Optimal Regret Bounds for Selecting the State Representation in Reinforcement Learning

Odalric-Ambrym Maillard[1] ODALRICAMBRYM.MAILLARD@GMAIL.COM
The Technion, Faculty of Electrical Engineering, 32000 Haifa, ISRAEL

Phuong Nguyen NMPHUONG@CECS.ANU.EDU.AU
Australian National University and NICTA, Canberra ACT 0200, AUSTRALIA

Ronald Ortner RORTNER@UNILEOBEN.AC.AT
Montanuniversität Leoben, Franz-Josef-Strasse 18, A-8700 Leoben, AUSTRIA

Daniil Ryabko DANIIL@RYABKO.NET
INRIA Lille - Nord Europe, 40 Avenue Halley, 59650 Villeneuve d'Ascq, FRANCE

Abstract

We consider an agent interacting with an environment in a single stream of actions, observations, and rewards, with no reset. This process is not assumed to be a Markov Decision Process (MDP). Rather, the agent has several representations (mapping histories of past interactions to a discrete state space) of the environment with unknown dynamics, only some of which result in an MDP. The goal is to minimize the average regret criterion against an agent who knows an MDP representation giving the highest optimal reward, and acts optimally in it. Recent regret bounds for this setting are of order $O(T^{2/3})$ with an additive term constant yet exponential in some characteristics of the optimal MDP. We propose an algorithm whose regret after T time steps is $O(\sqrt{T})$, with all constants reasonably small. This is optimal in T since $O(\sqrt{T})$ is the optimal regret in the setting of learning in a (single discrete) MDP.

1. Introduction

In Reinforcement Learning (RL), an agent has to learn a task through interactions with the environment. The standard RL framework models the interaction of the agent and the environment as a finite-state Markov decision process (MDP). Unfortunately, the real world is not (always) a finite-state MDP, and the learner often has to find a suitable *state-representation model*: a function that maps histories of actions, observations, and rewards provided by the environment into a finite space of *states*, in such a way that the resulting process on the state space is Markovian, reducing the problem to learning in a finite-state MDP. However, finding such a model is highly non-trivial. One can come up with several representation models, many of which may lead to non-Markovian dynamics. Testing which one has the MDP property one by one may be very costly or even impossible, as testing a statistical hypothesis requires a workable alternative assumption on the environment. This poses a challenging problem: find a generic algorithm that, given several state-representation models only some of which result in an MDP, gets (on average) at least as much reward as an optimal policy for any of the Markovian representations. Here we do not test the MDP property but propose to use models as long as they provide high enough rewards.

Motivation. One can think of specific scenarios where the setting of several state-representation models is applicable. First, these models can be discretisations of a continuous state space. Second, they may be discretisations of the parameter space: this scenario has been recently considered (Ortner & Ryabko, 2012) for learning in a continuous-state MDP with Lipschitz continuous rewards and transition probabilities where the Lipschitz constants are unknown; the models are discretisations of the parameter space. A

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

[1] This work has been done at Montanuniversität Leoben.

simple example is when the process is a second-order Markov process with discrete observations: in this case a model that maps any history to the last two observations is a Markov model; a detailed illustration of such an example can be found, e.g., in Section 4 of (Hutter, 2009). More generally, one can try and extract some high-level discrete features from (continuous, high-dimensional) observations provided by the environment. For example, the observation is a video input capturing a game board, different maps attempt to extract the (discrete) state of the game, and we assume that at least one map is correct. Some popular classes of models are context trees (McCallum, 1996), which are used to capture short-term memories, or probabilistic deterministic finite automata (Vidal et al., 2005), a very general class of models that can capture both short-term and long-term memories. Since only some of the features may exhibit Markovian dynamics and/or be relevant, we want an algorithm able to exploit whatever is Markovian and relevant for learning. For more details and further examples we refer to (Maillard et al., 2011).

Previous work. This work falls under the framework of providing performance guarantees on the average reward of a considered algorithm. In this setting, the optimal regret of a learning algorithm in a finite-state MDP is $O(\sqrt{T})$. This is the regret of `UCRL2` (Jaksch et al., 2010) and `Regal.D` (Bartlett & Tewari, 2009). Previous work on this problem in the RL literature includes (Kearns & Singh, 2002; Brafman & Tennenholtz, 2003; Strehl et al., 2006). Moreover, there is currently a big interest in finding practical state representations for the general RL problem where the environment's states and model are both unknown, e.g. U-trees (McCallum, 1996), MC-AIXI-CTW (Veness et al., 2011), ΦMDP (Hutter, 2009), and PSRs (Singh et al., 2004). Another approach in which possible models are known but need not be MDPs was considered in (Ryabko& Hutter, 2008).

For the problem considered in this paper, (Maillard et al., 2011) recently introduced the `BLB` algorithm that, given a finite set Φ of state-representation models, achieves regret of order $\sqrt{|\Phi|}T^{2/3}$ (where $|\Phi|$ is the number of models) in respect to the optimal policy associated with any model that is Markovian. `BLB` is based on uniform exploration of all representation models and uses the performance guarantees of `UCRL2` to control the amount of time spent on non-Markov models. It also makes use of some internal function in order to guess the MDP *diameter* (Jaksch et al., 2010) of a Markov model, which leads to an additive term in the regret bound that may be exponential in the true diameter, which means the order $T^{2/3}$ is only valid for possibly very large T.

Contribution. We propose a new algorithm called `OMS` (Optimistic Model Selection), that has regret of order $\sqrt{|\Phi|T}$, thus establishing performance that is optimal in terms of T, without suffering from an unfavorable additive term in the bound and without compromising the dependence on $|\Phi|$. This demonstrates that taking into consideration several possibly non-Markovian representation models does not significantly degrade the performance of an algorithm, as compared to knowing in advance which model is the right one. The proposed algorithm is close in spirit to the `BLB` algorithm. However, instead of uniform exploration it uses the principle of "optimism" for model selection, choosing the model promising the best performance.

Outline. Section 2 introduces the setting; Section 3 presents our algorithm `OMS`; its performance is analysed in Section 4; proofs are in Sections 5, and Section 6 concludes.

2. Setting

Environment. For each time step $t = 1, 2, \ldots$, let $\mathcal{H}_t := \mathcal{O} \times (\mathcal{A} \times \mathcal{R} \times \mathcal{O})^{t-1}$ be the set of histories up to time t, where $\mathcal{O}$ is the set of observations, $\mathcal{A}$ is a finite set of actions and $\mathcal{R} = [0, 1]$ is the set of possible rewards. We consider the problem of reinforcement learning when the learner interacts sequentially with some *unknown* environment: first some initial observation $h_1 = o_1 \in \mathcal{H}_1 = \mathcal{O}$ is provided to the learner, then at any time step $t > 0$, the learner chooses an action $a_t \in \mathcal{A}$ based on the current history $h_t \in \mathcal{H}_t$, then receives the immediate reward r_t and the next observation o_{t+1} from the environment. Thus, h_{t+1} is the concatenation of h_t with (a_t, r_t, o_{t+1}).

