EXPLOITING SIMILARITY INFORMATION IN REINFORCEMENT LEARNING

Similarity Models for Multi-Armed Bandits and MDPs

Ronald Ortner
Lehrstuhl für Informationstechnologie, Montanuniversität Leoben
rortner@unileoben.ac.at

Keywords: reinforcement learning, Markov decision process, multi-armed bandit, similarity, regret

Abstract: This paper considers reinforcement learning problems with additional similarity information. We start with the simple setting of multi-armed bandits in which the learner knows for each arm its *color*, where it is assumed that arms of the same color have close mean rewards. An algorithm is presented that shows that this color information can be used to improve the dependency of online regret bounds on the number of arms. Further, we discuss to what extent this approach can be extended to the more general case of Markov decision processes. For the simplest case where the same color for actions means similar rewards and identical transition probabilities, an algorithm and a corresponding online regret bound are given. For the general case where transition probabilities of same-colored actions imply only close but not necessarily identical transition probabilities we give upper and lower bounds on the error by action aggregation with respect to the color information. These bounds also imply that the general case is far more difficult to handle.

1 INTRODUCTION

Algorithms for reinforcement learning problems suffer from the curse of dimensionality when either the action space or the state space are large. Unlike that, in many of these problems humans have no difficulties in learning, as they are able to structure the state space and the action space in a favorable way. In many cases, this structure information regards *similarity* of states and actions.

Here we investigate to what extent similarity information can be exploited to improve over the performance in case no such information is given. Although our main interest lies in Markov decision processes (MDPs), we start with a multi-armed bandit problem with a simple similarity model: For each arm there is an additional *color* information available, where arms of the same color are assumed to have close mean rewards, that is, these deviate by at most θ, a parameter known to the learner. Indeed, a similar model has already been considered by (Pandey et al., 2007), who also give a typical application to an ad-selection problem on webpages, where ads with similar content are similarly attractive to the user and get comparable reward (i.e., user clicks). Also in other of the numerous applications of multi-armed bandits such as routing, wireless networks, design of experiments, or pricing (for references see e.g. (Kleinberg, 2005)), similarity information of the given kind seems to be natural.

In Section 2 below we present an algorithm that is able to exploit color information, as the derived bounds on the regret with respect to the best arm show: While online regret bounds for ordinary bandit problems (which usually are logarithmic in the number of steps taken) grow linearly with the number of actions, with color information the total number of actions can be replaced with the number of colors plus the number of arms with promising color.

In the subsequent Section 3 we consider the more general setting of Markov decision processes where color information for the actions is available. We start examining the simplest case where actions of the same color have similar rewards (again measured by a parameter θ) and identical transition probabilities. For this setting we give an adaptation of the UCRL2 algorithm of (Auer et al., 2009) for which we show regret bounds that demonstrate similarly to the bandit setting that the color information can be exploited to

get improved bounds.

When this setting is generalized so that actions with the same color have not identical but only similar transition probabilities, things get more complicated. In Section 3.2, we investigate *action aggregation* with respect to such colorings. We derive bounds on the error caused by working on the aggregated instead of the original MDP. Unlike in the simpler settings where this error is trivially bounded by the parameter θ, the error can be arbitrarily large, depending on the (aggregated) MDP. This indicates that similarity information regarding the transition probabilities cannot be as well exploited as for rewards, which is confirmed by an example at the end of Section 3 that shows that straightforward adaptations of the algorithm for the simpler setting fail.

2 COLORED BANDITS

In a multi-armed bandit problem the learner has a finite set of arms A at his disposal. Choosing an arm a from A gives a random reward bounded in the unit interval $[0,1]$ with mean $r(a)$. As performance measure for a learning algorithm one usually considers its *regret* with respect to choosing the optimal arm at each step. That is, setting $\tau(a)$ to be the number of steps where arm a has been chosen (up to some finite horizon T) the *regret* is defined as $\sum_{a\in A}\tau(a)\big(r^*-r(a)\big)$, where $r^*=\max_a r(a)$ is the optimal mean reward. The regret of established bandit algorithms such as UCB1 (Auer et al., 2002) is logarithmic in the number of steps, but grows linearly with the number of arms. This is also best possible (Mannor and Tsitsiklis, 2004).

Unlike in the general case, where the learner has no information apart from A, here we are interested in the question how given similarity information about different arms can be exploited to improve regret bounds with respect to the dependency on the number of arms. That is, the learner additionally knows the *color* of an arm, that is, a coloring function $c : A \to C$ that assigns each arm in A a color from a given set of colors C. (We assume that the function c is surjective, i.e., each color in C is assigned to an arm in A.) The color gives some similarity information about the rewards of arms according to the following assumption.

Assumption 1. *There is a $\theta > 0$ such that for each two arms $a,a' \in A$: If $c(a) = c(a')$ then $|r(a) - r(a')| < \theta$.*

We assume that the learner knows the parameter θ. This setting is similar to the one considered in (Pandey et al., 2007). However, there it is assumed that choosing an arm is a Bernoulli trial that gives reward 1 with some success probability p and reward 0 otherwise. Further, our Assumption 1 is replaced with the supposition that the success probabilities p of arms of the same color are distributed according to a common probability distribution.

2.1 Algorithm

An obvious idea is to adapt a standard bandit algorithm to first choose a color c and then in a second step to choose an arm with color c. This idea also underlies the TLP algorithm of (Pandey et al., 2007). However there is a problem with that direct approach when two colors are very close, as it takes $\Omega(\frac{1}{\varepsilon^2})$ steps to distinguish a distance of ε between two arms/colors (cf. the analysis of (Pandey et al., 2007), which does not derive regret bounds, but only considers the convergence behavior of the TLP algorithm). Our algorithm (shown in Figure 1) does not try to identify the *best* color c^* but instead forms a set of good colors C_t. A distance parameter β determines what the distance between the best color c^* and another color c should be in order to consider c to be suboptimal and exclude it from C_t. Unlike (Pandey et al., 2007) we do not maintain a single estimate value for each color c, but calculate a confidence interval for each color that w.h.p. contains the mean reward of the best arm of color c.

