# 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, ARL, ARO.

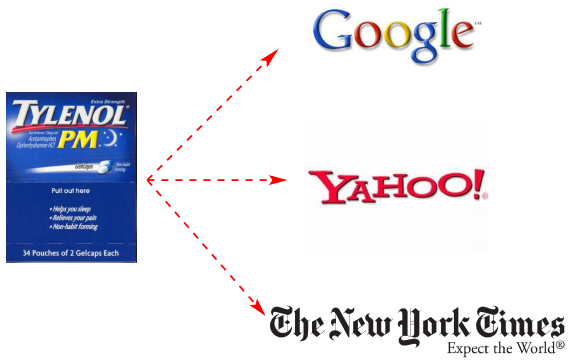
# Clinical Trial *(Thompson'33)*

**Two treatments with unknown effectiveness:**

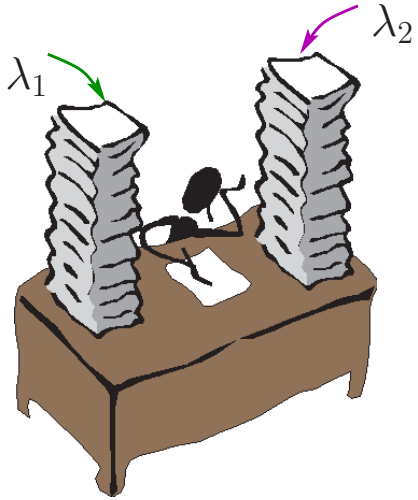
# Applications of MAB

## Web Search

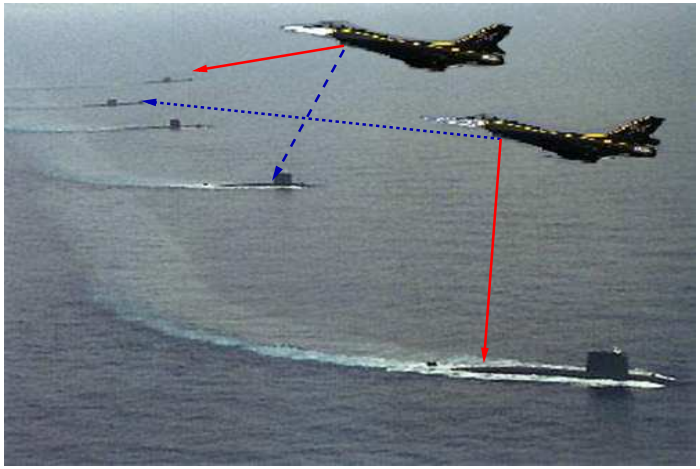

## Internet Advertising




## Queueing and Scheduling




## Multi-Agent Systems



# 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 $\theta_i$.
- Maximize the long-run reward.

# The Essential Tradeoff: Information vs. Immediate Payoff

**A Two-Armed Bandit:**

- Two coins with unknown bias $\theta_1$, $\theta_2$.
- Head: reward $= 1$; Tail: reward $= 0$.
- Objective: maximize total reward over $T$ flips.

*An Example (Berry&Fristedt'85):*

- $\theta_1 = \frac{1}{2}$, $\theta_2 = \begin{cases} 1, & \text{with probability } \frac{1}{4} \\ 0, & \text{with probability } \frac{3}{4} \end{cases}$
- To gain immediate payoff: flip Coin 1 indefinitely.
- To gain information: flip Coin 2 initially.

# Two Formulations

**Bayesian Formulation:**

- $\{\theta_i\}$ are random variables with *prior* distributions $\{f_{\theta_i}\}$.
- Policy $\pi$: choose an arm based on $\{f_{\theta_i}\}$ and the observation history.
- Objective: policies with good *average* (over $\{f_{\theta_i}\}$) performance.

# Two Formulations

**Bayesian Formulation:**

- $\{\theta_i\}$ are random variables with *prior* distributions $\{f_{\theta_i}\}$.
- Policy $\pi$: choose an arm based on $\{f_{\theta_1}\}$ and the observation history.
- Objective: policies with good *average* (over $\{f_{\theta_i}\}$) performance.

**Non-Bayesian Formulation:**

- $\{\theta_i\}$ are unknown deterministic parameters.
- Policy $\pi$: choose an arm based on the observation history.
- Objective: policies with *universally* (over all $\{\theta_i\}$) good performance.
- Key questions:
  - Is it possible to achieve the same average reward as in the known model case?
  - If yes, how fast is the convergence (learning efficiency)?

# Bayesian Formulation

# Bandit and MDP

**Multi-Armed Bandit as A Class of MDP:** *(Bellman'56)*

- $N$ *independent* arms with *fully observable* states $[Z_1(t), \cdots, Z_N(t)]$.
- *One* arm is activated at each time.
- Active arm changes state*(known Markov process)*; offers reward $R_i(Z_i(t))$.
- Passive arms are frozen and generate no reward.

# Bandit and MDP

**Multi-Armed Bandit as A Class of MDP:** *(Bellman'56)*

- $N$ *independent* arms with *fully observable* states $[Z_1(t), \cdots, Z_N(t)]$.
- *One* arm is activated at each time.
- Active arm changes state*(known Markov process)*; offers reward $R_i(Z_i(t))$.
- Passive arms are frozen and generate no reward.

*Why is sampling stochastic processes with unknown distributions an MDP?*

- The state of each arm is the *posterior* distribution $f_{\theta_i}(t)$*(information state)*.
- For an active arm, $f_{\theta_i}(t+1)$ is updated from $f_{\theta_i}(t)$ and the new observation.
- For a passive arm, $f_{\theta_i}(t+1) = f_{\theta_i}(t)$.

# Bandit and MDP

**Multi-Armed Bandit as A Class of MDP:** *(Bellman'56)*

- $N$ *independent* arms with *fully observable* states $[Z_1(t), \cdots, Z_N(t)]$.
- *One* arm is activated at each time.
- Active arm changes state*(known Markov process)*; offers reward $R_i(Z_i(t))$.
- Passive arms are frozen and generate no reward.

**Solving Multi-Armed Bandit using Dynamic Programming:**

- Exponential complexity with respect to $N$.

# Gittins Index

**The Index Structure of the Optimal Policy:** *(Gittins'74)*

- Assign each state of each arm a *priority* index.
- Activate the arm with highest current index value.

