across time in the density ratio, substantially decreasing its variance and eliminating the estimator's dependence on the horizon.

Our key challenge, of course, is to estimate the importance ratios of average visitation distributions. In practice, we often have access to both the target and behavior policies to compute their importance ratio of an action conditioned on a given state. But we typically have *no* access to transition probabilities of the environment, so estimating importance ratios of state visitation distributions has been very difficult, especially when only off-policy samples are available. In this paper, we develop a mini-max loss function for estimating the true stationary density ratio, which yields a closed-form representation similar to maximum mean discrepancy [9] when combined with a reproducing kernel Hilbert space (RKHS). We study the theoretical properties of our loss function, and demonstrate its empirical effectiveness on long-horizon problems.

2 Background

Problem Definition Consider a Markov decision process (MDP) [31] $M = \langle \mathcal{S}, \mathcal{A}, r, \boldsymbol{T} \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, reward function r, and transition probability function $\boldsymbol{T}$. Assume the environment is initialized at state $s_0 \in \mathcal{S}$, drawn from an unknown distribution $d_0(\cdot)$. At each time step t, an agent observes the current state s_t, takes an action a_t according to a possibly stochastic policy $\pi(\cdot|s_t)$, receives a reward r_t whose expectation is $r(s_t, a_t)$, and transitions to a next state s_{t+1} according to transition probabilities $\boldsymbol{T}(\cdot|s_t, a_t)$. To simplify exposition and avoid unnecessary technicalities, we assume $\mathcal{S}$ and $\mathcal{A}$ are finite unless otherwise specified, although our method extends to continuous spaces straightforwardly, as demonstrated in experiments.

We consider the *infinite horizon* problem in which the MDP continues without termination. Let $p_\pi(\cdot)$ be the distribution of trajectory $\boldsymbol{\tau} = \{s_t, a_t, r_t\}_{t=0}^{\infty}$ under policy π. The expected reward of π is

$$R_\pi := \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}[R^T(\boldsymbol{\tau})], \qquad\qquad R^T(\boldsymbol{\tau}) := (\sum_{t=0}^{T} \gamma^t r_t)/(\sum_{t=0}^{T} \gamma^t)\,,$$

where $R_\pi^T(\boldsymbol{\tau})$ is the reward of trajectory $\boldsymbol{\tau}$ up to time T. Here, $\gamma \in (0, 1]$ is a discount factor. We distinguish two reward criteria, the average reward ($\gamma = 1$) and discounted reward ($0 < \gamma < 1$):

$$\textit{Average:} \quad R(\boldsymbol{\tau}) := \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} r_t, \qquad \textit{Discounted:} \quad R(\boldsymbol{\tau}) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t r_t\,.$$

where $(1-\gamma) = 1/\sum_{t=0}^{\infty}\gamma^t$ is a normalization factor. The problem of *off-policy value estimation* is to estimate the expected reward R_π of a given *target* policy π, when we only observe a set of trajectories $\boldsymbol{\tau}^i = \{s_t^i, a_t^i, r_t^i\}_{t=0}^{T}$ generated by following a different *behavior* policy π_0.

Bellman Equation We briefly review the Bellman equation and the notation of value functions, for both average and discounted reward criteria. In the discounted case ($0 < \gamma < 1$), the value $V^\pi(s)$ is the expected total discounted reward when the initial state s_0 is fixed to be s: $V^\pi(s) = \mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s]$. Note that we *do not* normalize V^π by $(1-\gamma)$ in our notation. For the average reward ($\gamma = 1$) case, the expected average reward does not depend on the initial state if the Markov process is ergodic [31]. Instead, the value function $V^\pi(s)$ in the average case measures the *average adjusted* sum of reward: $V^\pi(s) = \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}[\sum_{t=0}^{T}(r_t - R_\pi)|s_0 = s]$. It represents the relative difference in total reward gained from starting in state $s_0 = s$ as opposed to R_π.

Under these definitions, V^π is the fixed-point solution to the respective Bellman equations:

$$\textit{Average:} \qquad V^\pi(s) - \ \mathbb{E}_{s',a|s\sim d_\pi}[V^\pi(s')] = \mathbb{E}_{a|s\sim\pi}[r(s,a) - R_\pi]\,, \tag{1}$$

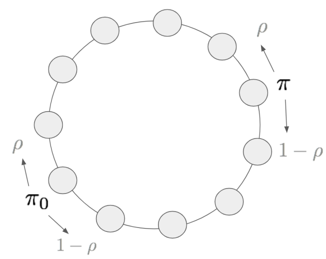
$$\textit{Discounted:} \qquad V^\pi(s) - \gamma\mathbb{E}_{s',a|s\sim d_\pi}[V^\pi(s')] = \mathbb{E}_{a|s\sim\pi}[r(s,a)]\,. \tag{2}$$

Importance Sampling IS represents a major class of approaches to off-policy estimation, which, in principle, only applies to the finite-horizon reward R_π^T when the trajectory is truncated at a finite time step $T < \infty$. IS-based estimators are based on the following change-of-measure equality:

$$R_\pi^T = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[w_{0:T}(\boldsymbol{\tau})R^T(\boldsymbol{\tau})], \quad \text{with} \quad w_{0:T}(\boldsymbol{\tau}) := \frac{p_\pi(\boldsymbol{\tau}_{0:T})}{p_{\pi_0}(\boldsymbol{\tau}_{0:T})} = \prod_{t=0}^{T} \beta_{\pi/\pi_0}(a_t|s_t), \tag{3}$$

where $\beta_{\pi/\pi_0}(a|s) := \pi(a|s)/\pi_0(a|s)$ is the single-step density ratio of policies π and π_0 evaluated at a particular state-action pair (s, a), and $w_{0:T}$ is the density ratio of the trajectory $\boldsymbol{\tau}$ up to time T. Methods based on (3) are called trajectory-wise IS, or weighted IS (WIS) when the weights are self-normalized [22, 30]. It is possible to improve trajectory-wise IS with the so called step-wise, or per-decision, IS/WIS, which uses weight $w_{0:t}$ for reward r_t at time t, yielding smaller variance [30]. More details about these estimators are given in Appendix A.

The Curse of Horizon The importance weight $w_{0:T}$ is a product of T density ratios, whose variance can grow exponentially with T. Thus, IS-based estimators have not been widely successful in long-horizon problems, let alone infinite-horizon ones where $w_{0:\infty}$ may not even be well-defined. While WIS estimators often have reduced variance, the exponential dependence on horizon is unavoidable in general. We call this phenomenon in IS/WIS-based estimators the *curse of horizon*.



Not all hope is lost, however. To see this, consider an MDP with n states and 2 actions, where states are arranged on a circle (see figure on the right). The two actions deterministically move the agent from the current state to the neighboring state counterclockwise and clockwise, respectively. Suppose we are given two policies with opposite effects: the behavior policy π_0 moves the agent clockwise with probability ρ, and the target policy π moves the agent counterclockwise with probability ρ, for some constant $\rho \in (0, 1)$. As shown in Appendix B, IS and WIS estimators suffer from exponentially large variance when estimating the average reward of π. However, a keen reader will realize that the two policies are symmetric, and thus their stationary state visitation distributions are identical. As we show in the sequel, this allows us to estimate the expected reward using a much more efficient importance sampling, whose importance weight equals the single-step density ratio $\beta_{\pi/\pi_0}(a_t|s_t)$, instead of the cumulative product weight $w_{0:T}$ in (3), allowing us to significantly reduce the variance. Such an observation inspired the approach developed in this paper.

3 Off-Policy Estimation via Stationary State Density Ratio Estimation

As shown in the example above, significant decrease in estimation variance is possible when we apply importance weighting on the state space, rather than the trajectory space. It eliminates the dependency on the trajectory length and is much more suited for long- or infinite-horizon problems. To realize this, we need to introduce an alternative representation of the expected reward. Denote by $d_{\pi,t}(\cdot)$ the distribution of state s_t when we execute policy π starting from an initial state s_0 drawn from an initial distribution $d_0(\cdot)$. We define the average visitation distribution to be

$$d_\pi(s) = \lim_{T\to\infty} \left(\sum_{t=0}^{T} \gamma^t d_{\pi,t}(s)\right) \Big/ \left(\sum_{t=0}^{T} \gamma^t\right). \tag{4}$$

We always assume the limit $T \to \infty$ exists in this work. When $\gamma \in (0, 1)$ in the discounted case, d_π is a discounted average of $d_{\pi,t}$, that is, $d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t d_{\pi,t}(s)$; when $\gamma = 1$ in the average reward case, d_π is the stationary distribution of s_t as $t \to \infty$ under policy π, that is, $d_\pi(s) = \lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} d_{\pi,t}(s) = \lim_{t\to\infty} d_{\pi,t}(s)$.