2.2 Analysis

In order to derive an upper bound on the regret of the algorithm, we will consider (i) for how many steps suboptimal colors are included in C_t, and (ii) when C_t contains only close to optimal colors, how often will a suboptimal arm be chosen? Question (ii) is answered by original UCB1 analysis taken from (Auer et al., 2002). For question (i) this has to be adapted. Let $r^+(c) := \max_{a:c(a)=c} r(a)$ and $r^-(c) := \min_{a:c(a)=c} r(a)$. We assume that at each step t of the algorithm

$$r^+(c) \geq \hat{r}_t(c) - \mathrm{conf}_t(c), \text{ and} \quad (1)$$

$$r^-(c) \leq \hat{r}_t(c) + \mathrm{conf}_t(c) \quad (2)$$

for each color c, and that

$$\hat{r}_t(a_t) \leq r(a_t) + \sqrt{\tfrac{\log(t^3/\delta)}{2n_t(a_t)}}, \text{ and} \quad (3)$$

$$\hat{r}_t(a^*) \geq r(a^*) - \sqrt{\tfrac{\log(t^3/\delta)}{2n_t(a^*)}}, \quad (4)$$

where $a^* = \arg\max_{a\in A} r(a)$ is an arm with maximal mean reward r^*. Application of Hoeffding's inequality shows that (1) as well as (2) holds with probability at least $1-\frac{\delta}{|C|t^3}$ for a fixed time step t, a fixed color

Input: A confidence parameter $\delta \in (0,1)$, and a distance parameter $\beta \in (0,1)$.

Initialization

For each color $c \in C$ sample an action $a \in A$ with $c(a) = c$.

For time steps $t = 1, 2, \ldots$ **do**

▷ *Calculate confidence intervals $I_t(c)$ for each color:*
For each color c in C calculate a confidence interval for $\max_{a:c(a)=c} r(a)$:

$$\hat{I}_t(c) := \left[\hat{r}_t(c) - \mathrm{conf}_t(c), \hat{r}_t(c) + \theta + \mathrm{conf}_t(c)\right],$$

where $\hat{r}_t(c) = \frac{1}{n_t(c)} \sum_{\tau<t:c(a_\tau)=c} r_\tau$ with r_τ being the random reward obtained at step τ for choosing arm a_τ, $n_t(a)$ being the number of times action a was chosen, and $n_t(c) := \sum_{a:c(a)=c} n_t(a)$ being the number of times an action with color c has been chosen. Further,

$$\mathrm{conf}_t(c) := \sqrt{\frac{\log(|C|t^3/\delta)}{2n_t(c)}}.$$

▷ *Determine relevant colors C_t:*
Let $c_t := \arg\max_{c \in C} \left\{\hat{r}_t(c) + \theta + \mathrm{conf}_t(c)\right\}$ be the color with maximal upper confidence bound value, and set $C_t := \left\{c \in C \,|\, \hat{I}_t(c) \cap \hat{I}_t(c_t) \neq \varnothing\right\}$. If $\mathrm{conf}_t(c_t) \geq \beta/4$ and $\mathrm{conf}_t(c) < \beta/4$ for some $c \in C_t$, reset $C_t := \{c_t\}$.

▷ *Arm selection:*
Use UCB1 to choose an arm from $A_t := \{a \in A \,|\, c(a) \in C_t\}$, i.e., if there is an unsampled arm a in A_t choose a, otherwise choose

$$a_t := \arg\max_{a \in A_t} \left\{\hat{r}_t(a) + \sqrt{\frac{\log(t^3/\delta)}{2n_t(a)}}\right\},$$

where $\hat{r}_t(a) = \frac{1}{n_t(a)} \sum_{\tau<t:a_\tau=a} r_\tau$.

Figure 1: The colored bandits algorithm.

c and a fixed value of $n_t(c)$. Thus, a union bound over all colors, all possible values of $n_t(c)$ and all t shows that (1) and (2) hold with probability at least $1 - 2\sum_t \frac{\delta}{t^2} > 1 - \frac{10}{3}\delta$ for all t. Similarly, (3) and (4) hold with probability at least $1 - \frac{10}{3}\delta$ for all t.

Note that under assumptions (1) and (2) an optimal color c^* (i.e., the color of an optimal arm a^*) is always in C_t, since

$$\hat{r}_t(c^*) + \mathrm{conf}_t(c^*) + \theta \geq r^-(c^*) + \theta \geq r^+(c^*) \geq r^+(c_t) \geq \hat{r}_t(c_t) - \mathrm{conf}_t(c_t).$$

Now, we establish sample complexity bounds both on (i) the number of times an arm of a color that is β-far from the optimal color c^* is chosen, and (ii) the number of times a suboptimal arm is chosen from C_t (assuming that C_t contains only colors β-close to the optimal color). For the bound on (ii), we may directly refer to (Auer et al., 2002), where it is shown that any suboptimal arm a is chosen at most $1 + \frac{8\log T}{(r^* - r(a))^2}$ times (w.h.p). As playing such an arm gives regret $r^* - r(a)$, this yields a bound of

$$\sum_{c \in C_\beta} \sum_{\substack{a:c(a)=c \\ r(a)<r^*}} \left(1 + \frac{8\log T}{r^* - r(a)}\right),$$

where $C_\beta := \{c \in C \,|\, r^* - r^+(c) \leq \beta + 2\theta\}$ is the set of colors that are β-close to the optimal reward r^*.

