Variational Regret Bounds for Reinforcement Learning

Ronald Ortner rortner@unileoben.ac.at
Pratik Gajane pratik.gajane@unileoben.ac.at
Peter Auer auer@unileoben.ac.at
Lehrstuhl für Informationstechnologie, Montanuniversität Leoben, Austria

Abstract

We consider undiscounted reinforcement learning in Markov decision processes (MDPs) where *both* the reward functions and the state-transition probabilities may vary (gradually or abruptly) over time. For this problem setting, we propose an algorithm and provide performance guarantees for the regret evaluated against the optimal non-stationary policy. The upper bound on the regret is given in terms of the total *variation* in the MDP. This is the first variational regret bound for the general reinforcement learning setting.

1 INTRODUCTION

A Markov decision process (MDP) is a discrete-time state-transition system in which the transition dynamics follow the Markov property (Puterman [1994], Bertsekas and Tsitsiklis [1996]). MDPs are a standard model to express uncertainty in reinforcement learning problems. In the classical MDP model, the transition dynamics and the reward functions are time-invariant. However, such fixed transition dynamics and reward functions are insufficient to model real world problems in which parameters of the world change over time. To deal with such problems, we consider a setting in which both the transition dynamics and the reward functions may vary over time. These changes can be either abrupt or gradual. As a motivation, consider the problem of deciding which ads to place on a webpage. The instantaneous reward is the payoff when viewers are redirected to an advertiser, and the state captures the details of the current ad. With a heterogeneous group of viewers, an invariant state-transition function cannot accurately capture the transition dynamics. The instantaneous reward, dependent on external factors, is also better represented by changing reward functions. For additional motivation and further applications, see [Yuan Yu and Mannor, 2009a,b, Abbasi et al., 2013].

1.1 MAIN CONTRIBUTION

For reinforcement learning in MDPs with changes in reward functions and transition probabilities, we provide an algorithm, Variation-aware UCRL, a variant of UCRL with restarts [Jaksch et al., 2010], which restarts according to a schedule dependent on the *variation* in the MDP (defined in Section 2 below). For reinforcement learning in an MDP with S states, A actions, diameter D, and changes with a variation of V we derive for our algorithm a high-probability upper bound on the cumulative regret after T steps of $\tilde{\mathcal{O}}(V^{1/3}T^{2/3}DS\sqrt{A})$. This bound is optimal with respect to time T and variation V and complements the known regret bound of $\tilde{\mathcal{O}}(L^{1/3}T^{2/3}DS\sqrt{A})$ for UCRL with restarts in MDPs with L abrupt changes, when using a restart schedule dependent on L [Jaksch et al., 2010]. In case when reward functions and transition probabilities change gradually, the latter bound becomes trivial when L is of order $T^{1/3}$, while our bound is still sublinear as long as the variation is sufficiently small. To the best of our knowledge, our bounds are the first variational bounds for the general reinforcement learning setting. So far, variational regret bounds have been derived only for simpler bandit settings [Besbes et al., 2014, Chen et al., 2019].

1.2 RELATED WORK

Nilim and El Ghaoui [2005] consider MDPs with arbitrarily changing state-transition probabilities but fixed reward functions where it is assumed that the uncertainty in the transition probabilities is state-wise independent. They provide a robust dynamic programming algorithm and prove that it is optimal with respect to the worst-case performance in terms of the expected total cost.

Even-Dar et al. [2005] and Dick et al. [2014] consider the problem of MDPs with fixed state-transition probabilities and changing reward functions and measure the performance of the learner against the best stationary policy in hindsight. Even-Dar et al. [2005] assume that the learner has complete knowledge of all the previous reward functions (i.e., also for states not visited) and provide regret bounds which depend on the mixing time. Dick et al. [2014] model learning in MDPs as an online linear optimization problem and propose solutions based on variants of mirror-descent.

Yuan Yu and Mannor [2009a] and Yuan Yu and Mannor [2009b] consider arbitrary changes in the reward functions and arbitrary, but bounded, changes in the state-transition probabilities. They also give regret bounds that scale with the proportion of changes in the state-transition kernel and which in the worst case grow linearly with time.

Abbasi et al. [2013] consider MDP problems with (oblivious) adversarial changes in state-transition probabilities and reward functions and provide an algorithm which guarantees $O(\sqrt{T})$ regret with respect to a comparison set of stationary (expert) policies.

2 SETTING

We start with collecting some basic facts about Markov decision processes (MDPs). In a (time-homogeneous) MDP $M = (\mathcal{S}, \mathcal{A}, \bar{r}, p, s_1)$ with a set $\mathcal{S}$ of S states, a set $\mathcal{A}$ of A actions the learner starts in some initial state s_1. At each time step $t = 1, 2, \ldots$ she chooses an action $a_t = a$ in the current state $s_t = s$, receives a random reward r_t with mean $\bar{r}(s, a)$ and observes a transition to the next state $s_{t+1} = s'$ according to transition probabilities $p(s'|s, a)$. Note that in a time-homogeneous MDP mean rewards and transition probabilities only depend on the current state and the chosen action.

An MDP is called *communicating*, if for any two states s, s', when starting in s it is possible to reach s' with positive probability choosing appropriate actions. In communicating MDPs we define the *diameter* to be the minimal expected time it takes to get from any state to any other state in the MDP, cf. [Jaksch et al., 2010].

For acting in an MDP one usually considers stationary policies $\pi : \mathcal{S} \to \mathcal{A}$ that fix for each state s the action $\pi(s)$ to choose. The average reward $\rho(M, \pi)$ of a stationary policy π is the limit of the expected average accumulated reward when following π, i.e.,

$$\rho(M, \pi) := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r_t\right].$$

The optimal average reward $\rho^*(M) = \max_\pi \rho(M, \pi)$ in communicating MDPs is independent of the initial state s_1 and cannot be increased when using nonstationary policies [Puterman, 1994].

In the problem setting we consider the underlying MDP is not time-homogeneous. Rather the mean rewards and transition probabilities depend on the current step t. Accordingly, we write them as $\bar{r}_t(s, a)$ and $p_t(s'|s, a)$, respectively, and denote the (time-homogeneous) MDP at step t by $M_t = (\mathcal{S}, \mathcal{A}, \bar{r}_t, p_t, s_1)$. We assume that all MDPs M_t are communicating with diameter D_t and denote by D a common upper bound on all D_t.

