IEOR 8100-001: Learning and Optimization for Sequential Decision Making 03/28/16

Lecture 18: Proof of Gittins Index theorem Part 2

Instructor: Shipra Agrawal Scribed by: Hong Bin Shim

Last time, we have seen that in Markovian Bandits, the optimal policy is Markovian and stationary. Today, we prove the Gittins Index Theorem which states:

Theorem. *For every arm a, and every state $s \in S_a$, the optimal policy is to choose arm*

$$a_t = \arg\max_a G_a(s)$$

where

$$G_a(s) = \sup_{\tau \geq 1} \frac{E[\sum_{t=1}^{\tau} r_a(s_t)\gamma^{t-1} | s_1 = s]}{E[\sum_{t=1}^{\tau} \gamma^{t-1} | s_1 = s]}$$

and τ is the stopping time.

Note that the numerator is the expected total discounted reward from playing the arm from time 1 to *tau*, and the denominator is the expected discount over time 1 to τ. When calculating Gittins Index, we only consider a single arm problem playing the arm from time 1 to τ.

Throughout this lecture, let (s^*, a^*) be the state-action pair with the maximum reward, i.e.

$$\begin{aligned}(s^*, a^*) &= \arg\max_{a, s \in S_a} r_a(s) \\ &\implies r_{a^*}(s^*) \geq r_a(s) \forall a, s \\ &\implies G_{a^*}(s^*) \geq G_a(s) \forall a, s\end{aligned}$$

To see this, note that $G_{a^*}(s^*) \geq r_{a^*}(s^*)$ when $tau = 1$. And for any a and $s \in S_a$,

$$\begin{aligned}G_a(s) &\leq \sup_{\tau \geq 1} \frac{E[\sum_{t=1}^{\tau} r_{a^*}(s^*)\gamma^{t-1}]}{E[\sum_{t=1}^{\tau} \gamma^{t-1}]} \\ &= r_{a^*}(s^*) \leq G_{a^*}(s^*)\end{aligned}$$

Before we start the proof of the theorem, let's first see a corollary of above theorem and its proof.

Corollary. *At any given time t, if a^* is in state s^*, then the optimal policy will play a^*.*

Proof of Corollary.
The corollary is equivalent to "At time 1, if a^* is in state s^*, then the optimal policy will play a^*" since the optimal policy is stationary. And we will prove this statement.
Assume that π is the optimal policy, and that, at time 1, a^* is in state s^*. If π picks a^* at time 1, there is nothing to prove. Assume for sake of contradiction that π does not pick a^* and it picks a^* for the first time at time τ (τ can be ∞). Let π' be another policy which picks a^* at time 1 and mimics π except for the pull at time τ. (See figure 1.)

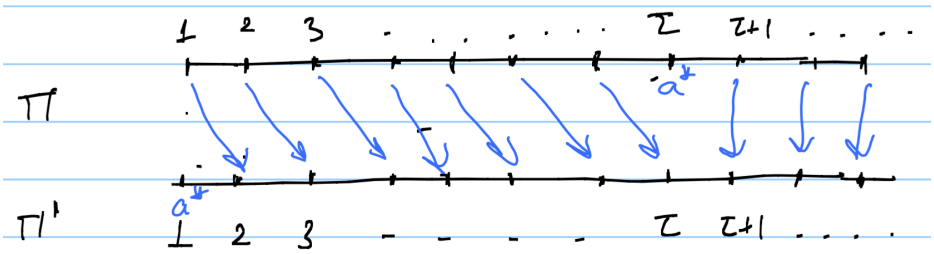

Figure 1: π' mimics actions from π

We want to show that π' is better than π to derive the contradiction. Note that not only the actions, but also the reward of π' at the mapped steps are the same as in π. This is because the reward depends on state and action and all the arms other than a^* have the same sequence of pulls and state of one arm does not depend on pulls of other arms. Consider the expected rewards for the two policies.

