Notes of [2]

Chao Tao

Mar. 13, 2020

1 Problem Setup

There is a tabular infinite undiscounted MDP $M^* = (\mathcal{S}, \mathcal{A}, \theta^*, c, x_1)$ where c is bounded within $[0, 1]$ and only the transition probability θ^* is *unknown*. x_1 is the initial state which can be either randomized or *adversarial*. For simplicity, we also assume the cost function c is *deterministic*. We want to find a policy such that the cost incurred by this policy after T time steps is minimized.

In infinite undiscounted MDP, the average cost per step for any policy π is defined as

$$\ell_\pi \stackrel{\text{def}}{=} \limsup_{T \to +\infty} \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} c(x_t, a_t)\right],$$

where x_t and a_t denotes the state and action pair at the tth time step. Note that we have removed the dependency on policy to simplify the notations. Let π^* be the optimal policy such that $\ell_{\pi^*} = \min_{\pi'} \ell_{\pi'}$. And the *frequentist* regret is defined by

$$\mathcal{R}_T^\pi \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} c(x_t, a_t)\right] - T\ell_{\pi^*}.$$

1.1 Weakly Communicating MDP

To make it possible to suffer a sub-linear regret, we also need to make some restrictions on the underlying MDP. Here, we assume the underlying MDP is *weakly communicating*.

Definition 1. *An MDP is weakly communicating iff the state space $\mathcal{S}$ can be decomposed into two parts $\mathcal{S}_1$ and $\mathcal{S}_2$ such that every state in $\mathcal{S}_1$ is reachable from other states in $\mathcal{S}_1$ under some policy, whereas all states in $\mathcal{S}_2$ are transient under all policies.*

The intuition to introduce such a concept is to avoid trap states. For example we can construct an MDP as the following (see Figure 1):

i. $\mathcal{S} = \{s_1, s_2\}$

ii. $\mathcal{A} = \{a_1, a_2\}$

iii. $\theta^*(s_1 \mid s_1, a_1) = 1$, $\theta^*(s_2 \mid s_1, a_2) = 1$, $\theta^*(s_2 \mid s_2, a_1) = 1$, $\theta^*(s_2 \mid s_2, a_2) = 1$

iv. $c(s_1, a_1) = 0.5$, $c(s_1, a_2) = 1$, $c(s_2, a_1) = 1$, $c(s_2, a_2) = 1$

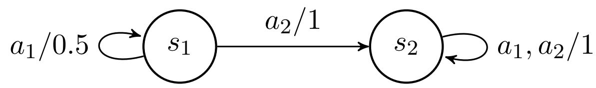

Figure 1: A counterexample

v. $x_1 = s_1$

It is clear that the optimal policy is $\pi^* = a_1$. However, no policy could achieve $\widetilde{\mathcal{O}}(\sqrt{T})$. We prove this by contradiction. Suppose such a policy exists. We call it π'. Then during the first $T/2$ steps, it must not try action a_2. Otherwise, the regret would be at least $0.25T = \Omega(T)$. A key observation is that when we change the cost function such that $c(s_1, a_1) = 1$, $c(s_1, a_2) = 0.5$, $c(s_2, a_1) = 0.5$ and $c(s_2, a_2) = 0.5$, π' will not change its behavior in the first $T/2$ steps since it does not even try action a_1. Note that during the first $T/2$ steps, it already incurs a $\Omega(T)$ regret. A contradiction happens.

1.2 Optimality

Theorem 1. *There always exists a stationary deterministic policy π^* achieving the optimal average cost and its average cost satisfies*

$$\ell_{\pi^*}^{M^*} + v(x, \theta^*) = \min_{a \in \mathcal{A}} \left\{ c(x, a) + \sum_{x' \in \mathcal{S}} \theta^*(x' \mid x, a) v(x', \theta^*) \right\},$$

where $v(\cdot, \theta^*)$ is called the bias vector of MDP M^*.

