Offline Evaluation of Online Reinforcement Learning Algorithms

Travis Mandel[1], **Yun-En Liu**[2], **Emma Brunskill**[3], **and Zoran Popović**[1,2]
[1]Center for Game Science, Computer Science & Engineering, University of Washington, Seattle, WA
[2]Enlearn™, Seattle, WA
[3]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA
{tmandel, zoran}@cs.washington.edu, yunliu@enlearn.org, ebrun@cs.cmu.edu

Abstract

In many real-world reinforcement learning problems, we have access to an existing dataset and would like to use it to evaluate various learning approaches. Typically, one would prefer not to deploy a fixed policy, but rather an algorithm that learns to improve its behavior as it gains more experience. Therefore, we seek to evaluate how a proposed algorithm *learns* in our environment, meaning we need to evaluate how an algorithm would have gathered experience if it were run online. In this work, we develop three new evaluation approaches which guarantee that, given some history, algorithms are fed samples from the distribution that they would have encountered if they were run online. Additionally, we are the first to propose an approach that is provably unbiased given finite data, eliminating bias due to the length of the evaluation. Finally, we compare the sample-efficiency of these approaches on multiple datasets, including one from a real-world deployment of an educational game.

1 Introduction

There is a growing interest in deploying reinforcement learning (RL) agents in real-world environments, such as healthcare or education. In these high-risk situations one cannot deploy an arbitrary algorithm and hope it works well. Instead one needs confidence in an algorithm before risking deployment. Additionally, we often have a large number of algorithms (and associated hyperparameter settings), and it is unclear which will work best in our setting. We would like a way to compare these algorithms without needing to collect new data, which could be risky or expensive.

An important related problem is developing testbeds on which we can evaluate new reinforcement learning algorithms. Historically, these algorithms have been evaluated on simple hand-designed problems from the literature, often with a small number of states or state variables. Recently, work has considered using a diverse suite of Atari games as a testbed for evaluating reinforcement learning algorithms (Bellemare et al. 2013). However, it is not clear that these artificial problems accurately reflect the complex structure present in real-world environments. An attractive alternative is to use precollected real-world datasets to evaluate new RL algorithms on real problems of interest in domains such as healthcare, education, or e-commerce.

These problems have ignited a recent renewal of interest in offline policy evaluation in the RL community (Mandel et al. 2014; Thomas, Theocharous, and Ghavamzadeh 2015), where one uses a precollected dataset to achieve high-quality estimates of the performance of a proposed policy. However, this prior work focuses only on evaluating a *fixed policy* learned from historical data. In many real-world problems, we would instead prefer to deploy a *learning algorithm* that continues to learn over time, as we expect that it will improve over time and thus (eventually) outperform such a fixed policy. Further, we wish to develop testbeds for RL algorithms which evaluate how they learn over time, not just the final policy they produce.

However, evaluating a learning algorithm is very different from evaluating a fixed policy. We cannot evaluate an algorithm's ability to learn by, for example, feeding it 70% of the precollected dataset as training data and evaluating the produced policy on the remaining 30%. Online, it would have collected a different training dataset based on how it trades off exploration and exploitation. In order to evaluate the performance of an algorithm as it learns, we need to simulate running the algorithm by allowing it to interact with the evaluator as it would with the real (stationary) environment, and record the resulting performance estimates (e.g. cumulative reward). See figure 1.[1]

A typical approach to creating such an evaluator is to build a model using the historical data, particularly if the environment is known to be a discrete MDP. However, this approach can result in error that accumulates at least quadratically with the evaluation length (Ross, Gordon, and Bagnell 2011). Equally important, in practice it can result in very poor estimates, as we demonstrate in our experiments section. Worse, in complex real-world domains, it is often unclear how to build accurate models. An alternate approach is to try to adapt importance sampling techniques to this problem, but the variance of this approach is unusably high if we wish to evaluate an algorithm for hundreds of timesteps (Dudík et al. 2014).

[1]In the bandit community, this problem setup is called *nonstationary policy evaluation*, but we avoid use of this term to prevent confusion, as these terms are used in many different RL contexts.

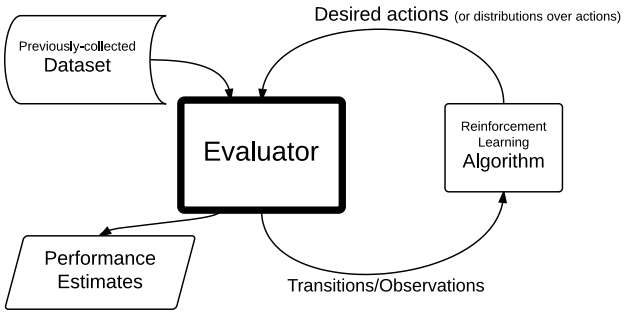


Figure 1: Evaluation process: We are interested in developing evaluators that use a previously-collected dataset to interact with an arbitrary reinforcement learning algorithm as it would interact with the true environment. As it interacts, the evaluator produces performance estimates (e.g. cumulative reward).

In this paper, we present, to our knowledge, the first methods for using historical data to evaluate how an RL algorithm would perform online, which possess both meaningful guarantees on the quality of the resulting performance estimates and good empirical performance. Building upon state-of-the-art work in offline bandit algorithm evaluation (Li et al. 2011; Mandel et al. 2015), we develop three evaluation approaches for reinforcement learning algorithms: queue-based (Queue), per-state rejection sampling (PSRS), and per-episode rejection sampling (PERS). We prove that given the current history, and that the algorithm receives a next observation and reward, that observation and reward is drawn from a distribution identical to the distribution the algorithm would have encountered if it were run in the real world. We show how to modify PERS to achieve stronger guarantees, namely per-timestep unbiasedness given a finite dataset, a property that has not previously been shown even for bandit evaluation methods. Our experiments, including those that use data from a real educational domain, show these methods have different tradeoffs. For example, some are more useful for short-horizon representation-agnostic settings, while others are better suited for long-horizon known-state-space settings. For an overview of further tradeoffs see Table 1. We believe this work will be useful for practitioners who wish to evaluate RL algorithms in a reliable manner given access to historical data.

2 Background and Setting

A discrete Markov Decision Process (MDP) is specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, s_I)$ where $\mathcal{S}$ is a discrete state space, $\mathcal{A}$ is a discrete action space, $\mathcal{R}$ is a mapping from state, action, next state tuples to a distribution over real valued rewards, $\mathcal{T}$ is a transition model that maps state, action, next state tuples to a probability between 0 and 1, and s_I denotes the starting state[2]. We consider an episode to end (and a new episode to begin) when the system transitions back to initial state s_I.

[2]Our techniques could still apply given multiple starting states, but for simplicity we assume a single starting state s_I

We assume, unless otherwise specified, that the domain consists of an episodic MDP $\mathcal{M}$ with a given state space $\mathcal{S}$ and action space $\mathcal{A}$, but unknown reward model $\mathcal{R}$ and transition model $\mathcal{T}$. As input we assume a set D of N transitions, (s, a, r, s') drawn from a fixed sampling policy π_e.

Our objective is to use this data to evaluate the performance of a RL algorithm $\mathbb{A}$. Specifically, without loss of generality[3] we will discuss estimating the discounted sum of rewards obtained by the algorithm $\mathbb{A}$ for a sequence of episodes, e.g. $R^{\mathbb{A}}(i) = \sum_{j=0}^{L(i)-1} \gamma^j r_j$ where $L(i)$ denotes the number of interactions in the i^{th} episode.

At each timestep $t = 1 \ldots \infty$, the algorithm $\mathbb{A}$ outputs a (possibility stochastic) policy π_b from which the next action should be drawn, potentially sampling a set of random numbers as part of this process. For concreteness, we refer to this (possibly random length) vector of random samples used by $\mathbb{A}$ on a given timestep with the variable χ. Let H_T be the history of (s, a, r, s') and χ consumed by $\mathbb{A}$ up to time T. Then, we can say that the behavior of $\mathbb{A}$ at time T depends only on H_T.

Our goal is to create evaluators (sometimes called replayers) that enable us to simulate running the algorithm $\mathbb{A}$ in the true MDP $\mathcal{M}$ using the input dataset D. One key aspect of our proposed evaluators is that they terminate at some timestep. To that end, let g_t denote the event that we do not terminate before outputting an estimate at timestep t (so g_t implies $g_1, \ldots, g_{t-1}$). In order to compare the evaluator to reality, let $P_{\mathcal{R}}(x)$ denote the probability (or pdf if x contains continuous rewards[4]) of generating x under the evaluator, and $P_{\mathcal{E}}(x)$ denote the probability of generating x in the true environment (the MDP $\mathcal{M}$). Similarly, $\mathbb{E}_{\mathcal{R}}[x]$ is the expected value of a random variable x under the evaluator, and $\mathbb{E}_{\mathcal{E}}[x]$ is the expected value of x under the true environment ($\mathcal{M}$).

We will shortly introduce several evaluators. Due to space limitations, we provide only proof sketches: full proofs are in the appendix (available at http://grail.cs.washington.edu/projects/nonstationaryeval).

What guarantees do we desire on the estimates our evaluators produce? Unbiasedness of the reward estimate on episode i is a natural choice, but it is unclear what this means if we do not always output an estimate of episode i due to termination caused by the limited size/coverage of our dataset. Therefore, we show a guarantee that is in some sense weaker, but applies given a finite dataset: Given some history, the evaluator either terminates or updates the algorithm as it would if run online. Given this guarantee, the empirical question is how early termination occurs, which we address experimentally. We now highlight some of the properties we would like an evaluator to possess, which are summarized in Table 1.

1. *Given some history, the (s, a, r, s') tuples provided to $\mathbb{A}$ have the same distribution as those the agent would*

[3]It is easy to modify the evaluators to compute other statistics of the interaction of the algorithm $\mathbb{A}$ with the evaluator, such as the cumulative reward, or the variance of rewards.

[4]In this case, sums should be considered to be integrals

	Samples true	Unbiased estimate of i-th episode performance	Allows unknown sampling distribution	Does not assume Markov	Computationally efficient
Queue	✓	×	✓	×	✓
PSRS	✓	×	×	×	✓
PERS	✓	×(Variants:✓)	×	✓	Not always

Table 1: Desired properties of evaluation approaches, and a comparison of the three evaluators introduced in this paper. We did not include the sample efficiency, because although it is a key metric it is typically domain-dependent.

receive in the true MDP $\mathcal{M}$. Specifically, we desire $P_{\mathcal{R}}(s, a, r, s', \chi | H_T, g_T) = P_{\mathcal{E}}(s, a, r, s', \chi | H_T)$ so that $P_{\mathcal{R}}(H_{T+1} | H_T, g_T) = P_{\mathcal{E}}(H_{T+1} | H_T)$. As mentioned above, this guarantee allows us to ensure that the algorithm is fed on-policy samples, guaranteeing the algorithm behaves similarly to how it would online.

2. *High sample efficiency*. Since all of our approaches only provide estimates for a finite number of episodes before terminating due to lack of data in D, we want to make efficient use of data to evaluate $\mathbb{A}$ for as long as possible.
3. *Given an input i, outputs an unbiased estimate of $R^{\mathbb{A}}(i)$*. Specifically, $\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(i)] = \mathbb{E}_{\mathcal{E}}[R^{\mathbb{A}}(i)]$. Note that this is non-trivial to ensure, since the evaluation may halt before the i-th episode is reached.
4. *Can leverage data D collected using an unknown sampling distribution π_e*. In some situations it may be difficult to log or access the sampling policy π_e, for example in the case where human doctors choose treatments for patients.
5. *Does not assume the environment is a discrete MDP with a known state space $\mathcal{S}$*. In many real world problems, the state space is unknown, partially observed, or continuous, so we cannot always rely on Markov assumptions.
6. *Computationally efficient*.

