On Robust Arm-Acquiring Bandit Problems

Shiqing Yu

Faculty Mentor: Xiang Yu

July 20, 2014

Abstract

In the classical multi-armed bandit problem, at each stage, the player has to choose one from N given projects (arms) to generate a reward depending on the arm played and its current state. The state process of each arm is modeled by a Markov chain and the transition probability is priorly known. The goal of the player is to maximize the expected total reward. One variant of the problem, the so-called arm-acquiring bandit, studies the case where at each stage new projects may arrive. Another recent extension of the classical bandit problem incorporates the uncertainty of the transition probabilities. This robust control problem considers an adversary, "nature", who aims to minimize the player's expected total reward by choosing a different transition probability measure each time after the player makes a decision. In this paper, we consider the robust arm-acquiring bandit problem, the combination of the two extensions above, and show that there exists an optimal state-by-state retirement policy. The extension to the robust arm-acquiring tax problem under some condition is also introduced.

1 Introduction

1.1 Multi-armed Bandit Problem

Since the 1940's, the multi-armed bandit problem has been a popular topic of stochastic optimization problems. At each time step $t = 0, 1, \ldots,$ one has to choose one among the N available machines (projects, or arms) $i = 1, 2, \ldots, N$ to operate. If $i(t)$ is selected, one receives a known one-step reward $\alpha^t r^{i(t)}\left(X^{i(t)}(t)\right)$, depending on $i(t)$ and its current state $X^{i(t)}(t)$, where $0 < \alpha < 1$ is a discount factor. Then $X^{i(t)}(t)$ undergoes a transition to state $X^{i(t)}(t+1)$ according to a time-homogeneous Markov chain with a known transition probability. The states of other machines remain frozen, i.e. $X^j(t+1) = X^j(t)$ for all $j \neq i(t)$. Then the bandit problem asks one to choose a project to operate at each stage in order to maximize the expected total reward

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t r^{i(t)}\left(X^{i(t)}(t)\right)\right]. \tag{1.1}$$

An optimal policy was introduced in Gittins [3]: A *dynamic allocation index* (DAI), also known as *Gittins index*, is assigned to each project i at each stage t, and the machine with the largest current index is operated. Suppose we are at time step 0, then the Gittins index is defined as

$$\nu(i) = \sup_{\tau>0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha^t r^i\left(X^i(t)\right) \middle| X^i(0) = x\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha^t \middle| X^i(0) = x\right]}, \tag{1.2}$$

where the supremum is actually proven to be attained at the stopping time

$$\tau^*(i) = \min\left\{t : \nu\left(X^i(t)\right) < \nu(x)\right\}. \tag{1.3}$$

The Gittins index policy for the multi-armed bandit problem, proven to be optimal, states that one always pulls the arm with the largest Gittins index.

Whittle [10] later gave a more explicit proof of the optimality of Gittins policy by proposing a conjecture and observing the properties of the optimal value function and the indices as well as their relationship. He introduced the retirement option, in which at any stage t one can stop playing all arms and receive an immediate reward M (discounted by α^t). The **optimal Gittins index policy for this M-process** is that at time step t, after assigning an analogous Gittins index which now depends on M to each project i, one operates the project with the highest Gittins index as long as it exceeds M, or retires from the whole system if all indices are less than M.

The retirement option M is introduced only to facilitate the proof of the policy's optimality. In real-life applications this option is generally not available, and the problem without the retirement option can be deduced by setting $M \to -\infty$.

1.2 Arm-Acquiring Bandit Problem

In real-life applications we often have cases where new projects may arrive at each stage. Nash [7] first considered this variant of the problem, which later was called the *Arm-acquiring bandit problem* by Whittle [11].

In his paper [11], Whittle formulated the problem by labeling the "projects" by "state" variables, each indicating the type of the project and its stage. Partially following his notations, we define the set of all possible values x of the states as $\mathbb{X}$ (assumed to be finite), and $n_t(x)$ or simply $n(x)$ as the number of projects in state x at time step t, then the vector of $n_t(x)$, $x \in \mathbb{X}$, the *state vector*,

can be written as $n_t = \{n_t(x); x \in \mathbb{X}\}$, or simply n. It is now clear that $n \in \mathbb{N}^{|\mathbb{X}|}$ where $|\mathbb{X}|$ is the number of elements, or different states, in $\mathbb{X}$. Write $n > \mathbf{0}$ if $n(x) > 0$ for some $x \in \mathbb{X}$.

