Problem Dependent Reinforcement Learning Bounds Which Can Identify Bandit Structure in MDPs

Andrea Zanette [1] **Emma Brunskill** [1]

Abstract

In order to make good decision under uncertainty an agent must learn from observations. To do so, two of the most common frameworks are Contextual Bandits and Markov Decision Processes (MDPs). In this paper, we study whether there exist algorithms for the more general framework (MDP) which automatically provide the best performance bounds for the specific problem at hand without user intervention and without modifying the algorithm. In particular, it is found that a very minor variant of a recently proposed reinforcement learning algorithm for MDPs already matches the best possible regret bound $\tilde{O}(\sqrt{SAT})$ in the dominant term if deployed on a tabular Contextual Bandit problem despite the agent being agnostic to such setting.

1. Introduction

For reinforcement learning (RL) to realize its huge potential benefit, we must create reinforcement learning algorithms that do not require extensive expertise and problem-dependent fine-tuning to achieve high performance in a particular domain of interest. Much exciting research is advancing this vision, such as alleviating the need for feature engineering using deep neural networks, and making it easier to specify the desired behavior through inverse reinforcement learning and reward design (Mnih et al., 2013; Abbeel & Ng, 2004). Here instead we consider the theoretical aspects of a key but understudied issue: what decision process framework to use, and how that choice impacts the resulting performance.

In reinforcement learning (learning to make good decisions under uncertainty), there are three common frameworks that allow learning from observations: multi-armed bandits (MABs) and contextual MABs, Markov decision processes (MDPs) and partially observable MDPs (POMDPs). Bandits assume that the actions taken do not impact the next state, MDPs assume actions impact the next state but the state is a sufficient statistic of prior history, and POMDPs assume that the true Markov state is latent, and in general the next state can depend on the full history of prior actions and observations. It is known that these three decision process frameworks differ significantly in computational complexity and statistical efficiency. In particular, when the decision process model is unknown and an agent must perform reinforcement learning, existing theoretical bounds illustrate that the best results possible in bandits, contextual bandits, MDPs and POMDPs may significantly differ. For example there exist *upper bounds* on the regret of algorithms for discrete state and action contextual bandits which scale as $\tilde{O}(\sqrt{SAT})$ (see (Bubeck & Cesa-Bianchi, 2012)) and *lower bounds* on the regret of algorithms for episodic discrete state and action MDPs which scale as $\Omega(\sqrt{HSAT})$ (Osband & Van Roy, 2016), here indicating there is a gap of at least a factor of $\sqrt{H}$ between the regret possible in the two settings. Such work suggests that to obtain good performance, it is of significant interest to have algorithms that either implicitly or explicitly use the simplest setting (of bandits, MDPs, POMDPs) that captures the domain of interest during reinforcement learning,

As (outside of simulated domains) the true decision process properties are unknown, choosing whether to model a problem using the bandit, MDP or POMDP frameworks is typically far from trivial. A software engineer working on a product recommendation engine may not know whether the product recommendations have a significant impact on the customers' later states and preferences, such that the engineer should model the problem as a MDP instead of a bandit in order to be able to use a reinforcement learning algorithm to learn a policy that best maximizes revenue. This may result in requiring prohibitive amounts of interaction data to learn a good decision policy. Ideally an engineer should be able to write down a problem in a very general way and be confident that the algorithm will inherit the best performance of the underlying domain and problem.

Here we work to create RL algorithms with strong setting / framework dependent bounds. Our hope is to create reinforcement learning methods that perform as well as the

[1]Stanford University, Stanford, California. Correspondence to: Andrea Zanette <zanette@stanford.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

underlying process allows but without the algorithm user having to specify in advance the process framework (bandit / MDP / POMDP) which is often unknown. In doing so we hope to alleviate the burden on the users, allowing them to inherit the benefits of more complex policies if the situation allows, without performance being harmed if the true process is simpler than the one specified.

Precisely here we consider the challenge of creating MDP algorithms that can inherit the best properties of tabular contextual bandits if the RL algorithm is operating in such setting. Our aim is similar in motivation to problem dependent theoretical analyses, that seek to provide tighter performance bounds by including an explicit dependence on some property of the domain, such as the mixing rate (Auer & Ortner, 2006), or the difference in rewards or optimal state-action values (Auer et al., 2002; Agrawal & Goyal, 2012; Even-Dar et al., 2006). However, existing problem dependent research has not yet enabled strong process-dependent learning bounds (e.g. bounds that depend on whether the domain is a MDP or a bandit). Prior problem dependent results are limited for our setting of interest because they typically make restrictive assumptions on the subset of Markov decision processes for which they hold (e.g., highly mixing for (Auer & Ortner, 2006)), require the user to explicitly provide domain properties (Bartlett & Tewari, 2009) or the provided bound does not yield strong guarantees when the MDP algorithm is deployed on a simpler bandit process (Maillard et al., 2014). A work with more similar intentions to ours is (Bubeck & Slivkins, 2012) where the authors propose an algorithm whose regret is optimal both for adversarial rewards and for stochastic rewards; by contrast here we consider a change in the learning framework (MDPs vs Bandits).

Perhaps the most closely related work is the recently introduced contextual decision process research (Jiang et al., 2017). The authors provide probably approximately correct (PAC) results for generic CDPs as a function of their Bellman rank; however their resulting bounds for tabular MDPs and CMABs do not provide the best or near-best PAC bounds (both have a worse dependence on the horizon). In contrast our work considers an algorithm for which we can achieve a near-optimal performance on MDPs and the best regret upper bound in the dominant terms for tabular contextual bandits.

