The Non-Bayesian Restless Multi-Armed Bandit: A Case of Near-Logarithmic Regret

Wenhan Dai[†], *Yi Gai*[‡], *Bhaskar Krishnamachari*[‡], *Qing Zhao*[§]

[†] Tsinghua University, Beijing, China

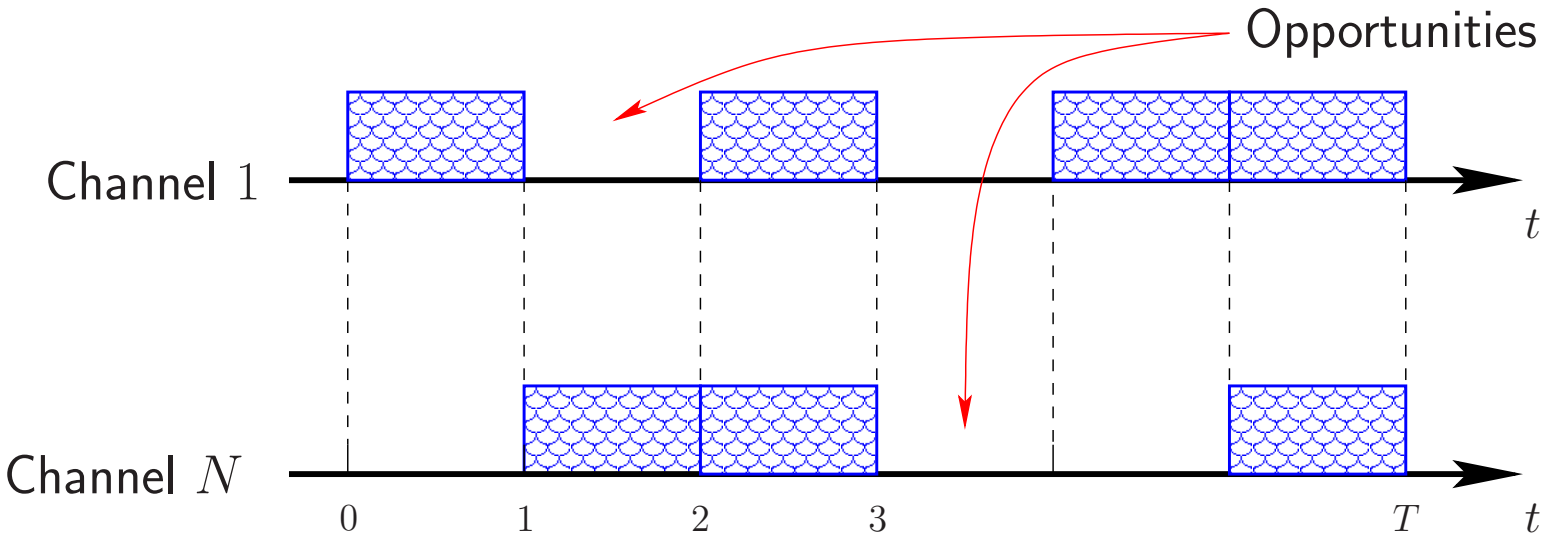
[‡] University of Southern California

[§] University of California at Davis

Supported by ARL and ARO.