For a bound on (i) we may easily adapt the mentioned proof as follows. Consider a β-*bad* color $c \notin C_\beta$. Then $r^+(c) + \beta + 2\theta < r^*$. According to the algorithm $c \in C_t$ only when

$$\hat{r}_t(c) + \mathrm{conf}_t(c) + \theta \;\geq\; \hat{r}_t(c_t) - \mathrm{conf}_t(c_t). \quad (5)$$

Further, if $\mathrm{conf}_t(c) < \beta/4$ then $c \in C_t$ only in case $\mathrm{conf}_t(c_t) < \beta/4$, too. But then we have from (1), (5), and the fact that $r^* \leq \hat{r}_t(c_t) + \mathrm{conf}_t(c_t) + \theta$

$$\begin{aligned} r^+(c) \;+\beta + 2\theta \;&\geq\; \hat{r}_t(c) - \mathrm{conf}_t(c) + \beta + 2\theta \\ &\geq\; \hat{r}_t(c_t) - 2\mathrm{conf}_t(c) - \mathrm{conf}_t(c_t) + \beta + \theta \\ &\geq\; r^* - 2\mathrm{conf}_t(c_t) - 2\mathrm{conf}_t(c) + \beta \;>\; r^*, \end{aligned}$$

contradicting our assumption that c is a β-bad color. Hence, whenever $\mathrm{conf}_t(c) < \beta/4$ we have $c \notin C_t$, so that $c(a_t) = c$ only at $\leq \left\lceil \frac{8\log(|C|T^3/\delta)}{\beta^2} \right\rceil$ time steps. Further we have to consider the case when setting $C_t := \{c_t\}$, which may be a suboptimal choice as well. However, this happens only when $\mathrm{conf}_t(c_t) \geq \beta/4$, that is, not more often than $\left\lceil \frac{8\log(|C|T^3/\delta)}{\beta^2} \right\rceil$ times. Summarizing (and also taking into account the regret of the initialization), we get the following result.

Theorem 2. *The regret of the colored bandits algorithm after T steps with probability at least $1 - \frac{20}{3}\delta$ is at most*

$$|C| + 2|C| \left\lceil \frac{8\log(|C|T^3/\delta)}{\beta^2} \right\rceil + \sum_{\substack{a:c(a) \in C_\beta \\ r(a)<r^*}} \left(1 + \frac{8\log T}{r^* - r(a)}\right).$$

As the regret at each step is at most 1, we may simply sum up the error probabilities for failing confidence intervals given in (1) and (3) to obtain a bound on the expected regret as well.

These bounds show that it is possible for the learner to exploit color information in order to eliminate the dependency on the total number of actions in the respective regret bounds.

3 COLORED ACTION MDPS

We continue dealing with the natural generalization of the problem to Markov decision processes. A *Markov decision process (MDP)* is a tuple $\mathcal{M} = \langle S, A, p, r \rangle$, where S is a finite set of *states* and A is a finite set of *actions*. Unlike in the usual setting where in each state from S each action from A is available, we consider that for each state s there is a nonempty subset $A(s) \subseteq A$ of actions available in s. Further, we assume that the sets $A(s)$ are a partition of A, i.e., $A(s) \cap A(s') = \varnothing$ for $s \neq s'$, and $\bigcup_{s \in S} A(s) = A$. The transition probabilities $p(s'|s,a)$ give the probability of reaching state s' when choosing action a in state s, and the payoff distributions with mean $r(s,a)$ and support in $[0,1]$ specify the random reward obtained for choosing action a in state s.

We are interested in the undiscounted, average reward $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r_t$, where r_t is the random reward obtained at step t. As possible strategies we consider (stationary) *policies* $\pi : S \to A$, where $\pi(s) \in A(s)$. This is justified by the fact that there is always such a policy π^* which gives optimal reward (Puterman, 1994). Let $\rho(\pi)$ denote the expected average reward of policy π. Then π^* is an optimal policy if $\rho(\pi) \leq \rho(\pi^*) =: \rho^*$ for all policies π.

In the analysis we will also need some transition parameters of the MDP at hand. Thus let $T(s'|\mathcal{M}, \pi, s)$ be the first (random) time step in which state s' is reached when policy π is executed on MDP $\mathcal{M}$ with initial state s. Then we define the *diameter* of the MDP to be the average time it takes to move from any state s to any other state s', using an appropriate policy, i.e.

$$D(\mathcal{M}) := \max_{s \neq s' \in S} \min_{\pi : S \to A} \mathbb{E}\big(T(s'|\mathcal{M}, \pi, s)\big).$$

We will consider only MDPs with finite diameter, which guarantees that there is always an optimal policy that achieves optimal average reward ρ^* independent of the initial state. Note that each policy π induces a Markov chain $\mathcal{M}_\pi$ on $\mathcal{M}$. If the Markov chain is *ergodic* (i.e., each state is reachable from each other state after a finite number of steps) this Markov chain has a state-independent *stationary distribution*. The *mixing time* of a policy π on an MDP $\mathcal{M}$ with induced stationary distribution μ_π is given by

$$\kappa_\pi(\mathcal{M}) := \sum_{s' \in S} \mathbb{E}\big(T(s'|\mathcal{M}, \pi, s)\big) \mu_\pi(s')$$

for an arbitrary $s \in S$. The definition is independent of the choice of s as shown in (Hunter, 2006). Finally, we remark that in case a policy π induces an ergodic Markov chain on $\mathcal{M}$ with stationary distribution μ_π, the *average reward of* π can be written as

$$\rho(\pi) := \sum_{s \in S} \mu_\pi(s)\, r(s, \pi(s)). \tag{6}$$

3.1 When the Color Determines the Transition Probabilities

The simplest case in the MDP scenario is when two actions of the same color have identical transition probabilities and close rewards, given a coloring function $c : A \to C$ on the action space.