In other words, we can use an MDP RL algorithm and if the real world is a bandit, the MDP RL algorithm automatically scales in performance about as well as a near-optimal algorithm that was designed *specifically* for bandit problems. Precisely, a small variant of UBEV (Dann et al., 2017) yields a $\sqrt{SAT}$ regret term if the MDP it is acting in is actually a tabular contextual bandit regardless of the prescribed MDP horizon H. Prior work in provably efficient RL algorithms (Jaksch et al., 2010; Dann & Brunskill, 2015; Azar et al., 2017) provide regret or PAC guarantees which depend on the MDP horizon H or diameter D for episodic and infinite-horizon MDPs, respectively. H is the MDP horizon and is specified to the algorithm. Therefore these analyses do not imply that "H" can be removed if the H-horizon MDP is actually generated from a CMAB problem.

The key insight of our analysis is to show that due to the bandit structure, the optimistic value function converges to the optimal value function fast enough that the regret bound terms due to the MDP framework contribute only to lower order terms with a logarithmic time dependence. In the rest of the paper, we first outline the setting, introduce the algorithm, and then provide our theoretical results and proofs before discussing future directions.

2. Notation and Setup

A finite horizon MDP is defined by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, H \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition function where $p(s' \mid s, a)$ is the probability of transitioning to state s' after taking action a in state s. The mean reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \in [0, 1]$ is the average instantaneous reward collected upon playing action a in state s, denoted by $r(s, a)$. The agent interacts with the environment in a sequence of episodes $k \in [1, \dots, K]$, each of a horizon of H time steps before resetting. As the optimal policy in finite-horizon domains is generally time-step-dependent, on each episode the agent selects a π_k which maps states s and timesteps t to actions. A policy π_k induces a value function for every state s and timestep $t \in [H]$ defined as $V_t^{\pi_k}(s_t) = \mathbb{E} \sum_{i=t}^{H} r(s_i, \pi_k(s_i, i))$ which is the expected return until the end of the episode (the expectation is over the states s_i encountered in the MDP). We denote the optimal policy with π^* and its value function as $V_t^*(s)$ and define the range of a vector V: $\operatorname{rng} V \stackrel{def}{=} \max_s V(s) - \min_s V(s)$.

There are multiple formal measures of RL algorithm performance. We focus on regret, which is frequently used in RL and very widely used in bandit research. Let the regret of the algorithm up to episode K from any sequence of starting states $s_{1k}, s_{2k}, \dots$ be:

$$\text{Regret}(K) \stackrel{def}{=} \sum_k V_1^*(s_{1k}) - V_1^{\pi_k}(s_{1k}). \qquad (1)$$

Since the policies depend on the history of observations, the regret is a random variable. Here we focus on a high probability bound on the regret.

We use the $\tilde{O}(\cdot)$ notation to indicate a quantity that depends on $(\cdot)$ up to a polylog expression of a quantity at most polynomial in $S, A, T, K, H, \frac{1}{\delta}$. We use the $\lesssim, \gtrsim, \simeq$ notation to mean $\leq, \geq, =$, respectively, up to a numerical constant.

Algorithm 1 UBEV-S for Stationary Episodic MDPs

1: **Input**: failure tolerance $\delta \in (0, 1]$
2: $n(s,a) = l(s,a) = m(s',s,a) = 0 \ \ \forall s', s, a \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}; \quad \tilde{V}_{H+1}(s) = 0 \ \ \forall s \in \mathcal{S}; \quad \phi^+ = 0$
3: **for** $k = 1, 2, \ldots$ **do**
4: **for** $t = H, H-1, \ldots, 1$ **do**
5: **for** $s \in \mathcal{S}$ **do**
6: **for** $a \in \mathcal{A}$ **do**
7: $\phi = \sqrt{\frac{2 \ln \ln(\max\{e, n(s,a)\}) + \ln(27HSA/\delta)}{n(s,a)}}$
8: $\hat{r} = \frac{l(s,a)}{n(s,a)}, \hat{V}_{next} = \frac{m(\cdot,s,a)^\top \tilde{V}_{t+1}}{n(s,a)}$
9: $Q(a) = \min\{1, \hat{r} + \phi\} + \min\{\max_s \tilde{V}_{t+1}(s), \hat{V}_{next} + \min\{(H-t), (\mathrm{rng}\, \tilde{V}_{t+1} + \phi^+)\}\phi\}$
10: **end for**
11: $\pi_k(s,t) = \arg\max_a Q(a); \quad \tilde{V}_t(s) = Q(\pi_k(s,t)); \quad \phi^+ = \max\{4\sqrt{S}H^2\phi(s, \pi_k(s,t)), \phi^+\}$
12: **end for**
13: **end for**
14: $s_1 \sim p_0$
15: **for** t=1,...H **do**
16: $a_t = \pi_k(s_t, t); \quad r_t \sim p_R(s_t, a_t); \quad s_{t+1} \sim p_P(s_t, a_t)$
17: $n(s_t, a_t) + +; \quad m(s_{t+1}, s_t, a_t) + +; \quad l(s_t, a_t) += r_t$
18: **end for**
19: **end for**

3. Mapping Contextual Bandits to MDPs

Tabular contextual multi-armed bandits are a generalization of the multiarmed bandit problem. They prescribe a set of contexts or states and the expected reward of an action depends on the state and action, $r(s, a)$. They can be alternatively viewed as a simplification of MDPs in which the next state is independent of the prior state and action. Let $\mathcal{M}_C$ be an episodic MDP with horizon H which is actually a contextual bandit problem: the transition probability is identical $p(s'|s, a) = \mu(s')$ for all states and actions, where μ is a fixed stationary distribution over states. Note that when doing RL in a $\mathcal{M}_C$ the agent does not know the transition model and therefore does not know the MDP can be viewed as a contextual bandit.

