An Adaptation of RLSVI with Explicit Action Sampling Probabilities

Ziping Xu*
zipingxu@fas.harvard.edu
Department of Statistics
Harvard University*

Iris Yan*
irisyan@college.harvard.edu
Department of Statistics
Harvard University

Susan A. Murphy
samurphy11@gmail.com
Department of Statistics
Department of Computer Science
Harvard University

Abstract

In real-world Reinforcement Learning (RL) deployment, the deployed online RL algorithms often need to collect datasets that enable offline policy evaluation for any target policy. Many offline policy evaluation approaches, use the Action Sampling Probabilities (ASPs), the conditional probabilities that the implemented RL algorithm used to select a particular action given all the previously observed states, actions and rewards. In the motivating digital health clinical trial, we originally planned to use the online Randomized Least Squares Value Iteration (RLSVI) algorithm for its robust empirical performance such settings. However RLSVI only has implicit ASPs as it utilizes external sources of randomness for exploration.

To harness RLSVI's effective exploration while providing explicit ASPs, we propose to approximate the implicit ASPs of RLSVI, and sample actions directly using these approximations during the online learning. Computing the implicit ASPs is an exact Bayesian computation problem. We address this through Monte Carlo integration with importance sampling. We call this method RLSVI-IS (Importance Sampling). We evaluate RLSVI-IS on a simulation testbed built for the mobile health clinical trial. Our results demonstrate that RLSVI-IS not only achieves cumulative rewards comparable to those of RLSVI but also provides explicit ASPs. Moreover, we propose a sufficient condition that enables rigorous control over the distance between the explicit ASPs for RLSVI-IS and the implicit ASPs for RLSVI.

1 Introduction

Recent advancements in online Reinforcement Learning (RL) have emphasized the critical need for after-study analyses, particularly when RL algorithms are deployed in fields like healthcare. In such applications, RL can be considered a form of treatment, and RL is implemented during a clinical trial, where the goal is to evaluate the treatment effects. Specifically, the treatment effect in an episodic RL environment is the difference between the value of the deployed RL algorithm and a baseline policy. This requires running offline policy evaluation (OPE) using datasets collected from the online RL implementation.

An importance sampling (IS) estimator, which requires an explicit form of action sampling probability (ASP), is used in almost all OPE approaches. To see this, let us consider an episodic MDP with state space $\mathcal{S}$, action space $\mathcal{A}$ and horizon H. The ASP of an online RL algorithm can be formally defined as a deterministic mapping π_b. This mapping takes current observations—comprising tuples $(s, a, s', r) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R}$– and outputs a distribution over the action space $\mathcal{A}$. A typical IS

*The first two authors have equal contributions.

estimator (Jiang & Li, 2016) of the value of a target Markovian policy π is represented by

$$V_{\text{IS}} = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\prod_{h=1}^{H}\frac{\pi(a_{i,h} \mid s_{i,h})}{\pi_b(a_{i,h} \mid \mathcal{D}_{i,h})}\right)\left(\sum_{h=1}^{H} r_{i,h}\right)\right], \tag{1}$$

where $\mathcal{D}_{i,h} = \{(s_{j,h'}, a_{j,h'}, s_{j,h'+1}, r_{j,h'})_{h'=1}^{H}\}_{j=1}^{i-1} \cup \{(s_{i,h'}, a_{i,h'}, s_{i,h'+1}, r_{i,h'})_{h'=1}^{h-1}\}$ is all the observations up to the h-th step in episode i, and N is the total number of episodes. In (1), ASPs are explicitly used in IS estimators.

Many online learning algorithms utilize external sources of randomness like posterior sampling algorithms. These algorithms are shown to have better empirical performance in real-world applications and they are commonly used in mobile health studies (Tomkins et al., 2021; Trella et al., 2022). However, algorithms with external randomness often do not provide an explicit form of ASPs. We take an example of RLSVI (Randomized Least Square Value Iteration) (Osband et al., 2016), a common choice (Li et al., 2023) for episodic RL. RLSVI samples parameters $\theta_{i,h}$ about the underlying optimal Q-value function Q_h^* from their posterior distribution at the start of episode i, which is used to make decisions throughout the episode. These randomness in sampling $\theta_{i,h}$'s is introduced for exploration, which is shown to help achieve minimax optimal regret guarantee (Agrawal et al., 2021). However, these random $\theta_{i,h}$'s hinder an explicit form of ASPs, as one will have to integrate over $\theta_{i,h}$'s according to its posterior. The posterior distribution given $\mathcal{D}_{i,h}$ do not align with the distribution RLSVI samples $\theta_{i,h}$ as shown in Figure 1. Thus, this integration, as an exact Bayesian computation, is computationally infeasible for real-time applications.

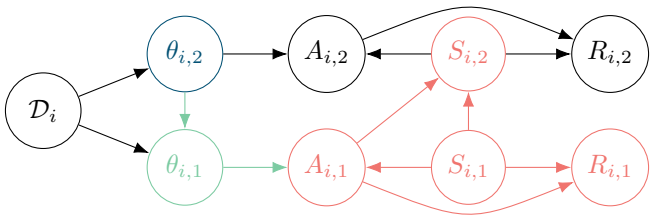


Figure 1: Direct acyclic graph (DAG) of one episode of RLSVI with horizon $H = 2$. The DAG shows the conditional dependence between $\theta_{i,2}$ (blue) and $(S_{i,1}, A_{i,1}, S_{i,2}, R_{i,1})$ (red) through $\theta_{i,1}$ (green).