Obviously, in a nonstationary MDP the optimal policy in general will not be stationary anymore. We are interested in online regret bounds after any T steps taken by the learner. Accordingly, we consider the optimal expected T-step reward $v_T^*(s_1)$ that can be achieved by any (time dependent) policy when starting in s_1, and define the regret after T steps as

$$R_T := v_T^*(s_1) - \sum_{t=1}^{T} r_t.$$

If there are no changes, this basically corresponds to the standard notion of regret as used e.g. by Jaksch et al. [2010] (apart from an additive constant of order D, cf. footnote 1 ibid.). In the following, we assume that the random rewards r_t are always bounded in $[0, 1]$.

2.1 DEFINITION OF VARIATION

We consider individual terms for the *variation* in mean rewards and transition probabilities, that is,

$$V_T^r := \sum_{t=1}^{T-1} \max_{s,a} \left|\bar{r}_{t+1}(s, a) - \bar{r}_t(s, a)\right|, \text{ and}$$

$$V_T^p := \sum_{t=1}^{T-1} \max_{s,a} \left\|p_{t+1}(\cdot|s, a) - p_t(\cdot|s, a)\right\|_1.$$

These "local" variation measures can also be used to bound a more "global" notion of variation in average reward defined as

$$V_T := \sum_{t=1}^{T-1} \left|\rho^*(M_{t+1}) - \rho^*(M_t)\right|.$$

Theorem 1. $V_T \le V_T^r + DV_T^p$.

The proof of Theorem 1 is given in Section 5.1 below. As an example of Ortner et al. [2014a] shows, the bound of Theorem 1 is best possible.

While V_T is a more straightforward adaptation of the notion of variation of Besbes et al. [2014] from the bandit

to the MDP setting, in the latter it seems more natural to work with the local variation measures for rewards and transition probabilities, as the learner does not have direct access to the average rewards of policies.

3 ALGORITHM

For reinforcement learning in the changing MDP setting, we propose Variation-aware UCRL (shown as Algorithm 1), which is based on the UCRL algorithm of Jaksch et al. [2010].

UCRL is an algorithm that is based on the idea of being optimistic in the face of uncertainty. It maintains estimates of rewards and transition probabilities (line 6) and employs confidence intervals to define a set $\mathcal{M}$ of MDPs that are plausible with respect to the observations so far (line 7). When computing a new policy the algorithm chooses the policy $\tilde{\pi}$ and the MDP $\tilde{M}$ in $\mathcal{M}$ that give the highest average reward (line 8). UCRL employs this optimistic policy $\tilde{\pi}$ until the state-action visits in some state have doubled (lines 9–10), when a new policy is computed. The time intervals in which the policy is fixed are called *episodes*.

For the changing MDP setting, we use adapted confidence intervals (1) and (2) to account for the variation in rewards and transition probabilities. The arising algorithm basically corresponds to the colored UCRL2 algorithm suggested by Ortner et al. [2014b] for reinforcement learning in MDPs with given similarities.

While the regret of Variation-aware UCRL contains a term that is linear in the number of steps (cf. Theorem 11 below), we can obtain sublinear regret bounds by restarting the algorithm according to a suitable scheme shown as Algorithm 2. Our restart schedule is optimized with respect to the variation, as the regret bounds presented in the next section will show. Note that the algorithm needs (upper bounds on) the local variations V_T^r and V_T^p as well as a bound D on the diameters D_t as an input. Alternatively, an upper bound on the global variation V_T (replacing the term $V_T^r + DV_T^p$ in the algorithm) could be used as well, cf. Theorem 1.

The idea of restarting UCRL in the changing MDP setting has already been considered by Jaksch et al. [2010]. When a bound L on the total number of changes is known, then using a restart schedule adapted to L gave the following regret bound.[1]

[1]Jaksch et al. [2010] consider a slightly different notion of regret defined as $\sum_t (\rho_t^* - r_t)$, where $\rho_t^* := \rho^*(M_t)$ is the optimal average reward at step t. However, when there are at most L changes, the difference to our notion of regret is only of order LD.

Algorithm 1 Variation-aware UCRL

1: **Input:** States $\mathcal{S}$, actions $\mathcal{A}$, confidence parameter δ, upper bounds $\tilde{V}^r$, $\tilde{V}^p$ on the variation of rewards and transition probabilities.
2: **Initialization:** Set current time step $t := 1$.
3: **for** episode $k = 1, \dots$ **do**
4: Set episode start $t_k := t$.
5: Let $v_k(s,a)$ denote the state-action counts for visits in current episode k, and $N_k(s,a)$ be the counts for visits before episode k.
6: For $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, compute estimates

$$\hat{r}_k(s,a) := \frac{\sum_\tau r_\tau \cdot \mathbb{1}_{s_\tau = s, a_\tau = a}}{\max(1, N_k(s,a))},$$

$$\hat{p}_k(s'|s,a) := \frac{\#\{\tau : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}}{\max(1, N_k(s,a))}.$$

Compute policy $\tilde{\pi}_k$:
7: Let $\mathcal{M}_k$ be the set of plausible MDPs $\tilde{M}$ with rewards $\tilde{r}(s,a)$ and transition probabilities $\tilde{p}(\cdot|s,a)$ satisfying

$$|\tilde{r}(s,a) - \hat{r}_k(s,a)| \le \tilde{V}^r + \sqrt{\tfrac{8\log(8SAt_k^3/\delta)}{\max(1,N_k(s,a))}}, \quad (1)$$

$$\left\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \le \tilde{V}^p + \sqrt{\tfrac{8S\log(8SAt_k^3/\delta)}{\max(1,N_k(s,a))}}. \quad (2)$$

8: Use extended value iteration (see Section 3.1.2 of Jaksch et al. [2010]) to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that

$$\tilde{\rho}_k := \rho(\tilde{M}_k, \tilde{\pi}_k) = \max_{M' \in \mathcal{M}_k} \rho^*(M').$$

Execute policy $\tilde{\pi}_k$**:**
9: **while** $v_k(s_t, \tilde{\pi}_k(s_t)) < \max(1, N_k(s_t, \tilde{\pi}_k(s_t)))$ **do**
10: Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward r_t, and observe s_{t+1}. Set $t = t+1$.
11: **end for**

Algorithm 2 Variation-aware UCRL with restarts

1: **Input:** States $\mathcal{S}$, actions $\mathcal{A}$, confidence parameter δ, variation parameters V_T^r and V_T^p, upper bound D on diameters D_t.
2: **Initialization:** Set current time step $\tau := 1$.
3: **for** phase $i = 1, \dots$ **do**
4: Perform UCRL with confidence parameter $\delta/2\tau^2$ for $\theta_i := \left\lceil \frac{i^2}{(V_T^r + DV_T^p)^2} \right\rceil$ steps.
5: Set $\tau = \tau + \theta_i$.
6: **end for**

Theorem 2 (Jaksch et al. [2010])**.** *In an MDP with at most L changes, after T steps the regret of UCRL restarted with confidence parameter $\frac{\delta}{L^2}$ at steps $\lceil \frac{i^3}{(L+1)^2} \rceil$ for $i = 1, 2, 3, \ldots$ is upper bounded as*

$$R_T \;\leq\; 65 \cdot (L+1)^{1/3} T^{2/3} DS \sqrt{A \log\left(\tfrac{T}{\delta}\right)}$$

with probability of at least $1 - \delta$.

