JMLR: Workshop and Conference Proceedings 24:89–101, 2012 10th European Workshop on Reinforcement Learning

An Empirical Analysis of Off-policy Learning in Discrete MDPs

Cosmin Păduraru COSMIN@CS.MCGILL.CA
Doina Precup DPRECUP@CS.MCGILL.CA
Joelle Pineau JPINEAU@CS.MCGILL.CA
Gheorghe Comănici GCOMAN@CS.MCGILL.CA
McGill University, School of Computer Science, Montreal, Canada

Editor: Marc Peter Deisenroth, Csaba Szepesvári, Jan Peters

Abstract

Off-policy evaluation is the problem of evaluating a decision-making policy using data collected under a different behaviour policy. While several methods are available for addressing off-policy evaluation, little work has been done on identifying the best methods. In this paper, we conduct an in-depth comparative study of several off-policy evaluation methods in non-bandit, finite-horizon MDPs, using randomly generated MDPs, as well as a Mallard population dynamics model [Anderson, 1975] . We find that un-normalized importance sampling can exhibit prohibitively large variance in problems involving look-ahead longer than a few time steps, and that dynamic programming methods perform better than Monte-Carlo style methods.

1. Introduction

One of the core competencies of most intelligent decision-making agents is the ability to properly evaluate their decision-making strategy. In a reinforcement learning context, this is the policy evaluation problem, which involves the estimation of the expected return associated with a policy. The ideal method for evaluating policies is to apply them in practice, observe the return, and estimate the expected return (and its uncertainty) using this data. However, if data is expensive (or rare), or the number of policies to evaluate is large, this may be infeasible. A popular alternative is to evaluate the decision-making policy of interest (called the *target policy*) using data collected under a different, *behaviour policy*. This method is known as *off-policy policy evaluation*. Off-policy learning has been used in a range of applications, such as energy systems [Hannah and Dunson, 2011], robotics [Riedmiller, 2005], clinical studies [Pineau et al., 2009], and tax collection [Abe et al., 2010].

We focus on two types of off-policy estimators: model-based [Sutton and Barto, 1998; Mannor et al., 2007], and importance sampling weighting of the returns [Precup et al., 2000; Robins et al., 2000; Murphy, 2005]. Several methods for continuous MDPs that would fall in neither category, such as LSTD, can be shown to reduce to a model-based estimator in the discrete setting [Boyan, 2002]. The tree backup algorithm proposed in Precup et al. [2000] also becomes, in expectation, equivalent to a model-based estimator. One class of algorithms which cannot be easily pegged into one of these categories is gradient-based temporal-difference methods [Sutton et al., 2009]. However, since these algorithms are de-

signed specifically to handle infinite-horizon problems in which function approximation is necessary, we will not discuss them in this paper. Three of the estimators we compare have been proposed in the literature. The other two (per-step importance sampling and normalized per-step importance sampling) are related to existing methods, but have not been studied in the form we propose.

Previous comparative studies for off-policy estimators considered single-step contextual bandit problems [Kang et al., 2007; Dudik et al., 2011]. We present an empirical study for finite-horizon discrete MDPs with arbitrary horizon length. We use a set of randomly generated MDPs, similar to those used by Castronovo and Ernst [2012], and a simulated model of the Mallard population dynamics, first proposed by Anderson [1975] and subsequently used by Fonnesbeck [2005].

2. Finite-horizon MDPs

A finite-horizon MDP is defined as a tuple $\langle S, A, P, R\rangle$, where S is a set of states; A is a set of actions; $P : S \times A \times S \to [0, 1]$ is the transition model, with $P_{sa}^{s'}$ denoting the conditional probability of a transition to state s' given current state s and action a; $R : S \times A \to [0, 1]$ is the reward function, with R_{sa} denoting the immediate expected reward for state s and action a. A policy $\pi : S \times A \to [0, 1]$ specifies how decisions are made. In a finite MDP, the model can be represented using matrices $P \in \mathbb{R}^{|S\times A|\times|S|}$ and $R \in \mathbb{R}^{|S\times A|}$. Similarly, policies can also be represented as block-diagonal matrices $\pi \in \mathbb{R}^{|S|\times|S\times A|}$.

