Competing with an Infinite Set of Models in Reinforcement Learning

Phuong Nguyen
Australian National University and NICTA
Canberra ACT 0200, AUSTRALIA

Odalric-Ambrym Maillard[1]
Technion, Faculty of Electrical Engineering,
32000 Haifa, ISRAEL

Daniil Ryabko
INRIA Lille - Nord Europe
59650 Villeneuve d'Ascq, FRANCE

Ronald Ortner
Montanuniversität Leoben
A-8700 Leoben, AUSTRIA

Abstract

We consider a reinforcement learning setting where the learner also has to deal with the problem of finding a suitable state-representation function from a given set of models. This has to be done while interacting with the environment in an online fashion (no resets), and the goal is to have small regret with respect to any Markov model in the set. For this setting, recently the BLB algorithm has been proposed, which achieves regret of order $T^{2/3}$, provided that the given set of models is finite. Our first contribution is to extend this result to a countably infinite set of models. Moreover, the BLB regret bound suffers from an additive term that can be exponential in the diameter of the MDP involved, since the diameter has to be guessed. The algorithm we propose avoids guessing the diameter, thus improving the regret bound.

1 Introduction

Motivation. In Reinforcement Learning (RL) an agent has to learn a task through interactions with the environment. The most well-studied fundamental framework for this problem is that of Markov decision processes (MDP). However, in reality most environments are non-Markovian. This poses a challenging problem: how to construct an efficient and generic agent that can deal with the non-Markovian property of such environments. Recently, [MMR11] introduced an algorithm called BLB (Best Lower Bound), whose regret is of order $T^{2/3}$ with respect to the optimal policy on a Markov state representation in the model set. However, this bound holds only under the assumption that the set of models (or *state representation functions*) is finite and contains at least one Markov model. Note that each model here is a state-representation function that maps histories to representative states; at time t, a model ϕ maps history h_t to state $s_t = \phi(h_t)$; we say ϕ is a Markov model if the process (s_t, r_t, a_t) is Markov for any time t. The second limitation of BLB lies in its use of a function for guessing the diameter of Markov models. The restriction to a finite set of models in BLB might hinder its flexibility in solving interesting RL tasks, while the dependence on the diameter-guessing function costs an additive term that can be exponential in the diameter in the overall regret. We propose an algorithm named IBLB (Infinite Best Lower Bound) that overcomes these two limitations of BLB. IBLB can be seen as a step towards solving the ultimate challenging continuous general RL (GRL) problem where observation and action spaces are continuous, and the environment's underlying model and states are both inaccessible.

Contributions. The contributions of this work are as follows: (1) the IBLB algorithm that can deal with a countably infinite set of models, and (2), unlike BLB, does not have the additive term that is exponential in the diameter of the MDP against which the performance of our algorithm is assessed; (3) the regret bound for IBLB is of order $T^{2/3}$ with respect to any algorithm that knows any of the Markov models. Finally, (4) we derive two lemmas for (4a) a model gener-

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

[1] This work was done while the second author was working at Montanuniversität Leoben.

ation setting where Markov models are available with some probability, and (4b) a model ordering setting in which we employ a map selection strategy based on the Occam's razor principle [Hut05, LV08].

Related work. Besides [MMR11], there are several other lines of research that are related to our approach. Both `BLB` and `IBLB` are constructed based on `UCRL2` [JOA10], an algorithm that efficiently learns undiscounted MDPs, and achieves regret of order $DT^{1/2}$ for any weakly communicating MDP having diameter D, with respect to the optimal policy on this MDP. There is a rich literature in RL on the regret bounds of MDP learning [BT02, SLW$^+$06, BT09, JOA10]. As already mentioned, the problem we address falls under the challenging GRL setting where both the environment dynamics and states are unknown. In GRL, to the best of our knowledge, Maillard et al's work [MMR11] is the first that offers a finite-time performance analysis given a set of state representations. Some other approaches attempting to solve the GRL problem include ΦMDP [Hut09], context-tree based methods [McC96, VNH$^+$11, NSH11], predictive state representations [LSS02, MB05, BG10], and learning to select from a countably infinite set of arbitrary models [RH08]. All this previous work offers general schemes and algorithms for constructing a GRL agent, but does not analyze the regret bounds with respect to the optimal model. In other words, the data-efficiency analysis of those methods is missing.

2 Preliminaries

Agent-Environment setup. Suppose that the agent interacts with some *unknown* environment. Denote the spaces of observations, actions, and rewards by $\mathcal{O}$, $\mathcal{A}$, and $\mathcal{R}$ respectively. We assume that $\mathcal{A}$ is finite. At $t = 0$, the agent gets some initial observation $h_0 = o_0 \in \mathcal{O}$, then at any time step $t > 0$, the agent takes action $a_t \in \mathcal{A}$ based on the current history $h_t = o_0 a_0 o_1 r_1 a_1 o_2 r_2 \dots\ o_t r_t$, and in return, it receives observation $o_{t+1} \in \mathcal{O}$ and reward $r_{t+1} \in \mathcal{R}$ from the environment.

State representation functions (models). A *state-representation function* ϕ, or shortly a *model* ϕ, is a function from the set of histories to a finite state set $\mathcal{S}$ of some (approximate) MDP. Let $\mathcal{S}_\phi$ be the set of states induced by model ϕ and $s_{t,\phi} := \phi(h_t)$ the state derived from ϕ at time t. In general we use $s_t := \phi(h_t)$ if the associated ϕ can be inferred from the context. Context trees [McC96, NSH11], looping suffix trees [HJ06], and probabilistic deterministic finite automata [VTH$^+$05] can be considered to be model classes in our sense. A state representation function ϕ is called a *Markov* model of the environment, if the process $(s_{t,\phi}, a_t, r_t), t \in \mathbb{N}$ is an MDP, which will be denoted by $M(\phi)$. We abbreviate $A := |\mathcal{A}|$, $S_\phi := |\mathcal{S}_\phi|$, and $S_j := S_{\phi_j}$.

We assume that MDPs induced from Markov models are *weakly communicating* [Put93], that is, for any two states u_1 and u_2 there is a non-zero probability that u_2 is reachable from u_1 after some finite number of actions. The *diameter* D of an MDP is defined as the expected minimum number of time steps needed to reach any state starting from any other state. The diameter of a Markov model ϕ is denoted by $D(\phi)$.