# Gittins Index

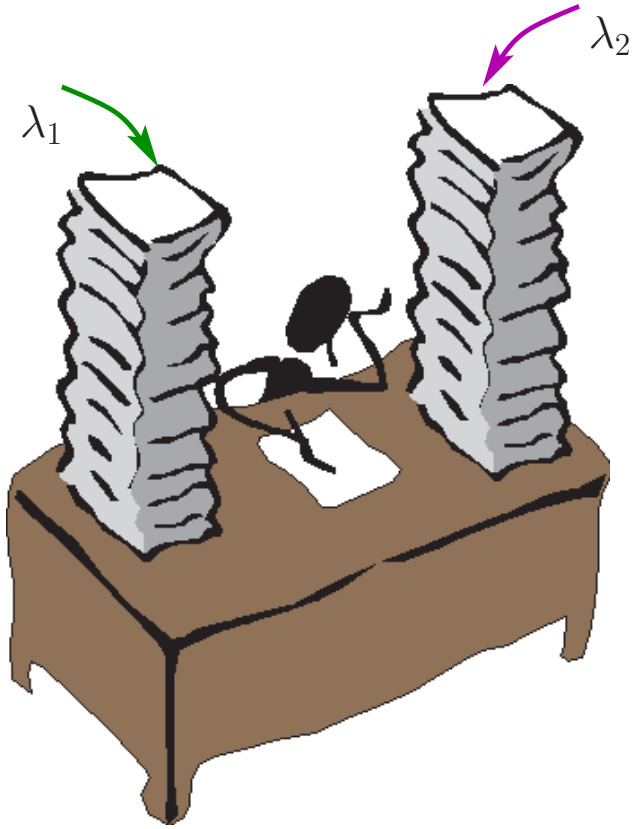
**The Index Structure of the Optimal Policy:** *(Gittins'74)*

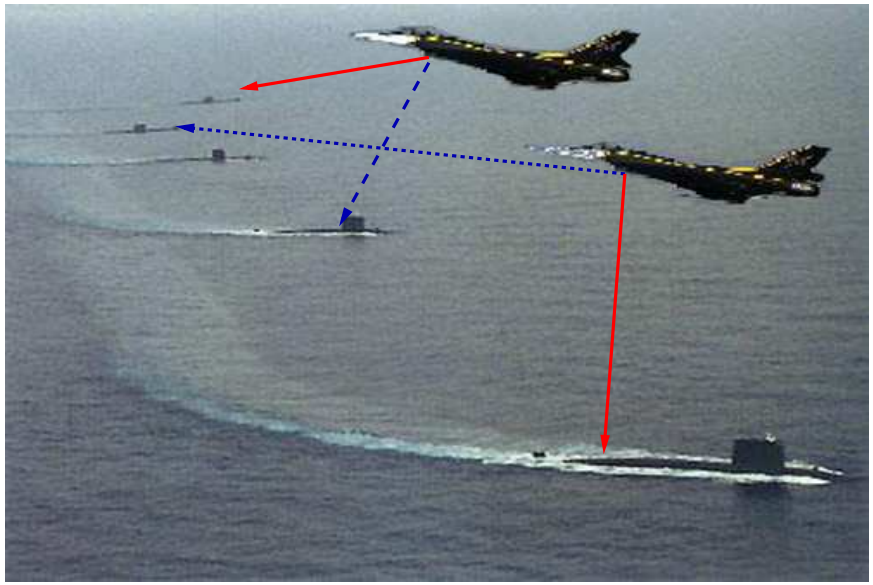
- Assign each state of each arm a *priority* index.
- Activate the arm with highest current index value.

**Complexity:**

- Arms are decoupled ($1$ $N$-dim to $N$ separate $1$-dim problems).
- *Linear* complexity with $N$.
- Polynomial (cubic) with the state space size of a *single* arm *(Varaiya&Walrand&Buyukkoc'85, Katta&Sethuraman'04)*.

# Restless Bandit



# Restless Bandit

**Restless Multi-Armed Bandit:** *(Whittle'88)*

- Passive arms also change state and offer reward.
- Activate $K$ arms simultaneously.

**Structure of the Optimal Policy:**

- Not yet found.

**Complexity:**

- PSPACE-hard *(Papadimitriou&Tsitsiklis'99)*.

# Whittle Index

**Whittle Index:** *(Whittle'88)*

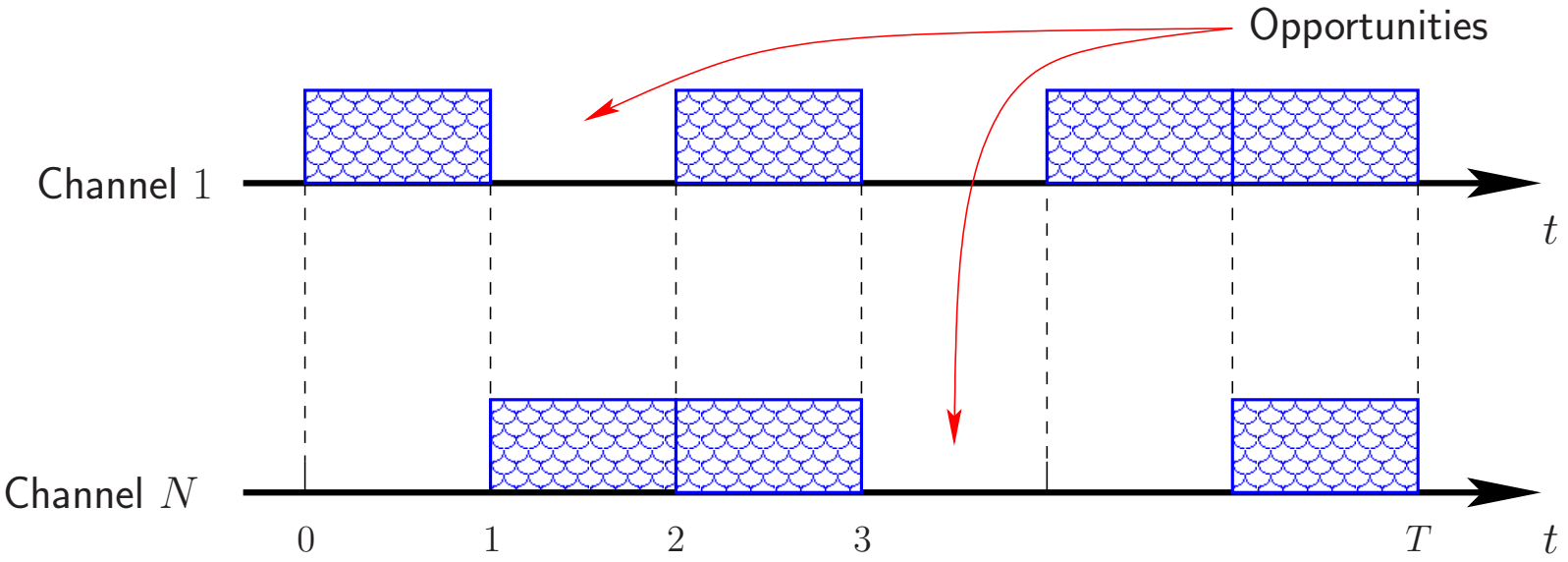
- Optimal under relaxed constraint on the average number of active arms.
- Asymptotically ($N \to \infty$) optimal under certain conditions *(Weber&Weiss'90)*.
- Near optimal performance observed from extensive numerical examples.

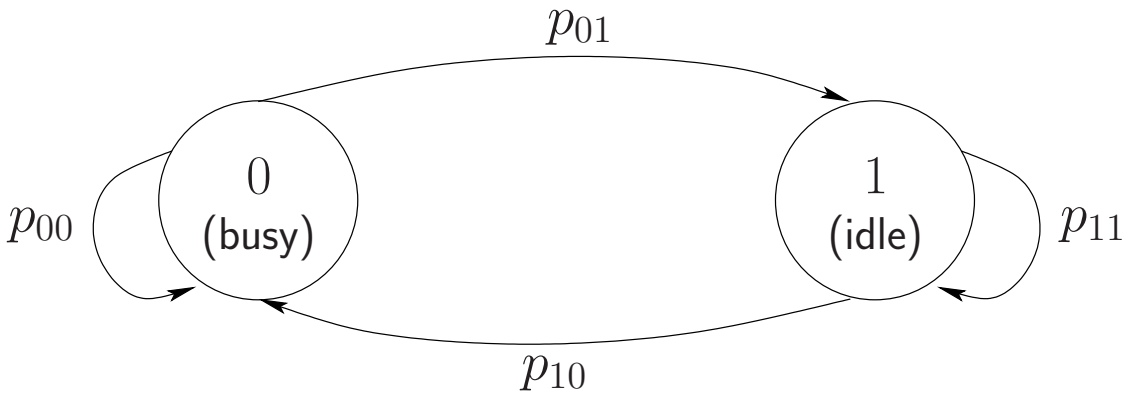
**Difficulties:**

- Existence (indexability) not guaranteed and difficult to check.
- Numerical index computation infeasible for infinite state space.
- Optimality in finite regime difficult to establish.

