

Notes of [2]

Chao Tao

Feb. 28, 2020

1 Problem Setup

There is a *tabular episodic* MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \theta^*, R, H, s_1)$ where we assume the reward function R is bounded within $[0, 1]$ and for simplicity we also assume R is *deterministic*. In other words, only the transition probability \mathbb{P} is *unknown*. We want to find a policy such that the *expected* regret incurred by this policy after K episodes is minimized.

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 beginning, 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 beginning of each episode $k \geq 1$, 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. The following pseudocode shows the aforementioned learning procedure.

Algorithm 1: Thompson Sampling

```
1 initialization: prior distribution  $\mathcal{P}_1$ 
2 for episode  $k = 1$  to  $K$  do
3   compute posterior distribution  $\mathcal{P}_k = \mathcal{P}_1 \mid \mathcal{H}_k$ 
4   sample  $\theta_k$  from  $\mathcal{P}_k$  and compute the optimal policy  $\pi_k$ 
5   for step  $h = 1$  to  $H$  do
6     observe state  $x_{k,h}$ 
7     take action  $a_{k,h} = \pi_k(x_{k,h})$ 
```

Denote the value function starting from time t under model M' using policy π' by $V_{\pi',t}^{M'}$. Given a prior distribution \mathcal{P}_1 on transition probability θ^* , the expected *Bayesian* regret is defined by

$$\mathcal{BR}_K^{\pi} \stackrel{\text{def}}{=} \mathbb{E}_{\theta^* \sim \mathcal{P}_1} \left[\mathbb{E} \left[\sum_{k=1}^K (V_{*,1}^{M^*} - V_{\pi_k,1}^{M^*})(x_{k,1}) \mid \theta^* \right] \right], \quad (1)$$

where the initial state for each episode can be either randomized or *adversarial*.

3 Notations and Definitions

$[n]$	$\{1, 2, \dots, n\}$
\mathcal{A}	action space
A	$ \mathcal{A} $
\mathcal{S}	state space
S	$ \mathcal{S} $
H	horizon
K	# of episodes
T	HK
$R : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$	<i>known</i> reward function
$\theta^* : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$	transition probability of the underlying MDP
$\pi = (\pi_1, \dots, \pi_K)$	an arbitrary policy where π_k is the policy in the k th episode
$V_{\pi', t}^{M'}$	value function starting from time t under model M' using policy π'
$x_{k,1}$	initial state of the k th episode
$(x_{k,h}, a_{k,h})$	state-action pair in the k th episode and at the h th time step
\mathcal{H}_k	history before the k th episode $(x_{1,1}, a_{1,1}, \dots, x_{1,H+1}, \dots, x_{k-1,1}, a_{k-1,1}, \dots, x_{k-1,H+1})$
$N_k(x, a)$	number of hits of state-action pair (x, a) <i>before</i> the k th episode
ρ	an arbitrary transition probability
V	an arbitrary value function
$(\rho V)(x, a)$	$\sum_{y \in \mathcal{S}} \rho(y x, a) V(y)$
\mathcal{BR}_K^π	Bayesian regret incurred by policy π
\mathcal{P}_k	posterior distribution right <i>before</i> the k th episode

4 Theorem

In this lecture, we are going to show

Theorem 1. When $T \geq \sqrt{SA}$, the expected Bayesian regret i.e., (1) of Algorithm 1 is bounded by $\tilde{\mathcal{O}}(HS\sqrt{AT})$.

Remark 2. 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. Rewrite \mathcal{BR}_K^π we have

$$\begin{aligned}
 (1) &= \sum_{k=1}^K \mathbb{E} \left[(V_{*,1}^{M^*} - V_{\pi_{k,1}}^{M^*})(x_{k,1}) \right] \\
 &= \sum_{k=1}^K \left(\mathbb{E} \left[(V_{*,1}^{M^*} - V_{\pi_{k,1}}^{M_k})(x_{k,1}) \right] + \mathbb{E} \left[(V_{\pi_{k,1}}^{M_k} - V_{\pi_{k,1}}^{M^*})(x_{k,1}) \right] \right) \\
 &= \sum_{k=1}^K \mathbb{E} \left[(V_{*,1}^{M^*} - V_{\pi_{k,1}}^{M_k})(x_{k,1}) \right] + \sum_{k=1}^K \mathbb{E}[\tilde{\Delta}_k], \tag{2}
 \end{aligned}$$

where we have defined $\tilde{\Delta}_k \stackrel{\text{def}}{=} (V_{\pi_{k,1}}^{M_k} - V_{\pi_{k,1}}^{M^*})(x_{k,1})$.

