The Curse of Passive Data Collection in Batch Reinforcement Learning

Chenjun Xiao[1,2]   Ilbin Lee[1]   Bo Dai[2]   Dale Schuurmans[1,2]   Csaba Szepesvari[1,3]
[1]University of Alberta   [2]Google Research, Brain Team   [3]DeepMind
{chenjun,ilbin,szepesva,daes}@ualberta.ca, bodai@google.com

Abstract

In high stake applications, active experimentation may be considered too risky and thus data are often collected passively. While in simple cases, such as in bandits, passive and active data collection are similarly effective, the price of passive sampling can be much higher when collecting data from a system with controlled states. The main focus of the current paper is the characterization of this price. For example, when learning in episodic finite state-action Markov decision processes (MDPs) with S states and A actions, we show that even with the best (but passively chosen) logging policy, $\Omega(A^{\min(S-1,H)}/\varepsilon^2)$ episodes are necessary (and sufficient) to obtain an ε-optimal policy, where H is the length of episodes. Note that this shows that the sample complexity blows up exponentially compared to the case of active data collection, a result which is not unexpected, but, as far as we know, have not been published beforehand and perhaps the form of the exact expression is a little surprising. We also extend these results in various directions, such as other criteria or learning in the presence of function approximation, with similar conclusions. A remarkable feature of our result is the sharp characterization of the exponent that appears, which is critical for understanding what makes passive learning hard.

1   Introduction

Batch reinforcement learning (RL) broadly refers to the problem of finding a policy with high expected

return in a stochastic control problem when only a batch of data collected from the controlled system is available. Here, we consider this problem for finite state-action Markov decision processes (MDPs), with or without function approximation, when the data is in the form of trajectories obtained by following some *logging policy*. In more details, the trajectories are composed of sequences of states, actions, and rewards, where the action is chosen by the logging policy, and the next states and rewards follow the distributions specified by the MDP's transition parameters.

There are two subproblems underlying batch RL: the *design problem*, where the learner needs to specify a data collection mechanism that will be used to collect the batch data; and the *policy optimization problem*, where the learner needs to specify the algorithm that produces a policy given the batch data. For the design problem, often times one can use an adaptive data collection process where the next action to be taken is determined by the past data. Another way to say this is that the data collection is done in an *active* way. Recent theoretical advances in reward-free exploration (e.g., Jin et al., 2020; Kaufmann et al., 2021; Zhang et al., 2021) show that one can design algorithms to collect a batch data set with only polynomial samples to have good coverage over all possible scenarios in the environment. Near-optimal policies can be obtained for any given reward functions using standard policy optimization algorithms with the collected data. A complication arises in applications, such as healthcare, education, autonomous driving, or hazard management, where active data collection is either impractical or dangerous (Levine et al., 2020). In these applications, the best one could do is to collect data using a fixed, logging policy, which needs to be chosen *a priori*, that is before the data collection process begins, so that the stakeholders can approve it. Arguably, this is the most natural problem setting to consider in batch learning. The fundamental questions are: how should one choose the logging policy so as to maximize the chance of obtaining a good policy with as little data as possible and how many samples are sufficient and necessary to

obtain a near optimal policy given a logging policy, and which algorithm to use to obtain such a policy?

Perhaps surprisingly, before our paper, these questions remained unanswered. In particular, while much work have studied the sample complexity of batch RL, the results in these works are focusing only on the policy optimization subproblem and as such fall short in providing an answer to our questions. In particular, some authors give sample complexity upper and lower bounds as a function of a parameter, d_m, which could be the smallest visit probability of state-action pairs under the logging policy (Chen and Jiang, 2019; Yin and Wang, 2020; Yin et al., 2021a; Ren et al., 2021; Uehara et al., 2021; Yin and Wang, 2021; Xie and Jiang, 2021; Xie et al., 2021a), or the smallest ratio of visit probabilities of the logging versus the optimal policies, again over all state-action pairs (Liu et al., 2019, 2020; Yin et al., 2021b; Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021b). The sample complexity results depend polynomially on $1/d_m$, S, A and H, where H is the episode length or the effective horizon. Although these results are valuable in informing us the policy optimization step of batch RL, they provide no clue as to how to choose the logging policy to get a high value for d_m and whether d_m will be uniformly bounded from below when adopting such a logging policy. In particular, if we take the first definition for d_m, in some MDPs d_m will be zero regardless of the logging policy if some state is not accessible from the initial distribution. While this predicts an infinite sample complexity for our problem, this is clearly too conservative, since if a state is not accessible under any policy, it is unimportant to learn about it. This is corrected by the second definition. However, even with this definition it remains unclear whether d_m will be uniformly bounded away from zero for an MDP with a fixed number of states and actions and the best instance-agnostic choice of the logging policy. The lower bounds in these work also fail to provide a lower bound for our setting. This is because in these lower bounds the instance will include an adversarially chosen logging policy, again falling short of helping us. Essentially, these are the gaps that we fill with this paper.

In particular, we first show that in tabular MDPs the number of transitions necessary and sufficient to obtain a good policy, *the sample complexity of learning*, is an exponential function of the minimum of the number of states and the planning horizon. In more details, we prove that the sample complexity of obtaining ε-optimal policies is at least $\Omega(\mathrm{A}^{\min(\mathrm{S}-1,H+1)})$ for γ-discounted problems, where S is the number of states, A is the number of actions (per state), and H is the effective horizon defined as $H = \lfloor \frac{\ln(1/\varepsilon)}{\ln(1/\gamma)} \rfloor$. For finite horizon problems with horizon H, we prove the ana-

logue $\Omega(\mathrm{A}^{\min(\mathrm{S}-1,H)}/\varepsilon^2)$ lower bound. These results for tabular MDPs immediately imply exponential lower bounds when linear value function approximation is applied with S replaced by d, the number of features. We also show that warm starts (when one starts with a policy which is achieving almost as much as the optimal policy) do not help either, crushing the hope that one the "curse" of passivity can be broken by adopting a straightforward two-phase data collection process (Bai et al., 2019; Zhang et al., 2020; Gao et al., 2021). We then establish nearly matching upper bounds for both the plug-in algorithm and pessimistic algorithm, showing that the sample complexity behaves essentially as shown in the lower bounds. While the upper bounds for these two algorithms may be off by a polynomial factor, we do not expect the pessimistic algorithm to have a major advantage over the plug-in method in the worst-case setting. In fact, the recent work of Xiao et al. (2021) established this in a rigorous fashion for the bandit setting by showing an algorithm independent lower bound that matched the upper bound for both the plug-in method and the pessimistic algorithm. In the average reward case we show that the sample complexity is infinite.

How do our results relate to the expectations of RL practitioners (and theoreticians)? We believe that most members of our community recognize that batch RL is hard at the fundamental level. Yet, it appears that many in the community are still highly optimistic about batch RL, as demonstrated by the numerous empirical papers that document successes of various levels and kinds (e.g., Laroche et al., 2019; Kumar et al., 2019; Wu et al., 2019; Jaques et al., 2019; Agarwal et al., 2020b; Kidambi et al., 2020; Yu et al., 2020; Gulcehre et al., 2020; Fu et al., 2020), or by the optimistic tone of the above-mentioned theoretical results. The enthusiasm of the community is of course commendable and nothing is farthest from our intentions than to break it. In connection to this, we would like to point out that our results show that if *either H, or* S (or d when we have d features) is fixed, batch RL is in fact tractable. Yet, the lower bound assures us that we cannot afford batch RL if *both* of these parameters are large, a result which one should not hide from. Perhaps the most important finding here is the curious interplay between the horizon and the number of states (more generally, we expect a complexity parameter of the MDP to stand here), which is reasonable yet we find it non-obvious. Certainly, the proof that shows that the interplay is "real" required some new ideas. Returning to the empirical works, recent studies identify the tandem effect from the issue of function approximation in the batch RL with passive data collection (Ostrovski et al., 2021). Our results suggest that there is a need to rethink how batch RL methods are benchmarked. In particular,

Chenjun Xiao[1,2]   Ilbin Lee[1]   Bo Dai[2]   Dale Schuurmans[1,2]   Csaba Szepesvari[1,3]

a tedious examination of the benchmarks shows that the promising results are almost always produced on data sets that are collected by a noisy version of a near-optimal policy. The problem is that this choice biases the development of algorithms towards those that exploit this condition (e.g., the pessimistic algorithm), yet, this mode of data collection is unrealistic.[1]

Another highlight of our result is that it implies an exponential gap between the sample complexity of passive and active learning. Recently, a similar conclusion is drawn in the paper of Zanette (2021), which served as the main motivation for our work. Zanette demonstrated an exponential separation for the case when batch learning is used in the presence of linear function approximation. A careful reading of the paper shows that the lower bound shown there does not apply to the tabular setting as the data collection method of this paper allows sampling from any distribution over the state-action space, which, in the tabular setting is sufficient for polynomial sample complexity.

2   Notation and Background

Notation   We let $\mathbb{R}$ denote the set of real numbers, and for a positive integer i, let $[i] = \{0, \ldots, i-1\}$ be the set of integers from 0 to $i-1$. We also let $\mathbb{N} = \{0, 1, \ldots\}$ be the set of nonnegative integers and $\mathbb{N}_+ = \{1, 2, \ldots\}$ be the set of positive integers. For a finite set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the set of probability distributions over $\mathcal{X}$. We also use the same notation for infinite sets when the set has a clearly identifiable measurability structure such as $\mathbb{R}$, which in this context is equipped by the σ-algebra of Borel measurable sets. We use $\mathbb{I}$ to denote the indicator function. We also use $\mathbf{1}$ to be the identically one function/vector; the domain/dimension is so that the expression that involves $\mathbf{1}$ is well-defined.

We consider finite Markov decision processes (MDPs) given by $M = (\mathcal{S}, \mathcal{A}, Q)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, Q is a stochastic kernel from the set $\mathcal{S} \times \mathcal{A}$ of state-action pairs to $\mathbb{R} \times \mathcal{S}$. In particular, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $Q(\cdot|s, a)$ gives a distribution over pairs composed of a real number and a state. If $(R, S') \sim Q(\cdot|s, a)$, R is interpreted as a random reward incurred and S' the random next state when action a is used in state s. Since the identity of the states and actions plays no role, without loss of

generality, in what follows we assume that $\mathcal{S} = [\mathrm{S}]$ and $\mathcal{A} = [\mathrm{A}]$ for some S, A positive integers.

Every MDP also induces a transition "function", $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, which is the marginal of Q with respect to S', and an immediate mean reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which is so that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $r(s, a)$ gives the mean of R above. Since $\mathcal{S}$ and $\mathcal{A}$ are finite, without loss of generality, we assume that the immediate mean rewards lie in the $[-1, 1]$ interval. We also assume that the reward distribution is ρ-subgaussian with a constant $\rho > 0$. For simplicity, we assume that $\rho = 1$. For any (memoryless) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we define P^π to be the transition matrix on state-action pairs induced by π, where $P^\pi_{(s,a),(s',a')} := \pi(a'|s')P(s'|s, a)$. We denote $\mathcal{M}(\mathcal{S}, \mathcal{A})$ the set of MDPs that satisfy the properties stated in this paragraph. Since the identity of states and actions is unimportant, we also use $\mathcal{M}(\mathrm{S}, \mathrm{A})$ to denote the set of MDPs with S states and A actions (say, over the canonical sets $\mathcal{S} = [\mathrm{S}]$ and $\mathcal{A} = [\mathrm{A}]$).

We use $\mathbb{E}^\pi$ to denote the expectation operator under the distribution $\mathbb{P}^\pi$ induced by the interconnection of policy π and the MDP M on trajectories $(S_0, A_0, R_0, S_1, A_1, R_1, \ldots)$ formed of an infinite sequence of state, action, reward triplets. Here, it is assumed that the initial state is randomly chosen from a distribution that is supported on the whole of the state space, but the identity of this distribution will be unimportant. As such the dependence on this distribution is suppressed. Further, under $\mathbb{P}^\pi$, the distribution of S_{t+1} follows $P(\cdot|S_t, A_t)$ given the history $H_t = (S_0, A_0, R_0, \ldots, R_{t-1}, S_t, A_t)$ while the distribution of A_t follows $\pi(\cdot|S_t)$ given H_t. To minimize clutter the notation also suppresses the dependence on the MDP whenever it is clear which MDP is referred.

For the *discounted total reward criterion* with discount factor $0 \le \gamma < 1$ the state value function $v^\pi : \mathcal{S} \to \mathbb{R}$ under π is defined as,

$$v^\pi(s) := \mathbb{E}^\pi \left[\sum_{t=0}^\infty \gamma^t r(S_t, A_t) \,\middle|\, S_0 = s \right].$$

For $\mu \in \Delta(\mathcal{S})$ and $v : \mathcal{S} \to \mathbb{R}$, we define $v(\mu) = \sum_{s \in \mathcal{S}} \mu(s)v(s) (= \mathbb{E}_{s \sim \mu}[v(s)])$. The state-action value function of π, $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is defined as,

$$q^\pi(s, a) := r(s, a) + \gamma \sum_{s'} P(s'|s, a)v^\pi(s').$$

The standard goal in a finite MDP under the discounted criterion is to identify the optimal policy π^* that maximizes the value function in every state $s \in \mathcal{S}$ such that $v^*(s) = \sup_\pi v^\pi(s)$. In this paper though, we consider the less demanding problem of finding a policy π that maximizes $v^\pi(\mu)$ for a fixed initial state distribution

---

[1]Qin et al. (2021) points to another problem with the benchmarks; namely that they fail to compare to the noise-free version of the near-optimal policy used in the data collection. Brandfonbrener et al. (2021) observe that simply doing one step of constrained/regularized policy improvement using a value estimate of the behavior policy performs surprisingly well in offline RL benchmarks. They hypothesize that the strong performance is due to a combination of favorable structure in the environment and behavior policy.

μ, i.e., finding a policy π which achieves, or nearly achieves $v^*(\mu)$.

For an initial state distribution $\mu \in \Delta(\mathcal{S})$ and a policy π, we define the (unnormalized) discounted occupancy measure ν_μ^π induced by μ, π, and the MDP as

$$\nu_\mu^\pi(s,a) := \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(S_t = s, A_t = a | S_0 \sim \mu).$$

The value of a policy can be represented as an inner product between the immediate reward function r and the occupancy measure ν_μ^π

$$v^\pi(\mu) = \sum_{s,a} r(s,a) \nu_\mu^\pi(s,a) = \langle \nu_\mu^\pi, r \rangle.$$

For $\varepsilon > 0$, we define the effective horizon $H_{\gamma,\varepsilon} := \lfloor \frac{\ln(1/\varepsilon)}{\ln(1/\gamma)} \rfloor$.[2] In the case of *fixed-horizon* policy optimization, instead of a discount factor, one is given a horizon $H > 0$ and the value of a policy π given μ is redefined to be $v^\pi(\mu) = \mathbb{E}^\pi[\sum_{t=0}^{H-1} r(S_t, A_t) | S_0 \sim \mu]$. As before, the goal is to identify a policy whose value is close to $v^*(\mu) = \sup_\pi v^\pi(\mu)$. Finally, in the *average reward* setting, $v^\pi(\mu)$ is redefined to be

$$v^\pi(\mu) = \liminf_{T \to \infty} \mathbb{E}^\pi \left[\frac{1}{T} \sum_{t=0}^{T-1} r(S_t, A_t) \,\Big|\, S_0 \sim \mu \right].$$

For a given $\varepsilon > 0$, in all the various settings, we say that π is ε-*optimal* in MDP M given μ if $v^\pi(\mu) \geq v^*(\mu) - \varepsilon$.

