The Sample-Complexity of General Reinforcement Learning

Tor Lattimore and **Marcus Hutter** and **Peter Sunehag**

Research School of Computer Science
Australian National University
{tor.lattimore,marcus.hutter,peter.sunehag}@anu.edu.au

July 2013

Abstract

We present a new algorithm for general reinforcement learning where the true environment is known to belong to a finite class of N arbitrary models. The algorithm is shown to be near-optimal for all but $O(N \log^2 N)$ time-steps with high probability. Infinite classes are also considered where we show that compactness is a key criterion for determining the existence of uniform sample-complexity bounds. A matching lower bound is given for the finite case.

Contents

1 Introduction **2**
2 Notation **3**
3 Finite Case **3**
4 Compact Case **12**
5 Unbounded Environment Classes **12**
6 Lower Bound **13**
7 Conclusions **13**
References **14**
A Technical Results **15**
B Constants **15**
C Table of Notation **16**

Keywords

Reinforcement learning; sample-complexity; exploration exploitation.

1 Introduction

Reinforcement Learning (RL) is the task of learning policies that lead to nearly-optimal rewards where the environment is unknown. One metric of the efficiency of an RL algorithm is sample-complexity, which is a high probability upper bound on the number of time-steps when that algorithm is not nearly-optimal that holds for all environment in some class. Such bounds are typically shown for very specific classes of environments, such as (partially observable/factored) Markov Decision Processes (MDP) and bandits. We consider more general classes of environments where at each time-step an agent takes an action $a \in A$ where-upon it receives reward $r \in [0, 1]$ and an observation $o \in O$, which are generated stochastically by the environment and may depend arbitrarily on the entire history sequence.

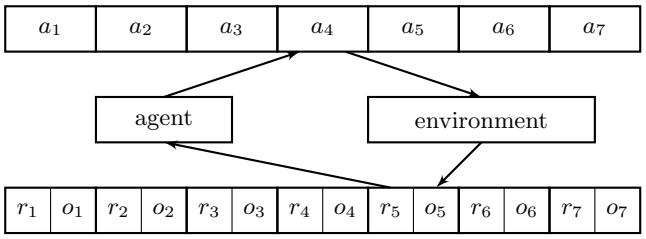

Figure 1: Agent/Environment Interaction

We present a new reinforcement learning algorithm, named Maximum Exploration Reinforcement Learning (MERL), that accepts as input a finite set $\mathcal{M} := \{\nu_1, \cdots, \nu_N\}$ of arbitrary environments, an accuracy ϵ, and a confidence δ. The main result is that MERL has a sample-complexity of

$$\tilde{O}\left(\frac{N}{\epsilon^2(1-\gamma)^3}\log^2\frac{N}{\delta\epsilon(1-\gamma)}\right),$$

where $1/(1-\gamma)$ is the effective horizon determined by discount rate γ. We also consider the case where $\mathcal{M}$ is infinite, but compact with respect to a particular topology. In this case, a variant of MERL has the same sample-complexity as above, but where N is replaced by the size of the smallest ϵ-cover. A lower bound is also given that matches the upper bound except for logarithmic factors. Finally, if $\mathcal{M}$ is non-compact then in general no finite sample-complexity bound exists.

Related work. Many authors have worked on the sample-complexity of RL in various settings. The simplest case is the multiarmed bandit problem that has been extensively studied with varying assumptions. The typical measure of efficiency in the bandit literature is regret, but sample-complexity bounds are also known and sometimes used. The next step from bandits is finite state MDPs, of which bandits are an example with only a single state. There are two main settings when MDPs are considered, the discounted case where sample-complexity bounds are proven and the undiscounted (average reward) case where regret bounds are more typical. In the discounted setting the upper and lower bounds on sample-complexity are now extremely refined. See Strehl et al. [2009] for a detailed review of the popular algorithms and theorems. More recent work on closing the gap between upper and lower bounds is by Szita and Szepesvári [2010], Lattimore and Hutter [2012], Azar et al. [2012]. In the undiscounted case it is necessary to make some form of ergodicity assumption as without this regret bounds cannot be given. In this work we avoid ergodicity assumptions and discount future rewards. Nevertheless, our algorithm borrows some tricks used by UCRL2 Auer et al. [2010]. Previous work for more general environment classes is somewhat limited. For factored MDPs there are known bounds, see Chakraborty and Stone [2011] and references there-in. Even-dar et al. [2005] give essentially unimprovable exponential bounds on the sample-complexity of learning in finite partially observable MDPs. Maillard et al. [2013] show

regret bounds for undiscounted RL where the true environment is assumed to be finite, Markov and communicating, but where the state is not directly observable. As far as we know there has been no work on the sample-complexity of RL when environments are completely general, but asymptotic results have garnered some attention with positive results by Hutter [2002], Ryabko and Hutter [2008], Sunehag and Hutter [2012] and (mostly) negative ones by Lattimore and Hutter [2011b]. Perhaps the closest related worked is Diuk et al. [2009], which deals with a similar problem in the rather different setting of learning the optimal predictor from a class of N experts. They obtain an $O(N \log N)$ bound, which is applied to the problem of structure learning for discounted finite-state factored MDPs. Our work generalises this approach to the non-Markov case and compact model classes.

2 Notation

The definition of environments is borrowed from the work of **?**, although the notation is slightly more formal to ease the application of martingale inequalities.

General. $\mathbb{N} = \{0, 1, 2, \cdots\}$ is the natural numbers. For the indicator function we write $[\![x = y]\!] = 1$ if $x = y$ and 0 otherwise. We use $\wedge$ and $\vee$ for logical and/or respectively. If A is a set then $|A|$ is its size and A^* is the set of all finite strings (sequences) over A. If x and y are sequences then $x \sqsubset y$ means that x is a prefix of y. Unless otherwise mentioned, log represents the natural logarithm. For random variable X we write $\mathbf{E}X$ for its expectation. For $x \in \mathbb{R}$, $\lceil x \rceil$ is the ceiling function.

Environments and policies. Let A, O and $R \subset \mathbb{R}$ be finite sets of actions, observations and rewards respectively and $\mathcal{H} := A \times O \times R$. $\mathcal{H}^\infty$ is the set of infinite history sequences while $\mathcal{H}^* := (A \times O \times R)^*$ is the set of finite history sequences. If $h \in \mathcal{H}^*$ then $\ell(h)$ is the number of action/observation/reward tuples in h. We write $a_t(h)$, $o_t(h)$, $r_t(h)$ for the tth action/observation/reward of history sequence h. For $h \in \mathcal{H}^*$, $\Gamma_h := \{h' \in \mathcal{H}^\infty : h \sqsubset h'\}$ is the cylinder set. Let $\mathcal{F} := \sigma(\{\Gamma_h : h \in \mathcal{H}^*\})$ and $\mathcal{F}_t := \sigma(\{\Gamma_h : h \in \mathcal{H}^* \wedge \ell(h) = t\})$ be σ-algebras. An environment μ is a set of conditional probability distributions over observation/reward pairs given the history so far. A policy π is a function $\pi : \mathcal{H}^* \to A$. An environment and policy interact sequentially to induce a measure, $P_{\mu,\pi}$, on filtered probability space $(\mathcal{H}^\infty, \mathcal{F}, \{\mathcal{F}_t\})$. For convenience, we abuse notation and write $P_{\mu,\pi}(h) := P_{\mu,\pi}(\Gamma_h)$. If $h \sqsubset h'$ then conditional probabilities are $P_{\mu,\pi}(h'|h) := P_{\mu,\pi}(h')/P_{\mu,\pi}(h)$. $R_t(h; d) := \sum_{k=t}^{t+d} \gamma^{k-t} r_k(h)$ is the d-step return function and $R_t(h) := \lim_{d\to\infty} R_t(h; d)$. Given history h_t with $\ell(h_t) = t$, the value function is defined by $V_\mu^\pi(h_t; d) := \mathbf{E}[R_t(h; d)|h_t]$ where the expectation is taken with respect to $P_{\mu,\pi}(\cdot|h_t)$. $V_\mu^\pi(h_t) := \lim_{d\to\infty} V_\mu^\pi(h_t; d)$. The optimal policy for environment μ is $\pi_\mu^* := \arg\max_\pi V_\mu^\pi$, which with our assumptions is known to exist Lattimore and Hutter [2011a]. The value of the optimal policy is $V_\mu^* := V_\mu^{\pi_\mu^*}$. In general, μ denotes the true environment while ν is a model. π will typically be the policy of the algorithm under consideration. $Q_\mu^*(h, a)$ is the value in history h of following policy π_μ^* except for the first time-step when action a is taken. $\mathcal{M}$ is a set of environments (models).