It is easy to see if $v(\cdot, \theta^*)$ is a bias vector of model M^*, so does $v(\cdot, \theta^*) - C$ where C is an arbitrary constant. Hence w.l.o.g., we assume $\min_{x \in \mathcal{S}} v(x, \theta^*) = 0$ also $\max_{x \in \mathcal{S}} v(x, \theta^*) \le D'$.

Remark 2. *We only assume the existence of D'. We do not assume D' is known beforehand.*

From now on, we only need to consider a MDP such that its bias vector is upper bounded by D' and stationary deterministic policies.

2 Thompson Sampling

Like *Optimism in the Face of Uncertainty*, *Thompson Sampling* dating back to [3] is another general principal guiding you how to operate in a poorly understood environment. Due to its superior empirical performance [1], it gains increasing popularity recently.

Thompson Sampling is a *Bayesian* method. Basically, at the very begining, the learner equipped with this policy assumes a prior distribution $\mathcal{P}_1$ on the unknown parameter of the underlying environment i.e., θ^*. At the begining of each episode $k \ge 1$, the learner just samples a virtual environment from the posterior distribution $\mathcal{P}_k$ on θ^* which is derived based on $\mathcal{P}_{k-1}$ and the history in the $(k-1)$th episode via Bayes' Theorem and then takes the optimal policy assuming the underlying model is the sampled one.

To apply Thompson Sampling, we need to design a stopping criteria for each episode. Before describing the stopping criteria, we introduce several notations. Let t_k and T_k denote the start time and the length of the kth episode respectively. Also let $N_t(x, a)$ be the number of visits of state-action pairs before time step t. In the algorithm, episode k finishes if one of the following situation happens:

i. $t - t_k > T_{k-1}$ or

ii. $\exists (x, a) \in \mathcal{S} \times \mathcal{A}$, s.t., $N_t(x, a) > 2N_{t_k}(x, a)$.

The details are decribed in the following Algorithm 1.

Algorithm 1: Thompson Sampling

initialization: prior distribution $\mathcal{P}_1$, start of episode $k = 1$ and start time $t = 1$
while $t \le T$ **do**
  $t_k = t$ `// start time of` k`th episode`
  compute posterior distribution $\mathcal{P}_k = \mathcal{P}_1 \mid \mathcal{H}_{t_k}$
  sample θ_k from $\mathcal{P}_k$ and compute the optimal policy π_k
  while $t \le T$ *and* $t - t_k \le T_{k-1}$ *and* $N_t(x, a) \le 2N_{t_k}(x, a)\ \forall (x, a) \in \mathcal{S} \times \mathcal{A}$ **do**
    observe state x_t and take action a_t according to policy π_k
    $t = t + 1$
  $k = k + 1$

Given a prior distribution $\mathcal{P}_1$ on transition probability θ^*, the *Bayesian* regret is defined by

$$\mathcal{BR}_T^\pi \stackrel{\text{def}}{=} \mathbb{E}_{\theta^* \sim \mathcal{P}_1} \left[\mathbb{E} \left[\mathcal{R}_T^\pi \mid \theta^* \right] \right]. \tag{1}$$

3 Notations and Definitions

$[n]$	$\{1, 2, \ldots, n\}$
$\mathcal{A}$	action space
A	$\lvert\mathcal{A}\rvert$
$\mathcal{S}$	state space
S	$\lvert\mathcal{S}\rvert$
T	horizon of the MDP
$c : \mathcal{S} \times \mathcal{A} \to [0, 1]$	*known* cost function
$\theta^* : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$	transition probability of the underlying MDP
π_k	policy in the kth episode
x_1	initial state
(x_t, a_t)	state-action pair at the tth time step
$\mathcal{H}_t$	history before the tth time step $(x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t)$
$N_t(x, a)$	number of hits of state-action pair (x, a) *before* the tth time step
t_k	start time of the kth episode
T_k	length of the kth episode
$\mathcal{P}_k$	posterior distribution right *before* the kth episode
$\mathcal{BR}_T^\pi$	Bayesian regret incurred by policy π

4 Theorem

In this lecture, we are going to show

Theorem 3. *The Bayesian regret i.e.,* (1) *incurred by Algorithm 1 is bouned by* $\widetilde{\mathcal{O}}(D'S\sqrt{AT})$.