2.1 Problem description

Given a countably infinite set $\Phi = \{\phi_1, \phi_2, \dots, \}$ of models which contains at least one Markov model, we want to construct a strategy that performs as well as the algorithm that knows any Markov model ϕ_j, and also knows all the parameters (transition probabilities and rewards) of the MDP corresponding to this model. Thus, we define the regret of any strategy at time T, like in [JOA10, BT09], as

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

where the r_t's are the rewards received when following the proposed strategy and $\rho^\star(\phi_j)$ is the average optimal value in the Markov model ϕ_j, i.e., $\rho^\star(\phi_j) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} r_t(\pi(\phi_j))\big]$ where $r_t(\pi(\phi_j))$ are the rewards received when following the optimal policy $\pi(\phi_j)$ for ϕ_j. Note that in a weakly communicating MDP, the average optimal value does not depend on the initial state. In general, $T\rho^\star(\phi_j)$ can be replaced with the expected sum of rewards obtained in T steps (following the optimal policy) at the expense of an additional (additive) term upper bounded by the diameter of the underlying MDP.

3 Algorithms

As the `UCRL2` algorithm of [JOA10] is an integral part of the proposed algorithm `IBLB`, we briefly recall the main features of `UCRL2` first.

3.1 The `UCRL2` algorithm

`UCRL2` is an efficient algorithm for learning in a finite MDP M with unknown rewards and transition probabilities. We first define some needed quantities for the description of `UCRL2`. Let $N(\tau, s, a)$ be the number of times action a has been taken in state s up to time τ; if a has not been chosen in s so far we set $N(\tau, s, a) := 1$. Denote the empirical state transition probabilities and empirical mean rewards of M up to

time τ as $\widehat{p}_\tau(\cdot|s,a)$ and $\widehat{r}_\tau(s,a)$. At time τ, we define $\mathcal{M}(\tau,\delta)$ as the set of so-called δ-*admissible* MDPs (where δ is a confidence parameter) with transition probabilities $p(\cdot|s,a)$ and mean rewards $r(s,a)$ such that

$$\|p(\cdot|s,a) - \widehat{p}_\tau(\cdot|s,a)\|_1 \leqslant \sqrt{\frac{14S\log(2A\tau/\delta)}{N(\tau,s,a)}}, \qquad (1)$$

$$|r(s,a) - \widehat{r}_\tau(s,a)| \leqslant \sqrt{\frac{7S\log(2A\tau/\delta)}{N(\tau,s,a)}}. \qquad (2)$$

The algorithm proceeds in periods $m = 1, 2, \ldots$. At the beginning time τ_m of each period m, the algorithm finds an optimistic MDP (that is, the one with the highest expected value) $M_m^+ \in \mathcal{M}(\tau_m,\delta)$ and a corresponding optimal policy π_m^+ using the extended iteration value (EVI) procedure with precision $1/\sqrt{\tau_m}$ [JOA10]. Then the policy π_m^+ is executed until the number of visits in some action-pair has doubled, that is, until $v_m(s,a) = N(\tau_m,s,a)$ for some (s,a), where $v_m(s,a)$ is the number of times the pair (s,a) has been visited from τ_m up to current time τ in period m.

Our own IBLB algorithm will apply UCRL2 given some model ϕ. Then τ in the confidence intervals (1) and (2) will not correspond to the current (absolute) time step, but to the (relative) time step of playing model ϕ. That is, any model ϕ has an individual counter τ_ϕ counting the number of steps when model ϕ is used to choose an action, and UCRL2 will consider the respective set of δ-admissible MDPs $\mathcal{M}(\phi,\tau_\phi,\delta)$.

Value span. In each period m, the EVI procedure computes approximate state values $u_m^+(s)$ for each state s. We define the *empirical value span* of the optimistic MDP M_m^+ in period m as

$$\mathbf{sp}(u_m^+) := \max_{s\in\mathcal{S}} u_m^+(s) - \min_{s\in\mathcal{S}} u_m^+(s)\,.$$

3.2 The IBLB algorithm

Algorithm description. Our IBLB algorithm (given in detail as Algorithm 1) proceeds in *episodes* $k = 1, 2, \ldots$ (not to be confused with the *periods* of UCRL2) each of deterministic length $\ell_k := 2^k$. In each episode k, the algorithm considers a finite set $\Phi_k \subset \Phi$ of $J_k := k^\beta$ ($\beta := 2$) many models by adding models to the J_{k-1} models of the model set Φ_{k-1} of the previous episode $k-1$. Each episode starts with an exploration phase of length ℓ_k^{explore} followed by an exploitation phase of length ℓ_k^{exploit}, such that $\ell_k = \ell_k^{\text{explore}} + \ell_k^{\text{exploit}}$.

The *exploration phase* of episode k consists of J_k *runs* of the UCRL2 algorithm, one for each $\phi \in \Phi_k$. The UCRL2 algorithm requires a confidence parameter that is chosen to be $\delta_k := \frac{75}{76}\cdot 2^{1-k}\delta$ where δ is the confidence

Algorithm 1 IBLB

Require: $\Phi = \{\phi_1, \phi_2, \ldots\}$, confidence parameter δ.