Following Definition 4, it can be verified that R_π can be expressed alternatively as

$$R_\pi = \sum_{s,a} d_\pi(s)\pi(a|s)r(s,a) = \mathbb{E}_{(s,a)\sim d_\pi}[r(s,a)], \tag{5}$$

where, abusing notation slightly, we use $(s, a) \sim d_\pi$ to denote draws from distribution $d_\pi(s, a) := d_\pi(s)\pi(a|s)$. Our idea is to construct an IS estimator based on (5), where the importance ratio is computed on state-action pairs rather than on trajectories:

$$R_\pi = \mathbb{E}_{(s,a)\sim d_{\pi_0}}\left[w_{\pi/\pi_0}(s)\beta_{\pi/\pi_0}(a,s)r(s,a)\right], \tag{6}$$

where $\beta_{\pi/\pi_0}(a,s) = \pi(a|s)/\pi_0(a|s)$ and $w_{\pi/\pi_0}(s) := d_\pi(s)/d_{\pi_0}(s)$ is the density ratio of the visitation distributions d_π and d_{π_0}; here, $w_{\pi/\pi_0}(s)$ is not known directly but can be estimated, as shown later. Eq 5 allows us to construct a (weighted-)IS estimator by approximating $\mathbb{E}_{(s,a)\sim d_{\pi_0}}[\cdot]$ with data $\{s_t^i, a_t^i, r_t^i\}_{i=1}^m$ obtained when running policy π_0,

$$\hat{R}_\pi = \sum_{i=1}^m \sum_{t=0}^T w_t^i r_t^i, \qquad \text{where} \qquad w_t^i := \frac{\gamma^t w_{\pi/\pi_0}(s_t^i)\beta_{\pi/\pi_0}(a_t^i|s_t^i)}{\sum_{t',i'}\gamma^{t'} w_{\pi/\pi_0}(s_{t'}^{i'})\beta_{\pi/\pi_0}(a_{t'}^{i'}|s_{t'}^{i'})}. \tag{7}$$

This IS estimator works in the space of (s,a), instead of trajectoris $\boldsymbol{\tau} = \{s_t, a_t\}_{t=0}^T$, leading to a potentially significant variance reduction. Returning to the example in Section 2 (see also Appendix B), since the two policies are symmetric and lead to the same stationary distributions, that is, $w_{\pi/\pi_0}(s) = 1$, the importance weight in (6) is simply $\pi(a|s)/\pi_0(a|s)$, independent of the trajectory length. This avoids the excessive variance in long horizon problems. In Appendix A, we provide a further discussion, showing that our estimator can be viewed as a type of *Rao-Backwellization* of the trajectory-wise and step-wise estimators.

3.1 Average Reward Case

The key technical challenge remaining is estimating the density ratio $w_{\pi/\pi_0}(s)$, which we address in this section. For simplifying the presentation, we start with estimating $d_\pi(s)$ for the average reward case and discuss the discounted case in Section 3.2.

Let $\boldsymbol{T}_\pi(s'|s) := \sum_a \boldsymbol{T}(s'|s,a)\pi(a|s)$ be the transition probability from s to s' following policy π. In the average reward case, d_π equals the stationary distribution of $\boldsymbol{T}_\pi$, satisfying

$$d_\pi(s') = \sum_s \boldsymbol{T}_\pi(s'|s)d_\pi(s), \quad \forall s'. \tag{8}$$

Assume the Markov chain of $\boldsymbol{T}_\pi$ is finite state and ergodic, d_π is also the unique distribution that satisfies (8). This simple fact can be leveraged to derive the following key property of $w_{\pi/\pi_0}(s)$.

Theorem 1. *In the average reward case ($\gamma = 1$), assume d_π is the unique invariant distribution of $\boldsymbol{T}_\pi$ and $d_{\pi_0}(s) > 0$, $\forall s$. Then a function $w(s)$ equals $w_{\pi/\pi_0}(s) := d_\pi(s)/d_{\pi_0}(s)$ (up to a constant factor) if and only if it satisfies*

$$\begin{aligned} &\mathbb{E}_{(s,a)|s'\sim d_{\pi_0}}[\Delta(w; s, a, s') \mid s'] = 0, \quad \forall\, s', \\ &\textit{with} \qquad \Delta(w; s, a, s') := w(s)\beta_{\pi/\pi_0}(a|s) - w(s'), \end{aligned} \tag{9}$$

where $\beta_{\pi/\pi_0}(a|s) = \pi(a|s)/\pi_0(a|s)$ and $(s,a)|s' \sim d_{\pi_0}$ denote the conditional distribution $d_{\pi_0}(s,a|s')$ related to joint distribution $d_{\pi_0}(s,a,s') := d_{\pi_0}(s)\pi_0(a|s)\boldsymbol{T}(s'|s,a)$. Note that this is a time-reserved *conditional probability, since it is the conditional distribution of (s,a) given that their next state is s' following policy π_0.*

Because the conditional distribution is time reversed, it is difficult to directly estimate the conditional expectation $\mathbb{E}_{(s,a)|s'}[\cdot]$ for a given s'. This is because we usually can observe only a single data point from $d_{\pi_0}(s,a|s')$ of a fixed s', given that it is difficult to see by chance two different (s,a) pairs transit to the same s'. This problem can be addressed by introducing a discriminator function and constructing a mini-max loss function. Specifically, multiplying (9) with a function $f(s')$ and averaging under $s' \sim d_{\pi_0}$ gives

$$\begin{aligned} L(w,f) &:= \mathbb{E}_{(s,a,s')\sim d_{\pi_0}}\left[\Delta(w; s, a, s')f(s')\right] \\ &= \mathbb{E}_{(s,a,s')\sim d_{\pi_0}}\left[\left(w(s)\beta_{\pi/\pi_0}(a|s) - w(s')\right)f(s')\right]. \end{aligned} \tag{10}$$

Following Theorem 1, we have $w \propto w_{\pi/\pi_0}$ if and only if $L(w,f) = 0$ for any function f. This motivates us to estimate w_{π/π_0} with a mini-max problem:

$$\min_w \left\{ D(w) := \max_{f\in\mathcal{F}} L\left(w/z_w,\ f\right)^2 \right\}, \tag{11}$$

where $\mathcal{F}$ is a set of discriminator functions and $z_w := \mathbb{E}_{s\sim d_{\pi_0}}[w(s)]$ normalizes w to avoid the trivial solution $w \equiv 0$. We shall assume $\mathcal{F}$ to be rich enough following the conditions to be discussed in Section 3.3. A promising choice of a rich function class is neural networks, for which the mini-max problem (11) can be solved numerically in a fashion similar to generative adversarial networks (GANs) [8]. Alternatively, we can take $\mathcal{F}$ to be a ball of a reproducing kernel Hilbert space (RKHS), which enables a closed form representation of $D(w)$ as we show in the following.

Theorem 2. *Assume $\mathcal{H}$ is a RKHS of functions $f(s)$ with a positive definite kernel $k(s, \bar{s})$, and define $\mathcal{F} := \{f \in \mathcal{H}\colon ||f||_{\mathcal{H}} \le 1\}$ to be the unit ball of $\mathcal{H}$. We have*

$$\max_{f\in\mathcal{F}} L(w, f)^2 = \mathbb{E}_{d_{\pi_0}}\left[\Delta(w;\ s, a, s')\Delta(w;\ \bar{s}, \bar{a}, \bar{s}')k(s', \bar{s}')\right], \tag{12}$$

where (s, a, s') and $(\bar{s}, \bar{a}, \bar{s}')$ are independent transition pairs obtained when running policy π_0, and $\Delta(w; s, a, s')$ is defined in (10)*. See Appendix C for more background on RKHS.*

In practice, we approximate the expectation in (12) using discounted empirical distribution of the transition pairs, yielding consistent estimates following standard results on V-statistics [33].

3.2 Discounted Reward Case

We now discuss the extension to the discount case of $\gamma \in (0, 1)$. Similar to the average reward case, we start with a recursive equation that characterizes $d_\pi(s)$ in the discounted case.

Lemma 3. *Following the definition of d_π in* (4)*, for any $\gamma \in (0, 1]$, we have*

$$\gamma \sum_s \boldsymbol{T}_\pi(s'|s)d_\pi(s) - d_\pi(s') + (1-\gamma)d_0(s') = 0, \quad \forall s'. \tag{13}$$

Denote by $(s, a, s') \sim d_\pi$ draws from $d_\pi(s)\pi(a|s)\boldsymbol{T}(s'|s, a)$. For any function f, we have

$$\mathbb{E}_{(s,a,s')\sim d_\pi}[\gamma f(s') - f(s)] + (1-\gamma)\mathbb{E}_{s\sim d_0}[f(s)] = 0. \tag{14}$$

One may view d_π as the invariant distribution of an *induced* Markov chain with transition probability of $(1-\gamma)d_0(s') + \gamma\boldsymbol{T}_\pi(s'|s)$, which follows $\boldsymbol{T}_\pi$ with probability γ, and restarts from initial distribution $d_0(s')$ with probability $1-\gamma$. We can show that d_π exists and is unique under mild conditions [31].

Theorem 4. *Assume d_π is the unique solution of* (13)*, and $d_{\pi_0}(s) > 0$, $\forall s$. Define*

$$L(w, f) = \gamma\mathbb{E}_{(s,a,s')\sim d_{\pi_0}}[\Delta(w; s, a, s')f(s')] + (1-\gamma)\mathbb{E}_{s\sim d_0}[(1 - w(s))f(s)]. \tag{15}$$

Assume $0 < \gamma < 1$, then $w(s) = w_{\pi/\pi_0}(s)$ if and only if $L(w, f) = 0$ for any test function f.

When $\gamma = 1$, the definition in (15) reduces to the average reward case in (10). A subtle difference is that $L(w, f) = 0$ only ensures $w \propto w_{\pi/\pi_0}$ when $\gamma = 1$, while $w = w_{\pi/\pi_0}$ when $\gamma \in (0, 1)$. This is because the additional term $\mathbb{E}_{s\sim d_0}[(1 - w(s))f(s)]$ in (15) forces w to be normalized properly. In practice, however, we still find it works better to pre-normalize w to $\tilde{w} = w/\mathbb{E}_{d_{\pi_0}}[w]$, and optimize the objective $L(\tilde{w},\ f)$.

3.3 Further Theoretical Analysis

In this section, we develop further theoretical understanding on the loss function $L(w, f)$. Lemma 5 below reveals an interesting connection between $L(w, f)$ and the Bellman equation, allowing us to bound the estimation error of density ratio and expected reward with the mini-max loss when the discriminator space $\mathcal{F}$ is chosen properly (Theorems 6 and 7). The results in this section apply to both discounted and average reward cases.