$$
\begin{aligned}
J(\pi) &:= E[\sum_{t=1}^{\infty} \gamma^{t-1} r_t] && \text{(where } r_t = r_{a^t}(s_t, a_t), a_t = \pi(s_t)\text{)} \\
&= E[\sum_{t=1}^{\tau-1} \gamma^{t-1} r_t + r_{a^*}(s^*)\gamma^{\tau-1} + \sum_{t=\tau+1}^{\infty} r_t \gamma^{t-1}] \\
J(\pi') &:= E[\sum_{t=1}^{\infty} \gamma^{t-1} r'_t] \\
&= E[r_{a^*}(s^*) + \sum_{t=2}^{\tau} \gamma^{t-1} r'_t + \sum_{t=\tau+1}^{\infty} r'_t \gamma^{t-1}]
\end{aligned}
$$

But we know that $r_t = r'_{t+1}$ for $t = 1, ..., \tau - 1$, and $r_t = r'_t$ for $t = \tau + 1, ..., \infty$. Therefore, with a change of index in the middle term, we can rewrite $J(\pi')$ as

$$
\begin{aligned}
J(\pi') &= E[r_{a^*}(s^*) + \sum_{t=1}^{\tau-1} \gamma^t r_t + \sum_{t=\tau+1}^{\infty} r_t \gamma^{t-1}] \\
\implies J(\pi') - J(\pi) &= r_{a^*}(s^*)(1 - \gamma^{t-1}) - (\sum_{t=1}^{\tau-1} \gamma^{t-1} r_t - \sum_{t=1}^{\tau-1} \gamma^t r_t) \\
&= r_{a^*}(s^*)(1 - \gamma^{t-1}) - (\sum_{t=1}^{\tau-1} \gamma^{t-1} r_t - \gamma \sum_{t=1}^{\tau-1} \gamma^{t-1} r_t) \\
\text{But } (\sum_{t=1}^{\tau-1} \gamma^{t-1} r_t - \gamma \sum_{t=1}^{\tau-1} \gamma^{t-1} r_t) &\leq r_{a^*}(s^*)(\sum_{t=1}^{\tau-1} \gamma^{t-1})(1 - \gamma) \\
&= r_{a^*}(s^*) \frac{1 - \gamma^{\tau-1}}{1 - \gamma}(1 - \gamma) \\
&= r_{a^*}(s^*)(1 - \gamma^{\tau-1}) \\
\implies J(\pi') - J(\pi) &\geq r_{a^*}(s^*)(1 - \gamma^{t-1}) - r_{a^*}(s^*)(1 - \gamma^{\tau-1}) \\
&= 0
\end{aligned}
$$

Since $J(\pi') - J(\pi) \geq 0$, π' gets better reward than the optimal policy π, which is contradiction. Therefore, it's optimal to play a^* at time 1 when a^* is in state s^*, which implies that it's always optimal to play a^* whenever a^*

is in state s^*.

For the proof of Gittins Index Theorem, we will let set $\Pi(s*)$ to be the set of all Markovian stationary policies that play a^* whenever a^* is in state s^*.

Proof of Gittins Index Theorem. We will prove by induction on m where $m = |S_1| + ... + |S_N|$, the total number of states. When $m = 1$, the case is trivial since the policy pulls the only arm. For $m + 1$, we will reduce the state space of arm a^* by removing s^* and apply the induction hypothesis for m.

We will reduce the state space by setting new state space s'_{a^*}, reward function r'_{a^*}, and transition probability p'_{a^*}, while keeping all other arms and their conditions unchanged. Our goal is to achieve this reduction while satisfying two properties:

1. For any $\pi \in \Pi(s^*)$, expected discounted reward for π on new problem is the same as that on the original problem.

2. Gittins index remains the same. That is, $G_a(s)$, for all $s \in S_{a^*}$, $s \neq s^*$ is the same as before. (It is trivial to see that $G_a(s)$, $a \neq a^*$, $s \in S_a$ is the same as before.)

Before discussing how to reduce the state space with the properties above, let's first argue how this reduction can lead us to prove the theorem.

Assume for sake of contradiction that some policy π' is better than the Gittins index policy π in the original problem (with $m + 1$ states), i.e.

$$J(\pi') > J(\pi)$$

Then,

1. For new problem with m states, expected discounted reward of π and π' are the same as the original problem. So π' is still better than π. This follows from the first property.