Sample-complexity. Policy π is ϵ-optimal in history h and environment μ if $V_\mu^*(h) - V_\mu^\pi(h) \le \epsilon$. The sample-complexity of a policy π in environment class $\mathcal{M}$ is the smallest Λ such that, with high probability, π is ϵ-optimal for all but Λ time-steps for all $\mu \in \mathcal{M}$. Define $L_{\mu,\pi}^\epsilon : \mathcal{H}^\infty \to \mathbb{N} \cup \{\infty\}$ to be the number of time-steps when π is not ϵ-optimal.

$$L_{\mu,\pi}^\epsilon(h) := \sum_{t=1}^{\infty} [\![V_\mu^*(h_t) - V_\mu^\pi(h_t) > \epsilon]\!] ,$$

where h_t is the length t prefix of h. The sample-complexity of policy π is Λ with respect to accuracy ϵ and confidence $1 - \delta$ if $\mathrm{P}\left\{L_{\mu,\pi}^\epsilon(h) > \Lambda\right\} < \delta, \forall \mu \in \mathcal{M}$.

3 Finite Case

We start with the finite case where the true environment is known to belong to a finite set of models, $\mathcal{M}$. The Maximum Exploration Reinforcement Learning algorithm is model-based in the sense that it maintains a set, $\mathcal{M}_t \subseteq \mathcal{M}$, where models are eliminated once they become implausible. The algorithm operates in phases of exploration and exploitation, choosing to exploit if it knows all plausible environments are reasonably close under all optimal policies and explore otherwise. This method of exploration essentially guarantees that MERL is nearly optimal whenever it is exploiting and the number of exploration phases is limited with high probability. The main difficulty is specifying what it means to be plausible. Previous authors working on finite environments, such as MDPs or bandits, have removed models for which the transition probabilities are not sufficiently close to their empirical estimates. In the more general setting this approach fails because states (histories) are never visited more than once, so sufficient empirical estimates cannot be collected. Instead, we eliminate environments if the reward we actually collect over time is not sufficiently close to the reward we expected given that environment.

Before giving the explicit algorithm, we explain the operation of MERL more formally in two parts. First we describe how it chooses to explore and exploit and then how the model class is maintained. See Figure 2 for a diagram of how exploration and exploitation occurs.

Exploring and exploiting. At each time-step t MERL computes the pair of environments $\underline{\nu}, \overline{\nu}$ in the model class $\mathcal{M}_t$ and the policy π maximising the difference

$$\Delta := V^\pi_{\overline{\nu}}(h;d) - V^\pi_{\underline{\nu}}(h;d), \quad d := \frac{1}{1-\gamma} \log \frac{8}{(1-\gamma)\epsilon}.$$

If $\Delta > \epsilon/4$, then MERL follows policy π for d time-steps, which we call an exploration phase. Otherwise, for one time-step it follows the optimal policy with respect to the first environment currently in the model class. Therefore, if MERL chooses to exploit, then all policies and environments in the model class lead to similar values, which implies that exploiting is near-optimal. If MERL explores, then either $V^\pi_{\overline{\nu}}(h;d) - V^\pi_\mu(h;d) > \epsilon/8$ or $V^\pi_\mu(h;d) - V^\pi_{\underline{\nu}}(h;d) > \epsilon/8$, which will allow us to apply concentration inequalities to eventually eliminate either $\overline{\nu}$ (the upper bound) or $\underline{\nu}$ (the lower bound).

The model class. An exploration phase is a κ-exploration phase if $\Delta \in [\epsilon 2^{\kappa-2}, \epsilon 2^{\kappa-1})$, where

$$\kappa \in \mathcal{K} := \left\{ 0, 1, 2, \cdots, \log_2 \frac{1}{\epsilon(1-\gamma)} + 2 \right\}.$$

For each environment $\nu \in \mathcal{M}$ and each $\kappa \in \mathcal{K}$, MERL associates a counter $E(\nu, \kappa)$, which is incremented at the start of a κ-exploration phase if $\nu \in \{\underline{\nu}, \overline{\nu}\}$. At the end of each κ-exploration phase MERL calculates the discounted return actually received during that exploration phase $R \in [0, 1/(1-\gamma)]$ and records the values

$$X(\overline{\nu}, \kappa) := (1-\gamma)(V^\pi_{\overline{\nu}}(h;d) - R)$$
$$X(\underline{\nu}, \kappa) := (1-\gamma)(R - V^\pi_{\underline{\nu}}(h;d)),$$

where h is the history at the start of the exploration phase. So $X(\overline{\nu}, \kappa)$ is the difference between the return expected if the true model was $\overline{\nu}$ and the actual return and $X(\underline{\nu}, \kappa)$ is the difference between the actual return and the expected return if the true model was $\underline{\nu}$. Since the expected value of R is $V^\pi_\mu(h;d)$, and $\overline{\nu}$,$\underline{\nu}$ are upper and lower bounds respectively, the expected values of both $X(\overline{\nu}, \kappa)$ and $X(\underline{\nu}, \kappa)$ are non-negative and at least one of them has expectation larger than $(1-\gamma)\epsilon/8$.

MERL eliminates environment ν from the model class if the cumulative sum of $X(\nu, \kappa)$ over all exploration phases where $\nu \in \{\underline{\nu}, \overline{\nu}\}$ is sufficiently large, but it tests this condition only when the counts $E(\nu, \kappa)$ has increased enough since the last test. Let $\alpha_j := \lceil \alpha^j \rceil$ for $\alpha \in (1, 2)$ as defined in

the algorithm. MERL only tests if ν should be removed from the model class when $E(\nu, \kappa) = \alpha_j$ for some $j \in \mathbb{N}$. This restriction ensures that tests are not performed too often, which allows us to apply the union bound without losing too much. Note that if the true environment $\mu \in \{\overline{\nu}, \underline{\nu}\}$, then $\mathbf{E}_{\mu,\pi} X(\mu, \kappa) = 0$, which will ultimately be enough to ensure that μ remains in the model class with high probability. The reason for using κ to bucket exploration phases will become apparent later in the proof of Lemma 3.