State representation models. Let Φ be a set of state-representation models. A *state-representation* model $\phi \in \Phi$ is a function from the set of histories $\mathcal{H} := \bigcup_{t \geqslant 1} \mathcal{H}_t$ to a finite set of states $\mathcal{S}_\phi$. For a model ϕ, the state at step t under ϕ is denoted by $s_{t,\phi} := \phi(h_t)$ or simply s_t when ϕ is clear from context. For the sake of simplicity, we assume that $\mathcal{S}_\phi \cap \mathcal{S}_{\phi'} = \emptyset$ for $\phi \neq \phi'$. Further, we set $\mathcal{S} := \bigcup_{\phi \in \Phi} \mathcal{S}_\phi$.

A particular role will be played by state-representation models that induce a *Markov decision process (MDP)*. An MDP is defined as a decision process in which at any discrete time t, given action a_t, the probability of immediate reward r_t and next observation o_{t+1}, given the past history h_t, only depends on the current observation o_t. That is, $P(o_{t+1}, r_t | h_t a_t) = P(o_{t+1}, r_t | o_t, a_t)$. Observations in this process are called *states* of the environment. We say that a state-representation model ϕ

is a *Markov model* of the environment, if the process $(s_{t,\phi}, a_t, r_t), t \in \mathbb{N}$ is an MDP. This MDP is denoted as $M(\phi)$. We will always assume that such MDPs are *weakly communicating*, that is, for each pair of states x_1, x_2 there exists $k \in \mathbb{N}$ and a sequence of actions $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$ such that $P(s_{k+1,\phi} = x_2 | s_{1,\phi} = x_1, a_1 = \alpha_1, \ldots, a_k = \alpha_k) > 0$. It should be noted that there may be infinitely many state-representation models under which an environment is Markov.

Problem description. Given a finite set Φ which includes at least one Markov model, we want to construct a strategy that performs as well as the algorithm that knows any Markov model $\phi \in \Phi$, including its rewards and transition probabilities. For that purpose we define for any Markov model $\phi \in \Phi$ the regret of any strategy at time T, cf. (Jaksch et al., 2010; Bartlett & Tewari, 2009; Maillard et al., 2011), as

$$\Delta(\phi, T) := T\rho^\star(\phi) - \sum_{t=1}^{T} r_t \,,$$

where r_t are the rewards received when following the proposed strategy and $\rho^\star(\phi)$ is the optimal average reward in ϕ, i.e., $\rho^\star(\phi) := \rho(M(\phi), \pi^\star_\phi) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} r_t(\pi^\star_\phi)\big]$ where $r_t(\pi^\star_\phi)$ are the rewards received when following the optimal policy $\pi^\star_\phi$ for ϕ. Note that for weakly communicating MDPs the optimal average reward indeed does not depend on the initial state. One could replace $T\rho^\star(\phi)$ with the expected sum of rewards obtained in T steps (following the optimal policy) at the price of an additional $O(\sqrt{T})$ term.

3. Algorithm

High-level overview. The `OMS` algorithm we propose (shown in detail as Algorithm 1) proceeds in episodes $k = 1, 2, \ldots$, each consisting of several runs $j = 1, 2, \ldots$. In each run j of some episode k, starting at time $t = t_{k,j}$, `OMS` chooses a policy $\pi_{k,j}$ applying the optimism in face of uncertainty principle twice. First, in line 6, `OMS` considers for each model $\phi \in \Phi$ a set of admissible MDPs $\mathcal{M}_{t,\phi}$ (defined via confidence intervals for the estimates so far), and computes a so-called optimistic MDP $M^+_t(\phi) \in \mathcal{M}_{t,\phi}$ and an associated optimal policy $\pi^+_t(\phi)$ on $M^+_t(\phi)$ such that the average reward $\rho(M^+_t(\phi), \pi^+_t(\phi))$ is maximized. Then (line 7) `OMS` chooses the model $\phi_{k,j} \in \Phi$ which maximizes the average reward $\pi_{k,j} := \pi^+_t(\phi_{k,j})$ penalized by a term intuitively accounting for the "complexity" of the model, similar to the `REGAL` algorithm of (Bartlett & Tewari, 2009).

The policy $\pi_{k,j}$ is then executed until either (i) run j reaches the maximal length of 2^j steps (line 19), (ii) episode k terminates when the number of visits in some state has been doubled (line 17), or (iii) the executed policy $\pi_{k,j}$ does not give sufficiently high average reward (line 12). Note that `OMS` assumes each model to be Markov, as long as it performs well. Otherwise the model is eliminated (line 15).

Algorithm 1 Optimistic Model Selection (`OMS`)

Require: Set of models Φ_0, parameter $\delta \in [0, 1]$.

1: Set $t := 1$, $k := 0$, and $\Phi := \Phi_0$.
2: **while true do**
3: $k := k + 1$, $j := 1$, sameEpisode := **true**
4: **while** sameEpisode **do**
5: $t_{k,j} := t$
6: $\forall \phi \in \Phi$, use EVI to compute optimistic MDP $M^+_t(\phi) \in \mathcal{M}_{t,\phi}$ and (near-)optimal policy $\pi^+_t(\phi)$ with approximate optimistic average reward $\widehat{\rho}^+_{t_{k,j}}(\phi)$.
7: Choose model $\phi_{k,j} \in \Phi$ such that

$$\phi_{k,j} = \operatorname*{argmax}_{\phi\in\Phi} \Big\{ \widehat{\rho}^+_{t_{k,j}}(\phi) - \mathbf{pen}(\phi; t_{k,j}) \Big\} . \quad (1)$$