Lemma 3. $\mathbb{E}[V_{*,1}^{M^*}(x_{k,1})] = \mathbb{E}[V_{\pi_{k,1}}^{M_k}(x_{k,1})]$.

Proof. Just note that

$$\begin{aligned}
 \mathbb{E}[V_{*,1}^{M^*}(x_{k,1})] &= \mathbb{E}[\mathbb{E}[V_{*,1}^{M^*}(x_{k,1}) \mid \mathcal{H}_k]] \\
 &= \mathbb{E}[\mathbb{E}[V_{\pi_{k,1}}^{M_k}(x_{k,1}) \mid \mathcal{H}_k]] \\
 &= \mathbb{E}[V_{\pi_{k,1}}^{M_k}(x_{k,1})].
 \end{aligned}$$

□

Applying Lemma 3 in (2), we obtain

$$(1) = \sum_{k=1}^K \mathbb{E}[\tilde{\Delta}_k].$$

Next we focus on bounding $\tilde{\Delta}_k$. Note that

$$\begin{aligned}
 \mathbb{E}[\tilde{\Delta}_k \mid M^*, M_k] &= \mathbb{E}[(V_{\pi_{k,1}}^{M_k} - V_{\pi_{k,1}}^{M^*})(x_{k,1}) \mid M^*, M_k] \\
 &= \mathbb{E}[(\theta_k V_{\pi_{k,2}}^{M_k})(x_{k,1}, a_{k,1}) - (\theta^* V_{\pi_{k,2}}^{M^*})(x_{k,1}, a_{k,1}) \mid M^*, M_k] \\
 &= \mathbb{E}[(\theta_k - \theta^*) V_{\pi_{k,2}}^{M_k}(x_{k,1}, a_{k,1}) + (V_{\pi_{k,2}}^{M_k} - V_{\pi_{k,2}}^{M^*})(x_{k,2}) \mid M^*, M_k] \\
 &\quad + \mathbb{E}[(\theta^* (V_{\pi_{k,2}}^{M_k} - V_{\pi_{k,2}}^{M^*}))(x_{k,1}, a_{k,1}) - (V_{\pi_{k,2}}^{M_k} - V_{\pi_{k,2}}^{M^*})(x_{k,2}) \mid M^*, M_k] \\
 &= \mathbb{E}[(\theta_k - \theta^*) V_{\pi_{k,2}}^{M_k}(x_{k,1}, a_{k,1}) + (V_{\pi_{k,2}}^{M_k} - V_{\pi_{k,2}}^{M^*})(x_{k,2}) \mid M^*, M_k],
 \end{aligned}$$

where in the second last inequality we have used

$$\mathbb{E}[(\theta^*(V_{\pi_k,2}^{M_k} - V_{\pi_k,2}^{M^*}))(x_{k,1}, a_{k,1}) - (V_{\pi_k,2}^{M_k} - V_{\pi_k,2}^{M^*})(x_{k,2}) \mid M^*, M_k] = 0.$$

By recursion and law of total expectation, we derive

$$\begin{aligned} \mathbb{E}[\tilde{\Delta}_k] &= \sum_{t=1}^H \mathbb{E}[(\theta_k - \theta^*)V_{\pi_k,t+1}^{M_k}(x_{k,t}, a_{k,t})] \\ &\leq \sum_{t=1}^H \mathbb{E} \left[\|(\theta_k - \theta^*)(x_{k,t}, a_{k,t})\|_1 \cdot \|V_{\pi_k,t+1}^{M_k}\|_\infty \right] \\ &\leq H \cdot \sum_{t=1}^H \mathbb{E}[\|(\theta_k - \theta^*)(x_{k,t}, a_{k,t})\|_1], \end{aligned} \quad (3)$$

where in the first inequality we have used Hölder's inequality and in the last inequality we apply $\|V_{\pi_k,t+1}^{M_k}\|_\infty \leq H$.

