Advanced Topics in Machine Learning and Algorithmic Game Theory

Lecture 7: Bayesian approach to MAB - Gittins index

Lecturer: Yishay Mansour *Scribe: Mariano Schain*

7.1 Introduction

In the Bayesian approach to the *Multi Armed Bandit* problem we assume a statistical model governing the rewards (or costs) observed upon sequentially choosing one of n possible *arms*. We consider a γ-*discounted* setting in which the value of a reward r at time t is $r\gamma^t$. We will see that although searching for an optimal policy (a rule for choosing the next arm, based on history, such that expected rewards are maximal) may be infeasible, the structure of an optimal policy is based on an *index* value that may be computed for each arm independently. The optimal policy will just choose next the arm of highest index, and update the index value (of the chosen arm only) based on the observed result, thereby breaking down the optimization problem to a small set of independent computations.

7.1.1 Example 1 - Single machine scheduling

There are n jobs to be completed but just a single machine. Each job $i \in \{1, \ldots, n\}$ requires S_i machine time to complete. Upon completion of job i at time t, a cost tC_i is charged (i.e a cost rate of C_i per unit of time for unfinished jobs). What is the optimal ordering of jobs to be completed by the machine such that the total cost is minimal?

Claim 7.1 *The optimal ordering of jobs in the single machine scheduling setting is by decreasing* $\frac{C_i}{S_i}$.

Proof: Consider j_1 and j_2, two of the n jobs to be performed sequentially by some policy. Since the costs related to the rest of the jobs are the same regardless of the order in which j_1 and j_2 are performed, we can assume that j_1 and j_2 are the only jobs (i.e. $n = 2$, $j_1 = 1$, $j_2 = 2$). Now, the total costs if performing first j_1 and then j_2 are $C_1S_1 + C_2(S_1 + S_2)$, and the total costs if performing the jobs in reversed order are $C_2S_2 + C_1(S_1 + S_2)$. Therefore, an optimal policy will perform j_1 before j_2 only if $C_1S_1 + C_2(S_1 + S_2) \leq C_2S_2 + C_1(S_1 + S_2)$, i.e. only if $\frac{C_1}{S_1} \geq \frac{C_2}{S_2}$. □

In this first example, we see that the optimal policy is an *index policy*, that is, a policy that is based on an index value function (that may be evaluated independently for each possible option) and at each decision time selects the option having highest index. In the single machine scheduling setting the options at each decision time are the jobs to be handled

next by the machine and the index function of job i is $\frac{C_i}{S_i}$. Also note the simple interchange argument in the proof - we will use similar interchange arguments throughout.

7.1.2 Example 2 - Gold mines

We have n gold mines, each with an initial amount of gold Z_i, $i \in \{1, 2, \ldots, n\}$. We have a single machine that may be used sequentially to extract gold from a mine. When the machine is used in mine i, with probability p_i it will extract a q_i portion of the remaining gold (and may be afterwards used again in the same mine or another), and with probability $1-p_i$ will break (ending the process). We are looking for an optimal policy to select the order of mines to use the machine, such that the expected amount of gold extracted is maximal.

Claim 7.2 *The optimal ordering of mines in the gold mine setting is by selecting the mine with the highest $\frac{p_i q_i x_i}{1-p_i}$, where x_i is the remaining amount of gold in mine $i \in \{1, 2, \ldots, n\}$.*

Proof: We again use an interchange argument. Assume that we consider using the machine in two gold mines 1 and 2, one after the other. Given the gold levels x_1 and x_2 in the mines, compare the expected amount of gold extracted for a policy that uses the machine in gold mine 1 first and (if the machine did not brake) mine 2 afterwards, to the expected amount of gold extracted for a policy that uses the machine in reversed order (note that the expected amount of gold remaining in the mines after using the machine on both mines does not depend on the order). To use first the machine in gold mine 1 we require that

$$p_1(q_1 x_1 + p_2 q_2 x_2) \geq p_2(q_2 x_2 + p_1 q_1 x_1)$$

which holds when $\frac{p_1 q_1 x_1}{1-p_1} \geq \frac{p_2 q_2 x_2}{1-p_2}$. Note that after using the machine in gold mine i (and assuming the machine did not break) the relevant index $\frac{p_i q_i x_i}{1-p_i}$ decreases and therefore the optimal policy will recompute and compare the indices after each usage, choosing to use the machine in the gold mine with higher index at every step. □