Lemma 5. *Given $L(w, f)$ in* (15)*, and assuming $\mathbb{E}_{d_{\pi_0}}[w] = 1$ in the average reward case, we have*

$$L(w, f) = \mathbb{E}_{s\sim d_{\pi_0}}\left[(w_{\pi/\pi_0}(s) - w(s))\Pi f(s)\right], \tag{16}$$

$$\textit{where} \qquad \Pi f(s) := f(s) - \gamma\mathbb{E}_{(s',a)|s\sim d_\pi}\left[f(s')\right]. \tag{17}$$

Note that Πf equals the left hand side of the Bellman equations (1) *and* (2)*, when $f = V^\pi$.*

Lemma 5 represents $L(w, f)$ as an inner product between $w_{\pi/\pi_0} - w$ and Πf (under base measure d_{π_0}). This provides an alternative proof of Theorem 4, since $L(w, f) = 0,\ \forall f \in \mathcal{F}$ implies that $w_{\pi/\pi_0} - w$ is orthogonal with all Πf and hence $w_{\pi/\pi_0} = w$ when $\{\Pi f \colon f \in \mathcal{F}\}$ is sufficiently rich.

In order to make $(w_{\pi/\pi_0} - w)$ orthogonal to a given function g, it requires "reversing" operator Π: finding a function f_g which solves $g = \Pi f_g$ for given g. Observing that $g = \Pi f_g$ can be viewed as a Bellman equation (Eqs. (1)–(2)) when taking g and f_g to be the reward and value functions, respectively, we can derive an explicit representation of f_g (Lemma 10 in Appendix). This allows one to gain insights into what discriminator set $\mathcal{F}$ would be a good choice, so that minimizing $\max_{f\in\mathcal{F}} L(w, f)$ yields good estimation with desirable properties. In the following, by taking $g(s) \propto \pm\mathbf{1}(s = \tilde{s}),\ \forall \tilde{s}$, we can characterize the conditions on $\mathcal{F}$ under which the mini-max loss upper bounds the estimation error of w_{π/π_0} or d_π.

Theorem 6. *Let $\boldsymbol{T}_\pi^t(s'|s)$ be the t-step transition probability of $\boldsymbol{T}_\pi(s'|s)$. For $\forall \tilde{s} \in \mathcal{S}$, define*

$$f_{\tilde{s}}(s) = \begin{cases} \sum_{t=0}^{\infty} \gamma^t \boldsymbol{T}_\pi^t(\tilde{s}|s) & \text{when } 0 < \gamma < 1, \\ \sum_{t=0}^{\infty} (\boldsymbol{T}_\pi^t(\tilde{s}|s) - d_\pi(\tilde{s})) & \text{when } \gamma = 1, \end{cases} \tag{18}$$

Assume Lemma 5 holds. We have

$$\max_{f\in\mathcal{F}} L(w, f) \geq \|d_\pi(s) - w(s) d_{\pi_0}(s)\|_\infty, \qquad \text{if} \quad \{\pm f_{\tilde{s}} \colon \ \forall \tilde{s} \in \mathcal{S}\} \subseteq \mathcal{F},$$

$$\max_{f\in\mathcal{F}} L(w, f) \geq \left\|w_{\pi/\pi_0} - w\right\|_\infty, \qquad \text{if} \quad \{\pm f_{\tilde{s}}/d_{\pi_0}(\tilde{s}) \colon \ \forall \tilde{s} \in \mathcal{S}\} \subseteq \mathcal{F}.$$

Since our main goal is to estimate the expected total reward R_π instead of the density ratio w_{π/π_0}, it is of interest to select $\mathcal{F}$ to directly bound the estimation error of the total reward. Interestingly, this can be achieved once $\mathcal{F}$ includes the true value function V^π.

Theorem 7. *Define $R_\pi[w]$ 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,*

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

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

4 Related Work

Our off-policy setting is related to, but different from, off-policy value-function learning [30, 29, 37, 12, 25, 21]. Our goal is to estimate a single scalar that *summarizes* the quality of a policy (a.k.a. off-policy value estimation as called by some authors [20]). However, our idea can be extended to estimating value functions as well, by using estimated density ratios to weight observed transitions (c.f., the distribution μ in LSTDQ [16]). We leave this as future work.

IS-based off-policy value estimation has seen a lot of interest recently for short-horizon problems, including contextual bandits [26, 13, 7, 42], and achieved many empirical successes [7, 34]. When extended to long-horizon problems, it faces an exponential blowup of variance, and variance-reduction techniques are used to improve the estimator [14, 39, 10, 42]. However, it can be proved that in the worst case, the mean squared error of *any* estimator has to depend exponentially on the horizon [20, 10]. Fortunately, many problems encountered in practical applications may present structures that enable more efficient off-policy estimation, as tackled by the present paper. An interesting open direction is to characterize theoretical conditions that can ensure tractable estimation for long horizon problems.

Few prior work directly target *infinite*-horizon problems. There exists approaches that use simulated samples to estimate stationary state distributions [1, Chapter IV]. However, they need a reliable model to draw such simulations, a requirement that is not satisfied in many real-world applications. To the best of our knowledge, the recently developed COP-TD algorithm [11] is the only work that attempts to estimate w_{π/π_0} as an intermediate step of estimating the value function of a target policy π. They take a stochastic-approximation approach and show asymptotic consistence. However, extending their approach to continuous state/action spaces appears challenging.

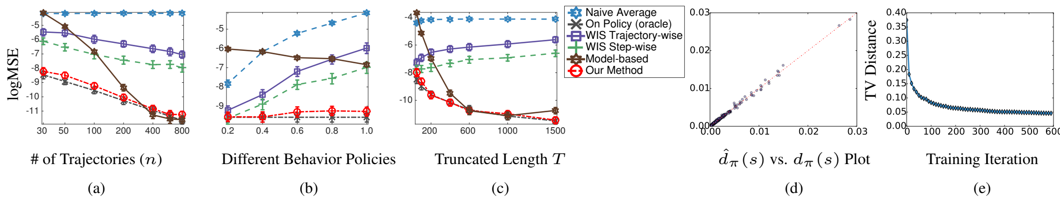


Figure 1: Results on Taxi environment with average reward ($\gamma = 1$). (a)-(b) show the performance of various methods as the number of trajectory (a) and the difference between behavior and target policies (b) vary. (c) shows the change of truncated length T. (d) shows that scatter plot of pairs $(\hat{d}_\pi(s), d_\pi(s)), \forall s$. The diagonal lines means exact estimation. (e) shows the weighted total variation distance between $\hat{d}_\pi := \hat{w} d_{\pi_0}$ and d_π along the training iteration of the ratio estimator $\hat{w}$. The number of trajectory is fixed to be 100 in (b,c,d). The potential behavior policy π_+ (the right most points in (b)) is used in (a,c,d,e).

Finally, there is a comprehensive literature of two-sample density ratio estimation [e.g., 27, 35], which estimates the density ratio of two distributions from pairs of their samples. Our problem setting is different in that we only have data from d_{π_0}, but not from d_π; this makes the traditional density ratio estimators inapplicable to our problem. Our method is made possible by taking the special temporal structure of MDP into consideration.

5 Experiment

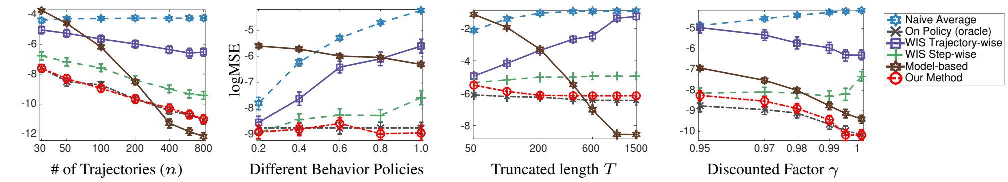
In this section, we conduct experiments on different environmental settings to compare our method with existing off-policy evaluation methods. We compare with the standard trajectory-wise and step-wise IS and WIS methods. We do not report the results of unnormalized IS because they are generally significantly worse than WIS methods [30, 22]. In all the cases, we also compare with an *on-policy oracle* and a *naive averaging* baseline, which estimates the reward using direct averaging over the trajectories generated by the target policy and behavior policy, respectively. For problems with discrete action and state spaces, we also compare with a standard model-based method, which estimates the transition and reward model and then calculates expected reward explicitly using the model up to the desired truncation length. When applying our method on problems with finite and discrete state space, we optimize w and f in the space of all possible functions (corresponding to using a delta kernel in terms of RKHS). For continuous state space, we assume w is a standard feed-forward neural network, and $\mathcal{F}$ is a RKHS with a standard Gaussian RBF kernel whose bandwidth equals the median of the pairwise distances between the observed data points.

Because we cannot simulate truly infinite steps in practice, we use the behavior policy to generate trajectories of length T, and evaluate the algorithms based on the mean square error (MSE) w.r.t. the T-step rewards of a large number of trajectories of length T from the target policy. We expect that our method gets better as T increases, since it is designed for infinite horizon problems, while the IS/WIS methods receive large variance and deteriorate as T increases.

Taxi Environment Taxi [6] is a 2D grid world simulating taxi movement along the grids. A taxi moves North, East, South, West or attends to pick up or drop off a passenger. It receives a reward of 20 when it successfully picks up a passenger or drops her off at the right place, and otherwise a reward of -1 every time step. The original taxi environment would stop when the taxi successfully picks up a passenger and drops her off at the right place. We modify the environment to make it infinite horizon, by allowing passengers to randomly appear and disappear at every corner of the map at each time step. We use a grid size of 5×5, which yields 2000 states in total ($25 \times 2^4 \times 5$, corresponding to 25 taxi locations, 2^4 passenger appearance status and 5 taxi status (empty or with one of 4 destinations)).

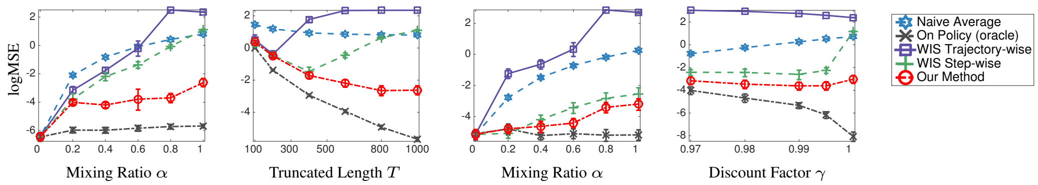
To construct target and behavior policies for testing our algorithm, we set our target policy to be the final policy π_* after running Q-learning for 1000 iterations, and set another policy π_+ after 950 iterations. The behavior policy is $\pi = (1-\alpha)\pi_* + \alpha\pi_+$, where α is a mixing ratio that can be varied.

