Minimax Regret Bounds for Reinforcement Learning

Mohammad Gheshlaghi Azar [1] Ian Osband [1] Rémi Munos [1]

Abstract

We consider the problem of provably optimal exploration in reinforcement learning for finite horizon MDPs. We show that an optimistic modification to value iteration achieves a regret bound of $\widetilde{O}(\sqrt{HSAT}+H^2S^2A+H\sqrt{T})$ where H is the time horizon, S the number of states, A the number of actions and T the number of time-steps. This result improves over the best previous known bound $\widetilde{O}(HS\sqrt{AT})$ achieved by the UCRL2 algorithm of Jaksch et al. (2010). The key significance of our new results is that when $T \geq H^3S^3A$ and $SA \geq H$, it leads to a regret of $\widetilde{O}(\sqrt{HSAT})$ that matches the established lower bound of $\Omega(\sqrt{HSAT})$ up to a logarithmic factor. Our analysis contains two key insights. We use careful application of concentration inequalities to the optimal value function as a whole, rather than to the transitions probabilities (to improve scaling in S), and we define Bernstein-based "exploration bonuses" that use the empirical variance of the estimated values at the next states (to improve scaling in H).

1. Introduction

We consider the reinforcement learning (RL) problem of an agent interacting with an environment in order to maximize its cumulative rewards through time (Burnetas & Katehakis, 1997; Sutton & Barto, 1998). We model the environment as a Markov decision process (MDP) whose transition dynamics are unknown from the agent. As the agent interacts with the environment it observes the states, actions and rewards generated by the system dynamics. This leads to a fundamental trade off: should the agent explore poorly-understood states and actions to gain information and improve future performance, or exploit its knowledge to optimize short-run rewards.

The most common approach to this learning problem is to separate the process of estimation and optimization. In this paradigm, point estimates of the unknown quantities are used in place of the unknown parameters and a plan is made with respect to these estimates. Naive optimization with respect to these point estimates can lead to premature exploitation and so may never learn the optimal policy. Dithering approaches to exploration (e.g., ϵ-greedy) address this failing through random action selection. However, as this exploration is not directed the resultant algorithms may take exponentially long to learn (Kearns & Singh, 2002). In order to learn efficiently it is necessary that the agent prioritizes potentially informative states and actions. To do this, it is important that the agent maintains some notion of its own uncertainty. In some sense, given any prior belief, the optimal solution to this exploration/exploitation dilemma is given by the dynamic programming in the extended Bayesian belief state (Bertsekas, 2007). However, the computational demands of this method become intractable for even small problems (Guez et al., 2013) while finite approximations can be arbitrarily poor (Munos, 2014).

To combat these failings, the majority of provably efficient learning algorithms employ a heuristic principle known as *optimism in the face of uncertainty* (OFU). In these algorithms, each state and action is afforded some “optimism” such that its imagined value is as high as statistically plausible. The agent then chooses a policy under this optimistic view of the world. This allows for efficient exploration since poorly-understood states and actions are afforded higher optimistic bonus. As the agent resolves its uncertainty, the effects of optimism will reduce and the agent’s policy will approach optimality. Almost all reinforcement learning algorithms with polynomial bounds on sample complexity employ optimism to guide exploration (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002; Strehl et al., 2006; Dann et al., 2017).

An alternative principle motivated by the Thompson sampling (Thompson, 1933) has emerged as a practical competitor to optimism. The algorithm *posterior sampling reinforcement learning* (PSRL) maintains a posterior distribution for MDPs and, at each episode of interaction, follows a policy which is optimal for a single random sample (Strens, 2000). Previous works have argue for the potential

[1] DeepMind, London, UK. Correspondence to: Mohammad Gheshlaghi Azar <mazar@google.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

benefits of such PSRL methods over existing optimistic approaches (Osband et al., 2013; Osband & Van Roy, 2016b) but they come with guarantees on the Bayesian regret only. However a very recent work (Agrawal & Jia, 2017) have shown that an optimistic version of posterior sampling (using a max over several samples) achieves a frequentist regret bound $\widetilde{O}(H\sqrt{SAT})$ (for large T) in the more general setting of weakly communicating MDPs.

In this paper we present a conceptually simple and computationally efficient approach to optimistic reinforcement learning in finite-horizon MDPs and report results for the frequentist regret. Our algorithm, *upper confidence bound value iteration* (UCBVI) is similar to *model-based interval estimation* (MBIE-EB) (Strehl & Littman, 2005) with a delicate alteration to the form of the "exploration bonus". In particular UCBVI replaces the universal scalar of the bonus in MBIE-EB with the empirical variance of the next-state value function of each state-action pair. This alteration is essential to improve the regret bound from $\widetilde{O}(H)$ to $\widetilde{O}(\sqrt{H})$.

Our key contribution is to establish a high probability regret bound $\widetilde{O}(\sqrt{HSAT}+H^2S^2A+H\sqrt{T})$ where S is the number of states, A is the number of actions, H is the episode length and T is the total number of time-steps (and where $\widetilde{O}$ ignores logarithmic factors). Importantly, for $T > H^3S^3A$ and $SA \geq H$ this bound is $\widetilde{O}(\sqrt{HSAT})$, which matches the established lower bound for this problem, up to logarithmic factors (Osband & Van Roy, 2016a).[1] This positive result is the first of its kind and helps to address an ongoing question about where the fundamental lower bounds lie for reinforcement learning in finite horizon MDPs (Bartlett & Tewari, 2009; Dann & Brunskill, 2015; Osband & Van Roy, 2016a). Our refined analysis contains two key ingredients:

- We use careful application of Bernstein and Freedman inequalities (Bernstein, 1927; Freedman, 1975) to the concentration of the *optimal value function* directly, rather than building confidence sets for the transitions probabilities and rewards, like in UCRL2 (Jaksch et al., 2010) and UCFH (Dann & Brunskill, 2015).
- We use empirical-variance exploration bonuses based on Bernstein's inequality, which together with a recursive Bellman-type Law of Total Variance (LTV) provide tight bounds on the expected sum of the variances of the value estimates, in a similar spirit to the analysis from (Azar et al., 2013; Lattimore & Hutter, 2012).

At a high level, this work addresses the noted shortcomings of existing RL algorithms (Bartlett & Tewari, 2009; Jaksch et al., 2010; Osband & Van Roy, 2016b), in terms of dependency on S and H. We demonstrates that it is possible to design a simple and computationally efficient optimistic algorithm that simultaneously address both the loose scaling in S and H to obtain the first regret bounds that match the $\Omega(\sqrt{HSAT})$ lower bounds as T becomes large.

We should be careful to mention the current limitations of our work, each of which may provide fruitful ground for future research. First, we study the setting of episodic, finite horizon MDPs and not the more general setting of weakly communicating systems (Bartlett & Tewari, 2009; Jaksch et al., 2010). Also we assume that the horizon length H is known to the learner. Further, our bounds only improve over previous scaling $\widetilde{O}(HS\sqrt{AT})$ for $T > H^3S^3A$.

We hope that this work will serve to elucidate several of the existing shortcomings of exploration in the tabular setting and help further the direction of research towards provably optimal exploration in reinforcement learning.

2. Problem formulation

In this section, we briefly review some notation, as well as some standard concepts and definitions from the theory of Markov decision processes (MDPs).

Markov Decision Problems We consider the problem of undiscounted episodic reinforcement learning (RL) (Bertsekas & Tsitsiklis, 1996), where an RL agent interacts with a stochastic environment and this interaction is modeled as a discrete-time MDP. An MDP is a quintuple $\langle \mathcal{S}, \mathcal{A}, P, R, H \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the set of states and actions, P is the state transition distribution, The function $R : \mathcal{S} \times \mathcal{A} \to \Re$ is a real-valued function on the state-action space and the horizon H is the length of episode. We denote by $P(\cdot|x,a)$ and $R(x,a)$ the probability distribution over the next state and the immediate reward of taking action a at state x, respectively. The agent interacts with the environment in a sequence of episodes. The interaction between the agent and environment at every episode[2] $k \in [K]$ is as follows: starting from $x_{k,1} \in \mathcal{S}$ which is chosen by the environment, the agent interacts with the environment for H steps by following a sequence of actions chosen in $\mathcal{A}$ and observes a sequence of next-states and rewards until the end of episode. The initial state $x_{k,1}$ may change arbitrarily from one episode to the next. We also use the notation $\|\cdot\|_1$ for the ℓ_1 norm throughout this paper.

Assumption 1 (MDP Regularity)**.** *We assume $\mathcal{S}$ and $\mathcal{A}$ are finite sets with cardinalities S, A, respectively. We also assume that the immediate reward $R(x,a)$ is deterministic*

[1] In fact the lower bound of (Jaksch et al., 2010) is for the more general setting of the weakly communicating MDPs and it doesn't directly apply to our setting. But a similar approach can be used to prove a lower bound of same order for the finite-horizon MDPs, as it is already used in (Osband & Van Roy, 2016a).

[2] We write $[n]$ for $\{i \in \mathbb{N} \mid 1 \leq i \leq n\}$.

and belongs to the interval $[0,1]$.[3]

In this paper we focus on the setting where the reward function R is known, but extending our algorithm to unknown stochastic rewards poses no real difficulty.

The policy during an episode is expressed as a mapping $\pi : \mathcal{S} \times [H] \to \mathcal{A}$. The *value* $V_h^\pi : \mathcal{S} \to \mathbb{R}$ denotes the value function at every step $h = 1, 2, \ldots, H$ and state $x \in \mathcal{S}$ such that $V_h^\pi(x)$ corresponds to the expected sum of $H - h$ rewards received under policy π, starting from $x_h = x \in \mathcal{S}$. Under Assumption 1 there exists always a policy π^* which attains the best possible values, and we define the *optimal value function* $V_h^*(x) \overset{\text{def}}{=} \sup_\pi V_h^\pi(x)$ for all $x \in \mathcal{S}$ and $h \geq 1$. The policy π at every step h defines the state transition kernel P_h^π and the reward function r_h^π as $P_h^\pi(y|x) \overset{\text{def}}{=} P(y|x, \pi(x,h))$ and $r_h^\pi(x) \overset{\text{def}}{=} R(x, \pi(x,h))$ for all $x \in \mathcal{S}$. For every $V : \mathcal{S} \to \mathbb{R}$ the right-linear operators $P\cdot$ and $P_h^\pi\cdot$ are also defined as $(PV)(x,a) \overset{\text{def}}{=} \sum_{y \in \mathcal{S}} P(y|x,a)V(y)$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ and $(P_h^\pi V)(x) \overset{\text{def}}{=} \sum_{y\in\mathcal{S}} P_h^\pi(y|s)V(y)$ for all $x \in \mathcal{S}$, respectively. The Bellman operator for the policy π, at every step $h > 0$ and $x \in \mathcal{S}$, is defined as $(\mathcal{T}_h^\pi V)(x) \overset{\text{def}}{=} r_h^\pi(x) + (P_h^\pi V)(x)$. We also define the state-action Bellman operator for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ as $(\mathcal{T}V)(x,a) \overset{\text{def}}{=} R(x,a) + (PV)(x,a)$ and the optimality Bellman operator for all $x \in \mathcal{S}$ as $(\mathcal{T}^*V)(x) \overset{\text{def}}{=} \max_{a\in\mathcal{A}} (\mathcal{T}V)(x,a)$. For ease of exposition, we remove the dependence on x and (x,a), e.g., writing PV for $(PV)(x,a)$ and V for $V(x)$, when there is no possible confusion.

We measure the performance of the learner over $T = KH$ steps[4] by the regret $\text{Regret}(K)$, defined as

$$\text{Regret}(K) \overset{\text{def}}{=} \sum_{k=1}^{K} V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1}),$$

where π_k is the control policy followed by the learner at episode k. Thus the regret measures the expected loss of following the policy produced by the learner instead of the optimal policy. So the goal of learner is to follow a sequence of policies $\pi_1, \pi_2, \ldots, \pi_K$ such that $\text{Regret}(K)$ is as small as possible.

3. Upper confidence bound value iteration

In this section we introduce two variants of the algorithm that we investigate in this paper. We call the algorithm *upper confidence bound value iteration* (UCBVI). UCBVI is an extension of value iteration which guarantees that the resultant value function is a (high-probability) upper confidence bound (UCB) on the optimal value function. This algorithm is related to the *model based interval estimation* (MBIE-EB) algorithm (Strehl & Littman, 2008). Our key contribution is the precise design of the upper confidence sets, and the analysis which lead to tight regret bounds.

`UCBVI`, described in Algorithm 1, calls `UCB-Q-values` (Algorithm 2) which returns UCBs on the Q-values computed by value iteration using an empirical Bellman operator to which is added a confidence bonus `bonus`. We consider two variants of UCBVI depending on the structure of `bonus`, which we present in Algorithms 3 and 4.

