Ming Hsieh Department of Electrical Engineering

USC Viterbi School of Engineering

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: $s \in [s_0, s_{\max}]$
- 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 $s_{\max}$ 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:

$$s_j(t+1) = \begin{cases} s_{\max}, & \text{w.p. } p(s_j(t)), \text{ if } j = a \\ s_0, & \text{w.p. } p(s_j(t)), \text{ if } j = a \\ \tau(s_j(t)), & \text{w.p. } 1, \text{ if } j \neq a \end{cases}$$

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.

Problem

- Policy vector: $\pi = [\pi(1),...,\pi(T)]$
- Selecting arm a to play: $\pi(t) = a \in \{1,...,n\}$
- Maximizing total discounted expected reward:

$$\max_{\pi} E^{\pi}[\sum_{t=1}^{T} \beta^{t-1} R(s_{\pi(t)}(t)) \mid \bar{s}(1) = \bar{s}]$$

- defining value function, i.e. maximum expected remaining reward starting from time t: $V_t(\bar{s})$

Recursive Equations (DP)

$$V_t(\bar{s}) = \max_{a=1,...,n} V_{a,t}(\bar{s}), \quad \forall t = 1,...,T$$

$$V_{a,T}(\bar{s}) = R(s_a),$$

$$V_{a,t}(\bar{s}) = R(s_a) + \beta p(s_a) V_{t+1}(\tau(\bar{s}_1^{a-1}), s_{\max}, \tau(\bar{s}_{a+1}^{n})) +$$

$$\beta(1-p(s_a)) V_{t+1}(\tau(\bar{s}_1^{a-1}), s_0, \tau(\bar{s}_{a+1}^{n})), \quad \forall t = 1,...,T-1$$

Myopic Policy

ignoring the impact of the current action on the future reward, myopic policy is given by

$$\pi^*(\bar{s}) = \arg\max_{a=1,...,n} R(s_a) = \arg\max_{a=1,...,n} s_a$$

Which selects an arm with the highest state at each time

Conditions:

C1-2: monotonically increasing and affine functions of state s, $R(s),\ p(s),\ \tau(s)$

C3: $\tau(s)$ is a contraction mapping

$$|\tau(s_1) - \tau(s_2)| \leq |s_1 - s_2|$$

Theorem

Under C1-C3, the myopic policy is optimal, i.e.,

$$V_{1,t}(\bar{s}) \geq V_{i,t}(\bar{s}), \forall t = 1,...,T$$

$$\text{if } s_1 \geq s_2,\ i = 2,...,n$$

Proof: using following lemmas, ….

Lemmas:

L1: Symmetry of $V_{a,t}(\bar{s})$

L2: Affine linearity

$$\lambda V_{a,t}(\bar{s}_1^{i-1}, s, \bar{s}_{i+1}^{n}) + (1-\lambda) V_{a,t}(\bar{s}_1^{i-1}, s', \bar{s}_{i+1}^{n})$$

$$= V_{a,t}(\bar{s}_1^{i-1}, \lambda s + (1-\lambda) s', \bar{s}_{i+1}^{n})$$

L3:

$$V_{1,t}(\bar{s}) - V_{i,t}(\bar{s}) = (\lambda_1 - \lambda_i) \times$$

$$[V_{1,t}(U, \bar{s}_2^{i-1}, L, \bar{s}_{i+1}^{n}) - V_{1,t}(L, \bar{s}_2^{i-1}, U, \bar{s}_{i+1}^{n})]$$

$$U = \tau^{-1}(s_{\max}),\ L = \tau^{-1}(s_0)$$

L4:

$$V_t(s_1, \bar{s}_2^{n}) - V_{i,t}(s'_1, \bar{s}_2^{n}) \leq \frac{R(s_1) - R(s'_1)}{1 - \beta(p_{\max} - p_0)}$$

Future works

- Generalizing to non-identical arms, non-affine evolution, or multi-dimensional states
- Indentifying 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

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering