Selecting Near-Optimal Approximate State Representations in Reinforcement Learning

Ronald Ortner[1], Odalric-Ambrym Maillard[2], and Daniil Ryabko[3]

[1] Montanuniversitaet Leoben, Austria
[2] The Technion, Israel
[3] Inria Lille-Nord Europe, équipe SequeL, France, and Inria Chile
`rortner@unileoben.ac.at, odalric-ambrym.maillard@ens-cachan.org,`
`daniil@ryabko.net`

Abstract. We consider a reinforcement learning setting introduced in [6] where the learner does not have explicit access to the states of the underlying Markov decision process (MDP). Instead, she has access to several models that map histories of past interactions to states. Here we improve over known regret bounds in this setting, and more importantly generalize to the case where the models given to the learner do not contain a true model resulting in an MDP representation but only approximations of it. We also give improved error bounds for state aggregation.

1 Introduction

Inspired by [3], in [6] a reinforcement learning setting has been introduced where the learner does not have explicit information about the state space of the underlying Markov decision process (MDP). Instead, the learner has a set of *models* at her disposal that map histories (i.e., observations, chosen actions and collected rewards) to states. However, only some models give a correct MDP representation. The first regret bounds in this setting were derived in [6]. They recently have been improved in [7] and extended to infinite model sets in [8]. Here we extend and improve the results of [7] as follows. First, we do not assume anymore that the model set given to the learner contains a true model resulting in an MDP representation. Instead, models will only approximate an MDP. Second, we improve the bounds of [7] with respect to the dependence on the state space.

For discussion of potential applications and related work on learning state representations in POMDPs (like predictive state representations [4]), we refer to [6–8]. Here we only would like to mention the recent work [2] that considers a similar setting, however is mainly interested in the question whether the true model will be identified in the long run, a question we think is subordinate to that of minimizing the regret, which means fast learning of optimal behavior.

1.1 Setting

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, $\mathcal{A}$ a finite set of

actions, and $\mathcal{R} = [0,1]$ the set of possible rewards. We consider the following reinforcement learning problem: The learner receives some initial observation $h_1 = o_1 \in \mathcal{H}_1 = \mathcal{O}$. Then at any time step $t > 0$, the learner chooses an action $a_t \in \mathcal{A}$ based on the current history $h_t \in \mathcal{H}_t$, and receives an immediate reward r_t and the next observation o_{t+1} from the unknown environment. Thus, h_{t+1} is the concatenation of h_t with (a_t, r_t, o_{t+1}).

State representation models. A *state-representation model* ϕ is a function from the set of histories $\mathcal{H} = \bigcup_{t \geq 1} \mathcal{H}_t$ to a finite set of states $\mathcal{S}_\phi$. 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 i.e., $P(o_{t+1}, r_t | h_t a_t) = P(o_{t+1}, r_t | o_t, a_t)$, and this probability is also independent of 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 $(\phi(h_t), a_t, r_t), t \in \mathbb{N}$ is an MDP. Note that such an MDP representation needs not be unique. In particular, we assume that we obtain a Markov model when mapping each possible history to a unique state. Since these states are not visited more than once, this model is not very useful from the practical point of view, however. In general, an MDP is denoted as $M(\phi) = (\mathcal{S}_\phi, \mathcal{A}, r, p)$, where $r(s, a)$ is the mean reward and $p(s'|s, a)$ the probability of a transition to state $s' \in \mathcal{S}_\phi$ when choosing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}_\phi$.

We assume that there is an *underlying true* Markov model ϕ° that gives a finite and *weakly communicating* MDP, that is, for each pair of states $s, s' \in \mathcal{S}^\circ := \mathcal{S}_{\phi^\circ}$ there is a $k \in \mathbb{N}$ and a sequence of actions $a_1, \ldots, a_k \in \mathcal{A}$ such that the probability of reaching state s' when starting in state s and taking actions $a_1, \ldots, a_k$ is positive. In such a weakly communicating MDP we can define the *diameter* $D := D(\phi^\circ) := D(M(\phi^\circ))$ to be the expected minimum time it takes to reach any state starting from any other state in the MDP $M(\phi^\circ)$, cf. [5]. In finite state MDPs, the *Poisson equation* relates the average reward ρ_π of any policy π to the single step mean rewards and the transition probabilities. That is, for each *policy* π that maps states to actions, it holds that

$$\rho_\pi + \lambda_\pi(s) = r(s, \pi(s)) + \sum_{s' \in \mathcal{S}^\circ} p(s'|s, \pi(s)) \cdot \lambda_\pi(s'), \tag{1}$$

where λ_π is the so-called *bias* vector of π, which intuitively quantifies the difference in accumulated rewards when starting in different states. Accordingly, we are sometimes interested in the *span* of the bias vector λ of an optimal policy defined as $\mathrm{span}(\lambda) := \max_{s \in \mathcal{S}^\circ} \lambda(s) - \min_{s' \in \mathcal{S}^\circ} \lambda(s')$. In the following we assume that rewards are bounded in $[0,1]$, which implies that $\mathrm{span}(\lambda)$ is upper bounded by D, cf. [5,1].

Problem setting. Given a finite set of models Φ (not necessarily containing a Markov model), we want to construct a strategy that performs as well as the algorithm that knows the underlying true Markov model ϕ°, including its rewards and transition probabilities. For that purpose we define for the Markov

model ϕ° the *regret* of any strategy at time T, cf. [5, 1, 6], as

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

where r_t are the rewards received when following the proposed strategy and $\rho^*(\phi^\circ)$ is the average optimal reward in ϕ°, i.e. $\rho^*(\phi^\circ) := \rho^*(M(\phi^\circ)) := \rho(M(\phi^\circ), \pi^*_{\phi^\circ}) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} r_t(\pi^*_{\phi^\circ})\big]$ where $r_t(\pi^*_{\phi^\circ})$ are the rewards received when following the optimal policy $\pi^*_{\phi^\circ}$ on $M(\phi^\circ)$. Note that for weakly communicating MDPs the average optimal reward does not depend on the initial state.

We consider the case when Φ is finite and the learner has no knowledge of the correct approximation errors of each model in Φ. Thus, while for each model $\phi \in \Phi$ there is an associated $\epsilon = \epsilon(\phi) \geq 0$ which indicates the aggregation error (cf. Definition 1 below), this ϵ is unknown to the learner for each model.

We remark that we cannot expect to perform as well as the unknown underlying Markov model, if the model set only provides approximations. Thus, if the best approximation has error ϵ we have to be ready to accept respective error of order ϵD per step, cf. the lower bound provided by Theorem 2 below.

Overview. We start with explicating our notion of approximation in Section 2, then introduce our algorithm in Section 3, present our regret bounds in Section 4, and conclude with the proofs in Section 5.

2 Preliminaries: MDP Approximations

Approximations. Before we give the precise notion of *approximation* we are going to use, first note that in our setting the transition probabilities $p(h'|h, a)$ for any two histories $h, h' \in \mathcal{H}$ and an action a are well-defined. Then given an arbitrary model ϕ and a state $s' \in \mathcal{S}_\phi$, we can define the aggregated transition probabilities $p^{\mathrm{agg}}(s'|h, a) := \sum_{h':\phi(h')=s'} p(h'|h, a)$. Note that the true transition probabilities under ϕ° are then given by $p(s'|s, a) := p^{\mathrm{agg}}(s'|h, a)$ for $s = \phi^\circ(h)$ and $s' \in \mathcal{S}^\circ$.

Definition 1. *A model ϕ is said to be an ϵ-approximation of the true model ϕ° if: (i) for all histories h, h' with $\phi(h) = \phi(h')$ and all actions a*

$$\left|r(\phi^\circ(h), a) - r(\phi^\circ(h'), a)\right| < \epsilon, \text{ and } \left\|p(\cdot|\phi^\circ(h), a) - p(\cdot|\phi^\circ(h'), a)\right\|_1 < \tfrac{\epsilon}{2}, \quad (2)$$

and (ii) there is a surjective mapping $\alpha : \mathcal{S}^\circ \to \mathcal{S}_\phi$ such that for all histories h and all actions a it holds that

$$\sum_{\dot{s}' \in \mathcal{S}_\phi} \left|p^{\mathrm{agg}}(\dot{s}'|h, a) - \sum_{s' \in \mathcal{S}^\circ:\alpha(s')=\dot{s}'} p^{\mathrm{agg}}(s'|h, a)\right| < \tfrac{\epsilon}{2}. \quad (3)$$

Intuitively, condition (2) assures that the approximate model aggregates only histories that are mapped to similar states under the true model. Complementary, condition (3) guarantees that the state space under the approximate model

resembles the true state space.[1] Note that any model will be an ϵ-approximation of the underlying true model ϕ° for sufficiently large ϵ.

A particular natural case are approximation models ϕ which also satisfy

$$\forall h, h' \in \mathcal{H} : \ \phi^\circ(h) = \phi^\circ(h') \Longrightarrow \phi(h) = \phi(h').$$

That is, intuitively, states in $\mathcal{S}^\circ$ are aggregated to meta-states in $\mathcal{S}_\phi$, and (3) holds trivially.

We may carry over our definition of ϵ-approximation to MDPs. This will turn out useful, since each approximate model can be interpreted as an MDP approximation, cf. Section 5.1 below.

Definition 2. An MDP $\bar{M} = (\bar{\mathcal{S}}, \mathcal{A}, \bar{r}, \bar{p})$ is an ϵ-approximation of another MDP $M = (\mathcal{S}, \mathcal{A}, r, p)$ if there is a surjective function $\alpha : \mathcal{S} \to \bar{\mathcal{S}}$ such that for all s in $\mathcal{S}$:

$$\left| \bar{r}(\alpha(s), a) - r(s, a) \right| < \epsilon, \ \ and \ \sum_{\dot{s}' \in \bar{\mathcal{S}}} \left| \bar{p}(\dot{s}' | \alpha(s), a) - \sum_{s' : \alpha(s') = \dot{s}'} p(s' | s, a) \right| < \epsilon. \quad (4)$$

Error Bounds for ϵ-Approximations. The following is an error bound on the error made by an ϵ-approximation. It generalizes bounds of [9] from ergodic to communicating MDPs. For a proof see Appendix A.

