| | |
|---|---|
| INSE6300 | Jia Yuan Yu |

# 8: Gittins index

| | |
|---|---|
| Concordia | April 5, 2016 |

In this lecture, we consider a special type of MDP often encountered in supply chains and an efficient "divide-and-conquer" solution for it.

## 1 Multiarmed bandit

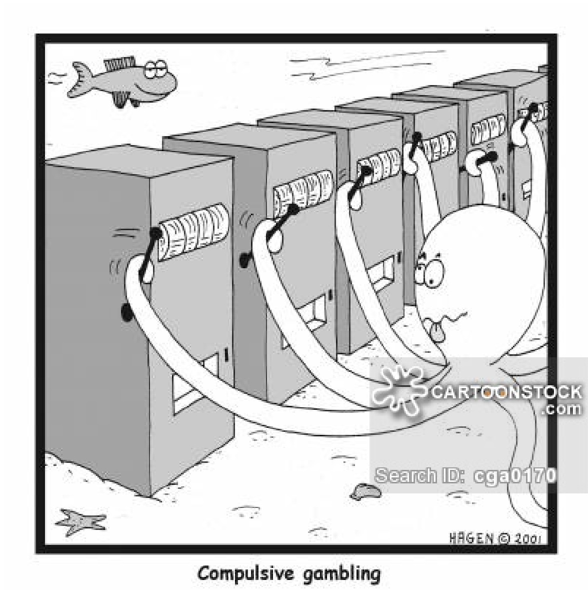


Figure 1: By Hagen, 2001.

Consider the following sequential decision problem (special case of an MDP). The time steps are $t = 1, 2, \ldots$ At every time step, the decision maker must choose one of $n$ alternative actions (imagine the case $n = 2$). The state of the system is $n$-dimensional random variable:

$$X_t = (X_t^1, \ldots, X_t^n) \in \mathbb{X}^n,$$

where $\mathbb{X}$ is a finite or countable set for simplicity.

Let $f^1, \ldots, f^n$ denote fixed deterministic functions, and let $\{W_1^1, W_2^1, \ldots\}, \ldots, \{W_1^n, W_2^n, \ldots\}$ denote i.i.d. sequences that are also mutually independent. If the $i$-th action is chosen, i.e., $A_t = i$, then the state transition is as follows

$$\begin{aligned}
X_{t+1}^1 &= X_t^1, \\
&\dots \\
X_{t+1}^i &= f^i(X_t^i, W_t^i), \\
&\dots \\
X_{t+1}^n &= X_t^n.
\end{aligned}$$

Observe that these transitions are Markovian, and only the $i$-th element changes.

The reward corresponding to action $A_t$ at time $t$ is

$$\sum_{k=1}^{n} r^k(X_t^k) 1_{[A_t=k]},$$

which depends only on the reward function for the arm $A_t$ that is chosen. The expected total discounted reward with discount factor $\beta \in (0,1)$ and given initial condition $X_0$ is

$$\mathbb{E} \sum_{t=0}^{\infty} \beta^t \sum_{k=1}^{n} r^k(X_t^k) 1_{[A_t=k]},$$

where the expectation is over the $\{W_t^i\}$ random variables.

**Example 1.1** (Advertising and pricing)**.** There are $n$ possible ads (prices). The state is the success frequency of each ad (price).

Due to the curse of dimensionality, when $n$ is large, typical dynamic programming solutions are not appropriate (complexity proportional to $\mathbb{X}^n$).

# 2 Gittins index

One approach to solve the above bandit problem is to assign a number to each pair $(x, i)$ of state $x \in \mathbb{X}$ and bandit $i$:

$$\nu^i(x) = \sup_{\tau} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau} \beta^t r^i(X_t^i) \mid X_0^i = x\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau} \beta^t \mid X_0^i = x\right]},$$

where the sup is taken over all stopping times[1], or alternatively:

$$\nu^i(x) = \sup \left\{ \lambda : \sup_{\ell > 0} \mathbb{E}\left[ \sum_{t=0}^{\ell} \beta^t (r^i(X_t^i) - \lambda) \mid X_0^i = x \right] \right\}.$$

This is the Gittins index of bandit $i$ at state $x$. Then, at every time step, the optimal action is to choose an action with the highest Gittins index.

[1]For all $t$, the event $\tau = t$ depends only on $X_0, X_1, \dots, X_t$.

There are various ways to compute the Gittins index. Consider a single bandit (with another action giving 0 reward). Instead of stopping rules, the Gittins index can also be defined in terms of stopping sets:

$$S(x) = \{x' \in \mathbb{X} : \nu(x') \leq \nu(x)\}.$$

The Gittins index is then:

$$\nu^i(x) = \sup_{S(x) \subseteq \mathbb{X}} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau(S(x))} \beta^t r^i(X_t^i) \mid X_0^i = x\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau(S(x))} \beta^t \mid X_0^i = x\right]},$$

where $\tau(S(x)) = \inf\{t > 0 : X_t \in S(x)\}$.

# 3 Computing the Gittins index

Consider a single bandit (with another action giving 0 reward).

- Input: $r$, $f$, distribution of $\{W_t\}$.
- Find the state with the highest index (with stopping set $S(x_{(1)}) = \mathbb{X}$): $x_{(1)} = \arg\max_{x \in \mathbb{X}} r(x)$.
- Output $\nu(x_{(1)}) = r(x_{(1)})$.
- Find $x_{(2)}$ with the second highest index, where the stopping set is $S(x_{(2)}) = \mathbb{X} \setminus x_{(1)}$. Compute the probability distribution of $\tau(S(x_{(2)})) = \inf\{t > 0 : X_t \neq x_{(1)}\}$.
- Output $\nu(x_{(2)})$.
- Continue for $x_{(3)}, x_{(4)}, \ldots$

# 4 References


- "Multi-armed bandits, Gittins index, and its calculation." J. Chakravorty and A. Mahajan, Methods and Applications of Statistics in Clinical Trials, 2014.