Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds

Andrea Zanette [1] **Emma Brunskill** [2]

Abstract

Strong worst-case performance bounds for episodic reinforcement learning exist but fortunately in practice RL algorithms perform much better than such bounds would predict. Algorithms and theory that provide strong problem-dependent bounds could help illuminate the key features of what makes a RL problem hard and reduce the barrier to using RL algorithms in practice. As a step towards this we derive an algorithm and analysis for finite horizon discrete MDPs with state-of-the-art worst-case regret bounds and substantially tighter bounds if the RL environment has special features but without apriori knowledge of the environment from the algorithm. As a result of our analysis, we also help address an open learning theory question (Jiang & Agarwal, 2018) about episodic MDPs with a constant upper-bound on the sum of rewards, providing a regret bound function of the number of episodes with no dependence on the horizon.

1. Introduction

In reinforcement learning (RL) an agent must learn how to make good decision without having access to an exact model of the world. Most of the literature for provably efficient exploration in Markov decision processes (MDPs) (Jaksch et al., 2010; Osband et al., 2013; Lattimore & Hutter, 2014; Dann & Brunskill, 2015; Dann et al., 2017; Osband & Roy, 2017; Azar et al., 2017; Kakade et al., 2018) has focused on providing near-optimal worst-case performance bounds. Such bounds are highly desirable as they do not depend on the structure of the particular environment considered and therefore hold for even extremely hard-to-learn MDPs.

Fortunately in practice reinforcement learning algorithms often perform far better than what these problem-independent bounds would suggest. While we may observe better or worse performance empirically on different MDPs, we would like to derive a more systematic understanding of what types of decision processes are inherently easier or more challenging for RL. This motivates our interest in deriving algorithms and theoretical analyses that provide problem-dependent bounds. Ideally, such algorithms will do as well as RL solutions designed for the worst case if the problem is pathologically difficult and otherwise match the performance bounds of algorithms specifically designed for a particular problem subclass. This exciting scenario might bring considerable saving in the time spent designing domain-specific RL solutions and in training a human expert to judge and recognize the complexity of different problems. An added benefit would include the robustness of the RL solution in case the actual model does not belong to the identified subclass, yielding increased confidence to deploying RL to high-stakes applications.

Towards this goal, in this paper we contribute with a new algorithm for episodic tabular reinforcement learning which automatically provides provably stronger regret bounds in many domains which have a small variance of the optimal value function (in the infinite horizon setting, this variance has been called the *environmental norm* (Maillard et al., 2014)). Indeed, there is good reason to believe that some features of the range or variability of the optimal value function should be a critical aspect of the hardness of reinforcement learning in a MDP. Many worst-case bounds for finite-state MDPs scale with a *worst case* bound on the range / magnitude of the value function, such as the diameter D for an infinite-horizon setting and the horizon H in an episodic problem. Note that here both D and H arise in the analyses as upper bounds on the (range of the) *optimistic* value function across the entire MDP[1]. As more samples are collected, one would hope that the agent's optimistic value function converges to the true optimal value function. Unfortunately this is not the case, see for example (Jaksch et al., 2010; Bartlett & Tewari, 2009; Zanette & Brunskill, 2018) for a

[1] Institute for Computational and Mathematical Engineering, Stanford University, USA [2] Department of Computer Science, Stanford University, USA. Correspondence to: Andrea Zanette <zanette@stanford.edu>, Emma Brunskill <ebrun@cs.stanford.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] Many RL algorithms with strong performance bounds rely on the principle of optimism under uncertainty and compute an optimistic value function.

discussion of this. As a result, most prior analyses bounded the optimistic value function by generic quantities like D or H regardless of the actual behaviour of the optimal value function.

While the majority of formal performance guarantees has focused on bounds for the worst case, there have been several contributions of algorithms and/or theoretical analyses focused on MDPs with particular structure. Such contributions have focused on the infinite horizon setting, which involves a number of subtleties that are not present in the finite horizon setting we consider, which is likely a cause of the less strong results in this setting which can require stronger input knowledge on a tighter range on the possible value function (Bartlett & Tewari, 2009; Fruit et al., 2018), or do not match in dominant terms strong bounds for the worst case setting (Maillard et al., 2014). We defer more detailed discussion of related work to Section 7, except to briefly highlight likely the most closely related recent result from (Talebi & Maillard, 2018). Like us, Talebi and Maillard provide a problem-dependent regret bound that scales as a function of the variance of the next state distribution. However, like the aforementioned references, their focus is on the infinite horizon setting. In this setting the authors achieve their resulting regret bound under an assumption that the mixing time of the MDP is such that all states are visited at a linear rate in expectation regardless of the agent's chosen policies. This mixing rate, that could be exponential in certain MDPs, appears in the regret bound. In our, arguably simpler finite horizon setting, we do not use an assumption on the mixing rate of the MDP and we instead pursue a different proof technique to obtain strong results for this setting.

More precisely, in this paper we derive an algorithm for finite horizon discrete MDPs and associated analysis that yields state-of-the art worst-case regret bounds of order $\tilde{O}(\sqrt{HSAT})$ in the leading term while improving if the environment has next-state value function variance (i.e., small environmental norm) or bounded total possible reward. Compared to the existing literature, our work

- Maintains state of the art worst-case guarantees (Azar et al., 2017) for episodic finite horizon settings,
- Improves the regret bounds of (Zanette & Brunskill, 2018) when deployed in the same settings,
- Provides demonstration that characterizing problems using environmental norm (Maillard et al., 2014) can yield substantially tighter theoretical guarantees in the finite horizon setting,
- Identifies problem classes with low environmental norm which are of significant interest, including deterministic domains, single-goal MDPs, and high stochasticity domains, and
- Helps address an open learning theory problem (Jiang & Agarwal, 2018), showing that for their setting, we obtain a regret bound that scales with no dependence on the planning horizon in the dominant terms.

