Near-optimal PAC Bounds for Discounted MDPs

Tor Lattimore[1] and **Marcus Hutter**[2]

[1]University of Alberta, Canada
tor.lattimore@gmail.com

[2] Australian National University, Australia
marcus.hutter@anu.edu.au

Abstract

We study upper and lower bounds on the sample-complexity of learning near-optimal behaviour in finite-state discounted Markov Decision Processes (MDPs). We prove a new bound for a modified version of Upper Confidence Reinforcement Learning (UCRL) with only cubic dependence on the horizon. The bound is unimprovable in all parameters except the size of the state/action space, where it depends linearly on the number of non-zero transition probabilities. The lower bound strengthens previous work by being both more general (it applies to all policies) and tighter. The upper and lower bounds match up to logarithmic factors provided the transition matrix is not too dense.

Contents

1 Introduction **1**
2 Notation **2**
3 Estimation **2**
4 Upper Confidence Reinforcement Learning Algorithm **3**
5 Upper PAC Bounds **5**
6 Lower PAC Bound **11**
7 Conclusion **16**
A Constants **18**
B Table of Notation **19**
C Technical Results **20**
D Proof of Lemma 7 **20**

Keywords

Sample-complexity; PAC bounds; Markov decision processes; Reinforcement learning.

1 Introduction

The goal of reinforcement learning is to construct algorithms that learn to act optimally, or nearly so, in unknown environments. In this paper we restrict our attention to finite state discounted MDPs with unknown transitions, but known rewards.[1] The performance of reinforcement learning algorithms in this setting can be measured in a number of ways, for instance by using regret or PAC bounds [Kak03]. We focus on the latter, which is a measure of the number of time-steps where an algorithm is not near-optimal with high probability. Many previous algorithms have been shown to be PAC with varying bounds [Kak03, SL05, SLW$^+$06, SLL09, SS10].

We construct a new algorithm, UCRLγ, based on Upper Confidence Reinforcement Learning (UCRL) [AJO10] and prove a PAC bound of

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

[1]Learning reward distributions is substantially easier than transitions, so is omitted for clarity as in [SS10].

where T is the number of non-zero transitions in the unknown MDP. Previously, the best published bound [SS10] is

$$\tilde{O}\left(\frac{|S \times A|}{\epsilon^2(1-\gamma)^6} \log \frac{1}{\delta}\right)$$

Our bound is substantially better in terms of the horizon, $1/(1-\gamma)$, but can be worse if the state-space is very large compared to the horizon and the transition matrix is dense. A bound with quartic dependence on the horizon has been shown in [Aue11], but this work is still unpublished.

We also present a matching (up to logarithmic factors) lower bound that is both larger and more general than the previous best given by [SLL09]. This article is an extended version of [LH12] with the most notable difference being the inclusion of all proofs omitted in that work. It is worth observing that sample-complexity bounds have been proven for larger classes than finite-state MDPs. The case where the true MDP has (possibly) infinitely many states, but is known to be contained in some finite set of infinite-state MDPs has been considered by a number of authors [DMS08, LHS13, LH14]. Factored and partial observable MDPs have also been studied under a variety of assumptions [CS11, EDKM05]. We expect that many of the improvements given in this paper can be translated with relative ease to the factored setting.

2 Notation

Proofs of the type found in this paper tend to use a number of complex magic constants. Readers will have an easier time if they consult the tables of constants and notation found in A and B.

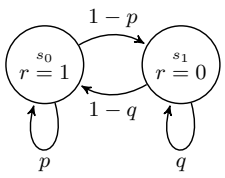
General. $\mathbb{N} = \{0, 1, 2, \cdots\}$ is the natural numbers. For the indicator function we write $\mathbb{1}\{x = y\} = 1$ if $x = y$ and 0 if $x \neq y$. We use $\wedge$ and $\vee$ for logical and/or respectively. If A is a set, then $|A|$ is its size and A^* is the set of all finite ordered subsets (possibly with repetition). Unless otherwise mentioned, $\log$ represents the natural logarithm. For random variable X we write $\mathbf{E}X$ and $\operatorname{Var} X$ for its expectation and variance respectively. We make frequent use of the progression defined recursively by $z_1 := 0$ and $z_{i+1} := \max\{1, 2z_i\}$. Define a set $\mathcal{Z}(a) := \{z_i : 1 \leq i \leq \arg\min_i \{z_i \geq a\}\}$. We write $\tilde{O}(\cdot)$ for big-O, but where logarithmic multiplicative factors are dropped.

Markov Decision Process. An MDP is a tuple $M = (S, A, p, r, \gamma)$ where S and A are finite sets of states and actions respectively, $r : S \to [0, 1]$ is the reward function, $p : S \times A \times S \to [0, 1]$ is the transition function, and $\gamma \in (0, 1)$ the discount factor. A stationary policy π is a function $\pi : S \to A$ mapping a state to an action. We write $p^{s'}_{s,a}$ as the probability of moving from state s to s' when taking action a and $p^{s'}_{s,\pi} := p^{s'}_{s,\pi(s)}$. The value of policy π in M and state s is $V^\pi_M(s) := r(s) + \gamma \sum_{s' \in S} p^{s'}_{s,\pi(s)} V^\pi_M(s')$. We view V^π_M either as a function $V^\pi_M : S \to \mathbb{R}$ or a vector $V^\pi_M \in \mathbb{R}^{|S|}$ and similarly $p_{s,a} \in [0, 1]^{|S|}$ is a vector. We often use the scalar product between vectors, for example, $p_{s,a} \cdot V^\pi_M := \sum_{s'} p^{s'}_{s,a} V^\pi_M(s')$. The optimal policy of M is defined $\pi^*_M := \arg\max_\pi V^\pi_M$. Common MDPs are M, $\widehat{M}$ and $\widetilde{M}$, which represent the true MDP, the estimated MDP using empirical transition probabilities and a model. We write $V := V_M$, $\widehat{V} := V_{\widehat{M}}$ and $\widetilde{V} := V_{\widetilde{M}}$ for their values respectively. Similarly, $\hat{\pi}^* := \pi^*_{\widehat{M}}$ and in general, variables with an MDP as a subscript will be written with a hat, tilde or nothing as appropriate and the subscript omitted.

3 Estimation

In the next section we will introduce the new algorithm, but first we give an intuitive introduction to the type of parameter estimation required to prove sample-complexity bounds for MDPs. The general idea is to use concentration inequalities to show the empiric estimate of a transition probability approaches the true probability exponentially fast in the number of samples gathered. There are many such inequalities, each catering to a slightly different purpose. We improve on previous work by using a version of Bernstein's inequality, which takes variance into account (unlike Hoeffding). The following example demonstrates the need for a variance dependent concentration inequality when estimating the value functions of MDPs. It also gives insight into the workings of the proof in the next two sections.

Consider the MDP on the right with two states and one action where rewards are shown inside the states and transition probabilities on the edges. We are only concerned with how well the value can be approximated. Assume $p > \gamma$, q arbitrarily large (but not 1) and let $\hat{p}$ be the empiric estimate of p. By writing out the definition of the value function one can show that



$$\left|V(s_0) - \widehat{V}(s_0)\right| \;\approx\; \frac{|\hat{p}-p|}{(1-\gamma)^2}. \tag{1}$$

Therefore if $V - \widehat{V}$ is to be estimated with ϵ accuracy, we need $|\hat{p}-p| < \epsilon(1-\gamma)^2$. Now suppose we bound $|\hat{p}-p|$ via a standard Hoeffding bound, then with high probability $|\hat{p}-p| \lesssim \sqrt{L/n}$ where n is the number of visits to state s_0 and $L = \log(1/\delta)$. Therefore to obtain an error less than $\epsilon(1-\gamma)^2$ we need $n > \frac{L}{\epsilon^2(1-\gamma)^4}$ visits to state s_0, which is already too many for a bound in terms of $1/(1-\gamma)^3$. If Bernstein's inequality is used instead, then $|\hat{p}-p| \lesssim \sqrt{Lp(1-p)/n}$ and so $n > \frac{Lp(1-p)}{\epsilon^2(1-\gamma)^4}$ is required, but Equation (1) depends on $p > \gamma$. Therefore $n > \frac{L}{\epsilon^2(1-\gamma)^3}$ visits are sufficient. If $p < \gamma$, then Equation (1) can be improved.

4 Upper Confidence Reinforcement Learning Algorithm

UCRL is based on the optimism principle for solving the exploration/exploitation dilemma. It is model-based in the sense that at each time-step the algorithm acts according to a model (in this case an MDP, $\widetilde{M}$) chosen from a model class. The idea is to choose the smallest model class guaranteed to contain the true model with high probability and act according to the most optimistic model within this class. With a good choice of model class this guarantees a policy that biases its exploration towards unknown states that may yield good rewards, while avoiding states that are known to be bad. The approach has been successful in obtaining uniform sample complexity (or regret) bounds in various domains where the exploration/exploitation problem is an issue [LR85, SL05, AO07, AJO10]. We modify UCRL2 of Auer and Ortner (2010) to a new algorithm, UCRLγ, given below.

We start our analysis by considering a restricted setting where for each state/action pair in the true MDP there are at most two possible next-states, which are known. We will then apply the algorithm and bound in this setting to solve the general problem.

Assumption 1. For each (s,a) pair the true unknown MDP satisfies $p^{s'}_{s,a} = 0$ for all but two $s' \in S$ denoted $\mathscr{s}a^+, \mathscr{s}a^- \in S$. Note that $\mathscr{s}a^+$ and $\mathscr{s}a^-$ *are* dependent on (s,a) and are known to the algorithm.

Extended value iteration. The function EXTENDEDVALUEITERATION is as used in [SL08]. The only difference is the definition of the confidence intervals, which are now tighter for small/large values of $\hat{p}$.