Set parameters: $\ell_k := 2^k$ length of episode k, $\ell_k^{\text{explore}} = 2(\frac{102^2}{3})^{1/3}\ell_k^{2/3}J_k^{1/3}$, $\ell_k^{\text{exploit}} := \ell_k - \ell_k^{\text{explore}}$, $J_k := k^\beta$ ($\beta := 2$), $\delta_k = \frac{75}{76}\cdot 2^{1-k}\delta$.
Initialize $t := 0$, $\Phi_0 := \emptyset$.
for episodes $k = 1, 2, \ldots$ **do**
  $\tilde{\Phi}_k := \{$get $J_k - J_{k-1}$ models ϕ from $\Phi\}$, $\Phi := \Phi\backslash\tilde{\Phi}_k$
  $\Phi_k := \Phi_{k-1} \cup \tilde{\Phi}_k$
  {**Exploration phase**}
  Run UCRL2 periods for each $\phi \in \Phi_k$, with parameter δ_k for $\ell_k^{\text{explore}}/J_k$ time steps.
  $t := t + \ell_k^{\text{explore}}$
  {**Exploitation phase**}
  $q := 0$, *stop* :=false
  while NOT *stop* **do**
    $q := q + 1$
    {**Exploitation run**}
    For each $\phi \in \Phi_k$, compute an optimistic MDP $M_{k,q}^+(\phi)$ and its corresponding policy $\pi_{k,q}^+(\phi)$ using the EVI procedure.
    $\Phi_k' := \Phi_k$
    while NOT *stop* AND $\Phi_k' \neq \emptyset$ **do**
      $\widehat{\phi} := \text{argmax}_{\phi\in\Phi_k'} \left\{\widehat{r}_{k,<q}(\phi) - 2B_{k,q}(\phi,\delta_k)\right\}$
      $\widehat{\pi} := \pi_{k,q}^+(\widehat{\phi})$
      {**Exploitation play**}
      $s := s_{t,\widehat{\phi}}$
      while $v_{k,q,t}(\widehat{\phi},s,\widehat{\pi}(s)) < N_{k,<q}(\widehat{\phi},s,\widehat{\pi}(s))$ **do**
        Choose action $a = \widehat{\pi}(s)$, get reward r' and next observation o'.
        $t := t + 1$
        $\ell_k^{\text{exploit}} := \ell_k^{\text{exploit}} - 1$
        {BLB **test**}
        if $\widehat{r}_{k,t}(\widehat{\phi}) < \widehat{r}_{k,<q}(\widehat{\phi}) - 2B_{k,q}(\widehat{\phi},\delta_k)$ **then**
          $\Phi_k' := \Phi_k'\backslash\{\widehat{\phi}\}$ {BLB test fails}
          break {end current play}
        end if
        if $\ell_k^{\text{exploit}} = 0$ **then**
          stop := true
          break {end current episode}
        end if
        $s := s_{t,\widehat{\phi}}$
      end while
      if $v_{k,q,t}(\widehat{\phi},s,\widehat{\pi}(s)) = N_{k,<q}(\widehat{\phi},s,\widehat{\pi}(s))$ **then**
        break {end current run}
      end if
    end while
  end while
end for

parameter for IBLB. Each of the exploration runs stops after precisely $\ell_k^{\text{explore}}/J_k$ steps (the same for each of the models).

The subsequent *exploitation phase* of episode k is split into several runs $q = 1, 2, \ldots$. In each run q in episode k, an optimistic MDP $M^+_{k,q}(\phi)$ together with an associated optimistic policy $\pi^+_{k,q}(\phi)$ is computed for each $\phi \in \Phi_k$ based on previous observations. Then, the algorithm repeatedly chooses a candidate model $\widehat{\phi}$ according to line 14, and executes its corresponding optimistic policy $\pi^+_{k,q}(\widehat{\phi})$ following a modified version of UCRL2 with additional stopping conditions. The model $\widehat{\phi}$ chosen is the one with the biggest empirical mean reward received at the beginning of run q in episode k, and penalized by some quantity accounting for a confidence interval on the would-be cumulated reward if $\widehat{\phi}$ is Markov. Each selection of $\widehat{\phi}$ together with an execution of the policy $\widehat{\pi}$ is called an (exploitation) *play*. A play is terminated when the current collected average reward is too low (cf. the *BLB test* in line 20 of the algorithm). Then the current $\widehat{\phi}$ is discarded, and the next best model (line 14) is selected. If the number of visits of some state-action pair in the current run q is equal to the total respective visits from the beginning of episode k to the beginning of run q in episode k (lines 16 and 30 of the algorithm), the current run q stops and the algorithm proceeds to the next run $q+1$ of the same episode. Finally, a new episode is started after ℓ_k steps (line 24 of the algorithm). Note that a model ϕ is selected at most once within an exploitation *run*, but can be chosen and executed within many different runs in the exploitation *phase* of the same episode.

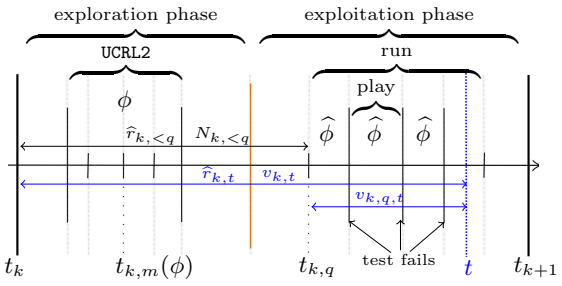

Figure 1: Illustration of some notations used.

Notation. We now introduce some notation needed for the description of Algorithm 1. We say a model ϕ is *active* at time t if it is the model used to generate the current action. This happens exactly $\ell_k^{\text{explore}}/J_k$ times during the exploration phase of episode k, and during the runs in the exploitation phase when ϕ is the chosen model $\widehat{\phi}$ (line 14 of IBLB). We write $t \in \{1, \ldots, T\}$ for the current time index, and respectively t_k and $t_{k,q}$ for the beginning time of episode k, and run q in the exploitation phase of episode k.

We write $N_{k,<q}(\phi) := \sum_{t'=t_k}^{t_{k,q}-1} \mathbb{I}\{\phi \text{ is active at } t'\}$ for the number of times ϕ is active from the beginning of episode k to the beginning of run q in the same episode. Furthermore, let $v_{k,t}(\phi) := \sum_{t'=t_k}^{t} \mathbb{I}\{\phi \text{ is active at } t'\}$ be the number of times ϕ has been active in episode k up to step t. Similarly, $v_{k,q,t}(\phi) := \sum_{t'=t_{k,q}}^{t} \mathbb{I}\{\phi \text{ is active at } t'\}$ is the number of times ϕ has been active in run q of episode k up to step t. Analogously, we write $N_{k,<q}(\phi, s, a)$, $v_{k,t}(\phi, s, a)$, and $v_{k,q,t}(\phi, s, a)$ for the respective number of time steps when ϕ was active and action a has been chosen in state s.