Results in Taxi Environment Figure 1(a)–(b) show results with average reward. We can see our method performs almost as well as the on-policy oracle, outperforming all the other methods. To evaluate the approximation error of the estimated density ratio $\hat{w}$, we plot in Figure 1(c) the weighted



(a) (b) (c) (d)

Figure 2: Results on Taxi with discounted reward ($0 < \gamma < 1$), as we vary the number of trajectory n (a), the difference between target and behavior policies (b), the truncated length T(c), the discount factor γ (d). The default values of the parameters, unless it is varying, are $\gamma = 0.99$, $n = 200$, $T = 400$. The potential behavior policy π_+ (the right most points in (b)) is used in (a,c,d).



(a) Average Reward Case (b) Average Reward Case (c) Discounted Reward Case (d) Discounted Reward Case

Figure 3: Results on Pendulum. (a)-(b) show the results in the average reward case when we vary the mixing ratio α in the behavior policies and the truncated length T, respectively. (c)-(d) show the results of the discounted reward case when we vary mixing ratio α in the behavior policies and discount factor γ, respectively. The default parameters are $n = 150$, $T = 1000$, $\gamma = 0.99$, $\alpha = 1$.

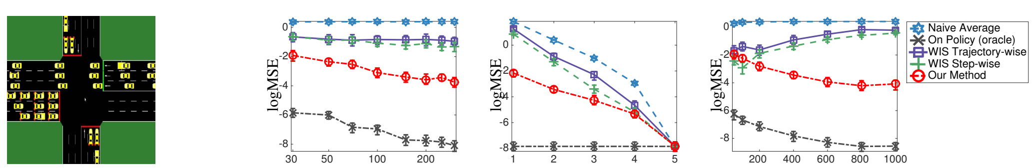
total variation distance between $\hat{d}_\pi = \hat{w} d_{\pi_0}$ with the true d_π with TV distance as we optimize the loss function. Figure 1(d) shows scatter plot of $\{(\hat{d}_\pi(s), d_\pi(s)) \ : \ \forall s \in \mathcal{S}\}$ at convergence, indicating our method correctly estimates the true density ratio over the state space.

Figure 2 shows similar results for discounted reward. From Figure 2(c) and (d), we can see that typical IS methods deteriorate as the trajectory length T and discount factor γ increase, respectively, which is expected since their variance grows exponentially with T. In contrast, our density ratio method performs better as trajectory length T increases, and is robust as γ increases.

Pendulum Environment The Taxi environment features discrete action and state spaces. We now test Pendulum, which has a continuous state space of $\mathbb{R}^3$ and action space of $[-2, 2]$. In this environment, we want to control the pendulum to make it stand up as long as possible (for the average case), or as fast as possible (for small discounted case). The policy is taken to be a truncated Gaussian whose mean is a neural network of the states and variance a constant.

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 3, where we again find that our method generally outperforms the standard trajectory-wise and step-wise WIS, and works favorably in long-horizon problems (Figure 3(b)).

SUMO Traffic Simulator SUMO [15] is an open source traffic simulator; see Figure 4(a) for an illustration. We consider the task of reducing traffic congestion by modelling traffic light control as a reinforcement learning problem [41]. We use TraCI, a built-in "Traffic Control Interface", to interact



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

Figure 4: Results on SUMO (a) with average reward, as we vary the number of trajectories (b), choose different behavior policies (c), and truncated size (d). When being fixed, the default parameters are $n = 250$, $T = 400$. The behavior policy in (c) with x-tick 2 is used in (b) and (d).

with the SUMO simulator. Full details of our environmental settings can be found in Appendix E. Our results are shown in Figure 4, where we again find that our method is consistently better than standard IS methods.

6 Conclusions

We study the off-policy estimation problem in infinite-horizon problems and develop a new algorithm based on direct estimation of the stationary state density ratio between the target and behavior policies. Our mini-max objective function enjoys nice theoretical properties and yields an intriguing connection with Bellman equations that is worth further investigation. Future directions include scaling our method to larger scale problems and extending it to estimate value functions and leverage off-policy data in policy optimization.

Acknowledgement

This work is supported in part by NSF CRII 1830161. We would like to acknowledge Google Cloud for their support.

References

[1] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57 of *Probability Theory and Stochastic Processes*. Springer-Verlag, 2007.

[2] Richard E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[3] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[4] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis Xavier Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[5] Olivier Chapelle, Eren Manavoglu, and Romer Rosales. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology*, 5(4):61:1–61:34, 2014.

[6] Thomas G Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303, 2000.

[7] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1097–1104, 2011.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 2672–2680, 2014.

[9] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[10] Zhaohan Guo, Philip S. Thomas, and Emma Brunskill. Using options and covariance testing for long horizon off-policy policy evaluation. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 2489–2498, 2017.

[11] Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1372–1383, 2017.

[12] Assaf Hallak, Aviv Tamar, Remi Munos, and Shie Mannor. Generalized emphatic temporal difference learning: Bias-variance analysis. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 1631–1637, 2016.

[13] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[14] Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 652–661, 2016.

[15] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International Journal On Advances in Systems and Measurements*, 5(3&4), 2012.

[16] Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

[17] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[18] Lihong Li, Shunbao Chen, Ankur Gupta, and Jim Kleban. 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*, pages 929–934, 2015.

[19] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 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)*, pages 297–306, 2011.

[20] Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 608–616, 2015.

[21] Hao Liu, Yihao Feng, Yi Mao, Dengyong Zhou, Jian Peng, and Qiang Liu. Action-dependent control variates for policy optimization via stein identity. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[22] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer-Verlag, 2001.

[23] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. 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)*, pages 1077–1084, 2014.

[24] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[25] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 1046–1054, 2016.

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

[27] XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *Information Theory, IEEE Transactions on*, 56(11):5847–5861, 2010.

[28] Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. `http://statweb.stanford.edu/~owen/mc`.

[29] Doina Precup, Richard S. Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with funtion approximation. In *Proceedings of the 18th Conference on Machine Learning (ICML)*, pages 417–424, 2001.

[30] Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 759–766, 2000.

[31] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, New York, 1994.

[32] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[33] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[34] Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems 23 (NIPS-10)*, pages 2217–2225, 2010.

[35] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[36] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, March 1998.

[37] Richard S. Sutton, A. Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17(73):1–29, 2016.

[38] Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM)*, pages 1587–1594, 2013.

[39] Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2139–2148, 2016.

[40] Philip S. Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, Ishan Durugkar, and Emma Brunskill. 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)*, pages 4740–4745, 2017.

[41] Elise Van der Pol and Frans A Oliehoek. Coordinated deep reinforcement learners for traffic light control. In *NIPS Workshop on Learning, Inference and Control of Multi-Agent Systems*, 2016.

[42] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3589–3597, 2017.

A Several Variants of IS- and WIS-based Estimators

Denote by $\gamma_t = \gamma^t / \sum_{t=0}^{T} \gamma^t$ for notation simplicity. Define

$$w_{0:T}(\boldsymbol{\tau}) := \prod_{t=0}^{T} \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}.$$

Then we have the following two key formulas, which derive the trajectory-wise, and step-wise importance sampling (IS) estimators, respectively.

$$R_\pi^T = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[\sum_{t=0}^{T} w_{0:T}(\boldsymbol{\tau})\gamma_t r_t\right] \qquad \textit{(Trajectory-wise)} \tag{19}$$

$$= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[\sum_{t=0}^{T} w_{0:t}(\boldsymbol{\tau})\gamma_t r_t\right] \qquad \textit{(Step-wise)} \tag{20}$$

where the only difference of (19) and (20) is that (20) replaces the $w_{0:T}$ in (19) with $w_{0:t}$, yielding smaller variance without changing the expectation. This is made possible because $w_{0:t} = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[w_{0:T}(\boldsymbol{\tau}) \mid \boldsymbol{\tau}_{0:t}]$. Therefore, step-wise estimator can be viewed as Rao-backwellizing each term $w_{0:T}(\boldsymbol{\tau})\gamma_t r_t$ in (19) by conditioning on $\boldsymbol{\tau}_{0:t}$.

Given a set of m observed trajectories $\boldsymbol{\tau}^i = \{s_t^i, a_t^i, r_t^i\}_{t=0}^T$, $\forall i = 1, \ldots, m$, drawn from p_{π_0}. The trajectory-wise and step-wise estimators are

$$\textit{Trajectory-wise:} \quad \hat{R}_\pi^T = \frac{1}{Z_T}\sum_{t=0}^{T}\sum_{i=1}^{m} \gamma_t w_{0:T}^i r_t^i\,, \qquad \textit{Step-wise:} \quad \hat{R}_\pi^T = \sum_{t=0}^{T}\sum_{i=1}^{m}\frac{1}{Z_t}\gamma_t w_{0:t}^i r_t^i\,,$$

where $w_{0:t}^i = w_{0:t}(\boldsymbol{\tau}^i)$ and Z_t is a normalization constant of the importance weights: when $Z_t = m, \quad \forall t$, the corresponding estimators (called Trajectory-wise IS and Step-wise IS, respectively) provide unbiased estimates of R_π^T; when $Z_t = \sum_{i=1}^m w_{0:t}^i$, the corresponding estimators are weighted (or self-normalized) importance sampling (called Trajectory-wise WIS and Step-wise WIS, respectively), which introduce bias but often have lower variance. It has been shown that the Step-wise WIS often performs the best among all these variants [30, 22].

