Multi-Armed Bandit:

Learning in Dynamic Systems with Unknown Models

Qing Zhao

Department of Electrical and Computer Engineering

University of California, Davis, CA 95616

Supported by NSF, ARO.

Multi-Armed Bandit

Multi-Armed Bandit:

- N arms and a single player.
- Select one arm to play at each time.
- i.i.d. reward with *Unknown* mean θ_i.
- Maximize the long-run reward.

Exploitation v.s. Exploration

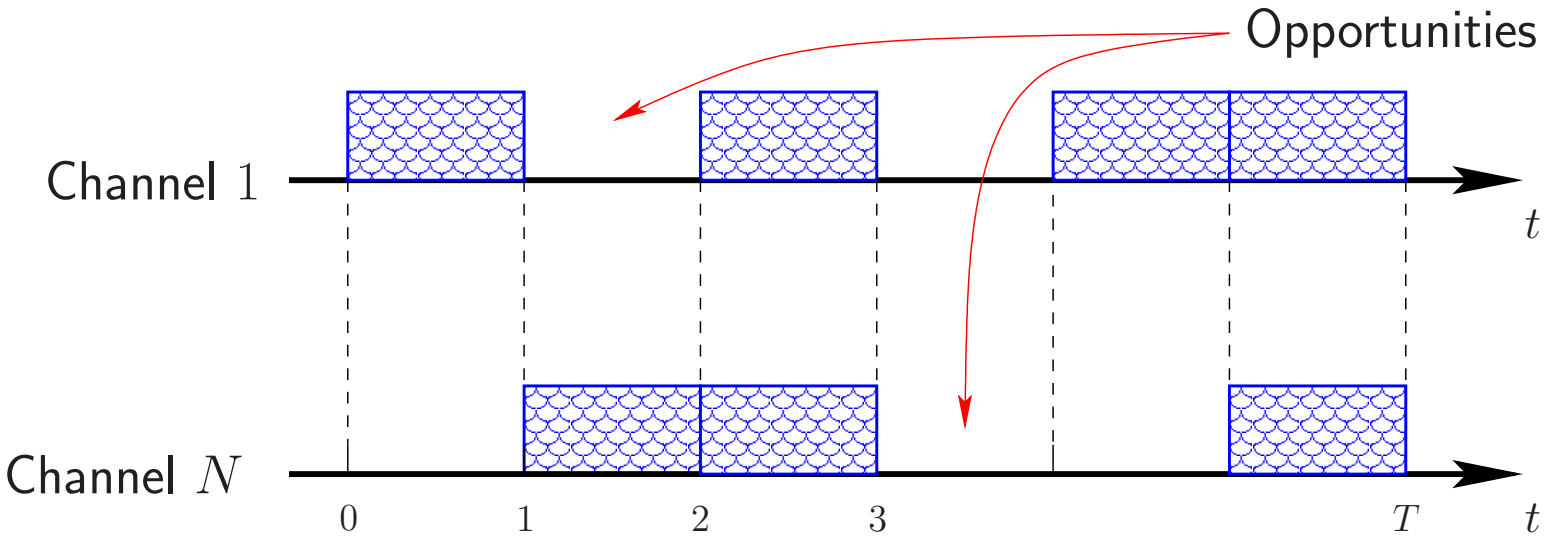
- Exploitation: play the arm with the largest sample mean.
- Exploration: play an arm to learn its reward statistics.

Clinical Trial *(Thompson'33)*

Two treatments with unknown effectiveness:

Dynamic Spectrum Access

Dynamic Spectrum Access under Unknown Model:

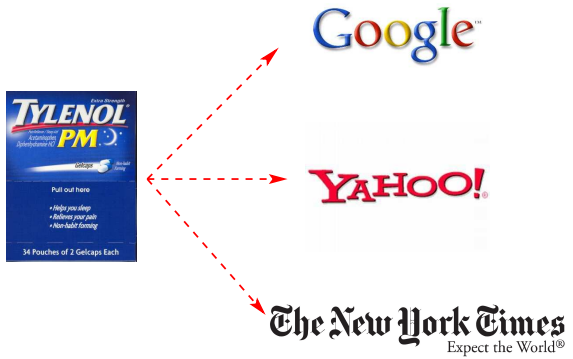


- N independent channels.
- Choose K channels to sense/access in each slot.
- Accessing an idle channel results in a unit reward.
- Channel occupancy: i.i.d. Bernoulli with unknown mean θ_i.

