UCRL with post-decision state: regret analysis

Yijun Wan

1 Introduction

The idea of upper confidence reinforcement learning (UCRL) algorithm was first introduced by Auer and Ortner (2007), which at each step chooses an optimal policy on an extended Markov decision process (MDP) with confidence bounds on rewards and transition probabilities of the true MDP. The regret analysis was greatly improved by Jaksch et al. (2010) for commuting MDPs with $S = |\mathcal{S}|$ states, $A = |\mathcal{A}|$ actions and finite diameter D. In this article, we restrict ourselves to MDPs with a set of post-decision state $\mathcal{Y}$ ($Y = |\mathcal{Y}|$ is in general much smaller than $S \times A$) such that the transition probabilities can be decomposed according to

$$p(s, a, s') = p(f(s, a), s'), \quad s, s' \in \mathcal{S}, a \in \mathcal{A}.$$

Here $f : \mathcal{S} \times \mathcal{A} \to \mathcal{Y}$ is a problem-dependent (known) deterministic transition function from state-action pairs to post-decision states (hence $\mathcal{S}$ is called the pre-decision state set) and $p : \mathcal{Y} \times \mathcal{S} \to [0, 1]$ is an appropriate probability kernel.

Heuristically, the computational benefits of post-decision state assumption comes from replacing the augmented state-action by post-decision state when calculating the expectation (with respect to the transition kernel). For some situations, $\mathcal{Y}$ has the same cardinality as $\mathcal{S}$. But in many settings, $\mathcal{Y}$ is dramatically simpler that $\mathcal{S}$, hence simplifying the value function approximation.

Without loss of generality, the reward is set to be within $[0, 1]$. Nevertheless, the results scale linearly with the rewards. If the reward is taken from $[r_{min}, r_{\max}]$, the rewards can be normalized to $[0, 1]$ and the regret bounds hold with additional factor $(r_{\max} - r_{\min})$.

1.1 Examples

Here are some typical examples where post-decision states can be introduced.

- *Invertory problem*: let R_t be the number of items in stock and D_t be the number of customer demands available to be moved at time t. The state is $S_t = (R_t, D_t)$. Once the store makes a decision, all the unserved demands in D_t are lost. After making decision on the number of items A_t added to the storage , the post-decision state is $R_t + A_t$.
- *Backgammon*: the state S_t at each step t is the number of pieces on the board. The transition from S_t to S_{t+1} depends on the player's decision a_t, the play of the opposing player and the next roll of the dice. The post-decision state variable is simply the state of the board after a player moves but before his opponent has moved.

1.2 Main results and contributions

The goal of this report is to give a self-contained regret and complexity analysis of UCRL2 with the presence of post-decision set, based on the framework of UCRL2 Jaksch et al. (2010) where the optimi bound Hoeffding inequality)

Theorem 1. *With probability* $1-3\delta$ *it holds that for any initial state* s *and for any* $T>1$*, the regret of* UCRL2 *with post-decision state is bounded by*

$$\Delta(M,s,T) \leq 25D\sqrt{SYT\log(2YT/\delta)} + 15\sqrt{SAT\log(SAT/\delta)}$$

In Theorem 1, the dominant term is $\Omega(D\sqrt{SYT})$ with a logarithmic correction in terms of T, which could be much smaller than the bound $\Omega(DS\sqrt{AT})$ given by Jaksch et al. (2010) if the post-decision set $\mathcal{Y}$ is much smaller than $\mathcal{S}\times\mathcal{A}$. Another improvement is the computational complexity by a factor of Y/S of *extended value iteration* in Section 2.2 using modified Bellman's equation for post-decision values.

1.3 Preliminaries

Definition 2. *Given an MDP* M *and a stationary policy* $\pi:\mathcal{Y}\to\mathcal{A}$*, consider the Markov process for which the transition probability is*

$$p(y,y') = \sum_{f(s,\pi(y))=y'} p_A(y,s) \text{ for all } y,y'\in\mathcal{Y}.$$

Define $T(y'|M,\pi,y)$ *to be the first time when hitting* y'' *for the above Markov process starting from* y*. Then the* diameter *of* M *is defined as*

$$D(M) := \max_{y\neq y'\in\mathcal{Y}} \min_{\pi:\mathcal{Y}\to\mathcal{A}} \mathbb{E}[T(y'|M,\pi,y)],$$