We shall also denote $e(x, x') = 1$ when $x = x'$ and 0 otherwise, then $e(x) = \{e(x, x'); x' \in \mathbb{X}\}$, $e^x(x') = e(x, x')e(x)$. In addition, it would be useful in our subsequent discussion if we let $\mathbb{U}(n_t) = \{x \in \mathbb{X} : n_t(x) > 0\}$ be the set of all available states in n_t.

To illustrate the notations above, consider the set of states $\mathbb{X} = \{1, 2, 3\}$. If at time t we have two projects in state 1, three in state 2, but none in state 3, then $n_t(1) = 2$, $n_t(2) = 3$, $n_t(3) = 0$, $n_t = (2, 3, 0) > \mathbf{0}$, $e(1) = e^1(1) = (1, 0, 0)$, $e^1(2) = (0, 0, 0)$, and $\mathbb{U}(n_t) = \{1, 2\}$.

The new-arriving states at time t are described by the vector $w_t \in \mathbb{N}^{|\mathbb{X}|}$, or simply w, with the set of its possible values denoted by $\mathbb{W}$. w is a non-negative random vector with time-homogeneous probability measure $q : (\mathbb{W}, \mathcal{W}) \to [0, 1]$, which is known and independent of the state processes and of the decision maker's choice.

The arm-acquiring bandit problem can now be stated as follows. At each time step $t = 0, 1, \ldots$, the decision maker chooses a project in state $u \in \mathbb{U}(n_t)$ and receives a reward $r(u) \leq B$ uniformly bounded for $u \in \mathbb{X}$. Then the state of the project makes a transition to state u' according to a time-homogeneous Markov chain with the known transition probability ζ^u. The state vector in the next time step becomes

$$n_{t+1} = n_t - e(u) + e(u') + w_t. \tag{1.4}$$

The *(optimal expected) value function* $V(n)$ then satisfies the Bellman equation

$$V(n) = \sup_{u \in \mathbb{U}(n)} \left\{ r(u) + \alpha \mathbb{E}_{\zeta^u} \mathbb{E}_q \left[V\left(n - e(u) + e(u') + w\right) \right] \right\}. \tag{1.5}$$

The decision maker's goal in the arm-acquiring bandit problem is to choose a series of projects by their current states in order to achieve the optimal expected total reward $V(n_0)$ where n_0 is the given initial state vector.

Whittle [11] introduced the *retirement option* just as in the simple multi-armed case, in which one can retire at any time with an immediate reward M. Then with the initial state vector n_0, the goal is to operate the projects and retire at some proper stage (possibly ∞) in order to achieve the optimal expected total reward $V_M(n_0)$ for the M-process that satisfies

$$V(n_0, M) = \max \left\{ M, \sup_{u \in \mathbb{U}(n_0)} \left\{ r(u) + \alpha \mathbb{E}_{\zeta^u} \mathbb{E}_q \left[V\left(n_0 - e(u) + e(u') + w\right) \right] \right\} \right\}. \tag{1.6}$$

Whittle [11] showed that if we assign the Gittins index $M(x) = \inf\left\{m : V(e(x), M) = M\right\}$ to each state x, then **the Gittins index policy for the arm-acquiring problem is optimal** in the sense that we retire any project whenever its state enters the set $\mathscr{G} = \{x : M(x) \leq M\}$, and we operate an available project in the state with the largest index if it exceeds M, or retire the whole system if none of the available states has an index greater than M.

2 Formulation

2.1 Simple Robust Bandit Problems

The *robust multi-armed bandit problems*, as introduced in Caro and Gupta [1] and in Kim and Lim [5], is a dynamic game between the decision maker and an adversary, called nature. In this case, after the player chooses an arm to play, nature chooses a transition probability for this arm to "hurt" the decision maker's expected total rewards, and the player receives a one-step reward depending on the arm operated by the player and the transition probability selected by nature.

