# Online Learning for Indexable Restless Multi-Armed Bandits

**Vishrant Tripathi**

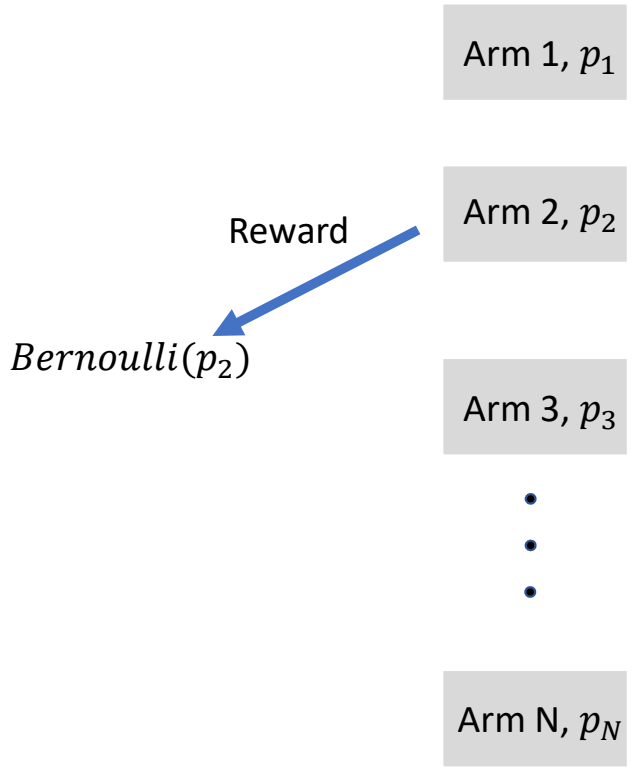
LIDS Student Conference, 2022, MIT

# Outline

1. Restless Multi-Armed Bandits (RMAB)
2. Indexability and the Whittle Index
3. An Online RMAB Formulation
   - **Follow-the-Perturbed-Whittle-Index**
4. Application to Wireless Scheduling

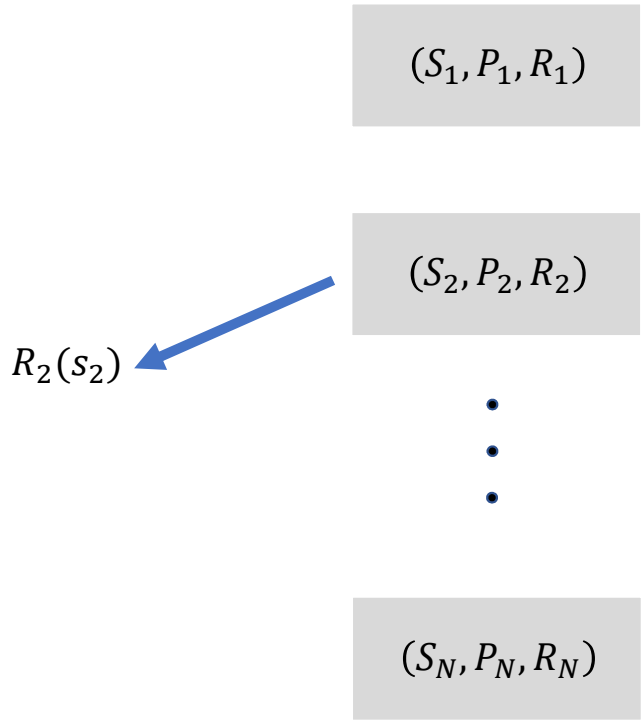
# (Bernoulli) Multi-Armed Bandits

- **N arms,** each of which gives a Bernoulli reward when activated
- Can activate **one arm at a time**
- If probabilities known, always **choose arm with highest reward**
- **Setting of interest**: **learning the best arm** without knowing probabilities a-priori