MDPs with finite diameter are called communicating.

long-term average reward and bias function:

Theorem 3. *If* $\mathcal{M}$ *contains at least one communicating MDP, then the extended value iteration in Section 2.2 converges.*

2 UCRL2 algorithm with post-decision state

In this section, we propose a variant of UCRL2 that uses optimism in the face of uncertainty, taking into consideration the post-decision state. Similarly, the algorithm proceeds in episodes. For each episode, first the confidence bounds are constructed based on empirical means and Hoeffding's inequality. And then an optimal policy $\tilde{\pi}_k$ can be found using *extended value iteration*. In the end we execute the policy $\tilde{\pi}_k$ to explore the true MDP M. Note that the lengths of episodes depends on the observation to have a square root growth rate of time.

2.1 Construct confidence bounds

At the beginning of each episode k, the empirical average of rewards and transitions are calculated:

$$\hat{p}_k(s|y) := \frac{1}{N_k^+(y)} \sum_{t=1}^{t_k-1} \mathbf{1}_{\{s_t^a=t, s_{t+1=s}\}} \text{ and } \hat{r}_k(s,a) := \frac{1}{N_k^+(s,a)} \sum_{t=0}^{t_k-1} \mathbf{1}_{(s_t,a_t)=(s,a)},$$

where $N_k(y)$ is the number of visits of post-decision state $y \in \mathcal{Y}$ before episode k, $N_k(s,a)$ is the number of visits of state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $N_k^+(y) = \max\{1, N_k(y)\}$, $N_k^+(s,a) = \max\{1, N_k(s,a)\}$. A set of plausible MDPs are defined around the empirical estimates as

$$\mathcal{M}_k = \{\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{Y}, \tilde{r}, \tilde{\pi} \rangle : \forall y \in \mathcal{Y},\ \|\tilde{p}(\cdot|y) - \hat{p}_k(\cdot|y)\|_1 \leq \beta_k^p(y), \sum_{s'} \tilde{p}(s'|y) = 1,$$
$$\forall s \in \mathcal{S}, a \in \mathcal{A},\ \tilde{p}(s|y) \in [0,1],\ \tilde{r}(s,a) \in [0,1],\ |\tilde{r}(s,a) - \hat{r}(s,a)| \leq \beta_k^r(s,a)\}.$$

If we set

$$\beta_k^p(y) := \sqrt{\frac{2S + 4\log(2Yt_k/\delta)}{\max\{1, N_k(y)\}}} \text{ and } \beta_k^r(s,a) := \sqrt{\frac{\log(2SAt_k/\delta)}{\max\{1, N_k(s,a)\}}},$$

then at each episode k, with high-probability the true MDP stays in $\mathcal{M}_k$.

Lemma 4. *For any $t \geq 1$, the probability that the true MDP M is not contained in the set of plausible MDPs $\mathcal{M}(t)$ is at most δ, which means that*

$$\mathbb{P}[M \notin \mathcal{M}(t),\ \forall t \geq 1] \leq \delta.$$

Proof. Let us first consider the probability that the confidence interval given in an episode k for the transition probabilities fails. Recall that we set $\beta_k^p(y) := \sqrt{\frac{2S+4\log(2Yt_k/\delta)}{\max\{1,N_k(y)\}}}$. During each episode k, by L^1-deviation of the true distribution and the empirical distribution over S states from $N_k(y)$ samples is bounded by

$$\begin{aligned}
\mathbb{P}[\|\hat{p}(\cdot|y) - p(\cdot|y)\|_1 \geq \beta_k^p(y)] &\leq (2^S - 2)\exp(-\frac{N_k(y)(\beta_k^p(y))^2}{2}) \\
&\leq 2^S \exp(-S - 2\log(2Yt_k/\delta)) \\
&\leq \exp(-2\log(2Yt_k/\delta)) \\
&\leq \frac{\delta}{4t_k^2 Y}
\end{aligned}$$

And in episode k, the same argument using Hoeffding's inequality as Jaksch et al. (2010) gives that

$$\mathbb{P}[|\hat{r}(s,a) - \overline{r}(s,a)| \geq \beta_k^r(s,a)] \leq \frac{\delta}{4t_k^2 SA}$$

The bounds follows by summing over all episodes before t and over all $\mathcal{A},\ \mathcal{S},\ \mathcal{Y}$. □