We will explain in more detail in the next subsection, where we will extend the results to the combination of the arm-acquiring and the robust multi-armed bandit problems. The most important result obtained in both Caro and Gupta [1] and Kim and Lim [5] is that there exists an optimal *project-by-project retirement policy* in which we "retire (abandon)" a project whenever its index does not exceed the retirement reward M, and we continue to work on some project as long as there exists some project that has an index larger than M. It is notable that the policy does not assume the order of the projects played, and the **Gittins index policy** in which we always choose the arm with the largest index **is not necessarily optimal.**

2.2 Robust Multi-Armed Bandit Problems

We now combine the two extensions discussed above, and consider the *robust arm-acquiring bandit problem*. Since Gittins index policy is optimal for the arm-acquiring problem and there exists an optimal project-by-project retirement policy (not necessarily Gittins) for the robust bandit problem, it is natural to guess that we equivalently have a *state-by-state retirement (SSR)* policy for the robust multi-armed bandit problem, and we want to show that this is correct.

Like the simple robust bandit problem, the robust arm-acquiring one is a dynamic game between the decision maker and nature. Suppose at stage $t = 0, 1, \cdots$ with state vector n_t, the decision maker chooses a project in state $u \in \mathbb{U}(n_t)$ to perform, and nature subsequently chooses

a transition probability ν^u from $\mathscr{N}^u = \{\pi \ll \kappa^u : (\mathbb{X} \times \mathcal{X}) \to [0,1]\}$ for u, the set of all probability measures defined on $(\mathbb{X} \times \mathcal{X})$ that are absolutely continuous with respect to κ^u, the nominal transition probability for u the player has some confidence in. Then the decision maker receives a one-step reward (discussed later) paid by nature, and the state vector n_t becomes

$$n_{t+1} = n_t - e(u) + e(u') + w_{t+1}, \tag{2.1}$$

where the transition from u to u' follows the time-homogeneous Markov chain with transition probability ν^u, and w_{t+1} is the non-negative random vector with acquired arms under probability measure q on $(\mathbb{W}, \mathcal{W})$.

As before, a natural assumption of q is that it is known and independent of the arm process, i.e. q is independent of the state u chosen by the decision maker and the transition probability ν^u selected by nature. Therefore, we can simplify our notation by defining the product probability measure $p^u = \nu^u \times q$ on $(\mathbb{X} \times \mathbb{W}, \mathcal{X} \times \mathcal{W})$, which exists and is unique, and correspondingly we define $\mathscr{P}^u = \mathscr{N}^u \times \{q\}$ as the set of all selectable p^u for nature. Then we say the state vector n_t undergoes the transition to n_{t+1} according to p^u chosen by nature from $\mathscr{P}^u$.

The one-step reward can be written as

$$c(u, p^u) = r(u) + \theta^u I\left(p^u \| \, \rho^u\right), \tag{2.2}$$

where the model uncertainty is taken into account by $\theta^u I\left(p^u \| \, \rho^u\right)$. $\rho^u = \kappa^u \times q$ is the nominal transition probability for *u and w* that the decision maker has some confidence in, which is the product probability between the nominal transition probability κ^u *for state* u and the probability measure q *for arrivals* w. $\theta^u > 0$, the *penalty parameter*, indicates the confidence in the accuracy of the nominal transition probability ρ^u, with a higher value of θ^u showing the player is more confident. I is the *Kullback-Leibler divergence*, or *relative entropy* that measures the distance between p^u and ρ^u, introduced to penalize nature for choosing p^u far from ρ^u.

Definition 2.1. *For two measures μ and λ on $(E, \mathcal{E})$ such that $\mu \ll \lambda$, i.e. μ is absolutely continuous with respect to λ, the* Kullback-Leibler divergence*, also known as the* relative entropy*, is defined as*

$$I(\mu \| \lambda) = \int_E \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right) \mathrm{d}\mu, \tag{2.3}$$

where $\dfrac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ is the Radon-Nikodym derivative of μ with respect to λ. If μ is not absolutely continuous

with respect to λ*, the* Kullback-Leibler divergence *is defined to be* $I(\mu \| \lambda) = +\infty$.

Having fully described the one-step reward, we will discuss the properties of the optimal expected total reward (hereafter, the value function) in the next section.

3 Properties of the Value Function

Analogous to the one in Kim and Lim [5], we have the following theorem.