7.1.3 Example 3 - Search

An item is placed in one of n boxes. We are given a prior probability $p \in \Delta_n$, where p_i is the prior probability that the item is in box i. At each step we choose one of the n boxes i and if the item is indeed in box i then we find it with probability q_i. If upon searching box i the item is not found, the probability p_i is updated according to bayes' rule:

$$p_i^{new} = Pr(\text{item in box } i | \text{item not found upon searching box } i) = \frac{(1-q_i)p_i}{1-q_i p_i}$$

The cost of searching box i is C_i. We are looking for a policy to sequentially choose boxes to be searched such that the average cost of finding the item is minimal.

Claim 7.3 *The optimal ordering of boxes in the search setting is by decreasing* $\frac{p_i q_i}{C_i}$.

Proof: Again, we use an interchange argument. A similar reasoning as in the previous examples indicates that we may restrict our attention to two boxes only, which without loss of generality we assume are boxes 1 and 2. The average cost of searching box 1 followed (if not found) by searching box 2 is $C_1 + (1 - p_1 q_1)C_2$, while the average cost of searching in the reversed order is $C_2 + (1 - p_2 q_2)C_1$. Therefore[1], we will prefer searching box 1 first if $C_1 + (1 - p_1 q_1)C_2 \leq C_2 + (1 - p_2 q_2)C_1$, that is if $\frac{p_1 q_1}{C_1} \geq \frac{p_2 q_2}{C_2}$. □

7.1.4 Example 4 - Multi Armed Bandit

We are given n arms $B_1 \dots B_n$. Each arm B_i when selected has an (unknown) probability of success θ_i. At a sequence of decision times $t = 0, 1, 2 \dots$ we select an arm i, and (if successful) earn a γ-discounted reward γ^t. Given a prior probability distribution on the values $\{\theta_i\}_{i=1}^n$, our goal is to find an optimal rule for the sequence of arms chosen such that the average of the $\gamma-$discounted sum of rewards over time is maximal. As before, the probability distribution of θ_i is updated according to bayes' rule after observing the result of every selection. For example, if the prior distribution of θ_i is Beta$(1, 1)$ (i.e. uniform over $[0, 1]$) then after observing a_i successes and b_i failures in $a_i + b_i$ selections of arm i, the posterior probability distribution for θ_i is Beta$(1 + a_i, 1 + b_i)$. Note that if the probability distributions for θ_i are Beta(α_i, β_i) then the obvious greedy policy that at each step choses the arm of highest index $\frac{\alpha_i}{\alpha_i+\beta_i}$ is not optimal. This is because given two arms of the same index value $\frac{\alpha_i}{\alpha_i+\beta_i} = \frac{\alpha_j}{\alpha_j+\beta_j}$ but different times used (e.g. $\alpha_i + \beta_i << \alpha_j + \beta_j$) an optimal policy will prefer arm i over j since the substantially larger information gain in observing B_i (which has much higher variance at this point) may be later used to achieve higher expected rewards.

To see how the expected total reward under the optimal policy may be calculated, consider the simple setting $n = 2$ with arm 2 having a fixed known success probability p. Now, $R(\alpha, \beta, p)$, the expected total reward under an optimal policy, when the probability of success of arm 1 is $\theta \sim$ Beta(α, β) satisfies the following recursion:

$$R(\alpha, \beta, p) = \max\{\frac{p}{1-\alpha}, \frac{\alpha}{\alpha+\beta}[1 + \gamma R(\alpha + 1, \beta, p)] + \frac{\beta}{\alpha+\beta}\gamma R(\alpha, \beta + 1, p)\} \qquad (7.1)$$

where $\frac{p}{1-\alpha}$ is the expected reward when choosing arm 2 indefinitely[2], and the other term sums two summands which are the optimal expected rewards when choosing arm 1 and observing a success, or a failure, respectively. We may therefore solve for $R(\alpha, \beta, p)$ iteratively, starting

[1]Note that a search of one of the boxes has no effect on the probability p_i of the other, and therefore the probabilities p_1 and p_2 after searching both boxes are independent of the searching order.