3 Related work

Work in reinforcement learning has typically focused on evaluating fixed policies using importance sampling techniques (Precup 2000). Importance sampling is widely-used in off-policy learning, as an objective function when using policy gradient methods (Levine and Koltun 2013; Peshkin and Shelton 2002) or as a way to re-weight samples in off-policy TD-learning methods (Mahmood, van Hasselt, and Sutton 2014; Sutton, Mahmood, and White 2015; Maei and Sutton 2010). Additionally, this approach has recently enabled practitioners to evaluate learned policies on complex real-world settings (Thomas, Theocharous, and Ghavamzadeh 2015; Mandel et al. 2014). However, this work focuses on evaluating fixed policies, we are not aware of work specifically focusing on the problem of evaluating how an RL algorithm would learn online, which involves feeding the algorithm new training samples as well as evaluating its current performance. It is worth noting that any of the above-mentioned off-policy learning algorithms could be evaluated using our methods.

Our methods do bear a relationship to off-policy learning work which has evaluated policies by synthesizing artificial trajectories (Fonteneau et al. 2010; 2013). Unlike our work, this approach focuses only on evaluating fixed policies. It also assumes a degree of Lipschitz continuity in some continuous space, which introduces bias. There are some connections: our queue-based estimator could be viewed as related to their work, but focused on evaluating learning algorithms in the discrete MDP policy case.

One area of related work is in the area of (possibly contextual) multi-armed bandits, in which the corresponding problem is termed "nonstationary policy evaluation". Past work has showed evaluation methods that are guaranteed to be unbiased (Li et al. 2011), or have low bias (Dudík et al. 2012; 2014), but only assuming an infinite data stream. Other work has focused on evaluators that perform well empirically but lack this unbiasedness (Mary, Preux, and Nicol 2014). Work by Mandel et al. 2015 in the non-contextual bandit setting show guarantees similar to ours, that issued feedback comes from the true distribution even with finite data. However, in addition to focusing on the more general setting of reinforcement learning, we also show stronger guarantees of unbiasedness even given a finite dataset.

Algorithm 1 Queue-based Evaluator

```
Input: Dataset D, RL Algorithm 𝔸, Starting state s_I
Output: R^𝔸 s.t. R^𝔸(i) is sum of rewards in ep. i
Q[s,a] = Queue(RandomOrder((s_i,a_i,r,s') ∈ D, s.t. s_i = s, a_i = a)), ∀s ∈ S, a ∈ A
for i = 1 to ∞ do
    s = s_I, t = 0, r_i = 0
    Let π_b be 𝔸's initial policy
    while ¬(t > 0 and s == s_I) do
        a_b ~ π_b(s)
        if Q[s,a_b] is empty then return R^𝔸
        (r, s') = Q[s,a_b].pop()
        Update 𝔸 with (s,a,r,s'), yields new policy π_b
        r_i = r_i + γ^t r, s = s', t = t + 1
    R^𝔸(i) = r_i
```

4 Queue-based Evaluator

We first propose a *queue-based* evaluator for evaluating algorithms for episodic MDPs with a provided state $\mathcal{S}$ and action $\mathcal{A}$ space (Algorithm 1). This technique is inspired by the queue-based approach to evaluation in non-contextual bandits (Mandel et al. 2015). The key idea is to place feedback (next states and rewards) in queues, and remove elements

based on the current state and chosen action, terminating evaluation when we hit an empty queue. Specifically, first we partition the input dataset D into queues, one queue per (s, a) pair, and fill each queue $Q(s, a)$ with a (random) ordering of all tuples $(r, s') \in D$ s.t. $(s_i = s, a_i = a, r_i = r, s'_i = s')$ To simulate algorithm $\mathbb{A}$ starting from a known state s_k, the algorithm $\mathbb{A}$ outputs a policy π_b, and selects an action a sampled from $\pi_b(s_k)$.[5]

The evaluator then removes a tuple (r, s') from queue $Q[s_k, a]$, which is used to update the algorithm $\mathbb{A}$ and its policy π_b, and simulate a transition to the next state s'. By the Markov assumption, tuples (r, s') are i.i.d. given the prior state and selected action, and therefore an element drawn without replacement from the queue has the same distribution as that in the true environment. The evaluator terminates and outputs the reward vector $R^{\mathbb{A}}$, when it seeks to draw a sample from an empty queue.[6]

Unlike many offline evaluation approaches (such as importance sampling for policy evaluation), our queue evaluator does not require knowledge of the sampling distribution π_e used to generate D. It can even use data gathered from a deterministic sampling distribution. Both properties are useful for many domains (for example, it may be hard to know the stochastic policy used by a doctor to make a decision).

Theorem 4.1. *Assuming the environment is an MDP with state space $\mathcal{S}$ and the randomness involved in drawing from π_b is treated as internal to $\mathbb{A}$, given any history of interactions H_T, if the queue-based evaluator produces a (s, a, r, s') tuple, the distribution of this tuple and subsequent internal randomness χ under the queue-based evaluator is identical to the true distribution the agent would have encountered if it was run online. That is, $P_{\mathcal{R}}(s, a, r, s', \chi | H_T, g_T) = P_{\mathcal{E}}(s, a, r, s', \chi | H_T)$, which gives us that $P_{\mathcal{R}}(H_{T+1} | H_T, g_T) = P_{\mathcal{E}}(H_{T+1} | H_T)$.*

Proof Sketch. The proof follows fairly directly from the fact that placing an (r, s') tuple drawn from $\mathcal{M}$ in $Q[s, a]$ and sampling from Q without replacement results in a sample from the true distribution. See the appendix (available at http://grail.cs.washington.edu/projects/nonstationaryeval).

Note that theorem 4.1 requires us to condition on the fact that $\mathbb{A}$ reveals no randomness, that is, we consider the randomness involved in drawing from π_b on line 8 to be considered as internal, that is (included in χ). This means the guarantee is slightly weaker than the approaches we will present in sections 5 and 6, which condition on general π_b.

5 Per-State Rejection Sampling Evaluator

Ideally, we would like an evaluator that can recognize when the algorithm chooses actions similarly to the sampling distribution, in order use more of the data. For example, in the extreme case where we know the algorithm we are evaluating always outputs the sampling policy, we should be able to make use of all data, or close to it. However, the queue method only uses the sampled action, and thus cannot determine directly whether or not the distribution over actions at each step (π_b) is similar to the sampling policy (π_e). This can make a major difference in practice: If π_b and π_e are both uniform, and the action space is large relative to the amount of data, we will be likely to hit an empty queue if we sample a fresh action from π_b. But, if we know the distributions are the same we can simply take the first sampled action from π_e. Being able to take advantage of stochastic distributions in this way is sometimes referred to as leveraging *revealed randomness* in the candidate algorithm (Dudík et al. 2012).

To better leverage this similarity, we introduce the Per-State Rejection Sampling (PSRS) evaluator (see Algorithm 2), inspired by approaches used in contextual bandits (Li et al. 2011; Dudík et al. 2012). PSRS divides data into streams for each state s, consisting of a (randomized) list of the subsequent (a, r, s') tuples that were encountered from s in the input data. Specifically, given the current state s, our goal is to sample a tuple (a, r, s') such that a is sampled from algorithm $\mathbb{A}$'s current policy $\pi_b(s)$, and r and s' are sampled from the true environment. We already know that given the Markov property, once we select an action a that r and s' in a tuple (s, a, r, s') represent true samples from the underlying Markov environment. The challenge then becomes to sample an action a from $\pi_b(s)$ using the actions sampled by the sampling distribution $\pi_e(s)$ for the current state s. To do this, a rejection sampling algorithm[7] samples a uniform number u between 0 and 1, and accepts a sample (s, a, r, s') from D if $u < \frac{\pi_b(a|s)}{M \pi_e(a|s)}$, where $\pi_b(a|s)$ is the probability under the candidate distribution of sampling action a for state s, $\pi_e(a|s)$ is the corresponding quantity for the sampling distribution, and M is an upper bound on their ratio, $M \geq \max_a \frac{\pi_b(a|s)}{\pi_e(a|s)}$. M is computed by iterating over actions[8] (line 8). It is well known that samples rejection sampling accepts represent true samples from the desired distribution, here π_b (Gelman et al. 2014).

Slightly surprisingly, even if $\mathbb{A}$ always outputs the sampling policy π_e, we do not always consume all samples (in other words PSRS is not *self-idempotent*), unless the original ordering of the streams is preserved (see appendix, available at http://grail.cs.washington.edu/projects/nonstationaryeval). Still, in the face of stochasticity PSRS can be significantly more data-efficient than the Queue-based evaluator.

Theorem 5.1. *Assume the environment is an MDP with state space $\mathcal{S}$, π_e is known, and for all a, $\pi_e(a) > 0$ if $\pi_b(a) > 0$. Then if the evaluator produces a (s, a, r, s') tuple, the distribution of (s, a, r, s') tuple returned by PSRS*

[5] Note that since we only use π_b to draw the next action, this does not prevent $\mathbb{A}$ from internally using a policy that depends on more than s (for example, s and t in finite horizon settings).

[6] For details about why this is necessary, see the appendix (available at http://grail.cs.washington.edu/projects/nonstationaryeval).

[7] One might wonder if we could reduce variance by using an importance sampling instead of rejection sampling approach here. Although in theory possible, one has to keep track of all the different states of the algorithm with and without each datapoint accepted, which is computationally intractable.

[8] This approach is efficient in the since that it takes time linear in $|\mathcal{A}|$, however in very large action spaces this might be too expensive. In certain situations it may be possible to analytically derive a bound on the ratio to avoid this computation.

Algorithm 2 Per-State Rejection Sampling Evaluator

Input: Dataset D, RL Algorithm $\mathbb{A}$, Start state s_I, π_e
Output: Output: $R^{\mathbb{A}}$ s.t. $R^{\mathbb{A}}(i)$ is sum of rewards in ep. i
$Q[s] = Queue(RandomOrder((s_i, a_i, r, s') \in D \text{ s.t. } s_i = s)), \forall s \in \mathcal{S}$
for $i = 1$ to ∞ **do**
    $s = s_I$,$t = 0, r_i = 0$
    Let π_b be $\mathbb{A}$'s initial policy
    while $\neg(t > 0$ and $s_t == s_I)$ **do**
        $M = \max_a \frac{\pi_b(a|s)}{\pi_e(a|s)}$
        (a, r, s') = Q[s].pop()
        if Q[s] is empty **then return** $R^{\mathbb{A}}$
        Sample $u \sim Uniform(0, 1)$
        if $u > \frac{\pi_b(a|s)}{M\pi_e(a|s)}$ **then**
            Reject sample, go to line 9
        Update $\mathbb{A}$ with (s, a, r, s'), yields new policy π_b
        $r_i = r_i + \gamma^t r, s = s', t = t + 1$
    $R^{\mathbb{A}}(i) = r_i$

(and subsequent internal randomness χ) given any history of interactions H_T is identical to the true distribution the agent would have encountered if was run online. Precisely, $P_{\mathcal{R}}(s, a, r, s', \chi|H_T, g_T) = P_{\mathcal{E}}(s, a, r, s', \chi|H_T)$, which gives us that $P_{\mathcal{R}}(H_{T+1}|H_T, g_T) = P_{\mathcal{E}}(H_{T+1}|H_T)$.

Proof Sketch. The proof follows fairly directly from the fact that given finite dataset, rejection sampling returns samples from the correct distribution (Lemma 1 in the appendix, available at http://grail.cs.washington.edu/projects/nonstationaryeval).

6 Per-Episode Rejection Sampling

The previous methods assumed the environment is a MDP with a known state space. We now consider the more general setting where the environment consists of a (possibly high dimensional, continuous) observation space $\mathcal{O}$, and a discrete action space $\mathcal{A}$. The dynamics of the environment can depend on the full history of prior observations, actions, and rewards, $h_t = o_0, \ldots, o_t, a_0, \ldots, a_{t-1}, r_0, \ldots, r_{t-1}$. Multiple existing models, such as POMDPs and PSRs, can be represented in this setting. We would like to build an evaluator that is *representation-agnostic*, i.e. does not require Markov assumptions, and whose sample-efficiency does not depend on the size of the observation space.