Algorithm 1 MERL

1: **Inputs:** ϵ, δ and $\mathcal{M} := \{\nu_1, \nu_2, \cdots, \nu_N\}$.
2: $t = 1$ and h empty history
3: $d := \frac{1}{1-\gamma} \log \frac{8}{(1-\gamma)\epsilon}$, $\delta_1 := \frac{\delta}{32|\mathcal{K}|N^{3/2}}$
4: $\alpha := \frac{4\sqrt{N}}{4\sqrt{N}-1}$ and $\alpha_j := \lceil \alpha^j \rceil$
5: $E(\nu, \kappa) := 0, \quad \forall \nu \in \mathcal{M}$ and $\kappa \in \mathbb{N}$
6: **loop**
7: **repeat**
8: $\Pi := \{\pi^*_\nu : \nu \in \mathcal{M}\}$
9: $\overline{\nu}, \underline{\nu}, \pi := \underset{\overline{\nu}, \underline{\nu} \in \mathcal{M}, \pi \in \Pi}{\arg\max} \; V^\pi_{\overline{\nu}}(h; d) - V^\pi_{\underline{\nu}}(h; d)$
10: **if** $\Delta := V^\pi_{\overline{\nu}}(h; d) - V^\pi_{\underline{\nu}}(h; d) > \epsilon/4$ **then**
11: $\tilde{h} = h$ and $R = 0$
12: **for** $j = 0 \to d$ **do**
13: $R = R + \gamma^j r_t(h)$
14: $\textsc{Act}(\pi)$
15: **end for**
16: $\kappa := \min\left\{\kappa \in \mathbb{N} : \Delta > \epsilon 2^{\kappa-2}\right\}$.
17: $E(\underline{\nu}, \kappa) = E(\underline{\nu}, \kappa) + 1$ and $E(\overline{\nu}, \kappa) = E(\overline{\nu}, \kappa) + 1$
18: $X(\overline{\nu}, \kappa)_{E(\overline{\nu}, \kappa)} = (1 - \gamma)(V^\pi_{\overline{\nu}}(\tilde{h}; d) - R)$
19: $X(\underline{\nu}, \kappa)_{E(\underline{\nu}, \kappa)} = (1 - \gamma)(R - V^\pi_{\underline{\nu}}(\tilde{h}; d))$
20: **else**
21: $i := \min\{i : \nu_i \in \mathcal{M}\}$ and $\textsc{Act}(\pi^*_{\nu_i})$
22: **end if**
23: **until** $\exists \nu \in \mathcal{M}, \kappa, j \in \mathbb{N}$ such that $E(\nu, \kappa) = \alpha_j$ and

$$\sum_{i=1}^{E(\nu,\kappa)} X(\nu, \kappa)_i \geq \sqrt{2E(\nu, \kappa) \log \frac{E(\nu, \kappa)}{\delta_1}}.$$

24: $\mathcal{M} = \mathcal{M} - \{\nu\}$
25: **end loop**
26: **function** $\textsc{Act}(\pi)$
27: Take action $a_t = \pi(h)$ and receive reward and observation r_t, o_t from environment
28: $t \leftarrow t + 1$ and $h \leftarrow h a_t o_t r_t$
29: **end function**

Subscripts. For clarity, we have omitted subscripts in the pseudo-code above. In the analysis we will refer to $E_t(\nu, \kappa)$ and $\mathcal{M}_t$ for the values of $E(\nu, \kappa)$ and $\mathcal{M}$ respectively at time-step t. We write ν_t for ν_i in line 21 and similarly $\pi_t := \pi^*_{\nu_t}$.

Phases. An *exploration phase* is a period of exactly d time-steps, starting at time-step t if

1. t is not currently in an exploration phase.
2. $\Delta := V^\pi_{\overline{\nu}}(h_t; d) - V^\pi_{\underline{\nu}}(h_t; d) > \epsilon/4$.

We say it is a ν-exploration phase if $\nu = \underline{\nu}$ or $\nu = \overline{\nu}$ and a κ-exploration phase if $\Delta \in [\epsilon 2^{\kappa-2}, \epsilon 2^{\kappa-1}) \equiv [\epsilon_\kappa, 2\epsilon_\kappa)$ where $\epsilon_\kappa := \epsilon 2^{\kappa-2}$. It is a (ν, κ)-exploration phase if it satisfies both

of the previous statements. We say that MERL is *exploiting* at time-step t if t is not in an exploration phase. A *failure* phase is also a period of d time-steps and starts in time-step t if

1. t is not in an exploration phase or earlier failure phase
2. $V^*_\mu(h_t) - V^\pi_\mu(h_t) > \epsilon$.

Unlike exploration phases, the algorithm does not depend on the failure phases, which are only used in the analysis, An exploration or failure phase starting at time-step t is *proper* if $\mu \in \mathcal{M}_t$. The effective horizon d is chosen to ensure that $V^\pi_\mu(h; d) \geq V^\pi_\mu(h) - \epsilon/8$ for all π, μ and h.

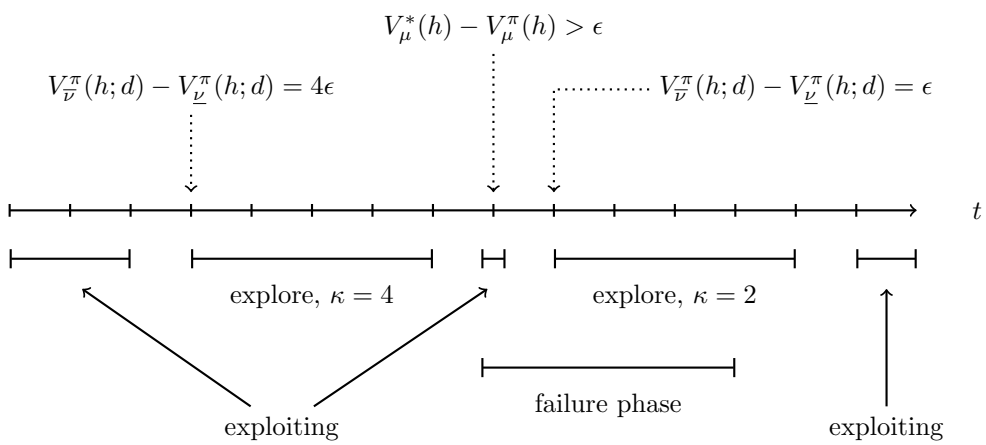

Figure 2: Exploration/exploitation/failure phases, $d = 4$

Test statistics. We have previously remarked that most traditional model-based algorithms with sample-complexity guarantees record statistics about the transition probabilities of an environment. Since the environments are assumed to be finite, these statistics eventually become accurate (or irrelevant) and the standard theory on the concentration of measure can be used for hypothesis testing. In the general case, environments can be infinite and so we cannot collect useful statistics about individual transitions. Instead, we use the statistics $X(\nu, \kappa)$, which are dependent on the value function rather than individual transitions. These satisfy $\mathbf{E}_{\mu,\pi}[X(\mu, \kappa)_i] = 0$ while $\mathbf{E}_{\mu,\pi}[X(\nu, \kappa)_i] \geq 0$ for all $\nu \in \mathcal{M}_t$. Testing is then performed on the statistic $\sum_{i=1}^{\alpha_k} X(\nu, \kappa)_i$, which will satisfy certain martingale inequalities.

