PAC Reinforcement Learning with an Imperfect Model

Nan Jiang
Microsoft Research
New York, NY 10011
nanjiang@umich.edu

Abstract

Reinforcement learning (RL) methods have proved to be successful in many simulated environments. The common approaches, however, are often too sample intensive to be applied directly in the real world. A promising approach to addressing this issue is to train an RL agent in a simulator and transfer the solution to the real environment. When a high-fidelity simulator is available we would expect significant reduction in the amount of real trajectories needed for learning. In this work we aim at better understanding the theoretical nature of this approach. We start with a perhaps surprising result that, even if the approximate model (e.g., a simulator) only differs from the real environment in a single state-action pair (but which one is unknown), such a model could be information-theoretically useless and the sample complexity (in terms of real trajectories) still scales with the total number of states in the worst case. We investigate the hard instances and come up with natural conditions that avoid the pathological situations. We then propose two conceptually simple algorithms that enjoy polynomial sample complexity guarantees with *no dependence* on the size of the state-action space, and prove some foundational results to provide insights into this important problem.

1 Introduction

Recently, Reinforcement learning (RL) methods have achieved impressive successes in many challenging domains (Mnih et al. 2015; Heess et al. 2015; Silver et al. 2016; Levine et al. 2016; Mnih et al. 2016). Many of these successes occur in simulated environments (e.g., video / board games, simulated robotics domains), and the state-of-the-art approaches require a large number of training samples, rendering them inapplicable in non-simulator problems where data acquisition may be costly.

A promising approach to addressing this issue is to train an RL agent in a simulator and transfer the solution to the real environment, which is particularly relevant but not limited to robotics domains (Koos, Mouret, and Doncieux 2010; Cutler and How 2015; Hanna and Stone 2017). The approach faces a significant challenge that the policy trained in a simulator may have degenerate performance in the real environment due to the imperfectness of the simulator (Kober, Bagnell, and Peters 2013).

There are many aspects from which one could address this challenge. For example, the simulator and the real environment may not share the same observation spaces and we may need to learn a transfer function that corrects the mismatch, or the simulator may be significantly different from the real environment that we should only transfer useful features instead of actual policies (Rusu et al. 2016), etc. While there has been active empirical research in this area, little in theory is known in terms of when transfer is possible and what guarantees we can have.

In this paper we focus on a particular angle of this problem, and provide some foundational results under stylized assumptions to help understand the theoretical nature of this approach. We start with a simple question: given an approximate model (e.g., a simulator) that only differs from the real environment in 1 state-action pair (but which one is unknown), can we always learn a near-optimal policy by collecting significantly fewer real trajectories compared to RL from scratch, i.e., without the model? Perhaps surprisingly, the answer is no due to a lower bound. We understand and draw insights from the hard instances, and come up with natural conditions that exclude the pathological scenarios (Sec.4). Under these conditions, we describe and analyze two algorithms whose sample complexity guarantees only depend on the number of incorrect state-action pairs and have no dependence on the size of the state and action spaces (Sec.5 and 6).

2 Preliminaries

We consider episodic RL problems where the real environment is specified by a finite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, P, R, H, \mu)$. $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and for simplicity we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite but can be arbitrarily large. $P: \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function ($\Delta(\mathcal{S})$ is the probability simplex over $\mathcal{S}$, i.e., the set of all probability distributions). $R: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function; we assume rewards are non-negative. H is the horizon, and $\mu \in \Delta(\mathcal{S})$ is the initial distribution.

In general, optimal policies in the finite-horizon setting are non-stationary, i.e., they are time dependent. To keep the notations simple, w.l.o.g. we assume that each state only appears in a particular time step (or *level*) $1 \le h \le H$, and the state space can be partitioned into disjoint sets $\mathcal{S} = \bigcup_{h=1}^{H} \mathcal{S}_h$, where μ is supported on $\mathcal{S}_1$ and states

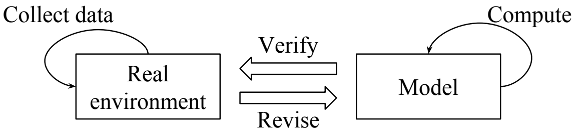


Figure 1: Protocol of how the learner interacts with the real environment and the approximate model.

in $\mathcal{S}_h$ only transition to those in $\mathcal{S}_{h+1}$. Assume that the total reward has bounded magnitude for any sequence of state-actions, i.e., $\sum_{h=1}^{H} R(s_h, a_h) \in [0, 1]$ holds for all $s_1 \in \mathcal{S}_1, a_1 \in \mathcal{A}, \ldots, s_H \in \mathcal{S}_H, a_H \in \mathcal{A}$.[1]

Given a policy $\pi : \mathcal{S} \to \mathcal{A}$, a random trajectory is generated by $s_1 \sim \mu$, and for $h = 1, \ldots, H$, $a_h \sim \pi(s_h)$, $r_h = R(s_h, a_h)$, $s_{h+1} \sim P(s_h, a_h)$. The ultimate measure of π's performance is $v_M^\pi := \mathbb{E}[\sum_{h=1}^{H} r_h \,|\, \pi]$. Also define the value function of π as $V_M^\pi(s) := \mathbb{E}[\sum_{h'=h}^{H} r_{h'} \,|\, \pi, s_h = s]$ where h is such that $s \in \mathcal{S}_h$. Note that all such value functions have bounded range $[0, 1]$.

Let $\pi_M^\star$ be the optimal policy in M, which maximizes V_M^π for all $s \in \mathcal{S}$. We use $V_M^\star$ as a shorthand for $V_M^{\pi_M^\star}$, which satisfies the Bellman optimality equation $V_M^\star = \mathcal{T} V_M^\star$, where $\mathcal{T} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is the Bellman update operator[2] $(\mathcal{T} f)(s) := \max_{a \in \mathcal{A}} \left[R(s, a) + \mathbb{E}_{s' \sim P(s,a)}[f(s')] \right]$.

3 Problem formulation

Our learning algorithm is given an approximate model $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, R, H, \mu)$ as input. For simplicity we assume that $\widehat{M}$ and M only differ in dynamics, and our analyses extend straightforwardly to approximate reward functions.

Abstraction of computation

Since this work focuses on the sample efficiency regarding the trajectories in the real environment, we abstract away the *computation* in the approximate model by assuming an oracle that can take any M' as input and return $\pi_{M'}^\star$ and $v_{M'}^\star$. Later in Algorithm 2 we will also need the oracle to return $V_{M'}^\star$, but that is often a side product of computing $\pi_{M'}^\star$.[3]

Protocol

We consider a learner that interacts with the real environment and the approximate model in an alternating manner (see Figure 1). The learner repeats the following steps until it finds a satisfying policy: (1) carrying out computation in the model, (2) collecting data in the real environment, and (3) revising the model. In Sec.7 we show that the interactivity in the protocol is crucial — under a non-interactive protocol no algorithm can achieve polynomial sample complexity.

Under this protocol, an algorithm needs to specify what computation to carry out in the approximate model, what actions to take and how much data to collect in the real environment, and how to revise the model based on the real data. For now assume that the learner can revise the model arbitrarily, i.e., it can change any entry of the transition function of $\widehat{M}$. Of course, this assumption can be unrealistic when the simulator is sophisticated and can only be accessed in a black-box manner; we relax this assumption in Sec.6.

Incorrect state-action pairs

When the approximate model is very close to the real environment, intuitively we would expect significantly reduced sample complexity (in terms of real trajectories) compared to RL from scratch. To formalize this intuition, we define a soft notion of incorrect state-action pairs in Definition 1, and use the number of incorrect state-action pairs $|\mathcal{X}_{\xi\text{-inc}}|$ to characterize the imperfectness of the model.

Definition 1 (ξ-correctness). *Given M and $\widehat{M}$, we say that (s, a) is ξ-incorrect if*[4]

$$d_{M,\widehat{M}}(s, a) := \|P(s, a) - \widehat{P}(s, a)\|_{TV} > \xi.$$

Let $\mathcal{X}_{\xi\text{-inc}}$ be the set of state-action pairs that are ξ-incorrect.

Fact 1. *$\mathcal{X}_{\xi\text{-inc}} \subseteq \mathcal{X}_{\xi'\text{-inc}}$ if $\xi \geq \xi'$.*

Remark 1. When $\mathcal{S}$ is large, it may be difficult to have a model $\widehat{M}$ that matches M on the transition probability to each next-state in most state-action pairs. In Sec.8 we give alternative definitions of $\mathcal{X}_{\xi\text{-inc}}$ that are more lenient and show that our analyses automatically extend.

Goal: no dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$

We are interested in the scenario where $|\mathcal{S}|$ and $|\mathcal{A}|$ may be very large but $|\mathcal{X}_{\xi\text{-inc}}|$ is small (for some reasonable choice of ξ). Our goal is to develop algorithms that can learn a near-optimal policy in M using polynomially many real trajectories, where the polynomial can depend on $|\mathcal{X}_{\xi\text{-inc}}|$ (and other parameters such as H) but *not* on $|\mathcal{S}|$ or $|\mathcal{A}|$.

4 Sufficient conditions for avoiding dependence on $|\mathcal{S}|$ and $|\mathcal{A}|$

Unfortunately, the goal stated above is impossible without further assumptions. In particular, there can be situations where $|\mathcal{X}_{0\text{-inc}}| = 1$ but the sample complexity is polynomial in $|\mathcal{S}|$. This result is formalized in Proposition 1. The proof builds on the lower bound from (Auer, Cesa-Bianchi, and Fischer 2002) and is deferred to Appendix A.