8: Define $\rho_{k,j} := \widehat{\rho}^+_{t_{k,j}}(\phi_{k,j}), \pi_{k,j} := \pi^+_{t_{k,j}}(\phi_{k,j})$.
9: sameRun := **true**.
10: **while** sameRun **do**
11: Choose action $a_t := \pi_{k,j}(s_t)$, get reward r_t, observe next state $s_{t+1} \in \mathcal{S}_{k,j} := \mathcal{S}_{\phi_{k,j}}$.
12: Set testFail := **true** iff the sum of the collected rewards so far from time $t_{k,j}$ is less than

$$\ell_{k,j}\rho_{k,j} - \mathbf{lob}_{k,j}(t), \quad (2)$$

where $\ell_{k,j} := t - t_{k,j} + 1$.
13: **if** testFail **then**
14: sameRun := **false**, sameEpisode := **false**
15: $\Phi := \Phi \setminus \{\phi_{k,j}\}$
16: **if** $\Phi = \emptyset$ **then** $\Phi := \Phi_0$ **end if**
17: **else if** $v_k(s_t, a_t) = N_{t_k}(s_t, a_t)$ **then**
18: sameRun := **false**, sameEpisode := **false**
19: **else if** $\ell_{k,j} = 2^j$ **then**
20: sameRun := **false**, $j := j + 1$
21: **end if**
22: $t := t + 1$
23: **end while**
24: **end while**
25: **end while**

Details. We continue with some details of the algorithm. In the following, $S_\phi := |\mathcal{S}_\phi|$ denotes the number of states under model ϕ, $S := |\mathcal{S}|$ is the total number of states, and $A := |\mathcal{A}|$ is the number of actions. Further, $\delta_t := \delta/36t^2$ is the confidence parameter for time t.

Admissible models. First, the set of *admissible*

MDPs $\mathcal{M}_{t,\phi}$ the algorithm considers at time t for each model $\phi \in \Phi$ is defined to contain all MDPs with state space $\mathcal{S}_\phi$ and with rewards r and transition probabilities p satisfying

$$\left\|p(\cdot|s,a) - \widehat{p}_t(\cdot|s,a)\right\|_1 \;\leqslant\; \sqrt{\tfrac{2\log(2^{S_\phi} S_\phi A t/\delta_t)}{N_t(s,a)}}\,, \quad (3)$$

$$\left|r(s,a) - \widehat{r}_t(s,a)\right| \;\leqslant\; \sqrt{\tfrac{\log(2S_\phi A t/\delta_t)}{2N_t(s,a)}}\,, \quad (4)$$

where $\widehat{p}_t(\cdot|s,a)$ and $\widehat{r}_t(s,a)$ are respectively the empirical transition probabilities and mean rewards (at time t) for taking action a at state s, and $N_t(s,a)$ is the number of times action a has been chosen in state s up to time t. (If a hasn't been chosen in s so far, we set $N_t(s,a)$ to 1.) It can be shown (cf. Appendix C.1 of Jaksch et al. (2010)) that the mean rewards r and the transition probabilities p of a Markovian state-representation ϕ satisfy (3) and (4) at time t for all $s \in \mathcal{S}_\phi$ and $a \in \mathcal{A}$, each with probability at least $1 - \delta_t$, making Markov models admissible with high probability.

Extended Value Iteration. For computing a near-optimal policy $\pi_t^+(\phi)$ and a corresponding optimistic MDP $M_t^+(\phi) \in \mathcal{M}_{t,\phi}$ (line 6), `OMS` applies for each $\phi \in \Phi$ extended value iteration (EVI) (Jaksch et al., 2010) with precision parameter $t^{-1/2}$. EVI computes optimistic approximate state values $\mathbf{u}_{t,\phi}^+ = (u_{t,\phi}^+(s))_s \in \mathbb{R}^{S_\phi}$ just like ordinary value iteration (Puterman, 1994) with an additional optimization step for choosing the transition kernel maximizing the average reward. The (approximate) average reward $\widehat{\rho}_t^+(\phi)$ of $\pi_t^+(\phi)$ in $M_t^+(\phi)$ then is given by

$$\widehat{\rho}_t^+(\phi) = \min\Big\{ r_t^+(s,\pi_t^+(\phi,s)) + \sum_{s'} p_t^+(s'|s)\, u_{t,\phi}^+(s') - u_{t,\phi}^+(s), s \in \mathcal{S}_\phi \Big\}\,, \quad (5)$$

where r_t^+ and p_t^+ are the rewards and transition probabilities of $M_t^+(\phi)$ under $\pi_t^+(\phi)$. It can be shown (Jaksch et al., 2010) that $\widehat{\rho}_t^+(\phi) \geqslant \rho^\star(\phi) - 2/\sqrt{t}$.

Penalization term. At time $t = t_{k,j}$, we define the empirical value span of the optimistic MDP $M_t^+(\phi)$ as $\mathbf{sp}(\mathbf{u}_{t,\phi}^+) := \max_{s\in\mathcal{S}_\phi} u_{t,\phi}^+(s) - \min_{s\in\mathcal{S}_\phi} u_{t,\phi}^+(s)$, and the penalization term considered in (1) for each model ϕ is given by

$$\mathbf{pen}(\phi;t) \;:=\; 2^{-j/2}\, c(\phi;t)\, \mathbf{sp}(\mathbf{u}_{t,\phi}^+) + 2^{-j/2}\, c'(\phi;t) + 2^{-j}\, \mathbf{sp}(\mathbf{u}_{t,\phi}^+),$$

where the constants are given by

$$c(\phi;t) \;:=\; 2\sqrt{2S_\phi A \log(2^{S_\phi} S_\phi A t/\delta_t)} + 2\sqrt{2\log(\tfrac{1}{\delta_t})},$$

$$c'(\phi;t) \;:=\; 2\sqrt{2S_\phi A \log(2S_\phi A t/\delta_t)}\,.$$

Deviation from the optimal reward. Let $\ell_{k,j} := t - t_{k,j} + 1$, and $v_{k,j}(s,a)$ be the total number of times a has been played in s during run j in episode k (or until current time t if j is the current run). Similarly, we write $v_k(s,a)$ for the respective total number of visits during episode k. (Note that by the assumption $\mathcal{S}_\phi \cap \mathcal{S}_{\phi'} = \emptyset$ for $\phi \neq \phi'$, the state implicitly determines the respective model.) Then for the test (2) that decides whether the chosen model $\phi_{k,j}$ gives sufficiently high reward, we define the allowed deviation from the optimal average reward in the optimistic model for any $t \geqslant t_{k,j}$ in run j as

$$\begin{aligned}\mathbf{lob}_{k,j}(t) := {} & 2\sum_{s\in\mathcal{S}_{k,j}}\sum_{a\in\mathcal{A}} \sqrt{2v_{k,j}(s,a)\log\big(\tfrac{2S_{k,j}At_{k,j}}{\delta_{t_{k,j}}}\big)} \\ & 2\mathbf{sp}_{k,j}^+ \sum_{s\in\mathcal{S}_{k,j}}\sum_{a\in\mathcal{A}} \sqrt{2v_{k,j}(s,a)\log\big(\tfrac{2^{S_{k,j}}S_{k,j}At_{k,j}}{\delta_{t_{k,j}}}\big)} \\ & + 2\mathbf{sp}_{k,j}^+\sqrt{2\ell_{k,j}\log(1/\delta_{t_{k,j}})} + \mathbf{sp}_{k,j}^+\,, \end{aligned} \quad (6)$$

where $\mathbf{sp}_{k,j}^+ := \mathbf{sp}(\mathbf{u}_{t_{k,j},\phi_{k,j}}^+)$ and $S_{k,j} := S_{\phi_{k,j}}$. Intuitively, the first two terms correspond to the estimation error of the transition kernel and the rewards, while the last one is due to stochasticity of the sampling process.

