CS 6840: Algorithmic Game Theory **Spring 2017**

Lecture 36: April 28

Lecturer: Robert Kleinberg *Scribe: Matvey Soloviev*

Apparently, the lecture is largely taught from a slide deck, which Bobby said he will make available later. I will note down any relevant-looking out-of-band remarks.

36.1 Multi-Armed Bandits and the Gittins Index

Unlike the games we have seen so far, multi-armed bandits have only one player and no notion of equilibria. However, the connections are manifold, and the MAB is one of the most influential models of sequential learning in use nowadays.

36.1.1 Example: updating beliefs about MAB arm states

If the reward that a given arm generates is distributed according to

$$r = \begin{cases} 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p \end{cases}$$

and so supported on $\{0, 1\}$. Suppose my prior belief is that

$$\Pr\left[p = \frac{1}{4}\right] = \frac{1}{2}$$

and

$$\Pr\left[p = \frac{3}{4}\right] = \frac{1}{2},$$

i.e. there are two possible types of arm, which I consider equally likely.

Now, the arm is pulled once. What posterior belief will I hold in the case of each nonzero-probability outcome?

Suppose we observe 0. We have

$$\Pr\left[0 \mid p = \frac{1}{4}\right] \cdot \Pr\left[p = \frac{1}{4}\right] = \frac{3}{8}$$

and

$$\Pr\left[0 \mid p = \frac{3}{4}\right] \cdot \Pr\left[p = \frac{3}{4}\right] = \frac{1}{8}.$$

So by Bayes,

$$\Pr\left[p = \frac{1}{4} \mid \text{observe } 0\right] = \frac{\frac{3}{8}}{\frac{1}{2}} = \frac{3}{4},$$

and

$$\Pr\left[p = \frac{3}{4} \mid \text{observe } 0\right] = \frac{\frac{1}{8}}{\frac{1}{2}} = \frac{1}{4}.$$

A similar calculation will show that conditioned on observing 1, the posterior probabilities of $p = \frac{3}{4}$ and $p = \frac{1}{4}$ are $(\frac{1}{4}, \frac{3}{4})$ instead. So if we consider "states" that encode our current beliefs as a pair of probabilities, pulling an arm at state $(\frac{1}{2}, \frac{1}{2})$ results in a transition to one of two states $(\frac{3}{4}, \frac{1}{4})$ and $(\frac{1}{4}, \frac{3}{4})$.

In general, the number of possible transitions out of each state will depend on the size of (union of) the support(s) of the outcome distribution of each arm type, and so in particular possibly be infinite.

36.1.2 Remarks on complexity of the problem

If the Markov chains for each arm are finite (which they, notably, are not in the above example!), we can show with some work that outputting the optimal policy is in PSPACE in general. The algorithm involves backpropagating the optimal policy (and its value) from the absorbing states of each chain, which is possible in PSPACE as each state only has a polynomial number of successors.

For examples with infinite Markov chains (even very structured ones such as the one we just saw!), we can't even assure that the Gittins Index has a closed-form representation in general, let alone output the optimal policy.

36.1.3 Example for when choosing greedily arm is suboptimal

As a corollary to the Gittins Index Theorem, if there is any counterexample to the greedy algorithm (that is, choosing the arm with the highest expected reward) being optimal, there must be a two-armed counterexample.

Set

$$\text{Arm 1:} \quad \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \text{ with } p = \begin{cases} 1 & \text{with prob. } 0.1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Arm 2:} \quad \frac{1}{2} \text{ guaranteed}$$

and $\gamma = 0.99$.

Then the first pull of Arm 1 will only give a payoff of 1 with probability 0.1, but if it does, we know we can get 1 forever; otherwise, after that first pull, we will know that it is of type $p = 0$ and so can just revert to pulling the second arm (for guaranteed payoffs of 0.5). So the sequence of expected payoffs from pulling the first arm and then pulling whichever arm is known to be better afterwards is

$$(0.1, 0.55, 0.55, 0.55, \ldots),$$

whereas the sequence of expected payoffs from always pulling the second arm greedily is

$$(0.5, 0.5, 0.5, 0.5, \ldots),$$

with the exponentially discounted reward for the first sequence being greater:

$$0.1 + \frac{0.55}{1-\gamma} = 55.1 > 0.5 + \frac{0.5}{1-\gamma} = 50.5.$$

Think of it as a decision between trying to go to grad school for a year and just taking that sweet Facebook job offer right after graduation: you certainly just expect to suffer in the short term, but assuming you don't prioritise short-term pleasure too much (γ is sufficiently high), the possibility of a bright degreed future (or, if it doesn't work out, dropping out and getting your perpetual 0.5 fallback utility starting a year later) is worth the sacrifice in the long run.

36.1.4 Computing the Gittins Index

There is an $O(n^6)$ algorithm for it. It's not obvious how to do it in $O(n^6)$, but it would be great if someone could come up with a better algorithm; the bound is not known to be tight nor does there seem to be any intuition that it should be that way.