B9120 Dynamic Programming Lecture 4 - 02/24/2020

Bandit Processes

Lecturer: Daniel Russo *Scribe: David Cheikhi, Christos Tsanikidis*

1 References

There are several beautiful proofs of the Gittins index theorem. We will follow a short inductive proof of Tsitsiklis. However it is recommended that the student goes through the proof from Weber, which is proceeds almost entirely in words and provides an interesting economic perspective.

- A short proof of the Gittins Index theorem (Tsitsiklis, 1994)
- On the Gittins Index for multiarmed bandit (Weber, 1992)
- A Note on the Equivalence of Upper Confidence Bounds and Gittins Indices for Patient Agents (Russo, 2020)

2 Gittins Index Theorem and MABs

2.1 The Multi-Armed Bandit in the literature

The Multi-Armed Bandit (MAB) is an old problem that was first described in the 30s. It can refer to a class of DP problems with nice decomposition structure (in the MDP literature). It can also refer to class of sequential learning problems with a tension between exploration & exploitation (in the ML/stats literature). There is a point of intersection, where a certain class of learning problems with independent beliefs can be formulated as a dynamic program with nice decomposition structure and in principle solved optimally.

2.2 Definition of the problem

MAB is a special case of our formulation of (discounted) DP. There are n risky projects (*the bandits*) and choosing to act on one of these projects doesn't modify the state of the other projects. The state $x_k = (x_k^1, \ldots, x_k^n)$ factors into the k states of each of the bandits, the control $u_k \in \{1, \ldots, n\}$ indicates project the decision maker chooses to work on, after which point only the state of the selected bandit evolves:

$$x_{k+1}^i = \begin{cases} f_k^i(x_k^i, w_k^i) & \text{if } u_k = i \\ u_k^i & \text{otherwise} \end{cases}$$

The cost function is $g(x_k, u) = -R^u(x_k^u)$, so our problem is to maximize discounted expected rewards where $R^u(x_k^u)$ depends only on the bandit that has been chosen.
We will focus on cases where $f_k^i = f_k$.

2.3 Applications

1. Golf with n balls: one ball is played at each time and we try to maximize the (discounted) number of balls that get into a hole. Only the state of the ball that was hit evolves.
2. Oil drilling: if oil wells are well separated enough, acting on one well doesn't affect the state of the other wells. In these problems, the state likely encodes both a physical state of the wells and a belief state (representing e.g. posterior beliefs that evolve due to seismic surveys etc.).

3. Pandoras Box problems, which have important applications as models of search in economic theory. See *Optimal search for the best alternative* (Weitzman, 1979).
4. **Multi-armed bandit problems** (in the ML sense) with independent prior beliefs about the arms
5. **Scheduling in queues**

The last two will be our focus.

2.4 The Gittins index theorem

Although many simple and natural problems can be formulated in this framework, multi-armed bandit problems were long believed to be intractable. The following colorful anecdote by Whittle captures this well:

> The [MAB] problem was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied scientists that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage."

In the 1970s on Gittins made a remarkable breakthrough on this problem. As shown in the theorem below, an optimal policy plays the bandit with the highest Gittins index, where $G^u(\cdot) : \mathcal{X}_u \to \mathbb{R}$ is a function mapping any state of bandit u to a real number and this function can be computed based only on the reward and transition probabilities of that bandit. The computation of a Gittins index can be complex when the state space of individual bandits is large, but the computational burden only scales linearly in the number of individual projects n, avoiding the curse of dimensionality.

Theorem 1 (Informal)**.** *Under some technical regularity conditions, there exists an optimal stationary policy that plays $u_k = \arg\max_{u\in\mathcal{U}}(G^u(x_k^u))$ where $G^u(\cdot)$ (called the Gittins index) is computed separately across arms.*

On a superficial level, one can say that this is intuitive as the arms evolve independently of each other. However, selecting one arm forgoes the opportunity to select any other arm in the same period, which complicates the situation.