The paper is organized as follows: we recall some basic definitions in Section 2 and describe the algorithm in Section 3. We state and comment the main result in Section 4, discuss how this helps address an open learning theory problem in Section 5 and then describe selected problem-dependent bounds in Section 6. The analysis is sketched in Section 4.1. Due to space constraints, most proofs are in the full report available at:
`https://arxiv.org/abs/1901.00210`.

2. Preliminaries and Definitions

In this section we introduce some notation and definitions. We consider undiscounted finite horizon MDPs (Sutton & Barto, 1998), which are defined by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, H \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces with cardinality S and A, respectively. We denote by $p(s' \mid s, a)$ the probability of transitioning to state s' after taking action a in state s while $r(s, a) \in [0, 1]$ is the average instantaneous reward collected. We label with $n_k(s, a)$ the visits to the (s, a) pair at the beginning of the k-th episode. The agent interacts with the MDP starting from arbitrary initial states in a sequence of episodes $k \in [K]$(where $[K] = \{j \in \mathbb{N} : 1 \leq j \leq K\}$) of fixed length H by selecting a policy $\tilde{\pi}_k$ which maps states s and timesteps t to actions. Each policy identifies a value function for every state s and timestep $t \in [H]$ defined as $V_t^{\tilde{\pi}_k}(s_t) = \mathbb{E}_{(s,a)\sim\tilde{\pi}_k} \sum_{i=t}^{H} r(s, a)$ which is the expected return until the end of the episode (the conditional expectation is over the pairs (s, a) encountered in the MDP upon starting from s_t). The optimal policy is indicated with π^* and its value function as $V_t^{\pi^*}$. We indicate with $\underline{V}_{t+1k}^{\tilde{\pi}_k}$ and $\overline{V}_{t+1k}^{\tilde{\pi}_k}$, respectively, a pointwise underestimate, respectively, overestimate, of the optimal value function and with $\hat{p}_k(\cdot \mid s, a)$ and $\hat{r}_k(s, a)$ the MLE estimates of $p(\cdot \mid s, a)$ and $r(s, a)$. We focus on deriving a high probability upper bound on the $\textsc{Regret}(K) \overset{def}{=} \sum_{k\in[K]} \left(V_1^{\pi^*}(s_k) - V_1^{\tilde{\pi}_k}(s_k)\right)$ to measure the agent's learning performance. 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 also use the $\lesssim, \gtrsim, \simeq$ notation to mean $\leq, \geq, =$, respectively, up to a numerical constant and indicate with $\|X\|_{2,p}$ the 2-norm of a random variable[2] under p, i.e., $\|X\|_{2,p} \overset{def}{=} \sqrt{\mathbb{E}_p X^2} \overset{def}{=} \sqrt{\sum_{s'} p(s')X^2(s')}$ if $p(\cdot)$ is its probability mass function.

[2] To be precise, this is a norm between classes of random variables that are almost surely the same

3. EULER

We define the maximum per-step conditional variance (conditioning is on the (s, a) pair) for a particular MDP as $\mathbb{Q}^*$:

$$\mathbb{Q}^* \overset{def}{=} \max_{s,a,t} \left(\operatorname{Var} R(s,a) + \operatorname*{Var}_{s^+ \sim p(s,a)} V_{t+1}^{\pi^*}(s^+) \right) \quad (1)$$

where $R(s, a)$ is the reward random variable in (s, a). This definition is identical to the environmental norm (Maillard et al., 2014) but here we will generally refer to it as the maximum conditional value variance, in order to connect with other work which explicitly bounds the variance.

We introduce the algorithm *Episodic Upper Lower Exploration in Reinforcement learning* (EULER) which adopts the paradigm of "optimism under uncertainty" to conduct exploration. Recent work (Dann & Brunskill, 2015; Dann et al., 2017; Azar et al., 2017) has demonstrated how the choice of the exploration bonus is critical to enabling tighter *problem-independent* performance bounds. Indeed minimax worst case regret bounds have been obtained by using a Bernstein-Friedman-type reward bonus defined over an empirical quantity related very closely to the conditional value variance $\mathbb{Q}^*$, plus an additional correction term necessary to ensure optimism (Azar et al., 2017).

Similarly, in our algorithm we use a bonus that combines an empirical Bernstein type inequality for estimating the $\mathbb{Q}^*$ conditional variance, coupled with a different correction term which explicitly accounts for the value function uncertainty. We provide pseudocode for EULER which details the main procedure in Figure 1. Notice that EULER has the same computational complexity as value iteration.

4. Main Result

Now we present our main result, which is a problem-dependent high-probability regret upper bound for EULER in terms of the underlying max conditional variance $\mathbb{Q}^*$ and maximum return. Crucially, EULER is **not** provided with $\mathbb{Q}^*$ and the value of the max return. We also prove a worst-case guarantee that matches the established (Osband & Van Roy, 2016; Jaksch et al., 2010) lower bound of $\Omega(\sqrt{HSAT})$ in the dominant term. We introduce the following definition:

Definition 1 (Max Return). *We define as $\mathcal{G} \in \mathbb{R}$ the maximum (random) return in an episode upon following any policy π from any starting state s_0, i.e., the deterministic upper bound to:*

$$\sum_{t=1}^{H} R(s_t, \pi(s_t)) \leq \mathcal{G}, \quad \forall \pi, s_0. \quad (2)$$

where the states $s_1, \dots, s_H$ are the (random) states generated upon following the trajectory identified by the policy π from s_0.