There are two directions in tackling the issue of no ASPs. First, one may simply run RLSVI and approximate the ASPs based on the collected dataset after the online learning is finished. Second, one may run an alternative online algorithm that provides explicit ASPs close to the implicit ASPs of RLSVI at the each step, and directly samples actions based on these explicit ASPs. Compared to the second choice, the first choice potentially results in better regret guarantees during the implementation but can result in after-implementation policy evaluation with greater bias/uncertainty. The second choice, though leads to potentially higher regrets, guarantees the explicit ASPs, making it the primary choice of this paper.

Our contribution. We introduce RLSVI-IS, a novel algorithm with explicit ASPs unlike RLSVI. The ASPs of RLSVI-IS are approximations to the implicit ASPs of RLSVI through Monte Carlo integration combined with an importance sampling estimator. We evaluate RLSVI-IS in a simulated mobile health environment. Our findings demonstrate that:

- RLSVI-IS achieves cumulative rewards comparable to those of RLSVI in simulation.
- Furthermore, we establish a sufficient condition that ensures tight control the distance between ASPs for RLSVI-IS and the implicit ASPs for the underlying RLSVI. Our simulation also demonstrates that the distance between ASPs of RLSVI-IS and these of RLSVI is relatively small.

2 Problem Formulation

We formally introduce our problem setup. We consider episodic MDPs with state space $\mathcal{S}$, action space $\mathcal{A}$, rewards in $[0,1]$ and horizon H. Each MDP can be denoted by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, H, s_0)$, where $\mathcal{P} = (P_h)_{h=1}^H$ is the collection of transition kernels with each $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{A})$, and $R = (R_h)_{h=1}^H$ is the collection of reward functions with each $R_h : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$. Here s_0 is the initial state of each episode. An agent interacts with the environment episodically. Within the i-th episode, the RL agent interacts with the environment for H steps. At the each step h, the agent chooses an action $a_{i,h}$ and the environment samples the next state $s_{i,h+1} \sim P_h(\cdot \mid s_{i,h}, a_{i,h})$ and reward $r_{i,h} = R_h(s_{i,h}, a_{i,h})$. We denote by $\mathcal{D}_i = \{(S_{j,h'}, A_{j,h'}, S_{j,h'+1}, R_{j,h'})_{h'=1}^H\}_{j=1}^{i-1}\}$ all the observations up to the i-th episode and by $\mathcal{D}_{i,h} = \mathcal{D}_i \cup \{(S_{i,h'}, A_{i,h'}, S_{i,h'+1}, R_{i,h'})_{h'=1}^{h-1}\}$ the observations up to step h of the i-th episode.

2.1 Randomized Least Square Value Iteration

We formally introduce RLSVI (Osband et al., 2016). RLSVI assumes the existence of a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ of the state and action pair and the underlying optimal Q-function can be written as $Q_h^*(s,a) = \langle \phi(s,a), \theta_h^* \rangle$ for some $\theta_h^* \in \mathbb{R}^d$. This assumption is satisfied by linear MDP class (Jin et al., 2020).

In the beginning of each episode i, RLSVI agent generates a random $\theta_i = (\theta_{i,h})_{h=1}^d$ that is the agent's current belief of the underlying true θ_i^*. Specifically, Algorithm 1 is called to generate θ_i. Each $\theta_{i,h}$ is sampled from a Gaussian distribution in a backward manner according to their step index h. The mean and variance of the Gaussian distributions are solved from Bayesian linear regression that minimizes the square loss of the Bellman errors (line 5).

Algorithm 1 Randomized Least Squares Value Iteration (RLSVI)

1: **Input:** previous dataset $\mathcal{D}_{i,1} = \{(s_{j,h}, a_{j,h}, s_{j,h+1}, r_{j,h})_{h=1}^H\}_{j=1}^{i-1}$, feature mapping ϕ, parameters $\lambda, \sigma > 0$
2: Set $\theta_{i,H+1} = \vec{0}$
3: **for** $h = H, \ldots, 1$ **do**
4: Generate regression problem:

$$X_h = (\phi(s_{j,h}, a_{j,h}))_{j=1}^{i-1}, \qquad y_h = \left(r_{j,h} + \max_\alpha \theta_{i,h+1}^\top \phi(s_{j,h+1}, \alpha)\right)_{j=1}^{i-1}$$

5: Bayesian linear regression:

$$\mu_{i,h} \leftarrow \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2} X_h^\top X_h + \lambda I\right)^{-1} X_h^\top y_h, \qquad \Sigma_{i,h} \leftarrow \left(\frac{1}{\sigma^2} X_h^\top X_h + \lambda I\right)^{-1} \tag{2}$$

6: Sample $\theta_{i,h} \sim \mathcal{N}(\mu_{i,h}, \Sigma_{i,h})$, Gaussian posterior formed by previous data
7: **end for**
8: **Output:** $\theta_i = (\theta_{i,1}, \ldots \theta_{i,H})$

Action sampling probability of RLSVI. Given the agent's belief $\theta_{i,h}$ of $\theta_{i,h}^*$ and the current state $S_{i,h}$, the algorithm chooses a deterministic action

$$A_{i,h} = A^*(S_{i,h}, \theta_{i,h}) := \arg\max_{\alpha \in \mathcal{A}} \langle \phi(S_{i,h}, \alpha), \theta_{i,h} \rangle . \tag{3}$$

According to the definition of the ASP, we want to calculate

$$\mathbb{P}(A_{i,h} = a \mid \mathcal{D}_{i,h}) = \int_\theta \mathbb{1}\{a = A^*(S_{i,h}, \theta_{i,h})\} \, \mathbb{P}(\theta_{i,h} = \theta \mid \mathcal{D}_{i,h}) \; d\theta, \tag{4}$$

where $\mathbb{P}(\theta_{i,h} = \cdot \mid \mathcal{D}_{i,h})$ is the density function for the conditional distribution of $\theta_{i,h}$ given $\mathcal{D}_{i,h}$. In other words, one has to integrate sampling probability of each $\theta_{i,h}$ over their posterior distribution.