Cognitive Radio for Dynamic Spectrum Access

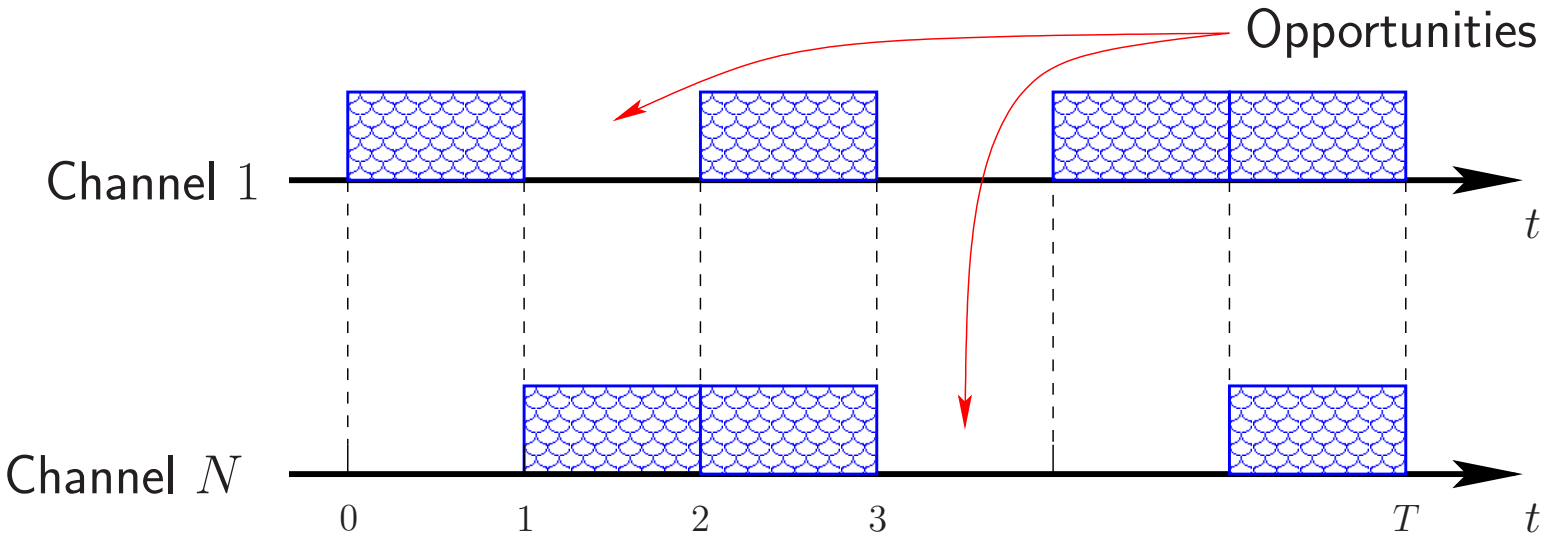
Dynamic Spectrum Access under Unknown Model:



- 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. Model

Dynamic Spectrum Access under Unknown Model:



- Primary occupancy of channel i: i.i.d. Bernoulli with unknown mean θ_i:

$$\theta_i = \Pr[\text{channel } i \text{ is idle}]$$

- Objective: a channel selection policy to achieve the max throughput $\theta^{(1)}$:

$$\theta^{(1)} = \max\{\theta_1, \cdots, \theta_N\}$$

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

Performance Measure: Regret

- $\Theta \triangleq (\theta_1, \cdots, \theta_N)$: unknown reward means.
- $\theta^{(1)}T$: max total reward (by time T) if Θ is known.
- $V_T^\pi(\Theta)$: total reward of policy π by time T.
- Regret (cost of learning):