[2]if it is optimal to choose arm 2 once, then it remains optimal thereafter since the information before choosing arm 2 is the same as the information after observing the result

with an approximation[3] for all values of α and β such that $\alpha+\beta=N$ and then calculating iteratively for all values of α and β such that $\alpha+\beta=N-1$ and so on. It can be shown that that the approximation error exponentially[4] decreases with N. An index value for arm 1 given a Beta(α,β) probability of success may be the value of p for which the max in (7.1) is over two expressions of the same value. In what follows we formalize this notion and prove the existence of the Gittins index and its form. We start with the formal model.

7.2 Model

Given n arms $B_1 \dots B_n$. At any time t, each arm B_i may be in a state $x_i(t) \in S_i$. At a sequence of decision times $t_0 = 0, t_1, \dots, t_l, \dots$ we select (control) an arm i. Upon choosing arm i at time t the state of arm B_i (and only B_i) transitions to state $y \in S_i$ according to $p_i(y|x_i(t))$ and we observe a bounded reward $r(x_i(t))$. The interval T until the next decision time $t+T$ is set according to a probability distribution that may also depend on $x_i(t)$.

Our goal is to find a policy (a rule that given the history and the problem parameters selects which arm to control at every decision time) that maximizes the average (over realizations[5]) of the $\gamma-$discounted sum of rewards over time:

$$\sum_{t_l} \gamma^{t_l} r(x_i(t_l)) \tag{7.2}$$

It will be convenient to consider the observed reward $r(s)$ (where s is the state of the selected arm at decision time t) as being 'spread' over the time interval ending in the subsequent decition time $t+T$. We therefore define the reward *rate* $\overline{r}(s)$ as follows:

$$\overline{r}(s) \triangleq \frac{r(s)}{E[\int_0^T \gamma^t dt | x(0)=s]}$$

Note that $E[\int_0^T \gamma^t \overline{r}(s) dt | x(0)=s] = r(s)$ and therefore the two reward methods are equivalent with respect to the target (7.2). It will also be convenient to refer to the arm choice process as being continuous between decision times - i.e. the arm is being chosen throughout the time period (resulting in $\overline{r}(s)$ reward per unit of time) until the next decision time. Now, we define for a fixed time interval $[0,T)$

$$w(T) \triangleq \int_0^T \gamma^t dt = \frac{1-\gamma^T}{\ln \frac{1}{\gamma}} \tag{7.3}$$

[3] larger values of $\alpha+\beta$ imply higher concentration around the true success probability θ, and therefore we are able to provide increasingly good approximations of R as we increase the initial $\alpha+\beta$

[4] an ϵ-approximation to R for $\alpha+\beta=N$ results in an $\epsilon\gamma$-approximation to R for $\alpha+\beta=N-1$

[5] all expectations are over realizations, unless explicitly indicated otherwise

And note that for such a fixed T we have

$$r(s) = w(T)\overline{r}(s) \tag{7.4}$$

It is assumed that at every decision time t all the states $x(t) = (x_1(t), \ldots, x_n(t))$ and problem parameters (e.g. the discount factor γ, the transition distributions p_i and reward function r) are known to the policy. Therefore, optimizing (7.2) is possible by state space evaluation methods such as dynamic programming. Such methods however are computationally infeasible due to the exponential size of the state space.

In what follows we will see that the optimal policy for (7.2) is an *index* policy - a policy that assigns to each arm an index value that only depends on its state (and not on the states of the other arms) and at each decision time selects the arm of highest index value. In doing so, we replace a problem of evaluating values of $\prod_i |S_i|$ states (exponential in n) with n independent computations of the values of $|S_i|$ states for each arm.

7.3 First proof: Finite number of states

Without loss of generality we may assume that all arms are identical, with the same state space $S = \bigcup S_i$, and only differ by their initial state (any underlying state independence is reflected in the state transition function). We first show that at any decision time it is optimal to choose the arm of maximal reward rate, and then we use this to prove (by induction on the number of states $|S|$) that an optimal index policy exists. Furthermore, the construction in the proof will serve to define the index.