When performing the BLB test in run q to decide whether the collected rewards are sufficiently high, the algorithm compares the empirical mean reward of the current model ϕ in episode k before the start of run q

$$\widehat{r}_{k,<q}(\phi) := \frac{1}{N_{k,<q}(\phi)} \sum_{t'=t_k}^{t_{k,q}-1} r_{t'}\, \mathbb{I}\{\phi \text{ is active at } t'\}$$

to the empirical mean reward in episode k up to the current time step t

$$\widehat{r}_{k,t}(\phi) := \frac{1}{v_{k,t}(\phi)} \sum_{t'=t_k}^{t} r_{t'}\, \mathbb{I}\{\phi \text{ is active at } t'\}\,.$$

Note that when UCRL2 is run in the exploration phase for some model ϕ (line 6 of Algorithm 1) in episode k, UCRL2 uses internal periods $m = 1, 2, \ldots$, and computes for each of these an optimistic value function $u^{+,\text{explore}}_{k,m}(\phi)$ at $t_{k,m}(\phi)$, the beginning time of period m of UCRL2 where ϕ is used. Similarly, we write $u^{+,\text{exploit}}_{k,q}(\phi)$ for the optimistic value function in run q of episode k when ϕ is active. Then we define the maximum span of the optimistic value functions computed up to time t in episode k as

$$\mathbf{sp}^+_{k,t}(\phi) := \max\Big\{\max\big\{\mathbf{sp}(u^{+,\text{explore}}_{k,m}(\phi))\,;\, m \text{ s.t. } t_{k,m}(\phi) \leqslant t\big\},\ \max\big\{\mathbf{sp}(u^{+,\text{exploit}}_{k,q}(\phi))\,;\, q \text{ s.t. } t_{k,q} \leqslant t\big\}\Big\}\,.$$

The penalization used in line 14 is defined as

$$B_{k,q}(\phi, \delta_k) := 34\mathbf{sp}^+_{k,t_{k,q}}(\phi)\, S_\phi \sqrt{\frac{A \log\left(N_{k,<q}(\phi)/\delta_k\right)}{N_{k,<q}(\phi)}}.$$

For convenience, we also introduce the notation

$$B(\phi, \ell, \delta) := 34 D(\phi) S_\phi \sqrt{\frac{A \log\left(\ell/\delta\right)}{\ell}}\,.$$

4 Performance bounds

The following upper bound on the regret of the IBLB algorithm is our main result. It shows that IBLB suffers

regret of order $T^{2/3}$ w.r.t the best Markov model in the given model set.

Theorem 1. *Let $\Phi = \{\phi_1, \phi_2, \ldots\}$ be an infinite set of state-representation functions containing at least one Markov model, and assume that in each episode k the* IBLB *algorithm chooses $\tilde{\Phi}_k := \{\phi_1, \ldots, \phi_{J_k}\}$. Then for any time horizon T, with probability at least $1-\delta$, the regret of the* IBLB *algorithm w.r.t the optimal policy on any Markov model $\phi_j \in \Phi$, is bounded by*

$$43(1+\sqrt{\log(\tfrac{1}{\delta})})D(\phi_j)S_j\sqrt{A}(\log(T+1))^{4/3}T^{2/3} + 27(\log(T+1))^2\rho^\star(\phi_j) + 2^{\lfloor\sqrt{j}\rfloor}\rho^\star(\phi_j)\,.$$

Remark. The last term $2^{K_0(\phi_j)}\rho^\star(\phi_j)$ in the regret bound of Theorem 1 depends on $K_0(\phi_j) = \lfloor j^{1/\beta}\rfloor$, the first episode when ϕ_j appears. In Theorem 1 we have chosen $\beta = 2$. In principle, the rate of taking new models —which is $(\log(T))^\beta$ for IBLB— can be increased to a polynomial order $T^{\beta'}$ $(0 < \beta' < 2/3)$, when $K_0(\phi_j) = \log(j)/\beta'$. However, this is at the expense of increasing the regret to order $T^{2/3+\beta'/2}$.

In the following, we consider two special cases, where an upper bound on the first episode K_0 providing a Markov model can be given with high probability.

Model generation. First, consider a model generation setting where the models are generated from an "infinite source" of models Φ. We assume that this source Φ has the property that whenever a new model is taken (at line 4 of IBLB), then with probability at least $\alpha \in (0,1)$ (w.r.t the source Φ), the presented or generated model is Markov. That is, the IBLB algorithm is unchanged, but Markov models from Φ are assumed to be available with probability at least α at any time. Under this assumption on the source Φ, we have Lemma 1.

Lemma 1. *Given an infinite model source Φ that at any time step generates a Markov model with probability at least α $(0 < \alpha < 1)$, with probability at least $1-\delta$ $(0 < \delta < 1)$, the first episode in which a Markov model appears is bounded as*

$$K_0 \leqslant 2^{\sqrt{\log_{1-\alpha}\delta}}.$$

Proof. The probability that a Markov model is generated within n_0 generations is lower bounded by $1-(1-\alpha)^{n_0}$, which is $\geqslant 1-\delta$ if $n_0 \geqslant \frac{1}{\log_\delta(1-\alpha)}$. The lemma follows after some simple derivations. □

Model ordering. We now consider a model ordering setting where we have access to an initial ordering of the countably infinite model set $\Phi = \{\phi_1, \phi_2, \ldots\}$. Based on this model ordering, we design a specific strategy for choosing new models (line 4 of IBLB), which is based on some descriptional complexity for models in Φ. Note again that the general working steps of IBLB are unchanged; thus, at line 4, we still add $J_k - J_{k-1}$ new models but these models are chosen based on our specific selection strategy. We prefer to select ϕ's that have low complexity; this strategy is inspired by the Occam's razor principle [LV08]. Our reason for proposing this strategy is that among all Markov models in Φ, the ones with short description length are likely to be selected first; and these small Markov models will help the IBLB algorithm early achieve small regret as IBLB is competitive to any Markov models in the current considered set. The interested reader is referred to [LV08] and [Grü07] for detailed treatment of description theory. Assume that the description method for all models $\phi \in \Phi$ gives prefix-free codes with code length $\mathbf{CL}(\phi)$; and suppose that models in Φ are enumerated as $\phi_1, \phi_2, \ldots, .$ Denote $\eta_i = \sum_{i'=1}^{i} 2^{-\mathbf{CL}(\phi_{i'})}$. Then since $\eta_i < 1$ (by Kraft's inequality [CT91, LV08]) and η_i is increasing in i, there exists a finite constant $\eta = \lim_{i\to\infty}\eta_i = \sum_{\phi\in\Phi} 2^{-\mathbf{CL}(\phi)}$ $(0 < \eta \leqslant 1)$.