The value of a policy π for a decision horizon of length K is defined as:

$$V_K^\pi(s) = E[r_0 + r_1 + \cdots + r_K | s_0 = s] = \sum_a \pi_{sa} \left(R_{sa} + \sum_{s'} P_{sa}^{s'} V_{K-1}^\pi(s') \right).$$

or, if we consider V_K^π to be the vector of all state values,

$$V_K^\pi = \pi(R + PV_{K-1}^\pi) = \cdots = \sum_{k=0}^{K} (\pi P)^k \pi R. \tag{1}$$

3. Off-policy value function estimation

In practice, the model of the MDP is usually unknown, so Eq. (1) cannot be applied directly. Furthermore, the user might not be able to obtain trajectories using policy π. In natural resource management, in particular, implementing a policy is not only expensive but also potentially unrealistic, given that interesting time horizons may span decades. Instead, the user may have access to data gathered under some existing policy b (or possibly, under several known policies from different geographic regions or different periods). Off-policy value function estimation methods are designed to deal with this case.

All estimators used in this paper are summarized in Table 1. We will now describe the notation and context for each of them.

Importance sampling [Rubinstein, 1981] is a technique for sampling from one distribution by weighting the samples generated from another distribution. It has been proposed as an off-policy estimator for MDPs both in reinforcement learning [Precup et al., 2000] and in the

	Un-normalized	Normalized
Per-trajectory importance sampling	$\hat{V}_K^{PTIS}(s) = \frac{1}{n_s}\sum_{i=1}^{n_s}\left(\prod_{j=1}^{K}\frac{\pi(s_{i:j},a_{i:j})}{b(s_{i:j},a_{i:j})}\right)\sum_{l=1}^{K} r_{i:l}$	Instead of n_s, divide by $\sum_{i=1}^{n_s}\left(\prod_{j=1}^{K}\frac{\pi(s_{i:j},a_{i:j})}{b(s_{i:j},a_{i:j})}\right)$
Per-step importance sampling	$\hat{V}_k^{PSIS}(s) = \frac{1}{n_s}\sum_a \frac{\pi_{sa}}{b_{sa}}\sum_{i\in B(s,a)}[r_i+\hat{V}_{k-1}^{PSIS}(s'_i)]$	Instead of n_s, divide by $\sum_{i\in B(s)}\frac{\pi(s,a_i)}{b(s,a_i)} = \sum_a n_{sa}\frac{\pi_{sa}}{b_{sa}}$
Model-based	$\hat{V}_K^{MB}(s) = \sum_a \pi_{sa}\left(\hat{R}_{sa} + \sum_{s'}\hat{P}_{sa}^{s'}\hat{V}_{K-1}^{MB}(s')\right) = \sum_{k=0}^{K}(\pi\hat{P})^k\pi\hat{R}$	

Table 1: Off-policy estimators for discrete MDPs. **All methods are consistent, but (un-normalized) per-trajectory importance sampling is the only one that is unbiased.**

clinical trial literature [Robins et al., 2000], where it is called inverse probability weighting. Importance sampling methods typically assume that the behaviour policy used to collect the data, denoted b, is known, and that $\pi_{sa} > 0 \implies b_{sa} > 0$. The naive implementation of importance sampling for off-policy evaluation weighs entire trajectories. This existing estimator is the (un-normalized) *per-trajectory importance sampling* (PTIS) in Table 1. It is computed based on n_s trajectory fragments of length K that start from state s in the batch, where the i^{th} trajectory is denoted by:

$$(s_{i:0} = s, a_{i:0}, r_{i:0}, s_{i:1}, a_{i:1}, r_{i:1}, ..., s_{i:K}, a_{i:K}, r_{i:K}),$$

The weights in the importance sampling estimator can be scaled to the $[0, 1]$ interval by normalizing over their sum. This is seen as a way to reduce estimator variance, and leads to the *normalized per-trajectory importance sampling estimator* in Table 1, which we will denote by $\hat{V}^{PTIS-N}$. This is also an existing estimator [Precup et al., 2000; Murphy, 2005].