In comparison, our method can be viewed as a further Rao-backwellization of the step-wise estimators. Define

$$w_{t:t}(a_t, s_t) = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[w_{0:T}(\boldsymbol{\tau}) \mid (s_t, a_t)\right] = \frac{d_\pi(s_t)}{d_{\pi_0}(s_t)}\frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}.$$

Then we have

$$R_\pi^T = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[\sum_{t=0}^{T} w_{t:t}(a_t, s_t)\gamma_t r_t\right] \qquad \textit{(Our method)}, \tag{21}$$

where we replace $w_{0:t}$ in (20) with $w_{t:t}$, based on Rao-backwellization conditioning on (s_t, a_t). This gives an empirical estimator:

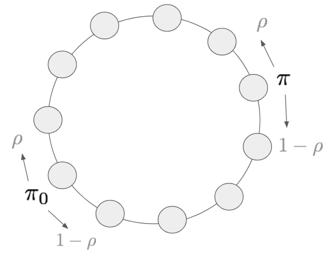
$$\textit{Our method:} \qquad \hat{R}_\pi^T = \sum_{t=0}^{T}\sum_{i=1}^{m}\frac{1}{Z_t}\gamma_t w_{t:t}^i r_t^i,$$

where $w_{t:t}^i = w_{t:t}(a_t^i, s_t^i)$ and $Z_t = m$ or $Z_t = \sum_{i=1}^m w_{t:t}^i$. Comparing this with the trajectory-wise and step-wise estimators, it is easy to expect that it yields smaller variance, when ignoring the estimation error of $w_{t:t}$.

B A motivating example

Here we provide an example when $w_{0:T}$ is exponential on the trajectory length T, yielding high variance in trajectory-wise and step-wise estimators in long horizon problems, while the variance of our stationary density ratio based importance weight $w_{t:t}$ stays to be a constant as T increases.

The MDP has n states: $\mathcal{S} = \{0, 1, \ldots, n-1\}$, arranged on a circle (see the figure on the right), where n is an odd number. There are two actions, left (L) and right (R). The left action moves the agent from the current state counterclockwise to the next state, and the right action has the opposite (clockwise) effect. The deterministic reward is 0 if taking action L and 1 otherwise. In summary, we have for any s and a that



$$
\begin{aligned}
\boldsymbol{T}(s'|s, \mathsf{L}) &= \mathbb{I}(s' = s - 1 \bmod n) \\
\boldsymbol{T}(s'|s, \mathsf{R}) &= \mathbb{I}(s' = s + 1 \bmod n) \\
r(s, a) &= \mathbb{I}(a = \mathsf{R})\,.
\end{aligned}
$$

Suppose we are given two policies. The behavior policy π_0 and target policy π choose action R with probability ρ and $1-\rho$, respectively. We focus on the average reward ($\gamma = 1$) here.

Claim #1. Stationary density ratio $w_{t:t}$ stays constant as $t \to \infty$. First, note that the MDP is ergodic under either policy, as n is odd. Since π_0 and π are symmetric, their stationary distributions are identical, that is, $d_\pi(s)/d_{\pi_0}(s) = 1$. In fact, both $d_\pi = d_{\pi_0}$ are uniform over $\mathcal{S}$. Therefore,

$$
w_{t:t}(s, \mathsf{R}) = \frac{d_\pi(s)\pi(\mathsf{R}|s)}{d_{\pi_0}(s)\pi_0(\mathsf{R}|s)} = \frac{\pi(\mathsf{R}|s)}{\pi_0(\mathsf{R}|s)} = \frac{\rho}{1-\rho}\,,
$$

and similarly $w_{t:t}(s, \mathsf{L}) = (1-\rho)/\rho$. Both ratios are *independent* of the trajectory length, and have *zero* variance.

Claim #2. Variance of trajectory-wise IS weight $w_{0:T}$ grows exponentially in T.

Proposition 8. *Under the setting above, let $\boldsymbol{\tau} = \{s_t, a_t, r_t\}_{0 \le t \le T}$ be a trajectory drawn from the behavior policy π_0, we have*

$$
\begin{aligned}
&\mathrm{var}_{p_{\pi_0}}[w_{0:T}(\boldsymbol{\tau})] = A_\rho^{T+1} - 1, \\
&\mathrm{var}_{p_{\pi_0}}\left[w_{0:T}(\boldsymbol{\tau})R^T(\boldsymbol{\tau})\right] = B_{\rho,T}A_\rho^{T-1} - (1-\rho)^2,
\end{aligned}
$$

where

$$
A_\rho := \frac{\rho^3 + (1-\rho)^3}{(1-\rho)\rho}, \qquad B_{\rho,T} = \frac{(1-\rho)\rho}{T+1} + \frac{(1-\rho)^4}{\rho}\,.
$$

Obviously, $A_\rho > 1$ for $\rho \neq 1/2$ and $A_\rho = 1$ for $\rho = 1/2$, and $B_{\rho,T} > 0$ for large enough T. Therefore, the variance of both the trajectory-wise importance weights and the corresponding estimator grow exponentially in the order of A_ρ^T.

Remark When $\rho = 1/2$, it reduces to the on-policy case of $\pi = \pi_0$, for which we can show that $\mathrm{var}_{p_{\pi_0}}[w_{0:T}(\boldsymbol{\tau})] = 0$ (since $w_{0:T}(\boldsymbol{\tau}) = 1$), and $\mathrm{var}_{p_{\pi_0}}\left[w_{0:T}(\boldsymbol{\tau})R^T(\boldsymbol{\tau})\right] = 1/(4(T+1))$.

Proof. From the definition of the setting, it is easy to show that

$$
R^T(\boldsymbol{\tau}) = \frac{F(\boldsymbol{\tau})}{T+1}, \qquad w_{0:T}(\boldsymbol{\tau}) = \prod_{t=0}^{T} \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)} = \left(\frac{1-\rho}{\rho}\right)^{2F(\boldsymbol{\tau})-(T+1)}
$$

where

$$
F(\boldsymbol{\tau}) = \sum_{t=0}^{T} \mathbb{I}(a_t = \mathsf{R}).
$$

Under policy π_0, $F(\boldsymbol{\tau})$ follows a Binomial distribution $Binomial(T+1, \rho)$. The first order moments can be easily calculated as follows

$$
\mathbb{E}_{\boldsymbol{\tau} \sim p_{\pi_0}}[w_{0:T}(\boldsymbol{\tau})] = 1, \qquad \mathbb{E}_{\boldsymbol{\tau} \sim p_{\pi_0}}[w_{0:T}(\boldsymbol{\tau})R^T(\boldsymbol{\tau})] = \mathbb{E}_{\boldsymbol{\tau} \sim p_\pi}[R^T(\boldsymbol{\tau})] = 1 - \rho.
$$

It remains to calculate the second order moments. We achieve this by leveraging the moment-generating function (MGF) of Binomial distribution:

$$\Phi(\lambda) := \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[\exp(\lambda F(\boldsymbol{\tau}))] = (1-\rho+\rho\exp(\lambda))^{T+1}, \quad \forall \lambda\in\mathbb{R}. \tag{22}$$

It will turn out be useful to consider the derivatives of $\Phi(\lambda)$:

$$\begin{aligned}\Phi'(\lambda) &= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[\exp(\lambda F(\boldsymbol{\tau}))F(\boldsymbol{\tau})] \\ &= (T+1)(1-\rho+\rho\exp(\lambda))^{T}\rho\exp(\lambda),\end{aligned}$$

and

$$\begin{aligned}\Phi''(\lambda) &= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[\exp(\lambda F(\boldsymbol{\tau}))F(\boldsymbol{\tau})^2] \\ &= (T+1)(1-\rho+\rho\exp(\lambda))^{T-1}(1-\rho+(T+1)\rho\exp(\lambda))\rho\exp(\lambda).\end{aligned} \tag{23}$$

For convenience, define $C=(1-\rho)/\rho$, and we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[w_{0:T}(\boldsymbol{\tau})^2] &= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[(C^{2F(\boldsymbol{\tau})-(T+1)})^2] \\ &= \Phi(4\log C)\cdot C^{-2(T+1)} \\ &= \left[(1-\rho+\rho C^4)C^{-2}\right]^{T+1} \\ &= A_\rho^{T+1},\end{aligned}$$

where we use the fact that $(1-\rho+\rho C^4)C^{-2} = \frac{\rho^3+(1-\rho)^3}{(1-\rho)\rho} = A_\rho$. Similarly, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[w_{\pi/\pi_0}(\boldsymbol{\tau})^2 R(\boldsymbol{\tau})^2\right] \\ &= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[C^{4F(\boldsymbol{\tau})-2(T+1)}F(\tau)^2\right]/(T+1)^2 \\ &= \Phi''(4\log C)C^{-2(T+1)}/(T+1)^2 \\ &= ((1-\rho+\rho C^4)C^{-2})^{T-1}(C/(T+1)+\rho C^4)\rho^2 \\ &= B_{\rho,T}A_\rho^{T-1}\end{aligned}$$

where we use the fact that $B_{\rho,T}=(C/(T+1)+\rho C^4)\rho^2$. It is then straightforward to calculate the variance from here. ☐

Claim #3. Variance of trajectory-wise WIS weight grows exponentially in T. Although weighted-IS (WIS) often improves over IS estimators by using self-normalized weights, it cannot eliminate the exponential dependence on the trajectory length. Here, we calculate the asymptotic variance of trajectory-wise WIS using delta method [28, Chapter 9].

Proposition 9. *Let $\hat{R}_{n,wis}$ be the trajectory-wise WIS estimator of R_π based on n copies of independent trajectories drawn from π_0, we have*

$$\mathbb{E}_{p_{\pi_0}}[(\hat{R}_{n,wis}-R_\pi)^2] = \frac{1}{n}D_{\rho,T}A_\rho^T + o\left(\frac{1}{n}\right),$$

where $D_{\rho,A} = B_{\rho,T}A_\rho^{-1} - 2(1-\rho)^3/\rho + (1-\rho)^2A_\rho$, with A_ρ and $B_{\rho,T}$ defined in Proposition 8.