One may simply verify that there is no closed form for the posterior distribution of $\theta_{i,h}$ given $\mathcal{D}_{i,h}$, neither could we generate samples from it. Computing (4) is a Bayesian computation problem that is computationally infeasible in general. Instead, we can generate $\theta_{i,h}$ from $\mathbb{P}(\theta_{i,h} = \cdot \mid \mathcal{D}_i)$ by running Algorithm 1.

The two posterior distributions do not align in general

$$\mathbb{P}(\theta_{i,h} = \cdot \mid \mathcal{D}_i) \neq \mathbb{P}(\theta_{i,h} = \cdot \mid \mathcal{D}_{i,h}) = \mathbb{P}(\theta_{i,h} = \cdot \mid \mathcal{D}_i \cup \{(S_{i,h'}, A_{i,h'}, S_{i,h'+1}, R_{i,h'})_{h'=1}^{h-1}\}),$$

because $(S_{i,h'}, A_{i,h'}, S_{i,h'+1}, R_{i,h'})_{h'=1}^{h-1}$ and $\theta_{i,h}$ are conditional dependent given $\mathcal{D}_i$.

This point is elucidated in Figure 1, a directed acyclic graph (DAG) of an episode with $H = 2$ steps. The DAG shows the dependence between $\theta_{i,2}$ (blue node) and $(S_{i,1}, A_{i,1}, S_{i,2}, R_{i,1})$ (red nodes) conditioned on $\mathcal{D}_i$ through the path of $\theta_{i,1}$ (green nodes).

3 RLSVI-IS

We propose a new online learning algorithm, namely RLSVI-IS (Importance Sampling), that samples action $A_{i,h}$ by approximating the implicit ASPs of the original RLSVI through Monte Carlo integration combined with an importance sampling estimator.

To approximate ASPs in (4) we sample $(\tilde{\theta}_{i,h}^{(1)}, \ldots, \tilde{\theta}_{i,h}^{(M)})$ by calling Algorithm 1 M times independently. We compute the sampling probability given each $\tilde{\theta}_{i,h}^{(m)}$ by adding importance weights:

$$\hat{\mathbb{P}}_{i,h}(a \mid \mathcal{D}_{i,h}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\left\{a = A^*(S_{i,h}, \tilde{\theta}_{i,h}^{(m)})\right\} \frac{\mathbb{P}(\theta_i = \tilde{\theta}_i^{(m)} \mid \mathcal{D}_{i,h})}{\mathbb{P}(\theta_i = \tilde{\theta}_i^{(m)} \mid \mathcal{D}_i)}. \tag{5}$$

We directly samples $A_t \sim \hat{\mathbb{P}}_{i,h}(a \mid \mathcal{D}_{i,h})$ during the course of online learning. It is well-known that importance sampling estimator $\hat{\mathbb{P}}_{i,h}$ is an unbiased estimator of $\mathbb{P}(A_{i,h} = a \mid \mathcal{D}_{i,h})$. RLSVI-IS reduces to ensemble sampling (Lu & Van Roy, 2017) when the underlying environment is a bandit. Note that the true ASP of RLSVI-IS is not $\hat{\mathbb{P}}_{i,h}(a \mid \mathcal{D}_{i,h})$ as it introduces another source of randomness from Monte Carlo sampling. However, the errors of Monte Carlo Integration can be controlled especially when $M \to \infty$.

Calculation of importance weights. The importance weights are given by the ratio between the conditional density $\mathbb{P}(\theta_i = \tilde{\theta}_i^{(m)} \mid \mathcal{D}_{i,h})/\mathbb{P}(\theta_i = \tilde{\theta}_i^{(m)} \mid \mathcal{D}_i)$. Proposition 1 indicates that the importance weight is nonzero only when the sampled θ reproduces the exact same action sequence $A_{i,1}, \ldots, A_{i,h-1}$ that has been observed so far, given the observed state sequence $S_{i,1}, \ldots, S_{i,h-1}$. Note that their is at least one $\theta_i^{(m)}$ that is consistent with the current action selections. In the worst case, it requires $M = \mathcal{O}(|\mathcal{A}|^H)$ many particles to ensure that there are more than one $\tilde{\theta}_i^{(m)}$'s with non-zero importance weight. However, as we demonstrate later in simulation study that the ASPs calculation converges for reasonable $M = 200$.

Proposition 1 (Importance Weight)**.** *The importance weight of a given $\theta = (\theta_h)_{h=1}^H$ admits*

$$\frac{\mathbb{P}(\theta_i = \theta \mid \mathcal{D}_{i,h})}{\mathbb{P}(\theta_i = \theta \mid \mathcal{D}_i)} \propto \prod_{h'=1}^{h-1} \mathbb{1}\left\{A_{i,h'} = A^*(S_{i,h}, \theta_h)\right\}, \tag{6}$$

where $\propto$ hides terms that do not depend on θ.

Combined with (5), the sampling probability satisfies

$$\hat{\mathbb{P}}_{i,h}(a \mid \mathcal{D}_{i,h}) \propto \frac{1}{M} \sum_{m=1}^{M} \left[\mathbb{1}\left\{a = A^*(S_{i,h}, \tilde{\theta}_{i,h}^{(m)})\right\} \prod_{h'=1}^{h-1} \mathbb{1}\left\{A_{i,h'} = A^*(S_{i,h'}, \tilde{\theta}_{i,h'}^{(m)})\right\}\right], \tag{7}$$

which is in fact the average of the products of h indicator functions.

The form in (6) allows us to update importance weights incrementally. We maintain the current set of θ's that align with the actions selected so far. Whenever, a new action $A_{i,h}$ is observed, we eliminate these θ's in the set that is not consistent with $A_{i,h}$. A pseudo code for this update is described in Algorithm 2.