In order to avoid the variance introduced by weighting entire trajectories, we introduce *per-step importance sampling* as an alternative. These estimators are very similar to the per-decision importance sampling [Precup et al., 2000]. However, we present them for state values and finite-horizon problems with batch data, rather than for action values and episodic problems with on-line data. Consider the more general setting where a sample i is composed of start state s_i, action a_i generated from b at s_i, and (r_i, s'_i) the response of the model (P, R) at (s_i, a_i). The (un-normalized) *per-step importance sampling estimator*, shown in Table 1, uses n_s, n_{sa}, and $n_{sas'}$ to denote the sizes of the subsets restricted by the start state s, action a and/or next state s', and $B(s)$ and $B(s, a)$ to denote the subsets of samples for which the

start state and/or action choice is s, a. If $n_s = 0$, there is no data at this state, so we have to pre-define $\hat{V}_k^{PSIS}(s)$. We also construct a normalized version, which we denote by $\hat{V}_K^{PSIS-N}$.

Model-based MDP estimators construct approximations $\hat{P}$ and $\hat{R}$ of the transition and reward model, and then use standard methods such as dynamic programming to compute the value function for the estimated model. For discrete MDPs, consistent estimators of the model are given by:

$$\hat{R}_{sa} = \frac{1}{n_{sa}} \sum_{i \in B(s,a)} r_i, \qquad \hat{P}_{sa}^{s'} = \frac{n_{sas'}}{n_{sa}}. \tag{2}$$

Similarly to per-step importance sampling, we have to use a pre-defined value if $n_s = 0$ or $n_{sa} = 0$. The finite-horizon value function can then be estimated using the approximate model $\hat{R}, \hat{P}$.

This estimator is intuitive and has a long history. However, we are only aware of one work on its statistical properties, by Mannor et al. [2007]. For infinite-horizon discrete MDPs with discounting, Mannor et al. [2007] compute second-order approximations for the bias and variance of the model-based estimator, and examine its empirical performance on a discretized version of a catalog ordering problem.

Note that the MB estimator can be expressed in the same form as the per-step importance sampling estimators, but with an estimate of b as surrogate:

$$\hat{V}_k^{MB}(s) = \frac{1}{n_s} \sum_a \frac{\pi_{sa}}{n_{sa}/n_s} \sum_{i \in B(s,a)} \left(r_i + \hat{V}_{k-1}^{MB}(s_i') \right). \tag{3}$$

Results from [Rubinstein, 1981] can be used to show that both PTIS and PTIS-N are consistent estimators. PTIS is also unbiased; however, normalization can introduce bias, while typically reducing variance. The dynamic programming estimators (PSIS, PSIS-N and MB) are also consistent. This can be proven using Slutsky's theorem [Dudewicz and Mishra, 1988], by noting that all of them can be written in the form

$$\hat{V}_K = \sum_{k=0}^{K} (\pi Z \hat{P})^k \pi Z \hat{R}, \tag{4}$$

where Z is a diagonal matrix with entries

$$Z_{sa}^{PSIS} = (n_{sa}/n_s)/b_{sa} \qquad Z_{sa}^{PSIS-N} = (n_{sa}/W(s))/b_{sa} \qquad Z_{sa}^{MB} = 1,$$

with $W(s)$ denoting the normalization term for PSIS-N.

4. Empirical results

In this section we study the empirical performance of the different off-policy estimators on two domains: a simulated model of a natural resource management problem, and a set of randomly generated MDPs. Note that, for the model-based method, we use a default reward value $R_{sa} = 0$ for the state-action pairs (s, a) for which $n_{sa} = 0$ (unless otherwise specified), and a default transition model that self-loops ($P_{sa}^s = 1$).

4.1. Mallard population model

Anderson's model is formulated as an MDP with yearly time increments, two-dimensional state, and continuous actions [Anderson, 1975]. The state variables are the adult population N_t and the number of ponds P_t (both expressed in millions), while the action H_t represents the proportion of animals to be harvested in year t. The state transitions are defined by the following equations:

$$N_{t+1} = N_t(1 - 0.37e^{2.78H_t}) + \left(\frac{1}{12.48}P_t^{0.851} + \frac{0.519}{N_t}\right)^{-1}(1 - 0.49e^{0.9H_t})$$
$$P_{t+1} = -2.76 + 0.391P_t + 0.233\epsilon_t$$