Proof. The asymptotic mean square error (MSE) of a self-normalized importance sampling estimator can be estimated using the delta method [28, Chapter 9]:

$$\begin{aligned}&\mathbb{E}_{p_{\pi_0}}[(\hat{R}_{n,\text{wis}}-R_\pi)^2] \\ &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[w_{\pi/\pi_0}(\boldsymbol{\tau})^2(R(\boldsymbol{\tau})-R_\pi)^2\right] + o\left(\frac{1}{n}\right).\end{aligned}$$

Note that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[w_{\pi/\pi_0}(\boldsymbol{\tau})^2(R(\boldsymbol{\tau})-R_\pi)^2\right] \\ &= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[w_{\pi/\pi_0}(\boldsymbol{\tau})^2R(\boldsymbol{\tau})^2\right] - 2R_\pi\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[w_{\pi/\pi_0}(\boldsymbol{\tau})^2R(\boldsymbol{\tau})] + R_\pi^2\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[w_{\pi/\pi_0}(\boldsymbol{\tau})^2\right],\end{aligned}$$

where the first and third terms have been calculated in the proof of Proposition 8. We just need to calculate the cross term:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}[w_{\pi/\pi_0}(\boldsymbol{\tau})^2 R(\boldsymbol{\tau})] &= \mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[C^{4F(\boldsymbol{\tau})-2(T+1)}F(\tau)\right]/(T+1)\\ &= \Phi'(4\log C)C^{-2(T+1)}/(T+1)\\ &= \left[(1-\rho+\rho C^4)C^{-2}\right]^T \rho C^2\\ &= (1-\rho)^2/\rho A_\rho^T.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\tau}\sim p_{\pi_0}}\left[w_{\pi/\pi_0}(\boldsymbol{\tau})^2(R(\boldsymbol{\tau})-R_\pi)^2\right] &= B_{\rho,T}A_\rho^{T-1} \;-\; 2R_\pi(1-\rho)^2/\rho A_\rho^T \;+\; R_\pi^2 A_\rho^{T+1}\\ &= D_{\rho,T}A_\rho^T,\end{aligned}$$

where

$$\begin{aligned}D_{\rho,T} &:= B_{\rho,T}A_\rho^{-1} \;-\; 2R_\pi(1-\rho)^2/\rho \;+\; R_\pi^2 A_\rho\\ &= B_{\rho,T}A_\rho^{-1} \;-\; 2(1-\rho)^3/\rho \;+\; (1-\rho)^2 A_\rho.\end{aligned}$$

We used $R_\pi = 1-\rho$ here.

C Proofs

Reproducing Kernel Hilbert Space (RKHS) We start with a brief, informal introduction of RKHS. A symmetric function $k(s,s')$ is called positive definite if all matrices of form $[k(s_i,s_j)]_{ij}$ are positive definite for any $\{s_i\}\subseteq\mathcal{S}$. Related to every positive definite kernel $k(s,s')$ is an unique RKHS $\mathcal{H}$ which is the closure of functions of form $f(s)=\sum_i a_i k(s,s_i)$, $\forall a_i\in\mathbb{R},\ s_i\in\mathcal{S}$, equipped with a norm and inner product defined as

$$\langle f,g\rangle_{\mathcal{H}} = \sum_{ij} a_i b_j k(s_i,s_j), \qquad\qquad \|f\|_{\mathcal{H}}^2 = \sum_{ij} a_i a_j k(s_i,s_j),$$

where we assume $g(x)=\sum_i b_i k(s,s_i)$. A simple yet important fact that our proof will leverage is that

$$\|f\|_{\mathcal{H}} = \max_{g\in\mathcal{F}}\langle f,\ g\rangle_{\mathcal{H}}, \qquad \text{where} \qquad \mathcal{F}=\{g\in\mathcal{H}\colon\ \|g\|_{\mathcal{H}}\le 1\}.$$

A key property of RKHS is the so called reproducing property, which says

$$f(s)=\langle f(\cdot),\ k(s,\cdot)\rangle_{\mathcal{H}}, \qquad \text{and hence} \qquad k(s,s')=\langle k(s,\cdot),\ k(s',\cdot)\rangle_{\mathcal{H}}.$$

In our proof, we will consider functions of form $f(s)=\mathbb{E}_{s'\sim d}[w(s')k(s,s')]$ for some function w and distribution d, for which one can show that

$$\max_{g\in\mathcal{F}}\langle f,\ g\rangle_{\mathcal{H}} = \|f\|_{\mathcal{H}} = \mathbb{E}_{s,s'\sim d}[w(s)w(s')k(s,s')]^{1/2};$$

this can be proved using the reproducing property as follows

$$\begin{aligned}\|f\|_{\mathcal{H}}^2 = \langle f,f\rangle_{\mathcal{H}} &= \langle \mathbb{E}_{s\sim d}[w(s)k(\cdot,s)],\ \ \mathbb{E}_{s'\sim d}[w(s')k(\cdot,s')]\rangle_{\mathcal{H}}\\ &= \mathbb{E}_{s,s'\sim d}[w(s)w(s')\langle k(\cdot,s),\ k(\cdot,s')\rangle_{\mathcal{H}}]\\ &= \mathbb{E}_{s,s'\sim d}[w(s)w(s')k(s,s')].\end{aligned}$$

For more introduction to RKHS, see [32, 3, 24], to name only a few.

Proof of Theorem 1. Note that $d_{\pi_0}(s,a|s') = \frac{d_{\pi_0}(s)\pi_0(a|s)\boldsymbol{T}(s'|s,a)}{d_{\pi_0}(s')}$. Therefore, (9) is equivalent to

$$\begin{aligned}w(s') &= \mathbb{E}_{(s,a)|s'\sim\pi_0}\left[w(s)\frac{\pi(a|s)}{\pi_0(a|s)}\ \middle|\ s'\right] = \sum_{s,a}\frac{d_{\pi_0}(s)\pi_0(a|s)\boldsymbol{T}(s'|s,a)}{d_{\pi_0}(s')}w(s)\frac{\pi(a|s)}{\pi_0(a|s)}\\ &= \frac{1}{d_{\pi_0}(s')}\sum_{s,a}\boldsymbol{T}(s'|s,a)\pi(a|s)d_{\pi_0}(s)w(s), \qquad \forall s'.\end{aligned}$$

Denote $g(s) := d_{\pi_0}(s)w(s)$. Since $d_{\pi_0}(s') > 0$ for all s', we find that (9) is equivalent to

$$g(s') = \sum_{s,a} \boldsymbol{T}(s'|s,a)\pi(a|s)g(s), \qquad \forall s'. \tag{24}$$

This implies that $g(s)$ is invariant under Markov transition $\boldsymbol{T}(s'|s,a)\pi(a|s)$. Because $d_\pi(s)$ is the unique stationary distribution under the same Markov transition, (24) holds if and only if $g(s) \propto d_\pi(s)$, or equivalently, $w(s) \propto w_{\pi/\pi_0}(s)$. This completes the proof. □

Proof of Theorem 2. By the reproducing property of RKHS, we have $f(s) = \langle f(\cdot), k(s,\cdot)\rangle\rangle_{\mathcal{H}}$. This gives $L(w,f) = \langle f, \phi^* \rangle_{\mathcal{H}}$, where $\phi^*(\cdot) = \mathbb{E}_{\pi_0}[\Delta(w; \bar{s}, \bar{a}, \bar{s}')k(\bar{s}', \cdot)]$. The results then follow by

$$\max_f L(w,f)^2 = \max_{f \in \mathcal{F}} \langle f, \phi^* \rangle_{\mathcal{H}}^2 = \|\phi^*\|_{\mathcal{H}}^2 = \mathbb{E}_{\pi_0}\left[\Delta(w;\ s,a,s')\Delta(w;\ \bar{s},\bar{a},\bar{s}')k(s',\bar{s}')\right].$$

□

Proof of Lemma 3. Assume $\gamma \in (0,1)$. The definition in (4) gives $d_\pi(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t d_{\pi,t}(s)$. Therefore,

$$\begin{aligned}
d_\pi(s') &= (1-\gamma)\sum_{t=0}^\infty \gamma^t d_{\pi,t}(s') \\
&= (1-\gamma)d_0(s') + (1-\gamma)\sum_{t=1}^\infty \gamma^t d_{\pi,t}(s') \\
&= (1-\gamma)d_0(s') + (1-\gamma)\gamma\sum_{t=0}^\infty \gamma^t d_{\pi,t+1}(s') \\
&= (1-\gamma)d_0(s') + (1-\gamma)\gamma\sum_{t=0}^\infty \gamma^t \sum_s \boldsymbol{T}_\pi(s'|s)d_{\pi,t}(s) \qquad // \ d_{\pi,t+1}(s') = \sum_{s,a} \boldsymbol{T}_\pi(s'|s)d_{\pi,t}(s) \\
&= (1-\gamma)d_0(s') + \gamma\sum_s \boldsymbol{T}_\pi(s'|s)\left((1-\gamma)\sum_{t=0}^\infty \gamma^t d_{\pi,t}(s)\right) \\
&= (1-\gamma)d_0(s') + \gamma\sum_s \boldsymbol{T}_\pi(s'|s)d_\pi(s) \\
&= (1-\gamma)d_0(s') + \gamma\sum_{s,a} \boldsymbol{T}(s'|s,a)\pi(a|s)d_\pi(s).
\end{aligned}$$

Multiplying both sides by $f(s')$ and summing over s', we get

$$\sum_{s'} d_\pi(s')f(s') = (1-\gamma)\sum_{s'} d_0(s')f(s') + \gamma\sum_{s,a,s'} \boldsymbol{T}(s'|s,a)\pi(a|s)d_\pi(s)f(s').$$

Recall that $(s,a,s') \sim d_\pi$ denotes sampling from the joint distribution of $d_\pi(s,a,s') = d_\pi(s)\boldsymbol{T}(s',a|s)\pi(a|s)$. Note that under this joint distribution, the marginal distribution of s' is different from $d_\pi(s)$.[1]

The above equation is equivalent to

$$\mathbb{E}_{s'\sim d_\pi}[f(s')] = (1-\gamma)\mathbb{E}_{s'\sim d_0}[f(s')] + \gamma\mathbb{E}_{(s,a,s')\sim d_\pi}[f(s')].$$

For notation, changing the dummy variable s' in $\mathbb{E}_{s'\sim d_\pi}[\cdot]$ and $\mathbb{E}_{s'\sim d_0}[\cdot]$ to s gives

$$\mathbb{E}_{s\sim d_\pi}[f(s)] = (1-\gamma)\mathbb{E}_{s\sim d_0}[f(s)] + \gamma\mathbb{E}_{(s,a,s')\sim d_\pi}[f(s')].$$

Therefore,

$$\mathbb{E}_{(s,a,s')\sim d_\pi}[\gamma f(s') - f(s)] + (1-\gamma)\mathbb{E}_{s\sim d_0}[f(s)] = 0.$$

□

[1] This is different from the average reward case, in which $d_\pi(s)$ is the stationary distribution of $\boldsymbol{T}_\pi$.

