Algorithms and Uncertainty, Winter 2017/18 Lecture 11 (4 pages)

Gittins Index Theorem

Instructor: Thomas Kesselheim

We will consider a special case of Markov decision process that we call *Markovian multi-armed bandit*. Of course, the theory from last lecture applies. However, as they have an easy structure, the optimal policies get particularly nice.

Let us first define a single-armed bandit. This is a Markov decision process that has only two actions $\mathcal{A} = \{\texttt{play}, \texttt{stop}\}$. The state transitions and rewards for action `play` are arbitrary, but $p_{\texttt{stop}}(s,s) = 1$, $r_{\texttt{stop}}(s) = 0$ for all $s \in \mathcal{S}$. That is, when using action `stop`, the process remains in its state and gives no reward.

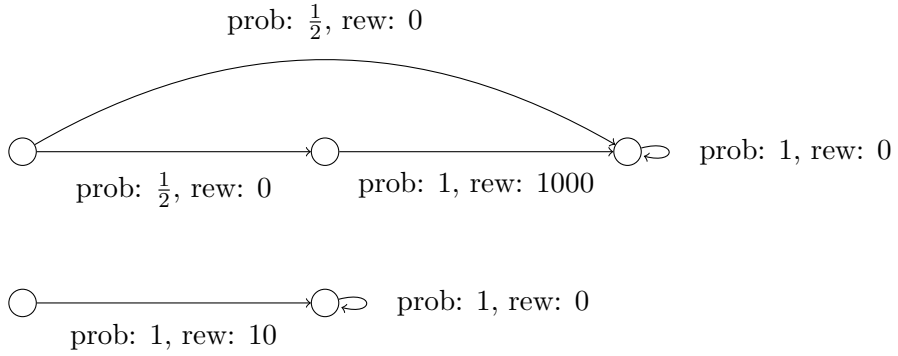

Figure 1: A simple example of two arms. For large values of γ, it is better to play the first arm first. Depending on the outcome, one then continues with the first or the second arm.

A multi-armed bandit bandit is a parallel composition of such single-armed bandits. We have $\mathcal{S} = \mathcal{S}_1 \times \ldots \times \mathcal{S}_n$, where $\mathcal{S}_i$ is the state space of the i^{th} single-armed bandits. Available actions are $\mathcal{A} = \{\texttt{play}_1, \ldots, \texttt{play}_n, \texttt{stop}\}$, where $\texttt{play}_i$ means that we run the `play` action on the i^{th} single-armed bandit and `stop` on any other. So the different single-armed bandits operate independently but we may only play one arm at a time.

We consider the infinite time-horizon setting with discounts, so for a policy π

$$V(\pi, s_0) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{a_t^\pi}(s_t^\pi)\right] \ .$$

Remember that the value of an optimal policy is given as

$$V^*(s) = \max_{a \in \mathcal{A}} \left(r_a(s) + \gamma \sum_{s' \in \mathcal{S}} p_a(s,s') V^*(s') \right) \ .$$

Note that once we have decided to stop, we will never play an arm again. If $\gamma = 1$ then it would be irrelevant in which order we play the arms. However, because $\gamma < 1$, time is the distinguishing factor.

We could always myopically choose the arm with the highest upcoming reward. However, in the example above, we would want to play the arm once without getting any reward and then play it again to get some big reward.

1 Single-Armed Bandit with Charges

To better understand what is happening, we first consider only the single-arm problem. Suppose you had to pay λ every time you played the arm. Then the value of the optimal policy starting at state s would be (see last lecture)

$$V^*(s,\lambda) = \max\left\{0, r_{\texttt{play}}(s) - \lambda + \gamma \sum_{s' \in \mathcal{S}} p_{\texttt{play}}(s,s') V^*(s',\lambda)\right\} \ .$$

Observe that for larger charges λ the value $V^*(s,\lambda)$ gets smaller and smaller. This means, there is some amount $\delta(s)$ that makes the optimal policy only exactly as good as not playing at all. Formally, $\delta(s) = \sup\{\lambda \mid V^*(s,\lambda) > 0\} = \inf\{\lambda \mid V^*(s,\lambda) = 0\}$. This is the *fair charge* of state s.