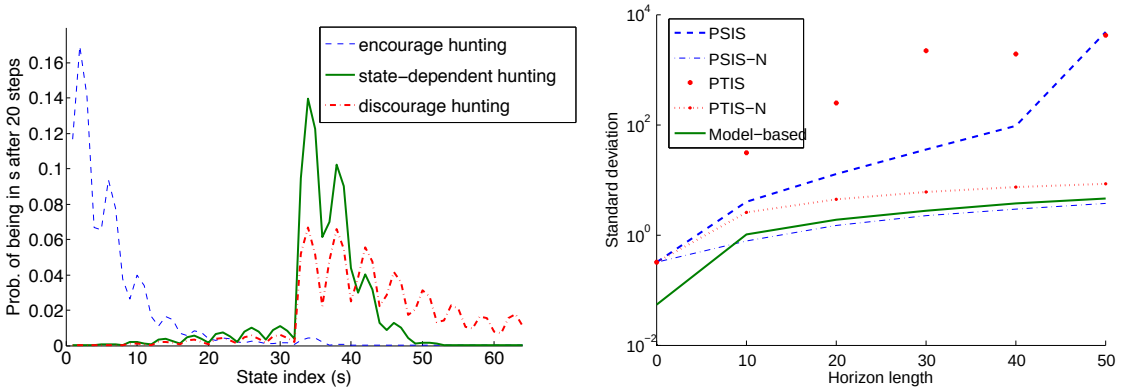
where $\epsilon_t \sim N(16.46, 4.41)$ is a normally distributed random variable describing the amount of precipitation during year t (in inches). The reward is defined as the number of birds harvested in a given year, computed as

$$R(N_t, P_t, H_t) = H_t\left(0.92N_t + \left(\frac{1}{12.48}P_t^{0.851} + \frac{0.519}{N_t}\right)^{-1}\right).$$

Anderson constructed and validated this model based on real data about the evolution of the Mallard population. For more details, including model justification, we refer the reader to [Anderson, 1975].

For our experiments, we used a discretized version of the model. Since the states where the bird population is close to 0 are particularly important, we used a discretization with higher resolution in that region of the state space. More precisely, we divided N_t into intervals of length 2 when $N_t > 2$, and length 0.25 when $N_t \leq 2$. P_t was divided into four intervals of unit length. We also assumed that state features are bounded, so $N_t \in [0, 17]$ and $P_t \in [0, 4]$. This resulted in 64 states. We generated 10 million transitions from the original MDP by sampling starting states uniformly randomly; then, we used the data to estimate a transition matrix and reward function for the discrete MDP. The transition function was estimated using maximum likelihood estimation, whereas the reward function was defined as a Gaussian for each discretized interval, with its mean and variance estimated from the generated data. This produced the MDP that we used as "ground truth"; that is, we investigated how well our methods estimate the value function for this discretized MDP.

We considered three policies, all selecting from two actions: a_1 representing $H_t = 0$ and a_2 representing $H_t = 0.3$. The first policy, which we call *discourage hunting*, selects a_1 with probability 0.8 and a_2 with probability 0.2 in every state. The second policy, which we call *state-dependent hunting*, prescribes reduced hunting when the mallard population or the number of ponds is low, and larger amounts of hunting otherwise; more precisely, it selects a_1 with probability 0.8 in the discrete states corresponding to $[0, 12] \times [0, 1]$, $[0, 8] \times [1, 2]$, $[0, 4] \times [2, 3]$, and $[0, 2] \times [3, 4]$, and a_2 with probability 0.8 for the rest of the state space. The third policy, called *encourage hunting*, selects a_2 with probability 0.8 in all states. Throughout the experiments, we used *discourage hunting* as the behaviour policy, and used the discrete state corresponding to $N_t = 7$ and $P_t = 1.5$ as the starting state s_0. Each batch of training data was generated as a single, uninterrupted trajectory starting in s_0, with actions selected according to the behaviour policy. Hence, the number of samples in a



(*a*) Probabilities of being in different states after 20 steps, under the three different policies

(*b*) Standard deviations of the three estimators based on a 500-sample trajectory with *state-dependent hunting* as the target.

Figure 4.1: State probabilities (left) and standard deviations of three estimators (right). Results averaged over 100, 000 runs. Note the logarithmic scale for the y axis in the right panel.

batch is the length of this trajectory. We present results estimating the value of the start state s_0 under all policies. Unless otherwise specified, the results are averages over 1000 batches.

As seen in Figure 1(a), the three policies tend to visit different regions of the state space. In particular, the distribution of states under *encourage hunting* leads predominantly to states corresponding to low population numbers, which is very different from the other two policies. Intuitively, this discrepancy should make estimating the value function for *encourage hunting* particularly challenging, given that *discourage hunting* is used as the behaviour policy. This intuition is confirmed by our empirical results.

Figure 4.1(*b*) contains a plot of the standard deviations of the different estimators when *state-dependent hunting* is the target policy. Even for a target policy that induces a state distribution fairly close to the one under the behaviour policy, PSIS and PTIS can have very large variance for horizons longer than a few time steps.

For the remainder of this section, we further investigate the performance of PSIS-N, PTIS-N, and MB. We examine the performance of these three estimators in terms of bias and root mean squared error (RMSE), when either *state-dependent hunting* or *encourage hunting* is the target policy.

For a particular horizon, we can examine the methods' performance as a function of the amount of data available, as illustrated in Figures 4.2 and 4.3. As expected, the performance of both methods improves when increasing the size of the batch, although much slower if the target policy is very different from the behaviour policy.

PTIS-N exhibits the poorest performance in all settings. The performance difference is most striking when *encourage hunting* is the target policy. *Encourage hunting* has the reversed action selection probabilities from the behaviour policy, and it induces a completely

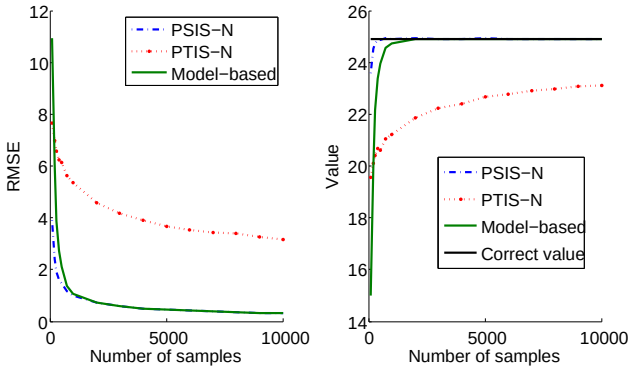
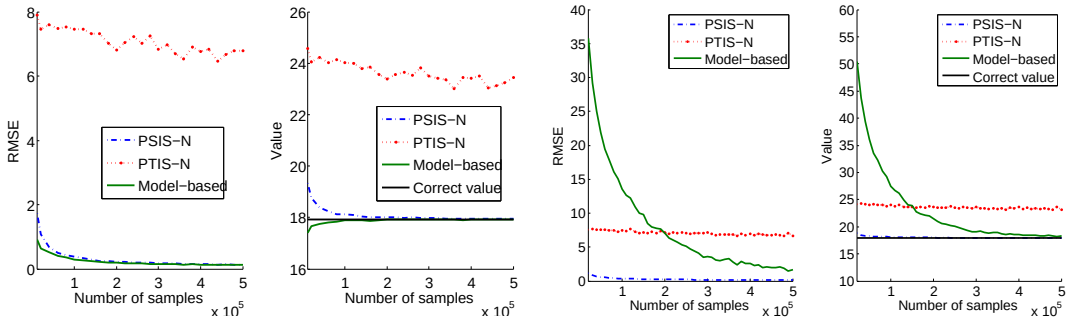


Figure 4.2: RMSE and bias of the 20-step value function estimate for various sample sizes, using *state-dependent hunting* as the target policy.



(*a*) Default MB reward = 0. (*b*) Default MB reward = 10.

Figure 4.3: RMSE and bias of the 20-step value function estimate for various sample sizes, using *encourage hunting* as the target policy. The only difference between the two graphs is the default value used by the model-based method for R_{sa} when $n_{sa} = 0$. Note that the x axis is different from Figure 4.2, allowing for batch sizes of up to 500,000 samples.

different state distribution (as seen in Figure 4.1). This suggests that PTIS-N's performance is particularly weak when the problem is highly off-policy.