Episodes and phases. UCRLγ operates in *episodes*, which are contiguous blocks of time-steps ending when UPDATE is called. The length of each episode is not fixed, instead, an episode ends when either the number of visits to a state/action pair reaches $mw_{\min}$ for the first time or has doubled since the end of the last episode. We often refer to time-step t and episode k and unless there is ambiguity we will not define k and just assume it is the episode in which t resides. A *delay phase* is the period of $H := \frac{1}{1-\gamma}\log\frac{8|S|}{\epsilon(1-\gamma)}$ contiguous time-steps where UCRLγ is in the function DELAY, which happens immediately after an update. An *exploration phase* is a period of H time-steps starting at time t that is not in a delay phase and where $\widetilde{V}^{\pi_k}(s_t) - V^{\pi_k}(s_t) \geq \epsilon/2$. Exploration phases do not overlap with each other, but may overlap with delay phases. More formally, the starts of exploration phases, $t_1, t_2, \cdots$, are defined inductively with $t_0 := -H$.

$$t_i := \min\left\{t : t \geq t_{i-1} + H \wedge \widetilde{V}^{\pi_k}(s_t) - V^{\pi_k}(s_t) \geq \epsilon/2 \wedge t \text{ not in a delay phase}\right\}$$

Note there need not, and with high probability will not, be infinitely many such t_i. The exploration phases are only used in the analysis, they are not known to UCRLγ. We will later prove that the maximum number of updates is $U_{\max} := |S \times A| \log_2 \frac{|S|}{w_{\min}(1-\gamma)}$ and that with high probability the number of exploration phases is bounded by

Algorithm 1 UCRLγ

$t = 1, k = 1, n(s,a) = n(s,a,s') = 0$ for all s, a, s' and s_1 is the start state.
$v(s,a) = v(s,a,s') = 0$ for all s, a, s'
$H := \frac{1}{1-\gamma} \log \frac{8|S|}{\epsilon(1-\gamma)}$ and $w_{\min} := \frac{\epsilon(1-\gamma)}{4|S|}$
$\delta_1 := \frac{\delta}{2|S\times A|} \left(\log_2 |S| \log_2 \frac{1}{w_{\min}(1-\gamma)}\right)^{-1}$ and $L_1 := \log \frac{2}{\delta_1}$
$m := \frac{1280 L_1}{\epsilon^2 (1-\gamma)^2} \left(\log\log \frac{1}{1-\gamma}\right)^2 \left(\log \frac{|S|}{\epsilon(1-\gamma)}\right) \log \frac{1}{\epsilon(1-\gamma)}$
loop
  $\hat{p}_{s,a}^{sa^+} := n(s,a,sa^+) / \max\{1, n(s,a)\}$
  $\mathcal{M}_k := \left\{\widetilde{M} : |\tilde{p}_{s,a}^{sa^+} - \hat{p}_{s,a}^{sa^+}| \le \text{CONFIDENCEINTERVAL}(\tilde{p}_{s,a}^{sa^+}, n(s,a)),\ \forall (s,a)\right\}$
  $\widetilde{M} = \text{EXTENDEDVALUEITERATION}(\mathcal{M}_k)$
  $\pi_k = \tilde{\pi}^*$
  repeat
    ACT
  until $v(s_{t-1}, a_{t-1}) \ge \max\{m w_{\min}, n(s_{t-1}, a_{t-1})\}$ and $n(s_{t-1}, a_{t-1}) < \frac{|S|m}{1-\gamma}$
  UPDATE(s_{t-1}, a_{t-1}) and DELAY and $k = k + 1$
end loop
function DELAY
  for $j = 1 \to H$ **do**
    ACT
  end for
end function
function UPDATE(s,a)
  $n(s,a) = n(s,a) + v(s,a)$ and $n(s,a,s') = n(s,a,s') + v(s,a,s') \forall s'$
  $v(s,a) = v(s,a,\cdot) = 0$
end function
function ACT
  $a_t = \pi_k(s_t)$
  $s_{t+1} \sim p_{s_t,a_t}$ ▷ Sample from MDP
  $v(s_t,a_t) = v(s_t,a_t) + 1$ and $v(s_t,a_t,s_{t+1}) = v(s_t,a_t,s_{t+1}) + 1$ and $t = t + 1$
end function
function EXTENDEDVALUEITERATION$(\mathcal{M})$
  return optimistic $\widetilde{M} \in \mathcal{M}$ such that $V^*_{\widetilde{M}}(s) \ge V^*_{\widetilde{M}'}(s)$ for all $s \in S$ and $\widetilde{M}' \in \mathcal{M}$.
end function
function CONFIDENCEINTERVAL(p,n)
  return $\min\left\{\sqrt{\frac{2L_1 p(1-p)}{n}} + \frac{2L_1}{3n},\ \sqrt{\frac{L_1}{2n}}\right\}$
end function

$E_{\max} := 4m|S \times A| \log_2 |S| \log_2 \frac{1}{w_{\min}(1-\gamma)}$. We write $n_t(s,a)$ to be the value of $n(s,a)$ at time-step t. The delay phases are introduced as a trick to ease the analysis. If UCRLγ enters an exploration phase we want to ensure that the policy is fixed throughout the phase. The delay phase guarantees this because exploration phases do not start in delay phases, and because updates occur only after the delay phase. In practise we expect the delay phases are unnecessary.

5 Upper PAC Bounds

We present two new PAC bounds. The first improves on all previous analyses, but relies on Assumption 1. The second is more general and optimal in all terms except the number of states, where it depends on the number of non-zero transition probabilities, T, rather than $|S \times A|$. This can be worse than the state-of-the-art if the transition matrix is dense, but by at most a factor of $|S|$. We denote the policy followed by UCRLγ by UCRLγ, which is non-stationary. Therefore the value depends not only on the current state, but the entire history. For this reason we write $V^{\text{UCRL}\gamma}(s_{1:t})$ to indicate the discounted expected cumulative future rewards obtained from following non-stationary policy UCRLγ given the history sequence $s_{1:t} = s_1 s_2 \cdots s_t$ (the policy is deterministic, so actions need not be included).

Theorem 2. *Let M be the true* MDP *satisfying Assumption 1 and $0 < \epsilon \leq 1$ and $s_{1:t}$ the sequence of states seen up to time t. Then*

$$\mathrm{P}\left\{\sum_{t=1}^{\infty} \mathbb{1}\{V^*(s_t) - V^{\text{UCRL}\gamma}(s_{1:t}) > \epsilon\} > HU_{\max} + HE_{\max}\right\} < \delta.$$

where $V^{\text{UCRL}\gamma}(s_{1:t})$ is the expected discounted value of UCRLγ *from $s_{1:t}$.*

If lower order terms are dropped, then

$$HU_{\max} + HE_{\max} \in \tilde{O}\left(\frac{|S \times A|}{\epsilon^2(1-\gamma)^3} \log \frac{1}{\delta}\right).$$

Theorem 3. *Let T be the unknown number of non-zero transitions in the true* MDP *with $0 < \epsilon \leq 1$. Then there exists a modification of* UCRLγ *(see end of this section) such that*

$$\mathrm{P}\left\{\sum_{t=1}^{\infty} \mathbb{1}\{V^*(s_t) - V^{\text{UCRL}\gamma}(s_{1:t}) > \epsilon\} > \frac{T}{|S \times A|} H\left(U_{\max} + E_{\max}\right)\right\} < \delta.$$

If the lower order terms are dropped, then the modified PAC bound is of order

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

Before the proofs, we briefly compare Thereom 3 with the more recent work on the sample complexity of reinforcement learning when a generative model is available [AMK12]. In that paper they obtain a bound equal (up to logarithmic factors) to that of Theorem 3, but where the dependence on the number of states is linear. The online version of the problem studied in this paper is harder in two ways. Firstly, access to a generative model allows you to obtain independent samples from any state/action pair without needing to travel through the model. Secondly, and more subtly, the difference bounded in [AMK12] is $|V^*(s) - \hat{V}^*(s)|$ rather than the more usual $|V^*(s) - V^{\hat{\pi}^*}(s)|$, which is closer to what we require. The second problem was resolved with a more subtle proof in [AMK13]. It may be possible that the techniques used in that work are transferable to the online setting, but not in a very straight-forward way. It may eventually be a surprising fact that learning with the generative model is no easier than the online case considered in this paper.

Proof overview. The proof of Theorem 2 borrows components from the work of [AJO10], [SL08] and [SS10]. It also shares similarities with the proofs in [AMK12], although these were independently and simultaneously discovered.

1. Bound the number of updates by $\tilde{O}\left(|S \times A| \log \frac{1}{\epsilon(1-\gamma)}\right)$, which follows from the algorithm. Since a delay phase only occurs after an update, the number of delaying phases is also bounded by this quantity.
2. Show that the true Markov Decision Process, M, remains in the model class $\mathcal{M}_k$ for all k with high probability.
3. Use the optimism principle to show that if $M \in \mathcal{M}_k$ and $V^* - V^{\text{UCRL}\gamma} > \epsilon$, then $\widetilde{V}^{\pi_k} - V^{\pi_k} > \epsilon/2$. This key fact shows that if UCRLγ is not nearly-optimal at some time-step t, then the true value and model value of π_k differ and so some information is (probably) gained by following this policy.
4. The most complex part of the proof is then to show that the information gain is sufficiently quick to tightly bound the number of exploration phases by $E_{\max}$.
5. Note that $V^*(s_t) - V^{\text{UCRL}\gamma}(s_{1:t}) > \epsilon$ implies t is in a delay or exploration phase. Since with high probability there are at most $U_{\max} + E_{\max}$ of these phases, and both phases are exactly H time-steps long, the number of time-steps when UCRLγ is not ϵ-optimal is at most $HU_{\max} + HE_{\max}$.

Weights and variances. We define the weight[2] of state/action pair (s, a) as follows.

$$\begin{aligned} w^\pi(s,a|s') &:= \mathbb{1}\{(s', \pi(s')) = (s,a)\} + \gamma \sum_{s''} p^{s''}_{s',\pi(s')} w^\pi(s,a|s'') \\ w_t(s,a) &:= w^{\pi_k}(s,a|s_t). \end{aligned}$$

As usual, $\tilde{w}$ and $\hat{w}$ are defined as above but with p replaced by $\tilde{p}$ and $\hat{p}$ respectively. Think of $w_t(s,a)$ as the expected number of discounted visits to state/action pair (s,a) while following policy π_k starting in state s_t. The important point is that this value is approximately equal to the expected number of visits to state/action pair (s,a) within the next H time-steps. We also define the local variances of the value function. These measure the variability of values while following policy π.

$$\sigma^\pi(s,a)^2 := p_{s,a} \cdot (V^\pi)^2 - (p_{s,a} \cdot V^\pi)^2 \quad \text{and} \quad \tilde{\sigma}^\pi(s,a)^2 := \tilde{p}_{s,a} \cdot \left(\widetilde{V}^\pi\right)^2 - \left(\tilde{p}_{s,a} \cdot \widetilde{V}^\pi\right)^2$$

where $(V^\pi)^2$ is define component-wise.

Knownness. We define the knownness index of state s at time t as

$$\kappa_t(s,a) := \max\left\{ z_i : z_i \leq \frac{n_t(s,a)}{m w_t(s,a)} \right\},$$

where m is as in the preamble of the algorithm above. The idea will be that if all states are sufficiently well known then UCRLγ will be ϵ-optimal. What we will soon show is that states with low weight need not have their transitions approximated as accurately as those with high weight. Therefore fewer visits to these states are required. Conversely, states with high weight need very accurate estimates of their transition probabilities. Fortunately, these states are precisely those we expect to visit often. By carefully balancing these factors we will show that all states become known after roughly the same number of exploration phases.

The active set. State/action pairs with very small $w_t(s,a)$ cannot influence the differences in value functions. Thus we define an *active* set of states where $w_t(s,a)$ is not tiny. At each time-step t define the *active* set X_t by

$$X_t := \left\{ (s,a) : w_t(s,a) > \frac{\epsilon(1-\gamma)}{4|S|} =: w_{\min} \right\}.$$

We further partition the active set by knownness and weights.

$$\begin{aligned} \iota_t(s,a) &:= \max\left\{ z_i : z_i \leq \frac{w_t(s,a)}{w_{\min}} \right\} \\ X_{t,\kappa,\iota} &:= \{(s,a) : (s,a) \in X_t \wedge \kappa_t(s,a) = \kappa \wedge \iota_t(s,a) = \iota\} \end{aligned}$$

[2] Also called the discounted future state-action distribution in [Kak03].