Theorem 3.1. *The value function* V *for the robust arm-acquiring bandit problem is the unique solution to the Bellman equation*

$$V(n) = \max\left\{M, \max_{u \in \mathbb{U}(n)} F^u(n, V)\right\}, \tag{3.1}$$

where

$$F^u(n, V) = r(u) + \inf_{p^u \ll \rho^u} \theta^u I(p^u \| \rho^u) + \mathbb{E}_{p^u}\left[\alpha V(n - e(u) + e(u') + w)\right] \tag{3.2}$$

In addition,

$$V = \lim_{N \to \infty} V_N, \tag{3.3}$$

where V_N *is the unique solution to the* N*-stage Bellman equation*

$$\begin{aligned} V_0(n) &= \max\{M, 0\} && (3.4)\\ V_k(n) &= \max\left\{M, \max_{u \in \mathbb{U}(n)} F^u(n, V_{k-1})\right\}, \quad k = 1, \ldots, N. && (3.5) \end{aligned}$$

See González-Trejo et al. [4], Theorem 4.2. for an analogous proof of Theorem 3.1. The function $F^u(n, V)$ in (3.2) has this form since

$$\begin{aligned} F^u(n, V) &= \inf_{p^u \in \mathscr{P}^u} r(u) + \theta^u I(p^u \| \rho^u) + \mathbb{E}_{p^u}\left[\alpha V(n - e(u) + e(u') + w)\right] \\ &= r(u) + \inf_{p^u \in \mathscr{P}^u} \theta^u I(p^u \| \rho^u) + \mathbb{E}_{p^u}\left[\alpha V(n - e(u) + e(u') + w)\right] \\ &= r(u) + \inf_{p^u \ll \rho^u} \theta^u I(p^u \| \rho^u) + \mathbb{E}_{p^u}\left[\alpha V(n - e(u) + e(u') + w)\right], \end{aligned} \tag{3.6}$$

where the last equality because $\mathscr{P}^u$ is the set of all measures absolutely continuous with respect to ρ^u.

We wish to choose p^u to minimize the function. Similar to Kim and Lim [5], we have the

following lemma.

Lemma 3.2. *Let η be a probability measure defined on $(E, \mathcal{E})$, and g: $E \to \mathbb{R}$ a bounded measurable function. If $\theta > 0$, then we have the following variational formula*

$$\inf_{\mu \ll \eta} \left\{ \int_E g \, \mathrm{d}\mu + \theta I(\mu \| \eta) \right\} = -\theta \log \int_E \exp\left(-\frac{g}{\theta}\right) \mathrm{d}\eta. \tag{3.7}$$

Furthermore, the infimum is uniquely attained at $\mu^* \ll \eta$ with

$$\frac{\mathrm{d}\mu^*}{\mathrm{d}\eta}(\omega) = \frac{\exp\left(-\frac{g(\omega)}{\theta}\right)}{\int_E \exp\left(-\frac{g}{\theta}\right) \mathrm{d}\eta}, \quad \omega \in E. \tag{3.8}$$

Proof. The proof can be found in Dupuis and Ellis [2], Proposition 1.4.2. □

Remark. *The proof in Dupuis and Ellis [2], Proposition 1.4.2. only has verification of the optimality of μ^*, so here we explain how to arrive at this "guess". Writing $f \triangleq \frac{\mathrm{d}\mu}{\mathrm{d}\eta}$ we have*

$$\begin{aligned} \int_E g \, \mathrm{d}\mu + \theta I(\mu \| \eta) &= \theta \int_E \frac{g}{\theta} + \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right) \mathrm{d}\mu \\ &= \theta \int_E \log\left(f \cdot \exp\left(\frac{g}{\theta}\right)\right) \mathrm{d}\mu \\ &= \theta \int_E f \cdot \log\left(f \cdot \exp\left(\frac{g}{\theta}\right)\right) \mathrm{d}\eta. \end{aligned} \tag{3.9}$$

We need to minimize (3.9) *subject to $\int_E f \, \mathrm{d}\eta = 1$, then the first-order condition formally implies that the partial derivatives of the integrand of*

$$\int_E f \cdot \log\left(f \cdot \exp\left(\frac{g}{\theta}\right)\right) \mathrm{d}\eta - \lambda \left(\int_E f \, \mathrm{d}\eta - 1 \right) \tag{3.10}$$

with respect to f and with respect to λ must be zero. Thus,

$$\log\left(f \cdot \exp\left(\frac{g}{\theta}\right)\right) + 1 - \lambda = 0, \tag{3.11}$$

$$\int_E f \, \mathrm{d}\eta = 1. \tag{3.12}$$

Solving the equation system (3.11) *and* (3.12) *we get the solution*

$$\frac{\mathrm{d}\mu^*}{\mathrm{d}\eta}(\omega) = f^*(\omega) = \frac{\exp\left(-\dfrac{g(\omega)}{\theta}\right)}{\displaystyle\int_E \exp\left(-\frac{g}{\theta}\right)\mathrm{d}\eta}, \quad \omega \in E. \tag{3.13}$$

Then we can verify the optimality by invoking Dupuis and Ellis [2], Proposition 1.4.2.

