Loop Estimator for Discounted Values in Markov Reward Processes

Falcon Z. Dai, Matthew R. Walter
Toyota Technological Institute at Chicago
Chicago, Illinois, USA 60637
{dai, mwalter}@ttic.edu

Abstract

At the working heart of policy iteration algorithms commonly used and studied in the discounted setting of reinforcement learning, the policy evaluation step estimates the value of states with samples from a Markov reward process induced by following a Markov policy in a Markov decision process. We propose a simple and efficient estimator called *loop estimator* that exploits the regenerative structure of Markov reward processes without explicitly estimating a full model. Our method enjoys a space complexity of $O(1)$ when estimating the value of a single positive recurrent state s unlike TD with $O(S)$ or model-based methods with $O\left(S^2\right)$. Moreover, the regenerative structure enables us to show, without relying on the generative model approach, that the estimator has an instance-dependent convergence rate of $\widetilde{O}\left(\sqrt{\tau_s/T}\right)$ over steps T on a single sample path, where τ_s is the maximal expected hitting time to state s. In preliminary numerical experiments, the loop estimator outperforms model-free methods, such as TD(k), and is competitive with the model-based estimator.

1 Introduction

The problem of policy evaluation arises naturally in the context of reinforcement learning (RL) (Sutton and Barto 2018) when one wants to evaluate the (action) values of a policy in a Markov decision process (MDP). In particular, policy iteration (Howard 1960) is a classic algorithmic framework for solving MDPs that poses and solves a policy evaluation problem during each iteration. Being motivated by the setting of reinforcement learning, i.e., the underlying MDP parameters are unknown and samples are obtained interactively, we focus on solving the policy evaluation problem given only a *single* sample path.

Following a stationary Markov policy in an MDP, i.e., actions are determined based solely on the current state, gives rise to a *Markov reward process* (MRP) (Puterman 1994). For the rest of the article, we focus on MRPs and consider the problem of estimating the infinite-horizon *discounted* state values of an unknown MRP.

A straightforward approach to policy evaluation is to estimate the parameters of the MRP and then the value by plugging them into the classic Bellman equation (5) (Bertsekas and Tsitsiklis 1996). We call this the model-based estimator in the sequel. This approach is recently proved to be minimax-optimal given a generative model (Pananjady and Wainwright 2019) and it provides excellent estimates of discounted values in the single sample path setting as well, as our numerical experiments show (Section 5). However, model-based estimators suffer from a space complexity of $O\left(S^2\right)$, where S is the number of states in the MRP. In contrast, *model-free* methods enjoy a lower space complexity of $O(S)$ by not explicitly estimating the model parameters (Sutton 1988) but tend to exhibit a greater estimation error.

A popular class of estimators, k-step bootstrapping temporal difference or TD(k)[1] estimates a state's value based on the estimated values of other states. Like the model-based estimator, TD(k) is based on the classic Bellman equation (5). The key property of the Bellman equation (5) is that the estimate of a state's value is tied to the estimates of other states which makes it hard to study the convergence of a specific state's value estimate in isolation and motivates the traditional approach of generative model in the literature.

Traditionally, prior works (Kearns and Singh 1999; Even-Dar and Mansour 2003; Gheshlaghi Azar, Munos, and Kappen 2013; Pananjady and Wainwright 2019) first show efficient estimation of *all* state values under the assumption that we can generate a sample of next states and rewards starting in each states, and then invoke an argument that such a batch of samples can be obtained over a single sample path when all states are visited for at least once, i.e., over cover times. In this work, we break with the traditional approach by directly studying the convergence of the value estimate of a *single* state over the sample path. The convergence over all states is obtained as a simple consequence of the union bound. Our key insight is that it is possible to circumvent the general difficulties of non-independent samples in the single sample path setting by recognizing the embedded regenerative structure of an MRP. We alleviate the reliance on estimates of other states by studying segments of the sample path that start and end in the same state, i.e., *loops*. This

[1] An important variant is TD(λ), but we do not include it in our experiments since there is not a canonical implementation of the idea of estimating λ-return (Sutton and Barto 2018). However, any implementation is expected to exhibit similar behaviors as TD(k) with large k corresponding to large λ (Kearns and Singh 2000).

results in a novel and simple algorithm we call the *loop estimator* (Algorithm 1) which is a plug-in estimator based on a novel loop Bellman equation (10). One important consequence is that the loop estimator can estimate the value of a single state with a space complexity of $O(1)$ which neither $TD(k)$ or the model-based estimator can achieve.

We first review the requisite definitions (Section 3) and then propose the loop estimator (Section 4.2). First, we analyze the algorithm's rate of convergence over visits to a single state (Theorem 4.2). Second, we study many steps it takes to visit a state. Using the exponential concentration of first return times (Lemma 4.3), we relate visits to their waiting times and establish the rate of convergence over steps (Theorem 4.5). Lastly, we obtain the convergence in ℓ_∞-norm over all states via the union bound as a consequence (Corollary 4.6). Besides theoretical analysis, we also compare the loop estimator to several other estimators numerically on a commonly used example (Section 5). Finally, we discuss the model-based vs. model-free status of the loop estimator (Section 6).

