Topics in Game Theory: Bandit Problems

Juuso Välimäki

Aalto BIZ and Helsinki GSE

March 2022

1 Introduction

In this Lecture, we specialize the setting of the previous lecture to *multi-armed bandit* problems (MAB). The problem was first introduced as an idealized model for. conducting medical trials by Thompson (1933). The suggestive name refers to the problem faced a gambler entering the floor for slot machines (one-armed bandits) at a casino. Each k of the K machines gives a random prize for each coin fed into the machine. If the reward distribution for a sequence of coins fed into machine k is a sequence of independent draws from $p(x \mid \theta_k)$, then we are in the parametric learning setting of the first lecture. For this application, it may make sense to assume that the parameters of the different machines are independently drawn. If we denote the gambler's posterior distribution on the parameter of machine k after using t_k coins on that machine (and after observing $(x_0, ..., x_{t_k})$ by $\mu^k_{t_k}$, we get a new posterior μ_{t_k+1} after inserting one more coin and observing reward x_{t_k+1} by Bayes' rule. The gambler's problem is to maximize the expected discounted reward from the machines over a sequence of coins. If you want, you can also include the strategy of walking out of the casino as an additional arm yielding 1 coin for sure for each

trial. Notice that the posteriors evolve according to a Markov process on the state space $\Delta(\Theta)$ of probability distributions on the parameter set. Perhaps you can see the connection to the sequential allocation of alternative treatments with initially uncertain effectiveness for the treatment of a sequence of patients.

To give you an idea of where multi-armed bandit models have been applied, I reproduce a list from Slivkins (2021):

Application domain	Action	Reward
medical trials	which drug to prescribe	health outcome.
web design	e.g., font color or page layout	#clicks.
content optimization	which items/articles to emphasize	#clicks.
web search	search results for a given query	#satisfied users.
advertisement	which ad to display	revenue from ads.
recommender systems	e.g., which movie to watch	1 if follows .
sales optimization	which products to offer at which prices	revenue.
procurement	which items to buy at which prices	#items procured.
auction/market design	e.g. which reserve price to use	revenue.
crowdsourcing	match tasks and workers, assign prices	#completed tasks.
datacenter	design e.g., which server to route	job completion time.
Internet	e.g., which TCP settings to use?	connection quality.
radio networks	which radio frequency to use?	#transmissions.
robot control	a "strategy" for a given task	job completion time.

2 The Model

2.1 Sequential Markov Decision Problem

We start by formulating a sequential Markov decision problem.

- A sequence of decisions to be taken over a discrete infinite horizon $t = 0, 1, ...$
- At each t, the decision maker chooses one alternative amongst a fixed set of alternatives K called arms and we denote this choice by $a_t \in \{1, ..., K\}$.
- If $a_t = k$, a random payoff x_t^k is realized and we denote the associated random variable by X_t^k. We assume bounded rewards: $\sup_t \mid X_t^k \mid < \infty$ for all k.
- The decision problem is a Markov decision problem with state variable $s_t \in S$. This means simply that for all current states $s_t = s \in S$, each action $a_t = k$ induces a Markovian transition probability $P^k(s' \mid s)$ for reaching the state $s' \in S$ and that the (distribution of the) reward depends only on (a_t, s_t).
- Write the distribution of X_t^k as $F^k(\cdot \mid s_t)$.
- The state transition function ϕ depends on the choice of the arm and the realized payoff:
$$s_{t+1} = \phi\left(x_t^k; s_t\right).$$
- A feasible Markov policy $a = \{a_t\}_{t=0}^{\infty}$ selects an available alternative for each conceivable state s_t, i.e.
$$a_t : S \to \{1, ..., K\}.$$

2.2 Bandit Problem

The following two assumptions must be met for the problem to qualify as a bandit problem.

1. Payoffs are evaluated according to the discounted expected payoff criterion where the discount factor δ satisfies $0 \leq \delta < 1$.

2. The payoff from each k depends only on outcomes if periods with $a_t = k$. In other words, we can decompose the state variable s_t into K components $\left(s_t^1, ..., s_t^K\right)$ such that for all k :

$$\begin{aligned} s_{t+1}^k &= s_t^k && \text{if} \quad a_t \neq k, \\ s_{t+1}^k &= \phi(s_t^k, x_t) && \text{if} \quad a_t = k, \end{aligned}$$

and

$$F^k\left(\cdot, s_t\right) = F^k\left(\cdot; s_t^k\right).$$

Notice that when the second assumption holds, the alternatives must be statistically independent.