4. Main result

We now provide the main result of this paper, an upper bound on the regret of our `OMS` strategy. The bound involves the *diameter* of a Markov model ϕ, $D(\phi)$, which is defined as the expected minimum time required to reach any state starting from any other state in the MDP $M(\phi)$ (Jaksch et al., 2010).

Theorem 1 *Let $\phi^\star$ be an optimal model, i.e. $\phi^\star \in \operatorname{argmax}\{\rho^\star(\phi) \,|\, \phi \in \Phi, \phi \text{ is Markovian}\}$. Then the regret $\Delta(\phi^\star, T)$ of `OMS` (with parameter δ) w.r.t. $\phi^\star$ after any $T \geqslant SA$ steps is upper bounded by*

$$\begin{aligned}&\big(8D^\star S^\star + 4\sqrt{S^\star}\big)\sqrt{A\log\big(\tfrac{48S^\star AT^3}{\delta}\big)}\log\big(\tfrac{2T}{SA}\big) \\ &\times\Big(\sqrt{\big(AS+|\Phi|\big)T} + \big(AS+|\Phi|\big)\log\big(\tfrac{2T}{SA}\big)\Big) \\ &\qquad + \big(\rho^\star + D^\star\big)\big(AS+|\Phi|\big)\log^2\big(\tfrac{2T}{SA}\big)\end{aligned}$$

with probability higher than $1-\delta$, where $\rho^\star := \rho^\star(\phi^\star)$, $S^\star := S_{\phi^\star}$, and $D^\star := D(\phi^\star)$.

In particular, if for all $\phi \in \Phi$, $S_\phi \leqslant B$, then $S \leqslant B|\Phi|$ and hence with high probability

$$\Delta(\phi^\star, T) = \tilde{O}\big(D^\star AB^{3/2}\sqrt{|\Phi|T}\big)\,.$$

Comparison with the `BLB` algorithm. Compared to the results obtained by (Maillard et al., 2011) the

regret bound in Theorem 1 has improved dependence of $T^{1/2}$ (instead of $T^{2/3}$) with respect to the horizon (up to logarithmic factors). Moreover, the new bound avoids a possibly large constant for guessing the diameter of the MDP representation, as unlike BLB, the current algorithm does not need to know the diameter. These improvements were possible since unlike BLB (which uses uniform exploration over all models, and applies UCRL2 as a "black box") we employ optimistic exploration of the models, and do a more in-depth analysis of the "UCRL2 part" of our algorithm.

On the other hand, we lose in lesser parameters: the multiplicative term in the new bound is $S^\star A\sqrt{S} \leqslant S^\star A\sqrt{|\Phi|B}$ (assuming that all representations induce a model with no more than $S_\phi \leqslant B$ states), whereas the corresponding factor in the bound of (Maillard et al., 2011) is $S^\star\sqrt{A|\Phi|}$. Thus, we currently lose a factor $\sqrt{AB}$. Improving on the dependency on the state spaces is an interesting question: one may note that the algorithm actually only chooses models not much more complex (in terms of the diameter and the state space) than the best model. However, it is not easy to quantify this in terms of a concrete bound.

Another interesting question is how to reuse the information gained on one model for evaluation of the others. Indeed, if we are able to propagate information to all models, a $\log(|\Phi|)$ dependency as opposed to the current $\sqrt{|\Phi|}$ seems plausible. However, in the current formulation, a policy can be completely uninformative for the evaluation of other policies in other models. In general, this heavily depends on the internal structure of the models in Φ. If all models induce state spaces that have strictly no point in common, then it seems hard or impossible to improve on $\sqrt{|\Phi|}$.

We also note that it is possible to replace the diameter in Theorem 1 with the span of the optimal bias vector just as for the REGAL algorithm (Bartlett & Tewari, 2009) by suitably modifying the OMS algorithm. However, unlike UCRL2 and OMS for which computation of optimistic model and respective (near-)optimal policy can be performed by EVI, this modified algorithm (as REGAL) relies on finding the solution to a constraint optimization problem, efficient computation of which is still an open problem.

5. Regret analysis of the OMS strategy

The proof of Theorem 1 is divided into two parts. In Section 5.1, we first show that with high probability all Markovian state-representation models will collect sufficiently high reward according to the test in (2). This also means that the regret of any Markov model is not too large. This in turn is used in Section 5.2 to show that also the optimistic model employed by OMS (which is not necessarily Markov) does not lose too much with respect to an optimal policy in an arbitrary Markov model. In our proof we use analysis similar to (Jaksch et al., 2010) and (Bartlett & Tewari, 2009).

5.1. Markov models pass the test in (2)

Assume that $\phi_{k,j} \in \Phi$ is a Markov model. We are going to show that $\phi_{k,j}$ will pass the test on the collected rewards in (2) of the algorithm at any step t w.h.p.

Initial decomposition. First note that at time t when the test is performed, we have $\sum_{s\in\mathcal{S}_{k,j}}\sum_{a\in\mathcal{A}} v_{k,j}(s,a) = \ell_{k,j} = t - t_{k,j} + 1$, so that

$$\ell_{k,j}\rho_{k,j} - \sum_{\tau=t_{k,j}}^{t} r_\tau = \sum_{s\in\mathcal{S}_{k,j}}\sum_{a\in\mathcal{A}} v_{k,j}(s,a)\Big(\rho_{k,j} - \widehat{r}_{t_{k,j}:t}(s,a)\Big), \quad (7)$$

where $\widehat{r}_{t_{k,j}:t}(s,a)$ is the empirical average reward collected for choosing a in s from time $t_{k,j}$ to the current time t in run j of episode k. Let $r^+_{k,j}(s,a)$ be the optimistic rewards of the model $M^+_{t_{k,j}}(\phi_{k,j})$ under policy $\pi_{k,j}$ and $\mathbf{P}^+_{k,j}$ the respective optimistic transition matrix. Set $\mathbf{v}_{k,j} := (v_{k,j}(s,\pi_{k,j}(s)))_s \in \mathbb{R}^{S_{k,j}}$, and let $\mathbf{u}^+_{k,j} := (u^+_{t_{k,j},\phi_{k,j}}(s))_s \in \mathbb{R}^{S_{k,j}}$ be the state value vector given by EVI. By (5) and noting that $v_{k,j}(s,a) = 0$ when $a \neq \pi_{k,j}(s)$ or $s \notin \mathcal{S}_{k,j}$, we get