Algorithm 2 RLSVI-IS in the i-th episode

Input: Previous dataset $\mathcal{D}_i$
Compute $(\tilde{\theta}_i^{(m)})_{m=1}^M$ by running Algorithm 1
Initialize $\mathcal{D}_{i,1} = \mathcal{D}_i$
Initialize importance weight $w_1^{(m)} = 1$ for each $m \in [M]$
for $h = 1, \ldots, H$ **do**
    Calculate sampling probability $\hat{\mathbb{P}}_{i,h}(a \mid \mathcal{D}_{i,h}) = \left(\sum_{m=1}^M A^*(S_{i,h}, \tilde{\theta}_{i,h}^{(m)}) w_h^{(m)}\right) / \sum_{m=1}^M w_h^{(m)}$
    Samples $A_{i,h} \sim \hat{\mathbb{P}}_{i,h}(\cdot \mid \mathcal{D}_{i,h})$, and observes $S_{i,h+1}, R_{i,h}$
    Update $\mathcal{D}_{i,h+1} = \mathcal{D}_{i,h} \cup \{(S_{i,h}, A_{i,h}, S_{i,h+1}, R_{i,h})\}$
    Set $w_{h+1}^{(m)} = 0$ for all m such that $A_{i,h} \notin \arg\max_{\alpha \in \mathcal{A}} \langle \phi(S_{i,h}, \alpha), \tilde{\theta}_{i,h}^{(m)} \rangle$
end for

4 ASPs Distance Analysis

The goal of this section is to understand the distance between the explicitly calculated ASPs $\hat{\mathbb{P}}$ and the implicit ASPs of RLSVI that RLSVI-IS tends to approximate. Define the squared distance as

$$\mathcal{E}_{i,h}(\mathcal{D}_{i,h}) = \sum_a \left(\hat{\mathbb{P}}_{i,h}(a \mid \mathcal{D}_{i,h}) - \mathbb{P}(A_{i,h} = a \mid \mathcal{D}_{i,h})\right)^2, \text{ and } \mathcal{E}_i = \mathbb{E}_{\mathcal{D}_{i,H}}\left[\sum_{h=1}^H \mathcal{E}_{i,h}(\mathcal{D}_{i,h}) \mid \mathcal{D}_i\right],$$

where the later is the expected sum of distance conditional on previous history $\mathcal{D}_i$ and the expectation is taken over both the randomness of $\mathcal{D}_{i,h}$ and the randomness in generating particles $\theta_{i,h}^{(m)}$'s.

The ASP distance for ensemble sampling in bandit is analyzed by Qin et al. (2022). For episodic RL, since the posterior distribution of $\theta_{i,h}$ is shifted once new states and actions are observed, we would not expect a strong worst-case control on per-step approximation error $\mathcal{E}_{i,h}$. Instead, we show that under mild conditions, the expected per-episode errors $\mathcal{E}_i$ can be controlled, and thus we can show that the Bayesian regret of actions sampled from RLSVI-IS is similar to that of RLSVI. The goal of the experimental results demonstrates the empirical performance if the actions are sampled from RLSVI-IS, and an evaluation of the average approximation errors.

We denote by $a(\theta, s) = \arg\max_\alpha \langle \phi(s, \alpha), \theta \rangle$ the optimal action given fixed parameter θ and state s. Let $\vec{s}_h = (s_1, \ldots, s_h)$, $\vec{a}_h = (a_1, \ldots, a_h)$ be a sequence of h states and actions, respectively. Let $\vec{a}_h(\theta, \vec{s}_h) = (a(\theta_1, s_1), \ldots, a(\theta_h, s_h))$. Let $\delta \in [0, 1]$. We define

$$N(\mathcal{D}_i, \delta) := \min_{\mathcal{A}' \in \mathbb{A}(\mathcal{D}_i, \delta)} |\mathcal{A}'|,$$

where $\mathbb{A}(\mathcal{D}_i, \delta)$ is the set of all $\mathcal{A}' \subset \mathcal{A}^h$ that satisfy $\mathbb{P}(\exists \vec{s}_h \text{ such that } \vec{a}_h(\theta_i, \vec{s}_h) \in \mathcal{A}' \mid \mathcal{D}_i) \geq 1 - \delta$.

In words, $N(\mathcal{D}_i, \delta)$ is the cardinality of the largest subset of $\mathcal{A}^h$ that contains the possible action sequence that could be generated by the random θ_i given previous dataset $\mathcal{D}_i$ and any state sequence $\vec{s}_h$ with a probability at least $1 - \delta$.

Theorem 1. *The per-step average approximate error can be upper bounded by*

$$\mathbb{E}[\mathcal{E}_{i,h}(\mathcal{D}_{i,h}) \mid \mathcal{D}_i] = \mathcal{O}\left(\inf_\delta \left(\delta + N(\mathcal{D}_i, \delta)/M\right)\right).$$

Theorem 1 states that if the posterior of θ_i are highly concentrated in terms of the selected action sequence, then the error can be better controlled. In the worst case, $N(\mathcal{D}_i, \delta) \approx (1-\delta)|\mathcal{A}|^h$ and to control the distance, one has to set $M \approx |\mathcal{A}|^h$.

5 Simulation Studies

In this section, we introduce the simulation environment we use to evaluate RLSVI-IS and empirical results.

