Online Restless Bandits with Unobserved States

Bowen Jiang[1], Bo Jiang[1], Jian Li [2], Tao Lin[3], Xinbing Wang[1], Chenghu Zhou[4]

[1]Shanghai Jiao Tong University, [2]SUNY-Binghamton University, [3]Communication University of China, [4]Chinese Academy of Science

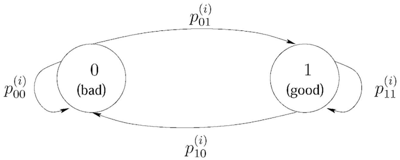
ICML, July 2023

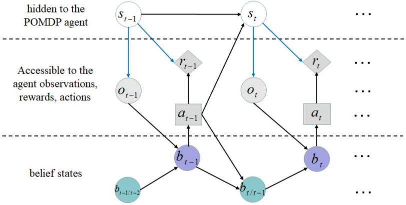
Background

- The restless multi-armed bandits (RMAB) is a general setup to model sequential decision making problems ranging from wireless communication, healthcare, etc.
- Genaral Setting:
  - considers one agent and N arms
  - each arm i is modulated by a Markov chain M^i with state transition function P^i and reward function R^i.
  - at each time, the agent decides which arm to pull
  - after the pulling, all arms make a state transition independently
  - the agent receives a reward r_t
- Goal: maximize the expected reward, i.e., $\mathbb{E}[\sum_{t=1}^{T} r_t]$, where r_t is the reward at time t and T is the time horizon.

Motivation

- The system parameters (transition functions and reward functions) are often unknown in advance
- The arms' states are often unobserved in real world application





- Challenges
  - 1. How to estimate the unknown parameters with unobserved states and control the estimation errors?
  - 2. How to design the algorithm which control the total regret well?
  - 3. How about the theoretical analysis without the observed state-action pairs?

Problem Formulation

- consider **unknown** P^i, R^i and **unobserved states** s_t^i for $\forall i, t$.
- Goal: Minimize the Bayesian regret of policy π as follows,

$$R_T := \mathbb{E}_{\theta^* \sim Q}\left[\sum_{t=1}^{T}(J(\theta^*) - r_t)\right], \tag{1}$$

where $J(\theta^*)$ is the optimal reward under the setting with unknown states and known parameters, Q is the prior distribution, θ^* is the true parameters(including P^i, R^i for all arms)

- existing results:
frequentist regret $\tilde{\mathcal{O}}(T^{2/3})$ (Zhou et al,2021) and Bayesian regret $\tilde{\mathcal{O}}(T^{2/3})$ (Jahromi et al, 2022)

Our contribution

- Assumptions:
  - 1. The smallest element ϵ_1 in the transition functions $P^i, i \in N$ is larger than zero.
  - 2. The smallest element ϵ_2 in the reward functions $R^i, i \in N$ is larger than zero.
- Our contribution:
  - Our solution to Challenge 1:
    update the posterior distribution as mixture of Dirichlet distributions
  - Our solution to Challenge 2:
    conduct the explore-then-commit learning in an episodic way and operate in episodes with increasing length
  - Our solution to Challenge 3:
    define the pseudo-count about the number of visits to state-action based on Dirichlet distribution and prove the first Bayesian regret $\tilde{\mathcal{O}}(\sqrt{T})$

Mixture of Dirichlet Distribution

- all state sequences (and their corresponding Dirichlet posteriors) should be considered, with some weight proportional to the likelihood of each state sequence
- assume that $P^i = \epsilon_1 \mathbf{1} + (1 - S\epsilon_1)\tilde{P}^i$, $R^i = \epsilon_2 \mathbf{1} + (1 - S\epsilon_2)\tilde{R}^i$, where $\tilde{P}^i, \tilde{R}^i$ follows the Dirichlet distribution and $\mathbf{1}$ is the vector with one in each position.

Algorithm 1 Posterior Update for $R^i(s, \cdot)$ and $P^i(s, \cdot)$

Input: the history length τ_1, the state space $\mathcal{S}$, the belief history $b^i_{0:\tau_1}$, the reward history $r^i_{0:\tau_1}$, the initial parameters $\phi^i_{s,s'}, \psi^i_{s,r}$, for $s, s' \in \mathcal{S}, r \in \mathcal{R}$,
generate $\mathcal{S}^{\tau_1}$ possible state sequences
calculate the weight $w(j) = \prod_{t=0}^{\tau_1 - 1} b^i_t(s, \theta), j \in \mathcal{S}^{\tau_1}$
for j in $1, \ldots, \mathcal{S}^{\tau_1}$ **do**
    count the occurence times of event (s, s') and (s, r) as $N^i_{s,s'}, N^i_{s,r}$ in sequence j
    update $\phi^i_{s,s'} \leftarrow \phi^i_{s,s'} + N^i_{s,s'}$, $\psi^i_{s,r} \leftarrow \psi^i_{s,r} + N^i_{s,r}$
    aggregate the $\phi^i_{s,s'}$ as $\phi(j)$, $\psi^i_{s,r}$ as $\psi(j)$ for all $s, s' \in \mathcal{S}, r \in \mathcal{R}$
end for
update the mixture Dirichlet distribution
$g_{\tau_1}(P^i) \propto \sum_{j=1}^{\mathcal{S}^{\tau_1}} w(j) f(\frac{P^i - \epsilon_1 \mathbf{1}}{1 - S\epsilon_1} \mid \phi(j))$,
$g_{\tau_1}(R^i) \propto \sum_{j=1}^{\mathcal{S}^{\tau_1}} w(j) f(\frac{R^i - \epsilon_2 \mathbf{1}}{1 - S\epsilon_2} \mid \psi(j))$