$$\begin{aligned}\ell_{k,j}\rho_{k,j} - \sum_{\tau=t_{k,j}}^{t} r_\tau &= \sum_{s,a} v_{k,j}(s,a)\big(\widehat{\rho}^+_{k,j}(\phi_{k,j}) - r^+_{k,j}(s,a)\big) \\ &\quad + \sum_{s,a} v_{k,j}(s,a)\big(r^+_{k,j}(s,a) - \widehat{r}_{t_{k,j}:t}(s,a)\big) \\ &\leqslant \mathbf{v}^\top_{k,j}\big(\mathbf{P}^+_{k,j} - I\big)\mathbf{u}^+_{k,j} \\ &\quad + \sum_{s,a} v_{k,j}(s,a)\Big(r^+_{k,j}(s,a) - \widehat{r}_{t_{k,j}:t}(s,a)\Big). \quad (8)\end{aligned}$$

We continue bounding each of the two terms on the right hand side of (8) separately.

Control of the second term. Writing $r(s,a)$ for the mean reward for choosing a in s (this is well-defined, since we assume the model is Markov), we have

$$\begin{aligned}r^+_{k,j}(s,a) - \widehat{r}_{t_{k,j}:t}(s,a) &= \big(r^+_{k,j}(s,a) - \widehat{r}_{t_{k,j}}(s,a)\big) \\ &\quad + \big(\widehat{r}_{t_{k,j}}(s,a) - r(s,a)\big) + \big(r(s,a) - \widehat{r}_{t_{k,j}:t}(s,a)\big).\end{aligned}$$

The terms of this decomposition are controlled. That is, using that $M(\phi_{k,j})$ is an admissible model according

to (4) with probability $1-\delta_{t_{k,j}}$ (by applying the results of measure concentration in Appendix C.1 of (Jaksch et al., 2010) to the quantity $\widehat{r}_{t_{k,j}}(s,a)$), and the mere definition of $r^+_{k,j}(s,a)$, and since $N_{t_k}(s,a) \leqslant N_t(s,a)$, we deduce that with probability higher than $1-\delta_{t_{k,j}}$,

$$\begin{aligned}\sum_{s,a} v_{k,j}(s,a)\Big(&\big(r^+_{k,j}(s,a) - \widehat{r}_{t_{k,j}}(s,a)\big) \\ &+\big(\widehat{r}_{t_{k,j}}(s,a) - r(s,a)\big)\Big) \\ \leqslant\ & 2\sum_{s,a}\frac{v_{k,j}(s,a)}{\sqrt{2N_{t_k}(s,a)}}\sqrt{\log\left(\tfrac{2S_{k,j}At_{k,j}}{\delta_{t_{k,j}}}\right)} \\ \leqslant\ & \sum_{s,a}\sqrt{2v_{k,j}(s,a)\log\left(\tfrac{2S_{k,j}At_{k,j}}{\delta_{t_{k,j}}}\right)}. \end{aligned} \tag{9}$$

On the other hand, using again the results of measure concentration in Appendix C.1 of (Jaksch et al., 2010), and that $v_{k,j}(s,a) \leqslant N_{t_k}(s,a) \leqslant t_{k,j}$, we deduce by a union bound over $S_{k,j}At_{k,j}$ events that with probability higher than $1-\delta_{t_{k,j}}$ we get

$$\begin{aligned}\sum_{s,a} v_{k,j}(s,a)\Big(r(s,a) - \widehat{r}_{t_{k,j}:t}(s,a)\Big) &\leqslant \sum_{s,a}\frac{v_{k,j}(s,a)}{\sqrt{2v_{k,j}(s,a)}}\sqrt{\log\left(\tfrac{2S_{k,j}At_{k,j}}{\delta_{t_{k,j}}}\right)} \\ &\leqslant \sum_{s,a}\sqrt{2v_{k,j}(s,a)\log\left(\tfrac{2S_{k,j}At_{k,j}}{\delta_{t_{k,j}}}\right)}. \end{aligned} \tag{10}$$

Control of the first term. For the first term in (8), let us first notice that, since the rows of $\mathbf{P}^+_{k,j}$ sum to 1, $\big(\mathbf{P}^+_{k,j} - I\big)\mathbf{u}^+_{k,j}$ is invariant under a translation of the vector $\mathbf{u}^+_{k,j}$. In particular, we can replace $\mathbf{u}^+_{k,j}$ with the quantity $\mathbf{h}^+_{k,j}$, where

$$h^+_{k,j}(s) := u^+_{k,j}(s) - \min\big\{\, u^+_{k,j}(s) \,|\, s \in \mathcal{S}_{k,j} \,\big\}.$$

Then, we make use of the decomposition

$$\begin{aligned}&\mathbf{v}^\top_{k,j}\big(\mathbf{P}^+_{k,j} - I\big)\mathbf{u}^+_{k,j} = \\ &\quad \mathbf{v}^\top_{k,j}\big(\mathbf{P}^+_{k,j} - \mathbf{P}_{k,j}\big)\mathbf{h}^+_{k,j} + \ \mathbf{v}^\top_{k,j}\big(\mathbf{P}_{k,j} - I\big)\mathbf{h}^+_{k,j}, \end{aligned} \tag{11}$$

where $\mathbf{P}_{k,j}$ denotes the transition matrix corresponding to the MDP $M(\phi_{k,j})$ under policy $\pi_{k,j}$. Since both matrices are close to the empirical transition matrix $\widehat{\mathbf{P}}_{t_{k,j}}$ at time $t_{k,j}$, we can control the first term of this expression.

