Reinforcement Learning in POMDPs Without Resets

Eyal Even-Dar
School of Computer Science
Tel-Aviv University
Tel-Aviv, Israel 69978
evend@post.tau.ac.il

Sham M. Kakade
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
skakade@linc.cis.upenn.edu

Yishay Mansour
School Computer Science
Tel-Aviv University
Tel-Aviv, Israel 69978
mansour@post.tau.ac.il

Abstract

We consider the most realistic reinforcement learning setting in which an agent starts in an unknown environment (the POMDP) and must follow one continuous and uninterrupted chain of experience with no access to "resets" or "offline" simulation. We provide algorithms for general connected POMDPs that obtain near optimal average reward. One algorithm we present has a convergence rate which depends *exponentially* on a certain horizon time of an optimal policy, but has *no dependence* on the number of (unobservable) states. The main building block of our algorithms is an implementation of an *approximate* reset strategy, which we show always exists in every POMDP. An interesting aspect of our algorithms is how they use this strategy when balancing exploration and exploitation.

1 Introduction

We address the problem of lifelong learning in a partially observable Markov decision process (a POMDP). We consider the most general setting where an agent begins in an unknown POMDP and desires to obtain near optimal reward. In this setting, the agent is forced to obey the dynamics of the environment, which, in general, do not permit resets.

The problem of lifelong learning has been well studied for observable MDPs. Kearns and Singh (1998) provide the E^3 algorithm, which has finite (polynomial) time guarantees until the agent obtains near optimal reward. Unfortunately, such an algorithm is not applicable in the more challenging POMDP setting. In fact, none of the guarantees in the literature for learning in the limit for MDPs apply to POMDPs, for reasons which are essentially due to the partially observability.

For POMDPs, the problem of balancing exploitation with exploration has received rather little attention in the literature — typically most results in POMDPs are on planning (see for example Sondik (1971); Lovejoy (1991a, 1991b); Hauskrecht (1997); Cassandra (1998)). Most of the existing learning algorithms such as Parr and Russell (1995); Peshkin et al. (2000) either assume a goal state or assume a reset button. In fact, to the best of our knowledge, the literature does not contain even asymptotic results for general POMDPs which guarantee that the average reward of an agent will be near optimal in the limit.

Part of the technical difficulty is that there are currently no general results for belief state tracking with an approximate model showing that divergence in the belief state does not eventually occur. Crudely, the issue is that if a belief state is being tracked in an approximate manner, then it is important to show that this approximation quality does not continually degrade with time — otherwise the agent will eventually loose track of the belief state in the infinite horizon (this of course is not an issue in an MDP where the current state is observable). Boyen and Koller (1998) address the issue of approximate belief state tracking, but in their setting the model is known perfectly and their goal is to keep a compact representation of the belief state. Note that approximate belief state tracking is much simpler if the agent is only acting over a fixed finite horizon, since then one can bound the error accumulation as a function of the horizon.

We present new algorithms for learning in POMDPs which guarantee that the agent will obtain the optimal average reward in the limit. Furthermore, we provide a finite time convergence rates for one of our algorithms which has an *exponential* dependence on a certain horizon time (of an optimal strategy) but has *no dependence* on the number of states in the POMDP. This result is reminiscent of the trajectory tree algorithm of Kearns et al. (1999) which has similar dependencies (though there they assumed access to a generative model, which allowed simulation of the POMDP). Given the plethora of complexity results in the literature on planning in POMDPs (see Lusena et al. (2001)), we feel these dependencies are the best one could hope for in the most general setting.

Central to our algorithms is the implementation of an approximate reset strategy or a *homing strategy*. The idea of a reset strategy is not new to the literature — homing sequences were used in the learning of deterministic finite automata (see Rivest and Schapire (1993), though there the sequence provided *exact* resets). Here, the agent follows a homing strategy in order to move *approximately* towards a reset. We show that such a strategy always exists, and our finite convergence rates also depend on a characteristic time it takes to approximately reset. However, note that existence of such a strategy alone does not imply that such a strategy will be useful.

