# Introduction to the domain of research
# Regret Bounds in Reinforcement Learning

JIAN QIAN
jian.qian@ens.fr
ADVISOR: ALEXANDER RAKHLIN

## Contents

**1 MDP as a Planning Problem** **2**
1.1 Basic Concepts . . . 2
1.2 Average Reward . . . 2
1.3 Evaluation Equations . . . 3
1.4 Classification of Markov Decision Processes . . . 4
1.5 The Average Reward Optimality Equation – Unichain Model . . . 5
1.6 Value Iteration in Unichain Models . . . 6

**2 MDP as a Reinforcement Learning Problem** **7**
2.1 Basic Concepts . . . 7
2.2 Algorithm . . . 8
2.3 Extended Value Iteration . . . 8
2.4 Analysis . . . 9

**3 Open Questions** **11**

# 1 MDP as a Planning Problem

The central goal of this section is to define the Markov Decision Process(MDP) and describe what we want to achieve regarding this subject as a planning problem.

## 1.1 Basic Concepts

**Definition 1.** *An Markov Decision Process(MDP) is a four-tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, such that $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{P} : \bigcup_{s\in\mathcal{S}} \{s\} \times \mathcal{A}_s \to \mathcal{D}(\mathcal{S})$, where $\mathcal{A}_s$ is the set of feasible actions at state $s$, $\mathcal{D}$ is the set of probability distributions on the set $\mathcal{S}$, and $r : \bigcup_{s\in\mathcal{S}} \{s\} \times \mathcal{A}_s \to [0, r_{max}]$ is the reward, where $r(s, a)$ is the immediate reward the agent gets performing action $a$ at state $s$.*

**Definition 2.** *A decision rule $d : \mathcal{S} \to \bigcup_s \mathcal{D}(\mathcal{A}_s)$ is a map from the states to the distributions on the set of actions.*

**Definition 3.** *A policy $\pi = (d_1, d_2, ...)$ is an infinite sequence of decision rules.*

We consider three sets of policies,

**Definition 4.** *A stationary randomized policy $\pi = (d, d, d, ...)$, is a sequence of identical decision rules. The set of all stationary randomized policy is denoted $\Pi^{SR}$.*
*A Markov randomized policy $\pi = (d_0, d_1, d_2, ...)$, is a sequence of decision rules that only depend on its index. The set of all Markov randomized policy is denoted $\Pi^{MR}$.*
*A history dependent randomized policy $\pi = (d_0(h_0), d_1(h_1), d_2(h_2), ...)$ is a sequence of decision rules which depend on the history $(h_0, h_1, h_2, ...)$ the agent has seen, where the history is defined as below in Def 5. The set of all history dependent randomized policy is denoted $\Pi^{HR}$.*

**Remark:** We sometimes denote $d^\infty = (d, d, ...) \in \Pi^{SR}$.

**Definition 5.** *A history at time $t \in \mathbb{N}$ is a sequence of states, actions and rewards that are generated according to a policy and an MDP, i.e. $h_t = (s_0, a_0, r_0, s_1, a_1, r_1, ...s_t, a_t, r_t)$, where $a_0$ is drown out of $d_0(s_0)$, and $s_1$ drown out of $\mathcal{P}(\cdot|s_0, a_0)$ and $r_0 = r(s_0, a_0)$, and so on.*

So the goal is to find an algorithm that can almost guarantee us with a very good reward in the long run, as well as not lose too much in the short run.

## 1.2 Average Reward

**Assumption 1.** *Stationary rewards and transition probabilities: $r(s, a)$ and $\mathcal{P}(j|s, a)$ do not depend on the stage.*

**Assumption 2.** *Bounded rewards: $|r(s, a)| \le M < \infty$ for all $a \in A_s$ and $s \in S$.*

**Assumption 3.** *Finite-state spaces: $|S| < \infty$.*

### 1.2.1 The average reward of a fixed policy

**Definition 6.** *We define the average reward of a policy $\pi$ as,*

$$g_\pi(s) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} r_i\right]\Bigg|_{s_0=s, a_i\sim\pi_i(\cdot|s_i), s_{i+1}\sim\mathcal{P}(\cdot|s_i,a_i)}$$