Note that the restart schedule of Theorem 2 basically corresponds to performing UCRL for $\sim \frac{i^2}{L^2}$ steps for $i = 1, 2, \ldots$, which is similar to our algorithm replacing L by the variation term $V_T^r + DV_T^p$.

4 MAIN RESULT

The following regret bound for Variation-aware UCRL with restarts is our main result. The proof is given in the next section.

Theorem 3. *After any T steps, the regret of the restart scheme for Variation-aware UCRL of Algorithm 2 is bounded as*

$$R_T \leq 74 \cdot DS(V_T^r + DV_T^p)^{1/3} T^{2/3} \sqrt{A \log\left(\tfrac{16 S^2 A T^5}{\delta}\right)}$$

with probability $1 - \delta$, provided that in each phase i the variation parameters $\tilde{V}_i^r$, $\tilde{V}_i^p$ are set to the respective true variation values for phase i.

Obviously, the bound of Theorem 3 is better than the bound of Theorem 2, when L is large but the variation small as e.g. when having small gradual changes at any time step. On the other hand, if there are L changes with maximal variation, the bound of Theorem 3 is worse by a factor of $D^{1/3}$.

With respect to the variation and the horizon, our bound is optimal, as bounds of $\tilde{\mathcal{O}}(V^{1/3} T^{2/3})$ are already best possible in the bandit setting as shown by Besbes et al. [2014].

5 ANALYSIS

We start with some preliminaries. First, we introduce the Poisson equation for the optimal policy in a communicating MDP. That is, the mean rewards $\bar{r}(s, \pi^*(s))$ under an optimal policy π^* and the respective optimal average reward ρ^* are related via the Poisson equation $\rho^*(M') - \bar{r}(s, \pi^*(s)) = \sum_{s'} p'(s'|s, \pi^*(s)) \cdot \lambda(s') - \lambda(s)$, where λ is the so-called bias function for π^*, cf. [Puterman, 1994]. It holds that each $\lambda(s)$ as well as the span of the bias function $\Lambda := \max_s \lambda(s) - \min_{s'} \lambda(s)$ is upper bounded by the diameter, cf. [Jaksch et al., 2010, Bartlett and Tewari, 2009].

We will frequently make use of Azuma-Hoeffding inequality, which we state here for convenience.

Lemma 4 (Azuma-Hoeffding inequality (Hoeffding [1963])). *Let $X_1, X_2, \ldots$ be a martingale difference sequence with $|X_i| \leq c$ for all i. Then for all $\epsilon > 0$ and $n \in \mathbb{N}$,*

$$\mathbb{P}\Big\{ \sum_{i=1}^{n} X_i \geq \epsilon \Big\} \leq \exp\Big(- \frac{\epsilon^2}{2nc^2} \Big).$$

5.1 A PERTURBATION BOUND

We continue with establishing a perturbation bound on the optimal average reward, which is a generalization of Lemma 8 of Ortner et al. [2014a].

Lemma 5. *Assume we have two MDPs $M = (\mathcal{S}, \mathcal{A}, \bar{r}_t, p_t, s_1), M' = (\mathcal{S}, \mathcal{A}, \bar{r}', p', s_1)$ on the same state and action space. The MDP M may be non-time-homogeneous so that its mean rewards $\bar{r}_t$ and transition probabilities p_t are allowed to depend on time t. We assume that M' is time-homogeneous and communicating with optimal policy π'^*, such that for all steps $t = 1, \ldots, T$,*

$$\max_s \big| \bar{r}_t(s, \pi'^*(s)) - \bar{r}'(s, \pi'^*(s)) \big| \leq \Delta_t^r(s),$$
$$\max_s \big\| p_t(\cdot|s, \pi'^*(s)) - p'(\cdot|s, \pi'^*(s)) \big\|_1 \leq \Delta_t^p(s).$$

If π'^* is performed on M for T steps, then denoting by s_t the state visited at step t it holds that

$$\begin{aligned} T\rho^*(M') &- \sum_{t=1}^{T} \bar{r}_t(s_t, \pi'^*(s_t)) \\ &\leq \sum_{t=1}^{T} \big(\Lambda' \Delta_t^p(s_t) + \Delta_t^r(s_t) \big) \\ &\quad + \sum_{t=1}^{T} \Big(\sum_{s'} p_t(s'|s_t) \cdot \lambda'(s') - \lambda'(s_t) \Big), \end{aligned}$$

where λ' is the bias function and Λ' the respective bias span of π'^* on M'.

Proof. The proof is a modification of the proof of Lemma 8 in Appendix A of [Ortner et al., 2014a]. Abbreviating $\bar{r}_t(s) := \bar{r}_t(s, \pi'^*(s))$, $\bar{r}'(s) := \bar{r}'(s, \pi'^*(s))$ and $p_t(s'|s) := p_t(s'|s, \pi'^*(s))$, $p'(s'|s) := p'(s'|s, \pi'^*(s))$ in the following, we can write

$$\begin{aligned} T\rho^*(M') - \sum_{t=1}^{T} \bar{r}_t(s_t) &= \sum_{t=1}^{T} \big(\rho^*(M') - \bar{r}_t(s_t) \big) \\ &\leq \sum_{t=1}^{T} \big(\rho^*(M') - \bar{r}'(s_t) \big) + \sum_{t=1}^{T} \big(\bar{r}'(s_t) - \bar{r}_t(s_t) \big) \end{aligned}$$

$$\leq \sum_{t=1}^{T} \big(\rho^*(M') - \bar{r}'(s_t)\big) + \sum_{t=1}^{T} \Delta_t^r(s_t). \qquad (3)$$