3 Batch Policy Optimization

We consider *policy optimization* in a *batch mode*, or, in short, *batch policy optimization* (BPO). A *BPO problem* for a fixed sample size n is given by the tuple $\mathcal{B} = (\mathcal{S}, \mathcal{A}, \mu, n, \mathcal{P})$ where $\mathcal{S}$ and $\mathcal{A}$ are finite sets, μ is a probability distribution over $\mathcal{S}$, n is a positive integer, and $\mathcal{P}$ is a set of MDP-distribution pairs of the form (M, G), where $M \in \mathcal{M}(\mathcal{S}, \mathcal{A})$ is an MDP over $(\mathcal{S}, \mathcal{A})$ and G is a probability distribution over $(\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S})^n$. In what follows a pair (M, G) of the above form will be called a BPO instance.

A *BPO algorithm* for a given sample size n and sets $\mathcal{S}, \mathcal{A}$ takes data $\mathcal{D} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S})^n$ and returns a policy π (possibly history-dependent). Ignoring computational aspects, we will identify BPO algorithms with (possibly randomized) maps $\mathcal{L} : (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S})^n \to \Pi$, where Π is the set of all policies. The aim is to find BPO algorithms that find near-optimal policies with high probability on every instance within a BPO problem:

---

[2]This is different with the normally used effective horizon, $\frac{\ln(1/\varepsilon(1-\gamma))}{\ln(1/\gamma)}$, with only a $\ln(1/(1-\gamma))$ factor.

Definition 1 ((ε, δ)-sound algorithm)**.** *Fix $\varepsilon > 0$ and $\delta \in (0,1)$. A BPO algorithm $\mathcal{L}$ is (ε, δ)-sound on instance (M, G) given initial state distribution μ if*

$$\mathbb{P}_{\mathcal{D} \sim G} \left(v^{\mathcal{L}(\mathcal{D})}(\mu) > v^*(\mu) - \varepsilon \right) > 1 - \delta,$$

where the value functions are for the MDP M. Further, we say that a BPO algorithm is (ε, δ)-sound on a BPO problem $\mathcal{B} = (\mathcal{S}, \mathcal{A}, \mu, n, \mathcal{P})$ if it is sound on any $(M, G) \in \mathcal{P}$ given the initial state distribution μ.

Data collection mechanisms A *data collection mechanism* is a way of arriving at a distribution G over the data given an MDP and some other inputs, such as the sample size. We consider two types of data collection mechanisms. One of them is governed by a distribution over the state-action pairs, the other is governed by a policy and a way of deciding how a fixed sample size n should be split up into episodes in which the policy is followed. We call the first SA-sampling, the second policy-induced data collection.

- *SA-sampling:* An SA-sampling scheme is specified by a probability distribution $\mu_{\log} \in \Delta(\mathcal{S} \times \mathcal{A})$ over the state-action pairs. For a given sample size n, $\mu_{\log}$ together with an MDP M induces a distribution $G_n(M, \mu_{\log})$ over n tuples $\mathcal{D} = (S_i, A_i, R_i, S_i')_{i=0}^{n-1}$ so that the elements of this sequence form an i.i.d. sequence such that for any $i \in [n]$, $(S_i, A_i) \sim \mu_{\log}$, $(R_i, S_i') \sim Q(\cdot|S_i, A_i)$.

- *Policy-induced data collection* A policy induced data collection scheme is specified by $(\pi_{\log}, \mathbf{h})$, where $\pi_{\log} : \mathcal{S} \to \Delta(\mathcal{A})$ is a policy, which we shall call the *logging policy*, and $\mathbf{h} = (\mathbf{h}_n)_{n \geq 1}$: For each $n \geq 1$, $\mathbf{h}_n$ is an m-tuple $(h_j)_{j \in [m]}$ of positive integers for some m, specifying the length of the m episodes in the data whose total length is n. Then, for any n, the pair $(\pi_{\log}, \mathbf{h}_n)$ together with an MDP M and an initial distribution μ induces a distribution $G(M, \pi_{\log}, \mathbf{h}_n, \mu)$ over the n tuples $\mathcal{D} = (S_i, A_i, R_i, S_i')_{i=0}^{n-1}$ as follows: The data consists of m episodes, with episode $j \in [m]$ having length h_j and taking the form $\tau_j = (S_0^{(j)}, A_0^{(j)}, R_0^{(j)}, \ldots, S_{h_j-1}^{(j)}, A_{h_j-1}^{(j)}, R_{h_j-1}^{(j)}, S_{h_j}^{(j)})$, where $S_0^{(j)} \sim \mu$, $A_t^{(j)} \sim \pi_{\log}(\cdot|S_t^{(j)})$, $(R_t^{(j)}, S_{t+1}^{(j)}) \sim Q(\cdot|S_t^{(j)}, A_t^{(j)})$. Then, for $i \in [n]$, $(S_i, A_i, R_i, S_i') = (S_t^{(j)}, A_t^{(j)}, R_t^{(j)}, S_{t+1}^{(j)})$ where $j \in [m]$, $t \in [h_j]$ are unique integers such that $i = \sum_{j' < j} h_{j'} + t$. We call $\mathbf{h}$ *data splitting scheme.*

Now, under SA-sampling, the sets $\mathcal{S}$, $\mathcal{A}$, a logging distribution $\mu_{\log}$ and state-distribution μ over the respective sets give rise to the BPO problem $\mathcal{B}(\mu_{\log}, \mu, n) = (\mathcal{S}, \mathcal{A}, \mu, n, \mathcal{P}(\mu_{\log}, n))$, where $\mathcal{P}(\mu_{\log}, n)$ is the set of

Chenjun Xiao[1,2]   **Ilbin Lee**[1]   **Bo Dai**[2]   **Dale Schuurmans**[1,2]   **Csaba Szepesvari**[1,3]

all pairs of the form $(M, G_n(M, \mu_{\log}))$, where $M \in \mathcal{M}(\mathcal{S}, \mathcal{A})$ is an MDP with the specified state-action spaces and $G_n(M, \mu_{\log})$ is defined as above. Similarly, a fixed policy $\pi_{\log}$, fixed episode lengths $\mathbf{h} \in \mathbb{N}_+^m$ for some m integer and a fixed state-distribution μ give rise to a BPO problem $\mathcal{B}(\pi_{\log}, \mu, \mathbf{h}) = (\mathcal{S}, \mathcal{A}, \mu, |\mathbf{h}|, \mathcal{P}(\pi_{\log}, \mathbf{h}))$, where $\mathcal{P}(\pi_{\log}, \mathbf{h})$ is the set of pairs of the form $(M, G(M, \pi_{\log}, \mathbf{h}, \mu))$ where $M \in \mathcal{M}(\mathcal{S}, \mathcal{A})$ and $G(M, \pi_{\log}, \mathbf{h}, \mu)$ is a distribution as defined above. Here, we use $|\mathbf{h}|$ to denote $\sum_{s=0}^{m-1} h_s$ which is the sample size specified by $\mathbf{h}$.

The *sample-complexity of BPO with SA-sampling* for a given pair (ε, δ) and a criterion (discounted, finite horizon, or average reward) is the smallest integer n such that for each μ there exists a logging distribution $\mu_{\log}$ and a BPO algorithm $\mathcal{L}$ for this sample size such that $\mathcal{L}$ is (ε, δ)-sound on the BPO problem $\mathcal{B}(\mu_{\log}, \mu, n)$. Similarly, *the sample-complexity of BPO with policy-induced data collection* for a given pair (ε, δ) and a criterion is the smallest integer n such that for each μ there exists a logging policy $\pi_{\log}$ and episode lengths $\mathbf{h} \in \mathbb{N}_+^m$ with $|\mathbf{h}| = n$ and a BPO algorithm that is (ε, δ)-sound on $\mathcal{B}(\pi_{\log}, \mu, \mathbf{h})$.

The *SA*-sampling based data collection is realistic when there is a simulator that allows this type of data collection (Agarwal et al., 2020a; Azar et al., 2013; Cui and Yang, 2020; Li et al., 2020). Besides this scenario, it is hard to imagine a case when *SA*-sampling can be realistically applied. Indeed, in most practical settings, data collection happens by following some policy, usually from the same initial state distribution that is used in the objective of policy optimization.

For policy-induced data collection, a key restriction on the logging policy is that it is chosen without any knowledge of the MDP. Moreover, that the logging policy is memoryless rules out any adaptation to the MDP. The intention here is to model a "tabula rasa" setting, which is relevant when one must find a good policy in a completely new environment but only passive data collection is available. However, our lower bound shows that there is not much to be gained even if the logging policy is known to be a good policy: If the goal is to improve the suboptimality level of the logging policy, by saying, a factor of two, the exponential sample complexity lower bound still applies.

From a statistical perspective, the main difference between these two data collection mechanisms is that for policy-induced data-collection the distribution of (S_i, A_i) will depend on the specific MDP instance, while this is not the case for *SA*-sampling. As we shall see, this makes BPO under *SA*-sampling provably exponentially more efficient.

4   Lower Bounds

We first give a lower bound on the sample complexity for BPO when the data available for learning is obtained by following some logging policy:

Theorem 1 (Exponential sample complexity with policy-induced data collection in discounted problems). *For any positive integers* S *and* A, *discount factor* $\gamma \in [0, 1)$ *and a pair* (ε, δ) *such that* $0 < \varepsilon < 1/2$ *and* $\delta \in (0, 1)$, *any* (ε, δ)-*sound algorithm needs at least* $\Omega(A^{\min(S-1, H_{\gamma, 2\varepsilon}+1)} \ln(1/\delta))$ *episodes of any length with policy-induced data collection for MDPs with* S *states and* A *actions under the* γ-*discounted total expected reward criterion. The result remains true if the MDPs are restricted to have deterministic transitions.*

Remark 1. *Random rewards are not essential in proving Theorem 1 as long as stochastic transitions are allowed: First, the proof can be modified to use Bernoulli rewards and stochastic transitions can be used to emulate Bernoulli rewards. Note also that for ρ-subgaussian random reward, the sample complexity in Theorem 1 becomes* $\Omega(\max\{1, \rho^2\} A^{\min(S-1, H+1)} \ln(1/\delta))$. *The maximum appears exactly because stochastic transitions can emulate Bernoulli rewards.*

Simplifying things a bit, the theorem states that the sample complexity is exponential as the number of states and the planning horizon grow together and without a limit. Note that this is in striking contrast to sample complexity of learning actively, or even with *SA*-sampling, as we shall soon discuss it. The hard MDP instance used to construct the lower bound is adopted from the combination lock problem (Whitehead, 1991). The detailed proof is provided in the supplementary material, as are the proofs of the other statements. By realizing that tabular MDPs can be considered as using one-hot features, the exponential lower bound still hold for linear function approximation.

Corollary 1 (Exponential sample complexity with linear function approximation in discounted problem). *Let* d, A *be positive integers. Then the same result as Theorem 1 with* S *replaced by* d *holds when the data collection happens for some MDP* $M \in \mathcal{M}(\mathcal{S}, [A])$ *and in addition to the dataset the learner is also given access to a featuremap* $\phi : \mathcal{S} \to \mathbb{R}^d$ *such that for every policy* π *of this MDP, there exists* $\theta \in \mathbb{R}^d$ *such that* $v^\pi(s) = \phi(s)^\top \theta, \forall s \in \mathcal{S}$, *where* v^π *is the value function of* π *in* M. *The result also remains true if the MDPs are restricted to have deterministic transitions.*

One may wonder about whether this exponential complexity can be avoided if more is assumed about the logging policy. In particular, one may hope that improving on an already good logging policy (i.e., one that is close to optimal) should be easier. Our next result shows that this is not the case.

Corollary 2 (Warm starts do not help). *Fix $\pi_{\log}$, $0 < \varepsilon < 1/2$, $\delta \in (0,1)$, S and A as before. Let $\mathcal{M}_{2\varepsilon}^{\pi_{\log}}$ denote a set of MDPs with deterministic transitions, state space $\mathcal{S} = [S]$ and action space $\mathcal{A} = [A]$ such that $\pi_{\log}$ is 2ε-optimal for all $M \in \mathcal{M}_{2\varepsilon}^{\pi_{\log}}$. Then for any length of sampled episodes, any (ε, δ)-sound algorithm needs at least $\Omega(A^{\min(S-1, H+1)} \ln(1/\delta))$ episodes, where $H = H_{\gamma, 2\varepsilon}$.*

The third corollary shows that when the logging policy is not uniform, the lower bound gets worse.

Corollary 3 (The uniform policy is the best logging policy). *If $\pi_{\log}$ is not uniform at every state, then the sample complexity in Theorem 1 increases, and in particular, A^u can be replaced by $\max_{\mathcal{S}' \subset \mathcal{S}, |\mathcal{S}'| = u} \prod_{s \in \mathcal{S}'} \max_a \frac{1}{\pi_{\log}(a|s)}$ where $u = \min(S - 1, H + 1)$.*

For fixed-horizon policy optimization, we have the following result similar to Theorem 1

Theorem 2 (Exponential sample complexity with policy-induced data collection in finite-horizon problems). *For any positive integers S and A, planning horizon $H > 0$ and a pair (ε, δ) such that $0 < \varepsilon < 1/2$ and $\delta \in (0,1)$, any (ε, δ)-sound algorithm needs at least $\Omega(A^{\min(S-1, H)} \ln(1/\delta)/\varepsilon^2)$ episodes with policy-induced data collection for MDPs with S states and A actions under the H-horizon total expected reward criterion. The result also remains true if the MDPs are restricted to have deterministic transitions.*

Remark 1 and Corollaries 1 to 3 also remain essentially true; we omit these to save space. As shown in the next result, the sample complexity could be even worse in average reward MDPs. The different sample complexities of the average reward problems and the two previous settings can be explained as follows. In discounted and finite-horizon problems, rewards beyond the planning horizon do not have to be observed in data to find a near optimal policy. In contrast, this is not the case for the average reward criterion: rewards at states that are "hard" to reach may have to be observed enough in data. Thus, the fact that the planning horizon is finite is crucial for a finite sample complexity.

Theorem 3 (Infinite sample complexity with policy-induced data collection in average reward problems). *For any positive integers S and A, any pair (ε, δ) such that $0 < \varepsilon < 1/2$ and $\delta \in (0,1)$, the sample complexity of BPO with policy-induced data collection for MDPs with S states and A actions under the average reward criterion is infinite.*

For SA-sampling, the sample complexity becomes polynomial in the relevant quantities: Staying with discounted problems, this is implied by the results of Agarwal et al. (2020a) who study plug-in methods when

a generative model is used to generate the same number of observations for each state-action pair. In particular, they show that in this setting, if N samples are available in each state-action pair then the plug-in algorithm will find a policy with $v^\pi \geq v^* - \varepsilon \mathbf{1}$ provided that $N = \Omega(\ln \frac{SA}{(1-\gamma)\delta}/(\varepsilon^2(1-\gamma)^3))$. This implies a sample complexity upper bound of size $\tilde{O}(SAH^3 \ln(1/\delta)/\varepsilon^2)$ where $H = 1/(1-\gamma)$, though for the stronger requirement that π is optimal not only from μ, but from each state. The upper bound is essentially matched by a lower bound by Sidford et al. (2018) who prove their result in Section D of their paper using a reduction to a result of Azar et al. (2013) that stated a similar sample complexity lower bound for estimating the optimal value. Our result is stronger than these results, which require the algorithm to produce a "globally good" policy, i.e., a policy that is near-optimal no matter the initial state, while in our result, the policy needs to be good only at a *fixed* initial state distribution.