*Notice the average reward is not defined for all cases, although for most of the cases that we are concerned it is, as to show that we introduce the following two limits.*

$$g_\pi^+(s) = \limsup_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} r_i\right]\Bigg|_{s_0=s, a_i\sim\pi_i(\cdot|s_i), s_{i+1}\sim\mathcal{P}(\cdot|s_i,a_i)}$$

$$g_\pi^-(s) = \liminf_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} r_i\right]\Bigg|_{s_0=s, a_i\sim\pi_i(\cdot|s_i), s_{i+1}\sim\mathcal{P}(\cdot|s_i,a_i)}$$

**Proposition 1.** *Let $d^\infty \in \Pi^{SR}$, $P_d(s) = \mathcal{P}(\cdot|s, d(s))$, $r_d(s) = r(s, d(s))$ and $P_d^* = \lim_{n\to\infty} P_d^\infty$, then,*

$$g_{d^\infty}(s) = P_d^* r_d(s).$$

*proof.*

$$\begin{aligned} g_{d^\infty}(s) &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} r_i\right]\Big|_{s_0=s, a_i\sim d(\cdot|s_i), s_{i+1}\sim\mathcal{P}(\cdot|s_i,a_i)} \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} P_d^i r_d(s) \\ &= P_d^* r_d(s). \end{aligned}$$

□

**Remark:** Generally if we want to force the existence of $\lim_{n\to\infty} P_d^\infty$ we generally assume aperiodic. Else we sometimes define as $P_d^* = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} P_d^n$.

**Theorem 1.** *For each $\pi \in \Pi^{HR}$ and $s \in S$, there exists a $\pi' \in \Pi^{MR}$ for which,*

- $g_{\pi'}^+ = g_\pi^+$
- $g_{\pi'}^- = g_\pi^-$

**Remark:** This theorem narrows the search of the optimal policy to the set of all Markov sets, although it is not quite clear how.

## 1.3 Evaluation Equations

Now if we fix a stationary policy $\pi = d^\infty$, we want to be able to evaluate this policy using the aforementioned criterion - average reward. While we already have the definition of it, it is not the easiest way in terms of calculation. Thus, we proceed to clarify the properties of a Markov Reward Process, which is the sequence $\{(X_t, r(X_t)) : t = 1, 2, ...\}$.

### 1.3.1 The gain and bias

**Proposition 2.** *If $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, fix a policy $\pi = d^\infty$, denote $P_d(s) = \mathcal{P}(\cdot|s, d(s))$, $r_d(s) = r(s, d(s))$, and $P_d^* = \lim_{n\to\infty} P_d^n$, if $j$ and $k$ are in the same recurrent class, i.e. $P_d^*(k|j) > 0$ , then $g_{d^\infty}(j) = g_{d^\infty}(k)$.*

*proof.* Because if $j$ and $k$ are in the same recurrent class, it means that the corresponding rows for $j$ and $k$ are the same, as the stationary distribution starting at $j$ and $k$ are the same. Thus the result. □

Now we define the bias, which characterize the advantage/disadvantage of being in the states.

**Definition 7.** *The bias is defined as:*

$$h_{d^\infty}(s) := \lim_{N\to\infty}\sum_{n=1}^{N}(P_d^{n-1} - P_d^*)r_d(s) = \lim_{N\to\infty}\left(\sum_{n=1}^{N} P_d^{n-1} r_d(s) - N g_{d^\infty}(s)\right)$$

### 1.3.2 Evaluation equations

To shorten our burden of notation, we denote $g_d = g_{d^\infty}$ and $h_d = h_{d^\infty}$.

**Theorem 2.** *All the notations inherited, we have, for any $(g, h)$ satisfying the equations on the left hand side, then we have the equations on the right hand side.*

$$\begin{cases} (I - P_d)g = 0 \\ r_d - g + (P_d - I)h = 0 \end{cases} \Rightarrow \begin{cases} g = g_d \\ h \in h_d + ker(I - P_d) \end{cases}$$

*proof.* Just have to show $g = P_d^* r_d$ and $(I - P_d)(h - h_d) = 0$.

$$g = P_d g = P_d^2 g = .. = P_d^* g$$

And,

$$\begin{aligned} 0 &= P_d^*(r_d - g + (P_d - I)h) \\ &= P_d^* r_d - P_d^* g \end{aligned}$$

Thus we have,

$$g = P_d^* r_d$$

And,

$$\begin{aligned} (I - P_d)(h - h_d) &= (I - P_d)h - (I - P_d)h_d \\ &= r_d - g - (I - P_d)h_d \\ &= r_d - g - (I - P_d) \lim_{N \to \infty} \left( \sum_{n=1}^{N} P_d^{n-1} r_d(s) - N g_{d^\infty}(s) \right) \\ &= r_d - g - (r_d - g) \\ &= 0 \end{aligned}$$

□

Then we know that for certain MDP and policy, the equation suffice to provide the essential information.