4. UBEV for Stationary MDPs

In this section we introduce the UBEV-S algorithm which is a slight variant of UBEV (Dann et al., 2017), a recent PAC algorithm designed for episodic non-stationary MDPs. Here we focus on a regret analysis due to its popularity in the bandit literature.

A large fraction of the literature for episodic MDPs considers stationary environments. If the MDP is truly stationary (i.e., with time-independent rewards and transition dynamics) then this assumption can be leveraged to produce $\sqrt{H}$-tighter regret bounds. For the purpose of our analysis on CMABs the rationale for removing the non-stationarity from UBEV is the following: if the MDP is transient the agent cannot "assume" that the same state s gives identical expected rewards $r(s, a)$ if visited at different timesteps, say t_1 and t_2. As a consequence, it would treat the same "context" s visited at t_1 and t_2 as different entities. We therefore adapt UBEV to handle stationary MDPs and modify the exploration bonus slightly. This second change preserves the original bounds in the MDP setting and enables us to obtain stronger bounds in the bandit setting. We call the resulting algorithm UBEV-S (Algorithm 1). Lines 4 through 13 refers to the planning step and lines 14 through 18 to the execution of the chosen policy in the MDP. UBEV-S is a minor variant of UBEV and it can be analyzed in the same way as the original UBEV to obtain a regret bound whose leading order term is $\tilde{O}(H\sqrt{SAT})$ on a generic (albeit stationary) MDP[1]. We outline such analysis in the appendix (in section A.4). The main difference from UBEV in (Dann et al., 2017) and UBEV-S here is the stated stationarity of the MDP. In stationary MDPs the transition dynamics $p(s'|s, a)$ and rewards $r(s, a)$ are assumed to be time-independent for a fixed (s, a) pair. This allows data aggregation for the same state-action pair (s, a) from different timesteps t in order to estimate the rewards and system dynamics, as seen in lines 2, 7, 8, 17. As a result, UBEV-S is more efficient on stationary environments because it does not need to estimate r and p for different timesteps but it will not handle transient MDPs as UBEV . This ultimately leads to a saving of $\sqrt{H}$ in the leading order regret term if the MDPs is time-invariant.

The other minor change is to make the exploration bonus

[1]Notice the difference in notation. Here T is the time elapsed; in (Dann et al., 2017) it is the number of episodes elapsed. The two differ by a factor of H.

(Algorithm 1 Line 9) depend on the range of the optimistic value function $(\operatorname{rng} \tilde{V}_{t+1})\phi(s,a)$ (defined in Algorithm 1) of the successor states. In contrast UBEV used a fixed overestimate $(H-t)\phi(s,a)$. A bonus dependent on the actual $\tilde{V}^{\pi_k}_{t+1}$ is the typical approach used in similar works (e.g. (Jaksch et al., 2010; Dann & Brunskill, 2015; Azar et al., 2017)). The rationale here is that if $\operatorname{rng} \tilde{V}_{t+1}$ is very small then the agent is not "too uncertain" about that transition, hence the exploration bonus should be smaller. Although this does not improve the MDP regret bound (which only considers a worst-case scenario), better practical performance should be expected and it will have important benefits for our bandit analysis. For the exploration bonus to be valid we require that optimism be guaranteed on any MDP. We ensure this by adding a correction term ϕ^+ which varies in different (s,a) pairs and is an estimate of the uncertainty of $\operatorname{rng} \tilde{V}_{t+1}$. The correction term ϕ^+ is continuously updated in line 11 of Algorithm 1 so that ϕ^+ keeps track of the largest bonus / confidence interval which is related to the least visited (s,a) pair (in subsequent states) under the agent's policy. In the appendix (section A.3) we carefully justify why this choice guarantees optimism on any MDP. This change does not affect the regret bound for stationary MDPs since our exploration bonus is still upper bounded by $H\phi(s,a)$ (this is the upper bound used to obtain the result on MDPs).

5. Theoretical Result

In this section we present the main result of the paper, which is an upper bound on the regret of UBEV-S on $\mathcal{M}_C$.

Theorem 1. *If* UBEV-S *is run on an H-horizon MDP with S states and A actions where the successor states s' is sampled from a fixed distribution μ then with probability at least $1-\delta$ the regret is bounded by the minimum between:*

$$\underbrace{\tilde{O}\left(\sqrt{SAT} + \frac{S^2AH^2\sqrt{H}}{\sqrt{\mu_{min}}} + \frac{SAH^2}{\mu_{min}}\right)}_{\textit{CMAB Analysis}} \tag{2}$$

and

$$\underbrace{\tilde{O}\left(H\sqrt{SAT} + S^2AH^2 + S\sqrt{S}AH^3\right)}_{\textit{MDP Analysis}} \tag{3}$$

jointly for all timesteps T.

Notice that equation 2 is obtained by the analysis that we discuss in this main paper while equation 3 is the regret bound that UBEV-S would achieve in *any* episodic stationary MDP (detailed the appendix). Since $\mathcal{M}_C$ is an MDP, the tighter bound applies.

The significance of this result is that the leading order term matches the lower bound $\Omega(\sqrt{SAT})$ previously established for tabular contextual bandit problems. The lower order terms of Equation 2 depend upon $\mu_{min} \overset{def}{=} \min_s \mu(s)$, which is the lowest probability of visiting any given context.