Algorithm 1 `UCBVI`

```
Initialize data $\mathcal{H} = \emptyset$
for episode $k = 1, 2, \ldots, K$ do
    $Q_{k,h} = \texttt{UCB} - \texttt{Q} - \texttt{values}(\mathcal{H})$
    for step $h = 1, \ldots, H$ do
        Take action $a_{k,h} = \arg\max_a Q_{k,h}(x_{k,h}, a)$
        Update $\mathcal{H} = \mathcal{H} \cup (x_{k,h}, a_{k,h}, x_{k,h+1})$
    end for
end for
```

Algorithm 2 `UCB-Q-values`

```
Require: Bonus algorithm bonus, Data $\mathcal{H}$
  Compute, for all $(x,a,y) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$,
  $N_k(x,a,y) = \sum_{(x',a',y')\in\mathcal{H}} \mathbb{I}(x'=x, a'=a, y'=y)$
  $N_k(x,a) = \sum_{y\in\mathcal{S}} N_k(x,a,y)$
  $N'_{k,h}(x,a) = \sum_{(x_{i,h},a_{i,h},x_{i,h+1})\in\mathcal{H}} \mathbb{I}(x_{i,h}=x, a_{i,h}=a)$
  Let $\mathcal{K} = \{(x,a)\in\mathcal{S}\times\mathcal{A}, N_k(x,a) > 0\}$
  Estimate $\widehat{P}_k(y|x,a) = \frac{N_k(x,a,y)}{N_k(x,a)}$ for all $(x,a)\in\mathcal{K}$
  Initialize $V_{k,H+1}(x) = 0$ for all $(x,a)\in\mathcal{S}\times\mathcal{A}$
  for $h = H, H-1, \ldots, 1$ do
    for $(x,a)\in\mathcal{S}\times\mathcal{A}$ do
      if $(x,a)\in\mathcal{K}$ then
        $b_{k,h}(x,a) = \texttt{bonus}(\widehat{P}_k, V_{k,h+1}, N_k, N'_{k,h})$
        $Q_{k,h}(x,a) = \min\big(Q_{k-1,h}(x,a), H,$
            $R(x,a) + (\widehat{P}_k V_{k,h+1})(x,a) + b_{k,h}(x,a)\big)$
      else
        $Q_{k,h}(x,a) = H$
      end if
      $V_{k,h}(x) = \max_{a\in\mathcal{A}} Q_{k,h}(x,a)$
    end for
  end for
  return Q-values $Q_{k,h}$
```

The first of these `UCBVI-CH` is based upon Chernoff-Hoeffding's concentration inequality, considers `UCBVI` with `bonus` = `bonus_1`. `bonus_1` is a very simple bound which only assumes that values are bounded in $[0, H]$. We

[3] For rewards in $[R_{\min}, R_{\max}]$ simply rescale these bounds.

[4] In this paper we will often substitute $T = KH$ to highlight various dependencies relative to the existing literature. This equivalence should be kept in mind by the reader.

will see in Theorem 1 that this very simple algorithm can already achieve a regret bound of $\widetilde{O}(H\sqrt{SAT})$, thus improving the best previously known regret bounds from a S to a $\sqrt{S}$ dependence. The intuition for this improved S-dependence is that our algorithm (as well as our analysis) does not consider confidence sets on the transition dynamics $P(y|x,a)$ like UCRL2 and UCFH do, but instead directly maintains confidence intervals on the optimal value function. This is crucial as, for any given (x,a), the transition dynamics are S-dimensional whereas the Q-value function is one-dimensional.

Algorithm 3 `bonus_1`

```
Require: $\widehat{P}_k(x,a), N_k(x,a)$
  $b(x,a) = 7HL\sqrt{\frac{1}{N_k(x,a)}}$ where $L = \ln(5SAT/\delta)$,
  return $b$
```

However, the loose form of UCB given by `UCBVI-CH` does not look at the value function of the next state, and just consider it as being bounded in $[0,H]$. However, much better bounds can be obtained by looking at the variance of the next state values. Our main result relies upon `UCBVI` with `bonus` = `bonus_2`, which we refer to as `UCBVI-BF` as it relies on Bernstein-Freedman's concentration inequalities to build the confidence set. `UCBVI-BF` builds upon the intuition for `UCBVI-CH` but also incorporates a variance-dependent exploration bonus. This leads to tighter exploration bonuses and an improved regret bound of $\widetilde{O}(\sqrt{HSAT})$.

Algorithm 4 `bonus_2`

```
Require: $\widehat{P}_k(x,a), V_{k,h+1}, N_k, N'_{k,h}$
  $b(x,a) = \sqrt{\frac{8L\mathrm{Var}_{Y\sim\widehat{P}_k(\cdot|x,a)}(V_{k,h+1}(Y))}{N_k(x,a)}} + \frac{14HL}{3N_k(x,a)} + \sqrt{\frac{8\sum_y \widehat{P}_k(y|x,a)\left[\min\left(\frac{100^2H^3S^2AL^2}{N'_{k,h+1}(y)}, H^2\right)\right]}{N_k(x,a)}}$
  where $L = \ln(5SAT/\delta)$
  return $b$
```

Compared to `UCBVI-BF` here we use a bonus built from the empirical variance of the estimated next values. The idea is that if we had knowledge of the optimal value V^*, we could build tight confidence bounds using the variance of the optimal value function at the next state in place of the loose bound of H. Since however V^* is unknown, here we use as a surrogate the empirical variance of the estimated values. As more data is gathered, this variance estimate will converge to the variance of V^*. Now we need to make sure our estimates $V_{k,h}$ are optimistic (i.e., that they upper bound V^*_h) at all times. This is achieved by adding an additional bonus (last term in $b(x,a)$), which guarantees that we upper bound the variance of V^*. Now, using an iterative -Bellman-type- Law of Total Variance, we have (see proof) that the sum of the next-state variances of V^* (over H time steps) (which is related to the sum of the exploration bonuses over H steps) is bounded by the variance of the H-steps return. Thus the size of the bonuses built by `UCBVI-BF` are constrained over the H steps. And we prove that the sum of those bonuses do not grow linearly in H but in $\sqrt{H}$ only. This is the key for our improved dependence from H to $\sqrt{H}$.

4. Main results

In this section we present the main results of the paper, which are upper bounds on the regret of `UCBVI-CH` and `UCBVI-BF` algorithms. We assume Assumption 1 holds.

Theorem 1 (Regret bound for `UCBVI-CH`)**.** *Consider a parameter $\delta > 0$. Then the regret of `UCBVI-CH` is bounded w.p. at least $1-\delta$, by*

$$\mathrm{Regret}(K) \le 20H^{3/2}L\sqrt{SAK} + 250H^2S^2AL^2,$$

where $L = \ln(5HSAT/\delta)$.

For $T \ge HS^3A$ and $SA \ge H$ this bound translates to a regret bound of $\widetilde{O}(H\sqrt{SAT})$, where $T = KH$ is the total number of time-steps at the end of episode K.