The poor performance of per-trajectory methods is particularly interesting, given that they are commonly used as a method for evaluating treatment effects from clinical trials [Robins et al., 2000; Murphy, 2005]. We conjecture that this happens because of at least two reasons. First, the existing evaluations of multi-stage treatments are based on clinical trials that typically have very short horizon lengths (two and three stage trials are common), and for such short horizons the difference between the methods' performance is not as large. Second, epidemiologists tend to include information about previous treatment in the state space, making the set of states accessible in k steps different for all k. In such a setup,

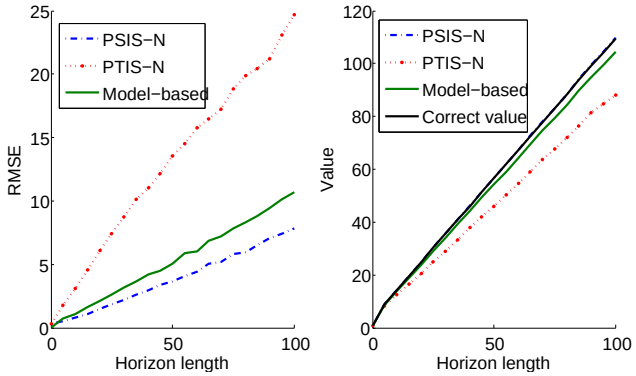


Figure 4.4: RMSE and bias for various horizons and batches of 500 samples each, using *state-dependent hunting* as the target policy. For the bias graph, the line for PSIS-N overlaps that for the correct value.

methods that take advantage of the Markov property (such as PSIS-N and MB) may offer fewer benefits.

PSIS-N performed strongly in all settings. The only time that the model-based method outperformed PSIS-N was when *encourage hunting* was the target policy, and the default value used by the model-based method for the state-action pairs (s, a) for which $n_{sa} = 0$ was $R_{sa} = 0$ (Figure 4.3(a)). The good performance of the model-based method in this setting is likely to be due to the fact that the reward for most states encountered under *encourage hunting* is actually very close to zero, because *encourage hunting* depletes the animal population. When a different default value was used ($R_{sa} = 10$, Figure 4.3(b)), the model-based method performed noticeably worse, due to the increased bias induced by using a default reward that was further from the true value.

Figure 4.4 illustrates how the length of the horizon affects performance. The performance of all methods degrades as the horizon increases. This is expected, as increasing the horizon while maintaining the same number of samples means that we effectively have fewer samples per time step, which increases the variance. However, the ranking of the methods' performance remains the same regardless of the horizon. We have observed a similar phenomenon when using *encourage hunting* as a target policy.

4.2. Randomly generated MDPs

In this section, we present experiments on a set of randomly generated MDPs. The experiments in the previous section indicated that the model-based method may have reduced performance due to high bias when there are zero samples for some of the state-action pairs. Since increasing the total number of available actions increases the probability of having no samples for a state-action pair, we use the random MDPs to illustrate how the bias of the model-based method is affected by the number of available actions.

The randomly generated MDPs are similar to those used by Castronovo and Ernst [2012]. Each of the MDPs has 20 states and 5 actions, except for one experiment where we varied the number of actions. The transition function is generated by randomly selecting, for each

state-action pair (s, a), 10% of the states as successor states, generating a uniform random variable in $N(s') \in [0, 1]$ for each of the successor states s', and then normalizing to obtain the transition probabilities: $P_{sa}^{s'} = \frac{N(s')}{\sum_{s'' \in Succ(s,a)} N(s'')}$.

For states one through 10, the reward for state-action pair (s, a) is equal to zero with 0.9 probability and to a number chosen uniformly randomly in $[0, 1]$ otherwise. For states 11 through 20, the probabilities are reversed (zero with probability 0.1 and a uniform number with probability 0.9). This is slightly different from Castronovo and Ernst [2012] - they used the first reward model (the one we used for states 1-10) as a prior for generating deterministic rewards at all the states in the MDP. The starting state is state 1.

We used a uniform behaviour policy that selected each action with equal probability at all states. The target policy was one that ascribed 60% of the probability mass to the first action, with the rest spread equally among the other actions.

Similar to the previous experiment, we computed the bias and RMSE of the different estimators with respect to the correct value function. We sampled 10 different MDPs, and for each of the MDPs we generated 1000 batches. The results we will present are averaged over the resulting 10000 batches.