Theorem 4. *Fix any $\gamma_0 > 0$. Then, there exist some constants $c_0, c_1 > 0$ such that for any $\gamma \in [\gamma_0, 1)$, any positive integers S and A, $\delta \in (0,1)$, and $0 < \varepsilon \leq c_0/(1-\gamma)$, the sample size n needed by any (ε, δ)-sound algorithm that produces as output a memoryless policy and works with SA-sampling for MDPs with S states and A actions under the γ-discounted expected reward criterion must be so that is at least $c_1 \frac{SA \ln(1/(4\delta))}{\varepsilon^2(1-\gamma)^3}$.*

Our proof for the lower bound essentially follows the ideas of Azar et al. (2013), but an effort was made to make the proof more streamlined. In particular, the new proof uses Le Cam's method. We leave it for future work to extend the result to algorithms whose output is not restricted to memoryless policies.

5 Upper Bounds

We now consider the "plug-in" algorithm for BPO and the discounted total expected reward criterion and will present a result for it that shows that this simple approach essentially matches the sample complexity lower bound of Theorem 1. For simplicity, we assume that the reward function is known.[3] Given a batch of data, the plug-in algorithm uses sample means to construct estimates for the transition probabilities. These can then be fed into any MDP solver to get a policy. The plug-in method is an obvious first choice that has proved its value in a number of different settings (Agarwal et al., 2020a; Azar et al., 2013; Cui and Yang, 2020; Li et al., 2020; Ren et al., 2021; Xiao et al., 2021).

---

[3]As noted also, e.g., by Agarwal et al. (2020a), the sample size requirements stemming from the need to obtain a sufficiently accurate estimate of the reward function is a lower order term compared to that needed to accurately estimate the transition probabilities.

Chenjun Xiao[1,2]   Ilbin Lee[1]   Bo Dai[2]   Dale Schuurmans[1,2]   Csaba Szepesvári[1,3]

For the details, let $\mathcal{D} = (S_i, A_i, R_i, S'_i)_{i=0}^{n-1}$ be the data available to the algorithm. We let

$$N(s, a, s') = \sum_{i=0}^{n-1} \mathbb{I}(S_i = s, A_i = a, S'_i = s')$$

denote the number of transitions observed in the data from s to s' while action a is used. We also let $N(s, a) = \sum_{s'} N(s, a, s')$ be the number of times the pair (s, a) is seen in the data. Provided that the visit count $N(s, a)$ is positive, we let

$$\hat{P}(s'|s, a) = \frac{N(s, a, s')}{N(s, a)}$$

be the estimated probability of transitioning to s' given that a is taken in state s. We let $\hat{P}(s'|s, a) = 0$ for all $s' \in \mathcal{S}$ when $N(s, a)$ is zero.[4] The plug-in algorithm returns a policy by solving the planning problem defined with $(\hat{P}, r)$, exploiting that planning algorithms need only the mean rewards and the transition probabilities (Puterman, 2014). By slightly abusing the definitions, we will hence treat $(\hat{P}, r)$ as an MDP and denote it by $\hat{M}$. In the result stated below we also allow a little slack for the planner; i.e., the planner is allowed to return a policy which is $\varepsilon_{\mathrm{opt}}$-optimal.

The main result for this section is as follows:

Theorem 5. *Fix* S, A, *an MDP* $M \in \mathcal{M}(S, A)$ *and a distribution* μ *on the state space of* M. *Suppose* $\delta > 0$, $\varepsilon > 0$, *and* $\varepsilon_{opt} > 0$. *Assume that the data is collected by following the uniform policy and it consists of* m *episodes, each of length* $H = H_{\gamma, (1-\gamma)\varepsilon/(2\gamma)}$. *Let* $\hat{\pi}$ *be any deterministic,* ε_{opt}-*optimal policy for* $\hat{M} = (\hat{P}, r)$ *where* $\hat{P}$ *is the sample-mean based empirical estimate of the transition probabilities based on the data collected. Then if*

$$m = \tilde{\Omega}\left(\frac{S^3 A^{\min(H,S)+2} \ln\frac{1}{\delta}}{(1-\gamma)^4 \varepsilon^2}\right),$$

where $\tilde{\Omega}$ *hides log factors of* S, A *and* H, *we have* $v^{\hat{\pi}}(\mu) \geq v^*(\mu) - 4\varepsilon - \varepsilon_{opt}$ *with probability at least* $1 - \delta$.

Remark 2. *Our proof technique for the upper bound can be directly applied to the fixed* H-*horizon setting and gives an identical result.*

In summary, when the logging policy is the uniform one, the plug-in algorithm will find an $O(\varepsilon)$ optimal policy with $\tilde{O}(S^3 A^{\min(H,S)+2} \ln(1/\delta)/(\varepsilon^2(1-\gamma)^4))$ episodes.

---

[4]We note that the particular values chosen here do not have an essential effect on the results. For example, when $\hat{P}(\cdot|s, a)$ is the uniform distribution over $\mathcal{S}$, it will only effect the constant factor in Theorem 5 (see Eq. (15) in the appendix).

We note that for BPO with policy-induced data collection, it is not possible to directly apply a reductionist approach based on analysis for SA-sampling, which requires a uniform lower bound on the number of transitions observed at all the state-action pairs, which could even be infinite. The key to avoid this is to show that the ratio of visit probabilities for an arbitrary policy vs. the uniform policy in step t is at most $A^{\min(t+1, S)}$. We provide the proof of Theorem 5, as well as an analogous result for the pessimistic policy (Jin et al., 2021; Buckman et al., 2021; Kidambi et al., 2020; Yu et al., 2020; Kumar et al., 2020; Liu et al., 2020; Xiao et al., 2021) in the supplementary material.

6 Related Work

Our work is motivated by that of Zanette (2021) who considers the sample complexity of BPO in MDPs and linear function approximation. One of the main results in this paper (Theorem 2) is that the $(1/2, 1/2)$ sample complexity with a "reinforced" policy-induced data collection in MDPs whose optimal action-value function is realizable with a d-dimensional feature map given to the learner is at least $\Omega((1/(1-\gamma))^{\frac{d-1}{2}})$. The "reinforced" data collection gives the learner full access to the transition kernel and rewards at states that are reachable from the start states with the policy (or policies) chosen. Thus, the learner here has more information than in our setting, but the problem is made hard by the presence of linear function approximators. As noted by Zanette, the same setting is trivially easy in the finite horizon setting, thus the result shows a separation between the infinite and finite horizon settings. The weakness of this separation is that the "reinforced data collection" mechanism is unrealistic. A second result in the paper (Theorem 3) shows that in the presence of function approximation, even under SA-sampling, the sample complexity is still exponential in d (as in Theorem 2 mentioned above) even when the features are so that the action-value functions of *any* policy can be realized. This exponential sample complexity is to be contrasted with the fully polynomial result available for the same setting when a generative model is available (Lattimore et al., 2020). Thus, this result shows an exponential separation between passive and active learning. It is interesting to note that this separation disappears in the tabular setting under SA-sampling.

For linear function approximation under SA-sampling a number of authors show related exponential (or infinite) sample complexity when the sampling distribution is chosen in a semi-adversarial way (Amortila et al., 2020; Wang et al., 2021; Chen et al., 2021) in the sense that it can be chosen to be the worst distribution among those that provide *good coverage in the feature space* (expressed as a condition on the minimum eigenvalue of

the feature second moment matrix). The main message of these results is that good coverage in the feature space is insufficient for sample-efficient BPO. Since the hard examples in these works are tabular MDPs with $O(d)$ state-action pairs, the uniform distribution over the state-action space is sufficient to guarantee polynomial sample complexity in the same "hard MDPs". Hence, these hardness results also have a distinctly different feature than the hardness result we present.

A different line of research can be traced back to the work of Li et al. (2015) who were concerned with statistically efficient batch policy evaluation (BPE) with policy-induced data-collection. The significance of this work for our paper is that at the end of the paper the authors of this work added a sidenote stating that the *sample complexity of finite horizon BPE must be exponential in the horizon*. Their example is a combination-lock type MDP, which served as an inspiration for the constructions we use in our lower bound proofs. No arguments are made for the suitability of the lower-bound for BPO, nor is a formal proof given for the exponential sample complexity for BPE. As such, our work can be seen as the careful examination of this remark in this paper and its adoption to BPO. A closely related, but weaker observation, is that the (vanilla) importance sampling estimators for BPE suffer an exponential blow-up of the variance (Guo et al., 2017), a phenomenon that Liu et al. (2018) call the *curse of horizon* in BPE. This exponential dependence is also pointed out by Jiang and Li (2016), who provide a lower bound on the asymptotic variance of any unbiased estimator in BPE.

Lately, much effort has been devoted to "breaking" this aforementioned curse. The basis of these works is the observation that if sufficient coverage for the state-action space is provided by the logging policy, the curse should not apply (i.e., plug-in estimators should work well). Considering finite-horizon problems for now, the coverage condition is usually expressed as a lower bound d_m on the smallest visit probabilities of the logging policy. The main result here, due to Yin et al. (2021a), is that the sample-complexity (or, better, episode-complexity) of BPO, with an inhomogeneous H-step transition structure and up to constant and logarithmic factors, is $H^3/(d_m\varepsilon^2)$, achieved by the plug-in estimator. According to Yin et al. (2021b), this complexity continues to hold for the discounted setting when it represents the "step complexity" (as opposed to "episode complexity"). The same work also removes a factor of H both from the lower and upper bounds for the finite horizon setting with homogeneous transitions. A further strengthening of the results for the homogeneous setting is due to Ren et al. (2021) who remove an additional H factor under the assumption that the total reward in every episode belongs

to the $[0, 1]$ interval. These results justify the use of coverage as a way of describing the inherent hardness of BPO. These results are complementary to ours. The lower bound in these works for fixed d_m is achieved by keeping the number of state-action pairs free, while we consider sample complexity for a fixed number of state-action pairs.

An alternative approach to characterize the sample-complexity of BPO is followed by Jin et al. (2021) who, for the inhomogeneous transition kernel, finite-horizon setting, consider a *weighted error metric*. While their primary interest is in obtaining results for linear function approximation, their result can be simplified back to the tabular setting. Their main result then shows that the minimax expected value of this weighted metric is lower bounded by a universal constant, while the pessimistic algorithm can match this bound with polynomial factors. Note that the results that are phrased with the help of the minimum coverage probability can also be rewritten as results on the minimax error for a weighted error, where the weights would include the minimum coverage probabilities. All these results are complementary to each other.

Average reward BPO with a parametric policy class for finite MDPs using policy-induced data is considered by Liao et al. (2020). The authors derive an "efficient" value estimator, and the policy returned is defined as the one that achieves the largest estimated value. An upper bound on the suboptimality of the policy returned is given in terms of a number of quantities that relate to the policy parameterization provided that a coverage condition is satisfied similar to the coverage assumption discussed above.

Finally, we note that there is extensive literature on BPE; the reader is referred to the works of (Yin et al., 2021a; Yin and Wang, 2020; Ren et al., 2021; Uehara et al., 2021; Pananjady and Wainwright, 2020) and the references therein. The most relevant works for SA-sampling are concerned with the sample complexity of planning with generative models; see, e.g., (Azar et al., 2013; Agarwal et al., 2020a; Yin and Wang, 2020) and the references therein.

7   Conclusion

The main motivation for our paper is to fill a substantial gap in the literature of batch policy optimization: While the most natural setting for batch policy optimization is when the data is obtained by following some policy, the sample complexity, the minimum number of observations necessary and sufficient to find a good policy, of batch policy optimization with data obtained this way has never been formally studied. Our results characterize how hard BPO under passive

Chenjun Xiao[1,2]   **Ilbin Lee**[1]   **Bo Dai**[2]   **Dale Schuurmans**[1,2]   **Csaba Szepesvari**[1,3]

data collection exactly is and how the difficulty scales as the problem parameter changes. While our main result that, with a finite planning horizon, the sample complexity scales exponentially is perhaps somewhat expected, this has never been formally established and should therefore be a valuable contribution to the field. In fact, both the lower and the upper bound required considerable work to be rigorously establish and that the sample complexity is finite is less obvious in light of the previous results that involved "minimum coverage" as a superficial argument with these results suggest that the sample complexity could grow without bound if some state-action pairs have arbitrary small visit probabilities. That these results, as far as the details are concerned, are non-obvious is also shown by the gap that we could not close between the upper and lower sample complexity bounds. Another non-obvious insight of our work is that warm starts provably cannot help in reducing the sample complexity. Our results should be given even more significance by the fact that the tabular setting provides the foundation for most of the insights that lead to better algorithms in RL.

8   Acknowledgments

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *COLT*, pages 67–83, 2020a.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR, 2020b.

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the Wang-Foster-Kakade lower bound for the discounted setting. *arXiv preprint 2011.01075*, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient q-learning with low switching cost.

In *Advances in Neural Information Processing Systems*, volume 32, 2019.

David Brandfonbrener, William F Whitney, Rajesh Ranganath, and Joan Bruna. Offline rl without off-policy evaluation. *arXiv preprint arXiv:2106.08909*, 2021.

Jacob Buckman, Carles Gelada, and Marc G. Bellemare. The importance of pessimism in fixed-dataset policy optimization. In *ICLR*, 2021.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Lin Chen, Bruno Scherrer, and Peter L. Bartlett. Infinite-horizon offline reinforcement learning with linear function approximation: Curse of dimensionality and algorithm. *arXiv preprint 2103.09847*, 2021.

Qiwen Cui and Lin Yang. Is plug-in solver sample-efficient for feature-based reinforcement learning? In *NeurIPS*, volume 33, pages 6015–6026, 2020.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gomez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, et al. Rl unplugged: A suite of benchmarks for offline reinforcement learning. *arXiv preprint arXiv:2006.13888*, 2020.

Zhaohan Daniel Guo, Philip S Thomas, and Emma Brunskill. Using options and covariance testing for long horizon off-policy policy evaluation. In *NeurIPS*, pages 2489–2498, 2017.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *ICML*, pages 652–661, 2016.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *ICML*, 2021.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages 865–891, 2021.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *NeurIPS*, volume 33, pages 21810–21823, 2020.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *NeurIPS*, volume 33, pages 1179–1191, 2020.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *ICML*, pages 5662–5670, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *NeurIPS*, 33:12861–12872, 2020.

L. Li, R. Munos, and Cs. Szepesvári. Toward minimax off-policy value estimation. In *AISTATS*, pages 608–616, 2015.

Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward Markov decision processes. *arXiv 2007.11771*, 2020.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *NeurIPS*, pages 5361–5371, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In *NeurIPS*, 2020.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

Georg Ostrovski, Pablo Samuel Castro, and Will Dabney. The difficulty of passive learning in deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Ashwin Pananjady and Martin J Wainwright. Instance-dependent bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585, 2020.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rongjun Qin, Songyi Gao, Xingyuan Zhang, Zhen Xu, Shengkai Huang, Zewen Li, Weinan Zhang, and Yang Yu. Neorl: A near real-world benchmark for offline reinforcement learning. *arXiv preprint arXiv:2102.00714*, 2021.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *arXiv preprint arXiv:2103.14077*, 2021.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *NeurIPS*, pages 5192–5202, 2018.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *ICLR*, 2021.

Steven D Whitehead. Complexity and cooperation in q-learning. In *International Conference on Machine Learning*, pages 363–367, 1991.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Chenjun Xiao, Yifan Wu, Tor Lattimore, Bo Dai, Jincheng Mei, Lihong Li, Csaba Szepesvári, and

Chenjun Xiao[1,2]   **Ilbin Lee**[1]   **Bo Dai**[2]   **Dale Schuurmans**[1,2]   **Csaba Szepesvari**[1,3]

Dale Schuurmans. On the optimality of batch policy optimization algorithms. In *ICML*, 2021.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021a.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *arXiv preprint arXiv:2106.04895*, 2021b.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *AISTATS*, pages 3948–3958, 2020.