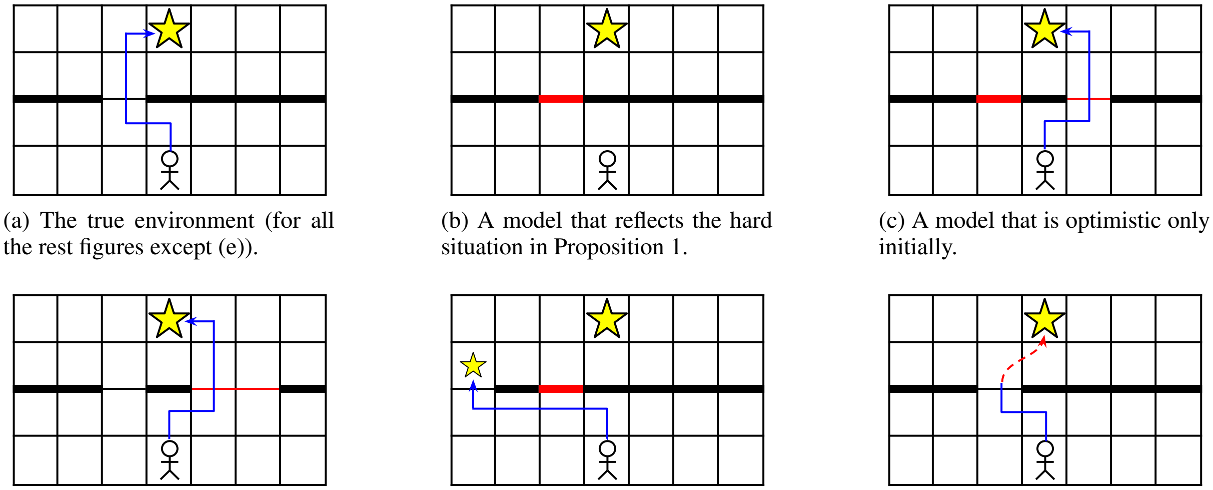
Proposition 1. *Without further assumptions, no algorithm can return an ϵ-optimal policy with a probability higher than $2/3$ and a sample complexity of poly$(|\mathcal{X}_{0\text{-inc}}|, H, 1/\epsilon)$ (recall from Fact 1 that $|\mathcal{X}_{\xi\text{-inc}}|$ is the largest when $\xi = 0$).*

[1] The assumption makes no reference to the transition dynamics, which leads to boundedness of total reward in the approximate model and its revisions to be introduced later.

[2] Let $f(s_{H+1}) \equiv 0$ so that the same equation applies to $s \in \mathcal{S}_H$.

[3] To approximate the oracle in problems with large state spaces, we can use Sparse Sampling (Kearns, Mansour, and Ng 2002) or any Monte-Carlo tree search methods that do not depend on the state branching factor (Bjarnason, Fern, and Tadepalli 2009; Grill, Valko, and Munos 2016). Practically speaking, deep RL methods, which are empirically state-of-the-art, are also reasonable approximations (Mnih et al. 2015).

[4] For distributions over finite spaces, $\|\cdot\|_{TV} = \frac{1}{2}\|\cdot\|_1$.



(a) The true environment (for all the rest figures except (e)).

(b) A model that reflects the hard situation in Proposition 1.

(c) A model that is optimistic only initially.

(d) A desired approximate model.

(e) A situation where Assumption 1 fails but Eq.(1) promises nontrivial value.

(f) A situation where Eq.(1) delivers nontrivial guarantee and Eq.(5) is vacuous.

Figure 2: An environment **(a)** and a series of models. See text in Sec.4 for the description of the domain and the models in **(b)** to **(d)**. The red lines and bars indicate the mistakes of the model, and the blue path shows an optimal policy of the model.
(e): Here we modify the real environment by removing the leftmost obstacle and putting a smaller reward behind it. Since an episode terminates at any star, the optimal policy is still to go through the middle gate, which is erroneously blocked in the model. In this case, Eq.(1) will guarantee that we can get the smaller reward, which is suboptimal but nontrivial.
(f): This model erroneously believes that the agent will be teleported to the reward upon passing through the gate (red dashed line). In this case, Definition 6 terminates a model episode when the agent goes through the gate, essentially blocking it. As a result, the value guaranteed by Theorem 2 (Eq.(5)) becomes vacuous in this case, while Theorem 1 still competes against $v^\star_M$.

Understand the hard instances

We explain the hardness result in Figure 2 and draw insights from it. Here the real environment is depicted in 2(a): the agent can move in 4 directions in this grid world, and the thick bars represent obstacles. An episode ends either when the agent runs into an obstacle or when it gets to a star (reward). The optimal policy is to go through the only gate to get the star, but the agent has no knowledge of where the gate is. Without additional information, the agent needs to try each obstacle one by one and incurs $\Omega(|\mathcal{S}|)$ sample complexity (note that we can easily scale up the problem).

The hard instance in Proposition 1 is similar to the one depicted in Figure 2(b): the model claims that there is no gate, hence no policy can get to the reward. While such a model is obviously useless, it indeed satisfies $|\mathcal{X}_{\text{0-inc}}| = 1$. Therefore, a small $|\mathcal{X}_{\text{0-inc}}|$ does not necessarily imply a useful model, and we need to impose additional conditions to exclude such degenerate cases.

Exclude the degenerate cases: a first attempt

One thing that we might notice in Figure 2(b) is that the optimal value in the model is very low ($v^\star_{\widehat{M}} = 0$). Intuitively, a model should claim that there exists a policy that achieves a high value to be any useful. Based on this observation, we might come up with the condition that $v^\star_{\widehat{M}} \geq v^\star_M$, which excludes the model in 2(b).

However, it is easy to construct another degenerate case without violating the condition; see 2(c). The model satisfies the condition by claiming the existence of a gate in the wrong location, which yields $|\mathcal{X}_{\text{0-inc}}| = 2$. Note, however, that the learner could generate such a model by choosing a location randomly, which does not require any external information hence the model is still useless.

A sufficient condition

An alternative explanation of the failure of 2(c) is that, if we follow the optimal policy suggested by the model in the real environment, we will realize that the red gate is actually an obstacle. Once this mistake is fixed, however, we literally get back to the model in 2(b).

Based on the above intuitions, we come up with the a sufficient condition that excludes all degenerate cases that obscure the desired sample complexity. Roughly speaking, we require the optimal value in the approximate model to always stay high whenever we replace the dynamics of any subset of state-action pairs in $\mathcal{X}_{\xi\text{-inc}}$ with the true dynamics (see Figure 2(d) for example). This idea is formalized in the following definitions and Assumption 1.

Definition 2 (Partially repaired model)**.** *Given M and $\widehat{M}$ which only differ in the transition dynamics P and $\widehat{P}$, and $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, define $\widehat{M}_{\mathcal{X}}$ as the MDP $(\mathcal{S}, \mathcal{A}, \widehat{P}_{\mathcal{X}}, R, H, \mu)$ where*

$$\widehat{P}_{\mathcal{X}}(s,a) := \begin{cases} P(s,a), & \textit{if } (s,a) \in \mathcal{X}, \\ \widehat{P}(s,a), & \textit{otherwise.} \end{cases}$$

Definition 3. $\mathcal{M} := \{\widehat{M}_{\mathcal{X}} : \mathcal{X} \subseteq \mathcal{X}_{\xi\text{-}inc}\}$.

Assumption 1 (Always optimistic). $\forall M' \in \mathcal{M}$, $v^\star_{M'} \geq v^\star_M$.

While Assumption 1 is sufficient, it is also too strict since it fails if any $v^\star_{M'}$ is slightly below $v^\star_M$. In the next section we will not make any explicit assumptions but rather use an agnostic version of Assumption 1: instead of requiring the algorithm to compete against $v^\star_M$ (i.e., return a policy with at least $v^\star_M - \epsilon$ value), we only require the algorithm to compete against

$$\inf_{M' \in \mathcal{M}} v^\star_{M'}. \tag{1}$$

If Assumption 1 holds, we compete against $v^\star_M$ as usual; when Assumption 1 fails, our optimality guarantee degrades gracefully with the violation. Figure 2(e) illustrates a situation where Eq.(1) delivers nontrivial guarantee while Assumption 1 fails (see caption for details).

5 Repair the model

In this section, we describe a conceptually simple algorithm whose sample complexity has no dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$. The pseudocode is given in Algorithm 1. We first walk through the pseudocode and give some intuitions, and then state and prove the sample complexity guarantee.

The outer loop of the algorithm computes $\pi_t = \pi^\star_{M_t}$, the optimal policy of the current model M_t (initialized as $\widehat{M}$), and use Monte-Carlo policy evaluation to estimate its return. If the policy's performance is satisfying, we simply output it. Otherwise, the inner loop uses the same policy to collect trajectories, and add every next-state to the dataset associated with the preceding state-action pair.

The inner loop stops whenever the size of a dataset $D_{s,a}$ for some (s,a) increases to a pre-determined threshold, n_{est}, which will be set later in the analysis. This triggers a model revision: we replace $\widehat{P}(s,a)$ with the empirical frequency of states observed in $D_{s,a}$, and produce model M_{t+1} for the next iteration of outer loop.