Theorem 1 is significant in that, for large T, it improves the regret dependence from S to $\sqrt{S}$, compared to the best known bound of (Jaksch et al., 2010). The main intuition for this improved S-dependence is that we bound the estimation error of the next-state value function directly, instead of the transition probabilities. More precisely, instead of bounding the estimation error $(\widehat{P}^{\pi_k}_k - P^{\pi_k})V_{k,h+1}$ by $\|\widehat{P}^{\pi_k}_k - P^{\pi_k}\|_1\|V_{k,h+1}\|_\infty$ (as is done in (Jaksch et al., 2010) for example), we bound $(\widehat{P}^{\pi_k}_k - P^{\pi_k})V^*_{h+1}$ instead (for which a bound with no dependence on S can be achieved since V^* is deterministic) and handle carefully the correction term $(\widehat{P}^{\pi_k}_k - P^{\pi_k})(V_{k,h+1} - V^*_{h+1})$.

Our second result, Theorem 2, demonstrates that we can improve upon the H-dependence by using a more refined, Bernstein-Friedman-type, exploration bonus.

Theorem 2 (Regret bound for `UCBVI-BF`)**.** *Consider a parameter $\delta > 0$. Then the regret of `UCBVI-BF` is bounded w.p. $1-\delta$, by*

$$\begin{aligned}\mathrm{Regret}(K) \quad \le \quad & 30HL\sqrt{SAK} + 2500H^2S^2AL^2 \\ & +4H^{3/2}\sqrt{KL},\end{aligned}$$

where $L = \ln(5HSAT/\delta)$.

We note that for $T \geq H^3S^3A$ and $SA \geq H$ this bound translates to a regret bound of $\widetilde{O}(\sqrt{HSAT})$. This result is particularly significant since, for T large enough (i.e., $T \geq H^3S^3A$), our bound is $\widetilde{O}(\sqrt{HSAT})$ which matches the established lower bound $\Omega(\sqrt{HSAT})$ of (Jaksch et al., 2010; Osband & Van Roy, 2016a) up to logarithmic factors.

The key insight is to apply concentration inequalities to bound the estimation errors and the exploration bonuses in terms of the variance of V^* at the next state. We then use the fact that the sum of these variances is bounded by the variance of the return (see e.g., Munos & Moore, 1999; Azar et al., 2013; Lattimore & Hutter, 2012), which shows that the estimation errors accumulate as $\sqrt{H}$ instead of linearly in H, thus implying the improved H-dependence.

Computational efficiency

Theorems 1 and 2 guarantee the statistical efficiency of UCBVI. In addition, both `UCBVI-CH` and `UCBVI-BF` are computationally tractable. Each episode both algorithms perform an optimistic value iteration with computational cost of the same order as solving a known MDP. In fact, the computational cost of these algorithms can be further reduced by only selectively recomputing UCBVI after sufficiently many observations. This technique is common to the literature (Jaksch et al., 2010; Dann & Brunskill, 2015) and would not affect the $\widetilde{O}$ statistical efficiency. The computational cost of this variant of UCBVI then amounts to $\widetilde{O}(SA\min(SA,T)\min(T,S))$ as it only needs to update the model $\widetilde{O}(SA)$ times (Jaksch et al., 2010).

Weakly communicating MDPs

In this short paper we focus on the setting of finite horizon MDPs. By comparison, previous optimistic approaches to exploration, such as UCRL2, provide bounds for the more general setting of weakly communicating MDPs (Jaksch et al., 2010; Bartlett & Tewari, 2009). However, we believe that much of the insight from the UCBVI algorithm (and its analysis) will carry over to this more general setting using existing techniques such as ‘the doubling trick‘ (Jaksch et al., 2010).

5. Proof sketch

Here we provide the sketch proof of our results. The full proof is deferred to the appendix.

5.1. Sketch Proof of Theorem 1

Let $\Omega = \{V_{k,h} \geq V_h^*, \forall k, h\}$ be the event under which all computed $V_{k,h}$ values are upper bounds on the optimal value function. Using backward induction on h (and standard concentration inequalities) one can prove that Ω holds with high probability (see Lem. 18 in the appendix). To simplify notations in this sketch of proof we will not make the numerical constants explicit, and instead we will denote by $\Box$ a numerical constant which can vary from line to line. The exact values of these constants are provided in the full proof. We will also make use of simplified notations, such as using L to represent the logarithmic term $L = \ln(\Box HSAT/\delta)$.

The cumulative regret at episode K is $\mathrm{Regret}(K) \overset{\text{def}}{=} \sum_{1\leq k\leq K} V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1})$. Define $\widetilde{\mathrm{Regret}}(K) \overset{\text{def}}{=} \sum_{1\leq k\leq K} V_{k,1}(x_{k,1}) - V_1^{\pi_k}(x_{k,1})$. Under Ω we have $\mathrm{Regret}(K) \leq \widetilde{\mathrm{Regret}}(K)$, so we now bound $\widetilde{\mathrm{Regret}}(K)$. Define $\Delta_{k,h} \overset{\text{def}}{=} V_h^* - V_h^{\pi_k}$ and $\widetilde{\Delta}_{k,h} \overset{\text{def}}{=} V_{k,h} - V_h^{\pi_k}$. Thus

$$\begin{aligned}\Delta_{k,h} &\leq \widetilde{\Delta}_{k,h} = \widehat{P}_k^{\pi_k} V_{k,h+1} + b_{k,h} - P^{\pi_k} V_{h+1}^{\pi_k} \\ &= (\widehat{P}_k^{\pi_k} - P^{\pi_k})V_{k,h+1} + P^{\pi_k}\widetilde{\Delta}_{k,h+1} + b_{k,h}.\end{aligned}$$

The difficulty in bounding $(\widehat{P}_k^{\pi_k} - P^{\pi_k})V_{k,h+1}$ is that both $V_{k,h+1}$ and $\widehat{P}_k^{\pi_k}$ are random variables and are not independent (the value function $V_{k,h+1}$ computed at $h+1$ may depend on the samples collected from state $x_{h,k}$), thus a straightforward application of Chernoff-Hoeffding (CH) inequality does not work here. In (Jaksch et al., 2010), this issue is addressed by bounding it by $\|\widehat{P}_k^{\pi_k} - P^{\pi_k}\|_1 \|V_{k,h+1}\|_\infty$ at the price of an additional $\sqrt{S}$.