For bounding the first term in (3) we use that by the Poisson equation for policy π'^* on M' we have that $\rho^*(M') - \bar{r}'(s) = \sum_{s'} p'(s'|s) \cdot \lambda'(s') - \lambda'(s)$. Accordingly, it holds that

$$\begin{aligned}
&\sum_{t=1}^{T} \big(\rho^*(M') - \bar{r}'(s_t)\big) \\
&= \sum_{t=1}^{T} \Big(\sum_{s'} p'(s'|s_t) \cdot \lambda'(s') - \lambda'(s_t) \Big) \\
&\leq \sum_{t=1}^{T} \Big(\sum_{s'} p_t(s'|s_t) \cdot \lambda'(s') - \lambda'(s') \Big) \\
&\quad + \sum_{t=1}^{T} \sum_{s'} \big(p'(s'|s_t) - p_t(s'|s_t)\big) \lambda'(s_t) \\
&\leq \sum_{t=1}^{T} \Big(\sum_{s'} p_t(s'|s_t) \cdot \lambda'(s') - \lambda'(s_t) \Big) + \sum_{t=1}^{T} \Lambda' \Delta_t^p(s_t),
\end{aligned}$$

whence the lemma follows together with (3). □

For the analysis of the last term in the bound of Lemma 5 we can use the following result, which is a simple generalization of a technique used by Jaksch et al. [2010].

Lemma 6. *Consider some MDP $M = (\mathcal{S}, \mathcal{A}, \bar{r}, p, s_1)$ and let $f : \mathcal{S} \to \mathbb{R}$ be some function on the state space of M. Then for any (possibly nonstationary) policy choosing at each step τ action a_τ in the current state s_τ, it holds with probability at least $1 - \delta$,*

$$\sum_{\tau=1}^{T} \Big(\sum_{s'} p(s'|s_\tau, a_\tau) \cdot f(s') - f(s_\tau) \Big) \leq F \sqrt{2T \log\left(\tfrac{1}{\delta}\right)} + F,$$

where $F := \max_s f(s) - \min_s f(s)$ is the span of f.

Proof. Following an argument due to Jaksch et al. [2010] we write

$$\begin{aligned}
&\sum_{\tau=1}^{T} \Big(\sum_{s'} p(s'|s_\tau) \cdot f(s') - f(s_\tau) \Big) \\
&= \sum_{\tau=1}^{T} \Big(\sum_{s'} p(s'|s_\tau) \cdot f(s') - f(s_{\tau+1}) \Big) \\
&\quad + f(s_{T+1}) - f(s_1). \qquad (4)
\end{aligned}$$

Now $f(s_{T+1}) - f(s_1) \leq F$, while the sum is a martingale difference sequence $\sum_\tau X_\tau$ with $|X_\tau| \leq F$. The lemma follows by an application of Azuma-Hoeffding (Lemma 4). □

The following corollary is a variant of Lemma 9 contained in the (unpublished) appendix of [Ortner et al., 2014a].[2]

Corollary 7. *For two communicating MDPs M, M' that satisfy the assumptions of Lemma 5 for time and state independent values Δ^r, Δ^p (i.e., $\Delta_t^r(s) \leq \Delta^r$ and $\Delta_t^p(s) \leq \Delta^p$ for all s and all t) it holds that*

$$\rho^*(M') - \rho^*(M) \leq \Lambda' \Delta^p + \Delta^r.$$

Proof. From Lemmata 5 and 6 we have that with probability $1 - \delta$

$$\begin{aligned}
&\rho^*(M') - \frac{1}{T} \sum_{t=1}^{T} \bar{r}(s_t, \pi'^*(s_t)) \\
&\leq \Lambda' \Delta^p + \Delta^r + \frac{1}{T} \big(\Lambda' \sqrt{2T \log(1/\delta)} + \Lambda' \big).
\end{aligned}$$

Choosing $\delta = 1/T$ this yields for $T \to \infty$ and taking expectations that

$$\begin{aligned}
\rho^*(M') - \rho^*(M) &\leq \rho^*(M') - \rho(M, \pi'^*) \\
&\leq \Lambda' \Delta^p + \Delta^r. \qquad \square
\end{aligned}$$

Corollary 7 allows to give the following quick proof of Theorem 1.

Proof of Theorem 1. Let $\Delta_t^r := \max_{s,a} |\bar{r}_{t+1}(s,a) - \bar{r}_t(s,a)|$ and $\Delta_t^p := \|p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a)\|_1$. Then by Corollary 7 and the assumption that the diameters of all M_t are bounded by D we have for $t = 1, \ldots, T-1$

$$|\rho^*(M_{t+1}) - \rho^*(M_t)| \leq D\Delta_t^p + \Delta_t^r$$

and Theorem 1 follows by summing over all t. □

5.2 OPTIMISM

We show that the set of plausible MDPs with high probability contains each MDP M_t the learner acts on in step t. This is the theoretical justification for optimism, as it will allow us to show in the next section that the true reward can be upper bounded by the optimistic value $\tilde{\rho}$.

Lemma 8. *With probability $1 - \frac{5\delta}{6}$, the set $\mathcal{M}(t)$ of plausible MDPs computed at any time step t contains all MDPs M_τ for $\tau = 1, \ldots, T$.*

Proof. The proof is similar to the handling of failing confidence intervals for the colored UCRL algorithm given in Appendix A.2 of Ortner et al. [2014b].

[2] We note that a bound on the absolute value of the difference in average reward as stated in [Ortner et al., 2014a] will in general depend on the bias spans resp. the diameters of *both* MDPs, that is, the maximum of both values.

Fix a state-action pair (s,a), and let $\tau_1, \tau_2, \dots$ be the $N_t(s,a)$ time steps at which action a has been chosen in state s, i.e., $(s_{\tau_i}, a_{\tau_i}) = (s,a)$ for all i. For the analysis of the transition probability estimates $\hat{p}_t$ computed at step t we consider all vectors $\mathbf{x}$ indexed over the states with entries ± 1. Then writing $x(s)$ for the entry in $\mathbf{x}$ with index s we have

$$\begin{aligned}
&\left\|\hat{p}_t(\cdot|s,a) - \mathbb{E}\big[\hat{p}_t(\cdot|s,a)\big]\right\|_1 \\
&= \sum_{s'} \left|\hat{p}_t(s'|s,a) - \mathbb{E}\big[\hat{p}_t(s'|s,a)\big]\right| \\
&\leq \max_{\mathbf{x}\in\{-1,1\}^S} \sum_{s'} \Big(\hat{p}_t(s'|s,a) - \mathbb{E}\big[\hat{p}_t(s'|s,a)\big]\Big)\, x(s') \\
&\leq \max_{\mathbf{x}\in\{-1,1\}^S} \frac{1}{N_t(s,a)} \sum_{i=1}^{N_t(s,a)} X_i(\mathbf{x}),
\end{aligned}$$

where we set

$$X_i(\mathbf{x}) := x(s_{\tau_i+1}) - \sum_{s'} p_{\tau_i}(s'|s_{\tau_i}, a_{\tau_i})\, x(s').$$