Proof of Theorem 4. Define

$$\delta(g,\ s') := \gamma \sum_{s} \boldsymbol{T}_\pi(s'|s)g(s) - g(s') + (1-\gamma)d_0(s'),$$

where g is any function. Then by assumption, we have $g(s) = d_\pi(s)$ if and only if $\delta(g,\ s') = 0$ for any s'. Replacing d_π with d_{π_0} and $f(s)$ with $w(s)f(s)$ in (14) gives

$$\mathbb{E}_{(s,a,s')\sim d_{\pi_0}}[w(s)f(s) - \gamma w(s')f(s')] = (1-\gamma)\mathbb{E}_{s\sim d_0}[w(s)f(s)]\,.$$

Plugging it into the definition of $L(w, f)$ in (15), we get

$$\begin{aligned}
&L(w, f)\\
&= \gamma\mathbb{E}_{(s,a,s')\sim d_{\pi_0}}[(\beta_{\pi/\pi_0}(a|s)w(s) - w(s'))f(s')] + (1-\gamma)\mathbb{E}_{s\sim d_0}[(1-w(s))f(s)]\\
&= \gamma\mathbb{E}_{(s,a,s')\sim d_{\pi_0}}[(\beta_{\pi/\pi_0}(a|s)w(s)f(s')] - \mathbb{E}_{s\sim d_{\pi_0}}[w(s)f(s)] + (1-\gamma)\mathbb{E}_{s\sim d_0}[f(s)] \qquad (25)\\
&= \gamma\mathbb{E}_{(s,a,s')\sim d_\pi}[w_{\pi/\pi_0}(s)^{-1}w(s)f(s')] - \mathbb{E}_{s\sim d_\pi}[w_{\pi/\pi_0}(s)^{-1}w(s)f(s)] + (1-\gamma)\mathbb{E}_{s\sim d_0}[f(s)]\\
&= \sum_{s'} \delta(g, s')f(s')\,,
\end{aligned}$$

where we have defined $g(s) := d_\pi(s)w_{\pi/\pi_0}(s)^{-1}w(s)$. Therefore, $L(w, f) = 0$ for $\forall f$ is equivalent to $\delta(g, s') = 0$ for $\forall s'$, which is in turn equivalent to $g(s) = d_\pi(s)$. Therefore, we have $w(s) = w_{\pi/\pi_0}(s)$ when $0 < \gamma < 1$, and $g(s) \propto d_\pi(s)$, or equivalently, $w(s) \propto w_{\pi/\pi_0}(s)$, when $\gamma = 1$. $\square$

Proof of Lemma 5. Note that

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

Following the proof of Theorem 4 up to (25), we have

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

Since $L(w_{\pi/\pi_0}, f) = 0$, we have

$$\begin{aligned}
L(w, f) &= L(w, f) - L(w_{\pi/\pi_0}, f)\\
&= \mathbb{E}_{s\sim d_{\pi_0}}[(w_{\pi/\pi_0}(s) - w(s))\Pi f(s)]\,.
\end{aligned}$$

$\square$

Lemma 10. *For any function $g(s)$, define $\bar{g} = \mathbb{E}_{s\sim d_\pi}[g(s)]$ and*

$$f_g(s) = \begin{cases} \mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}[\sum_{t=0}^{\infty}\gamma^t g(s_t) \mid s_0 = s] & \textit{when } 0 < \gamma < 1,\\ \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}[\sum_{t=0}^{T} g(s_t) - \bar{g} \mid s_0 = s] & \textit{when } \gamma = 1, \end{cases} \qquad (26)$$

assuming the limits above exist. Then, when $0 < \gamma < 1$, $f = f_g$ is the unique solution of $g = \Pi f$; when $\gamma = 1$ and $\boldsymbol{T}_\pi$ is irreducible, all the solutions of $g - \bar{g} = \Pi f$ satisfies $f = f_g + \mathrm{constant}$.

Proof of Lemma 10. Consider first the discounted case $\gamma \in (0, 1)$, we have

$$\begin{aligned}
\Pi f_g(s) &= f_g(s) - \gamma \mathbb{E}_{(s',a)|s\sim d_\pi}[f_g(s')] \\
&= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(s_t) \mid s_0 = s] - \gamma \mathbb{E}_{(s',a)|s\sim d_\pi} \left[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(s_t) \mid s_0 = s]\right] \\
&= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(s_t) \mid s_0 = s] - \mathbb{E}[\sum_{t=0}^{\infty} \gamma^{t+1} g(s_{t+1}) \mid s_0 = s]] \\
&= \mathbb{E}[g(s_0) \mid s_0 = s] \\
&= g(s)\,.
\end{aligned}$$

For the uniqueness, assume $g = \Pi f_1$ and $g = \Pi f_2$, and $\delta f = f_1 - f_2$, then $\Pi \delta f = 0$, where

$$\delta f(s) = \gamma \sum_{s'} \boldsymbol{T}_\pi(s'|s) \delta f(s').$$

If $0 < \gamma < 1$, we have

$$\|\delta f\|_\infty = \left\| \gamma \sum_{s'} \boldsymbol{T}_\pi(s'|s) \delta f(s') \right\|_\infty \le \gamma \|\delta f\|_\infty \,,$$

which implies $\|\delta f\|_\infty = 0$.

For the average reward case $\gamma = 1$, we have

$$\begin{aligned}
\Pi f_g(s) &= f_g(s) - \mathbb{E}_{(s',a)|s\sim d_\pi}[f_g(s')] \\
&= \lim_{T\to\infty} \mathbb{E}[\sum_{t=0}^{T} (g(s_t) - \bar{g}) \mid s_0 = s] - \mathbb{E}_{(s',a)|s\sim d_\pi}[\mathbb{E}[\sum_{t=0}^{T} (g(s_t) - \bar{g}) \mid s_0 = s]] \\
&= \lim_{T\to\infty} \mathbb{E}[\sum_{t=0}^{T} (g(s_t) - \bar{g}) \mid s_0 = s] - \mathbb{E}[\sum_{t=0}^{T} (g(s_{t+1}) - \bar{g}) \mid s_0 = s] \\
&= \mathbb{E}[g(s_0) - \bar{g} \mid s_0 = s] \\
&= g(s) - \bar{g}\,.
\end{aligned}$$

For the uniqueness, assume $g = \Pi f_1$ and $g = \Pi f_2$, and $\delta f = f_1 - f_2$, then $\delta f = \sum_{s'} \boldsymbol{T}_\pi(s'|s)\delta f(s')$, which implies $\delta f = \sum_{s'} \boldsymbol{T}^n_\pi(s'|s)\delta f(s')$, where $\boldsymbol{T}^n_\pi$ is the n-step transition probability function. If δf is not a constant, there must exists a state $\tilde{s}$ such that $\delta f(\tilde{s}) < \|\delta f\|_\infty$. Since $\boldsymbol{T}_\pi$ is irreducible, there exists a $n > 0$ such that $\boldsymbol{T}^n_\pi(\tilde{s}|s) > 0$. Therefore,

$$\|\delta f\|_\infty = \left\| \boldsymbol{T}^n_\pi(\tilde{s}|s)\delta f(\tilde{s}) + \sum_{s' \ne \tilde{s}} \boldsymbol{T}^n_\pi(s'|s)\delta f(s') \right\|_\infty < \|\delta f\|_\infty \,,$$

which is contradictory. Therefore, δf must be a constant. In fact, functions that satisfies $\delta f = \sum_{s'} \boldsymbol{T}_\pi(s'|s)\delta f(s')$ is called harmonic [17, Lemma 1.16]. □

Proof of Theorem 6. By taking f_g such that $g(s) = \mathbf{1}(s = \tilde{s})$, we have

$$L(w, f_g) = \mathbb{E}_{s\sim d_{\pi_0}}[(w_{\pi/\pi_0}(s) - w(s))g(s)] = d_\pi(\tilde{s}) - w(\tilde{s})d_{\pi_0}(\tilde{s}).$$

We just need to calculate f_g, following Lemma 10.

Note that $\boldsymbol{T}^t_\pi(\tilde{s} \mid s) = \mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}[\mathbf{1}(s_t = \tilde{s}) \mid s_0 = s)]$. When $0 < \gamma < 1$, we have

$$\begin{aligned}
f_g(s) &= \mathbb{E}_{\boldsymbol{\tau}\sim p_\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}(s_t = \tilde{s}) | s_0 = s\right] \\
&= \sum_{t=0}^{\infty} \gamma^t \boldsymbol{T}^t_\pi(\tilde{s}|s).
\end{aligned}$$

Algorithm 1 Main Algorithm (Average Reward Case)

Input: Transition data $\mathcal{D} = \{s_t, a_t, s'_t, r_t\}_t$ from simulator from the behavior policy π_0; a target policy π for which we want to estimate the expected reward. Denote by $\beta_{\pi/\pi_0}(a|s) = \pi(a|s)/\pi_0(a|s)$.

Initial the density ratio $w(s) = w_\theta(s)$ to be a neural network parameterized by θ.

for iteration = 1, 2, ... **do**

Randomly choose a batch $\mathcal{M}$ of size m from the transition data $\mathcal{D}$, $\mathcal{M} \subset \{1, \ldots, n\}$.