ADAPTS HCT simulation testbed. As we mentioned above, explicit action sampling probability is crucial for policy evaluation in the after-study analysis that is commonly conducted in health care applications. In light of the health care applications, we evaluate RLSVI-IS under a simulation environment built for a mobile health clinical trial, called ADAPTS HCT (Li et al., 2023), based on the real dataset Roadmap 2.0 (Rozwadowski et al., 2020). ADAPTS HCT provides digital interventions to pairs of adolescents who undergone bone marrow transplantation and also to their care-partners. The goal of ADAPTS HCT is to improve the medication adherence of adolescents by leveraging the dyadic structure formed by the adolescents and their care-partners. This testbed is an ideal environment to evaluate RLSVI-IS as it replicates the noise level and structure that is typically encountered in practical health care applications.

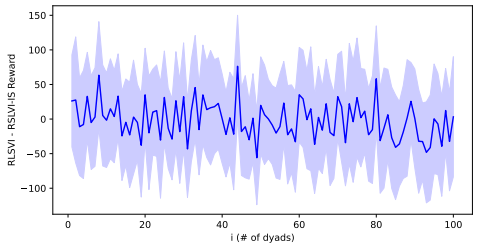
We gave a brief overview of the simulation testbed and readers may find the details in Li et al. (2023). The testbed interacts with N simulated dyads, each staying for 98 days. Each dyad is an episodic MDP (Li et al., 2023) with in total $T = 196$ decision times. The agent makes twice-daily binary decisions $A_t \in \{0, 1\}$ for $t \in [196]$. The state S_t encompasses the past-24-hour *Heart Rate*, *Sleep* and *Step Count*, and the past-week *Mood* measurements of both adolescent and their carepartner. The state transitions $\mathbb{P}(S_{t+1} \mid S_t, A_t)$ follow a linear model with AR(1) working correlation fitted from Roadmap 2.0 dataset for each dyad individually. The reward is a binary version of the step count, which is used as a proxy to medication adherence.

Experimental setups. We run RLSVI-IS and RLSVI over $N = 100$ dyads. Though each dyad is in fact an MDP with long horizon 196, Li et al. (2023) observes that there is very weak dependence across weeks and a more efficient learning is to treat each week as an independent MDP. Therefore, we run RLSVI-IS and RLSVI with horizon $H = 14$ for 100×14 episodes.

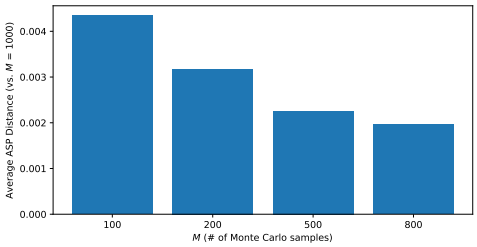
We first evaluate the difference in the sum of rewards for each dyad between RLSVI-IS and RLSVI. The result is demonstrated in Figure 2a, where we observe no significant difference between RLSVI and RLSVI-IS in terms of sum of rewards. We suspect that in the early episodes both the RLSVI and RLSVI-IS have low rewards due to exploration, and RLSVI-IS still explores sufficiently due to the stochastic nature of the digital health environment. In the later episodes, since the posterior distribution of θ_i becomes more concentrated, the distance between ASPs of RLSVI and these of RLSVI-IS becomes very small. Therefore, they both have similar rewards. This conjecture is complimented by Figure 2c, where we observe that ASPs distance significantly decreases over the number of dyads.

5.1 Evaluating ASP Distance

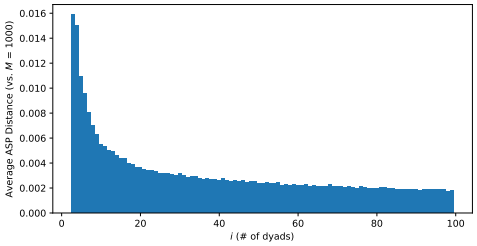
We evaluate the distance between ASPs of RLSVI and these of RLSVI-IS. The goal is to understand how Monte Carlo integration approximates the implicit ASPs of RLSVI. Since there is no explicit form of ASPs for RLSVI, we use ASPs of RLSVI with $M = 1000$ as a proxy. Figure 2 (b) summarizes the average distance between ASPs with each of $M = 100, 200, 500, 800$ and ASPs with $M = 1000$. Recall that M is the total number of Monte Carlo samples. We observe that the ASPs becomes relatively small, less then 0.003, for $M \geq 200$. This demonstrates that the ASPs distance can be well-controlled for reasonable number of Monte Carlo samples.



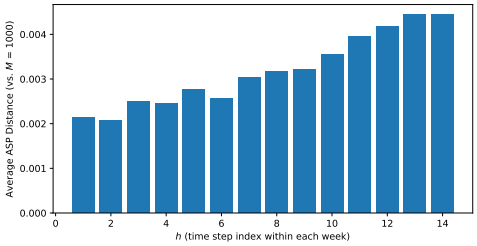
(a) The difference between cumulative rewards for RLSVI and these for RLSVI-IS by dyad count i.



(b) Average ASP Distance between RLSVI with $M = 100, 200, 500, 800$ and RLSVI with $M = 1000$.



(c) Average ASP Distance grouped by dyad count $i = 1, \ldots, 100$.



(d) Average ASP Distance grouped by step count $h = 1, \ldots, 14$ within each week.

Figure 2: Results on RLSVI-IS under the simulation environment for ADAPTS HCT.

In Figure 2 (c) and (d), we study the average ASP distance between $M = 200$ and $M = 1000$ grouped by dyad counts i and step counts h, respectively. In Figure 2c, we observe an decreasing trend in average distance when the number of dyads increases. As implied by Theorem 1, this may due to the fact that the random θ_i generated from the posterior distribution is more consistent when more data is observed, which leads to a smaller ASP distance.

