Indexability and Whittle Index for Restless Bandit Problems Involving Reset Processes

Keqin Liu*, Richard Weber§, Qing Zhao*

*Department of Electrical and Computer Engineering, University of California, Davis
§Statistical Laboratory, University of Cambridge

Supported by ARO and NSF.
Reset Processes

- Given a set of $N$ independent random processes.
- The state—1 or 0—of each process stochastically evolves over time.
- At each time, activate $K$ ($K < N$) processes for observation and obtain a reward $r_n$ from the observed process $n$ if in state 1.
- After observing a process, a state-adaptive action is made to reset its state evolution: the probability that the process is in state 1 only depends on the last observed state and the time elapsed.
- No assumption on the time-correlation of the state of each process.
- Objective: maximize the expected long-run average reward.
Interactive Application: Supervisory Control/Monitoring

- $N$ physical (chemical/biomedical) processes.
- At each time, each process is either in abnormal state $1$ or normal state $0$.
- After process $n$ is observed in state $i \in \{0, 1\}$, an adaptive action (repair or maintenance) is applied such that the probability of being in state $1$ after $t$ time steps is given by $p^{(n)}_{i1}(t)$, where $p^{(n)}_{i1}(t)$ can be an arbitrary probability sequence over $t$.
- Under the objective of targeting at abnormal processes, a reward $r_n$ is obtained when process $n$ is observed in the abnormal state.
Non-Interactive Application: Dynamic Multi-Channel Access

- Dynamic Access: adapt to time-varying channel state.
- Channel State: “good (1)” or “bad (0)”.
- Applications: Cognitive Radios, Downlink Scheduling in cellular network, Opportunistic transmission, Jamming/Anti-jamming, target tracking.
- Limited Sensing: can only sense and access $K$ out of $N$ channels in each slot.

Which channels to sense in each slot?
Channel Sensing: Reset Information

Gilbert-Elliot Channel Model

After a channel $n$ is sensed in state $i \in \{0,1\}$, the probability that the channel is in state 1 after $t$ slots is given by the $t$-step transition probability $p_{01}^{(n)}(t)$ of the Markov chain.

If a channel $n$ is observed in state 1, a reward $r_n$ given by the channel capacity is obtained.
The Classic Multi-Armed Bandit (MAB)

- $N$ independent arms.
- Fully observable arm states $\{Z_n(t)\}$.
- Only one arm can be activated at each time.
- Activating arm $n$ yields a reward $R_n(Z_n(t))$.
- Activated arm changes state (Markovian).
- Passive arms are frozen.

Exploitation v.s. Exploration

- Exploitation: select the arm with the highest immediate reward.
- Exploration: select an arm for future benefit.

Optimal Policy—Gittins Index (Gittins’74)

- Attach an index to each state of each arm.
- Select the arm with the largest index at current time.
Restless MAB

Restless Multi-Armed Bandits (RMAB) \((Whittle'88)\)

- Passive arms also \textit{change states} and offer reward.
- Activate \(K\) \((K < N)\) arms simultaneously at each time.

Structure of the Optimal Policy:

- Not yet found.

Complexity:

- PSPACE-hard \((Papadimitriou&Tsitsiklis'99)\).
Whittle Index Policy

**Whittle Index** *(Whittle’88)*:
- Provide a *subsidy* $\lambda$ for passivity whenever the arm is made passive.
- Whittle index: the subsidy $\lambda$ that makes active and passive actions equally attractive at the current state.

**Performance:**
- Optimal under relaxed constraint on the average number of active arms.
- Asymptotically 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 general difficult to establish.
RMAB formulation of Reset Processes

- The state 0/1 of each process is only partially observable. Can not use it as the arm state.

- **Sufficient Statistic:** \( \{(i_n, t_n)\}_{n=1}^N \) where \( i_n \in \{0, 1\} \) is the last observed state of process \( n \) and \( t_n \) the time elapsed since that observation.