2.5 Deriving an expression for the Gittins Index

The Gittins index is calculated by considering each bandit arm in isolation. We can reverse engineer a formula for the Gittins index by considering very simple problem instances and using that the Gittins index must prescribe an optimal policy. Consider a "1.5" armed bandit problem, where

- Arm 0 yields rewards $\lambda, \alpha\lambda, \alpha^2\lambda, \ldots$ (safe arm that gives a known reward)
- Arm 1 yields rewards $R(x_0), \alpha R(x_1), \alpha^2 R(x_2), \ldots$

Note that, if there is an optimal policy selects arm 0 initially, by the stationarity of the optimal policy and the fact that plays of arm 0 do not chance the state of the bandit, there is an optimal policy that plays are 0 perpetually. Similarly, there is always an optimal policy that plays arm 1 up until a time τ and state x_τ at which it retires and plays arm 0 thereafter. It is optimal to play arm 1 at least once if and only if

$$\sup_{\tau\geq 1}\mathbb{E}\left[\sum_{t=0}^{\tau}\alpha^t R(x_t) + \sum_{t=\tau+1}^{\infty}\alpha^t\lambda \,\middle|\, x_0 = x\right] \geq \mathbb{E}\left[\sum_{t=0}^{\infty}\alpha^t\lambda\right]$$

where the supremum is over stopping times τ with respect to $(x_0, x_1, \cdots)$. By subtracting the right hand side, this is equivalent to the inequality $V_\lambda(x) \geq 0$ where

$$V_\lambda(x) := \sup_{\tau\geq 1}\mathbb{E}\left[\sum_{t=0}^{\tau}\alpha^t(R(x_t) - \lambda) \,\middle|\, x_0 = x\right].$$

Suppose that the Gittins index assigns value λ to any safe arm [1] with known reward λ. The Gittins index policy will stop playing arm 1 in state x if and only if $G(x) \geq \lambda$. This must be optimal for any value of λ, so we define the Gittins index to be the *highest* value of λ at which it is optimal to play arm 1 in state x:

$$G(x) = \sup\{\lambda | V_\lambda(x) \geq 0\} \tag{1}$$

Weber (1992) interprets this λ as tax per use of arm 1 that effectively changes the reward from $R(x)$ to $R(x) - \lambda$.

3 Semi-Markov MABs and a proof of the Gittins Index Theorem

Tsitsiklis (1994) observed that moving to a more general Semi-Markov formulation allows for a simple inductive proof of the Gittins index theorem. We follow his formulation:

- There are n bandit processes
- The ith such process is a semi-markov process with state-space with **finite** state space $\mathcal{X}_i$
- The $\mathcal{X}_i$'s are disjoint and $\mathcal{X} = \cup_i \mathcal{X}_i$
    - This is without loss of generality. You could imagine that every element of $\mathcal{X}_i$ is of the form $(\tilde{x}, i)$ where $\tilde{x}$ is the "true" description of the state and i represents the identification of the arm. In this way, we don't need to track the index i in all our notation and x suffices.
- Playing bandit i in state x realizes an outcome:

  $$(R(x), T(x), Y(x)) = (\text{Reward, Time elapsed, Next state})$$

  Drawn from a known distribution that depends on the past only through x (ie. conditioning on x or the whole history is the same)
- The objective is to maximize $\mathbb{E}\left[\sum_{i=1}^{\infty} R_i e^{-t_i \beta}\right]$ where t_i is the time of the ith play and R_i is the corresponding reward.

Comment on coupling: Tsitsiklis implicitly has in mind the following model for the generation of the stochastic outcomes. For each arm i and each sample path there is a realization $(x_0^i, T_0^i, R_0^i), (x_1^i, T_1^i, R_1^i), \ldots$ where (x_k^i, T_k^i, R_k^i) is the outcome the agent receives after playing an arm for the k+1st time. You can think of this sequence as being chosen randomly at the beginning, regardless of the applied policy, and every time a policy chooses one arm, it will consume the corresponding next element of the sequence. This construction allows for a nice coupling between the discounted reward earned by different policies.

Reformulation as a problem with a reward rate: While rewards are stochastic, the objective value attained by any policy depends only on the expected rewards at any sate. Rather than receive a lump sum reward $\mathbb{E}[R(x)]$ upon playing a bandit in state x, we can equivalently provide the decision maker with rewards that are received at the known constant rate

$$r(x) = \frac{\mathbb{E}[R(x)]}{\mathbb{E}\left[\int_0^{T(x)} e^{-\beta t} dt\right]}$$

[1] Note that many different formulas for the index may induce the same policy. In particular, scaling the index of every arm and state by a fixed constant does not change the decision rule.

throughout the uncertain duration $T(x)$ of play. To see the equivalence, multiply each side by $\mathbb{E}\left[\int_0^{T(x)} e^{-\beta t} dt\right]$ and bring $r(x)$ inside the integral. The idea of this reformulation plays an important role in the proof.

Our objective is to maximize

$$\mathbb{E}\left[\int_0^{\infty} r(x(t)) e^{-\beta t} dt\right]$$

where $x(t)$ is the state of the bandit played at time t. This maximization is over (non-preemptive) policies which select an action to play at each discrete decision point.