Figure 1: Posterior Update for $R^i(s, \cdot)$ and $P^i(s, \cdot)$

TSEETC-exploration phase

- TSEETC operates in episodes with different lengths
- In episode k, for the exploration phase

- Step 1: initialize R_{t_k}, P_{t_k}
- Step 2: pull each arm for τ_1/N times in a round-robin way
- Step 3: update the arm's belief state with pulled or not
- Then the reward and belief history of each arm are input into Algorithm 1 to update the posterior distribution

Algorithm 2 Thompson Sampling with Episodic Explore-Then-Commit

Input: prior $g_0(P)$,$g_0(R)$, initial belief b_0, exploration length τ_1, the first episode length T_1
for episode $k = 1, 2, \ldots,$ **do**
  start the first time of episode k, $t_k := t$
  sample $R_{t_k} \sim g_{t_{k-1}+\tau_1}(R)$ and $P_{t_k} \sim g_{t_{k-1}+\tau_1}(P)$
  for $t = t_k, t_k + 1, ..., t_k + \tau_1$ **do**
    pull the arm i for τ_1/N times in a round robin way
    receive the reward r_t
    update the belief b_t^i using R_{t_k}, P_{t_k} according to (4)
    update the belief $b_t^j, j \in N \setminus \{i\}$ using P_{t_k} according to (5)
  end for
  for $i = 1, 2, \ldots, N$ **do**
    input the obtained $r_{t_1:t_1+\tau_1}, ..., r_{t_k:t_k+\tau_1}$, $b_{t_1:t_1+\tau_1}, ..., b_{t_k:t_k+\tau_1}$ to Algorithm 1 to update the posterior distribution $g_{t_k+\tau_1}(P)$, $g_{t_k+\tau_1}(R)$
  end for

Figure 2: Thompson Sampling with Episodic Explore-Then-Commit

TSEETC-exploitation phase

- Then we sample the new $R_{t_k+\tau_1}$, $P_{t_k+\tau_1}$ from the posterior distribution
- re-calibrate the belief b_t based on the most recent sampled $R_{t_k+\tau_1}$, $P_{t_k+\tau_1}$.
- Next, we enter into the exploitation phase (line 18-23)
  - Step 1: use an Oracle to derive the optimal policy π_k for the sampled parameters $R_{t_k+\tau_1}$, $P_{t_k+\tau_1}$
  - Step 2: use policy π_k for the rest of the episode k

```
14:   sample R_{t_k+τ_1} ~ g_{t_k+τ_1}(P), P_{t_k+τ_1} ~ g_{t_k+τ_1}(R)
15:   for i in 0, 1, ..., N do
16:     re-update the belief b_t^i from time 0 to t_k + τ_1 according to R_{t_k+τ_1} and P_{t_k+τ_1}
17:   end for
18:   compute π_k^*(·) = Oracle (·, R_{t_k+τ_1}, P_{t_k+τ_1})
19:   for t = t_k + τ_1 + 1, ··· , t_{k+1} − 1 do
20:     apply action a_t = π_k^* (b_t)
21:     observe new reward r_{t+1}
22:     update the belief b_t of all arms using (4), (5)
23:   end for
24: end for
```

Figure 3: Thompson Sampling with Episodic Explore-Then-Commit

Performance Analysis

Theorem

Suppose Assumptions 1,2 hold and the Oracle returns the optimal policy in each episode. The Bayesian regret of our algorithm satisfies

$$R_T \leq 48 C_1 C_2 S \sqrt{NT \log(NT)} + C_1 C_2 + (\tau_1 \Delta R + H + 4 C_1 C_2 SN)\sqrt{T},$$

where C_1, C_2, L_1, L_2 are constants independent with T, τ_1 is the fixed exploration length in each episode, ΔR is is the gap between the maximum and the minimum rewards, H is the bounded span, r_{max} is the maximum reward obtain each time.

- first to achieves the $\tilde{O}(\sqrt{T})$ Bayesian regret bound on average

Numerical Experiments

- Basic setting: two arms with two hidden states (0 and 1), the reward set $\{10, 20\}$ at state 1, the reward set $\{-10, 10\}$ at state 0, learning horizon $T = 50000$
- Baselines: ϵ-greedy, Sliding-Window UCB , RUCB, Q-learning, SEEU

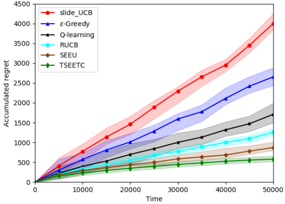


Figure 4: The cumulative regret

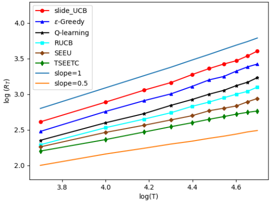


Figure 5: The log-log regret

- TSEETC has the minimum regret among these algorithms.
- the slope of TSEETC is close to 0.5, which is consistent with our theoretical result

Conclusion

- consider restless bandits with unknown states and unknown dynamics.
- propose the TSEETC algorithm to estimate these unknown parameters and derive the optimal policy
- establish the Bayesian regret of our algorithm as $\tilde{\mathcal{O}}(\sqrt{T})$.
- future work
  - consider the setting where the transition functions are action dependent
  - discuss the impact of approximation errors on the posterior distribution in relation to the regret bound

Thank you!