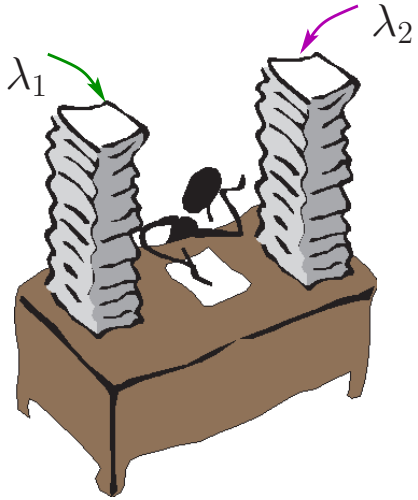
Other Applications of MAB

Web Search

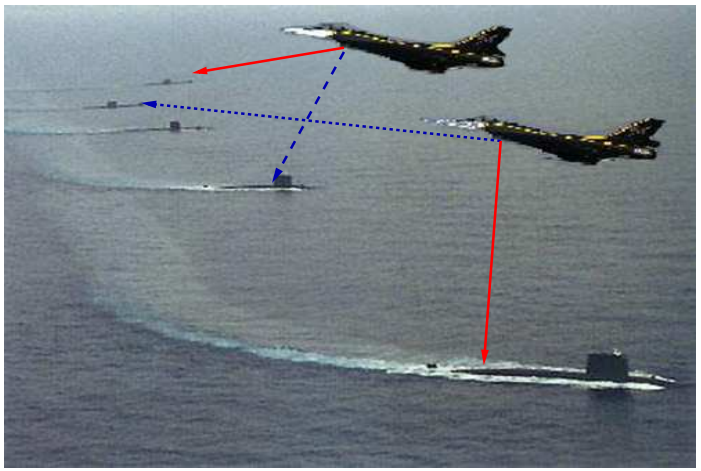
Internet Advertising/Investment



Queueing and Scheduling



Multi-Agent Systems



Non-Bayesian Formulation

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**:

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

  - Optimal policies explicitly constructed for Gaussian, Bernoulli, Poisson, and Laplacian distributions.

- **Agrawal'95**:

  - Order-optimal index policies explicitly constructed for Gaussian, Bernoulli, Poisson, Laplacian, and Exponential distributions.

- **Auer&Cesa-Bianchi&Fischer'02**:

  - Order-optimal index policies for distributions with finite support.

Classic Policies

Key Statistics:

- Sample mean $\bar{\theta}_i(t)$ (*exploitation*);
- Number of plays $\tau_i(t)$ (*exploration*);

In the classic policies:

- $\bar{\theta}_i(t)$ and $\tau_i(t)$ are combined together for arm selection at each t:

  UCB Policy(*Auer et al. :02*): index $= \bar{\theta}_i + \sqrt{\frac{2\log t}{\tau_i(t)}}$

- A fixed form difficult to adapt to different reward models.

Limitations

- **Limitations of the Classic Policies**:
  - Reward distributions limited to finite support or specific cases;
  - A single player (equivalently, centralized multiple players);
  - i.i.d. or *rested* Markov reward over successive plays of each arm.

Recent Results

- **Limitations of the Classic Policies**:
  - Reward distributions limited to finite support or specific cases;
  - A single player (equivalently, centralized multiple players);
  - i.i.d. or *rested* Markov reward over successive plays of each arm.

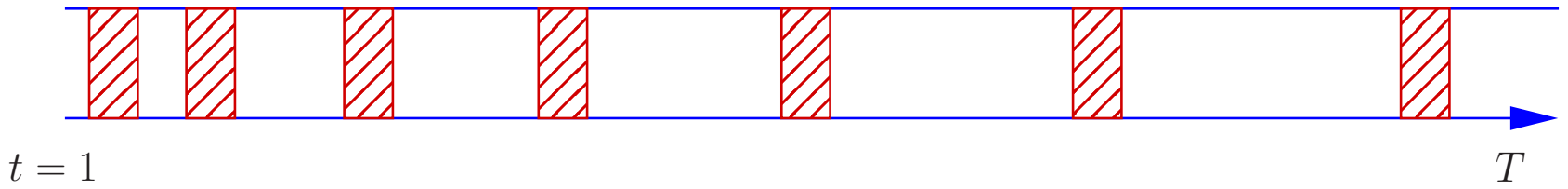
- **Recent results: policies with a tunable parameter capable of handling**
  - a more general class of reward distributions (including heavy-tailed);
  - decentralized MAB with partial reward observations;
  - *restless* Markovian reward model.