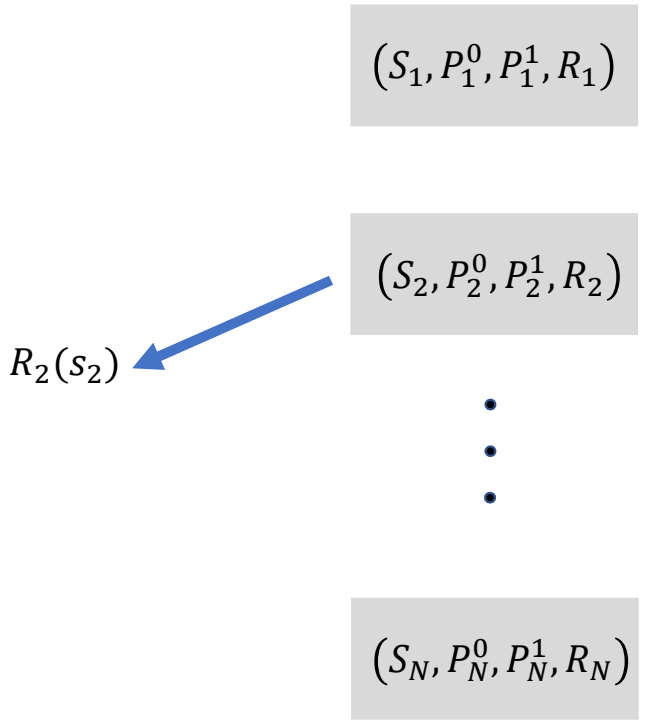
# (Markov) Multi-Armed Bandits

- Arm $i$ is a Markov chain over state space $S_i$ with law $P_i$ and rewards $R_i: S_i \to [0,1]$
- Arm state **only evolves when activated**, at rest otherwise
- **Setting of interest**: optimal policy to activate arms, **all information** about arms is **known**
- **Gittins (1970s)** – optimal policy can be computed **efficiently** using an **index structure**



# (Restless) Multi-Armed Bandits

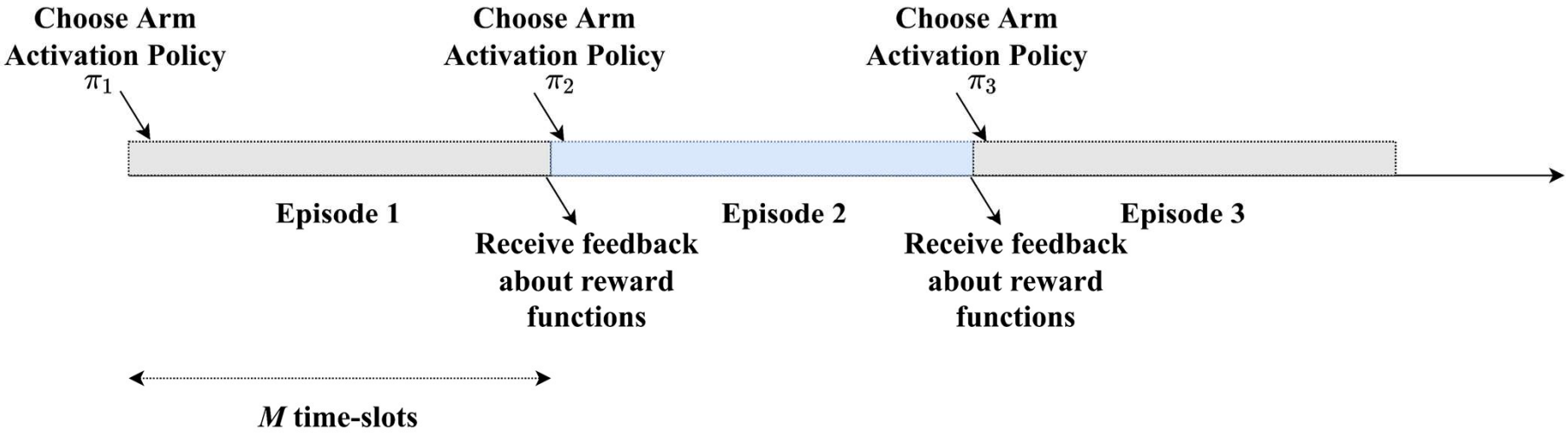
- Arm $i$ is a Markov chain over state space $S_i$ with **two laws** $P_i^0$ and $P_i^1$ and rewards $R_i: S_i \to [0,1]$
- Arm state **evolves using** $P_i^1$ **when activated**, and using $P_i^0$ when at rest
- **Setting of interest**: optimal policy to activate arms, all information **known**
- **Papadimitriou and Tsitsiklis (1994):** RMAB are **PSPACE complete**