Our main contributions in this paper are two-fold:

- By recognizing the embedded regenerative structure in MRPs, we derive a new Bellman equation over loops, segments that start and end in the same state.
- We introduce *loop estimator*, a novel algorithm that can provably efficiently estimate the discounted values of a single state in an MRP from a single sample path.

In the interest of a concise presentation, we defer detailed proofs to Appendix A of our full technical report (Dai and Walter 2020) with fully expanded logarithmic factors and constants. Similarly, see Appendix B of (Dai and Walter 2020) for extra results. An implementation of the proposed loop estimator and presented experiments is publicly available.[2]

2 Related Works

Much work that formally studies the convergence of value estimators (particularly the TD estimators) relies on having access to independent trajectories that start in *all* states (Dayan and Sejnowski 1994; Even-Dar and Mansour 2003; Jaakkola, Jordan, and Singh 1994; Kearns and Singh 2000). This is called a *generative model* or sometimes a parallel sampling model (Kearns and Singh 1999). Given a convergence over batches of generative samples, we still need some reduction arguments to actually obtain a batch of generative (or parallel) samples over the sample path of a MRP. Kearns and Singh (1999) consider how a set of independent trajectories can be obtained via mixing, i.e., approximately samples from the stationary distribution. This suggests on *average* it takes $O(t_{\text{mix}}/p^*)$-many steps where t_{mix} is the expected steps to get close to the stationary distribution ($1/4$ in total variation distance) and p^* is the smallest probability in the stationary distribution.

This reduction can be improved by considering the steps the chain takes to visit all states at least once, i.e., *cover times*, which is exactly when we have a batch of generative samples. This is an improved reduction in that we can study its convergence rate with high probability instead of the average behavior. But the cover time of a Markov chain can be quite large: its concentration can be related to that of the hitting times to *all* states. In contrast, for a single state, our results scale more favorably with the maximal expected hitting time of that state by a factor of $\log S$. To ensure consistency of estimation is at all possible, we assume that the specific state to estimate is positive recurrent (Assumption 3.1), otherwise we cannot hope to (significantly) improve its value estimate after the final visit (see Appendix B.1 of (Dai and Walter 2020) for an illustrative example). We think that this assumption is reasonable as recurrence is a key feature of many Markov chains and it connects naturally to the online (interactive) setting where we cannot arbitrarily restart the chain. Moreover, this assumption is no stronger than the assumption used in the cover time reduction which assumes that we can repeatedly visit all states. If a resetting mechanism is available, values of transient states can be estimated from values of the recurrent states. Furthermore, in a finite MRP, there is at least one recurrent state due to the infinite length of a trajectory.

Besides the interest in the RL community to study the policy evaluation problem, operation researchers were also motivated to study estimation in order to leverage simulations as a computational tool. In such settings, the restriction of estimating only from a single sample path is usually not a concern. Classic work in simulations by Fox and Glynn (1989) deals with estimating discounted value in a continuous time setting, including an estimator using regenerative structure. In comparison to their work, we provides an instance-dependent rate based on the transition structure which is relevant for the single sample path setting. Haviv and Puterman (1992) and Derman (1970) propose unbiased value estimators whereas the loop estimator is biased due to inversion.

Outside of the studies on reward processes, the regenerative structure of Markov chains has found application in the *local* computation of PageRank (Lee, Ozdaglar, and Shah 2013). We make use of a lemma (Lemma 4.3, whose proof is included in the Appendix A.3 of (Dai and Walter 2020) for completeness) from this work to establish an upper bound on waiting times (Corollary 4.4). Furthermore, we provide an example to support why hitting times do not exponentially concentrate over its expectation in general (see Appendix B.2 of (Dai and Walter 2020)). Similar in spirit to the concept of locality studied by Lee, Ozdaglar, and Shah (2013), our loop estimator enables space-efficient estimation of a single state value with a space complexity of $O(1)$ and an error bound without explicit dependency on the size of the state space. As a consequence, the loop estimator can provably estimate the value of a state with a finite maximal expected hitting time even if the state space is infinite.

Recently, an independent work by Subramanian and Mahajan (2019) makes a similar observation of the regenerative structure and studies using estimates similar to the loop estimator in the context of a policy gradient algorithm. It provides promising experimental results that complement our novel theoretical guarantees on the rates of convergence.

[2] https://github.com/falcondai/loop-estimator

Taken together, these works show that regenerative structure is a promising direction in RL.

