Logarithmic Weak Regret of Non-Bayesian
Restless Multi-Armed Bandit

Haoyang Liu, Keqin Liu, Qing Zhao

Department of Electrical and Computer Engineering
University of California, Davis, CA 95616

Supported by NSF and ARO.
**Cognitive Radio for Dynamic Spectrum Access**

**Dynamic Spectrum Access under Unknown Model:**

- $N$ independent channels.
- Choose $K$ channels to sense/access in each slot.
- Accessing an idle channel results in a unit reward.
- Channel occupancy: a stochastic process with unknown parameters.
  - i.i.d. Bernoulli with unknown mean $\theta_i$.
  - Markovian with unknown transition probabilities.
Multi-Armed Bandit

Multi-Armed Bandit:

- $N$ arms and a single player.
- select one arm to play at each time.
- Unknown reward statistics.
- Maximize the long-run reward.

Exploitation v.s. Exploration

- Exploitation: play the arm with the largest sample mean.
- Exploration: play an arm to learn its reward statistics.
i.i.d. Reward Model
i.i.d. Reward Model

Performance Measure: Regret

- $\Theta \overset{\Delta}{=} (\theta_1, \cdots, \theta_N)$: unknown reward means.

- $\theta^{(1)}T$: max total reward (by time $T$) if $\Theta$ is known (always play the best arm).

- $V_T^\pi(\Theta)$: total reward of policy $\pi$ by time $T$.

- Regret (cost of learning):

$$R_T^\pi(\Theta) \overset{\Delta}{=} \theta^{(1)}T - V_T^\pi(\Theta) = \sum_{i=2}^{N} (\theta^{(1)} - \theta^{(i)}) \mathbb{E}[\text{time spent on } \theta^{(i)}].$$

Objective: minimize the growth rate of $R_T^\pi(\Theta)$ with $T$.

\[\text{sublinear regret} \implies \text{maximum average reward } \theta^{(1)}\]
Classic Results

- Lai&Robbins’85:
  \[ R_T^*(\Theta) \sim \sum_{i=2}^{N} \frac{\theta^{(1)} - \theta^{(i)}}{I(\theta^{(i)}, \theta^{(1)})} \log T \quad \text{as } T \to \infty. \]
  
  *KL divergence*

- Agrawal’95, Auer&Cesa-Bianchi&Fischer&Informatik’02:
  - Sample-mean based index policies.
  - UCB-1: index = \( \bar{s}_i + \sqrt{\frac{2 \log t}{t_i}} \).
Rested Markovian Reward Model
Restured Markovian Reward Model

Restured Markovian Reward Model:
- Rewards from successive plays form a MC with unknown transition $P_i$.
- Arm state is frozen when not played.

Optimal Performance under Known Model:
- Best arm: the largest reward mean $\theta^{(1)}$ in steady state.
- Optimal policy: play the best arm except a finite time to exploit transient states of other arms.
- Optimal performance: $\theta^{(1)}T + O(1)$.

Regret:
\[
R_T^\pi = \sum_{i=2}^{N} (\theta^{(1)} - \theta^{(i)})\mathbb{E}[\text{time spent on arm } i] + O(1).
\]

Policies Achieving Logarithmic Regret:
- Extension of Lai-Robbins policy (Anantharam&Varaiya&Walrand’87).
- Extension of Auer’s UCB1 policy (Tekin&M.Liu’10)
Restless Multi-Armed Bandit
Restless MAB under Unknown Dynamics

Restless MAB under Unknown Dynamics:

- Rewards from successive plays form a MC with unknown transition $P_i$.
- When passive, arm evolves a.t. an arbitrary unknown random process.

Difficulty:

- Restless MAB under known model itself is intractable in general.
- The optimal policy under known model is no longer staying on one arm.

Weak Regret:

- Defined with respect to the optimal single-arm policy under known model:
  \[ R_T^\pi = T\theta^{(1)} - V_T^\pi + O(1). \]
- The optimal performance under known model $\geq T\theta^{(1)}$. 
Restless MAB under Unknown Dynamics

Challenges:

- Need to learn $\{\theta_i\}$ from contiguous segments of the sample path.
- Need to limit arm switching to bound the transient effect.
Restless UCB (RUCB):

- Epoch structure with geometrically growing epoch length
  $\Rightarrow$ arm switching limited to $\log$ order.

- Exploration and exploitation epochs interleaving for fast error decay:
  - In exploration epochs, play all arms in turn.
  - In exploitation epochs, play the arm with the largest index $\bar{s}_i + \sqrt{\frac{L \log t}{t_i}}$.
  - Start an exploration epoch iff total exploration time $< D \log t$.
The Logarithmic Regret of RUCB

Logarithmic regret of RUCB:

- Uniformly bounded leading constant determined by $D$ and $L$.
- Choosing $D$ and $L$ requires
  - an arbitrary (nontrivial) lower bound on the eigenvalue gaps of $P_i$.
  - an arbitrary (nontrivial) lower bound on $\theta^{(1)} - \theta^{(2)}$.

Near logarithmic regret in the absence of system knowledge:

- For any increasing sequence $f(t)$,
  $$R_{RUCB}(t) \sim O(f(t) \log t)$$
- by choosing $D(t)$ and $L(t)$ as increasing sequences satisfying
  $$D(t) = f(t), \quad \frac{L(t)}{D(t)} \to 0.$$
Conclusion

Restless MAB under Unknown Dynamics:
- Reward from successive plays forms a MC with unknown transition $P_i$.
- When passive, arm state evolves according to an arbitrary unknown random process.

Restless UCB Policy:
- Logarithmic regret with arbitrary (nontrivial) bounds on certain parameters.
- Arbitrarily close to log regret in the absence of any knowledge.

Related Work on Single-Player Restless MAB:
- Tekin&Liu’11: logarithmic weak regret.
- Dai&Gai&Krishnamachari&Zhao’11: near-log strict regret (a special RMAB).
- Extension to decentralized RMAB with multiple players (Liu&Liu&Zhao’11)
  - Distributed decision-making using only local observations.
  - Players activating the same arm collide with reward penalty.
  - Exogenous and endogenous restless models.