- State of each arm: \((i_n, t_n)\) transiting over time under the following Markov rule:

  \[
  (i_n, t_n) \rightarrow \begin{cases} 
  (0, 1), & \text{if observed in state } 0 \\
  (1, 1), & \text{if observed in state } 1 \\
  (i_n, t_n + 1), & \text{if unobserved}
  \end{cases}
  \]

- Activating an arm in state \((i, t)\) leads to immediate reward \( p_i(t) r_n \).
Consider the case that there is only one arm in the system.

A subsidy $\lambda$ is obtained whenever the arm is made passive.

Consider the problem of whether or not to activate the arm at each time to maximize the long-run average reward.

**Passive set**: The set of states on which the optimal action is being passive.

*Definition*: The arm is indexable if the passive set is monotonically increasing with subsidy $\lambda$. For each state of the arm, Whittle index is then defined as the minimum subsidy to draw the state into the passive set. An RMAB is indexable if all arms are indexable.
Indexability of RMAB with Reset Processes

► Sufficient to focus on one arm, drop the arm index for simplicity.
► Define $g$ as the maximum average reward over the infinite horizon and $\phi(i)$ the transient reward starting from state $i$ under the optimal policy.
► Define stopping time $t_i$ as the first activation time after observing the arm in state $i$.
► The system dynamic equation is given by

$$
\phi(i) = \max_{t_i \geq 1} \left\{ -gt_i + \lambda(t_i - 1) + (1 - p_{i1}(t_i))\phi(0) + p_{i1}(t_i)(r + \phi(1)) \right\}, \quad i = 1, 2.
$$

► Let $t_i^* \geq 1$ be the optimal first activation time after observing the arm in state $i$.
► Passive set: $\{(i, 1), (i, 2), \ldots, (i, t_i^* - 1)\}$.
► By analyzing the system dynamic equation, we have that $t_i^*$ is monotonically increasing with $\lambda$, indexability thus holds.
Whittle Index

Monotone Condition: $p_{11}(t)$ is decreasing with $t$ and $p_{01}(t)$ is increasing with $t$; $p_{11}(1) \geq p_{01}(t)$ for all $t \geq 1$; $p_{01}(t+1) - p_{01}(t)$ is strictly decreasing with $t$.

- The monotone condition always holds for positively-correlated Markov chains ($p_{11}(1) > p_{01}(1)$).

Under the monotone condition, Whittle index can be solved in closed form:

\begin{align*}
W(0, t) &= \frac{p_{01}(t)(t + 1) - p_{01}(t + 1)t}{1 - p_{11}(1) + tp_{01}(t) - (t - 1)p_{01}(t + 1)} r , \\
W(1, 1) &= p_{11}(1)r .
\end{align*}
Performance of Whittle Index Policy

- Near-optimal performance observed through numerical examples.

- Optimal for statistically identical arms, i.e., arms with the same parameters $p_{i1}(t)$ and $r$. 
Optimality of Whittle Index Policy for Statistically Identical Arms

For statistically identical arms, Whittle index policy is equivalent to the myopic policy that always chooses the arm with the largest expected immediate reward $p_i(t)r$.

Simple structure: stay the arm if observed in state $1$; otherwise switch to the arm observed longest time ago.

Whittle index policy is asymptotically optimal as $K/N \to 0$.

For Markov processes (non-interactive application), Whittle index policy is optimal for any $K$ and $N$ while both $K$ and $p_i(1)$ can be time varying (inhomogeneous Markov chains).
Related Work

For homogeneous Markov processes (non-interactive reset processes):

- Indexability and closed-form Whittle index: *Liu&Zhao’08.*

- Optimality of myopic policy for statistically identical arms:
  - $N = 2$: *Zhao&Krishnamachari’07.*
  - $K = N - 1$: *Liu&Zhao’08.*
  - General $N$ and $K = 1$ for positive correlated Markov processes: *Ahmad&Liu&Javidi&Zhao&Krishnamachari’09.*
  - General $N$ and $K$ for positive correlated Markov processes: *Ahmad&Liu’09.*
  - $N = 2, 3$ for negative correlated Markov processes: *Ahmad&Liu&Javidi&Zhao&Krishnamachari’09.*
Conclusion

- An RMAB formulation of reset processes.
- Indexability and closed-form Whittle index.
- Asymptotic optimality for statistically identical arms.
- Optimality for general inhomogeneous Markov processes in finite regime.