Put differently, for T sufficiently large and not too small μ_{min}, the leading order term dominates and the bound matches the lower bound for contextual bandits up to $\operatorname{polylog}(\cdot)$ factor. Problems where a large T is most critical for the regret are those where the optimal actions are barely distinguishable from the suboptimal ones. Our result shows that in this case there is little penalty for using a more general approach like UBEV-S which is designed for MDPs and is unaware of the problem structure. By the time the agent has identified which actions have maximum instantaneous reward the structure of the underlying problem is already clear to the agent. The key insight to obtain the result of theorem 1 is to examine the rate at which the optimistic value function $\tilde{V}^{\pi_k}_t$ converges to the true one V^*_t. While such convergence does not necessarily occur in a generic MDP, the highly mixing nature of contextual bandits ensures that enough information is collected in every context / state that convergence of the value function does occur for all states. The rate of convergence is high enough that the "price" for using an MDP algorithm on CMABs gets transferred to lower order terms without any T dependence.

6. Analysis on $\mathcal{M}_C$

We begin our analysis by looking at the main source of regret for UBEV-S when deployed on a generic MDP. We do this to identify the leading order term contributing to the regret. Next, we provide a tighter analysis of such term when the process is a CMAB.

Optimistic RL agents work by computing with high probability an optimistic value function $\tilde{V}^{\pi_k}_1(s_0)$ for any starting state s_0. This overestimates the true optimal value function $V^*_1(s_0)$ and allows to estimate the regret of an agent by evaluating the same policy on two different MDPs which get closer and closer to each other as more data is collected:

$$\begin{aligned}
\text{Regret}(K) &\overset{def}{=} \sum_k V^*_1(s_0) - V^{\pi_k}_1(s_0) \\
&\overset{Opt.}{\leq} \sum_k \tilde{V}^{\pi_k}_1(s_0) - V^{\pi_k}_1(s_0) \\
&= \underbrace{\sum_{k\leq K}\sum_{t\in[H]}\sum_{s,a} w_{tk}(s,a)\left(\tilde{r}_k(s,a) - r(s,a)\right)}_{\tilde{O}(\sqrt{SAT})} \\
&+ \underbrace{\sum_{k\leq K}\sum_{t\in[H]}\sum_{s,a} w_{tk}(s,a)\left(\tilde{p}_k(s,a) - p(s,a)\right)^\top \tilde{V}^{\pi_k}_{t+1}}_{\tilde{O}(H\sqrt{SAT})}
\end{aligned} \tag{4}$$

In the above expression the last equality follows from a standard decomposition, see for example lemma E.15 in (Dann et al., 2017). We indicated with $\tilde{p}_k(s,a)$ the optimistic transition probability vector implicitly computed by UBEV-S along with the optimistic value function $\tilde{V}_t^{\pi_k}$. Here $w_{tk}(s,a)$ is the probability of visiting state s and taking action a there at timestep t of the k-th episodes. Finally, $\tilde{r}_k(s,a)$ is the instantaneous optimistic reward collected upon taking action a in state s.

Below each term we have reported the regret that UBEV-S would obtain on a generic MDP. Estimating the rewards alone implies a regret contribution of order $\tilde{O}(\sqrt{SAT})$, which is what a (near) optimal CMAB algorithm achieves. Thus, to obtain a tighter bound on $\mathcal{M}_C$ we need to address the regret due to the transition dynamics which is of order $\tilde{O}(H\sqrt{SAT})$ for UBEV-S on a generic MDP. A careful examination of the proof for that regret bound of that term reveals that H appears because it is a deterministic upper bound on the range of $\tilde{V}_t^{\pi_k}$ and V_t^*. The optimistic value function is a random variable, but under the assumption that $r(s,a) \in [0,1]$ the agent maintains an optimistic estimate of such reward with the same constraint $\tilde{r}(s,a) \in [0,1]$, leading to $\operatorname{rng} \tilde{V}_t^{\pi_k} \leq H$ when the rewards are summed over H timesteps; likewise $V_t^* \leq H$. As we show next, $\mathcal{M}_C$ is characterized by $\operatorname{rng} V_t^* \leq 1$, which means there is not a big advantage for being in one context (i.e., state) versus another. This happens because the agent's current mistake only affects the instantaneous reward; the agent can never make "costly mistakes" that lead it to a sequence of contexts / states with low payoff as a result of that mistake as may happen on a generic MDP. Unfortunately this consideration need not be true in the "optimistic" MDP that the agent computes, that is, it is not true that $\operatorname{rng} \tilde{V}^{\pi_k} \leq 1$. However, we can relate $\operatorname{rng} \tilde{V}_t^{\pi_k}$ to $\operatorname{rng} V_t^*$ and show that $\operatorname{rng} \tilde{V}_t^{\pi_k}$ is of order 1 plus a quantity that shrinks fast enough so that the regret contribution due to uncertain system dynamics is of the same order as the rewards plus a term that does not depend on $\sqrt{T}$.

Remark: the convergence of the optimistic value function to the true one is not a property generally enjoyed by these algorithms, see for example (Bartlett & Tewari, 2009) for an extensive discussion for UCRL2 -style approaches in the infinite horizon case. However, said convergence does occur here due to the highly mixing nature of the contextual bandit problem.