First part of the first term. Indeed, since $\mathbf{sp}^+_{k,j} = \|\mathbf{h}^+_{k,j}\|_\infty$, we have for the first term in (11), using the decomposition $p^+_{k,j}(\cdot|s) - p_{k,j}(\cdot|s) = \big(p^+_{k,j}(\cdot|s) - \widehat{p}_{t_{k,j}}(\cdot|s)\big) + \big(\widehat{p}_{t_{k,j}}(\cdot|s) - p_{k,j}(\cdot|s)\big)$ together with a concentration result and the definition of $p^+_{k,j}$, that with probability higher than $1-\delta_{t_{k,j}}$

$$\begin{aligned}&\mathbf{v}^\top_{k,j}\big(\mathbf{P}^+_{k,j} - \mathbf{P}_{k,j}\big)\mathbf{h}^+_{k,j} \\ &\quad= \sum_{s,a,s'} v_{k,j}(s,a)\Big(p^+_{k,j}(s'|s) - p_{k,j}(s'|s)\Big)h^+_{k,j}(s') \\ &\quad\leqslant \sum_{s,a} v_{k,j}(s,a)\left\|p^+_{k,j}(\cdot|s) - p_{k,j}(\cdot|s)\right\|_1 \cdot \left\|\mathbf{h}^+_{k,j}\right\|_\infty \\ &\quad\leqslant \sum_{s,a} 2\, v_{k,j}(s,a)\sqrt{\tfrac{2\log(2^{S_{k,j}}S_{k,j}At_{k,j}/\delta_{t_{k,j}})}{N_{t_k}(s,a)}}\left\|\mathbf{h}^+_{k,j}\right\|_\infty \\ &\quad\leqslant 2\,\mathbf{sp}^+_{k,j}\sum_{s,a}\sqrt{2v_{k,j}(s,a)\log\left(\tfrac{2^{S_{k,j}}S_{k,j}At_{k,j}}{\delta_{t_{k,j}}}\right)}. \end{aligned} \tag{12}$$

Second part of the first term. The second term of (11) can be rewritten using a martingale difference sequence. That is, let $\mathbf{e}_s \in \mathbb{R}^{S_{k,j}}$ be the unit vector with coordinates 0 for all $s' \neq s$. Following (Jaksch et al., 2010) we set $X_\tau := \big(p(\cdot|s_\tau,a_\tau) - \mathbf{e}^\top_{s_{\tau+1}}\big)\mathbf{h}^+_{k,j}$ and get

$$\begin{aligned}&\mathbf{v}^\top_{k,j}\big(\mathbf{P}_{k,j} - I\big)\mathbf{h}^+_{k,j} \\ &\quad= \sum_{\tau=t_{k,j}}^{t}\Big(p(\cdot|s_\tau,a_\tau) - \mathbf{e}^\top_{s_\tau}\Big)\,\mathbf{h}^+_{k,j} \\ &\quad= \Big(\mathbf{e}^\top_{s_{t+1}} - \mathbf{e}^\top_{s_{t_{k,j}}} + \sum_{\tau=t_{k,j}}^{t}\Big(p(\cdot|s_\tau,a_\tau) - \mathbf{e}^\top_{s_{\tau+1}}\Big)\Big)\,\mathbf{h}^+_{k,j} \\ &\quad= \sum_{\tau=t_{k,j}}^{t} X_\tau + h^+_{k,j}(s_{t+1}) - h^+_{k,j}(s_{t_{k,j}}) \\ &\quad= \sum_{\tau=t_{k,j}}^{t} X_\tau + u^+_{k,j}(s_{t+1}) - u^+_{k,j}(s_{t_{k,j}}). \end{aligned} \tag{13}$$

Now the sequence $\{X_\tau\}_{t_{k,j}\leqslant\tau\leqslant t}$ is a martingale difference sequence with

$$|X_\tau| \ \leqslant\ \left\|p(\cdot|s_\tau,a_\tau) - \mathbf{e}^\top_{s_{\tau+1}}\right\|_1 \mathbf{sp}^+_{k,j} \ \leqslant\ 2\mathbf{sp}^+_{k,j}.$$

Thus, an application of Azuma-Hoeffding's inequality (cf. Lemma 10 and its application in Jaksch et al. (2010)) to (13) yields

$$\begin{aligned}&\mathbf{v}^\top_{k,j}\big(\mathbf{P}_{k,j} - I\big)\mathbf{h}^+_{k,j} \\ &\qquad\leqslant 2\mathbf{sp}^+_{k,j}\sqrt{2\ell_{k,j}\log(1/\delta_{t_{k,j}})} + \mathbf{sp}^+_{k,j} \end{aligned} \tag{14}$$

with probability higher than $1-\delta_{t_{k,j}}$. Together with (12) this concludes the control of the first term of (8).

Putting all steps together. Combining (8), (9), (10), (11), (12), and (14), we deduce that at each time t

of run j in episode k, any Markovian model $\phi_{k,j}$ passes the test in (2) with probability higher than $1-4\delta_{t_{k,j}}$. Further, it passes all the tests in run j with probability higher than $1-4\delta_{t_{k,j}}2^j$.

5.2. Regret analysis

Next, let us consider a model $\phi_{k,j}\in\Phi$, not necessarily Markovian, that has been chosen at time $t_{k,j}$. Let $t+1$ be the time when one of the three stopping conditions in the algorithm (lines 12, 17, and 19) is met. Thus OMS employs the model $\phi_{k,j}$ between $t_{k,j}$ and $t+1$, until a new model is chosen after the step $t+1$. Noting that $r_\tau\in[0,1]$ and that the total length of the run is $(t+1)-t_{k,j}+1=\ell_{k,j}+1$ we can bound the regret $\Delta_{k,j}$ of run j in episode k by

$$\begin{aligned}\Delta_{k,j} &:= (\ell_{k,j}+1)\rho^\star - \sum_{\tau=t_{k,j}}^{t+1} r_\tau \\ &\leqslant \ell_{k,j}\big(\rho^\star-\rho_{k,j}\big)+\rho^\star+\ell_{k,j}\rho_{k,j}-\sum_{\tau=t_{k,j}}^{t} r_\tau\,.\end{aligned}$$

Since by assumption the test in (2) has been passed for all steps $\tau\in[t_{k,j},t]$, we have

$$\Delta_{k,j}\leqslant \ell_{k,j}\big(\rho^\star-\rho_{k,j}\big)+\rho^\star+\mathbf{lob}_{k,j}(t), \tag{15}$$

and we continue bounding the terms of $\mathbf{lob}_{k,j}(t)$.