3 Preliminaries

3.1 Markov Reward Processes and Markov Chains

Consider a finite state space $\mathcal{S} \coloneqq \{1, \cdots, S\}$ whose size is $S = |\mathcal{S}|$, a transition probability matrix $\mathbf{P} : \mathcal{S} \times \mathcal{S} \to [0, 1]$ that specifies the transition probabilities between consecutive states X_t and X_{t+1}, i.e., (strong) Markov property $\mathbb{P}[X_{t+1} = s' | X_t = s, \cdots, X_0] = \mathbb{P}[X_{t+1} = s' | X_t = s] = P_{ss'}$, and a reward function $r : \mathcal{S} \to \mathcal{P}([0, r_{\max}])$ where $R_t \sim r(X_t)$, then $(X_t, R_t)_{t \geq 0}$ is called a discrete-time finite *Markov reward process* (MRP) (Puterman 1994). Note that $(X_t)_{t \geq 0}$ is an embedded Markov chain with transition law $\mathbf{P}$. Furthermore, we denote the mean rewards as $\bar{\mathbf{r}} : s \mapsto \mathbb{E}[r(s)]$. As conventions, we denote $\mathbb{E}_s[\cdot] \coloneqq \mathbb{E}[\cdot | X_0 = s]$ and $\mathbb{P}_s[\cdot] \coloneqq \mathbb{P}[\cdot | X_0 = s]$.

The first step when a Markov chain visits a state s is called the *hitting time to* s, i.e., $H_s \coloneqq \inf\{t : X_t = s\}$. Note that if a chain starts at s, then $H_s = 0$. We refer to the first time a chain returns to s as the *first return time to* s

$$H_s^+ \coloneqq \inf\{t > 0 : X_t = s\}. \tag{1}$$

Definition 3.1 (Expected recurrence time)**.** *Given a Markov chain, we define the* expected recurrence time of state s *as the expected first return time of* s *starting in* s

$$\rho_s \coloneqq \mathbb{E}_s \left[H_s^+ \right]. \tag{2}$$

A state s is *positive recurrent* if its expected recurrence time is finite, i.e., $\rho_s < \infty$.

Definition 3.2 (Maximal expected hitting time)**.** *Given a Markov chain, we define the* maximal expected hitting time of state s *as the maximal expected first return time over starting states*

$$\tau_s \coloneqq \max_{s' \in \mathcal{S}} \mathbb{E}_{s'}[H_s^+]. \tag{3}$$

3.2 Discounted Total Rewards

In RL, we are generally interested in some expected long-term rewards that will be collected by following a policy. In the infinite-horizon discounted total reward setting, following a Markov policy on an MDP induces an MRP and the *state value* of state s is

$$v(s) \coloneqq \mathbb{E}_s \left[\sum_{t=0}^{\infty} \gamma^t R_t \right], \tag{4}$$

where $\gamma \in [0, 1)$ is the *discount factor*. Note that since the reward is bounded by $r_{\max}$, state values are also bounded by $r_{\max}/1-\gamma$. A fundamental result relating values to the MRP parameters $(\mathbf{P}, \bar{\mathbf{r}})$ is the Bellman equation for each state $s \in \mathcal{S}$ (Sutton and Barto 2018)

$$v(s) = \bar{r}_s + \gamma \sum_{s' \in \mathcal{S}} P_{ss'} v(s'). \tag{5}$$

3.3 Problem Statement

Suppose that we have a sample path $(X_t, R_t)_{0 \leq t < T}$ of length T from an MRP whose parameters $(\mathbf{P}, \bar{\mathbf{r}})$ are unknown. Given a state s and discount factor γ, we want to estimate $v(s)$.

Assumption 3.1 (State s is reachable)**.** *We assume state* s *is reachable from all states, i.e.,* $\tau_s < \infty$.

Otherwise there is some non-negligible probability that state s will not be visited from some starting state. This will prevent the convergence in probability (in the form of a PAC-style error bound) that we seek (see Appendix B.1 of (Dai and Walter 2020)).

Remark 3.1. *Assumption 3.1 can be weakened to the assumption that* s *is positive recurrent and the MRP starts in the recurrent class containing* s*. All following results can be recovered by restricting* $\mathcal{S}$ *in the definition of* τ_s *to the recurrent class containing* s*. However, for ease of presentation, we will adopt Assumption 3.1 in the rest of the article without loss of generality.*

Note that Assumption 3.1 implies the positive recurrence of s, i.e., $\rho_s < \infty$, by definition, and that the MRP visits state s for infinitely many times with probability 1.

3.4 Renewal Theory and Loops

Stochastic processes in general can exhibit complex dependencies between random variables at different steps, and thus often fall outside of the applicability of approaches that rely on independence assumptions. Renewal theory (Ross 1996) focuses on a class of stochastic processes where the process restarts after a renewal event. Such regenerative structure allows us to apply results from the independent and identical distribution (IID) settings.