A straightforward analysis of the above procedure, however, would incur polynomial dependence on $|\mathcal{S}|$: we recover the multinomial distributions $\{P(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$ from sample data, and such distributions are supported on $\mathcal{S}$. In general, if we want to guarantee low estimation error (measured by e.g., total variation), we would incur dependence on size of the support.

To overcome this difficulty, note that there is no need to recover the detailed transition distribution over states. Instead, we simply need to guarantee that when we use the estimated probabilities in the Bellman update operators, the value function(s) of interest are updated correctly. That is, when we have enough samples for some $D_{s,a}$, we would like to guarantee that

$$d^f_{M,D}(s,a) := \left|\mathbb{E}_{s' \sim P(s,a)}[f(s')] - \mathbb{E}_{s' \sim D_{s,a}}[f(s')]\right| \tag{2}$$

is small for some careful choice(s) of f. In standard RL literature, f is often chosen to be $V^\star_M$, the optimal value function which we compete against (see e.g., (Kearns, Mansour, and Ng 2002)). For such a fixed f, we can use Hoeffding's inequality to guarantee concentration, and the necessary size of $D_{s,a}$ has no dependence on $|\mathcal{S}|$.

Algorithm 1 MODEL_REPAIR($\widehat{M}$)

```
$M_0 \leftarrow \widehat{M}$. $D_{s,a} \leftarrow \{\}$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$.
for $t = 0, 1, \ldots$ do
    $\pi_t \leftarrow \pi^\star_{M_t}$.
    Collect $n_{\text{eval}}$ trajectories using $\pi_t$, and let $\hat{v}^{\pi_t}_M$ be the Monte-Carlo estimate of value.
    if $\hat{v}^{\pi_t}_M \geq v^{\pi_t}_{M_t} - 7\epsilon/10$ then return $\pi_t$.
    repeat
        Collect trajectory $s_1, a_1, \ldots, s_H, a_H$ using $\pi_t$.
        $\forall\, h$, add $s_{h+1}$ to $D_{s_h,a_h}$ if $|D_{s_h,a_h}| < n_{\text{est}}$.
    until some $|D_{s,a}|$ increases to $n_{\text{est}}$ for the 1st time
    Construct $M_{t+1}$ from $\widehat{M}$ by plugging in $D_{s,a}$ for each $(s,a)$ in $\mathcal{X}'_{t+1} := \{|D_{s,a}| = n_{\text{est}}\}$.
end for
```

In our case, however, we compete against multiple value functions (Eq.(1)), and need to guarantee that all of them are updated correctly. In particular, when we have enough samples in $D_{s,a}$, we want Eq.(2) to be small for any f that is the optimal value function of some partially repaired model (Definition 2). Interestingly, there are at most $2^{|\mathcal{X}_{\xi\text{-inc}}|}$ such models (Fact 2), and by union bound we pay logarithmic dependence on the number of functions, which is $O(|\mathcal{X}_{\xi\text{-inc}}|)$. Below we state the formal guarantees for the algorithm and prove it using the intuitions described above.

Theorem 1. *Given any $\delta \in (0,1), \epsilon \in (0,1)$, run Algorithm 1 with parameters $\xi = \frac{\epsilon}{10H^2}$, $n_{eval} = \tilde{O}(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$, $n_{est} = \tilde{O}(\frac{H^4}{\epsilon^2}(|\mathcal{X}_{\xi\text{-}inc}| + \log\frac{1}{\delta}))$. With probability at least $1-\delta$ the algorithm will return a policy π_T such that $v^{\pi_T}_M \geq \inf_{M' \in \mathcal{M}} v^\star_{M'} - \epsilon$ after acquiring $\tilde{O}\left(|\mathcal{X}_{\xi\text{-}inc}|(|\mathcal{X}_{\xi\text{-}inc}| + \log(1/\delta))\frac{H^4}{\epsilon^3}\right)$ sample trajectories.*[5]

To prove Theorem 1, we introduce some further definitions and helping lemmas. We first define the value function class of interest, $\mathcal{F}$, and establish some basic properties.

Definition 4. *Given $M, \widehat{M}, \xi$, let $\mathcal{F} := \{V^\star_{M'} : M' \in \mathcal{M}\}$ (recall the definition of $\mathcal{M}$ from Definition 3).*

Fact 2. $|\mathcal{F}| \leq |\mathcal{M}| \leq 2^{|\mathcal{X}_{\xi\text{-}inc}|}$.

Definition 5. *Given value function f and M_1 and M_2 that differ only in dynamics P_1 and P_2, define*

$$d^f_{M_1,M_2}(s,a) := \left|\mathbb{E}_{s' \sim P_1(s,a)}[f(s')] - \mathbb{E}_{s' \sim P_2(s,a)}[f(s')]\right|,$$

and $d^{\mathcal{F}}_{M_1,M_2} := \sup_{f \in \mathcal{F}} d^f_{M_1,M_2}(s,a)$.

Fact 3. *If $f \in [0,1]\ \forall f \in \mathcal{F}$, then for any M_1, M_2, s, a,*

$$d^f_{M_1,M_2}(s,a) \leq d^{\mathcal{F}}_{M_1,M_2}(s,a) \leq d_{M_1,M_2}(s,a) \leq 1$$

Lemma 1. *In Algorithm 1, for any fixed t and any $\delta \in (0,1)$, w.p. at least $1-\delta$, $|\hat{v}^{\pi_t}_M - v^{\pi_t}_M| \leq \sqrt{\frac{1}{2n_{eval}}\log\frac{2}{\delta}}$.*

[5] We use $\tilde{O}$ to suppress logarithmic dependence on $|\mathcal{X}_{\xi\text{-inc}}|$, H, $1/\epsilon$, and $\log(1/\delta)$. However, no $\log|\mathcal{S}|$ or $\log|\mathcal{A}|$ is incurred here.

Lemma 2. *In Algorithm 1, for any fixed $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any $\delta \in (0,1)$, when $|D_{s,a}| = n_{est}$, w.p. at least $1-\delta$, $d^{\mathcal{F}}_{M,D}(s,a) \le \sqrt{\frac{1}{2n_{est}} \log \frac{2|\mathcal{F}|}{\delta}}$, where $d^{\mathcal{F}}_{M,D}(s,a) := \sup_{f \in \mathcal{F}} d^{f}_{M,D}(s,a)$.*

The proofs of the above two lemmas are elementary and deferred to Appendix B. Our argument that $D_{s,a}$ only needs to update certain value functions correctly is supported by Lemma 3. Its proof is deferred to Appendix C.

Lemma 3. *Given any M_1 and M_2 where $V^{\star}_{M_1} \in \mathcal{F}$, we have $\|V^{\star}_{M_1} - V^{\star}_{M_2}\|_\infty \le H \|d^{\mathcal{F}}_{M_1,M_2}\|_\infty$.*

The last lemma in this section is Lemma 4, which is a fine-grained version of the *simulation lemma* (Kearns and Singh 2002), a result commonly found in the analyses of PAC exploration algorithms. The proof is deferred to Appendix D.

Lemma 4. *Suppose M_1 and M_2 only differ in dynamics. If there exists f' such that $\|f' - V^{\pi}_{M_2}\|_\infty \le \xi'$, we have*

$$|v^{\pi}_{M_1} - v^{\pi}_{M_2}| \le \langle \eta^{\pi}_{M_1}, d^{f'}_{M_1,M_2} \rangle + 2H\xi',$$

where $\eta^{\pi}_{M_1}$ is the state-action occupancy of π in M_1, defined as $\eta^{\pi}_{M_1}(s,a) := \sum_{h=1}^{H} \mathbb{P}[s_h = s, a_h = a \,|\, \pi, P_1]$.

With all the preparation, we are ready to prove Theorem 1.

Proof of Theorem 1. Throughout the analysis we assume that two type of events always hold, which are later guaranteed by concentration inequalities and union bound: (A) whenever a (s,a) pair satisfies $|D_{s,a}| = n_{\text{est}}$ we have $d^{\mathcal{F}}_{M,D}(s,a) \le \xi$; (B) $|v^{\pi_t}_{M} - \hat{v}^{\pi_t}_{M}| \le \epsilon/10$ (Line 5).

Given these high probability events, we first show the correctness of the algorithm. That is, when it terminates at $t = T$, the returned policy π_T satisfies the theorem statement. Define $\mathcal{X}'_t$ as on Line 10 (i.e., the set of (s,a) with sufficient samples at the beginning of round t) and $\mathcal{X}_t := \mathcal{X}_{\xi\text{-inc}} \bigcap \mathcal{X}'_t$. Since $\widehat{M}_{\mathcal{X}_t} \in \mathcal{M}$ and $V^{\star}_{\widehat{M}_{\mathcal{X}_t}} \in \mathcal{F}$, claim that

$$\forall t,\ \|d^{\mathcal{F}}_{M_t,\widehat{M}_{\mathcal{X}_t}}\|_\infty \le 2\xi. \tag{3}$$

This is because, for $(s,a) \in \mathcal{X}_t$, $\widehat{P}_{\mathcal{X}_t}(s,a) = P(s,a)$, and $P_t(s,a)$ uses the empirical estimate which is guaranteed to be ξ-close to $P(s,a)$ with respect to $\mathcal{F}$; for $(s,a) \notin \mathcal{X}'_t$, $P_t(s,a) = \widehat{P}_{\mathcal{X}_t}(s,a) = \widehat{P}(s,a)$; for the remaining case, both $\widehat{P}_{\mathcal{X}_t}(s,a)$ and $P_t(s,a)$ are ξ-close to $P(s,a)$ with respect to $\mathcal{F}$, so they differ by at most 2ξ.