It is easy to see that many situations of economic interest are special cases of the above formulation.

- First, it could be that $F^k\left(\cdot; \theta^k\right)$ is a fixed distribution with an unknown parameter θ^k. The state variable is then the vector of posterior probability distributions on θ^k for $k \in \{1, ..., K\}$.

- Alternatively, $F^k\left(\cdot; s^k\right)$ could denote the random yield per period from resource k after extracting s^k units (think about mining or harvesting etc.).

The value function $V\left(s_0\right)$ of the bandit problem can be written as follows. Let $X^k\left(s_t^k\right)$ denote the random reward with distribution $F^k\left(\cdot; s_t^k\right)$. Then the problem of finding an optimal allocation policy is the solution to the following intertemporal optimization problem:

$$V(s_0) = \sup_a \left\{ \mathbb{E}_a \sum_{t=0}^{\infty} \delta^t X^{a_t}\left(s_t^{a_t}\right) \right\}.$$

The celebrated index theorem due to Gittins and Jones (1974) transforms the problem of finding the optimal policy into a collection of k stopping problems. Stopping problems are amongst the simplest dynamic stochastic optimization problems. At each point in time, the decision maker has to decide whether to stop or continue. To formalize this idea, let X_t be the (bounded) reward from stopping in period t. Deterministic stopping problems are quite easy: just pick the t at which $\overline{X} := \sup_t X_t$ is reached. If no such t exists, then you can get an ϵ-optimal solution for all $\epsilon > 0$ by stopping at the first t with $X_t > \overline{X} - \epsilon$.

Stochastic stopping problems allow for a stochastic process $X_{tt} = 0^{\infty}$. The decision to stop or not must be based on information observed at or before t, i.e. the σ-algebra generated by $(X_0, ..., X_t)$. Even if you know the statistical properties of the process (i.e. the probability measure P on X^{∞}, you do not know the realizations of $(X_{t+1}, ...)$. We denote stopping times by τ and since they depend on the realizations of the (X_t), they are random variables. The requirement that you can use only information available at t just says that the event $\tau = t$ must be measurable with respect to $\sigma(X_0, ..., X_t)$. With these preliminaries, a stopping problem is just to find

$$\sup_{\tau} \mathbb{E} X_{\tau}.$$

For each alternative k, we calculate the following index $m^k\left(s_t^k\right)$, which only depends on the state variable of alternative k:

$$m^k\left(s_t^k\right) = \sup_{\tau} \left\{ \frac{\mathbb{E}\sum_{u=t}^{\tau} \delta^t X^k\left(s_u^k\right)}{\mathbb{E}\sum_{u=t}^{\tau} \delta^t} \right\}, \tag{1}$$

where τ is a stopping time with respect to $\{s_t^k\}$.

The idea is to find for each k the stopping time τ that results in the highest discounted expected return per discounted expected number of

periods in operation. It is a good exercise to show that a stopping time achieving the supremum exists (hint: discounting and bounded returns).

The Gittins index theorem then states that the optimal way of choosing arms in a bandit problem is to select in each period the arm with the highest Gittins index, $m^k\left(s_t^k\right)$.

Theorem 2.1 (Gittins-Jones (1974))**.**
The optimal policy satisfies $a_t = k$ for some k such that

$$m^k\left(s_t^k\right) \geq m^j\left(s_t^j\right) \text{ for all } j \in \{1, ..., K\}.$$

Proof. The theorem is proved by using the principle of optimality: a policy is optimal if there is no profitable one-shot deviation from it. Starting at an arbitrary state $s_0 = s$, let π^* be a policy that activates an arm with maximal index at each point (in case of ties choose the arm with the smallest index). We claim that any policy $\pi^{(0)}$ starting with $a_0 = k$ in the first period and then according to π^* yields a payoff no larger than the payoff from policy π^*. I.e. we need to show that for all s, k

$$V_{\pi^*}(s) \geq V_{\pi^{(0)}}(s).$$

To this effect, we construct a sequence of policies $\pi^{(n)}$ such that the associated value $V_{\pi^{(n)}}(s)$ satisfies:

$$\lim_n V^n_{\pi^{(n)}}(s) = V_{\pi^*}(s) \text{ for all } s,$$

$$V_{\pi^{(n)}}(s) \geq V_{\pi^{(n-1)}}(s) \text{ for all } s.$$

The sequence $\pi^{(n)}$ is defined recursively. For $s_0 = s$, let k^*, m^*, τ^* be the arm with the highest index, the value of the highest index and the stopping time achieving the index. Then $\pi^{(n)}$ is a concatenation of the optimal policy up to τ^* (reaching the random state s_{τ^*}) with the policy $\pi^{(0)}$ with initial state s_{τ^*}.