The reason is that the agent must take actions to reset, which might otherwise be better spent exploring or exploiting. It turns out that our algorithms use the homing strategy while *both* exploring and exploiting. In fact, they use the homing strategies infinitely often, which, unfortunately, detracts from exploiting. However, we are able to show that the ratio of the time these homing strategies are used compared to the time spent exploiting is decreasing sufficiently rapidly, such that near optimal average reward can be obtained.

2 Preliminaries

A Partially Observable Markov Decision Process (POMDP) is defined by a finite set of states S, an initial state s_0, a set of actions A, a set of observations O, with an output model Q, where $Q(o, r|s, a)$ is the probability of observing o and reward r after performing action a at state s (we assume that $r \in [0, 1]$), and a set of transitions probabilities P, where $P(s'|a, s)$ is the transition probability from s to s' after performing action a. We define $r(s, a)$ as the expected reward under $Q(\cdot|s, a)$ after performing action a in state s.

A *history* h is a sequence of actions, rewards and observations of some finite length, *i.e.* $h = \{(a_1, r_1, o_1), ..., (a_t, r_t, o_t)\}$. A *strategy* or *policy* in a POMDP is defined as a mapping from histories to actions. We define a *belief state* B to be a distribution over states. Given an initial belief state B_0 let $\Pr[h|B_0] = \Pr[r_1, o_1 \ldots, r_t, o_t|a_1, \ldots, a_t, B_0]$ be the probability of observing the sequence of reward-observations $(r_1, o_1, \ldots r_t, o_t)$ after performing the actions $a_1 \ldots a_t$.

For each strategy π we define its t-horizon expected reward from a belief state B as $R_t^\pi(B) = (1/t)E_{h\sim\pi}[\sum_{i=1}^t r(s_i, a_i)|B_0 = B]$. A *t-Markov* strategy is a strategy that depends only on the last t observations. The optimal t-Markov strategy's expected return from initial belief state B is defined as $R_t^*(B)$.

The only assumption we make is that the POMDP is *connected*, *i.e.* for all states s, s', there exists a strategy π which can reach s' with positive probability starting from s. (We do not make any ergodicity assumptions, since strategies are by definition non-stationary). Note that if the POMDP is disconnected, then the best statement we could hope for is to obtain the optimal average reward for one of its connected components.

Connectivity implies that there exists a strategy π^* that maximizes the average reward. More formally, there exists a π^* such that: i) for every B, $\lim_{t\to\infty} R_t^{\pi^*}(B)$ exists and does not depend on B, which we denote by R^*, and ii) for all π and B, $R^* \geq \lim_{t\to\infty} \sup R_t^\pi(B)$. Hence, for all $\epsilon > 0$ there exist a τ, such that for all B and $t \geq \tau$:

$$|R^* - R_t^{\pi^*}(B)| \leq \epsilon$$

and we refer to τ as the *ϵ-horizon time* of the optimal strategy. Essentially, τ is the timescale in which the optimal strategy achieves close to its average reward.

When we say that we *restart* a t-Markov strategy π from a belief state B we mean specifically that we reset the history, i.e., $h = \emptyset$, and run π starting from a state s distributed according to B.

3 Homing Strategies

Clearly, having an action which resets the agent to some designated state would be useful, as it would allow us to test and compare the performance of various policies, starting at the same start state. However, in general, such an action is not at our disposal.

Instead our algorithms utilize an approximate reset, which we show always exists. There are a few subtle points when designing such a reset. First, we must select actions to achieve the approximate reset, *i.e.* the approximate reset is done through the use of a *homing strategy*. Hence, while homing, the agent is neither exploring nor exploiting. Second, rather than moving to a fixed state, the homing strategy can only hope to move to toward a fixed (unknown) belief state. Third, as we shall see, since the POMDP might be periodic the stopping time must be a random variable[1]. To implement this randomized stopping time, we introduce *fictitious* 'stay' actions, in which the agent does not take an action that period. By this, we mean that if the homing strategy decides to take a 'stay' action at some time — which may not be possible if the true POMDP does not permit 'stay' actions — then the agent just ignores this 'stay' action and obtains another action from the homing strategy to execute. Hence, after the agent has taken t *homing actions* (which are either real or 'stay' actions), the agent has taken $t - m$ real actions in the POMDP and m stay actions. We now define an approximate reset strategy.