Substituting (3.7) into (3.2) we get

$$\begin{aligned} F^u(n,V) &= r(u) + \inf_{p^u \ll \rho^u} \theta^u I\left(p^u \| \rho^u\right) + \mathbb{E}_{p^u}\left[\alpha V(n - e(u) + e(u') + w)\right] \\ &= r(u) - \theta^u \log \int_{\mathbb{X}\times\mathbb{W}} \exp\left(-\frac{\alpha V(n - e(u) + e(u') + w)}{\theta^u}\right)\mathrm{d}\rho^u \\ &= r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\frac{\alpha V(n - e(u) + e(u') + w}{\theta^u}\right)\right]. \end{aligned} \tag{3.14}$$

Similar to Kim and Lim [5], if now we think of $V(n)$ as a function $V(n, M)$ in M instead of treating M as a constant, we have the following lemma.

Lemma 3.3. *The value function $V(n, M)$ is a non-decreasing function in M. Furthermore, $V(n, M)$ is constant for all $M \le -\dfrac{B}{1-\alpha}$, and $V(n, M) = M$ for all $M \ge \dfrac{B}{1-\alpha}$, where constant $B = \sup_{u\in\mathbb{U}(n)} |r(u)|$.*

The proof is given in Kim and Lim [5], Lemma 4.1.

Consider now the problem with retirement option where there is only one state u that can be worked on. If there is no u at one certain stage in the future, we can wait or equivalently play a project in another state until u arrives, and then continue to work on u. If u is transient so that the time before u arrives is ∞ with a positive probability, we just play projects in other states. We let n^u denote the state vector conditional on state u only, that is, $n^u = n(u)e(u)$ (see section 1.2 for the definitions). Then $V^u(n^u)$, the value function of that problem, satisfies

$$V^u(n^u) = \max\left\{M, r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\frac{\alpha V^u(n - e(u) + e(u') + w)^u}{\theta^u}\right)\right]\right\} \tag{3.15}$$

if there is at least one project in state u in n, i.e. $n^u > \mathbf{0}$ or $n(u) \ge 1$, and

$$V^u(n^u) \ge \max\{M, 0\} \tag{3.16}$$

otherwise since the retirement option is always available. Then the k-stage Bellman equations are

$$V_0^u(n^u) = \max\{M, 0\} \tag{3.17}$$

$$V_k^u(n^u) = \max\left\{M, r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\frac{\alpha V_{k-1}^u(n - e(u) + e(u') + w)^u}{\theta^u}\right)\right]\right\} \text{ if } n^u > \mathbf{0} \tag{3.18}$$

$$V_k^u(n^u) \geq M \text{ otherwise,} \tag{3.19}$$

for $k = 1, \ldots, N$, and $V^u(n^u)$ can be obatined as the limit $V^u = \lim_{N\to\infty} V_N^u$ since intuitively $V^u(0) = \lim_{N\to\infty} V_N^u(0)$.

Similarly, we can define $V^{u-}(n^{u-})$ and $V_k^{u-}(n^{u-})$ as the value function and k-stage value function associated with the problem where all projects in states *other than* u can be selected, with the retirement option. As in the case for $V^u(n^u)$, if there is no project in any state other than u, we can wait, or equivalently choose a project in u to play until an arm with another state arrives. Also, we let n^{u-} denote the state vector n without state u, i.e. $n^{u-}(x) = n(x)$ if $x \neq u$, but $n^{u-}(u) = 0$.

We now have the following lemma.

Lemma 3.4. *(1) If $n^u > \mathbf{0}$ (or equivalently $u \in \mathbb{U}(n)$), then for $\forall u' \in \mathbb{X}, w \in \mathbb{W}$ such that $(n - e(u) + e(u') + w)^u > \mathbf{0}$, we have $V_{k+1}^u(n^u) \geq V_k^u((n - e(u) + e(u') + w)^u)$ for any $k \in \mathbb{N}$.*
(2) If $n^x > 0$ for some $x \neq u$ (or equivalently $x \in \mathbb{U}(n)$ for some $x \neq u$), then for $\forall x' \in \mathbb{X}, w \in \mathbb{W}$ such that $(n - e(x) + e(x') + w)^{u-} > \mathbf{0}$, we have $V_{k+1}^{u-}(n^{u-}) \geq V_k^{u-}((n - e(x) + e(x') + w)^{u-})$ for any $k \in \mathbb{N}$.