Theorem 1 (Problem Dependent High Probability Regret Upper Bound for EULER). *With probability at least $1 - \delta$ the regret of* EULER *is bounded for any time $T \leq KH$ by the minimum between*

$$\tilde{O}\left(\sqrt{\mathbb{Q}^* SAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})\right) \quad (3)$$

and

$$\tilde{O}\left(\sqrt{\frac{\mathcal{G}^2}{H} SAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})\right), \quad (4)$$

jointly for all episodes $k \in [K]$.

While the maximum conditional variance $\mathbb{Q}^*$ is always upper bounded by $\mathcal{G}$ if rewards are positive and bounded, we include both forms of regret bound for two reasons. First, the second bound is tighter than naively upper bounding $\mathbb{Q}^* \leq \mathcal{G}^2$ by a factor of H. Second, we will shortly see that both quantities can provide insights into which instances of MDP domains can have lower regret.

In addition, since the rewards are in $[0, 1]$, we immediately have that $\mathcal{G}^2 \leq H^2$, and thereby obtain a worst-case regret bound expressed in the following corollary:

Corollary 1.1. *With probability at least $1 - \delta$ the regret of* EULER *is bounded for any time $T \leq KH$ by*

$$\tilde{O}\left(\sqrt{HSAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})\right). \quad (5)$$

This matches in the dominant term the minimax regret problem independent bounds for tabular episodic RL settings (Azar et al., 2017). Therefore, the importance of our theorem 1 lies in providing problem dependent bounds (equation 3,4) while simultaneously matching the existing best worst case guarantees (equation 5). We shall shortly show that our results help address a recent open question on the performance dependence of episodic MDPs on the horizon (Jiang & Agarwal, 2018).

4.1. Sketch of the Theoretical Analysis

We devote this section to the sketch of the main point of the regret analysis that yields problem dependent bounds. Readers that wish to focus on how our results yield insight into the complexity of solving different problems may skip ahead to the next section. Central to the analysis is the relation between the agent's optimistic MDP and the "true" MDP. A more detailed overview of the proof is given in section C of the appendix, while the rest of the appendix presents the detailed analysis under a more general framework.

Regret Decomposition Denote with $\mathbb{E}_{(s,a)\sim\tilde{\pi}_k}$ the expectation taken along the trajectories identified by the agent's policy $\tilde{\pi}_k$. A standard regret decomposition is given below

Algorithm 1 EULER for Stationary Episodic MDPs

```
Input: $\delta' = \frac{1}{7}\delta$, $b^r_k(s,a) = \sqrt{\frac{2\widehat{\mathrm{Var}}R(s,a)\ln\frac{4SAT}{\delta'}}{n_k(s,a)}} + \frac{7\ln\frac{4SAT}{\delta'}}{3(n_k(s,a)-1)}$, $\phi(s,a) = \sqrt{\frac{2\widehat{\mathrm{Var}}_{\widehat{p}_k(s,a)}(\overline{V}^{\tilde{\pi}_k}_{t+1k})\ln\frac{4SAT}{\delta'}}{n_k(s,a)}} + \frac{H\ln\frac{4SAT}{\delta'}}{3(n_k(s,a)-1)}$,
    $B_p = H\sqrt{\frac{2\ln(4SAT)}{\delta'}}$, $B_v = \sqrt{\frac{2\ln(4SAT)}{\delta'}}$, $J = \frac{H\ln(4SAT)/\delta'}{3}$.
for $k = 1, 2, \ldots$ do
  for $t = H, H-1, \ldots, 1$ do
    for $s \in \mathcal{S}$ do
      for $a \in \mathcal{A}$ do
        $\hat{p} = \frac{p_{sum}(\cdot,s,a)}{n_k(s,a)}$
        $b^{pv}_k = \phi(\hat{p}(s,a), \overline{V}_{t+1}) + \frac{1}{\sqrt{n(s,a)}}\left(\frac{4J+B_p}{\sqrt{n_k(s,a)}} + B_v\|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}}\right)$
        $Q(a) = \min\{H - t, \hat{r}_k(s, \tilde{\pi}_k(s,t)) + b^r_k(s,a) + \hat{p}^\top\overline{V}_{t+1} + b^{pv}_k\}$
      end for
      $\tilde{\pi}_k(s,t) = \arg\max_a Q(a)$
      $\overline{V}_t(s) = Q(\tilde{\pi}_k(s,t))$
      $b^{pv}_k = \phi(\hat{p}(s, \tilde{\pi}_k(s,t)), \underline{V}_{t+1}) + \frac{1}{\sqrt{n(s,\tilde{\pi}_k(s,t))}}\left(\frac{4J+B_p}{\sqrt{n_k(s,\tilde{\pi}_k(s,t))}} + B_v\|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}}\right)$
      $\underline{V}_t(s) = \max\{0, \hat{r}_k(s, \tilde{\pi}_k(s,t)) - b^r_k(s, \tilde{\pi}_k(s,t)) + \hat{p}^\top\underline{V}_{t+1} - b^{pv}_k\}$
    end for
  end for
  Evaluate policy $\tilde{\pi}_k$ and update MLE estimates $\hat{p}(\cdot,\cdot)$ and $\hat{r}(\cdot,\cdot)$
end for
```