Stopping criterion based on the visit counter. Since $\sum_{s,a}v_{k,j}(s,a)=\ell_{k,j}\leqslant 2^j$, by Cauchy-Schwarz inequality $\sum_{s,a}\sqrt{v_{k,j}(s,a)}\leqslant 2^{j/2}\sqrt{S_{k,j}A}$. Plugging this into the definition (6) of $\mathbf{lob}_{k,j}$, we deduce from (15) that

$$\begin{aligned}\Delta_{k,j} \leqslant\ & \ell_{k,j}\big(\rho^\star-\rho_{k,j}\big)+\rho^\star \\ &+\mathbf{sp}^+_{k,j}+2^{j/2}\mathbf{sp}^+_{k,j}c(\phi_{k,j};t_{k,j})+2^{j/2}c'(\phi_{k,j};t_{k,j})\,.\end{aligned} \tag{16}$$

Selection procedure with penalization. Now, by definition of the algorithm, for any optimal Markov model $\phi^\star$ defined in the statement of Theorem 1, whenever $M(\phi^\star)$ is admissible, i.e. $M(\phi^\star)\in\mathcal{M}_{t_{k,j},\phi^\star}$ and was not eliminated during all runs before run j in episode k, we have $\rho_{k,j}-\mathbf{pen}(\phi_{k,j};t_{k,j})\geqslant\widehat{\rho}^+_{k,j}(\phi^\star)-\mathbf{pen}(\phi^\star;t_{k,j})\geqslant\rho^\star-\mathbf{pen}(\phi^\star;t_{k,j})-2t_{k,j}^{-1/2}$, or equivalently

$$\begin{aligned}\rho^\star-\rho_{k,j}&\leqslant \mathbf{pen}(\phi^\star;t_{k,j})-\mathbf{pen}(\phi_{k,j};t_{k,j})+2t_{k,j}^{-1/2}\\ &\leqslant\ 2^{-j/2}c(\phi^\star;t_{k,j})\,\mathbf{sp}(\mathbf{u}^+_{t_{k,j},\phi^\star})\\ &\quad+2^{-j/2}c'(\phi^\star;t_{k,j})+2^{-j}\mathbf{sp}(\mathbf{u}^+_{t_{k,j},\phi^\star})\\ &\quad-2^{-j/2}c(\phi_{k,j};t_{k,j})\,\mathbf{sp}^+_{k,j}\\ &\quad-2^{-j/2}c'(\phi_{k,j};t_{k,j})-2^{-j}\mathbf{sp}^+_{k,j}+2t_{k,j}^{-1/2}.\end{aligned} \tag{17}$$

Noting that $\ell_{k,j}\leqslant 2^j$ and recalling that when $M(\phi^\star)$ is admissible, the span of the corresponding optimistic model is less than the diameter of the true model, i.e. $\mathbf{sp}(\mathbf{u}^+_{t_{k,j},\phi^\star})\leqslant D^\star$, see (Jaksch et al., 2010), and we obtain from (16), (17), and a union bound that

$$\begin{aligned}\Delta_{k,j} \leqslant\ & \rho^\star+D^\star+2^{j/2}D^\star c(\phi^\star;t_{k,j})\\ &+2^{j/2}c'(\phi^\star;t_{k,j})+2^{j+1}t_{k,j}^{-1/2}\end{aligned} \tag{18}$$

with probability higher than

$$1-\sum_{k',j';t_{k',j'}<t_{k,j}} 4\delta_{t_{k',j'}}2^{j'}-2\delta_{t_{k,j}}\,. \tag{19}$$

The sum in (19) comes from the event that $\phi^\star$ passes all tests (and is admissible) for all runs in all episodes previous to time $t_{k,j}$, and $2\delta_{t_{k,j}}$ comes from the event that $\phi^\star$ is admissible at time $t_{k,j}$. We conclude in the following by summing $\Delta_{k,j}$ over all runs and episodes.

Summing over runs and episodes. Let J_k be the total number of runs in episode k, and let K_T be the total number of episodes up to time T. Noting that $c(\phi^\star;t_{k,j})\leqslant c(\phi^\star;T)$ and $c'(\phi^\star;t_{k,j})\leqslant c'(\phi^\star;T)$ as well as using that $2t_{k,j}\geqslant 2^j$ (so that $2^{j+1}t_{kj}^{-1/2}\leqslant 2\sqrt{2}\cdot 2^{j/2}$), summing (18) over all runs and episodes gives

$$\begin{aligned}\Delta(\phi^\star,T) \;=\; \sum_{k=1}^{K_T}\sum_{j=1}^{J_k}\Delta_{k,j} &\leqslant \big(\rho^\star+D^\star\big)\sum_{k=1}^{K_T}J_k \\ &+\Big(D^\star c(\phi^\star;T)+c'(\phi^\star;T)+2\sqrt{2}\Big)\sum_{k=1}^{K_T}\sum_{j=1}^{J_k}2^{j/2},\end{aligned} \tag{20}$$

with probability higher than $1-\sum_{k=1}^{K_T}\sum_{j=1}^{J_k}4\delta_{t_{k,j}}2^j$, where we used a union bound over all events considered in (19) for the control of all the $\Delta_{k,j}$ terms, avoiding redundant counts (such as the admissibility of $\phi^\star$ at time $t_{k,j}$). Now, using the definition of $\delta_{t_{k,j}}$ and the fact that $2t_{k,j}\geqslant 2^j$, we get that

$$\begin{aligned}4\delta_{t_{k,j}}2^j=\frac{2^j\delta}{9t_{k,j}^2}&\leqslant\frac{2^j\delta}{2t_{k,j}(t_{k,j}+2^j)}\\ &=\frac{\delta}{2t_{k,j}}-\frac{\delta}{2(t_{k,j}+2^j)}\leqslant\sum_{t=t_{k,j}}^{t_{k,j}+2^j-1}\frac{\delta}{2t^2}\,,\end{aligned}$$

where the last inequality follows by a series-integral comparison, using that $t\mapsto t^{-2}$ is a decreasing function. Thus, we deduce that the bound (20) is valid with probability at least $1-\sum_{t=1}^{\infty}\frac{\delta}{2t^2}\geqslant 1-\delta$ for all T, and it remains to bound the double sum $\sum_k\sum_j 2^{j/2}$.

From the number of runs... First note that by definition of the total number of episodes K_T we have

$$T\geqslant\sum_{k=1}^{K_T}\sum_{j=1}^{J_k-1}2^j=\sum_{k=1}^{K_T}\Big(2^{J_k}-2\Big)\,, \tag{21}$$

which implies also that we have the bound

$$\sum_{k=1}^{K_T}\sum_{j=1}^{J_k} 2^j = 2\sum_{k=1}^{K_T}\Big(2^{J_k}-2\Big) + 2K_T \leqslant 2T + 2K_T.$$

Further, by Jensen's inequality we get

$$\begin{aligned}\sum_{k=1}^{K_T}\sum_{j=1}^{J_k-1} 2^{j/2} &\leqslant \sqrt{\textstyle\sum_{k=1}^{K_T} J_k}\,\sqrt{\textstyle\sum_{k=1}^{K_T}\sum_{j=1}^{J_k} 2^j} \\ &\leqslant \sqrt{\textstyle\sum_{k=1}^{K_T} J_k}\,\sqrt{2T+2K_T}\,. \quad (22)\end{aligned}$$

Now, to bound the total number of runs $\sum_{k=1}^{K_T} J_k$, using Jensen's inequality and (21), we deduce

$$\begin{aligned}\sum_{k=1}^{K_T} J_k &= \sum_{k=1}^{K_T}\log_2(2^{J_k}) \leqslant K_T\log_2\Big(\frac{1}{K_T}\sum_{k=1}^{K_T}2^{J_k}\Big) \\ &\leqslant K_T\log_2\big(\tfrac{T}{K_T}+2\big) \leqslant K_T\log_2\big(\tfrac{2T}{K_T}\big), \quad (23)\end{aligned}$$

and thus it remains to deal with K_T.