Proof. We will prove by mathematical induction. The base cases $V_1^u(n^u) \geq \max\{M, 0\} = V_0^u(n - e(u) + e(u') + w)^u$ and $V_1^{u-}(n^{u-}) \geq \max\{M, 0\} = V_0^{u-}((n - e(x) + e(x') + w)^{u-})$ clearly satisfy the inequality. Suppose that the inequalities also hold for some $k \geq 1$. Then applying the induction hypothesis, it follows that for $\forall u' \in \mathbb{X}, w \in \mathbb{W}$,

$$\begin{aligned}
&V_k^u\left(\left(n - e(u) + e(u') + w\right)^u\right) \\
=&\max\left\{M, r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\frac{\alpha V_{k-1}^u((n - 2e(u) + e(u') + e(u'') + w + w')^u)}{\theta^u}\right)\right]\right\} \\
\leq&\max\left\{M, r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\frac{\alpha V_k^u((n - e(u) + e(u'') + w')^u)}{\theta^u}\right)\right]\right\} \\
=&V_{k+1}^u(n^u),
\end{aligned}$$

where the expectations are taken for u'' and w'.

Likewise, we have $V_{k+1}^{u-}\left(n^{u-}\right) > V_k^{u-}\left((n - e(x) + e(x') + w)^{u-}\right)$.

Thus, it follows by mathematical induction that the inequalities hold for all k. □

4 State-by-state Retirement Policy

Analogous to the *project-by-project retirement* policy in Kim and Lim [5], we now define a *state-by-state retirement (SSR)* policy.

Definition 4.1. *Suppose at an arbitrary stage t, the current state vector is n. A policy is of state-by-state retirement (SSR) type if there exist index functions $g^u : \mathbb{N}^{|\mathbb{X}|} \to \mathbb{R}$ for each $u \in \mathbb{U}(n)$, such that the policy*

(1) permanently retires state u if $g^u(n^u) \leq M$, and

(2) continues to work on some project , if $\max_{u \in \mathbb{U}(n)} g^u(n^u) > M$.

For $u \in \mathbb{U}(n)$, we define

$$g^u(n^u) = \begin{cases} \inf\{M : V^u(n^u, M) = M\} & \text{if } n^{u-} > \mathbf{0}, \\ r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha V(n - e(u) + e(u') + w)}{\theta^u}\right)\right] & \text{if } n^{u-} = \mathbf{0}. \end{cases} \tag{4.1}$$

Then we have the following theorem.

Theorem 4.1. *There exists an optimal SSR policy.*

Proof. To show the existence of an optimal state-by-state retirement policy, it suffices to prove the following results for any $n > 0$ and $u \in \mathbb{U}(n)$ (or equivalently $n^u > \mathbf{0}$):

(1) If $n^{u-} = \mathbf{0}$, we only have one state u in n, and thus we have to choose between operating a project in state u and retiring, which is equivalent to comparing M and

$$r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\frac{\alpha V(n - e(u) + e(u') + w)}{\theta^u}\right)\right]. \tag{4.2}$$

Thus the policy is optimal in this case.

(2) If $n^{u-} > \mathbf{0}$, we need to show that

$$V^{u-}\left(n^{u-}\right) \leq V(n) \leq V^{u-}\left(n^{u-}\right) + \left(V^u\left(n^u\right) - M\right). \tag{4.3}$$

Recall that in this case $n^u > \mathbf{0}$ and $n^{u-} > \mathbf{0}$. The first inequality clearly holds. The proof of the second inequality follows by mathematical induction. The base case

$$V_0(n) = V_0^{u-}\left(n^{u-}\right) = V_0^u\left(n^u\right) = \max\{M, 0\} \tag{4.4}$$

clearly satisfies the inequality. Assume now that for some k,

$$V_k(n) \leq V_k^{u-}\left(n^{u-}\right) + \left(V_k^u(n^u) - M\right). \tag{4.5}$$