In particular, we consider the visits to state s as renewal events and define *waiting times* $W_n(s)$ for $n = 1, 2, \cdots$, to be the number of steps before the n-th visit

$$W_n(s) \coloneqq \inf \left\{ w : n \leq \sum_{t=0}^{w} \mathbb{1}[X_t = s] \right\}, \tag{6}$$

and the *interarrival times* $I_n(s)$ to be the steps between the n-th and $(n + 1)$-th visit

$$I_n(s) \coloneqq W_{n+1}(s) - W_n(s). \tag{7}$$

Remark 3.2. *The random times relate to each other in a few intuitive relations. The waiting time of the first visit is the same as the hitting time* $W_1(s) = H_s \leq H_s^+$*. Waiting times relate to interarrival times* $W_{n+1}(s) = W_1(s) + \sum_{i=1}^{n} I_i(s)$.

To justify treating visits to s as renewal events, consider the sub-processes starting at $W_1(s)$ and at $W_2(s)$—both MRPs start in state s—due to Markov property of MRP, they are statistical replica of each other. Since segments $(X_t, R_t)_{W_n(s) \leq t < W_{n+1}(s)}$ start and end in the same state, we call them *loops*. It follows that loops are independent of each other and obey the same statistical law. Intuitively speaking, an MRP is (probabilistically) specified by its starting state.

Definition 3.3 (Loop γ-discounted rewards)**.** *Given a Markov reward process and a positive recurrent state* s*, we*

define the n-th loop γ-discounted rewards *as the discounted total rewards over the n-th loop*

$$G_n(s) \coloneqq \sum_{u=0}^{I_n(s)-1} \gamma^u R_{W_n(s)+u}. \tag{8}$$

Definition 3.4 (Loop γ-discount)**.** *Given a Markov reward process and a positive recurrent state s, we define the n-th* loop γ-discount *as the total discounting over the n-th loop*

$$\Gamma_n(s) \coloneqq \gamma^{I_n(s)}. \tag{9}$$

$(I_n(s), G_n(s))_{n>0}$ forms a regenerative process that has nice independence relations. Specifically, $I_n(s) \perp\!\!\!\perp I_m(s)$, $G_n(s) \perp\!\!\!\perp G_m(s)$, and $G_n(s) \perp\!\!\!\perp I_m(s)$ when $n \neq m$. Furthermore, $(I_n(s))_{n>0}$ are identically distributed the same as H_s^+ when starting in s. Similarly, $(G_n(s))_{n>0}$ are identically distributed. Note however that $G_n(s) \not\perp\!\!\!\perp I_n(s)$.

4 Main Results

4.1 Bellman Equations over Loops

Given the regenerative process $(I_n(s), G_n(s))_{n>0}$, we derive a new Bellman equation over the loops for state value $v(s)$.

Theorem 4.1 (Loop Bellman equations)**.** *Suppose the expected loop γ-discount is $\alpha(s) \coloneqq \mathbb{E}_s[\Gamma_1(s)]$ and the expected loop γ-discounted rewards is $\beta(s) \coloneqq \mathbb{E}_s[G_1(s)]$, we can relate the state value $v(s)$ to itself*

$$v(s) = \beta(s) + \alpha(s)\, v(s). \tag{10}$$

Remark 4.1. *The key difference between the loop Bellman equations* (10) *and the classic Bellman equations* (5) *is the state values involved. Only state value $v(s)$ appears on the right-hand side of* (10)*.*

4.2 Loop Estimator

We plug in the empirical means for the expected loop γ-discount $\alpha(s)$ and the expected loop γ-discounted rewards $\beta(s)$ into the loop Bellman equation (10) and define the n-th *loop estimator* for state value $v(s)$

$$\hat{v}_n(s) \coloneqq \hat{\beta}_n(s)/(1 - \hat{\alpha}_n(s)), \tag{11}$$

where $\hat{\alpha}_n(s) \coloneqq \frac{1}{n}\sum_{i=1}^n \gamma^{I_i(s)}$ and $\hat{\beta}_n(s) \coloneqq \frac{1}{n}\sum_{i=1}^n G_i(s)$. Furthermore, we have visited state s for $(N+1)$ times before step T where N is a random variable that counts the number of loops before step T

$$N \coloneqq \sup\{n : W_{n+1}(s) \leq T\}, \tag{12}$$

and the estimate $\hat{v}_N(s)$ would be the last estimate before step T. Hence, with a slight abuse of notations, we define

$$\hat{v}_T(s) \coloneqq \hat{v}_N(s). \tag{13}$$

By using incremental updates to keep track of empirical means, Algorithm 1 implements the loop estimator $\hat{v}_T(s)$ with a space complexity of $O(1)$. Running S-many copies of loop estimators, one for each state $s \in \mathcal{S}$, takes a space complexity of $O(S)$.