# Indexability and the Whittle Index

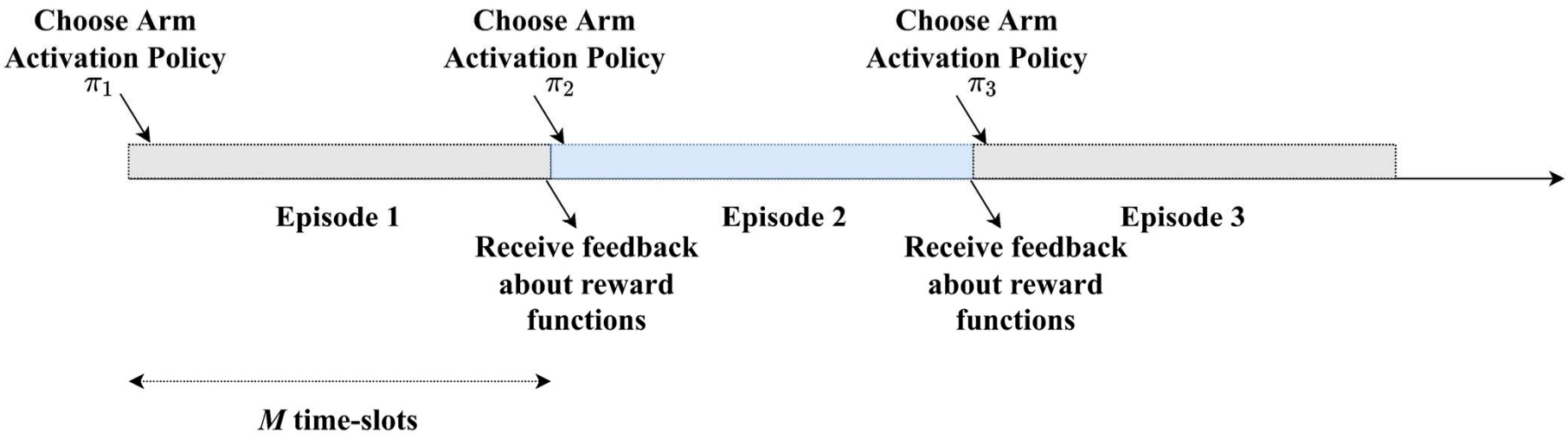
- **Whittle (1980s)** – near-optimal policy can be computed **efficiently,** given special **indexability** property
- **Indexability of Arm $i$:** Given activation cost $C > 0$, the set of states for which it is optimal to activate the arm decreases **monotonically** as $C$ increases
- Compute index functions $W_i$: $S_i \to \mathbb{R}$, which denote how "valuable" it is to activate arm $i$ at state $s_i$
- **Whittle Index policy** $\pi(t) = \arg\max_i \{W_i(s_i(t))\}$

# RMAB: An Online Learning Formulation




- Episodes of length $M$, each episode involves solving a RMAB problem
- Arms' **state-spaces and transition laws remain fixed**
- **Reward functions change across episodes** in an unknown manner *while maintaining indexability*

# RMAB: An Online Learning Formulation




- **Q:** Can we design a scheme that learns the best scheduling policy in an online manner?
- **Answer: Yes!**

# Follow The Perturbed Leader

- Viewing arm activation policies as experts:

    1. Maintain the sum of rewards observed in the past
    2. **Perturb i.i.d.** the history of rewards **for each scheduling policy**
    3. **Find the best policy** using this perturbed history

- The number of policies scales exponentially in the length of the epoch $\Theta\left(N^M\right)$

- Thus **traditional online learning methods are infeasible**

# Follow The Perturbed ***Whittle Index***

**An Alternative**:

1. Accumulate the history of reward functions observed
$$R_i = R_i + r_i(t)$$
2. Perturb these reward functions while maintaining indexability
$$\tilde{R}_i = R_i + \eta_i$$
3. Compute the Whittle-Index Policy
$$\pi(t+1) = Whittle(\tilde{R}_1, \tilde{R}_2, \ \ldots, \tilde{R}_N)$$