Now we invoke Lemma 3 on $\widehat{M}_{\mathcal{X}_T}$ and M_T, and obtain $|v^{\star}_{\widehat{M}_{\mathcal{X}_T}} - v^{\pi_T}_{M_T}| \le 2H\xi$. Hence,

$$\begin{aligned} v^{\pi_T}_{M} &\ge \hat{v}^{\pi_T}_{M} - \tfrac{\epsilon}{10} \ge v^{\pi_T}_{M_T} - \tfrac{4\epsilon}{5} \\ &\ge v^{\star}_{\widehat{M}_{\mathcal{X}_T}} - 2H\xi - \tfrac{4\epsilon}{5} \ge \inf_{M' \in \mathcal{M}} v^{\star}_{M'} - \epsilon. \end{aligned}$$

Next we show that when $t < T$, π_t always puts significant occupancy on unlearned state-action pairs in $\mathcal{X}_{\xi\text{-inc}}$; we will refer to the visits to such state-action pairs as *effective visits* in the remainder of the proof. $\forall\, 0 \le t < T$, $v^{\pi_t}_{M_t} - v^{\pi_t}_{M} \ge \hat{v}^{\pi_t}_{M} + 7\epsilon/10 - \hat{v}^{\pi_t}_{M} - \epsilon/10 = 3\epsilon/5$. Let p be the occupancy of $\pi^{\star}_{M_t}$ on the subset of $\mathcal{X}_{\epsilon\text{-inc}}$ that are unlearned, i.e., p is the expected number of effective visits. We would like to upper bound $v^{\pi_t}_{M_t} - v^{\pi_t}_{M}$ via Lemma 4 by letting $M_1 := M$, $M_2 := M_t$. To do that we first have to bound ξ' in Lemma 4 by approximating $V^{\pi_t}_{M_t}$ with $V^{\star}_{\widehat{M}_{\mathcal{X}_t}}$ (the latter is in $\mathcal{F}$): recall that $\pi_t = \pi^{\star}_{M_t}$, and Lemma 3 implies that $\|V^{\pi_t}_{M_t} - V^{\star}_{\widehat{M}_{\mathcal{X}_t}}\|_\infty \le 2H\xi$. Now we can invoke Lemma 4 on M and M_t with $f' = V^{\star}_{\widehat{M}_{\mathcal{X}_t}}$ and $\xi' = 2H\xi$, and

$$\begin{aligned} v^{\pi_t}_{M_t} - v^{\pi_t}_{M} &\le p + (H-p)\xi + 4H^2\xi \qquad (4) \\ &\le p + 5H^2\xi = p + \epsilon/2. \end{aligned}$$

From this we conclude that $p + \epsilon/2 \ge 3\epsilon/5$, so $p \ge \epsilon/10$. In other words, in expectation we have $\epsilon/10$ effective visits per trajectory. If we ignore the randomness in effective visits, $10 n_{\text{est}} |\mathcal{X}_{\xi\text{-inc}}|/\epsilon$ sample trajectories would guarantee successful termination of the algorithm, which matches the theorem statement. The remainder of the proof applies concentration inequalities to deal with the randomness and is deferred to Appendix E. □

Before concluding the section, we comment on the guarantee in Theorem 1. Perhaps the most outstanding term is H^4, which seems unreasonably high. The main difficulty here is that M_t is random, and it is hard to capture it in a model class with reasonable size specified a priori. What we do is to approximate its optimal value function using $V^{\star}_{\widehat{M}_{\mathcal{X}_t}}$ before invoking Lemma 4, which blows up one-step transition error twice (see the $H^2\xi$ term in Eq.(4)).[6] In the next section we consider a similar but slightly different setting, where M_t can be exactly captured in a deterministic model class and the sample complexity is quadratic in H.

6 What if we cannot change the model?

As discussed in Sec.3, one of the unrealistic assumptions so far is that we can manipulate the model and make arbitrary changes to the transition function, which is seldom possible for a sophisticated simulator. To remove the assumption, we need to incorporate the knowledge learned from real trajectories without changing the model itself. But how?

We borrow intuitions from Figure 2 again. An alternative way of explaining 2(b) to 2(e) is that, we should actually compete against policies that only visit (s,a) pairs where the model dynamics are correct. Such an objective is also consistent with the optimality criterion of Eq.(1) for the models in 2(b) to 2(e).

To make this objective more robust, we may want to allow the agent to visit incorrect (s,a) pairs with small probabilities. Instead of imposing constraints on the state-action occupancy of a policy, a more lenient solution is to penalize a policy for visiting incorrect (s,a) pairs. A natural penalty would be to fix the future value of incorrect state-action pairs as the minimum value 0: the future value predicted by the

[6] The recent work of (Azar, Osband, and Munos 2017) faces a similar difficulty and they avoid heavy dependence on H by bounding a particular residual (see their Sec 5.1). However, their technique incurs dependence on $|\mathcal{S}|$ and cannot be applied here.

Algorithm 2 MODEL_PENALIZE($\widehat{M}$)

```
 1: $M_0 \leftarrow \widehat{M}$. $\mathcal{X}_0 \leftarrow \{\}$. $D_{s,a} \leftarrow \{\}$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$.
 2: for $t = 0, 1, \ldots$ do
 3:     $\pi_t \leftarrow \pi^\star_{M_t}$, $f_t \leftarrow V^\star_{M_t}$.
 4:     Collect $n_{\text{eval}}$ trajectories using $\pi_t$, and let $\hat{v}^{\pi_t}_M$ be the
        Monte-Carlo estimate of value.
 5:     if $\hat{v}^{\pi_t}_M \ge v^{\pi_t}_{M_t} - 4\epsilon/5$ then return $\pi_t$.
 6:     while $\forall (s,a) \notin \mathcal{X}_t$, ISFINE($s$, $a$, $f_t$) do
 7:         Collect a trajectory $s_1, a_1, \ldots, s_H, a_H$ using $\pi_t$.
 8:         $\forall\, h$, add $s_{h+1}$ to $D_{s_h,a_h}$ if $|D_{s_h,a_h}| < n_{\text{est}}$.
 9:     end while
10:     $\mathcal{X}_{\text{tmp}} \leftarrow \{\}$.
11:     for $h = H-1, \ldots, 2, 1$ do
12:         $M_{\text{tmp}} \leftarrow \widehat{M}_{\setminus \mathcal{X}_{\text{tmp}}}$, $f_{\text{tmp}} \leftarrow V^\star_{M_{\text{tmp}}}$.
13:         $\mathcal{X}_{\text{tmp}} \leftarrow \mathcal{X}_{\text{tmp}} \bigcup$
                $\{(s,a) \in \mathcal{S}_h \times \mathcal{A} : \neg$ISFINE$(s, a, f_{\text{tmp}})\}$.
14:     end for
15:     $\mathcal{X}_{t+1} \leftarrow \mathcal{X}_{\text{tmp}}$, $M_{t+1} \leftarrow \widehat{M}_{\setminus \mathcal{X}_{t+1}}$.
16: end for

17: function ISFINE($s$, $a$, $f$)
18:     if $|D_{s,a}| < n_{\text{est}}$ then return true.
19:     if $d^f_{\widehat{M},D}(s,a) \le 1.5\xi$ then return true.
        // recall definition of $d^f_{\widehat{M},D}$ from Eq.(2)
20:     return false.
21: end function
```

model is not trustworthy due to incorrect dynamics and we replace it with a pessimistic guess.

Finally, due to our assumption of non-negative rewards, the penalty can be simply implemented by terminating a model episode upon running into an incorrect (s,a) pair. We will still treat the penalized model as a new MDP to facilitate analysis, with the understanding that such changes can be implemented in a black-box manner. The penalized model is formally defined below.

Definition 6 (Partially penalized model)**.** *Given MDP* $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, R, H, \mu)$ *and* $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$*, define* $\widehat{M}_{\setminus \mathcal{X}}$ *as the MDP* $(\mathcal{S}, \mathcal{A}, \widehat{P}_{\setminus \mathcal{X}}, R, H, \mu)$ *where*

$$\widehat{P}_{\setminus \mathcal{X}}(s,a) := \begin{cases} \texttt{termination}, & \textit{if } (s,a) \in \mathcal{X}, \\ \widehat{P}(s,a), & \textit{otherwise.} \end{cases}$$

We define $\widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{F}}$, the analogies of $\mathcal{M}$ and $\mathcal{F}$.

Definition 7. *Given* M*,* $\widehat{M}$*, and* ξ*, define* $\widetilde{\mathcal{M}} := \{\widehat{M}_{\setminus \mathcal{X}} : \mathcal{X} \subseteq \mathcal{X}_{\xi\text{-}inc}\}$, $\widetilde{\mathcal{F}} := \{V^\star_{M'} : M' \in \widetilde{\mathcal{M}}\}$.