Let $\bar{\theta}_k(\cdot \mid s, a)$ be the empirical transition probability before the k th episode. We define \mathcal{M}_k as the set of models such that its transition probability θ satisfies $|\bar{\theta}_k(\cdot \mid s, a) - \theta(\cdot \mid s, a)| \leq C\sqrt{\frac{S \ln(SAT)}{1 \vee N_k(s, a)}}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ where C is a universal constant which will be defined later. According to Theorem 4, we know that there exists a constant $C > 0$ such that $\Pr(M_k \notin \mathcal{M}_k) \leq 1/K$ and $\Pr(M^* \notin \mathcal{M}_k) \leq 1/K$. Hence

$$\begin{aligned} (1) &= \sum_{k=1}^K \mathbb{E}[\tilde{\Delta}_k] \\ &\leq \sum_{k=1}^K \mathbb{E}[\tilde{\Delta}_k \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k)] + H \cdot \sum_{k=1}^K (\Pr(M_k \notin \mathcal{M}_k) + \Pr(M^* \notin \mathcal{M}_k)), \end{aligned}$$

according to $\tilde{\Delta}_k \leq H$ and a union bound. Recall $\Pr(M_k \notin \mathcal{M}_k) \leq 1/K$ and $\Pr(M^* \notin \mathcal{M}_k) \leq 1/K$, we further obtain

$$\begin{aligned} (1) &\lesssim \sum_{k=1}^K \mathbb{E}[\tilde{\Delta}_k \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k)] \\ &= \sum_{k=1}^K \mathbb{E}[\tilde{\Delta}_k \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k)] \end{aligned} \quad (4)$$

Putting (3) back into (4), we have

$$\begin{aligned} (1) &\lesssim \sum_{k=1}^K \mathbb{E} \left[\sum_{t=1}^H ((\theta_k - \theta^*)V_{\pi_k,t+1}^{M_k})(x_{k,t}, a_{k,t}) \cdot \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k) \right] \\ &\leq \sum_{k=1}^K \mathbb{E} \left[\sum_{h=1}^H CH \sqrt{\frac{S \ln(SAT)}{1 \vee N_k(x_{k,h}, a_{k,h})}} \cdot \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k) \right] \\ &= \mathbb{E} \left[CH \cdot \sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{S \ln(SAT)}{1 \vee N_k(x_{k,h}, a_{k,h})}} \cdot \mathbb{1}(M_k \in \mathcal{M}_k, M^* \in \mathcal{M}_k) \right], \end{aligned} \quad (5)$$

where in the second last inequality we have used $|\bar{\theta}_k(\cdot | s, a) - \theta_k(\cdot | s, a)| \leq C\sqrt{\frac{S \ln(SAT)}{1 \vee N_k(s, a)}}$ and $|\bar{\theta}_k(\cdot | s, a) - \theta^*(\cdot | s, a)| \leq C\sqrt{\frac{S \ln(SAT)}{1 \vee N_k(s, a)}}$ when $M_k \in \mathcal{M}_k$ and $M^* \in \mathcal{M}_k$.

Note that

$$\begin{aligned} \sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{S \ln(SAT)}{1 \vee N_k(x_{k,h}, a_{k,h})}} &\lesssim \sqrt{S \ln(SAT)} \cdot \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sum_{t=0}^{N_K(s,a)} \sqrt{\frac{1}{1 \vee t}} \\ &\leq \sqrt{S \ln(SAT)} \cdot \left(\sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} 2\sqrt{N_K(s,a)} + SA \right) \\ &\leq 2S\sqrt{AT \ln(SAT)} = \tilde{\mathcal{O}}(S\sqrt{AT}), \end{aligned} \tag{6}$$

where in the third last inequality we have used $\sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sqrt{N_K(s,a)} \leq \sqrt{SAT}$ which is due to Cauchy-Schwarz inequality and $T \geq \sqrt{SA}$.

Putting (6) back into (5), we prove this theorem. \square

5 Tools

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

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

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] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011, 2013.
- [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.