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

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

sublinear regret $\Longrightarrow$ maximum average reward $\theta^{(1)}$

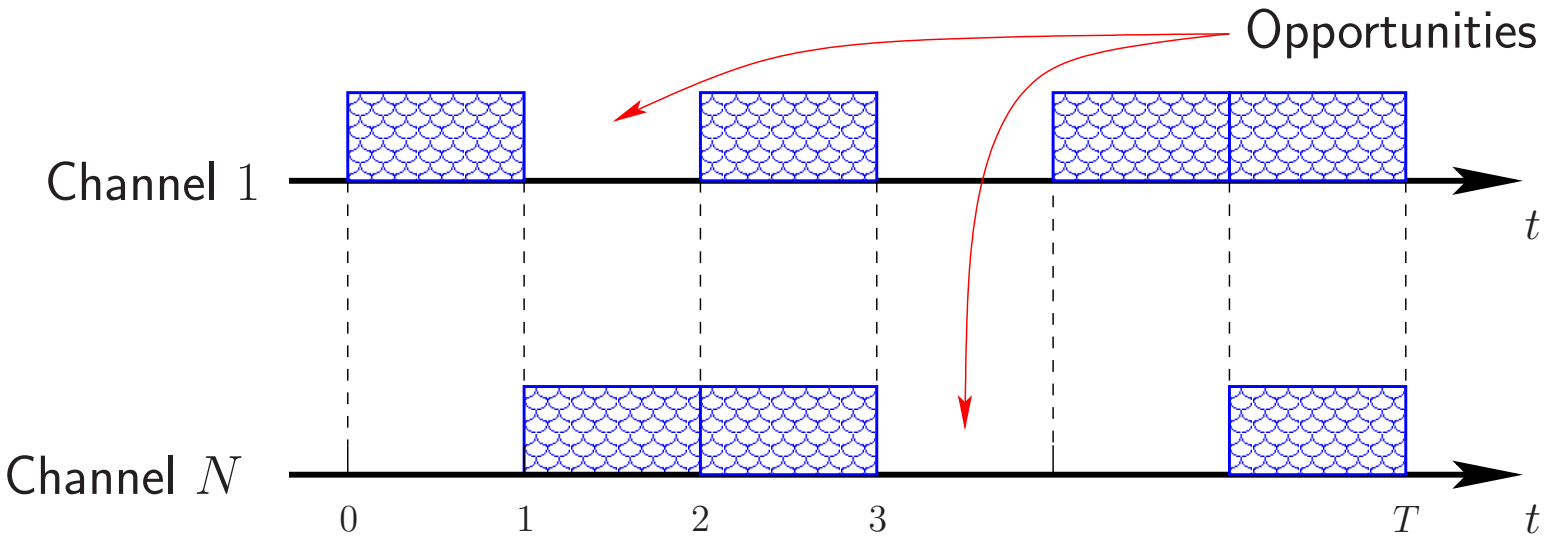
Classic Results

- Lai&Robbins'85:

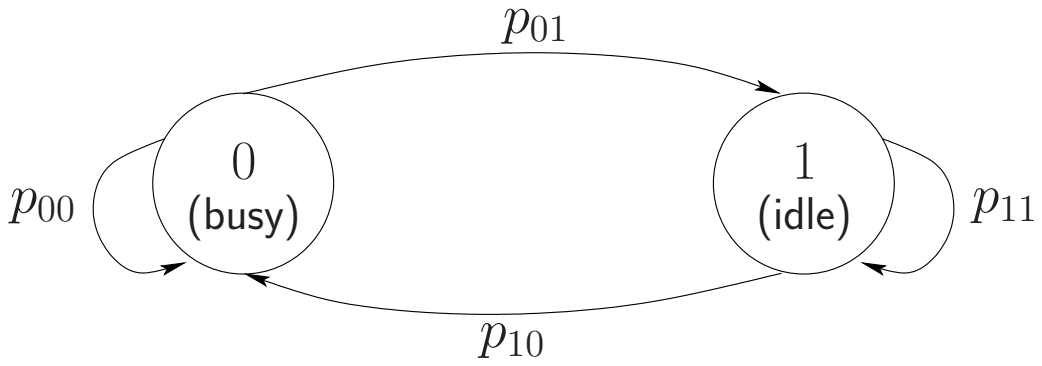
$$R_T^*(\Theta) \sim \sum_{i=2}^{N} \frac{\theta^{(1)} - \theta^{(i)}}{\underbrace{I(\theta^{(i)}, \theta^{(1)})}_{\text{KL divergence}}} \log T \quad \text{as } T \to \infty.$$

- Agrawal'95, Auer&Cesa-Bianchi&Fischer&Informatik'02:
  - Sample-mean based index policies.
  - UCB-1: index $= \bar{s}_i + \sqrt{\frac{2 \log t}{t_i}}$.