The main contribution of our $\widetilde{O}(H\sqrt{SAT})$ bound (which removes a $\sqrt{S}$ factor compared to the previous bound of (Jaksch et al., 2010)) is to handle this term more properly. Instead of directly bounding $(\widehat{P}_k^{\pi_k} - P^{\pi_k})V_{k,h+1}$, we bound $(\widehat{P}_k^{\pi_k} - P^{\pi_k})V_{h+1}^*$, using straightforward application of CH (which removes the $\sqrt{S}$ factor since V_{h+1}^* is deterministic), and deal with the correction term $(\widehat{P}_k^{\pi_k} - P^{\pi_k})(V_{k,h+1} - V_{h+1}^*)$. We have

$$\begin{aligned}\widetilde{\Delta}_{k,h} &= (\widehat{P}_k^{\pi_k} - P^{\pi_k})(V_{k,h+1} - V_{h+1}^*) \\ &\quad + P^{\pi_k}\widetilde{\Delta}_{k,h+1} + b_{k,h} + e_{k,h},\end{aligned}$$

where $e_{k,h} \overset{\text{def}}{=} (\widehat{P}_k^{\pi_k} - P^{\pi_k})V_{h+1}^*(x_{k,h})$ is the estimation error of the optimal value function at the next state. Defining $\widetilde{\delta}_{k,h} \overset{\text{def}}{=} \widetilde{\Delta}_{k,h}(x_{k,h})$, we have

$$\begin{aligned}\widetilde{\delta}_{k,h} &\leq (\widehat{P}_k^{\pi_k} - P^{\pi_k})\Delta_{k,h+1}(x_{k,h}) + \widetilde{\delta}_{k,h+1} \\ &\quad + \epsilon_{k,h} + b_{k,h} + e_{k,h},\end{aligned}$$

where $\epsilon_{k,h} \overset{\text{def}}{=} P^{\pi_k}\Delta_{k,h+1}(x_{k,h}) - \Delta_{k,h+1}(x_{k,h+1})$.

Step 1: bound on the correction term $(\widehat{P}_k^{\pi_k} - P^{\pi_k})\Delta_{k,h+1}(x_{k,h})$. Using Bernstein's inequality (B), this term is bounded by

$$\sum_y P^{\pi_k}(y|x_{k,h})\sqrt{\frac{\Box L}{P^{\pi_k}(y|x_{k,h})n_{k,h}}}\Delta_{k,h+1}(y) + \frac{\Box SHL}{n_{k,h}},$$

where $n_{k,h} \overset{\text{def}}{=} N_k(x_{k,h}, \pi_k(x_{k,h}))$. Now considering only the y such that $P^{\pi_k}(y|x_{k,h})n_{k,h} \geq \square H^2 L$, and since $0 \leq \Delta_{k,h+1} \leq \widetilde{\Delta}_{k,h+1}$, then $(\widehat{P}_k^{\pi_k} - P^{\pi_k})\Delta_{k,h+1}(x_{k,h})$ is bounded by

$$\begin{aligned} &\bar{\epsilon}_{k,h} + \sqrt{\frac{\square L}{P^{\pi_k}(x_{k,h+1}|x_{k,h})n_{k,h}}}\widetilde{\delta}_{k,h+1} + \frac{\square SHL}{n_{k,h}} \\ \leq \quad &\bar{\epsilon}_{k,h} + \frac{1}{H}\widetilde{\delta}_{k,h+1} + \frac{\square SHL}{n_{k,h}}, \end{aligned}$$

where $\bar{\epsilon}_{k,h} \overset{\text{def}}{=} \sqrt{\frac{\square L}{n_{k,h}}}\Big(\sum_y P^{\pi_k}(y|x_{k,h})\frac{\widetilde{\Delta}_{k,h+1}(y)}{\sqrt{P^{\pi_k}(y|x_{k,h})}} - \frac{\widetilde{\delta}_{k,h+1}}{\sqrt{P^{\pi_k}(x_{k,h+1}|x_{k,h})}}\Big)$.

The sum over the neglected y such that $P^{\pi_k}(y|x_{k,h})n_{k,h} < \square H^2 L$ contributes to an additional term

$$\sum_y \sqrt{\frac{\square P^{\pi_k}(y|x_{k,h})n_{k,h}L}{n_{k,h}^2}}\Delta_{k,h+1}(y) \quad \leq \quad \frac{\square SH^2 L}{n_{k,h}}.$$

Neglecting this term (and the smaller order term $\square SHL/n_{k,h}$) for now (by the pigeon-hole principle we can prove that these terms contribute to the final regret by a constant at most $\square S^2 AH^2 L^2$), we have

$$\begin{aligned} \widetilde{\delta}_{k,h} &\leq \Big(1+\frac{1}{H}\Big)\widetilde{\delta}_{k,h+1} + \epsilon_{k,h} + \bar{\epsilon}_{k,h} + b_{k,h} + e_{k,h} \\ &\leq \underbrace{\Big(1+\frac{1}{H}\Big)^H}_{\leq e}\sum_{i=h}^{H-1}\Big(\epsilon_{k,i} + \bar{\epsilon}_{k,i} + b_{k,i} + e_{k,i}\Big). \end{aligned}$$

The regret is thus bounded by

$$\widetilde{\text{Regret}}(K) \leq \square \sum_{k,h}(\epsilon_{k,h} + \bar{\epsilon}_{k,h} + b_{k,h} + e_{k,h}). \qquad (1)$$

We now bound those 4 terms. It is easy to check that $\sum_{k,h}\epsilon_{k,h}$ and $\sum_{k,h}\bar{\epsilon}_{k,h}$ are sums of martingale differences, which are bounded using Azuma's inequality, and lead to a regret of $\widetilde{O}(H\sqrt{T})$ without dependence on the size of state and action space. The leading terms in the regret bound comes from the sum of the exploration bonuses $\sum_{k,h} b_{k,h}$ and the estimation errors $\sum_{k,h} e_{k,h}$.

Step 2: Bounding the martingales $\sum_{k,h}\epsilon_{k,h}$ and $\sum_{k,h}\bar{\epsilon}_{k,h}$. Using Azuma's inequality we deduce

$$\sum_{k,h}\epsilon_{k,h} \overset{(Az)}{\leq} \square H\sqrt{TL}, \qquad \sum_{k,h}\bar{\epsilon}_{k,h} \overset{(Az)}{\leq} \sqrt{\square TL}. \qquad (2)$$

Step 3: Bounding the exploration bonuses $\sum_{k,h} b_{k,h}$: Using the pigeon-hole principle, we have

$$\begin{aligned} \sum_{k,h} b_{k,h} &= \square HL\sum_{k,h}\sqrt{\frac{1}{n_{k,h}}} \\ &= \square HL\sum_{x,a}\sum_{n=1}^{N_K(x,a)}\sqrt{\frac{1}{n}} \\ &\leq \square HL\sqrt{SAT}. \end{aligned} \qquad (3)$$

Step 4: Bounding on the estimation errors $\sum_{k,h} e_{k,h}$. Using CH, w.h.p. we have $e_{k,h} = (\widehat{P}_k^{\pi_k} - P^{\pi_k})V_{h+1}^* \overset{(CH)}{\leq} \square H\sqrt{\frac{L}{n_{k,h}}}$. Thus this bound on the estimation errors are of the same order as the exploration bonuses (which is the reason we choose those bonuses...).