Ming Yin and Yu-Xiang Wang. Optimal uniform ope and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings. *arXiv preprint arXiv:2105.06029*, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *AISTATS*, pages 1567–1575, 2021a.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*, 2021b.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: model-based offline policy optimization. In *NeurIPS*, pages 14129–14142, 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. COMBO: conservative offline model-based policy optimization. *arXiv preprint arXiv:2102.08363*, 2021.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. In *ICML*, 2021.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, pages 12402–12412, 2021.

Appendix: The Curse of Passive Data Collection in Batch Reinforcement Learning

The purpose of this appendix is to give the proofs for the results in the main paper.

A   An absolute bound on the state-action probability ratios under the uniform logging policy

For a policy π, $t \geq 0$, $(s, a) \in \mathcal{S} \times \mathcal{A}$ let

$$\nu_{\mu,t}^{\pi}(s, a) := \mathbb{P}^{\pi}(S_t = s, A_t = a | S_0 \sim \mu).$$

As noted beforehand, ratios of these marginal probabilities appear in previous upper (and lower) bounds on how well the value of a target policy π_{trg} can be estimated given data from a logging policy π_{log}. To minimize clutter, let $\nu_{\mu,t}^{\mathrm{trg}}$ stand for $\nu_{\mu,t}^{\pi_{\mathrm{trg}}}$ and, similarly, let $\nu_{\mu,t}^{\mathrm{log}}$ stand for $\nu_{\mu,t}^{\pi_{\mathrm{log}}}$. The purpose of this section is to present a short calculation that bounds $\frac{\nu_{\mu,t}^{\mathrm{trg}}(s,a)}{\nu_{\mu,t}^{\mathrm{log}}(s,a)}$, which is the ratio that appears in the previously mentioned bounds. In particular, we bound this ratio for the uniform logging policy when $\pi_{\mathrm{log}}(a|s) = 1/\mathrm{A}$.

Proposition 1. *When π_{log} is the uniform policy, for any $t \geq 0$, $(s, a) \in \mathcal{S} \times \mathcal{A}$ and π_{trg} is any target policy,*

$$\frac{\nu_{\mu,t}^{\mathrm{trg}}(s, a)}{\nu_{\mu,t}^{\mathrm{log}}(s, a)} \leq \mathrm{A}^{\min(t+1,\mathrm{S})}. \tag{1}$$

Proof. First we prove that the ratio is bounded by A^{t+1}. Fix an arbitrary pair $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$. Let $s_{0:t}$ denote a sequence $(s_0, \ldots, s_t)$ of states and let $a_{0:t}$ denote a sequence $(a_0, \ldots, a_t)$ of actions. We have

$$\nu_{\mu,t}^{\mathrm{trg}}(s_t, a_t) = \sum_{\substack{s_{0:t-1} \\ a_{0:t-1}}} \mu(s_0)\pi_{\mathrm{trg}}(a_0|s_0)p(s_1|s_0, a_0) \ldots \pi_{\mathrm{trg}}(a_{t-1}|s_{t-1})p(s_t|s_{t-1}, a_{t-1})\pi_{\mathrm{trg}}(a_t|s_t)$$

$$\leq \sum_{\substack{s_{0:t-1} \\ a_{0:t-1}}} \mathrm{A}^{t+1} \mu(s_0)\pi_{\mathrm{log}}(a_0|s_0)p(s_1|s_0, a_0) \ldots \pi_{\mathrm{log}}(a_{t-1}|s_{t-1})p(s_t|s_{t-1}, a_{t-1})\pi_{\mathrm{log}}(a_t|s_t)$$

$$= \mathrm{A}^{t+1} \nu_{\mu,t}^{\mathrm{log}}(s_t, a_t).$$

Dividing both sides by $\nu_{\mu,t}^{\mathrm{log}}(s_t, a_t)$ gives the desired bound. The inequality is tight when there is only one possible path $(s_0, a_0, s_1, a_1, \ldots, s_t, a_t)$ to (s_t, a_t) in an MDP and the target policy is the deterministic policy taking the actions in the unique path.

We now show that the ratio on the left-hand side of Eq. (1) is also bounded by A^{S}. For this, let DET be the set of stationary deterministic policies over $\mathcal{S}$ and $\mathcal{A}$. Then, for $(s, a) \in \mathcal{S} \times \mathcal{A}$ we have

$$\nu_{\mu,t}^{\mathrm{trg}}(s, a) \leq \sum_{\pi \in \mathrm{DET}} \nu_{\mu,t}^{\pi}(s, a) = \mathrm{A}^{\mathrm{S}} \frac{1}{\mathrm{A}^{\mathrm{S}}} \sum_{\pi \in \mathrm{DET}} \nu_{\mu,t}^{\pi}(s, a) = \mathrm{A}^{\mathrm{S}} \nu_{\mu,t}^{\pi_{\mathrm{log}}}(s, a), \tag{2}$$

where the first inequality follows because $\nu_{\mu,t}^{\pi_{\mathrm{trg}}}(s, a) \leq \max_{\pi} \nu_{\mu,t}^{\pi}(s, a) = \max_{\pi \in \mathrm{DET}} \nu_{\mu,t}^{\pi}(s, a)$, while the last follows because the uniform policy and the uniform mixture of all deterministic policies are the same. To see the latter, note that if $\mathbb{P}$ is the probability distribution induced by the interconnection of the uniform mixture of deterministic policies and the MDP over the space of state-action histories $(S_0, A_0, S_1, A_1, \ldots) \in (\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$, for any t, $a \in \mathcal{A}$, and $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$, letting $H_t = (S_0, A_0, \ldots, S_{t-1}, A_{t-1}, S_t)$, we have

$$\mathbb{P}(A_t = a | H_t = h_t) = \frac{1}{\mathrm{A}^{\mathrm{S}}} \sum_{\pi \in \mathrm{DET}} \mathbb{P}^{\pi}(A_t = a | H_t = h_t) = \frac{1}{\mathrm{A}^{\mathrm{S}}} \sum_{\pi \in \mathrm{DET}} \mathbb{I}(a = \pi(s_t)) = \frac{1}{\mathrm{A}^{\mathrm{S}}} \mathrm{A}^{\mathrm{S}-1} = \frac{1}{\mathrm{A}}.$$

The statement follows because t, h_t, and a were arbitrary and the probability measure induced by the interconnection of a policy and the MDP is unique over the canonical probability space of the MDP whose sample set is the set of state-action trajectories.

The inequality in Eq. (2) is tight under the same condition as before: when there is only one possible path to (s, a) in an MDP and the target policy is the deterministic policy taking the actions in the unique path. $\square$

From the proof it is clear that the result continues to hold even if the target policy depends on the full history.

Chenjun Xiao[1,2]   Ilbin Lee[1]   Bo Dai[2]   Dale Schuurmans[1,2]   Csaba Szepesvari[1,3]
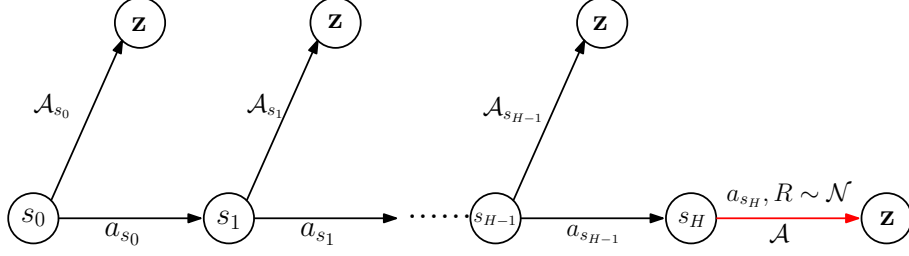
Figure 1: Illustration of the MDPs used in the proof of Theorem 1. For $\varepsilon > 0$, Let $H = H_{\gamma, 2\varepsilon}$. The state space consists of two parts $\mathcal{S} = \{s_0, s_1, \ldots, s_H\} \cup \{z\}$, where s_0 is the initial state, z is a self-absorbing state. For any $s \in \mathcal{S}$, let $a_s = \arg\min_a \pi_{\log}(a|s)$ be the action with minimal chance of being selected by $\pi_{\log}$, and $\mathcal{A}_s = \mathcal{A} \setminus \{a_s\}$. The transitions and rewards are as follows: State z is absorbing under any action. For $i \in \{0, \ldots, H-1\}$, at state s_i under action a_{s_i} the MDP transits to s_{i+1}, while it transits to z under any other actions. From s_H, the next state is also z under any action. The rewards are deterministically zero for any state-action pair except when the state is s_H, and action a_{s_H} is taken, when it is random with either a positive or negative mean.

B   Lower Bound Proofs

Before these proofs, an equivalent form of (ε, δ)-soundness will be useful to consider. Recall that $\mathcal{L}$ is (ε, δ)-sound on instance (M, G) if

$$\mathbb{P}_{\mathcal{D} \sim G} \left(v^{\mathcal{L}(\mathcal{D})}(\mu) > v^*(\mu) - \varepsilon \right) > 1 - \delta \,,$$

Now, $\mathbb{P}_{\mathcal{D} \sim G} \left(v^{\mathcal{L}(\mathcal{D})}(\mu) > v^*(\mu) - \varepsilon \right) = 1 - \mathbb{P}_{\mathcal{D} \sim G} \left(v^{\mathcal{L}(\mathcal{D})}(\mu) \leq v^*(\mu) - \varepsilon \right)$. Hence, $\mathcal{L}$ is (ε, δ)-sound on instance (M, G) if and only if

$$\mathbb{P}_{\mathcal{D} \sim G} \left(v^{\mathcal{L}(\mathcal{D})}(\mu) \leq v^*(\mu) - \varepsilon \right) < \delta \,.$$

Finally, by reordering, this last display is equivalent to

$$\mathbb{P}_{\mathcal{D} \sim G} \left(v^*(\mu) - v^{\mathcal{L}(\mathcal{D})}(\mu) \geq \varepsilon \right) < \delta \,.$$

Thus, $\mathcal{L}$ is *not* (ε, δ) sound on (M, G) if

$$\mathbb{P}_{\mathcal{D} \sim G} \left(v^*(\mu) - v^{\mathcal{L}(\mathcal{D})}(\mu) \geq \varepsilon \right) \geq \delta \,. \tag{3}$$

We will need some basic concepts, definitions, and results from information theory. For two probability measures, P and Q over a common measurable space, we use $D_{\mathrm{KL}}(P, Q)$ to denote the relative entropy (or Kullback-Leibler divergence) of P with respect to Q, which is infinite when P is not absolutely continuous with respect to Q, and otherwise it is defined as $D(P||Q) = \int \log(\frac{dP}{dQ})dP$, where dP/dQ is the Radon-Nikodym derivative of P with respect to Q. By abusing notation, we will use $P(X)$ to denote the probability distribution $P(X \in \cdot)$ of a random element X under probability measure P. For jointly distributed random elements X and Y, we let $P(X|Y)$ denote the conditional distribution of X given Y, $P(X \in \cdot|Y)$, which is Y-measurable. With this, the *chain rule* for relative entropy states that

$$D_{\mathrm{KL}}(P(X, Y), Q(X, Y)) = \int D_{\mathrm{KL}}(P(X|Y), Q(X|Y))dP + D_{\mathrm{KL}}(P(Y), Q(Y)) \,,$$

which, of course, extends to any number of jointly distributed random elements.

We will also need the following result, which is given, for example, as Theorem 14.2 in the book of Lattimore and Szepesvári (2020).

Lemma 1 (Bretagnolle–Huber inequality). *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-D_{\mathrm{KL}}(P, Q) \right) \,, \tag{4}$$

where $A^c = \Omega \setminus A$ is the complement of A.

Proof of Theorem 1. We first consider the case where $S \geq H := H_{\gamma, 2\varepsilon} + 2$.

We let $\{s_0, s_1, \ldots, s_H, z\}$ be arbitrary, distinct states and choose μ to be the distribution that is concentrated at s_0. Let $\pi_{\log}$ be the distribution used to construct the batch (which could depend on μ). We now define two MDPs, $M_1, M_{-1} \in \mathcal{M}(S, A)$ (cf. Fig. 1). For any $s \in \mathcal{S}$, let $a_s = \arg\min_a \pi_{\log}(a|s)$ be the action with the minimal chance of being selected by $\pi_{\log}$. Note that $\pi_{\log}(a_s|s) \leq 1/A$.

The transition structure in the two MDPs are identical, the transitions are deterministic and almost all the rewards are also the same with the exception of one transition. The details are as follows. State z is absorbing: For any action taken at z, the next state is z. For $i < H$, s_i is followed by s_{i+1} when a_{s_i} is taken, while the next state is z when any other action is taken at this state. At s_H under any action, the next state is z. The rewards are deterministically zero for any state-action pair except when the state is s_H and action a_{s_H} is taken at this state. In this case, the reward R is drawn from a Gaussian with mean $\alpha \in \{-1, +1\}$ in MDP M_α.

We will use v_α^π, v_α^* and $\nu_{\mu,\alpha}$ to denote the value function of a policy π on M_α, the optimal value function on M_α, and the discounted occupancy measure on M_α with μ as the initial state distribution, respectively. Note that $v_1^*(s_0) = \gamma^H \geq \gamma^{\frac{\ln(1/(2\varepsilon))}{\ln(1/\gamma)}} = 2\varepsilon$, where the first inequality is because $\gamma \leq 1$ and $H \leq \frac{\ln(1/(2\varepsilon))}{\ln(1/\gamma)}$ by its definition. Note also that $v_{-1}^*(s_0) = 0$.

Fix $\pi_{\log}$ and the episode lengths $\mathbf{h} = (h_0, \ldots, h_{m-1})$. We show that if the number of episodes m is too small, then no algorithm will be sound both on M_1 and M_{-1}.

For this fix an arbitrary BPO algorithm $\mathcal{L}$. Let the data collected by following the logging policy $\pi_{\log}$ be $\mathcal{D} = (S_i, A_i, R_i, S_i')_{i=0}^{n-1}$. Let π be the output of $\mathcal{L}$. Let $\mathbb{P}_\alpha$ be the distribution over $(\mathcal{D}, \pi)$ induced by using $\pi_{\log}$ on M_α with episode lengths $\mathbf{h}$ and μ and then running $\mathcal{L}$ on $\mathcal{D}$ to produce π. Note that both $\mathbb{P}_1$ and $\mathbb{P}_{-1}$ share the same measure space. Let $\mathbb{E}_\alpha$ be the expectation operator for $\mathbb{P}_\alpha$.

Define the event $E = \{v_1^\pi(s_0) < \varepsilon\}$. Let E^c be the complement of E. Let $a \vee b$ denote the maximum of a and b. We first prove the following claim:

Claim: If

$$\mathbb{P}_1(E) \vee \mathbb{P}_{-1}(E^c) \geq \delta \tag{5}$$

then $\mathcal{L}$ is not (ε, δ)-sound.

