About this class

An example

Bandit problems in general

Two-armed bandits

Multi-armed bandits and Gittins indices

An Example

[Most of this lecture from Berry & Fristedt]

You want to maximize the sum of two observations. The process works as follows. At time 1, you can select either "Arm 1," whose payoff is a random variable, or you can select "Arm 2," whose payoff is some fixed and known λ. You will face the same choice at time 2.

For the moment, let's assume that the payoff of Arm 1 is $N(\theta, 1)$ and your prior on θ is $N(\mu, \rho^2), \rho^2 > 0$

What is the difference in the decisions you would make at times 1 and 2?

At time 2 it always makes sense to be myopic.

What is a strategy in this case? A mapping from a history of observations to an action.

Let's find the best strategy that chooses arm 2 at time 1.

At Time 2, what should we choose? Arm 1 if $\mu > \lambda$, Arm 2 otherwise. Then the value of the process under this strategy is $\lambda + \max(\lambda, \mu)$

Here's something interesting. If it makes sense to choose Arm 2 at Time 1 then it must make sense to choose Arm 2 at Time 2 as well. Why? We'll show this in a somewhat more general framework a little bit later...we don't actually need it right now, though

Now the best strategy that chooses Arm 1 at Time 1

First, the update of the mean of my belief about Arm 1 given that I observe X_1 when I pull it is:

$$\frac{\mu + \rho^2 X_1}{1 + \rho^2}$$

So what will I do at Time 2? I'll choose Arm 2 iff

$$\frac{\mu + \rho^2 X_1}{1 + \rho^2} \leq \lambda$$

So now what do these two things taken together tell us about what action to take at Time 1? Well, the value of pulling Arm 1 is:

$$\mu + E[\max(\frac{\mu + \rho^2 X_1}{1 + \rho^2}, \lambda)]$$

The value of pulling Arm 2 is:

$$\lambda + \max(\lambda, \mu)$$

We only need to compare with 2λ in this case because the second value $(\mu + \lambda)$ could then be achieved by pulling Arm 1 at Time 1 and then Arm 2 at Time 2.

So in order to choose Arm 1, we need:

$$\mu + E[\max(\frac{\mu + \rho^2 X_1}{1+\rho^2}, \lambda)] > 2\lambda$$

$$\Rightarrow \mu - \lambda + E[\max(\frac{\mu + \rho^2 X_1}{1+\rho^2} - \lambda, 0)] > 0$$

We won't go into the details of solving this, but it is doable, and in fact, the solution is of the following form.

Let

$$t = (\lambda - \mu)\frac{\sqrt{1+\rho^2}}{\rho^2}$$

$$\Psi(t) = \int_t^\infty (x-t)N(x)dx$$

$$= N(t) - t(1 - \Phi(t))$$

So basically the breakeven point will come for some t_0 where $\Psi(t_0) = t_0$. Numerically $t_0 \simeq 0.2760$

Then, if $t < t_0$, at Time 1, play Arm 1, otherwise play Arm 2. Then update your beliefs, and at Time 2, only play Arm 1 if the mean of your new belief is $> \lambda$.

What can we say about μ and λ?

Well, if $\mu > \lambda$ then it always makes sense to play Arm 1. But if μ is smaller, it depends on p. In fact, note that $\frac{\sqrt{1+\rho^2}}{\rho^2} \to 0$ as $\rho \to \infty$. This means that for sufficiently large uncertainty it always makes sense to play the uncertain Arm at Time 1!

Bandit Problems: A More General Description

You can have many arms. In general we'll assume they're independent and work with a few different reward structures. Each arm can also be thought of as having a Markovian structure, but we won't worry about that complication for the most part.

What is the problem with just thinking about states and using value functions? Our posteriors have to somehow be folded into the state description. This is not necessarily easy.

We'll see some remarkable things in the multi-armed bandit case for independent arms, but first let's look at some very simple approaches.

ϵ-greedy Methods

Greedy methods: Pull the arm with the best historical reward that has been achieved so far

Problem: may not learn enough about arms that initially seem suboptimal

ϵ-greedy: with probability ϵ, pull an arm uniformly at random

Flow utility vs. asymptotic learning for different ϵ values

Can also use ϵ declining over time to try and make the best of all worlds

Other methods: use an exploration schedule, and then always exploit after that.

ϵ-soft methods:

$$\frac{\exp(Q_t(a)/\tau)}{\sum_b \exp(Q_t(b)/\tau}$$

where τ is the temperature

These methods are surprisingly effective in general and in real-world problems