By construction, π^n and π^* agree for at least the first n periods. Because of discounting (and bounded returns), this implies that $\lim_n V^n_{\pi^{(n)}}(s) = V_{\pi^*}(s)$. For $n > 1$, $\pi^{(n)}$ and $\pi^{(n-1)}$ agree over the first τ^* periods and:

$$V_{\pi^{(n)}}(s) - V_{\pi^{(n-1)}}(s) = \mathbb{E}\left\{\delta^{\tau^*}\mathbb{E}\left\{V_{\pi^{(n-1)}}(s_{\tau^*}) - V_{\pi^{(n-2)}}(s_{\tau^*})|s_{\tau^*}\right\}\right\}.$$

To prove that $V_{\pi^{(n)}}(s) \geq V_{\pi^{(n-1)}}(s)$ inductively, it is enough to show that $V_{\pi^{(1)}}(s) \geq V_{\pi^{(0)}}(s)$.

If $k = k^*$, then $\pi^{(1)} = \pi^{(0)} = \pi^*$ and there is nothing to prove, so suppose $k \neq k^*$. Define a stopping time σ for arm k as the first t where $m^k\left(s^k_t\right) < m^*$. Then we see that $\pi^{(0)}$ activates arm k for a duration of σ periods. At stopping time σ, the highest index is m^* of arm k^*. Since $\pi^{(0)}$ continues according to π^*, it activates k^* for the next τ^* periods up to stopping time $\sigma + \tau^* - 1$. From $\sigma + \tau^*$ onwards, $\pi^{(0)}$ proceeds according to π^*.

Policy $\pi^{(1)}$ starts with k^* for τ^* periods, then selects k for a period and then proceeds according to π^*. $\pi^{(1)}$ activates arm k for at least σ periods from τ^* onwards. This follows from the fact that the continuation is acoording to π^* and the indices of $k' \neq k, k^*$ are unchanged (and at most m^*) and $m^k(s_{\tau^*}) \leq m*$ by the definition of τ^*.

To sum, $\pi^{(0)}$ starts with k for σ periods followed by $k*$ for τ^* periods. $\pi^{(1)}$ starts with $k*$ for τ^* periods followed by k for σ periods. At the end of $\sigma + \tau*$ periods, both policies have reached the same state and their continuation is according to π^*. Note that for given $k, k*$, the stopping times σ and τ are independent. Hence the difference in expected returns is:

$$\begin{aligned}
&\mathbb{E}\left\{\sum_{t=0}^{\sigma-1}\delta^t X^k(s^k_t) + \delta^\sigma \sum_{t=0}^{\tau^*-1}\delta^t X^{k^*}(s^{k^*}_t)|s_0\right\} \\
&-\mathbb{E}\left\{\sum_{t=0}^{\tau^*-1}\delta^t X^{k^*}(s^{k^*}_t) + \delta^{\tau^*} \sum_{t=0}^{\sigma-1}\delta^t X^k(s^k_t)|s_0\right\} \\
&= \mathbb{E}(1-\delta^{\tau^*})\mathbb{E}\left\{\sum_{t=0}^{\sigma-1}\delta^t X^k(s^k_t)\right\} - \mathbb{E}(1-\delta^\sigma)\mathbb{E}\left\{\sum_{t=0}^{\tau^*-1}\delta^t X^{k^*}(s^{k^*}_t)\right\}
\end{aligned}$$

$$= \frac{1}{1-\delta}\mathbb{E}(1-\delta^{\tau^*})\mathbb{E}(1-\delta^{\sigma})(m^k_{\sigma}(s_k) - m^*)$$
$$\leq \frac{1}{1-\delta}\mathbb{E}(1-\delta^{\tau^*})\mathbb{E}(1-\delta^{\sigma})(m^k(s_k) - m^*) \leq 0,$$

where we have written:

$$m^k_{\sigma}(s_k) = \left\{ \frac{\mathbb{E}\sum_{t=0}^{\sigma} \delta^t X^k\left(s^k_t\right)}{\mathbb{E}\sum_{t=0}^{\sigma}\delta^t} \right\}.$$