(see (Dann et al., 2017; Azar et al., 2017)):

$$
\begin{aligned}
\textsc{Regret}(K) \leq \sum_{\substack{k\in[K]\\ t\in[H]\\ (s,a)\in\mathcal{S}\times\mathcal{A}}} \mathbb{E}_{(s,a)\sim\tilde{\pi}_k}\Bigg(& \underbrace{\tilde{r}_k(s,a) - r(s,a)}_{\substack{\text{Reward}\\ \text{Estimation}\\ \text{and Optimism}}} \\
& + \underbrace{\left(\tilde{p}_k(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a)\right)^\top \overline{V}^{\tilde{\pi}_k}_{t+1k}}_{\substack{\text{Transition}\\ \text{Dynamics}\\ \text{Optimism}}} \\
& + \underbrace{\left(\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a)\right)^\top V^{\pi^*}_{t+1}}_{\substack{\text{Transition}\\ \text{Dynamics}\\ \text{Estimation}}} \\
& + \underbrace{\left(\hat{p}_k(\cdot \mid s,a) - p(\cdot \mid s,a)\right)^\top \left(\overline{V}^{\tilde{\pi}_k}_{t+1k} - V^{\pi^*}_{t+1}\right)}_{\substack{\text{Lower}\\ \text{Order}\\ \text{Term}}}\Bigg)
\end{aligned}
\tag{6}
$$

Here, the "tilde" quantities $\tilde{r}$ and $\tilde{p}$ represent the agent's optimistic estimate. Of the terms in equation 6, the "Transition Dynamics Estimation" and "Transition Dynamics Optimism" are the leading terms to bound as far as the regret is concerned. The former is expressed through MDP quantities (i.e, the true transition dynamics $p(\cdot \mid s,a)$ and the optimal value function $V^{\pi^*}_{t+1}$) and hence it can be readily bounded using Bernstein Inequality, giving rise to a problem dependent regret contribution. More challenging is to show that a similar simplification can be obtained for the "Transition Dynamics Optimism" term which relies on the agent's optimistic estimates $\tilde{p}_k(\cdot \mid s,a)$ and $\overline{V}^{\tilde{\pi}_k}_{t+1k}$.

Optimism on the System Dynamics Said term $\left(\tilde{p}_k(\cdot \mid s,a) - \hat{p}_k(\cdot \mid s,a)\right)^\top \overline{V}^{\tilde{\pi}_k}_{t+1k}$ represents the difference between the agent's imagined (i.e., optimistic) transition $\tilde{p}_k(\cdot \mid s,a)$ and the maximum likelihood transition $\hat{p}_k(\cdot \mid s,a)$ weighted by the next-state optimistic value function $\overline{V}^{\tilde{\pi}_k}_{t+1k}$. By construction, this is the exploration bonus which incorporates an estimate of the conditional variance over the value function. This bonus reads:

$$
\substack{\text{Transition}\\ \text{Dynamics}\\ \text{Optimism}} = \substack{\text{Exploration}\\ \text{Bonus}} \approx \overbrace{\underbrace{\sqrt{\frac{\mathrm{Var}_{s\sim\hat{p}_k(\cdot|s,a)}\,\overline{V}^{\tilde{\pi}_k}_{t+1k}}{n_k(s,a)}}}_{}}^{\substack{\text{Dominant Term}\\ \text{of Exploration Bonus}}}\underbrace{\vphantom{\sqrt{\frac{a}{b}}} + \frac{H}{n_k(s,a)}}_{\substack{\text{Empirical Bernstein evaluated}\\ \text{with Empirical Value Function}}}
\tag{7}
$$

$$
+ \underbrace{\left(\frac{\|\overline{V}^{\tilde{\pi}_k}_{t+1k} - \underline{V}^{\tilde{\pi}_k}_{t+1k}\|_{\hat{p}_k(\cdot|s,a)}}{\sqrt{n_k(s,a)}} + \frac{H}{n_k(s,a)}\right)}_{\text{Correction Bonus}}
\tag{8}
$$

In the above expression the "Correction Bonus" is needed to ensure optimism because the "Empirical Bernstein" contribution is evaluated with the agent's estimate $\overline{V}^{\tilde{\pi}_k}_{t+1k}$ as opposed to the real $V^{\pi^*}_{t+1}$. If we assume that $\|\overline{V}^{\tilde{\pi}_k}_{t+1k} - \underline{V}^{\tilde{\pi}_k}_{t+1k}\|_{\hat{p}_k(\cdot|s,a)}$ shrinks quickly enough, then the "Dominant Term" in equation 7 is the most slowly decaying term with a rate $1/\sqrt{n}$. If that term involved the true transition dynamics $p(\cdot \mid s,a)$ and value function $V^{\pi^*}_{t+1}$ (as opposed to the agent's estimates $\hat{p}_k(\cdot \mid s,a)$ and $\overline{V}^{\tilde{\pi}_k}_{t+1k}$) then problem dependent bounds would follow in the same way as they could be proved for the "Transition Dynamics Estimation". Therefore we wish to study the relation between such "Dominant Term" evaluated with the agent's MDP estimates vs the MDP's true parameters.

