Algorithms and Uncertainty, Summer 2021 Lecture 17 (5 pages)

Gittins Index Theorem

Thomas Kesselheim Last Update: June 10, 2021

Suppose you are the boss of a large company. You company is producing a well-running product out of which you make a profit of 10 units of money per year. You can keep the production running as it is, or you can invest the profits into research and development. Chances are that after five years of development, you can decrease the production cost and make a profit of 11 units of money per year. However, you have to spend your entire profits during this time and there is still a 50 % chance that after five years you realize that the development failed. Then, of course, you can still keep your production running, keeping the profit of 10 units of money per year. Under what circumstances would you decide to invest?

Such a problem can be modeled via *Markovian multi-armed bandits*. We will define them formally and then derive what optimal policies look like. The theory of Markov decision processes applies and, as we have an infinite time horizon, one could compute an optimal policy via linear programming, value iteration, and so on. However, the state space will be quite huge and the optimal policies have a particularly nice structure.

1 Single-Armed Bandit

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

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

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

Already finding an optimal policy of such a single-armed bandit is a nice exercise.

Example 17.1. *Consider the following single-armed bandit in Figure 1.*

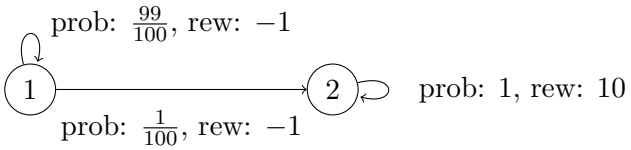

Figure 1: A simple example of one arm. In state 2, it is clearly better to play. But what about the state 1?

For state 2, it is pretty clear that we should play. Formally, we can show this as follows. If π is a policy that plays in state 2, then we have $V(\pi, 2) = 10 + \gamma V(\pi, 2)$, and so $V(\pi, 2) = \frac{10}{1-\gamma} \geq 0$. So, this is better than not playing in state 2, regardless of γ.

For state 1, we can do the same comparison. Let's consider a policy π that chooses to play both states. Then we have $V(\pi, 1) = -1 + \gamma\frac{99}{100}V(\pi, 1) + \gamma\frac{1}{100}V(\pi, 2)$. This implies $\left(1 - \frac{99}{100}\gamma\right) V(\pi, 1) = -1 + \gamma\frac{1}{100}V(\pi, 2) = \frac{\gamma}{10(1-\gamma)} - 1$. So, as we see, depending on γ, $V(\pi, 1)$ will be positive or not. If it is positive or zero, then it is an optimal policy to play. Otherwise, the optimal policy chooses to stop in state 1.

2 Multi-Armed Bandit

A *multi-armed 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 $\mathcal{A} = \{\texttt{play}_1, \ldots, \texttt{play}_n, \texttt{pause}\}$, where $\texttt{play}_i$ means that we run the `play` action on the i^{th} single-armed bandit and `pause` on any other. So the different single-armed bandits operate independently but we may only play one arm at a time.

Note that when a Markovian policy (for example an optimal policy) decides to pause, it remains in the state and therefore will keep pausing from now on and not resume playing 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 in Figure 2, we would want to play the first arm first without getting any reward. Depending on which state we are in, we would play it again to get some big reward.

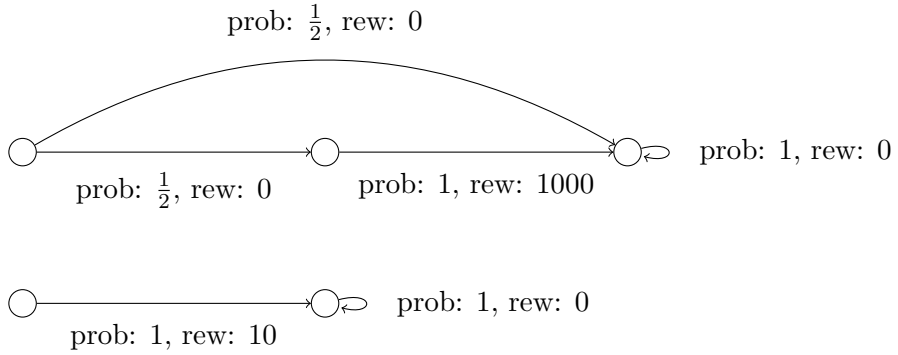

Figure 2: A simple example of two arms. For large values of γ, it is better to play the first arm first. Depending on the outcome, one then continues with the first or the second arm.