Assumption 3. *There is a $\theta > 0$ such that for each two actions $a, a' \in A(s)$ with $s \in S$: If $c(a) = c(a')$ then (i) $p(\cdot|s,a) = p(\cdot|s,a')$, and (ii) $|r(s,a) - r(s,a')| < \theta$.*

We will try to exploit the color information only for same-colored actions in the same state, so that we may assume without loss of generality that actions available in distinct states have distinct colors. Thus the set of colors C is partitioned by the sets $C(s)$ of colors of actions that are available in state s. Further, we have to distinguish between colors having distinct transition probability distributions. Thus, we write $C(s,p)$ for the set of colors of actions available in state s and having transition probabilities $p(\cdot)$.

3.1.1 Algorithm

The algorithm we propose (shown in Figure 2) is a straightforward adaptation of the UCRL2 algorithm (Auer et al., 2009). For the sake of simplicity, we only consider the case where the transition structure of the MDP is known. The general case can be handled analogously to (Auer et al., 2009). We use the color information just as in the bandit case. That is, in each state a set of promising colors is defined. Then an optimal policy is calculated where the action set is restricted to actions with promising colors, and the actions' rewards are set to their upper confidence values. The algorithm proceeds in episodes i, and the chosen policy π_i is executed until a state s is arrived where the action $\pi_i(s)$ has been played *in* the episode as often as *before* the episode.

3.1.2 Analysis

Again we are interested in the algorithm's regret after T steps, defined as[1] $T\rho^* - \sum_t r_t$. Furthermore, we also consider the regret with respect to an ε-optimal policy, i.e., with respect to $\rho^* - \varepsilon$ instead of ρ^*. The analysis of the algorithm's regret is a combination of the respective proofs of Theorem 2 and of the logarithmic regret bounds in (Auer et al., 2009). That is, we first establish a sample complexity bound on the number of steps in episodes where

[1] Unlike in the bandit case, this regret definition also considers the deviations of the achieved rewards from the mean rewards. Actually, the regret bounds for the bandit case can be adapted to this alternative regret definition.

Input: A confidence parameter $\delta \in (0,1)$, and a distance parameter $\beta \in (0,1)$.

Notation: Let t denote the current time step.

For episodes $i = 1,2,\ldots$ **do**

▷ *Initialize episode i:*
Set $t_i := t$. Calculate estimates $\hat{r}_t(s,a)$ for the mean reward $r(s,a)$ for state-action pairs (s,a) with $a \in A(s)$, and determine confidence intervals $I_t(c)$ for each color as follows. For each state s, each transition probability distribution $p(\cdot)$ and each color c in $C(s,p)$ calculate a confidence interval for $\max_{a:c(a)=c} r(s,a)$:

$$\hat{I}_t(s,c) := \big[\hat{r}_t(s,c) - \mathrm{conf}_t(s,c), \hat{r}_t(c) + \theta + \mathrm{conf}_t(s,c)\big],$$

where $\hat{r}_t(s,c) = \frac{1}{n_t(s,c)} \sum_{\tau: c(a_\tau)=c, s_\tau = s} r_\tau$ with r_τ being the random reward obtained at step τ for choosing action a_τ in state s, $n_t(s,a)$ being the number of times action a was chosen in state s, and $n_t(s,c)$ being the number of times an action with color c has been chosen in s.[2] Further,

$$\mathrm{conf}_t(s,c) := \sqrt{\tfrac{7\log(2|C|t/\delta)}{2n_t(s,c)}}.$$

▷ *Determine relevant colors C_t:*
For each state s and each $p(\cdot)$ let $c_t(s,p)$ be the color with maximal upper confidence bound value, i.e.,

$$c_t(s,p) := \operatorname*{arg\,max}_{c \in C(s,p)} \big\{\hat{r}_t(s,c) + \theta + \mathrm{conf}_t(s,c)\big\}.$$

Set $C_t(s,p) := \big\{c \in C(s,p) \,|\, \hat{I}(s,c_t(s,p)) \cap \hat{I}(s,c) \neq \varnothing\big\}$. If $\mathrm{conf}_t(s,c_t(s,p)) \geq \beta/4$ and $\mathrm{conf}_t(s,c) < \beta/4$ for some $c \in C_t(s,p)$, reset $C_t(s,p) := \{c_t(s,p)\}$.

▷ *Policy selection and execution:*
Choose an optimal policy[3] π_i in the MDP with transition structure as given and action sets $A_t(s) := \{a \in A(s) \,|\, c(a) \in \bigcup_p C_t(s,p)\}$ with rewards

$$\widetilde{r}_t(s,a) := \hat{r}_t(s,a) + \sqrt{\tfrac{7\log(2|A|t/\delta)}{2n_t(s,a)}}.$$

Play π_i as long as $n_t(s,\pi_i(s_t)) < 2n_{t_i}(s,\pi_i(s_t))$ in the current state s_t.

Figure 2: The colored MDP algorithm for MDPs with known transition structure and colored action set.

[2] If the color or action count is 0, reset it to 1.

[3] Such an optimal policy can be calculated using ordinary value iteration (Puterman, 1994).