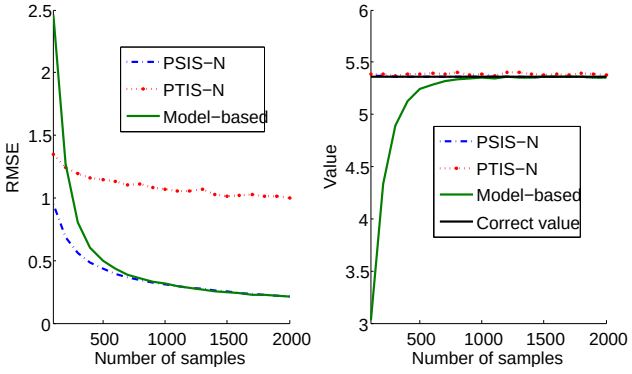
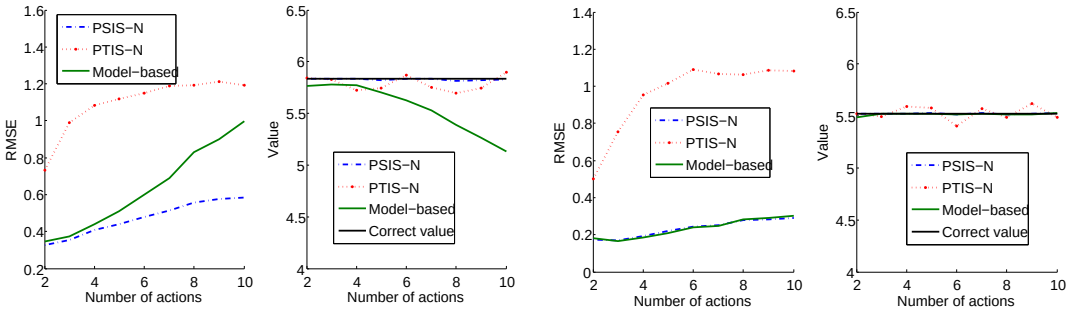


Figure 4.5: RMSE and bias of the 20-step value function estimate for various sample sizes on the random MDPs.

Figure 4.5 illustrates the performance of the methods as a function of the sample size on a 20-step problem. The results are similar to those observed on the mallard domain, with PSIS-N having the best performance, the model-based estimator having high RMSE for small sample sizes due to bias induced by having to use default parameter values when $n_{sa} = 0$, and PTIS-N being very slow to converge.

In order to emphasize the effect that having to use default parameter values when $n_{sa} = 0$ has on the model-based method, we conducted an experiment where we varied the number of actions for our random MDPs. Our hypothesis was that the probability that some n_{sa} is zero will increase as the number of actions increases, and therefore the bias of the model-based method will increase as well. The results, displayed in Figure 4.6, illustrate this phenomenon. For small sample sizes (Figure 4.6(a)), the bias of the model-based method increases steeply as the number of available actions increases. In contrast, PSIS-N appears to be more robust to changes in the size of the action set. For larger sample sizes, $P(n_{sa} = 0)$



(*a*) Batch size = 500 transitions. (*b*) Batch size = 2000 transitions.

Figure 4.6: RMSE and bias of the 20-step value function estimate for various action set sizes on the random MDPs. The two graphs differ in the number of samples in the each batch.

decreases, and all methods are affected by the change in the size of the action set in a similar way (Figure 4.6(*b*)).

5. Discussion

We studied several off-policy learning algorithms, including two new estimators, PSIS and PSIS-N, that are per-step versions of importance sampling which take advantage of Markov assumptions about the model. We briefly discussed the estimators' bias and consistency, and presented a detailed empirical analysis of their performance in a case study pertaining to the management of an animal species. We found that the model-based estimator and the normalized per-step estimator (PSIS-N) performed particularly well. We also found that the model-based estimator can suffer from significant bias if no samples are available for some of the state-action pairs, particularly for problems with many available actions.

We emphasize that the importance sampling estimators require a fixed and known behaviour policy. If the behaviour policy is instead estimated from data, we obtain the model-based estimator (as shown). Cases in which the data is gathered according to multiple behaviour policies (e.g. gathered from different geographic locations), could also be easily incorporated in the estimators by appropriate weighting of the different data batches.