# Follow The Perturbed ***Whittle Index***

- **Key Idea 1:** Whittle Index acts like a low complexity optimization oracle for the RMAB problem, so incorporate it in FTPL

- **Key Idea 2:** Instead of perturbing the costs of policies, perturb the reward functions themselves

- **New Challenges** introduced:

    1. Create perturbations to **maintain indexability** structure
    2. Perturbations are **no longer i.i.d.** per expert/policy
    3. Whittle Index is an approximate but **not exact** maximizer

- **Our Contribution:** resolving these challenges!

**Algorithm 2:** Follow the Perturbed Whittle Leader

**Input** : parameter $\epsilon > 0$

Set $F_1^{(i)}(j) = j, \forall i \in \{1, ..., N\}, \forall j \in \{1, ..., M\}$

**while** $t \in 1, ..., T$ **do**

  Set $A^{(1)}, ..., A^{(N)} = \mathbf{1}$

  Sample $\delta_t^{(i)}(j) \sim$ uniform in $[0, 1/\epsilon]$, i.i.d. $\forall i \in \{1, ..., N\}$ and $\forall j \in \{1, ..., M\}$

  Compute $\gamma_t^{(i)}(j) = \sum_{k=1}^{j} \delta_t^{(i)}(k), \forall i, j$

  Choose scheduling policy

  $$\pi_t = \text{Whittle}\left(F_t^{(1)} + \gamma_t^{(1)}, ..., F_t^{(N)} + \gamma_t^{(N)}\right)$$

  Incur loss = $C_t(\pi_t)$ over epoch $t$ and observe feedback on $f_t^{(1)}, ..., f_t^{(N)}$

  In case of bandit feedback, construct cost estimates $\hat{f}_t^{(i)}, \forall i \in \{1, ..., N\}$ using linear interpolation

  Update

  $$F_{t+1}^{(i)} = \begin{cases} F_t^{(i)} + f_t^{(i)}, \forall i \in \{1, ..., N\}, & \text{if full feedback} \\ F_t^{(i)} + \hat{f}_t^{(i)}, \forall i \in \{1, ..., N\}, & \text{if bandit feedback.} \end{cases}$$

**end**

Monotone Perturbation

Whittle Index Scheduling

Accumulate Cost Functions

# Regret of FTPL

Given $N$ sources, $T$ epochs, $M$ time-slots per epoch and upper-bound $D$ on cost

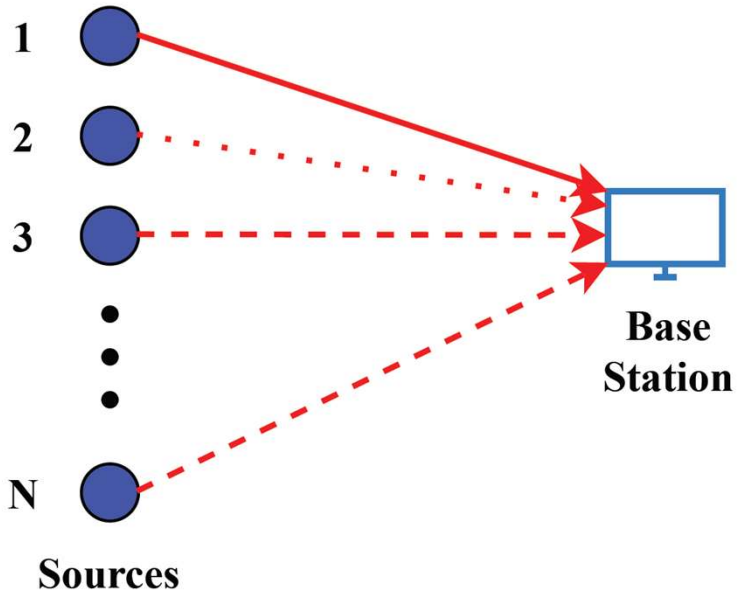
$$\boxed{\mathbb{E}[\text{Regret}_T(\text{FPWL})] \leq \alpha T + 2D\sqrt{2MNT}}$$