Theorem 1 also implies that the ASP distance increases with the number of steps h per week. In the worst-case, ASP distance scales with $|\mathcal{A}|^h$, because the posterior distribution $\mathbb{P}(\theta_{i,h} \mid \mathcal{D}_{i,h})$ can potentially shift more significantly from $\mathbb{P}(\theta_{i,h} \mid \mathcal{D}_i)$–the distribution we use to generate Monte Carlo samples. This is verified in Figure 2d, where we observe an increase in ASP distance for larger h in a week. However, this effect appears to have a linear form, which may imply that the worst case characterized in Theorem 1 does not always happen in practice.

6 Discussion

Enabling after-study policy evaluation is crucial for real-world RL deployment. This paper contributes significantly to the field by adapting existing online RL algorithms to provide explicit action sampling probabilities (ASPs). Despite these advancements, several challenges remain in optimizing after-study OPE. Our analysis focuses on the discrepancy between the ASPs used by RLSVI-IS and traditional RLSVI, and we have shown that they may scale exponentially with the horizon H. A further importance question is to understand their impact on cumulative regrets in online learning. Although the worst-case distances appear extensive, our results demonstrate comparable cumulative regrets in the specific simulation tested. Identifying conditions where approximation distances remain benign—ensuring similar regret guarantees between RLSVI-IS and RLSVI—is an important ongoing challenge. Furthermore, ensuring that ASPs are strictly bounded away from 0 and 1 for any action taken by the evaluation policy is essential. High variances can arise, as highlighted in Equation (1), when the action sampling probabilities $\pi_b(a \mid \mathcal{D}_{i,h})$ are low. A preliminary solution involves clipping the ASPs with a constant γ, but the it is unclear how such adjustments may impact the theoretical guarantees such as regret bounds. To approximate the implicit ASPs of RLSVI, we

utilized a sequential Monte Carlo method with an importance weight estimator. However, the field of approximate Bayesian computation is rich with alternative approaches, such as Variational Inference and Markov chain Monte Carlo, which have proven effective in various applications. Future work will explore these methods further, assessing their performance in typical mobile health scenarios to determine the most effective approach.

References

Priyank Agrawal, Jinglin Chen, and Nan Jiang. Improved worst-case regret bounds for randomized least-squares value iteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6566–6573, 2021.

Malay Ghosh, N Reid, and DAS Fraser. Ancillary statistics: A review. *Statistica Sinica*, pp. 1309–1332, 2010.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Shuangning Li, Lluis Salvat Niell, Sung Won Choi, Inbal Nahum-Shani, Guy Shani, and Susan Murphy. Dyadic reinforcement learning. *arXiv preprint arXiv:2308.07843*, 2023.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386. PMLR, 2016.

Chao Qin, Zheng Wen, Xiuyuan Lu, and Benjamin Van Roy. An analysis of ensemble sampling. *Advances in Neural Information Processing Systems*, 35:21602–21614, 2022.

Nancy Reid and David R Cox. On some principles of statistical inference. *International Statistical Review*, 83(2):293–308, 2015.

Michelle Rozwadowski, Manasa Dittakavi, Amanda Mazzoli, Afton L Hassett, Thomas Braun, Debra L Barton, Noelle Carlozzi, Srijan Sen, Muneesh Tewari, David A Hanauer, et al. Promoting health and well-being through mobile health technology (roadmap 2.0) in family caregivers and patients undergoing hematopoietic stem cell transplantation: protocol for the development of a mobile randomized controlled trial. *JMIR research protocols*, 9(9):e19288, 2020.

Sabina Tomkins, Peng Liao, Predrag Klasnja, and Susan Murphy. Intelligentpooling: Practical thompson sampling for mhealth. *Machine learning*, 110(9):2685–2727, 2021.

Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8):255, 2022.

A Further Discussions on RLSVI-IS

One may argue that RLSVI-IS also introduces the external source of randomness in the process of generating Monte Carlo samples. In this section, we discuss in depth how these randomness are different in nature from that introduced for RLSVI and why we should condition on all these randomness in the after-study analysis like OPE.

We first observe that one running RLSVI-IS could generate a whole set of particles from standard normal distribution prior the online implementation, and conditioning on these particles, the Monte Carlo samples $\theta_{i,h}^{(m)}$, are simply deterministic mappings of the history.

Generate particles and pseudo samples. Let M be the number of particles we use for each approximation. Before running RLSVI-IS, we generate a set of particles $(\eta_{i,h}^{(m)})_{i,h\in[N]\times[H]}^{m\in[M]}$ i.i.d from the standard normal distribution $\mathcal{N}(0, I_d)$. In the beginning of each episode i, we compute pseudo samples $\tilde{\theta}_i^{(1)}, \ldots, \tilde{\theta}_i^{(m)}$ by running Algorithm 1 but replacing the random Gaussian sample in line 6 with the deterministic mapping $\tilde{\theta}_{i,h}^{(m)} = \mu_{i,h} + \Sigma_{i,h}^{1/2} \eta_{i,h}^{(m)} \Sigma_{i,h}^{1/2}$. The details can be found in Algorithm 3. As we mentioned above, the pseudo sample $\tilde{\theta}_i^{(m)}$ is a deterministic mapping of particles $(\eta_{i,h}^{(m)})_{h\in[H]}$.