The bias and variance of the off-policy estimators were illustrated through the empirical results. From a theoretical standpoint, there are challenges in providing a formal analysis of the weighted estimators in the sequential case (horizon > 1). This is an interesting area for future work, though we expect it may be difficult to obtain closed-form expressions for these quantities.

As shown in our experiments, it is crucial to assess values for longer time horizons, as the horizon impacts the value of a policy, as well as the ordering of policies. Our results suggest that the horizon length should also be an important factor when choosing an estimator. Some decision-making domains, notably in medicine, deal with relatively short horizons, and in those cases estimators such as PTIS, which have large variance over long horizons

but are unbiased, may be preferable. In domains with longer decision horizons, estimators such as PSIS-N tend to have lower error (though the error increases with horizon length).

One limitation of this work is that the empirical domains are discrete, simulated environments. Additional experiments with real data would undoubtedly make the analysis more compelling. However, gathering real data under a new policy in domains such as natural resource management tends to be expensive, or even impossible. Therefore, ground truth (V^π in our case) is difficult to establish, potentially rendering comparative analyses less meaningful.

In continuous MDPs, off-policy learning can be applied, but generates further complications. In particular, the discrepancy between the probability of a trajectory under the behaviour and the target policy can lead to divergence. Several algorithms have been proposed in order to account for trajectory distribution discrepancies. Precup et al. [2001] use importance sampling weights to correct for the probability of reaching a specific point in a trajectory. The resulting estimators are consistent (in the space of representable value functions) but tend to have high variance. Sutton et al. [2009] address the problem of off-policy learning from on-line data. The main idea is to estimate a secondary set of parameters (in addition to those describing the value function), which are used to stabilize the value function weights and prevent divergence. In discrete MDPs, however, all these estimators are more conservative and hence less sample-efficient than those on which we focused.

Acknowledgements

Funding for this research was provided by the National Institutes of Health (grant R21 DA019800) and the Natural Sciences and Engineering Council Canada (Discovery Grant program).

References

Naoki Abe, P. Melville, C. Pendus, C.K. Reddy, D.L. Jensen, V.P. Thomas, J.J. Bennett, G.F. Anderson, B.R. Cooley, M. Kowalczyk, and Others. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2010.

D.R. Anderson. Optimal exploitation strategies for an animal population in a Markovian environment: a theory and an example. *Ecology*, 56(6):1281–1297, 1975.

J.A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49(2):233–246, 2002.

M. Castronovo and D. Ernst. Learning Exploration / Exploitation Strategies for Single Trajectory Reinforcement Learning. In *10th European Workshop on Reinforcement Learning*, 2012.

Edward J. Dudewicz and Satya N. Mishra. *Modern Mathematical Statistics*. Wiley, New York, NY, 1988.

M Dudik, J Langford, and L Li. Doubly Robust Policy Evaluation and Learning. In *International Conference on Machine Learning*, 2011.

C.J. Fonnesbeck. Solving dynamic wildlife resource optimization problems using reinforcement learning. *Natural Resource Modeling*, 18(1):1–40, 2005.

L Hannah and D Dunson. Approximate Dynamic Programming for Storage Problems. In *International Conference on Machine Learning*, 2011.

JDY Kang, J L Schafer, Anastasios a Tsiatis, and Marie Davidian. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):569–573, January 2007.

S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and Variance Approximation in Value Function Estimates. *Management Science*, 53(2):308–322, February 2007.

S A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–81, May 2005.

J Pineau, A Guez, R D Vincent, G Panuccio, and M Avoli. Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach. *International journal of neural systems*, 19(4):227–40, August 2009.

D Precup, RS Sutton, and S Singh. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Doina Precup, RS Sutton, and S Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

M Riedmiller. Neural fitted Q-iteration - first experiences with a data efficient neural reinforcement learning method. In *Proceedings of the European Conference on Machine Learning*, volume 3720. Springer, 2005.

J M Robins, M a Hernán, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–60, September 2000.

Reuven Y. Rubinstein. *Simulation and the Monte Carlo Method*. Wiley, New York, NY, 1981.

R S Sutton and A G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvari, and Eric Wiewiora. Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation. In *International Conference on Machine Learning*, 2009.