3.1 Priority rule theorem

Theorem 2. *There exists a priority rule which is optimal.*

Definition 3 (Priority rule)**.** *A policy is a priority rule if there exists an ordering of the elements of $\mathcal{X}$ such that at each decision point the bandit with highest whose state is ordered highest is played.*

Proof. The proof will be completed by induction on the state space size which denote with $N = |\mathcal{X}|$

Base : $N = 1$ a priority rule is trivially optimal.
Inductive Step: Assume that a priority rule is optimal for any $\mathcal{X}$ such that $|\mathcal{X}| = k$, we'll show a priority rule is optimal for any $\mathcal{X}$ with $|\mathcal{X}| = k + 1$.

Let $s^\star \in \arg\max_{x \in \mathcal{X}} r(x)$ where $i^\star$ is such that $s^\star \in \mathcal{X}_{i^\star}$. Using Lemma 4, there is an optimal policy in the class $\Pi(s^\star)$ of policies that play $i^\star$ whenever it is in state $s^\star$. We now search for an optimal policy in $\Pi(s^\star)$:

- If $\mathcal{X}_i = \{s^\star\}$, we're done.
- Suppose $|\mathcal{X}_i| \geq 2$. Take $x \in \mathcal{X}_{i^\star}, x \neq s^\star$. If a policy $\pi \in \Pi(s^\star)$ plays $i^\star$, it continues playing $i^\star$ if it transitions to $s^\star$, and continues until some different state is reached. We call this behavior a "composite play" which:
  - Has random duration $\hat{T}(x)$
  - Generates discounted expected reward

$$\mathbb{E}\left[\int_0^{\hat{T}(x)} \overline{r}(t) e^{-\beta t} dt\right] = \mathbb{E}\left[\int_0^{\hat{T}(x)} \hat{r}(t) e^{-\beta t} dt\right]$$

  where

$$\hat{r}(t) = \frac{\mathbb{E}\left[\int_0^{\hat{T}(x)} \overline{r}(t) e^{-\beta t} dt\right]}{\mathbb{E}\left[\int_0^{\hat{T}(x)} e^{-\beta t} dt\right]}$$

  Finding an optimal policy within $\Pi(s^\star)$ is a bandit problem where the size of $X_{i^\star}$ has been reduced by one by removing s^* and plays of other states in $x \in X_{i^*}$ are associated with a modified reward rate $\hat{r}(x)$ and a random duration $\hat{T}(x)$ with modified distribution. We call this process **reducing** bandit i^*. This modification is possible because every time we transition in state $s^\star$ we know the behavior of the optimal policy until the first time the bandit leaves this state. The construction of a composite play reflects that transitions to s^* no longer reflect a decision point so much as uncertain events that influence our calculation of expected rewards and time elapsed.). An optimal policy overall:
  - gives maximum priority to $s^\star$
  - Follows the priority rule over the state space $\mathcal{X} \backslash \{s^\star\}$ (… which exists by the induction hypothesis)

□

3.1.1 Warm up for Lemma 4

To warm up for Lemma 4, let us review what is called an interchange argument. The idea is easier to grasp in a simpler setting:

- There are n projects
- Each project can only be played once
- Each generates a known reward
- In this case we argue that playing the projects in decreasing order of their rewards is optimal

Proof. It is easy to see this claim through an example. In the following sequence the projects are not played in decreasing order of rewards (see 10).

$$6 + \alpha 9 + \alpha^2 5 + \alpha^3 4 + \alpha^4 10 + \alpha^5 3$$

By simply exchanging 10 with 6, we can improve the total reward. In fact, for any non-sorted order cannot be optimal, since swapping elements in this way would increase the total discounted reward. We conclude that the optimal ordering is the sorted one. The argument we are about to give is more akin to moving 10 to the first position to obtain

$$10 + \alpha 6 + \alpha^2 9 + \alpha^3 5 + \alpha^4 4 + \alpha^5 3,$$

which is an improvement because 10 is the highest possible instantaneous reward. □

Lemma 4. *Let $s^\star \in \arg\max_{x \in \mathcal{X}} r(x)$ and let $i^\star$ be such that $s^\star \in \mathcal{X}_{i^\star}$. There is an optimal policy that plays $i^\star$ whenever the state is $s^\star$*

Proof. Suppose π is an optimal policy and bandit $i^\star$ is in state $s^\star$ at time 0. Suppose π does not immediately choose $i^\star$ (otherwise we would be done). Let τ be the first time π plays $i^\star$ (could be $\tau = \infty$).
We now construct a new policy π' which is no worse than π but that chooses $i^\star$:

- π' plays $i^\star$ immediately
- Thereafter, π' mimics π except it skips the first time π plays $i^\star$.
  - This mimickry is possible because the play of bandit i^* does not influence the other bandits. The decision maker knows that had she been following π, the bandit i^* would still be in its initial state and her plays of the other bandits reveal the same state transitions she would have seen under π.