Claim 7.4 *It is optimal to choose an arm which is in state* $s_N = \arg\max_{s \in S} \overline{r}(s)$

Proof: Note that it is not necessarily the case that there *is* an arm in state s_N, the claim is that *if* there is then any optimal policy will choose it right away. Assume that it is arm B_1 in state s_N at time 0 ($x_1(0) = s_N$). We use a simple interchange argument: assume there is an optimal policy π that does not choose s_N at time 0, and instead chooses at a sequence of decision times a sequence of arms in states different than s_N until eventually (after a period of length τ, collecting an accumulated reward R) chooses B_1 until the next decision time $\tau + T$. The reward observed by π during the interval $[0, \tau + T)$ is $R + \gamma^\tau r(s_N) = R + \gamma^\tau w(T)\overline{r}(s_N)$.

We will compare the accumulated reward of π with that of a policy π' that chooses B_1 at time 0 for a period of length T and then chooses the same sequence as π during a period of length τ and is identical to π thereafter (note that the states of the arms at time $T + \tau$ is the same for both policy realizations). The reward observed by π' during the interval $[0, \tau + T)$ is $r(s_N) + \gamma^T R = w(T)\overline{r}(s_N) + \gamma^T R$. We consider the difference between the reward of π' and the reward of π:

$$w(T)\overline{r}(s_N) + \gamma^T R - (R + \gamma^\tau w(T)\overline{r}(s_N)) = w(T)\overline{r}(s_N) - R(1 - \gamma^T) - \gamma^\tau w(T)\overline{r}(s_N)$$

Now, by the definition of s_N we have that $R \leq w(\tau)\overline{r}(s_N)$ and therefore the above difference is at least

$$\overline{r}(s_N)[w(T) - w(\tau)(1-\gamma^T) - \gamma^\tau w(T)] = \overline{r}(s_N)[w(T)(1-\gamma^\tau) - w(\tau)(1-\gamma^T)] = 0$$

where the last equality is by (7.3). We conclude that choosing the state of global maximum reward rate is optimal. □

We now use the claim to constructively prove that an optimal index policy exists:

Theorem 7.5 *If the number of state S is finite ($|S| = N$), then there exists an optimal index policy. Furthermore, the index values may be iteratively computed as follows:*

$$v(s_j) = \frac{E[\sum_{t_l<\tau} r(x(t_l))\gamma^{t_l} dt | x(0) = s_j]}{E[\int_0^\tau \gamma^t dt | x(0) = s_j]}, \qquad j = N, N-1, \ldots, 1 \tag{7.5}$$

Where the expectations above are over realizations that start with an arm at state $x(0) = s_j$ and continue (arm chosen again and again at decision times t_l) until a decision time τ in which the state of the arm is no longer in the set of already computed 'higher priority' values $\{s_N, \ldots, s_{j+1}\}$

Proof: First we prove by induction on the number of states that there is an optimal index policy (i.e. that there is an ordering of the states such that it is optimal to choose the state of highest order). When there is a single state this is trivial. Now, assume the existence of such an ordering for a problem of $N-1$ states. We can now consider a modification of the given problem to a problem of $N-1$ states such that the rewards and decision times of an optimal policy for the original setting are the same as the rewards and decision times of an optimal index policy for the modified setting:

We eliminate the state of highest reward rate (s_N) by modifying the probabilities of transitions $p(y|s)$, reward rates $\overline{r}(s)$, and decision times $T(s)$ such that whenever an arm reaches state s_N at a decision time it is automatically selected (therefore the actual decision times in the modified setting are until no arm is in state s_N). By the inductive assumption, there is an optimal index policy for the modified setting (implying an ordering of the $N-1$ states at every decision time, that only depends on the state). By the claim above, any optimal policy for the original setting of N states selects an arm at state s_N when available. Therefore, the combination of the selection rule of state s_N with the optimal index policy for the other $N-1$ states forms an optimal index policy for the original setting.