An easy computation shows that the indices κ and ι are contained in $\mathcal{Z}(|S|)$ and $\mathcal{Z}(\frac{1}{(1-\gamma)w_{\min}})$ respectively. We write the joint index set,

$$\mathcal{K} \times \mathcal{I} := \mathcal{Z}(|S|) \times \mathcal{Z}(\frac{1}{(1-\gamma)w_{\min}}).$$

For $\iota \in \mathcal{I}$ we define w_ι by $w_\iota := z_\iota w_{\min}$. Therefore if $(s,a) \in X_{t,\kappa,\iota}$, then by definition we $\iota(s,a) = \iota$ and $w_\iota \leq w_t(s,a) \leq w_{\iota+1} = 2w_\iota$. Additionally,

$$\kappa \equiv \kappa_t(s,a) \leq \frac{n_t(s,a)}{m w_t(s,a)} \leq 2\kappa$$

Therefore

$$\kappa w_\iota m \leq \kappa m w_t(s,a) \leq n_t(s,a) \leq 2\kappa m w_t(s,a) \leq 4\kappa w_\iota m.$$

Analysis. The proof of Theorem 2 follows easily from three key lemmas.

Lemma 4. *The following hold:*

1. *The total number of updates is bounded by* $U_{\max} := |S \times A| \log_2 \frac{|S|}{w_{\min}(1-\gamma)}$.
2. *If* $M \in \mathcal{M}_k$ *and* t *is not in a delay phase and* $V^*(s_t) - V^{\text{UCRL}\gamma}(s_{1:t}) > \epsilon$*, then*

$$\widetilde{V}^{\pi_k}(s_t) - V^{\pi_k}(s_t) > \epsilon/2.$$

Lemma 5. $M \in \mathcal{M}_k$ *for all* k *with probability at least* $1 - \delta/2$.

Lemma 6. *The number of exploration phases is bounded by* $E_{\max}$ *with probability at least* $1 - \delta/2$.

The proofs of the lemmas are delayed while we apply them to prove Theorem 2.

Proof of Theorem 2. By Lemma 5, $M \in \mathcal{M}_k$ for all k with probability $1-\delta/2$. By Lemma 6 we have that the number of exploration phases is bounded by $E_{\max}$ with probability $1 - \delta/2$. Now if t is not in a delaying or exploration phase and $M \in \mathcal{M}_k$, then by Lemma 4, UCRLγ is nearly-optimal. Finally note that the number of updates is bounded by $U_{\max}$ and so the number of time-steps in delaying phases is at most $HU_{\max}$. Therefore UCRLγ is nearly-optimal for all but $HU_{\max} + HE_{\max}$ time-steps with probability $1 - \delta$. □

We now turn our attention to proving Lemmas 4, 5 and 6. Of these, only Lemma 6 presents a substantial challenge.

Proof of Lemma 4. For part 1 we note that no state/action pair is updated once it has been visited more than $|S|m/(1-\gamma)$ times. Since updates happen only when the visit counts would double, and only start when they are at least $mw_{\min}$, the number of updates to pair (s,a) is bounded by $\log_2 \frac{|S|}{w_{\min}(1-\gamma)}$. Therefore the total number of updates is bounded by $U_{\max} := |S \times A| \log_2 \frac{|S|}{w_{\min}(1-\gamma)}$.

The proof of part 2 is closely related to the approach taken by [SL08]. Recall that $\widetilde{M}$ is chosen optimistically by extended value iteration. This generates an MDP, $\widetilde{M}$, such that $V^*_{\widetilde{M}}(s) \geq V^*_{\widetilde{M}'}(s)$ for all $\widetilde{M}' \in \mathcal{M}_k$. Since we have assumed $M \in \mathcal{M}_k$ we have that $\widetilde{V}^{\pi_k}(s) \equiv V^*_{\widetilde{M}}(s) \geq V^*_M(s)$. Therefore $\widetilde{V}^{\pi_k}(s_t) - V^{\text{UCRL}\gamma}(s_{1:t}) > \epsilon$. By the assumption that t is a non-delaying time-step we have that the policy of UCRLγ will remain stationary and equal to π_k for at least H time-steps. Using the definition of the horizon, H, we have that $|V^{\text{UCRL}\gamma}(s_{1:t}) - V^{\pi_k}(s_t)| < \epsilon/2$. Therefore $\widetilde{V}^{\pi_k}(s_t) - V^{\pi_k}(s_t) > \epsilon/2$ as required. □

Proof of Lemma 5. In the previous lemma we showed that there are at most $U_{\max}$ updates where exactly one state/action pair is updated. Therefore we only need to check $M \in \mathcal{M}_k$ after each update. For each update let (s,a) be the updated state/action pair and apply the best of either Bernstein or Hoeffding inequalities[3] to show that

[3]The application of these inequalities is somewhat delicate since although the samples from state action pair (s,a) are independent by the Markov property, they are not independent given the number of samples from (s,a). For a detailed discussion, and a proof that using these bounds is theoretically sound, see [SL08].

$|\hat{p}^{sa^+}_{s,a} - p^{sa^+}_{s,a}| \leq \textsc{ConfidenceInterval}(p^{sa^+}_{s,a}, n(s,a)))$ with probability $1-\delta_1$. Setting $\delta_1 := \frac{\delta}{2U_{\max}}$ and applying the union bound completes the proof. □

We are now ready to work on the Lemma 6, which gives a high-probability bound on the number of exploration phases. First we will show that if t is the start of an exploration phase, then there exists a (κ, ι) such that $|X_{t,\kappa,\iota}| > \kappa$. Since $X_{t,\kappa,\iota}$ consists of active states with similar weights, we expect their visit counts to increase at approximately the same rate. More formally we show that:

1. If t is the start of an exploration phase, then there exists (κ, ι) such that $|X_{t,\kappa,\iota}| > \kappa$.
2. If $|X_{t,\kappa,\iota}| > \kappa$ for sufficiently many t, then sufficient information is gained for an update occur.
3. Combining the results above with the fact that there are at most $U_{\max}$ updates completes the result.

Lemma 7. *Let t be a non-delaying time-step and assume $M \in \mathcal{M}_k$. If $|X_{t,\kappa,\iota}| \leq \kappa$ for all (κ, ι), then $|\widetilde{V}^{\pi_k}(s_t) - V^{\pi_k}(s_t)| \leq \epsilon/2$.*

The full proof is long, technical and may be found in Appendix D. We provide a sketch, but first we need some useful results about MDPs and the differences in value functions. The first shows that less accurate transition probabilities are required for low-weight states than their high-weight counter parts. The second lemma formalises our intuitions in Section 3, and motivates the use of Bernstein's inequalities.

Lemma 8. *Let M and $\widetilde{M}$ be two Markov decision processes differing only in transition probabilities and π be a stationary policy. Then*

$$V^\pi(s_t) - \widetilde{V}^\pi(s_t) \;=\; \gamma \sum_{s,a} w^\pi(s,a|s_t)(p_{s,a} - \tilde{p}_{s,a}) \cdot \widetilde{V}^\pi. \tag{2}$$

Proof sketch. Drop the π superscript and write $V(s_t) = r(s_t) + \gamma \sum_{s_{t+1}} p^{s_{t+1}}_{s_t,\pi} V(s_{t+1})$. Then $V(s_t) - \widetilde{V}(s_t) = \gamma[p_{s_t,\pi} - \tilde{p}_{s_t,\pi}] \cdot \widetilde{V} + \gamma \sum_{s_{t+1}} p^{s_{t+1}}_{s_t,\pi} [V(s_{t+1}) - \widetilde{V}(s_{t+1})]$. The result is obtained by continuing to expand the second term of the right hand side. □

Lemma 9 (Sobel 1982)**.** *For any MDP $\widetilde{M}$, stationary policy π and state s',*

$$\sum_{s,a} \tilde{w}^\pi(s,a|s') \tilde{\sigma}^\pi(s,a)^2 \leq \frac{1}{\gamma^2(1-\gamma)^2}. \tag{3}$$

Proof sketch of Lemma 7. For ease of notation we drop references to π_k. We approximate $w_t(s,a) \approx \tilde{w}_t(s,a)$ and $|(p_{s,a} - \tilde{p}_{s,a}) \cdot \widetilde{V}| \lesssim \sqrt{\frac{L_1 \tilde{\sigma}(s,a)^2}{n_t(s,a)}}$. Using Lemma 8

$$|\widetilde{V}(s_t) - V(s_t)| \;\lesssim\; \left| \sum_{s,a \in X_t} w_t(s,a)(p_{s,a} - \tilde{p}_{s,a}) \cdot \widetilde{V} \right| \tag{4}$$

$$\lesssim \sum_{s,a \in X_t} w_t(s,a) \sqrt{\frac{L_1 \tilde{\sigma}(s,a)^2}{n_t(s,a)}} \;\lesssim \sum_{\kappa,\iota} \sum_{s,a \in X_{t,\kappa,\iota}} \sqrt{\frac{L_1 \tilde{w}_t(s,a) \tilde{\sigma}(s,a)^2}{\kappa m}} \tag{5}$$

$$\leq \sum_{\kappa,\iota} \sqrt{\frac{L_1 |X_{t,\kappa,\iota}|}{\kappa m} \sum_{s,a \in X_{t,\kappa,\iota}} \tilde{w}_t(s,a) \tilde{\sigma}_t(s,a)^2} \;\leq\; \sqrt{\frac{L_1 |\mathcal{K} \times \mathcal{I}|}{m\gamma^2(1-\gamma)^2}}, \tag{6}$$

where in Equation (4) we used Lemma 8 and the fact that states not in X_t are visited very infrequently. In Equation (5) we used the approximations for $(p - \tilde{p}) \cdot \widetilde{V}$, the definition of $X_{t,\kappa,\iota}$ and the approximation $w \approx \tilde{w}$. In Equation (6) we used the Cauchy-Schwarz inequality,[4] the fact that $\kappa \geq |X_{t,\kappa,\iota}|$ and Lemma 9. Substituting

$$m \;:=\; \frac{1280 L_1}{\epsilon^2(1-\gamma)^2} \left(\log\log \frac{1}{1-\gamma} \right)^2 \left(\log \frac{|S|}{\epsilon(1-\gamma)} \right) \log \frac{1}{\epsilon(1-\gamma)}$$

[4] $|\langle \mathbb{1}, v \rangle| \leq \|\mathbb{1}\|_2 \|v\|_2$.

completes the proof. The extra terms in m are needed to cover the errors in the approximations made here. □

The full proof requires formalising the approximations made at the start of the sketch above. The second approximation is comparatively easy and follows from the definition of the confidence intervals. Showing that $w(s,a) \approx \tilde{w}(s,a)$ requires substantial work.

We have shown in Lemma 7 that if the value of UCRLγ is not ϵ-optimal, then $|X_{t,\kappa,\iota}|$ must be greater than κ for some (κ, ι). Now we show that this cannot happen too often except with low probability. This will be sufficient to bound the number of exploration phases and therefore bound the number of times UCRLγ is not ϵ-optimal. Let t be the start of an exploration phase and define $\nu_t(s,a)$ to be the number of visits to state s within the next H time-steps. Formally,

$$\nu_t(s,a) := \sum_{i=t}^{t+H-1} \mathbb{1}\{s_i = s \wedge \pi_k(s_i) = a\}\,.$$

The following lemma captures our intuition that state/action pairs with high $w_t(s,a)$ will, in expectation, be visited more often.

