On Learning and Information Acquisition with Respect to Future Availability of Alternatives

Kazutoshi Yamazaki

Center for the Study of Finance and Insurance
Osaka University

The Econometric Society
2009 North American Summer Meeting

June 5, 2009
Value of Learning with respect to Availability

There are two coins, risky coin and safe coin:

- risky coin has been flipped twice: 1 head and 1 tail,
- safe coin has been flipped 20 times: 10 heads and 10 tails.

Consider a game where you sequentially choose to flip either risky coin or safe coin and you gain

- $1 if it comes up a head,
- $0 if it comes up a tail.

**Question 1:**

Assuming that you are risk-neutral, which coin would you flip?

**Question 2:**

What if risky coin is only available this time and NOT available from next time?
We study how to solve the tradeoff between \textbf{Exploitation} and \textbf{Exploration} where

- \textbf{Exploitation}: maximize the current immediate reward,
- \textbf{Exploration}: sacrifice the immediate reward in order to get more information for the future rewards,

using the \textit{multi-armed bandit} framework.

In particular, we study how the solution changes with respect to the availability of alternatives.
There are $N$ arms, and at each decision time $t = 0, 1, 2, \ldots$, an arm is played,
- a reward from the arm is obtained,
- state (or information) of the arm (active arm) is updated,
- states of the other arms (passive arms) are NOT updated.

Goal: maximize the expected total $\beta$-discounted reward over an infinite-time horizon for some $\beta \in (0, 1)$. 
Bandit Problems in Economics

The Bandit Problem is directly related to

- Learning by Doing,
- Research and Development (R & D),

and has applications in

- Job Search/Matching (e.g. Jovanovic (1979) and Miller (1984)),
- Asset Pricing (e.g. Rothschild (1974)),
- Mechanism Design (e.g. Bergemann and Valimaki (2006)).
The **Multi-Armed Bandit problem** is a **Markov Decision Process (MDP)** problem with states and actions at time $t$ denoted by $X(t) = (X_N(t), \ldots, X_N(t))$ and $a(t) = (a_1(t), \ldots, a_N(t))$, respectively, where for every $n = 1, \ldots, N$,

$$X_n(t) := \text{state of arm } n \text{ at time } t,$$

$$a_n(t) := \begin{cases} 1, & \text{if arm } n \text{ is played at time } t, \\ 0, & \text{otherwise,} \end{cases}$$

$X_n(t)$ evolves in a **Markov fashion** only when $a_n(t) = 1$ according to a transition matrix $P^{(n)} = (p_{xx'})_{x,x' \in \mathcal{X}_n}$ where $\mathcal{X}_n$ is the state space of $(X_n(t))_{t \geq 1}$.
Rewards are state-dependent:

\[ R_n(x) := \text{(expected) one-time reward} \]

obtained from arm \( n \) when its state is \( x \in X_n \),

and a (stationary) policy is a function that maps from \( X_1 \times \cdots \times X_N \) to

\[ A := \{(a_1, \ldots, a_N) \in (0, 1)^N : a_1 + \cdots + a_N = 1\} . \]

The objective is to obtain a policy \( \pi^* \) that maximizes the \( \beta \)-discounted total rewards:

\[
\max_{\pi} \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \beta^t \left( \sum_{n=1}^{N} R_n(X_n(t))1\{a_n(t)=1\} \right) \right] = \text{reward at time } t .
\]
The Gittins index policy plays at each time the arm with the largest Gittins index,

\[
\arg\max_{1 \leq n \leq N} G_n(X_n(t))
\]

where the index for fixed arm \( n \) is

\[
G_n(x) := \sup_{\tau > 0} \frac{\mathbb{E}^1 \left[ \sum_{t=0}^{\tau-1} \beta^t R_n(X_n(t)) \right] | X_n(0) = x}{\mathbb{E}^1 \left[ \sum_{t=0}^{\tau-1} \beta^t | X_n(0) = x \right]},
\]

- \( \mathbb{P}^1 \) is the probability measure induced by the policy where arm \( n \) is always played, and
- supremum is taken over all strictly positive \((X_n(t))_{t \geq 0}\)-adapted stopping times.

Note \( G_n(x) \) can be computed separately for each arm.
Introducing Availability

What if arms may become unavailable in the future?

- jobs whose openings are subject to their economic conditions;
- outdoor projects that can be engaged only on sunny days;
- medications that can be applied only to certain patients;
- baseball players who get injured frequently;
- oil-based products whose supply is unstable and unpredictable.

If arms are potentially unavailable in the future, the Gittins index policy is no longer optimal;

- rational decision makers do not use the Gittins index policy,
- how can we take into account the pessimism about the availability?
Exploitation vs Exploration with respect to Availability

1. If there exists an optimal policy, how does it look like?

where $\theta_n$ is the probability of availability of arm $n$.

2. Does there exist an optimal index policy just like Gittins?

Remark: It has been shown by Banks and Sundaram (1992) that an optimal index policy does NOT exist in the presence of switching costs.

3. If there is no optimal index policy, what is the best policy among all the index policies?
Main Results

There does not exist an optimal index policy.