Now $\sum_i^{N_t(s,a)} X_i(\mathbf{x})$ is a martingale difference sequence with $|X_i(\mathbf{x})| \leq 2$ for any fixed $\mathbf{x}$ and fixed $N_t(s,a) = n$ so that by Azuma-Hoeffding (Lemma 4)

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i(\mathbf{x}) \geq \sqrt{8Sn\log\left(\tfrac{8SAt^3}{\delta}\right)}\right\} \leq \frac{\delta}{8^S SAt^3}.$$

A union bound over all 2^S vectors $\mathbf{x}$ and all possible values of $N_t(s,a)$ shows that with probability $1 - \frac{\delta}{4SAt^2}$

$$\left\|\hat{p}_t(\cdot|s,a) - \mathbb{E}\big[\hat{p}_t(\cdot|s,a)\big]\right\|_1 \leq \sqrt{\frac{8S\log(8SAt^3/\delta)}{\max(1, N_t(s,a))}}. \tag{5}$$

Finally, note that for any fixed $N_t(s,a)$ we have

$$\mathbb{E}\big[\hat{p}_t(\cdot|s,a)\big] = \frac{1}{N_t(s,a)} \sum_{i=1}^{N_t(s,a)} p_{\tau_i}(\cdot|s,a),$$

so that for all τ

$$\left\|\mathbb{E}\big[\hat{p}_t(\cdot|s,a)\big] - p_\tau(\cdot|s,a)\right\|_1 \leq V_T^p,$$

which together with (5) shows that with probability $1 - \frac{\delta}{4SAt^2}$ the transition probabilities $p_\tau(\cdot|s,a)$ at each step τ are contained in the confidence intervals (2) at step t.

For the rewards, $\hat{r}_t(s,a) - \mathbb{E}\big[\hat{r}_t(s,a)\big]$ as well as $\mathbb{E}\big[\hat{r}_t(s,a)\big] - \hat{r}_t(s,a)$ can be written as martingale difference sequences and Azuma-Hoeffding can be used to show that with probability $1 - \frac{\delta}{4SAt^2}$, the rewards $\bar{r}_\tau(s,a)$ at each step τ are contained in the confidence intervals (1) for step t.

A union bound over all t and all state-action pairs concludes the proof, noting that $\sum_t \frac{\delta}{2t^2} \leq \frac{5\delta}{6}$. □

In the following, we assume that the statement of Lemma 8 holds, so that we need to consider the respective error probability only once.

5.3 T-STEP VS. AVERAGE REWARD

In this section we consider the difference between the optimal T-step reward and the optimal average reward. First we recall the well-known fact that the optimal T-step policy does not deviate by much from the optimal policy in respect to average reward, see e.g. Exercise 38.17 of Lattimore and Szepesvári [2019].

Lemma 9. *Let M be a time-homogeneous and communicating MDP with diameter D and rewards in $[0,1]$. Further let $v_T^*(M,s)$ be the optimal T-step reward when starting in state s. Then for any state s,*

$$v_T^*(M,s) \;\leq\; T\rho^*(M) + D.$$

Accordingly, the T-step reward in the changing MDP setting can also bounded by the optimistic average reward $\tilde{\rho}$.

Lemma 10. *Under the assumption of Lemma 8, for all k and all s,*

$$v_T^*(s) \;\leq\; T\tilde{\rho}_k + D.$$

Proof. Fix any k. As in Section 3.1 of [Jaksch et al., 2010] we consider the following extended MDP $\tilde{M}_k^+$ that corresponds to the set of plausible MDPs $\mathcal{M}_k$. That is, for any state s in $\mathcal{S}$ the extended action set contains for each a in the original action set $\mathcal{A}$, each value $\tilde{r}(s,a)$ in (1), and each transition probability distribution $\tilde{p}(\cdot|s,a)$ in (2) a respective action with reward $\tilde{r}(s,a)$ and transition probability distribution $\tilde{p}(\cdot|s,a)$. By the assumption that Lemma 8 holds, the true values for rewards and transition probabilities at any step τ are contained in these confidence intervals, so that there is a nonstationary T-step policy on $\tilde{M}_k^+$ whose expected T-step reward is $v_T^*(s)$. Therefore, for the optimal nonstationary T-step reward on $\tilde{M}_k^+$, denoted by $\tilde{v}_T$, we have

$$\tilde{v}_T \;\geq\; v_T^*(s). \tag{6}$$

Further, as $\tilde{M}_k^+$ contains the original transition probabilities of each M_τ, its diameter is bounded by D. Hence, by Lemma 9, $\tilde{v}_T \leq T\rho^*(\tilde{M}_k^+) + D$. As $\rho^*(\tilde{M}_k^+) = \tilde{\rho}_k$ (cf. Section 3.1 of Jaksch et al. [2010]) this shows together with (6) the lemma. □

5.4 REGRET WITHOUT RESTARTS

As a next step, we derive the following regret bound for Variation-aware UCRL without restarts (i.e. Algorithm 1).

Theorem 11. *If the variation parameters are set to their true values, that is, $\tilde{V}^r := V_T^r$ and $\tilde{V}^p := V_T^p$, then after any T steps the regret of Variation-aware UCRL without restarts (i.e., Algorithm 1) is upper bounded by*

$$32DS\sqrt{AT\log\left(\tfrac{8SAT^3}{\delta}\right)} + 2T(DV_T^p + V_T^r)$$

with probability $1-\delta$. This bound also holds when the algorithm starts in an initial state s_1 that is different from the initial state s_1^* of the optimal T-step policy we compare to.

For the proof we will use the following two basic facts about UCRL (Proposition 18 and Lemma 19 of Jaksch et al. [2010]), which can be derived from its episode termination criterion. As we use the same criterion these results also hold for our variation-aware adaptation.