Definition 3.1 *H is an (ϵ, k)-approximate reset (or homing) strategy if for every two belief states B_1 and B_2, we have $\|H_E(B_1) - H_E(B_2)\|_1 \leq \epsilon$, where $H_E(B)$ is the expected belief state reached from B after k homing actions of H (so at most k real actions have been taken) and $H(B)$ is a random variable such that $H_E(B) = E_{H(B)\sim H,B}[H(B)]$.*

The above definition only states that H will approximately reset, but this approximation quality could be poor. We now show how to amplify the accuracy of an approximate homing strategy, and then we show that an approximate homing strategy always exists.

Lemma 3.1 *Suppose that H is an (ϵ, k) approximate reset then H^ℓ is an $(\epsilon^\ell, k\ell)$ approximate reset, where H^ℓ consecutively implements H for ℓ times. Furthermore, this implies there exists a unique belief state B_H such that $H_E(B_H) = B_H$.*

Proof: The proof is a standard contraction argument, and we use induction. For $l = 1$, the claim follows by definition. Assume now that $\|H_E^{\ell-1}(B_1) - H_E^{\ell-1}(B_2)\|_1 \leq \epsilon^{\ell-1}$. Let $H_E^{l-1}(B_1) = Q_1$ and $H_E^{l-1}(B_2) = Q_2$, so $\sum_s |Q_1(s) - Q_2(s)| \leq \epsilon^{\ell-1}$. For an arbitrary state s', and using the fact

[1] Without randomizing over the stopping times (*i.e.* allowing 'stay' actions), the state transition matrix may be periodic and no stationary distribution may exist, *e.g.* if the states deterministically alternate between states 1 and 2.

that H is a linear operator, we have

$$\begin{aligned}
&\|H_E^\ell(B_1) - H_E^\ell(B_2)]\|_1 \\
&= \|H_E(Q_1) - H_E(Q_2)]\|_1 \\
&= \|\sum_s (Q_1(s) - Q_2(s))H_E(s)\|_1 \\
&= \|\sum_s (Q_1(s) - Q_2(s))H_E(s')\|_1 \\
&\quad +\|\sum_s (Q_1(s) - Q_2(s))(H_E(s) - H_E(s'))\|_1 \\
&= 0 + \sum_s |Q_1(s) - Q_2(s)|\epsilon \\
&\leq \epsilon^\ell
\end{aligned}$$

where the first term is 0 since for any two distributions $\sum_s Q_1(s) - Q_2(s) = 1 - 1 = 0$ (and the vector $H(s')$ is constant in this sum), and we have used the fact that $\|H(s) - H(s')\|_1 \leq \epsilon$ (by definition of H). □

We now show that the random walk strategy (including 'stay' actions) is an approximate reset strategy in every POMDP (including periodic ones), though with prior knowledge we might have better approximate reset strategies at our disposal.

Lemma 3.2 *For all POMDPs, the random walk strategy (using 'stay' actions) constitutes an (ϵ, k) approximate reset strategy for some $k \geq 1$ and $0 < \epsilon < \frac{1}{2}$.*

Proof: By our connectivity assumption, for all states s and s', there exists some strategy that reaches s' from s with positive probability. This implies that under H (the random walk strategy), there is positive probability of moving from one state to another, *i.e.* the Markov chain is irreducible. Furthermore, since H performs 'stay' actions, then the Markov chain is aperiodic. Thus, there exists a unique stationary distribution. We choose k to be the time at which the error in convergence is less than $1/2$, from all starting states. Hence, by linearity of expectation, the error is less than $1/2$ from all belief states in k steps. □

4 Reinforcement Learning with Homing

We now provide two algorithms which demonstrate how near-optimal average reward can be obtained, with different rates of convergence. The key to the success of these algorithms is their use of homing sequences in *both* exploration and in exploitation. For exploration, the idea is that each time we attempt some exploration trajectory we do it after implementing our reset strategy — hence our information is (approximately) grounded with respect to the belief state B_H (recall $H_E(B_H) = B_H$). The idea of exploration is to find a good t-Markov strategy $\hat{\pi}_t^*$ from B_H. During exploitation, the goal is to use this t-Markov strategy. Unfortunately, we have only guaranteed that $\hat{\pi}_t^*$ performs well *starting from* B_H and only for t steps. Hence, after each time we exploit with $\hat{\pi}_T^*$, we run our homing sequence to get back close to B_H (and then we rerun $\hat{\pi}_t^*$). We gradually increase t in the process.