Algorithm 1 Loop estimator (for a specific state)

Input: discount factor γ, state s, sample path $(X_t, R_t)_{0\leq t<T}$ of some length T.
Return: an estimate of the discounted value $v(s)$.
Initialize the empirical mean of loop discounts $\hat{\alpha} \leftarrow 0$.
Initialize the empirical mean of loop discounted rewards $\hat{\beta} \leftarrow 0$.
Initialize the loop count $n \leftarrow 0$.
for each loop in $(X_t, R_t)_{0\leq t<T}$ **do**
  Increment visit count $n \leftarrow n + 1$.
  Compute the length of the interarrival time $I_n(s) \leftarrow W_{n+1}(s) - W_n(s)$.
  Compute the partial discounted sum of rewards, $G_n(s) \leftarrow \sum_{u=0}^{I_n(s)-1} \gamma^u R_{W_n(s)+u}$.
  Update the empirical means incrementally, $\hat{\alpha} \leftarrow \frac{1}{n}\gamma^{I_n(s)} + \left(1 - \frac{1}{n}\right)\hat{\alpha}$, and $\hat{\beta} \leftarrow \frac{1}{n}G_n(s) + \left(1 - \frac{1}{n}\right)\hat{\beta}$.
end for
return $\hat{\beta}/(1 - \hat{\alpha})$

4.3 Rates of Convergence

Now we investigate the convergence of the loop estimator, first over visits, i.e., $\hat{v}_n(s) \xrightarrow{p} v(s)$ as $n \to \infty$, then over steps, i.e., $\hat{v}_T(s) \xrightarrow{p} v(s)$ as $T \to \infty$. By applying Hoeffding bound to the definition of loop estimator (11), we obtain a PAC-style upper bound on the estimation error.

Theorem 4.2 (Convergence rate over visits)**.** *Given a sample path from an MRP $(X_t, R_t)_{t\geq 0}$, a discount factor $\gamma \in [0, 1)$, and a positive recurrent state s, with probability of at least $1 - \delta$, the loop estimator converges to $v(s)$*

$$|\hat{v}_n(s) - v(s)| = O\left(\frac{r_{max}}{(1-\gamma)^2}\sqrt{\frac{1}{n}\log\frac{1}{\delta}}\right).$$

To determine the convergence rate over steps, we need to study the concentration of waiting times which allows us to lower-bound the random visits with high probability. As an intermediate step, we use the fact that the tail of the distribution of first return times is upper-bounded by an exponential distribution per the Markov property of MRP (Lee, Ozdaglar, and Shah 2013; Aldous and Fill 1999).

Lemma 4.3 (Exponential concentration of first return times (Lee, Ozdaglar, and Shah 2013; Aldous and Fill 1999))**.** *Given a Markov chain $(X_t)_{t\geq 0}$ defined on a finite state space $\mathcal{S}$, for any state $s \in \mathcal{S}$ and any $t > 0$, we have*

$$\mathbb{P}\left[H_s^+ \geq t\right] \leq e \cdot e^{-t/e\tau_s}.$$

Secondly, since by Remark 3.2 we have $W_{n+1}(s) = W_1(s) + \sum_{i=1}^n I_i(s)$, we apply the union bound to upper-bound the tail of waiting times.

Corollary 4.4 (Upper bound on waiting times)**.** *With probability of at least $1 - \delta$, $W_n(s) = O\left(n\tau_s \log\frac{n}{\delta}\right)$.*

Remark 4.2. *Note that the waiting time $W_n(s)$ is nearly linear in n with a dependency on the Markov chain structure via the maximal expected hitting time of s, namely $\widetilde{O}(n\tau_s)$.*

In contrast, the expected *waiting time scales with the expected recurrence time* $\mathbb{E}[W_n(s)] = \Theta(n\rho_s)$. *However, an exponential concentration with the expected recurrence time is not possible in general (see Appendix B.2 of (Dai and Walter 2020) for a counterexample).*

Using Lambert W function, we invert Corollary 4.4 to lower-bound the visits by step T with high probability. Finally, the convergence rate of $\hat{v}_T(s)$ follows from Theorem 4.2.

Theorem 4.5 (Convergence rate over steps)**.** *With probability of at least* $1-\delta$, *for any* $T > e\,\delta\,\tau_s$, *the MRP* $(X_t, R_t)_{0\le t<T}$ *visits state* s *for at least* $\widetilde{\Omega}(T/\tau_s)$ *many times, and the last loop estimate converges to* $v(s)$

$$|\hat{v}_T(s) - v(s)| = \widetilde{O}\left(\frac{r_{max}}{(1-\gamma)^2}\sqrt{\frac{\tau_s}{T}\log\frac{1}{\delta}}\right).$$

Suppose we run a copy of loop estimator to estimate each state's value in $\mathcal{S}$, and denote them with a vector $\hat{\mathbf{v}}_T : s \mapsto \hat{v}_T(s)$. Convergence of the estimation error $\hat{\mathbf{v}}_T - \mathbf{v}$ in terms of the ℓ_∞-norm follows immediately by applying the union bound.

