Course Notes On Dynamic Optimization (Fall 2023)
Lecture 5B: Bandit processes and the Gittins Index Theorem

Instructor: Daniel Russo

Email: djr2174@gsb.columbia.edu

Graduate Instructor: David Cheikhi
Email: d.cheikhi@columbia.edu

These notes are based of scribed notes from a previous edition of the class. I have done some follow up light editing, but there may be typos or errors.

Topics:

- Statement of the gittins index theorem
- Applications to priority rules in scheduling
- Application to upper confidence bound type strategies in learning problems.

1 Gittins Index Theorem and MABs

1.1 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.

1.2 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 indicates $u_k \in \{1, \ldots, n\}$, indicates which 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 immediate rewards depend only only on the state of the bandit that has been chosen.

We will focus on cases where $f_k^i = f_k$.

1.3 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.

1.4 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 $G_u(\cdot) : \mathcal{X}_u$ßℝ 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.

Proposition 1 (Informal)**.** *Under some 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.

1.5 Deriving an expression for the Gittins index

The Gittins index is calculated by considering each bandit arm in isolation. Knowing that the theorem holds, 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 $\lambda, \alpha\lambda, \alpha^2\lambda, \ldots$, (i.e. a safe arm that gives a known reward)
- Arm 1 yields rewards $R(x_0)$, $\alpha R(x_1)$, $\alpha^2 R(x_2), \ldots$ (i.e. a risky project whose state evolves stochastically)

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

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

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

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

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 $G(x) \geqslant \lambda$. 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:

$$G(x) = \sup\{\lambda \mid V_\lambda(x) \geqslant 0\} \tag{1}$$

Weber (1992) interprets this λ as tax per use of arm 1 that effectively changes the reward from $R(x)$ to $R(x) - \lambda$. The Gittins index is the "fair tax" at which a risk neutral decision maker is precisely indifferent between paying the tax to play arm 1 at least once and quitting immediately

1.6 Proofs

This result has fascinated researchers and several innovative proofs have been provided. The textbook gives a sort of explicit argument. Two others that I personally enjoy greatly are:

[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.

1. Richard Weber. "On the Gittins Index for Multiarmed Bandits." Ann. Appl. Probab. 2 (4) 1024 - 1033, November, 1992

2. John N. Tsitsiklis. "A Short Proof of the Gittins Index Theorem." Ann. Appl. Probab. 4 (1) 194 - 199, February, 1994

The first provides a kind of proof in words. It's beautiful, but I think it takes a lot of sophistication on the part of the reader to be 100% confident that it is rigorously true — i.e. it's probably not the best for the course. We'll focus on the second paper, which moves to continuous time which enables a very nice inductive argument.

2 Application to scheduling

See slides.

3 Application to exploration vs exploitation with independent beliefs

In this problem the decision maker is faced with with n options ('arms'). They start with a prior belief on the payoff of every arm $\theta_i \sim \mathcal{N}(\mu_{i,0}, \sigma^2_{i,o})$, $i = 1, \ldots, n$, which is drawn independently across arms. If they choose to play $u_k = i$, they 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

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

The state is not the physical state but the state of our beliefs. The parameters of the posterior distribution $x_t = \left((\mu_{t,i}, \sigma^2_{t,i})_{i=1,\ldots,n}\right)$ evolve according to Bayes rule:

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

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 simply μ. Note that by the tow property 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, 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

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

[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

at each 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.

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 increases, 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, note that $\Phi^{-1}(\alpha) \approx \sqrt{2\log(1/(1-\alpha))}$ and $\sigma \approx \sigma_W/\sqrt{n}$ if an arm has been sampled about n times.