We now turn to explicitly formulate the index value based on the above construction. First note that $\overline{r}(s_N) \geq \overline{r}_1(s_{N-1})$ where $\overline{r}_1(s_{N-1})$ is the maximal reward rate of the best arm s_{N-1} in the modified setting not including s_N. Therefore the list of non-increasing, iteratively computed values

$$v(s_j) = \overline{r}_{N-j}(s_j), \quad j = N, N-1, \ldots, 1$$

may serve as the index values of the states in S, where $\overline{r}_{N-j}(s_j)$ is the maximal reward rate of the best arm s_j in the modified setting not including $\{s_N, \ldots, s_{j+1}\}$. By the construction of the modified settings we have (7.5). □

7.4 Gittins Index

In this section we will explore the general form of an optimal index policy assuming that it exists. Two additional existence proofs (not assuming finite state space) are given in subsequent sections. To simplify notation we assume from now on that the decision times are fixed at times $t = 0, 1, 2....$ The results apply and are easy to generalize to the case of random decision times.

We start by observing that the infinite horizon accumulated rewards of a single state fixed $\lambda-$reward arm is $\frac{\lambda}{1-\gamma}$. We denote such an arm by $B(\lambda)$. In a setting of two arms, B and $B(\lambda)$, an optimal policy that switches from arm B (that started in state s_0) to arm $B(\lambda)$ at some decision time $\tau > 0$ will never switch back to B (the information regarding B in future decision times is the same as the information that was available at time τ and resulted in choosing $B(\lambda)$). We conclude that the maximal average reward is the optimal choice of the *stopping time* τ:

$$\sup_{\tau>0} E[\sum_{t=0}^{\tau-1} \gamma^t r(x(t)) + \gamma^\tau \frac{\lambda}{1-\gamma} | x(0) = s_0] \tag{7.6}$$

where the average is over all realizations of the state transitions and rewards of arm B, and the supremum is over all functions τ that associate a stopping time in $\{1, 2, \ldots\}$ to a realized states history[6]. We are looking for the fixed reward λ^* that makes the two arms equivalent (equally optimal to switch to $B(\lambda^*)$ initially, or wait for the optimal switch time, and therefore may serve as the index value of arm B at state s_0), that is, satisfying

$$\sup_{\tau>0} E[\sum_{t=0}^{\tau-1} \gamma^t r(x(t)) + \gamma^\tau \frac{\lambda^*}{1-\gamma} | x(0) = s_0] = \frac{\lambda^*}{1-\gamma}$$

or equivalently

$$\sup_{\tau>0} E[\sum_{t=0}^{\tau-1} \gamma^t r(x(t)) - (1-\gamma^\tau) \frac{\lambda^*}{1-\gamma} | x(0) = s_0] = 0$$

The left hand side of the above equation, the supremum of a decreasing linear function of λ is convex and decreasing in λ. Therefore, the above equation has a single root that may

[6] A stopping time is a mapping from histories to a decision of either to continue or to stop

also be expressed as follows (since $\frac{1-\gamma^\tau}{1-\gamma} = \sum_{t=1}^{\tau-1}\gamma^t$):

$$\lambda^* = \sup\{\lambda | \sup_{\tau>0} E[\sum_{t=0}^{\tau-1}\gamma^t[r(x(t)) - \lambda]|x(0) = s_0] \geq 0\} \tag{7.7}$$

The above provides an economic interpretation of λ^* as the highest rent (per period) someone (who has an optimal stopping policy τ) may be willing to pay for receiving the rewards of B. From (7.7) we get that λ^* (the index value of arm B at state s_0) is of the following form:

$$v(B, s_0) \triangleq \lambda^* = \sup_{\tau>0} \frac{E[\sum_{t=0}^{\tau-1}\gamma^t r(x(t))|x(0) = s_0]}{E[\sum_{t=0}^{\tau-1}\gamma^t|x(0) = s_0]} \tag{7.8}$$

Note that it is a legitimate index since it only depends on the state and parameters of B. Note also that (7.8) coincides with (7.5) since the optimal stopping time τ is inherent in the construction described in the proof of Theorem 7.5.

Finally, consider the optimal stopping time τ in (7.7), which is characterized by the set of stopping states $\Theta(s_0)$. It can be shown that any state s having index value $v(B, s) < v(B, s_0)$ must be a stopping state, and any stopping state s must satisfy $v(B, s) \leq v(B, s_0)$:

$$\{s|v(B, s) < v(B, s_0)\} \subseteq \Theta(s_0) \subseteq \{s|v(B, s) \leq v(B, s_0)\}$$

This implies that an optimal policy will not stop at a state having higher index value than the index value of the initial state, and will always switch if reaching a state of lower index value than that of the initial state. The following example illustrates the power of using the index.