We introduce the Per-Episode Rejection Sampler (PERS) evaluator (Algorithm 3) that evaluates RL algorithms in these more generic environments. In this setting we assume that the dataset D consists of a stream of episodes, where each episode e represents an ordered trajectory of actions, rewards and observations, $(o_0, a_0, r_0, o_1, a_1, r_1, \ldots, r_{l(e)})$ obtained by executing the sampling distribution π_e for $l(e)-1$ time steps in the environment. We assume that π_e may also be a function of the full history h_t in this episode up to the current time point. For simplicity of notation, instead of keeping track of multiple policies π_b, we simply write π_b (which could implicitly depend on χ).

Algorithm 3 Per-Episode Rejection Sampling Evaluator

Input: Dataset of episodes D, RL Algorithm $\mathbb{A}$, π_e
Output: Output: $R^{\mathbb{A}}$ s.t. $R^{\mathbb{A}}(i)$ is sum of rewards in ep. i
Randomly shuffle D
Store present state $\mathcal{A}$ of algorithm $\mathbb{A}$
$M = calculateEpisodeM(\mathcal{A}, \pi_e)$ (see the appendix)
$i = 1$, Let π_b be $\mathbb{A}$'s initial policy
for $e \in D$ **do**
    $p = 1.0, h = [], t = 0, r_i = 0$
    for $(o, a, r) \in e$ **do**
        $h \to (h, o)$
        $p = p\frac{\pi_b(a|h)}{\pi_e(a|h)}$
        Update $\mathbb{A}$ with (o, a, r), output new policy π_b
        $h \to (h, a, r)$, $r_i = r_i + \gamma^t r$
    Sample $u \sim Uniform(0, 1)$
    if $u > \frac{p}{M}$ **then**
        Roll back algorithm: $\mathbb{A} = \mathcal{A}$
    else
        Store present state $\mathcal{A}$ of algorithm $\mathbb{A}$
        $M = calculateEpisodeM(\mathcal{A}, \pi_e)$
        $R^{\mathbb{A}}(i) = r_i, i = i + 1$
return $R^{\mathbb{A}}$

PERS operates similarly to PSRS, but performs rejection sampling at the episode level. This involves computing the ratio of $\frac{\Pi_{t=0}^{l(e)}\pi_b(a_t|h_t)}{M\Pi_{t=0}^{l(e)}\pi_e(a_t|h_t)}$, and accepting or rejecting the episode according to whether a random variable sampled from the uniform distribution is lower than the computed ratio. As M is a constant that represents the maximum possible ratio between the candidate and sampling *episode* probabilities, it can be computationally involved to compute M exactly. Due to space limitations, we present approaches for computing M in the appendix (available at http://grail.cs.washington.edu/projects/nonstationaryeval). Note that since the probability of accepting an episode is based only on a ratio of action probabilities, one major benefit to PERS is that its sample-efficiency does not depend on the size of the observation space. However, it does depend strongly on the episode length, as we will see in our experiments.

Although PERS works on an episode-level, to handle algorithms that update after every timestep, it updates $\mathbb{A}$ throughput the episode and "rolls back" the state of the algorithm if the episode is rejected (see Algorithm 3).

Unlike PSRS, PERS is self-idempotent, meaning if $\mathbb{A}$ always outputs π_e we accept all data. This follows since if $\pi_e(a_t|h_t) = \pi_b(a_t|h_t)$, $M = 1$ and $\frac{\Pi_{t=0}^{l(e)}\pi_b(a_t|h_t)}{M\Pi_{t=0}^{l(e)}\pi_e(a_t|h_t)} = 1$.

Theorem 6.1. *Assuming π_e is known, and $\pi_b(e) > 0 \to \pi_e(e) > 0$ for all possible episodes e and all π_b, and PERS outputs an episode e, then the distribution of e (and subsequent internal randomness χ) given any history of episodic interactions H_T using PERS is identical to the true distribution the agent would have encountered if it was run online. That is, $P_{\mathcal{E}}(e, \chi|H_T) = P_{\mathcal{R}}(e, \chi|H_T, g_T)$, which gives us*

that $P_{\mathcal{R}}(H_{T+1}|H_T, g_T) = P_{\mathcal{E}}(H_{T+1}|H_T)$.

Proof Sketch. The proof follows fairly directly from the fact that given finite dataset, rejection sampling returns samples from the correct distribution (Lemma 1 in the appendix, available at http://grail.cs.washington.edu/projects/nonstationaryeval).

7 Unbiasedness Guarantees in the Per-Episode case

Our previous guarantees stated that if we return a sample, it is from the true distribution given the history. Although this is fairly strong, it does not ensure $R^A(i)$ is an unbiased estimate of the reward obtained by A in episode i. The difficulty is that across multiple runs of evaluation, the evaluator may terminate after different numbers of episodes. The probability of termination depends on a host of factors (how random the policy is, which state we are in, etc.). This can result in a bias, as certain situations may be more likely to reach a given length than others.

For example, consider running the queue-based approach on a 3-state MDP: s_I is the initial state, if we take action a_0 we transition to state s_1, if we take action a_1 we transition to s_2. The episode always ends after timestep 2. Imagine the sampling policy chose a_1 99% of the time, but our algorithm chose a_1 50% of the time. If we run the queue approach many times in this setting, runs where the algorithm chose a_1 will be much more likely to reach timestep 2 than those where it chose a_0, since s_2 is likely to have many more samples than s_1. This can result in a bias: if the agent receives a higher reward for ending in s_2 compared to s_1, the average reward it receives at timestep 2 will be overestimated.

One approach proposed by past work (Mandel et al. 2015; Dudík et al. 2014) is to assume T (the maximum timestep/episode count for which we report estimates) is small enough such that over multiple runs of evaluation we usually terminate after T; however it can be difficult to fully bound the remaining bias. Eliminating this bias for the state-based methods is difficult, since the the agent is much more likely to terminate if it transitions to a sparsely-visited state, and so the probability of terminating is hard to compute as it depends on the unknown transition probabilities.

However, modifying PERS to use a fixed M throughout its operation allows us to show that if PERS outputs an estimate, that estimate is unbiased (Theorem 7.1). In practice one will likely have to overestimate this M, for example by bounding $p(x)$ by 1 (or $(1-\epsilon)$ for epsilon-greedy) and calculating the minimum $q(x)$.

Theorem 7.1. *If M is held fixed throughout the operation of PERS,* π_e *is known, and* $\pi_b(e) > 0 \rightarrow \pi_e(e) > 0$ *for all possible episodes* e *and all* π_b*, then if PERS outputs an estimate of some function* $f(H_T)$ *at episode T, that estimate is an unbiased estimator of* $f(H_T)$ *at episode T, in other words,* $\mathbb{E}_{\mathcal{R}}[f(H_T)|g_T, \ldots, g_1] = \sum_{H_T} f(H_T) P_{\mathcal{E}}(H_T) = \mathbb{E}_{\mathcal{E}}[f(H_T)]$*. For example, if* $f(H_T) = R^{\mathbb{A}}(T)$*, the estimate is an unbiased estimator of* $R^{\mathbb{A}}(T)$ *given* $g_T, \ldots, g_1$.

Proof Sketch. We first show that if M is fixed, the probability that each episode is accepted is constant $(1/M)$. This allows us to show that whether we continue or not (g_T) is conditionally independent of H_{T-1}. This lets us remove the conditioning on H_{T-1} in Theorem 6.1 to give us that $P_{\mathcal{R}}(H_T|g_T, \ldots, g_1) = P_{\mathcal{E}}(H_T)$, meaning the distribution over histories after T accepted episodes is correct, from which conditional unbiasedness is easily shown.

Although useful, this guarantee has the downside that the estimate is still conditional on the fact that our approach does not terminate. Theorem 7.2 shows that it is possible to use a further modification of Fixed-M PERS based on importance weighting to always issue unbiased estimates for N total episodes. For a discussion of the empirical downsides to this approach, see the appendix (available at http://grail.cs.washington.edu/projects/nonstationaryeval).

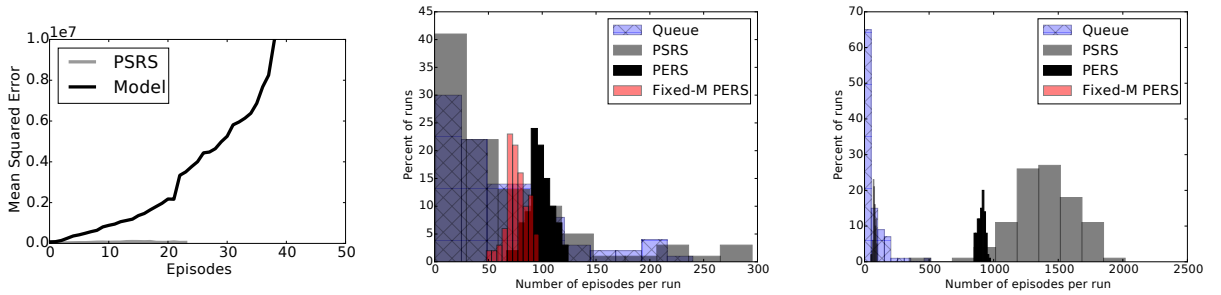
Theorem 7.2. *Assuming for each* T, $R^{\mathbb{A}}(T)$ *is divided by by* $\phi = 1 - Binomial(N, 1/M).cdf(k-1)$*, and after terminating at timestep k we output 0 as estimates of reward for episodes* $k+1, \ldots, N$*, and M is held fixed throughout the operation of PERS, and* π_e *is known, and* $\pi_b(e) > 0 \rightarrow \pi_e(e) > 0$ *for all possible episodes* e *and all* π_b*, then the estimate of reward output at each episode* $T = 1 \ldots N$ *is an unbiased estimator of* $R^{\mathbb{A}}(T)$.

Proof Sketch. Outputting an estimate of the reward at an episode T by either dividing the observed reward by the probability of reaching T (aka $P(g_T, ..., g_1)$), for a run of the evaluator that reaches at least T episodes, or else outputting a 0 if the evaluation has terminated, is an importance weighting technique that ensures the expectation is correct.

8 Experiments

Any RL algorithm could potentially be run with these evaluators. Here, we show results evaluating Posterior Sampling Reinforcement Learning (PSRL) (Osband et al. 2013, Strens 2000), which has shown good empirical and theoretical performance in the finite horizon case. The standard version of PSRL creates one deterministic policy each episode based on a single posterior sample; however, we can sample the posterior multiple times to create multiple policies and randomly choose between them at each step, which allows us to test our evaluators with more or less revealed randomness.

Comparison to a model-based approach We first compare PSRS to a model-based approach on SixArms (Strehl and Littman 2004), a small MDP environment. Our goal is to evaluate the cumulative reward of PSRL run with 10 posterior samples, given a dataset of 100 samples collected using a uniform sampling policy. The model-based approach uses the dataset to build an MLE MDP model. Mean squared error was computed against the average of 1000 runs against the true environment. For details see the appendix (available at http://grail.cs.washington.edu/projects/nonstationaryeval).
In Figure 2a we see that the model-based approach starts fairly accurate but quickly begins returning very poor estimates. In this setting, the estimates it returned indicated that PSRL was learning much more quickly than it would in reality. In contrast, our PSRS approach returns much more accurate estimates and ceases evaluation instead of issuing poor estimates.



(a) PSRS tends to be much more accurate than a model-based approach.

(b) Comparing on Treefrog Treasure with 3 timesteps and 1 PSRL posterior sample.

(c) Comparing on Treefrog with 3 timesteps and 10 PSRL posterior samples.

Figure 2: Experimental results.

Figure 3: Treefrog Treasure: players guide a frog through a dynamic world, solving number line problems.