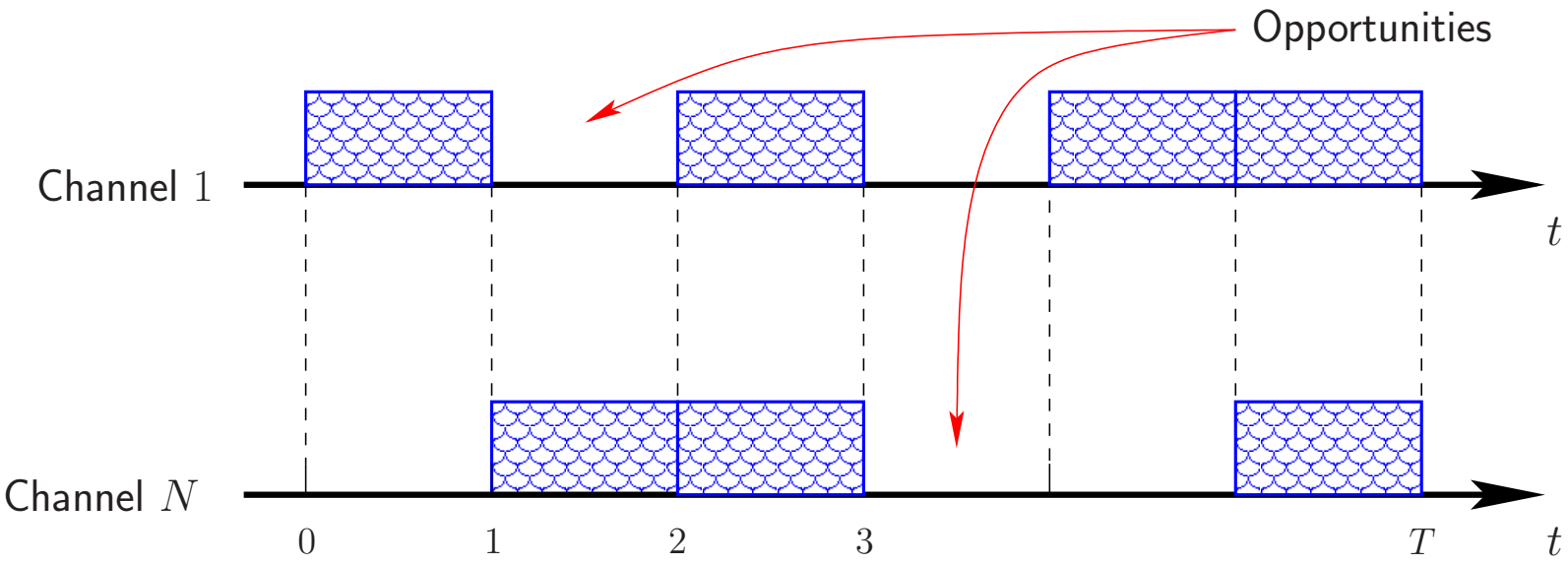
# Spectrum Opportunity Tracking

**Sense $K$ of $N$ channels:**



# Spectrum Opportunity Tracking

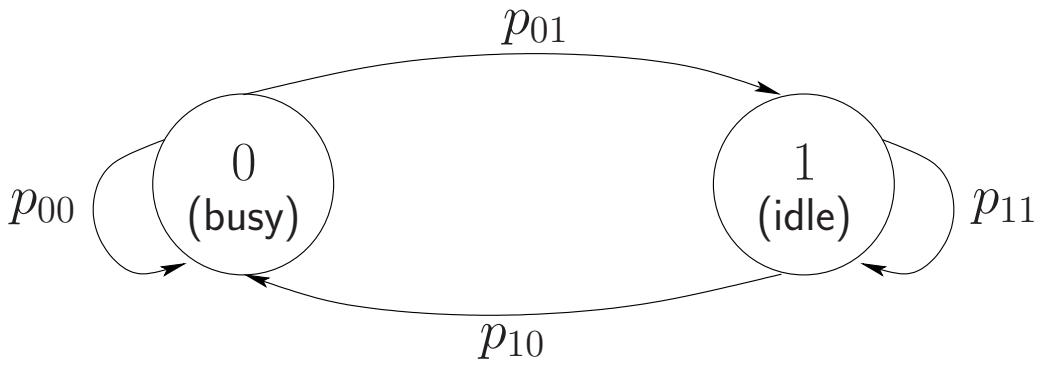



- Each channel is considered as an arm.
- State of arm $i$: *posterior* probability that channel $i$ is idle.

$$\omega_i(t) = \Pr[\text{channel } i \text{ is idle in slot } t \mid \underbrace{O(1), \cdots, O(t-1)}_{\textit{observations}}]$$

- The expected immediate reward for activating arm $i$ is $\omega_i(t)$

# Markovian State Transition




- If channel $i$ is activated in slot $t$:

$$\omega_i(t+1) = \begin{cases} p_{11}, & \text{if} \quad O_i(t) = 1 \\ p_{01}, & \text{if} \quad O_i(t) = 0 \end{cases}.$$

- If channel $i$ is made passive in slot $t$:

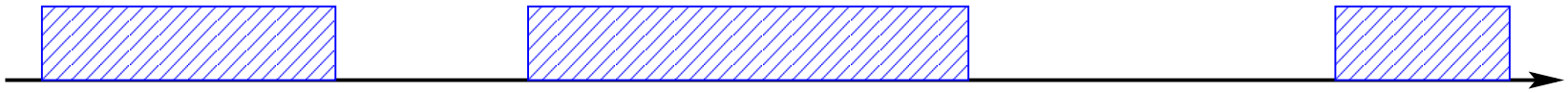
$$\omega_i(t+1) = \omega_i(t)p_{11} + (1 - \omega_i(t))p_{01}.$$

# Structure of Whittle Index Policy

**The Semi-Universal Structure of Whittle Index Policy:**

- No need to compute the index.
- No need to know $\{p_{01}, p_{11}\}$ except their order *(robust to model mismatch)*.