General Reward Distributions

DSEE

Deterministic Sequencing of Exploration and Exploitation (DSEE):

- Time is partitioned into interleaving exploration and exploitation sequences.



  - Exploration: play all arms in round-robin.
  - Exploitation: play the arm with the largest sample mean.

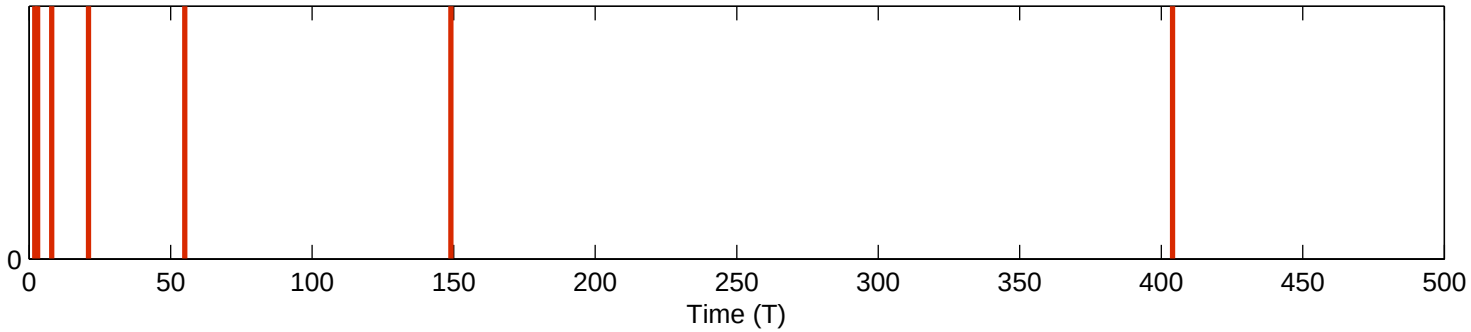
- A tunable parameter: the cardinality of the exploration sequence
  - can be adjusted according to the "hardness" of the reward distributions.

The Optimal Cardinality of Exploration

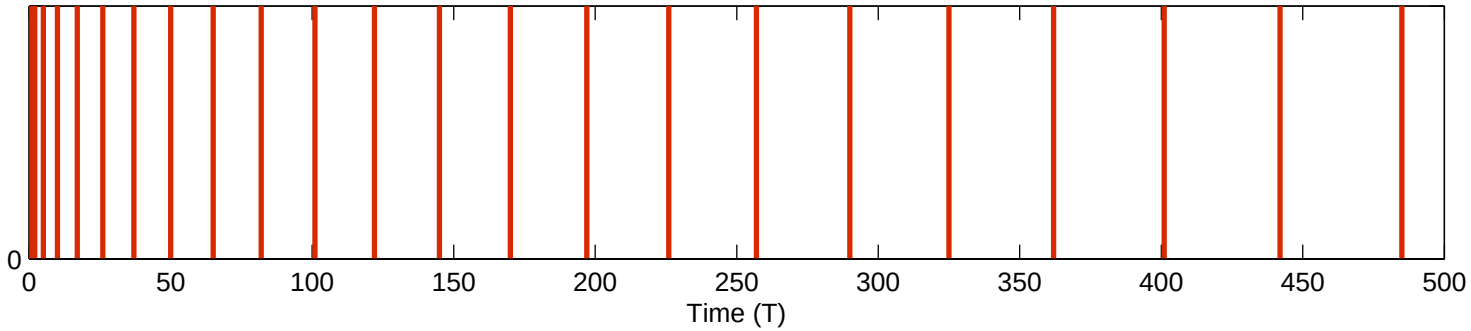
The Cardinality of Exploration:

- a lower bound of the regret order;
- should be the min x so that regret in exploitation is no larger than x.

- $O(\log T)$?



- $O(\sqrt{T})$?