Markovian Model



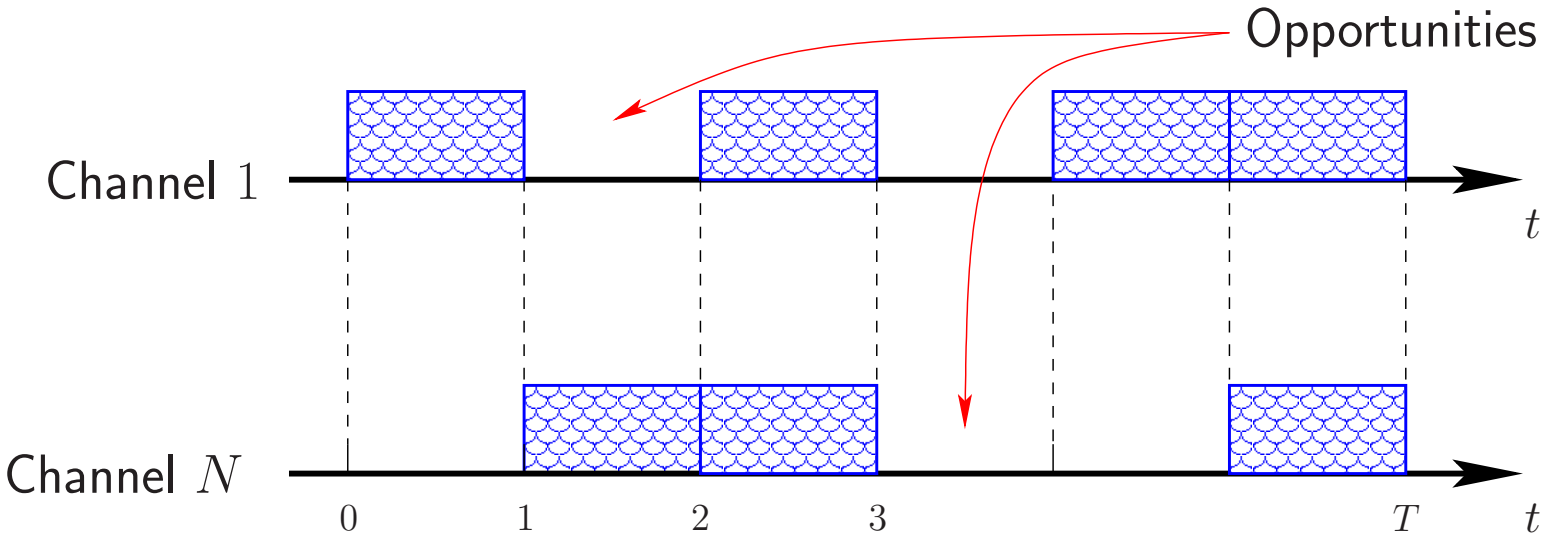
- Channel occupancy: Markovian with unknown transition probabilities:



- Objective: a channel selection policy to achieve max throughput.

Markovian Model

Dynamic Spectrum Access under Unknown Model:



- Challenges:
  - The optimal policy under known model is not staying on one channel.
  - Need to learn the best way to switch among channels based on observations (infinite possibilities).