Convergence of the System Dynamics in the Dominant Term of the Exploration Bonus Theorem 10 of (Maurer & Pontil, 2009) gives the high probability statement:

$$\left| \sqrt{\operatorname*{Var}_{\hat{p}_k(\cdot|s,a)} V^{\pi^*}_{t+1}} - \sqrt{\operatorname*{Var}_{p(\cdot|s,a)} V^{\pi^*}_{t+1}} \right| \lessapprox \frac{H}{n_k(s,a)} \tag{9}$$

to quantify the rate of convergence of the empirical variance using the true value function (this leads to the empirical version of Bernstein's inequality). Next, two basic computations yield:

$$\begin{aligned} &\left| \sqrt{\operatorname*{Var}_{\hat{p}_k(\cdot|s,a)} V^{\pi^*}_{t+1}} - \sqrt{\operatorname*{Var}_{\hat{p}_k(\cdot|s,a)} \overline{V}^{\tilde{\pi}_k}_{t+1k}} \right| \\ &\leq \|\overline{V}^{\tilde{\pi}_k}_{t+1k} - V^{\pi^*}_{t+1}\|_{\hat{p}_k(\cdot|s,a)} \leq \|\overline{V}^{\tilde{\pi}_k}_{t+1k} - \underline{V}^{\tilde{\pi}_k}_{t+1k}\|_{\hat{p}_k(\cdot|s,a)} \end{aligned} \tag{10}$$

Together, equation 9 and 10 quantify the rate of convergence of $\operatorname{Var}_{s\sim\hat{p}_k(\cdot|s,a)} \overline{V}^{\tilde{\pi}_k}_{t+1k}$ to $\operatorname{Var}_{s\sim p(\cdot|s,a)} V^{\pi^*}_{t+1}$, yielding the following upper bound for the dominant term of the exploration bonus:

$$\begin{aligned} \text{Dominant Term of Exploration Bonus} = \sqrt{\frac{\operatorname{Var}_{\hat{p}_k(\cdot|s,a)} \overline{V}^{\tilde{\pi}_k}_{t+1k}}{n_k(s,a)}} \lessapprox \underbrace{\sqrt{\frac{\operatorname{Var}_{p(\cdot|s,a)} V^{\pi^*}_{t+1}}{n_k(s,a)}}}_{\text{Gives Problem Dependent Bounds}} \\ + \underbrace{\frac{H}{n_k(s,a)} + \frac{\|\overline{V}^{\tilde{\pi}_k}_{t+1k} - \underline{V}^{\tilde{\pi}_k}_{t+1k}\|_{\hat{p}_k(\cdot|s,a)}}{\sqrt{n_k(s,a)}}}_{\text{Shrinks Faster}} \end{aligned} \tag{11}$$

In words, we have decomposed the "Dominant Term of the Exploration Bonus" (which is constructed using the agent's available knowledge) as a problem-dependent contribution (that is equivalent to Bernstein Inequality evaluated as if the model was known) and a term that accounts for the distance between the the true and empirical model, expressed as (computable) upper and lower bounds on the value function. This additional term shrinks faster that the former. It is precisely this "Correction Bonus" that we use in equation 7 and in the definition of the Algorithm itself.

What gives rise to problem dependent bounds? Our analysis highlights EULER uses a Bernstein inequality on the empirical estimate of the conditional variance of the next state values, with a correction term $\|\overline{V}^{\tilde{\pi}_k}_{t+1k} - \underline{V}^{\tilde{\pi}_k}_{t+1k}\|_{\hat{p}_k(\cdot|s,a)}$ function of the inaccuracy of the value function estimate at the next-step states re-weighted by their relative importance as encoded in the experienced transitions $\hat{p}_k(\cdot \mid s,a)$. Said correction term is of high value only if the successor states do not have an accurate estimate for the value function *and* they are going to be visited with high probability. A pigeonhole argument guarantees that this situation cannot happen for too long ensuring fast decay of $\|\overline{V}^{\tilde{\pi}_k}_{t+1k} - \underline{V}^{\tilde{\pi}_k}_{t+1k}\|_{\hat{p}_k(\cdot|s,a)}$ and therefore of the whole "Correction Bonus" of eq. 7.

Our primary analysis yields a regret bound that scales directly with the (unknown to the algorithm) problem-dependent $\mathbb{Q}^*$ max conditional variance of the next state values. We further extend this to a bound directly in terms of the max returns $\mathcal{G}$ by using a law of total variance argument.

Notice that such considerations and results would not be achievable by a naive application of an Hoeffding-like inequality as the latter would put equal weight on all successor states, but the accuracy in the estimation of $V^{\pi^*}_{t+1}(\cdot)$ only shrinks in a way that depends on the visitation frequency of said successor states as encoded in $\hat{p}_k(\cdot \mid s,a)$. The key to enable problem dependent bound is, therefore, to re-weight the importance of the uncertainty on the value function of the successor states by the corresponding visitation probability, which Bernstein Inequality implicitly does.

There exist other algorithms (e.g. (Dann & Brunskill, 2015; Azar et al., 2017)) which are based on Bernstein's inequality but to our knowledge they have not been analyzed in a way that provably yields problem dependent bounds as those presented here.