6.1. Range of the True Value Function

On $\mathcal{M}_C$ a policy that greedily maximizes the instantaneous reward is optimal. Let $\overline{s}_t \overset{def}{=} \arg\max V_t^*(s)$ and $\underline{s}_t \overset{def}{=} \arg\min V_t^*(s)$ and recall that the transition dynamics $P(s,a) = \mu$ depends nor on the action a nor on the current state s:

$$\begin{cases} V_t^*(\overline{s}_t) = \max_a \left(r(\overline{s}_t, a) + \mu^\top V_{t+1}^* \right) \\ V_t^*(\underline{s}_t) = \max_a \left(r(\underline{s}_t, a) + \mu^\top V_{t+1}^* \right) \end{cases} \tag{5}$$

Since the rewards are bounded $r(\cdot,\cdot) \in [0,1]$ subtracting the two equations in 5 yields:

$$\operatorname{rng} V_t^* = \max_a r(\overline{s}_t, a) - \max_a r(\underline{s}_t, a) \leq 1. \tag{6}$$

6.2. Range of the Optimistic Value Function

Now we relate $\operatorname{rng} \tilde{V}_t^{\pi_k}$ to $\operatorname{rng} V_t^*$ by a quantity that is naturally shrinking. Our reasoning assumes that we are outside the failure event so that confidence intervals hold (confidence intervals are essentially the same as UBEV and are discussed in the appendix in section A.1). We use the notation $n_k(s,a)$ to indicate the number of visit to the (s,a) pair at the beginning of the k-th episode.

Lemma 1. *If* UBEV-S *is run on $\mathcal{M}_C$ then outside of the failure event it holds that:*

$$\operatorname{rng} \tilde{V}_t^{\pi_k} \leq 1 + \tilde{O}\left(\frac{H\sqrt{S}}{\sqrt{\min_{(s',t')} n_k(s', \pi_k(s',t'))}} \right). \tag{7}$$

Proof. We denote by $\hat{p}_k(s,a)$ the maximum likelihood vector for the transitions from (s,a). For simplicity redefine $\underline{s}_{tk} = \arg\min_s \tilde{V}_t^{\pi_k}(s)$ and $\overline{s}_{tk} = \arg\max_s \tilde{V}_t^{\pi_k}(s)$. Neglecting the reward $\tilde{r}_k(s, \pi_k(s,t))$ and the optimistic bonus ϕ while planning at timestep t (line 9 of the algorithm) yields a lower bound on the optimistic value function:

$$\min_s \tilde{V}_t^{\pi_k}(s) \overset{def}{=} \tilde{V}_t^{\pi_k}(\underline{s}_{tk}) \geq \hat{p}_k(\underline{s}_{tk}, \pi_k(\underline{s}_{tk}, t))^\top \tilde{V}_{t+1}^{\pi_k}. \tag{8}$$

Recalling that $\tilde{r}(s,a) \leq 1$, an upper bound on $\tilde{V}_t^{\pi_k}$ can also be obtained (from planning in line 9):

$$\begin{aligned} &\max_s \tilde{V}_t^{\pi_k}(s) \overset{def}{=} \tilde{V}_t^{\pi_k}(\overline{s}_{tk}) \\ &\leq \underbrace{1}_{\text{Reward}} + \hat{p}_k(\overline{s}_{tk}, \pi_k(\overline{s}_{tk}, t))^\top \tilde{V}_{t+1}^{\pi_k} + \underbrace{H\phi\left(\overline{s}_k, \pi_k(\overline{s}_k, t)\right)}_{\substack{\text{Bonus} \\ \text{(Overestimate)}}}. \end{aligned} \tag{9}$$

Subtracting 8 from 9 yields (a) below:

$$\begin{aligned} &\operatorname{rng} \tilde{V}_t^{\pi_k} \overset{def}{=} \max_s \tilde{V}_t^{\pi_k}(s) - \min_s \tilde{V}_t^{\pi_k}(s) \leq \\ &\overset{(a)}{\leq} 1 + \left(\hat{p}_k(\overline{s}_{tk}, \pi_k(\overline{s}_{tk}, t))^\top - \hat{p}_k(\underline{s}_{tk}, \pi_k(\underline{s}_{tk}, t))^\top \right) \tilde{V}_{t+1}^{\pi_k} \\ &\quad + H\phi\left(\overline{s}_k, \pi_k(\overline{s}_k, t)\right) \\ &\overset{(b)}{\leq} 1 + \|\hat{p}_k(\overline{s}_{tk}, \pi_k(\overline{s}_{tk}, t)) - \hat{p}_k(\underline{s}_{tk}, \pi_k(\underline{s}_{tk}, t))\|_1 \|\tilde{V}_{t+1}^{\pi_k}\|_\infty \\ &\quad + H\phi\left(\overline{s}_k, \pi_k(\overline{s}_k, t)\right) \end{aligned}$$

$$\overset{(c)}{\leq} 1 + H\|\hat{p}_k(\overline{s}_{tk}, \pi_k(\overline{s}_{tk}, t)) - \mu\|_1 + H\|\hat{p}_k(\underline{s}_{tk}, \pi_k(\underline{s}_{tk}, t)) - \mu\|_1 + H\phi\left(\overline{s}_k, \pi_k(\overline{s}_k, t)\right) \tag{10}$$

In (b) we used Holder's inequality and the hard bound $\operatorname{rng} \tilde{V}_{t+1}^{\pi_k} \leq H$ coupled with the triangle inequality for step (c). Before continuing the development we pause and notice that we have upper bounded $\operatorname{rng} \tilde{V}_t^{\pi_k}$ by 1 plus two concentration terms (for the transition probabilities) and the optimistic bonus, which are quantities that are shrinking on $\mathcal{M}_C$. In particular, being outside of the failure event ensures a bound on the system dynamics (this is made precise by referring to the concentration inequality of the failure event $F_k^{L_1}$ as explained in our appendix in section A.1):

$$\|\hat{p}_k(s,a)) - \mu\|_1 = \tilde{O}\left(\sqrt{\frac{S}{n_k(s,a)}}\right) \tag{11}$$