Optimal Policy under Known Model

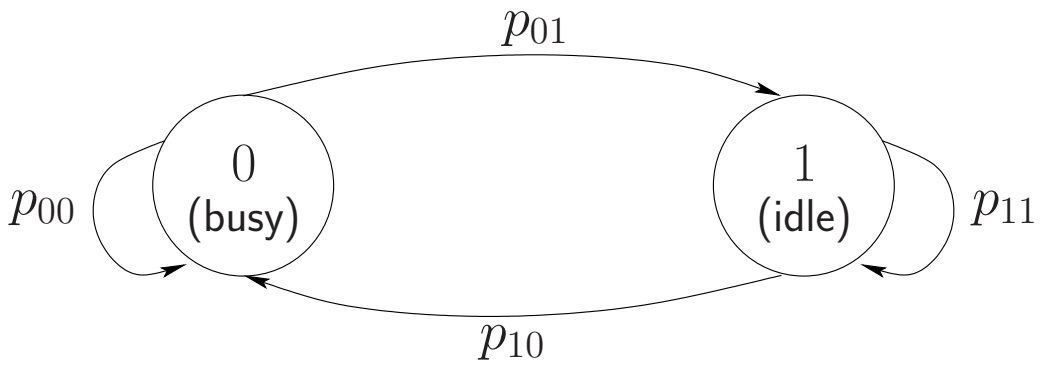
Restless Multi-Armed Bandit:

- Formulated by Whittle in 1988.
- PSPACE-hard in general *(Papadimitriou&Tsitsiklis:99)*.

Optimality of the Myopic Policy:

(Zhao&Krishnamachari:07,Ahmad&Liu&Javidi&Zhao&Krishnamachari:09)

- When $p_{11} \geq p_{01}$, holds for all N and K;
- When $p_{11} < p_{01}$, holds for $N = 2, 3$ or $K = N - 1$ (conjectured for all N).



Optimal Policy under Known Model

Semi-Universal Structure of the Myopic Policy: *(Zhao&Krishnamachari:07)*

- When $p_{11} \geq p_{01}$, stay at "idle" and switch at "busy" to the channel visited longest time ago.

- When $p_{11} < p_{01}$, stay at "busy" and switch at "idle" to the channel most recently visited among all channels visited an even number of slots ago or the channel visited longest time ago.

Achieving Optimal Throughput under Unknown Model

Achieving Optimal Throughput under Unknown Model:

- Treat each way of channel switching as an arm.
- Learn which arm is the good arm.

Challenges in Achieving Sublinear Regret:

- How long to play each arm: the optimal length L^* depends on the transition probabilities.
- Rewards are not i.i.d. in time or across arms.

Achieving Optimal Throughput under Unknown Model

Approach:

- Play each arm with increasing length $L_n \to \infty$ at arbitrarily slow rate.
- Modified Chernoff-Hoeffding bound to handle non-i.i.d. samples:

Assume $|E[X_i|X_1, \cdots, X_{i-1}] - \mu| \leq C$ ($0 < C < \mu$). Let $S_n = \sum_{i=1}^{n} X_i$.
Then $\forall a \geq 0$,

$$\Pr\{S_n \geq n(\mu + C) + a\} \leq e^{-2(\frac{a(\mu-C)}{b(\mu+C)})^2/n}$$
$$\Pr\{S_n \leq n(\mu - C) - a\} \leq e^{-2(a/b)^2/n}$$

Regret Order:

- Near-logarithmic regret: $G(T) \log T$

$$G(T): \quad \underbrace{L_1, \cdots, L_1}_{L_1 \text{times}}, \quad \underbrace{L_2, \cdots, L_2}_{L_2 \text{times}}, \quad \underbrace{L_3, \cdots, L_3}_{L_3 \text{times}}, \quad \underbrace{L_4, \cdots, L_4}_{L_4 \text{times}}, \cdots$$

Conclusion

Finite-Option Restless Multi-Armed Bandit with Unknown Dynamics:

- Finite-option: depending on the parameters, the optimal policy takes one of a finite possible forms.
- Treat each form as an arm and learn which arm is the best.

In the Context of Dynamic Spectrum Access:

- When channels are stochastically identical, the optimal (myopic) policy takes two possible forms.
- Near-logarithmic regret is achieved, leading to the optimal throughput defined by known models.

General Restless Multi-Armed Bandit with Unknown Dynamics:

- Logarithmic weak regret: *Liu&Liu&Zhao:11*