... to the number of episodes. First recall that an episode is terminated when either the number of visits in some state-action pair (s,a) has been doubled (line 17 of the algorithm) or when the test on the accumulated rewards has failed (line 12). We know that with probability at least $1-\delta$ the optimal Markov model is not eliminated from Φ, while non-Markov models failing the test are deleted from Φ. Therefore, with probability $1-\delta$ the number of episodes terminated with a model failing the test is upper bounded by $|\Phi|-1$.

Next, let us consider the number of episodes which are ended since the number of visits in some state-action pair (s,a) has been doubled. Let $K(s,a)$ be the number of episodes which ended after the number of visits in (s,a) has been doubled, and let $T(s,a)$ be the number of steps in these episodes. As it may happen that in an episode the number of visits is doubled in more than one state-action pair, we assume that $K(s,a)$ and $T(s,a)$ count only the episodes/steps where (s,a) is the first state-action pair for which this happens. It is easy to see that $K(s,a) \leqslant 1+\log_2 T(s,a) = \log_2 2T(s,a)$ for $T(s,a)>0$. Then the bound $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\log_2 2T(s,a)$ on the total number of these episodes is maximal under the constraint $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} T(s,a) = T$ when $T(s,a) = \frac{T}{SA}$ for all (s,a). This shows that the total number of episodes K_T is upper bounded by

$$K_T \leqslant SA\log_2\big(\tfrac{2T}{SA}\big) + |\Phi| - 1 \quad (24)$$

with probability $1-\delta$, provided that $T \geqslant SA$.

Putting all steps together. Combining (20), (22) and (23) we get $\Delta(\phi^\star,T) \leqslant \big(\rho^\star + D^\star\big)K_T\log_2\big(\frac{2T}{K_T}\big) + \big(D^\star c(\phi^\star;T) + c'(\phi^\star;T) + 2\sqrt{2}\big)\sqrt{2K_T\log_2\big(\frac{2T}{K_T}\big)\big(T+K_T\big)}$. Hence, by (24) and the definition of c, c', the regret of `OMS` is, with probability higher than $1-\delta$, bounded by

$$\begin{aligned}&\Delta(\phi^\star,T) \leqslant \big(\rho^\star+D^\star\big)\big(SA+|\Phi|\big)\log_2^2\big(\tfrac{2T}{SA}\big) \\ &+\Big(\,2D^\star\sqrt{2S^\star A\log\big(\tfrac{2^{S^\star}24S^\star AT^3}{\delta}\big)} + 2D^\star\sqrt{2\log\big(\tfrac{24T^2}{\delta}\big)} \\ &+2\sqrt{2S^\star A\log\big(\tfrac{48S^\star AT^3}{\delta}\big)} + 2\sqrt{2}\,\Big) \\ &\times\log_2\big(\tfrac{2T}{SA}\big)\Big(\sqrt{\big(SA+|\Phi|\big)2T} + \big(SA+|\Phi|\big)\log_2\big(\tfrac{2T}{SA}\big)\Big),\end{aligned}$$

and we may conclude the proof with some minor simplifications.

6. Outlook

The first natural question about the performance guarantees obtained is whether they are optimal. We know from the corresponding lower-bounds for learning MDPs (Jaksch et al., 2010) that the dependence on T we get for `OMS` is indeed optimal. Among other parameters, perhaps the most important one is the number of models $|\Phi|$; here we conjecture that the $\sqrt{|\Phi|}$ dependence we obtain is optimal, but this remains to be proven. Other parameters are the size of the action and state spaces for each model; here we lose with respect to the precursor `BLB` algorithm (see the remark after Theorem 1), and thus have room for improvement. It may be possible to obtain a better dependence for `OMS` at the expense of more sophisticated analysis. Note, however, that so far there are no known algorithms for learning even a single MDP that would have known optimal dependence on all these parameters.

Another important direction for future research is infinite sets Φ of models; perhaps, countably infinite sets is the natural first step, with separable — in a suitable sense — continuously-parametrized general classes of models being a foreseeable extension. A problem with the latter formulation is that one would need to formalize the notion of a model being close to a Markovian model and quantify the resulting regret.

Acknowledgments

This work was supported by the French National Research Agency (ANR-08-COSI-004 project EXPLORA), by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 270327 (CompLACS), 216886 (PASCAL2) and 306638 (SUPREL), the Nord-Pas-de-Calais Regional Council and FEDER through CPER 2007-2013, the Austrian Science Fund (FWF): J 3259-N13, the Australian Research Council Discovery Project DP120100950, and NICTA.

References

Bartlett, P. L. and Tewari, A. REGAL: A regularization based algorithm for reinforcement learning in weakly-communicating MDPs. In *UAI 2009, Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pp. 35–42, 2009.

Brafman, R.I., and Tennenholtz, M. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2003.

Hutter, M. Feature reinforcement learning: Part I. Unstructured MDPs. *Journal of General Artificial Intelligence*, 1:3–24, 2009.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, 2010.

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

Maillard, O., Munos, R., and Ryabko, D. Selecting the state-representation in reinforcement learning. In *Advances in Neural Information Processing Systems* 24: 2627–2635, 2011.

McCallum, R. A. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, Department of Computer Science, U. Rochester, 1996.

Ortner, R. and Ryabko, D. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems* 25: 1772–1780, 2012.

Puterman, M. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

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

Singh, S. P., James, M. R., and Rudary, M. R. Predictive state representations: A new theory for modeling dynamical systems. In *UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*, pp. 512–518, 2004.

Strehl, A. L., Li, L., Wiewiora, Eric, Langford, J., and Littman, M. L. PAC model-free reinforcement learning. In *Machine Learning, Proceedings of the 23rd International Conference (ICML 2006)*, pp. 881–888, 2006.

Veness, J., Ng, K. S., Hutter, M., Uther, W., and Silver, D. A Monte-Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, 40(1):95–142, 2011.

Vidal, E., Thollard, F., Higuera, C. D. L., Casacuberta, F., and Carrasco, R.C. Probabilistic finite-state machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1013–1025, 2005.