Based on the fair charges $\delta(s)$, it is very easy to describe an optimal policy for the arm with charge λ: Whenever in a state s with $\delta(s) \geq \lambda$ choose `play`, whenever in a state s with $\delta(s) < \lambda$ choose `stop`.[1]

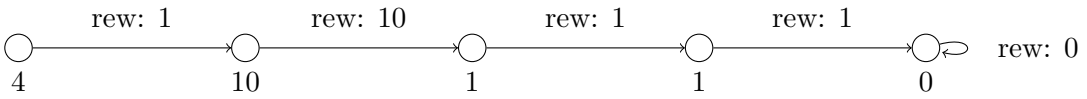

Figure 2: A simple example of one deterministic arm with fair charges for $\gamma = \frac{1}{2}$.

We can bound the reward of a policy by the fair charges of the states during its execution.

Lemma 11.1. *Consider a policy for a single arm that first only chooses* `play` *and then only chooses* `stop`*. Let τ be the index of the step in which it chooses* `play` *for the last time. Then*

$$\mathbf{E}\left[\sum_{t=0}^{\tau} \gamma^t r_{\texttt{play}}(s_t)\right] \leq \mathbf{E}\left[\sum_{t=0}^{\tau} \gamma^t \min_{t' \leq t} \delta(s_{t'})\right]$$

with equality if $\delta(s_\tau) = \min_{t' \leq \tau} \delta(s_{t'})$ with probability 1.

Proof. Let us first consider the case of a policy for which $\delta(s_\tau) = \min_{t' \leq \tau} \delta(s_{t'})$ with probability 1. An alternative way to understand such a policy is as follows: If $\delta(s_0) \geq x$ for some x, first start the optimal policy for the arm with charges $\delta(s_0)$. It stops at some time τ_1. Then run the optimal policy for the arm with charges $\delta(s_{\tau_1})$ until τ_2 and so on until $\delta(s_t) < x$, which is the time that we stop.

Observe that the steps τ_k are exactly the times in which $\delta(s_t)$ is smaller than it has ever been before. Otherwise, the optimal policy for the charged setting would continue playing. Furthermore, the expected reward in the setting with charges overall is exactly 0. Therefore, the expected sum of charges matches the expected sum of rewards in the setting without charges. That is,

$$\mathbf{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}} \gamma^t r_{\texttt{play}}(s_t) \,\middle|\, \tau_k\right] = \mathbf{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}} \gamma^t \delta(s_{\tau_k}) \,\middle|\, \tau_k\right] = \mathbf{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}} \gamma^t \min_{t' \leq t} \delta(s_{t'}) \,\middle|\, \tau_k\right] \ .$$

Taking the sum over all k, the equality in the lemma follows.

[1] For $\delta(s) = \lambda$ actually both choices are equally good.

To get the upper bound for a general policy, we can follow the argument above with the exception that the policy might stop one of its sub-phases early. In this case, the fair cost of the current state is higher than the charge. This means that expected reward for this sub-phase is at most 0. □

Lemma 11.2. *Consider an arbitrary policy for a single arm and let the indices of steps in which it plays the arm be denoted by T (possibly random, depending on previous states). Then*

$$\mathbf{E}\left[\sum_{t\in T}\gamma^t r_{\texttt{play}}(s_t)\right] \le \mathbf{E}\left[\sum_{t\in T}\gamma^t \min_{t'\le t}\delta(s_{t'})\right]$$

with equality if $\delta(s_t) = \min_{t'\le t}\delta(s_{t'})$ for all $t \notin T$ with probability 1.

Proof. Note that Lemma 11.1 is exactly the case that $T = \{0, 1, \ldots, \tau\}$.

It is easy to extend the lemma to the case $T = \{t', \ldots, t'+\tau\}$ because then $\delta(s_0) = \ldots = \delta(s_{t'})$ and both sides get multiplied by the same $\gamma^{t'}$.