□

An alternative formulation of the main theorem, based on dynamic programming can be found in Whittle (1982). The basic idea is to find for every arm a retirement value M^k_t, and then to choose in every period the arm with the highest retirement value. Formally, for every arm k and retirement value M, we can compute the optimal retirement policy given by:

$$V^k\left(s^k_t, M\right) \triangleq \max\left\{\mathbb{E}\left[X^k\left(s^k_u\right) + \delta V^k\left(s^{k+1}_t, M\right)\right], M\right\} \tag{2}$$

The auxiliary decision problem given by (2) compares in every period the trade-off between continuation with the reward process generated by arm k or stopping with a fixed retirement value M. The index of arm k in the state s^k_t is the highest retirement value at which the decision maker is just indifferent between continuing with arm k or retiring with $M = M\left(s^k_t\right)$:

$$M^k\left(s^k_t\right) = V^k\left(s^k_t, M^k\left(s^k_t\right)\right).$$

The resulting index $M^k\left(s^k_t\right)$ is equal to the discounted sum of flow index $m^k\left(s^k_t\right)$, or $M^k\left(s^k_t\right) = m^k\left(s^k_t\right)/(1-\delta)$.

We write $V(s)$ for the value function of the original bandit problem starting at s without retirement options. Since rewards are bounded, we have $V(s, M) = V(s)$ for low enough M, and the problems coincide. From the definition it is clear that $V(s, M)$ is non-decreasing in M. Consider an arbitrary policy of retiring at a stopping time τ. The value from this policy is

$$V(s, M; \tau) = \sum_{t=0}^{\tau} \delta^t \mathbb{E} X_{s_t}^{a_t} + \mathbb{E}\delta^{\tau} M.$$

Since each $V(s, M; \tau)$ is linear in M and $V(s, M) = \sup_{\tau} V(s, M; \tau)$, we know that $V(s, M)$ is convex in M. Therefore it has a derivative almost everywhere, and by envelope theorem,

$$\frac{\partial V(s, M)}{\partial M} = \mathbb{E}\delta^{\tau_M},$$

where τ_M is the optimal retirement policy for retirement value M.

It seems reasonable to conjecture that arm k is permanently abandoned if at the states where the individual arm is retired, i.e. at $\tau_{k,M}$. By the independence of the arms, we have:

$$\mathbb{E}\delta^{\tau_M} = \prod_k \mathbb{E}\delta^{\tau_{k,M}},$$

and

$$\frac{\partial V(s, M)}{\partial M} = \prod_k \frac{\partial V^k(s^k, M)}{\partial M}.$$

Notice that $\prod_k \frac{\partial V^k(s^k, M)}{\partial M}$ is non-decreasing in M since each $V^k(s^k, M)$ is convex and nondecreasing in M. Furthermore, $\prod_k \frac{\partial V^k(s^k, M)}{\partial M}$ is zero for $M < -L$ and unity for $M \geq L$ (recall that L is the bound on the absolute value of the rewards). Therefore it has the properties of a distribution function. Integrating gives:

$$V(s, M) = L - \int_M^L \prod_k \frac{\partial V^k(s^k, m)}{\partial m} dm.$$

The remaining step is to verify that Gittins index rule is optimal given this value function. This is done by a standard verification argument. Under the Gittins Index policy, the above conjectured value function satisfies the Bellman equation of the problem.

Computing Gittins Index for Examples
Pandora's boxes

- Weitzman (1979), Econometrica (Pandora's boxes) asks how one schedules the search for a prize (only one prize can be claimed) when there are k statistically independent boxes characterized by the value of the prize and the probability of finding a price in the box (v^k, p^k).
- By now we know how to answer this. Compute Gittins indices for all boxes and open them in the decreasing sequence of the indices.
- Properties of the optimal sequence: Suppose $p^k v^k = p^l v^l$ for some k, l, i.e. if these were the only boxes, the decision maker would be indifferent. Which should be opened first?
- Formulate the problem so that it fits our model above.
- Each arm starts in s_0^k
- If arm k is chosen, it gives an immediate expected return of $p^k v^k$
- If k is chosen, then $s_t^k = H$ for all t with probability p^k and $s_t^k = L$ for all t with probability $1 - p^k$
- $x_t^k = v^k$ for all t if $s_t^k = H$ and $x_t^k = 0$ for all t if $s_t^k = L$.
- Observe that if the arm is tried once, all uncertainty for that arm is immediately resolved.
- $\delta < 1$ the discount factor.
- Compute the Gitttins index as follows: Let $V^k(s_0^k, M)$ be the value function of the auxiliary problem.

