Optimality of Myopic Policy for a Class of Monotone Affine RMAB

Parisa Mansourifard, Tara Javidi, Bhaskar Krishnamachari

Introduction

- We consider a class of Restless Multi-Armed Bandits with n independent and stochastically identical arms.
- Only one arm can be played at each time (easy to generalize to more than one arm).
- Each arm is in a real-valued state: \( s \in [s_0, s_{\text{max}}] \).
- Selecting an arm with state \( s \) yields an immediate reward with expectation \( R(s) \).
- The state of selected arm stochastically jumps from its current value \( s \) to \( s_{\text{max}} \) or \( s_0 \), with probability \( p(s) \) or \( 1 - p(s) \), respectively.
- The state of not-played arms evolve according to a function \( \tau(s) \).
- Finite horizon \( T \), time steps \( t = 1, ..., T \).
- the state of arm \( j \) at time \( t \), \( s_j(t) \).
- State transition of arm \( j \) upon playing arm \( a \): \( s_j(t+1) = \begin{cases} s_{\text{max}}, & \text{w.p. } p(s_j(t)), \text{iff } a = a_j \\ s_j, & \text{w.p. } 1 - p(s_j(t)), \text{iff } a = a_j \\ \tau(s_j(t)), & \text{w.p. } 1, \text{iff } a \neq a_j \end{cases} \)

Problem

- Policy vector: \( \pi = [\pi(1), ..., \pi(T)] \).
- Selecting arm \( a \) to play: \( \pi(t) = a \in \{1, ..., n\} \).
- Maximizing total discounted expected reward:
  \[ \max_{\pi} E^{\pi} \left[ \sum_{t=1}^{T} \beta^{t-1} R(s_{\pi(t)}(t)) \right] | \tilde{s}(1) = \bar{s} \]
  - defining value function, i.e. maximum expected remaining reward starting from time \( t \): \( V_t(\bar{s}) \)

Recursive Equations (DP)

\[ V_t(\bar{s}) = \max_{a=1, ..., n} V_{a,t}(\bar{s}), \quad \forall t = 1, ..., T \]
\[ V_{a,t}(\bar{s}) = R(s_a), \]
\[ V_{a,t}(\bar{s}) = R(s_a) + \beta p(s_a) V_{a+1,t}(\tau(s_{a+1}) - s_{a+1}, \tau(s_{a+1})) + \beta (1 - p(s_a)) V_{a+1,t}(\tau(s_{a+1}) - s_{a+1}, s_{a+1}, \tau(s_{a+1})), \quad \forall t = 1, ..., T - 1 \]

Goal

We prove that under some conditions, the simple myopic policy, which selects at each time the arm with the highest immediate reward, is optimal.

Evolution of states

Myopic Policy

- ignoring the impact of the current action on the future reward, myopic policy is given by
  \[ \pi^*(\bar{s}) = \arg \max_{a \in \{1, ..., n\}} R(s_a) = a_\tau + b_\tau \cdot \bar{s} \]
  - Which selects an arm with the highest state at each time

Conditions:

- **C1**: \( \text{monotonically increasing and affine functions of state } s \), \( R(s), p(s), \tau(s) = a_\tau + b_\tau \cdot s \)
- **C3**: \( \tau(s) \) is a contraction mapping \( |\tau(s_1) - \tau(s_2)| \leq |s_1 - s_2| \implies b_\tau \leq 1 \)

Theorem

Under C1-C3, and \( b_\tau \leq \frac{1}{\beta (1 + \beta (p_{\text{max}} - p_0))} \),

- the myopic policy is optimal, i.e.,
  - \( V_{a,t}(\bar{s}) \geq V_{a,t}(\bar{s}), \forall t = 1, ..., T \)
  - if \( s_i \geq s_j, i = 2, ..., n \)

Future works

- Generalizing to non-identical arms, non-affine evolution, or multi-dimensional states
- Identifying conditions for related problems where myopic policy is not optimal, but other efficient, possibly index-based, policy is optimal

parisama@usc.edu, tjavidi@ucsd.edu, bkrishna@usc.edu