Input: Confidence parameter $\delta \in]0,1[$, $\mathcal{S}$, $\mathcal{A}$
Initialization: Set $t := 1$, observe the initial state s_1 and choose arbitrarily an action a_1. And for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, set $\widehat{r}_1(s,a) = 0$ and $N_1(s,a) = 0$. For any $(y,s) \in \mathcal{Y} \times \mathcal{S}$, set $\widehat{p}_1(s|y) = 0$.
For episodes $k = 1, 2, ...$ **do**

1. Set $t_k \leftarrow t$ and initialize for episode k the state-action counters $\nu_k(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and post-decision counter $\nu_k(y)$.
2. Let $\mathcal{M}_k$ be the extended MDPs with transition probabilities $\widetilde{p}(\cdot|y)$ close to $\hat{p}_k(\cdot|y)$ and rewards $\widetilde{r}(s,a) \in [0,1]$ close to $\hat{r}_k(s,a)$, that is
$$\|\tilde{p}(\cdot|y) - \hat{p}_k(\cdot|y)\|_1 \leq \beta_k^p(y)$$
and
$$|\tilde{r}(s,a) - \hat{r}(s,a)| \leq \beta_k^r(s,a).$$
Use extended value iteration on the extended MDP to compute (an approximation of) the optimal policy $\widetilde{\pi}_k$ and the optimistic MDP $\widetilde{M}_k$.
3. Sample action $a_t \sim \pi_k(\cdot|s_t)$.
4. **While** $\nu_k(s_t,a_t) \leq \max\{1, N_k(s_t,a_t)\}$ **do**
   (a) Execute a_t, obtain reward r_t, and observe next state s_{t+1}.
   (b) Update $\nu_k(s_t,a_t) \leftarrow \nu_k(s_t,a_t) + 1$
   (c) Sample action $a_{t+1} \sim \pi_k(\cdot|s_{t+1})$ and set $t \leftarrow t+1$.
5. Update statistics $(\widehat{p}_{k+1}, \widehat{r}_{k+1})$.
6. Set $N_{k+1}(s,a) \leftarrow N_k(s,a) + \nu_k(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Figure 1: UCRL2B algorithm.

2.2 Find optimal policy for extended MDPs: extended value iteration

In the second step of each episode, we need to find a near-optimal policy $\tilde{\pi}_k$ for an approximate optimistic MDP $\tilde{M}_k$ in the set of plausible MDPs for the optimization problem

$$\rho_k^* := \sup_{M' \in \mathcal{M}_k} \max_{\pi} g_{M'}^{\pi} = \sup_{M' \in \mathcal{M}_k} g_{M'}^*.$$

Here the post-decision set $\mathcal{Y}$ brings advantage for complexity: the cardinality of value set is reduced from $|\mathcal{S}|$ to $|\mathcal{Y}|$, therefore $\boldsymbol{u}_{i+1}$ can be computed from $\boldsymbol{u}_i$ in at most $O(SAY)$ steps (S for finding the optimal MDP and S for finding the optimal value).

3 Regret analysis

The regret after T steps is defined as $\Delta(M,s,T) = \sum_{t=1}^{T}(\rho^* - r_t)$. To adopt the post-decision set in our regret analysis, let us denote the average reward of a post-decision state y with respect to policy π by

$$\overline{r}^{\pi}(y) := \mathbb{E}_y[r(s,\pi(s))] = \sum_s p(s|y)\overline{r}(s,\pi(s)).$$

Input: The empirical estimate $\hat{p}_k(\cdot|y)$ before episode k for each post-decision state $y \in \mathcal{Y}$ and $\hat{r}_k(s,a)$ for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$; optimal bounds $\beta_k^p(y)$ for each $y \in \mathcal{Y}$ and $\beta_k^r(s,a)$ for each $(s,a) \in \mathcal{S} \times \mathcal{Y}$.
Initialization: $n = 0$, $u_0(y) = 0$ for all $y \in \mathcal{Y}$.
While $\max_{y\in\mathcal{Y}}\{u_{i+1}(y) - u_i(y)\} - \min_{y\in\mathcal{Y}}\{u_{i+1}(y) - u_i(y)\} < 1/\sqrt{t_k}$ **do**