Putting everything together: Plugging (2) and (3) into (1) (and adding the smaller order term) we deduce

$$\text{Regret}(K) \leq \widetilde{\text{Regret}}(K) \leq \square\big(H^{\frac{3}{2}}L\sqrt{SAK} + H^2 S^2 AL^2\big).$$

5.2. Sketch Proof of Theorem 2

The proof of Theorem 1 relied on proving by a straightforward induction over h that $\Omega = \{V_{k,h} \geq V_h^*, \forall k,h\}$ hold with high probability. In the case of exploration bonuses defined by:

$$\begin{aligned} b_{k,h}(x,a) = &\underbrace{\sqrt{\frac{\square L\mathbb{V}_{Y\sim\widehat{P}_k^{\pi_k}(\cdot|x,a)}\big(V_{k,h+1}(Y)\big)}{N_k(x,a)}} + \frac{\square HL}{N_k(x,a)}}_{\text{empirical Bernstein}} \\ &+ \underbrace{\sqrt{\frac{\min\Big(\square H^3 S^2 AL^2\sum_y\frac{\widehat{P}_k(y|x,a)}{N'_{k,h+1}(y)}, H^2\Big)}{N_k(x,a)}}}_{\text{additional bonus}}, \end{aligned} \qquad (4)$$

the backward induction over h is not straightforward. Indeed, if the $V_{k,h+1}$ are upper bounds on V_{h+1}^*, it is not necessarily the case that the empirical variance of $V_{k,h+1}$ are upper bound on the empirical variance of V_{h+1}^*. However we can prove by (backward) induction over h that $V_{k,h+1}$ is sufficiently close to V_{h+1}^* to guarantee that the variance of those terms are sufficiently close to each other so that the additional bonus (additional bonus in (4)) will make sure that $V_{k,h}$ is still an upper-bound on V_h^*. More precisely, define the set of indices:

$$\begin{aligned} [k,h]_{hist} \quad \overset{\text{def}}{=} \quad &\{(i,j), s.t. (1 \leq i \leq k \wedge 1 \leq j \leq H) \\ &\vee(i = k \wedge h < j \leq H)\}, \end{aligned}$$

and the event $\Omega_{k,h} \overset{\text{def}}{=} \{V_{i,j} \ge V_h^*, (i,j) \in [k,h]_{hist}\}$. Our induction is the following:

- Assume that $\Omega_{k,h}$ holds. Then we prove that $(V_{k,h+1} - V^*_{h+1})(y) \le \Box H\sqrt{\frac{SAL}{N'_{k,h+1}(y)}}$.
- We deduce that $\mathbb{V}_{Y\sim\widehat{P}_k(\cdot|x,a)}\big(V_{k,h+1}(Y)\big) + \Box H^3S^2AL^2\sum_y \frac{\widehat{P}_k(y|x,a)}{N'_{k,h+1}(y)} \ge \mathbb{V}_{Y\sim\widehat{P}_k(\cdot|x,a)}\big(V^*_{h+1}(Y)\big)$, so the additional bonus compensates for the possible variance difference. Thus $V_{k,h} \ge V_h^*$ and $\Omega_{k,h-1}$ holds.

So in order to prove that all values computed by the algorithm are upper bounding V^*, we just need to prove that under $\Omega_{k,h}$, we have $(V_{k,h+1} - V^*_{h+1})(y) \le \min(\Box H^{1.5}SL\sqrt{\frac{A}{N'_{k,h+1}(y)}}, H)$, which is obtained by deriving the following regret bound on

$$\begin{aligned}\widetilde{R}_{k,h}(y) &\overset{\text{def}}{=} \sum_{i\le k}(V_{i,h+1} - V^{\pi_i}_{h+1})(x_{i,h+1})\mathbb{I}\{x_{i,h+1} = y\}\\ &\le \Box HL\sqrt{SAN'_{k,h+1}(y)}. \qquad (5)\end{aligned}$$

Indeed, since $\{V_{i,h}\}_i$ is a decreasing sequence in i, we have

$$\begin{aligned}(V_{k,h+1} - V^*_{h+1})(y) &\le \widetilde{R}_{k,h+1}(y)/N'_{k,h+1}(y)\\ &\le \Box HL\sqrt{SA/N'_{k,h+1}(y)}.\end{aligned}$$

Once we have proven that w.h.p., all computed values are upper bounds on V^* (i.e. event Ω), then we prove that under Ω, the following regret bound holds:

$$\text{Regret}(K) \le \widetilde{\text{Regret}}(K) \le \Box(HL\sqrt{SAK} + H^2S^2AL^2). \qquad (6)$$

The proof of (5) relies on the same derivations as those used for proving (6). The only two differences being that (i) HK is replaced by $N'_{k,h+1}(y)$, the number of times a state y was reached at time $h+1$, up to episode k, and (ii) the additional $\sqrt{H}$ factor which comes from the fact that at any episode, $N'_{k,h+1}(y)$ can only tick once, whereas the total number of transitions from y during any episode can be as large as H. The full proof of (5) will be given in details in the appendix. We now give a proof sketch of (6) under Ω.

Similar steps used for proving Theorem 1 apply. The main difference compared to Theorem 1 is the bound on the sum of the exploration bonuses and the estimation errors (which we consider in Steps 3' and 4' below). This is where we can remove the $\sqrt{H}$ factor. The use of the Bernstein inequality makes it possible to bound both of those terms in terms of the expected sum of variances (under the current policy π_k at any episode k) of the next-state values (for that policy), and then using recursively the Law of Total Variance to conclude that this quantity is nothing but the variance of the returns. This step is detailed now. For simplicity of the exposition of this sketch we neglect second order terms.