Our model selection strategy here attempts to choose any state representation ϕ in $\Phi = \{\phi_1, \phi_2, \ldots\}$ with probability at least $p_\phi := 2^{-\mathbf{CL}(\phi)}/\eta$ where $\eta = \sum_{\phi\in\Phi} 2^{-\mathbf{CL}(\phi)}$ is assumed to be known. We describe how this model selection (sampling) process is executed based on the proposed distribution p_ϕ. The first model ϕ' is selected as follows. First generate a random number u in $(0, \eta]$, then incrementally construct intervals $(\sum_{h=0}^{i} p_{\phi_h}, \sum_{h=0}^{i+1} p_{\phi_h}]$ $(i = 0, 1, \ldots$ and adding a fictitious model ϕ_0 and suppressing $p_{\phi_0} = 0)$ until u falls within the current interval $(\sum_{h=0}^{i} p_{\phi_h}, \sum_{h=0}^{i+1} p_{\phi_h}]$. So the model selected is $\phi' = \phi_{i+1}$; and this happens with probability $p_{\phi'}$. The selection process for subsequent models is repeated with $\eta := \eta - 2^{-\mathbf{CL}(\phi')}$, and $\Phi := \Phi\backslash\{\phi'\}$. With this so-called "simplicity-biased" strategy, the IBLB algorithm gives the following result.

Lemma 2. *Let Φ be a countably infinite set of models. Then for time horizon T, with probability at least $1-\delta$ $(0 < \delta < 1)$, the first episode in which the Markov model ϕ appears is bounded as*

$$K_0(\phi) \leqslant 2^{\sqrt{\log_{1-p_\phi}\delta}},$$

where $p_\phi = 2^{-\boldsymbol{CL}(\phi)}/\eta$ is a constant with $\eta = \sum_{\phi'\in\Phi} 2^{-\boldsymbol{CL}(\phi')}$.

Proof. The first model ϕ' is chosen with probability $p_{\phi'}$. After ϕ' is selected, the probability of choosing any other model $\phi \in \Phi \setminus \{\phi'\}$ in the next step increases as the new normalizing constant $\eta' =$

$\sum_{\phi\in\Phi\setminus\{\phi'\}} 2^{-\mathbf{CL}(\phi)} < \eta$. This process is repeated for all other subsequent models. Hence, at any time the probability to select any model ϕ is at least p_ϕ. So the probability that any Markov model ϕ is selected after n_0 model selections from Φ is $P_{n_0} \geqslant 1-(1-\frac{1}{\eta}p_\phi)^{n_0}$. It can be verified that with $n_0 \geqslant \frac{1}{\log_\delta(1-p_\phi)}$, $P_{n_0} \geqslant 1-(1-p_\phi)^{n_0} \geqslant 1-\delta$. Now the claimed result follows similarly as Lemma 1. □

Lemma 2 shows the potential of a selection strategy based on the Occam's razor principle. If there exists a Markov model that has a short description length compared to other models, then the IBLB algorithm is likely to perform well w.r.t this model.

Getting tight and explicit bounds on K_0 that depend on the characteristics of the underlying MDPs induced from Markov models is an interesting open question. One direction to attack it is to use notions of complexity introduced in [Hut09] that take into account characteristics of state representation functions, the resulting MDPs, and previous histories.

5 Proof of Theorem 1

Overview. Let ϕ_j be an arbitrary Markov model. First, we express the total regret Δ w.r.t ϕ_j by the regret Δ_k suffered in episodes k. Each Δ_k then is further divided into the regret of models only active in the exploration phases, denoted by Φ_k^{explore}, and the regret of models active in the exploitation phase, denoted by Φ_k^{exploit}. In general, models in $\Phi_k^{\text{explore}}\setminus\{\phi_j\}$ are assumed to suffer the worst case regret since we do not know whether they are Markov. When analyzing the regret suffered in the exploitation phases, we distinguish between "good" and "bad" models in Φ_k^{exploit}. Good models are those better than ϕ_j according to the criterion in line 14 of IBLB, while all other models in $\Phi_k^{\text{exploit}}\setminus\{\phi_j\}$ are classified as bad. We show that with high probability any Markov model will pass the BLB test; hence it is highly probable that we only models at least as good as ϕ_j are active. This crucial fact allows to achieve a bound on the regret for each episode.

5.1 Performance guarantee for Markov models

Understanding the behavior of Markov models in IBLB is essential in our regret analysis. Lemma 3 below shows that at any time when a Markov model ϕ_j is active, its associated MDP $M(\phi_j)$ is admissible (cf. equations (1) and (2)) with high probability. Furthermore, Lemma 4 provides an episode-wise regret bound for the regret incurred when a Markov model ϕ_j is active.

Lemma 3. *For any Markov model $\phi_j \in \Phi_k$, any $0 < \delta' < 1$, with probability at least $1-\frac{\delta'}{75}$, for all time steps t it holds that $M(\phi_j) \in \mathcal{M}(\phi_j, v_t(\phi_j), \delta')$, where $v_t(\phi)$ denotes the number of times model ϕ has been active up to time t. Hence, with probability $1-\delta_k$ we have $\mathbf{sp}^+_{k,t}(\phi_j) \leqslant D(\phi_j)$ for all time steps t in any episode k.*

Proof sketch. Lemma 17 in Appendix C.1 of [JOA10] shows that $M(\phi_j) \in \mathcal{M}(\phi_j, v_t(\phi_j), \delta')$ with probability $1-\frac{\delta'}{v^6}$ under the assumption that $v_t(\phi_j) = v$. Then a union bound over all possible values of $v_t(\phi_j) = 1, 2, \ldots, t$ gives the claimed result. That the diameter $D(\phi_j)$ bounds the optimistic value span $\mathbf{sp}^+_{k,t}(\phi_j)$ can be looked up in Section 4.3 of [JOA10]. □