Updates. As MERL explores, it updates its model class, $\mathcal{M}_t \subseteq \mathcal{M}$, by removing environments that have become implausible. This is comparable to the updating of confidence intervals for algorithms such as MBIE (Strehl and Littman, 2005) or UCRL2 (Auer et al., 2010). In MBIE, the confidence interval about the empirical estimate of a transition probability is updated after every observation. A slight theoretical improvement used by UCRL2 is to only update when the number of samples of a particular statistic doubles. The latter trick allows a cheap application of the union bound over all updates without wasting too many samples. For our purposes, however, we need to update slightly more often than the doubling trick would allow. Instead, we check if an environment should be eliminated if the number of (ν, κ)-exploration phases is exactly α_j for some j where $\alpha_j := \lceil \alpha^j \rceil$ and $\alpha := \frac{4\sqrt{N}}{4\sqrt{N}-1} \in (1, 2)$. Since the growth of α_j is still exponential, the union bound will still be applicable.

Probabilities. For the remainder of this section, unless otherwise mentioned, all probabilities and expectations are with respect to $P_{\mu,\pi}$ where π is the policy of Algorithm 1 and $\mu \in \mathcal{M}$ is the true environment.

Analysis. Define $G_{\max} := \frac{2^{16}N|\mathcal{K}|}{\epsilon^2(1-\gamma)^2}\log^2\frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$ and $E_{\max} := \frac{2^{16}N}{\epsilon^2(1-\gamma)^2}\log^2\frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$, which are high probability bounds on the number of failure and exploration phases respectively.

Theorem 1. *Let $\mu \in \mathcal{M} = \{\nu_1, \nu_2, \cdots \nu_N\}$ be the true environment and π be the policy of Algorithm 1. Then*

$$\mathrm{P}\left\{L^{\epsilon}_{\mu,\pi}(h) \geq d \cdot (G_{\max} + E_{\max})\right\} \leq \delta.$$

If lower order logarithmic factors are dropped then the sample-complexity bound of MERL given by Theorem 1 is $\tilde{O}\left(\frac{N}{\epsilon^2(1-\gamma)^3}\log^2\frac{N}{\delta\epsilon(1-\gamma)}\right)$. Theorem 1 follows from three lemmas.

Lemma 2. *$\mu \in \mathcal{M}_t$ for all t with probability $1-\delta/4$.*

Lemma 3. *The number of proper failure phases is bounded by*

$$G_{\max} := \frac{2^{16}N|\mathcal{K}|}{\epsilon^2(1-\gamma)^2}\log^2\frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$$

with probability at least $1-\frac{\delta}{2}$.

Lemma 4. *The number of proper exploration phases is bounded by*

$$E_{\max} := \frac{2^{16}N}{\epsilon^2(1-\gamma)^2}\log^2\frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$$

with probability at least $1-\frac{\delta}{4}$.

Proof of Theorem 1. Applying the union bound to the results of Lemmas 2, 3 and 4 gives the following with probability at least $1-\delta$.

1. There are no non-proper exploration or failure phases.
2. The number of proper exploration phases is at most $E_{\max}$.
3. The number of proper failure phases is at most $G_{\max}$.

If π is not ϵ-optimal at time-step t then t is either in an exploration or failure phase. Since both are exactly d time-steps long the total number of time-steps when π is sub-optimal is at most $d \cdot (G_{\max} + E_{\max})$. ■

We now turn our attention to proving Lemmas 2, 3 and 4. Of these, Lemma 4 is more conceptually challenging while Lemma 3 is intuitively unsurprising, but technically difficult.

Proof of Lemma 2. If μ is removed from $\mathcal{M}$, then there exists a κ and $j \in \mathbb{N}$ such that

$$\sum_{i=1}^{\alpha_j} X(\mu,\kappa)_i \geq \sqrt{2\alpha_j \log\frac{\alpha_j}{\delta_1}}.$$

Fix a $\kappa \in \mathcal{K}$, $E_\infty(\mu,\kappa) := \lim_t E_t(\mu,\kappa)$ and $X_i := X(\mu,\kappa)_i$. Define a sequence of random variables

$$\tilde{X}_i := \begin{cases} X_i & \text{if } i \leq E_\infty(\mu,\kappa) \\ 0 & \text{otherwise.} \end{cases}$$

Now we claim that $B_n := \sum_{i=1}^n \tilde{X}_i$ is a martingale with $|B_{i+1} - B_i| \leq 1$ and $\mathbf{E}B_i = 0$. That it is a martingale with zero expectation follows because if t is the time-step at the start of the exploration phase associated with variable X_i, then $\mathbf{E}[X_i|\mathcal{F}_t] = 0$. $|B_{i+1} - B_i| \leq 1$ because discounted returns are bounded in $[0, 1/(1-\gamma)]$ and by the definition of X_i.

For all $j \in \mathbb{N}$, we have by Azuma's inequality that

$$\mathrm{P}\left\{B_{\alpha_j} \geq \sqrt{2\alpha_j \log \frac{\alpha_j}{\delta_1}}\right\} \leq \frac{\delta_1}{\alpha_j}.$$

Apply the union bound over all j.

$$\mathrm{P}\left\{\exists j \in \mathbb{N} : B_{\alpha_j} \geq \sqrt{2\alpha_j \log \frac{\alpha_j}{\delta_1}}\right\} \leq \sum_{j=1}^{\infty} \frac{\delta_1}{\alpha_j}.$$

Complete the result by the union bound over all κ, applying Lemma 10 (see Appendix) and the definition of δ_1 to bound $\sum_{\kappa \in \mathcal{K}} \sum_{j=1}^{\infty} \frac{\delta_1}{\alpha_j} \leq \delta/4$. ■

We are now ready to give a high-probability bound on the number of proper exploration phases. If MERL starts a proper exploration phase at time-step t then at least one of the following holds:

1. $\mathbf{E}[X(\underline{\nu}, \kappa)_{E(\underline{\nu},\kappa)} | \mathcal{F}_t] > (1-\gamma)\epsilon/8$.
2. $\mathbf{E}[X(\overline{\nu}, \kappa)_{E(\overline{\nu},\kappa)} | \mathcal{F}_t] > (1-\gamma)\epsilon/8$.

This contrasts with $\mathbf{E}[X(\mu, \kappa)_{E(\mu,\kappa)} | \mathcal{F}_t] = 0$, which ensures that μ remains in $\mathcal{M}$ for all time-steps. If one could know which of the above statements were true at each time-step then it would be comparatively easy to show by means of Azuma's inequality that all environments that are not ϵ-close are quickly eliminated after $O(\frac{1}{\epsilon^2(1-\gamma)^2})$ ν-exploration phases, which would lead to the desired bound. Unfortunately though, the truth of (1) or (2) above cannot be determined, which greatly increases the complexity of the proof.