Theorem 1. Let M be a communicating MDP and $\bar{M}$ be an ϵ-approximation of M. Then

$$\left| \rho^*(M) - \rho^*(\bar{M}) \right| \ \leq \ \epsilon \left(D(M) + 1 \right).$$

The following is a matching lower bound on the error by aggregation. This is an improvement over the results in [9], which only showed that the error approaches 1 when the diameter goes to infinity.

Theorem 2. For each $\epsilon > 0$ and each $2 < D < \frac{4}{\epsilon}$ there are MDPs M, $\bar{M}$ such that $\bar{M}$ is an ϵ-approximation of M, M has diameter $D(M) = D$, and

$$|\rho^*(M) - \rho^*(\bar{M})| > \tfrac{1}{56} \epsilon D(M).$$

Proof. Consider the MDP M shown in Figure 1 (left), where the (deterministic) reward in states s_0, s_0' is 0 and 1 in state s_1. We assume that $0 < \varepsilon := \frac{\epsilon}{2} < \delta := \frac{2}{D}$. Then the diameter $D(M)$ is the expected transition time from s_0' to s_0 and equal to $\frac{2}{\delta} = D$. Aggregating states s_0, s_0' gives the MDP $\bar{M}$ on the right hand side of Figure 1. Obviously, $\bar{M}$ is an ϵ-approximation of M. It is straightforward to check that the stationary distribution μ (of the only policy) in M is $(\mu(s_0), \mu(s_0'), \mu(s_1)) = \left(\frac{\delta}{3\varepsilon + 4\delta}, \frac{\varepsilon + \delta}{3\varepsilon + 4\delta}, \frac{2\varepsilon + 2\delta}{3\varepsilon + 4\delta} \right)$, while the stationary distribution in $\bar{M}$ is $(\frac{1}{2}, \frac{1}{2})$. Thus, the difference in average reward is

$$|\rho^*(M) - \rho^*(\bar{M})| = \tfrac{2\varepsilon + 2\delta}{3\varepsilon + 4\delta} - \tfrac{1}{2} \ = \ \tfrac{\varepsilon}{2(3\varepsilon + 4\delta)} \ > \ \tfrac{\varepsilon}{14\delta} = \tfrac{1}{56} \epsilon D(M). \qquad \square$$

---

[1] The allowed error in the conditions for the transition probabilities is chosen to be $\frac{\epsilon}{2}$ so that the total error with respect to the transition probabilities is ϵ. This matches the respective condition for MDP approximations in Definition 2, cf. also Section 5.1.
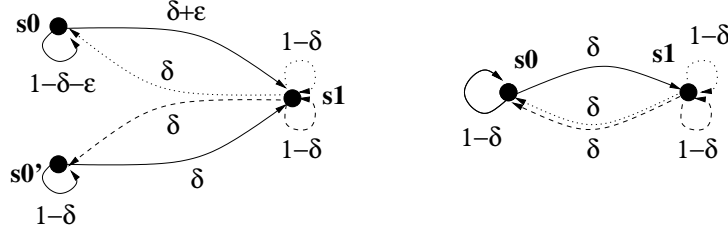
Fig. 1. The MDPs M (left) and $\bar{M}$ (right) in the proof of Theorem 2. Solid, dashed, and dotted arrows indicate different actions.

Theorems 1 and 2 compare the optimal policies of two different MDPs, however it is straightforward to see from the proofs that the same error bounds hold when comparing on some MDP M the optimal average reward $\rho^*(M)$ to the average reward when applying the optimal policy of an ϵ-approximation $\bar{M}$ of M. Thus, when we approximate an MDP M by an ϵ-approximation $\bar{M}$, the respective error of the optimal policy of $\bar{M}$ on M can be of order $\epsilon D(M)$ as well. Hence, we cannot expect to perform below this error if we only have an ϵ-approximation of the true model at our disposal.

3 Algorithm

The `OAMS` algorithm (shown in detail as Algorithm 1) we propose for the setting introduced in Section 1 is a generalization of the `OMS` algorithm of [7]. Application of the original `OMS` algorithm to our setting would not work, since `OMS` compares the collected rewards of each model to the reward it would receive if the model were Markov. Models not giving sufficiently high reward are identified as non-Markov and rejected. In our case, there may be no Markov model in the set of given models Φ. Thus, the main difference to `OMS` is that `OAMS` for each model estimates and takes into account the possible approximation error with respect to a closest Markov model.

`OAMS` 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 $t = t_{kj}$, `OAMS` chooses a policy π_{kj} applying the *optimism in face of uncertainty* principle twice.

Plausible models. First, `OAMS` considers for each model $\phi \in \Phi$ a set of *plausible* MDPs $\mathcal{M}_{t,\phi}$ defined to contain all MDPs with state space $\mathcal{S}_\phi$ and with rewards r^+ and transition probabilities p^+ satisfying

$$\left| r^+(s,a) - \widehat{r}_t(s,a) \right| \le \tilde{\epsilon}(\phi) + \sqrt{\frac{\log(48 S_\phi A t^3/\delta)}{2 N_t(s,a)}}, \tag{5}$$

$$\left\| p^+(\cdot|s,a) - \widehat{p}_t(\cdot|s,a) \right\|_1 \le \tilde{\epsilon}(\phi) + \sqrt{\frac{2 S_\phi \log(48 S_\phi A t^3/\delta)}{N_t(s,a)}}, \tag{6}$$

where $\tilde{\epsilon}(\phi)$ is the estimate for the approximation error of model ϕ (cf. below), $\widehat{p}_t(\cdot|s,a)$ and $\widehat{r}_t(s,a)$ are respectively the empirical state-transition probabilities

and the mean reward at time t for taking action a in state $s \in \mathcal{S}_\phi$, $S_\phi := |\mathcal{S}_\phi|$ denotes the number of states under model ϕ, $A := |\mathcal{A}|$ is the number of actions, 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.) The inequalities (5) and (6) are obviously inspired by Chernov bounds that would hold with high probability in case the respective model ϕ is Markov, cf. also Lemma 1 below.

Optimistic MDP for each model ϕ. In line 4, the algorithm computes for each model ϕ 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. This can be done by extended value iteration (EVI) [5]. Indeed, if $r_t^+(s, a)$ and $p_t^+(s'|s, a)$ denote the optimistic rewards and transition probabilities of $M_t^+(\phi)$, then EVI computes optimistic state values $\mathbf{u}_{t,\phi}^+ = (u_{t,\phi}^+(s))_s \in \mathbb{R}^{S_\phi}$ such that (cf. [5])

$$\widehat{\rho}_t^+(\phi) := \min_{s \in \mathcal{S}_\phi} \left\{ r_t^+(s, \pi_{t,\phi}^+(s)) + \sum_{s'} p_t^+(s'|s, \pi_{t,\phi}^+(s))\, u_{t,\phi}^+(s') - u_{t,\phi}^+(s) \right\} \qquad (7)$$

is an approximation of $\rho^*(M_t^+(\phi))$, that is,

$$\widehat{\rho}_t^+(\phi) \geq \rho^*(M_t^+(\phi)) - 2/\sqrt{t}. \qquad (8)$$

Optimistic model selection. In line 5, `OAMS` chooses a model $\phi_{kj} \in \Phi$ with corresponding MDP $M_{kj} = M_t^+(\phi_{kj})$ and policy $\pi_{kj} := \pi_{t,\phi_{kj}}^+$ that maximizes the average reward penalized by the term $\mathbf{pen}(\phi, t)$ defined as

$$\mathbf{pen}(\phi, t) := 2^{-j/2} \left(\left(\lambda(\mathbf{u}_{t,\phi}^+)\, \sqrt{2 S_\phi} + \tfrac{3}{\sqrt{2}} \right) \sqrt{S_\phi A \log\left(\tfrac{48 S_\phi A t^3}{\delta} \right)} \right. \qquad (9)$$

$$\left. + \lambda(\mathbf{u}_{t,\phi}^+)\sqrt{2 \log(\tfrac{24 t^2}{\delta})} \right) + 2^{-j} \lambda(\mathbf{u}_{t,\phi}^+) + \tilde{\epsilon}(\phi)\left(\lambda(\mathbf{u}_{t,\phi}^+) + 3 \right),$$

where we define $\lambda(\mathbf{u}_{t,\phi}^+) := \max_{s \in \mathcal{S}_\phi} u_{t,\phi}^+(s) - \min_{s \in \mathcal{S}_\phi} u_{t,\phi}^+(s)$ to be the empirical value span of the optimistic MDP $M_t^+(\phi)$. Intuitively, the penalization term is an upper bound on the per-step regret of the model ϕ in the run to follow in case ϕ is chosen, cf. eq. (35) in the proof of the main theorem. Similar to the REGAL algorithm of [1] this shall prefer simpler models (i.e., models having smaller state space and smaller value span) to more complex ones.