Lemma 10. *Let t be the start of an exploration phase and $w_t(s,a) \geq w_{\min}$. Then $\mathbf{E}\nu_t(s,a) \geq w_t(s,a)/2$.*

Proof sketch. Use the definition of the horizon H to show that $w_t(s,a)$ is not much larger than a bounded-horizon version. Compare $\mathbf{E}\nu_t(s,a)$ and the definition of $w_t(s,a)$.

$$w_t(s,a) \equiv \mathbf{E}\sum_{i=0}^{\infty} \gamma^i \mathbb{1}\{(s_{t+i}, \pi(s_{t+i})) = (s,a)\} \leq \mathbf{E}\sum_{i=0}^{H-1} \mathbb{1}\{(s_{t+i}, \pi(s_{t+i})) = (s,a)\} + w_{\min}/2.$$

Rearranging completes the result. □

Proof of Lemma 6. Let $N := 3|S \times A|m$, where m is as in the proof of Lemma 7 or the appendix. If $\kappa, \iota \in \mathcal{K} \times \mathcal{I}$, then we call a visit to state-action pair (s,a) at time-step t (κ,ι)-*useful* if $\kappa m w_\iota \leq n_t(s,a) \leq 4\kappa m w_\iota$. Therefore the total number of (κ,ι)-useful visits is bounded by $3|S \times A|m\kappa w_\iota = N\kappa w_\iota$. By definition, if $n_t(s,a) > 4\kappa m w_\iota$, then $(s,a) \notin X_{t',\kappa,\iota}$ for all $t' \geq t$.

Bounding the number of exploration phases. Let t be the start of an exploration phase. Therefore $\widetilde{V}^{\pi_k}(s_t) - V^{\pi_k}(s_t) > \epsilon/2$ and so by Lemma 7 there exists a (κ,ι) such that $|S| \geq |X_{t,\kappa,\iota}| > \kappa$. For each (κ,ι), let $E_{\kappa,\iota}$ be the number of exploration phases where $|X_{t,\kappa,\iota}| > \kappa$. We shortly show that $\mathrm{P}\{E_{\kappa,\iota} > 4N\} < \delta_1$, which allows us to apply the union bound over all (κ,ι) pairs to show there are at most $E_{\max} := 4N|\mathcal{K} \times \mathcal{I}|$ exploration phases with probability at least $1 - \delta_1|\mathcal{K} \times \mathcal{I}| \equiv 1 - |\mathcal{K} \times \mathcal{I}|\frac{\delta}{2U_{\max}} > 1 - \delta/2$.

Bounding $\mathbf{P}\{E_{\kappa,\iota} > 4N\}$. Consider the sequence of exploration phases, $t_1, t_2, \cdots, t_{E_{\kappa,\iota}}$, such that $|X_{t_i,\kappa,\iota}| > \kappa$. We make the following observations:

1. $\{t_i\}$ is a (finite with probability 1) sequence of random variables depending on the MDP and policy.
2. The first part of this proof shows that the sequence necessarily ends after an exploration phase if the total number of (κ,ι)-useful visits is at least $Nw_\iota\kappa$. The sequence may end early for other reasons, such as states becoming unreachable or being visited while not exploring.
3. Define $\nu_i := \sum_{s,a \in X_{t_i,\kappa,\iota}} \nu_{t_i}(s,a)$, which is the number of (κ,ι)-useful visits in exploration phase t_i. By Lemma 10 we know that for $(s,a) \in X_{t_i,\kappa,\iota}$, the expected number of visits to state-action pair (s,a) is at least $w_{t_i}(s,a)/2 \geq w_\iota/2$. Therefore since $|X_{t_i,\kappa,\iota}| > \kappa$ we have

   $$\mathbf{E}[\nu_i | \nu_1 \cdots \nu_{i-1}] \geq (\kappa+1)w_\iota/2$$

   and $\mathrm{Var}[\nu_i | \nu_1 \cdots \nu_{i-1}] \leq \mathbf{E}[\nu_i | \nu_1 \cdots \nu_{i-1}]H$.[5]

[5] If $X \in [0,H]$, then $\mathrm{Var}\, X < H\mathbf{E}X$. $\nu_i \in [0,H]$.

We now wish to show the sequence has length at most $4N$ with probability at least $1 - \delta_1$. Define auxiliary sequences of length $4N$ by

$$\nu_i^+ \;:=\; \begin{cases} \nu_i & \text{if } i \le E_{\kappa,\iota} \\ w_\iota(\kappa+1)/2 & \text{otherwise} \end{cases} \qquad\qquad \bar{\nu}_i \;:=\; \frac{\nu_i^+ w_\iota(\kappa+1)}{2\mathbf{E}[\nu_i^+ | \nu_1^+ \cdots \nu_{i-1}^+]} \le v^+,$$

which are chosen such that

$$\mathbf{E}\bar{\nu}_i = \mathbf{E}[\bar{\nu}_i | \bar{\nu}_1 \cdots \bar{\nu}_{i-1}] = w_\iota(\kappa+1)/2. \tag{7}$$

Both equalities follow from the definition of $\bar{\nu}_i$, which is normalised. It is straightforward to verify that $\mathrm{P}\left\{E_{\kappa,\iota} > 4N\right\} \le \mathrm{P}\left\{\sum_{i=1}^{4N} \bar{\nu}_i \le N w_\iota(\kappa+1)\right\}$. We now use the method of bounded differences and the martingale version of Bernstein's inequality [CL06, §6] applied to $\sum \bar{\nu}_i$. Let $B_i := \mathbf{E}[\sum_{j=1}^{4N} \bar{\nu}_j | \bar{\nu}_1 \cdots \bar{\nu}_i]$, which forms a Doob martingale with $B_{4N} = \sum_{i=1}^{4N} \bar{\nu}_i$, $B_0 = 2N w_\iota(\kappa+1)$ and $|B_{i+1} - B_i| \le H$. The expression for B_0 from Equation 7. Letting $\sigma^2 := \sum_{i=1}^{4N} \mathrm{Var}[B_i | B_1 \cdots B_{i-1}] \le 2NH w_\iota(\kappa+1)$, which follows by the definitions of B, $\bar{\nu}$ and by point 3 above. Then

$$\begin{aligned} \mathrm{P}\left\{E_{\kappa,\iota} > 4N\right\} \;&\le\; \mathrm{P}\left\{\sum_{i=1}^{4N} \bar{\nu}_i \le N w_\iota(\kappa+1)\right\} \;=\; \mathrm{P}\left\{B_{4N} - B_0 \le -B_0/2\right\} \\ &\le\; 2\exp\left(-\frac{\frac{1}{4}B_0^2}{2\sigma^2 + \frac{HB_0}{3}}\right) \;=\; 2\exp\left(-\frac{N^2 w_\iota^2(\kappa+1)^2}{2\sigma^2 + \frac{2HNw_\iota(\kappa+1)}{3}}\right) \\ &\le\; 2\exp\left(-\frac{N w_\iota(\kappa+1)}{4H + \frac{2H}{3}}\right). \end{aligned}$$

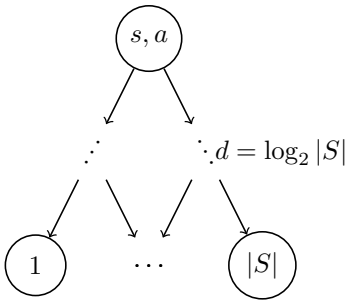
Setting this equal to δ_1, solving for N and noting that $w_\iota(\kappa+1) \ge w_{\min}$ gives

$$N \;\ge\; \frac{5H}{w_{\min}} \log \frac{2}{\delta_1} \;\in\; \tilde{O}\left(\frac{|S|}{\epsilon(1-\gamma)^2} \log \frac{1}{\delta_1}\right)$$

Since N satisfies this, the result is complete. □

Note that although in the final step we did not use a factor of $|A|$ in N, this cannot be omitted from the definition as it is used elsewhere in the proof. The result above completes the proof of Theorem 2. We now drop the assumption on the number of next-states by proving the more general Theorem 3. While we believe it is possible to do this directly, we take the simplest approach by applying the algorithm above to an augmented MDP. A more direct approach would be to use the Good-Turing estimator to estimate the missing probability, an approach recently used successfully in [DTC13].

Proof sketch of Theorem 3. The idea is to augment each state/action pair of the original MDP with $|S| - 2$ states in the form of a binary tree as pictured in the diagram below. The intention of the tree is to construct an MDP, $\overline{M}$, that with appropriate transition probabilities is functionally equivalent to the true MDP while satisfying Assumption 1. If we naively add the states as described above, then we will add an unnecessary number of addition state/action pairs because the new states need only one action. This problem is fixed by modifying the definition of an MDP to allow a varying number of actions for each state. This adds no difficulty to the proof and means the augmented MDP now has $O(|S|^2|A|)$ state-action pairs. The rewards in the added states are set to zero.



Since the tree has depth $d = \log_2 |S|$, it now takes d time-steps in the augmented MDP to change states once in the original MDP. Therefore we must also rescale γ by passing the algorithm $\bar{\gamma}$, which satisfies $\gamma > \bar{\gamma}^d = \gamma$. Then the rewards received at time-steps $d, 2d, \cdots, kd$ in the augmented MDP are discounted by

$$\bar{\gamma}^{kd} = \left(\bar{\gamma}^d\right)^k = \gamma^k,$$

Policies and values can easily be translated between the two and importantly the augmented MDP now satisfies Assumption 1. Before we apply UCRLγ to $\overline{M}$ we note that the sample-complexity bound of the algorithm depends on the inputted $\bar{\gamma}$, rather than the true γ. Since $\bar{\gamma} < \gamma$, the values $1/(1-\bar{\gamma}) \geq 1/(1-\gamma)$. Fortunately the effect is not substantial since $\frac{1}{1-\bar{\gamma}} < \frac{\log |S|}{1-\gamma}$. Therefore the scaling loses at most $\log^3 |S|$ in the final PAC bound.

Now if we simply apply UCRLγ to $\overline{M}$ and use Theorem 2 to bound the number of mistakes, then we obtain a PAC bound in the general case. Unfortunately, this leads to a bound depending on all the state/action pairs in $\overline{M}$, which total $|S|^2|A|$. To obtain dependence on the number of non-zero transitions, T, requires a little more justification. Let $T(s,a) := \sum_{s'} \mathbb{1}\left\{p^{s'}_{s,a} > 0\right\}$ be the number of non-zero transitions from state/action pair (s,a). It is easy to show the number of reachable states in the tree associated with (s,a) is at most $T(s,a) \log |S|$. Therefore the total number of reachable state/action pairs is $\log |S| \sum_{s,a} T(s,a) = T \log |S|$. Finally note that by Equation (2) from Lemma 8, state/action pairs that are not reachable do not contribute to the error and need no visits. More specifically, state/action pairs that are not reachable have $w_t(s,a) = 0$ for all t and so are never part of the active set. So in the proof of Lemma 6 we can replace $N := 3|S \times A|m$ by $N = 3Tm \log |S|$. Note that N was not used by the algorithm, so only the proof must be altered. □

6 Lower PAC Bound

We now turn our attention to the lower bound. The approach is similar to that of [SLL09]. In that work it was assumed that policies were deterministic and that the action selection policy between one time-step and the next depends only on the previously visited state. As the authors point out, neither assumption feels too restrictive in the counter-example used. Nevertheless, in this work we eliminate the second assumption and, more importantly, refine the dependence on the horizon from quadratic to cubic. We make essentially two modifications to the counter-example used in that paper. The first is to add a delaying state where no information can be gained, but where an algorithm may still fail to be PAC. The second is more subtle and will be described in the proof.