Proof of the claim. By Eq. (3), $\mathcal{L}$ is not (ε, δ)-sound if

$$\mathbb{P}_1(v_1^*(s_0) - v_1^\pi(s_0) \geq \varepsilon) \vee \mathbb{P}_{-1}(v_{-1}^*(s_0) - v_{-1}^\pi(s_0) \geq \varepsilon) \geq \delta \,.$$

By $v_1^*(s_0) \geq 2\varepsilon$, we have

$$\mathbb{P}_1(v_1^*(s_0) - v_1^\pi(s_0) \geq \varepsilon) \geq \mathbb{P}_1(v_1^\pi(s_0) \leq \varepsilon) \geq \mathbb{P}_1(v_1^\pi(s_0) < \varepsilon) = \mathbb{P}_1(E) \,.$$

Similarly, by $v_{-1}^*(s_0) = 0$, we have

$$\mathbb{P}_{-1}(v_{-1}^*(s_0) - v_{-1}^\pi(s_0) \geq \varepsilon) = \mathbb{P}_{-1}(v_{-1}^\pi(s_0) \leq -\varepsilon) \geq \mathbb{P}_{-1}(v_1^\pi(s_0) \geq \varepsilon) = \mathbb{P}_{-1}(E^c) \,,$$

where the inequality follows because if $v_1^\pi(s_0) \geq \varepsilon$ holds then since $v_1^\pi(s_0) = \langle \nu_1^\pi, r_1^\pi \rangle = \nu_1^\pi(s_H, a_{s_H}) r_1^\pi(s_H, a_{s_H}) = \nu_1^\pi(s_H, a_{s_H})$ and since the transitions in M_1 and M_{-1} are same, we have $\nu_{-1}^\pi(s_H, a_{s_H}) = \nu_1^\pi(s_H, a_{s_H}) \geq \varepsilon$ and therefore $v_{-1}^\pi(s_0) = -\nu_{-1}^\pi(s_H, a_{s_H}) \leq -\varepsilon$.

Putting things together, we get that

$$\mathbb{P}_1(v_1^*(s_0) - v_1^\pi(s_0) \geq \varepsilon) \vee \mathbb{P}_{-1}(v_{-1}^*(s_0) - v_{-1}^\pi(s_0) \geq \varepsilon) \geq \mathbb{P}_1(E) \vee \mathbb{P}_{-1}(E^c) \geq \delta,$$

where the last inequality follows by our assumption. $\qquad\square$

Chenjun Xiao[1,2] **Ilbin Lee**[1] **Bo Dai**[2] **Dale Schuurmans**[1,2] **Csaba Szepesvari**[1,3]

It remains to prove that Eq. (5) holds. For this, note that by the Bretagnolle-Huber inequality (Lemma 1) we have,

$$\mathbb{P}_1(E) \vee \mathbb{P}_{-1}(E^c) \geq \frac{\mathbb{P}_1(E) + \mathbb{P}_{-1}(E^c)}{2} \geq \frac{1}{4}\exp(-D_{\mathrm{KL}}(\mathbb{P}_1, \mathbb{P}_{-1})). \tag{6}$$

It remains to upper bound $D_{\mathrm{KL}}(\mathbb{P}_1, \mathbb{P}_{-1})$. Let $U_0 = S_0, U_1 = A_0, U_2 = R_0, U_3 = S_0', U_4 = S_1, \ldots, U_{4(n-1)} = S_{n-1}'$. Further, for $0 \leq j \leq 4n-1$ let $U_{0:j} = (U_0, \ldots, U_j)$ and let $U_{0:-1}$ stand for a "dummy" (trivial) random element. By the chain rule for relative entropy,[5]

$$D_{\mathrm{KL}}(\mathbb{P}_1, \mathbb{P}_{-1}) = \mathbb{E}_1[D_{\mathrm{KL}}(\mathbb{P}_1(\pi|U_{0:4(n-1)}), \mathbb{P}_{-1}(\pi|U_{0:4(n-1)}))]$$
$$+ \sum_{j=0}^{4(n-1)} \mathbb{E}_1[D_{\mathrm{KL}}(\mathbb{P}_1(U_j|U_{0:j-1}), \mathbb{P}_{-1}(U_j|U_{0:j-1}))].$$

Note that, $\mathbb{P}_1$-almost surely, $\mathbb{P}_1(\pi|U_{0:4(n-1)}) = \mathbb{P}_{-1}(\pi|U_{0:4(n-1)})$ since, by definition, $\mathcal{L}$ assigns a fixed probability distribution over the policies to any possible dataset. For $0 \leq j \leq 4(n-1)$, let $D_j = D_{\mathrm{KL}}(\mathbb{P}_1(U_j|U_{0:j-1}), \mathbb{P}_{-1}(U_j|U_{0:j-1}))$. Since the only difference between M_1 and M_{-1} is in the reward distribution corresponding to taking action a_{s_H} in state s_H, unless $j = 4i+2$ for some $i \in [n]$ and $S_i = s_H, A_i = a_{s_H}$, we have $D_j = 0$ $\mathbb{P}_1$-almost surely. Further, when $j = 4i + 2$, $\mathbb{P}_1$-almost surely we have $D_j = \mathbb{I}\{S_i = s_H, A_i = a_{s_H}\}(1 - (-1))^2/2 = 2\mathbb{I}\{S_i = s_H, A_i = a_{s_H}\}$ by the formula for the relative entropy between $\mathcal{N}(1,1)$ and $\mathcal{N}(-1,1)$. Therefore,

$$D_{\mathrm{KL}}(\mathbb{P}_1, \mathbb{P}_{-1}) = 2\mathbb{E}_1\left[\sum_{i=0}^{n-1} \mathbb{I}\{S_i = s_H, A_i = a_{s_H}\}\right] \leq 2m\mathbb{P}_1(S_H = s_H, A_H = a_{s_H}) \leq \frac{2m}{\mathrm{A}^{H+1}},$$

where the first inequality follows from that, by the construction of M_1, s_H can be visited only in the Hth step of *any* episode, the data in distinct episodes are identically distributed, and there are at most m episodes. The second inequality follows because

$$\mathbb{P}_1(S_H = s_H, A_H = a_{s_H}) = \mathbb{P}_1(A_H = a_{s_H}|S_H = s_H)\mathbb{P}_1(S_H = s_H)$$
$$= \mathbb{P}_1(A_H = a_{s_H}|S_H = s_H)\mathbb{P}_1(A_{H-1} = a_{s_{H-1}}, S_{H-1} = s_{H-1})$$
$$= \mathbb{P}_1(A_H = a_{s_H}|S_H = s_H)\mathbb{P}_1(A_{H-1} = a_{s_{H-1}}|S_{H-1} = s_{H-1}) \ldots \mathbb{P}_1(A_0 = a_{s_0}|S_0 = s_0)$$
$$= \pi_{\log}(a_{s_0}|s_0) \ldots \pi_{\log}(a_{s_H}|s_H) \leq \frac{1}{\mathrm{A}^{H+1}},$$

where the last inequality follows by the choice of a_{s_i}, $i \in [H+1]$. Plugging the upper bound on $D_{\mathrm{KL}}(\mathbb{P}_1, \mathbb{P}_{-1})$ into Eq. (6), we get that

$$\mathbb{P}_1(E) \vee \mathbb{P}_{-1}(E^c) \geq \frac{1}{4}\exp(-2m\mathrm{A}^{-(H+1)})$$

which is larger than δ if $m \leq (\mathrm{A}^{H+1}\ln\frac{1}{4\delta})/2$. The result then follows by our previous claim.

To prove the result for $\mathrm{S} < H_{\gamma,2\varepsilon} + 2$, we use the same construction as described above with $H_{\gamma,2\varepsilon'} = \mathrm{S} - 2 < H_{\gamma,2\varepsilon}$ for some $\varepsilon' \geq \varepsilon$. Then any learning algorithm $\mathcal{L}$ needs at least $(\mathrm{A}^{H_{\gamma,2\varepsilon'}+1}\ln\frac{1}{4\delta})/2$ episodes to be (ε', δ)-sound. To be (ε, δ)-sound it needs at least the same amount of data. This finishes the proof. □

Proof of Corollary 2. The result directly follows from the lower bound construction in Theorem 1. □

Proof of Corollary 3. We first consider the case $\mathrm{S} \geq H + 2$. Recall that $H = H_{\gamma,2\varepsilon}$. Define $\mathcal{S}_{\min}$ to be the set of $H + 1$ states that have the smallest $\pi_{\log,\min}(s)$ values, where we let $\pi_{\log,\min}(s) = \min_{a \in \mathcal{A}} \pi_{\log}(a|s)$. Construct the same MDPs as in the proof of Theorem 1 using the states in $\mathcal{S}_{\min}$ to form the chain. Then, the same proof holds with A^{H+1} replaced by

$$\prod_{s \in \mathcal{S}_{\min}} \frac{1}{\pi_{\log,\min}(s)}. \tag{7}$$

---

[5]Here, we use a notation common in information theory, which uses $P(X)$ ($P(X|Y)$) to denote the distribution of X induced by P (the conditional distribution of X, given Y, induced by P, respectively).

Since $\pi_{\log}$ is not uniform, the above value is strictly greater than A^{H+1}.

For the case $S < H + 2$, let $\mathcal{S}_{\min}$ to be the set of $S - 1$ states that have the smallest $\pi_{\log,\min}(s)$ values. Construct the same MDPs as above. Then the same arguments hold as in the last part of the proof of Theorem 1 except that A^{S-1} is replaced by equation 7, which is strictly greater than A^{S-1}. This concludes the proof of the corollary. $\quad\square$

Proof of Theorem 2. This proof is similar to the proof of Theorem 1. We first consider the case where $S \geq H + 1$. We construct the same MDPs as in the proof of Theorem 1 except that the chain consists of H states, that is, ending at s_{H-1} and the hidden reward R is at $(s_{H-1}, a_{s_{H-1}})$. The logging policy $\pi_{\log}$ collects m trajectories with length H as the dataset $\mathcal{D} = (S_i, A_i, R_i, S_i')_{i=0}^{mH-1}$, where $S_0 = S_H = \cdots = S_{(m-1)H} = s_0$. Now we consider two MDPs $M_\alpha \in \mathcal{M}, \alpha \in \{2\varepsilon, -2\varepsilon\}$, where the reward $R \sim \mathcal{N}(\alpha, 1)$ on M_α.

We use the same notation as in the proof of Theorem 1. Define the event $E = \{v_{2\varepsilon}^\pi(s_0) < \varepsilon\}$. Then, by following the same arguments we can show that $\mathcal{L}$ is not (ε, δ)-sound on $M_{2\varepsilon}$ if $\mathbb{P}_{2\varepsilon}(E) \geq \delta$ and that $\mathcal{L}$ is not (ε, δ)-sound on $M_{-2\varepsilon}$ if $\mathbb{P}_{-2\varepsilon}(E^c) \geq \delta$.

By the Bretagnolle–Huber inequality, we have

$$\max\{\mathbb{P}_{2\varepsilon}(E), \mathbb{P}_{-2\varepsilon}(E^c)\} \geq \frac{\mathbb{P}_{2\varepsilon}(E) + \mathbb{P}_{-2\varepsilon}(E^c)}{2} \geq \frac{1}{4}\exp(-D_{\mathrm{KL}}(\mathbb{P}_{2\varepsilon}, \mathbb{P}_{-2\varepsilon})).$$

Similarly as in the proof of Theorem 1, we obtain

$$
\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_{2\varepsilon}, \mathbb{P}_{-2\varepsilon}) &= 8\varepsilon^2 \mathbb{E}_{2\varepsilon}\left[\sum_{i=0}^{mH-1} \mathbb{I}\{S_i = s_{H-1}, A_i = a_{s_{H-1}}\}\right] \\
&= 8m\varepsilon^2 \mathbb{E}_{2\varepsilon}\left[\sum_{i=0}^{H-1} \mathbb{I}\{S_i = s_{H-1}, A_i = a_{s_{H-1}}\}\right] \\
&= 8m\varepsilon^2 \mathbb{P}_{2\varepsilon}(S_{H-1} = s_{H-1}, A_{H-1} = a_{s_{H-1}}) \leq \frac{8m\varepsilon^2}{A^H},
\end{aligned}
$$

where the second equality is obtained by the fact that the episodes are independently sampled. Combining the above together we have that if $m \leq \frac{A^H \ln \frac{1}{4\delta}}{8\varepsilon^2}$, $\max\{\mathbb{P}_{2\varepsilon}(E), \mathbb{P}_{-2\varepsilon}(E^c)\} \geq \delta$, which means that $\mathcal{L}$ is not (ε, δ)-sound on either $M_{2\varepsilon}$ or $M_{-2\varepsilon}$.

To prove the result for $S \leq H$, we use the same construction as described above with $H' = S - 1 < H$. Then any learning algorithm $\mathcal{L}$ needs at least $\frac{A^{H'} \ln \frac{1}{4\delta}}{8\varepsilon^2}$ trajectories to be (ε, δ)-sound. This finishes the proof. $\quad\square$

Proof of Theorem 3. We use MDPs similar to those in the proof of Theorem 1 but with some key differences. Let the state space consist of three parts $\mathcal{S} = \{s_0, s_1, \ldots, s_{H-1}\} \cup \{y\} \cup \{z\}$, where $H = S - 2$. Consider μ concentrated on s_0. For any $s \in \mathcal{S}$, let $a_s = \arg\min_a \pi_{\log}(a|s)$. At s_i for $i \in \{0, \ldots, H-2\}$, it transits to s_{i+1} by taking a_{s_i} and transits to z by taking any other actions, where z is an absorbing state. At s_{H-1}, by taking any action it transits to y with probability $p > 0$ and goes back to s_0 with probability $1 - p$. y is also an absorbing state, but there is a reward R for any action in y. The rewards are deterministically zero for any other state-action pairs.

Now consider two such MDPs $M_\alpha \in \mathcal{M}, \alpha \in \{2\varepsilon, -2\varepsilon\}$, where the reward $R \sim \mathcal{N}(\alpha, 1)$ on M_α. We keep using the same notation v_α and $\nu_{\mu,\alpha}$, the latter of which denotes the occupancy measure on M_α with μ as the initial state distribution. Also, we use the rest of notation in the proof of Theorem 1. Recall that π is the output policy of a learning algorithm $\mathcal{L}$.

Define the event $E = \{v_{2\varepsilon}^\pi(s_0) < \varepsilon\}$.

Claim: If

$$\mathbb{P}_{2\varepsilon}(E) \vee \mathbb{P}_{-2\varepsilon}(E^c) \geq \delta \tag{8}$$

then $\mathcal{L}$ is not (ε, δ)-sound.

Chenjun Xiao[1,2] Ilbin Lee[1] Bo Dai[2] Dale Schuurmans[1,2] Csaba Szepesvari[1,3]

Proof of the claim. By Eq. (3), $\mathcal{L}$ is not (ε, δ)-sound if

$$\mathbb{P}_{2\varepsilon}(v_{2\varepsilon}^*(s_0) - v_{2\varepsilon}^\pi(s_0) \geq \varepsilon) \vee \mathbb{P}_{-2\varepsilon}(v_{-2\varepsilon}^*(s_0) - v_{-2\varepsilon}^\pi(s_0) \geq \varepsilon) \geq \delta \,.$$

By the definition of $M_{2\varepsilon}$, the optimal policy is choosing a_{s_i} at s_i for $i \in \{0, \ldots, H-2\}$. We have $v_{2\varepsilon}^*(s_0) = 2\varepsilon$, because p is positive, and thus, the optimal policy reaches y in finite steps with probability one. Thus, we have

$$\mathbb{P}_{2\varepsilon}(v_{2\varepsilon}^*(s_0) - v_{2\varepsilon}^\pi(s_0) \geq \varepsilon) = \mathbb{P}_{2\varepsilon}(2\varepsilon - v_{2\varepsilon}^\pi(s_0) \geq \varepsilon) \geq \mathbb{P}_{2\varepsilon}(v_{2\varepsilon}^\pi(s_0) < \varepsilon) = \mathbb{P}_{2\varepsilon}(E) \,.$$

Similarly, by $v_{-2\varepsilon}^*(s_0) = 0$, we have

$$\mathbb{P}_{-2\varepsilon}(v_{-2\varepsilon}^*(s_0) - v_{-2\varepsilon}^\pi(s_0) \geq \varepsilon) = \mathbb{P}_{-2\varepsilon}(v_{-2\varepsilon}^\pi(s_0) \leq -\varepsilon) \geq \mathbb{P}_{-2\varepsilon}(v_{2\varepsilon}^\pi(s_0) \geq \varepsilon) = \mathbb{P}_{-2\varepsilon}(E^c) \,,$$

where the inequality follows because if $v_{2\varepsilon}^\pi(s_0) \geq \varepsilon$ holds then since $v_{2\varepsilon}^\pi(s_0) = \langle \nu_{2\varepsilon}^\pi, r_{2\varepsilon}^\pi \rangle = 2\varepsilon \nu_{2\varepsilon}^\pi(y)$ and since the transitions in $M_{2\varepsilon}$ and $M_{-2\varepsilon}$ are same, we have $\nu_{-2\varepsilon}^\pi(y) = \nu_{2\varepsilon}^\pi(y) \geq 1/2$ and therefore $v_{-2\varepsilon}^\pi(s_0) = -2\varepsilon \nu_{-2\varepsilon}^\pi(y) \leq -\varepsilon$.