Termination of runs and episodes. The chosen policy π_{kj} is then executed until either (i) run j reaches the maximal length of 2^j steps, (ii) episode k terminates when the number of visits in some state has been doubled (line 12), or (iii) the executed policy π_{kj} does not give sufficiently high rewards (line 9). That is, at any time t in run j of episode k it is checked whether the total reward in the current run is at least $\ell_{kj} \rho_{kj} - \mathbf{lob}_{kj}(t)$, where $\ell_{kj} := t - t_{kj} + 1$ is the (current) length of run j in episode k, and $\mathbf{lob}_{kj}(t)$ is defined as

$$\mathbf{lob}_{kj}(t) := \left(\lambda_{kj}^+ \sqrt{2 S_{kj}} + \tfrac{3}{\sqrt{2}} \right) \sum_{s \in \mathcal{S}_{kj}} \sum_{a \in \mathcal{A}} \sqrt{v_{kj}(s, a) \log\left(\tfrac{48 S_{kj} A t_{kj}^3}{\delta} \right)}$$

$$+ \lambda_{kj}^+ \sqrt{2 \ell_{kj} \log\left(\tfrac{24 t_{kj}^2}{\delta} \right)} + \lambda_{kj}^+ + \tilde{\epsilon}(\phi_{kj}) \ell_{kj} (\lambda_{kj}^+ + 3), \qquad (10)$$

Algorithm 1 Optimal Approximate Model Selection (`OAMS`)

---

input set of models Φ, confidence parameter $\delta \in (0, 1)$, precision parameter $\epsilon_0 \in (0, 1)$

1: Let t be the current time step, and set $\tilde{\epsilon}(\phi) := \epsilon_0$ for all $\phi \in \Phi$.

2: **for** episodes $k = 1, 2, \ldots$ **do**

3:    **for** runs $j = 1, 2, \ldots$ **do**

4:       $\forall \phi \in \Phi$, use EVI to compute an optimistic MDP $M_t^+(\phi)$ in $\mathcal{M}_{t,\phi}$ (the set of *plausible* MDPs defined via the confidence intervals (6) and (5) for the estimates so far), a (near-)optimal policy $\pi_{t,\phi}^+$ on $M_t^+(\phi)$ with approximate average reward $\widehat{\rho}_t^+(\phi)$, and the empirical value span $\lambda(\mathbf{u}_{t,\phi}^+)$.

5:       Choose model $\phi_{kj} \in \Phi$ such that

$$\phi_{kj} = \operatorname*{argmax}_{\phi \in \Phi} \left\{ \widehat{\rho}_t^+(\phi) - \mathbf{pen}(\phi, t) \right\}. \tag{11}$$

6:       Set $t_{kj} := t$, $\rho_{kj} := \widehat{\rho}_t^+(\phi_{kj})$, $\pi_{kj} := \pi_{t,\phi_{kj}}^+$, and $\mathcal{S}_{kj} := \mathcal{S}_{\phi_{kj}}$.

7:       **for** 2^j steps **do**

8:          Choose action $a_t := \pi_{kj}(s_t)$, get reward r_t, observe next state $s_{t+1} \in \mathcal{S}_{kj}$.

9:          **if** the total reward collected so far in the current run is less than

$$(t - t_{kj} + 1)\rho_{kj} - \mathbf{lob}_{kj}(t), \tag{12}$$

         then

10:           $\tilde{\epsilon}(\phi_{kj}) := 2\tilde{\epsilon}(\phi_{kj})$

11:           Terminate current episode.

12:          **else if** $\sum_{j'=1}^{j} v_{kj'}(s_t, a_t) = N_{t_{k1}}(s_t, a_t)$ **then**

13:           Terminate current episode.

14:          **end if**

15:       **end for**

16:    **end for**

17: **end for**

---

where $\lambda_{kj}^+ := \lambda(\mathbf{u}_{t_{kj},\phi_{kj}}^+)$, $S_{kj} := S_{\phi_{kj}}$, and $v_{kj}(s, a)$ are the (current) state-action counts of run j in episode k. That way, `OAMS` assumes each model to be Markov, as long as it performs well. We will see that $\mathbf{lob}_{kj}(t)$ can be upper bounded by $\ell_{kj}\mathbf{pen}(\phi_{kj}, t_{kj})$, cf. eq. (35) below.

    Guessing the approximation error. The algorithm tries to guess for each model ϕ the correct approximation error $\epsilon(\phi)$. In the beginning the guessed value $\tilde{\epsilon}(\phi)$ for each model $\phi \in \Phi$ is set to the precision parameter ϵ_0, the best possible precision we aim for. Whenever the reward test fails for a particular model ϕ, it is likely that $\tilde{\epsilon}(\phi)$ is too small and it is therefore doubled (line 10).

4   Regret Bounds

The following upper bound on the regret of `OAMS` is the main result of this paper.

Theorem 3. *There are $c_1, c_2, c_3 \in \mathbb{R}$ such that in each learning problem where the learner is given access to a set of models Φ not necessarily containing the*

true model ϕ°, the regret $\Delta(\phi^\circ, T)$ of `OAMS` *(with parameters δ, ϵ_0) with respect to the true model ϕ° after any $T \geq SA$ steps is upper bounded by*

$$c_1 \cdot DSA(\log(\tfrac{1}{\epsilon_0})\log T + \log^2 T) + c_2 \cdot D\max\{\epsilon_0, \epsilon(\phi)\}T$$

$$+c_3 \cdot \left(DS_\phi\sqrt{S}A\log^{3/2}(\tfrac{T}{\delta}) + \sqrt{|\Phi|\log(\tfrac{1}{\epsilon_0})\log T}\right)\sqrt{T}$$

with probability at least $1 - \delta$, where $\phi \in \Phi$ is an $\epsilon(\phi)$-approximation of the true underlying Markov model ϕ°, $D := D(\phi^\circ)$, and $S := \sum_{\phi \in \Phi} S_\phi$.

As already mentioned, by Theorem 2 the second term in the regret bound is unavoidable when only considering models in Φ. Note that Theorem 3 holds for *all* models $\phi \in \Phi$. For the best possible bound there is a payoff between the size S_ϕ of the approximate model and its precision $\epsilon(\phi)$.

When the learner knows that Φ contains a Markov model ϕ°, the original `OMS` algorithm of [7] can be employed. In case when the total number $S = \sum_\phi S_\phi$ of states over all models is large, i.e., $S > D^2|\Phi|S^\circ$, we can improve on the state space dependence of the regret bound given in [7] as follows. The proof (found in Appendix H) is a simple modification of the analysis in [7] that exploits that by (11) the selected models cannot have arbitrarily large state space.

Theorem 4. *If Φ contains a Markov model ϕ°, with probability at least $1 - \delta$ the regret of* `OMS` *is bounded by $\tilde{O}(D^2 S^{\circ 3/2} A\sqrt{|\Phi|T})$.*

Discussion. Unfortunately, while the bound in Theorem 3 is optimal with respect to the dependence on the horizon T, the improvement on the state space dependence that we could achieve in Theorem 4 for `OMS` is not as straightforward for `OAMS` and remains an open question just as the optimality of the bound with respect to the other appearing parameters. We note that this is still an open question even for learning in MDPs (without additionally selecting the state representation) as well, cf. [5].

Another direction for future work is the extension to the case when the underlying true MDP has continuous state space. In this setting, the models have the natural interpretation of being discretizations of the original state space. This could also give improvements over current regret bounds for continuous reinforcement learning as given in [10]. Of course, the most challenging goal remains to generate suitable state representation models algorithmically instead of assuming them to be given, cf. [3]. However, at the moment it is not even clear how to deal with the case when an infinite set of models is given.

5 Proof of Theorem 3

The proof is divided into three parts and follows the lines of [7], now taking into account the necessary modifications to deal with the approximation error. First, in Section 5.1 we deal with the error of ϵ-approximations. Then in Section 5.2, we show that all state-representation models ϕ which are an $\epsilon(\phi)$-approximation of a Markov model pass the test in (12) on the rewards collected so far with high probability, provided that the estimate $\tilde{\epsilon}(\phi) \geq \epsilon(\phi)$. Finally, in Section 5.3 we use this result to derive the regret bound of Theorem 3.

5.1 Error Bounds for ϵ-Approximate Models

We start with some observations about the empirical rewards and transition probabilities our algorithm calculates and employs for each model ϕ. While the estimated rewards $\widehat{r}$ and transition probabilities $\widehat{p}$ used by the algorithm do in general not correspond to some underlying true values, the expectation values of $\widehat{r}$ and $\widehat{p}$ are still well-defined, given the history $h \in \mathcal{H}$ so far. Indeed, consider some $h \in \mathcal{H}$ with $\phi(h) = \dot{s} \in \mathcal{S}_\phi$, $\phi^\circ(h) = s \in \mathcal{S}^\circ$, and an action a, and assume that the estimates $\widehat{r}(\dot{s}, a)$ and $\widehat{p}(\cdot|\dot{s}, a)$ are calculated from samples when action a was chosen after histories $h_1, h_2, \ldots, h_n \in \mathcal{H}$ that are mapped to the same state $\dot{s}$ by ϕ. (In the following, we will denote the states of an approximation ϕ by variables with dot, such as $\dot{s}$, $\dot{s}'$, etc., and states in the state space $\mathcal{S}^\circ$ of the true Markov model ϕ° without a dot, such as s, s', etc.) Since rewards and transition probabilities are well-defined under ϕ°, we have

$$\mathbb{E}[\widehat{r}(\dot{s}, a)] = \tfrac{1}{n} \sum_{i=1}^n r(\phi^\circ(h_i), a), \text{ and } \mathbb{E}[\widehat{p}(\dot{s}'|\dot{s}, a)] = \tfrac{1}{n} \sum_{i=1}^n \sum_{h':\phi(h')=\dot{s}'} p(h'|h_i, a). \quad (13)$$

Since ϕ maps the histories $h, h_1, \ldots, h_n$ to the same state $\dot{s} \in \mathcal{S}_\phi$, the rewards and transition probabilities in the states $\phi^\circ(h), \phi^\circ(h_1), \ldots, \phi^\circ(h_n)$ of the true underlying MDP are ϵ-close, cf. (2). It follows that for $s = \phi^\circ(h)$ and $\dot{s} = \phi(h)$

$$\left| \mathbb{E}[\widehat{r}(\dot{s}, a)] - r(s, a) \right| = \left| \tfrac{1}{n} \sum_{i=1}^n \left(r(\phi^\circ(h_i), a) - r(\phi^\circ(h), a) \right) \right| < \epsilon(\phi). \quad (14)$$