Length Results All three of our estimators produce samples from the correct distribution at every step. However, they may provide different length trajectories before termination. To understand the data-efficiency of each evaluator, we tested them on a real-world educational game dataset, as well as a small but well-known MDP example.

Treefrog Treasure is an educational fractions game (Figure 3). The player controls a frog to navigate levels and jump through numberlines. We have 11 actions which control parameters of the numberlines. Our reward is based on whether students learn (based on pretest-to-postest improvement) and whether they remain engaged (measured by whether the student quit before the posttest). We used a state space consisting of the history of actions and whether or not the student took more than 4 tries to pass a numberline (note that this space grows exponentially with the horizon). We varied the considered horizon between 3 and 4 in our experiments. We collected a dataset of 11,550 players collected from a child-focused educational website, collected using a semi-uniform sampling policy. More complete descriptions of the game, experimental methodology, method of calculating M, and details of PSRL can be found in the appendix (available at http://grail.cs.washington.edu/projects/nonstationaryeval).

Figure 2 shows results on Treefrog Treasure, with histograms over 100 complete runs of each evaluator. The graphs show how many episodes the estimator could evaluate the RL algorithm for, with more being better. PERS does slightly better in a short-horizon deterministic setting (Figure 2b). Increasing the posterior samples greatly improves performance of rejection sampling methods (Figure 2c).

We also examined an increased horizon of 4 (graphs provided in appendix, available at http://grail.cs.washington.edu/projects/nonstationaryeval). Given deterministic policies on this larger state space, all three methods are more or less indistinguishable; however, revealing more randomness causes PERS to overtake PSRS (mean 260.54 vs. 173.52). As an extreme case, we also tried a random policy: this large amount of revealed randomness benefits the rejection sampling methods, especially PERS, which evaluates for much longer than the other approaches. PERS outperforms PSRS here because there are small differences between the random candidate policy and the semi-random sampling policy, and thus if PSRS enters a state with little data it is likely to terminate.

The fixed-M PERS method does much worse than the standard version, typically barely accepting any episodes, with notable exceptions when the horizon is short (Figure 2b). Since it does not adjust M it cannot take advantage of revealed randomness (Figure 2c). However, we still feel that this approach can be useful when one desires truly unbiased estimates, and when the horizon is short. Finally, we also note that PERS tends to have the lowest variance, which makes it an attractive approach since to reduce bias one needs to have a high percentage of runs terminating after the desired length.

The state space used in Treefrog Treasure grows exponentially with the horizon. To examine a contrasting case with a small state space (6 states), but a long horizon (20), we also test our approaches in Riverswim (Osband, Russo, and Van Roy 2013), a standard toy MDP environment. The results can be found in the appendix (available at http://grail.cs.washington.edu/projects/nonstationaryeval), but in general we found that PERS and its variants suffer greatly from the long horizon, while Queue and PSRS

do much better, with PSRS doing particularly well if randomness is revealed.

Our conclusion is that the PERS does quite well, especially if randomness is revealed and the horizon is short. It appears there is little reason to choose Queue over PSRS, except if the sampling distribution is unknown. This is surprising because it conflicts with the results of Mandel et al. 2015. They found a queue-based approach to be more efficient than rejection sampling in a non-contextual bandit setting, since data remained in the queues for future use instead of being rejected. The key difference is that in bandits there is only one state, so we do not encounter the problem that we happen to land on an undersampled state, hit an empty queue by chance, and have to terminate the whole evaluation procedure. If the candidate policy behaves randomly at unvisited states, as is the case with 10-sample PSRL, PSRS can mitigate this problem by recognizing the similarity between sampling and candidate distributions to accept the samples at that state, therefore being much less likely to terminate evaluation when a sparsely-visited state is encountered.

9 Conclusion

We have developed three novel approaches for evaluating how RL algorithms perform online: the most important differences are summarized in Table 1. All methods have guarantees that, given some history, if a sample is output it comes from the true distribution. Further, we developed a variant of PERS with even stronger guarantees of unbiasedness. Empirically, there are a variety of tradeoffs to navigate between the methods, based on horizon, revealed randomness in the candidate algorithm, and state space size. We anticipate these approaches will find wide use when one wishes to compare different reinforcement learning algorithms on a real-world problem before deployment. Further, we are excited at the possibility of using these approaches to create real-world testbeds for reinforcement learning problems, perhaps even leading to RL competitions similar to those which related contextual bandit evaluation work (Li et al. 2011) enabled in that setting (Chou and Lin 2012). Future theoretical work includes analyzing the sample complexity of our approaches and deriving tight deviation bounds on the returned estimates. Another interesting direction is developing more accurate estimators, e.g. by using doubly-robust estimation techniques (Dudík et al. 2012).

Acknowledgments This work was supported by the NSF BIGDATA grant No. DGE-1546510, the Office of Naval Research grant N00014-12-C-0158, the Bill and Melinda Gates Foundation grant OPP1031488, the Hewlett Foundation grant 2012-8161, Adobe, Google, and Microsoft.

Appendix

A Algorithm Details

Queue Based Evaluation: Termination Upon Empty Queue

One question raised by the queue-based method is why do we terminate as soon as we hit an empty stream (meaning, run out of (r, s') tuples at the current (s, a))? Especially in an episodic setting, why not just throw away the episode, and start a new episode? The reason is that this would no longer draw samples from the true environment. To see why, imagine that after starting in s_I, half the time the agent goes to s_1, otherwise it goes to s_2. The candidate algorithm $\mathbb{A}$ picks action 1 in s_1, but the sampling policy avoids it 99% of the time. In s_2, both sampling and candidate approaches pick each action with equal probability. If we run a "restart if no data" approach, it is very unlikely to ever include a transition from s_I to s_1, since there aren't many episodes going through s_1 that picked the action we chose, causing us to quickly run out of data. So the algorithm will almost always be fed s_2 after s_I, leading to the incorrect distribution of states. Our approach does not have this problem, since in this case it will stop as soon as it hits a lack of data at s_2, leading to a balanced number of samples. However, a downside of this approach is that with very large spaces, we may not be able to produce long sequences of interactions, since in these scenarios we may very quickly run into an empty queue.

Calculating M for PERS

Algorithm 4 Efficient M calculation

Input: Candidate policy π_b, Exploration policy π_e, common state space $\mathcal{S}$,
Binary transition matrix T denoting whether a nonzero transition probability from s and s' is possible a priori, maximum horizon H.
Initialize $M_s = 1.0$ for all $s \in \mathcal{S}$
for t = 1 to H **do**
    for $s \in \mathcal{S}$ **do**
        $M'_s = max_a \frac{\pi_b(s,a)}{\pi_e(s,a)} max_{s'} T(s,a,s') M_{s'}$
    $M = M'$
return M_{s_I}

Let $\pi_b(e)$ be shorthand for $\Pi_{t=0}^{l(e)} \pi_b(a_t|h_t)$, and similarly for $\pi_e(e)$

In order to ensure that rejection sampling returns a sample from the candidate distribution, it is critical that M be set correctly. One way to understand the necessity of M is to consider that since the ratio $\frac{\pi_b(e)}{\pi_e(e)}$ can grow extremely large, we need an M such that $\frac{\pi_b(e)}{M \pi_e(e)}$ is a probability between 0 and 1. Therefore, $M \geq \max_e \frac{\pi_b(e)}{\pi_e(e)}$ Obviously, taking the maximum over all possible episodes is a bit worrisome from a computational standpoint, although in some domains it may be feasible. Alternatively, one can always use an overestimate. Some examples that may be useful are: $\max_T \frac{\max_{es.t.l(e)=T} \pi_b(e)}{\min_{es.st.l(e)=T} \pi_e(e)}$ or $\frac{\max_e \pi_b(e)}{\min_e \pi_e(e)}$ or even $\frac{1.0}{\min_e \pi_e(e)}$. However, one needs to be careful that the overestimate is not too extreme, or else rejection sampling will accept very few samples, since an overly large M lowers all probabilities (for example, doubling M means all probabilities will be normalized to [0,0.5]).

In certain cases, even if we are not willing to assume a state space for the purposes of evaluation, we know that π_b uses a discrete state space $\mathcal{S}_1$ and π_e uses a discrete state space $\mathcal{S}_2$. In this case we can formulate a common state space, either as a cross product of the two spaces or by observing that one is contained within the other, which allows us to avoid an exponential enumeration of histories as follows. Assume $\mathcal{A}$ does make use of additional internal randomness over the course of a single episode, and the horizon is upper bounded, and we have access to a binary transition matrix T denoting whether a nonzero transition probability from s and s' is possible a priori. Then the maximum from an episode starting at each state M_s, satisfies the following recursion:

$$M_{s,0} = 1.0$$

$$M_{s,t} = max_a \frac{\pi_b(s,a)}{\pi_e(s,a)} max_{s'} T(s,a,s') M_{s',t-1}$$

One can use a dynamic programming approach to efficiently solve this recurrence. See algorithm 4.

PSRS: Self-Idempotent or Not?

If we randomize the data in our streams and use PSRS, scenarios such as the following can occur. Imagine we have three states, s_1, s_2, and s_3, and a candidate policy that is identical to the sampling policy, so that rejection sampling accepts samples with probability 1. In our initial dataset, assume we took 1 transition from s_1 to s_2, N transitions from s_2 to s_2, and one transition from s_2 to s_3. If we keep the order fixed, the trace will accept every transition in the order they occurred and thus behave exactly the same as the sampling policy, accepting all data. But, if we randomize the order, as will happen in general, we are very likely to spend less than N transitions in s_2 before we draw the tuple from $Q[s_2]$ that causes us to transition to s_3, leaving the remaining samples at s_2 uncollected.

B Sample from True Distribution Guarantees

Properties of Rejection Sampling

Rejection sampling is guaranteed to return a sample from $p(x)$ given an infinite stream of samples from $q(x)$. However, in practice we only have a finite stream of size N and would like to know whether, conditioned on the fact that rejection sampling outputs an estimate before consuming the dataset, the accepted sample is distributed according to $p(x)$. Or formally,

Lemma 1. *$P_R(x = a | r_1 \vee \cdots \vee r_N) = P_E(a)$, where P_R denotes the probability (or pdf) under rejection sampling, P_E denotes the probability under the candidate distribution, and r_i means all samples before the i^{th} are rejected, and x is accepted on the i^{th} sample.*

Proof. Proof by induction on N. The base case (N=0) is trivial because we never return an estimate given zero data. Assume $P_R(x = a | r_1 \vee \cdots \vee r_{N-1}) = P_E(x = a)$ and show for N.

$$\begin{aligned} P_R(x = a | r_1 \vee \ldots \vee r_N) \\ &= P_R(r_N) P_R(x = a | r_N) \\ &+ (1 - P_R(r_N)) P_R(x = a | r_1 \vee \cdots \vee r_{N-1}) \end{aligned}$$

By the inductive hypothesis we have:

$$\begin{aligned} P_R(x = a | r_1 \vee \cdots \vee r_N) \\ &= P_R(r_N) P_R(x = a | r_N) + (1 - P_R(r_N)) P_E(x = a) \end{aligned}$$

So it suffices to show $P_R(x = a | r_N) = P_E(x = a)$.

$$P_R(x = a | r_N) = \frac{P_R(x = a, r_N)}{P_R(r_N)}$$

$$P_R(x = a | r_N) = \frac{P_R(x = a, r_N)}{\sum_b P_R(x = b, r_N)}$$

Since we perform rejection sampling:

$$= \frac{q(a) \frac{P_E(a)}{M q(a)}}{\sum_b q(b) \frac{P_E(b)}{M q(b)}}$$

where q is the sampling distribution.

$$= \frac{\frac{P_E(a)}{M}}{\sum_b \frac{P_E(b)}{M}}$$

$$= \frac{\frac{P_E(a)}{M}}{\frac{1}{M}}$$

$$= \frac{M P_E(a)}{M}$$

$$= P_E(a)$$

□

A note on the base case

Note that our guarantees say that, if our evaluators do not terminate, our the produced pair of tuple (or episode) and χ comes from the correct distribution given some history. However, these guarantees does not explicitly address how the initial history H_0 is chosen. The following lemma (with trivial proof) addresses this issue explicitly:

Lemma 2. *Under evaluators Queue, PSRS, and PERS, $P_{\mathcal{R}}(H_0) = P_{\mathcal{E}}(H_0)$.*

Proof. For all of our evaluators, the initial state is correctly initialized to s_I, and the initial χ is drawn correctly according to $\mathbb{A}$, so the initial history H_0 (consisting of initial state and initial χ) is drawn from the correct distribution. □

Queue-based evaluator guarantees

Theorem 4.1. *Assuming the environment is an MDP with state space $\mathcal{S}$ and the randomness involved in drawing from π_b is treated as internal to $\mathbb{A}$, given any history of interactions H_T, if the queue-based evaluator produces a (s, a, r, s') tuple, the distribution of this tuple and subsequent internal randomness χ under the queue-based evaluator is identical to the true distribution the agent would have encountered if it was run online. That is, $P_{\mathcal{R}}(s, a, r, s', \chi | H_T, g_T) = P_{\mathcal{E}}(s, a, r, s', \chi | H_T)$, which gives us that $P_{\mathcal{R}}(H_{T+1} | H_T, g_T) = P_{\mathcal{E}}(H_{T+1} | H_T)$.*

Proof. Recall that for the queue-based evaluator we treat the randomness involved in drawing an action from the distribution π_b as part of the χ stored in the history. Therefore, given H_T (which includes χ), $\mathcal{A}$ deterministically selects a_T given H_T. Given an MDP and the history of interactions H_T at timestep t, and assuming for convenience $s_{-1} = s_I$, $P_{\mathcal{E}}(s, a, r, s' | H_T) = \mathbb{I}(s = s'_{t-1}, a = a_t) P_{\mathcal{E}}(r, s' | s, a)$, where the conditional independences follow from the fact that in an MDP, the distribution of (r, s') only depends on s, a. Under our evaluator, $P_{\mathcal{R}}(s, a, r, s' | H_T, g_T) = \mathbb{I}(s = s'_{t-1}, a = a_T) P_{\mathcal{R}}(Q[s, a].pop() = (r, s'))$, since the state is properly translated from one step to the next, and the action a is fixed to a_T. So we just need to show $P_{\mathcal{R}}(Q[s, a].pop() = (r, s')) = P_{\mathcal{E}}(r, s' | s, a)$. But since the (r, s') tuple at the front of our $Q[s, a]$ was drawn from the true distribution[9] given s, a, but independent of the samples in our history, it follows immediately that $P_{\mathcal{R}}(Q[s, a].pop() = (r, s')) = P_{\mathcal{E}}(r, s' | s, a)$, and thus $P_{\mathcal{R}}(s, a, r, s' | H_T, g_T) = P_{\mathcal{E}}(s, a, r, s' | H_T)$. Given an (s, a, r, s'), the algorithm's internal randomness χ is drawn from the correct distribution and thus $P_{\mathcal{R}}(s, a, r, s', \chi | H_T, g_T) = P_{\mathcal{E}}(s, a, r, s', \chi | H_T)$. □

[9] For further discussion of this property of queues see Joulani, Gyorgy, and Szepesvari 2013

Per-state RS evaluator Guarantees

Theorem 5.1. *Assume the environment is an MDP with state space $\mathcal{S}$, π_e is known, and and for all a, $\pi_e(a) > 0$ if $\pi_b(a) > 0$. Then if the evaluator produces a (s, a, r, s') tuple, the distribution of (s, a, r, s') tuple returned by PSRS (and subsequent internal randomness χ) given any history of interactions H_T is identical to the true distribution the agent would have encountered if was run online. Precisely, in the case that we accept a tuple, $P_{\mathcal{R}}(s, a, r, s', \chi|H_T) = P_{\mathcal{E}}(s, a, r, s', \chi|H_T)$, which gives us that $P_{\mathcal{R}}(H_{T+1}|H_T) = P_{\mathcal{E}}(H_{T+1}|H_T)$.*

Proof. Note that the candidate distribution π_b is deterministically output given H_T, since H_T includes any internal randomness χ by definition.

Given some H_T at timestep T, and assuming for convenience s_{-1} is a fixed start state, we know $s = s'_{t-1}$. So, we accept each (a, r, s') tuple with probability:

$$\frac{\pi_b(a|s)}{M * \pi_e(a|s)}$$

where $M = \max_a \frac{\pi_b(a|s)}{\pi_e(a|s)}$ always. Equivalently, we can say we accept each tuple with probability:

$$\frac{\pi_b(a|s)P_{\mathcal{E}}(r, s'|s, a)}{M * \pi_e(a|s)P_{\mathcal{E}}(r, s'|s, a)}$$

Let $P_{explore}$ denote the probability in the true environment under the sampling policy. Then we have the probability of accepting this tuple is:

$$\frac{P_{\mathcal{E}}(a, r, s'|s, H_T)}{M * P_{explore}(a, r, s'|s)}$$

where the conditioning on H_T is introduced because the specific π_b chosen depends on H_T. We can write $M = \max_a \frac{\pi_b(a|s)}{\pi_e(a|s)} = \max_{a,r,s'} \frac{\pi_b(a|s)}{\pi_e(a|s)} = \max_{a,r,s'} \frac{\pi_b(a|s)P_{\mathcal{E}}(r,s'|s,a)}{\pi_e(a|s)P_{\mathcal{E}}(r,s'|s,a)} = \max_{a,r,s'} \frac{P_{\mathcal{E}}(a,r,s'|s,H_T)}{P_{explore}(a,r,s'|s)}$. The (a, r, s') tuples in each PSRS stream Q[s] are drawn according to $P_{explore}(a, r, s'|s)$, since the actions in each stream are drawn according to $\pi_e(a|s)$ and r and s' are then drawn according to $P_{\mathcal{E}}(r, s'|s, a)$ by the Markov assumption. Therefore, since we draw (a, r, s') according to $P_{explore}(a, r, s'|s)$ but we wish to draw from $P_{\mathcal{E}}(a, r, s'|s, H_T)$, this is a straightforward application of rejection sampling. Since $M \geq \max_{a,r,s'} \frac{P_{\mathcal{E}}(a,r,s'|s,H_T)}{P_{explore}(a,r,s'|s)}$, rejection sampling guarantees that a returned sample is sampled according to $P_{\mathcal{E}}(a, r, s'|s, H_T)$, even if conditioned on only having a finite dataset (Lemma 1). In other words, $P_{\mathcal{R}}(a, r, s'|s, H_T, g_T) = P_{\mathcal{E}}(a, r, s'|s, H_T)$. Since in both cases s is deterministically extracted from the last tuple of H_T, this implies $P_{\mathcal{R}}(s, a, r, s'|H_T, g_T) = P_{\mathcal{E}}(s, a, r, s'|H_T)$

Given an (s, a, r, s'), the algorithm's internal randomness χ is drawn from the correct distribution and thus $P_{\mathcal{R}}(s, a, r, s', \chi|H_T, g_T) = P_{\mathcal{E}}(s, a, r, s', \chi|H_T)$. □

Per-Episode Rejection Sampling Guarantees

Theorem 6.1. *Assuming π_e is known, and $\pi_b(e) > 0 \rightarrow \pi_e(e) > 0$ for all possible episodes e and all π_b, and PERS outputs an episode e, then the distribution of e (and subsequent internal randomness χ) given any history of episodic interactions H_T using PERS is identical to the true distribution the agent would have encountered if it was run online. That is, $P_{\mathcal{E}}(e, \chi|H_T) = P_{\mathcal{R}}(e, \chi|H_T, g_T)$, which gives us that $P_{\mathcal{R}}(H_{T+1}|H_T, g_T) = P_{\mathcal{E}}(H_{T+1}|H_T)$.*

Proof. An episode e consists of some sequence of actions, observations and rewards $o_0, a_0, r_0, o_1, a_1, r_1, \ldots$. Let χ_t denote the internal randomness generated by algorithm $\mathbb{A}$ after receiving r_t and o_{t+1}, so that $\chi = \chi_0, \ldots, \chi_{l(e)}$. Let h_t denote the within-episode history at time t, namely $h_t = o_0, \ldots, o_t, a_0, \ldots, a_{t-1}, r_0, \ldots, r_{t-1}, \chi_0, \ldots, \chi_{t-1}$. Given H_T, for notational convenience we assume the algorithm $\mathbb{A}$ outputs a single policy π_b which maps h_t (recall this includes any internal randomness) to action probabilities, that is $\mathbb{A}$ chooses actions according to $\pi_b(a_t|h_t)$. The sampling policy likewise chooses actions according to $\pi_e(a_t|h_t)$ (however the χ component of h_t is not used by the sampling policy). The environment generates observations and rewards from some unknown distribution $P_{\mathcal{E}}$ given the past history, in other words according to $P_{\mathcal{E}}(r_t, o_{t+1}|a_t, h_t)$.[10] We accept episodes with probability:

$$\frac{\prod_{t=0}^{l(e)} \pi_b(a_t|h_t)}{M \prod_{t=0}^{l(e)} \pi_b(a_t|h_t)}$$

where $M \geq \frac{\prod_{t=0}^{l(e)} \pi_b(a_t|h_t)}{\prod_{t=0}^{l(e)} \pi_e(a_t|h_t)}$ always. This can also be written as:

$$\frac{\prod_{t=0}^{l(e)} \pi_b(a_t|h_t)P_{\mathcal{E}}(r_t, o_{t+1}|a_t, h_t)P_E(\chi_t|h_t, r_t, o_{t+1}, H_T)}{M \prod_{t=0}^{l(e)} \pi_e(a_t|h_t)P_{\mathcal{E}}(r_t, o_{t+1}|a_t, h_t)P_E(\chi_t|h_t, r_t, o_{t+1}, H_T)} \tag{1}$$

Let $P_{explore}(e)$ denote the probability of episode e under the exploration policy π_e in the true environment. Then we have the probability of accepting this episode is:

$$\frac{P_{\mathcal{E}}(e, \chi|H_T)}{M * P_{explore}(e) \prod_{t=0}^{l(e)} P_E(\chi_t|h_t, r_t, o_{t+1}, H_T)} \tag{2}$$

where the conditioning on H_T in the numerator is introduced because how $\mathbb{A}$ updates π_b depends on H_T. We can write $M \geq \max_e \frac{\prod_{t=0}^{l(e)} \pi_b(a_t|h_t)P_{\mathcal{E}}(r_t,o_{t+1}|a_t,h_t)P_E(\chi_t|h_t,r_t,o_{t+1},H_T)}{\prod_{t=0}^{l(e)} \pi_e(a_t|h_t)P_{\mathcal{E}}(r_t,o_{t+1}|a_t,h_t)P_E(\chi_t|h_t,r_t,o_{t+1},H_T)} = \max_e \frac{P_{\mathcal{E}}(e,\chi|H_T)}{P_{explore}(e)\prod_{t=0}^{l(e)} P_E(\chi_t|h_t,r_t,o_{t+1},H_T)}$. The episodes e in our dataset are drawn according to $P_{explore}(a, r, s'|s)$, since the action at each timestep is drawn according to $\pi_e(a_t|h_t)$, and then the next reward and observation are drawn according to $P_{\mathcal{E}}(r_t, o_{t+1}|a_t, h_t)$ by the episodic assumption. And during the operation of PERS $\mathbb{A}$ draws the internal randomness χ according to the correct distribution at each step, $P_E(\chi_t|h_t, r_t, o_{t+1}, H_T)$. Since we draw e, χ according to $P_{explore}(e) \prod_{t=0}^{l(e)} P_E(\chi_t|h_t, r_t, o_{t+1}, H_T)$, but wish to draw from $P_{\mathcal{E}}(e, \chi)$, this is a straightforward application of rejection sampling. Since $M \geq \max_e \frac{P_{\mathcal{E}}(e,\chi|H_T)}{P_{explore}(e)\prod_{t=0}^{l(e)} P_E(\chi_t|h_t,r_t,o_{t+1},H_T)}$, rejection sampling guarantees that any returned sample is sampled according to $P_{\mathcal{E}}(e, \chi|H_T)$, even if conditioned on only having a finite dataset (Lemma 1). So $P_{\mathcal{R}}(e, \chi|H_T, g_T) = P_{\mathcal{E}}(e, \chi|H_T)$. □

C Empirical Performance of Importance-Weighted Fixed-M PERS

Despite the stronger guarantees, one should be cautious about interpreting the empirical results generated by the variant proposed in Theorem 7.2. Although the expectation is correct, for a single run of the algorithm, the estimates will tend to rise above their true

[10] We assume whether the episode continues or not is an observation, i.e. is it also drawn from an unknown distribution given the history.

value due to the importance weights (leading one to believe the algorithm is learning more than it truly is) before abruptly dropping to zero. Averaging together multiple unbiased estimates is more likely to give a better picture of behavior.

