Project: Multi-armed bandit

Student name: *Yacine Thabet*

Course: *ECSE 506: Stochastic Control and Decision Theory* – Professor: *Dr. Aditya Mahajan*
Due date: *April* 30^{th}*, 2020*

Abstract

The multi-armed bandit problem is a popular problem presenting different proofs, the one that is widely explored is the proofs using the Gittins index which is presenting many formulation. In this work we will focus on Whittle [4] proof to show the optimality of the process.

1. Introduction

The multi-armed bandit (MAB) problems are a set of problems which are defined by a single agent, the decision maker which have a choice of taking action between multiple different possibilites called jobs by Gittins et al. in 1979 [3] or projects by Whittle et al. in 1980 [4] sequentially, meaning that at each time step the agent takes an action on one and only one project among all the projects. Taking an action on one project could give an immediate reward. The projects that are not involved at a time step bring no reward and the overall goal is to maximize the expected total reward by deciding which project we should activate at each time step. This is what we can call "sequential resessource allocation problems", because at each time step the decision maker faces a decision by allocating or not the limited ressource available to complete a project. These problems are mainly found in ad placement, crown sourcing, sensor management, search theory, clinical trials and now in reinforcement learning etc.

Basically the MAB problems are focusing on the constant conflict of choosing the project that brings the best immediate reward or exploring and expect having a better option.

This basic version of the multi-armed bandit problem involves a sequence of decisions which can be based only on known informations. Then it could be solved using the dynamic programming presented by Bellman in 1957 [1] and an optimal policy can be found as presented by Gittins et al. in 1979 [3] using the Ginttins index. There are multiple proofs, but for this work we will focus on presenting the one done by Whittle et al. in 1980 [4].

2. The multi-armed bandit problem

In this section we will present the MAB problem by firstly introducing the case of one arm bandit and then extend it to the multi-armed following the scheme done by Gittins et al. in 1979 [3].

2.1. The single-armed bandit

A simple analogy of the single-armed bandit problem, is the exemple of the slot machine with one arm, at each time step we invest a certain amount of money and pull the arm expecting a reward.

We can express the stat of the single-armed bandit for our problem at a time $t \in \{1, 2, \ldots T\}$ which can be denoted by $X_t \in \mathcal{X} = \mathbb{R}$. This state is a function of the policy $U_t \in \mathcal{U} = \{1, 0\}$, 1 meaning to pull the arm and 0 to not pull the arm at time t.

Thus we have:

$$X_{t+1} = f_t(X_0, X_1, \ldots, X_t, W_t) \tag{1}$$

With f_t a given function and $W_t \in \mathbb{R}$ a process noise independant of the state.

At each time step t we expect a reward wich is a function of the state:

$$R : \mathcal{X} \to \mathbb{R}^+ \tag{2}$$

$$X_t \mapsto R(X_t, U_t) \tag{3}$$

That characterizes our problem with the pairs of state-reward overtime :
$(\{X_0, R(X_0, U_0)\}, \{X_1, R(X_1, U_1)\}, \ldots, \{X_T, R(X_T, U_T)\})$.

2.2. The multi-armed bandit problem

The multi-armed bandit problem can be analogeous to an extansion of the example we presented previously with a slot machine, but instead of having only one arm to pull, we have a number of k arms and at each time step $t \in \{1, 2, \ldots, T\}$ we can pull only on of them, that is why the MAB problem can be called the k-armed bandit problem.

Then our system is described by its state vector $\mathbf{X}_t \in \mathcal{X} = \mathbb{R}^k$, with $i \in \{1, 2, \ldots, k\}$ denoting the i^{th} arm, and $U_t^i \in \{0, 1\}$ denoting the policy taken on the i^{th} arm at time t, 1 if the arm is pulled and 0 if not. Knowing that at one time step t only one arm can be pulled and the orthers remain frozen.