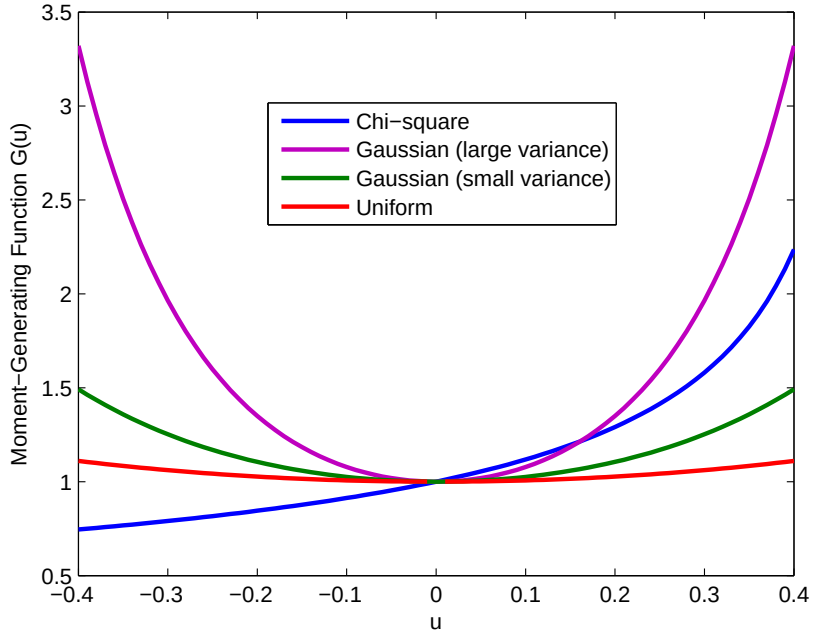
Performance of DSEE

When moment generating functions of $\{f_i(x)\}$ are properly bounded around 0:

- $\exists \zeta > 0,\ u_0 > 0$ s.t. $\forall u$ with $|u| \le u_0$,

$$\mathbb{E}[\exp((X-\theta)u)] \le \exp(\zeta u^2/2)$$

- DSEE achieves the optimal regret order $O(\log T)$.
- Achieve a regret *arbitrary close to* logarithmic w.o. any knowledge.



When $\{f_i(x)\}$ are heavy-tailed distributions:

- The moments of $\{f_i(x)\}$ exist only up to the pth order;
- DSEE achieves regret order $O(T^{1/p})$.

Basic Idea in Regret Analysis

Convergence Rate of the Sample Mean:

- Chernoff-Hoeffding Bound ('63): for distributions w. finite support $[a, b]$,

$$\Pr(|\overline{X_s} - \theta| \geq \delta) \leq 2\exp(-2\delta^2 s/(b-a)^2).$$

- Chernoff-Hoeffding-Agrawal Bound ('95): for distributions w. bounded MGF around 0,

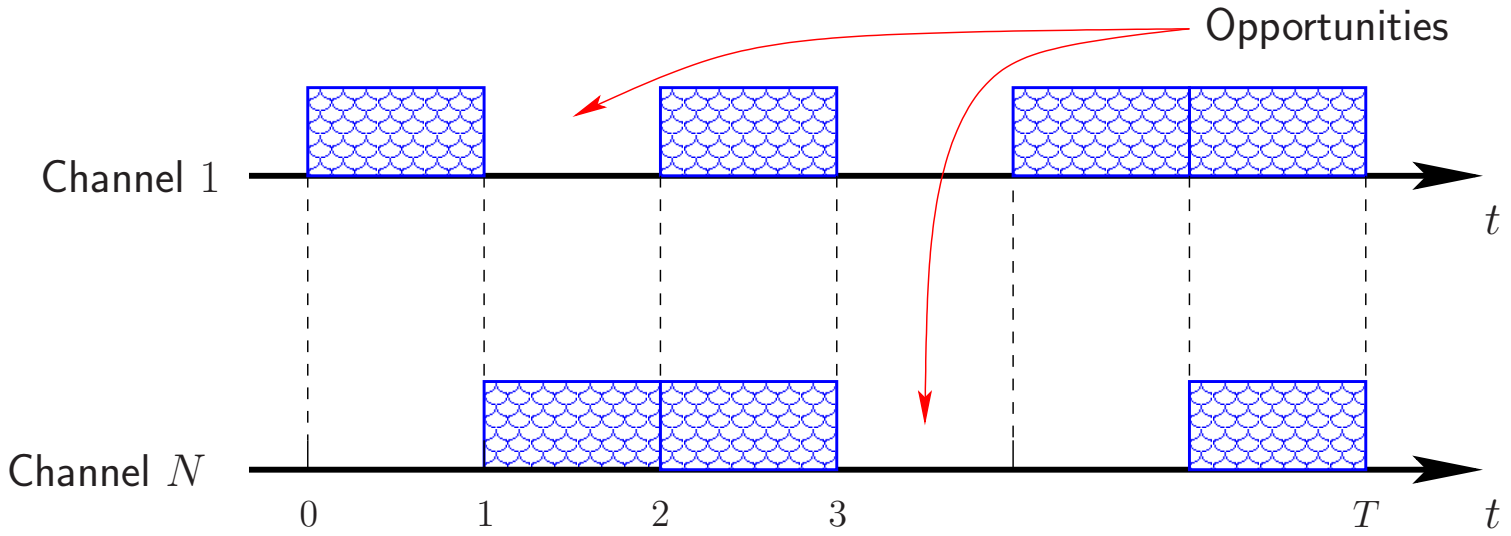
$$\Pr(|\overline{X_s} - \theta| \geq \delta) \leq 2\exp(-c\delta^2 s), \ \forall \ \delta \in [0, \zeta u_0], c \in (0, 1/(2\zeta)].$$