Proof of Lemma 4. Fix a $\kappa \in \mathcal{K}$ and let $E_{\max,\kappa}$ be a constant to be chosen later. Let h_t be the history at the start of some κ-exploration phase. We say an $(\underline{\nu}, \kappa)$-exploration phase is $\underline{\nu}$-effective if

$$\begin{aligned}\mathbf{E}[X(\underline{\nu}, \kappa)_{E(\underline{\nu},\kappa)} | \mathcal{F}_t] &\equiv (1-\gamma)(V_\mu^\pi(h_t; d) - V_{\underline{\nu}}^\pi(h_t; d)) \\ &> (1-\gamma)\epsilon_\kappa/2\end{aligned}$$

and $\overline{\nu}$-effective if the same condition holds for $\overline{\nu}$. Now since t is the start of a proper exploration phase we have that $\mu \in \mathcal{M}_t$ and so

$$\begin{aligned}&V_{\overline{\nu}}^\pi(h_t; d) \geq V_\mu^\pi(h_t; d) \geq V_{\underline{\nu}}^\pi(h_t; d) \\ &V_{\overline{\nu}}^\pi(h_t; d) - V_{\underline{\nu}}^\pi(h_t; d) > \epsilon_\kappa.\end{aligned}$$

Therefore every proper exploration phase is either $\underline{\nu}$-effective or $\overline{\nu}$-effective. Let $E_{t,\kappa} := \sum_\nu E_t(\nu, \kappa)$, which is twice the number of κ-exploration phases at time t and $E_{\infty,\kappa} := \lim_t E_{t,\kappa}$, which is twice the total number of κ-exploration phases.[1] Let $F_t(\nu, \kappa)$ be the number of ν-effective (ν, κ)-exploration phases up to time-step t. Since each proper κ-exploration phase is either $\underline{\nu}$-effective or $\overline{\nu}$-effective or both, $\sum_\nu F_t(\nu, \kappa) \geq E_{t,\kappa}/2$. Applying Lemma 8 to $y_\nu := E_t(\nu, \kappa)/E_{t,\kappa}$ and $x_\nu := F_t(\nu, \kappa)/E_t(\nu, \kappa)$ shows that if $E_{\infty,\kappa} > E_{\max,\kappa}$ then there exists a t' and ν such that $E_{t',\kappa} = E_{\max,\kappa}$ and

$$\frac{F_{t'}(\nu, \kappa)^2}{E_{\max,\kappa} E_{t'}(\nu, \kappa)} \geq \frac{1}{4N}, \tag{1}$$

which implies that

$$F_{t'}(\nu, \kappa) \geq \sqrt{\frac{E_{\max,\kappa} E_{t'}(\nu, \kappa)}{4N}} \overset{(a)}{\geq} \frac{E_{t'}(\nu, \kappa)}{\sqrt{4N}}, \tag{2}$$

[1] Note that it is never the case that $\overline{\nu} = \underline{\nu}$ at the start of an exploration phase, since in this case $\Delta = 0$.

where (a) follows because $E_{\max,\kappa} = E_{t',\kappa} \geq E_{t'}(\nu,\kappa)$. Let $Z(\nu)$ be the event that there exists a t' satisfying (1). We will shortly show that $\mathrm{P}\{Z(\nu)\} < \delta/(4N|\mathcal{K}|)$. Therefore

$$\begin{aligned}\mathrm{P}\{E_{\infty,\kappa} > E_{\max,\kappa}\} &\leq \mathrm{P}\{\exists \nu : Z(\nu)\} \leq \sum_{\nu\in\mathcal{M}} \mathrm{P}\{Z(\nu)\} \\ &\leq \delta/(4|\mathcal{K}|)\end{aligned}$$

Finally take the union bound over all κ and let

$$E_{\max} := \sum_{\kappa\in\mathcal{K}} \frac{1}{2} E_{\max,\kappa},$$

where we used $\frac{1}{2}E_{\max,\kappa}$ because $E_{\max,\kappa}$ is a high-probability upper bound on $E_{\infty,\kappa}$, which is *twice* the number of κ-exploration phases.

Bounding $\mathrm{P}\{Z(\nu)\} < \delta/(4N|\mathcal{K}|)$. Fix a $\nu \in \mathcal{M}$ and let $X_1, X_2, \cdots, X_{E_\infty(\nu,\kappa)}$ be the sequence with $X_i := X(\nu,\kappa)_i$ and let t_i be the corresponding time-step at the start of the ith (ν,κ)-exploration phase. Define a sequence

$$Y_i := \begin{cases} X_i - \mathbf{E}[X_i|\mathcal{F}_{t_i}] & \text{if } i \leq E_\infty(\nu,\kappa) \\ 0 & \text{otherwise} \end{cases}$$

Let $\lambda(E) := \sqrt{2E\log\frac{E}{\delta_1}}$. Now if $Z(\nu)$, then the largest time-step $t \leq t'$ with $E_t(\nu,t) = \alpha_j$ for some $j \in \mathbb{N}$ is

$$t := \max\{t \leq t' : \exists j \in \mathbb{N} \text{ s.t. } \alpha_j = E_t(\nu,t)\},$$

which exists and satisfies

1. $E_t(\nu,\kappa) = \alpha_j$ for some j.
2. $E_\infty(\nu,\kappa) > E_t(\nu,\kappa)$.
3. $F_t(\nu,\kappa) \geq \sqrt{E_t(\nu,\kappa)E_{\max,\kappa}/(16N)}$.
4. $E_t(\nu,\kappa) \geq E_{\max,\kappa}/(16N)$.

where parts 1 and 2 are straightforward and parts 3 and 4 follow by the definition of $\{\alpha_j\}$, which was chosen specifically for this part of the proof. Since $E_\infty(\nu,\kappa) > E_t(\nu,\kappa)$, at the end of the exploration phase starting at time-step t, ν must remain in $\mathcal{M}$. Therefore

$$\begin{aligned}\lambda(\alpha_j) &\overset{(a)}{\geq} \sum_{i=1}^{\alpha_j} X_i \overset{(b)}{\geq} \sum_{i=1}^{\alpha_j} Y_i + \frac{\epsilon_\kappa(1-\gamma)F_t(\nu,\kappa)}{2} \\ &\overset{(c)}{\geq} \sum_{i=1}^{\alpha_j} Y_i + \frac{\epsilon_\kappa(1-\gamma)}{8}\sqrt{\frac{\alpha_j E_{\max,\kappa}}{N}},\end{aligned} \tag{3}$$

where in (a) we used the definition of the confidence interval of MERL. In (b) we used the definition of Y_i and the fact that $\mathbf{E}X_i \geq 0$ for all i and $\mathbf{E}X_i \geq \epsilon_\kappa(1-\gamma)/2$ if X_i is effective. Finally we used the lower bound on the number of effective ν-exploration phases, $F_t(\nu,\kappa)$ (part 3 above). If $E_{\max,\kappa} := \frac{2^{11}N}{\epsilon_\kappa^2(1-\gamma)^2}\log^2\frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$, then by applying Lemma 9 with $a = \frac{2^9 N}{\epsilon_\kappa^2(1-\gamma)^2}$ and $b = 1/\delta_1$ we obtain

$$E_{\max,\kappa} \geq \frac{2^9 N}{\epsilon_\kappa^2(1-\gamma)^2}\log\frac{E_{\max,\kappa}}{\delta_1} \geq \frac{2^9 N}{\epsilon_\kappa^2(1-\gamma)^2}\log\frac{\alpha_j}{\delta_1}$$

Multiplying both sides by α_j and rearranging and using the definition of $\lambda(\alpha_j)$ leads to

$$\frac{\epsilon_\kappa(1-\gamma)}{8}\sqrt{\frac{\alpha_j E_{\max,\kappa}}{N}} \geq 2\lambda(\alpha_j).$$

Inserting this into Equation (3) shows that $Z(\nu)$ implies that there exists an α_j such that $\sum_{i=1}^{\alpha_j} Y_i \leq -\lambda(\alpha_j)$. Now by the same argument as in the proof of Lemma 2, $B_n := \sum_{i=1}^{n} Y_i$ is a martingale with $|B_{i+1} - B_i| \leq 1$. Therefore by Azuma's inequality

$$\mathrm{P}\left\{\sum_{i=1}^{\alpha_j} Y_i \leq -\lambda(\alpha_j)\right\} \leq \frac{\delta_1}{\alpha_j}.$$

Finally apply the union bound over all j. ■

Recall that if MERL is exploiting at time-step t, then π_t is the optimal policy with respect to the first environment in the model class. To prove Lemma 3 we start by showing that in this case π_t is nearly-optimal.