Similarly to Eq.(1), in this section we compete against:

$$\inf_{M' \in \widetilde{\mathcal{M}}} v^\star_{M'}. \tag{5}$$

It is worth noting that the value in Eq.(5) is always less than or equal to that in Eq.(1). Intuitively, whenever an incorrect state-action pair is discovered, we allow the agent to avoid it as opposed to fixing the incorrect dynamics since we refrain ourselves from modifying the model. This comes with the cost that we give up the opportunity of reusing these state-action pairs in future policies. This intuition is formalized in Fact 4, and we give a concrete example in Figure 2(f) where Eq.(1) and (5) have a nontrivial gap.

Fact 4. *For any* $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$ *and* $\pi : \mathcal{S} \to \mathcal{A}$*,* $v^\pi_{\widehat{M}_{\setminus \mathcal{X}}} \le v^\pi_{\widehat{M}_{\mathcal{X}}}$.

In the remainder of this section, we introduce Algorithm 2 and state the sample complexity result. Overall Algorithm 2 is similar to Algorithm 1, but with a few differences:

1. Unlike Algorithm 1 where we blindly replace $\widehat{P}(s,a)$ with $D_{s,a}$ (which is valid as $D_{s,a}$ is always unbiased), here we penalize (s,a) selectively as penalizing a correct (s,a) may affect the value we could obtain.
2. While we would like to compare $\widehat{P}(s,a)$ and $D_{s,a}$ against all functions in $\widetilde{\mathcal{F}}$, this is impossible as we do not know $\widetilde{\mathcal{F}}$ in advance (because $\mathcal{X}_{\xi\text{-inc}}$ is unknown). Fortunately, it is sufficient to compare them against the current value function f_t, which gives the criterion on Line 6.
3. The for-loop computes a set of incorrect (s,a) pairs from the bottom up. This is necessary because penalizing an (s,a) pair at a lower level (i.e., a later time step) may trigger a change in value function, which affects whether some other (s,a) at a higher level should be penalized or not. The bottom-up procedure guarantees that `isfine` is never invoked on an outdated f.
4. Thanks to the binary nature of the penalty, any M_t that we could run into is a member of $\widetilde{\mathcal{M}}$ (with high probability), hence M_t and π_t are much more deterministic objects than in Algorithm 1. As a result, we avoid the difficulty in Theorem 1 and can show a quadratic dependence on H.

Theorem 2. *Given any* $\delta \in (0,1), \epsilon \in (0,1)$*, we run Algorithm 2 with parameters* $\xi = \frac{\epsilon}{5H}$*,* $n_{eval} = \tilde{O}(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$*,* $n_{est} = \tilde{O}(\frac{H^2}{\epsilon^2}(|\mathcal{X}_{\xi\text{-}inc}| + \log \frac{1}{\delta}))$*. W.p. at least* $1 - \delta$ *the algorithm will return a policy* π_T *such that* $v^{\pi_T}_M \ge \inf_{M' \in \widetilde{\mathcal{M}}} v^\star_{M'} - \epsilon$ *after acquiring* $\tilde{O}(|\mathcal{X}_{\xi\text{-}inc}|(|\mathcal{X}_{\xi\text{-}inc}| + \log(1/\delta))\frac{H^2}{\epsilon^3})$ *sample trajectories.*

The proof of Theorem 2 is similar to that of Theorem 1 and is deferred to Appendix F.

7 Non-interactive algorithms are inefficient

In the previous sections we give two algorithms that enjoy polynomial sample complexity guarantee without any dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$. Both algorithms fit into the abstract protocol we introduce at the beginning, that is, they alternate between computation in the approximate model and data collection in the real environment.

In this section we show that such interactivity is crucial for our purpose. In particular, we prove a hardness result that, if all the data are collected before we perform any computation in $\widehat{M}$, no algorithm can achieve the desired polynomial sample complexity even if the conditions required by Algorithms 1 and 2 are satisfied. We briefly sketch the proof idea and the full proof is deferred to Appendix G.

Theorem 3. *If the data collection strategy is independent of $\widehat{M}$, no algorithm can learn an ϵ-optimal policy with probability $2/3$ using $poly(|\mathcal{X}_0|, H, 1/\epsilon)$ sample trajectories, even if $\inf_{M' \in \mathcal{M}} v^\star_{M'} = \inf_{M' \in \widetilde{\mathcal{M}}} v^\star_{M'} = v^\star_M$.*

Proof sketch. Assume towards contradiction that such an algorithm exists. We can solve the hard instance of best arm identification with $poly(\log|\mathcal{A}|, 1/\epsilon)$ samples, which is against the known lower bound. Concretely, we design $|\mathcal{A}|^2$ models, where each model claims that a pair of arms are more rewarding than others. Applying the hypothetical algorithm to each model allows us to make reliable pair-wise comparison between the arms using only $O(\log|\mathcal{A}|)$ independent datasets, each of size $O(1/\epsilon^2)$. □

8 Relax the definition of ξ-correctness

In this section we relax the definition of ξ-correctness as promised in Remark 1. In Definition 1, whether an (s, a) pair is correct is determined by how large $d_{M,\widehat{M}}(s, a)$ is. The key observation here is that in the proof of Theorem 1, we only use the fact $d_{M,\widehat{M}}(s, a) \le \xi$ for $(s, a) \notin \mathcal{X}_{\xi\text{-inc}}$ in Eq.(3) through $d^{\mathcal{F}}_{M,\widehat{M}}$ (recall Fact 3). Therefore, we can simply re-define $\mathcal{X}_{\xi\text{-inc}}$ based on $d^{\mathcal{F}}_{M,\widehat{M}}$ instead of $d_{M,\widehat{M}}$, and all the analyses and guarantees extend straightforwardly. Due to Fact 3, the new definition will result in a smaller number of incorrect state-action pairs, and hence yield improved sample complexity in Theorem 1. The same thing happens to Theorem 2, where we can re-define $\mathcal{X}_{\xi\text{-inc}}$ based on $d^{\widetilde{\mathcal{F}}}_{M,\widehat{M}}$.

The complication here, however, is that the definition of $d^{\mathcal{F}}_{M,\widehat{M}}$ ($d^{\widetilde{\mathcal{F}}}_{M,\widehat{M}}$) depends on $\mathcal{F}$ ($\widetilde{\mathcal{F}}$), which further depends on $\mathcal{X}_{\xi\text{-inc}}$ (recall Definitions 3, 4, and 7), and we are now modifying the definition of $\mathcal{X}_{\xi\text{-inc}}$ to make it depend on $\mathcal{F}$ ($\widetilde{\mathcal{F}}$). The resolution to the recursive dependence is to define things in a bottom-up order; see Algorithm 3. This procedure is very similar to the for-loop in Algorithm 2, where the same issue has already been encountered. The formal statement of the tightened results is given below. In Appendix H we also describe a scenario where $|\overline{\mathcal{X}}_{\xi\text{-inc}}|$ and $|\widetilde{\mathcal{X}}_{\xi\text{-inc}}|$ are substantially smaller than $|\mathcal{X}_{\xi\text{-inc}}|$.

Proposition 2. *Let $\overline{\mathcal{X}}_{\xi\text{-inc}}$ and $\widetilde{\mathcal{X}}_{\xi\text{-inc}}$ be defined via Algorithm 3. We have: (1) $\overline{\mathcal{X}}_{\xi\text{-inc}} \subseteq \mathcal{X}_{\xi\text{-inc}}$, $\widetilde{\mathcal{X}}_{\xi\text{-inc}} \subseteq \mathcal{X}_{\xi\text{-inc}}$.*
(2) Theorems 1 and 2 still hold if we replace $|\mathcal{X}_{\xi\text{-inc}}|$ in the theorem statements by $|\overline{\mathcal{X}}_{\xi\text{-inc}}|$ and $|\widetilde{\mathcal{X}}_{\xi\text{-inc}}|$ respectively.

9 Conclusions and discussions

In this paper, we investigate the theoretical properties of reinforcement learning with an approximate model as side information. We believe that there are 3 high-level insights that can be drawn from the paper:

1. We need the model to *always* stay optimistic (Assumption 1), otherwise there are degenerate cases where the model is useless even if it is correct in all but a constant number of state-action pairs (Proposition 1).

Algorithm 3 ConstructBadSet(M, $\widehat{M}$, ξ)

```
$\overline{\mathcal{X}}_{\xi\text{-inc}} \leftarrow \{\}$.
for $h = H-1, \ldots, 2, 1$ do
    $\mathcal{M} \leftarrow \{\widehat{M}_{\mathcal{X}} : \mathcal{X} \subseteq \overline{\mathcal{X}}_{\xi\text{-inc}}\}$. // use $\widehat{M}_{\setminus\mathcal{X}}$ for $\widetilde{\mathcal{X}}_{\xi\text{-inc}}$
    $\mathcal{F} \leftarrow \{V^\star_{M'} : M' \in \mathcal{M}\}$.
    $\overline{\mathcal{X}}_{\xi\text{-inc}} \leftarrow \overline{\mathcal{X}}_{\xi\text{-inc}} \bigcup$
            $\{(s, a) \in \mathcal{S}_h \times \mathcal{A} : d^{\mathcal{F}}_{M,\widehat{M}}(s, a) > \xi\}$.
end for
```

2. It is important that the learner interacts with the environment and the model in an alternating manner (Theorem 3).
3. Under 1+2, we can achieve polynomial sample complexity in the number of incorrect state-action pairs and incur no dependence on $|\mathcal{S}|$ and $|\mathcal{A}|$ (Theorems 1 and 2).

We conclude the paper by discussing related work, limitations of our assumptions and results, and open questions.