- Chow's Bound ('75): for distributions having the pth $(p > 1)$ moment,

$$\Pr(|\overline{X_s} - \theta| > \epsilon) = o(s^{1-p}).$$

Decentralized Bandit with Multiple Players

Distributed Spectrum Sharing



- N channels, M $(M < N)$ **distributed** secondary users (no info exchange).
- Primary occupancy of channel i: i.i.d. Bernoulli with unknown mean θ_i:
- Users accessing the same channel collide; no one receives reward.
- Objective: decentralized policy for optimal **network-level** performance.

Decentralized MAB with Multiple Players

Decentralized MAB with Multiple Players:

- M $(M < N)$ *distributed* players.
- Each player selects one arm to play.
- Players make decisions based on local observations w.o. info. exchange.
- Colliding players receive no or partial reward.
- Collisions may not be observable.

System Regret:

- Total reward with **known** $(\theta_1, \cdots, \theta_N)$ and **centralized scheduling**:

$$T \ \Sigma_{i=1}^{M} \theta^{(i)}$$

- Regret:

$$R_T^{\pi}(\Theta) = T\Sigma_{i=1}^{M} \theta^{(i)} - V_T^{\pi}(\Theta)$$

Decentralized MAB with Multiple Players

Decentralized MAB with Multiple Players:

- M $(M < N)$ *distributed* players.
- Each player selects one arm to play.
- Players make decisions based on local observations w.o. info. exchange.
- Colliding players receive no or partial reward.
- Collisions may not be observable.

Difficulties:

- Need to learn arms with different ranks for sharing.
- Collisions affect not only immediate reward, but also learning ability.

MAB under Various Objectives

Targeting at Arms with Arbitrary Ranks:

- The classic policies cannot be directly extended, *e.g.*,

  UCB Policy(*Auer et al. :02*): index $= \bar{\theta}_i + \sqrt{\frac{2\log t}{\tau_i(t)}}$

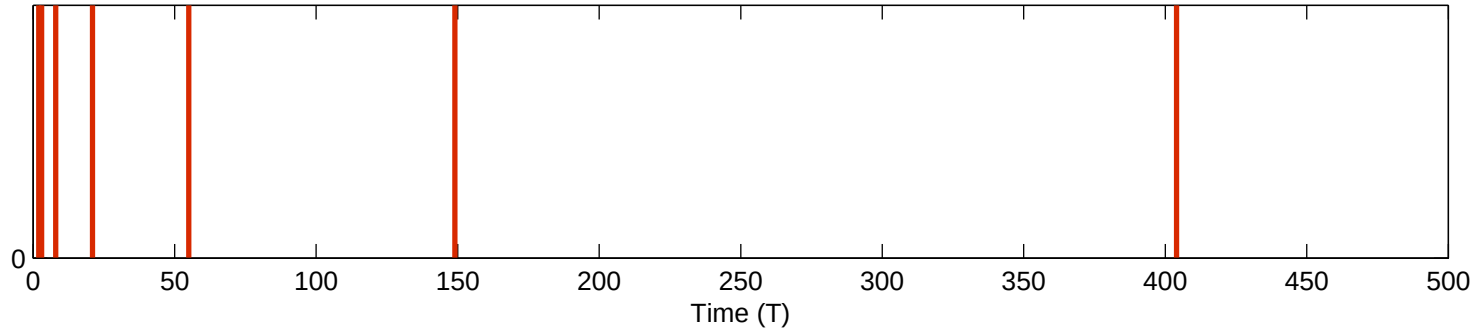
  If the index of the desired arm is too large to be selected, its index tends to become even larger.