**Corollary 1.** *Suppose, $P_d$ induces an irreducible Markov Chain, then*

$$r_d - ge + (P_d - I)h = 0 \Rightarrow \begin{cases} ge = g_d \\ h = h_d + ke \end{cases}$$

*where* $e = (1, ..., 1)^T$.

## 1.4 Classification of Markov Decision Processes

Now we provide a classification scheme for the MDPs for taxonomy.

### 1.4.1 Classify a Markov Decision Process

- Recurrent: All stationary deterministic policy induces only one single recurrent class.
- Unichain: All stationary deterministic policy induces a single recurrent class plus a possibly empty set of transient states.
- Communicating: There exists a stationary randomized policy that induces only a single recurrent class.
- Weakly Communicating: There exists a stationary randomized policy that induces a single recurrent class plus a set of transient states.
- Multichain: There exists a stationary policy that induces two or more recurrent classes.

The relationship between all the categories are shown below in the Venne diagram - Figure 1.

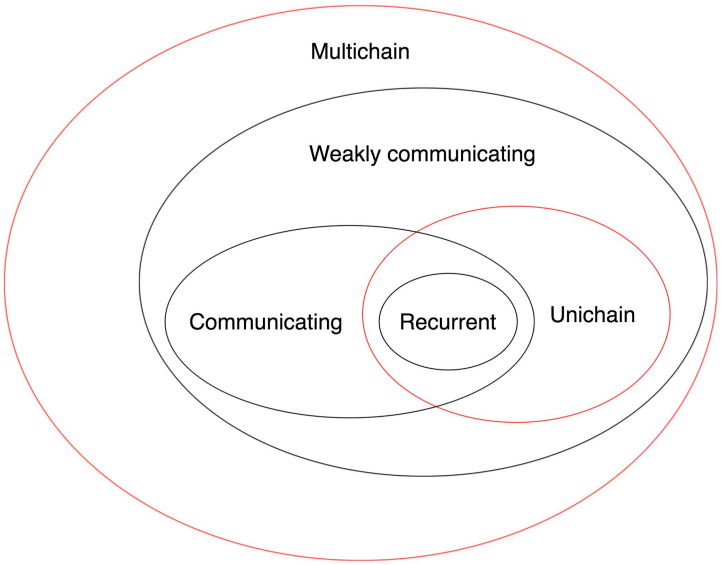


Figure 1: Venne diagram

### 1.4.2 Model Classification and the Average Reward Criterion

**Definition 8.** *A stationary optimal policy $\pi^*$ for an MDP $\mathcal{M}$ is a stationary randomized policy that for any other stationary randomized policy $\pi'$ we have,*

$$g_{\pi^*} \geq g_{\pi'}$$

**Theorem 3.** *For a weakly communicating model, if there is a stationary optimal policy $\pi$, then the average reward vector is constant, i.e., $g_\pi = ge$.*

*proof.* Let's suppose there is a stationary optimal policy $\delta$ then we can alter it to another policy $\pi$ which is with constant average reward across the states. Let's suppose the recurrent class with the highest average reward under policy $\delta$ is $C_\delta$. For any states in $C_\delta$, we don't change the action, for all the other states, we change the action to have them able to go to $C_\delta$ to get $\pi$. □

## 1.5 The Average Reward Optimality Equation – Unichain Model

Now we restrict ourselves to the unichain MDP, where we can have a constant optimal average reward across the states, we always denote the optimal policy $\pi^*$, the optimal gain $g^*$.