The problem is that while homing, we are wasting time and neither exploiting nor exploring. Furthermore, since we use the homing sequence between *every* run of $\hat{\pi}_t^*$, asymptotically we never stop homing. Nonetheless, we able to show that there exists an algorithm which obtains near optimal reward in a POMDP, since the ratio of the time spent exploiting vs. homing decreases sufficiently fast.

```
Input  : $H$ /*a $(1/2, K_H)$ approximate reset strategy */
for $t = 1$ to $\infty$ do
    /*Exploration in Phase $t$ */;
    $k_1^t = O\left(\frac{1}{\epsilon_t^2}\log(t^2|\Pi_t|)\right)$;
    foreach Policy $\pi$ in $\Pi_t$ do
        for $i = 1$ to $k_1^t$ do
            Run $\pi$ for $t$ steps;
            Repeatedly run $H$ for $\log(1/\epsilon_t)$ times;
        end
        Let $v^\pi$ be the average return of $\pi$ in from these
        $k_1^t$ trials;
    end
    /*Exploitation in Phase $t$ */;
    Let $\hat{\pi}_t^* = \arg\max_{\pi \in \Pi_t} v^\pi$;
    $k_2^t = O\Big(\frac{1}{\epsilon_t}($[current time T]
            $+$[time in $t+1$-th exploration phase]$)\Big)$;
    for $i = 1$ to $k_2^t$ do
        Run $\hat{\pi}_t^*$ for $t$ steps;
        Repeatedly run $H$ for $\log(1/\epsilon_t)$ times;
    end
end
```

Algorithm 1: Policy Search

Theorem 4.1 *There exists an algorithm $\mathcal{A}$, such that in any connected POMDP, $\mathcal{A}$ obtains the optimal average reward in the limit, with probability* 1.

We later provide an algorithm with a better convergence rate (see Theorem 4.5). However, we start with a simpler policy search algorithm which establishes the above Theorem.

4.1 Policy Search

Algorithm 1 takes as input a $(1/2, K_H)$-approximate reset strategy, which could be the random walk strategy with a very crude reset. The algorithm works in phases, interleaving exploration phases with exploitation phases. Let us start by describing the exploration phases. Let Π_t be the set of all t-Markov strategies. An estimate of the value of a policy $\pi \in \Pi_t$ can be found by first resetting and then running π for t steps. The exploration phase consists of obtaining an estimate v^π of the return of each policy $\pi \in \Pi_t$, where each estimate consists of an average of k_1 trials (followed by approximate resets).

These estimates have both bias and variance. The variance is just due to the stochastic nature of the POMDP. The bias is due to the fact that we never can exactly reset to B_H. However, if we run H for $\log 1/\epsilon_t$ times (where ϵ_t is an error parameter in the t-th phase, which will be fixed latter) then, by lemma 3.1, all expected belief states we could approach will be $(1/2)^{\log 1/\epsilon_t} = \epsilon_t$ close to B_H. The following lemma shows that accurate estimates can be obtained.

Lemma 4.2 *In phase t, if $k_1^t = O\left(\frac{1}{\epsilon_t^2}\log(t^2|\Pi_t|)\right)$ and each reset consists of using the homing sequence $\log 1/\epsilon_t$ times, then for all policies $\pi \in \Pi_t$, the estimated t-horizon reward, v^π, satisfies $|R_t^\pi(B_H) - v^\pi| \le 2\epsilon_t$ with probability greater than $1 - \frac{1}{2t^2}$.*

Proof: First, let us deal with the bias. Any expected belief states b that results from using the homing sequence $\log 1/\epsilon_t$ times must satisfy $\|b - B_H\| \le (1/2)^{\log 1/\epsilon_t} \le \epsilon_t$. Now it is straightforward to see that if b and B_H are belief states such that $\|b - B_H\|_1 \le \epsilon$, then for every strategy π, $|R_t^\pi(b) - R_t^\pi(B_H)| \le \epsilon$. To see this, let the belief states at time t be b^t and B_H^t, which result from following either π starting from b or B_H, respectively. By linearity of expectation, it follows that $\|b^t - B_H^t\|_1 \le \epsilon$. This directly implies that $|R_t^\pi(b) - R_t^\pi(B_H)| \le \epsilon$.