The exploration bonus defined in line 7 of algorithm 1 is also similar in magnitude:

$$H\phi(s,a) = \tilde{O}\left(\frac{H}{\sqrt{n_k(s,a)}}\right) \tag{12}$$

By definition, $\min_{(s',t')} n_k(s', \pi_k(s',t')) \leq n_k(s, \pi_k(s,t))$ for any s, t pair which allows us to combine equation 11 and 12 above to rewrite 10 as:

$$1 + \tilde{O}\left(H \frac{\sqrt{S}+1}{\sqrt{\min_{(s',t')} n_k(s', \pi_k(s',t'))}}\right) \tag{13}$$

which can be simplified to obtain the statement. □

6.3. Regret Analysis on $\mathcal{M}_C$

Lemma 1 shows that the optimistic value function on $\mathcal{M}_C$ is of order 1 plus a quantity which is related to the confidence interval of the least visited (s,a) pair under the policy selected by the agent. On $\mathcal{M}_C$ we know that the states are sampled from μ. This ensures that all states are going to be visited at a linear rate so that $\min_{(s',t')} n_k(s', \pi_k(s',t'))$ must be increasing at a linear rate. The above consideration together with lemma 1 allows us to sketch the analysis that leads to the result of theorem 1.

6.3.1. Regret Decomposition

Outside of the failure event we can use optimism to justify the first inequality below that leads to the regret decomposition for the first K episodes:

$$\begin{aligned}
\textsc{Regret}(K) &\overset{def}{=} \sum_{k=1}^{K} V_1^{\pi^*}(s) - V_1^{\pi_k}(s) \\
&\overset{Optimism}{\leq} \sum_{k=1}^{K} \tilde{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s) \\
&= \sum_{k=1}^{K} \sum_{t\in[H]} \sum_{(s,a)} w_{tk}(s,a) \Bigg(\underbrace{(\tilde{r}(s,a) - r(s,a))}_{\text{Reward Estimation and Optimism}} + \\
&+ \underbrace{(\tilde{p}(s,a) - \hat{p}(s,a))^\top \tilde{V}_{t+1}^{\pi_k}}_{\text{Transition Dynamics Optimism}} + \underbrace{(\hat{p}(s,a) - p(s,a))^\top V_{t+1}^*}_{\text{Transition Dynamics Estimation}} \\
&+ \underbrace{(\hat{p}(s,a) - p(s,a))^\top \left(\tilde{V}_{t+1}^{\pi_k} - V_{t+1}^*\right)}_{\text{Lower Order Term}} \Bigg) .
\end{aligned} \tag{14}$$

The decomposition is standard in recent RL literature (Azar et al., 2017; Dann et al., 2017).

6.3.2. The "Good" Episodes on $\mathcal{M}_C$

In the original paper (Dann et al., 2017), the authors introduce the notion of "nice" and "friendly" episodes to relate the probability of visiting a state-action pair $w_{tk}(s,a)$ to the actual number of visits there $n_k(s,a)$ (the latter is a random variable). Here we do a similar distinction directly for a regret analysis (as opposed to a PAC analysis) and we leverage the structure of $\mathcal{M}_C$. In particular we partition the set of all episodes into two, namely the set G of *good episodes* and the set of episodes that are "not good". Under good episodes we require that:

$$n_k(s,a) \geq \frac{1}{4} \sum_{i<k} \sum_{\tau\in[H]} w_{\tau i}(s,a) \tag{15}$$

holds true for *all* states s and actions a chosen by the agent's policy. In other words, we require that the number of visits $n_k(s,a)$ to the (s,a) pair is at least $\frac{1}{4}$ times its expectation. In lemma 12 in the appendix we examine the regret under non-good episodes, which can be bounded by $\tilde{O}(\frac{SAH^2}{\mu_{min}})$.

6.3.3. Regret Bound for the Optimistic Transition Dynamics (Leading Order Term)

Equipped with lemma 1 we are ready to bound the leading order term contributing to the regret under good episodes. This is the regret due to the optimistic transition dynamics which appear in equation 14. While planning for state s and timestep t (see line 9 of Algorithm 1), UBEV-S implicitly finds an optimistic transition dynamics $\tilde{p}_k(s,a)$. In particular the "optimistic" MDP satisfies the following upper bound on $\tilde{p}_k(s,a)^\top \tilde{V}_{t+1}^{\pi_k}$:

$$\overset{line\ 9}{\leq} \hat{p}_k(s,a)^\top \tilde{V}_{t+1}^{\pi_k} + (\operatorname{rng} \tilde{V}_{t+1}^{\pi_k} + \phi^+)\phi_{tk}(s,a). \tag{16}$$

Notice that line 9 of the algorithm provides additional constraints enforced by taking $\min\{\cdot,\cdot\}$, but equation 16 al-

ways remains an upper bound. Rearranging the inequality above and summing over the "good episodes", the timesteps $t \in [H]$ and all the (s,a) pairs yields an upper bound on the regret due to the optimistic transition dynamics that appears in equation 14:

$$\sum_{k\in G}\sum_{t\in[H]}\sum_{(s,a)} w_{tk}(s,a)\left(\tilde{p}_k(s,a)-\hat{p}_k(s,a)\right)^\top \tilde{V}_{t+1}^{\pi_k} \leq \sum_{k\in G}\sum_{t\in[H]}\sum_{(s,a)} w_{tk}(s,a)(\mathrm{rng}\,\tilde{V}_{t+1}^{\pi_k}+\phi^+)\phi_{tk}(s,a). \tag{17}$$

Next, notice that the correction factor ϕ^+ is updated in line 11 of the algorithm and depends on the state with the lowest visit count $\min_{(s',t')} n_k(s',\pi_k(s',t'))$. This implies the following upper bound on ϕ^+.