For the transition probabilities we have by (3) for $i = 1, \ldots, n$

$$\sum_{\dot{s}' \in \mathcal{S}_\phi} \left| p^{\mathrm{agg}}(\dot{s}'|h_i, a) - \sum_{s' \in \mathcal{S}^\circ : \alpha(s')=\dot{s}'} p^{\mathrm{agg}}(s'|h_i, a) \right| < \tfrac{\epsilon(\phi)}{2}. \quad (15)$$

Further, all h_i as well as h are mapped to $\dot{s}$ by ϕ so that according to (2) and recalling that $s = \phi^\circ(h)$ we have for $i = 1, \ldots, n$

$$\sum_{\dot{s}' \in \mathcal{S}_\phi} \left| \sum_{s' \in \mathcal{S}^\circ : \alpha(s')=\dot{s}'} p^{\mathrm{agg}}(s'|h_i, a) - \sum_{s' \in \mathcal{S}^\circ : \alpha(s')=\dot{s}'} p(s'|s, a) \right|$$

$$\leq \sum_{s' \in \mathcal{S}^\circ} \left| p^{\mathrm{agg}}(s'|h_i, a) - p(s'|s, a) \right| < \tfrac{\epsilon(\phi)}{2}. \quad (16)$$

By (15) and (16) for $i = 1, \ldots, n$

$$\sum_{\dot{s}' \in \mathcal{S}_\phi} \left| p^{\mathrm{agg}}(\dot{s}'|h_i, a) - \sum_{s' \in \mathcal{S}^\circ : \alpha(s')=\dot{s}'} p(s'|s, a) \right| < \epsilon(\phi), \quad (17)$$

so that from (13) and (17) we can finally bound

$$\sum_{\dot{s}' \in \mathcal{S}_\phi} \left| \mathbb{E}[\widehat{p}(\dot{s}'|\dot{s}, a)] - \sum_{s' \in \mathcal{S}^\circ : \alpha(s')=\dot{s}'} p(s'|s, a) \right|$$

$$\leq \tfrac{1}{n} \sum_{i=1}^n \sum_{\dot{s}' \in \mathcal{S}_\phi} \left| p^{\mathrm{agg}}(\dot{s}'|h_i, a) - \sum_{s' \in \mathcal{S}^\circ : \alpha(s')=\dot{s}'} p(s'|s, a) \right| < \epsilon(\phi). \quad (18)$$

Thus, according to (14) and (18) the ϵ-approximate model ϕ gives rise to an MDP $\bar{M}$ on $\mathcal{S}_\phi$ with rewards $\bar{r}(\dot{s}, a) := \mathbb{E}[\widehat{r}(\dot{s}, a)]$ and transition probabilities $\bar{p}(\dot{s}'|\dot{s}, a) := \mathbb{E}[\widehat{p}(\dot{s}'|\dot{s}, a)]$ that is an ϵ-approximation of the true MDP $M(\phi^\circ)$. Note that $\bar{M}$ actually depends on the history so far.

The following lemma gives some basic confidence intervals for the estimated rewards and transition probabilities. For a proof sketch see Appendix B.

Lemma 1. *Let t be an arbitrary time step and $\phi \in \Phi$ be the model employed at step t. Then the estimated rewards $\widehat{r}$ and transition probabilities $\widehat{p}$ satisfy for all $\dot{s}, \dot{s}' \in \mathcal{S}_\phi$ and all $a \in \mathcal{A}$*

$$\widehat{r}(\dot{s}, a) - \mathbb{E}[\widehat{r}(\dot{s}, a)] \leq \sqrt{\frac{\log(48 S_\phi A t^3/\delta)}{N_t(\dot{s}, a)}},$$

$$\left\| \widehat{p}(\cdot|\dot{s}, a) - \mathbb{E}[\widehat{p}(\cdot|\dot{s}, a)] \right\|_1 \leq \sqrt{\frac{2 S_\phi \log(48 S_\phi A t^3/\delta)}{N_t(\dot{s}, a)}},$$

each with probability at least $1 - \frac{\delta}{24 t^2}$.

The following is a consequence of Theorem 1, see Appendix C for a detailed proof.

Lemma 2. *Let ϕ° be the underlying true Markov model leading to MDP $M = (\mathcal{S}^\circ, \mathcal{A}, r, p)$, and ϕ be an ϵ-approximation of ϕ°. Assume that the confidence intervals given in Lemma 1 hold at step t for all states $\dot{s}, \dot{s}' \in \mathcal{S}_\phi$ and all actions a. Then the optimistic average reward $\widehat{\rho}_t^+(\phi)$ defined in (7) satisfies*

$$\widehat{\rho}_t^+(\phi) \geq \rho^*(M) - \epsilon(D(M) + 1) - \frac{2}{\sqrt{t}}.$$

5.2 Approximate Markov models pass the test in (12)

Assume that the model $\phi_{kj} \in \Phi$ employed in run j of episode k is an $\epsilon_{kj} := \epsilon(\phi_{kj})$-approximation of the true Markov model. We are going to show that ϕ_{kj} will pass the test (12) on the collected rewards with high probability at any step t, provided that $\tilde{\epsilon}_{kj} := \tilde{\epsilon}(\phi_{kj}) \geq \epsilon_{kj}$.

Lemma 3. *For each step t in some run j of some episode k, given that $t_{kj} = t'$ the chosen model ϕ_{kj} passes the test in (12) at step t with probability at least $1 - \frac{\delta}{6 t'^2}$ whenever $\tilde{\epsilon}_{kj}(\phi_{kj}) \geq \epsilon(\phi_{kj})$.*

Proof. In the following, $\dot{s}_\tau := \phi_{kj}(h_\tau)$ and $s_\tau := \phi^\circ(h_\tau)$ are the states at time step τ under model ϕ_{kj} and the true Markov model ϕ°, respectively.

Initial decomposition. First note that at time t when the test is performed, we have $\sum_{\dot{s} \in \mathcal{S}_{kj}} \sum_{a \in \mathcal{A}} v_{kj}(\dot{s}, a) = \ell_{kj} = t - t' + 1$, so that

$$\ell_{kj} \rho_{kj} - \sum_{\tau=t'}^{t} r_\tau = \sum_{\dot{s} \in \mathcal{S}_{kj}} \sum_{a \in \mathcal{A}} v_{kj}(\dot{s}, a) \left(\rho_{kj} - \widehat{r}_{t':t}(\dot{s}, a) \right),$$

where $\widehat{r}_{t':t}(\dot{s}, a)$ is the empirical average reward collected for choosing a in $\dot{s}$ from time t' to the current time t in run j of episode k.

Let $r_{kj}^+(\dot{s}, a)$ be the rewards and $p_{kj}^+(\cdot|\dot{s}, a)$ the transition probabilities of the optimistic model $M_{t_{kj}}^+(\phi_{kj})$. Noting that $v_{kj}(\dot{s}, a) = 0$ when $a \neq \pi_{kj}(\dot{s})$, we get

$$\ell_{kj}\rho_{kj} - \sum_{\tau=t'}^{t} r_\tau = \sum_{\dot{s},a} v_{kj}(\dot{s}, a)\big(\widehat{\rho}_{kj}^+(\phi_{kj}) - r_{kj}^+(\dot{s}, a)\big) \tag{19}$$

$$+ \sum_{\dot{s},a} v_{kj}(\dot{s}, a)\big(r_{kj}^+(\dot{s}, a) - \widehat{r}_{t':t}(\dot{s}, a)\big). \tag{20}$$

We continue bounding the two terms (19) and (20) separately.

Bounding the reward term (20). Recall that $r(s, a)$ is the mean reward for choosing a in s in the true Markov model ϕ°. Then we have at each time step $\tau = t', \dots, t$

$$\begin{aligned}
r_{kj}^+(\dot{s}_\tau, a) - \widehat{r}_{t':t}(\dot{s}_\tau, a) &= \big(r_{kj}^+(\dot{s}_\tau, a) - \widehat{r}_{t'}(\dot{s}_\tau, a)\big) \\
&+ \big(\widehat{r}_{t'}(\dot{s}_\tau, a) - \mathbb{E}[\widehat{r}_{t'}(\dot{s}_\tau, a)]\big) + \big(\mathbb{E}[\widehat{r}_{t'}(\dot{s}_\tau, a)] - r(s_\tau, a)\big) \\
&+ \big(r(s_\tau, a) - \mathbb{E}[\widehat{r}_{t':t}(\dot{s}_\tau, a)]\big) + \big(\mathbb{E}[\widehat{r}_{t':t}(\dot{s}_\tau, a)] - \widehat{r}_{t':t}(\dot{s}_\tau, a)\big) \\
&\leq \tilde{\epsilon}_{kj} + 2\sqrt{\frac{\log(48 S_{kj} A t'^3/\delta)}{2 N_{t'}(\dot{s}, a)}} + 2\epsilon_{kj} + \sqrt{\frac{\log(48 S_{kj} A t'^3/\delta)}{2 v_{kj}(\dot{s}, a)}},
\end{aligned} \tag{21}$$

where we bounded the first term in the decomposition by (5), the second term by Lemma 1, the third and fourth according to (14), and the fifth by an equivalent to Lemma 1 for the rewards collected so far in the current run. In summary, with probability at least $1 - \frac{\delta}{12t'^2}$ we can bound (20) as

$$\sum_{\dot{s},a} v_{kj}(\dot{s}, a)\big(r_{kj}^+(\dot{s}, a) - \widehat{r}_{t':t}(\dot{s}, a)\big) \leq 3\tilde{\epsilon}_{kj}\ell_{kj} + \frac{3}{\sqrt{2}}\sum_{\dot{s},a}\sqrt{v_{kj}(\dot{s}, a)\log\big(\tfrac{48 S_{kj} A t'^3}{\delta}\big)}, \tag{22}$$

where we used the assumption that $\tilde{\epsilon}_{kj} \geq \epsilon_{kj}$ as well as $v_{kj}(\dot{s}, a) \leq N_{t'}(\dot{s}, a)$.