For the variance, the Hoeffding bound and our choice of k_1^t imply that the average return of each policy is ϵ_t close to its expectation (the expectation is both over the initial state and on the policy trajectory) with probability $1 - 1/(2t^2)$. □

Now during exploitation, the algorithm uses the policy $\hat{\pi}_t^*$ which had the highest return in the exploration phase (and by the previous lemma this is close to the policy with largest return). Note that we have only guaranteed a large return for executing $\hat{\pi}_t^*$ *from* B_H for t steps. However, we would like to exploit for a longer period of time than t. The key is that we again reset $\log(1/\epsilon_t)$ times after each time we run $\hat{\pi}_t^*$, which resets us close ϵ_t close to B_H. Unfortunately, this means we spend $K_H \log 1/\epsilon_t$ steps between *each* run of $\hat{\pi}^*$. Hence, our average return could be $O(\frac{1}{t}K_H \log 1/\epsilon_t)$ less than we would like, since $O(\frac{1}{t}K_H \log 1/\epsilon_t)$ is the fraction of time we spend resetting. Note that this fraction could be large if we desire that ϵ_t be very small (thought this would guarantee very accurate resets).

Now, when we do exploit (and reset), we run the exploitation phase long enough (for k_2^t time) such that our *overall* average reward is comparable to the average reward in the last exploitation phase.

Lemma 4.3 *At any time T after phase t, the average reward from time 1 to time T satisfies: $\frac{1}{T}\sum_{i=1}^{T} r_i \ge R_t^*(B_H) - O(\epsilon_t + \frac{1}{t}K_H \log(1/\epsilon_t))$ with probability at least $1 - \frac{1}{t^2}$.*

Before proving this lemma, let us state a corollary from which Theorem 4.1 follows.

Corollary 4.4 *Let $\epsilon_t = 1/t$. Then $\frac{1}{T}\sum_{i=1}^{T} r_i \ge R_t^*(B_H) - O(\frac{K_H \log(t)}{t})$ with probability at least $1 - \frac{1}{t^2}$.*

Importantly, note the loss term goes to 0 as t goes to infinity. Furthermore, for a large enough phase t, we know that $R_t^*(B_H)$ will approach the optimal average reward R^* (since R^* is independent of the starting B). Theorem 4.1 follows. Essentially, although we have to home infinitely often, the ratio of time spent homing to the time spent using our t-step exploitation policies is going as $O(\frac{\log t}{t})$, which goes to 0.

Proof: First, let us show that the average reward, $\frac{1}{T}\sum_{i=1}^{T} r_i \ge R_t^*(B_H)$, is no less than ϵ_t from the average reward obtained in the t-th exploitation phase. To do this, we set the time of exploitation phase, k_2^t, to be $1/\epsilon_t$ times greater than previous amount of time spent in the MDP time plus the amount of time that will be spent in the next exploration phase (this latter factor accounts for the case in which time T lies in the exploration phase immediately after t).

Now we bound the average reward obtained in the exploitation phase. First. let us show that the t-average reward of the policy used, $R_t^{\hat{\pi}_t^*}$, satisfies $R_t^*(B_H) - R_t^{\hat{\pi}_t^*}(B_H) \le 4\epsilon_t$ with probability at least $1 - \frac{1}{2t^2}$. By Lemma 4.2 for each policy $\pi \in \Pi_t$, we have $|R_t^\pi(B_H) - v^\pi| \le 2\epsilon_t$ with probability at least $1 - \frac{1}{2|\Pi_t|t^2}$. Therefore, $\hat{\pi}_t^*$ is $4\epsilon_t$-optimal with probability $1 - 1/(2t^2)$. Now the observed average return of $\hat{\pi}_t^*$ in the exploitation period is $2\epsilon_t$ close to $R_t^*(B_H)$ with probability at least $1 - \frac{1}{2t^2}$, since our observed average return in exploitation is least as good as those used to find $v^{\hat{\pi}_t^*}$ (since $k_2^t > k_1^t$).