5. Horizon Dependence in Dominant Term

In this section we show that our result can help address a recently posed open question in the learning theory community (Jiang & Agarwal, 2018). The question posed centers on the whether there should exist a necessary dependence of sample complexity and regret lower bounds on the planning horizon H for episodic tabular MDP reinforcement learning tasks. Existing lower bound results for sample complexity (Dann & Brunskill, 2015) depend on the horizon, as do the best existing minimax regret bounds under asymptotic assumptions (Azar et al., 2017). However, such results have been derived under the common assumption of reward uniformity, that per-time-step rewards are bounded between 0 and 1, yielding a total value bounded by 0 and H. Jiang & Agarwal (2018) instead pose a more general setting, in which they assume that the rewards are positive and $\sum_{h=1}^{H} r_h \in [0,1]$ holds almost surely: note the standard setting of reward uniformity can be expressed in this setting by first normalizing all rewards by dividing by H. The authors then ask that if in this new, more general setting of tabular episodic RL there is necessarily a dependence on the planning horizon in the lower bounds. Note that in this setting, the prior existing lower bounds on the sample complexity (Dann & Brunskill, 2015) would yield no dependence on the horizon.

For our work, the setting of Jiang and Agarwal immediately implies that

$$0 \le V_t^{\pi^*}(s) \le \mathcal{G} \le 1,\ \forall (s,t) \in \mathcal{S} \times [H]. \tag{12}$$

Further, since $V_t^{\pi^*}(s) \le 1$ and $r(s,a) \ge 0$ we must have $r(s,a) \in [0,1]$, which is the assumption of this work. Therefore our main result (theorem 1) applies here. Recalling that $T = KH$, we obtain an upper bound on regret as

$$\tilde{O}\left(\sqrt{SAK} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})\right). \tag{13}$$

Note that the planning/episodic horizon H does not appear in the dominant regret term which scales polynomially with the number of episodes[3] K, and only appears in transient lower order terms that are independent of K.

In other words, **up to logarithmic dependency and transient terms, we have an upper bound on episodic regret that is independent of the horizon** H. This result answers part of Jiang and Agarwal's open question: for their setting, the regret primarily scales independently of the horizon.

Surprisingly, while EULER uses a common problem-agnostic bound on the maximum possible optimal value function (H), it does not need to be provided with information about the domain-dependent maximum possible value function to attain the improved bound in the setting of the COLT conjecture of Jiang & Agarwal (2018).

It remains an open question whether we could further avoid either a dependence on the planning horizon in the transient terms as well as obtaining a PAC result. In Appendix B we further discuss this direction. However, these results are promising: they suggest that the hardness of learning in sparse reward, and long horizon episodic MDPs may not be fundamentally much harder than shorter horizon domains if the total reward is bounded.

6. Problem dependent bounds

We now focus on deriving regret bounds for selected MDP classes that are very common in RL. We emphasize that such setting-dependent guarantees are obtained with the same algorithm that is not informed of a particular MDP's values of $\mathbb{Q}^*$ and $\mathcal{G}$. Although the described settings share common features and are sometimes subclasses of one another, they are in separate subsections due to their important relation to the past literature and their practical relevance. Importantly, they are *all characterized by low* $\mathbb{Q}^*$.

6.1. Bounds using the range of optimal value function

To improve over the worst case bound in infinite horizon RL there have been approaches that aim at obtaining stronger problem dependent bounds if the value function does not vary much across different states of the MDP. If $\operatorname{rng} V^{\pi^*}$ is smaller than the worst-case (either H or D for the fixed horizon vs recurrent RL), the reduced variability in the expected return suggests that performance can benefit from constructing tighter confidence intervals. This is achieved by Bartlett & Tewari (2009) by providing this range to their algorithm REGAL, achieving a regret bound:

$$\tilde{O}(\Phi S\sqrt{AT}) \tag{14}$$

where $\Phi \ge \operatorname{rng} V^{\pi^*}$ is an overestimate of the optimal value function range and is an input to the algorithm described in that paper. This means that if domain knowledge is available and is supplied to the algorithm the regret can be substantially reduced. This line of research was followed in (Fruit et al., 2018) which derived a computationally-tractable variant of REGAL. However, they still require knowledge of a value function range upper bound $\Phi \ge \operatorname{rng} V^{\pi^*}$. Specifying a too high value for Φ would increase the regret and a too low value would cause the algorithm to fail.

Our analysis shows that, in the episodic setting, it is possible to achieve at least the same but potentially much better level of performance *without knowing the optimal value function range*. This follows as an easy corollary of our main regret upper bound (Theorem 1) after bounding the environmental norm, as we discuss below.

Let $\mathcal{S}_{s,a}$ be the set of immediate successor states after one transition from state s upon taking action a there, that is, the states in the support of $p(\cdot \mid s,a)$ and define

$$\Phi_{succ} \overset{def}{=} \max_{s,a} \operatorname*{rng}_{s^+ \in \mathcal{S}_{s,a}} V_{t+1}^{\pi^*}(s^+) \tag{15}$$

as the maximum value function range *when restricted to the immediate successor states*. Since the variance is upper bounded by (one fourth of) the square range of a random variable we have that:

$$\mathbb{Q}^* \overset{def}{=} \max_{s,a,t}(\operatorname{Var}(R(s,a) \mid (s,a)) + \operatorname*{Var}_{s^+ \sim p(s,a)} V_{t+1}^{\pi^*}(s^+))$$
$$\le \max_{s,a,t}\left(1 + (\operatorname*{rng}_{s^+ \in \mathcal{S}_{s,a}} V_{t+1}^{\pi^*}(s^+))^2\right) \le 1 + \Phi_{succ}^2.$$

This immediately yields:

Corollary 1.2 (Bounded Range of V^{π^*} Among Successor States)**.** *With probability at least $1-\delta$, the regret of* EULER *is bounded by:*

$$\tilde{O}(\Phi_{succ}\sqrt{SAT} + \sqrt{S}SAH^2(\sqrt{S} + \sqrt{H})). \tag{16}$$

[3] This is stronger than scaling polynomially with the time T

A few remarks are in order:

- EULER does not need to know the value of Φ_{succ} or of the environmental norm or of the value function range to attain the improved bound;
- Φ_{succ} can be much smaller than $\operatorname{rng} V^{\pi^*}$ because it is the range of V^{π^*} restricted to few successor states as opposed to across the whole domain, and therefore it is always smaller than Φ, in other words: $\Phi \geq \operatorname{rng} V^{\pi^*} \geq \Phi_{succ}$.
- (Bartlett & Tewari, 2009; Fruit et al., 2018) consider the more challenging infinite horizon setting, while our results holds for fixed horizon RL.

6.2. Bounds on the next-state variance of V^{π^*} and empirical benchmarks

The environmental norm also can empirically characterize the hardness of RL in single problem instances. This was one of the key contributions of the work that introduced the environmental norm (Maillard et al., 2014), which evaluated the environmental norm for a number of common RL benchmark simulation tasks including mountain car, pinball, taxi, bottleneck, inventory and red herring In these domains the environmental norm is correlated with the complexity of reinforcement learning in these environments, as evaluated empirically. Indeed, in these settings, the environmental norm is often much smaller then the maximum value function range, which can itself be much smaller than the worst-case bound D or H. Our new results provide solid theoretical justification for the observed empirical savings.

This measure of MDP complexity also intriguingly allows us to gain more insight on another important simulation domain, chain MDPs like that in Figure 1. Chain MDPS have been considered a canonical example of challenging hard-to-learn RL domains, since naive strategies like ϵ greedy can take exponential time to attain satisfactory performance. By setting for simplicity $N \overset{def}{=} S = H$ EULER provides an upper regret bound of $\tilde{O}(\sqrt{NAK} + \dots)$ that is substantially tighter than a worst case bound $\tilde{O}(\sqrt{N^3 AK} + \dots)$, at least for large K. This is intriguing because it suggests pathological MDPs may be even less common than expected. More details about this example are in appendix A.1.

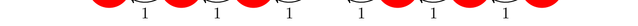

Figure 1. Classical hard-to-learn MDP

6.3. Stochasticity in the system dynamics

In this section we consider two important opposite classes of problems: deterministic MDPs and MDPs that are highly stochastic in that the successor state is sampled from a fixed distribution. These bounds are also a direct consequence of Theorem 1 and can be deduced from Corollary 1.2.

Deterministic domains Many problems of practical interest, for example in robotics, have low stochasticity, and this immediately yields low value for $\mathbb{Q}^*$. As a limit case, we consider domains with deterministic rewards and dynamics models. An agent designed to learn deterministic domains only needs to experience every transition *once* to reconstruct the model, which can take up to $O(SA)$ episodes with a regret at most $O(SAH)$(Wen & Van Roy, 2013).

Note in deterministic domains $\mathbb{Q}^* = 0$. Therefore if EULER is run on *any* deterministic MDP then the regret expression exhibits a $\log(T)$ dependence. This is a substantial improvement over prior RL regret bounds for problem-independent settings all have at least a $\sqrt{T}$ dependence. Further, a refined analysis (Appendix Section H.3) shows EULER is close to the lower bound except for a factor in the horizon and logarithmic terms:

Proposition 1. *If* EULER *is run on a deterministic MDP then the regret is bounded by* $\tilde{O}(SAH^2)$.

Highly mixing domains Recently, (Zanette & Brunskill, 2018) show that it is possible to design an algorithm that can switch between the MDP and the contextual bandit framework while retaining near-optimal performance in both without being informed of the setting. They consider mapping contextual bandit to an MDP whose transitions to different states (or contexts) are sampled from a fixed underlying distribution over which the agent has no control.

The Bandit-MDP considered in Zanette & Brunskill (2018) is an environment with high stochasticity (the MDP is highly mixing since every state can be reached with some probability in one step). Since the transition function is unaffected by the agent, an easy computation yields $\operatorname{rng} V_t^{\pi^*} \leq 1$, as replicated in Appendix A.2. A regret guarantee in the leading order term of order $\tilde{O}(\sqrt{SAT})$ for EULER which matches the established lower bound for tabular contextual bandits (Bubeck & Cesa-Bianchi, 2012) follows from corollary 1.2. This is useful since in many practical applications it is unclear in advance if the domain is best modeled as a bandit or a sequential RL problem. Our results improve over (Zanette & Brunskill, 2018) since EULER has better worst-case guarantees by a factor of $\sqrt{H}$. Our approach is also feasible with next-state distributions that have zero or near zero mass over some of the next states: in contrast to prior work, the inverse minimum visitation probability does not show up in our analysis.

7. Related Literature

In infinite horizon RL, prior empirical evaluation of $\mathbb{Q}^*$ in (Maillard et al., 2014) has shown encouraging performance in a number of common benchmarks that $\mathbb{Q}^*$ has small value and its size relates to the hardness of solving the RL task. The theoretical results provides a regret bound whose leading order term is $\tilde{O}\left(\frac{1}{p_0}DS\sqrt{\mathbb{Q}^* AT}\right)$ (where p_0 is the minimum (non-zero) transition probability), and generally does not improve over worst case analysis for the infinite horizon setting. Our algorithm operates in an easier setting (finite horizon) where it can improve over the worst case, but it is an open question whether an improvement is possible in infinite horizon.