1. Increment $i \leftarrow i + 1$
2. For each $a \in \mathcal{A}$ and $y \in \mathcal{Y}$, find

$$\arg\max_{\substack{|p(\cdot)-\hat{p}(\cdot|y)|\le\beta_k^p(y) \\ \sum_{s\in\mathcal{S}} p(s)=1}} \sum_{s\in\mathcal{S}} p(s)\cdot\left(\hat{r}_k(s,a) + \beta_k^r(s,a) + u_i(f(s,a))\right) :$$

   (a) Sort $\mathcal{S} = \{s_1', s_2' \cdots, s_n'\}$ in descending order according to their values

$$\tilde{r}(\cdot,a) + u_i(f(\cdot,a))$$

   (b) Set $p(s_1') = \min\{1, \hat{p}(s_1'|y) + \beta_k^p(y)/2\}$ and

$$p(s_j') = \hat{p}(s_j'|y) \text{ for all states } s_j' \text{ with } j > 1.$$

   (c) Set $l = n$.

   (d) **While** $\sum_{s_j'\in\mathcal{S}} p(s_j') > 1$ **do**

      i. Reset $p(s_l') = \max\{0, 1 - \sum_{s_j'\neq s_l'} p(s_j')\}$.

      ii. Set $l \leftarrow l - 1$.

3. Find the optimal value

$$u_{i+1}(y) = \max_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} p(s)\cdot\left(\hat{r}_k(s,a) + \beta_k^r(s,a) + u_i(f(s,a))\right)$$

Return gain u_i and policy $\tilde{\pi}_k(s) = \arg\max_{a\in\mathcal{A}}\{\hat{r}_k(s,a) + \beta_k^r(s,a) + u_i(f(s,a))\}$.

Figure 2: Extended Value Iteration.

First we split the total reward into the sum of rewards in each episode. Write $\overline{r}_k(y) := \overline{r}^{\tilde{\pi}_k}(y)$ and denote the pseudo-regret accumulated in episode k by

$$\Delta_k := \sum_y \nu_k(y)(\rho^* - \overline{r}_k(y)),$$

where $\nu_k(y) = \sum_{(s,a)\in f^{-1}(y)} \nu_k(s,a)$.

Lemma 5. *With probability $1-\delta$, the total regret can be bounded the sum of Δ_k with a square root error of T,*

$$\Delta(M,s,T) \le \sum_{k=1}^{k_T} \Delta_k + \sqrt{T\ln(2T/\delta)} + 1 + k_T,$$

where $k_T = \sup\{k \ge 1 : t_k \le T\}$ is the number of episodes started up to step T.

Proof. Based on the observation up to time T,

$$\Delta(M,s,T)=\sum_{t=1}^{T}(\rho^*-r_t)\leq 1+\sum_{t=1}^{T-1}(\rho^*-\overline{r}^{\pi_{t+1}}(y_t))+\sum_{t=1}^{T-1}(\overline{r}^{\pi_{t+1}}(y_t)-r_{t+1})$$

By Hoeffding's inequality,

$$\mathbb{P}[\sum_{t=1}^{T-1}(\overline{r}^{\pi_t}(y_t)-r_{t+1})\leq\sqrt{T\ln(2T/\delta)}]\geq 1-\exp(-2\ln(2T/\delta))\leq 1-\frac{\delta}{2T^2}.$$

Summing over T, we have that with high probability $1-\delta$,

$$\begin{aligned}\Delta(M,s,T)&\leq 1+k_T+\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-2}(\rho^*-\overline{r}^{\pi_{t+1}}(y_t))+\sqrt{T\log(2T/\delta)}\\&\leq 1+k_T+\sum_{k=1}^{k_T}\Delta_k+\sqrt{T\log(2T/\delta)}\end{aligned}$$

□

During each episode, the convergence rate of reward functions is encoded in $\rho^*-\overline{r}_k(y)$. Then we will use approximative Bellman's equation to write the first term in terms of optimal transition matrix. Denote by $\boldsymbol{u}_i=(u_i(\cdot))_s$ the value function obtained after i iterations of Algorithm 2 in episode k for the set of all plausible MDPs (we drop the notation k for convenience), then for undiscounted reward function, denote by $\tilde{\rho}_k$ the maximal reward of all possible choice of the extended MDP (expected reward of $\tilde{\pi}_k$ chosen in this iteration on the optimistic MDP $\tilde{M}_k$), when the convergence criterion

$$\max_{s\in\mathcal{S}}\{u_{i+1}(s)-u_i(s)\}-\min_{s\in\mathcal{S}}\{u_{i+1}(s)-u_i(s)\}<\frac{1}{\sqrt{t_k}}$$

holds, the error bounds when terminating the EVI algorithm are

$$|\rho^*-\tilde{\rho}_k|\leq\frac{1}{\sqrt{t_k}}\text{ and }|u_{i+1}(y)-u_i(y)-\tilde{\rho}_k|\leq\frac{1}{\sqrt{t_k}}\text{ for all }y\in\mathcal{Y}.$$