However, the average return during the exploitation phase is not the observed average return of $\hat{\pi}_t^*$, since we reset after each t exploitation steps for a number of steps that is $K_H \log(1/\epsilon_t)$. The resets in the exploitation period can change the average reward by at most a fraction $\frac{1}{t}K_H \log(1/\epsilon_t)$. □

4.2 A Model Based Algorithm

The previous algorithm was the simplest way to demonstrate Theorem 4.1. However, it is very inefficient, since it is testing all t-Markov policies — there are doubly exponential, in t, such polices[2]. Here, we provide a more efficient model based algorithm, which resembles the algorithms given in Kearns et al. (1999); McAllester and Singh (1999), and is exponential in the horizon time, yet it still has no dependence on the number of states in the POMDP.

We now state a convergence rate in terms of τ, the ϵ-horizon time of an optimal policy (see Section 2) and and in terms of the homing time K_H (recall, such a time exists for every POMDP using a random walk policy).

Theorem 4.5 *There exists an algorithm $\mathcal{A}$, such that in any connected POMDP and with probability greater than $1 - \delta$, $\mathcal{A}$ achieves an average reward that is 2ϵ close to the optimal average reward in a number of steps in the POMDP which is polynomial in $|A|$,$|O|$,K_H and $\log(1/\delta)$ and exponential in τ. Furthermore the computational runtime of this algorithm is polynomial in $|A|$,$|O|$, and $\log(1/\delta)$ and exponential τ.*

We provide such an algorithm in the next page. In exploration phase, the algorithm builds an approximate model of the transition probabilities after some history has occurred starting from B_H. In the t-th phase, it builds a model with respect to the set of all t-length histories, which we denote by $\mathcal{H}_t$. In the exploitation phase, it uses the best t Markov strategy with respect to this model. The use of homing strategies is similar to that in the previous algorithm.

Let $L = |A||O|$, and note that $2L^t \ge |\mathcal{H}_t|$. In the exploration phase, the algorithm takes actions uniformly at random for t steps and then resets (running the homing strategy

[2] The number of histories of length t is exponential in t, and the number of t-Markov polices is exponential in the number of t-length histories

```
Input : H /*an $(1/2, K_H)$ approximate reset strategy */
Let $L = |A| \cdot |O|$;
for $t = 1$ to $\infty$ do
    $k_1^t = O\left(\frac{L^{4t}}{\epsilon_t^2}\log(t^2|\mathcal{H}_t|)\right)$;
    for $k_1^t$ times do
        Run RANDOM for $t$ steps;
        Run $H$ for $K_H \log(L^t/\epsilon_t)$ steps.
    end
    for $h \in \mathcal{H}_t$, $a \in A$ and $o \in O$ do
        $\hat{\Pr}[o|h, B_0, a] = 0$;
        if $\hat{\Pr}[h(a,o)|B_0] \geq \frac{\epsilon_t}{L^t}$ then
            $\hat{\Pr}[o|h, B_0, a] = \frac{\hat{\Pr}[h(a,o)|B_0]}{\hat{\Pr}[h|B_0]\hat{\Pr}[a]}$
        end
    end
    Compute $\hat{\pi}_t^*$ using $\hat{\Pr}[o|B_0, h, a]$;
    $k_2^t = O\Big(\frac{1}{\epsilon_t}($[current time T]
            $+$[time in $t+1$-th exploration phase]$)\Big)$;
    for $k_2^t$ times do
        Run $\hat{\pi}_t^*$ for $t$ steps;
        Run $H$ for $K_H \log(1/\epsilon_t)$ steps;
    end
end
```

Algorithm 2: Model based

for $\log(L^t/\epsilon_t)$ times). This is done k_1^t times.[3] Then using the empirical frequencies in these trajectories the algorithm forms estimates $\hat{\Pr}[o|h, B_H, a]$, which is just the empirical probability of observing o conditioned on history h followed by taking action a. For histories h which are unlikely, these empirical estimates could be very bad, though, as we shall see, we do not need accurate estimates of $\hat{\Pr}[o|h, B_H, a]$ for such histories. Let $h(a,o)$ be a history with h followed by (a,o).