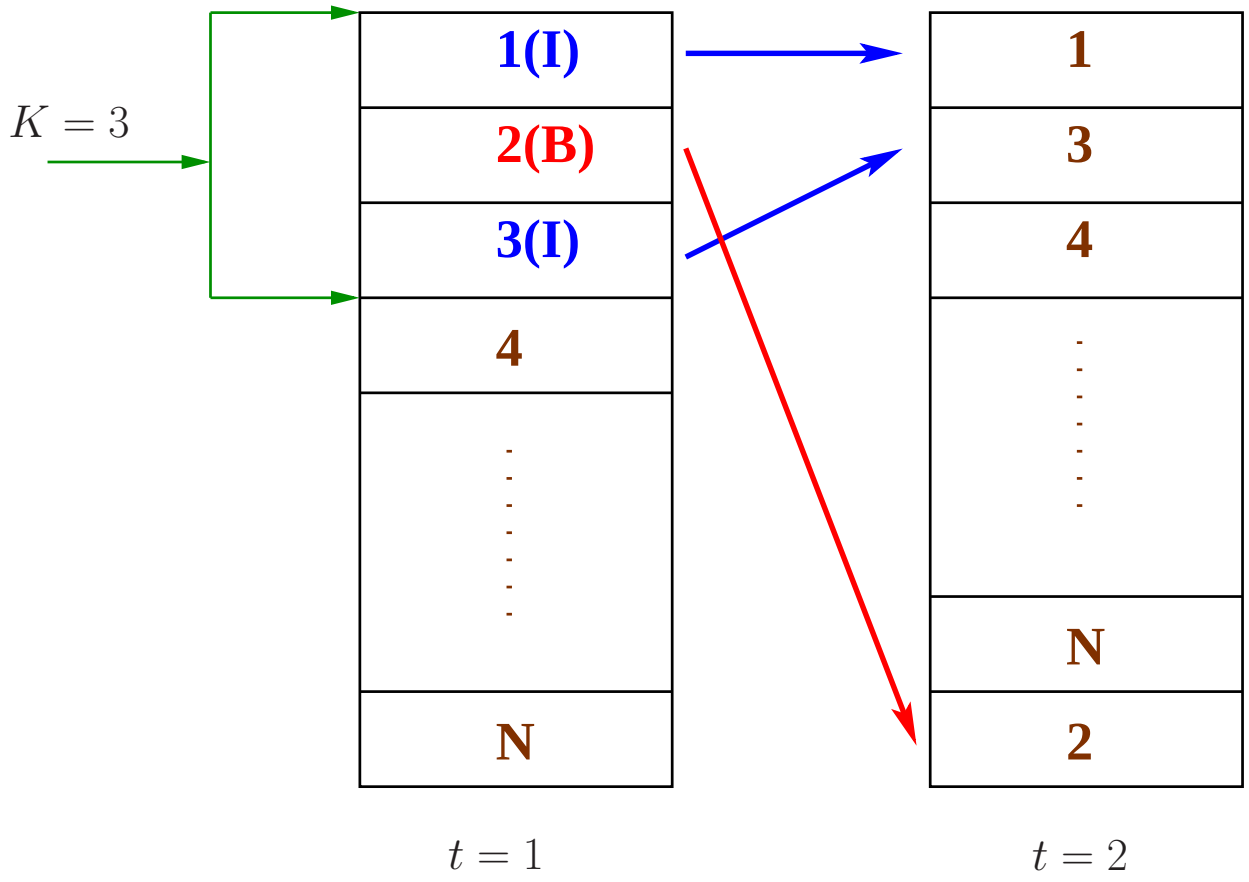
$p_{11} \geq p_{01}$ *(positive correlation):*



$p_{11} < p_{01}$ *(negative correlation):*

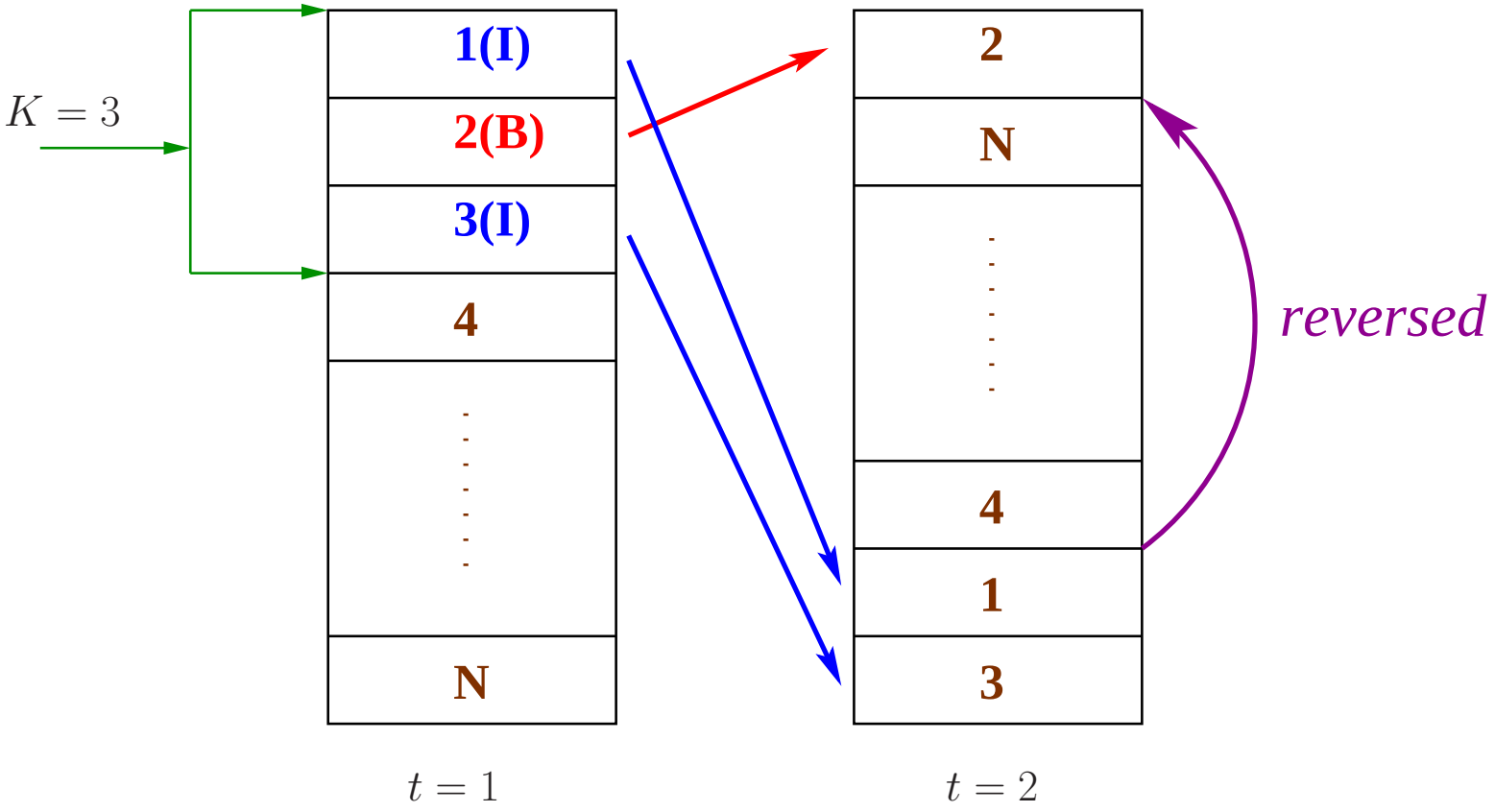
# Structure of Whittle Index Policy: Positive Correlation

- Stay with idle **(I)** channels and leave busy **(B)** ones to the end of the queue.



# Structure of Whittle Index Policy: Negative Correlation

- Stay with busy **(B)** channels and leave idle **(I)** ones to the end of the queue.
- *Reverse* the order of unobserved channels.