ε-suboptimal reward is received. Thus, let T_ε be the number of steps in episodes where the average per-step reward is less than $\rho^* - \varepsilon$, and let M_ε be the respective indices of these episodes. Note that setting $\Delta_i := \sum_{t=t_i}^{t_{i+1}-1} (\rho^* - r_t)$ to be the regret in episode i we have that

$$\Delta_\varepsilon := \textstyle\sum_{i \in M_\varepsilon} \Delta_i \;\geq\; \varepsilon T_\varepsilon. \tag{7}$$

Having this lower bound on Δ_ε, we now aim at achieving also an *upper* bound on Δ_ε in terms of T_ε. These two bounds then will give us the desired regret bound. The main part of the derivation of this upper bound is mainly the same as given in the extended version of (Auer et al., 2009), so we will not repeat it here and only state that it can be shown that

$$\Delta_\varepsilon \;\leq\; 1 + \sqrt{\tfrac{T_\varepsilon}{2}\log\tfrac{T}{\delta}} + 2D\sqrt{T_\varepsilon \log\tfrac{T}{\delta}} + D \cdot \#\text{episodes} + \sqrt{\log\tfrac{2|A|T}{\delta}} \sum_{s,a} \sqrt{n_\varepsilon(s,a)}. \tag{8}$$

with probability $1 - 3\delta$, where $n_\varepsilon(s,a)$ is the total number of times action a is chosen in s in episodes in M_ε. Now we split $\sum_{s,a} \sqrt{n_\varepsilon(s,a)}$ into one sum handling the actions of β-bad color $\notin C_\beta$ and another sum for all other actions, where

$$C_\beta := \textstyle\bigcup_{s,p} \big\{c \in C(s,p) \,|\, r^*(s,p) - r^+(c) \geq \beta + 2\theta\big\}$$

with $r^*(s,p) := \max_{a:c(a) \in C(s,p)} r(a)$ and $r^+(c) := \max_{a:c(a)=c} r(a)$. Then similarly to the bandit case, whenever the confidence interval $\mathrm{conf}_t(s,c)$ of such a β-bad color c is smaller than $\beta/4$ at the beginning of an episode, the respective color will not be part of $C_t(s,p)$. Consequently, by definition of $\mathrm{conf}_t(s,c)$ the number of times an action with that color is chosen in state s is upper bounded by $\frac{112\log(2|C|T)}{\beta^2}$, the additional factor 2 stemming from the fact that the confidence intervals are only updated at the beginning of an episode (in which the number of times the respective action is chosen may be doubled). Consequently, for any β-bad color $c \notin C_\beta$

$$\sum_{s,a:c(a)=c} n_\varepsilon(s,a) \;\leq\; \tfrac{112\log(2|C|T)}{\beta^2}, \tag{9}$$

whence one obtains by Jensen's inequality that

$$\sum_{s,a} \sqrt{n_\varepsilon(s,a)} \;\leq\; \sqrt{|A_\beta| T_\varepsilon} + \tfrac{11}{\beta}\sqrt{|A||C|\log(2|C|T)},$$

where $A_\beta := \{a \in A \,|\, c(a) \in C_\beta\}$. This yields from (8) that

$$\Delta_\varepsilon \;\leq\; \sqrt{\tfrac{T_\varepsilon}{2}\log\tfrac{T}{\delta}} + 2D\sqrt{T_\varepsilon\log\tfrac{T}{\delta}} + D\cdot\#\text{episodes} + \sqrt{|A_\beta| T_\varepsilon \log\tfrac{2|A|T}{\delta}} + \tfrac{11}{\beta}\sqrt{|A||C|\log\tfrac{2|A|T}{\delta}\log(2|C|T)} + 1, \tag{10}$$

so that it remains to upper bound the number of episodes. By the doubling criterion for episode termination it is easy to see that generally there are not more than $|A|\log_2 \frac{8T}{|A|}$ episodes (cf. Appendix A.2 of the extended version of (Auer et al., 2009) for details). However, again considering actions of β-good color $\in C_\beta$ and others separately, according to (9) this can be improved to a bound of $|A_\beta|\log\frac{8T}{|A_\beta|} + |A|\log\frac{896\log(2|C|T)}{|A|\beta^2}$. Putting this into (10) we get in combination with (7)

$$T_\varepsilon \;\le\; c_1\cdot\frac{(D^2+|A_\beta|)\log(T/\delta)}{\varepsilon^2} + c_2\cdot\frac{\sqrt{|A||C|}\log\log(T/\delta)}{\varepsilon\beta} + c_3\cdot\frac{D|A_\beta|\log T + D|A|\log\log T}{\varepsilon}. \tag{11}$$

As Δ_ε is an upper bound on the regret with respect to an ε-optimal policy, we may plug (11) into (10) to obtain after some calculations the following result.

Theorem 4. *The regret of the colored MDP algorithm with respect to an ε-optimal policy after T steps is with probability at least $1-3\delta$ upper bounded by (ignoring terms sublogarithmic in T)*

$$c_1'\cdot\frac{\left(D^2+|A_\beta|\right)\log\frac{T}{\delta}}{\varepsilon} + c_2'\cdot\frac{\sqrt{|A||C|}\log\frac{T}{\delta}}{\beta} + c_3'\cdot\frac{(D+\sqrt{|A_\beta|})\sqrt[4]{|A||C|}\log\frac{T}{\delta}}{\sqrt{\beta\varepsilon}} + D|A_\beta|\log T.$$

For sufficiently small ε, an ε-optimal policy is also optimal, which yields a corresponding bound with ε replaced by the difference between the optimal and the highest suboptimal average reward, i.e., $g := \rho^* - \max_{\pi:\rho(\pi)<\rho^*}\rho(\pi)$.

Thus, as in the bandit case the learner can benefit from the color information (as can be seen when comparing the bounds to the case without color information, i.e. $|C|=1$ and $A_\beta = A$). The reason why there is still some dependency on the total number of actions is that the doubling criterion for episode termination concerns the actions and not their colors. However, adapting the episode termination criterion to apply to colors instead of actions, some other parts of the proof do not work anymore.

3.2 ACTION AGGREGATION

Now let us consider the case where actions of the same color have not identical but only similar transition probabilities. As before we are interested only in similar actions that are available in the same state. Thus we again assume for the sake of simplicity that only actions contained in the same set $A(s)$ have the same color.