Algorithm 3 Generate pseudo samples with Particles

1: **Input:** particles $(\eta_{i,h}^{(m)})_{m\in[M]}$, previous dataset $\mathcal{D}_{i,1} = \{(s_{j,h}, a_{j,h}, s_{j,h+1}, r_{j,h})_{h=1}^{H}\}_{j=1}^{i-1}$, feature mapping ϕ, parameters $\lambda, \sigma > 0$
2: **for** $m = 1, \ldots, M$ **do**
3: Set $\tilde{\theta}_{i,H+1}^{(m)} = \vec{0}$
4: **for** $h = H, \ldots, 1$ **do**
5: Generate regression problem:

$$X_h = (\phi(s_{j,h}, a_{j,h}))_{j=1}^{i-1}, \qquad y_h = \left(r_{j,h} + \max_\alpha (\tilde{\theta}_{i,h+1}^{(m)})^\top \phi(s_{j,h+1}, \alpha)\right)_{j=1}^{i-1}$$

6: Bayesian linear regression:

$$\mu_{i,h} \leftarrow \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2} X_h^\top X_h + \lambda I\right)^{-1} X_h^\top y_h, \qquad \Sigma_{i,h} \leftarrow \left(\frac{1}{\sigma^2} X_h^\top X_h + \lambda I\right)^{-1} \tag{8}$$

7: Compute $\tilde{\theta}_{i,h}^{(m)} = \mu_{i,h} + \Sigma_{i,h}^{1/2} \eta_{i,h}^{(m)} \Sigma_{i,h}^{1/2}$
8: Set $\tilde{\theta}_i^{(m)} = (\tilde{\theta}_{i,h}^{(m)})_{h\in[H]}$
9: **end for**
10: **Output:** $(\tilde{\theta}_i^{(m)})_{m\in[M]}$
11: **end for**

Conditional inference. All of our sampling probabilities and OPE will condition on the generated set of particles $(\eta_{i,h}^{(m)})_{i,h\in[N]\times[H]}^{m\in[M]}$. These particles are ancillary random variables (independent of the any parameter of interest), and an important principle of statistical inference is the conditionality on ancillary variables (Ghosh et al., 2010; Reid & Cox, 2015). However, we should not think the same way for RLSVI, who also generates random Gaussian variables in the beginning of each episode. On one hand, if we had condition on these randomness, the ASPs of RLSVI becomes either 0 or 1. This will extremely limit the OPE since the ASPs on the denominator being 0 can leads to large variance in the estimator. Secondly, these randomness are introduced to do exploration, which should not be considered ancillary especially when the evaluation policy is the deployed online learning algorithm.

B Proofs

B.1 Proof of Proposition 1

Proposition 1. [Importance Weight] *The importance weight of a given $\theta = (\theta_h)_{h=1}^H$ admits*

$$\frac{\mathbb{P}(\theta_i = \theta \mid \mathcal{D}_{i,h})}{\mathbb{P}(\theta_i = \theta \mid \mathcal{D}_i)} \propto \prod_{h'=1}^{h-1} \mathbb{1}\left\{A_{i,h'} = A^*(S_{i,h}, \theta_h)\right\},$$

where $\propto$ hides terms that do not depend on θ.

Proof. Denote by $O_{i,h} = \{(S_{i,h',A_{i,h'},R_{i,h'},S_{i,h'+1}})\}_{h'=1}^{h-1}$.

$$\frac{\mathbb{P}(\theta_i = \tilde{\theta}_i^{(m)} \mid \mathcal{D}_{i,h})}{\mathbb{P}(\theta_i = \tilde{\theta}_i^{(m)} \mid \mathcal{D}_i)} = \frac{\mathbb{P}(O_{i,h} \mid \mathcal{D}_i, \theta_i = \theta)}{\mathbb{P}(O_{i,h} \mid \mathcal{D}_i)} \propto \mathbb{P}(O_{i,h} \mid \mathcal{D}_i, \theta_i = \theta)$$

To proceed,

$$\begin{aligned}
\mathbb{P}(O_{i,h} \mid \mathcal{D}_i, \theta_i = \theta) &= \mathbb{P}(O_{i,h} \mid \theta_i = \theta) \\
&= \prod_{h'=1}^{h-1} \mathbb{P}(A_{i,h'} \mid S_{i,h'}, \theta)\mathbb{P}(S_{i,h'+1} \mid S_{i,h'}, A_{i,h'})\mathbb{P}(R_{i,h'} \mid S_{i,h'}, A_{i,h'}) \\
&\propto \prod_{h'=1}^{h-1} \mathbb{P}(A_{i,h'} \mid S_{i,h'}, \theta) \\
&= \prod_{h'=1}^{h-1} P_{\mathcal{L}}(A_{i,h'} \mid S_{i,h'}, \theta)
\end{aligned}$$

□

B.2 Proof of Theorem 1

Theorem 1. *The per-step average approximate error can be upper bounded by*

$$\mathbb{E}[\mathcal{E}_{i,h}(\mathcal{D}_{i,h}) \mid \mathcal{D}_i] = \mathcal{O}\left(\inf_{\delta}\left(\delta + N(\mathcal{D}_i, \delta)/M\right)\right).$$

Proof. We first focus on the analysis of the error $\mathcal{E}_{i,h}$ for a given episode and step. Define $\mathcal{O}_{i,h} = \{(S_{i,h'}, A_{i,h'}, S_{i,h'+1}, R_{i,h'})_{h'=1}^{h-1}\}$ as the new observations received in the current episode up to step h. Since we focus on a fixed i and h, we omit the subscript of i, h for a simpler notation. That is $\mathcal{D}_h = \mathcal{D}_{i,h}$ and $\mathcal{E}_h = \mathcal{E}_{i,h}$, and $\hat{\mathbb{P}}_h = \hat{\mathbb{P}}_{i,h}$.

$$\begin{aligned}
\mathbb{E}[\mathcal{E}_h(\mathcal{D}_h)] &= \sum_a \mathbb{E}\left[\left(\hat{\mathbb{P}}_h^a(\mathcal{D}_h) - \mathbb{P}(A_h = a \mid \mathcal{D}_h)\right)^2\right] &(9)\\
&= \sum_a \mathbb{E}[\mathrm{Var}(\hat{\mathbb{P}}_h^a(\mathcal{D}_h) \mid \mathcal{D}_h)], &(10)
\end{aligned}$$

where the expectation in the second line is taken over the randomness of $\mathcal{D}_h$.