Bounding the bias term (19). First, notice that we can use (7) to bound

$$\sum_{\dot{s},a} v_{kj}(\dot{s}, a)\big(\widehat{\rho}_{kj}^+(\phi_{kj}) - r_{kj}^+(\dot{s}, a)\big) \leq \sum_{\dot{s},a} v_{kj}(\dot{s}, a)\Big(\sum_{\dot{s}'} p_{kj}^+(\dot{s}'|\dot{s}, a)\, u_{kj}^+(\dot{s}') - u_{kj}^+(\dot{s})\Big),$$

where $u_{kj}^+(\dot{s}) := u_{t_{kj},\phi_{kj}}^+(\dot{s})$ are the state values given by EVI. Further, since the transition probabilities $p_{kj}^+(\cdot|\dot{s}, a)$ sum to 1, this is invariant under a translation of the vector $\mathbf{u}_{kj}^+$. In particular, defining $w_{kj}(\dot{s}) := u_{kj}^+(\dot{s}) - \frac{1}{2}\big(\min_{\dot{s}\in\mathcal{S}_{kj}} u_{kj}^+(\dot{s}) + \max_{\dot{s}\in\mathcal{S}_{kj}} u_{kj}^+(\dot{s})\big)$, so that $\|\mathbf{w}_{kj}\|_\infty = \lambda_{kj}^+/2$, we can replace $\mathbf{u}_{kj}^+$ with $\mathbf{w}_{kj}$, and (19) can be bounded as

$$\sum_{\dot{s},a} v_{kj}(\dot{s}, a)\big(\widehat{\rho}_{kj}^+(\phi_{kj}) - r_{kj}^+(\dot{s}, a)\big)$$

$$\leq \sum_{\dot{s},a}\sum_{\tau=t'}^{t} \mathbb{1}\big\{(\dot{s}_\tau, a_\tau) = (\dot{s}, a)\big\}\Big(\sum_{\dot{s}'\in\mathcal{S}_{kj}} p_{kj}^+(\dot{s}'|\dot{s}_\tau, a)\, w_{kj}(\dot{s}') - w_{kj}(\dot{s}_\tau)\Big). \tag{23}$$

Now we decompose for each time step $\tau = t', \ldots, t$

$$\sum_{\dot{s}' \in \mathcal{S}_{kj}} p_{kj}^+(\dot{s}'|\dot{s}_\tau, a)\, w_{kj}(\dot{s}') - w_{kj}(\dot{s}_\tau) =$$

$$\sum_{\dot{s}' \in \mathcal{S}_{kj}} \left(p_{kj}^+(\dot{s}'|\dot{s}_\tau, a) - \widehat{p}_{t'}(\dot{s}'|\dot{s}_\tau, a) \right) w_{kj}(\dot{s}') \tag{24}$$

$$+ \sum_{\dot{s}' \in \mathcal{S}_{kj}} \left(\widehat{p}_{t'}(\dot{s}'|\dot{s}_\tau, a) - \mathbb{E}[\widehat{p}_{t'}(\dot{s}'|\dot{s}_\tau, a)] \right) w_{kj}(\dot{s}') \tag{25}$$

$$+ \sum_{\dot{s}' \in \mathcal{S}_{kj}} \left(\mathbb{E}[\widehat{p}_{t'}(\dot{s}'|\dot{s}_\tau, a)] - \sum_{s':\alpha(s')=\dot{s}'} p(s'|s_\tau, a) \right) w_{kj}(\dot{s}') \tag{26}$$

$$+ \sum_{\dot{s}' \in \mathcal{S}_{kj}} \sum_{s':\alpha(s')=\dot{s}'} p(s'|s_\tau, a)\, w_{kj}(\dot{s}') - w_{kj}(\dot{s}_\tau) \tag{27}$$

and continue bounding each of these terms individually.

Bounding (24)**:** Using $\|\mathbf{w}_{kj}\|_\infty = \lambda_{kj}^+/2$, (24) is bounded according to (6) as

$$\sum_{\dot{s}' \in \mathcal{S}_{kj}} \left(p_{kj}^+(\dot{s}'|\dot{s}_\tau, a) - \widehat{p}_{t'}(\dot{s}'|\dot{s}_\tau, a) \right) w_{kj}(\dot{s}') \le \left\| p_{kj}^+(\cdot|\dot{s}_\tau, a) - \widehat{p}_{t'}(\cdot|\dot{s}_\tau, a) \right\|_1 \|\mathbf{w}_{kj}\|_\infty$$

$$\le \frac{\tilde{\epsilon}_{kj}\lambda_{kj}^+}{2} + \frac{\lambda_{kj}^+}{2}\sqrt{\frac{2S_{kj}\log(48 S_{kj} A t'^3/\delta)}{N_{t'}(s,a)}}. \tag{28}$$

Bounding (25)**:** Similarly, by Lemma 1 with probability at least $1 - \frac{\delta}{24t'^2}$ we can bound (25) at all time steps τ as

$$\sum_{\dot{s}' \in \mathcal{S}_{kj}} \left(\widehat{p}_{t'}(\dot{s}'|\dot{s}_\tau, a) - \mathbb{E}[\widehat{p}_{t'}(\dot{s}'|\dot{s}_\tau, a)] \right) w_{kj}(\dot{s}') \le \frac{\lambda_{kj}^+}{2}\sqrt{\frac{2S_{kj}\log(48 S_{kj} A t'^3/\delta)}{N_{t'}(s,a)}}. \tag{29}$$

Bounding (26)**:** By (18) and using that $\|\mathbf{w}_{kj}\|_\infty = \lambda_{kj}^+/2$, we can bound (26) by

$$\sum_{\dot{s}' \in \mathcal{S}_{kj}} \left(\mathbb{E}[\widehat{p}_{t'}(\dot{s}'|\dot{s}_\tau, a)] - \sum_{s':\alpha(s')=\dot{s}'} p(s'|s_\tau, a) \right) w_{kj}(\dot{s}') < \frac{\epsilon_{kj}\lambda_{kj}^+}{2}. \tag{30}$$

Bounding (27)**:** We set $w'(s) := w_{kj}(\alpha(s))$ for $s \in \mathcal{S}^\circ$ and rewrite (27) as

$$\sum_{\dot{s}' \in \mathcal{S}_{kj}} \sum_{s':\alpha(s')=\dot{s}'} p(s'|s_\tau, a)\, w_{kj}(\dot{s}') - w_{kj}(\dot{s}_\tau) = \sum_{s' \in \mathcal{S}^\circ} p(s'|s_\tau, a)\, w'(s') - w'(s_\tau). \tag{31}$$

Summing this term over all steps $\tau = t', \ldots, t$, we can rewrite the sum as a martingale difference sequence, so that Azuma-Hoeffding's inequality (e.g., Lemma 10 of [5]) yields that with probability at least $1 - \frac{\delta}{24t'^3}$

$$\sum_{\tau=t'}^{t} \sum_{s' \in \mathcal{S}^\circ} p(s'|s_\tau, a)\, w'(s') - w'(s_\tau) = \sum_{\tau=t'}^{t} \left(\sum_{s'} p(s'|s_\tau, a)\, w'(s') - w'(s_{\tau+1}) \right)$$

$$+ w'(s_{t+1}) - w'(s_{t'}) \le \lambda_{kj}^+ \sqrt{2\ell_{kj}\log(\frac{24t'^3}{\delta})} + \lambda_{kj}^+, \tag{32}$$

since the sequence $X_\tau := \sum_{s'} p(s'|s_\tau, a)\, w'(s') - w'(s_{\tau+1})$ is a martingale difference sequence with $|X_t| \leq \lambda_{kj}^+$.

Wrap-up. Summing over the steps $\tau = t', \ldots, t$, we get from (23), (27), (28), (29), (30), (31), and (32) that with probability at least $1 - \frac{\delta}{12t'^2}$

$$\sum_{\dot{s},a} v_{kj}(\dot{s}, a)\big(\widehat{\rho}_{kj}^+(\phi_{kj}) - r_{kj}^+(\dot{s}, a)\big) \ \leq \ \tilde{\epsilon}_{kj}\lambda_{kj}^+\ell_{kj}$$

$$+\lambda_{kj}^+ \sum_{\dot{s},a} \sqrt{2v_{kj}(\dot{s}, a)\, S_{kj} \log\big(\tfrac{48 S_{kj} A t'^3}{\delta}\big)} + \lambda_{kj}^+ \sqrt{2\ell_{kj} \log\big(\tfrac{24 t'^2}{\delta}\big)} \ + \ \lambda_{kj}^+, \quad (33)$$

using that $v_{kj}(\dot{s}, a) \leq N_{t'}(\dot{s}, a)$ and the assumption that $\epsilon_{kj} \leq \tilde{\epsilon}_{kj}$. Combining (20), (22), and (33) gives the claimed lemma. $\qquad\square$

Summing Lemma 3 over all episodes gives the following lemma, for a detailed proof see Appendix D.

Lemma 4. *With probability at least $1 - \delta$, for all runs j of all episodes k the chosen model ϕ_{kj} passes all tests, provided that $\tilde{\epsilon}_{kj}(\phi_{kj}) \geq \epsilon(\phi_{kj})$.*

5.3 Preliminaries for the proof of Theorem 3

We start with some auxiliary results for the proof of Theorem 3. Lemma 5 bounds the bias span of the optimistic policy, Lemma 6 deals with the estimated precision of ϕ_{kj}, and Lemma 7 provides a bound for the number of episodes. For proofs see Appendix E, F, and G.