D Unbiasedness Proofs

Theorem 7.1. *If M is held fixed throughout the operation of the per-episode rejection sampling replayer, π_e is known, and $\pi_b(e) > 0 \rightarrow \pi_e(e) > 0$ for all possible episodes e and all π_b, then if the evaluator outputs an estimate of some function $f(H_T)$ at episode T, that estimate is an unbiased estimator of $f(H_T)$ at episode T, in other words, $\mathbb{E}_{\mathcal{R}}[f(H_T)|g_T, \ldots, g_1] = \sum_{H_T} f(H_T) P_{\mathcal{E}}(H_T) = \mathbb{E}_{\mathcal{E}}[f(H_T)]$. For example, if $f(H_T) = R^{\mathbb{A}}(T)$, the estimate is an unbiased estimator of $R^{\mathbb{A}}(T)$ given $g_T, \ldots, g_1$.*

Proof Sketch. We first show that if M is fixed, the probability that each episode is accepted is constant ($1/M$). This allows us to show that whether we continue or not (g_T) is conditionally independent of H_{T-1}. This lets us remove the conditioning on H_{T-1} in Theorem 6.1 to give us that $P_{\mathcal{R}}(H_T|g_T) = P_{\mathcal{E}}(H_T)$, meaning the distribution over histories after T accepted episodes is correct, from which conditional unbiasedness is easily shown.

Proof. First, we will calculate the probability (over randomization in the dataset, algorithm, and evaluator) that we accept the i^{th} episode in our dataset, e_i, given some H_T (over some sequence $\mathbb{S}$ of acceptances/rejections in the first $i-1$ episodes).

$$P(\text{Accept } e_i|H_T, \mathbb{S}) = \sum_e P(e_i = e) P(e \text{ is accepted } |H_T, \mathbb{S}) \tag{3}$$

If we let $q(e)$ refer to the probability of the episode under the sampling distribution,

$$P(\text{Accept } e_i|H_T, \mathbb{S}) = \sum_e q(e) P(e \text{ is accepted } |H_T, \mathbb{S}) \tag{4}$$

Recall that as part of Theorem 6.1 (equation (1)) we showed the per-episode rejection replayer accepted an episode with probability

$$\frac{\prod_{t=0}^{l(e)} \pi_b(a_t|h_t) P_{\mathcal{E}}(r_t, o_{t+1}|a_t, h_t) P_E(\chi_t|h_t, r_t, o_{t+1}, H_T)}{M \prod_{t=0}^{l(e)} \pi_e(a_t|h_t) P_{\mathcal{E}}(r_t, o_{t+1}|a_t, h_t) P_E(\chi_t|h_t, r_t, o_{t+1}, H_T)}$$

$$= \frac{\prod_{t=0}^{l(e)} \pi_b(a_t|h_t) P_{\mathcal{E}}(r_t, o_{t+1}|a_t, h_t)}{M \prod_{t=0}^{l(e)} \pi_e(a_t|h_t) P_{\mathcal{E}}(r_t, o_{t+1}|a_t, h_t)}$$

$$= \frac{P_{\mathcal{E}}(e|H_T)}{M q(e)}$$

where $P_{\mathcal{E}}(e|H_T)$ refers to the probability of episode e under the $\mathbb{A}$ given H_T, and M denotes the normalizer, which is constant in this variant of the algorithm. So we have:

$$P(\text{Accept } e_i|H_T, \mathbb{S}) = \sum_e P(e \text{ sampled } |H_T, \mathbb{S}) \frac{P_{\mathcal{E}}(e|H_T)}{M q(e)} \tag{5}$$

$$P(\text{Accept } e_i|H_T, \mathbb{S}) = \sum_e q(e) \frac{P_{\mathcal{E}}(e|H_T)}{M q(e)} \tag{6}$$

$$P(\text{Accept } e_i|H_T, \mathbb{S}) = \frac{1}{M} \sum_e P_{\mathcal{E}}(e|H_T) \tag{7}$$

And since $\sum_e P_{\mathcal{E}}(e|H_T) = 1$,

$$P(\text{Accept } e_i|H_T, \mathbb{S}) = \frac{1}{M} \tag{8}$$

$P(\text{Accept } e_i|H_T, \mathbb{S})$ is a constant it is independent of H_T and $\mathbb{S}$, so:

$$P(\text{Accept } e_i|H_T, \mathbb{S}) = P(\text{Accept } e_i) = \frac{1}{M} \tag{9}$$

So we accept the i^{th} episode with probability $1/M$, where M is a constant, and therefore independent of $\mathbb{S}$ and H_T. We will now proceed to show that since this is a constant, it does not cause any histories to be more likely than others, from which the unbiased property will follow.

Next, we will prove by induction that $P_{\mathcal{R}}(H_T|g_T, \ldots, g_0) = P_{\mathcal{E}}(H_T)$. Recall that H_T denotes a trajectory of T episodic interactions, g_T denotes the event that we continue[11] (i.e. do not terminate evaluation) from time $T-1$ to time T, $P_{\mathcal{R}}(x)$ denotes the probability of x under the replayer, and $P_{\mathcal{E}}(x)$ denotes the probability of x in the true (online) environment under the candidate policy. The base case is trivial (Since g_0 always occurs, $P_{\mathcal{R}}(H_0|g_0) = P_{\mathcal{R}}(H_0)$, and by Lemma 2, $P_{\mathcal{R}}(H_0) = P_{\mathcal{E}}(H_0)$). We now assume this holds for $T-1$ and we will show it holds for T.

$$P_{\mathcal{R}}(H_T|g_T, \ldots, g_0) = \sum_{H_{T-1}} P_{\mathcal{R}}(H_T|H_{T-1}, g_T, \ldots, g_0) P_{\mathcal{R}}(H_{T-1}|g_T, \ldots, g_0) \tag{10}$$

Our next step is to show that $g_T \perp H_{T-1}|g_{T-1}$, so that we can turn the right term of equation (10) into the inductive hypothesis ($\perp$ denotes independance). Let i be the number of episodes consumed after accepting the $(T-1)^{th}$ episode, so that there are exactly $N-i$ episodes remaining after accepting the $(T-1)^{th}$. Since, as we showed in equation (9), $P(\text{Accept } e_i) = \frac{1}{M}$, given g_{T-1} and $N-i$ episodes remaining, g_T is drawn from a $Bernoulli(1 - (1 - \frac{1}{M})^{N-i})$ distribution. So, since M and N are both constants, $g_T \perp H_{T-1}|g_{T-1}, i$. So we can write:

$$P(g_T, H_{T-1}|g_{T-1}) = \sum_i P(g_T, H_{T-1}|g_{T-1}, i) P(i|g_{T-1}) \tag{11}$$

$$= \sum_i P(g_T|g_{T-1}, i) P(H_{T-1}|g_{T-1}, i) P(i|g_{T-1}) \tag{12}$$

So we need to show $H_{T-1} \perp i|g_{T-1}$.

$$P(H_{T-1}, i|g_{T-1}) = \frac{1}{P(g_{T-1})} P(H_{T-1}, g_{T-1}, i) \tag{13}$$

Now, we know the i^{th} episode must have been accepted, but the $(T-2)^{th}$ acceptances could have occurred anywhere in the $(i-1)^{th}$ steps. The probability of the j^{th} e, χ pair in the history, $H_{T-1}[j]$ being accepted on the next episode given some sequence of previously accepted/rejected episodes $\mathbb{S}$ is

[11] Since we always generate the initial history H_0, for the purposes of induction we here condition on g_0 even though that is not strictly necessary since it always occurs.

$$
\begin{aligned}
&P(H_{T-1}[j] \text{ produced}|H_{j-1}, \mathbb{S}) \\
&= P(H_{T-1}[j] \text{ sampled}|H_{j-1}, \mathbb{S}) P(H_{T-1}[j] \text{ accepted}|H_{j-1}, \mathbb{S}) \\
&= q(H_{T-1}[j]|H_{j-1}) \frac{P_{\mathcal{E}}(H_{T-1}[j]|H_{j-1})}{M q(H_{T-1}[j]|H_{j-1})},
\end{aligned} \tag{14}
$$

where $q(H_{T-1}[j]|H_{j-1})$ denotes the probabilities of producing the j^{th} e, χ tuple in the history given H_{j-1} and sampling actions according to π_e, and $P(H_{T-1}[j]$ accepted) is derived as per equation (2) in Theorem 6.1. Equation (14) can be simplified:

$$
P(H_{T-1}[j] \text{ produced}|H_{j-1}, \mathbb{S}) = \frac{P_{\mathcal{E}}(H_{T-1}[j]|H_{j-1})}{M} \tag{15}
$$

Note that this depends on H_{j-1} but is independent of $\mathbb{S}$. The probability of rejection given $\mathbb{S}, H_{j-1}$ is a constant $1 - \frac{1}{M}$ (see equation (9)), so returning to equation 13 we have:

$$
P(H_{T-1}, i|g_{T-1}) \tag{16}
$$

$$
= \frac{1}{P(g_{T-1})} \sum_{\mathbb{S} \text{ s.t. } \mathbb{S} \text{ compat } (i, T-1)} P(\mathbb{S}, H_{T-1}) \tag{17}
$$

where ($\mathbb{S}$ compat $(i, T-1)$), means $|\mathbb{S}| = i$, there are $T-1$ acceptances in $\mathbb{S}$, and $\mathbb{S}[i-1]$ is not a rejection. Now, computing $P(\mathbb{S}, H_{T-1})$ would consist of multiplying quantities like P(reject episode k | $\mathbb{S}$ up to $k-1$)
$*$ P(accept $H_{T-1}[3]$ on episode k+1 | H_2, $\mathbb{S}$ up to k) $* \ldots$. Note, however, that the probability of rejecting an episode is independent of the particular past sequence of acceptances and rejections (see equation (9)), and so is the probability of accepting the next item in the history (see equation (15)). Therefore, for every $\mathbb{S}$, $P(\mathbb{S}, H_{T-1})$ is the multiplication of the probability of the initial history $P(H_0)$ with $i-T+1$ rejection probabilities together with the probabilities of accepting the $T-1$ items in the history. There are $\binom{i-1}{T-2}$ possible different sequences $\mathbb{S}$ such that ($\mathbb{S}$ compat $(i, T-1)$), since the last element is always accepted. Therefore:

$$
\begin{aligned}
&P(H_{T-1}, i|g_{T-1}) \\
&= \frac{P(H_0)}{P(g_{T-1})} \binom{i-1}{T-2} (1 - \frac{1}{M})^{i-T+1} \prod_{j=1}^{T-1} \frac{P_{\mathcal{E}}(H_{T-1}[j]|H_{j-1})}{M}
\end{aligned} \tag{18}
$$

since the probability of rejecting the k^{th} episode is $(1 - \frac{1}{M})$ from (9), and from (15) the probability of accepting the j^{th} element in the history on the k^{th} episode is $\frac{P_{\mathcal{E}}(H_{T-1}[j]|H_{j-1})}{M}$.