Then we have the dynamics for one arm:

$$X_{t+1}^i = \begin{cases} f_t(X_0^i, X_1^i, \ldots, X_t^i, W_t^i), & \text{if } U_{t+1}^i = 1 \\ X_t^i, & \text{if } U_{t+1}^i = 0 \end{cases} \tag{4}$$

With the following reward for that arm:

$$R^i(X_t^i, U_t^i) = \begin{cases} R^i(X_t^i, U_t^i), & \text{if } U_{t+1}^i = 1 \\ 0, & \text{if } U_{t+1}^i = 0 \end{cases} \tag{5}$$

We can denote our control action at a time step t by a vector $\mathbf{U}_t = (U_t^1, U_t^2, \ldots, U_t^k)$ and the sequence of actions over time $\{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_T\}$ if we suppose that we run our experience until a certain time T. Thus we have what we can call a scheduling policy $\mathbf{g} := (g_1, g_2, \ldots, g_T)$ which is a decision rule such that at time step t, the control action $\mathbf{U}_t$ takes value in the vectors of the standard basis $(e_1, \ldots, e_k)$. Our decision policy is a

Markov decision policy which can be written as follows:

$$\mathbf{U}_t = g_t(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_t, \mathbf{u}_1, \mathbf{U}_2, \dots, \mathbf{U}_{t-1}) \tag{6}$$

Then we can state our MAB prolem as: we need to determine the scheduling policy g that maximzes

$$J^g := \mathbb{E}[\sum_{t=0}^{T} \beta^t \sum_{i=1}^{k} R^i(X_t^i, U_t^i)|\mathbf{X}_0] \tag{7}$$

With β the discounting factor taking values in $[0, 1[$.

The MAB problem, as a sequential decision making problem can be solved using dynamic programming with backward induction. But no real solution could be made because of the high dimensionality of the state space, which can cause complexity problems. Until that Gittins et al. in 1979 [3] found that we can prove the optimality of the solution using using an index type solution. This solution is done using a dynamic allocation index (DAI) later called the Gittins index which is a priority index calculated for each project independantly from the others, and the optimal action is given by activating the project which has the highest index. Thus this reduces the problem of finding the solution for k single-armed bandit instead of solving the problem for a k-armed bandit.

3. The multi-armed bandit solution

As shown by Frostig et al. in 2016 [2], there are different ways to prove the optimality of the Gittins index solution. Several proofs were found in the past decades but as explained by Frostig et al. in 2016 [2] these proofs follow a common structure. They all start with the study of the single-armed bandit problem and find its solution,then they extend it to solve the multi-armed bandit problem using the Gittins index rule and showing that it is optimal.

3.1. The MAB formulation

As presented by Frostig et al. in 2016 [2], there are multiple ways to formulate the problem and these formulation are the roots of various proofs.

3.1.1. Gittins formulation: Playing against a standard arm. We have two single-armed bandit, one with a state X_t and the other one, the standard arm, which never changes state, but whenever it is playes gives a constant reward γ. This is called $1\frac{1}{2}$ by Gittins [3] because one arm is fixed. The optimality equations are given by:

$$V_t^s(X_t, \gamma) = \max\{R(X_t, U_t) + \beta\mathbb{E}[V_{t+1}^s(X_{t+1})|X_t], \gamma + \beta V_t^s(X_t)\} \tag{8}$$

3.1.2. Weber formulation: The fixed charge problem. For this formulation we have a single-armed bandit. At each time step t we could either pull the arm by spending a fixed cost γ and get the reward $R(X_t, U_t)$ or not to pull the arm and wait for the next time step $t+1$. The optimality equations are given by:

$$V_t^f(X_t, \gamma) = \max\{R(X_t, U_t) - \gamma + \beta\mathbb{E}[V_{t+1}^f(X_{t+1})|X_t], \beta V_t^f(X_t)\} \tag{9}$$