- $\alpha$ measures how close the Whittle-index solution is to optimality in the offline problem
- Specifically, for any two sets of cost functions $f_1, f_2, \ldots, f_N$ and $g_1, g_2, \ldots, g_N$ **assume**

$$\left| C_g\big(\text{Whittle}(f)\big) - C_g\big(\text{Opt}(f)\big) \right| \leq \alpha$$

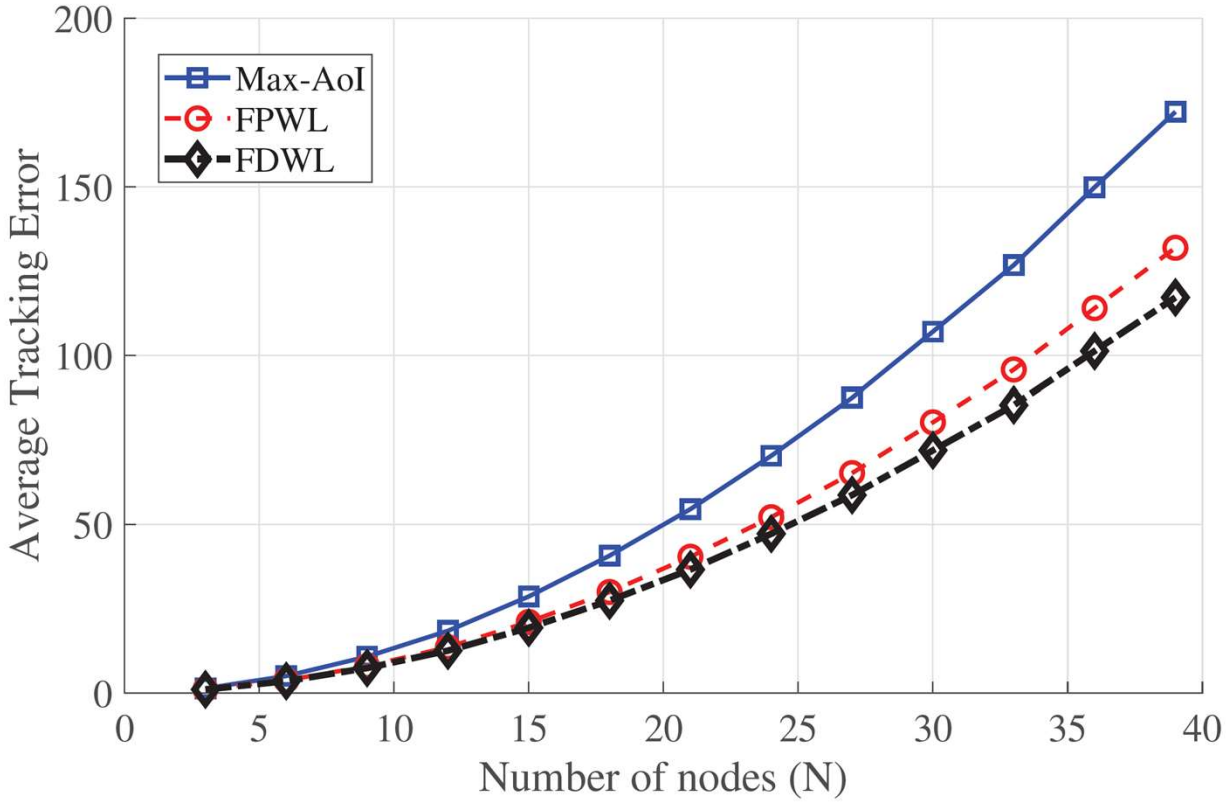
# Application to Wireless Scheduling

- Monitoring sources, with time-varying relative importance, over a wireless network, e.g. - mobility tracking
- Cost of stale information changes with time
- The static problem is an indexable RMAB