(Note that, like the warmup argument above, we have effectively shifted the reward from playing s^* earlier and simply delayed some of the other rewards)

Let $\overline{r}(t) = r(x(t))$. By the definition of $s^\star$ we have $\overline{r}(t) \leq r(s^\star)$. The expected he expected total discounted reward of policy π is

$$J(\pi) = \mathbb{E}\left[\underbrace{\int_0^\tau \overline{r}(t)e^{-\beta t}dt}_{\text{Initial Value}} + \underbrace{e^{-\beta\tau}\int_0^{T(s^\star)} e^{-\beta t} r(s^\star)dt}_{\text{First time I play } s^\star} + \underbrace{\int_{\tau+T(s^\star)} \overline{r}(t)e^{-\beta t}dt}_{\text{After}}\right]$$

The expected total discounted reward of policy π' is

$$J(\pi') = \mathbb{E}\left[\int_0^{T(s^\star)} e^{-\beta t} r(s^\star)dt + e^{-\beta T(s^\star)}\int_0^\tau \overline{r}(t)e^{-\beta t}dt + \int_{\tau+T(s^\star)} \overline{r}(t)e^{-\beta t}dt\right]$$

We notice that the first two terms differ, whereas the third term is equal in the two rewards. It's intuitive that this exchange should improve the reward for π' as we are placing a higher reward earlier before it is affected by heavier discounting. Computing the difference we have:

$$\begin{aligned} J(\pi') - J(\pi) &= \mathbb{E}\left[(1-e^{-\beta\tau}\int_0^{T(s^\star)} r(s^\star)e^{-\beta t}dt\right] - \mathbb{E}\left[(1-e^{-\beta T(s^\star)})\int_0^{\tau} \overline{r}(t)e^{-\beta t}dt\right] \\ &\geq \mathbb{E}\left[(1-e^{-\beta\tau}\int_0^{T(s^\star)} r(s^\star)e^{-\beta t}dt\right] - \mathbb{E}\left[(1-e^{-\beta T(s^\star)})\int_0^{\tau} r(s^\star)e^{-\beta t}dt\right] \quad \text{using } \overline{r}(t) \leq r(s^\star) \\ &= 0 \end{aligned}$$

To justify the inequality notice that for $\overline{r}(t) = r(s^\star)$ the inequality yields equality. The computations can be verified by the reader. We saw that playing $i^\star$ in that case does not decrease the value of the policy thus it wouldn't decrease the value of the optimal policy either. Therefore there is an optimal policy in the class $\Pi(s^\star)$ consisting of policies that play $i^\star$ whenever it is in state $s^\star$.

□

4 Index Algorithm

The proof produced the following algorithm for computing a Gittins index:

1. Pick a state $s^\star$ with $r(s^\star) = \max_{x\in\mathcal{X}} r(x)$. Define $G(s^\star) = r(s^\star)$ and take $i^\star$ s.t. $s^\star \in \mathcal{X}_{i^\star}$.
2. If $x_{i^\star}$ is a singleton, we remove bandit $i^\star$. Else, we "reduce" bandit $i^\star$ by removing $s^\star$ and go back to 1.

where the "reduce" is from the inductive step of the proof where we removed the state by applying the composite play.

In this procedure, we started with the best state in *any* bandit. We assign the the Index G and remove this state. But since removal of a state only effects the bandit from which it was removed, there is no reason to do this assignment in the whole state space rather than separately accross arms. The following Seperable Index Algorithm produces the same results as the index algorithm above, but makes clear that all computations are carried out separately across the individual bandits.

5 Separable Index Algorithm

1. For $i = 1...n$
   (a) Pick $s^\star$ s.t. $r(s^\star) = \max_{x\in\mathcal{X}_i} r(x)$ and set $G(s^\star) = r(s^\star)$ (index of $s^\star$)
   (b) If $\mathcal{X}_i$ is NOT a singleton, "reduce" Bandit i by removing $s^\star$ and goto (a)

The algorithm works in the following way: We find the best state for every bandit. We "imagine" that we play that bandit whenever they are in that state, and assign an Index for the state. Then we calculate out the effective reward rate of composite plays at any other state of the bandits. We then find the state with highest effective reward rate, assign an index to that state, remove it, and repeat.