- DSEE ensures efficient learning of arms at any rank.

- The objective can be time-varying: allows dynamic prioritized sharing.

Distributed Learning and Sharing Using DSEE

Distributed Learning and Sharing Using DSEE:



- Exploration: play all arms in round robin with different offsets;
- Exploitation: play the top M arms (in sample mean) with prioritized or fair sharing.

Regret:

- achieves the same regret order as in the centralized case;
- pre-agreement among players can be eliminated when collisions are observable: learn from collisions to achieve orthogonalization.

Restless Markov Reward Model

General Restless MAB with Unknown Dynamics

General Restless MAB with Unknown Dynamics:

- Rewards from successive plays form a MC with unknown transition $\mathbf{P}_i$.
- When passive, arm evolves a.t. an arbitrary unknown random process.

Difficulties:

- The optimal policy under known model is no longer staying on one arm.
- PSPACE-hard in general *(Papadimitriou-Tsitsiklis:99)*.

Weak Regret:

- Defined with respect to the optimal single-arm policy under known model:

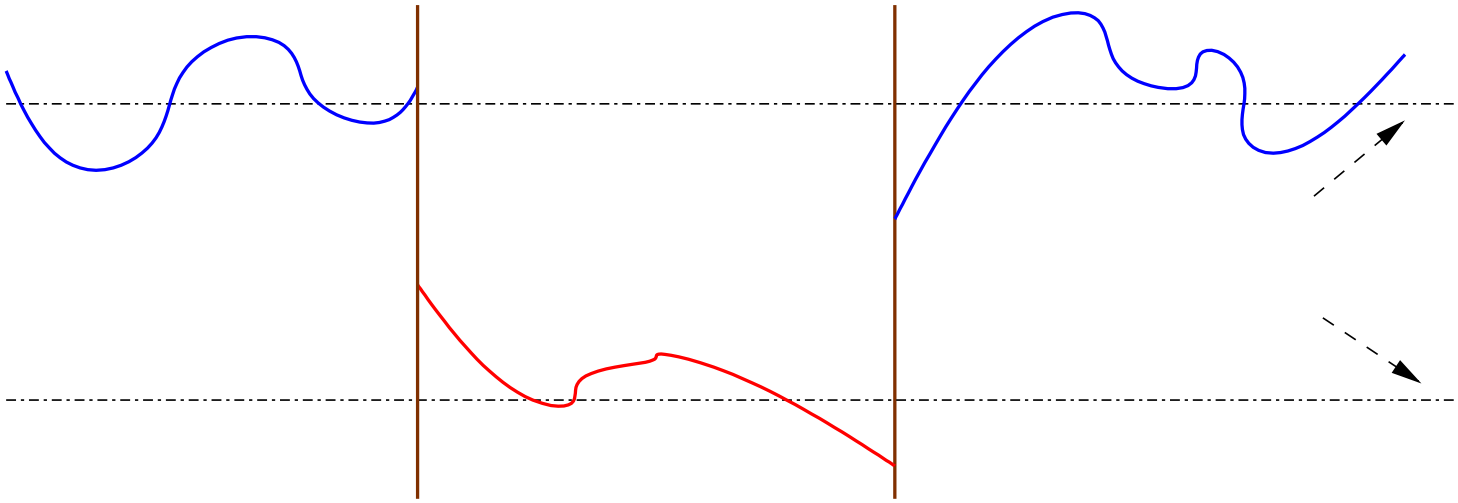
$$R_T^\pi = T\theta^{(1)} - V_T^\pi + O(1).$$

- The best arm: the largest reward mean $\theta^{(1)}$ in steady state.