# Optimality of Whittle Index Policy

**Optimality for positively correlated channels:**

- holds for general $N$ and $K$.
- holds for both finite and infinite horizon (discounted/average reward).

**Optimality for negatively correlated channels:**

- holds for all $N$ with $K = N - 1$.
- holds for $N = 2,\ 3$.

# Inhomogeneous Channels

## Whittle Index in Closed Form:

- Positive correlation $(p_{11} \geq p_{01})$:

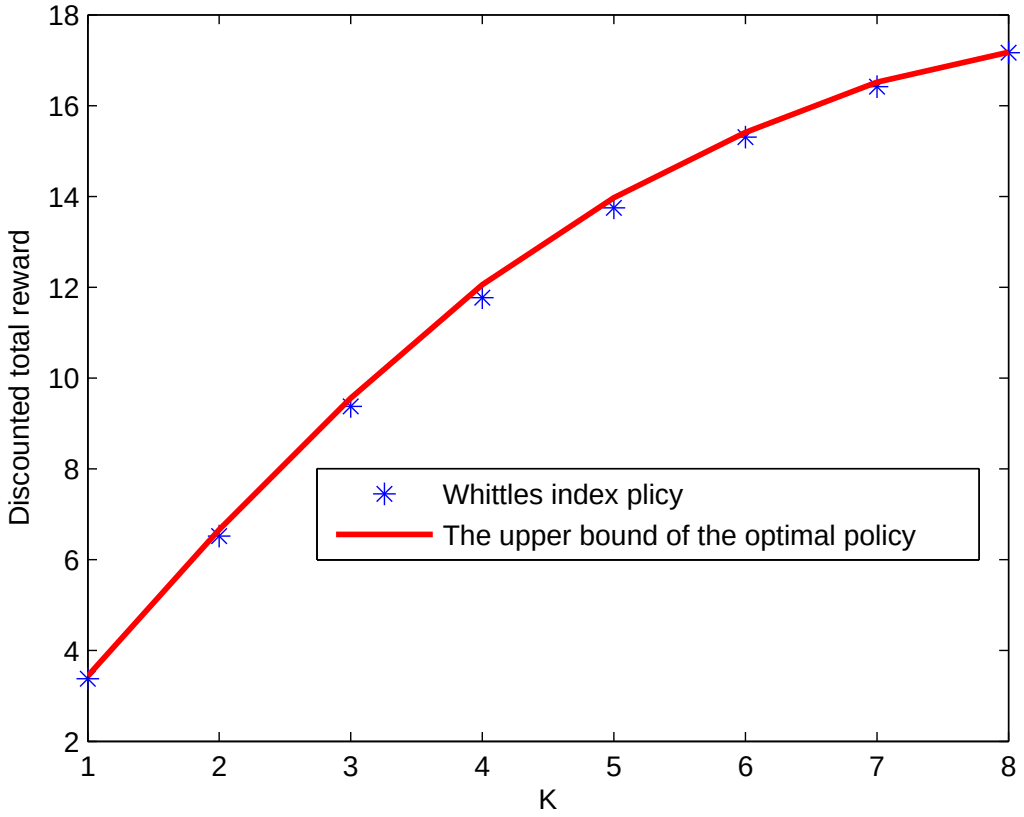
$$I(\omega) = \begin{cases} \omega, & \omega \leq p_{01} \text{ or } \omega \geq p_{11} \\ \frac{\omega}{1-p_{11}+\omega}, & \omega_o \leq \omega < p_{11} \\ \frac{(\omega-\mathcal{T}^1(\omega))(L+2)+\mathcal{T}^{L+1}(p_{01})}{1-p_{11}+(\omega-\mathcal{T}^1(\omega))(L+1)+\mathcal{T}^{L+1}(p_{01})}, & p_{01} < \omega < \omega_o \end{cases}$$

- Negative correlation $(p_{11} < p_{01})$:

$$I(\omega) = \begin{cases} \omega, & \omega \leq p_{11} \text{ or } \omega \geq p_{01} \\ \frac{p_{01}}{1+p_{01}-\omega}, & \mathcal{T}^1(p_{11}) \leq \omega < p_{01} \\ \frac{p_{01}}{1+p_{01}-\mathcal{T}^1(p_{11})}, & \omega_o \leq \omega < \mathcal{T}^1(p_{11}) \\ \frac{\omega+p_{01}-\mathcal{T}^1(\omega)}{1+p_{01}-\mathcal{T}^1(p_{11})+\mathcal{T}^1(\omega)-\omega}, & p_{11} < \omega < \omega_o \end{cases}$$

# Performance for Inhomogeneous Channels

- The tightness of the performance upper bound ($\mathcal{O}(N(\log N)^2)$ running time).
- The near-optimal performance of Whittle's index policy



# Non-Bayesian Formulation

# 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 $\Theta$ is known.
- $V_T^\pi(\Theta)$: total reward of policy $\pi$ 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:

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

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

# Classic Results

- Lai&Robbins'85:

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

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

- Limitations:
  - Reward distributions limited to exponential family or finite support.
  - Assume known distribution type or support range.
  - i.i.d. reward.
  - A single player (equivalently, centralized multiple players).

# General Reward Distributions

# MAB with A General Reward Model

**Multi-Armed Bandit:**

- $N$ arms and a single player.
- Select one arm to play at each time.
- i.i.d. reward w. *Unknown* distribution $f_i$.
- Reward mean $\theta_i$ exists.

# Sufficient Statistics

**Sufficient 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 $t$:

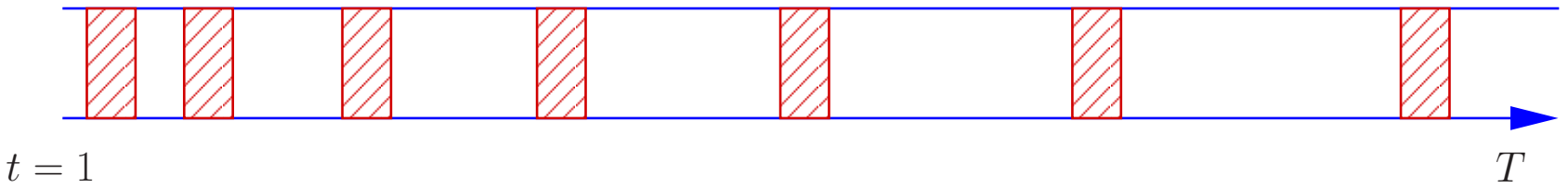
$$\text{index } = \bar{\theta}_i + \sqrt{\frac{2\log t}{\tau_i(t)}}$$

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

# 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.

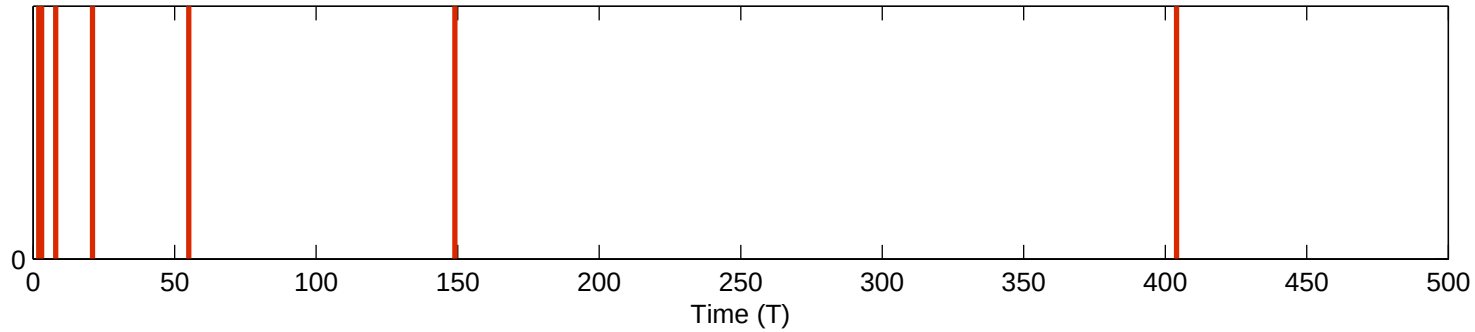
- A tunable parameter: the cardinality of the exploration sequence
  - can be adjusted according to the "hardness" of the reward distributions.