Theorem 11. *Let $\mathcal{A}$ be a (possibly non-stationary) policy depending on $S, A, r, \gamma, \epsilon$ and δ where $\epsilon(1-\gamma)$ is sufficiently small. Then there exists a Markov decision process M_{hard} such that with probability at least δ the number of time-steps where $V^*(s_t) - V^{\mathcal{A}}(s_{1:t}) > \epsilon$ is larger than*

$$\frac{c_1 |S \times A|}{\epsilon^2 (1-\gamma)^3} \log \frac{c_2 |S|}{\delta}$$

where $c_1, c_2 > 0$ are independent of the policy $\mathcal{A}$ as well as all inputs $S, A, \epsilon, \delta, \gamma$.

In light of the recent results in [AMK13], this should not come as a surprise. They show a comparable lower bound, but in the easier problem where a generative model is available. Since we study a harder problem it is unremarkable that our lower bound is at least as bad. Perhaps most surprising is that in the worst case both the online setting studied here and the setting for which a generative model is available are equally difficult.

Counter Example. We prove Theorem 11 by using a class of MDPs where $S = \{0, 1, \oplus, \ominus\}$ and $A = \{1, 2, \cdots, |A|\}$. The rewards and transitions for a single action are depicted in Figure 1 where $\epsilon(a^*) = 16\epsilon(1-\gamma)$ for some $a^* \in A$ and $\epsilon(a) = 0$ for all other actions. Some remarks:

1. States $\oplus$ and $\ominus$ are almost completely absorbing with

$$(\forall a \in A) \qquad p^{\oplus}_{\oplus,a} = p^{\ominus}_{\ominus,a} = q_2 := (1 - 2\gamma + 2\gamma^2)/\gamma.$$

   The states $\oplus$ and $\ominus$ confer maximum/minimum reward respectively.

2. The transitions are independent of actions for all states except state 1. From this state, actions lead uniformly to $\oplus$/$\ominus$ except for one action, a^*, which has a slightly higher probability of transitioning to state $\oplus$ and so a^* is the optimal action in state 1.

3. State 0 has an absorption rate $p^0_{0,a} = q_1 := 1(2-\gamma)$ such that, on average, a policy will stay there for $1/(1-\gamma)$ time-steps.

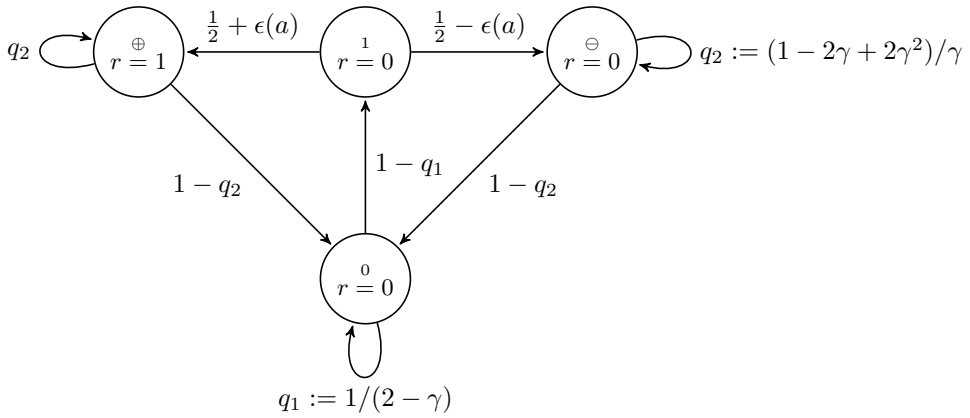


Figure 1: Hard MDP

Intuition. The MDP in Figure 1 is very bandit-like in the sense that once a policy reaches state 1 it should choose the action most likely to lead to state $\oplus$ whereupon it will either be rewarded or punished (visit state $\oplus$ or $\ominus$). Eventually it will return to state 1 when the whole process repeats. This suggests a PAC-MDP algorithm can be used to learn the bandit with $p(a) := p^{\oplus}_{1,a}$. We then make use of a theorem of Mannor and Tsitsiklis on bandit sample-complexity [MT04] to show that with high probability the number of times a^* is not selected is at least

$$\tilde{O}\left(\frac{|A|}{\epsilon^2(1-\gamma)^2}\log\frac{1}{\delta}\right). \tag{8}$$

Improving the bound to depend on $1/(1-\gamma)^3$ is intuitively easy, but technically somewhat annoying. The idea is to consider the value differences in state 0 as well as state 1. State 0 has the following properties:

1. The absorption rate is sufficiently large that any policy remains in state 0 for around $1/(1-\gamma)$ time-steps.
2. The absorption rate is sufficiently small that the difference in values due to bad actions planned in state 1 still matter while in state 0.

While in state 0 an agent cannot make an error in the sense that $V^*(0) - Q^*(0,a) = 0$ for all a. But we are measuring $V^*(0) - V^{\mathcal{A}}(0)$ and so an agent can be penalised if its policy upon reaching state 1 is to make an error. Suppose the agent is in state 0 at some time-step before moving to state 1 and making a mistake. On average it will stay in state 0 for roughly $1/(1-\gamma)$ time-steps during which time it will *plan* a mistake upon reaching state 1. Thus the bound in Equation (8) can be multiplied by $1/(1-\gamma)$. The proof is harder because an agent need not plan to make a mistake in all future time-steps when reaching state 1 before eventually doing so in one time-step. Dependence on $|S|$ can be added easily by chaining together $|S|/4$ copies of the counter-example MDP with arbitrarily low transitions between them. Note that [SLL09] proved their theorem for a specific class of policies while Theorem 11 holds for all policies.

The proof makes use of a simple form of bandit and Theorem 12 below, which lower bounds the sample-complexity of bandit algorithms. We need some new notation required for non-stationary policies and bandits.

History Sequences. We write $s_{1:t} = s_1, s_2, \cdots, s_t$ for the history sequence of length t. Histories can be concatenated, so $s_{1:t}\oplus = s_1, s_2, \cdots, s_t, \oplus$ where $\oplus \in S$.

Bandits. An *A-armed stationary bandit* is a vector $p : A \to [0,1]$. A bandit policy is a function $\pi : \{0,1\}^* \to A$. A policy interacts with a bandit sequentially with $a_t = \pi(r_{1:t-1})$ and r_t sampled from a Bernoulli distribution with parameter $p(a_t)$. The optimal arm is $a^* := \arg\max_a p(a)$. A policy dependent on ϵ, δ and A has sample-complexity $T := T(A, \epsilon, \delta)$ if for all bandits the arm chosen on time-step T satisfies $p(a^*) - p(a_T) \leq \epsilon$ with probability at least $1 - \delta$.

Theorem 12 (Mannor and Tsitsiklis, 2004)**.** *There exist positive constants c_1, c_2, ϵ_0, and δ_0, such that for every $A \geq 2$, $\epsilon \in (0, \epsilon_0)$ and $\delta \in (0, \delta_0)$ there exists a bandit $p \in [0,1]^A$ such that*

$$T(A, \epsilon, \delta) \;\geq\; c_1 \frac{|A|}{\epsilon^2} \log \frac{c_2}{\delta}$$

with probability at least δ.

Remark 13. The bandit used in the proof of Theorem 12 satisfies $p(a) = \frac{1}{2}$ for all a except a^* which has $p(a^*) := \frac{1}{2} + \epsilon$.

We now prepare to prove Theorem 11. For the remainder of this section let $\mathcal{A}$ be an arbitrary policy and M_{hard} be the MDP of Figure 2. As in previous work we write $V^{\mathcal{A}} := V^{\mathcal{A}}_{M_{\text{hard}}}$. The idea of the proof will be to use Theorem 12 to show that $\mathcal{A}$ cannot be approximately correct in state 1 too often. Then use this to show that while in state 0 before-hand it is also not approximately correct.

Definition 14. Let $s_{1:\infty} \in S^\infty$ be the sequence of states seen by policy $\mathcal{A}$ and for arbitrary history $s_{1:t}$ let

$$\Delta(s_{1:t}) := V^*(s_{1:t}) - V^{\mathcal{A}}(s_{1:t}).$$

Lemma 15. *If $\gamma \in (0,1)$, $q_1 := 1/(2-\gamma)$ and $q_2 := (1 - 2\gamma + 2\gamma^2)/\gamma$, then*

$$q_1^{\frac{1}{4(1-\gamma)}} \;>\; 3/4 \quad \textit{and} \quad \sum_{t=0}^{\infty} q^t(1-q)\gamma^t \;=\; \frac{1}{2} \quad \textit{and} \quad \sum_{n=0}^{\infty} q_2^n(1-q_2)\sum_{t=0}^{n}\gamma^t \geq \frac{1}{2\gamma(1-\gamma)}.$$

Note that $0 < q_2 < 1$ is true under the assumption that $1 > \gamma > 1/2$.

Proof sketch. All results follow from the geometric series and easy calculus. □

Lemma 16. *Let $s_{1:t}$ be such that $s_t = 1$ and $a := \mathcal{A}(s_{1:t}) \neq a^*$. Then*

$$\Delta(s_{1:t}) \geq 8\epsilon.$$

Proof. The result essentially follows from the definition of the value function.

$$\begin{aligned}
\Delta(s_{1:t}) &\equiv V^*(s_{1:t}) - V^{\mathcal{A}}(s_{1:t}) \\
&= \gamma\left[p^{\oplus}_{1,a^*}V^*(s_{1:t}\oplus) + p^{\ominus}_{1,a^*}V^*(s_{1:t}\ominus)\right] - \gamma\left[p^{\oplus}_{1,a}V^{\mathcal{A}}(s_{1:t}\oplus) + p^{\ominus}_{1,a}V^{\mathcal{A}}(s_{1:t}\ominus)\right] \\
&= \frac{\gamma}{2}\left[V^*(s_{1:t}\oplus) - V^{\mathcal{A}}(s_{1:t}\oplus) + V^*(s_{1:t}\ominus) - V^{\mathcal{A}}(s_{1:t}\ominus)\right] + \gamma\epsilon(a^*)\left[V^*(s_{1:t}\oplus) - V^*(s_{1:t}\ominus)\right] \\
&\geq 16\epsilon\gamma(1-\gamma)\left[V^*(s_{1:t}\oplus) - V^*(s_{1:t}\ominus)\right] \\
&= 16\epsilon\gamma(1-\gamma)\sum_{n=0}^{\infty} q_2^n(1-q_2)\sum_{t=0}^{n}\gamma^t \\
&\geq 8\epsilon,
\end{aligned}$$

where we used the definition of the value function and MDP, M_{hard}. The final step follows from Lemma 15. □

We now define the blocks of time where the policy is looping in state 0. Recall we chose the absorption in this state such that the expected number of time-steps a policy remains there is approximately $1/(1-\gamma)$. We define the intervals starting when a policy arrives in state 0 and ending when it leaves to state 1.

Definition 17. Define $t_1^0 := 1$ and

$$t_i^0 := \min\left\{t : t > t_{i-1}^0 \wedge s_t = 0 \wedge s_{t-1} \neq 0\right\} \qquad t_i^1 := \min\left\{t-1 : s_t = 1 \wedge t > t_i^0\right\}.$$

Define the intervals $I_i := [t_i^0, t_i^1] \subseteq \mathbb{N}$. We call interval I_i the ith *phase*.

Note the following facts:

1. Since all transition probabilities are non-zero, t_i^0 and t_i^1 exist for all $i \in \mathbb{N}$ with probability 1.
2. $|I_i|$ forms a sequence of i.i.d random variables with distribution independent of $\mathcal{A}$.