### 1.5.1 The Optimality Equation

**Lemma 1.** *The cumulative reward at step $N$ denoted $v_N$ has the following expression:*

$$v_N^\pi = \mathbb{E}\left[\sum_{n=1}^{N} r(s_n, a_n)\right] = \sum_{n=1}^{N} P_{d_1}..P_{d_{n-1}} r_{d_n}$$

**Definition 9.** *The Bellman operator $L : \mathbb{R}^S \to \mathbb{R}^S$ is defined as,*

$$LV(s) = \max_{a \in \mathcal{A}_s} (r(s,a) + \mathcal{P}(\cdot|s,a)V)$$

**Theorem 4.** *Let $g_+^* = \max_\pi g_\pi^+$, $g_-^* = \min_\pi g_\pi^+$*

- $Lh \leq h + ge \Rightarrow g_+^* \leq g$
- $Lh \geq h + ge \Rightarrow g_-^* \geq g$
- $Lh = h + ge \Rightarrow g = g^* = g_+^* = g_-^*$

*proof.* For any $\pi = (d_1, d_2, ...)$, we will have,

$$ge \geq r_{d_1} + (P_{d_1} - I)h.$$

Then,

$$ge \geq r_{d_2} + (P_{d_2} - I)h,$$
$$ge = gP_{d_1}e \geq P_{d_1}r_{d_2} + P_{d_1}(P_{d_2} - I)h.$$

And so on,

$$ge \geq P_{d_1}P_{d_2}...P_{d_{n-1}}r_{d_n} + P_{d_1}P_{d_2}...P_{d_{n-1}}(P_{d_n} - I)h$$

And then we sum all of them up,

$$nge \geq r_{d_1} + ... + P_{d_1}P_{d_2}...P_{d_{n-1}}r_{d_n} + P_{d_1}P_{d_2}...P_{d_{n-1}}P_{d_n}h - h$$

Thus, we have

$$g \geq g^*_+$$

Because $Lh \geq h + ge$, so there is a $d$ s.t. $P_d h + r_d \geq h + ge$, then,

$$P_d^n h + P_d^{n-1} r_d \geq P_d^{n-1} h + ge$$

So as above, we will have,

$$g^*_- \geq g^-_{d^\infty} \geq g.$$

□

### 1.5.2 Existence of Solution to the Optimality Equation

**Theorem 5.** *For all $s \in \mathcal{S}, |A_s| \leq \infty$, the model is unichain, then $\exists\ (g, h)$ such that*

$$L(h) = h + ge$$

### 1.5.3 Identification and Existence of Optimal Policies

**Theorem 6.** *Suppose $h^*, g^*$ satisfies $Lh^* = h^* + g^*e$, and a policy $d^* \in arg\max_d L_d h^*$, then $(d^*)^\infty$ is average optimal.*

*proof.* We know that $L_{d^*}h^* = h^* + g^*$, thus we know that $g_{(d^*)^\infty} = g^*$, thus $(d^*)^\infty$ is average optimal. □

**Corollary 2.** *If for alls $\in \mathcal{S}, |A_s| \leq \infty$, the model is unichain, then we have solutions for optimal gain, bias and policy.*

## 1.6 Value Iteration in Unichain Models

Then we try to develop an algorithm for solving the Optimality Equation.

### 1.6.1 Bounds on the Gain

**Theorem 7.** *If for all $s \in \mathcal{S}, |A_s| \leq \infty$, the model is unichain, then we have for $v \in V$,*

$$\min_{s\in S}\left[Lv(s) - v(s)\right] \leq g^{d^\infty} \leq g^* \leq \max_{s\in S}\left[Lv(s) - v(s)\right],$$

*where $d \in arg\max_d L_d v$*

*proof.* This is actually an immediate corollary from Theorem 2 and 4. □

**Theorem 8.** *If for all $s \in \mathcal{S}, |A_s| \leq \infty$, the model is unichain, and $h$ exists, then we have for $v \in \mathbb{R}_+^S$,*

$$\frac{1}{n} L^n v = g^*$$

*proof.* Find a $C, D$ s.t. $Ce \geq h, De > g$, then we have,

$$\begin{gathered} v \geq h - Ce \\ Lv \geq L(h - Ce) \\ \cdots \\ L^n v \geq L^n(h - Ce) = nge + h - Ce \end{gathered}$$

And

$$\begin{gathered} v \leq De + h \\ \cdots \\ L^n v \leq De + nge + h \end{gathered}$$

Thus the result. □

With Theorem 7 and 8, we might have(and under certain condition indeed we have), that hopefully, $L^n v - L^{n-1} v$, $L^n v$ converges to $g^*$, $h^*$ respectively.