Step 3': Bounding the sum of exploration bonuses $b_{k,h}$. We have

$$\sum_{k,h} b_{k,h} = \Box\sqrt{L}\underbrace{\sum_{k,h}\sqrt{\frac{\mathbb{V}_{Y\sim\widehat{P}^{\pi_k}_h(\cdot|x_{k,h})}\big(V_{k,h+1}(Y)\big)}{n_{k,h}}}}_{\text{main term}}$$
$$+\underbrace{\sqrt{\frac{\min\Big(\Box H^3S^2AL^2\sum_y\frac{\widehat{P}_k(y|x,a)}{N'_{k,h+1}(y)}, H^2\Big)}{N_k(x,a)}} + \sum_{k,h}\frac{\Box L}{N_k(x,a)}}_{\text{second order term}}.$$

By Cauchy-Schwarz, the main term is bounded by $\Big(\sum_{k,h}\widehat{\mathbb{V}}_{k,h+1}\sum_{k,h}\frac{1}{n_{k,h}}\Big)^{1/2}$, where $\widehat{\mathbb{V}}_{k,h+1} \overset{\text{def}}{=} \mathbb{V}_{Y\sim\widehat{P}^{\pi_k}_h(\cdot|x_{k,h})}\big(V_{k,h+1}(Y)\big)$. Since $\sum_{k,h}\frac{1}{n_{k,h}} \le \Box SA\ln(T)$ by the pigeon-hole principle, we now focus on the term $\sum_{k,h}\widehat{\mathbb{V}}_{k,h+1}$.

We now prove that $\widehat{\mathbb{V}}_{k,h+1}$ is close to $\mathbb{V}^{\pi_k}_{k,h+1} \overset{\text{def}}{=} \mathbb{V}_{Y\sim P^{\pi_k}_h(\cdot|x_{k,h})}\big(V^{\pi_k}_{h+1}(Y)\big)$ by bounding the following quantity:

$$\begin{aligned}&\widehat{\mathbb{V}}_{k,h+1} - \mathbb{V}^{\pi_k}_{k,h+1}\\ =\ &\widehat{P}^{\pi_k}V^2_{k,h+1} - (\widehat{P}^{\pi_k}V_{k,h+1})^2\\ &-P^{\pi_k}(V^{\pi_k}_{h+1})^2 + (P^{\pi_k}V^{\pi_k}_{h+1})^2\\ \overset{(i)}{\le}\ &\widehat{P}^{\pi_k}V^2_{k,h+1} - P^{\pi_k}(V^{\pi_k}_{h+1})^2 + 2H(P^{\pi_k} - \widehat{P}^{\pi_k})V^*_{h+1}\\ \overset{(ii)}{\le}\ &\underbrace{(\widehat{P}^{\pi_k} - P^{\pi_k})V^2_{k,h+1}}_{(a_{k,h})} + \underbrace{P^{\pi_k}(V^2_{k,h+1} - (V^{\pi_k}_{h+1})^2)}_{(a'_{k,h})}\\ &+\Box H^2\sqrt{\frac{L}{n_{k,h}}}, \qquad (7)\end{aligned}$$

where (i) holds since under $\Omega_{k,h}$, $V_{k,h} \ge V_h^* \ge V_h^{\pi_k}$ and (ii) holds due to Chernoff Hoeffding.

Step 3'-a: bounding $\sum_{k,h}\widehat{\mathbb{V}}_{k,h+1} - \mathbb{V}^{\pi_k}_{k,h+1}$. Using similar argument as those used in (Jaksch et al., 2010), we have that

$$a_{k,h} \le H^2\|\widehat{P}^{\pi_k} - P^{\pi_k}\|_1 \le \Box H^2\sqrt{SL/n_{k,h}},$$

(where $n_{k,h} \overset{\text{def}}{=} N_k(x_{k,h}, \pi_k(x_{k,h}))$). Thus from the pigeon-hole principle, $\sum_{k,h} a_{k,h} \le \Box H^2S\sqrt{ATL}$.

Now $a'_{k,h}$ is bounded as

$$\begin{aligned} a'_{k,h} &\leq 2HP^{\pi_k}(V_{k,h} - V_h^{\pi_k}) \\ &= 2HP^{\pi_k}\widehat{\Delta}_{k,h}. \end{aligned}$$

Thus using Azuma's inequality,

$$\begin{aligned} \sum_{k,h} a'_{k,h} &\overset{(Az)}{\leq} 2H\sum_{k,h}\widehat{\delta}_{k,h+1} + \Box H^2\sqrt{TL} \\ &\leq 2H^2U + \Box H^2\sqrt{TL}, \end{aligned}$$

where U is defined as an upper-bound on the pseudo regret: $U \overset{\text{def}}{=} \sum_{k,h}(b_{k,h} + e_{k,h}) + \Box H\sqrt{T}$ (an upper bound on the r.h.s. of (1)).

Step 3'-b: bounding $\sum_{k,h}\mathbb{V}^{\pi_k}_{k,h+1}$. (Dominant term)

For any episode k, $\mathbb{E}[\sum_h \mathbb{V}^{\pi_k}_{k,h+1}|\mathcal{H}_k]$ is the expected sum of variances of the value function $V_k^\pi(y)$ at the next state $y \sim P^{\pi_k}(\cdot|x_{k,h})$ under the true transition model for the current policy. A recursive application of the law of total variance (see e.g., Munos & Moore, 1999; Azar et al., 2013; Lattimore & Hutter, 2012) shows that this quantity is nothing else than the variance of the return (sum of H rewards) under policy π_k: $\mathbb{V}\big(\sum_h r(x_{k,h}, \pi_k(x_{k,h}))\big)$, which is thus bounded by H^2. Finally, using Freedman's (Fr) inequality to bound $\sum_{k,h}\mathbb{V}^{\pi_k}_{k,h+1}$ by its expectation (see the exact derivation in the appendix), we deduce

$$\begin{aligned} \sum_{k,h}\mathbb{V}^{\pi_k}_{k,h+1} &\overset{(Fr)}{\leq} \sum_k \mathbb{E}\Big[\sum_h \mathbb{V}^{\pi_k}_{k,h+1}|\mathcal{H}_k\Big] + \Box H^2\sqrt{TL} \\ &\leq TH + \Box H^2\sqrt{TL}. \end{aligned} \tag{8}$$

Thus, using (8), (7) and the bounds on $\sum a_{k,h}$ and $\sum a'_{k,h}$, we deduce that

$$\sum_{k,h} b_{k,h} \leq \Box L\sqrt{(TH + H^2U)SA}.$$

Step 4': Bounding the sum of estimation errors $\sum_{k,h} e_{k,h}$. We now use Bernstein inequality to bound the estimation errors

$$\begin{aligned} \sum_{k,h} e_{k,h} &= \sum_{k,h}(\widehat{P}_k^{\pi_k} - P^{\pi_k})V^*_{h+1}(x_{k,h}) \\ &\leq \sum_{k,h}\Box\sqrt{\frac{\mathbb{V}^*_{k,h+1}}{n_{k,h}}} + \Box\frac{HL}{n_{k,h}}, \end{aligned}$$

where $\mathbb{V}^*_{k,h+1} \overset{\text{def}}{=} \mathbb{V}_{Y\sim P^{\pi_k}(\cdot|x_{k,h})}\big(V^*_{h+1}(Y)\big)$. Now, in a very similar way as in Step 3' above, we relate $\mathbb{V}^*_{k,h+1}$ to $\mathbb{V}^{\pi_k}_{k,h+1}$ and use the Law of total variance to bound $\sum_{k,h}\mathbb{V}^{\pi_k}_{k,h+1}$ by HT and deduce that

$$\sum_{k,h} e_{k,h} \leq \Box L\sqrt{(TH + H^2U)SA}.$$

From (1) we see that $U \leq \Box L\sqrt{(TH + H^2U)SA}$ thus $U \leq \Box(L\sqrt{HSAT} + H^2SAL^2)$. This implies (6).

