# MS&E 235, Internet Commerce

### Stanford University, Winter 2007-08

**Instructor: Prof. Ashish Goel, Notes Scribed by Shrikrishna Shrin**

# Lecture 5: Multi Armed Bandit Problem

## Description of the problem

A multi-armed bandit, also sometimes called a K-armed bandit, is a simple machine learning problem based on an analogy with a traditional slot machine (one-armed bandit) but with more than one lever. When pulled, each lever provides a reward drawn from a distribution associated to that specific lever. The objective of the gambler is to maximize the collected reward sum through iterative pulls. It is classically assumed that the gambler has no initial knowledge about the levers. The crucial trade off the gambler faces at each trial is between "exploitation" of the lever that has the highest expected payoff and "experimentation" to get more information about the expected payoffs of the other levers. (From Wikipedia.org)

In the context of Internet Commerce, advertisements can be modeled as the arms of slot machines. Henceforth, the words machines, arms and advertisements will be used interchangeably.

The multi armed bandit problem can also be used to determine which products to offer to a customer.

As mentioned in the description of the problem, we have a choice between which machine to play. Let $\theta$ be the discount factor (how much a dollar tomorrow is worth today). Our goal is to maximize the total discounted expected reward. What this means is that we want to choose a strategy of which machine to play and at what time such that we maximize the present value of expected rewards that our strategy will generate.

In order to model the distribution associated with each lever, let us assume that arms have something called the $(\alpha, \beta)$ beliefs or priors associated with them which represents your belief about the chances of a win on them.

For an $(\alpha, \beta)$ machine, $(\alpha + \beta)$ represents the total number of experiments that have been carried out for that machine resulting in $\alpha$ successes and $\beta$ failures.

If we assume that the reward is 1 each time you succeed, then from a state of $(\alpha, \beta)$, with a probability of $\alpha/(\alpha + \beta)$, the machine transitions to the state $(\alpha+1, \beta)$ and a reward of 1 is earned during the transition and with probability of $\beta/(\alpha + \beta)$, the machine transitions to a state $(\alpha, \beta + 1)$ and no reward is earned during this transition (that is reward = 0).

The pair $(\alpha, \beta)$ represents one of the states in which a machine can be.

## Example of a machine

A machine can be thought of in terms of the various states it can occupy, the transition rules between those states and the reward that the machine generates during those transitions. For example, consider a machine in an initial state (state A). Suppose this machine is played (its lever is pulled), we get a reward of $r$ and the machine transitions to another state (state B) and on subsequent plays, it remains in this state (state B) without generating any reward.

## Deterministic Machine/Reference Machine

A reference machine of type $r$ is one that begins in an initial state and when played, it generates a reward $r$ and returns to its initial state. That is, its final state is the same as its initial state and in each cycle it generates a reward of $r$.

A reference machine of type $r$ can be thought of as an $(\alpha, \beta)$ machine in the limiting case, $\alpha \to \infty$, $\beta \to \infty$ and $\alpha/(\alpha + \beta) = r$

**How to determine which machine is the best?**

1. A) Pair-wise comparison: One way to find the best machine is by pair-wise comparison. Given a method to compare two machines, machines can be compared pair-wise to find the best machine.

2. B) A more efficient way to determine which machine is the best is by computing the value of Gittins' index ($g$) for each machine using the concept of a reference machine. In this approach, we compute the value of $g$ for all machines and the machine with the highest value of $g$ is the best machine. This is discussed in the following section.

## Gittins' Index

Theorem: There exists a function $g(\alpha, \beta, \theta)$ such that it is an optimum strategy to play the arm $i$ with the highest value of $g(\alpha_i, \beta_i, \theta)$.

The value $g$ for a machine is also known as the Gittins' index for that machine.

## Computing Gittins' index

Consider two reference machines machine A with $r_A = .5$ and machine B with $r_B = .25$. Among these two machines, it is always better to play machine A as it has a higher value of $r$.

Consider a case where $\theta = .9$ and we are given an $(\alpha, \beta)$ machine with $\alpha = 1$, $\beta = 1$ and a reference machine with $r = .5$. In the beginning, the $(\alpha, \beta)$ machine has an expected reward of 0.5 (same as the reference machine). It is therefore better to play the $(\alpha, \beta)$ machine as there is a chance that it might be better than the reference machine. If on playing the machine, the machine moves to the state $(\alpha + 1, \beta)$, then we play this machine again. If it moves to the state

$(\alpha, \beta + 1)$, we have the option of playing the reference machine. We can always come back to the reference machine so in this case it is better to explore the $(\alpha, \beta)$ machine first.

Suppose in the example the reference machine had a $r = 1$, then it would always be a better strategy to play the reference machine as the expected reward of the $(\alpha, \beta)$ machine can never be greater than 1.

In general, if at a value of $r$ greater than or equal to $r = m$, it is a better strategy to play the reference machine and for a value of $r$ lesser than or equal to $r = n$, it is a better strategy to play the $(\alpha, \beta)$ machine, then there is a value of $r \in (m, n)$ for which you are indifferent between playing the $(\alpha, \beta)$ machine or the reference machine during the first step. Let us denote this value by $r^*$.

Now, the first time we play this reference machine we get a reward of $r^*$, the second time we again get a reward of $r^*$ which has to be discounted by $\theta$ making the discounted reward equal to $r^*\theta$ and so on.

Therefore the total discounted expected reward obtained by playing the reference machine again and again equals:

$$r^* + r^*\theta + r^*\theta^2 + r^*\theta^3 + .....$$
$$= r^*/(1 - \theta)$$

This is the value of the Gittins' index for the $(\alpha, \beta)$ machine.

Gittins' index of an $(\alpha, \beta)$ arm denoted $g(\alpha, \beta, \theta)$ is $r^*/(1-\theta)$ where $r$ is the reference arm such that given the reference arm and the $(\alpha, \beta)$ arm, an optimum strategy is indifferent between which one to play at time 1.

Gittins' index is $g = r^*/(1 - \theta)$.

The total expected profit given an $(\alpha, \beta)$ arm and a reference arm with $r = I(1 - \theta)$ is given by:

$$V(\alpha, \beta, I) = \max(I, (\alpha/(\alpha+\beta))(1+\theta*V(\alpha+1, \beta, I))+(\beta/(\alpha+\beta))(\theta*V(\alpha, \beta+1, I)))$$

If this equation is implemented in Excel, then the indifference point or the Gittins' index for an $(\alpha, \beta)$ machine is the smallest value of the index at which the table entry $V(\alpha, \beta) = I$. This is the value $g(\alpha, \beta, \theta)$ of the Gittins' index.