$$V(L, M) = M \text{ for all } M \geq 0, \qquad V(H, M) = \frac{v}{1-\delta} \text{ for all } M \leq \frac{v}{1-\delta}.$$

$$M = p^k v^k + \delta(p^k V(H, M) + (1 - p^k)V(L, M))$$
$$= p^k v^k + \delta \left(p^k \frac{v^k}{1-\delta} + (1 - p^k)M \right),$$

or

$$M = \frac{p^k v^k (1 + \delta)}{(1 - \delta)(1 - \delta + p^k \delta)},$$

or

$$(1 - \delta)M = \frac{p^k v^k}{1 - \delta + p^k \delta}.$$

- Observe that as $\delta \to 1$, $(1 - \delta)M \to v^k$.
- Observe also that

$$(1 - \delta)M - p^k v^k = \frac{\delta(1 - p^k)p^k v^k}{1 - \delta + p^k \delta} > 0.$$

- Therefore there is always value to experimentation.
- How far does this generalize?
- Key property required for Index Theorem to work: Outside option for each alternative must stay constant.
- Simple generalizations fails this property: Choice of multiple arms simultaneously. Switching costs between arms.
- Generalizations that can be handled: Arms branching into different arms.

3 Comments

I have only talked about the standard Bayesian optimization approach to the bandit problem. While this is the standard approach in economic theory, the most active area of related research is within the computer science community.

That literature is more interested in finding a good solution that is easily implemented (i.e. using fast algorithms) and that performs well under a set of different scenarios. For anybody interested in models of learning (from the cs point of view), I strongly recommend two recent books:

Bandit Algorithms by Tor Lattimore and Csaba Szepesvári (available at https://tor-lattimore.com/downloads/book/book.pdf)

and

Introduction to Multi-Armed Bandits by Aleksandrs Slivkins (available at https://arxiv.org/pdf/1904.07272.pdf).

This literature uses as a performance criterion the minimization of regret: Find a policy that does well relative to a class of reference policies. For the original casino problem of the introduction to this lecture, a good set of reference policies might be the policies under full information. Regret then measures the difference between a suggested policy and the best reference policy (finite horizon, but without discounting). If the per period regret converges fast to zero as the horizon of the problem increases, then we can say that the policy performs well.

Thompson sampling is a very nice algorithm that does very well in terms of its expected regret. The idea is very simple, For each arm, compute the posterior probability that it is the best alternative and choose the arm for the next period with this posterior probability. The challenge is of course, how to implement the Bayesian updating computationally eficiently for this case.

Even though the Gittins Index policy is very nice, I should mention that the class of problems where it can be applied is quite limited. Bergemann and Välimäki (2001) show that if two or more alternatives are selected simultaneously, then following Gittins Indices is not the optimal policy. Bergemann and Välimäki (2008) contains discussions of economic models using bandits. Banks and Sundaram (1994) shows how switching costs between arms cause problems and Doval (2018) is a recent example showing that the problem is hard with inspection costs.

Classic references on the Bayesian approach are Berry and Fristedt (1985), and Gittins, Glazerbrook, and Weber (2011).

References

BANKS, J., AND R. SUNDARAM (1994): "Switching Costs and the Gittins Index," *Econometrica*, 62, 687–694.

BERGEMANN, D., AND J. VÄLIMÄKI (2001): "Stationary Multi Choice Bandit Problems," *Journal of Economic Dynamics and Control*, 25, 1585–1594.

——— (2008): "Bandit Problems," in *The New Palgrave Dictionary*, ed. by S. Durlauf, and L. Blume. Palgrave Macmillan, New York.

BERRY, D., AND B. FRISTEDT (1985): *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London.

DOVAL, L. (2018): "Whether or not to open Pandora's box," *Journal of Economic Theory*, 175, 127–158.

GITTINS, J., K. GLAZERBROOK, AND R. WEBER (2011): *Multi-armed Bandi tAllocation Indices*. London, Wiley.