Lemma 12. *The number of episodes K of Variation-aware UCRL after any $T > SA$ steps is bounded by $SA\log_2\left(\frac{T}{SA}\right)$.*

Lemma 13.

$$\sum_{s,a}\sum_{k=1}^{K}\frac{v_k(s,a)}{\sqrt{\max(1,N_k(s,a))}} \;\le\; \left(\sqrt{2}+1\right)\sqrt{SAT}.$$

Proof of Theorem 11. First, let $\tilde{\rho}_{\min} := \min_k \tilde{\rho}_k$. Then denoting the length of episode k by $T_k := t_{k+1} - t_k$ (setting $t_{K+1} := T$), by Lemma 10 we have

$$v_T^*(s_1^*) \;\le\; T\tilde{\rho}_{\min} + D \;\le\; \sum_{k=1}^{K} T_k\tilde{\rho}_k + D. \qquad (7)$$

Further, another application of Azuma-Hoeffding (Lemma 4) shows that for the rewards r_t collected by the algorithm with probability $1-\frac{\delta}{12}$

$$\sum_{t=1}^{T}\left(\bar{r}(s_t,a_t) - r_t\right) \;\le\; \sqrt{2T\log\left(\tfrac{12}{\delta}\right)}. \qquad (8)$$

Combining (7) and (8) the regret is bounded as

$$\begin{aligned}
R_T \;&=\; v_T^*(s_1^*) - \sum_{t=1}^{T} r_t \\
&= v_T^*(s_1^*) - \sum_{t=1}^{T}\bar{r}(s_t,a_t) + \sum_{t=1}^{T}\left(\bar{r}(s_t,a_t) - r_t\right) \\
&\le \sum_{k=1}^{K}\Big(T_k\tilde{\rho}_k - \sum_{t=t_k}^{t_{k+1}-1}\bar{r}(s_t,\tilde{\pi}_k(s_t))\Big) \\
&\qquad + \sqrt{2T\log\left(\tfrac{12}{\delta}\right)} + D. \qquad (9)
\end{aligned}$$

Now we are going to bound the term in the sum for each episode k using Lemma 5. Indeed, we perform the optimal policy $\tilde{\pi}_k$ of the optimistic MDP $\tilde{M}_k$ with average reward $\tilde{\rho}_k$ on the underlying (non-time-homogeneous) true MDP, and the rewards $\tilde{r}_k$ and the transition probabilities $\tilde{p}_k$ of $\tilde{M}_k$ satisfy the confidence intervals (1) and (2) according to Lemma 8 so that

$$\begin{aligned}
&\left|\bar{r}_t(s,a) - \tilde{r}_k(s,a)\right| \\
&\quad\le \left|\bar{r}_t(s,a) - \hat{r}_k(s,a)\right| + \left|\hat{r}_k(s,a) - \tilde{r}_k(s,a)\right| \\
&\quad\le V_T^r + \tilde{V}^r + 2\sqrt{\tfrac{8\log\left(8SAt_k^3/\delta\right)}{\max\left(1,N_k(s,a)\right)}}
\end{aligned}$$

as well as

$$\begin{aligned}
&\left\|p_t(\cdot|s,a) - \tilde{p}_k(\cdot|s,a)\right\|_1 \\
&\le \left\|p_t(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 + \left\|\hat{p}_k(\cdot|s,a) - \tilde{p}_k(\cdot|s,a)\right\|_1 \\
&\le V_T^p + \tilde{V}^p + 2\sqrt{\tfrac{8S\log\left(8SAt_k^3/\delta\right)}{\max\left(1,N_k(s,a)\right)}}.
\end{aligned}$$

By assumption $\tilde{V}^r = V_T^r$ and $\tilde{V}^p = V_T^p$, and Lemma 5 gives

$$\begin{aligned}
&T_k\tilde{\rho}_k - \sum_{t=t_k}^{t_{k+1}-1}\bar{r}(s_t,\tilde{\pi}_k(s_t)) \\
&\le 2\tilde{\Lambda}_k\sum_{t=t_k}^{t_{k+1}-1}\Big(V_T^p + \sqrt{\tfrac{8S\log\left(8SAt_k^3/\delta\right)}{\max\left(1,N_k(s_t,\tilde{\pi}_k(s_t))\right)}}\Big) \qquad (10)\\
&\quad + 2\sum_{t=t_k}^{t_{k+1}-1}\Big(V_T^r + \sqrt{\tfrac{8\log\left(8SAt_k^3/\delta\right)}{\max\left(1,N_k(s_t,\tilde{\pi}_k(s_t))\right)}}\Big) \qquad (11)\\
&\quad + \sum_{t=t_k}^{t_{k+1}-1}\Big(\sum_{s'}p_t(s'|s_t,\tilde{\pi}_k(s_t))\cdot\tilde{\lambda}_k(s') - \tilde{\lambda}_k(s_t)\Big), \\
&\qquad\qquad (12)
\end{aligned}$$

where $\tilde{\lambda}_k$ is the bias function of $\tilde{\pi}_k$ on $\tilde{M}_k$ and $\tilde{\Lambda}_k$ is the respective bias span. Since by Lemma 8 the set of plausible MDPs contains each MDP M_t and the diameter of each M_t is bounded by D, the bias span $\tilde{\Lambda}_k \le D$, cf. [Jaksch et al., 2010, Bartlett and Tewari, 2009].[3]

Accordingly, the sum of (10) and (11) over all episodes is bounded by Lemma 13 as

$$\begin{aligned}
&\le\; 2T(DV_T^p + V_T^r) \\
&\quad + 2(D+1)\sqrt{8S\log\left(\tfrac{8SAT^3}{\delta}\right)}\sum_{s,a}\sum_{k=1}^{K}\tfrac{v_k(s,a)}{\sqrt{\max\left(1,N_k(s,a)\right)}} \\
&\le\; 2T(DV_T^p + V_T^r) \\
&\qquad + 2\left(\sqrt{2}+1\right)(D+1)S\sqrt{8AT\log\left(\tfrac{8SAT^3}{\delta}\right)}. \qquad (13)
\end{aligned}$$

[3] Note that for the argument it would be sufficient if just one of the MDPs M_t is plausible. However, for the restart scheme of ALgorithm 2 we would need a plausible M_t in each phase.