6 Exploration vs exploitation with independent beliefs

(Gaussian MAB)
In this problem the decision maker is faced with with n options (arms). She begins with a prior belief on

the true mean payoff of every arm where $\theta_i \sim \mathcal{N}(\mu_{i,0}, \sigma^2_{i,o})$, $i = 1, \ldots, n$ which is drawn independently across arms. If she chooses to play $u_k = i$, she observe a reward $R_k^{u_k} = \theta_{u_k} + w_k$ where $w_k \sim \mathcal{N}(0, \sigma_w^2)$ are iid. The objective is to maximize

$$\mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k R_k^{u_k}\right] \tag{2}$$

This problem can be viewed as a special case of our formulation where the state is not a physical state but the state of our beliefs. The parameters of the posterior distribution $x_t = \big((\mu_{t,i}, \sigma^2_{t,i})_{i=1,\ldots,n}\big)$ evolve according to Bayes rule:

$$x_{t+1,i} = \begin{cases} x_{t,i} & \text{if } u_t \neq i \\ \left(\dfrac{\sigma_{t,i}^{-2}\mu_{t,i} + \sigma_w^{-2} R_t^i}{\sigma_{t,i}^{-2} + \sigma_w^{-2}}, (\sigma_{t,i}^{-2} + \sigma_w^{-2})^{-1}\right) & \text{if } u_t = i \end{cases}$$

Because we began with an independent prior, the (belief) state of bandits that are not played does not evolve in any way since we don't get any information about them. On the other hand if we do play them we update our beliefs using Bayes Rule. The expected rewards of playing bandit i in state (μ, σ^2) is μ. Note that by the tower property of conditional expectation we could replace the random reward $R_k^{u_k}$ in (2) with its conditional mean μ_{k,u_k}.

Because the state-space is infinite, so the proof of we gave of the Gittins index theorem does not apply directly. Thankfully, the Gittins index theorem does apply in this case[2]. The optimal policy plays the action:

$$u_t^\star \in \underset{u \in \{1,\ldots,n\}}{\arg\max}\ G((\mu_{t,u}, \sigma^2_{t,u}))$$

at each time period. The Gittins Index evaluates the quality of an arm considering its posterior mean and variance as well and noise level σ_W. But this evaluation doesn't take into consideration the quality of the other arms, which is surprising.

Unfortunately, the Gittins index in (1) cannot be computed exactly because the state space of each bandit is infinite. It could be carefully approximated. In practice, this would likely be done offline and stored for later use, just as we have done for values of the Gaussian CDF.

But what is G? Generally It's complex . . . but in effect, the fair tax problem defining 1 evaluates the potential upside of the arm. The decision-maker has the option of playing the arm many times if she learns it offers great rewards or abandoning it quickly if she learns otherwise. The fair price for this option is higher when the arm's payout is more uncertain –especially is the problem is has a long time horizon. As one would expect from this interpretation, $G(\mu, \sigma)$ is an increasing function of both the mean and variance of the posterior beliefs.

As the horizon goes to 1, the Gittins index simplifies in a manner that makes this interpretation clear. In particular, as α goes to 1 (. . . an increasingly long horizon),

$$G(\mu, \sigma^2) = \mu + \Phi^{-1}(\alpha)\sigma + o(1)$$

where $\Phi^{-1}(\alpha)$ is the α–quantile of the standard normal distribution and $o(1)$ represents any function that goes to 0 as $\alpha \to 1$. This is precisely a Bayesian upper confidence bound, drawing sharp a connection with a popular heuristic bandit strategy known as UCB. In particular, the Gittins index implements the principle of *optimism in the face of uncertainty*, playing the arm not with the highest expected performance (i.e. highest posterior mean) but the arm that offers the highest payout in the best plausible world (where each arm's true mean is a high quantile of its posterior). To draw a tighter connection with familiar formulas for UCB algorithms, note that $\Phi^{-1}(\alpha) \approx \sqrt{2\log\left(\frac{1}{1-\alpha}\right)}$, and $\sigma \approx \sigma_W \sqrt{\frac{1}{n}}$ if the arm has been sampled n times.

[2] This should not be too surprising since one could approximate the true decision problem arbitrarily well by problems with finite (but enormously large) state spaces, the same way we do throughout all of mathematical analysis.