3 Optimal Policy

In order to describe the optimal policy, consider the following problem with a single arm i. Suppose you can only choose between playing this arm or pausing and you had to pay λ every time you played the arm. Then the Bellman equation tells us that the maximum expected reward that you could get from arm i with an optimal policy is

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

Note that $V_i(s, \lambda)$ only depends on the state of arm i, not on the states of the other arms.

Observe that for larger charges λ the value $V_i(s, \lambda)$ 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,

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

This is the *fair charge* or the *Gittins index* of arm i in state s.[1]

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

Example 17.2. *In Figure 3, we see one arm with deterministic transitions. In state E, the fair change is* 0*. In states C and D, it is* 1*: We play once (or maybe twice from C) and then stop. For state B, we play once if the charge is* 10*. Finally, state A is interesting. The fair charge is* 4*: The reward from playing twice with charge* λ *is* $1 - \lambda + \frac{1}{2}(10 - \lambda)$*, which is* 0 *exactly for* $\lambda = 4$*.*

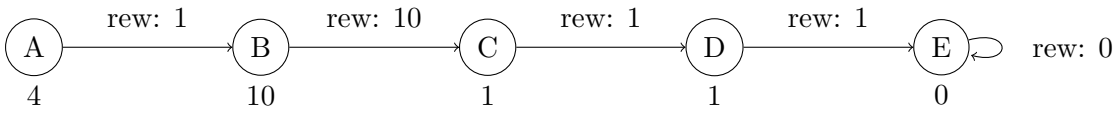

Figure 3: A simple example of one deterministic arm with fair charges for $\gamma = \frac{1}{2}$.

Our main result for today is the following theorem.

Theorem 17.3. *It is an optimal policy to always play the arm with the highest Gittins index.*

4 Single-Armed Bandit with Charges

We first consider the problem with only a single arm. 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 `pause`.[2]

We can also bound the reward of any policy by the fair charges of the states $s_0, s_1, \ldots$ that it enters during its execution.

Lemma 17.4. *Consider a policy for a single arm that first only chooses* `play` *and then only chooses* `pause`*. Let* τ *be the index of the step in which it chooses* `pause` *for the first time. Then*

$$\mathbf{E}\left[\sum_{t=0}^{\tau-1} \gamma^t r_{\texttt{play}}(s_t)\right] \leq \mathbf{E}\left[\sum_{t=0}^{\tau-1} \gamma^t \min_{t' \leq t} \delta(s_{t'})\right]$$

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. So, this is a policy which only stops playing when reaching a state that has the smallest fair charge it has seen so far. Any execution of such a policy naturally decomposes into phases of random length: Let $\tau_0 = 0$ and let τ_{k+1} be the first time step $t > \tau_k$ at which $\delta(s_t) < \delta(s_{\tau_k})$ (see Figure 4). The policy chooses `pause` at some point in time τ_k.

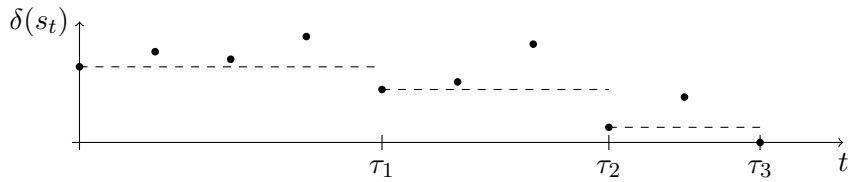

Figure 4: Phases in the proof of Lemma 17.4.