Lemma 5. *Assume that the confidence intervals given in Lemma 1 hold at some step t for all states $\dot{s} \in \mathcal{S}_\phi$ and all actions a. Then for each ϕ, the set of plausible MDPs $\mathcal{M}_{t,\phi}$ contains an MDP $\widetilde{M}$ with diameter $D(\widetilde{M})$ upper bounded by the true diameter D, provided that $\tilde{\epsilon}(\phi) \geq \epsilon(\phi)$. Consequently, the respective bias span $\lambda(\mathbf{u}_{t,\phi}^+)$ is bounded by D as well.*

Lemma 6. *If all chosen models ϕ_{kj} pass all tests in run j of episode k whenever $\tilde{\epsilon}(\phi_{kj}) \geq \epsilon(\phi_{kj})$, then $\tilde{\epsilon}(\phi) \leq \max\{\epsilon_0, 2\epsilon(\phi)\}$ always holds for all models ϕ.*

Lemma 7. *Assume that all chosen models ϕ_{kj} pass all tests in run j of episode k whenever $\tilde{\epsilon}(\phi_{kj}) \geq \epsilon(\phi_{kj})$. Then the number of episodes K_T after any $T \geq SA$ steps is upper bounded as $K_T \leq SA \log_2\big(\tfrac{2T}{SA}\big) + \sum_{\phi:\epsilon(\phi)>\epsilon_0} \log_2\big(\tfrac{\epsilon(\phi)}{\epsilon_0}\big)$.*

5.4 Bounding the regret (Proof of Theorem 3)

Now we can finally turn to showing the regret bound of Theorem 3. We will assume that all chosen models ϕ_{kj} pass all tests in run j of episode k whenever $\tilde{\epsilon}(\phi_{kj}) \geq \epsilon(\phi_{kj})$. According to Lemma 4 this holds with probability at least $1 - \delta$.

Let $\phi_{kj} \in \Phi$ be the model that has been chosen at time t_{kj}, and consider the last but one step t of run j in episode k. The regret Δ_{kj} of run j in episode k with respect to $\rho^* := \rho^*(\phi^\circ)$ is bounded by

$$\Delta_{kj} := (\ell_{kj} + 1)\rho^* - \sum_{\tau=t_{kj}}^{t+1} r_\tau \;\leq\; \ell_{kj}\big(\rho^* - \rho_{kj}\big) + \rho^* + \ell_{kj}\rho_{kj} - \sum_{\tau=t_{kj}}^{t} r_\tau\,,$$

where as before $\ell_{kj} := t - t_{kj} + 1$ denotes the length of run j in episode k up to the considered step t. By assumption the test (12) on the collected rewards has been passed at step t, so that

$$\Delta_{kj} \leq \ell_{kj}\big(\rho^* - \rho_{kj}\big) + \rho^* + \mathbf{lob}_{kj}(t), \tag{34}$$

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

Bounding the regret with the penalization term. Since we have $v_{kj}(\dot{s}, a) \leq N_{t_{k1}}(\dot{s}, a)$ for all $\dot{s} \in \mathcal{S}_{kj}, a \in \mathcal{A}$ and also $\sum_{\dot{s},a} v_{kj}(\dot{s}, a) = \ell_{kj} \leq 2^j$, by Cauchy-Schwarz inequality $\sum_{\dot{s},a} \sqrt{v_{kj}(\dot{s}, a)} \leq 2^{j/2}\sqrt{S_{kj}A}$. Applying this to the definition (10) of $\mathbf{lob}_{kj}$, we obtain from (34) and by the definition (9) of the penalty term that

$$\Delta_{kj} \;\leq\; \ell_{kj}\big(\rho^* - \rho_{kj}\big) + \rho^* + 2^{j/2}\big(\lambda_{kj}^+\sqrt{2S_{kj}} + \tfrac{3}{\sqrt{2}}\big)\sqrt{S_{kj}A\log\big(\tfrac{48S_{kj}At_{kj}^3}{\delta}\big)}$$
$$+ \lambda_{kj}^+\sqrt{2\ell_{kj}\log\big(\tfrac{24t_{kj}^2}{\delta}\big)} + \lambda_{kj}^+ + \tilde{\epsilon}\big(\phi_{kj}\big)\ell_{kj}(\lambda_{kj}^+ + 3)$$
$$\leq\; \ell_{kj}\big(\rho^* - \rho_{kj}\big) + \rho^* + 2^j\,\mathbf{pen}(\phi_{kj}, t_{kj}). \tag{35}$$

The key step. Now, by definition of the algorithm and Lemma 2, for any approximate model ϕ we have

$$\rho_{kj} - \mathbf{pen}(\phi_{kj}, t_{kj}) \geq \widehat{\rho}_{t_{kj}}^+(\phi) - \mathbf{pen}(\phi, t_{kj}) \tag{36}$$
$$\geq \rho^* - (D+1)\epsilon(\phi) - \mathbf{pen}(\phi, t_{kj}) - 2t_{kj}^{-1/2},$$

or equivalently $\rho^* - \rho_{kj} + \mathbf{pen}(\phi_{kj}, t_{kj}) \leq \mathbf{pen}(\phi, t_{kj}) + (D+1)\epsilon(\phi) + 2t_{kj}^{-1/2}$. Multiplying this inequality with 2^j and noting that $\ell_{kj} \leq 2^j$ then gives

$$\ell_{kj}\big(\rho^* - \rho_{kj}\big) + 2^j\mathbf{pen}(\phi_{kj}, t_{kj}) \leq 2^j\mathbf{pen}(\phi, t_{kj}) + 2^j(D+1)\epsilon(\phi) + 2^{j+1}t_{kj}^{-1/2}.$$

Combining this with (35), we get by application of Lemma 5, i.e., $\lambda(\mathbf{u}_{t_{kj},\phi}^+) \leq D$, and the definition of the penalty term (9) that

$$\Delta_{kj} \leq \rho^* + 2^{j/2}\Big(\big(D\sqrt{2S_\phi} + \tfrac{3}{\sqrt{2}}\big)\sqrt{S_\phi A\log\big(\tfrac{48S_\phi At_{kj}^3}{\delta}\big)} + D\sqrt{2\log\big(\tfrac{24t_{kj}^2}{\delta}\big)}\Big)$$
$$+ D + 2^j\tilde{\epsilon}(\phi)\big(D+3\big) + 2^j(D+1)\epsilon(\phi) + 2^{j+1}t_{kj}^{-1/2}\,.$$

By Lemma 6 and using that $2t_{kj} \geq 2^j$ (so that $2^{j+1}t_{kj}^{-1/2} \leq 2\sqrt{2}\cdot 2^{j/2}$) we get

$$\Delta_{kj} \leq \rho^* + 2^{j/2}\Big(\big(D\sqrt{2S_\phi} + \tfrac{3}{\sqrt{2}}\big)\sqrt{S_\phi A\log\big(\tfrac{48S_\phi At_{kj}^3}{\delta}\big)} + D\sqrt{2\log\big(\tfrac{24t_{kj}^2}{\delta}\big)}\Big)$$
$$+ D + \tfrac{3}{2}\cdot 2^j\max\{\epsilon_0, 2\epsilon(\phi)\}\big(D + \tfrac{7}{3}\big) + 2\sqrt{2}\cdot 2^{j/2}. \tag{37}$$

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 $t_{kj} \leq T$ and summing (37) over all runs and episodes gives

$$\Delta(\phi^\circ, T) = \sum_{k=1}^{K_T} \sum_{j=1}^{J_k} \Delta_{kj} \leq \left(\rho^* + D\right) \sum_{k=1}^{K_T} J_k + \tfrac{3}{2} \max\{\epsilon_0, 2\epsilon(\phi)\}\left(D + \tfrac{7}{3}\right) \sum_{k=1}^{K_T} \sum_{j=1}^{J_k} 2^j$$

$$+ \left(\left(D\sqrt{2S_\phi} + \tfrac{3}{\sqrt{2}}\right)\sqrt{S_\phi A \log\left(\tfrac{48 S_\phi A T^3}{\delta}\right)} + D\sqrt{2\log\left(\tfrac{24 T^2}{\delta}\right)} + 2\sqrt{2}\right) \sum_{k=1}^{K_T} \sum_{j=1}^{J_k} 2^{j/2}.$$

As shown in Section 5.2 of [7], $\sum_k J_k \leq K_T \log_2(2T/K_T)$, $\sum_k \sum_j 2^j \leq 2(T + K_T)$ and $\sum_k \sum_j 2^{j/2} \leq \sqrt{2K_T \log_2(2T/K_T)(T + K_T)}$, and we may conclude the proof applying Lemma 7 and some minor simplifications. $\qquad\square$

References

1. Bartlett, P.L., Tewari, A.: REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In: UAI 2009, Proc. 25th Conf. on Uncertainty in Artificial Intelligence, pp. 25–42. AUAI Press (2009)
2. Hallak, A., Castro, D.D., Mannor, S.: Model selection in Markovian processes. In: 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, KDD 2013, pp. 374–382. ACM (2013)
3. Hutter, M.: Feature Reinforcement Learning: Part I: Unstructured MDPs. J. Artificial General Intelligence 1, 3–24 (2009)
4. Littman, M, Sutton, R., Singh S.: Predictive representations of state. In: Adv. Neural Inf. Process. Syst. 15, pp. 1555–1561 (2002)
5. Jaksch, T., Ortner, R., Auer, P.: Near-optimal regret bounds for reinforcement learning. J. Mach. Learn. Res. 11, 1563–1600 (2010)
6. Maillard, O.A., Munos, R., Ryabko, D.: Selecting the state-representation in reinforcement learning. In: Adv. Neural Inf. Process. Syst. 24, pp. 2627–2635 (2012)
7. Maillard, O.A., Nguyen, P., Ortner, R., Ryabko, D.: Optimal regret bounds for selecting the state representation in reinforcement learning. In: Proc. 30th Int'l Conf. on Machine Learning, ICML 2013, JMLR Proc., vol. 28, pp. 543–551 (2013)
8. Nguyen, P., Maillard, O.A., Ryabko, D., Ortner, R.: Competing with an infinite set of models in reinforcement learning. In: Proc. 16th Int'l Conf. on Artificial Intelligence and Statistics, AISTATS 2013, JMLR Proc., vol. 31, pp. 463–471 (2013)
9. Ortner, R.: Pseudometrics for state aggregation in average reward Markov decision processes. In: ALT 2007. LNCS (LNAI), vol. 4754, pp. 373–387. Springer (2007)
10. Ortner, R., Ryabko, D.: Online Regret Bounds for Undiscounted Continuous Reinforcement Learning, In: Adv. Neural Inf. Process. Syst. 25, pp. 1772–1780 (2012)

A Proof of Theorem 1

We start with an error bound for approximation, assuming we compare two MDPs over the same state space.