3.1.3. Whittle formulation: The retirement option problem.

Whittle formulation assume that there is a single-armed bandit which can pull the arm as much as he wants, but can retire forever at any time and get a fixed reward M. The optimality equations are given by:

$$V_t^r(X_t, M) = \max\{R(X_t, U_t)) + \beta\mathbb{E}[V_{t+1}^r(X_{t+1})|X_t], M\} \tag{10}$$

All these problems fall in the category of the optimal stopping problems, because at some points it might be optimal to stop playing the standard arm in the Gittins formulation, or not to spend the cost in the Weber formulation or either to stop playing in the Whittle formulation. And even if the formulations are quite different, if we fixe $M = \frac{\gamma}{1-\beta}$ thus by rewritting the formulations we have:

$$V_t^s(x, \gamma) = V_t^f(x, \gamma) + \frac{\gamma}{1-\beta} = V_t^r(x, \frac{\gamma}{1-\beta}) \tag{11}$$

3.2. Gittins index solution

As explained, since the backward induction solution is computationally heavy, the solution using the Gittins index with a forward induction process is way less complex, but a problem arise.The forward induction processes are are usually suboptimal because they use "myopic" policies.

To get an overview of how the forward induction process works, let's suppose that our experience runs over time. Our goal is to find the policy which maximises the exepected reward over the next τ steps, this is called a τ look ahead policy. The biggest is τ the most information we will get about a policy and about its optimality, but there is a trade off with the computational ressources.

But using the following assumptions we will show that forward induction processes can be optimal for the MAB case:

1. Only one project is actionned at each time step
2. The projects that are not actionned remain frozen
3. All the projects are independant from another
4. The frozen project brings no reward

Thus, by forward induction the process to find the optimal policy is as follows:

Step 1: At the inital state, for each arm $i \in \{1,2,\ldots,k\}$, we need to maximize the following reward rate (expected discounted reward per unit of expected discounted time):

$$v_t(X_0^i) := \max_{\tau \geq 0} \frac{\mathbb{E}[\sum_{t_s=t}^{t+\tau-1} \beta^{t_s} R^i(X_{t+t_s}^i)|X_0^i]}{\mathbb{E}[\sum_{t_s=t}^{t+\tau-1} \beta^{t_s}|X_0^i]} \tag{12}$$

Step 2: Select the arm j with the highest reward:

$$j = arg\max_i \{v_t(X_t^i)\} \tag{13}$$

Step 3: Calling τ^j the corresponding stopping time for this arm, repeat the process by pulling the arm j for τ^j

Step 4: $t = t + \tau^j$ and we repeat from the process from step 1.

3.3. An example

We can present a small example given by Gittins et al. in 1979 [3] to show why the forwards induction policies are optimal for simple families of alternative processes but not for all Markov decision process. Considering an example of a car travelling on the road r_1 in one direction with many roads r_i intersecting with this road. Each road has its own speed limit v_i and our goal is to maximize the total discounted distance traveled ovenr the discounted time (which could be infinite). The first option would to take the road with the highest speed limit before the discounting factor becomes smaller with time. A forward induction policy in that case would pick the road with the highest ratio between distance and speed limit as long as we do not cross a road with a higher speed limit. We could prefer this road over a much road with a smaller speed limit but could lead to a road with a much higher speed limit. But we could end in to a unique intersection which leads us to a road with a smaller speed limit than the previous roads we crossed in our journey. In conclusion using forward unduction could lead us to a suboptimal policy. But we still have access to roads that we previously rejected we could explore more options that could lead us to optimal policies. That is the setup of the multi-armed bandit, because playing an arm do not affect the orther ams since they are frozen.

4. Whittle proof of Gittins index rule optimality

As presented by Frostig et al. in 2016 [2] there a many proofs to proove the optimality of the solution done by Gittins et al in 1979 [3], we will present a simple proof done by Whittle et al. in 1980 [4] using his formulation of the retirement option problem by presenting at first some preliminary lemmas and then prove the main theorem.