Remark 4. *The theorem holds for any prior distribution.*

Proof. In the subsequent part, unless otherwise specified, the expectation operator is taken over all random variables. Note that θ^* is treated as a random variable. Let K be the random variable denoting the total number of episodes. W.o.l.g., we assume $t_{K+1} = T + 1$. Rewrite $\mathcal{BR}_T^{\pi}$ we have

$$\begin{aligned}
(1) &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} c(x_t, a_t)\right] - T\cdot\mathbb{E}[\ell_{\pi^*}^{M^*}] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left[\ell_{\pi_k}^{M_k} + v(x_t,\theta_k) - \sum_{x'\in\mathcal{S}}\theta_k(x'\mid x_t,a_t)v(x',\theta_k)\right]\right] - T\cdot\mathbb{E}[\ell_{\pi^*}^{M^*}] \\
&= \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\ell_{\pi_k}^{M_k}\right] - T\cdot\mathbb{E}[\ell_{\pi^*}^{M^*}]}_{(I)} + \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}[v(x_t,\theta_k) - v(x_{t+1},\theta_k)]\right]}_{(II)} \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left[v(x_{t+1},\theta_k) - \sum_{x'\in\mathcal{S}}\theta_k(x'\mid x_t,a_t)v(x',\theta_k)\right]\right]}_{(III)},
\end{aligned} \tag{2}$$

where in the second last equality we have applied Theorem 1.

In the following part, we will try to bound (I), (II) and (III) separately.

Lemma 5. $(I) \le \mathbb{E}[K]$.

Proof. First we note that

$$\begin{aligned}
(I) &= \mathbb{E}\left[\sum_{k=1}^{K} T_k\ell_{\pi_k}^{M_k}\right] - T\cdot\mathbb{E}[\ell_{\pi^*}^{M^*}] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}\mathbb{E}[T_k\ell_{\pi_k}^{M_k}\mid\mathcal{H}_{t_k}]\right] - T\cdot\mathbb{E}[\ell_{\pi^*}^{M^*}] \\
&\le \mathbb{E}\left[\sum_{k=1}^{K}\mathbb{E}[(T_{k-1}+1)\ell_{\pi_k}^{M_k}\mid\mathcal{H}_{t_k}]\right] - T\cdot\mathbb{E}[\ell_{\pi^*}^{M^*}],
\end{aligned}$$

where in the last inequality we have used $T_k \le T_{k-1}$ which is enforced by the algorithm. Since conditioned on $\mathcal{H}_{t_k}$, T_{k-1} is a constant, we have $\mathbb{E}[(T_{k-1}+1)\ell_{\pi_k}^{M_k}\mid\mathcal{H}_{t_k}] = (T_{k-1}+1)\mathbb{E}[\ell_{\pi_k}^{M_k}\mid\mathcal{H}_{t_k}]$. Further utilizing the relation that $\theta_k \mid \mathcal{H}_{t_k} = \theta^* \mid \mathcal{H}_{t_k}$, we derive

$$\begin{aligned}
(I) &\le \mathbb{E}\left[\sum_{k=1}^{K}(T_{k-1}+1)\mathbb{E}[\ell_{\pi^*}^{M^*} \mid \mathcal{H}_{t_k}]\right] - T\cdot\mathbb{E}[\ell_{\pi^*}^{M^*}] \\
&= \mathbb{E}\left[\sum_{k=1}^{K}(T_{k-1}+1)\ell_{\pi^*}^{M^*}\right] - T\cdot\mathbb{E}[\ell_{\pi^*}^{M^*}] \\
&\le \mathbb{E}[K\ell_{\pi^*}^{M^*}] \le \mathbb{E}[K],
\end{aligned}$$

where the last inequality is due to $\ell_{\pi^*}^{M^*} \le 1$. ☐

Lemma 6. $(II) \le D'\mathbb{E}[K]$.