$$\phi^+ \lesssim \frac{H^2\sqrt{S}}{\sqrt{\min_{(s',t')} n_k(s',\pi_k(s',t'))}}\,\mathrm{polylog}(\cdot). \tag{18}$$

At this point we can substitute the definition of $\phi_{tk}(s,a)$ (line 7 of Algorithm 1) and put all the constants and logarithmic quantities in $\mathrm{polylog}(\cdot)$ to upper bound 17 as follows:

$$\lesssim \sum_{k\in G}\sum_{t\in[H]}\sum_{(s,a)} w_{tk}(s,a)\frac{\mathrm{rng}\,\tilde{V}_{t+1}^{\pi_k}}{\sqrt{n_k(s,a)}}\,\mathrm{polylog}(\cdot) + \sum_{k\in G}\sum_{t\in[H]}\sum_{(s,a)} \frac{w_{tk}(s,a)\sqrt{S}H^2\,\mathrm{polylog}(\cdot)}{\sqrt{\min_{(s',t')} n_k(s',\pi_k(s',t'))\times n_k(s,a)}}. \tag{19}$$

Finally we substitute lemma 1:

$$\lesssim \underbrace{\sum_{k\in G}\sum_{t\in[H]}\sum_{(s,a)} w_{tk}(s,a)\frac{1}{\sqrt{n_k(s,a)}}\,\mathrm{polylog}(\cdot)}_{\text{Leading Order Term}} + \underbrace{\sum_{\substack{k\in G\\ t\in[H]\\ (s,a)}} \frac{w_{tk}(s,a)\sqrt{S}H^2\,\mathrm{polylog}(\cdot)}{\sqrt{\min_{(s',t')} n_k(s',\pi_k(s',t'))\times n_k(s,a)}}}_{\text{Lower Order Term}}. \tag{20}$$

and apply Cauchy-Schwartz to get (omitting $\mathrm{polylog}(\cdot)$ factors):

$$\sqrt{\underbrace{\sum_{k\in G}\sum_{t\in[H]}\sum_{(s,a)} w_{tk}(s,a)}_{\leq T}}\sqrt{\underbrace{\sum_{k\in G}\sum_{t\in[H]}\sum_{(s,a)} \frac{w_{tk}(s,a)}{n_k(s,a)}}_{\tilde{O}(SA)}}+$$

$$\sqrt{S}H^2\sqrt{\underbrace{\sum_{\substack{k\in G\\ t\in[H]\\ (s,a)} } \frac{w_{tk}(s,a)}{n_k(s,a)}}_{\tilde{O}(SA)}}\sqrt{\underbrace{\sum_{\substack{k\in G\\ t\in[H]\\ (s,a)} } \frac{w_{tk}(s,a)}{\min\limits_{(s',t')} n_k(s',\pi_k(s',t'))}}_{(\star)}}. \tag{21}$$

The sum of the "visitation ratios" $\frac{w_{tk}(s,a)}{n_k(s,a)}$ under good episodes can be bounded in the usual way by $\tilde{O}(SA)$ by using a pigeonhole argument and will not be discussed further (details are in the appendix). To bound $(\star)$ we need to work a little more. The main problem is that the ratio

$$\frac{w_{tk}(s,a)}{\min_{(s',t')} n_k(s',\pi_k(s',t'))} \tag{22}$$

is a ratio between the visitation probability of a certain state (s,a) pair and the visit count of a *different* pair. For a general MDP these two quantities are not related as there can be states that are clearly suboptimal and are visited finitely often by PAC algorithms. As a result, $\sqrt{(\star)}$ can grow like $\sqrt{T}$ and it is not a lower order term. This is the key step where we leverage the underlying structure of the problem. With contextual bandits all contexts are going to be visited with probability at least μ_{min}. Since the analysis is under good episodes, for a fixed (s',t') pair we know that $n_k(s',\pi_k(s',t'))$ must increase by at least $\frac{1}{4}\mu_{min}$ every episodes. There are only $S\times A$ possible candidates for the (s',a') pair with the lowest visit count. Recalling $\sum_{t\in[H]}\sum_{(s,a)} w_{tk}(s,a) = H$, the final result then follows by pigeonhole (the computation is in the appendix).

$$(\star) = \sum_{k\in G}\frac{H}{\min_{(s',t')} n_k(s',\pi_k(s',t'))} = \tilde{O}\left(\frac{SAH}{\mu_{min}}\right). \tag{23}$$

This completes the sketch of the regret bound for the "Optimistic Transition Dynamics" with a regret contribution of order:

$$\tilde{O}\left(\sqrt{SAT}+\sqrt{SH}H^2\times\frac{SA}{\sqrt{\mu_{min}}}\right). \tag{24}$$

Remark: Although for simplicity we conduct here the analysis for the regret only, UBEV-S is still a uniformly-PAC algorithm and strong PAC guarantees can be obtained on $\mathcal{M}_C$ as well. The analysis for the regret due to the rewards, the estimation of the transition dynamics and the lower oder term can be found in the appendix. Together with the regret in non-good episodes they imply the regret bound of theorem 1.