In his formulation Whittle in 1980 [4] presents a formulation of the MAB with k arms with the state vector $\mathbf{X_t} = (X_t^1, X_t^2, \ldots, X_t^k) \in \mathcal{X} = \mathbb{R}^k$ evolving over time with $t \in \{1,2,\ldots\}$. We have the vector of actions taken $\mathbf{U_t} = (U_t^1, U_t^2, \ldots, U_t^k) \in (e_1, e_2, \ldots, e_k)$.

The reward for each arm $i \in \{1, 2, \ldots, k\}$ is assumed bounded:

$$k(1-\beta) \leq R_t^i(X_t^i) \leq K(1-\beta), \tag{14}$$

with $k, K \in \mathbb{R}$ constants, and $\beta \in [0, 1[$ the discounting factor. $R_t^i(X_t^i, U_t^i)$ represents the reward given by the arm i at time t which can be simplified by R_t because only the actionned arm brings a reward. The total discounted reward can be written as :

$$r_t = \sum_t \beta^t R_t \tag{15}$$

With the retirement reward M, for the arm $i \in \{1, 2, \ldots, k\}$ we define the value function as follows:

$$V_t^r(\mathbf{X_t}, M) = \max\{R(\mathbf{X_t}, \mathbf{U_t}) + \beta\mathbb{E}[V_{t+1}(\mathbf{X_{t+1}})|\mathbf{X_t}], M\} \tag{16}$$

4.1. Premilinary lemmas

Lemma 1:

1. $V_t^r(\mathbf{X_t}, M) = R(\mathbf{X_t}, \mathbf{U_t}) + \beta\mathbb{E}[V_{t+1}(\mathbf{X_{t+1}})|\mathbf{X_t}]$ for $M \leq k$
2. $V_t^r(\mathbf{X_t}, M) = M$ for $M \geq K$
3. $V_t^r(\mathbf{X_t}, M)$ is a non decreasing convex function of M

Proof. Properties 1 and 2 are obvious because it is optimal to continue if $M \geq k$ or to stop if $M \leq K$. For propertie 3, we fix $\mathbf{x} \in \mathcal{X}$, if we increase the retirement reward M, the function is increasing, since we can only increase the retirement reward the non-decreasing property is obvious. For the convexity, we start by taking any policy g with T_r the retirement time, we define:

$$V^{r,g}(\mathbf{x}, M) = (r_{T_r}(\mathbf{x}) + M\mathbb{E}^g[\beta^{T_r}]), \tag{17}$$

where $r_{T_r}(\mathbf{x})$ represents the reward up to T_r. $V^r(\mathbf{x}, M)$ is the supremem of these function over all policies, and it is convex a supremum of a convection function of M.

□

Let T_r denote the retirement time, at that time we retire and get the reward M. $V^r(\mathbf{x}, M)$ is differentiable for almost all $M \in R$ thus it has subgradients.

Lemma 2: For almost all M we have:

$$\frac{\partial V^r(\mathbf{x}, M)}{\partial(M)} = \mathbb{E}[\beta^{T_r}|\mathbf{X}_0 = \mathbf{x}] \tag{18}$$

Proof. For simplicity of notation with the Bellman operator let's denote $V^r(x, M) = V^{r,M}(x)$. Let g denote any policy and g_M denote the optimal policy for the retirement option when the retirement reward is M. Then, we have:

$$V^{r,M} = \mathcal{B}V^{r,M} \tag{19}$$

Using the porperty of the Bellman operator:

$$\mathcal{B}V^{r,M} = \mathcal{B}_{g_M}V^{r,M} \tag{20}$$

Let's denote $M' = M + \delta$.

$$V^{r,M'} - V^{r,M} = \mathcal{B}V^{r,M'} - \mathcal{B}V^{r,M} \geq \mathcal{B}_{g_M}V^{r,M'} - \mathcal{B}_{g_M}V^{r,M} \tag{21}$$

Since we are using the optimal policy g_m for the stopping process with reward M, which is not optimal for the stopping process with reward $M + \delta$.