Lemma 4. *For any episode k, any Markov model ϕ_j and any $0 < \delta' < 1$: With probability at least $1-\delta'$, for all time steps t in episode k with $v_{k,t}(\phi_j) > 1$, the empirical average $\widehat{r}_{k,t}(\phi_j)$ of the rewards collected from the beginning of episode k up to time t when model ϕ_j is active satisfies*

$$v_{k,t}(\phi_j)\left|\rho^\star(\phi_j) - \widehat{r}_{k,t}(\phi_j)\right| \leqslant 34\,\mathbf{sp}^+_{k,t}(\phi_j)\; S_j\sqrt{A v_{k,t}(\phi_j)\log(v_{k,t}(\phi_j)/\delta')}\,.$$

Proof sketch. The basic idea is to consider only those time steps when model ϕ_j is active and repeat the original UCRL2 analysis given in Section 4 of [JOA10], but replacing $D(\phi_j)$ with $\mathbf{sp}^+_{k,t}(\phi_j)$, and adding absolute values where needed. There are two potential problems with this approach: First, since, in general, other models will be employed between phases in which ϕ_j is active, unlike in the original UCRL2 setting, it will not be the case that each period starts in the state in which the previous period has terminated. Second, UCRL2 periods within IBLB may be terminated prematurely (i.e., not because of the original UCRL2 criterion that visits in some state-action pair have doubled, cf. lines 16 and 30 in IBLB).

However, neither of these two issues is crucial to our analysis. First, the regret analysis of UCRL2 does not make any assumptions on the initial state of each period and hence also holds when each period starts at any arbitrary state (what is important, however, is that the same model is consistently used within one period). Second, the analysis for UCRL2 can be adapted to the case when original UCRL2periods are terminated prematurely: one only has to take into account a bound on the total number of periods (as given for UCRL2 in Proposition 18 in Appendix 2 of [JOA10]). It is straightforward to see that there will

be at most one additional period in our analysis, which can be shown not to damage the regret bounds given in [JOA10]. Indeed, the only premature termination of a UCRL2 period that may cause another period is when the exploration phase of ϕ_j ends. If a play in which ϕ_j is active is terminated because the BLB test fails, or since the total number of exploitation steps in the current episode has been reached, then the model ϕ_j is not active in the rest of the episode and hence cannot cause another UCRL2 period in which ϕ_j is active.

Summarizing, adapting the UCRL2 analysis as indicated above results in that, analogous to the original regret bounds for UCRL2, for each episode k with probability $1-\delta'$, the claimed regret bound holds for all time steps t (i.e., for all possible values of $v_{k,t}(\phi_j)$). □

5.2 Regret analysis

For any Markov model $\phi_j \in \Phi$, the cumulative regret $\Delta(\phi_j, T)$ of the IBLB algorithm w.r.t the best strategy in the model ϕ_j can be decomposed into

$$\Delta(\phi_j, T) = \sum_{k<K_0(\phi_j)} \Delta_k + \sum_{k\geqslant K_0(\phi_j)} \Delta_k \,,$$

where $K_0(\phi_j)$ is the index of the first episode in which ϕ_j appears ($\phi_j \notin \Phi_{K_0(\phi_j)-1}$ and $\phi_j \in \Phi_{K_0(\phi_j)}$), and Δ_k is the total regret suffered in episode k.

5.2.1 Regret in episodes $k < K_0(\phi_j)$

When ϕ_j has not been selected yet (i.e., $\phi_j \notin \Phi_k$), we consider the worst case regret. Since the time when ϕ_j first appears is $1 + \sum_{k=1}^{K_0(\phi_j)-1} 2^k = 2^{K_0(\phi_j)} - 1$, the respective regret $\sum_{k<K_0(\phi_j)} \Delta_k$ is upper bounded by $(2^{K_0(\phi_j)} - 2)\rho^\star(\phi_j)$. As ϕ_j first appears in episode $K_0(\phi_j)$, its index j satisfies $J_{K_0(\phi_j)-1} < j \leqslant J_{K_0(\phi_j)}$. Since $J_k = k^\beta$, j satisfies $(K_0(\phi_j)-1) < j^{1/\beta} \leqslant K_0(\phi_j)$ and we have $j^{1/\beta} \leqslant K_0(\phi_j) < j^{1/\beta} + 1$ and consequently $K_0(\phi_j) = \lfloor j^{1/\beta} \rfloor$. It follows that

$$\sum_{k<K_0(\phi_j)} \Delta_k \leqslant (2^{\lfloor j^{1/\beta} \rfloor} - 2)\rho^\star(\phi_j). \tag{3}$$

5.2.2 Regret in episodes $k \geqslant K_0(\phi_j)$

Let us define in run q of episode k the set of *good* models with respect to model ϕ_j as

$$\begin{aligned}\mathcal{G}_{k,q}(\phi_j) \;:=\; \big\{\phi \in \Phi_k \backslash \{\phi_j\} : \widehat{r}_{k,<q}(\phi) - 2B_{k,q}(\phi, \delta_k) \\ \geqslant \widehat{r}_{k,<q}(\phi_j) - 2B_{k,q}(\phi_j, \delta_k)\big\}.\end{aligned}$$

All other models in $\Phi_k \backslash \{\phi_j\}$ will be considered as *bad*. Let Φ_k^{exploit} be the set of all models active in the exploitation phase of episode k.

For $\tilde{\Phi} \subset \Phi_k$ let $\Delta_k(\tilde{\Phi})$ be the total regret resulting when a $\phi \in \tilde{\Phi}$ is active in episode k, and set $\Delta_k(\phi_j) := \Delta_k(\{\phi_j\})$. Then we can decompose the regret Δ_k in episode k into three components:

$$\begin{aligned}\Delta_k(\Phi_k) \;=\; \Delta_k(\Phi_k^{\text{exploit}} \backslash \{\phi_j\}) + \Delta_k(\phi_j) \\ + \Delta_k(\Phi \backslash (\Phi_k^{\text{exploit}} \cup \{\phi_j\}))\,,\end{aligned} \tag{4}$$

the regret term corresponding to the exploited models without the comparative model ϕ_j, the regret for model ϕ_j itself, and the remaining models that are explored but not exploited in episode k.

Bad models. We prove that with high probability, there are no bad models active in any exploitation run, that is, Φ_k^{exploit} is a subset of $\bigcup_q \mathcal{G}_{k,q}(\phi_j)$. More precisely, we will show that a Markov model ϕ_j will pass all BLB tests with high probability, so that each model active in an exploitation run q must be at least as good as ϕ_j and consequently in $\mathcal{G}_{k,q}(\phi_j)$.