Putting things together, we get that

$$\mathbb{P}_{2\varepsilon}(v_{2\varepsilon}^*(s_0) - v_{2\varepsilon}^\pi(s_0) \geq \varepsilon) \vee \mathbb{P}_{-2\varepsilon}(v_{-2\varepsilon}^*(s_0) - v_{-2\varepsilon}^\pi(s_0) \geq \varepsilon) \geq \mathbb{P}_{2\varepsilon}(E) \vee \mathbb{P}_{-2\varepsilon}(E^c) \geq \delta,$$

where the last inequality follows by our assumption. $\square$

Following the same arguments in the proof of Theorem 1, we have

$$D_{\mathrm{KL}}(\mathbb{P}_{2\varepsilon}, \mathbb{P}_{-2\varepsilon}) = 8\varepsilon^2 \mathbb{E}_{2\varepsilon}\left[\sum_{i=0}^{n-1} \mathbb{I}\{S_i = y\}\right] = 8\varepsilon^2 \sum_{i=0}^{n-1} \mathbb{P}_{2\varepsilon}\{S_i = y\}$$

$$= 8\varepsilon^2 p \sum_{i=1}^{n-1} \mathbb{P}_{2\varepsilon}\{S_{i-1} = s_{H-1}\} \leq \frac{8\varepsilon^2 np}{\mathrm{A}^{H-1}} \,.$$

Combining the above together and using the Bretagnolle–Huber inequality (Lemma 1) as we did in the proof of Theorem 1, we have that if $n \leq \frac{\mathrm{A}^{H-1} \ln \frac{1}{4\delta}}{8\varepsilon^2 p}$, then $\mathcal{L}$ is not (ε, δ)-sound on either $M_{2\varepsilon}$ or $M_{-2\varepsilon}$. We obtain the result by sending p to zero from the right hand side. $\square$

For the proof of Theorem 4, we will need some results on the relative entropy between Bernoulli distributions, which we present now. Let $\mathrm{Ber}(p)$ denote the Bernoulli distribution with parameter $p \in [0, 1]$. As it is well known (and not hard to see from the definition),

$$D(\mathrm{Ber}(p), \mathrm{Ber}(q)) = d(p, q)$$

where $d(p, q)$ is the so-called *binary relative entropy function*, which is defined as

$$d(p, q) = p \log(p/q) + (1-p) \log((1-p)/(1-q)) \,.$$

Proposition 2. *For $p, q \in (0, 1)$, defining p^* to be p or q depending on which is further away from $1/2$,*

$$d(p, q) \leq \frac{(p-q)^2}{2p^*(1-p^*)} \,. \tag{9}$$

Proof. Let R be the unnormalized negentropy over $[0, \infty)^2$. Then, by Theorem 26.12 of the book of Lattimore and Szepesvári (2020), for any $x, y \in (0, \infty)^2$,

$$D_R(x, y) = \frac{1}{2}\|x - y\|_{\nabla R(z)}^2$$

for some z on the line segment connecting x to y. We have $R(z) = z_1 \log(z_1) + z_2 \log(z_2) - z_1 - z_2$. Hence, $\nabla R(z) = [\log(z_1), \log(z_2)]^\top$ and $\nabla R(z) = \mathrm{diag}(1/z_1, 1/z_2)$, both defined for $z \in (0, \infty)^2$. Thus,

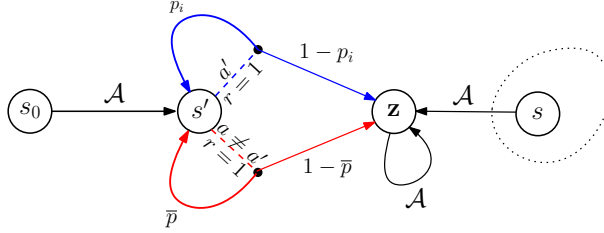$$D_R(x, y) = \frac{(x_1 - y_1)^2}{2z_1} + \frac{(x_2 - y_2)^2}{2z_2} \,.$$

Figure 2: Illustration of the MDPs used in the proof of Theorem 4. The initial distribution μ concentrates on state $\{s_0\}$. The pair (s', a') is the one where $\mu_{\log}$ takes on the smallest value (which is below $1/(SA)$) and without loss of generality $s' \neq s_0$ and taking any action in s_0 makes the next state s'. We have $p_0 < \bar{p} < p_1$, all in the $[1/2, 1)$ interval. In MDP M_i with $i \in \{0, 1\}$, the probability of transitioning under action a' from s' to z, an absorbing state, is p_i, while with probability $1 - p_i$, the next state is s'. All other actions use probability $\bar{p}$ at this state. All other states under any action lead to z. The rewards are deterministically zero except at state s', when all actions yield a reward of one, regardless of the identity of the next state.

Now choosing $x = (p, 1 - p)$, $y = (q, 1 - q)$, we see that $x, y \in (0, \infty)^2$ if $p, q \in (0, 1)$. In this case, with some $\alpha \in [0, 1]$, $z = \alpha x + (1 - \alpha)y = (\alpha p + (1 - \alpha)q, \alpha(1 - p) + (1 - \alpha)(1 - q))^\top = (\alpha p + (1 - \alpha)q, 1 - (\alpha p + (1 - \alpha)q))^\top$. Hence, $z_2 = 1 - z_1$ and

$$d(p, q) = \frac{(p - q)^2}{2z_1} + \frac{(p - q)^2}{2(1 - z_1)} = \frac{(p - q)^2}{2z_1(1 - z_1)}.$$

Now, $z_1(1 - z_1) \geq p^*(1 - p^*)$ (the function $z \mapsto z(1 - z)$ has a maximum at $z = 1/2$ and is decreasing on "either side" of the line $z = 1/2$). Putting things together, we get

$$d(p, q) = \frac{(p - q)^2}{2z_1(1 - z_1)} \leq \frac{(p - q)^2}{2p^*(1 - p^*)}.$$

$\square$

Now we re-state Theorem 4 and prove it.

Theorem 6 (Restatement of Theorem 4). *Fix any $\gamma_0 > 0$. Then, there exist some constants $c_0, c_1 > 0$ such that for any $\gamma \in [\gamma_0, 1)$, any positive integers S and A, $\delta \in (0, 1)$, and $0 < \varepsilon \leq c_0/(1 - \gamma)$, the sample size n needed by any (ε, δ)-sound algorithm that produces as output a memoryless policy and works with SA-sampling for MDPs with S states and A actions under the γ-discounted expected reward criterion must be so that is at least $c_1 \frac{SA \ln(1/(4\delta))}{\varepsilon^2(1-\gamma)^3}$.*

Proof of Theorem 6. The proof also uses Le Cam's method, just like Theorem 1. At the heart of the proof is a gadget with a self-looping state which was introduced by Azar et al. (2013) to give a lower bound on the sample complexity of estimating the optimal value function in the simulation setting where the estimate's error is measured with its worst-case error.

The idea of the proof is illustrated by Fig. 2. Fix an initial state distribution μ concentrated on an arbitrary state $s_0 \in \mathcal{S}$. Let $\mu_{\log}$ be the logging distribution chosen based on μ and let (s', a') be any state-action pair that has the minimum sampling probability under $\mu_{\log}$. Note that $\mu_{\log}(s', a') \leq 1/(SA)$. Assume that $s' \neq s_0$. As we shall see by the end of the proof, there is no loss of generality in making this assumption (when $s' = s_0$, the lower bound would be larger).

We construct two MDPs as follows. The reward structures of the two MDPs are completely identical and the transition structures are also identical except for when action a' is taken at state s'. In particular, in both MDPs, the rewards are identically zero except at state s', where for any action the reward incurred is one. The transition structures are as follows: Let $p_0 < \bar{p} < p_1$ be in $(0, 1)$, to be determined later. At s_0, by taking any action the system transits to s' deterministically. At s', for any $a \in \mathcal{A} \setminus \{a'\}$, taking action a leads to s' as the next state with probability $\bar{p}$ and to z with probability $1 - \bar{p}$, where z is an absorbing state. The transition under a' at s' is similar, except that in M_i ($i \in \{0, 1\}$), the probability that the next state is s' is p_i (and the probability that the

Chenjun Xiao[1,2]   Ilbin Lee[1]   Bo Dai[2]   Dale Schuurmans[1,2]   Csaba Szepesvari[1,3]

next state is z is $1 - p_i$). At any state $s \in \mathcal{S} \setminus \{s_0, s'\}$, taking any action moves the system to z deterministically. The optimal policy at s' in M_1 is to pull the action a', while in M_0 all the other actions are optimal. It is easy to see that in any of these MDPs, for $a \in \mathcal{A}$,

$$q^*(s', a) = \frac{1}{1 - \gamma P(s'|s', a)} \ .$$

Now we select $p_0, \bar{p}$, and p_1. Let b be a constant such that $1 < b < \frac{1 - \gamma/2}{1 - \gamma}$. We will choose a specific value for b at the end of the proof. The values of $p_0, \tilde{p}$ and p_1 will depend on b. In particular, we choose $\tilde{p} \in (1/2, 1)$ so that

$$b = \frac{1 - \gamma \tilde{p}}{1 - \gamma},$$

while we set $p_0 = \tilde{p}$. Then $p_0 = (1 - b + \gamma b)/\gamma$. Note that $p_0 > 1/2$ by its choice. Let $f(p) = \frac{\gamma}{1 - \gamma p}$. Note that for any deterministic policy π, $v^{\pi}(\mu) = f(P(s'|s', \pi(s')))$ and also $v^*(\mu) = f(\bar{p})$ in MDP M_0 and $v^*(\mu) = f(p_1)$ in MDP M_1.

By Taylor's theorem, for some $p \in [p_0, p_1]$, we have

$$f(p_1) = f(p_0) + f'(p)(p_1 - p_0) \geq f(p_0) + f'(p_0)(p_1 - p_0),$$

where the inequality follows by $p_1 > p_0$ and the fact that f' is increasing. Thus, if $p_1 - p_0 \geq 4\varepsilon/f'(p_0)$, we have $f(p_1) \geq f(p_0) + 4\varepsilon$. Because of the choice of p_0,

$$f'(p_0) = \frac{\gamma^2}{(1 - \gamma p_0)^2} = \frac{\gamma^2}{(1 - \gamma)^2 b^2} \ .$$

We let $p_1 = p_0 + 4\varepsilon/f'(p_0)$. Then, we have $p_1 < 1$ given that $\varepsilon \leq c_0/(1 - \gamma) := \frac{\gamma(b-1)}{8(1-\gamma)b^2}$, because

$$p_1 - 1 = p_0 + 4\varepsilon/f'(p_0) - 1 = \frac{1 - b + \gamma b}{\gamma} + \frac{4(1-\gamma)^2 \varepsilon b^2}{\gamma^2} - 1$$

$$= \frac{4(1-\gamma)^2 \varepsilon b^2 + \gamma(\gamma - 1)(b - 1)}{\gamma^2} \leq \frac{(\gamma - 1)(b - 1)}{2\gamma} < 0, \tag{10}$$

where the first inequality is due to the choice of ε. Lastly, we set $\bar{p}$ so that $f(\bar{p}) = [f(p_0) + f(p_1)]/2$ (such $\bar{p}$ uniquely exists because f is increasing and continuous). Note that $f(p_1) - f(\bar{p}) \geq 2\varepsilon$ and $f(\bar{p}) - f(p_0) \geq 2\varepsilon$.

Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be the joint probability distribution on the data and the output policy of any given learning algorithm $\mathcal{L}$, induced by μ, $\mu_{\log}$, $\mathcal{L}$, and the two MDPs M_0 and M_1, respectively. For any algorithm $\mathcal{L}$, let $E = \{\pi(a'|s') \geq 1/2\}$, where π is the output of $\mathcal{L}$.

If E is true, in M_0,

$$v^{\pi}(\mu) = \pi(a'|s')f(p_0) + (1 - \pi(a'|s'))f(\bar{p}) \leq \frac{f(p_0) + f(\bar{p})}{2} \leq \frac{(f(\bar{p}) - 2\varepsilon) + f(\bar{p})}{2} = f(\bar{p}) - \varepsilon \ .$$

Thus, $\mathcal{L}$ is not (ε, δ)-sound for M_0 if $\mathbb{P}_0(E) \geq \delta$. If E^c holds, in M_1,

$$v^{\pi}(\mu) = \pi(a'|s')f(p_1) + (1 - \pi(a'|s'))f(\bar{p}) \leq \frac{f(p_1) + f(\bar{p})}{2} \leq \frac{f(p_1) + (f(p_1) - 2\varepsilon)}{2} = f(p_1) - \varepsilon \ .$$

Therefore, if $\mathbb{P}_1(E^c) \geq \delta$, then $\mathcal{L}$ is not (ε, δ)-sound for M_1.

By the Bretagnolle-Huber inequality (Lemma 1) we have,

$$\max\{\mathbb{P}_0(E), \mathbb{P}_1(E^c)\} \geq \frac{\mathbb{P}_0(E) + \mathbb{P}_1(E^c)}{2} \geq \frac{1}{4} \exp\left(-D_{\mathrm{KL}}(\mathbb{P}_0 || \mathbb{P}_1)\right) \ .$$

Recall that n is the number of samples. Since M_0 and M_1 differ only in the state transition from (s', a'), by the chain rule of relative entropy, with a reasoning similar to that used in the proof of Theorem 1, we derive

$$D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_1) = n\mathbb{P}_0(S_i = s', A_i = a')D_{\mathrm{KL}}(\mathrm{Ber}(p_0), \mathrm{Ber}(p_1)) \leq \frac{n}{\mathrm{SA}} \cdot \frac{(p_0 - p_1)^2}{2p_1(1 - p_1)}$$

$$= \frac{n}{\text{SA}} \cdot \frac{16\varepsilon^2(1-\gamma)^4 b^4}{2\gamma^4 p_1(1-p_1)}$$
$$< \frac{n}{\text{SA}} \cdot \frac{16\varepsilon^2(1-\gamma)^3 b^4}{\gamma^3(b-1)p_0} \,,$$

where the first inequality is due to Proposition 2 and the second inequality is due to Eq. (10) and the fact that $p_0 < p_1$.

Now fix $\gamma_0 \in (0,1)$ and let $\gamma \geq \gamma_0$ and choose $b = 0.5(1 + \frac{1-\gamma_0/2}{1-\gamma_0}) \in (1, \frac{1-\gamma/2}{1-\gamma})$. Then, combining the above together and reordering show that if $n \leq c_1 \frac{\text{SA}\ln(1/(4\delta))}{\varepsilon^2(1-\gamma)^3}$ where $c_1 = \frac{\gamma_0^3(b-1)p_0}{16b^4}$, we can guarantee that $\mathcal{L}$ is not (ε, δ)-sound on either M_0 or M_1, concluding the proof. $\square$

C Upper bound proofs

We start with some extra notation. We identify the transition function P as an SA $\times$ S matrix, whose entries $P_{sa,s'}$ specify the conditional probability of transitioning to state s' starting from state s and taking action a, and the reward function r as an SA $\times 1$ reward vector. We use $\|x\|_1$ to denote the 1-norm $\sum_i |x_i|$ of $x \in \mathbb{R}^n$.