# Application to Wireless Scheduling

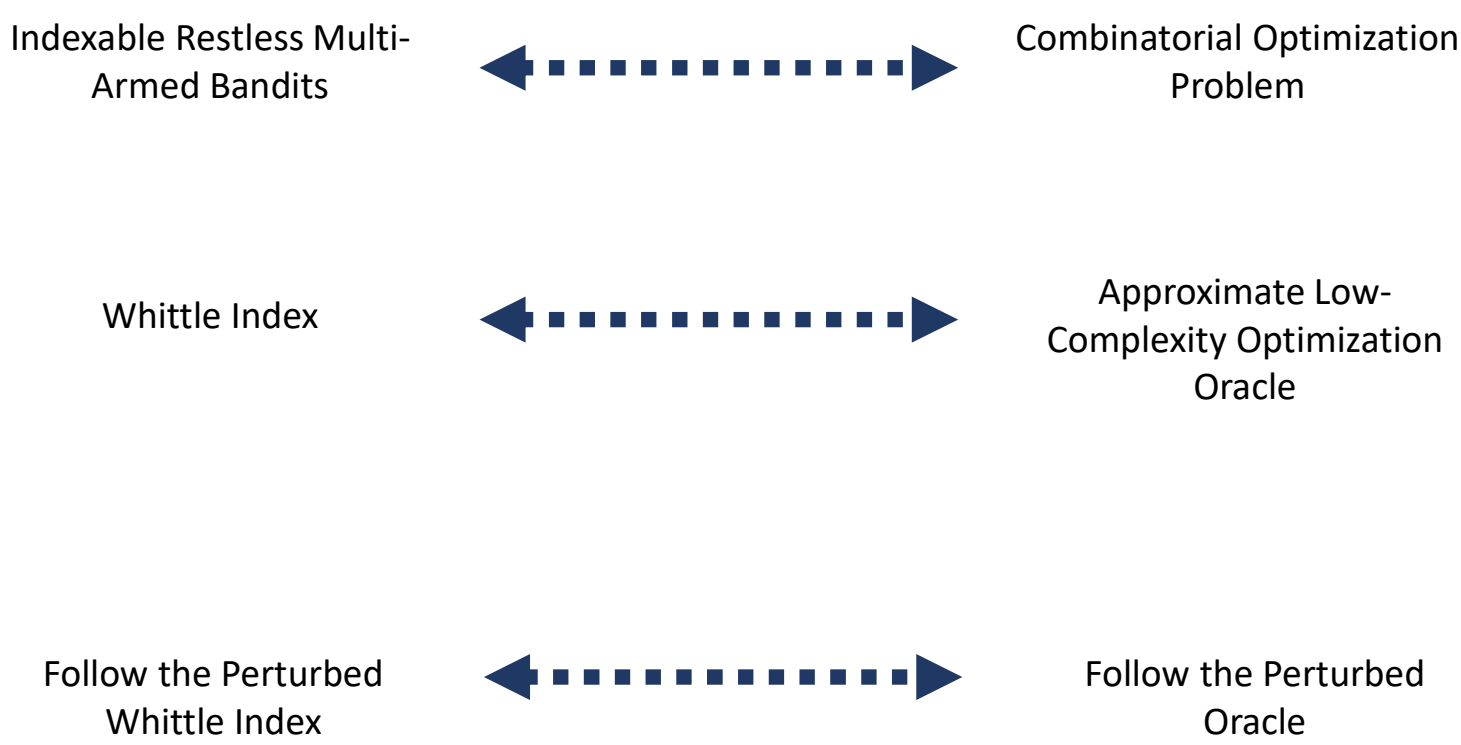
For time-varying Levy mobility ***Follow-the-perturbed-Whittle-Leader*** outperforms simple static policies



# The End

Questions?

# An Extension



# Restless Multi-Armed Bandits

## Restless Bandits: Activity Allocation in a Changing World

P. WHITTLE

**Abstract**

We consider a population of $n$ projects which in general continue to evolve whether in operation or not (although by different rules). It is desired to choose the projects in operation at each instant of time so as to maximise the expected rate of reward, under a constraint upon the expected number of projects in operation. The Lagrange multiplier associated with this constraint defines an index which reduces to the Gittins index when projects not being operated are static. If one is constrained to operate $m$ projects exactly then arguments are advanced to support the conjecture