Assumption 5. *There are $\theta_r,\theta_p > 0$ such that for each two actions $a,a' \in A(s)$ with $s\in S$: If $c(a)=c(a')$ then (i) $\left|r(s,a)-r(s,a')\right| < \theta_r$, and (ii) $\sum_{s'\in S}\left|p(s'|s,a)-p(s'|s,a')\right| < \theta_p$.*

Unlike in the settings considered so far, it is by no means clear what happens if one simply chooses a representative of each color and works on the *aggregated MDP*. In this section we derive error bounds that answer this question.

Definition 6. *Given an MDP $\mathcal{M}=\langle S,A,p,r\rangle$ and a coloring function $c: A\to C$ for the actions, an MDP $\widehat{\mathcal{M}}=\langle S,C,\widehat{p},\widehat{r}\rangle$ is called an* aggregation of $\mathcal{M}$ with respect to c *if for $a\in A(s)$ with $c(a)=c$: $\left|\widehat{r}(s,c)-r(s,a)\right| < \theta_r$, and $\sum_{s'\in S}\left|\widehat{p}(s'|s,c)-p(s'|s,a)\right| < \theta_p$.*

Thus beside picking an arbitrary reference action a for each color c one may also set e.g. $\widehat{r}(s,c) := \frac{1}{|A_c|}\sum_{a\in A_c} r(s,a)$ and $\widehat{p}(s'|s,c) := \frac{1}{|A_c|}\sum_{a\in A_c} p(s'|s,a)$, where $A_c := \{a\in A \,|\, c(a)=c\}$.

A policy $\widehat{\pi}$ on $\widehat{\mathcal{M}}$ is the *aggregation of a policy π on $\mathcal{M}$ with respect to a coloring function $c: A\to S$*, if $c(\pi(s)) = \widehat{\pi}(s)$ for all states s. In the following we consider only *ergodic* MDPs, where all policies induce ergodic Markov chains. However, if an aggregation conserves the ergodicity structure of the MDP the following results can be adapted to the general case.

Theorem 7. *Let $\mathcal{M}=\langle S,A,p,r\rangle$ be an MDP and $\widehat{\mathcal{M}}=\langle S,C,\widehat{p},\widehat{r}\rangle$ an aggregation of $\mathcal{M}$ with respect to a coloring function $c: A\to C$. Then for each policy π on $\mathcal{M}$ and the aggregation $\widehat{\pi}$ of π on $\widehat{\mathcal{M}}$ with respect to c, we have for the difference of the average reward $\rho(\pi)$ of π in $\mathcal{M}$ and the average reward $\widehat{\rho}(\widehat{\pi})$ of $\widehat{\pi}$ in $\widehat{\mathcal{M}}$*

$$\left|\rho(\pi)-\widehat{\rho}(\widehat{\pi})\right| \;<\; \theta_r + (\kappa_\pi - 1)\,\theta_p,$$

where κ_π is the mixing time of the Markov chain induced by π on $\mathcal{M}$.

As for the bounds on the error of state aggregation of (Ortner, 2007), in the proof of Theorem 7 we use the following result of (Hunter, 2006) on perturbations of Markov chains.

Theorem 8. *(Hunter, 2006) Let C, $\widetilde{C}$ be two ergodic Markov chains on the same state space S with transition probabilities $p(\cdot,\cdot)$, $\widetilde{p}(\cdot,\cdot)$ and stationary distributions μ, $\widetilde{\mu}$, and let κ_C be the mixing time of C. Then*

$$\left\|\mu-\widetilde{\mu}\right\|_1 \;\le\; (\kappa_C-1)\max_{s\in S}\sum_{s'\in S}\left|p(s,s')-\widetilde{p}(s,s')\right|.$$

Proof of Theorem 7: Writing μ_π and $\mu_{\widehat{\pi}}$ for the stationary distributions of π on $\mathcal{M}$ and $\widehat{\pi}$ on $\widehat{\mathcal{M}}$, respectively, and abbreviating $r(s,\pi(s))$ with $r_\pi(s)$, we have

by (6)

$$\begin{aligned}&\left|\rho(\pi)-\widehat{\rho}(\widehat{\pi})\right| = \left|\textstyle\sum_s \mu_\pi(s)\, r_\pi(s) - \sum_s \mu_{\widehat{\pi}}(s)\, r_{\widehat{\pi}}(s)\right| \\ &\leq \textstyle\sum_s \left|\mu_\pi(s)-\mu_{\widehat{\pi}}(s)\right| r_\pi(s) + \sum_s \mu_{\widehat{\pi}}(s)\left|r_{\widehat{\pi}}(s)-r_\pi(s)\right|.\end{aligned}$$

As $r_\pi(s) \leq 1$, using Theorem 8 together with our assumptions on aggregation gives

$$\begin{aligned}\left|\rho(\pi)-\widehat{\rho}(\widehat{\pi})\right| &< \sum_{s\in S}\left|\mu_\pi(s)-\mu_{\widehat{\pi}}(s)\right| + \sum_{s\in S}\mu_{\widehat{\pi}}(s)\,\theta_r \\ &< (\kappa_\pi - 1)\theta_p + \theta_r. \qquad \square\end{aligned}$$

Corollary 9. *Let π^* be an optimal policy on an MDP $\mathcal{M}$ with optimal average reward ρ^*, and let $\widehat{\pi}^*$ be an optimal policy with optimal average reward $\widehat{\rho}^*$ on an aggregation $\widehat{\mathcal{M}}$ of $\mathcal{M}$ with respect to some coloring function $c : A \to C$. Then*

$$\begin{aligned}&\textit{(i)} \quad |\rho^* - \widehat{\rho}^*| < \theta_r + (\kappa_{\mathcal{M}} - 1)\,\theta_p, \\ &\textit{(ii)} \quad \rho^* < \rho(\widehat{\pi}^{*e}) + 2\theta_r + 2(\kappa_{\mathcal{M}} - 1)\,\theta_p,\end{aligned}$$

where $\kappa_{\mathcal{M}} := \max_\pi \kappa_\pi$, and $\widehat{\pi}^{*e}$ is any policy such that $\widehat{\pi}^*$ is the aggregation of $\widehat{\pi}^{*e}$ with respect to c.