The optimal policy g_M plays the bandit process until the stopping time T_r (which could be infinite if the retirement option is never taken).

$$V^{r,M}(\mathbf{x}) = \mathbb{E}^{g_M}[\sum_{t=1}^{T_r-1} \beta^t R_t + \beta^{T_r} M | \mathbf{X}_0 = \mathbf{x}] \tag{22}$$

Hence,

$$\mathcal{B}_{g_M}V^{r,M'} - \mathcal{B}_g V^{r,M} \geq \mathbb{E}^{g_M}[\beta^T \underbrace{(M' - M)}_{=\delta}] \tag{23}$$

That gives us:

$$V^{r,M'} - V^{r,M} \geq \delta \mathbb{E}^{g_M}[\beta_r^T | \mathbf{X}_0 = \mathbf{x}] \tag{24}$$

That gives us that $\mathbb{E}^{g_M}[\beta_r^T | \mathbf{X}_0 = \mathbf{x}]$ is a subgradient of $V^r(\mathbf{x}, M)$ as a function of M and then coincides with the gradient of $V^r(\mathbf{x}, M)$ whenever it exist. Since $V^r(\mathbf{x}, M)$ is convex, we will have a gradient almost on every point, that gives us the existence of the optimal policy as well.

□

4.2. Whittle proof of the retirement option formulation

Whittle et al. in 1980 [4] presented a proof which was re-explained later by Frostig et al. 2016 [2] with the k multi-armed bandit with any state $\mathbf{X} = (X^1, X^2, \ldots, X^k)$ by defining for the i^{th} arm:

$$M^i(X^i) = \inf_M \{V^r(X^i, M) = M\} \tag{25}$$

Wich represents the smallest retirement reward that could lead us to a retirement option.

Whittle Theorem: For the multi-armed bandit problem with a retirement option, the optimal policy is:

1. If $M \geq M^i(X^i)$ for all $i \in \{1, 2, \ldots, k\}$ retire.

2. Otherwise activate the arm i^* for which $M^{i^*}(X^{i^*}) = \max_i\{M^i(X^i)\}$

Proof. The value function for the MAB with retirement option for any state $\mathbf{X}_t \in \mathcal{X}$ is given by:

$$V^r(\mathbf{X}_t, M) = \max_i\{R(X_t^i) + \beta\mathbb{E}[V^r(\mathbf{X}_{t+1}, M)|\mathbf{X}_t], M\} \tag{26}$$

Let $\tau^i(X_t^i, M)$ denote the retirement time for the i^{th} bandit with a terminal reward M. We denote by T_r the retirement time for the entire multi-armed bandit process with the retirement option.

Then we have:

$$T_r(M) = \sum_{i=1}^{k} \tau^i(X_t^i, M) \tag{27}$$

Whittle [4] introduces the following expression:

$$\frac{\partial V^r(\mathbf{X}, M)}{\partial M} \overset{(a)}{=} \mathbb{E}[\beta^{T_r}] \overset{(b)}{=} \mathbb{E}[\beta^{\sum_{i=1}^{k} \tau^i(X^i,M)}] \overset{(c)}{=} \prod_{i=1}^{k} \mathbb{E}[\beta^{\tau^i(X^i,M)}] \overset{(d)}{=} \prod_{i=1}^{k} \frac{\partial V^r(X^i, M)}{\partial M} \tag{28}$$

(a) is given by lemma 2, (b) is given by expression (27), (c) is given by the fact that the random variables $\tau^i(X^i, M)$ are independent for any i and finally (d) is given by expression (27).

Besides, we have $V^r(\mathbf{X}, M) = M$ for $M \geq K$.