Then one can replace ρ^* by $u_{i+1}(y_t)-u_i(y_t)$ in each term of the sum Δ_k by introducing a $2/\sqrt{t_k}$ error

$$\Delta_k\leq\sum_s\nu_k(y)(u_{i+1}(y)-u_i(y)-\overline{r}_k(y))+2\sum_y\frac{\nu_k(y)}{\sqrt{t_k}}.\tag{1}$$

Besides, the iteration relation gives

$$u_{i+1}(y)\leq\sum_{s\in\mathcal{S}}\tilde{p}_k(s|y)\cdot(\tilde{r}_k(s,\tilde{\pi}_k(s))+u_i(f(s,\tilde{\pi}_k(s)))).$$

Since the $\tilde{p}_k(\cdot|y)$ sum to 1, we can substract from $u_k(\cdot)$ a constant such that the module is smaller. In the following we will replace $u_i(\cdot)$ by $w_k(\cdot)$ where

$w_k(y) := u_i(y) - (\min_{y'} u_i(y') + \max_s u_i(y'))/2$. For simplicity, we write $w_k(s) := w_k(f(s, \tilde{\pi}_k(s)))$ for all $s \in \mathcal{S}$. Combined with (1), this leads to

$$\Delta_k \leq \Delta_k^p + \Delta_k^r + 2 \sum_y \nu_k(y)/\sqrt{y}, \tag{2}$$

where

$$\Delta_k^p := \sum_y \nu_k(y) \left(\sum_s \tilde{p}_k(s|y) w_k(s) - w_k(y) \right),$$

$$\Delta_k^r := \sum_y \nu_k(y) \left(\sum_{s \in \mathcal{S}} \tilde{p}_k(s|y) \tilde{r}_k(s, \tilde{\pi}_k(s)) - \overline{r}_k(y) \right).$$

In the confidence regium that $M \in \mathcal{M}_k$, the total regret comes from the difference of the optimal transition kernel and reward between the true transition kernel and reward functions. Thus we can estimate Δ_k^p and Δ_k^r individually. First,

$$\begin{aligned}
\Delta_k^r &= \sum_y \nu_k(y) \left(\sum_{s \in \mathcal{S}} \tilde{p}_k(s|y) \tilde{r}_k(s, \tilde{\pi}_k(s)) - \overline{r}_k(y) \right) \\
&\leq \sum_y \nu_k(y) \left(\sum_{s \in \mathcal{S}} \big(\tilde{p}_k(s|y) - p(s|y) \big) \tilde{r}_k(s, \tilde{\pi}_k(s)) + p(s|y)(\tilde{r}_k - \overline{r})(s, \tilde{\pi}_k(s)) \right) \\
&\leq \sum_y \nu_k(y) \big(\|\tilde{p}_k(\cdot|y) - p(\cdot|y)\|_{L^1} + \sum_{t=t_k}^{t_{k+1}-1} \sum_s p(s|y_t)(\tilde{r}_k - \overline{r})(s, \tilde{\pi}_k(s)) \right)
\end{aligned} \tag{3}$$

And we want to estimate the term Δ_k^p, which is dominant because of the product term of the transition kernel error and the span of the value function. Note that if $M \in \mathcal{M}_k$, for all $y \in \mathcal{Y}$,

$$\begin{aligned}
\Delta_k^p &= \sum_y \nu_k(y) \left(\sum_s \tilde{p}_k(s|y) w_k(s) - w_k(y) \right) \\
&= \sum_y \nu_k(y) \left(\sum_s (\tilde{p}_k(s|y) - p(s|y)) w_k(s) + \sum_s p(s|y) w_k(s) - w_k(y) \right) \\
&\leq \sum_y \nu_k(y) \left(\|w_k\|_\infty \cdot \|\tilde{p}_k(\cdot|y) - p(\cdot|y)\|_{L^1} + \sum_s p(s|y) w_k(s) - w_k(y) \right) \\
&= \|w_k\|_\infty \sum_y \nu_k(y) \|\tilde{p}_k(\cdot|y) - p(\cdot|y)\|_{L^1} + w_k(y_{t_{k+1}}) - w_k(y_{t_k}) \\
&\quad + \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_s p(s|y_t) w_k(f(s, \tilde{\pi}_k(s))) - w_k(y_{t+1}) \right)
\end{aligned} \tag{4}$$