Putting Lemma 5 and Lemma 6 together, we get $(I)+(II) \le (D'+1)\mathbb{E}[K]$. We next take care of $\mathbb{E}[K]$ and try to give an upper bound of the expected number of episodes.

Lemma 7. $\mathbb{E}[K] = \mathcal{O}(\sqrt{SAT\ln(T)})$.

Proof. According to the stopping condition, we divide K episodes into M meta episodes such that within meta episode m, except for the last episode, all the other episodes ends due to the first condition i.e., $t-t_k > T_{k-1}$. Let τ_m be the start episode of meta episode m. By default, we set $\tau_{M+1} = K+1$.

Hence for any meta episode m, the total number of time steps $\sum_{k=\tau_m}^{\tau_{m+1}-1} T_k$ satisfies $\sum_{k=\tau_m}^{\tau_{m+1}-1} T_k \ge \sum_{k=\tau_m}^{\tau_{m+1}-2}(T_{\tau_m}+k-\tau_m) = (\tau_{m+1}-\tau_m-1)(2T_{\tau_m}+\tau_{m+1}-\tau_m-2)/2$. Since $T_{\tau_m} \ge 1$, we further derive $\tau_{m+1}-\tau_m \le 1+\sqrt{2\sum_{k=\tau_m}^{\tau_{m+1}-1} T_k} \le 2\sqrt{2\sum_{k=\tau_m}^{\tau_{m+1}-1} T_k}$. Next by Cauchy-Schwarz inequality, we get

$$K = \tau_M - 1 = \sum_{m=1}^{M}(\tau_{m+1}-\tau_m) \le \sum_{m=1}^{M} 2\sqrt{2\sum_{k=\tau_m}^{\tau_{m+1}-1} T_k} \le \sqrt{8MT}.$$

Note that M is at most the total number of episodes which ends due to visit number of state-action pair doubles. Hence $M = \mathcal{O}(SA\ln T)$. Using this inequality, we prove $K = \mathcal{O}(\sqrt{SAT}\ln T)$ and finishes the proof of this lemma. ☐

In the remaining part of the proof, we focus on bounding (III). Expand $v(x_{t+1},\theta_k)$ we derive

$$(III) = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left[\sum_{x'\in\mathcal{S}}\theta^*(x' \mid x_t,a_t)v(x',\theta_k) - \sum_{x'\in\mathcal{S}}\theta_k(x' \mid x_t,a_t)v(x',\theta_k)\right]\right].$$

Since $v(\cdot,\cdot) \le D'$, further by Hölder's inequality, we have

$$(III) \le D'\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\|\theta^*(\cdot \mid x_t,a_t) - \theta_k(\cdot \mid x_t,a_t)\|_1\right].$$

Let $\bar{\theta}_k(\cdot \mid x,a)$ be the empirical transition probability before the kth episode. We define $\mathcal{M}_k$ as the set of models such that its transition probability θ satisfies $|\bar{\theta}_k(\cdot \mid x,a) - \theta(\cdot \mid x,a)| \le C\sqrt{\frac{S\ln(SAT)}{1\vee N_{t_k}(x,a)}}$ for all

$(x,a) \in \mathcal{S} \times \mathcal{A}$ where C is a universal constant which will be defined later. According to Theorem 8, we know that there exists a constant $C > 0$ such that $\mathbf{Pr}(M_k \notin \mathcal{M}_k) \leq 1/T$ and $\mathbf{Pr}(M^* \notin \mathcal{M}_k) \leq 1/T$.