By Lemma 2, $P_{\mathcal{R}}(H_0) = P_{\mathcal{E}}(H_0)$, so:

$$
\begin{aligned}
&P(H_{T-1}, i|g_{T-1}) \\
&= \frac{P_{\mathcal{E}}(H_0)}{P(g_{T-1})} \binom{i-1}{T-2} (1 - \frac{1}{M})^{i-T+1} \prod_{j=1}^{T-1} \frac{P_{\mathcal{E}}(H_{T-1}[j]|H_{j-1})}{M}
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
&= (P_{\mathcal{E}}(H_0) \prod_{j=1}^{T-1} P_{\mathcal{E}}(H_{T-1}[j]|H_{j-1})) \frac{1}{P(g_{T-1})} \\
&\quad * \binom{i-1}{T-2} (1 - \frac{1}{M})^{i-T+1} \prod_{j=1}^{T-1} \frac{1}{M}
\end{aligned} \tag{20}
$$

Where (20) follows by simply reordering the multiplication. Now by definition, $P_{\mathcal{E}}(H_{T-1}) = P_{\mathcal{E}}(H_0) \prod_{j=1}^{T-1} P_{\mathcal{E}}(H_{T-1}[j]|H_{j-1})$, so we have:

$$
\begin{aligned}
&P(H_{T-1}, i|g_{T-1}) \\
&= P_{\mathcal{E}}(H_{T-1}) \frac{1}{P(g_{T-1})} \binom{i-1}{T-2} (1 - \frac{1}{M})^{i-T+1} \prod_{j=1}^{T-1} \frac{1}{M}
\end{aligned} \tag{21}
$$

Now by induction, $P_{\mathcal{E}}(H_{T-1}) = P_{\mathcal{R}}(H_{T-1}|g_{T-1})$, so:

$$
\begin{aligned}
&P(H_{T-1}, i|g_{T-1}) \\
&= P(H_{T-1}|g_{T-1}) \frac{1}{P(g_{T-1})} \binom{i-1}{T-2} (1 - \frac{1}{M})^{i-T+1} \prod_{j=1}^{T-1} \frac{1}{M}
\end{aligned} \tag{22}
$$

Note that $\binom{i-1}{T-2} (1 - \frac{1}{M})^{i-T+1} \prod_{j=1}^{T-1} \frac{1}{M}$ is just the probability of consuming i elements and accepting exactly $T-1$ of them (namely the binomial formula, where the probability of success (acceptance) is $\frac{1}{M}$), so:

$$
P(H_{T-1}, i|g_{T-1}) = P(H_{T-1}|g_{T-1}) \frac{P(i, g_{T-1})}{P(g_{T-1})} \tag{23}
$$

By the definition of conditional probability:

$$
P(H_{T-1}, i|g_{T-1}) = P(H_{T-1}|g_{T-1}) P(i|g_{T-1}) \tag{24}
$$

So since $P(H_{T-1}, i|g_{T-1}) = P(H_{T-1}|g_{T-1}) P(i|g_{T-1})$, $H_{T-1} \perp i|g_{T-1}$. Now returning to equation (12) we have:

$$
\begin{aligned}
&P(g_T, H_{T-1}|g_{T-1}) \\
&= \sum_i P(g_T|g_{T-1}, i) P(H_{T-1}|g_{T-1}, i) P(i|g_{T-1})
\end{aligned} \tag{25}
$$

Since $H_{T-1} \perp i|g_{T-1}$:

$$
P(g_T, H_{T-1}|g_{T-1}) \tag{26}
$$

$$
= \sum_i P(g_T|g_{T-1}, i) P(H_{T-1}|g_{T-1}) P(i|g_{T-1}) \tag{27}
$$

$$
= P(H_{T-1}|g_{T-1}) \sum_i P(g_T|g_{T-1}, i) P(i|g_{T-1}) \tag{28}
$$

$$
= P(H_{T-1}|g_{T-1}) \sum_i \frac{P(g_T, g_{T-1}, i)}{P(i, g_{T-1})} \frac{P(i, g_{T-1})}{P(g_{T-1})} \tag{29}
$$

$$
= P(H_{T-1}|g_{T-1}) \frac{\sum_i P(g_T, g_{T-1}, i)}{P(g_{T-1})} \tag{30}
$$

$$
= P(H_{T-1}|g_{T-1}) \frac{P(g_T, g_{T-1})}{P(g_{T-1})} \tag{31}
$$

$$
= P(H_{T-1}|g_{T-1}) P(g_T|g_{T-1}) \tag{32}
$$

Where equation (31) follows since i is marginalized out, and (32) follows by the definition of conditional probability. So equation (32) tells us that $g_T \perp H_{T-1}|g_{T-1}$.

Given this conditional independence, $P_{\mathcal{R}}(H_{T-1}|g_T, g_{T-1}, \ldots, g_1) = P_{\mathcal{R}}(H_{T-1}|g_{T-1}, \ldots, g_0)$, and $P_{\mathcal{R}}(H_{T-1}|g_{T-1}, \ldots, g_0) = P_{\mathcal{E}}(H_{T-1})$ by induction. So $P_{\mathcal{R}}(H_{T-1}|g_T, \ldots, g_0) = P_{\mathcal{E}}(H_{T-1})$. Therefore, picking up from (10), we have:

$$P_{\mathcal{R}}(H_T|g_T,\ldots,g_0)$$
$$= \sum_{H_{T-1}} P_{\mathcal{R}}(H_T|H_{T-1},g_T,\ldots,g_0)P_{\mathcal{R}}(H_{T-1}|g_T,\ldots,g_0) \tag{33}$$
$$= \sum_{H_{T-1}} P_{\mathcal{R}}(H_T|H_{T-1},g_T,\ldots,g_0)P_{\mathcal{E}}(H_{T-1}) \tag{34}$$

Now observe that, as we showed in Theorem 6.1, given that we update, the update is drawn from the true distribution given the current history, namely[12] $P_{\mathcal{R}}(H_T|H_{T-1},g_T,\ldots,g_0) = P_{\mathcal{E}}(H_T|H_{T-1})$. Plugging this in we have:

$$P_{\mathcal{R}}(H_T|g_T,\ldots,g_0) = \sum_{H_{T-1}} P_{\mathcal{E}}(H_T|H_{T-1})P_{\mathcal{E}}(H_{T-1}) \tag{35}$$
$$P_{\mathcal{R}}(H_T|g_T,\ldots,g_0) = \sum_{H_{T-1}} P_{\mathcal{E}}(H_T,H_{T-1}) \tag{36}$$

Marginalizing out H_{T-1}:

$$P_{\mathcal{R}}(H_T|g_T,\ldots,g_0) = P_{\mathcal{E}}(H_T) \tag{37}$$

Which completes the induction. All that remains is to show that the expectation is correct. For any function f over histories of interactions (such as cumulative reward, final reward, etc.),

$$\mathbb{E}_{\mathcal{R}}[f(H_T)|g_T,\ldots,g_1] = \sum_{H_T} f(H_T)P_{\mathcal{R}}(H_T|g_T,\ldots,g_1) \tag{38}$$

Since g_0 always occurs we can freely condition on it:

$$\mathbb{E}_{\mathcal{R}}[f(H_T)|g_T,\ldots,g_1] = \sum_{H_T} f(H_T)P_{\mathcal{R}}(H_T|g_T,\ldots,g_0) \tag{39}$$

By (37):

$$\mathbb{E}_{\mathcal{R}}[f(H_T)|g_T,\ldots,g_1] = \sum_{H_T} f(H_T)P_{\mathcal{E}}(H_T) \tag{40}$$
$$= \mathbb{E}_{\mathcal{E}}[f(H_T)] \tag{41}$$

Therefore the estimate is unbiased. □

Theorem 7.2. *Assuming for each T, $R^{\mathbb{A}}(T)$ is divided by by $\phi = 1 - Binomial(N,1/M).cdf(T-1)$, and after terminating after k episodes are produced we output 0 as estimates of reward for episodes $k+1,\ldots,N$, and M is held fixed throughout the operation of the per-episode rejection sampling replayer, and π_e is known, and $\pi_b(e) > 0 \rightarrow \pi_e(e) > 0$ for all possible episodes e and all π_b, then the estimate of reward output at each episode $T = 1\ldots N$ is an unbiased estimator of $R^{\mathbb{A}}(T)$.*

Proof. The expectation of reward at episode T can be written as:

$$\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(T)] = P(g_T,...,g_1)\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(T)|g_T,...,g_1] + P(\neg(g_T,...,g_1))\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(T)|\neg(g_T,...,g_1)] \tag{42}$$

[12]Recall that g_T implies $g_{T-1},\ldots,g_0$.

where, as above, $g_T,...,g_1$ denotes the probability of not terminating before episode T. If we do not reach epsiode T (the second case), modified PERS outputs 0, so:

$$\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(T)] = P(g_T,...,g_1)\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(T)|g_T,...,g_1] \tag{43}$$

Now, in the remaining case (that we reach T), we divide the original value by ϕ. In theorem 7.1 we showed that the expectation of the the unweighted estimates conditioned on reaching T was unbiased, giving us:

$$\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(T)] = P(g_T,...,g_1)\frac{\mathbb{E}_{\mathcal{E}}[R^{\mathbb{A}}(T)]}{\phi} \tag{44}$$

Now, since the probability of accepting each episode i is $1/M$ (see Theorem 7.1, equation (9)) and there are N total episodes, the probability that we reach episode T (aka $P(g_T,...,g_1)$) is $P(Binomial(N,1/M) \geq T) = 1 - Binomial(N,1/M).cdf(T-1) = \phi$. So:

$$\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(T)] = \phi\frac{\mathbb{E}_{\mathcal{E}}[R^{\mathbb{A}}(T)]}{\phi} \tag{45}$$
$$\mathbb{E}_{\mathcal{R}}[R^{\mathbb{A}}(T)] = \mathbb{E}_{\mathcal{E}}[R^{\mathbb{A}}(T)] \tag{46}$$

□

E Experiments

PSRL PSRL leaves open the choice of prior. Additionally, in many cases, one has additional knowledge that can help speed learning (for example one may know there is only a certain set of discrete reward possibilities, or one may know certain transitions are impossible). A convenient way to incorporate this kind of knowledge, and also simplify the choice of prior, is the relative outcomes framework (Asmuth et al. 2009). In this setting, at each state action pair, one of k discrete outcomes occurs. Given a state, action and outcome, it is possible to deterministically extract the reward and next state. In theory any MDP can be encoded in this framework, but it gives us the ability to limit the space of possibilities to speed learning. Given a set of k discrete outcomes, we must clear a categorical distribution (probability) vector over outcomes. Then we can use the conjugate prior, a k-dimensional Dirichlet. Since we did not have any additional information the prior was flat ($\alpha_i = 1$ for all outcomes i).

Details of comparison to model-based The comparison shown in Figure 2a was done in the SixArms environment (Strehl and Littman 2004) (Figure 4), an environment with six actions and seven states. We defined 14 relative outcomes for use with PSRL, one for remaining put in a state, six for moving to each outer state, six for staying in the same state and receiving one of the non-zero rewards, and one for moving to the start state. We treated this a finite horizon problem, where episodes could be at most length 10. The algorithm evaluated was PSRL (Osband, Russo, and Van Roy 2013), with 10 posterior samples after each episode.

The model-based approach works by building the MLE MDP model from data, and sampling from it to generate experience. Specifically, in the relative outcome setting it need only estimate the probability of getting any outcome given a state an action, as well as the distribution over initial outcomes. These were estimated in the MLE sense, so for example if we had only 1 observed outcome of o at (s,a), 100% probability was put on o at (s,a). For state-action pairs with no data, a uniform distribution over outcomes was used.