Lemma 5. *Let t be a time-step and h_t be the corresponding history. If $\mu \in \mathcal{M}_t$ and MERL is exploiting (not exploring), then $V_\mu^*(h_t) - V_\mu^{\pi_t}(h_t) \leq 5\epsilon/8$.*

Proof of Lemma 5. Since MERL is not exploring

$$\begin{aligned} V_\mu^*(h_t) - V_\mu^{\pi_t}(h_t) &\overset{(a)}{\leq} V_\mu^*(h_t;d) - V_\mu^{\pi_t}(h_t;d) + \frac{\epsilon}{8} \\ &\overset{(b)}{\leq} V_{\nu_t}^{\pi_\mu^*}(h_t;d) - V_{\nu_t}^{\pi_t}(h_t;d) + 5\epsilon/8 \\ &\overset{(c)}{\leq} 5\epsilon/8, \end{aligned}$$

(a) follows by truncating the value function. (b) follows because $\mu \in \mathcal{M}_t$ and MERL is exploiting. (c) is true since π_t is the optimal policy in ν_t. ■

Lemma 5 is almost sufficient to prove Lemma 3. The only problem is that MERL only follows $\pi_t = \pi_{\nu_t}^*$ until there is an exploration phase. The idea to prove Lemma 3 is as follows:

1. If there is a low probability of entering an exploration phase within the next d time-steps following policy π_t, then π is nearly as good as π_t, which itself is nearly optimal by Lemma 5.

2. The number of time-steps when the probability of entering an exploration phase within the next d time-steps is high is unlikely to be too large before an exploration phase is triggered. Since there are not many exploration phases with high probability, there are also unlikely to be too many time-steps when π expects to enter one with high probability.

Before the proof of Lemma 3 we remark on an easier to prove (but weaker) version of Theorem 1. If MERL is exploiting then Lemma 5 shows that $V_\mu^*(h) - Q_\mu^*(h, \pi(h)) \leq 5\epsilon/8 < \epsilon$. Therefore if we cared about the number of time-steps when this is not the case (rather than $V_\mu^* - V_\mu^\pi$), then we would already be done by combining Lemmas 4 and 5.

Proof of Lemma 3. Let t be the start of a proper failure phase with corresponding history, h. Therefore $V_\mu^*(h) - V_\mu^\pi(h) > \epsilon$. By Lemma 5, $V_\mu^*(h) - V_\mu^\pi(h) = V_\mu^*(h) - V_\mu^{\pi_t}(h) + V_\mu^{\pi_t}(h) - V_\mu^\pi(h) \leq 5\epsilon/8 + V_\mu^{\pi_t} - V_\mu^\pi(h)$ and so

$$V_\mu^{\pi_t}(h) - V_\mu^\pi(h) \geq \frac{3\epsilon}{8}. \tag{4}$$

We define set $\mathcal{H}_\kappa \subset \mathcal{H}^*$ to be the set of extensions of h that trigger κ-exploration phases. Formally $\mathcal{H}_\kappa \subset \mathcal{H}^*$ is the prefix free set such that h' in $\mathcal{H}_\kappa$ if $h \sqsubset h'$ and h' triggers a κ-exploration phase for the first time since t. Let $\mathcal{H}_{\kappa,d} := \{h' : h' \in \mathcal{H}_\kappa \wedge \ell(h') \leq t + d\}$, which is the set of extensions of h that are at most d long and trigger κ-exploration phases. Therefore

$$\begin{aligned}
\frac{3\epsilon}{8} &\overset{(a)}{\leq} V_\mu^{\pi_t}(h) - V_\mu^{\pi}(h) \\
&\overset{(b)}{=} \sum_{\kappa \in \mathcal{K}} \sum_{h' \in \mathcal{H}_\kappa} P(h'|h)\gamma^{\ell(h')-t} \left(V_\mu^{\pi_t}(h') - V_\mu^{\pi}(h')\right) \\
&\overset{(c)}{\leq} \sum_{\kappa \in \mathcal{K}} \sum_{h' \in \mathcal{H}_{\kappa,d}} P(h'|h) \left(V_\mu^{\pi_t}(h') - V_\mu^{\pi}(h')\right) + \frac{\epsilon}{8} \\
&\overset{(d)}{\leq} \sum_{\kappa \in \mathcal{K}} \sum_{h' \in \mathcal{H}_{\kappa,d}} P(h'|h) \left(V_\mu^{*}(h';d) - V_\mu^{\pi}(h';d)\right) + \frac{\epsilon}{4} \\
&\overset{(e)}{\leq} \sum_{\kappa \in \mathcal{K}} \sum_{h' \in \mathcal{H}_{\kappa,d}} P(h'|h) 4\epsilon_\kappa + \frac{\epsilon}{4},
\end{aligned}$$

(a) follows from Equation (4). (b) by noting that that $\pi = \pi_t$ until an exploration phase is triggered. (c) by replacing $\mathcal{H}_\kappa$ with $\mathcal{H}_{\kappa,d}$ and noting that if $h' \in \mathcal{H}_\kappa - \mathcal{H}_{\kappa,d}$, then $\gamma^{\ell(h')-t} \leq (1-\gamma)\epsilon/8$. (d) by substituting $V_\mu^*(h') \geq V_\mu^{\pi_t}(h')$ and by using the effective horizon to truncate the value functions. (e) by the definition of a κ-exploration phase.

Since the maximum of a set is greater than the average, there exists a $\kappa \in \mathcal{K}$ such that $\sum_{h' \in \mathcal{H}_{\kappa,d}} P(h'|h) \geq 2^{-\kappa-3}/|\mathcal{K}|$, which is the probability that MERL encounters a κ-exploration phase within d time-steps from h. Now fix a κ and let $t_1, t_2, \cdots, \cdots, t_{G_\kappa}$ be the sequence of time-steps such that t_i is the start of a failure phase and the probability of a κ-exploration phase within the next d time-steps is at least $2^{-\kappa-3}/|\mathcal{K}|$. Let $Y_i \in \{0,1\}$ be the event that a κ-exploration phase does occur within d time-steps of t_i and define an auxiliary infinite sequence $\tilde{Y}_1, \tilde{Y}_2, \cdots$ by $\tilde{Y}_i := Y_i$ if $i \leq G_\kappa$ and 1 otherwise. Let E_κ be the number of κ-exploration phases and $G_{\max,\kappa}$ be a constant to be chosen later and suppose $G_\kappa > G_{\max,\kappa}$, then $\sum_{i=1}^{G_{\max,\kappa}} \tilde{Y}_i = \sum_{i=1}^{G_{\max,\kappa}} Y_i$ and either $\sum_{i=1}^{G_{\max,\kappa}} \tilde{Y}_i \leq E_{\max,\kappa}$ or $E_\kappa > E_{\max,\kappa}$, where the latter follows because $Y_i = 1$ implies a κ-exploration phase occurred. Therefore

$$\begin{aligned}
&\mathrm{P}\left\{G_\kappa > G_{\max,\kappa}\right\} \\
&\leq \mathrm{P}\left\{\sum_{i=1}^{G_{\max,\kappa}} \tilde{Y}_i < E_{\max,\kappa}\right\} + \mathrm{P}\left\{E_\kappa > E_{\max,\kappa}\right\} \\
&\leq \mathrm{P}\left\{\sum_{i=1}^{G_{\max,\kappa}} \tilde{Y}_i < E_{\max,\kappa}\right\} + \frac{\delta}{4|\mathcal{K}|}.
\end{aligned}$$