Update the parameter θ by $\theta \leftarrow \theta - \epsilon \nabla_\theta \hat{D}(w_\theta / z_{w_\theta})$, where

$$\hat{D}(w) = \frac{1}{|\mathcal{M}|} \sum_{i,j \in \mathcal{M}} \Delta(w, s_i, a_i, s'_i) \Delta(w, s_j, a_j, s'_j) k(s'_i, s'_j),$$

and z_{w_θ} is a normalization constant $z_{w_\theta} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} w_\theta(s_i)$.

end for

Output: Estimate the expected reward of π by $\hat{R}_\pi = \sum_{i=1}^n v_i r_i / \sum_{i=1}^n v_i$, where $v_i = w_\theta(s_i) \beta_{\pi/\pi_0}(a_i, s_i)$.

For the average reward case, note that $\bar{g} = \mathbb{E}_{s \sim d_\pi}[\mathbf{1}(s = \tilde{s})] = d_\pi(\tilde{s})$, so

$$\begin{aligned} f_g(s) &= \mathbb{E}_{\boldsymbol{\tau} \sim p_\pi}\left[\sum_{t=0}^\infty \mathbf{1}(s_t = \tilde{s}) - d_\pi(\tilde{s}) | s_0 = s\right] \\ &= \sum_{t=0}^\infty (\boldsymbol{T}_\pi^t(\tilde{s}|s) - d_\pi(\tilde{s})). \end{aligned}$$

Similarly, we take $g(s) = \mathbf{1}(s = \tilde{s})/d_{\pi_0}(\tilde{s})$, and obtain bounds for $\left\| w_{\pi/\pi_0} - w \right\|_\infty$. □

Proof of Theorem 7. Define $r_\pi(s) = \mathbb{E}_{a|s\sim\pi}[r(s,a)] = E_{a|s\sim\pi_0}[\beta_{\pi/\pi_0}(a|s) r(s,a)]$, then

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

We consider the average reward case first. Following the definition of the operator Π in (17) and the average reward Bellman equation, we have

$$\Pi V^\pi(s) = r_\pi(s) - R_\pi.$$

Following Lemma 10, we have

$$L(w, f) = \mathbb{E}_{s \sim d_{\pi_0}}[(w(s) - w_{\pi/\pi_0}(s))(r_\pi(s) - R_\pi(s))] = R_\pi[w_{\pi/\pi_0}] - R[w] = R_\pi - R_\pi[w].$$

For the discounted case, following the definition of Π and the discounted Bellman equation (2), we have $\Pi V_\pi(s) = r_\pi$, which gives

$$L(w, f) = \mathbb{E}_{s \sim \pi_0}[(w_{\pi/\pi_0}(s) - w(s)) r_\pi(s)] = R_\pi[w_{\pi/\pi_0}] - R[w] = R_\pi - R_\pi[w].$$

□

D Algorithm Details

Algorithm 1 summarizes our main algorithm for the average reward case, where we approximate the mini-max loss function in (12) using empirical averaging of observed data.

The algorithm for the discounted case follows the same idea, but requires some modification due to the additional term in (15). To handle it in a notionally convenient way, we find it is useful to introduce a dummy transition pair $\{s_{-1}, a_{-1}, s'_{-1}, r_{-1}\}$ at time $t = -1$, for which we define $s'_{-1} = s_0$, $r_{-1} = 0$ and $\Delta(w;\ s_{-1}, a_{-1}, s'_{-1}) := 1 - w(s_0) f(s_0)$. Related, we define an augmented discounted visitation distribution via

$$\tilde{d}_\pi(s) = \gamma d_{\pi,t}(s) + (1-\gamma) d_{\pi,-1}(s) = (1-\gamma) \sum_{t=-1}^\infty \gamma^{t+1} d_{\pi,t}(s). \quad (27)$$

Algorithm 2 Main Algorithm (Discounted Reward Case)

Input: Transition data $\mathcal{D} = \{s_t, a_t, s'_t, r_t\}_t$ from the behavior policy π_0; a target policy π for which we want to estimate the expected reward. Denote by $\beta_{\pi/\pi_0}(a|s) = \pi(a|s)/\pi_0(a|s)$. Discount factor $\gamma \in (0, 1]$.

Augment the data with dummy data $\{s_{-1}, a_{-1}, s'_{-1}, r_{-1}\}$ for which $r_{-1} = 0$, $s'_{-1} = s_0$ and $\Delta(w;\ s_{-1}, a_{-1}, s'_{-1}) := 1 - w(s_0)f(s_0)$. Add them to $\mathcal{D}$ to form an augment dataset $\tilde{\mathcal{D}}$.

Initial the density ratio $w(s) = w_\theta(s)$ to be a neural network parameterized by θ.

for iteration = 1, 2, ... **do**

Randomly choose a batch $\mathcal{M} \subseteq \{1, \ldots, n\}$ from the augmented transition data $\tilde{\mathcal{D}}$, by selecting time t with probability proportional to γ^{t+1}.

Update the parameter θ by $\theta \leftarrow \theta - \epsilon \nabla_\theta \hat{D}(w_\theta / z_{w_\theta})$, where

$$\hat{D}(w) = \frac{1}{|\mathcal{M}|} \sum_{i,j \in \mathcal{M}} \Delta(w, s_i, a_i, s'_i)\Delta(w, s_j, a_j, s'_j)k(s'_i, s'_j),$$

and z_{w_θ} is a normalization constant $z_{w_\theta} = \frac{1}{|\mathcal{M}|}\sum_{i \in \mathcal{M}} w_\theta(s_i)$.

end for

Output: Estimate the expected reward of π by $\hat{R}_\pi = \sum_{i=1}^n v_i r_i / \sum_{i=1}^n v_i$, where $v_i = w_\theta(s_i)\beta_{\pi/\pi_0}(a_i, s_i)$.

Under this notation, the loss (15) of discounted case is rewritten into a form identical to the average reward case:

$$\begin{aligned} L(w, f) &= \gamma \mathbb{E}_{(s,a,s') \sim d_{\pi_0}}[\Delta(w; s, a, s')f(s')] + (1-\gamma)\mathbb{E}_{s \sim d_0}[(1 - w(s))f(s)] \\ &= \mathbb{E}_{(s,a,s') \sim \tilde{d}_{\pi_0}}[\Delta(w; s, a, s')f(s')]. \end{aligned}$$

Therefore, following Theorem 2, we have

$$\max_{f \in \mathcal{F}} L(w, f)^2 = \mathbb{E}_{\tilde{d}_{\pi_0}}\left[\Delta(w;\ s, a, s')\Delta(w;\ \bar{s}, \bar{a}, \bar{s}')k(s', \bar{s}')\right], \tag{28}$$

when $\mathcal{F}$ is the ball of RKHS with kernel $k(s', \bar{s}')$.

We can further approximate the expectation $\mathbb{E}_{\tilde{d}_{\pi_0}}[\cdot]$ given a set of augmented trajectories $\tilde{\mathcal{D}} = \{s_t, a_t, s'_t, r_t\}_{t=-1}^T$. Following (27), this can be done by randomly drawing (with replacement) data at time t with probability proportional to γ^t. Let $\{s_t, a_t, s'_t, r_t\}_{t \in \mathcal{M}}$ be a subset of $\tilde{\mathcal{D}}$ generated this way, and the mini-max loss in (28) can be approximated by

$$\max_{f \in \mathcal{F}} L(w, f)^2 \approx \frac{1}{|\mathcal{M}|} \sum_{i,j \in \mathcal{M}} \Delta(w, s_i, a_i, s'_i)\Delta(w, s_j, a_j, s'_j)k(s'_i, s'_j).$$

This equation is identical to the one in Algorithm 1 for the average case, but differs in the way the minibatch $\mathcal{M}$ is generated: it includes the dummy transition at time $t = -1$ with probability $(1 - \gamma)$ and select time t with discounted probability γ^{t+1}. See Algorithm 2 for the summary of the procedure.

E Information on SUMO Traffic Simulator

We provide details of the SUMO traffic simulator and how we formulate it as a standard reinforcement learning problem.

States for SUMO A states of a traffic should provide us with enough information to control the traffic light. A complex way is an image-like representation of the traffic vehicle around the traffic light intersection [41]. Here, to simplify the problem, we add lane detectors around traffic light intersections, and count the total number of vehicles on each lane as states s_t. This should give us enough, though not perfect, information to guide the traffic light agent to choose its action.

Actions For a standard crossing intersection, its traffic light will have a program for 8 phases: "Straight signal for North-South", "Turn-left signal for North-south", "Straight signal for East-West", "Turn-left signal for East-west" and their corresponding "yellow light" slow down signals. Here, we simplify these 4 phases into actions a_t for each traffic light, where we let one big time step t in reinforcement learning setting to be 6 real time steps in SUMO simulator. Within each big time step t, we add a transition of 3 real time steps "yellow light" phase as a buffer to prevent vehicles for "emergency stop" if our agent decides to change light status ($a_t \neq a_T$).

Rewards Our goal is to minimize the total travelling time for all vehicles. Thus, we could set the negative of current aggregate total number of vehicles during the one big time step as reward r_t. To simplify, we can just consider 6 times the current total number of vehicle as a approximation of r_t to make our system simpler.

Policy We use linear policy with the final softmax layer as probability for each action. We train a policy π_* using Cross entropy(CE) method for 10 iterations and set it to be the target policy. And we set the policies at the training iteration 6, 7, 8, 9 as behavior policies, which correspond to x-ticks 1-4 in Figure 4(c).

Other details To simulate on our given network, we also need to design route documents for a vehicle to follow. Each route is a set of roads that connect any two exit nodes from the map. To make simple but reasonable routes for the vehicle, we constrain our routes with at most one turn in the network to avoid detours. We control each route with a fixed probability (different from each route) every time step to generate a vehicle, to guarantee a randomized environment.