Corollary 4.6 (Convergence rate over all states)**.** *With probability of at least* $1-\delta$, *for any* $T > e\,\delta\,\max_s \tau_s$, *the MRP* $(X_t, R_t)_{0\le t<T}$ *visits each state* s *for at least* $\widetilde{\Omega}(T/\tau_s)$ *many times, and the last loop estimates converge to state values* $\mathbf{v}$

$$\|\hat{\mathbf{v}}_T - \mathbf{v}\|_\infty = \widetilde{O}\left(\frac{r_{max}}{(1-\gamma)^2}\sqrt{\frac{\max_s \tau_s}{T}\log\frac{S}{\delta}}\right).$$

5 Numerical Experiments

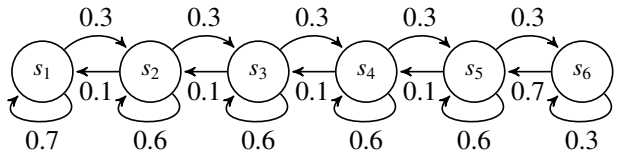
We consider `RiverSwim`, an MDP proposed by Strehl and Littman (2008) that is often used to illustrate the challenge of exploration in RL. The MDP consists of six states $\mathcal{S} = \{s_1, \cdots, s_6\}$ and two actions $\mathcal{A} = \{$"swim downstream", "swim upstream"$\}$. Executing the "swim upstream" action often fails due to the strong current, while there is a high reward for staying in the most upstream state s_6. For our experiments, we use the MRP induced by always taking the "swim upstream" action (see Figure 1a for numerical details).

The most relevant aspect of the induced MRP is that the maximal expected hitting times are very different for different states: $\tau_{s_1} \approx 752$, $\tau_{s_2} \approx 237$, $\tau_{s_3} \approx 68$, $\tau_{s_4} \approx 15$, $\tau_{s_5} \approx 17$, $\tau_{s_6} \approx 22$. Figure 1b shows a plot of the estimation errors of the loop estimator for each state over the *square root* of maximal expected hitting times $\sqrt{\tau_s}$ of that state. The observed linear relationship between the two quantities (supported by a good linear fit) is consistent with the instance-dependence in our result of $|\hat{v}_T(s) - v(s)| = \widetilde{O}\left(\sqrt{\tau_s}\right)$, c.f., Theorem 4.5.

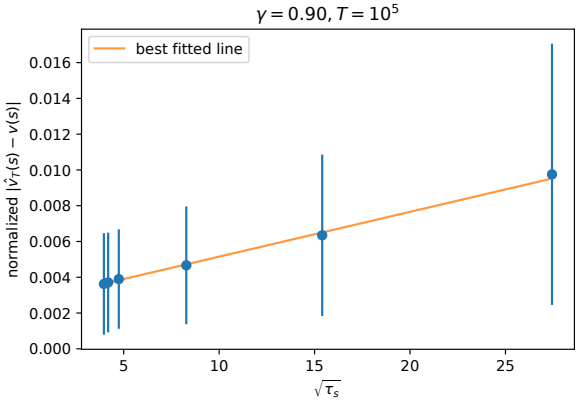
5.1 Alternative Estimators

We define several alternative estimators for $v(s)$ and briefly mention their relevance for comparison.

Model-based. We compute add-1 smoothed maximum likelihood estimates (MLE) of the MRP parameters $(\mathbf{P}, \bar{\mathbf{r}})$



(a)



(b)

Figure 1: (a) The induced `RiverSwim` MRP. The arrows are labeled with transition probabilities. The rewards are all zero except for state s_6, where $r(s_6) = 1$. (b) With discount factor $\gamma = 0.9, T = 10^5$. The estimation error of each state (normalized by $\max_s v(s)$) is plotted over the square root of maximal expected hitting times $\sqrt{\tau_s}$ of that state. Error bars show the standard deviations over 200 runs.

from the sample path

$$\hat{P}_{s\,s'} := \frac{\frac{1}{S} + \sum_{t=0}^{T-1} \mathbb{1}\left[X_{t+1} = s', X_t = s\right]}{1 + \sum_{t=0}^{T-1} \mathbb{1}\left[X_t = s\right]} \tag{14}$$

and

$$\hat{\bar{r}}_s := \frac{\sum_{t=0}^{T-1} R_t \mathbb{1}\left[X_t = s\right]}{1 + \sum_{t=0}^{T-1} \mathbb{1}\left[X_t = s\right]}. \tag{15}$$

We then solve for the discounted state values from the Bellman equation (5) for the MRP parameterized by $\left(\hat{\mathbf{P}}, \hat{\bar{\mathbf{r}}}\right)$, i.e., the (column) vector of estimated state values

$$\hat{\mathbf{v}}_{\text{MB}} := \left(\mathbf{I} - \gamma\hat{\mathbf{P}}\right)^{-1}\hat{\bar{\mathbf{r}}} \tag{16}$$

where $\mathbf{I}$ is the identity matrix.