7.4.1 Example 5 - Coins

Consider the following coins problem: given n biased coins (coin i having probability of heads p_i) we earn a reward γ^t for a head tossed at time t. It is easy to see that the optimal tossing order is by decreasing p_i. Now, assume that the heads probability of coin i is p_{ij} when tossed for the j^{th} time. If p_{ij} is nonincreasing (i.e. $p_{i1} \geq p_{i2} \geq \ldots$) for every i then again tossing by decreasing p_{ij} is optimal. However, in the general case (where p_{ij} is not necessarily decreasing) we can use the index (7.8) to define for each coin i its index value:

$$v_i = \max_{\tau\geq 1} \frac{\sum_{j=0}^{\tau-1}\gamma^j p_{ij}}{\sum_{j=0}^{\tau-1}\gamma^j}$$

Note that state transitions are deterministic and the expectations over realizations (of rewards) are reflected in the values p_{ij} in the expression above. The optimal policy will identify the optimal stopping time τ^* of the coin with highest index value $i^* = \arg\max_i v_i$, will toss τ^* times coin i^* , and advance its state accordingly. The policy may now recompute the index value of coin i^* and repeat.

7.5 Proof by economic interpretation

In this section and the following we present two proof of the index theorem (no longer assuming finite number of states):

Theorem 7.6 *An Index policy with respect to*

$$v(B_i, s) = \sup_{\tau>0} \frac{E[\sum_{t=0}^{\tau-1} \gamma^t r(x_i(t))|x_i(0) = s]}{E[\sum_{t=0}^{\tau-1} \gamma^t |x_i(0) = s]}$$

is optimal.

Proof: We use the economic interpretation following (7.7): assume that to use an arm B_i that is in state $x_i(t)$ at time t, a *prevailing charge* $\lambda_{i,t}$ must be paid. A too low charge will result in endless usage of the arm, while a too high charge will result in an abandoned arm. Let the *fair charge* be the charge for which we are indifferent between using the arm (for a sequence of times, until an optimal future stopping time τ) or not. The fair charge $\lambda_i(x_i(t))$ is given by λ^* of (7.7) and the related optimal usage time (given the state of the arm is $x_i(t)$) is the τ that attains the supremum, denoted $\tau(x_i(t))$.

Now, we set the prevailing charges of arm B_i as follows: initially ($t_0 = 0$) set $\lambda_{i,t_0} = \lambda_i(x_i(t_0))$. Thereafter, the prevailing charge is kept constant until time $t_1 = t_0 + \tau(x_i(t_0))$. By optimality of t_1, at that time the prevailing charge was (for the first time) higher than the fair charge, so we reduce the prevailing charge and set $\lambda_{i,t_1} = \lambda_i(x_i(t_1))$, keeping it constant until time $t_2 = t_1 + \tau(x_i(t_1))$. And so on, creating a nonincreasing series of prevailing charges $\lambda_{i,t} = \min_{t' \le t} \lambda_i(x_i(t'))$. By the construction, for arm B_i, the prevailing charges are never more than the fair charges: $\lambda_{i,t} \le \lambda_i(x_i(t)))$.

Finally, consider a setting of n arms $B_1, \ldots, B_n$ with prevailing charges $\lambda_{i,t}$ set as previously described (where t represents for each arm its process time - the number of times the arm has been selected). Note the perfect analogy to the setting of section 7.4.1 with nonincreasing probabilities p_{ij}. Now, since at any time no profit can be made from any selected arm, the expected total discounted sum of rewards is upper bounded by the discounted sum of prevailing charges paid by any policy that selects one of the n arms sequentially. However, those two quantities are equal for the policy that at each time selects the arm of highest prevailing charge, and therefore such a policy is optimal. We conclude that the prevailing charge $\lambda_{i,t}$ (which is always equal to the fair charge when selected) is the Gittins index as defined in (7.7) and (7.8). □