# The Optimal Cardinality of Exploration

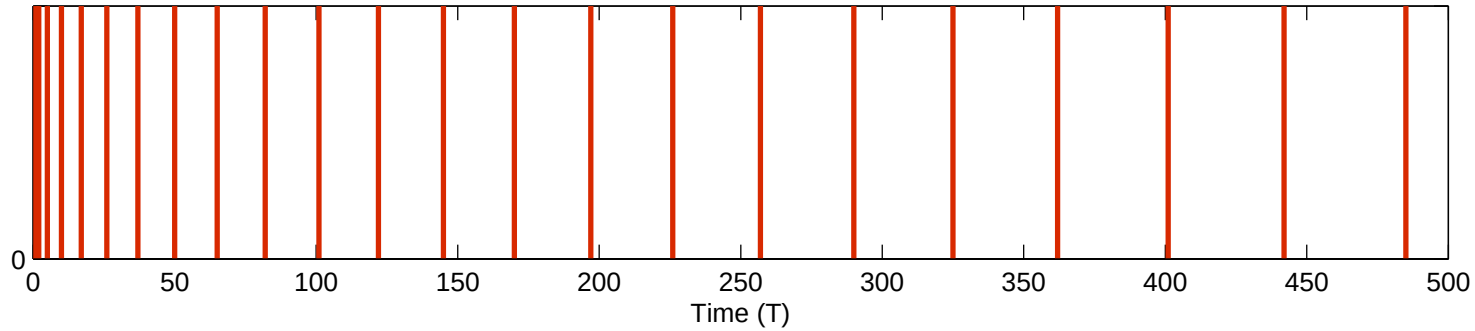
**The Cardinality of Exploration:**

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

- $O(\log T)$?