- The most related work is (Cutler, Walsh, and How 2015), who consider multiple simulators with varying levels of fidelity. They implicitly assume that a simulator's quality is homogeneous across the state-action space, evidenced by their fidelity defined as the worst-case error over states and actions. Consequently, they incur dependence on $|\mathcal{S}|$ and $|\mathcal{A}|$. In another related work, (Ha and Yamane 2015) propose an algorithm for linear control problems that is similar in spirit to our Algorithm 1. While no sample complexity guarantee is given, their algorithm is empirically validated and produces promising results.
- A major limitation of our work is that we assume $|\mathcal{S}|$ (and $|\mathcal{A}|$) is large but $|\mathcal{X}_{\xi\text{-inc}}|$ is small, which can be unrealistic. While it is possible to extend the analyses to accommodate continuous $\mathcal{X}_{\xi\text{-inc}}$ by a covering argument (Kakade, Kearns, and Langford 2003; Pazis and Parr 2013), such arguments incur dependence on the covering numbers, which are typically exponential in the dimension. Taking our analyses forward to a practical scenario may require satisfying theoretical solutions to exploration in large state spaces, an important research direction that is relatively understudied by itself despite a few very recent advances (Krishnamurthy, Agarwal, and Langford 2016; Jiang et al. 2017).
- The conditions we have identified (e.g., Assumption 1 and the agnostic version) are sufficient. Are they necessary and are there weaker conditions?
- There is a gap between the interactivity of our algorithms (polynomially many alternations) and the non-interactive lower bound (no alternation at all; see Theorem 3). While we might expect a stronger lower bound to exclude even a small (e.g., constant) number of alternations, the current formulation cannot prevent the agent from loading the entire $\widehat{M}$ into its memory to consult the model at any future time. Obtaining the stronger lower bound (if it exists) would require a careful characterization of what kind of computation is allowed within each round.

- The sample complexity guarantees obtained in Theorems 1 and 2 may not be optimal. In particular, it might be possible to remove one $1/\epsilon$ term by carefully distinguishing important states from the unimportant ones (Dann and Brunskill 2015). It might also be possible to reduce the dependence on $|\mathcal{X}_{\xi\text{-inc}}|$ by determinizing the order in which we fix / penalize incorrect (s, a) pairs, e.g., by delaying model revision and updating multiple state-action pairs at once. Tightening the upper bounds and finding matching lower bounds are interesting directions for future work.
- Our algorithms explore with the optimal policy of the model. Are there more sophisticated strategies that improve sample complexity? Since we want to avoid dependence on $|\mathcal{A}|$, standard operations such as taking actions uniformly are prohibited. Any intelligent exploration in this case might have to be heavily informed by the model.

Acknowledgements

The author thanks the anonymous reviewers for the insightful comments. The research question was inspired by a conversation with David Meger at McGill University.

References

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, 263–272.

Bjarnason, R.; Fern, A.; and Tadepalli, P. 2009. Lower bounding Klondike solitaire with Monte-Carlo planning. In *Proceedings of International Conference on Automated Planning and Scheduling*, 26–33.

Cutler, M., and How, J. P. 2015. Efficient reinforcement learning for robots using informative simulated priors. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, 2605–2612. IEEE.

Cutler, M.; Walsh, T. J.; and How, J. P. 2015. Real-world reinforcement learning via multifidelity simulators. *IEEE Transactions on Robotics* 31(3):655–671.

Dann, C., and Brunskill, E. 2015. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 2818–2826.

Grill, J.-B.; Valko, M.; and Munos, R. 2016. Blazing the trails before beating the path: Sample-efficient monte-carlo planning. In *Advances in Neural Information Processing Systems*, 4680–4688.

Ha, S., and Yamane, K. 2015. Reducing hardware experiments for model learning and policy optimization. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2620–2626.

Hanna, J. P., and Stone, P. 2017. Grounded action transformation for robot learning in simulation. In *AAAI*, 3834–3840.

Heess, N.; Wayne, G.; Silver, D.; Lillicrap, T.; Erez, T.; and Tassa, Y. 2015. Learning continuous control policies by stochastic value gradients. In *Advances in Neural Information Processing Systems*, 2944–2952.

Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J.; and Schapire, R. E. 2017. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 1704–1713.

Kakade, S.; Kearns, M. J.; and Langford, J. 2003. Exploration in metric state spaces. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 306–312.

Kearns, M., and Singh, S. 2002. Near-optimal reinforcement learning in polynomial time. *Machine Learning* 49(2-3):209–232.

Kearns, M.; Mansour, Y.; and Ng, A. Y. 2002. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning* 49(2-3):193–208.

Kober, J.; Bagnell, J. A.; and Peters, J. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 0278364913495721.

Koos, S.; Mouret, J.-B.; and Doncieux, S. 2010. Crossing the reality gap in evolutionary robotics by promoting transferable controllers. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, 119–126. ACM.

Krishnamurthy, A.; Agarwal, A.; and Langford, J. 2016. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, 1840–1848.

Levine, S.; Finn, C.; Darrell, T.; and Abbeel, P. 2016. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research* 17(39):1–40.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 1928–1937.

Pazis, J., and Parr, R. 2013. PAC optimal exploration in continuous space Markov Decision Processes. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*.

Rusu, A. A.; Vecerik, M.; Rothörl, T.; Heess, N.; Pascanu, R.; and Hadsell, R. 2016. Sim-to-real robot learning from pixels with progressive nets. *arXiv preprint arXiv:1610.04286*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.

A Proof of Proposition 1

We construct a family of MDPs with $H = 2$ to emulate the hard instances in multi-armed bandit (Figure 3): there is a single start date, and each action leads to a different state at level 2. Each level 2 state has only 1 action and transitions to a Bernoulli distribution over a rewarding state and non-rewarding state at level 3, mostly half-half. There is one special state that has $1/2+\epsilon$ probability to the rewarding state, and different M's in the MDP family differ in the identity of this state. For any MDP in the family, the $\widehat{M}$ that predicts half-half transition for all level 2 states is always near-perfect, in the sense that it is only wrong by one state. However, $\widehat{M}$ is information-theoretically useless and we essentially have the hard instances of best arm identification, which is known to have $\Omega(|\mathcal{A}|/\epsilon^2)$ lower bound (Krishnamurthy, Agarwal, and Langford 2016, Theorem 2). An even stronger $\Omega(|\mathcal{A}|^H/\epsilon^2)$ lower bound is obtainable by embedding a multi-armed bandit with exponentially many arms in a tree-structured MDP (Jiang et al. 2017, Proposition 11). In this case, the only variable that can "explain away" this exponential as a polynomial is $|\mathcal{S}|$ as $|\mathcal{S}| = |\mathcal{A}|^H$. Therefore, the sample complexity is polynomial in $|\mathcal{S}|$. □

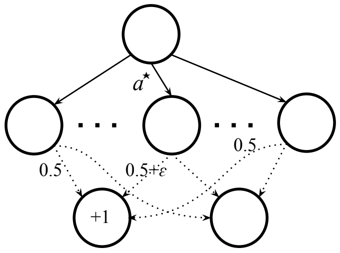


Figure 3: Lower bound construction for Proposition 1 and Theorem 3, which essentially emulates the well-known hard instances in multi-armed bandits (Auer, Cesa-Bianchi, and Fischer 2002).

B Proof of Lemmas 1 and 2

Proof of Lemma 1. $\hat{v}_M^{\pi_t}$ is the average of $\sum_{h=1}^H r_h$ over n_{eval} i.i.d. trajectories with actions taken according to π_t. By definition the estimation is unbiased. Since our boundedness assumption states that $\sum_{h=1}^H r_h \in [0,1]$, Hoeffding's inequality applies and immediately yields the result. □

Proof of Lemma 2. Fix any (s,a) and $f \in \mathcal{F}$. When $|D_{s,a}| = n_{\text{est}}$, we have n_{est} i.i.d. samples of $s' \sim P(s,a)$. When we plug s' into f, we get $f(s')$ which is an unbiased estimate of $\mathbb{E}_{s'\sim P(s,a)}[f(s')]$. Since f has bounded range $[0,1]$ (see Footnote 1), by Hoeffding's inequality we have $d_{M,D}^f(s,a) \le \sqrt{\frac{1}{2n_{\text{est}}}\log\frac{2}{\delta}}$. The lemma follows from union bounding over $\mathcal{F}$. □

C Proof of Lemma 3

Let $\mathcal{T}_1, \mathcal{T}_2$ be the Bellman update operator of M_1 and M_2 respectively.

$$\begin{aligned}&\|V_{M_1}^\star - \mathcal{T}_2 V_{M_1}^\star\|_\infty = \|\mathcal{T}_1 V_{M_1}^\star - \mathcal{T}_2 V_{M_1}^\star\|_\infty \\ &= \max_{s,a\in\mathcal{S}\times\mathcal{A}} \left|\mathbb{E}_{s'\sim P_1(s,a)}[V_{M_1}^\star(s')] - \mathbb{E}_{s'\sim P_2(s,a)}[V_{M_1}^\star(s')]\right| \le \|d_{M_1,M_2}^{\mathcal{F}}\|_\infty.\end{aligned}$$