Recall that we denote by $a(\theta, s) = \arg\max_\alpha \langle \phi(s, \alpha), \theta\rangle$ the optimal action given fixed parameter θ and state s. Let $\vec{s}_h = (s_1, \ldots, s_h)$, $\vec{a}_h = (a_1, \ldots, a_h)$ be a sequence of h states and actions, respectively. Let $\vec{a}_h(\theta, \vec{s}_h) = (a(\theta_1, s_1), \ldots, a(\theta_h, s_h))$. Let $\delta \in [0, 1]$. We define

$$N(\mathcal{D}_i, \delta) := \min_{\mathcal{A}' \in \mathbb{A}(\mathcal{D}_i, \delta)} |\mathcal{A}'|,$$

where $\mathbb{A}(\mathcal{D}_i, \delta)$ is the set of all $\mathcal{A}' \subset \mathcal{A}^h$ that satisfy $\mathbb{P}(\exists \vec{s}_h$ such that $\vec{a}_h(\theta_i, \vec{s}_h) \in \mathcal{A}' \mid \mathcal{D}_i) \geq 1 - \delta$. For any choice of $\delta \in [0, 1]$, let

$$\mathcal{A}'(\mathcal{D}_i, \delta) \in \arg\min_{\tilde{\mathcal{A}} \in \mathbb{A}(\mathcal{D}_i, \delta)} |\tilde{\mathcal{A}}|.$$

We choose a subset $\Theta(\mathcal{D}_i, \delta) \subset \mathbb{R}^d$ such that

$$\forall \theta \in \Theta(\mathcal{D}_i, \delta), \exists \vec{s}_h, \vec{a}_h(\theta, \vec{s}_h) \in \mathcal{A}'(\mathcal{D}_i, \delta),$$

and

$$\mathbb{P}(\theta_i \in \Theta(\mathcal{D}_i, \delta)) \geq 1 - \delta.$$

For the above conditional variance of a given action a, we have

$$\mathrm{Var}(\hat{\mathbb{P}}_h^a(\mathcal{D}_h) \mid \mathcal{D}_h) \tag{11}$$

$$= \frac{1}{4}\mathbb{P}(\theta_i \notin \Theta(\mathcal{D}_i, \delta)) + \frac{1}{M}\int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \frac{\left(\mathbb{1}\{a = A^*(S_h,\theta_h)\} f_{\theta_h|\mathcal{D}_h}(\theta, \mathcal{D}_h) - \mathbb{P}(A_h = a \mid S_h, \theta_h) f_{\theta_h}(\theta)\right)^2}{f_{\theta_h}(\theta)}\,\mathrm{d}\theta \tag{12}$$

$$\leq \frac{\delta}{4}\frac{1}{M}\left(\int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \mathbb{P}^2(A_h = a \mid S_h, \theta_h) f_{\theta_h}(\theta)\,\mathrm{d}\theta + \int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \mathbb{1}\{a = A^*(S_h,\theta_h)\}\frac{f^2_{\theta_h|\mathcal{D}_h}(\theta,\mathcal{D}_h)}{f_{\theta_h}(\theta)}\,\mathrm{d}\theta\right) \tag{13}$$

$$\leq \frac{\delta}{4} + \frac{1}{M} + \frac{1}{M}\int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \frac{f^2_{\theta_h|\mathcal{D}_h}(\theta,\mathcal{D}_h)}{f_{\theta_h}(\theta)}\,\mathrm{d}\theta \tag{14}$$

To proceed, we further bound the third term

$$\mathbb{E}\left[\int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \frac{f^2_{\theta_h|\mathcal{D}_h}(\theta,\mathcal{D}_h)}{f_{\theta_h}(\theta)}\,\mathrm{d}\theta\right] \tag{15}$$

$$= \int_D \int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \frac{f^2_{\theta_h|\mathcal{D}_h}(\theta,D)}{f_{\theta_h}(\theta)} f_{\mathcal{D}_h}(D)\,\mathrm{d}\theta\,\mathrm{d}D \tag{16}$$

$$= \int_D \int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \frac{f^2_{\theta_h|\mathcal{D}_h}(\theta,D)}{f_{\theta_h}(\theta)} f_{\mathcal{D}_h}(D)\,\mathrm{d}\theta\,\mathrm{d}D \tag{17}$$

$$= \int_D \int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \frac{f_{D_h|\theta_h}(D,\theta)}{f_{D_h}(D)} f_{D_h|\theta_h}(D,\theta) f_{\theta_h}(\theta)\,\mathrm{d}\theta\,\mathrm{d}D \tag{18}$$

(Assuming they are both finite, so we can exchange) (19)

$$= \int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \left(\int_D \frac{f_{D_h|\theta_h}(D,\theta)}{f_{D_h}(D)} f_{D_h|\theta_h}(D,\theta)\,\mathrm{d}D\right) f_{\theta_h}(\theta)\,\mathrm{d}\theta \tag{20}$$

$$= \int_{\theta\in\Theta(\mathcal{D}_i,\delta)} \left(\int_D \frac{\prod_{h'=1}^h \mathbb{1}\{a_h = A^*(s_h,\theta)\}}{\int_{\theta'\in\Theta(\mathcal{D}_i,\delta)} \prod_{h'=1}^h \mathbb{1}\{a_h = A^*(s_h,\theta')\} f_{\theta_h}(\theta')\,\mathrm{d}\theta'} f_{D_h|\theta_h}(D,\theta)\,\mathrm{d}D\right) f_{\theta_h}(\theta)\,\mathrm{d}\theta \tag{21}$$

$$\leq N(\mathcal{D}_i, \delta). \tag{22}$$

□