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



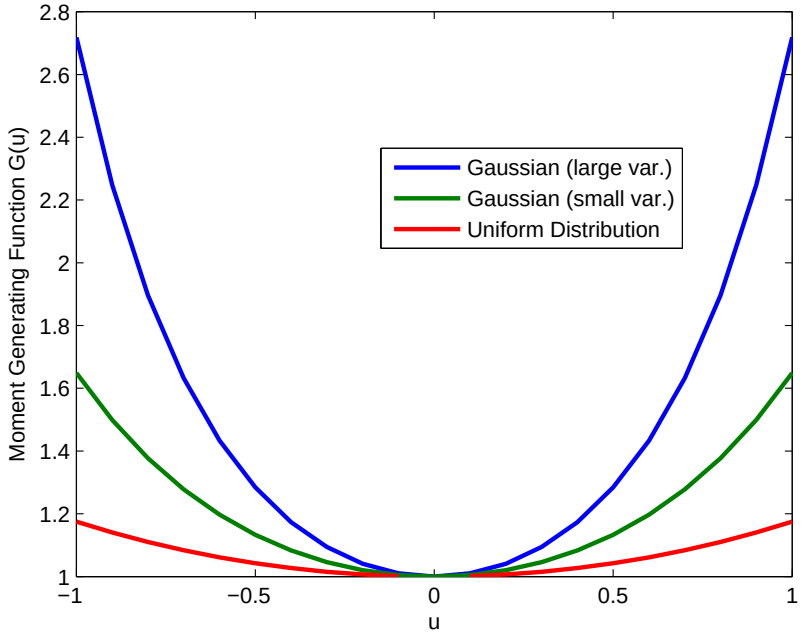
# Performance of DSEE

**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 $O(\log T)$.

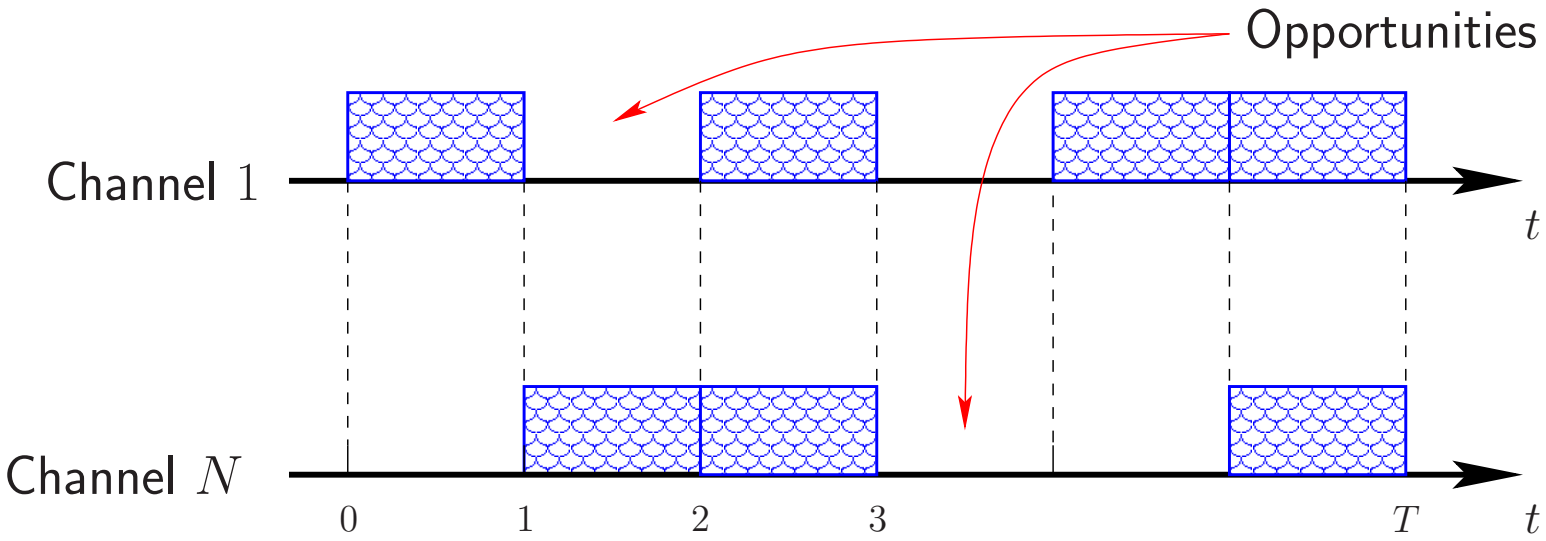



**When $\{f_i(x)\}$ are heavy-tailed distributions:**

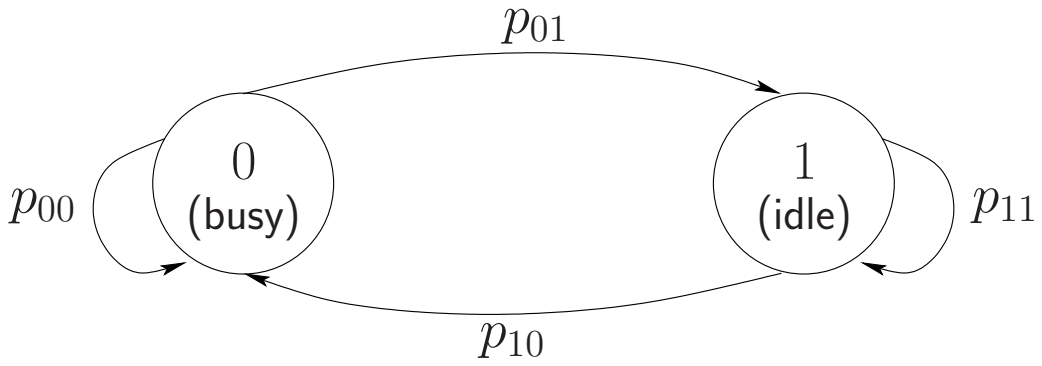
- The moments of $\{f_i(x)\}$ exist only up to the $p$th order;
- DSEE achieves regret order $O(T^{1/p})$.

# Restless Markov Reward Model

# Markov Model



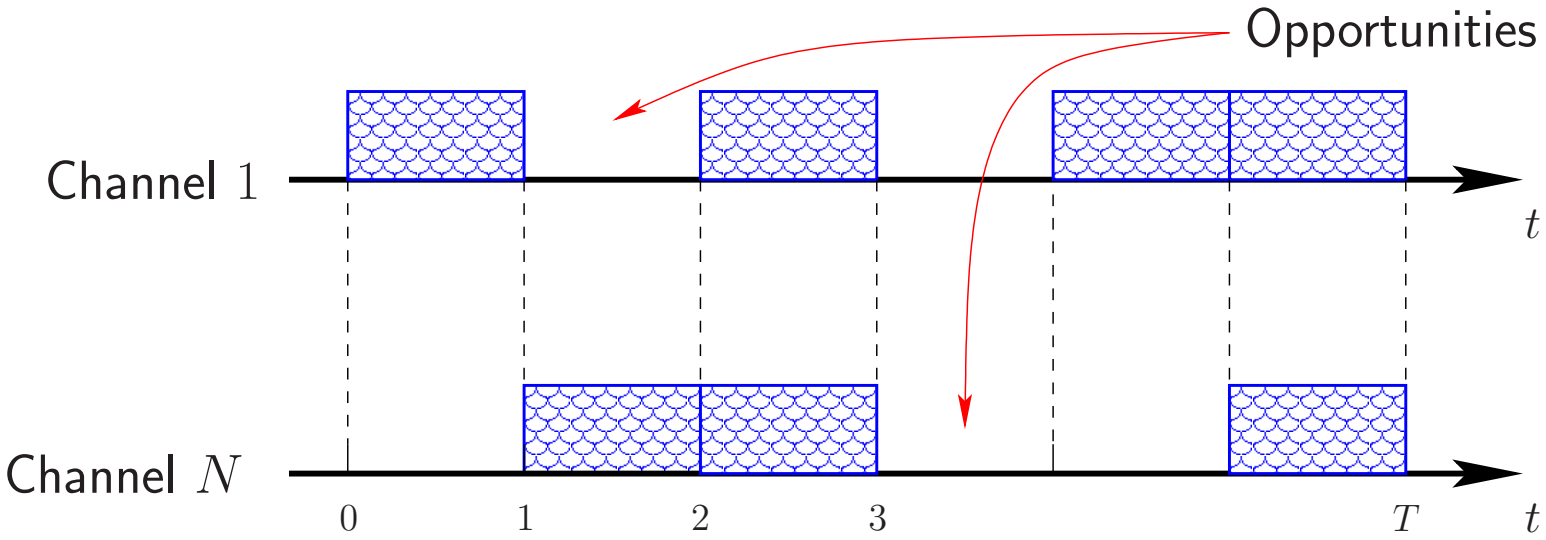

- Channel occupancy: Markovian with unknown transition probabilities:




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

# 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

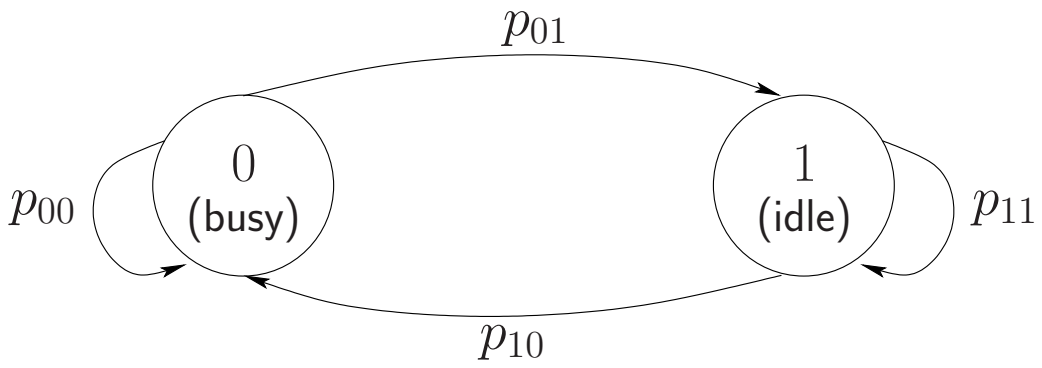
**Restless Multi-Armed Bandit:**

- Whittle index *(Whittle:88)*.
- PSPACE-hard in general *(Papadimitriou-Tsitsiklis:99)*.

**Optimality of Whittle Index:**

*(Zhao-Krishnamachari:07,Ahmad-Liu-Javidi-Zhao-Krishnamachari:09,Liu-Zhao:10)*

- 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 Whittle Index Policy:** *(Zhao-Krishnamachari:07,Liu-Zhao:10)*

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

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

**Regret Order:**

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

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

# 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.

**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:

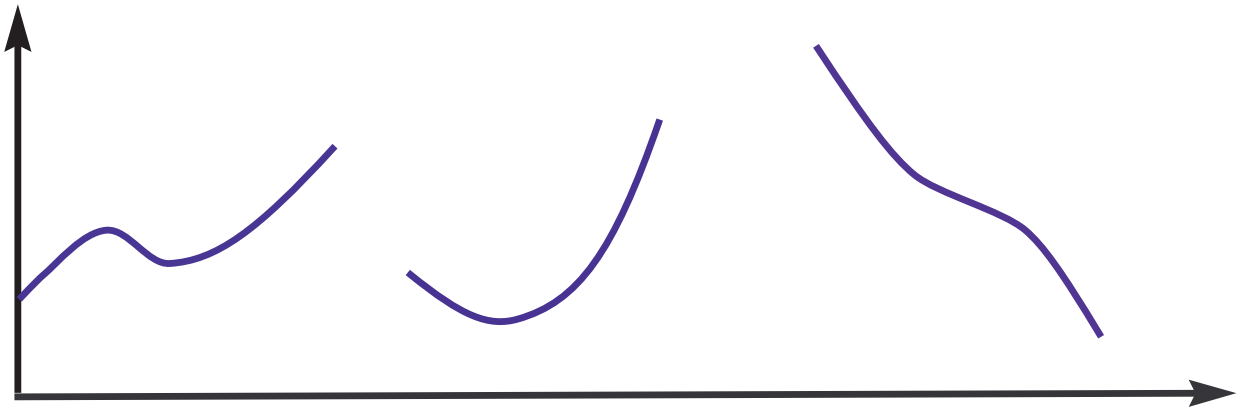
$$R_T^\pi = T\theta^{(1)} - V_T^\pi + O(1).$$