We now choose $G_{\max,\kappa}$ sufficiently large to bound the first term in the display above by $\delta/(4|\mathcal{K}|)$. By the definition of $\tilde{Y}_i$ and Y_i, if $i \leq G_\kappa$ then $\mathbf{E}[\tilde{Y}_i|\mathcal{F}_{t_i}] \geq 2^{-\kappa-3}/|\mathcal{K}|$ and for $i > G_\kappa$, $\tilde{Y}_i$ is always 1. Setting

$$\begin{aligned}
G_{\max,\kappa} &:= 2^{\kappa+4}|\mathcal{K}|E_{\max,\kappa} \\
&= \frac{2^{17}N|\mathcal{K}|}{\epsilon\epsilon_\kappa(1-\gamma)^2} \log^2 \frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}
\end{aligned}$$

is sufficient to guarantee $\mathbf{E}[\sum_{i=1}^{G_{\max,\kappa}} \tilde{Y}_i] > 2E_{\max,\kappa}$ and an application of Azuma's inequality to the martingale difference sequence completes the result. Finally we apply the union bound over all κ and set $G_{\max} := \sum_{\kappa \in \mathbb{N}} G_{\max,\kappa} > \sum_{\kappa \in \mathcal{K}} G_{\max,\kappa}$. ■

4 Compact Case

In the last section we presented MERL and proved a sample-complexity bound for the case when the environment class is finite. In this section we show that if the number of environments is infinite, but compact with respect to the topology generated by a natural metric, then sample-complexity bounds are still possible with a minor modification of MERL. The key idea is to use compactness to cover the space of environments with ϵ-balls and compute statistics on these balls rather than individual environments. Since all environments in the same ϵ-ball are sufficiently close, the resulting statistics cannot be significantly different and all analysis goes through identically to the finite case. Define a topology on the space of all environments induced by the pseudo-metric

$$d(\nu_1, \nu_2) := \sup_{h,\pi} |V^{\pi}_{\nu_1}(h) - V^{\pi}_{\nu_2}(h)|.$$

Theorem 6. *Let $\mathcal{M}$ be compact and coverable by N ϵ-balls then a modification of Algorithm 1 satisfies*

$$\mathrm{P}\left\{L^{2\epsilon}_{\mu,\pi}(h) \geq d \cdot (G_{\max} + E_{\max})\right\} \leq \delta.$$

The main modification is to define statistics on elements of the cover, rather than specific environments.

1. Let $U_1, \cdots, U_N$ be an ϵ-cover of $\mathcal{M}$.
2. At each time-step choose $\underline{U}$ and $\overline{U}$ such that $\underline{\nu} \in \underline{U}$ and $\overline{\nu} \in \overline{U}$.
3. Define statistics $\{X\}$ on elements of the cover, rather than environments, by

$$X(\underline{U}, \kappa)_{E(\underline{U},\kappa)} := \inf_{\underline{\nu} \in \underline{U}} (1-\gamma)(R - V^{\pi}_{\underline{\nu}}(h))$$
$$X(\overline{U}, \kappa)_{E(\overline{U},\kappa)} := \inf_{\overline{\nu} \in \overline{U}} (1-\gamma)(V^{\pi}_{\overline{\nu}}(h) - R)$$

4. If there exists a U where the test fails then eliminate all environments in that cover.

The proof requires only small modifications to show that with high probability the U containing the true environment is never discarded, while those not containing the true environment are if tested sufficiently often.

5 Unbounded Environment Classes

If the environment class is non-compact then we cannot in general expect finite sample-complexity bounds. Indeed, even asymptotic results are usually not possible.

Theorem 7. *There exist non-compact $\mathcal{M}$ for which no agent has a finite PAC bound.*

The obvious example is when $\mathcal{M}$ is the set of all environments. Then for any policy $\mathcal{M}$ includes an environment that is tuned to ensure the policy acts sub-optimally infinitely often. A more interesting example is the class of all computable environments, which is non-compact and also does not admit algorithms with uniform finite sample-complexity. See negative results by Lattimore and Hutter [2011b] for counter-examples.

6 Lower Bound

We now turn our attention to the lower bound. In specific cases, the bound in Theorem 1 is very weak. For example, if $\mathcal{M}$ is the class of finite MDPs with $|S|$ states then a natural covering leads to a PAC bound with exponential dependence on the state-space while it is known that the true dependence is at most quadratic. This should not be surprising since information about the transitions for one state gives information about a large subset of $\mathcal{M}$, not just a single environment. We show that the bound in Theorem 1 is unimprovable for general environment classes except for logarithmic factors. That is, there exists a class of environments where Theorem 1 is nearly tight.

The simplest counter-example is a set of MDPs with four states, $S = \{0, 1, \oplus, \ominus\}$ and N actions, $A = \{a_1, \cdots, a_N\}$. The rewards and transitions are depicted in Figure 3 where the transition probabilities depend on the action. Let $\mathcal{M} := \{\nu_1, \cdots, \nu_N\}$ where for ν_k we set $\epsilon(a_i) = [\![i = k]\!]\epsilon(1-\gamma)$. Therefore in environment ν_k, a_k is the optimal action in state 1. $\mathcal{M}$ can be viewed as a set of bandits with rewards in $(0, 1/(1-\gamma))$. In the bandit domain tight lower bounds on sample-complexity are known and given in Mannor and Tsitsiklis [2004]. These results can be applied as in Strehl et al. [2009] and Lattimore and Hutter [2012] to show that no algorithm has sample-complexity less than $O(\frac{N}{\epsilon^2(1-\gamma)^3}\log\frac{1}{\delta})$.

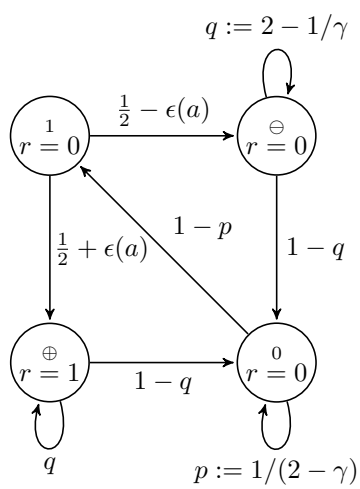

Figure 3: Counter-example

7 Conclusions

Summary. The Maximum Exploration Reinforcement Learning algorithm was presented. For finite classes of arbitrary environments a sample-complexity bound was given that is linear in the number of environments. We also presented lower bounds that show that in general this cannot be improved except for logarithmic factors. Learning is also possible for compact classes with the sample complexity depending on the size of the smallest ϵ-cover where the distance between two environments is the difference in value functions over all policies and history sequences. Finally, for non-compact classes of environments sample-complexity bounds are typically not possible.