To carry out the evaluation, we first calculated the "true" cumulative reward by averaging 1000 runs of PSRL against the true

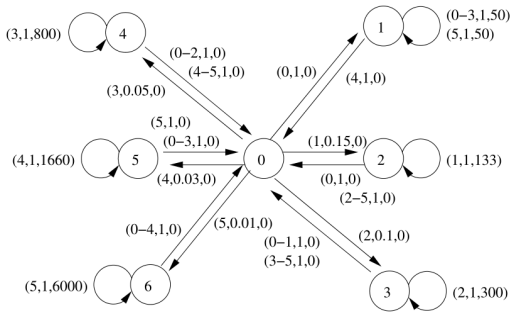


Figure 4: The SixArms environment. The labels on each edge are (a, p, r) tuples, where p is the probability of taking that transition. Image taken from Strehl et al. 2004.

environment. Then, for each evaluator (PSRS and model-based) we sampled 100 different datasets of 100 episodes from SixArms using a uniform sampling policy. We then ran 10 complete runs of the evaluator on that dataset, each of which returned estimates of cumulative reward. For PSRS, we only reported estimates up to the minimum evaluation length across those ten runs[13]. The squared error was computed between the mean of the 10 runs and the true cumulative reward curve. Finally, to compute the MSE, the average over the 100 runs was taken (ignoring cases where no estimate was returned).

Treefrog Treasure Experiment Setup Treefrog Treasure is an educational fractions game (Figure 3). The player controls a frog to navigate levels and jump through numberlines. Our ***action set*** is defined as follows: After giving an in-game pretest we give a first set of lines, then we want to either increase the difficulty level of one parameter (e.g. removing tick marks), stay, or go back, which we encoded as 11 actions. Our ***reward*** is terminal and ranges from -2 to 4. It measures engagement (measured by whether the student quit before the posttest) and learning (measured by pretest-to-posttest improvement). After each step we additionally receive a binary observation informing us whether the student took more than 4 attempts on the last pair of lines. The relative outcomes defined for PSRL included either terminating with some reward (-2 - 4) or continuing with one of the two observations. We used a **state space** consisting of the history of actions and the last observation. The true horizon is 5, but for reasons of data sparsity (induced by a large action space and thus large state space) we varied the horizon between 3 and 4 in our experiments.

Our dataset of 11,550 players was collected from BrainPOP.com, an educational website focused on school-aged children. The sampling policy used was semi-uniform and changed from day-to-day. We logged the probabilities for use with the rejection-sampling evaluator, modifying the rejection sampling approaches slightly to recalculate M at every step based on the changing sampling policy.

For PERS we use[14] Algorithm 4 to calculate M, updating it based both on the change in the algorithm and the change in the sampling policy. We also tried a variant which fixed M so as to achieve the unbiasedness guarantees in Theorem 7.1. To calculate this M we upper bounded the probability of an episode under our candidate distribution by 1, and calculated the minimum probability our sampling distribution could place on any episode.[15]

Further Treefrog Treasure Results We also examined an increased horizon of 4. Given deterministic policies on this larger state space, all three methods are more or less indistinguishable (Figure 5a) ; however, revealing more randomness causes PERS to overtake PSRS (mean 260.54 vs. 173.52), Figure 5b). As an extreme case, we also tried a random policy in Figure 5c: this large amount of revealed randomness benefits the rejection sampling methods, especially PERS, who evaluates for much longer than the other approaches. PERS outperforms PSRS here because there are small differences between the random candidate policy and the semi-random sampling policy, and thus if PSRS enters a state with little data it is likely to terminate.

Riverswim RiverSwim is described in Figure 6a. In this domain we defined a space of 5 relative outcomes: {*MovedRight, MovedLeft, Staywith0reward, Staywith5/1000reward,Staywith1reward*}. To generate histograms over 100 sizes, for each evaluator we sampled 10 datasets of 10000 episodes from this MDP, using a uniform sampling policy, and used each dataset for 10 runs of evaluation.

The results with 10-sample PSRL are shown in figure 6b. The per-episode rejection sampler was only able to accept a few episodes due to their length, while the two state-based evaluators leveraged the known state space to accept a substantially larger number of episodes. To more clearly understand sensitivity to revealed randomness, we tested two versions of a random policy: one which revealed its randomness and one which did not. In the case where the randomness was hidden (Figure 6c), we see that the per-episode rejection sampling approach does so poorly it never accepts any episodes and thus is hard to see on the graph, the per-state rejection sampling has fairly mediocre performance, and the queue-based method does the best. This is because the per-state rejection sampler, by treating the data as a stream, discards a lot of data for arms it did not choose to pull, unlike the queue-based approach which only discards data when it is consumed by the algorithm. We suspect the effect is particularly visible here for two reasons: (a) There are a small number of states, making it more similar to the bandit case where Queue does well (Mandel et al. 2015) and (b) the policy is not explicitly driving exploration towards poorly-visited states, which often causes both state-based methods to fail fairly quickly. When the randomness is revealed (Figure 6d), the two rejection sampling approaches improve greatly in performance. One can observe how the per-episode version is self-idempotent (since the sampling policy was the same as the candidate policy, all 10,000 episodes were accepted) while the per-state is not.

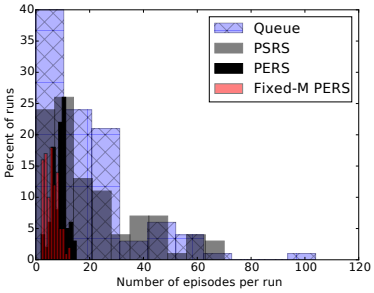
References

Asmuth, J.; Li, L.; Littman, M. L.; Nouri, A.; and Wingate, D. 2009. A bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 19–26. AUAI Press.

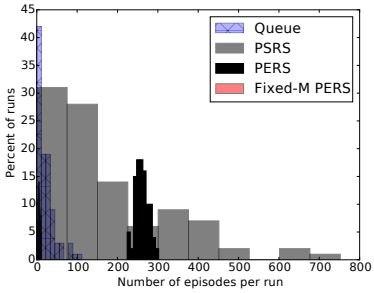
[13]This was following the approach proposed in (Mandel et al. 2015) to mitigate the bias introduced by having a wide variance in evaluation lengths, a problem we discuss in section 7.

[14]With straightforward extensions to handle policies that depend on both s and t.

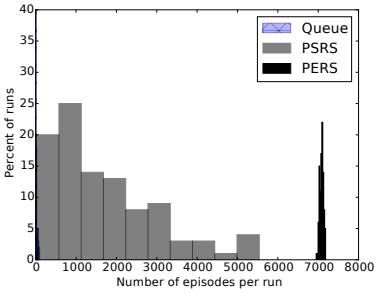
[15]This is probably fairly close to the minimum fixed M, since PSRL randomly generates distributions by sampling N times from the posterior, and we have to consider the worst case over that randomization, which always has a small probability of placing probability 1 on the minimum-probability episode under the sampling distribution.



(a) Comparing with 4 timesteps and 1 PSRL posterior sample.

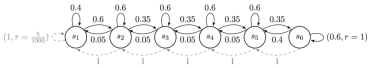


(b) Comparing with 4 timesteps and 10 PSRL posterior samples.

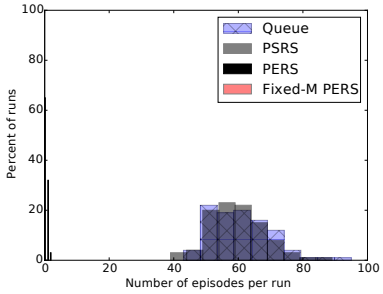


(c) Comparing a revealed, uniformly-random policy with 4 timesteps.

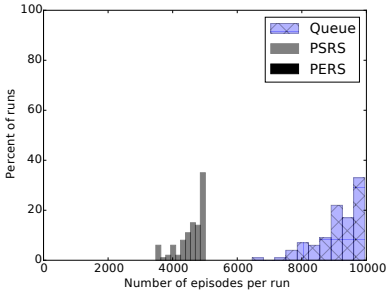
Figure 5: Further treefrog experimental results. Recall that in Treefrog treasure, more timesteps means a larger state space.



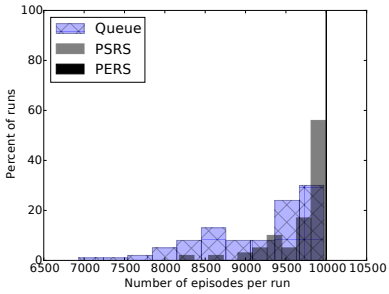
(a) RiverSwim: A 6-state MDP where one can either stay to the left for small reward, or venture nosily to the right for higher reward. (Osband et al. 2013)



(b) Comparing on RiverSwim with 10 PSRL posterior samples.



(c) Comparing on RiverSwim with a uniformly-random policy which does not reveal its randomness. Per-episode rejection sampling is invisible because it always accepts 0 samples.



(d) Comparing on RiverSwim with a uniformly-random policy which reveals its randomness. PERS is the fine line at 10,000, since it always accepts all data.

Figure 6: Evaluator data-efficiency results on RiverSwim.

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The arcade learning environment: an evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47(1):253–279.

Chou, K.-C., and Lin, H.-T. 2012. Balancing between estimated reward and uncertainty during news article recommendation for ICML 2012 exploration and exploitation challenge. In *ICML 2012 Workshop: Exploration and Exploitation*, volume 3.

Dudík, M.; Erhan, D.; Langford, J.; and Li, L. 2012. Sample-efficient nonstationary policy evaluation for contextual bandits. *UAI*.

Dudík, M.; Erhan, D.; Langford, J.; Li, L.; et al. 2014. Dou-

bly robust policy evaluation and optimization. *Statistical Science* 29(4):485–511.

Fonteneau, R.; Murphy, S.; Wehenkel, L.; and Ernst, D. 2010. Model-free monte carlo-like policy evaluation. In *Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*.

Fonteneau, R.; Murphy, S. A.; Wehenkel, L.; and Ernst, D. 2013. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research* 208(1):383–416.

Gelman, A.; Carlin, J. B.; Stern, H. S.; and Rubin, D. B. 2014. *Bayesian data analysis*, volume 2. Taylor & Francis.

Joulani, P.; Gyorgy, A.; and Szepesvari, C. 2013. Online learning under delayed feedback. In *Proceedings of The 30th International Conference on Machine Learning*, 1453–1461.

Levine, S., and Koltun, V. 2013. Guided policy search. In *Proceedings of The 30th International Conference on Machine Learning*, 1–9.

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*, 297–306. ACM.

Maei, H. R., and Sutton, R. S. 2010. GQ (λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the Third Conference on Artificial General Intelligence*, volume 1, 91–96.

Mahmood, A. R.; van Hasselt, H. P.; and Sutton, R. S. 2014. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 3014–3022.

Mandel, T.; Liu, Y.-E.; Levine, S.; Brunskill, E.; and Popović, Z. 2014. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, 1077–1084. IFAAMAS.

Mandel, T.; Liu, Y.-E.; Brunskill, E.; and Popović, Z. 2015. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. *AAAI*.

Mary, J.; Preux, P.; and Nicol, O. 2014. Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 172–180.

Osband, I.; Russo, D.; and Van Roy, B. 2013. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, 3003–3011.

Peshkin, L., and Shelton, C. R. 2002. Learning from scarce experience. *ICML*.

Precup, D. 2000. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series* 80.

Ross, S.; Gordon, G. J.; and Bagnell, J. A. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. *AISTATS*.

Strehl, A. L., and Littman, M. L. 2004. An empirical evaluation of interval estimation for Markov decision processes. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, 128–135. IEEE.

Strens, M. 2000. A Bayesian framework for reinforcement learning. In *ICML*, 943–950.

Sutton, R. S.; Mahmood, A. R.; and White, M. 2015. An emphatic approach to the problem of off-policy temporal-difference learning. *arXiv preprint arXiv:1503.04269*.

Thomas, P. S.; Theocharous, G.; and Ghavamzadeh, M. 2015. High confidence off-policy evaluation. *AAAI*.