# 2 MDP as a Reinforcement Learning Problem

## 2.1 Basic Concepts

**Definition 10.** *The accumulated reward of an algorithm $\mathfrak{A}$ after $T$ steps in an MDP $\mathcal{M}$ with initial state $s$, is defined as:*

$$R(\mathcal{M}, \mathfrak{A}, s, T) := \sum_{t=1}^{T} r_t$$

**Definition 11.** *Consider the stochastic process defined by a stationary policy $\pi$ and an MDP $\mathcal{M}$. Let $T(s'|\mathcal{M}, \pi, s)$ be the random variable for the first time step in which state $s'$ is reached in this process. Then the diameter of $\mathcal{M}$ is defined as:*

$$D(\mathcal{M}) := \max_{s \neq s'} \min_{\pi} \mathbb{E}[T(s'|\mathcal{M}, \pi, s)]$$

**Lemma 2.** *$sp\{h^*(M)\} \leq r_{\max} D(M)$, where $sp\{V\} = \max_s V(s) - \min_s V(s)$.*

**Definition 12.** *The total regret of $\mathfrak{A}$ after $T$ steps as:*

$$\Delta(\mathcal{M}, \mathfrak{A}, s, T) = Tg^*(\mathcal{M}) - R(\mathcal{M}, \mathfrak{A}, s, T)$$

**Notations:** $S = |\mathcal{S}|$, $A = \max_s |\mathcal{A}_s|$ and $\mathcal{H} := \bigcup_{s \in \mathcal{S}} \{s\} \times \mathcal{A}_s$

**Theorem 9.** *(Azuma-Hoeffding inequality) Suppose $\{X_k : k = 0, 1, 2, ...\}$ is a martingale and $|X_k - X_{k-1}| < c$ almost surely. Then for all positive integers $n$ and all positive reals $\varepsilon$,*

$$\mathbb{P}(|X_n - X_0| \leq \varepsilon) \geq 1 - 2\exp(-\frac{\varepsilon^2}{2nc^2})$$

## 2.2 Algorithm

**Assumption 4.** $D := D(\mathcal{M}) \leq \infty$, $r_{\max} \leq 1$

**Algorithm 1** UCRL

**Input:** $\mathcal{M}$

**Output:** $\pi_K$

**for** episode $k = 1, 2, ..$ **do**

▷ **Initialize episode** $k$**:**

Set the start time of episode $k$, $t_k := t$.

For all $(s, a) \in \mathcal{H}$ initialize the state-action counts for episode $k$, $v_k(s, a) = 0$. Further, set the state-action counts prior to episode $k$,

$$N_k(s, a) := \#\{\tau < t_k : s_\tau = s, a_\tau = a\}$$

For $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}_s$ set the observed accumulated rewards and the transition counts prior to episode $k$,

$$R_k(s, a) := \sum_{\tau=1}^{t_k-1} r_\tau \mathbb{1}(s_\tau = s, a_\tau = a),$$
$$P_k(s, a, s') := \#\{\tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}$$

Compute estimates $\hat{r}_k(s, a) := \frac{R_k(s,a)}{\max\{1, N_k(s,a)\}}, \hat{p}_k(s'|s, a) := \frac{P_k(s,a,s')}{\max\{1, N_k(s,a)\}}$

▷ **Compute policy** $\tilde{\pi}_k$**:**

Let $M_k$ be the set of all MDPs with states and actions as in $\mathcal{M}$, and with transition probabilities $\tilde{p}(\cdot|s, a)$ close to $\hat{p}_k(\cdot|s, a)$, and rewards $\tilde{r}(s, a) \in [0, 1]$ close to $\hat{r}_k(s, a)$, that is,

$$|\tilde{r}(s, a) - \hat{r}_k(s, a)| \leq \sqrt{\frac{7 \log(2SAt_k/\delta)}{2 \max\{1, N_k(s, a)\}}}$$
$$||\tilde{p}(\cdot|s, a) - \hat{p}_k(s, a)||_1 \leq \sqrt{\frac{14S \log(2At_k/\delta)}{\max\{1, N_k(s, a)\}}}$$

Use **Extended Value Iteration** to find a policy $\tilde{\pi}_k$ such that

$$\tilde{g}_k := \min_s g(\tilde{\mathcal{M}}_k, \tilde{\pi}_k, s) \geq \max_{\pi, s'} g(\tilde{\mathcal{M}}_k, \pi, s') - \frac{1}{\sqrt{t_k}}$$

▷ **Execute policy** $\tilde{\pi}_k$**:**

**while** $v_k(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_k(s_t, \tilde{\pi}_k(s_t))\}$ **do**

Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward $r_t$, and observe next state $s_{t+1}$.