General Restless MAB with Unknown Dynamics

Challenges:

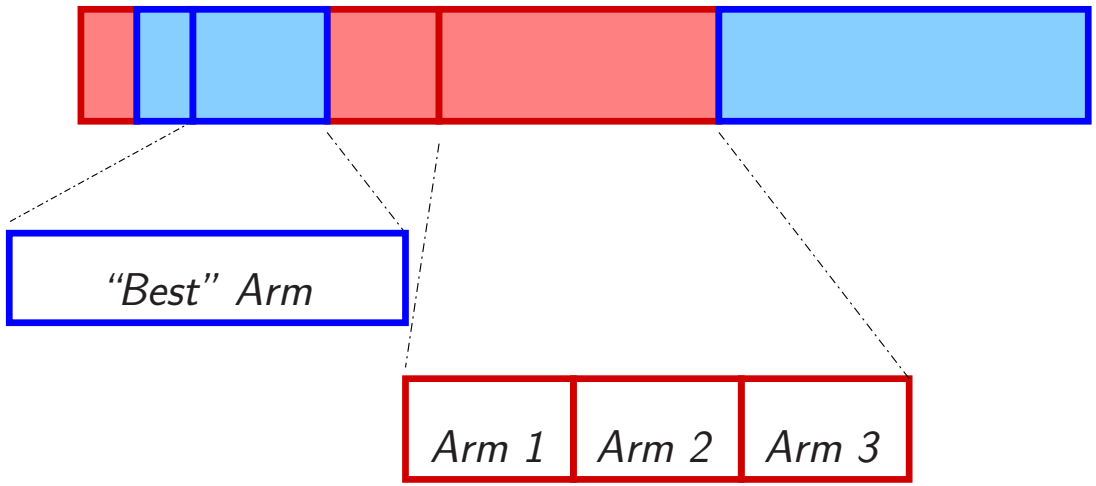
- Need to learn $\{\theta_i\}$ from contiguous segments of the sample path.
- Need to limit arm switching to bound the transient effect.



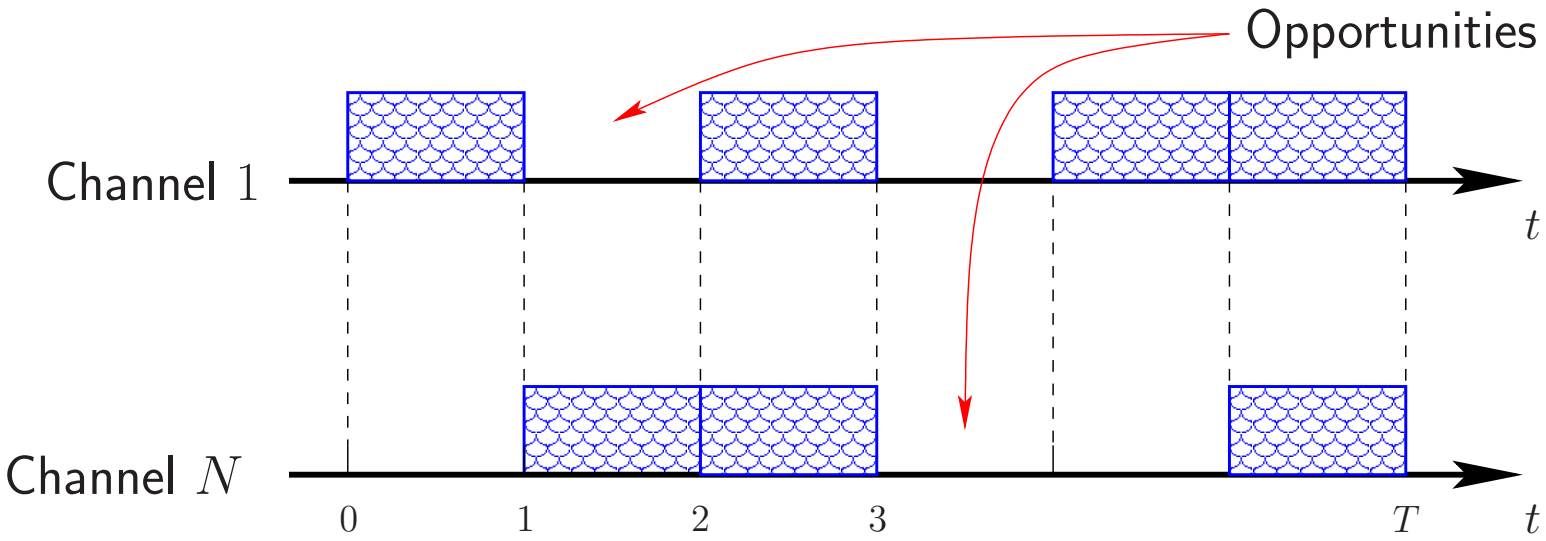
DSEE with An Epoch Structure

DSEE with an epoch structure:

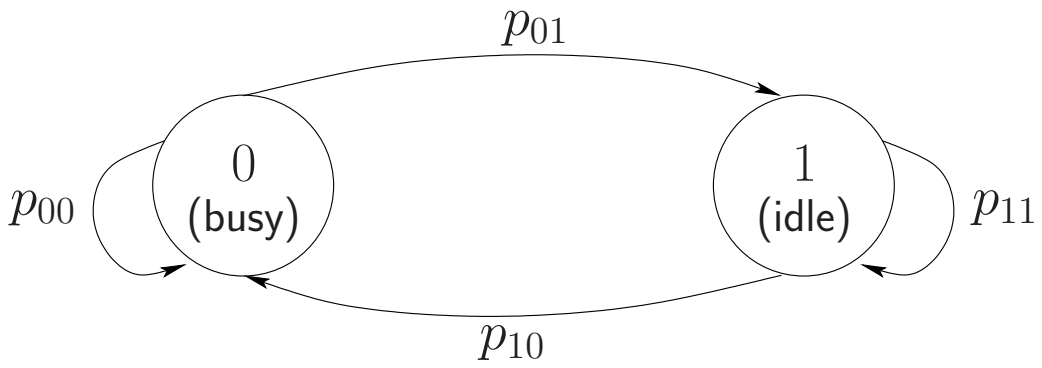
- Epoch structure with geometrically growing epoch length
  $\Longrightarrow$ arm switching limited to $\log$ order.
- Exploration and exploitation epochs interleaving:
  - In exploration epochs, play all arms in turn.
  - In exploitation epochs, play the arm with the largest sample mean.
  - Start an exploration epoch iff total exploration time $< D \log t$.
- Achieves *logarithmic* regret order.



Dynamic Spectrum Access Under Unknown Model

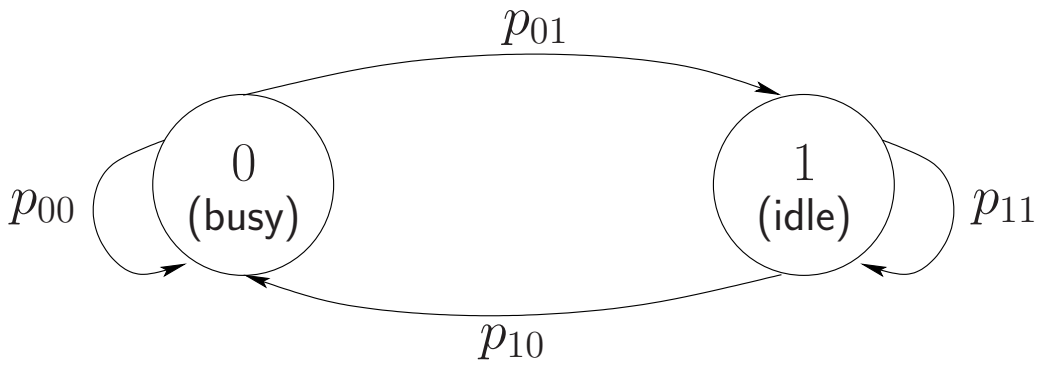


- Channel occupancy: Markovian with unknown transition probabilities:



- Objective: a channel selection policy to achieve max average reward.

Optimal Policy under Known Model



Semi-Universal Structure of the Optimal Policy:

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

- 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 $|E[X_i|X_1,\cdots,X_{i-1}]-\mu| \le C$ $(0 < C < \mu)$. Then $\forall a \ge 0$,

$$\Pr\{\overline{X_n} \ge n(\mu+C)+a\} \le e^{-2(\frac{a(\mu-C)}{b(\mu+C)})^2/n}$$
$$\Pr\{\overline{X_n} \le n(\mu-C)-a\} \le e^{-2(a/b)^2/n}$$

Regret Order:

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

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

Conclusion and Acknowledgement

- **Limitations of the Classic Results**:
  - Reward distributions limited to finite support or specific cases;
  - A single player (equivalently, centralized multiple players);
  - i.i.d. or *rested* Markov reward over successive plays of each arm.

- **Contributions: policies with a tunable parameter capable of handling**
  - a more general class of reward distributions (including heavy-tailed): *K.Liu-Q.Zhao:11*;
  - decentralized MAB with multiple players: *K.Liu-Q.Zhao:10, K.Liu-Q.Zhao:11*;
  - restless Markovian reward model: *H.Liu-K.Liu-Q.Zhao:11, Dai-Gai-Krishnamachari-Zhao:11*.