Definition 18. Suppose $t \in \mathbb{N}$ and $s_t = 0$ and define $\psi_t(a)$ to be the discounted probability that action a is chosen by algorithm $\mathcal{A}$ when state 1 is finally reached.

$$\psi_t(a) := \sum_{k=0}^{\infty} q_1^k(1-q_1)\gamma^k \mathbb{1}\left\{\mathcal{A}(s_{1:t}0^k1) = a\right\}.$$

Lemma 19. $\sum_{a\in A}\psi_t(a) = \frac{1}{2}$ *for all* t *where* $s_t = 0$.

Proof. We use Lemma 15.

$$\begin{aligned}
\sum_{a\in A}\psi_i(a) &\equiv \sum_{a\in A}\sum_{k=0}^{\infty} q_1^k(1-q_1)\gamma^k\mathbb{1}\left\{\mathcal{A}(s_{1:t}0^k1) = a\right\} \\
&= \sum_{k=0}^{\infty} q_1^k(1-q_1)\gamma^k = \frac{1}{2}
\end{aligned}$$

as required. □

Definition 20. Define random variables Y_i by $Y_i := \mathbb{1}\{|I_i| \geq 1/[4(1-\gamma)]\}$.

Intuitively, Y_i is the event that the ith phase lasts at least $1/[4(1-\gamma)]$ time-steps. The following lemma shows that at least two thirds of all phases have $Y_i = 1$ with high probability.

Lemma 21. *For all* $n \in \mathbb{N}$, $\mathrm{P}\left\{\sum_{i=1}^{n} Y_i \leq \frac{2}{3}n\right\} \leq 2e^{-n/72}$.

Proof. Preparing to use Hoeffding's bound,

$$\mathrm{P}\left\{Y_i = 1\right\} := \mathrm{P}\left\{|I_i| \geq 1/[4(1-\gamma)]\right\} = q_1^{1/[4(1-\gamma)]} > 3/4,$$

where we used the definitions of Y_i, I_i and Lemma 15. Therefore $\mathbf{E}Y_i > 3/4$.

$$\mathrm{P}\left\{\sum_{i=1}^{n} Y_i \leq \frac{2}{3}n\right\} \leq \mathrm{P}\left\{\sum_{i=1}^{n} Y_i \leq \frac{1}{12}n + n\mathbf{E}Y_i\right\} = \mathrm{P}\left\{\sum_{i=1}^{n} Y_i - \mathbf{E}Y_i \leq \frac{1}{12}n\right\} \leq 2e^{-n/72}$$

where we applied basic inequalities followed by Hoeffding's bound. □

Lemma 22. *If $\gamma > \frac{3}{4}$ and $\sum_{a\neq a^*} \psi_t(a) \geq \frac{1}{4}$, then $\sum_{a\neq a^*} \psi_{t+k}(a) \geq \frac{1}{8}$ for all $t \in \mathbb{N}$ and k satisfying $0 \leq k \leq 1/[16(1-\gamma)]$.*

Proof. Working from the definitions.

$$\begin{aligned}
\frac{1}{4} \leq \sum_{a\neq a^*} \psi_{t_i^0}(a) &\equiv \sum_{j=0}^{\infty} q_1^j(1-q_1)\gamma^j \mathbb{1}\Big\{\mathcal{A}(s_{1:t_i^0}0^j1) \neq a^*\Big\} \\
&= \sum_{j=0}^{k-1} q_1^j(1-q_1)\gamma^j \mathbb{1}\Big\{\mathcal{A}(s_{1:t_i^0}0^j1) \neq a^*\Big\} + q_1^k\gamma^k \sum_{a\neq a^*} \psi_a(s_{1:t_i^0}0^k) \\
&\leq (1-q_1)\sum_{j=0}^{k-1} q_1^j\gamma^j + q_1^k\gamma^k \sum_{a\neq a^*} \psi_a(s_{1:t_i^0}0^k)
\end{aligned}$$

Rearranging, setting $0 \leq k \leq 1/[16(1-\gamma)]$ and using the geometric series completes the proof. □

So far, none of our results have been especially surprising. Lemma 21 shows that at least two thirds of all phases have length exceeding $1/[4(1-\gamma)]$ with high probability. Lemma 22 shows that if at the start of a phase $\mathcal{A}$ assigns a high weight to the sub-optimal actions, then it does so throughout the initial part of the phase. The following lemma is more fundamental. It shows that the number of phases where $\mathcal{A}$ assigns a high weight to the sub-optimal actions is of order $\frac{1}{\epsilon^2(1-\gamma)^2}\log\frac{1}{\delta}$ with high probability.

Lemma 23. *Let $F := \frac{c_1 A}{\epsilon^2(1-\gamma)^2}\log\frac{c_2}{\delta}$ with constants as in Theorem 12, then there exists some $a^* \in A$ such that*

$$\left|\left\{i : \sum_{a\neq a^*} \psi_{t_i^0}(a) > \frac{1}{4} \wedge i < 2F+1\right\}\right| > F$$

with probability at least δ.

The idea is similar to that in [SLL09]. Assume a policy exists that doesn't satisfy the condition above and then use it to learn the bandit defined by $p(a) := p^{\oplus}_{1,a}$.

Proof. Define $p \in [0,1]^{|A|}$ by $p(a) := p^{\oplus}_{1,a}$, which characterises a bandit. Now we use the MDP algorithm $\mathcal{A}$ to learn bandit p. Since the only unknown quantity in the MDP described in Figure 1 we simply simulate the outcomes in states state 0, $\oplus$ and $\ominus$ and sample from the bandit when in state 1. After $2F+1$ visits to state 1 we return the action

$$a_{\mathrm{best}} := \arg\max_a \sum_{i=1}^{2F+1} \mathbb{1}\{\bar{a}_i = a\}, \quad \bar{a}_i := \arg\max_{a'} w_{t_i^0}(a').$$

By Theorem 12, the strategy in Algorithm 2 must fail with probability at least δ. Therefore with probability at least δ, $a_{\mathrm{best}} \neq a^*$. However a_{best} is defined as the majority action of all the $\bar{a}_i$ and so for at least F time-steps $\bar{a}_i \neq a^*$.

Suppose $\psi_{t_i^0}(a^*) > \frac{1}{4}$, then by Lemma 19, $\sum_{a \neq a^*} \psi_{t_i^0}(a) < \frac{1}{4}$ and $\bar{a}_i \equiv \arg\max_a \psi_{t_i^0}(a) = a^*$. This implies that with probability δ, for at least F time-steps $\sum_{a \neq a^*} \psi_{t_i^0}(a) > \frac{1}{4}$ as required. □

Proof of Theorem 11. Define Z_i to be the event that the ith phase lasts at least $1/[16(1-\gamma)]$ time-steps and the combined weight of sub-optimal actions at the start of a phase is at least $1/4$.

$$Z_i \;:=\; \mathbb{1}\left\{ |I_i| \geq 1/[16(1-\gamma)] \wedge \sum_{a \neq a^*} \psi_{t_i^0}(a) \geq 1/4 \right\}.$$

Suppose $Z_i = 1$ and $0 \leq k \leq 1/[16(1-\gamma)]$, then $s_{1:t_i^0+k} = s_{1:t_i^0} 0^k$ and

$$\begin{aligned}
\Delta(s_{1:t_i+k}) &= \sum_{t=0}^{\infty} q_1^t (1-q_1) \gamma^t \Delta(s_{1:t_i+k} 0^t 1) && (9)\\
&\geq \sum_{t=0}^{\infty} q_1^t (1-q_1) \gamma^t \sum_{a \neq a^*} \mathbb{1}\Big\{ \mathcal{A}(s_{1:t_i^0+k} 0^t 1) = a \Big\}\, 8\epsilon && (10)\\
&\geq \sum_{a \neq a^*} \psi_{t_i^0+k}(a) 8\epsilon && (11)\\
&\geq \epsilon, && (12)
\end{aligned}$$

where Equation (9) follows from the definition of M_{hard} and the value function. Equation (10) by Lemma 16. Equation (11) by the definition of $\psi_{t_i+k}(a)$ and Equation (12) by Lemma 22. Thus for each i where $Z_i = 1$, policy $\mathcal{A}$ makes at least $1/[16(1-\gamma)]$ ϵ-errors. We now show that $Z_i = 1$ for at least $F/3$ time-steps with probability at least 2δ, By Lemma 23 the number of phases within the first $2F+1$ in which $\psi_{t_i^0}(a) \geq 1/4$ is at least F with probability at least δ. By Lemma 21, the probability that less than two thirds of the intervals satisfy $|I_i| \geq 1/(16(1-\gamma))$ is at most $2e^{-2F/72} \leq \delta/2$ where we used the fact that $\epsilon(1-\gamma)$ are sufficiently small to guarantee that $2F \geq 72 \log \frac{2}{\delta}$. Therefore by the union bound the number of phases in which Z_i occurs must be at least $F/3$ with probability at least $\delta/2$.

Dependence on $|S| \log |S|$ is added as by chaining $|S|/4$ copies of the MDP depicted in Figure 1 together. Let $S = \big\{0_1, 1_1, \oplus_1, \ominus_1, \cdots, 0_{|S|/4}, 1_{|S|/4}, \oplus_{|S|/4}, \ominus_{|S|/4}\big\}$. Where each set of four states $\{0_k, 1_k, \oplus_k, \ominus_k\}$ is described by Figure 1 except that we modify $q_1 := (1 - 2\gamma + 2\gamma^2)/\gamma - q_3$ and $p_{0_k,a}^{0_{k+1}} = q_3$ for some arbitrarily small q_3. Then apply the analysis in [SLL09] to obtain a bound with on the sample-complexity of

$$\Omega\left(\frac{|S \times A|}{\epsilon^2 (1-\gamma)^3} \log \frac{|S|}{\delta} \right)$$

as required. □

Remark 24. Dependence on $S \log S$ can possibly be added by a similar technique used by [SLL09], but details could be messy.

7 Conclusion

Summary. We presented matching upper and lower bounds on the number of time-steps when a reinforcement learning algorithm can be nearly-optimal with high probability. We now compare the bound proven in Theorem 2 with the current state-of-the-art, MORMAX [SS10].

$$\underbrace{\tilde{O}\left(\frac{T}{\epsilon^2 (1-\gamma)^3} \log \frac{1}{\delta} \right)}_{\text{UCRL}\gamma} \qquad\qquad \underbrace{\tilde{O}\left(\frac{|S \times A|}{\epsilon^2 (1-\gamma)^6} \log \frac{1}{\delta} \right)}_{\text{MORMAX}}$$

The dependence on ϵ and δ match the lower bound for both algorithms. UCRLγ is optimal in terms of the horizon where MORMAX loses by three factors. On the other hand, MORMAX has a bound that is linear in the state space where UCRLγ can depend quadratically. Nevertheless, UCRLγ will be prefered unless the state/action space is both dense and extremely large relative to the effective horizon. Importantly, the new upper and lower bounds now match up to logarithmic factors if the MDP has at most $|S \times A| \log |S \times A|$ non-zero transitions, so at least for this class UCRLγ is now unimprovable. Additionally, UCRLγ combined with Theorem 2 is the first demonstration of a PAC reinforcement learning algorithm with cubic dependence on the effective horizon.

Running time. We did not analyze the running time of UCRLγ, but expect analysis similar to that of [SL08] can be used to show that UCRLγ can be approximated to run in polynomial time with no cost to sample-complexity.

References

References

[AJO10] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, August 2010.

[AMK12] Mohammad Azar, Rémi 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.