Lemma 8. *Consider a communicating MDP $M = (\mathcal{S}, \mathcal{A}, r, p)$, and another MDP $\bar{M} = (\mathcal{S}, \mathcal{A}, \bar{r}, \bar{p})$ over the same state-action space which is an ϵ-approximation of M (for $\alpha = \text{id}$). Assume that an optimal policy π^* of M is performed on $\bar{M}$ for ℓ steps, and let $\bar{v}^*(s)$ be the number of times state s is visited state among these ℓ steps. Then*

$$\ell\rho^*(M) - \sum_{s\in\mathcal{S}} \bar{v}(s) \cdot \bar{r}(s, \pi^*(s)) \; < \; \ell\epsilon(D+1) + D\sqrt{2\ell\log(1/\delta)} + D$$

with probability at least $1 - \delta$, where $D = D(M)$ is the diameter of M.

Proof. We abbreviate $r^*(s) := r(s, \pi^*(s))$ and $p^*(s'|s) := p(s'|s, \pi^*(s))$, and use $\bar{r}^*(s)$ and $\bar{p}^*(s'|s)$ accordingly. Then

$$\ell\rho^*(M) - \sum_{s\in\mathcal{S}} \bar{v}^*(s) \cdot \bar{r}^*(s) = \sum_s \bar{v}^*(s)\big(\rho^*(M) - \bar{r}^*(s)\big)$$

$$= \sum_s \bar{v}^*(s)\big(\rho^*(M) - r^*(s)\big) + \sum_s \bar{v}^*(s)\big(r^*(s) - \bar{r}^*(s)\big). \qquad (38)$$

Now, for the first term in (38) we can use the Poisson equation (for the optimal policy π^* on M) to replace $\rho^*(M) - r^*(s) = \sum_{s'} p^*(s'|s) \cdot \lambda^*(s') - \lambda^*(s)$, writing $\lambda^* := \lambda_{\pi^*}$ for the bias of π^* on M. Concerning the second term in (38) we can use (4) and the fact that $\sum_s \bar{v}^*(s) = \ell$. In summary, we get

$$\ell\rho^*(M) - \sum_{s\in\mathcal{S}} \bar{v}^*(s) \cdot \bar{r}^*(s) < \sum_s \bar{v}^*(s)\Big(\sum_{s'} p^*(s'|s)\,\lambda^*(s') - \lambda^*(s)\Big) + \ell\epsilon$$

$$= \ell\epsilon + \sum_s \bar{v}^*(s)\Big(\sum_{s'} \bar{p}^*(s'|s)\,\lambda^*(s') - \lambda^*(s)\Big)$$

$$+ \sum_s \bar{v}^*(s)\Big(\sum_{s'} p^*(s'|s)\,\lambda^*(s') - \sum_{s'} \bar{p}^*(s'|s)\,\lambda^*(s')\Big). \qquad (39)$$

By (4) and using that $\text{span}(\lambda^*) \le D$, the last term in (39) is bounded as

$$\sum_s \bar{v}^*(s) \sum_{s'} \Big(p^*(s'|s,a) - \bar{p}^*(s'|s,a)\Big)\lambda^*(s') < \ell\epsilon D. \qquad (40)$$

On the other hand, for the second term in (39), writing s_τ for the state visited at time step τ we have

$$\sum_s \bar{v}^*(s)\Big(\sum_{s'} \bar{p}^*(s'|s)\,\lambda^*(s') - \lambda^*(s)\Big) = \sum_{\tau=1}^{\ell}\Big(\sum_{s'} \bar{p}^*(s'|s_\tau)\,\lambda^*(s') - \lambda^*(s_\tau)\Big)$$

$$= \sum_{\tau=1}^{\ell}\Big(\sum_{s'} \bar{p}^*(s'|s_\tau)\,\lambda^*(s') - \lambda^*(s_{\tau+1})\Big) + \lambda^*(s_{\ell+1}) - \lambda^*(s_1). \quad (41)$$

Now $\lambda^*(s_{\ell+1}) - \lambda^*(s_1) \leq \mathrm{span}(\lambda^*) \leq \mathrm{D}$, while the sequence

$$X_\tau := \sum_{s'} \bar{p}^*(s'|s_\tau) \lambda^*(s') - \lambda^*(s_{\tau+1})$$

is a martingale difference sequence with $|X_t| \leq D$. Thus, an application of Azuma-Hoeffding's inequality (e.g., Lemma 10 of [5]) to (41) yields that

$$\sum_s \bar{v}^*(s)\Big(\sum_{s'} \bar{p}^*(s'|s) \lambda^*(s') - \lambda^*(s)\Big) \leq D\sqrt{2\ell \log(1/\delta)} + D \qquad (42)$$

with probability higher than $1 - \delta$. Summarizing, (39), (40), and (42) give the claimed

$$\ell \rho^*(M) - \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \bar{v}(s)^* \cdot \bar{r}^*(s) < \ell\epsilon(D+1) + D\sqrt{2\ell \log(1/\delta)} + D.$$

$\square$

As a corollary to Lemma 8 we can also bound the approximation error in average reward, which we will need below to deal with the error of ϵ-approximate models.

Lemma 9. *Let M be a communicating MDP with optimal policy π^*, and $\bar{M}$ an ϵ-approximation of M over the same state space. Then*

$$\big|\rho^*(M) - \rho^*(\bar{M})\big| \ \leq \ \big|\rho^*(M) - \rho(\bar{M}, \pi^*)\big| \ \leq \ \epsilon\,(D(M) + 1).$$

Proof. Divide the result of Lemma 8 by ℓ, choose $\delta = 1/\ell$, and let $\ell \to \infty$. Since the average reward of a policy is no random value, the result holds surely and not just with probability 1. $\square$

A.1 Proof of Theorem 1

The idea is to define a new MDP $M' = (\mathcal{S}, \mathcal{A}, r', p')$ on $\mathcal{S}$ whose rewards r' and transition probabilities p' are ϵ-close to M and that has the same optimal average reward as $\bar{M}$. Thus, for each state $s \in \mathcal{S}$ and each action a we set $r'(s,a) := \bar{r}(\alpha(s), a)$ and

$$p'(s'|s,a) := \frac{p(s'|s,a) \cdot \bar{p}(\alpha(s')|\alpha(s), a)}{\sum_{s'':\alpha(s'')=\alpha(s')} p(s''|s,a)}.$$

Note that $p'(\cdot|s,a)$ is indeed a probability distribution over $\mathcal{S}$, that is, in particular it holds that

$$\sum_{s' \in \mathcal{S}} p'(s'|s,a) = \sum_{s' \in \mathcal{S}} \frac{p(s'|s,a)}{\sum_{s'':\alpha(s'')=\alpha(s')} p(s''|s,a)} \cdot \bar{p}(\alpha(s')|\alpha(s), a)$$

$$= \sum_{\dot{s}' \in \bar{\mathcal{S}}} \sum_{s':\alpha(s')=\dot{s}'} \frac{p(s'|s,a)}{\sum_{s'':\alpha(s'')=\dot{s}'} p(s''|s,a)} \cdot \bar{p}(\dot{s}'|\alpha(s), a)$$

$$= \sum_{\dot{s}' \in \bar{\mathcal{S}}} \bar{p}(\dot{s}'|\alpha(s), a) \ = \ 1.$$

Now by definition, the rewards $r'(s,a) = \bar{r}(\alpha(s),a)$ and aggregated transition probabilities

$$\sum_{s':\alpha(s')=\dot{s}'} p'(s'|s,a) = \sum_{s':\alpha(s')=\dot{s}'} \frac{p(s'|s,a)}{\sum_{s'':\alpha(s'')=\dot{s}'} p(s''|s,a)} \cdot \bar{p}(\dot{s}'|\alpha(s),a) = \bar{p}(\dot{s}'|\alpha(s),a)$$

in M' have the same values for all states s that are mapped to the same meta-state by α. It follows that $\rho^*(M') = \rho^*(\bar{M})$.

Further by assumption, according to (4) we have

$$|r(s,a) - r'(s,a)| = |r(s,a) - \bar{r}(\alpha(s),a)| < \epsilon$$

and

$$
\begin{aligned}
\sum_{s'\in\mathcal{S}} |p(s'|s,a) - p'(s'|s,a)| &= \sum_{s'\in\mathcal{S}} p(s'|s,a)\left|1 - \frac{\bar{p}(\alpha(s')|\alpha(s),a)}{\sum_{s'':\alpha(s'')=\alpha(s')} p(s''|s,a)}\right| \\
&= \sum_{\dot{s}'\in\bar{\mathcal{S}}} \sum_{s':\alpha(s')=\dot{s}'} p(s'|s,a) \cdot \left|\frac{\sum_{s'':\alpha(s'')=\dot{s}'} p(s''|s,a) - \bar{p}(\dot{s}'|\alpha(s),a)}{\sum_{s'':\alpha(s'')=\dot{s}'} p(s''|s,a)}\right| \\
&= \sum_{\dot{s}'\in\bar{\mathcal{S}}} \left|\sum_{s'':\alpha(s'')=\dot{s}'} p(s''|s,a) - \bar{p}(\dot{s}'|\alpha(s),a)\right| < \epsilon.
\end{aligned}
$$

Thus, M' is an ϵ-approximation of M that has the same optimal average reward as $\bar{M}$ so that application of Lemma 9 to M and M' gives the claimed result. $\square$