Combining (2), (3) and (4) and note that

$$\|\tilde{p}_k(s'|y) - p(s'|y)\|_{L^1} \leq \|\tilde{p}_k(s'|y) - \hat{p}_k(s'|y)\|_{L^1} + \|\hat{p}_k(s'|y) - p(s', y)\|_{L^1} \leq 2\beta_k^p(s'|y),$$

$$|\tilde{r}_k(s,a) - \overline{r}(s,a)| \leq |\tilde{r}_k(s,a) - \hat{r}_k(s,a)| + |\hat{r}_k(s,a) - \overline{r}(s,a)| \leq 2\beta_k^r(s,a)$$

$$\|w_k\|_\infty \le \max_y |w_k(y)| \le D/2,$$

we have that

$$\Delta_k \le (D+2)\sum_y \nu_k(y)\beta_k^p(y) + 2\sum_{t=t_k}^{t_{k+1}-1}\sum_s p(s|y_t)\beta_k^r(s,\tilde\pi_k(s)) + \sum_{t=t_k}^{t_{k+1}-1}\left(\sum_s p(s|y_t)w_k(s,\tilde\pi_k(s)) - w_k(y_{t+1})\right) + D + 2\sum_y \frac{\nu_k(y)}{\sqrt{t_k}} \quad (5)$$

We can upper-bound the second and the third expectation term using Azuma's inequality of concentration for martingale difference sequence. For the martingale difference sequence $(\sum_s p(s|y_t)\beta^r_{k(t)}(s,\tilde\pi_{k(t)}(s)) - \beta^r_{k(t)}(s_{t+1},\tilde\pi_{k(t)}(s_{t+1})))_{t\ge 1}$ where $k(t)$ is such that $t_{k(t)} \le t \le t_{k(t)+1}-1$,

$$\mathbb{P}\left[\sum_{t=1}^T\sum_s p(s|y_t)\beta^r_{k(t)}(s,\tilde\pi_{k(t)}(s)) \ge \sum_{t=1}^T \beta^r_{k(t)}(s_{t+1},\tilde\pi_{k(t)}(s_{t+1})) + 2\sqrt{T\log(2T/\delta)}\right] \le \frac{\delta}{2T^2}.$$

For the martingale difference sequence $(\sum_s p(s|y_t)w_{k(t)}(f(s,\tilde\pi_k(s))) - w_{k(t)}(y_{t+1}))$ for any $T \ge 1$,

$$\mathbb{P}\left[\sum_{t=1}^T\left(\sum_s p(s|y_t)w_{k(t)}(f(s,\tilde\pi_k(s))) - w_{k(t)}(y_{t+1})\right) \ge 2\sqrt{DT\log(2T/\delta)}\right] \le \frac{\delta}{2T^2}$$

Summing over k in (5) to k_T, with probability $1-3\delta$, we have that

$$\Delta(M,s,T) \le 1 + (D+2)k_T + (5+2\sqrt{D})\sqrt{T\log(2T/\delta)} + 2\sum_y \frac{\nu_k(y)}{\sqrt{t_k}} + (D+2)\sum_{k=1}^{k_T}\sum_y \nu_k(y)\beta_k^p(y) + 2\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)\beta_k^r(s,a) \quad (6)$$

In the final step, we are going to plug in the confidence bound

$$\beta_k^p(y) := \sqrt{\frac{2S + 4\log(2Yt_k/\delta)}{N_k^+(y)}} \le \sqrt{\frac{2S}{N_k^+(y)}} + 2\sqrt{\frac{\log(2Yt_k/\delta)}{N_k^+(y)}}$$

and

$$\beta_k^r(s,a) := \sqrt{\frac{\log(2SAt_k/\delta)}{N_k^+(s,a)}}.$$

It suffers to apply Lemma 19 and Proposition 18 in Jaksch et al. (2010).

