PAC Bounds for Discounted MDPs

Tor Lattimore and Marcus Hutter

Australian National University
{tor.lattimore,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.

Keywords: Reinforcement learning, sample-complexity, exploration exploitation, PAC-MDP, Markov decision processes.

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, Aue11].

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

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

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

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

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

N.H. Bshouty et al. (Eds.): ALT 2012, LNAI 7568, pp. 320–334, 2012.

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 A bound with quartic dependence on the horizon has been shown in [Aue11], but this work is still unpublished.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].

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 table of constants found in the appendix.

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 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. 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.

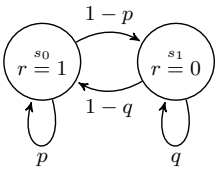
Markov Decision Process. An MDP is a tuple $M = (S, A, p, r, \gamma)$ where S and A are finite sets of states and actions respectively. $r : S \to [0, 1]$ is the reward function. $p : S \times A \times S \to [0, 1]$ is the transition function and $\gamma \in (0, 1)$ the discount rate. A stationary policy 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. $p_{s,a} \cdot V^\pi_M := \sum_{s'} p^{s'}_{s,a} V^\pi_M(s')$ is the scalar product. 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, Aue11]. 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 $p_{s,a}^{s'} = 0$ for all but two $s' \in S$ denoted $sa^+, sa^- \in S$. Note that sa^+ and sa^- are dependent on (s,a) and are known to the algorithm.*

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
  $\hat{p}_{s,a}^{sa^+} := n(s,a,sa^+) / \max\{1, n(s,a)\}$
  $\mathcal{M}_k := \left\{ \widetilde{M} : |\tilde{p}_{s,a}^{sa^+} - \hat{p}_{s,a}^{sa^+}| \leq \text{ConfidenceInterval}(\tilde{p}_{s,a}^{sa^+}, n(s,a)),\ \forall (s,a) \right\}$
  $\widetilde{M} = \text{ExtendedValueIteration}(\mathcal{M}_k)$
  $\pi_k = \tilde{\pi}^*$
  repeat
    Act
  until $v(s_{t-1}, a_{t-1}) \geq \max\{m w_{min}, n(s_{t-1}, a_{t-1})\}$ and $n(s_{t-1}, a_{t-1}) < \frac{|S|m}{1-\gamma}$
  Update(s_{t-1}, a_{t-1}) and Delay and $k = k + 1$
function Delay
  for $j = 1 \to H$ **do**
    Act
function Update(s,a)
  $n(s,a) = n(s,a) + v(s,a)$ and $n(s,a,s') = n(s,a,s') + v(s,a,s') \forall s'$
  $v(s,a) = v(s,a,\cdot) = 0$
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 $t = t + 1$
function ExtendedValueIteration$(\mathcal{M})$
  return optimistic $\widetilde{M} \in \mathcal{M}$ such that $V_{\widetilde{M}}^*(s) \geq V_{\widetilde{M}'}^*(s)$ for all $s \in S$ and $\widetilde{M}' \in \mathcal{M}$.
function ConfidenceInterval(p,n)
  return $\min \left\{ \sqrt{\frac{2L_1 p(1-p)}{n}} + \frac{2L_1}{3n},\ \sqrt{\frac{L_1}{2n}} \right\}$

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 $E_{\max} := 4m|S \times A| \log_2 |S| \log_2 \frac{1}{w_{min}(1-\gamma)}$. We write $n_t(s,a)$ to be the value of $n(s,a)$ at time-step t.

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 $|S|$.

Theorem 1. *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*

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

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 2. *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*

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

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 2 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 2, 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. Unfortunately, bounding the latter quantity appears to be somewhat more challenging due to subtle additional dependencies. Note that one can easily translate from the first type of bound to the second, but a naive method costs a factor of $1/(1-\gamma)$. In fact, it seems there is no clear way to modify the work in either this paper or theirs to achieve a bound on $|V^*(s) - V^{\hat{\pi}^*}(s)|$ that is both linear in the state space and cubic in the horizon, although either is possible at the expense of the other. 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 1 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.

$$w^\pi(s,a|s') := [\![(s', \pi(s')) = (s,a)]\!] + \gamma \sum_{s''} p^{s''}_{s',\pi(s')} w^\pi(s,a|s'')$$

$$w_t(s,a) := w^{\pi_k}(s,a|s_t).$$

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

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

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.

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

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.

$$\iota_t(s,a) := \min \left\{ z_i : z_i \geq \frac{w_t(s,a)}{w_{min}} \right\}$$
$$X_{t,\kappa,\iota} := \{(s,a) : (s,a) \in X_t \wedge \kappa_t(s,a) = \kappa \wedge \iota_t(s,a) = \iota\}$$

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,

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

Analysis. Space does not permit us to provide proofs for all results. Simple proofs are omitted while time-consuming ones are often only sketched. All details can be found in the technical report [LH12]. The proof of Theorem 1 follows easily from three key lemmas.