B   Proof of Lemma 1

For any given number of observations n it holds that (cf. Appendix C.1 of [5]) for any $\theta > 0$

$$\mathbb{P}\left\{\hat{r}(\dot{s},a) - \mathbb{E}[\hat{r}(\dot{s},a)] \leq \sqrt{\frac{\log(2/\theta)}{n}}\right\} < \theta,$$

$$\mathbb{P}\left\{\left\|\hat{p}(\cdot|\dot{s},a) - \mathbb{E}[\hat{p}(\cdot|\dot{s},a)]\right\|_1 \leq \sqrt{\frac{2S_\phi \log(2/\theta)}{n}}\right\} < \theta.$$

Choosing suitable values for θ, a union bound over all states $\dot{s}$, all actions a and all possible values for $N_t(\dot{s},a) = 1,\ldots,t$ shows the lemma. $\square$

C   Proof of Lemma 2

Let $\bar{M}$ be the MDP on $\mathcal{S}_\phi$ whose rewards and transition probabilities are given by the expectation values $\mathbb{E}[\hat{r}(\dot{s},a)]$ and $\mathbb{E}[\hat{p}(\dot{s}'|\dot{s},a)]$, respectively. We have already seen in (14) and (18) that $\bar{M}$ is an ϵ-approximation of the true MDP M, so that by Theorem 1

$$\left|\rho^*(M) - \rho^*(\bar{M})\right| \leq \epsilon(D(M) + 1). \tag{43}$$

It remains to deal with the difference between $\rho^*(\bar{M})$ and $\widehat{\rho}_t^+(\phi)$. By assumption, the confidence intervals of Lemma 1 hold for all state-action-pairs so that $\bar{M}$ is contained in the set of plausible MDPs $\mathcal{M}_{t,\phi}$ (defined via the empirical rewards and transition probabilities $\widehat{r}(\dot{s}, a)$ and $\widehat{p}(\dot{s}'|\dot{s}, a)$). It follows together with (8) that

$$\widehat{\rho}_t^+(\phi) \ \geq \ \rho^*(M_t^+(\phi)) - \tfrac{2}{\sqrt{t}} \ \geq \ \rho^*(\bar{M}) - \tfrac{2}{\sqrt{t}}, \tag{44}$$

which together with (43) proves the claimed inequality. $\qquad\square$

D   Proof of Lemma 4

By Lemma 3, at each step t of a run j in an episode k starting at step $t_{kj} = t'$ the test is passed with probability at least $1 - \frac{\delta}{6t'^2}$. Assuming that t'' is the last step in that run and setting $\ell := t'' - t' + 1$ to be the total number of steps in that run, the test is passed in all steps of the run with error probability bounded by (using that $2t' \geq 2^j \geq \ell$)

$$\ell \cdot \frac{\delta}{6t'^2} \ \leq \ \frac{\ell\delta}{2t'(t'+\ell)} = \frac{\delta}{2t'} - \frac{\delta}{2(t'+\ell)} = \int_{t'}^{t''} \frac{\delta}{2\tau^2} d\tau \ \leq \ \sum_{\tau=t'}^{t''} \frac{\delta}{2\tau^2}.$$

Summing over all episodes and runs shows that the test is passed in all time steps with probability at least $1 - \sum_{\tau=1}^{\infty} \frac{\delta}{2\tau^2} \geq 1 - \delta$. $\qquad\square$

E   Proof of Lemma 5

We define an MDP $\widetilde{M}$ on state space $\mathcal{S}_\phi$ as follows. First let $\beta : \mathcal{S}_\phi \to \mathcal{S}$ be an arbitrary mapping that maps states in $\mathcal{S}_\phi$ to some state in $\mathcal{S}$ such that $\alpha(\beta(\dot{s})) = \dot{s}$. Intuitively, $\beta(\dot{s})$ is an arbitrary reference state that is mapped to $\dot{s}$ by α. Then for $\dot{s}, \dot{s}'$ in $\mathcal{S}_\phi$ we set the transition probabilities of $\widetilde{M}$ as

$$\widetilde{p}(\dot{s}'|\dot{s}, a) := \sum_{s':\alpha(s')=\dot{s}'} p(s'|\beta(\dot{s}), a). \tag{45}$$

Then by (18) and Lemma 1 we obtain

$$\left\|\widetilde{p}(\cdot|\dot{s}, a) - \widehat{p}_t(\cdot|\dot{s}, a)\right\|_1 = \sum_{\dot{s}'} \Big| \sum_{s':\alpha(s')=\dot{s}'} p(s'|\beta(\dot{s}), a) - \widehat{p}_t(\cdot|\dot{s}, a)\Big|$$

$$\leq \sum_{\dot{s}'} \Big(\Big| \sum_{s':\alpha(s')=\dot{s}'} p(s'|\beta(\dot{s}), a) - \mathbb{E}[\widehat{p}_t(\dot{s}'|\dot{s}, a)]\Big| + \Big| \mathbb{E}[\widehat{p}_t(\dot{s}'|\dot{s}, a)] - \widehat{p}_t(\dot{s}'|\dot{s}, a)\Big| \Big)$$

$$\leq \epsilon(\phi) + \sqrt{\frac{2S_\phi \log(48 S_\phi A t^3/\delta)}{N_t(s,a)}},$$

showing that $\widetilde{M}$ is contained in $\mathcal{M}_{t,\phi}$. To see that $D(\widetilde{M}) \leq D$, note that β maps all transitions in $\widetilde{M}$ to transitions of the the same or lower probability in the

true MDP. That is, for any $\dot{s}, \dot{s}' \in \mathcal{S}_\phi$ it holds that $\widetilde{p}(\dot{s}'|\dot{s}, a) \geq p(\beta(\dot{s}')|\beta(\dot{s}), a)$. Thus, each trajectory in $\widetilde{M}$ can be mapped to a trajectory in the true MDP that cannot have higher probability, which proves the first claim of the lemma. The second claim follows immediately along the lines of Section 4.3.1 in [5]. $\square$

F Proof of Lemma 6

By definition of the algorithm, $\tilde{\epsilon}(\phi)$ for each model ϕ has initial value ϵ_0 and is doubled whenever ϕ fails a test. Thus, by assumption if $\epsilon_0 \leq \epsilon(\phi)$, then as soon as $\epsilon(\phi) \leq \tilde{\epsilon}(\phi) < 2\epsilon(\phi)$ the value of $\tilde{\epsilon}(\phi)$ will not change anymore, and consequently $\tilde{\epsilon}(\phi) < 2\epsilon(\phi)$ always holds.

On the other hand, if $\epsilon_0 > \epsilon(\phi)$ then also $\tilde{\epsilon}(\phi) \geq \epsilon(\phi)$ for the initial value $\tilde{\epsilon}(\phi) = \epsilon_0$ and again by assumption $\tilde{\epsilon}(\phi) = \epsilon_0$ remains unchanged, so that $\tilde{\epsilon}(\phi) \leq \epsilon_0$ holds. $\square$

G Proof of Lemma 7

First recall that an episode is terminated when either the number of visits in some state-action pair (s, a) has been doubled (line 12 of the algorithm) or when the test on the accumulated rewards has failed (line 9). By assumption, the test is passed provided that $\tilde{\epsilon}(\phi) \geq \epsilon(\phi)$. If $\epsilon(\phi) \leq \epsilon_0$, then $\tilde{\epsilon}(\phi) \geq \epsilon(\phi)$ holds trivially. Otherwise, ϕ will fail the test only $\log_2\left(\frac{\epsilon(\phi)}{\epsilon_0}\right)$ times until $\tilde{\epsilon}(\phi) \geq \epsilon(\phi)$ (after which the test is passed w.h.p. and $\tilde{\epsilon}(\phi)$ remains unchanged by Lemma 3). Therefore, the number of episodes terminated due to failure of the test is upper bounded by $\sum_{\phi:\epsilon(\phi)>\epsilon_0} \log_2\left(\frac{\epsilon(\phi)}{\epsilon_0}\right)$.

For the number of episodes terminated since the number of visits in some state-action pair (s, a) has been doubled, one can show that it is bounded by $SA \log_2\left(\frac{2T}{SA}\right)$, cf. Appendix C.2 of [5] or Section 5.2 of [7], and the lemma follows. $\square$

H Proof of Theorem 4

As the proof of the regret bound for OMS given in [7] follows the same lines as the proof of Theorem 3 given here, we only sketch the key step leading to the improvement of the bound. Note that by (36) and since average rewards are by assumption in $[0, 1]$, for the model ϕ_{kj} chosen in some run j of some episode k it holds that $\mathbf{pen}(\phi_{kj}, t_{kj}) \leq \mathbf{pen}(\phi^\circ, t_{kj}) - 1$. Hence, by definition of the penalization term (9) and since $1 \leq \lambda(\mathbf{u}_{t,\phi^\circ}^+) \leq D$, the chosen model ϕ_{kj} always satisfies

$$\left(2\sqrt{2S_{kj}} + \tfrac{3}{\sqrt{2}}\right)\sqrt{S_{kj}A \log\left(\tfrac{48S_{kj}At^3}{\delta}\right)} + 2\sqrt{2\log\left(\tfrac{24t_{kj}^2}{\delta}\right)}$$
$$\leq \left(2D\sqrt{2S^\circ} + \tfrac{3}{\sqrt{2}}\right)\sqrt{S^\circ A \log\left(\tfrac{48S^\circ At^3}{\delta}\right)} + 2D\sqrt{2\log\left(\tfrac{24t_{kj}^2}{\delta}\right)} + 2^{-j/2}D.$$

Some simplifications then show that S_{kj} is $\tilde{O}(D^2 S^\circ)$, so that one can replace the total number of all states $S = \sum_\phi S_\phi$ in Theorem 3 (respectively in the regret bound of [7]) by the total number of states of models ϕ with $S_\phi = \tilde{O}(D^2 S^\circ)$ and consequently by $\tilde{O}(|\Phi| D^2 S^\circ)$. $\qquad\square$