The sum in (12) can be written as in the proof of Lemma 6 as

$$\begin{aligned}
&\sum_{t=t_k}^{t_{k+1}-1} \Big(\sum_{s'} p_t(s'|s_t, \tilde{\pi}_k(s_t)) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s_t) \Big) \\
&= \sum_{t=t_k}^{t_{k+1}-1} \Big(\sum_{s'} p_t(s'|s_t, \tilde{\pi}_k(s_t)) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s_{t+1}) \Big) \\
&\quad + \tilde{\lambda}_k(s_{t_{k+1}}) - \tilde{\lambda}_k(s_{t_k}).
\end{aligned}$$

Now $\tilde{\lambda}_k(s_{t_{k+1}}) - \tilde{\lambda}_k(s_{t_k}) \le D$, so that summing over all episodes gives by Azuma Hoeffding (Lemma 4) and Lemma 12 that with probability $1 - \frac{\delta}{12}$

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \Big(\sum_{s'} p_t(s'|s_t, \tilde{\pi}_k(s_t)) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s_t) \Big) \\
&\le \sum_{t=1}^{T} \Big(\sum_{s'} p_t(s'|s_t, \tilde{\pi}_k(s_t)) \cdot \tilde{\lambda}_{k(t)}(s') - \tilde{\lambda}_{k(t)}(s_{t+1}) \Big) \\
&\quad + KD \\
&\le D\sqrt{2T \log\big(\tfrac{\delta}{12}\big)} + DSA \log_2\big(\tfrac{T}{SA}\big), \qquad (14)
\end{aligned}$$

where $k(t)$ denotes the episode in which time step t occurs.

Thus, combining (9)–(14) taking into account the error probabilities for (8), (14), and Lemma 8, we yield that with probability $1 - \delta$ the regret is bounded by

$$\begin{aligned}
R_T \;\le\; &\sqrt{2T \log\big(\tfrac{12}{\delta}\big)} + D + 2T(DV_T^p + V_T^r) \\
&+ 2\big(\sqrt{2}+1\big)(D+1)S\sqrt{8AT \log\big(\tfrac{8SAT^3}{\delta}\big)}. \\
&+ D\sqrt{2T \log\big(\tfrac{12}{\delta}\big)} + DSA \log_2\big(\tfrac{T}{SA}\big)
\end{aligned}$$

and some simplifications analogous to Appendix C.4 of Jaksch et al. [2010] give the claimed regret bound. □

5.5 PROOF OF THEOREM 3

Finally, we are ready to give the proof of the regret bound for the restart scheme of Algorithm 2. Abusing notation we write V_i^r and V_i^p for the variation of rewards and transition probabilities in phase i and abbreviate $V_i := V_i^r + DV_i^p$, $V := V_T^r + DV_T^p$ and $\theta_i := \lceil \frac{i^2}{V^2} \rceil$.

First, let us bound the number of phases N. Obviously, step T is reached in phase N when

$$\sum_{i=1}^{N-1} \Big\lceil \frac{i^2}{V^2} \Big\rceil < T \le \sum_{i=1}^{N} \Big\lceil \frac{i^2}{V^2} \Big\rceil.$$

Recalling that $\sum_{i=1}^{N} i^2 = \frac{1}{6}N(N+1)(2N+1) > \frac{1}{3}N^3$ we obtain

$$T > \sum_{i=1}^{N-1} \Big\lceil \frac{i^2}{V^2} \Big\rceil > \sum_{i=1}^{N-1} \frac{i^2}{V^2} > \frac{(N-1)^3}{3V^2},$$

so that the number of phases is bounded as

$$N < 1 + \sqrt[3]{3V^2T}. \qquad (15)$$

Writing τ_i for the initial step of phase i and $s^*_{\tau_i}$ for the (random) state visited by the optimal T-step policy at step τ_i, we can decompose the regret as

$$v_T^*(s_1) - \sum_{t=1}^{T} r_t = \sum_{i=1}^{N} \Big(\mathbb{E}\big[v_{\theta_i}^*(s^*_{\tau_i})\big] - \sum_{t=\tau_i}^{\tau_i - 1} r_t \Big). \qquad (16)$$

By Theorem 11 and a union bound over all possible values for state $s^*_{\tau_i}$, the i-th summand ($i = 1, \ldots, N$) in (16) with probability $1 - \frac{\delta}{2\tau_i^2}$ is bounded by

$$32DS\sqrt{A \log\big(\tfrac{16S^2AT^5}{\delta}\big)} \cdot \sqrt{\theta_i} \;+\; 2V_i \cdot \theta_i.$$

If $\sqrt[3]{3V^2T} < 1$, then we also have $3V^2T < 1$ and hence $3V^2T^2 < T$, so that

$$VT < \sqrt{3} \cdot VT < \sqrt{T}.$$

Further, in this case by (15) we have $N = 1$ with $\theta_1 = T$ and $V_1 = V$, so that the regret is bounded by

$$\begin{aligned}
&32DS\sqrt{A \log\big(\tfrac{16S^2AT^5}{\delta}\big)} \cdot \sqrt{T} + 2VT \\
&< \Big(32DS\sqrt{A \log\big(\tfrac{16S^2AT^5}{\delta}\big)} + 2\Big) \cdot \sqrt{T},
\end{aligned}$$

which is upper bounded by the claimed regret bound.

On the other hand, if $\sqrt[3]{3V^2T} \ge 1$, then $N < 2\sqrt[3]{3V^2T}$ from (15) and summing over all N phases yields from (16) that with error probability $\sum_i \frac{\delta}{2\tau_i^2} < \sum_t \frac{\delta}{2t^2} < \delta$ the regret is bounded by

$$32DS\sqrt{A \log\big(\tfrac{16S^2AT^5}{\delta}\big)} \cdot \sum_{i=1}^{N} \sqrt{\theta_i} + 2\sum_{i=1}^{N} V_i\Big(\frac{i^2}{V^2} + 1\Big).$$

Noting that using Jensen's inequality

$$\sum_{i=1}^{N} \sqrt{\theta_i} \;\le\; \sqrt{NT} \;\le\; 1.7 \cdot V^{1/3}T^{2/3}$$

and that also

$$\begin{aligned}
&\sum_{i=1}^{N} V_i\Big(\frac{i^2}{V^2} + 1\Big) \;\le\; \sum_{i=1}^{N} V_i\Big(\frac{N^2}{V^2} + 1\Big) \;\le\; \frac{N^2}{V} + V \\
&< \; 8.33 \cdot V^{1/3}T^{2/3} + V,
\end{aligned}$$

concludes the proof, noting that the claimed bound holds trivially if $V \ge T$, so that we may assume that $V < T$ and hence $V < V^{1/3}T^{2/3}$. □

6 DISCUSSION AND EXTENSIONS

The regret bound of Theorem 11 relies on the assumption that the variation for rewards and transition probabilities are known in advance. Accordingly it is necessary for the restart scheme to know the respective variation terms for each single phase. It is easy to check that if upper bounds on these values are used to set $\tilde{V}^r$ and $\tilde{V}^p$ instead, the regret bounds of Theorems 2 and 11 simply depend on these upper bounds instead of the true values.