Lemma 3. *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 4. *$M \in \mathcal{M}_k$ for all k with probability at least $1 - \delta/2$.*

Lemma 5. *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 1.

Proof of Theorem 1. By Lemma 4, $M \in \mathcal{M}_k$ for all k with probability $1-\delta/2$. By Lemma 5 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 3, 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 3, 4 and 5. Of these, only Lemma 5 presents a substantial challenge.

Proof of Lemma 3. 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 $\widetilde{V}^{\pi_k}(s_t) - V^{\text{UCRL}\gamma}(s_{1:t}) > \epsilon$. Finally note that t is a non-delaying time-step and so 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 4. 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 $|\hat{p}^{sa^+}_{s,a} - p^{sa^+}_{s,a}| \leq \text{ConfidenceInterval}(p^{sa^+}_{s,a}, n(s, a)))$ 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 5, 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 $X_{t,\kappa,\iota}$

[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].

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 at most $U_{\max}$ updates completes the result.

Lemma 6. *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 the associated technical report [LH12]. 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, motivates the use of Bernstein's inequalities and is the key observation to improve on the unpublished work in [Aue11], which has quartic dependence on the horizon.

Lemma 7. *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. Expand and rearrange the definition of the value functions. ■

Lemma 8 (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 6. 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 7

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

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

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

where in Equation (4) we used Lemma 7 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-Schwartz inequality,[4] the fact that $\kappa \geq |X_{t,\kappa,\iota}|$ and Lemma 8. Substituting

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

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 6 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,

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

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

Lemma 9. *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 of Lemma 5. Let $N := |S \times A| m$, where m is as in the proof of Lemma 6 or the appendix. We proceed in two stages. First we bound the total number of *useful* visits before $|X_{t,\kappa,\iota}| \leq \kappa$. Note that if the knownness, κ, is equal to $|S|$ then $|X_{t,\kappa,\iota}| \leq \kappa$ is vacuously true because the number of active state/action pairs is bounded by $|S|$. We then use this show that the number of exploration phases is at most $\tilde{O}(N)$ with high probability.

Bounding the Number of Useful Visits. A visit to state/action pair (s,a) in exploration phase starting at time-step t is (κ,ι)-*useful* if $(s,a) \in X_{t,\kappa,\iota}$ and $|X_{t,\kappa,\iota}| > \kappa$. Fixing a (κ,ι) we bound the number of (κ,ι)-useful visits to state/action pair (s,a). Suppose $t_1 < t_2$ with t_1 the start of an exploration phase and $(s,a) \in X_{t_1,\kappa,\iota}$. Therefore $n_{t_1}(s,a) < \kappa w_\iota m$. Now if $n_{t_2}(s,a) - n_{t_1}(s,a) \geq \kappa w_\iota m$ then an update ocurrs and for every $t_3 \geq t_2$ such that $\iota_t(s,a) = \iota$, $\kappa_t(s,a) > \kappa$. Therefore for each (κ,ι) pair there at most $|S \times A| m w_\iota \kappa \equiv N w_\iota \kappa$ visits that are (κ,ι)-useful.

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 6 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

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

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. Since $|X_{t_i,\kappa,\iota}| > \kappa$ and by Lemma 9, we have that $\mathbf{E}[\nu_i|\nu_1\cdots\nu_{i-1}] \geq (\kappa+1)w_\iota/2$ and $\mathrm{Var}[\nu_i|\nu_1\cdots\nu_{i-1}] \leq \mathbf{E}[\nu_i|\nu_1\cdots\nu_{i-1}]H$.[5]

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

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

which are chosen such that $\mathbf{E}\bar{\nu}_i = \mathbf{E}[\bar{\nu}_i|\bar{\nu}_1\cdots\bar{\nu}_{i-1}] = w_\iota(\kappa+1)/2$. It is straightforward to verify that $\mathrm{P}\{E_{\kappa,\iota} > 4N\} \leq \mathrm{P}\left\{\sum_{i=1}^{4N}\bar{\nu}_i \leq Nw_\iota(\kappa+1)\right\}$. 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 $B_{4N} = \sum_{i=1}^{4N}\bar{\nu}_i$, $B_0 = 2Nw_\iota(\kappa+1)$ and $|B_{i+1}-B_i| \leq H$. Letting $\sigma^2 := \sum_{i=1}^{4N}\mathrm{Var}[B_i|B_1\cdots B_{i-1}] \leq 2NHw_\iota(\kappa+1)$, which follows by the definitions of B, $\bar{\nu}$ and by point 3 above. Then

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

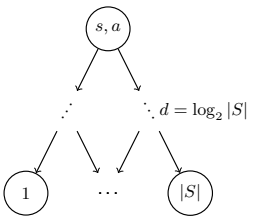
Setting this equal to δ_1, solving for N and noting that $w_\iota(\kappa+1) \geq w_{\min}$ gives

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

Since N satisfies this, the result is complete. ■

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

The result above completes the proof of Theorem 1. We now drop the assumption on the number of next-states by proving the more general Theorem 2. While it is possible to do this directly, we use the algorithm above.