In general, T can be considered a union of disjoint interval, each of which has the form $\{t', \ldots, t'+\tau\}$. By adding up the resulting inequalities, the lemma follows. □

2 Gittins Index Theorem

If we have multiple arms, then we get a different value function for each arm

$$V_i(s,\lambda) = \max\left\{0, r_{\texttt{play},i}(s) - \lambda + \gamma\sum_{s'\in\mathcal{S}} p_{\texttt{play},i}(s,s')V(s',\lambda)\right\} \ .$$

So $V_i(s,\lambda)$ is the maximum expected reward that we could get out of arm i if each play costs an additional λ. Note that $V_i(s)$ only depends on the state of the i^{th} arm, not on the states of the other arms.

For each arm i and each state, we again get a fair charge

$$\delta_i(s) = \sup\{\lambda \mid V_i(s,\lambda) > 0\} = \inf\{\lambda \mid V_i(s,\lambda) = 0\} \ .$$

We call this fair charge the *Gittins index* of the arm in state s.[2]

Our main result for today is as follows.

Theorem 11.3. *It is an optimal policy to always play the arm with the highest Gittins index.*

Proof. To prove the theorem, let T_i be the set of steps in which the Gittins index policy plays arm i. Let us observe how the Gittins index $\delta_i(s_t)$ changes over time. If $t \notin T_i$, then $\delta_i(s_{t+1}) = \delta_i(s_t)$. If $t \in T_i$ then $\delta_i(s_{t+1})$ can differ from $\delta_i(s_t)$. If it gets larger, then we keep playing the arm. We only stop playing the arm when its index falls below the value that we started from, meaning it is an all-time low. In other words, if $t \notin T_i$ then $\delta_i(s_t) \le \min_{t'\le t}\delta_i(s_{t'})$.

This allows us to invoke Lemma 11.2. We know that the expected reward from playing arm i is exactly

$$\mathbf{E}\left[\sum_{t\in T_i}\gamma^t \min_{t'\le t}\delta_i(s_t)\right]$$

[2]The original definition by Gittins and Jones is a little different but has the same consequences.

and so the overall expected reward is exactly

$$Q = \sum_{i=1}^{n} \mathbf{E}\left[\sum_{t \in T_i} \gamma^t \min_{t' \le t} \delta_i(s_t)\right]$$

For any other policy, we get different values of T_i and $\delta_i(s_t)$ but by Lemma 11.2 its expected reward is still upper-bounded by the respective Q. Therefore, it is sufficient to show the following proposition.

Proposition 11.4. *Among all policies, the Gittins index policy maximizes*

$$Q = \mathbf{E}\left[\sum_{i=1}^{n} \sum_{t \in T_i} \gamma^t \min_{t' \le t} \delta_i(s_t)\right] \quad .$$

We compare an arbitrary policy π to the Gittins index policy. For simplicity of the argument, we assume that both policies play each arm infinitely often. The spirit of the argument does not change without this assumption but things get much more messy.

For policy π as well as for the Gittins index policy, let us denote by x_t or y_t respectively, the value of $\delta_{\min,i}(t)$ for the arm chosen in step t.

Arm i randomly transitions from one state in $\mathcal{S}_i$ to another one, every time it is played. Let us fix these transitions arbitrarily. This way, the sequences $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ are not random anymore but fixed. Furthermore, they contain exactly the same numbers because each arm makes the same state transitions, only the order varies.

For the Gittins index policy, we have $y_1 \ge y_2 \ge \ldots$, so the sequence in non-increasing. Therefore, now

$$Q^\pi = \sum_{t=0}^{\infty} \gamma^t x_t \le \sum_{t=0}^{\infty} \gamma^t y_t = Q^{\text{Gittins}} \quad .$$

This holds for any fixed transitioning of each single arm, so it also holds in expectation. □

Further Reading

- On the Gittins Index for Multiarmed Bandits, R. Weber, Ann. Appl. Probab. (This proof without formulas)
- Four proofs of Gittins' multiarmed bandit theorem, E. Frostig, G. Weiss, Applied Probability Trust (This and other proofs, with heavy notation)