2. But π is the Gittins index policy for the new problem with m states as well. This follows from the second property.

This implies that, in the reduced problem, some policies are better than Gittins index policy, which is a contradiction. Therefore, we have a contradiction and this proves the theorem.

The details of reduction to achieve the properties above are as follows.

If, at time t, a^* is in some state s reachable to s^* in the original problem (as in figure 2), we know the policy $\pi \in \Pi(s^*)$ will pull a^* again for some l times until it goes to $s' \neq s^*$.

Figure 2: state reachable to s^* in the original problem

Now we want to reduce the original problem by replacing the sequence by self-loop to state s instead of the transition to the removed state s^* (as in figure 3).

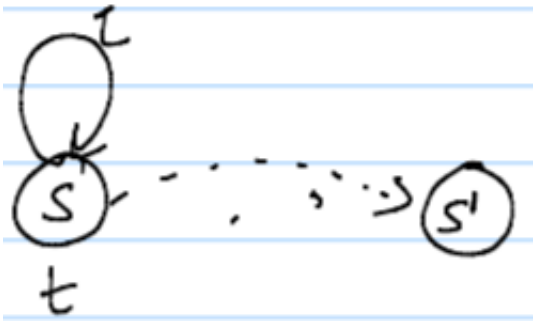

Figure 3: state in the new reduced problem

We update the transition probability to be

$$p'_{a^*}(s,s) = p_{a^*}(s,s) + \sum_{k=0}^{\infty} p_{a^*}(s,s^*)p_{a^*}(s^*,s^*)^k$$

as well as transition probability of moving to any other adjacent states other accordingly.

To update the reward function, we'd like to change it such that the new function merges all the rewards from pulling a^* at state s as well as those from l consecutive pulls at state s^* into one state, s, in the new problem.

Also, we update the reward function in the new problem to be

$$r'_{a^*}(s) = \frac{E[r_{a^*}(s) + r_{a^*}(s^*)(\gamma + \gamma^2 + ... + \gamma^l)|s]}{E[1 + \gamma + \gamma^2 + ... + \gamma^l|s]}$$

To see that this reduction, in fact, satisfies the required properties, we compare the expected rewards (first property) and Gittins index (second property).

From time t to time $t + l$, the reward from the original problem is

$$E[\gamma^{t-1} r_{a^*}(s) + \gamma^t r_{a^*}(s^*)(1 + \gamma + \gamma^2 + ... + \gamma^{l-1})]$$

In the new problem,

$$\begin{aligned} &r'_{a^*}(s)\gamma^{t-1}E[1 + \gamma + \gamma^2 + ... + \gamma^l] \\ &= E[\gamma^{t-1} r_{a^*}(s) + \gamma^t r_{a^*}(s^*)(1 + \gamma + \gamma^2 + ... + \gamma^{l-1})] \end{aligned}$$

Therefore, we have the first property satisfied.

The same logic can be used to show that Gittins index of arm a^* does not change. Only thing to check is that the "best" stopping time, τ, in the definition of Gittins index will not break the sequence of the l consecutive pulls. Consider new Gittins index, G', with stopping time l,

$$\begin{aligned} G' &= \frac{E[\sum_{t=1}^{\tau} \gamma^{t-1} r_t + \sum_{t=\tau+1}^{l} \gamma^{t-1} r_{a^*}(s^*)]}{E[1 + \gamma + ... + \gamma^{\tau+l-1}]} \\ &= \frac{\alpha}{\alpha+\beta} G + \frac{\beta}{\alpha+\beta} r_{a^*}(s^*) \\ &\geq \frac{\alpha}{\alpha+\beta} G + \frac{\beta}{\alpha+\beta} G \\ &\geq G \end{aligned}$$

where $\alpha = E[1 + \gamma + ... + \gamma^{\tau-1}]$, and $\beta = E[\gamma^{\tau} + \gamma^{\tau+1} + ... + \gamma^{\tau+l-1}]$. We can see that stopping time l is better or

at least as good. Therefore, this reduction also satisfies the second property.