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

# 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.



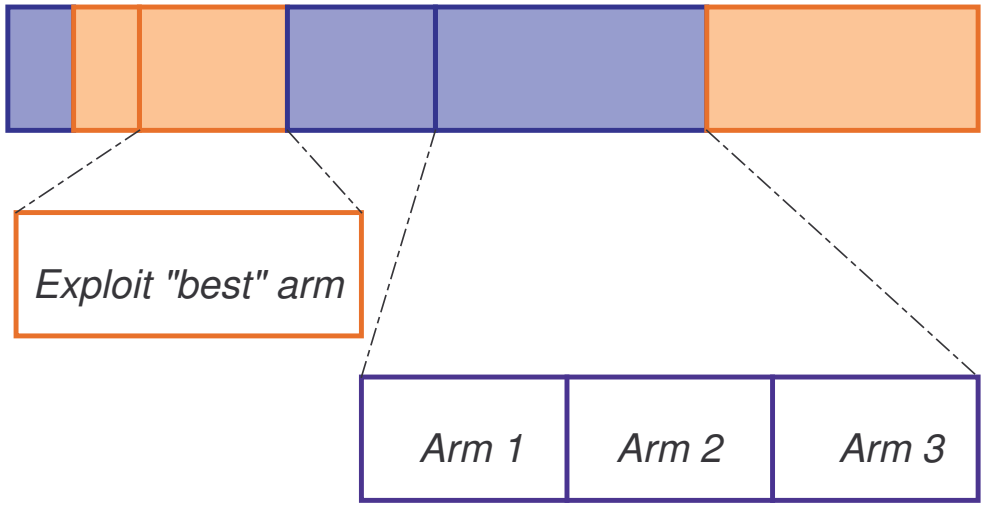
# Restless UCB Policy

## Restless UCB (RUCB):

- Epoch structure with geometrically growing epoch length

  $\Longrightarrow$ 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 $\mathbf{P}_i$.
  - 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

$$D(t) = f(t), \quad \frac{L(t)}{D(t)} \to 0.$$

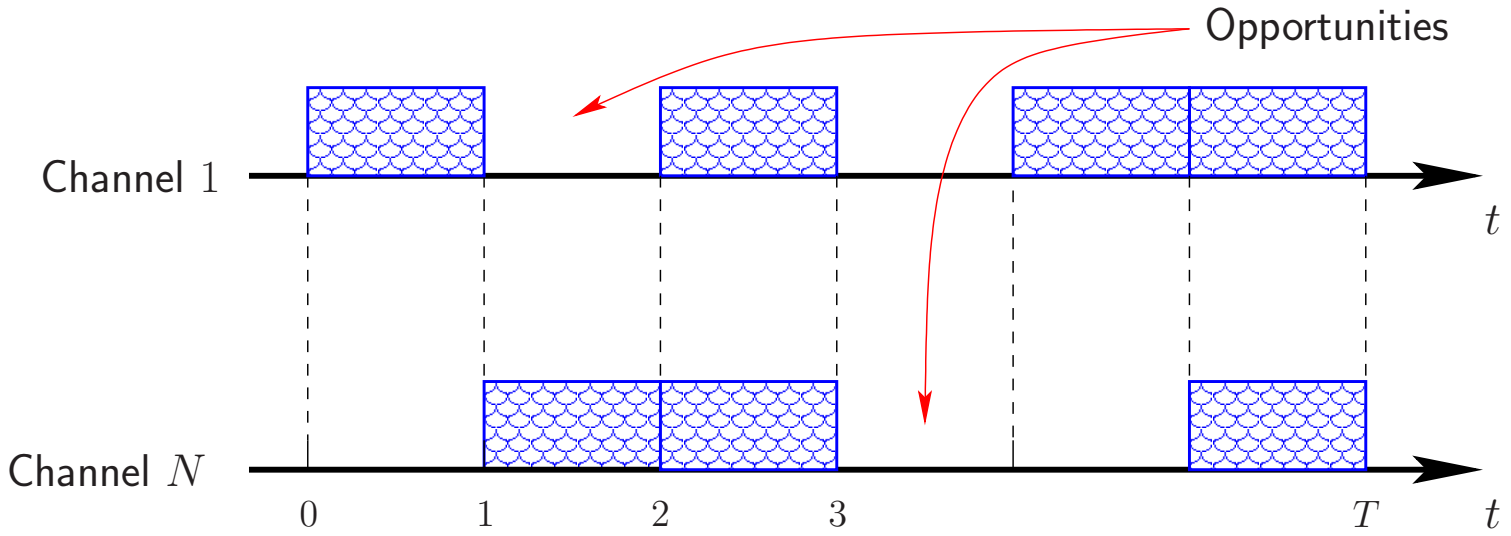
# Decentralized Bandit with Multiple Players

# Decentralized Multi-Armed Bandit

**Decentralized Bandit with Multiple Players:**

- $N$ arms with **unknown** reward statistics $(\theta_1, \cdots, \theta_N)$.
- $M$ $(M < N)$ distributed players.
- Each player selects one arm to play and observes the reward.
- Distributed decision making using only **local** observations.
- Colliding players either share the reward or receive no reward.

# 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 $\theta_i$:
- Users accessing the same channel collide; no one receives reward.
- Objective: decentralized policy for optimal **network-level** performance.

# Decentralized Multi-Armed Bandit

**System Regret:**

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

$$T \sum_{i=1}^{M} \underbrace{\theta^{(i)}}_{i\text{th best}}$$

- $V_T^{\pi}(\Theta)$ : total system reward under a decentralized policy $\pi$.
- System regret:

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

# The Minimum Regret Growth Rate

**The Minimum regret rate in Decentralized MAB is logarithmic.**

$$R_T^*(\Theta) \sim C(\Theta) \log T$$

**The Key to Achieving $\log T$ Order:**

- Learn from local observations which channels are the most rewarding.
- Learn from **collisions** to achieve efficient sharing with other users.

# Conclusion and Acknowledgement

- Bayesian Formulation: A Class of Restless Multi-Armed Bandit:
  - Indexability and optimality of Whittle index: *K.Liu&Q.Zhao:08-10*.
  - Extension to non-Markovian systems: *K.Liu&R.Weber&Q.Zhao:11*.
  - Structure and optimality of myopic policy:
    *Q.Zhao&B.Krishnamachari:07-08,*
    *S.Ahmad&M.Liu&T.Javidi&Q.Zhao&B.Krishnamachari:09*.

- Non-Bayesian Formulation:
  - From exponential family to general (heavy-tailed) unknown distributions:
    *K.Liu-Q.Zhao:11*.
  - From i.i.d. to restless Markov reward models:
    *Dai-Gai-Krishnamachari-Zhao:11, H.Liu-K.Liu-Q.Zhao:11*.
  - From single player to multiple distributed players:
    *K.Liu-Q.Zhao:10, H.Liu-K.Liu-Q.Zhao:11*