Indeed, let t be a time step in episode k and run q. Then $N_{k,t}(\phi_j) \geqslant N_{k,<q}(\phi_j)$ and also $\mathbf{sp}^+_{k,t}(\phi_j) = \mathbf{sp}^+_{k,t_{k,q}}(\phi_j) \geqslant \mathbf{sp}^+_{k,t_{k,q}-1}(\phi_j)$. Applying Lemma 4 twice it follows that with probability at least $1-\delta_k$ for all time steps t in episode k,

$$\begin{aligned}&\widehat{r}_{k,t}(\phi_j) - \widehat{r}_{k,<q}(\phi_j) \\ &\quad= \Big(\widehat{r}_{k,t}(\phi_j) - \rho^\star(\phi_j)\Big) + \Big(\rho^\star(\phi_j) - \widehat{r}_{k,<q}(\phi_j)\Big) \\ &\quad\geqslant -34\,\mathbf{sp}^+_{k,t}(\phi_j)\; S_j \sqrt{\frac{A\log(N_{k,t}(\phi_j)/\delta_k)}{N_{k,t}(\phi_j)}} \\ &\qquad -34\,\mathbf{sp}^+_{k,t_{k,q}-1}(\phi_j)\; S_j \sqrt{\frac{A\log(N_{k,<q}(\phi_j)/\delta_k)}{N_{k,<q}(\phi_j)}} \\ &\quad\geqslant -2B_{k,q}(\phi_j, \delta_k)\,.\end{aligned} \tag{5}$$

This means that in any episode k the Markov model ϕ_j passes the BLB test in any exploitation play with probability $1-\delta_k$.

Good models. Let $v_k(\phi)$ be the total number of times $\phi \in \Phi_k$ is active in episode k. Then the regret of exploited models in episode k (except ϕ_j) can be expressed as

$$\begin{aligned}&\Delta_k(\Phi_k^{\text{exploit}} \backslash \{\phi_j\}) \\ &\quad= \sum_{\phi \in \Phi_k^{\text{exploit}} \backslash \{\phi_j\}} v_k(\phi)\big(\rho^\star(\phi_j) - \widehat{r}_k(\phi)\big),\end{aligned} \tag{6}$$

writing $\widehat{r}_k(\phi) := \widehat{r}_{k,t_{k+1}-1}(\phi)$ for the average reward collected in episode k when ϕ was active. In the following, we also use $\widehat{r}'_k(\phi)$ to denote the respective average reward when ignoring the last time step when ϕ was active in episode k. Then, using that ϕ passed the BLB test before the last step when ϕ was active (in some

run q'), and since only good models are active,

$$\begin{aligned}
&v_k(\phi)\big(\rho^\star(\phi_j) - \widehat{r}_k(\phi)\big) \\
&\leqslant (v_k(\phi)-1)\big(\rho^\star(\phi_j) - \widehat{r}'_k(\phi)\big) + \rho^\star(\phi_j) \\
&\leqslant (v_k(\phi)-1)\big(\rho^\star(\phi_j) - \widehat{r}_{k,<q'}(\phi) + 2B_{k,q'}(\phi,\delta_k)\big) \\
&\quad +\rho^\star(\phi_j) \\
&\leqslant (v_k(\phi)-1)\big(\rho^\star(\phi_j) - \widehat{r}_{k,<q'}(\phi_j) + 2B_{k,q'}(\phi_j,\delta_k)\big) \\
&\quad +\rho^\star(\phi_j) \\
&\leqslant 3\big(v_k(\phi)-1\big)B_{k,q'}(\phi_j,\delta_k) + \rho^\star(\phi_j) \\
&\leqslant 3\big(v_k(\phi)-1\big)B(\phi_j,\ell_k^{\text{explore}}/J_k,\delta_k) + \rho^\star(\phi_j), \quad (7)
\end{aligned}$$

using Lemma 4, Lemma 3, and $N_{k,<q'}(\phi) \geqslant \ell_k^{\text{explore}}/J_k$ in the final two steps.

Comparative Markov model ϕ_j. If $v_k(\phi_j) > 1$ for the considered Markov model ϕ_j, then by Lemmas 3 and 4 and the fact that $\ell_k^{\text{explore}}/J_k \leqslant v_k(\phi_j)$, we have

$$\begin{aligned}
\Delta_k(\phi_j) &= v_k(\phi_j)\big(\rho^\star(\phi_j) - \widehat{r}_k(\phi_j)\big) \\
&\leqslant 34\,\mathbf{sp}_{k,t_{k+1}-1}(\phi_j)\; S_j\sqrt{A v_k(\phi_j)\log(v_k(\phi_j)/\delta_k)} \\
&\leqslant 34 D(\phi_j) v_k(\phi_j)\, S_j\sqrt{\frac{A\log(v_k(\phi_j)/\delta_k)}{v_k(\phi_j)}} \\
&\leqslant v_k(\phi_j) B(\phi_j,\ell_k^{\text{explore}}/J_k,\delta_k)
\end{aligned}$$

with probability at least $1-\frac{76}{75}\delta_k$. On the other hand, if $v_k(\phi_j) = 1$, then $\Delta_k(\phi_j) \leqslant \rho^\star(\phi_j)$ holds trivially and it follows that with probability at least $1-\frac{76}{75}\delta_k$,

$$\Delta_k(\phi_j) \leqslant v_k(\phi_j) B(\phi_j,\ell_k^{\text{explore}}/J_k,\delta_k) + \rho^\star(\phi_j)\,. \quad (8)$$

Remaining models. Finally, let us consider the models which are not active in the exploitation phase of episode k. We have

$$\begin{aligned}
&\Delta_k(\Phi_k \backslash (\Phi_k^{\text{exploit}} \cup \{\phi_j\})) \\
&\quad \leqslant \sum_{\phi\in(\Phi_k\backslash(\Phi_k^{\text{exploit}}\cup\{\phi_j\})} \rho^\star(\phi_j)\;\ell_k^{\text{explore}}/J_k\,. \quad (9)
\end{aligned}$$