7. Discussion, Related Work and Future Work

A natural question is whether there is something special about the UBEV algorithm, or if other MDP RL algorithms

with theoretical bounds can also be shown to have provably better or optimal regret bounds on contextual bandit problems. While we focused on UBEV because it matched (in the dominant terms) the best regret bounds for contextual bandits when run in such settings, we do think other MDP algorithms can yield strong (though not optimal) regret bounds when run in contextual bandits. For example, (Jiang et al., 2017) proposes OLIVE, a probably approximately correct algorithm with bounds for a broad number of settings which can potentially adapt to a CMAB problem if the Bellman rank is known. If the bellman rank is not known in advance (as is our case) a way around this issue is to use the "doubling trick". However, the resulting PAC bound of OLIVE on CMABs would scale in a way which is suboptimal in H. Another interesting candidate for our analysis on CMABs is given in (Bartlett & Tewari, 2009) the authors propose REGAL, a UCRL2-variant which can potentially achieve a $\tilde{O}(S\sqrt{AT})$ bound on CMABs while retaining a worst-case $\tilde{O}(DS\sqrt{AT})$ regret in generic MDPs (here D is the MDP diameter). The simplification on CMABs follows directly from the computation of the span (which is equivalent to the range here) of the optimal bias vector. Still, this result is not completely satisfactory because the lower bound is not achieved and REGAL *must know* the range of the bias vector in advance. Another noteworthy variant of UCRL2 is discussed in (Maillard et al., 2014). There the authors introduce a new norm and its dual (instead of the classical 1-norm and ∞-norm, respectively) to better capture the effect of the MDP transition dynamics. The result that they obtain does depend on a measure of the MDP complexity (constant C in their regret bound). This is essentially the variance of the value function, so $C = O(1)$ on CMABs; despite moving in the right direction, the resulting bound is still of order $\tilde{O}(DS\sqrt{AT})$ on CMABs.

By contrast, our analysis of vanilla UCRL2 (Jaksch et al., 2010) (see appendix C for extensive details) shows an improved regret bound of $\tilde{O}(S\sqrt{AT})$ if UCRL2 is run on CMABs which is better (although not optimal) than the UCRL2 worst-case bound for MDPs $\tilde{O}(DS\sqrt{AT})$. The key insight to obtain this result is that the MDP diameter D is an upper bound to a key quantity in the analysis of UCRL2 , and can be more tightly bounded in contextual bandit domains. This analysis suggests that if an algorithm for infinite-horizon MDPs is constructed using $\sqrt{S}$-tighter confidence intervals like in UBEV or UCBVI from (Azar et al., 2017) then a bound of order $\tilde{O}(\sqrt{SAT})$ should be achievable on an infinite horizon $\mathcal{M}_C$.

This work raises a number of interesting questions, in particular whether similar results are possible for other pairings of algorithms and domains: can we have algorithms designed for partially observable reinforcement learning that inherit the best performance of the setting they operate in, whether it is a bandit, contextual bandit, MDP or POMDP? As a step towards such exploration, we analyzed whether a MDP RL algorithm operating in a multi-armed bandit could match the upper bound on regret for such settings. In a multi-armed bandit there are no states, and the reward is solely a function of the arm (action) played. Regret for MABs must scale at least as $\Omega(\sqrt{AT})$, the lower bound for such setting. In our preliminary investigations, our analysis of UCRL2 when operating in a MAB (still in section C in the appendix) yielded an additional $\sqrt{S}$ dependence. It is a very interesting question whether existing or new MDP algorithms that explicitly or implicitly perform state aggregation (Mandel et al., 2016; Doshi-Velez, 2009) can yield a performance that matches the dominant terms of a bandit-specific regret analysis. Another important question is whether similar analyses are possible for reinforcement learning algorithms designed for very large or infinite state spaces, as well as an empirical investigation to see whether existing RL algorithms for more complex settings experimentally match algorithms designed for simpler settings when executing in said simpler settings.

Finally, our analysis for UBEV-S highlights a dependence on the minimum visitation probability μ_{min} which is absent in bandit analyses. We think that this can be avoided by a more careful design of the exploration bonus that re-weights the next-state uncertainty by the transition probability estimated empirically, see for example (Dann & Brunskill, 2015; Azar et al., 2017). For simplicity in this paper we focused on tabular bandits and therefore UBEV-S cannot handle general Contextual Bandits which use function approximations (e.g, (Abbasi-Yadkori et al., 2011)).

8. Conclusion

The ultimate goal of Reinforcement Learning is to design algorithms that can learn online and achieve the best performance afforded by the difficulty of the underlying domain. In this work we have introduced a minor variant of an existing RL algorithm that automatically provides strong regret guarantees whether it is deployed in a MDP or if the domain actually belongs to a simpler setting, a tabular contextual bandit, matching the lower bound in the dominant terms in the second setting. Note that the algorithm is not informed of this structure. This work suggests that already existing RL algorithms can inherit tighter theoretical guarantees if the domain turns out to have additional structure and yields many interesting next steps for the analysis and creation of algorithms for other settings, particularly the function approximation case.

Acknowledgements

Christopher Dann and the anonymous reviewers are acknowledged for providing very useful feedback which improved the quality of this paper.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Improved algorithms for linear stochastic bandits. In *NIPS*, 2011.

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 2012.

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. In *NIPS*, 2006.

Auer, P., Bianchi, N. C., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *ICML*, 2017.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *NIPS*, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *NIPS*, 2017.

Doshi-Velez, F. The infinite partially observable markov decision process. In *NIPS*, 2009.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.

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

Jiang, N., Krishnamurthy, A., Agarwal, A., Langforda, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *ICML*, 2017.

Maillard, O.-A., Mann, T. A., and Mannor, S. "how hard is my mdp?" the distribution-norm to the rescue. In *NIPS*, 2014.

Mandel, T., Liu, Y.-E., Brunskill, E., , and Popovic, Z. Efficient bayesian clustering for reinforcement learning. In *IJCAI*, 2016.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. In *Neural Information Processing Systems*, 2013.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. In *Arxiv*, 2016. URL https://arxiv.org/pdf/1608.02732.pdf. https://arxiv.org/pdf/1608.02732.pdf.

Shaked, M. and Shanthikumar, J. G. *Stochastic Orders*. Springer Series in Statistics, 2007.