Update $v_k(s_t, a_t) := v_k(s_t, a_t) + 1$

Set $t := t + 1$

**end while**

**end for**

## 2.3 Extended Value Iteration

The goal of Extended Value Iteration is to extend our search space according to our current data to find an optimistic (but not ridiculous) estimation of the average return.

The mathematical formulation is now we consider an extended MDP $\mathcal{M}_k = (\mathcal{S}, \tilde{\mathcal{A}}, \tilde{\mathcal{P}}, \tilde{r})$, where state set $\mathcal{S}$ is the same, but the action set is augmented:

$$\tilde{\mathcal{A}}_s = \left\{ (a, p(\cdot|s, a), r(s, a)) \middle| a \in \mathcal{A}_s, ||p(\cdot|s, a) - \hat{p}_k(s, a)||_1 \leq \sqrt{\frac{14S \log(2At_k/\delta)}{\max\{1, N_k(s, a)\}}}, |r(s, a) - \hat{r}_k(s, a)| \leq \sqrt{\frac{7 \log(2SAt_k/\delta)}{2 \max\{1, N_k(s, a)}} \right.$$

The transition is as:

$$\tilde{\mathcal{P}}(\cdot|s,(a,p(\cdot|s,a),r(s,a))) = p(\cdot|s,a)$$

The reward is such:

$$\tilde{r}(s,(a,p(\cdot|s,a),r(s,a))) = r(s,a)$$

It is easy to verify that this is an MDP, and so we can do value iteration, and we call it the Extended Value Iteration, i.e.:

$$u_0(s) = 0,$$
$$u_{i+1}(s) = \max_{a\in\tilde{\mathcal{A}}}\left\{\tilde{r}(s,a) + \max_{p\in\mathcal{B}(\hat{p}_k(\cdot|s,a))}\left\{\sum_{s'\in\mathcal{S}} p(s')u_i(s')\right\}\right\}$$

**Algorithm 2** Computing the Inner Maximum in the Extended Value Iteration

**Input:** Estimates $\hat{p}(\cdot|s,a)$ and distance $d(s,a)$ for a state-action pair $(s,a)$, and the states in $\mathcal{S}$ sorted descendingly according to their $u_i$ value. That is, let $\mathcal{S} = \{s'_1, s'_2, ..., s'_n\}$ with $u_i(s'_1) \geq u_i(s'_2) \geq ... \geq u_i(s'_n)$

1: Set

$$p(s'_1) := \min\left\{1, \hat{p}(s'_1|s,a) + \frac{d(s,a)}{2}\right\}$$
$$p(s'_j) := \hat{p}(s'_j|s,a) \text{ for all states } s'_j \text{ with } j > 1.$$

2: Set $l := n$

3: **while** $\sum_{s'_j\in\mathcal{S}} p(s'_j) > 1$ **do**

4: Reset $p(s'_l) := \max\left\{0, 1 - \sum_{s'_j\neq s'_l} p(s'_j)\right\}$.

5: Set $l := l - 1$.

6: **end while**

## 2.4 Analysis

**Theorem 10.** *By Azuma-Hoeffding inequality, with a very large probability, we have, for all $k, t$, the true probability and true reward are in the estimated interval, i.e.,*

$$||\mathcal{P}(\cdot|s,a) - \hat{p}_k(s,a)||_1 \leq Z_k^{sa} := \sqrt{\frac{14S\log(2At_k/\delta)}{\max\{1, N_k(s,a)\}}}$$
$$|r(s,a) - \hat{r}_k(s,a)| \leq \delta_k^{sa} := \sqrt{\frac{7\log(2SAt_k/\delta)}{2\max\{1, N_k(s,a)\}}}$$

**Definition 13.**

$$Q_k = \bigcap_{i\leq k}\{||\mathcal{P}(\cdot|s,a) - \hat{p}_i(s,a)||_1 \leq Z_i^{sa}, |r(s,a) - \hat{r}_i(s,a)| \leq \delta_i^{sa}\}$$
$$X_t = p(\cdot|s,a)\tilde{h}_k - \tilde{h}_k(s_{t+1})$$