Lemma 4.6 *In phase t, if $k_1^t = O\left(\frac{L^{4t}}{\epsilon_t^2}\log(t^2|\mathcal{H}_t|)\right)$ and each reset consists of using the homing sequence $\log(L^t/\epsilon_t)$ times, then: (1) $|\hat{\Pr}[h|B_0] - \Pr[h|B_0]| \leq \frac{\epsilon_t}{L^{2t}}$, and (2) for every $h(a,o) \in \mathcal{H}_t$ such that $\Pr[h(a,o)|B_0] \geq \frac{\epsilon_t}{L^t}$, we have $|\hat{\Pr}[o|h, B_0, a] - \Pr[o|h, B_0, a]| \leq \frac{2|A|}{L^t}$, with probability at least $1 - \frac{1}{t^2}$.*

Proof: We first note that, with probability $1 - 1/t^2$, for every history $h \in \mathcal{H}_t$ we have $|\hat{\Pr}[h|B_0] - \Pr[h|B_0]| \leq \frac{\epsilon_t}{L^{2t}}$ (using the Hoeffding bound). The error, $|\Pr[o|h, B_0, a] - \hat{\Pr}[o|h, B_0, a]|$, is then

$$\begin{aligned}
&\left|\frac{\Pr[h(a,o)|B_0]}{\Pr[h|B_0]\hat{\Pr}[a]} - \frac{\hat{\Pr}[h(a,o)|B_0]}{\hat{\Pr}[h|B_0]\hat{\Pr}[a]}\right| \\
&\leq \quad \frac{1}{\hat{\Pr}[a]}\left|\frac{\Pr[h(a,o)|B_0] + \frac{\epsilon_t}{L^{2t}}}{\Pr[h|B_0] - \frac{\epsilon_t}{L^{2t}}} - \frac{\Pr[h(a,o)|B_0]}{\Pr[h|B_0]}\right| \\
&= \quad \frac{1}{\hat{\Pr}[a]}\left|\frac{\frac{\epsilon_t}{L^{2t}}}{\Pr[h|B_0] - \frac{\epsilon_t}{L^{2t}}} + \frac{2\Pr[h(a,o)|B_0]\frac{\epsilon_t}{L^{2t}}}{\Pr[h|B_0](\Pr[h|B_0] - \frac{\epsilon_t}{L^{2t}})}\right| \\
&\leq \quad \frac{2|A|}{L^t},
\end{aligned}$$

where the first inequality holds with probability $1 - \frac{1}{t^2}$, and in the last inequality we used the fact that $\Pr[h|B_0] \geq \frac{\epsilon_t}{L^t}$. □

The exploitation policy can be found using dynamic programming with the model. Note that the POMDP is equivalent to an MDP where the histories are states. In the exploitation phase, the algorithm uses the best t-Markov policy, $\hat{\pi}_t^*$, (with respect to the approximate model) interleaving it with $K_H \log(1/\epsilon_t)$ homing steps.

Lemma 4.7 *In phase t, the exploitation policy $\hat{\pi}_t^*$, satisfies $|R_t^*(B_H) - R_t^{\hat{\pi}_t^*}(B_H)| \leq t(\epsilon_t + \frac{2|A|}{L^t}) + (2\epsilon_t + \frac{2\epsilon_t}{L^t})$ with probability at least $1 - \frac{1}{t^2}$.*

Proof: (sketch) We observe that by ignoring all histories (which we view as nodes in a tree) such that $\hat{\Pr}(h|B_0) \leq \frac{\epsilon_t}{L^t}$, the return of an optimal strategy in this empirical model is decreased by at most $2t(\epsilon_t + \frac{\epsilon_t}{L^t})$, due to the fact that the true history probability is bounded by $\frac{\epsilon_t}{L^t} + \frac{\epsilon_t}{L^{2t}}$, the return from each node is bounded by t and the total number of such nodes is bounded by $2L^t$. Next we prove that the return of the optimal policy in the empirical model loses at most $t^2(\epsilon_t + \frac{2|A|}{L^t})$ due to the tree approximation on the other nodes (the other histories). Using backward induction, we show that the policy $\hat{\pi}_t^*$ has return not less than $k^2(\epsilon_t + \frac{2|A|}{L^t})$ in comparison to the true optimal value, starting from $(t-k+1)$-length histories. The base case, for the leaves (the t-length histories), holds since the reward (which is encoded through the observations) is within $\epsilon_t + \frac{2|A|}{L^t}$, where the first error is due to the imperfect reset and the second is due to the marginal distribution error that is bounded by $\frac{2|A|}{L^t}$ by Lemma 4.6. Assume the induction assumption holds for $k-1$. There are two sources of error, the first is due to the current estimation error (of both the marginal distribution and the immediate reward) which is bounded by $(\epsilon_t + \frac{2|A|}{L^t})k$ and the second is due to errors from the previous levels and is bounded by $(k-1)^2(\epsilon_t + \frac{2|A|}{L^t})$ by the induction assumption. Summing the terms completes the induction step. □

