CS261 Winter 2016 - 2017
Lecture 11 & 12: Multi-Armed Bandit Problem

Instructor: Ashish Goel
Scribe: Kaidi Yan

Feburary 16, 2017

1 Problem Definition

We are going to study the multi-armed bandit problem, which is also useful in making online sequential decision making. Recall the second teaser problem in Lecture #1: there is a gambler and n slot machines (or "one-armed bandit"s) — each arm, when played, will give some reward from a probability distribution specific to that arm. The gambler keeps on playing for infinitely long time, and for each unit of time the gambler chooses to play exactly one of the N arms. The gambler's goal is to decide which arm to play each time so as to maximize the total expected reward at present with discount factor θ. Here θ is a constant between 0 and 1 determining the present value of a reward in the future – if an arm gives a reward r at time t in the future, then its present value is $\theta^t r$.

As a concrete example, consider a simple Bayesian reward model for arms. Each arm A_i is represented by (α_i, β_i), where α_i is the total number of successes of A_i the gambler has observed, and β_i is the total number of failures of A_i the gambler has observed[1]. We also define a [0,1] reward: if the arm succeeds then the reward is 1; otherwise the reward is 0.

Suppose the gambler plays an arm (α, β) — with probability $\alpha/(\alpha+\beta)$ the gambler gets reward 1 and the arm become $(\alpha+1, \beta)$; with probability $\beta/(\alpha+\beta)$ the gambler gets reward 0 and the arm become $(\alpha, \beta+1)$. This model for an arm is called the *Beta Prior*.

If there are two arms A and B with Beta Prior $(1, 2)$ and $(10^5, 10^5)$ respectively, which arm should the gambler choose to play at time 0? Instead of delving into too much detail, let's just consider two extreme cases:

1. If $\theta = 0$, then future rewards really don't matter to the gambler since $\theta^t r = 0$ for all $t > 0$. Thus the gambler should play B as B gives an expected reward of $1/2$ while A only gives an expected reward of $1/3$ (we don't care about what A and B become after time 0). Here we say the gambler "exploits" B as the choice is completely based on the current bayesian reward model for each arm.

[1]For technical reasons, $\alpha_i - 1$ is often described as the number of successes and $\beta_1 - 1$ as the number of failures, but those reasons are not important here.

2. If $\theta = 1 - 10^{-40}$, then future rewards are almost as important as the present reward. We can show that the gambler, in this case, should choose to play A at time 0 to see whether A succeeds or not. In other words, when future rewards matter a lot we should "explore" A first in order to find out a more accurate reward model for A.

This example, though simple, demonstrates a trade-off between exploitation and exploration which is very important in design of algorithms for sequential decision making.

2 General Setting

A more general Bayesian multi-armed bandit problem is defined as follow. We are given n arms and the discount factor θ ($0 \leq \theta \leq 1$). Assume arm A_i has a finite or countable state space S_i and is currently at state a_i. For each state $a \in S_i$, playing A_i gives an expected reward $r_i(a)$ with $0 \leq r_i(a) \leq 1$. Arm A_i is also associated with a transition matrix P_i: $P_i(a, b)$ indicates the probability of A_i going from state a to b when played in state a (Note that this probability is independent from time; hence we can think of it as a Markov process). The gambler chooses one arm to play each time, and the goal is to maximize $\sum_{t=0}^{\infty} \theta^t E[\text{reward at time } t]$, which is the total expected reward, by picking the "optimal" arm to play at each time.

In 1979, J.C. Gittins and D.M. Jones proved the following theorem:

Theorem 1 *There exists a function $g(S_i, a_i, r_i, P_i, \theta)$ such that alway playing arm A_{i^*} with the highest $g(S_{i^*}, a_{i^*}, R_{i^*}, P_{i^*}, \theta)$ value is an optimum solution for the multi-armed bandit problem.*

This function g is called the *Gittins Index*. Here we skip the proof to this theorem. Notice that this result is non-trivial since g's value only depends on each individual arm, meaning that we can make the optimal choice by ignoring interactions between different arms. This is much more efficient than the straightforward dynamic programming approach which needs to take all arms into consideration, and hence will have an exploding state space.

Interestingly, Theorem 1 is not just an existential theorem — it also indicates a practical way for the gambler to choose which arm to play at each time. To show that, we first define a family of *reference arms* $\{R_p : 0 \leq p \leq 1\}$:

Definition 2.1 *A reference arm R_p ($0 \leq p \leq 1$) is an arm that always gives reward 1 with probability p and reward 0 with probability $1 - p$.*

It's easy to see that if we are only given a list of arms R_p's in the original multi-armed bandit problem, then it's always best to play R_p with the largest p. Hence, we can simply define Gittins Index g for these arms to be any strictly increasing function of p and Theorem 1 is satisfied immediately. For convenience we define $g(p, \theta) = p/(1 - \theta)^2$.

Suppose we now have two arms A_i and R_p with A_i in state a at time 0. For some fixed value p^*, if somehow we know the optimal choice at time 0 is to play R_p whenever $p > p^*$

[2]We are overloading the notation of g here.

and to play A_i whenever $p < p^*$, then we know $g(A_i, \theta) = g(R_{p^*}, \theta)$ from Theorem 1. The value p^* is called the *indifference point* for arm A_i in state a, namely the optimal strategy is indifferent between playing A_i in state a and R_{p^*} at time 0. Thus, finding the Gittins Index g for A_i is reduced to finding the indifference point p^* for A_i; after that we can simply set $g(A_i, \theta) = g(R_{p^*}, \theta) = p^*/(1-\theta)$.

Now let's fix p, and let $v(a)$ be the optimal total expected rewards the gambler can achieve given arm A_i and R_p, assuming that A_i is in state a at time 0.

For the optimal strategy at time 0, there are two possibilities: the gambler chooses to play either A_i or R_p. Below we analyze these two cases:

1. If the gambler chooses to play R_p at time 0, then at time 1 the two arms are still A_i and R_p with A_i in state a. Hence the gambler will keep on playing R_p. Therefore

$$v(a) = p + \theta p + \theta^2 p + ... = p/(1-\theta) = g(R_p, \theta)$$

2. If the gambler chooses to play A_i at time 0, then with probability $P_i(a, b)$ arm A_i will be in state b at time 1. Hence

$$v(a) = r_i(a) + \theta \sum_{b \in S_i} v(b)$$

Overall, we choose the arm which gives a higher total expected reward. Therefore:

$$v(a) = max(p/(1-\theta), r_i(a) + \theta \sum_{b \in S_i} P_i(a, b) v(b)) \tag{1}$$

If playing R_p is at least as good as playing A_i in state a, then we should also have:

$$v(a) = g(R_p, \theta) = p/(1-\theta) \tag{2}$$

Thus, if we can find the smallest p such that constraints (1) and (2) are both satisfied then the Gittins Indx of A_i in state a is $g(R_p, \theta) = p/(1-\theta)$. In Homework 3 you are going to translate these constraints into a linear program and solve for p (or for $g(R_p, \theta)$ directly).

While this result looks away from our regular class flow, it is important to see how slightly different models of very similar underlying problems can lead to very different algorithms and approaches. The "experts-algorithm" makes very few assumptions and gets near-zero regret but only against one expert. Gittins Index gets optimality but with stronger assumptions.