[AMK13] M. Azar, R. Munos, and H. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

[AO07] P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems 19*, pages 49–56. MIT Press, 2007.

[Aue11] P. Auer. Upper confidence reinforcement learning. *Unpublished, keynote at European Workshop of Reinforcement Learning*, 2011.

[CL06] F. Chung and L. Lu. Concentration inequalities and martingale inequalities a survey. *Internet Mathematics*, 3:1, 2006.

[CS11] Doran Chakraborty and Peter 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*, 2011.

[DMS08] Kirill Dyagilev, Shie Mannor, and Nahum Shimkin. Efficient reinforcement learning in parameterized models: Discrete parameter case. In *Recent Advances in Reinforcement Learning*, pages 41–54. Springer, 2008.

[DTC13] Thomas Dietterich, Majid Alkaee Taleghan, and Mark Crowley. PAC optimal planning for invasive species management: Improved exploration for reinforcement learning from simulator-defined MDPs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[EDKM05] Eyal Even-Dar, Sham Kakade, and Yishay Mansour. Reinforcement learning in POMDPs without resets. In *International Joint Conference on Artificial Intelligence*, pages 690–695, 2005.

[Kak03] Sham Kakade. *On The Sample Complexity Of Reinforcement Learning*. PhD thesis, University College London, 2003.

[LH12] Tor Lattimore and Marcus Hutter. PAC bounds for discounted MDPs. In *Proceedings of the 23rd international conference on Algorithmic Learning Theory*, ALT'12, pages 320–334, Berlin, Heidelberg, 2012. Springer-Verlag.

[LH14] Tor Lattimore and Marcus Hutter. Bayesian reinforcement learning with exploration. In *Proceedings of the 25rd international conference on Algorithmic Learning Theory*, ALT'14. Springer-Verlag, Berlin, Heidelberg, 2014.

[LHS13] Tor Lattimore, Marcus Hutter, and Peter Sunehag. The sample-complexity of general reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[LR85] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[MT04] Shie Mannor and John Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, December 2004.

[SL05] Alexander Strehl and Michael Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 856–863, New York, NY, USA, 2005. ACM.

[SL08] Alexander Strehl and Michael Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[SLL09] Alexander Strehl, Lihong Li, and Michael Littman. Reinforcement learning in finite MDPs: PAC analysis. *J. Mach. Learn. Res.*, 10:2413–2444, December 2009.

[SLW+06] Alexander Strehl, Lihong Li, Eric Wiewiorac, John Langford, and Michael Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 881–888, New York, NY, USA, 2006. ACM.

[Sob82] M. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

[SS10] Istvan Szita and Csaba 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 Constants

$\lvert\mathcal{K}\times\mathcal{I}\rvert := \log_2\lvert S\rvert \log_2 \frac{1}{w_{\min}(1-\gamma)}$	$\tilde{O}\left(\log\lvert S\rvert \log \frac{1}{\epsilon(1-\gamma)}\right)$
$H := \frac{1}{1-\gamma}\log\frac{8\lvert S\rvert}{\epsilon(1-\gamma)}$	$\tilde{O}\left(\frac{1}{1-\gamma}\log\frac{\lvert S\rvert}{\epsilon}\right)$
$w_{\min} := \frac{\epsilon(1-\gamma)}{4\lvert S\rvert}$	$\tilde{\Omega}\left(\frac{\epsilon(1-\gamma)}{\lvert S\rvert}\right)$
$\delta_1 := \frac{\delta}{2U_{\max}}$	$\tilde{\Omega}\left(\frac{\delta}{\lvert S\times A\rvert\log\frac{1}{\epsilon(1-\gamma)}}\right)$
$L_1 := \log\frac{2}{\delta_1}$	$\tilde{O}\left(\log\frac{\lvert S\times A\rvert}{\delta\epsilon(1-\gamma)}\right)$
$m := \frac{1280L_1}{\epsilon^2(1-\gamma)^2}\left(\log\log\frac{1}{1-\gamma}\right)^2\left(\log\frac{\lvert S\rvert}{\epsilon(1-\gamma)}\right)\log\frac{1}{\epsilon(1-\gamma)}$	$\tilde{O}\left(\frac{1}{\epsilon^2(1-\gamma)^2}\log\frac{\lvert S\times A\rvert}{\delta}\right)$
$N := 3\lvert S\times A\rvert m$	$\tilde{O}\left(\frac{\lvert S\times A\rvert}{\epsilon^2(1-\gamma)^2}\log\frac{1}{\delta}\right)$
$E_{\max} := 4N\lvert\mathcal{K}\times\mathcal{I}\rvert$	$\tilde{O}\left(\frac{\lvert S\times A\rvert}{\epsilon^2(1-\gamma)^2}\log\frac{1}{\delta}\right)$
$U_{\max} := \lvert S\times A\rvert\log_2\frac{\lvert S\rvert}{w_{\min}(1-\gamma)}$	$\tilde{O}\left(\lvert S\times A\rvert\log\frac{1}{\epsilon(1-\gamma)}\right)$

B Table of Notation

S, A	Finite sets of states and actions respectively.
γ	The discount factor. Satisfies $\gamma \in (0, 1)$.
ϵ	The required accuracy.
δ	The probability that an algorithm makes more mistakes than its sample-complexity.
$\mathbb{N}$	The natural numbers, starting at 0.
$\log$	The natural logarithm.
$\wedge, \vee$	Logical and/or respectively.
$\mathbf{E}X, \operatorname{Var} X$	The expectation and variance of random variable X respectively.
z_i	$z_1 = 0$ and $z_i = \max\{1, 2z_{i-1}\}$.
$\mathcal{Z}(a)$	Defined as a set of all z_i up to and including a. Formally $\mathcal{Z}(a) := \{z_i : i \leq \arg\min_i \{z_i \geq a\}\}$. Contains approximately $\log a$ elements.
π	A policy.
p	The transition function, $p : S \times A \times S \to [0, 1]$. We also write $p^{s'}_{s,a} := p(s, a, s')$ for the probability of transitioning to state s' from state s when taking action a. $p^{s'}_{s,\pi(s)} := p^{s'}_{s,\pi(s)}$. $p_{s,a} \in [0, 1]^{\lvert S \rvert}$ is the vector of transition probabilities.
$\hat{p}, \tilde{p}$	Other transition probabilities, as above.
r	The reward function $r : S \to A$.
M	The true MDP. $M := (S, A, p, r, \gamma)$.
$\widehat{M}$	The MDP with empirically estimated transition probabilities. $\widehat{M} := (S, A, \hat{p}, r, \gamma)$.
$\widetilde{M}$	An MDP in the model class, $\mathcal{M}$. $\widetilde{M} := (S, A, \tilde{p}, r, \gamma)$.
V^π_M	The value function for policy π in MDP M. Can either be viewed as a function $V^\pi_M : S \to \mathbb{R}$ or vector $V^\pi_M \in \mathbb{R}^{\lvert S \rvert}$.
$\widetilde{V}^\pi, \widehat{V}^\pi$	The values of policy π in MDPs $\widetilde{M}$ and $\widehat{M}$ respectively.
$\pi^* \equiv \pi^*_M$	The optimal policy in MDP M.
$\tilde{\pi}^* \equiv \pi^*_{\widetilde{M}}$	The optimal policy in $\widetilde{M}$.
$\hat{\pi}^* \equiv \pi^*_{\widehat{M}}$	The optimal policy in $\widehat{M}$.
π_k	The (stationary) policy at used in episode k.
$n_t(s, a)$	The number of visits to state/action pair (s, a) at time-step t.
$n_t(s, a, s')$	The number of visits to state s' from state s when taking action a at time-step.
$v_{t_k}(s, a)$	If t_k is the start of an exploration phase, then this is the total number of visits to state (s, a) in that exploration phase.
s_t, a_t	The state and action in time-step t respectively.
V^π_d	A higher "moment" value function. See Definition 26.

$\sigma_d^\pi(s,a)^2$	The variance of $V_d^\pi(s')$ when taking action a in state s. Defined in Definition 26.
L_1	Defined as $\log(2/\delta_1)$.
$\mathcal{D}$	Defined as $\mathcal{Z}(\beta)$.
$w_t(s,a)$	The expected discounted number of visits to state/action pair (s,a) while following policy π_k from state s_t.
X_t	The set of state/action pairs s,a where $w_t(s,a) \geq w_{\min}$.
$X_{t,\kappa,\iota}$	The set of state/action pairs where $\kappa_t(s,a) = \kappa$ and $\iota_t(s,a) = \iota$. Note that $\bigcup_{\kappa,\iota} K_t(\kappa,\iota)$ contains all states with $w(s,a) \geq w_{\min}$.

C Technical Results

We can use Hoeffding and Bernstein inequalities to bound $|p-\hat{p}|$ and $|\hat{p}-\tilde{p}|$. We now want to combine these to bound $|p-\tilde{p}|$. The following lemma is easily proven with simple algebra reminiscent of the derivation of the empiric Bernstein inequality.

Lemma 25. *Let $p,\hat{p},\tilde{p} \in [0,1]$ satisfy*

$$|p-\hat{p}| \leq \min\{CI_1(p), CI_2(p)\} \qquad |\tilde{p}-\hat{p}| \leq \min\{CI_1(\tilde{p}), CI_2(\tilde{p})\}$$

where

$$CI_1(p) := \sqrt{\frac{2p(1-p)}{n}\log\frac{2}{\delta}} + \frac{2}{3n}\log\frac{2}{\delta} \qquad CI_2(p) := \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.$$

Then

$$|p-\tilde{p}| \;\leq\; \sqrt{\frac{8\tilde{p}(1-\tilde{p})}{n}\log\frac{2}{\delta}} + 2\left(\frac{1}{n}\log\frac{2}{\delta}\right)^{\frac{3}{4}} + \frac{4}{3n}\log\frac{2}{\delta}$$

D Proof of Lemma 7

We need to define some higher "moments" of the value function. This is somewhat unfortunate as it complicates the proof, but may be unavoidable.

Definition 26. We define the space of bounded value/reward functions $\mathcal{R}$ by

$$\mathcal{R}(i) \;:=\; \left[0, \left(\frac{1}{1-\gamma}\right)^i\right]^{|S|} \subset \mathbb{R}^{|S|}.$$

Let π be some stationary policy. For $r_d \in \mathcal{R}(d)$ define values V_d^π by the Bellman equations

$$V_d^\pi(s) \;=\; r_d(s) + \gamma \sum_{s'} p_{s,\pi(s)}^{s'} V_d^\pi(s').$$

Additionally,

$$\sigma_d^\pi(s,a)^2 \;:=\; p_{s,a}\cdot(V_d^\pi)^2 - (p_{s,a}\cdot V_d^\pi)^2.$$

Note that $V_d^\pi \in \mathcal{R}(d+1)$ and $\sigma_d^2 \in \mathcal{R}(2d+2)$. Let $r_0 \in \mathcal{R}(0)$ be the true reward function $r_0(s) := r(s)$ and define a *recurrence* by $r_{2d+2}(s) := \sigma_d^\pi(s,\pi(s))^2$. We define $\tilde{r}_d$, $\hat{r}_d$, $\widetilde{V}_d^\pi$, $\widehat{V}_d^\pi$ and $\tilde{\sigma}_d^\pi$, $\hat{\sigma}_d^\pi$ similarly but where all parameters have hat/tilde.