Therefore,

$$\begin{aligned}\|V_{M_1}^\star - V_{M_2}^\star\|_\infty &= \|V_{M_1}^\star - \mathcal{T}_2 V_{M_1}^\star + \mathcal{T}_2 V_{M_1}^\star - \mathcal{T}_2 V_{M_2}^\star\|_\infty \\ &\le \|d_{M_1,M_2}^{\mathcal{F}}\|_\infty + \|\mathcal{T}_2 V_{M_1}^\star - \mathcal{T}_2 V_{M_2}^\star\|_\infty \\ &= \|d_{M_1,M_2}^{\mathcal{F}}\|_\infty + \max_{s\in\mathcal{S},a\in\mathcal{A}} \left|\mathbb{E}_{s'\sim P_2(s,a)}[V_{M_1}^\star(s')] - \mathbb{E}_{s'\sim P_2(s,a)}[V_{M_2}^\star(s')]\right|.\end{aligned}$$

The lemma follows from noticing that for $s \in \mathcal{S}_1$, $P_2(s,a)$ is supported on $\mathcal{S}\setminus\mathcal{S}_1$, and expanding the above inequality H times yields the desired result. □

D Proof of Lemma 4

Prove by induction at each level, where we treat each state at that level as the initial distribution (point mass). At bottom level, both sides are 0 as reward is known, so the statement holds. Assume that the statement holds at level $h+1$ and below. Let $\eta_{M_1,s}^\pi$ be the occupancy when the starting state is s. Then for any state $s_h \in \mathcal{S}_h$, let $a_h = \pi(s_h)$, and we have

$$\begin{aligned}&|V_{M_1}^\pi(s_h) - V_{M_2}^\pi(s_h)| \\ &= |\langle P_1(s_h,a_h), V_{M_1}^\pi\rangle - \langle P_2(s_h,a_h), V_{M_2}^\pi\rangle| \\ &\le |\langle P_1(s_h,a_h), V_{M_2}^\pi\rangle - \langle P_2(s_h,a_h), V_{M_2}^\pi\rangle| + |\langle P_1(s_h,a_h), V_{M_1}^\pi - V_{M_2}^\pi\rangle|.\end{aligned}$$

To bound the first term,

$$\begin{aligned}
&|\langle P_1(s_h, a_h), V^\pi_{M_2}\rangle - \langle P_2(s_h, a_h), V^\pi_{M_2}\rangle| \\
\le\ & |\langle P_1(s_h, a_h), f'\rangle - \langle P_2(s_h, a_h), f'\rangle| + |\langle P_1(s_h, a_h), V^\pi_{M_2} - f'\rangle| + \langle P_2(s_h, a_h), V^\pi_{M_2} - f'\rangle| \\
\le\ & d^{f'}_{M_1,M_2}(s_h, a_h) + 2\|V^\pi_{M_2} - f'\|_\infty \le d^{f'}_{M_1,M_2}(s_h, a_h) + 2\xi'.
\end{aligned} \tag{6}$$

So

$$\begin{aligned}
&|V^\pi_{M_1}(s_h) - V^\pi_{M_2}(s_h)| \\
\le\ & d^{f'}_{M_1,M_2}(s_h, a_h) + 2\xi' + \mathbb{E}_{s_{h+1}\sim P_1(s_h,a_h)}\left[|V^\pi_{M_1}(s_{h+1}) - V^\pi_{M_2}(s_{h+1})|\right] \\
\le\ & d^{f'}_{M_1,M_2}(s_h, a_h) + 2\xi' + \mathbb{E}_{s_{h+1}\sim P_1(s_h,a_h)}\left[\langle \eta^\pi_{M_1,s_{h+1}}, d^{f'}_{M_1,M_2}\rangle\right] + 2(H-h)\xi' && \text{(induction assumption)} \\
=\ & \langle \eta^\pi_{M_1,s_h}, d^{f'}_{M_1,M_2}\rangle + 2(H-h+1)\xi'. && \square
\end{aligned}$$

E Proof of Theorem 1 (continued)

Now we apply concentration inequalities to guarantee the high probability events and establish sample complexity. We split the failure probability δ evenly into 3 pieces, and assign two of them to events (A) and (B) specified at the beginning of the proof. Let N be the total number of trajectories we collect via Line 7. Since a state-action pair has to be filled with n_{est} transitions to increase t by 1, we have $T \le NH/n_{\text{est}}$. By Lemma 1, we can guarantee event (B) by letting $n_{\text{eval}} = \frac{50}{\epsilon^2}\log(\frac{6NH}{n_{\text{est}}\delta})$, so the total trajectories spent on Line 5 is at most $NHn_{\text{eval}}/n_{\text{est}}$.

Next we set n_{est} to guarantee (A). For each individual (s, a) pair, to guarantee (A) we need $n_{\text{est}} = \frac{1}{2\xi^2}\log\frac{2|\mathcal{F}|}{\delta'}$ according to Lemma 2, where δ' is the failure probability for each individual (s, a). As common practice in PAC-MDP literature, here we could take a union bound over all states and actions and let $\delta' = \delta/(3|\mathcal{S}\times\mathcal{A}|)$, but that would incur logarithmic dependence on $|\mathcal{S}\times\mathcal{A}|$. To avoid any dependence, we require the following set of events to hold simultaneously, which include the events in (A) as a subset: for any $n = 1, \dots, N$, let $s^{(n)}_h, a^{(n)}_h$ be the state-action pair encountered on the n-th trajectory at the h-th time step, and we require the subsequent n_{est} transitions from this particular state-action pair to form a ξ-accurate estimate as in (A). In this case, we would union bound over NH events, so $n_{\text{est}} = \frac{1}{2\xi^2}\log\frac{6NH|\mathcal{F}|}{\delta} = \frac{50H^4}{\epsilon^2}\log\frac{6NH|\mathcal{F}|}{\delta}$.

Finally we leverage $p \ge \epsilon/10$ to give an upper bound on N. The intuition is that, roughly speaking the algorithm should terminate when $N \approx 10|\mathcal{X}_{\xi\text{-inc}}|n_{\text{est}}/\epsilon$, if the actual number of effective visits is equal to the expectation. To deal with the randomness, we set $N \ge 64|\mathcal{X}_{\xi\text{-inc}}|n_{\text{est}}/\epsilon$, and argue that the actual number of effective visits is at least $N\epsilon/64$. To show that, consider the following process:

$$\begin{aligned}
\sum_{n=1}^{N} \Big(&\#\text{ effective visits on } n\text{-th trajectory conditioned on the previous trajectories} \\
&- \text{expected effective visits on } n\text{-th trajectory conditioned on the previous trajectories}\Big).
\end{aligned} \tag{7}$$

The complication is that we cannot condition on the success of (A) here as that would interfere with the martingale property. Note, however, that the above process enjoys concentration regardless of whether (A) holds or not: the partial sum of Eq.(7) is a martingale with difference bounded in $[-H, H]$, and we can apply Azuma's inequality (one-sided version) to show that this difference is bounded from below by $-2H\sqrt{2N\log\frac{3}{\delta}}$. When (A) holds, total expected effective visits is at least $N\epsilon/10$, therefore

$$\text{total effective visits} \ge N\epsilon/10 - 2H\sqrt{2N\log\frac{3}{\delta}}.$$

Thanks to our choice of N,

$$\sqrt{N} \ge \sqrt{|\mathcal{X}_{\xi\text{-inc}}|\frac{64\cdot 50H^4}{\epsilon^2}\log\frac{6NH}{\delta}} \ge \frac{40\sqrt{2}H}{\epsilon}\sqrt{\log\frac{3}{\delta}}.$$

So

$$\text{total effective visits} \ge N\epsilon/10 - N\frac{2H\sqrt{2\log\frac{3}{\delta}}}{\frac{40\sqrt{2}H}{\epsilon}\sqrt{\log\frac{3}{\delta}}} = N\epsilon/20 > N\epsilon/64.$$

This guarantees $\mathcal{X}_{\xi\text{-inc}} n_{\text{est}}$ effective visits in total, therefore the algorithm will terminate. The last step is to resolve the issue that N and n_{est} depends on each other. In particular, we need

$$N \ge \frac{3200|\mathcal{X}_{\xi\text{-inc}}|H^4}{\epsilon^3} \log \frac{6H|\mathcal{F}|N}{\delta}.$$

We claim that the following value of N satisfies the above inequality.

$$N_0 = \frac{3200|\mathcal{X}_{\xi\text{-inc}}|H^4}{\epsilon^3} \log \frac{21H|\mathcal{F}|}{\delta}, \qquad N = N_0(1 + \log N_0).$$

To verify, it suffices to show that

$$\frac{(N - N_0)\epsilon^3}{3200|\mathcal{X}_{\xi\text{-inc}}|H^4} \ge \log N.$$

Our choice of N guarantees that the LHS is $\log \frac{21H|\mathcal{F}|}{\delta} \log N_0 \ge 3 \log N_0$. On the other hand,

$$\begin{aligned} \log N = &\ \log N_0 + \log(1 + \log N_0) \le \log N_0 + \log(2 \log N_0) \\ = &\ \log N_0 + \log 2 + \log\log N_0 \le 3 \log N_0. \end{aligned}$$

So the choice of N suffices. Note that the total trajectories collected on Line 5 is substantially fewer ($\tilde{O}(|\mathcal{X}_{\xi\text{-inc}}|H/\epsilon^3)$). The sample complexity guarantee follows from recalling that $|\mathcal{F}| \le 2^{|\mathcal{X}_{\xi\text{-inc}}|}$. □

F Proof of Theorem 2

Throughout the analysis we assume that two type of events always hold, which are later guaranteed by concentration inequalities and union bound: (A) whenever a (s,a) pair satisfies $|D_{s,a}| = n_{\text{est}}$ we have $d^{\widetilde{\mathcal{F}}}_{M,D}(s,a) \le \xi/2$, (B) $|v^{\pi_t}_M - \hat{v}^{\pi_t}_M| \le \epsilon/5$.