Remark: As the role of $\mathcal{C}$ and $\widetilde{\mathcal{C}}$ in Theorem 8 is symmetric, Theorem 7 and Corollary 9 hold also when the mixing time of $\mathcal{M}$ is replaced with the mixing time of the aggregated MDP. Hence, the results also hold for the minimum of the two mixing times.

The following theorem shows that the error in average reward indeed becomes arbitrarily large when the mixing time approaches infinity.

Theorem 10. *For each $\theta_p > 0$ and each sufficiently small $\eta > 0$ there is an MDP $\mathcal{M} = \langle S, A, p, r\rangle$ and a coloring $c : A \to C$ of the action space such that in each aggregation $\widehat{\mathcal{M}}$ of $\mathcal{M}$ with respect to c there is some policy π on $\mathcal{M}$ such that for the respective aggregated policy $\widehat{\pi}$ on $\widehat{\mathcal{M}}$,*

$$|\rho(\pi) - \widehat{\rho}(\widehat{\pi})| \geq 1 - \eta.$$

Proof. Fix some $\theta_p > 0$ and consider for $\delta \in (0, \frac{\theta_p}{2})$ an MDP with $S = \{s_1, s_2\}$, $A(s_1) = \{a_1, a_2\}$, and $A(s_2) = \{a_3\}$. Define the transition probabilities (cf. Figure 3) in s_1 as $p(s_1|s_1,a_1) = 1-\delta$, $p(s_2|s_1,a_1) = \delta$, $p(s_1|s_1,a_2) = 1 - \frac{\delta}{n^2}$, and $p(s_2|s_1,a_2) = \frac{\delta}{n^2}$, and those in s_2 as $p(s_1|s_2,a_3) = \frac{\delta}{n}$ and $p(s_2|s_2,a_3) = 1 - \frac{\delta}{n}$, where $n \in \mathbb{N}$. Then the stationary distribution of the policy π with $\pi(s_1) = a_2$ and $\pi(s_2) = a_3$ is $\mu_\pi = \left(\frac{n}{n+1}, \frac{1}{n+1}\right)$, which for $n \to \infty$ converges to $(1,0)$. On the other hand, the policy π' with $\pi'(s_1) = a_1$ and $\pi'(s_2) = a_3$ has stationary distribution $\mu_{\pi'} = \left(\frac{1}{n+1}, \frac{n}{n+1}\right)$, which for $n \to \infty$ converges to $(0,1)$.

Now, as $\delta < \frac{\theta_p}{2}$, a coloring function c may assign (choosing an arbitrary $\theta_r > 0$, cf. the choice of the rewards below) the same color c_1 to actions a_1 and a_2 in s_1. We consider transition probabilities $\widehat{p}(\cdot|s_1,c_1)$ which are a convex combination of the respective probabilities $p(\cdot|s_1,a_1)$ and $p(\cdot|s_1,a_2)$, i.e.

$$\begin{aligned}\widehat{p}(s_2|s_1,c_1) &= \lambda\delta + (1-\lambda)\tfrac{\delta}{n^2} \text{ and} \\ \widehat{p}(s_1|s_1,c_1) &= 1 - \lambda\delta - (1-\lambda)\tfrac{\delta}{n^2}\end{aligned}$$

for $\lambda \in [0,1]$. (For transition probabilities outside this convex set the error will clearly be larger.) Then the stationary distribution of the aggregated policy $\widehat{\pi}$ with $\widehat{\pi}(s_1) = c_1$ and $\widehat{\pi}(s_2) = a_3$ is $\mu_{\widehat{\pi}} = \left(\frac{n}{n^2\lambda+n-\lambda+1}, \frac{n^2\lambda-\lambda+1}{n^2\lambda+n-\lambda+1}\right)$. Thus, if $\lambda > 0$ then $\mu_{\widehat{\pi}}$ converges to $(0,1)$ for $n \to \infty$. Otherwise for $\lambda = 0$ we have $\mu_{\widehat{\pi}} = \left(\frac{n}{n+1}, \frac{1}{n+1}\right)$, which for $n \to \infty$ converges to $(1,0)$. Now setting the rewards $r(s_1,a_1) := r(s_1,a_2) := 1$ and $r(s_2,a_3) := 0$ (and choosing appropriate rewards $\widehat{r}$) we obtain an error arbitrarily close to 1 either with respect to π or to π', which proves the theorem. $\square$

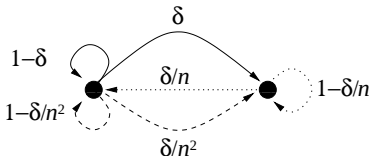

Figure 3: The MDP in the proof of Theorem 10. Solid arrows correspond to action a_1, dashed ones to a_2, and dotted ones to a_3.

The lower bound of Theorem 10 indicates that exploiting similarity of transition probabilities is harder than for rewards. Here we confirm this by showing that an optimistic algorithm in the style of UCRL2 fails already in very simple situations.