However, there exists a strong index policy in the class of the Whittle index policy (Whittle 1988) such that

1. it generalizes the Gittins index policy;
2. the index converges to
   - the Gittins index as $\theta \uparrow 1$,
   - the immediate one-time reward as $\theta \downarrow 0$;
3. no index policy performs uniformly better;
4. numerically shown to be near-optimal.
Modeling Availability

We modify the multi-armed bandit problem as in the following:

Objective: \[ \max_{\pi} \mathbb{E}^{\pi} \left[ \sum_{t=0}^{\infty} \beta^t \left( \sum_{n=1}^{N} R_n(X_n(t), Y_n(t)) 1\{a_n(t)=1\} \right) \right] \]

where

\[ Y_n(t) := \begin{cases} 1, & \text{if arm } n \text{ is available at time } t, \\ 0, & \text{otherwise}, \end{cases} \]

\[ R_n(x, y) := \text{reward obtained from arm } n \]

with \( X_n = x, Y_n = y \) if it is played.

Because unavailable arms cannot be played, \( R_n(x, 0) = -\infty \).
Whittle Index Policy

The Whittle index policy plays the arm with the largest Whittle index

\[ W_n(x, 1) := \sup_{\tau \in \mathcal{S}} \mathbb{E}^{1,0} \left[ \sum_{t=0}^{\tau-1} \beta^t R_n(X_n(t), 1) 1\{Y_n(t)=1\} \bigg| X_n(0) = x, Y_n(0) = 1 \right] \]

\[ W_n(x, 0) := -\infty, \]

where

- \( \mathbb{P}^{1,0} \) is the probability measure induced by the policy where arm \( n \) is always played whenever available,

- \( \mathcal{S} \) is the set of positive stopping times at which the arm is available \( \mathbb{P}^{1,0} \)-a.s.
Relationship with Gittins Index/Immediate Reward

The **Whittle index** $W_n(x, 1)$

\[
\sup_{\tau \in \mathcal{T}} \mathbb{E}^{1, 0} \left[ \sum_{t=0}^{\tau-1} \beta^t R_n(X_n(t), 1) 1 \{Y_n(t) = 1\} \middle| X_n(0) = x, Y_n(0) = 1 \right]
\]

coincides with

- the **Gittins index** $G_n(x)$

\[
\sup_{\tau > 0} \mathbb{E}^1 \left[ \sum_{t=0}^{\tau-1} \beta^t R_n(X_n(t)) \middle| X_n(0) = x \right]
\]

when $Y_n(t) = 1$ a.s. for all $t \geq 1$,

- and the **immediate reward**

$R_n(x, 1)$

when $Y_n(t) = 0$ a.s. for all $t \geq 1$. 
Exploitation vs Exploration with respect to the Probability of Availability

Let $\theta_n$ be the probability of availability of arm $n$.

For every arm $n$, we have
- $W_n(x, 1) \uparrow G_n(x)$ as $\theta_n \uparrow 1$,
- $W_n(x, 1) \downarrow R_n(x)$ as $\theta_n \downarrow 0$,
uniformly over $\mathcal{X}_n$ (set of states arm $n$ can take).
Non Existence of Optimal Index Policies

1. If there exists an optimal index policy, its index must be a strict monotone transformation of the Whittle index;

2. No index policy performs uniformly better than the Whittle index policy in every instance of the problem;

3. An optimal index policy does not exist because there are some examples where the Whittle index policy fails to be optimal.
The Coin Example Revisited

Risky Coin and Safe Coin currently have unknown success probabilities $\lambda_R$ and $\lambda_S$, respectively, such that

$$\lambda_R \sim Beta(1, 1) \quad \text{and} \quad \lambda_S \sim Beta(10, 10).$$

Their states form (controlled) Markov chains $(a_R(t), b_R(t))_{t \geq 0}$ and $(a_S(t), b_S(t))_{t \geq 0}$ such that

- $(a_R(0), b_R(0)) = (1, 1)$ and $(a_S(0), b_S(0)) = (10, 10)$

with common transition probabilities (if flipped)

- $(a, b) \rightarrow (a + 1, b)$ with probability $a/(a + b)$;
- $(a, b) \rightarrow (a, b + 1)$ with probability $b/(a + b)$;

and (if NOT flipped)

- $(a, b) \rightarrow (a, b)$ with probability 1.

Moreover, we have

$$R_i((a_i(t), b_i(t))) = \frac{a_i(t)}{a_i(t) + b_i(t)}, \quad i \in \{R, S\}.$$
Numerical Results

Comparison with the Gittins index policy where Gittins index policy chooses the arm with the largest Gittins index, and ties are broken randomly. The Gittins index is calculated regardless of the probability of availability.

- Case 1: \(N/2\) arms w/ \(\theta = (1, 0, 0.5)\)
- Case 2: \(N/3\) arms w/ \(\theta = (1.0, 0.7, 0.3)\),
- Case 3: \(N/6\) arms w/ \(\theta = (1.0, 0.9, 0.7, 0.5, 0.3, 0.1)\).

<table>
<thead>
<tr>
<th></th>
<th>Case 1</th>
<th></th>
<th>Case 2</th>
<th></th>
<th>Case 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Whittle</td>
<td>Gittins</td>
<td>Whittle</td>
<td>Gittins</td>
<td>Whittle</td>
<td>Gittins</td>
</tr>
</tbody>
</table>
Thank you!

Questions?