TD(k). k-step temporal difference (or k-step backup) estimators are commonly recursively defined (Kearns and Singh 2000) with TD(0) being a textbook classic (Bertsekas and Tsitsiklis 1996; Sutton and Barto 2018). Let $\hat{v}_{\text{TD}}(0, s) := 0$ for all states $s \in \mathcal{S}$. And for $t > 0$

$$\hat{v}_{\text{TD}}(t, s) := \begin{cases} (1-\eta_t)\,\hat{v}_{\text{TD}}(t-1, s) \\ \quad + \eta_t\big(\gamma^0 R_t + \cdots + \gamma^k R_{t+k} \\ \quad + \gamma^{k+1}\hat{v}_{\text{TD}}(t-1, X_{t+k+1})\big), & \text{if } s = X_t \\ \hat{v}_{\text{TD}}(t-1, s), & \text{otherwise} \end{cases}$$

where η_t is the learning rates. A common choice is to set $\eta_t = 1/\left(\sum_{u=0}^{t} \mathbb{1}[X_u = s]\right)$ which satisfies the Robbins-Monro conditions (Bertsekas and Tsitsiklis 1996). But it has been shown to lead to slower convergence than $\eta_t = 1/\left(\sum_{u=0}^{t} \mathbb{1}[X_u = s]\right)^d$ where $d \in (1/2, 1)$ (Even-Dar and Mansour 2003).

It is more accurate to consider TD methods as a large family of estimators each with different choices of k, η_t. Choosing these parameters can create extra work and sometimes confusion for practitioners. Whereas the loop estimator, like the model-based estimator, has no parameters to tune. In any case, it is not our intention to compare with the TD family exhaustively (see more results on TD on (Kearns and Singh 2000; Even-Dar and Mansour 2003)). Instead, we will compare with TD(0) and TD(10), both with $d = 1$, and TD(0)* with $d = 1/2$.

5.2 Comparative Experiments

We experiment with different values for the discount factor γ, because, roughly speaking, $1/(1-\gamma)$ sets the horizon beyond which rewards are discounted too heavily to matter. We compare the estimation errors measured in ∞-norm, which is important in RL. The results are shown in Figure 2.

- The model-based estimator dominates all estimators for every discount setting we tested.
- TD(k) estimators perform well if $k \geq 1/(1-\gamma)$.
- The loop estimator performs worse than, but is competitive with, the model-based estimator. Furthermore, similar to the model-based estimator and unlike the TD(k) estimators, its performance seems to be less influenced by discounting.

6 Discussions

The elementary identity below relates the expected first return times $Y_{s\,s'} := \mathbb{E}_s\left[H^+_{s'}\right]$ to the transition probabilities $P_{s\,s'}$ for a finite Markov chain. Using the matrix notations, suppose that the expected first return times are organized in a matrix $\mathbf{Y}$, and $\mathbf{P}$ the transition matrix of the Markov chain, then we have $\mathbf{Y} = \mathbf{P}\,(\mathbf{Y} - \mathrm{diag}\mathbf{Y} + \mathbf{E})$ where $\mathrm{diag}\mathbf{Y}$ is a matrix with the same diagonal as $\mathbf{Y}$ and zero elsewhere, and $\mathbf{E}$ is a matrix with all ones. Thus, knowing $\mathbf{Y}$ is equivalent to knowing the full model, as we can compute $\mathbf{P}$ using this identity. Recall that by definition $\mathbb{E}\,[I_1(s)] = \mathbb{E}_s\,[H^+_s]$, which is exactly the diagonal of $\mathbf{Y}$. But only knowing the diagonal is not sufficient to determine the entire set of model parameters, namely $\mathbf{Y}$, the loop estimator based on $(I_n)_{n>0}$ indeed falls short of being a model-based method. It may be considered a *semi*-model-based method as it estimates some but not all of the model parameters.

For large MRPs, a natural extension of our work is to consider recurrence of features instead of states, e.g., a video game screen might not repeat itself completely but the same items might reappear. After all, without repetition exactly or approximately, it would not be possible for an agent to learn and improve its decisions.