Then by integrating $\frac{\partial V^r(\mathbf{X},M)}{\partial M} = \prod_{i=1}^{k} \frac{\partial V^r(X^i,M)}{\partial M}$ for $M \in \mathbb{R}$ we have:

$$\hat{V}^r(\mathbf{X}, M) = K - \int_M^K \prod_{i=1}^{k} \frac{\partial V^r(X^i, m)}{\partial m} dm \tag{29}$$

To simplify our expressions, Whittle [4] defined :

$$Q^i(\mathbf{X}, M) = \prod_{j=1; j\neq i}^{k} \frac{\partial V^r(X^j, M)}{\partial M} \tag{30}$$

Lemma 1 implies that Q^i is non-negative and non-decreasing function of M and:

$Q^i : \mathbb{R} \to [0, 1]$ as function of M.

Then by substituting expression (30) in expression (29) and doing an integration by parts we have:

$$\hat{V}^r(\mathbf{X}, M) = V^r(X^i, M)Q^i(\mathbf{X}, M) + \int_M^K V^r(X^i, m)dQ^i(\mathbf{X}, m) \tag{31}$$

Finally we need to show that $\hat{V}^r$ defined in expression (29) satisfies the optimality of the value function of the MAB with retirement option, meaning that we need to show $\hat{V}^r = V^r$, this is done in 3 steps as shown by Whittle [4] and Frostig [2].

Step 1: At first, we need to show that $\hat{V}^r(\mathbf{X}, M) \geq M$
Since we have $V^r(X^i, m)$ as a monotone function in m, by expression (31) we have:

$$\hat{V}^r(\mathbf{X}, M) \geq \underbrace{V^r(X^i, M)Q^i(\mathbf{X}, M) + V^r(X^i, M)\int_M^K dQ^i(\mathbf{X}, m)}_{=V^r(X^i,M)Q^i(\mathbf{X},M)+V^r(X^i,M)[Q^i(\mathbf{X},K)-V^r(X^i,M)Q^i(\mathbf{X},M)]} \tag{32}$$

$$\hat{V}^r(\mathbf{X}, M) \geq V^r(X^i, M)\underbrace{Q^i(\mathbf{X}, K)}_{=1} \geq M \tag{33}$$

Step 2: We define:

$$\Delta^i = \hat{V}^r(\mathbf{X}, M) - (R(X^i) + \beta\mathbb{E}[\hat{V}^r(\mathbf{X}_{t+1}, M)|\mathbf{X}_t = \mathbf{X}]) \tag{34}$$

We know that $Q^i(\mathbf{X}, M) + \int_M^K dQ^i(\mathbf{X}, m) = Q^i(\mathbf{X}, M) + [\underbrace{Q^i(\mathbf{X}, K)}_{=1} - Q^i(\mathbf{X}, M)] = 1$

We substitute (34) in (31), we have:

$$\begin{aligned}\Delta^i = V^r(X^i, M)Q^i(\mathbf{X}, M) + \int_M^K V^r(X^i, m)dQ^i(\mathbf{X}, m) - R(x^i)(Q^i(\mathbf{X}, M) + \int_M^K dQ^i(\mathbf{X}, m))\\ - \beta\mathbb{E}[V^r(X^i, M)Q^i(\mathbf{X}, M) + \int_M^K V^r(X^i, m)dQ^i(\mathbf{X}, m)]\end{aligned} \tag{35}$$

Which gives us by rearranging the terms :

$$\begin{aligned}&\Delta^i = V^r(X^i, M)Q^i(x, M) - \beta\mathbb{E}[V^r(X^i, M)Q^i(\mathbf{X}, M)] - R(x^i)Q^i(\mathbf{X}, M)\\ &+ \int_M^K V^r(X^i, m)dQ^i(x, m) - R(X^i)\int_M^K dQ^i(\mathbf{X}, m)) - \beta\mathbb{E}[\int_M^K V^r(X^i, m)dQ^i(\mathbf{X}, m)]\end{aligned} \tag{36}$$

$$\begin{aligned}\Delta^i = Q^i(\mathbf{X}, M)\left[V^r(X^i, M) - \beta\mathbb{E}[V^r(X^i, M)] - R(X^i)\right]\\ + \int_M^K \left[V^r(X^i, m) - R(x^i) - \beta\mathbb{E}[V^r(X^i, m)]\right] dQ^i(\mathbf{X}, m)\end{aligned} \tag{37}$$