*And*

$$Q = \bigcap_k (Q_k \bigcap \{\sum_{t=1}^{T} X_t \mathbb{1}(Q_k) \leq \widetilde{O}(D\sqrt{T})\})$$

**Remark:** By union law we know that $Q$ also happens with a very large probability.

**Theorem 11.** *(Optimism) Given Q, we have,*

$$g^*(\tilde{\mathcal{M}}_k) \geq g^*(\mathcal{M})$$

*proof.*

$$g^*(\tilde{\mathcal{M}}_k) = \lim_{n\to\infty} \frac{1}{n}\tilde{L}0 \geq \lim_{n\to\infty} \frac{1}{n}L0 = g^*(\mathcal{M})$$

□

**Notation:** $\tau_k$ is the time when the $k$-th episode start.

**Lemma 3.** *Given Q, we have,*

$$\sum_{t=1}^{T} g^* \leq \sum_{t=1}^{T} \tilde{g}_k$$

$$-\sum_{t=1}^{T} r(s_t, a) \leq -\sum_{t=1}^{T} \tilde{r}_k(s_t, a) + 2\delta_k^{sa}(\tau_{k+1} - \tau_k)$$

**Lemma 4.** *Given Q, then we have,*

$$\tilde{D} = D(\tilde{M}_k) \leq D$$

*Thus,*

$$sp\{\tilde{h}_k\} \leq r_{\max} D \leq D.$$

**Lemma 5.** *Given Q, then we have,*

$$\sum_{t=1}^{T} \frac{(\tau_{k+1} - \tau_k)}{\sqrt{\tau_k}} \leq O(\sqrt{T\log(T)}) = \widetilde{O}(\sqrt{T}), \quad \sum_k 2\delta_k^{sa} v_k(s,a) \leq \widetilde{O}(\sqrt{SAT})$$

$$DZ_k^{sa} v_k(s,a) \leq \widetilde{O}(DS\sqrt{AT}), \quad \sum_{t=1}^{T} X_t \mathbb{1}(Q_k) \leq \widetilde{O}(D\sqrt{T})$$

**Theorem 12.** *Given Q, we have,*

$$\mathcal{R}_T \leq O\left(DS\sqrt{AT}\right)$$

*proof.*

$$\begin{aligned}
R_T &= \sum_{t=1}^{T}(g^* - r(s_t, a_t)) \\
&\leq \sum_{t=1}^{T}(\tilde{g}_k - \tilde{r}(s_t, \pi_k(s_t))) + O\left(\sum_k 2\delta_k^{sa} v_k(s,a) + \sum_{t=1}^{T} \frac{(\tau_{k+1} - \tau_k)}{\sqrt{\tau_k}}\right) \\
&= \sum_{t=1}^{T}(p_k(\cdot|s_t, \tilde{a}) - 1_{s_t})\tilde{h}_k + \widetilde{O}\left(\sqrt{SAT}\right) \\
&= \sum_{t=1}^{T}(p_k(\cdot|s_t, \tilde{a}) - p(\cdot|s,a))\tilde{h}_k + \sum_{t=1}^{T}(p(\cdot|s,a)\tilde{h}_k - \tilde{h}_k(s_t)) + \widetilde{O}\left(\sqrt{SAT}\right) \\
&\leq \sum_{t=1}^{T} X_t + \sum_{k=1}^{m}(\tilde{h}_k(s_{\tau_{t+1}}) - \tilde{h}_k(s_{\tau_t})) + \widetilde{O}\left(\sqrt{SAT} + DZ_k^{sa} v_k(s,a)\right) \\
&\leq \sum_{t=1}^{T} X_t + \widetilde{O}\left(DSA + DS\sqrt{AT}\right) \\
&= \sum_{t=1}^{T} X_t \mathbb{1}(Q_k) + \widetilde{O}\left(DSA + DS\sqrt{AT}\right) \\
&\leq \widetilde{O}(DS\sqrt{AT})
\end{aligned}$$

□

# 3 Open Questions

- The theoretical lower bound for the regret bound is $\Omega(\sqrt{DAST})$, so is it possible to find an explicit algorithm that is able to achieve such bounds.
- The method we are using now is model based, and it requires a large memory and computation, so can we find a better way to achieve the same regret bound.
- Weakly communicating model might not be the suitable model regarding the scenarios of application, so how can we define some regret bounds for general models, and find some algorithm to solve them.