Running time. The running time of MERL can be arbitrary large since computing the policy maximising Δ depends on the environment class used. Even assuming the distribution of observation/rewards given the history can be computed in constant time, the values of optimal policies can still only be computed in time exponential in the horizon.

Future work. MERL is close to unimprovable in the sense that there exists a class of environments where the upper bound is nearly tight. On the other hand, there are classes of environments where the bound of Theorem 1 scales badly compared to the bounds of tuned algorithms (for example, finite state MDPs). It would be interesting to show that MERL, or a variant thereof, actually performs comparably to the optimal sample-complexity even in these cases. This question is likely to be subtle since there are unrealistic classes of environments where the algorithm minimising sample-complexity should take actions leading directly to a trap where it receives low reward eternally, but is never (again) sub-optimal. Since MERL will not behave this way it will tend to have poor sample-complexity bounds in this type of environment class. This is really a failure of the sample-complexity optimality criterion rather than MERL, since jumping into non-rewarding traps is clearly sub-optimal by any realistic measure.

Acknowledgements. This work was supported by ARC grant DP120100950.

References

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 99:1563–1600, August 2010. ISSN 1532-4435.

M. Azar, R. Munos, and B. Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th international conference on machine learning*, New York, NY, USA, 2012. ACM.

D. Chakraborty and P. Stone. Structure learning in ergodic factored mdps without knowledge of the transition function's in-degree. In *Proceedings of the Twenty Eighth International Conference on Machine Learning (ICML'11)*, 2011.

C. Diuk, L. Li, and B. Leffler. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pages 249–256. ACM, 2009.

E. Even-dar, S. Kakade, and Y. Mansour. Reinforcement learning in POMDPs without resets. In *In IJCAI*, pages 690–695, 2005.

M. Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proc. 15th Annual Conf. on Computational Learning Theory (COLT'02)*, volume 2375 of *LNAI*, pages 364–379, Sydney, 2002. Springer, Berlin. URL `http://arxiv.org/abs/cs.AI/0204040`.

T. Lattimore and M. Hutter. Time consistent discounting. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011a.

T. Lattimore and M. Hutter. Asymptotically optimal agents. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011b.

T. Lattimore and M. Hutter. PAC bounds for discounted MDPs. Technical report, 2012. http://tor-lattimore.com/pubs/pac-tech.pdf.

Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *Proceedings of the Thirtieth International Conference on Machine Learning (ICML'13)*, 2013.

S. Mannor and J. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, December 2004. ISSN 1532-4435.

D. Ryabko and M. Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theoretical Computer Science*, 405(3):274–284, 2008.

A. Strehl and M. Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 856–863, 2005.

A. Strehl, L. Li, and M. Littman. Reinforcement learning in finite MDPs: PAC analysis. *J. Mach. Learn. Res.*, 10:2413–2444, December 2009.

P. Sunehag and M. Hutter. Optimistic agents are asymptotically optimal. In *Proceedings of the 25th Australasian AI conference*, 2012.

I. Szita and C. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th international conference on Machine learning*, pages 1031–1038, New York, NY, USA, 2010. ACM.

A Technical Results

Lemma 8. *Let* $x, y \in [0,1]^N$ *satisfy* $\sum_{i=1}^N y_i = 1$ *and* $\sum_{i=1}^N x_i y_i \geq 1/2$. *Then* $\max_i x_i^2 y_i > 1/(4N)$.

Proof. The result essentially follows from the fact that a maximum is greater than an average.

$$\begin{aligned}\sum_{i=1}^N x_i^2 y_i &= \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i y_i (1 - x_i) \\ &\geq \frac{1}{2} - \sum_{i=1}^N x_i y_i (1 - x_i) \geq \frac{1}{2} - \sum_{i=1}^N \frac{y_i}{4} = \frac{1}{4}\end{aligned}$$

Therefore there exists an i such that $x_i^2 y_i \geq 1/(4N)$ as required. ■

Lemma 9. *Let* $a, b > 2$ *and* $x := 4a(\log ab)^2$. *Then* $x \geq a \log bx$.

Lemma 10. *Let* $\alpha_j := \lceil \alpha^j \rceil$ *where* $\alpha := \frac{4\sqrt{N}}{4\sqrt{N}-1}$. *Then* $\sum_{j=1}^\infty \alpha_j^{-1} \leq 4\sqrt{N}$.

Proof. We have $\frac{1}{\alpha_j} \leq \left(\frac{1}{\alpha}\right)^j < 1$. Therefore by the geometric series,

$$\sum_{j=1}^\infty \frac{1}{\alpha_j} \leq \frac{1}{1 - \frac{1}{\alpha}} \equiv \frac{1}{1 - \frac{4\sqrt{N}-1}{4\sqrt{N}}} = 4\sqrt{N}$$

as required. ■

B Constants

d	$\frac{1}{1-\gamma} \log \frac{8}{(1-\gamma)\epsilon}$
ϵ_κ	$2^{\kappa-2}\epsilon$
$G_{\max}$	$\frac{2^{16} N \lvert\mathcal{K}\rvert}{\epsilon^2(1-\gamma)^2} \log^2 \frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$
$G_{\max,\kappa}$	$\frac{2^{17} N \lvert\mathcal{K}\rvert}{\epsilon\epsilon_\kappa(1-\gamma)^2} \log^2 \frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$
$E_{\max}$	$\frac{2^{16} N}{\epsilon^2(1-\gamma)^2} \log^2 \frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$
$E_{\max,\kappa}$	$\frac{2^{11} N}{\epsilon_\kappa^2(1-\gamma)^2} \log^2 \frac{2^9 N}{\epsilon^2(1-\gamma)^2\delta_1}$
α	$\frac{4\sqrt{N}}{4\sqrt{N}-1}$
δ_1	$\frac{\delta}{32\lvert\mathcal{K}\rvert N^{3/2}}$
$\lvert\mathcal{K}\rvert$	$\log_2 \frac{1}{\epsilon(1-\gamma)} + 2$

C Table of Notation

N	number of candidate models
ϵ	required accuracy
δ	probability that an algorithm makes more mistakes than its sample-complexity
t	time-step
h_t	history at time-step t
$V_\mu^\pi(h)$	value of policy π in environment μ given history h
d	effective horizon
μ	true environment
ν	an environment
$\overline{\nu}, \underline{\nu}$	models achieving upper and lower bounds on the value of the exploration policy
γ	discount factor Satisfies $\gamma \in (0, 1)$
$E_{\max,\kappa}$	high probability bound on the number of κ-exploration phases
$E_{\max}$	high probability bound on the number of exploration phases
E_∞	number of exploration phases
$E_\infty(\nu, \kappa)$	number of ν-exploration phases
$E_t(\nu, \kappa)$	number of (ν, κ)-exploration phases at time-step t
$F_t(\nu, \kappa)$	number of effective (ν, κ)-exploration phases at time-step t
$X(\nu, \kappa)_i$	ith test statistic for (ν, κ) pair