But since we have:

$$\begin{aligned}V^r(X^i, M) = \max\{R^i(X^i) + \beta\mathbb{E}[V(X^i_{t+1})|X^i_t = X^i], M\}\\ \Rightarrow V^r(X^i, M) \geq R^i(X^i) + \beta\mathbb{E}[V(X^i_{t+1})|X^i_t = X^i]\end{aligned} \tag{38}$$

Which gives us $\Delta^i \geq 0$.

Step 3: Finally Whittle in 1980 [4] showed as presented by Frostig in 2016 [2] the equality in expression (35) (same as (34)) holds the equality under Whittle's policy.

Considering $M \in \mathbb{R}$, and any $m \geq M^{i^*}(X^{i^*})$ giving us: $Q^i(\mathbf{X}, M) = 1$ and $Q^i(\mathbf{X}, m) = 0$, giving us the equality $\hat{V}^r(\mathbf{X}, M) = M$ for $M \geq M^{i^*}(X^{i^*})$.

Now for the case of $M \leq m \leq M^{i^*}(X^{i^*})$, we have:

$$V^r(X^i, M) = R^i(X^i) + \beta\mathbb{E}[V(X^i_{t+1})|X^i_t = X^i] \tag{39}$$

That gives us by substiting this expression in expression (37):

$$\Delta^i = \int_{M^i(X^i)}^{K} \left[V^r(X^i, m) - R(X^i) - \beta\mathbb{E}[V^r(X^i, m)]\right] dQ^i(\mathbf{X}, m) \tag{40}$$

And for the particular case of i^*:

$$\Delta^{i^*} = \int_{M^i(X^{i^*})}^{K} \left[V^r(X^{i^*}, m) - R(X^{i^*}) - \beta\mathbb{E}[V^r(X^{i^*}, m)]\right] dQ^{i^*}(\mathbf{X}, m) \tag{41}$$

Since we have $Q^{i^*}(\mathbf{X}, m) = 0$ for the case of $m \geq M^{i^*}(X^{i^*})$, that gives us $\Delta^{i^*} = 0$.

That concludes the proof as explained by Whittle [4] and Frostig [2], as it shows that $\hat{V}$ satisfies the optimality equations. We showed that for $M \geq M^{i^*}(X^{i^*})$ we have $V^r(\mathbf{X}, M) = M$ which means that the optimal action is to retire and get the retirement reward M giving us the statement 1. of the theorem. For $M \leq M^{i^*}(X^{i^*})$ we have $V^r(\mathbf{X}, M) = R(X^{i^*}) - \beta\mathbb{E}[V_{t+1}(\mathbf{X}_{t+1})|\mathbf{X}_t = \mathbf{X}]$ meaning the optimal choice is to pull the arm i^*, giving us the statement 2. of the theorem.

□

References

[1] Richard Bellman. *Dynamic Programming*. Dover Publications, 1957. ISBN: 9780486428093.

[2] Esther Frostig and Gideon Weiss. "Four proofs of Gittins' multiarmed bandit theorem". In: *Annals of Operations Research* 241.1-2 (2016), pp. 127–165. ISSN: 15729338. DOI: 10.1007/s10479-013-1523-0. URL: http://dx.doi.org/10.1007/s10479-013-1523-0.

[3] J. C. Gittins. "Bandit Processes and Dynamic Allocation Indices". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (1979), pp. 148–164. DOI: 10.1111/j.2517-6161.1979.tb01068.x.

[4] P. Whittle. "Multi-armed Bandits and the Gittins Index By". In: 42.2 (2012), pp. 143–149.