So the reason we are able to remove the $\sqrt{H}$ factor from the regret bound comes from the fact that the sum, over H steps, of the variances of the next state values (which define the amplitude of the confidence intervals) is at most bounded by the variance of the return. Intuitively this means that the size of the confidence intervals do not add up linearly over H steps but grows as $\sqrt{H}$ only. Although the sequence of estimation errors are not independent over time, we are able to demonstrate a concentration of measure phenomenon that shows that those estimation errors concentrate as if they were independent.

6. Conclusion

In this paper we refine the familiar concept of optimism in the face of uncertainty. Our key contribution is the design and analysis of the algorithm UCBVI-BF , which addresses two key shortcomings in existing algorithms for optimistic exploration in finite MDPs. First we apply a concentration to the value as a whole, rather than the transition estimates, this leads to a reduction from S to $\sqrt{S}$. Next we apply a recursive law of total variance to couple estimates across an episode, rather than at each time step individually, this leads to a reduction from H to $\sqrt{H}$.

Theorem 2 provides the first regret bounds which, for sufficiently large T, match the lower bounds for the problem $\widetilde{O}(\sqrt{HSAT})$ up to logarithmic factors. It remains an open problem whether we can match the lower bound using this approach for small T. We believe that the higher order term can be improved from $\widetilde{O}(H^2S^2A)$ to $\widetilde{O}(HS^2A)$ by a more careful analysis, i.e., a more extensive use of Freedman-Bernstein inequalities. The same applies to the term of order $H\sqrt{T}$ which can be improved to $\sqrt{HT}$.

These results are particularly significant because they help to estabilish the information-theoretic lower bound of reinforcement learning at $\Omega(\sqrt{HSAT})$ (Osband & Van Roy, 2016a), whereas it was suggested in some previous work that lower-bound should be of $\Omega(H\sqrt{SAT})$. Moving from this big-picture insight to an analytically rigorous bound is non-trivial. Although we push many of the technical details to the appendix, our paper also makes several contributions in terms of analytical tools that may be useful in subsequent work. In particular we believe that the way we construct the exploration bonus and confidence intervals in UCBVI-CH is novel to the literature of RL. Also the constructive approach in the proof of UCBVI-CH, which bootstraps the regret bounds to prove that $V_{k,h}$s are ucbs, is another analytical contribution of this paper.

Acknowledgements

The authors would like to thank Marc Bellemare and all the other wonderful colleagues at DeepMind for many hours of discussion and insight leading to this research. We are also grateful for the anonymous reviewers for their helpful comments and for fixing several mistakes in an earlier version of this paper.

References

Agrawal, Shipra and Jia, Randy. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.

Azar, Mohammad Gheshlaghi, Munos, Rémi, and Kappen, Hilbert J. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

Bartlett, Peter L. and Tewari, Ambuj. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, pp. 35–42, June 2009.

Bernstein, S. Theory of probability, 1927.

Bertsekas, D. P. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, Belmount, Massachusetts, third edition, 2007.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

Brafman, Ronen I. and Tennenholtz, Moshe. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Bubeck, Sébastien and Cesa-Bianchi, Nicolò. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR*, abs/1204.5721, 2012. URL http://arxiv.org/abs/1204.5721.

Bubeck, Sébastien, Munos, Rémi, Stoltz, Gilles, and Szepesvári, Csaba. X-armed bandits. *Journal of Machine Learning Research*, 12:1587–1627, 2011.

Burnetas, Apostolos N and Katehakis, Michael N. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

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

Dann, Christoph and Brunskill, Emma. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 2015.

Dann, Christoph, Lattimore, Tor, and Brunskill, Emma. Ubev-a more practical algorithm for episodic rl with near-optimal pac and regret guarantees. *arXiv preprint arXiv:1703.07710*, 2017.

Freedman, David A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Guez, Arthur, Silver, David, and Dayan, Peter. Scalable and efficient bayes-adaptive reinforcement learning based on monte-carlo tree search. *Journal of Artificial Intelligence Research*, pp. 841–883, 2013.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Kearns, Michael J. and Singh, Satinder P. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

Lattimore, Tor and Hutter, Marcus. PAC bounds for discounted MDPs. *CoRR*, abs/1202.3890, 2012.

Maurer, Andreas and Pontil, Massimiliano. Empirical bernstein bounds and sample variance penalization. *stat*, 1050:21, 2009.

Munos, R. and Moore, A. Influence and variance of a Markov chain : Application to adaptive discretizations in optimal control. In *Proceedings of the 38th IEEE Conference on Decision and Control*, 1999.

Munos, Rémi. From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

Osband, Ian and Van Roy, Benjamin. On lower bounds for regret in reinforcement learning. *stat*, 1050:9, 2016a.

Osband, Ian and Van Roy, Benjamin. Why is posterior sampling better than optimism for reinforcement learning. *arXiv preprint arXiv:1607.00215*, 2016b.

Osband, Ian, Russo, Dan, and Van Roy, Benjamin. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

Strehl, Alexander L and Littman, Michael L. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, pp. 856–863. ACM, 2005.

Strehl, Alexander L and Littman, Michael L. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74 (8):1309–1331, 2008.

Strehl, Alexander L., Li, Lihong, Wiewiora, Eric, Langford, John, and Littman, Michael L. PAC model-free reinforcement learning. In *ICML*, pp. 881–888, 2006.

Strens, Malcolm J. A. A Bayesian framework for reinforcement learning. In *ICML*, pp. 943–950, 2000.

Sutton, Richard and Barto, Andrew. *Reinforcement Learning: An Introduction*. MIT Press, March 1998.

Thompson, W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Weissman, Tsachy, Ordentlich, Erik, Seroussi, Gadiel, Verdu, Sergio, and Weinberger, Marcelo J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.