- For any sequence of numbers $z_1,\cdots,z_n$ with $0 \le z_k \le Z_{k-1} := \max\left\{1, \sum_{i=1}^{k-1} z_i\right\}$

$$\sum_{k=1}^n \frac{z_k}{\sqrt{Z_{k-1}}} \le (\sqrt{2}+1)\sqrt{Z_n}. \quad (7)$$

- The number k_T up to step $T > SA$ can be upper bounded as

$$k_T \le SA\log_2\left(\frac{8T}{SA}\right). \quad (8)$$

Note that

$$\nu_k(y) = \sum_{(s,a)\in f^{-1}(y)} \nu_k(s,a) \le \sum_{(s,a)\in f^{-1}(y)} N_k(s,a) = N_k(y).$$

By (7) and Cauchy-Schwarz inequality,

$$\sum_{k=1}^{k_T} \sum_y \frac{\nu_k(y)}{\sqrt{N_k^+(y)}} \le \sum_y \left(\sqrt{2}+1\right) \sqrt{N_{k+1}^+(y)} \le \left(\sqrt{2}+1\right) \sqrt{YT} \tag{9}$$

Similarly,

$$\sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) / \sqrt{N_k^+(s,a)} \le (\sqrt{2}+1)\sqrt{SAT}. \tag{10}$$

Combining (6), (8), (9) and (10), with probability $1-3\delta$, we arrive at

$$\begin{aligned}
\Delta(M,s,T) &\le 1 + (D+2)SA\log_2\left(\frac{8T}{SA}\right) + (5+2\sqrt{D})\sqrt{T\log(2T/\delta)} + 2\sum_y \frac{\nu_k(y)}{\sqrt{t_k}} \\
&\quad + (D+2)\sum_{k=1}^{k_T}\sum_y \nu_k(y)\beta_k^p(y) + 2\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)\beta_k^r(s,a) \\
&\le (D+2)\left(4\sqrt{SYT} + 5\sqrt{YT\log(2YT/\delta)} + SA\log_2(8T/SA)\right) \\
&\quad + (5+2\sqrt{D})\sqrt{T\log(2T/\delta)} + 5\sqrt{SAT\log(2SAT/\delta)} + 5\sqrt{YT} \\
&\le 25D\sqrt{SYT\log(2YT/\delta)} + 15\sqrt{SAT\log(SAT/\delta)}
\end{aligned}$$

4 Discussion

It is well-known that the worst-case lower bound expect regret of any algorithm is $\Omega(\sqrt{DSAT})$, and UCRL2 can be improved theoretically by a factor of $\sqrt{DS}$. Many variants have been presented to improve this gap by choosing tighter confidence bounds. Among these works, Filippi et al. (2010) introduces UCRL-KL, which is a variant of UCRL2 that uses the KL divergence to define confidence bounds. Similarly to UCRL2, KL-UCRL achieves a regret of $\Omega(DS\sqrt{AT})$ for the class of communicating MDPs with improved log-log correction of T. Recently, this gap was closed using empirical Berstein inequality by Tossou et al. (2019). Another approach presented by Bourel et al. (2020) is to tighten exploration in UCRL motivated by practical efficiency. Although our work does not aim on this improvement as it is for a set specific MDPS (where the worst case is not included), it is nevertheless useful to mention that those improvements for general MDPs can be adapted to our methodology for MDPs with post-decision states, therefore reduce the regret by a factor of $\Omega(SA/Y)$.

References

Peter Auer and Ronald Ortner. Online regret bounds for a new reinforcement learning algorithm. *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pages 49-56, 2007.

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

Liyuan Zheng and Lillian Ratliff. Constrained Upper Confidence Reinforcement Learning. *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, pages 620–629, 2020.

Hippolyte Bourel, Odalric Maillard and Mohammad Sadegh Talebi. Tightening Exploration in Upper Confidence Reinforcement Learning. *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:1056-1066, 2020.

Sarah Filippi, Olivier Cappe, and Aurélien Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122, 2010.

Aristide Tossou, Debabrota Basu and Christos Dimitrakakis Near-optimal Optimistic Reinforcement Learning using Empirical Berstein Inequalities. *ArXiv e-prints*, page arXiv: 1905.12425, May 2019.