Lemma 27. *Let $M \in \mathcal{M}_k$ at time-step t and $\pi = \pi_k$. Then*

$$|(p_{s,\pi} - \tilde{p}_{s,\pi}) \cdot \widetilde{V}_d^{\pi}| \;\leq\; \sqrt{\frac{8L_1\tilde{\sigma}_d^{\pi}(s,\pi(s))^2}{n_t(s,\pi(s))}} + 2\left(\frac{L_1}{n_t(s,\pi(s))}\right)^{\frac{3}{4}} \frac{1}{(1-\gamma)^{d+1}} + \frac{4L_1}{3n_t(s,\pi(s))(1-\gamma)^{d+1}} \tag{13}$$

Proof. Drop references to π and let $p := p_{s,\pi(s)}^{sa^+}$, $\tilde{p} := \tilde{p}_{s,\pi(s)}^{sa^+}$ and $n := n_t(s,\pi(s))$. Since $M, \widetilde{M} \in \mathcal{M}_k$ then apply Lemma 25 to obtain

$$|p - \tilde{p}| \leq \sqrt{\frac{8L_1\tilde{p}(1-\tilde{p})}{n}} + 2\left(\frac{L_1}{n}\right)^{\frac{3}{4}} + \frac{4L_1}{3n}$$

Assume without loss of generality that $\widetilde{V}_d(sa^+) \geq \widetilde{V}_d(sa^-)$. Therefore we have

$$\begin{aligned}|(p_{s,\pi(s)} - \tilde{p}_{s,\pi(s)}) \cdot \widetilde{V}_d| \leq \sqrt{\frac{8L_1\tilde{p}(1-\tilde{p})}{n}}\left(\widetilde{V}_d(sa^+) - \widetilde{V}_d(sa^-)\right) + 2\left(\frac{L_1}{n}\right)^{\frac{3}{4}} \frac{1}{(1-\gamma)^{d+1}} \\ + \;\; \frac{4L_1}{3n(1-\gamma)^{d+1}},\end{aligned} \tag{14}$$

where we used Assumption 1 and the fact that $V_d \in \mathcal{R}_{d+1}$.

$$\begin{aligned}\tilde{p}(1-\tilde{p})\left(\widetilde{V}_d(sa^+) - \widetilde{V}_d(sa^-)\right)^2 &= \tilde{p}(1-\tilde{p})\left(\widetilde{V}_d(sa^+)^2 + \widetilde{V}_d(sa^-)^2 - 2\widetilde{V}_d(sa^+)\widetilde{V}_d(sa^-)\right) \\ &= \tilde{\sigma}_d(s,\pi(s))^2.\end{aligned}$$

Substituting into Equation (14) completes the proof. □

Proof of Lemma 7. For ease of notation we drop π and t super/subscripts. Let

$$\Delta_d := \left|\sum_{s,a}[w(s,a) - \tilde{w}(s,a)]r_d(s)\right| \equiv |\widetilde{V}_d(s_t) - V_d(s_t)|.$$

Using Lemmas 8 and 27:

$$\begin{aligned}\Delta_d \;&=\; \gamma\left|\sum_{s,a} w(s,a)(p_{s,a} - \tilde{p}_{s,a}) \cdot \widetilde{V}_d\right| \\ &\leq\; \frac{\epsilon}{4(1-\gamma)^d} + \left|\sum_{s,a\in X} w(s,a)(p_{s,a} - \tilde{p}_{s,a}) \cdot \widetilde{V}_d\right| \\ &\leq\; \frac{\epsilon}{4(1-\gamma)^d} + A_d + B_d + C_d,\end{aligned}$$

where

$$A_d := \sum_{s,a\in X} w(s,a)\sqrt{\frac{8L_1\tilde{\sigma}_d(s,a)^2}{n(s,a)}} \qquad B_d := \sum_{s,a\in X} w(s,a)\frac{4L_1}{3n(s,a)(1-\gamma)^{d+1}}$$

$$C_d := \sum_{s,a\in X} w(s,a)\frac{2}{(1-\gamma)^{d+1}}\left(\frac{L_1}{n(s,a)}\right)^{3/4}.$$

The expressions B_d and C_d are substantially easier to bound than A_d. First we give a naive bound on A_d, which we use later.

$$A_d \;\le\; \sum_{s,a\in X}\sqrt{\frac{8w(s,a)\tilde{\sigma}_d(s,a)^2L_1}{n(s,a)}} \;\equiv\; \sum_{\kappa,\iota\in\mathcal{K}\times\mathcal{I}}\sum_{s,a\in X_{\kappa,\iota}}\sqrt{\frac{8w(s,a)\tilde{\sigma}_d(s,a)^2L_1}{n(s,a)}} \tag{15}$$

$$\le \sum_{\kappa,\iota\in\mathcal{K}\times\mathcal{I}}\sqrt{\frac{8L_1|X_{\kappa,\iota}|}{m\kappa}\sum_{s,a\in X_{\kappa,\iota}}w(s,a)\tilde{\sigma}_d(s,a)^2} \;\le\; \sum_{\kappa,\iota\in\mathcal{K}\times\mathcal{I}}\sqrt{\frac{8L_1}{m}\sum_{s,a\in X_{\kappa,\iota}}w(s,a)\tilde{\sigma}_d(s,a)^2} \tag{16}$$

$$\le \sqrt{\frac{8|\mathcal{K}\times\mathcal{I}|L_1}{m}\sum_{\kappa,\iota\in\mathcal{K}}\sum_{s,a\in X_{\kappa,\iota}}w(s,a)\tilde{\sigma}_d(s,a)^2} \;\le\; \sqrt{\frac{8|\mathcal{K}\times\mathcal{I}|L_1}{m}\sum_{s,a\in X}w(s,a)\tilde{\sigma}_d(s,a)^2} \tag{17}$$

$$\le \sqrt{\frac{8|\mathcal{K}\times\mathcal{I}|L_1}{m(1-\gamma)^{2d+3}}}, \tag{18}$$

where in Equation (15) we used the definition of A_d. In Equation (16) we applied Cauchy-Schwarz and the assumption that $|X_{\kappa,\iota}| \le \kappa$. In Equation (17) we used Cauchy-Schwarz again and the definition of $\mathcal{K}$. Finally we apply the trivial bound of $\sum w(s,a)\tilde{\sigma}_d(s,a)^2 \le 1/(1-\gamma)^{2d+3}$. Unfortunately, this bound is not sufficient for our needs. The solution is approximate $w(s,a)$ by $\tilde{w}(s,a)$ and use Lemma 9 to improve the last step above.

$$A_d \;\le\; \sqrt{\frac{8|\mathcal{K}\times\mathcal{I}|L_1}{m}\sum_{s,a}w(s,a)\tilde{\sigma}_d(s,a)^2} \tag{19}$$

$$\equiv \sqrt{\frac{8|\mathcal{K}\times\mathcal{I}|L_1}{m}\sum_{s,a}\tilde{w}(s,a)\tilde{\sigma}_d(s,a)^2 + \frac{8|\mathcal{K}\times\mathcal{I}|L_1}{m}\sum_{s,a}(w(s,a)-\tilde{w}(s,a))\tilde{\sigma}_d(s,a)^2} \tag{20}$$

$$\le \sqrt{\frac{8|\mathcal{K}\times\mathcal{I}|L_1}{m(1-\gamma)^{2d+2}} + \frac{8|\mathcal{K}\times\mathcal{I}|L_1}{m}\Delta_{2d+2}}, \tag{21}$$

where Equation (19) is as in the naive bound. Equation (20) is substituting $w(s,a)$ for $\tilde{w}(s,a)$ and Equation (20) uses the definition of Δ. Therefore

$$\Delta_d \;\le\; \frac{\epsilon}{4(1-\gamma)^d} + B_d + C_d + \sqrt{\frac{8L_1|\mathcal{K}\times\mathcal{I}|}{m(1-\gamma)^{2d+2}}} + \sqrt{\frac{8L_1|\mathcal{K}\times\mathcal{I}|}{m}\Delta_{2d+2}}. \tag{22}$$

The bounds on B_d and C_d are somewhat easier, and follow similar lines to the naive bound on A_d.

$$B_d \;\equiv\; \sum_{s,a\in X}w(s,a)\frac{4L_1}{3n(s,a)(1-\gamma)^{d+1}} \;=\; \frac{4L_1}{3(1-\gamma)^{d+1}}\sum_{\kappa,\iota}\frac{|X_{\kappa,\iota}|}{m\kappa} \;\le\; \frac{4|\mathcal{K}\times\mathcal{I}|L_1}{3m(1-\gamma)^{d+1}}$$

and

$$\begin{aligned}
C_d &\equiv 2\sum_{s,a\in X} w(s,a)\left(\frac{L_1}{n(s,a)}\right)^{\frac{3}{4}}\frac{1}{(1-\gamma)^{d+1}} \\
&\leq \frac{2}{(1-\gamma)^{d+1+1/4}}\sum_{\kappa,\iota\in\mathcal{K}\times\mathcal{I}}\sum_{s,a\in X_{\kappa,\iota}}\left(\frac{w(s,a)L_1}{n(s,a)}\right)^{\frac{3}{4}} \\
&\leq \frac{2}{(1-\gamma)^{d+1+1/4}}\sum_{\kappa,\iota\in\mathcal{K}\times\mathcal{I}}\left(\frac{L_1|X_{\kappa,\iota}|}{m\kappa}\right)^{\frac{3}{4}} \\
&\leq \frac{2}{(1-\gamma)^{d+1+1/4}}\sum_{\kappa,\iota\in\mathcal{K}\times\mathcal{I}}\left(\frac{L_1}{m}\right)^{\frac{3}{4}} \\
&\leq \frac{2|\mathcal{K}\times\mathcal{I}|}{(1-\gamma)^{d+1+1/4}}\left(\frac{L_1}{m}\right)^{\frac{3}{4}}
\end{aligned}$$

where in the second line we used Cauchy-Schwarz. For sufficiently large m (to be chosen shortly) it can be checked that $C_d \leq A_d$ and $B_d \leq A_d$. Expanding the recurrence in Equation (22) up to depth $d = \beta := \left\lceil \frac{1}{2\log 2}\log\frac{1}{1-\gamma}\right\rceil$ with $\mathcal{D} = \{0, 2, 6, 14, \cdots, \beta\}$leads to

$$\begin{aligned}
\Delta_0 \leq \sum_{d\in\mathcal{D}-\{\beta\}}&\left(\frac{8L_1|\mathcal{K}\times\mathcal{I}|}{m}\right)^{d/(d+2)}\left[\frac{\epsilon}{4(1-\gamma)^d}+3\sqrt{\frac{8L_1|\mathcal{K}\times\mathcal{I}|}{m(1-\gamma)^{2d+2}}}\right]^{2/(d+2)} \\
&+\left(\frac{8L_1|\mathcal{K}\times\mathcal{I}|}{m}\right)^{\beta/(\beta+2)}\left[\frac{\epsilon}{4(1-\gamma)^\beta}+3\sqrt{\frac{8L_1|\mathcal{K}\times\mathcal{I}|}{m(1-\gamma)^{2\beta+3}}}\right]^{2/(\beta+2)}, \qquad (23)
\end{aligned}$$

where we used the naive bound to control A_β.

Letting $m := \frac{1280L_1}{\epsilon^2(1-\gamma)^2}\left(\log\log\frac{1}{1-\gamma}\right)^2\left(\log\frac{|S|}{\epsilon(1-\gamma)}\right)\log\frac{1}{\epsilon(1-\gamma)}$ completes the proof. □