An alternative way to view the phases is as follows. Fix everything that happens until time τ_k. At time τ_k, start an optimal policy for the arm with charges $\delta(\tau_k)$. This optimal policy

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

stops exactly at time τ_{k+1} because this is the first time that the charge exceeds the fair charge. The expected reward of this optimal policy starting at time τ_k with charge $\delta(s_{\tau_k})$ is exactly 0. So,

$$\mathbf{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1}\gamma^{t-\tau_k}\left(r_{\texttt{play}}(s_t)-\delta(s_{\tau_k})\right)\;\middle|\;\tau_k\right]=0\ .$$

Equivalently,

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

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

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 17.5. *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 17.4 is exactly the case that $T=\{0,1,\ldots,\tau-1\}$.

It is easy to extend the lemma to the case $T=\{t',\ldots,t'+\tau-1\}$ 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 intervals, each of which has the form $\{t',\ldots,t'+\tau-1\}$. By adding up the resulting inequalities, the lemma follows. □

5 Proof of Gittins Index Theorem

Proof of Theorem 17.3. To prove the theorem, let π be the Gittins index policy and let T_i be the set of steps in which it plays arm i. Let us observe how the Gittins index $\delta_i(s_t^\pi)$ changes over time. If $t\notin T_i$, then $\delta_i(s_{t+1}^\pi)=\delta_i(s_t^\pi)$. If $t\in T_i$ then $\delta_i(s_{t+1}^\pi)$ can differ from $\delta_i(s_t^\pi)$. 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 $\delta_i(s_t^\pi)\le\min_{t'\le t}\delta_i(s_{t'}^\pi)$.

This allows us to invoke Lemma 17.5. We know that the expected reward from playing arm i is exactly

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

and so the overall expected reward is exactly

$$V(\pi,s_0)=\sum_{i=1}^n\mathbf{E}\left[\sum_{t\in T_i}\gamma^t\min_{t'\le t}\delta_i(s_t^\pi)\right]=\mathbf{E}\left[\sum_{i=1}^n\sum_{t\in T_i}\gamma^t\min_{t'\le t}\delta_i(s_t^\pi)\right]\ .$$

Now, consider any other policy π^* and let T_i^* be the set of steps in which it plays arm i. By Lemma 17.5, we get

$$V(\pi, s_0) \leq \mathbf{E}\left[\sum_{i=1}^{n}\sum_{t\in T_i^*}\gamma^t \min_{t'\leq t}\delta_i(s_t^{\pi^*})\right] .$$

Proposition 17.6. *Fix the transitions of all arms each time they are played arbitrarily. Then the Gittins index policy π guarantees that*

$$\sum_{i=1}^{n}\sum_{t\in T_i}\gamma^t \min_{t'\leq t}\delta_i(s_t^{\pi}) \geq \sum_{i=1}^{n}\sum_{t\in T_i^*}\gamma^t \min_{t'\leq t}\delta_i(s_t^{\pi^*})$$

compared to any policy π^*.

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.

Let us denote by $x_t = \min_{t'\leq t}\delta_i(s_t^{\pi})$ or $y_t = \min_{t'\leq t}\delta_i(s_t^{\pi^*})$ respectively, the value of the smallest Gittins index for the arm i chosen by the respective policy in step t.

As we fixed arm i's random transitions from one state in $\mathcal{S}_i$ to another one, the sequences $x_0, x_1, \ldots$ and $y_0, y_1, \ldots$ are not random anymore but fixed. Furthermore, because we assume that each arm is played infinitely often, 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 $y_0 \geq y_1 \geq \ldots$, so the sequence is non-increasing. Therefore, now

$$\sum_{i=1}^{n}\sum_{t\in T_i^*}\gamma^t \min_{t'\leq t}\delta_i(s_t^{\pi^*}) = \sum_{t=0}^{\infty}\gamma^t x_t \leq \sum_{t=0}^{\infty}\gamma^t y_t = \sum_{i=1}^{n}\sum_{t\in T_i}\gamma^t \min_{t'\leq t}\delta_i(s_t^{\pi}) .$$

□

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)