In principle, it is also possible to set the variation parameters $\tilde{V}^r$ and $\tilde{V}^p$ in Algorithm 1 to 0. Then Lemma 8 need not hold anymore, that is, it is not guaranteed that the set of plausible MDPs contains any of the MDPs M_t. Accordingly, we cannot rely on Lemma 10 anymore, which is based on Lemma 8 and guarantees that the optimistic average reward is an upper bound on the true reward. However, taking into account the true variation one can still establish an upper bound on the true reward.

Lemma 14. *Let $\tilde{\rho}$ be the optimistic average reward computed when using the true variations V_T^r and V_T^p and $\tilde{\rho}^0$ be the optimistic average reward computed with variation parameters $\tilde{V}^r = 0$ and $\tilde{V}^p = 0$. Then*

$$\tilde{\rho} \leq \tilde{\rho}^0 + V_T^r + DV_T^p.$$

Proof. The rewards and transition probabilities of the respective optimistic MDPs $\tilde{M}$ and $\tilde{M}^0$ with $\tilde{\rho} = \rho^*(\tilde{M})$ and $\tilde{\rho}^0 = \rho^*(\tilde{M}^0)$ differ by at most V_T^r and V_T^p, respectively. Hence the claim of the lemma follows by Corollary 7, recalling that the bias span of the optimal policy in $\tilde{M}$ is bounded by the diameter D, cf. the proof of Theorem 11. □

Thus, in principle one could use Lemma 14 to replace Lemma 10 in the proof of Theorem 11. It is easy to check that this works fine except for the second application of Lemma 8 used to show that the bias span $\tilde{\Lambda}$ is bounded by D. Indeed, the set of plausible MDPs $\mathcal{M}^0$ computed with $\tilde{V}^r = 0$ and $\tilde{V}^p = 0$ need not contain any of the MDPs M_t and hence there is no guarantee that it contains an MDP with diameter bounded by D.

Still it is possible to obtain an alternative bound on the bias span by observing that $\mathcal{M}^0$ contains with high probability an MDP where for each state-action pair (s,a) there is a subset $\mathcal{T}$ of $\{1,2,\ldots,T\}$ such that the transition probabilities under (s,a) are of the form

$$p(\cdot|s,a) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} p_t(\cdot|s,a). \tag{17}$$

The intution here is that the set $\mathcal{T}$ corresponds to the time steps when a sample for the state-action pair (s,a) has been taken. Let $\hat{\mathcal{M}}$ be the set of MDPs with transition probabilities of the form specified in (17). Then defining

$$\hat{D} := \max_{M\in\hat{\mathcal{M}}} D(M)$$

to be the maximal diameter over all MDPs in $\hat{\mathcal{M}}$, one has $\tilde{\Lambda} \leq \hat{D}$ and the following regret bound holds when setting the variation parameters to 0.

Theorem 15. *With probability $1-\delta$, the regret of Variation-aware UCRL with variation parameters set to 0 is upper bounded by*

$$R_T \leq 32\hat{D}S\sqrt{AT\log\left(\tfrac{8SAT^3}{\delta}\right)} + 2T(DV_T^p + V_T^r).$$

Using Theorem 15, one can derive the following regret bound for the restart scheme of Algorithm 2. Note that the restart scheme still needs $V_T^r + DV_T^p$ as input.

Theorem 16. *With probability $1-\delta$, the regret of the restart scheme of Algorithm 2 with variation parameters set to 0 in each phase is bounded by*

$$R_T \leq 74\hat{D}S(V_T^r + DV_T^p)^{1/3}T^{2/3}\sqrt{A\log\left(\tfrac{16S^2AT^5}{\delta}\right)}.$$

Similar bounds obviously also hold when the variation parameters are set to any value smaller than the true variation. Note also that $D \leq \hat{D}$, as the set $\hat{\mathcal{M}}$ also contains the MDPs M_t for each t. However, the following example shows that $\hat{D}$ in general cannot be bounded by D and in some even simple cases can be unbounded.

Example. Consider two MDPs M_1, M_2 over the same state space $\{s,s'\}$ and the same action space $\{a,a'\}$. In M_1 the nonzero transition probabilities are given by $p_1(s|s',a) = 1$, $p_1(s'|s,a) = 1/D$, $p_1(s|s,a) = 1-1/D$, and $p_1(s|s,a') = p_1(s'|s',a') = 1$. In M_2 the roles of a and a' are swapped, that is, we have $p_2(s|s,a) = p_2(s'|s',a) = 1$, $p_2(s|s',a') = 1$, $p_2(s'|s,a') = 1/D$, and $p_2(s|s,a') = 1-1/D$. Obviously, both MDPs have diameter D. However, the MDP M with nonzero transition probabilities $p(s|s,a) := p_2(s|s,a) = 1$, $p(s'|s',a) := p_2(s'|s',a) = 1$, $p(s|s,a') := p_1(s|s,a') = 1$, and $p(s'|s',a') := p_1(s'|s',a') = 1$ is contained in $\hat{\mathcal{M}}$, but does not have finite diameter, as the states s, s' are not connected.

To conclude, we note that recently variational bounds for the (contextual) bandit setting have been derived also for the case when the variation is unknown [Chen et al., 2019]. Achieving such bounds in our setting seems not easy, as sampling a particular state-action pair usually causes some transition costs.

Acknowledgements. This work has been supported by the Austrian Science Fund (FWF): I 3437-N33 in the framework of the CHIST-ERA ERA-NET (DELTA project).

References

Yasin Abbasi, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems 26*, pages 2508–2516, 2013.

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

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems 27*, pages 199–207, 2014.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. In *Proceedings of the 32nd Conference On Learning Theory, COLT 2019*, 2019.

Travis Dick, András György, and Csaba Szepesvári. Online learning in Markov decision processes with changing cost sequences. In *Proceedings of the International Conference on Machine Learning, ICML 2014*, pages 512–520, 2014.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Experts in a Markov decision process. In *Advances in Neural Information Processing Systems 17*, pages 401–408, 2005.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, August 2010.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019. to appear.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, September 2005.

Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko. Selecting near-optimal approximate state representations in reinforcement learning. In *Algorithmic Learning Theory – 25th International Conference, ALT 2014*, pages 140–154, 2014a.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless Markov bandits. *Theoretical Computer Science*, 558:62–76, 2014b.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

Jia Yuan Yu and Shie Mannor. Arbitrarily modulated Markov decision processes. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2946–2953, 2009a.

Jia Yuan Yu and Shie Mannor. Online learning in Markov decision processes with arbitrarily changing rewards and transitions. In *2009 International Conference on Game Theory for Networks*, pages 314–322, 2009b.