We wish to show that the inequality holds for $k+1$. Applying the above induction hypothesis, it then follows that

$$\begin{aligned} & V_{k+1}(n) \\ &= \max\left\{M, \max_{x\in\mathbb{U}(n)} r(x) - \theta^x \log \mathbb{E}_{\rho^x}\left[\exp\left(-\frac{\alpha V_k(n - e(x) + e(x') + w)}{\theta^x}\right)\right]\right\} \qquad (4.6) \\ &\leq \max\left\{\begin{array}{l} M, \max\limits_{x\in\mathbb{U}(n)}\left\{ r(x) - \right. \\ \left. \theta^x \log \mathbb{E}_{\rho^x}\left[\exp\left(-\frac{\alpha\left(V_k^{u-}\left((n - e(x) + e(x') + w)^{u-}\right) + V_k^u\left((n - e(x) + e(x') + w)^u\right) - M\right)}{\theta^x}\right)\right]\right\} \end{array}\right\}. \qquad (4.7) \end{aligned}$$

Splitting the $\max\limits_{x\in\mathbb{U}(n)}$ into the max of state u and $\max\limits_{x\in\mathbb{U}(n), x\neq u}$, and taking constants out of expectations, the right hand side of the above inequality is equal to

$$\max\left\{\begin{array}{l} M, r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha\left(V_k^{u-}\left((n-e(u)+e(u')+w)^{u-}\right)+V_k^u\left((n-e(u)+e(u')+w)^u\right)-M\right)}{\theta^u}\right)\right], \\ \max\limits_{x\in\mathbb{U}(n),x\neq u} r(x) - \theta^x \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha\left(V_k^{u-}\left((n-e(x)+e(x')+w)^{u-}\right)+V_k^u\left((n-e(x)+e(x')+w)^u\right)-M\right)}{\theta^x}\right)\right] \end{array}\right\} \tag{4.8}$$

$$\leq \max\left\{\begin{array}{l} M, r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha\left(V_{k+1}^{u-}\left(n^{u-}\right)-M\right)+\alpha V_k^u\left((n-e(u)+e(u')+w)^u\right)}{\theta^u}\right)\right], \\ \max\limits_{x\in\mathbb{U}(n),x\neq u} r(x) - \theta^x \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha\left(V_{k+1}^{u}\left(n^{u}\right)-M\right)+\alpha V_k^{u-}\left((n-e(x)+e(x')+w)^{u-}\right)}{\theta^x}\right)\right] \end{array}\right\} \tag{4.9}$$

$$= \max\left\{\begin{array}{l} M, r(u) + \alpha\left(V_{k+1}^{u-}\left(n^{u-}\right)-M\right) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha V_k^u\left((n-e(u)+e(u')+w)^u\right)}{\theta^u}\right)\right], \\ \max\limits_{x\in\mathbb{U}(n),x\neq u} r(x) + \alpha\left(V_{k+1}^{u}\left(n^{u}\right)-M\right) - \theta^x \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha V_k^{u-}\left((n-e(x)+e(x')+w)^{u-}\right)}{\theta^x}\right)\right] \end{array}\right\} \tag{4.10}$$

$$\leq \max\left\{\begin{array}{l} \max\left\{M, r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha V_k^u\left((n-e(u)+e(u')+w)^u\right)}{\theta^u}\right)\right]\right\} + \alpha\left(V_{k+1}^{u-}\left(n^{u-}\right)-M\right), \\ \max\left\{M, \max\limits_{x\in\mathbb{U}(n),x\neq u}\left\{r(x) - \theta^x \log \mathbb{E}_{\rho^u}\left[\exp\left(-\dfrac{\alpha V_k^{u-}\left((n-e(x)+e(x')+w)^{u-}\right)}{\theta^x}\right)\right]\right\}\right\} + \alpha\left(V_{k+1}^{u}\left(n^{u}\right)-M\right) \end{array}\right\}, \tag{4.11}$$

where the first inequality follows from Lemma 3.4 and the last inequality since $V_k^u, V_k^{u-} \geq M$. Finally, recognizing that

$$V_{k+1}\left(n^u\right) = \max\left\{M, r(u) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(-\frac{\alpha V_k^u\left((n-e(u)+e(u')+w)^u\right)}{\theta^u}\right)\right]\right\} \tag{4.12}$$