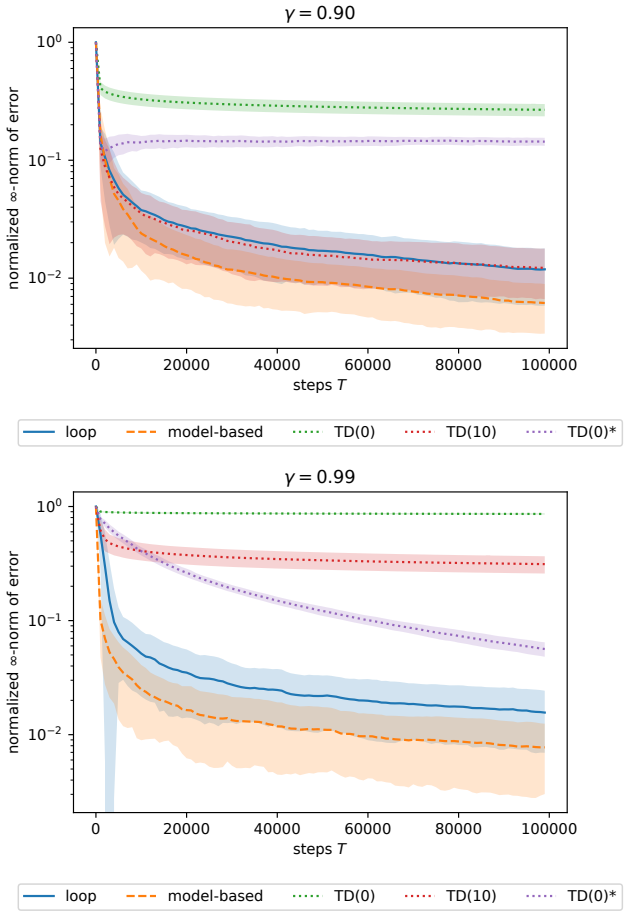


Figure 2: Estimation errors (normalized by $\max_s v(s)$ to be comparable across discount factors) of different estimators at different discount factors (left) $\gamma = 0.9$ and (right) $\gamma = 0.99$. Shaded areas represent the standard deviations over 200 runs. Note the vertical log scale.

We believe that regenerative structure can be further exploited in RL (particularly in the form of the loop Bellman equation (10)) and we think this article provides the fundamental results for future study in this direction.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. 1830660. We thank Mesrob I. Ohannessian for a helpful discussion on Markov chains, and Christina Lee Yu for discussing an early version of this work. We also thank anonymous reviewers for their constructive feedback, in particular, for bringing an independent work (Subramanian and Mahajan 2019) to our attention.

References

Aldous, D.; and Fill, J. 1999. Reversible Markov chains and random walks on graphs. Book in preparation (available at http://www.stat.berkeley.edu/~aldous/RWG/Chap2.pdf).

Bertsekas, D. P.; and Tsitsiklis, J. N. 1996. *Neuro-dynamic programming*. Athena Scientific Belmont, MA.

Dai, F. Z.; and Walter, M. R. 2020. Loop Estimator for Discounted Values in Markov Reward Processes. *arXiv preprint* URL https://arxiv.org/abs/2002.06299.

Dayan, P.; and Sejnowski, T. J. 1994. TD (λ) converges with probability 1. *Machine Learning* 14(3): 295–301.

Derman, C. 1970. *Finite state Markovian decision processes*. Academic Press.

Even-Dar, E.; and Mansour, Y. 2003. Learning rates for Q-learning. *Journal of machine learning Research* 5(Dec): 1–25.

Fox, B. L.; and Glynn, P. W. 1989. Simulating discounted costs. *Management Science* 35(11): 1297–1315.

Gheshlaghi Azar, M.; Munos, R.; and Kappen, H. J. 2013. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning* 91(3): 325–349. ISSN 1573-0565. doi:10.1007/s10994-013-5368-1.

Haviv, M.; and Puterman, M. L. 1992. Estimating the value of a discounted reward process. *Operations Research Letters* 11(5): 267–272. ISSN 01676377. doi:10.1016/0167-6377(92)90002-K.

Howard, R. A. 1960. *Dynamic programming and Markov processes.* John Wiley.

Jaakkola, T.; Jordan, M. I.; and Singh, S. P. 1994. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, 703–710.

Kearns, M. J.; and Singh, S. P. 1999. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems*, 996–1002.

Kearns, M. J.; and Singh, S. P. 2000. Bias-Variance Error Bounds for Temporal Difference Updates. In *Conference on Learning Theory*, 142–147.

Lee, C. E.; Ozdaglar, A.; and Shah, D. 2013. Approximating the Stationary Probability of a Single State in a Markov chain. *arXiv preprint* URL https://arxiv.org/abs/1312.1986.

Pananjady, A.; and Wainwright, M. J. 2019. Value function estimation in Markov reward processes: Instance-dependent ℓ_∞-bounds for policy evaluation. *arXiv preprint* URL https://arxiv.org/abs/1909.08749.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Ross, S. M. 1996. *Stochastic processes*. John Wiley, 2nd edition.

Strehl, A. L.; and Littman, M. L. 2008. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences* 74(8): 1309–1331.

Subramanian, J.; and Mahajan, A. 2019. Renewal Monte Carlo: Renewal theory based reinforcement learning. *IEEE Transactions on Automatic Control* 1–1.

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

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press, 2nd edition.