Our connection with (Bartlett & Tewari, 2009; Fruit et al., 2018; Zanette & Brunskill, 2018) has already been described. Here we focus on the remaining literature. We again note that the infinite horizon setting offers a number of important complexities and comparisons to the finite horizon setting (as considered here) cannot be directly made; however, as some of the closest related work lies in the infinite horizon setting, we briefly discuss it here.

- Bounds that depend on gap between policies: In the infinite horizon setting, (Even-Dar et al., 2006) has bounds dependent on the minimum gap in the optimal state action values between the best and second best action, and UCRL2 (Jaksch et al., 2010) has bounds as function of the gap in the difference in the average reward between the best and second best policies. Such gaps reflect an interesting alternate structure in the problem domain: note that in prior work as these gaps become arbitrarily small, the bound approaches infinity: even in such settings, if the next state variance is small, our bound will stay bounded. An interesting future direction is to consider bounds that consider both forms of structure.
- Regret bounds with value function approximation: In finite horizon settings, (Osband & Roy, 2014) uses the Eluder dimension as a measure of the dimensionality of the problem and (Jiang et al., 2017) proposes the Bellman rank to measure the learning complexity. Such measures capture a different notion of hardness than ours and do not match the lower bound in tabular settings.
- Infinite horizon results with additional properties of the transition model: the most closely related work to ours is (Talebi & Maillard, 2018) who also develop tighter regret bounds as a function of the next-state variance, but for infinite horizon settings. Exploration in such settings is nontrivial and the authors leverage an important assumption of ergodicity (which has also been considered in (Auer & Ortner, 2006). Specifically the agent will visit every state regardless of the current policy, and the rate of this mixing appears in the regret bound. An interesting and nontrivial question is whether our results can be extended to this setting without additional assumptions on the mixing structure of the domain.

A natural additional question is whether prior algorithms also inherit strong problem dependent rounds. Indeed, recent work by (Dann & Brunskill, 2015; Azar et al., 2017) has also used the variance of the value function at the next state in their analysis, though their final results are expressed as worst case bounds. However, the actual bonus terms used in their algorithms are distinct from our bonus terms, perhaps most significantly in that we maintain and leverage point-wise upper and lower bounds on the value function. While it is certainly possible that their algorithms or others already attain some form of problem dependent performance, they have not been analyzed in a way that yields problem dependent bounds. This is a technical area, and performing such analyses is a non-trivial deviation from a worst-case analysis. For example, the current worst case bounds from Azar et al. (2017) yield a regret bound that scales as $\tilde{O}(\sqrt{HSAT} + \sqrt{H^2T} + S^2AH^2)$ and it is a non-trivial extension to analyze how each of these terms might change to reflect problem-dependent quantities. One of our key contributions is an analysis of the rate of convergence of the empirical quantities about properties of the underlying MDP to the real ones in determining the regret bound.

8. Future Work and Conclusion

In this paper we have proposed EULER, an algorithm for episodic finite MDPs that matches the best known worst-case regret guarantees while provably obtaining much tighter guarantees if the domain has a small variance of the value function over the next-state distribution $\mathbb{Q}^*$ or a small bound in the possible achievable reward. EULER does not need to know these MDP-specific quantities in advance. We show that $\mathbb{Q}^*$ is low for a number of important subclasses of MDPs, including: MDPs with sparse rewards, (near) determinisitic MDPs, highly mixing MDPs (such as those closer to bandits) and some classical empirical benchmarks. We also show how our result helps answer a recent open learning theory question about the necessary dependence of regret results on the episode horizon. Possible interesting directions for future work would be to examine problem-dependent bounds in the infinite horizon setting, incorporate a gap-dependent analysis, or see if such ideas could be extended to the continuous state setting.

Acknowledgments

The authors are grateful for the high quality feedback of the reviewers, and for the comments of Yonathan Efroni and his colleagues, including Mohammad Ghavamzadeh and Shie Mannor, who helped find a mistake in an earlier draft.

This work was partially supported by the Total Innovation Fellowship program, a NSF CAREER award and an ONR Young Investigator Award.

References

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

Azar, M., Munos, R., and Kappen, H. J. On the sample complexity of reinforcement learning with a generative model. In *ICML*, 2012.

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.

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.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *ICML*, 2019.

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.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. https://arxiv.org/abs/1802.04020, 2018.

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

Jiang, N. and Agarwal, A. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pp. 3395–3398, 2018.

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.

Kakade, S., Wang, M., and Yang, L. F. Variance reduction methods for sublinear reinforcement learning. *Arxiv*, 2018.

Lattimore, T. and Hutter, M. Near-optimal pac bounds for discounted mdps. In *Theoretical Computer Science*, 2014.

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

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. In *COLT*, 2009.

Osband, I. and Roy, B. V. Model-based reinforcement learning and the eluder dimension. In *NIPS*, 2014.

Osband, I. and Roy, B. V. Why is posterior sampling better than optimism for reinforcement learning? In *ICML*, 2017.

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.

Osband, I., Van Roy, B., and Russo, D. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, 2013.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Talebi, M. and Maillard, O. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *ALT*, 2018.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. Technical report, Hewlett-Packard Labs, 2003.

Wen, Z. and Van Roy, B. Efficient exploration and value function generalization in deterministic systems. In *NIPS*, 2013.

Zanette, A. and Brunskill, E. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *ICML*, 2018.