Recall first that we defined P^π to be the transition matrix on state-action pairs induced by the policy π. Define the H-step action-value function for $H > 0$ by

$$q_H^\pi = \sum_{h=0}^{H-1} (\gamma P^\pi)^h r \,.$$

We let v_H^π denote the H-step state-value function. In what follows we will need quantities for $\hat{M}$, which, in general could be any MDP that differs from M from only its transition kernel. Quantities related to $\hat{M}$ receive a "hat". For example, we use $\hat{P}$ for the transition kernel of $\hat{M}$, $\hat{P}^\pi$ for the state-action transition matrix induced by a policy π and $\hat{P}$, etc.

In subsequent proofs, we will need the following lemma, which gives two decompositions of the difference between the action-value functions on two MDPs, M and $\hat{M}$:

Lemma 2. *For any policy π, transition model $\hat{P}$, and $H > 0$,*

$$q_H^\pi - \hat{q}_H^\pi = \gamma \sum_{h=0}^{H-1} (\gamma P^\pi)^h (P - \hat{P})\hat{v}_{H-h-1}^\pi \,, \tag{11}$$

$$\hat{q}_H^\pi - q_H^\pi = \gamma \sum_{h=0}^{H-1} (\gamma \hat{P}^\pi)^h (\hat{P} - P)v_{H-h-1}^\pi \,. \tag{12}$$

Proof. By symmetry, it suffices to prove Eq. (11). For convenience, we re-express a policy π as an S $\times$ SA matrix/operator Π. In particular, as a left linear operator, Π maps $q \in \mathbb{R}^{\text{SA}}$ to $\sum_a \pi(a|\cdot)q(\cdot, a) \in \mathbb{R}^{\text{S}}$. Note that with this $P^\pi = P\Pi$, $\hat{P}^\pi = \hat{P}\Pi$, $v_h^\pi = \Pi q_h^\pi$ and $\hat{v}_h^\pi = \Pi \hat{q}_h^\pi$. To reduce clutter, as π is fixed, for the rest of the proof we drop the upper indices and just use v_h, $\hat{v}_h$, q_h and $\hat{q}_h$.

Note that for $H > 0$,

$$q_H = r + \gamma P\Pi q_{H-1}\,, \qquad \text{and}$$
$$\hat{q}_H = r + \gamma \hat{P}\Pi \hat{q}_{H-1}\,.$$

Hence,

$$q_H - \hat{q}_H = \gamma(P\Pi q_{H-1} - \hat{P}\Pi \hat{q}_{H-1})$$
$$= \gamma(P - \hat{P})\Pi \hat{q}_{H-1} + \gamma P\Pi(q_{H-1} - \hat{q}_{H-1})\,.$$

Then using $\Pi \hat{q}_{H-1} = \hat{v}_{H-1}$ and recursively expanding $q_{H-1} - \hat{q}_{H-1}$ in the same way gives the result, noting that $q_0 = r = \hat{q}_0$. $\square$

Chenjun Xiao[1,2]   **Ilbin Lee**[1]   **Bo Dai**[2]   **Dale Schuurmans**[1,2]   **Csaba Szepesvari**[1,3]

We need two standard results from the concentration of binomial random variables.

Lemma 3 (Multiplicative Chernoff Bound for the Lower Tail, Theorem 4.5 of Mitzenmacher and Upfal (2005))**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter p, $S_n = \sum_{i=1}^{n} X_i$. Then, for any $0 \leq \beta < 1$,*

$$\mathbb{P}\left(\frac{S_n}{n} \leq (1-\beta)p\right) \leq \exp\left(-\frac{\beta^2 np}{2}\right).$$

Lemma 4. *Let n be a positive integer, $p > 0$, $\delta \in (0, 1)$ such that*

$$\frac{2}{np} \ln \frac{1}{\delta} \leq \frac{1}{4}. \tag{13}$$

Let S_n be as in the previous lemma, $\hat{p} = S_n/n$. Then, with probability at least $1 - \delta$, it holds that

$$\hat{p} \geq p/2 > 0$$

while we also have

$$\frac{1}{\hat{p}} \leq \frac{1}{p} + \frac{2}{p}\sqrt{\frac{2}{np} \ln \frac{1}{\delta}}.$$

on the same $(1 - \delta)$-probability event.

In what follows, we will only need the first lower bound, $\hat{p} \geq p/2$ from above; the second is useful to optimize constants only.

Proof. According to the multiplicative Chernoff bound for the low tail (cf. Lemma 3), for any $0 < \delta \leq 1$, with probability at least $1 - \delta$, we have

$$\hat{p} \geq p - \sqrt{\frac{2p}{n} \ln \frac{1}{\delta}}.$$

Denote by $\mathcal{E}_\delta$ the event when this inequality holds. Using Eq. (13), on $\mathcal{E}_\delta$ we have

$$\hat{p} \geq p - \sqrt{\frac{2p}{n} \ln \frac{1}{\delta}} = p\left(1 - \sqrt{\frac{2}{pn} \ln \frac{1}{\delta}}\right) \geq p\left(1 - \frac{1}{2}\right) = \frac{p}{2} > 0,$$

and thus, thanks to $1/(1-x) \leq 1 + 2x$ which holds for any $x \in [0, 1/2]$,

$$\frac{1}{\hat{p}} \leq \frac{1}{p}\frac{1}{1 - \sqrt{\frac{2}{np} \ln \frac{1}{\delta}}} \leq \frac{1}{p} + \frac{2}{p}\sqrt{\frac{2}{np} \ln \frac{1}{\delta}}.$$

$\square$

Our next lemma bounds the deviation between the empirical transition kernel and the "true" one:

Lemma 5. *With probability $1 - \delta$, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\|\hat{P}(\cdot|s, a) - P(\cdot|s, a)\|_1 \leq \beta(N(s, a), \delta) \tag{14}$$

where for $u \geq 0$,

$$\beta(u, \delta) = 2\sqrt{\frac{S \ln 2 + \ln \frac{u_+(u+1)SA}{\delta}}{2u_+}},$$

where $u_+ = u \vee 1$.

Proof. By abusing notation, for $u \geq 0$, let $\beta(u) = 2\sqrt{\frac{\mathsf{S}\ln 2 + \ln\frac{u_+(u+1)}{\delta}}{2u_+}}$, where $u_+ = u \vee 1$. We will prove below the following claim:

<u>Claim:</u> For any fixed state-action pair (s, a), with probability $1 - \delta$,

$$\|\hat{P}(\cdot|s, a) - P(\cdot|s, a)\|_1 \leq \beta(N(s, a)).$$

Clearly, from this claim the lemma follows by a union bound over the state-action pairs. Hence, it remains to prove the claim.

For this fix $(s, a) \in \mathcal{S} \times \mathcal{A}$. Recall that the data $\mathcal{D} = ((S_i, A_i, R_i, S'_i)_{i \in [n]})$ that is used to estimate $\hat{P}(\cdot|s, a)$ consists of m trajectories of length H obtained by following the uniform policy $\pi_{\log}$ while the initial state is selected from μ at random. In particular, for $j \in [m]$, the jth trajectory is $(S_0^{(j)}, A_0^{(j)}, R_0^{(j)}, \dots, S_{h_j-1}^{(j)}, A_{H-1}^{(j)}, R_{H-1}^{(j)}, S_H^{(j)})$, where $S_0^{(j)} \sim \mu$, $A_t^{(j)} \sim \pi_{\log}(\cdot|S_t^{(j)})$, $(R_t^{(j)}, S_{t+1}^{(j)}) \sim Q(\cdot|S_t^{(j)}, A_t^{(j)})$. Clearly, if $q := \mathbb{P}\left(\exists 0 \leq i \leq H - 1 : S_i^{(0)} = s, A_i^{(0)} = a\right) = 0$ then $N(s, a) = 0$ holds with probability one. The claim then follows since when $N(s, a) = 0$, $\hat{P}(\cdot|s, a)$ is identically zero, hence,

$$\|\hat{P}(\cdot|s, a) - P(\cdot|s, a)\|_1 = \|P(\cdot|s, a)\|_1 = 1 \leq 1.177\ldots \leq \beta(0). \tag{15}$$

Hence, it remains to prove the claim for the case when $q > 0$, which we assume from now on. For convenience, append to the data infinitely many further trajectories, giving rise to the infinite sequence $(S_i, A_i, R_i, S'_i)_{i \geq 0}$. Let $\tau_0 = 0$ and for $u \geq 1$, let $\tau_u = \min\{i \in \mathbb{N} : i > \tau_{u-1} \text{ and } S_i = s, A_i = a\}$ be the "time" indices when (s, a) is visited, where we define the minimum of an empty set to be infinite. Since $q > 0$, almost surely $(\tau_u)_{u \geq 0}$ is a well-defined sequence of *finite* random variables. Now let $X_u = S'_{\tau_u}$ be the "next state" upon the uth visit of (s, a). Let $\hat{p}_u(s') = \frac{\sum_{v=1}^u \mathbb{I}\{X_v = s'\}}{u}$. Note that

$$\hat{P}(\cdot|s, a) = \hat{p}_{N(s,a)}(\cdot) \tag{16}$$

provided that $N(s, a) > 0$. By the Markov property, it follows that $(X_v)_{v \geq 1}$ is an i.i.d. sequence of categorical variables with common distribution $p(\cdot) := P(\cdot|s, a)$.

Now,

$$\|\hat{p}_u - p\|_1 = \max_{y \in \{-1, +1\}^{\mathcal{S}}} \langle \hat{p}_u - p, y \rangle,$$

while

$$\langle \hat{p}_u - p, y \rangle = \frac{1}{u} \sum_{v=1}^u \underbrace{y(X_v) - \sum_{s'} p(s')y(s')}_{\Delta_v}.$$

Now, $(\Delta_v)_{1 \leq v \leq u}$ is an i.i.d. sequence, $|\Delta_v| \leq 2$ for any v and $\mathbb{E}\Delta_v = 0$. Hence, by Hoeffding's inequality, with probability $1 - \delta$,

$$\frac{1}{u} \sum_{v=1}^u \Delta_v \leq 2\sqrt{\frac{\ln\frac{1}{\delta}}{2u}}.$$

Since the cardinality of $\{-1, +1\}^{\mathcal{S}}$ is 2^{S}, applying a union bound over $y \in \{-1, +1\}^{\mathcal{S}}$, we get that with probability $1 - \delta$,

$$\|\hat{p}_u - p\|_1 \leq 2\sqrt{\frac{\mathsf{S}\ln 2 + \ln\frac{1}{\delta}}{2u}}.$$

Chenjun Xiao[1,2]   Ilbin Lee[1]   Bo Dai[2]   Dale Schuurmans[1,2]   Csaba Szepesvari[1,3]

Applying another union bound over u, owing to that $\sum_{u=1}^{\infty} \frac{1}{u(u+1)} = 1$, we get that with probability $1 - \delta$, for any $u \geq 1$,

$$\|\hat{p}_u - p\|_1 \leq 2\sqrt{\frac{S \ln 2 + \ln \frac{u(u+1)}{\delta}}{2u}} = \beta(u) \,.$$

Since $\|\hat{p}_0 - p\|_1 \leq 1 \leq \beta(0)$ (cf. Eq. (15)), the claim follows by Eq. (16). $\qquad\square$

We now state a lemma that bounds, with high probability, the error of predicting the value of some fixed policy when the prediction is based on a transition kernel P' which is "close" to the true transition kernel P, where closeness is based on how often the individual state-action pairs have been visited. This notion of closeness is motivated by Lemma 5; this lemma can be used when $P' = \hat{P}$, or some other transition kernel in the vicinity of $\hat{P}$. The former will be needed in the analysis of the plug-in method presented here; while the latter will be used in the next section where we analyze the pessimistic algorithm.

Lemma 6. *Let $\delta \in (0, 1)$ and m be the number of episodes collected by the logging policy and fix any policy π. For any P' such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\|P'(\cdot|s, a) - P(\cdot|s, a)\|_1 \leq \frac{C}{\sqrt{N(s, a) \vee 1}},$$

with probability at least $1 - \delta$ for $C > 0$, we have

$$v^\pi(\mu) - v_{P'}^\pi(\mu) \leq \frac{4\gamma C S A^{\frac{\min(H,S)}{2}+1}}{(1-\gamma)^2 \sqrt{m}} + \frac{8\gamma S A}{(1-\gamma)^2} \frac{\ln \frac{SA}{\delta}}{m} + \varepsilon \,. \tag{17}$$

Proof. Note that

$$\gamma^{H_{\gamma,\varepsilon}} \leq \gamma^{1 + \frac{\ln(1/\varepsilon)}{\ln(1/\gamma)}} = \gamma\varepsilon \,.$$

Hence, for $H = H_{\gamma,(1-\gamma)\varepsilon/(2\gamma)}$,

$$\gamma^H \leq \tfrac{1}{2}\varepsilon(1-\gamma) \,.$$

Owning to that the immediate rewards belong to $[-1, 1]$, it follows that for any policy π,

$$q^\pi - q_{P'}^\pi \leq q_H^\pi - q_{P',H}^\pi + \varepsilon\mathbf{1} \,, \tag{18}$$

where we use $q_{P',H}^\pi$ to denote the H-step value function under transition model P'. Define $N_h(s, a)$ as the number of episodes when the hth state-action pair in the episode is (s, a). Note that $N(s, a) \geq N_h(s, a)$. Let $Z_h = \{(s, a) \in \mathcal{S} \times \mathcal{A} : \nu_{\mu,h}^\pi(s, a) > \frac{8}{m} \ln \frac{SA}{\delta}\}$ and let $\mathcal{F}$ be the event when

$$\frac{N_h(s, a)}{m} \geq \frac{\nu_{\mu,h}^{\pi_{\log}}(s, a)}{2}$$

holds for any $(s, a) \in Z_h$.[6] By Lemma 4, $\mathbb{P}(\mathcal{F}) \geq 1 - \delta$.