5.3 Summary and fine-tuning of parameters

Summarizing, and noting that $\sum_\phi v_k(\phi) = \ell_k$, we get from equations (3), (4), (6), (7), (8), and (9), by a union bound over the J_k models of episode k that with probability at least $1-\frac{76}{75}\delta_k$, the total regret in episode k satisfies

$$\begin{aligned}
\Delta_k(\Phi_k) &\leqslant 3\ell_k B(\phi_j,\ell_k^{\text{explore}}/J_k,\delta_k) \\
&\quad +(J_k-1)\,\rho^\star(\phi_j)\,\ell_k^{\text{explore}}/J_k + J_k\,\rho^\star(\phi_j) \\
&\leqslant 3\ell_k B(\phi_j,\ell_k^{\text{explore}}/J_k,\delta_k) + (\ell_k^{\text{explore}}+J_k)\rho^\star(\phi_j) \\
&= 102\ell_k(\ell_k^{\text{explore}})^{-1/2} D(\phi_j) S_j\sqrt{A J_k\log(\tfrac{\ell_k^{\text{explore}}}{J_k\delta_k})} \\
&\quad +(\ell_k^{\text{explore}}+J_k)\rho^\star(\phi_j)\,. \quad (10)
\end{aligned}$$

Note that considering the error probability of Lemma 3 and Lemma 4 once is sufficient, as the claims of both Lemma 3 and Lemma 4 hold with the given error probability for all time steps t in each episode k.

Recall that parameters for `IBLB` (line 1 of the algorithm) are chosen as follows: $\ell_k = 2^k$, $\ell_k^{\text{explore}} = 2(\frac{102^2}{3})^{1/3}\ell_k^{2/3}J_k^{1/3}$, $J_k = k^2$ ($\beta = 2$), $\delta_k = \frac{75}{76}\cdot 2^{1-k}\delta$. Actually, ℓ_k^{explore} is tuned to get the best possible regret bound of the form $aT^bJ_k^c$. Denoting the episode at time T of `IBLB` by k_T, it can be seen that $\sum_{k=1}^{k_T-1}\ell_k = \sum_{k=1}^{k_T-1}2^k = 2^{k_T}-1 \leqslant T$; hence, $k_T \leqslant \log(T+1)$. With all of these notes, we deduce that with probability at least $1-\delta$

$$\begin{aligned}
\Delta(\phi_j,T) &= \sum_{k=1}^{k_T}\Delta_k(\Phi_k) \\
&\leqslant 43(\log(T+1))^{4/3}T^{2/3}D(\phi_j)S_j\sqrt{A}\big(1+\sqrt{\log(\tfrac{1}{\delta})}\big) \\
&\quad +27(\log(T+1))^2\rho^\star(\phi_j) + 2^{\lfloor\sqrt{j}\rfloor}\rho^\star(\phi_j).
\end{aligned}$$

□

6 Outlook

A natural direction for future research is to develop an algorithm that has regret of the optimal order $\sqrt{T}$ (as in the case of learning an MDP), without losing on the order of other factors in the regret. A first step in this direction is the recent work [MNOR], where regret bounds of order $\sqrt{T}$ are achieved for the case of a finite set of models, yet with a worse dependence on the state space. Another important direction, both challenging and of practical interest, is the extension to continuous state-action domains, using, for instance, discretization or aggregation maps as candidate models. Finally, an important assumption that we would like to relax (or to get rid of) is that the model set contains a true Markov model; instead, we would like to assume that we only have some approximation of such a model.

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

[BG10] B. Boots and G. Gordon. Predictive state temporal difference learning. In *Advances in Neural Information Processing Systems 23*, pages 271–279, 2010.

[BT02] R. I. Brafman and M. Tennenholz. R-max – A general polynomial time algorithm for near-optimal reinforcementlearning. *Journal of Machine Learing Research*, 3:213–231, 2002.

[BT09] P. L. Bartlett and A. Tewari. 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*, pages 35–42, 2009.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Willey and Sons, 1991.

[Grü07] P. D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, 2007.

[HJ06] M. P. Holmes and C. L. Isbell Jr. Looping suffix tree-based inference of partially observable hidden state. In *Machine Learning, Proceedings of the 23rd International Conference (ICML 2006)*, pages 409–416, 2006.

[Hut05] M. Hutter. *Universal Articial Intelligence: Sequential Decisions based on AlgorithmicProbability*. Springer, Berlin, 2005.

[Hut09] M. Hutter. Feature reinforcement learning: Part I. Unstructured MDPs. *Journal of General Artificial Intelligence*, 1:3–24, 2009.

[JOA10] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, 2010.

[LSS02] M. Littman, R. Sutton, and S. Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems 14*, pages 1555–1561, 2002.

[LV08] M. Li and P. Vitani. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2008.

[MB05] P. McCracken and M. H. Bowling. Online discovery and learning of predictive state representations. In *Advances in Neural Information Processing Systems 18*, pages 875–882, 2005.

[McC96] R. Andrew McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, Department of Computer Science, University of Rochester, 1996.

[MMR11] O. Maillard, R. Munos, and D. Ryabko. Selecting the state-representation in reinforcement learning. In *Advances in Neural Information Processing Systems 24*, pages 2627–2635, 2011.

[MNOR] O. Maillard, P. Nguyen, R. Ortner, and D. Ryabko. Optimal Regret Bounds for Selecting the State Representation in Reinforcement Learning. In *JMLR Workshop and Conference Proceedings Volume 28 : Proceedings of The 30th International Conference on Machine Learning*, pages 543–551, 2013.

[NSH11] P. Nguyen, P. Sunehag, and M. Hutter. Feature reinforcement learning in practice. In *Proceedings of the 9th European Workshop in Reinforcement Learning*. Springer, 2011.

[Put93] M. L. Puterman. *Markov Decision Processes*. Kluwer Academic Publishing, 1993.

[RH08] D. Ryabko and M. Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theoretical Computer Science*, 405(3):274–284, 2008.

[SLW+06] A. L. Strehl, L. Li, Eric Wiewiora, J. Langford, and M. L. Littman. PAC model-free reinforcement learning. In *Machine Learning, Proceedings of the 23rd International Conference (ICML 2006)*, pages 881–888, 2006.

[VNH+11] J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. A Monte-Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, 40(1):95–142, 2011.

[VTH+05] E. Vidal, F. Thollard, C. D. L. Higuera, F. Casacuberta, and R.C. Carrasco. Probabilistic finite-state machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1013–1025, 2005.