7.6 Proof by interchange arguments

In this section we present yet another proof of Theorem 7.6. Using the notation established in the previous secion and denoting the numerator and denominator of the index defined in

theorem 7.6 by $R_\tau(B_i, s)$ and $W_\tau(B_i, s)$ respectively, we have $\lambda_i(x_i) = \sup_{\tau>0} \frac{R_\tau(B_i,x_i)}{W_\tau(B_i,x_i)}$. We first prove the following interchange claim:

Claim 7.7 *For two arms B_1 and B_2 at states x_1 and x_2 respectively at time t, if $\lambda_1(x_1) > \lambda_2(x_2)$ with $\tau = \tau(x_1)$ the optimal stopping time of B_1 at state x_1, and σ an arbitrary stopping time for B_2 at time state x_2 then the expected reward is higher when selecting B_1 for a period τ and then selecting B_2 for a period σ than the expected reward when the order is reversed.*

Proof: $\lambda_1(x_1) > \lambda_2(x_2) \Rightarrow \frac{R_\tau(B_1,x_1)}{W_\tau(B_1,x_1)} > \frac{R_\sigma(B_2,x_2)}{W_\sigma(B_2,x_2)}$. Now, since for any $\sigma > 0$ we have $W_\sigma(B_i, s) = \frac{1-E[\gamma^\sigma|x_0=s]}{1-\gamma}$, the last inequality us equivalent to $\frac{R_\tau(B_1,x_1)}{1-E[\gamma^\tau|x_1]} > \frac{R_\sigma(B_2,x_2)}{1-E[\gamma^\sigma|x_2]}$ which in turn is equivalent to $R_\tau(B_1, x_1) + E[\gamma^\tau|x_1]R_\sigma(B_2, x_2) > R_\tau(B_2, x_2) + E[\gamma^\sigma|x_2]R_\tau(B_1, x_1)$. The left side of this last inequality is the expected reward when selecting B_1 for a period τ and then selecting B_2 for a period σ, while the right side is the expected reward when the order is reversed. □

We are now ready to prove the theorem:

Proof:(of theorem 7.6) For a given setting and the index (7.8) define a parameterized class of policies Π_k. A policy π is in Π_k if it makes at most k arm selections that are not the arm of highest index value (at decision time). We will show by induction on k that an optimal policy belongs to Π_0. First, consider $\pi \in \Pi_1$. We use the interchange claim 7.7 to show that π is not optimal. Indeed, consider the time t_0 in which π deviates and selects arm B_2 (having index λ_{2,t_0}) instead of arm B_1 of maximal index[7] $\lambda_{1,t_0} > \lambda_{2,t_0}$ (without loss of generality we may assume $t_0 = 0$). Since π may not deviate again, arm B_1 will get selected as soon as $\lambda_{2,\sigma} < \lambda_{1,0}$, and remain selected for the optimal period τ. By the interchange claim 7.7, the reward of π during time $\sigma + \tau$ is less than the reward of a policy π' that reverses the arms order and selects arm B_1 first for a period of length τ followed by arm B_2 for a period of length σ (and is identical to π thereafter). Note that the states of B_1 and B_2 at time $\tau + \sigma$ do not depend on which policy was used. We conclude that π is not optimal and that optimal policies restricted to Π_1 should never exercise the (single) option to deviate. Therefore, optimal policies restricted to Π_k should never exercise their last option to deviate, and (inductively restricting attention to $\Pi_{k-1}, \Pi_{k-2}, \ldots$) we conclude that the Gittins index policy is optimal in Π_k. We are not done since there might be a better policy in Π_∞, which is not accounted for in the induction. Assume that the optimal policy Π^* is in Π_∞ and not Π_0. Given any $\epsilon > 0$, for a sufficiently large k there exists an ϵ-optimal policy in Π_k (since ϵ determines a time horizon after which the discounted rewards are of negligible influence) which, by the above reasoning belongs to Π_0. Since Π_0 holds an optimal policy for any $\epsilon > 0$, it also holds the optimal policy. □

[7]Note that if multiple arms have maximal index (i.e. in case B_1 is not unique) it does not matter which arm of maximal index is selected first, and therefore without loss of generality we may assume that B_1 is selected.