Assume that $\mathcal{F}$ holds. Combining Eq. (18) with Lemma 2, we get that on this event

$$v^\pi(\mu) - \hat{v}^\pi(\mu) \leq (\mu^\pi)^\top (q_H^\pi - q_{P',H}^\pi) + \varepsilon$$

$$= \gamma \sum_{h=0}^{H-1} \gamma^h (\nu_{\mu,h}^\pi)^\top (P - P') v_{P',H-h-1}^\pi + \varepsilon \qquad \text{(by Eq. (11))}$$

$$\leq \frac{\gamma}{1-\gamma} \sum_{h=0}^{H-1} \gamma^h \sum_{s,a} \nu_{\mu,h}^\pi(s, a) \|P(\cdot|s, a) - P'(\cdot|s, a)\|_1 + \varepsilon \qquad \text{(by } \|\hat{v}_{H-h-1}^\pi\|_\infty \leq 1/(1-\gamma)\text{)}$$

---

[6] Note that Z_h, and thus also $\mathcal{F}$ depends on π, which is the reason that the result, as stated, holds only for a fixed policy.

$$\leq \frac{\gamma}{1-\gamma} \sum_{h=0}^{H-1} \gamma^h \sum_{(s,a) \in Z_h} \nu_{\mu,h}^\pi(s,a) \|P(\cdot|s,a) - P'(\cdot|s,a)\|_1 + \frac{8\gamma SA}{m(1-\gamma)^2} \ln \frac{SA}{\delta} + \varepsilon \quad \text{(by the definition of } Z_h)$$

$$\leq \frac{\gamma}{1-\gamma} \sum_{h=0}^{H-1} \gamma^h \sum_{(s,a) \in Z_h} \sqrt{\nu_{\mu,h}^\pi(s,a)} \|P(\cdot|s,a) - P'(\cdot|s,a)\|_1 + \frac{8\gamma SA}{m(1-\gamma)^2} \ln \frac{SA}{\delta} + \varepsilon \quad \text{(by } \nu_{\mu,h}^\pi(s,a) \leq 1)$$

$$\leq \frac{\gamma A^{\min(H,S)/2}}{1-\gamma} \sum_{h=0}^{H-1} \gamma^h \sum_{(s,a) \in Z_h} \sqrt{\nu_{\mu,h}^{\pi_{\log}}(s,a)} \|P'(\cdot|s,a) - P(\cdot|s,a)\|_1 + \frac{8\gamma SA}{m(1-\gamma)^2} \ln \frac{SA}{\delta} + \varepsilon$$

$$\text{(by Proposition 1)}$$

$$\leq \frac{2\gamma A^{\min(H,S)/2} C}{1-\gamma} \sum_{h=0}^{H-1} \gamma^h \sum_{(s,a) \in Z_h} \sqrt{\nu_{\mu,h}^{\pi_{\log}}(s,a)} \frac{1}{\sqrt{N_h(s,a) \vee 1}} + \frac{8\gamma SA}{m(1-\gamma)^2} \ln \frac{SA}{\delta} + \varepsilon$$

$$\text{(by the definition of } P')$$

$$\leq \frac{2\gamma A^{\min(H,S)/2} C}{1-\gamma} \sum_{h=0}^{H-1} \gamma^h \sum_{(s,a) \in Z_h} \sqrt{\nu_{\mu,h}^{\pi_{\log}}(s,a)} \sqrt{\frac{2}{m\nu_{\mu,h}^{\pi_{\log}}(s,a)}} + \frac{8\gamma SA}{m(1-\gamma)^2} \ln \frac{SA}{\delta} + \varepsilon$$

$$\text{(by the definitions of } \mathcal{F} \text{ and } Z_h)$$

$$\leq \frac{4\gamma A^{\min(H,S)/2} C}{(1-\gamma)^2 \sqrt{m}} SA + \frac{8\gamma SA}{m(1-\gamma)^2} \ln \frac{SA}{\delta} + \varepsilon.$$

This finishes the proof. $\qquad \square$

For the plug-in method we use the previous lemma with $P' = \hat{P}$, resulting in the following corollary:

Corollary 4. *Let $\delta \in (0,1)$ and m be the number of episodes collected by the logging policy and fix any policy π. With probability at least $1 - \delta$, we have*

$$v^\pi(\mu) - v_{\hat{P}}^\pi(\mu) \leq \frac{8\gamma SA^{\frac{\min(H,S)}{2}+1}}{(1-\gamma)^2} \sqrt{\frac{S \ln 2 + \ln \frac{2n(n+1)SA}{\delta}}{2m}} + \frac{8\gamma SA}{(1-\gamma)^2} \frac{\ln \frac{2SA}{\delta}}{m} + \varepsilon.$$

Proof. Fix $\delta \in (0,1)$. Let $\mathcal{E}_\delta$ be the event when for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \leq \beta(N(s,a), \delta), \tag{19}$$

where β is defined in Lemma 5, which also gives that $\mathbb{P}(\mathcal{E}_\delta) \geq 1 - \delta$. Further, defining

$$C_\delta = 2\sqrt{\frac{S \ln 2 + \ln \frac{n(n+1)SA}{\delta}}{2}},$$

note that $\beta(u,\delta) \leq C_\delta/\sqrt{u \vee 1}$. Now, let $\mathcal{F}_\delta$ be the event when the conclusion of Lemma 6 holds. Then, on the one hand, by a union bound, $\mathbb{P}(\mathcal{E}_{\delta/2} \cap \mathcal{F}_{\delta/2}) \geq 1 - \delta$, while on the other hand on $\mathcal{E}_{\delta/2} \cap \mathcal{F}_{\delta/2}$, the condition of Lemma 6 holds for P' defined so that

$$P'(\cdot|s,a) = \begin{cases} \hat{P}(\cdot|s,a), & \text{if } \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \leq \beta(N(s,a), \delta/2); \\ P(\cdot|s,a), & \text{otherwise}. \end{cases}$$

with $C := C_{\delta/2}$.

Furthermore, on $\mathcal{E}_{\delta/2}$, $\hat{P}(\cdot|s,a) = P(\cdot|s,a)$ holds for any (s,a) pair. Hence, the result follows by replacing δ with $\delta/2$ in Eq. (17) and plugging in $C_{\delta/2}$ in place of C. $\qquad \square$

We now are ready to prove the upper bound of plug-in algorithm.

Theorem 7 (Restatement of Theorem 5). *Fix S, A, an MDP $M \in \mathcal{M}(S, A)$ and a distribution μ on the state space of M. Suppose $\delta > 0$, $\varepsilon > 0$, and $\varepsilon_{opt} > 0$. Assume that the data is collected by following the uniform policy and it consists of m episodes, each of length $H = H_{\gamma,(1-\gamma)\varepsilon/(2\gamma)}$. Let $\hat{\pi}$ be any deterministic, ε_{opt}-optimal policy*

Chenjun Xiao[1,2]   Ilbin Lee[1]   Bo Dai[2]   Dale Schuurmans[1,2]   Csaba Szepesvari[1,3]

for $\hat{M} = (\hat{P}, r)$ where $\hat{P}$ is the sample-mean based empirical estimate of the transition probabilities based on the data collected. Then if

$$m = \tilde{\Omega}\left(\frac{S^3 A^{\min(H,S)+2} \ln \frac{1}{\delta}}{(1-\gamma)^4 \varepsilon^2}\right),$$

where $\tilde{\Omega}$ hides log factors of S, A and H, we have $v^{\hat{\pi}}(\mu) \geq v^*(\mu) - 4\varepsilon - \varepsilon_{opt}$ with probability at least $1 - \delta$.

Proof. We upper bound the suboptimality gap of $\hat{\pi}$ as follows:

$$v^*(\mu) - v^{\hat{\pi}}(\mu) = v^*(\mu) - \hat{v}^{\pi^*}(\mu) + \hat{v}^{\pi^*}(\mu) - \hat{v}^{\hat{\pi}}(\mu) + \hat{v}^{\hat{\pi}}(\mu) - v^{\hat{\pi}}(\mu)$$

$$\leq v^*(\mu) - \hat{v}^{\pi^*}(\mu) + \hat{v}^{\hat{\pi}}(\mu) - v^{\hat{\pi}}(\mu) + \varepsilon_{\text{opt}}. \qquad (\hat{\pi} \text{ is } \varepsilon_{\text{opt}}\text{-optimal in } \hat{M})$$

By Corollary 4 and a union bound, with probability at least $1 - \delta$, for any deterministic policy π obtained from the data $\mathcal{D}$ we have

$$v^\pi(\mu) - \hat{v}^\pi(\mu)$$

$$\leq \frac{8\gamma S A^{\frac{\min(H,S)}{2}+1}}{(1-\gamma)^2}\sqrt{\frac{S\ln 2 + \ln\frac{2n(n+1)SA}{\delta} + S\ln A}{2m}} + \frac{8\gamma SA}{(1-\gamma)^2}\frac{\ln\frac{2SA}{\delta} + S\ln A}{m} + \varepsilon$$

$$\leq \frac{8\gamma S^{\frac{3}{2}} A^{\frac{\min(H,S)}{2}+1}}{(1-\gamma)^2}\sqrt{\frac{\ln 2 + \ln\frac{H^2 SA}{\delta} + \ln 2m + \ln A}{2m}} + \frac{8\gamma SA}{(1-\gamma)^2}\frac{\ln\frac{2SA}{\delta} + S\ln A}{m} + \varepsilon,$$

Thus, given that

$$m = \tilde{\Omega}\left(\frac{S^3 A^{\min(H,S)+2} \ln \frac{1}{\delta}}{(1-\gamma)^4 \varepsilon^2}\right),$$

where $\tilde{\Omega}$ hides log-factors, with probability at least $1 - \delta$ we have,

$$v^*(\mu) - v^{\hat{\pi}}(\mu) \leq v^*(\mu) - \hat{v}^{\pi^*}(\mu) + \hat{v}^{\hat{\pi}^*}(\mu) - v^{\hat{\pi}^*}(\mu) + \varepsilon_{\text{opt}} \leq 4\varepsilon + \varepsilon_{\text{opt}}.$$

$\square$

C.1   Pessimistic Algorithm

We present a result in this section for the "pessimistic algorithm" in the discounted total expected reward criterion to complement the results in the main text (Jin et al., 2021; Buckman et al., 2021; Kidambi et al., 2020; Yu et al., 2020; Kumar et al., 2020; Liu et al., 2020; Yu et al., 2021). The sample complexity we show is the same as for the plug-in method. While this may be off by a polynomial factor, we do not expect the pessimistic algorithm to have a major advantage over the plug-in method in the worst-case setting. In fact, the recent work of Xiao et al. (2021) established this in a rigorous fashion for the bandit setting by showing an algorithm independent lower bound that matched the upper bound for both the plug-in method and the pessimistic algorithm. As argued by Xiao et al. (2021) (and proved by Jin et al. (2021) in the context of linear MDPs, which includes tabular MDPs), the advantage of the pessimistic algorithm is that it is weighted minimax optimal with respect to a special criterion.

The pessimistic algorithm with parameters $\delta \in (0, 1)$ and $\varepsilon_{\text{opt}} > 0$ chooses a deterministic ε_{opt} policy $\tilde{\pi}$ of the MDP with reward r and transition kernel $\tilde{P}$, the latter of which is obtained via

$$\tilde{P} = \arg\min_{P' \in \mathcal{P}_\delta} v^*_{P'}(\mu),$$

where for a transition kernel P' we use $v^*_{P'}$ to denote the optimal value function in the MDP with immediate rewards r and transition kernel P', and $\mathcal{P}_\delta$ is is defined as

$$\mathcal{P}_\delta = \left\{ P' : \text{for any } (s, a) \in \mathcal{S} \times \mathcal{A}, \|\hat{P}(\cdot|s, a) - P'(\cdot|s, a)\|_1 \leq \beta(N(s, a), \delta) \right\},$$

where β is defined in Lemma 5. Recall that the same result ensures that P, the "true" transition kernel belongs to $\mathcal{P}_\delta$ with probability at least $1 - \delta$.

Theorem 8 (Pessimistic algorithm). *Fix* S, A, *an MDP* $M \in \mathcal{M}(S, A)$ *and a distribution* μ *on the state space of* M. *Suppose* $\delta > 0$, $\varepsilon > 0$, *and* $\varepsilon_{\text{opt}} > 0$. *Assume that the data is collected by following the uniform policy and it consists of* m *episodes, each of length* $H = H_{\gamma, (1-\gamma)\varepsilon/(2\gamma)}$. *Then, if*

$$m = \tilde{\Omega} \left(\frac{S^3 A^{\min(H, S)+2} \ln \frac{1}{\delta}}{(1-\gamma)^4 \varepsilon^2} \right),$$

where $\tilde{\Omega}$ *hides log factors of* S, A *and* H, *we have* $v^{\tilde{\pi}}(\mu) \geq v^*(\mu) - 2\varepsilon - \varepsilon_{opt}$ *with probability at least* $1 - \delta$, *where* $\tilde{\pi}$ *is the output of the pessimistic algorithm run with parameters* $(\delta, \varepsilon_{\text{opt}})$.

Proof. Let us denote by $v_{P'}^{\pi}$ the value function of policy π in the MDP with reward r and transition kernel P'. Let $\Delta^{\pi} = v_{\tilde{P}}^{\pi}(\mu) - v^{\pi}(\mu)$ and let π^* is an deterministic optimal policy in the "true" MDP. Such a policy exists (e.g., see Theorem 6.2.10 of Puterman (2014)). We have

$$
\begin{aligned}
v^*(\mu) - v^{\tilde{\pi}}(\mu) &= v^*(\mu) - v_{\tilde{P}}^{\tilde{\pi}}(\mu) + \Delta^{\tilde{\pi}} \\
&\leq v^*(\mu) - v_{\tilde{P}}^*(\mu) + \Delta^{\tilde{\pi}} + \varepsilon_{\text{opt}} && \text{(by the definition of } \tilde{\pi}) \\
&\leq v^*(\mu) - v_{\tilde{P}}^{\pi^*}(\mu) + \Delta^{\tilde{\pi}} + \varepsilon_{\text{opt}} && \text{(because } v_{\tilde{P}}^{\pi^*} \leq v_{\tilde{P}}^*) \\
&\leq \Delta^{\pi^*} + \Delta^{\tilde{\pi}} + \varepsilon_{\text{opt}}. && \text{(because } v^*(\mu) = v^{\pi^*}(\mu))
\end{aligned}
$$

Hence, it remains to upper bound Δ^{π^*} and $\Delta^{\tilde{\pi}}$. For this, we make the following claim:

<u>Claim:</u> Fix any policy π. Then, with probability at least $1 - \delta$, we have

$$v^{\pi}(\mu) - v_{\tilde{P}}^{\pi}(\mu) \leq \frac{16\gamma SA^{\frac{\min(H,S)}{2}+1}}{(1-\gamma)^2} \sqrt{\frac{S \ln 2 + \ln \frac{2n(n+1)SA}{\delta}}{2m}} + \frac{8\gamma SA}{(1-\gamma)^2} \frac{\ln \frac{2SA}{\delta}}{m} + \varepsilon.$$

Proof of Claim. For the latter, note that the proof of Corollary 4 can be repeated with the only change that now instead of Eq. (19), we have

$$\|P(\cdot|s,a) - \tilde{P}(\cdot|s,a)\|_1 \leq \|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1 \|\hat{P}(\cdot|s,a) - \tilde{P}(\cdot|s,a)\|_1 \leq 2\beta(N(s,a), \delta).$$

$\square$

From this claim, by a union bound over all the A^S deterministic policies, we get that with probability $1 - \delta$, for any deterministic policy π,

$$v^{\pi}(\mu) - v_{\tilde{P}}^{\pi}(\mu)$$

$$\leq \frac{16\gamma SA^{\frac{\min(H,S)}{2}+1}}{(1-\gamma)^2} \sqrt{\frac{\left(S \ln 2 + \ln \frac{2n(n+1)SA}{\delta} + S \ln A\right)}{2m}} + \frac{8\gamma SA}{(1-\gamma)^2} \frac{\ln \frac{2SA}{\delta} + S \ln A}{m} + \varepsilon.$$

Since $\tilde{\pi}$, by definition is also a deterministic policy, the last display holds with probability $1 - \delta$ for $\tilde{\pi}$ as well. Putting things together gives that

$$v^*(\mu) - v^{\tilde{\pi}}(\mu)$$

$$\leq \frac{32\gamma SA^{\frac{\min(H,S)}{2}+1}}{(1-\gamma)^2} \sqrt{\frac{S \ln 2 + \ln \frac{2n(n+1)SA}{\delta} + S \ln A}{2m}} + \frac{16\gamma SA}{(1-\gamma)^2} \frac{\ln \frac{2SA}{\delta} + S \ln A}{m} + 2\varepsilon + \varepsilon_{\text{opt}}.$$

The proof is finished by a calculation similar to that done at the end of the proof of Theorem 7. $\square$