Example: Consider a two state MDP with $S = \{s_1, s_2\}$, $A(s_1) = \{a_1, a_2\}$, and $A(s_2) = \{s_3\}$ as shown in Figure 4. The transition probabilities in s_1 are set to $p(s_1|s_1,a_1) = \frac{3}{4}, p(s_2|s_1,a_1) = \frac{1}{4}, p(s_1|s_1,a_2) = \frac{1}{2}$, and $p(s_2|s_1,a_2) = \frac{1}{2}$. In state s_2 the transition probabilities are $p(s_1|s_2,a_3) = p(s_2|s_2,a_3) = \frac{1}{2}$. The mean rewards are given by $r(s_1,a_1) = \frac{1}{2}$, $r(s_1,a_2) = \frac{5}{12}$, and $r(s_2,a_3) = \frac{3}{4}$. Thus, setting $\theta_p > 1$ and $\theta_r > \frac{1}{12}$ we may color actions a_1 and a_2 with the same color c_1. The intervals of possible values of actions with color c_1 are then $\widetilde{p}(s_1|s_1,c_1) \in \left[\frac{1}{2}, \frac{3}{4}\right]$, $\widetilde{p}(s_2|s_1,c_1) \in \left[\frac{1}{4}, \frac{1}{2}\right]$, and $\widetilde{r}(s_1,c_1) \in \left[\frac{5}{12}, \frac{1}{2}\right]$.

An optimistic algorithm that handles this information as in a *bounded parameter MDP* (Tewari and Bartlett, 2007) now would assume values $\widetilde{p}(s_1|s_1,c_1) = \widetilde{p}(s_2|s_1,c_1) = \frac{1}{2}$ and $\widetilde{r}(s_1,c_1) = \frac{1}{2}$ in order to maximize the average reward of a policy playing an action with color c_1 in state s_1, which then is

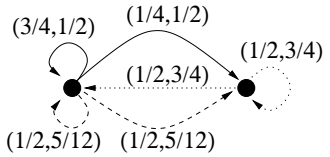

Figure 4: The MDP from the example. Solid arrows correspond to action a_1, dashed ones to a_2, and dotted ones to a_3. For each arrow the respective transition probability and the reward are displayed.

$\widetilde{\rho} = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{3}{4} = \frac{5}{8}$, since the stationary distribution in this case is obviously $(\frac{1}{2}, \frac{1}{2})$. However, the real achievable average reward for actions a_1 (that gives stationary distribution $(\frac{2}{3}, \frac{1}{3})$) and a_2 (with stationary distribution $(\frac{1}{2}, \frac{1}{2})$) of color c_1 is $\rho(a_1) = \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{3}{4} = \frac{7}{12}$ and $\rho(a_2) = \frac{1}{2} \cdot \frac{5}{12} + \frac{1}{2} \cdot \frac{3}{4} = \frac{7}{12}$.

Hence, if we add another action a_4 to state s_1 with $p(s_1|s_1,a_4) = 1$ and $r(s_1,a_4) \in \left(\frac{7}{12}, \frac{5}{8}\right)$, the optimal policy would choose a_4 in s_1. However, as the algorithm expects the larger average reward of $\widetilde{\rho}$, action a_4 would not be chosen.

4 DISCUSSION AND PROBLEMS

Our aim was to investigate a simplest possible similarity model for discrete bandits / action spaces. We think that the shift of analysis from the bandit to the MDP case may serve as a blueprint for much more general settings where the action space of an MDP is e.g. a metric space with Lipschitz condition, when an appropriate bandit algorithm like the zooming algorithm (Kleinberg et al., 2008) may be similarly adapted. Indeed, this particular setting would be a proper generalization of the scenario considered in this paper, as the zooming algorithm can handle the colored bandits case by defining a special metric d, where $d(a,a') := 0$ if $c(a) = c(a')$ and $d(a,a') := 1$ otherwise (although it is not quite clear whether the same regret bounds are achievable). However, these considerations need further investigation.

We have tried to exploit similarity information only with respect to the actions. As many real-world problems (also) have large state spaces, it is a natural question whether a similar approach would work for coloring the state space of an MDP. However, it is not quite clear how similarity for states could be used in principle. The most natural thing to do would be to choose in a state s (of color c) that has not been visited before an action that proved to be successful in other states of the same color. This would obviously lead to some sort of state aggregation similarly to the action aggregation concept considered in Section 3.2. However, in this setting lower bounds that resemble that of Theorem 10 have already been shown that state aggregation may cause arbitrarily large error as well (Ortner, 2007). Still, regret bounds that consider mixing time parameters of the MDP may be possible.

ACKNOWLEDGMENTS

The author is grateful for the input of the anonymous reviewers of an earlier version of (parts of) this paper. In particular, the remark before Theorem 8 and the ideas about the zooming algorithm are due to two reviewers. This work was supported in part by the Austrian Science Fund FWF (S9104-N13 SP4). The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 216886 (PASCAL2), and n° 216529 (PinView). This publication only reflects the authors' views.

REFERENCES

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47:235–256.

Auer, P., Jaksch, T., and Ortner, R. (2009). Near-optimal regret bounds for reinforcement learning. In *Adv. Neural Inf. Process. Syst. 21*, pages 89–96. (full version http://www.unileoben.ac.at/∼infotech/publications/TR/CIT-2009-01.pdf).

Hunter, J. J. (2006). Mixing times with applications to perturbed Markov chains. *Linear Algebra Appl.*, 417:108–123.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings STOC 2008*, pages 681–690.

Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *Adv. Neural Inf. Process. Syst. 17*, pages 697–704.

Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648.

Ortner, R. (2007). Pseudometrics for state aggregation in average reward Markov decision processes. In *Proceedings of ALT 2007*, pages 373–387.

Pandey, S., Chakrabarti, D., and Agarwal, D. (2007). Multi-armed bandit problems with dependent arms. In *Proceedings of ICML 2007*, pages 721–728.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA.

Tewari, A. and Bartlett, P. L. (2007). Bounded parameter Markov decision processes with average reward criterion. In *Proceedings of COLT 2007*, pages 263–277.