Given these high probability events, we first show the correctness of the algorithm. First, notice that (A) and the threshold of 1.5ξ on Line 19 guarantees that whenever isfine returns false, the (s,a) is ξ-incorrect. Hence, $\mathcal{X}_{\text{tmp}} \subseteq \mathcal{X}_{\xi\text{-inc}}$, $\mathcal{X}_t \subseteq \mathcal{X}_{\xi\text{-inc}}$, $M_{\text{tmp}} \in \widetilde{\mathcal{M}}$, $M_t \in \widetilde{\mathcal{M}}$, $f_{\text{tmp}} \in \widetilde{\mathcal{F}}$, $f_t \in \widetilde{\mathcal{F}}$ always hold, and when the algorithm terminates at $t = T$,

$$v^{\pi_T}_M \ge \hat{v}^{\pi_T}_M - \tfrac{\epsilon}{5} \ge v^{\pi_T}_{M_T} - \epsilon \ge \inf_{M' \in \widetilde{\mathcal{M}}} v^\star_{M'} - \epsilon.$$

On the other hand, for all $(s,a) \notin \mathcal{X}_t$ and $|D_{s,a}| = n_{\text{est}}$, if isfine returns true on Line 13, we know that $d^{f_t}_{M,D}(s,a) \le 2\xi$.

Next we show that when $t < T$, π_t always explores with nontrivial probability. $\forall\, 0 \le t < T$,

$$\begin{aligned} v^{\pi_t}_{M_t} - v^{\pi_t}_{M_{\setminus\mathcal{X}_t}} &\ge v^{\pi_t}_{M_t} - v^{\pi_t}_{M} && \text{(pessimism of } M_{\setminus\mathcal{X}_t}\text{)} \\ &\ge (\hat{v}^{\pi_t}_M + 4\epsilon/5) - \hat{v}^{\pi_t}_M - \epsilon/5 = 3\epsilon/5. && \text{(Line 5 of Alg 2 and Event (B))} \end{aligned}$$

We then argue that M_t and $M_{\setminus\mathcal{X}_t}$ can only differ significantly w.r.t. f_t on unlearned (s,a) in $\mathcal{X}_{\xi\text{-inc}}$:

1. For $(s,a) \in \mathcal{X}_t$, M_t and $M_{\setminus\mathcal{X}_t}$ are identical , as such (s,a) leads to termination in both cases.

2. For $(s,a) \in \mathcal{X}_{\xi\text{-inc}} \setminus \mathcal{X}_t$, if $|D_{s,a}| = n_{\text{est}}$ ("learned"), we know that (s,a) is 2ξ-correct with respect to the current f_t.

3. For $(s,a) \notin \mathcal{X}_{\xi\text{-inc}}$, by definition it is ξ-correct w.r.t. all $f \in \widetilde{\mathcal{F}}$, which includes f_t.

The only remaining case is unlearned (s,a) in $\mathcal{X}_{\xi\text{-inc}}$, which corresponds to "effective visits" (we will use this term in the same way as in the previous proof).

Let p be the expected effective visits of π_t as in the proof of Theorem 1. We can upper bound $v^{\pi_t}_{M_t} - v^{\pi_t}_{M_{\setminus\mathcal{X}_t}}$ using p via Lemma 4. Unlike the previous proof, however, this time we have $V^{\pi_t}_{M_t} = V^\star_{M_t} \in \widetilde{\mathcal{F}}$, so ξ' in Lemma 4 is 0 if we set $f' = f_t = V^\star_{M_t}$! Using the lemma, we have

$$\begin{aligned} v^{\pi_t}_{M_t} - v^{\pi_t}_{M_{\setminus\mathcal{X}_t}} &\le \langle \eta^{\pi_t}_{M_{\setminus\mathcal{X}_t}}, d^{f_t}_{M_{\setminus\mathcal{X}_t}, M_t} \rangle \le \langle \eta^{\pi_t}_{M}, d^{f_t}_{M_{\setminus\mathcal{X}_t}, M_t} \rangle \\ &\le p + (H - p) \cdot 2\xi \le p + 2\epsilon/5. \end{aligned}$$

The second step uses the fact that $M_{\setminus\mathcal{X}_t}$ is identical to M up to forced terminations, so a policy's occupancy in $M_{\setminus\mathcal{X}_t}$ is always upper bounded by that in M. From this we conclude that $p + 2\epsilon/5 \ge 3\epsilon/5$, so $p \ge \epsilon/5$. The remainder of the proof follows from exactly the same arguments as in Theorem 1 hence is omitted here. □

G Full proof of Theorem 3

Assume towards contradiction that we have such an algorithm, and its sample complexity is $g(|\mathcal{X}_{\text{0-inc}}|, H, 1/\epsilon)$ where g is some polynomial. We will again use the construction in Proposition 1 to emulate the hard instance of best arm identification, and use the hypothetical algorithm to solve it with success probability $2/3$ and sample complexity $poly(\log|\mathcal{A}|, 1/\epsilon)$, which is clearly against the known lower bound $\Omega(|\mathcal{A}|/\epsilon^2)$.

Since data collection is independent of $\widehat{M}$, we will collect a dataset first, and run the hypothetical algorithm on the dataset with different $\widehat{M}$'s. In particular, we create a class of approximate models $\{M_{a,a'} : a, a' \in \mathcal{A}, a \neq a'\}$, where $M_{a,a'}$ predicts that both a and a' at the initial state give $1/2 + \epsilon$ value and all the other actions give $1/2$ value.

Let the true optimal action be $a^\star$. For $\widehat{M} = M_{a,a^\star}$ where $a \neq a^\star$, by construction $|\mathcal{X}_{\text{0-inc}}| = 1$ and $\inf_{M' \in \mathcal{M}} v^\star_{M'} = \inf_{M' \in \widetilde{\mathcal{M}}} v^\star_{M'} = v^\star_M = 1/2 + \epsilon$. So if we run the hypothetical algorithm on any such $\widehat{M}$ with a pre-collected dataset of size $g(1, 3, \frac{1}{3\epsilon})$, the algorithm returns an $\epsilon/3$-optimal policy with probability at least $2/3$. Note that we can identify $a^\star$ from an $\epsilon/3$-optimal policy, as any stochastic policy that puts less than $1/2$ probability on $a^\star$ would incur at least $\epsilon/2$ loss.

The next step is to collect $18\log(3|\mathcal{A}|)$ independent datasets, and apply the algorithm on each of them to identify $a^\star$. We aggregate their predictions by majority vote; the portion of votes that $a^\star$ gets is an average of i.i.d. Bernoulli random variables. By Hoeffing's inequality, the probability that the actual average portion deviates from the expectation by $1/6$ is at most $\exp\{-2 \cdot (1/6)^2 \cdot 18\log(3|\mathcal{A}|)\} = \frac{1}{3|\mathcal{A}|}$, so given a fixed $M_{a,a^\star}$ we can identify $a^\star$ with at least $1 - \frac{1}{3|\mathcal{A}|}$ probability.

Of course, we would not be able to know which pair of actions contains $a^\star$ in advance. What we will do is to run the above procedure on all models in $\{M_{a,a'} : a, a' \in \mathcal{A}, a \neq a'\}$. By union bound, with probability at least $2/3$, the algorithm returns $a^\star$ on $M_{a^\star,a}$ for all $a \in \mathcal{A}$ simultaneously. (There is no guarantee on the results from $M_{a,a'}$ when $a, a' \neq a^\star$ but we do not care about them either.) When this happens, there exists a unique action that beats every other action in pair-wise comparison, which is $a^\star$. The lower bound follows by noticing that the sample complexity of this procedure is $18\log(3|\mathcal{A}|) \cdot g(1, 3, \frac{1}{3\epsilon}) = poly(\log|\mathcal{A}|, 1/\epsilon)$. □

H A situation where $|\overline{\mathcal{X}}_{\xi\text{-inc}}| = |\widetilde{\mathcal{X}}_{\xi\text{-inc}}| = 0$ but $|\mathcal{X}_{\xi\text{-inc}}|$ is arbitrarily large

Let M_1 and M_2 be two MDPs such that $v^\star_{M_1} = v^\star_{M_2}$. We create a bigger MDP M by taking the union of M_1's and M_2's state spaces and extending the horizon to $H + 2$. We add 3 states on the top: there is a single initial state at level 1; there are 2 states at level 2, each of which has only 1 action that transitions to the initial distribution of M_1 and M_2 respectively. The initial state has K actions, each of which transitions to both s_1 and s_2 with more than ξ probabilities (assuming $\xi < 0.5$). Now we construct $\widehat{M}$: $\widehat{M}$ only errs at the initial state, where each action transitions to s_1 deterministically. It is easy to verify that $|\overline{\mathcal{X}}_{\xi\text{-inc}}| = |\widetilde{\mathcal{X}}_{\xi\text{-inc}}| = 0$ and $|\mathcal{X}_{\xi\text{-inc}}| = K$.