Proof sketch of Theorem 2. The idea is to augment each state/action pair of the original MDP with $|S| - 2$ states in the form of a 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.

To make the augmented MDP functionally equivalent to the true one we must also rescale γ. Let d be the depth of the tree then γ must be rescaled to $\bar{\gamma}$ such that $\bar{\gamma}^d = \gamma$. The augmented MDP is now functionally equivalent to the original in the obvious way. 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 rescaling of γ has the potential to damange the bound. This is true, but 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 1 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 $T(s,a) := \sum_{s'} [\![p^{s'}_{s,a} > 0]\!]$ 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 $|S \times A| + \log|S| \sum_{s,a} T(s,a) < 2T\log|S|$. Finally note that by Equation (2) from Lemma 7, state/action pairs that are not reachable do not contribute to the error and need no visits. This allows the analysis in Lemma 5 to be tightened, which completes the proof. ■

6 Lower PAC Bound

We now turn our attention to the lower bound. The approach is similar to that of [SLL09], but we make two refinements to improve the bound to depend on $1/(1-\gamma)^3$ and remove the policy restrictions. 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 10. *Let $\mathcal{A}$ be a (possibly non-stationary) policy depending on $S, A, r, \gamma, \epsilon$ and δ, then there exists a Markov decision process M_{hard} such that $V^*(s_t) - V^{\mathcal{A}}(s_{1:t}) > \epsilon$ for at least N time-steps with probability at least δ where*

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

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

The proof is omitted, but we give the counter-example and intuition.

Counter Example. We prove Theorem 10 for 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 and confer maximum/minimum rewards 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 such that, on average, a policy will stay there for $1/(1-\gamma)$ time-steps.

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

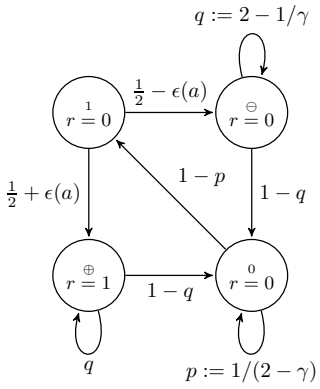


Fig. 1. Hard MDP

$$\tilde{O}\left(\frac{|A|}{\epsilon^2(1-\gamma)^2} \log \frac{1}{\delta}\right). \tag{7}$$

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 (7) 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 10 holds for all policies.

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 1 with the current state-of-the-art, MORMAX [SS10].

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

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 1 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

[AJO10] Auer, P., Jaksch, T., Ortner, R.: Near-optimal regret bounds for reinforcement learning. J. Mach. Learn. Res. 99, 1563–1600 (2010)

[AMK12] Azar, M., Munos, R., Kappen, B.: On the sample complexity of reinforcement learning with a generative model. In: Proceedings of the 29th International Conference on Machine Learning. ACM, New York (2012)

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

[Aue11] Auer, P.: Upper confidence reinforcement learning. Unpublished, keynote at European Workshop of Reinforcement Learning (2011)
[CL06] Chung, F., Lu, L.: Concentration inequalities and martingale inequalities a survey. Internet Mathematics 3, 1 (2006)
[Kak03] Kakade, S.: On The Sample Complexity of Reinforcement Learning. PhD thesis, University College London (2003)
[LH12] Lattimore, T., Hutter, M.: PAC bounds for discounted MDPs. Technical report (2012), http://arxiv.org/abs/1202.3890
[LR85] Lai, T., Robbins, H.: Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics 6(1), 4–22 (1985)
[MT04] Mannor, S., Tsitsiklis, J.: The sample complexity of exploration in the multi-armed bandit problem. J. Mach. Learn. Res. 5, 623–648 (2004)
[SL05] Strehl, A., Littman, M.: A theoretical analysis of model-based interval estimation. In: Proceedings of the 22nd International Conference on Machine Learning, ICML 2005, pp. 856–863 (2005)
[SL08] Strehl, A., Littman, M.: An analysis of model-based interval estimation for Markov decision processes. Journal of Computer and System Sciences 74(8), 1309–1331 (2008)
[SLL09] Strehl, A., Li, L., Littman, M.: Reinforcement learning in finite MDPs: PAC analysis. J. Mach. Learn. Res. 10, 2413–2444 (2009)
[SLW+06] Strehl, A., Li, L., Wiewiorac, E., Langford, J., Littman, M.: PAC model-free reinforcement learning. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 881–888. ACM, New York (2006)
[Sob82] Sobel, M.: The variance of discounted Markov decision processes. Journal of Applied Probability 19(4), 794–802 (1982)
[SS10] Szita, I., Szepesvári, C.: Model-based reinforcement learning with nearly tight exploration complexity bounds. In: Proceedings of the 27th International Conference on Machine Learning, pp. 1031–1038. ACM, New York (2010)

A Constants

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