Similarly to Subsection 4.1, we exploit long enough such that the overall average reward is essentially the average reward in the last exploitation period.

Lemma 4.8 *At any time T after phase t, the average reward from time 1 to time T satisfies: $\frac{1}{T}\sum_{i=1}^{T} r_i \geq R_t^*(B_H) - O(t\epsilon_t + \frac{|A|t}{L^t} + (1/t)K_H \log(1/\epsilon_t))$ with probability at least $1 - \frac{1}{t^2}$.*

[3] We can use in the algorithm any approximate homing strategy H. However, if H is simply the random policy, then the reset and exploration would both use the same policy, and the algorithm would slightly simplify.

Using the above lemma, Theorem 4.5 follows immediately if we set $\epsilon_t = 1/t^2$.

Proof: (sketch) We first note that all exploitations and explorations from phases 1 to T and from the next, $(T+1)$-th, exploration phase can effect the average reward by at most ϵ_t. By Lemma 4.7, the exploitation policy is near optimal and satisfies, $|R_t^*(B_H) - R_t^{\hat{\pi}_t^*}(B_H)| \leq t(\epsilon_t + \frac{2|A|}{L^t}) + (\epsilon_t + \frac{\epsilon_t}{L^t})$ with probability $1 - 1/(2t^2)$. As in Lemma 4.2, we observe that the bias of the exploitation policy is ϵ_t and the variance due to Hoeffding's bound and the large exploitation time is bounded by ϵ_t with probability $1 - 1/(2t^2)$. The last source for loss is the resets in the exploitation period and its effect can be bounded by $\frac{K_H \log(1/\epsilon_t)}{t}$. □

A direct and simple corollary from which Theorem 4.1 follows as well.

Corollary 4.9 *Let $\epsilon_t = 1/t$. Then $\frac{1}{T}\sum_{i=1}^{T} r_i \geq R_t^*(B_H) - O(\frac{K_H \log(t)}{t})$ with probability at least $1 - \frac{1}{t^2}$.*

Acknowledgements

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and by a grant from the Israel Science Foundation and an IBM faculty award. This publication only reflects the authors' views.

References

X. Boyen and D. Koller. Tractable inference for complex stochastic processes. *In UAI*, 1998.

A. Cassandra. *Exact and approximate algorithms for partially observable markov decision processes*. PhD thesis, Brown University, 1998.

M. Hauskrecht. A heuristic variable-grid solution method for pomdps. In *AAAI-97*, pages 734–739, 1997.

M. Kearns, Y. Mansour, and A. Ng. Approximate planning in large pomdps via reusable trajectories. In *NIPS 12*, 1999.

M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Proceedings of ICML*, 1998.

W. S. Lovejoy. Computationally feasible bounds for partially observed markov decision processes. *Operations Research*, 39(1):162–175, 1991a.

W. S. Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28:47–66, 1991b.

C. Lusena, J. Goldsmith, and M. Mundhenk. Nonapproximability results for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 14: 83–103, 2001.

D. McAllester and S. Singh. Approximate planning for factored pomdps using belief state simplification. In *In Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 409–416, 1999.

R. Parr and S. Russell. Approximating optimal policies for partially observable stochastic domains. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1995.

L. Peshkin, K. Kim, N. Meuleau, and L.P. Kaelbling. Learning to cooperate via policy search. In *16th Proceedings of UAI*, pages 307–314, 2000.

R. Rivest and R. Schapire. Inference of finite automata using homing sequences. *Information and Computation*, 103(2): 299–347, 1993.

E. Sondik. *The optimal control of partially observable processes over a finite horizon*. PhD thesis, Stanford University, Stanford, California, 1971.