Plugging in events $M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k$, we get

$$
\begin{aligned}
(III) \leq & D'\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} \|\theta^*(\cdot \mid x_t, a_t) - \theta_k(\cdot \mid x_t, a_t)\|_1 \cdot \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k)\right] \\
& + D'\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k)\right] \\
\leq & D'\mathbb{E}\left[\underbrace{\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} C\sqrt{\frac{S\ln(SAT)}{1 \vee N_{t_k}(x_t, a_t)}}}_{(*)}\right] + \underbrace{D'\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k)\right]}_{(**)}.
\end{aligned} \tag{3}
$$

Note that for any $t_k \leq t < t_{k+1}$, we have $N_t(x,a) \leq 2N_{t_k}(x,a)$ holds for any state-action pair (x,a). Hence $(t_{k+1} - t_k)\sqrt{\frac{1}{1 \vee N_{t_k}(s,a)}} \leq 2 \cdot \sum_{t=N_{t_k}(x,a)}^{N_{t_{k+1}-1}(x,a)} \sqrt{\frac{1}{1 \vee t}}$. Using this inequality in $(*)$, we have

$$
\begin{aligned}
(*) &\leq \sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} C\sqrt{\frac{S\ln(SAT)}{1 \vee N_{t_k}(x_t, a_t)}} \\
&\leq 2C \cdot \sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} \sqrt{\frac{S\ln(SAT)}{1 \vee N_t(x_t, a_t)}} \\
&= 2C \cdot \sum_{t=1}^{T} \sqrt{\frac{S\ln(SAT)}{1 \vee N_t(x_t, a_t)}} \\
&= 2C\sqrt{S\ln(SAT)} \cdot \sum_{(x,a)}\sum_{t=0}^{N_T(x,a)} \sqrt{\frac{1}{1 \vee t}}.
\end{aligned}
$$

Since $\sum_{t=0}^{t'} \sqrt{\frac{1}{1 \vee t}} \leq 2\sqrt{t'} + 1$, we further derive $(*) \leq 2C\sqrt{S\ln(SAT)} \cdot (SA + \sum_{(x,a)} \sqrt{N_T(x,a)}) \leq 2C\sqrt{S\ln(SAT)} \cdot (SA + \sqrt{SAT}) = \mathcal{O}(S\sqrt{AT\ln(SAT)})$, where the second last inequality is due to Cauchy-Schwarz inequality and the last inequality is due to $T \geq \sqrt{SA}$.

Recall that $\mathbf{Pr}(M_k \notin \mathcal{M}_k) \leq 1/T$ and $\mathbf{Pr}(M^* \notin \mathcal{M}_k) \leq 1/T$. Hence we have

$$
\begin{aligned}
(**) &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} \mathbb{E}[\mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k) \mid \mathcal{H}_{t_k}]\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1} \mathbb{E}[\mathbf{Pr}(M_k \notin \mathcal{M}_k) + \mathbf{Pr}(M^* \notin \mathcal{M}_k) \mid \mathcal{H}_{t_k}]\right] \\
&\leq 2.
\end{aligned} \tag{4}
$$

Plugging in inequality $(*) \leq \mathcal{O}(S\sqrt{AT\ln(SAT)})$ and (4) back to (3), we prove this theorem.

□

5 Tools

Theorem 8 ([4])**.** *Let P be a probability distribution on the set $\mathcal{S} = \{1, \ldots, S\}$. Let X_1, X_2, ..., X_m be i.i.d. random variables distributed according to P. Then, for all $\epsilon > 0$, it holds that*

$$\mathbf{Pr}(\left\|P - \bar{P}\right\|_1 \geq \epsilon) \leq (2^S - 2)\exp(-m\epsilon^2/2),$$

where $\bar{P}$ is the empirical estimation of P defined as $\bar{P}(i) = \frac{\sum_{j=1}^{m} \mathbb{1}(X_j = i)}{m}$.

References

[1] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *NIPS*, pages 2249–2257, 2011.

[2] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *NeurIPS*, 2017.

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

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