$$V_{k+1}^{u-}\left(n^{u-}\right) = \max\left\{M, \max_{x\in\mathbb{U}(n),x\neq u}\left\{r(x) - \theta^x \log \mathbb{E}_{\rho^u}\left[\exp\left(-\frac{\alpha V_k^{u-}\left((n-e(x)+e(x')+w)^{u-}\right)}{\theta^x}\right)\right]\right\}\right\}, \tag{4.13}$$

the right-hand side of the inequality equals to

$$\begin{aligned} & \max\left\{V_{k+1}^{u}\left(n^{u}\right)+\alpha\left(V_{k+1}^{u-}\left(n^{u-}\right)-M\right), V_{k+1}^{u-}\left(n^{u-}\right)+\alpha\left(V_{k+1}^{u}\left(n^{u}\right)-M\right)\right\} \\ \leq & \max\left\{V_{k+1}^{u}\left(n^{u}\right)+\left(V_{k+1}^{u-}\left(n^{u-}\right)-M\right), V_{k+1}^{u-}\left(n^{u-}\right)+\left(V_{k+1}^{u}\left(n^{u}\right)-M\right)\right\} \\ = & V_{k+1}^{u}\left(n^{u}\right)+V_{k+1}^{u-}\left(n^{u-}\right)-M, \end{aligned} \tag{4.14}$$

which ends the proof. □

5 Extension And Conjecture

5.1 Tax Problem

Now consider a new variant of the robust arm-acquiring bandit problem, the *robust arm-acquiring tax problem*. The only difference between these two problems is that after one chooses a project in state $u \in \mathbb{U}(n)$, all other projects are charged taxes depending on their states, i.e. the one-step nominal tax charged is

$$t(u,n)=\sum_{x\in\mathbb{U}(n)} t(x)n(x)-t(u), \tag{5.1}$$

where $t(x)$ is the "increase" of the player's money after tax is charged for *every* project in state x, which is negative if the tax is paid by the player to nature rather than paid to the player as a "subsidy".

We may equivalently write $t(u,n)=\sum\limits_{x\in\mathbb{X}} t(x)n(x)-t(u)$ since $n(x)=0$ if $x\notin\mathbb{U}(n)$.

We can introduce the retirement option with $M(n)$ depending on n, which in this case is that one can pay $-M(n)$ at one time and quit the game, without the need to pay any taxes in the future. This is because naturally we need to charge different one-time fees for different n's.

Denote $T(n)$ as the total expected "profit" for the player, which decreases by the amount charged each time a tax is incurred if the tax is "charged" from the player, and increases if it is "paid" to the player as a "subsidy". Then the Bellman equation is

$$T(n)=\max\left\{M(n), \max_{u\in\mathbb{U}(n)}\left(\sum_{x\in\mathbb{X}} t(x)n(x)-t(u)\right)+\inf_{p^u\ll\rho^u}\theta^u I\left(p^u\|\,\rho^u\right)+\mathbb{E}_{\rho^u}\left[T(n-e(u)+e(u')+w)\right]\right\}, \tag{5.2}$$

since if nature chooses p^u further from ρ^u, or if the player has more confidence in ρ^u, then $\theta^u I\left(p^u\|\,\rho^u\right)$ and thus $T(n)$ is more positive and the player is better-off.

This Bellman equation has exactly the same form as (3.1) with $r(u)$ replaced by $t(u,n)$ and M by $M(n)$. So similar to (3.14), by invoking Lemma 3.2 we have

$$T(n) = \max\left\{M(n), \max_{u\in\mathbb{U}(n)} t(u,n) - \theta^u \log \mathbb{E}_{\rho^u}\left[\exp\left(\frac{\alpha T(n-e(u)+e(u')+w)}{\theta^u}\right)\right]\right\}. \tag{5.3}$$

This similarity characterizes the relationship between the robust arm-acquiring bandit and tax problems. But since now $M(n)$ and $t(u,n) = \sum_{x\in\mathbb{X}} t(x)n(x) - t(u)$ depend on n, Lemmas 3.3 and 3.4 are no longer valid under the tax case; we do not have a "state-by-state" policy for the tax problem either, since even if we wish to retire projects in one certain state, which in the bandit problem means never playing those arms again, we are still charged taxes on these projects at each future stage. Therefore, here we only discuss the following special case.

Theorem 5.1. *If $w \geq \sum_{x\in\mathbb{X}} e(x)$, for the problem without the retirement option, the Gittins policy is optimal.*

Proof. Since $w \geq \sum_{x\in\mathbb{X}} e(x)$, i.e. at each stage at least one project in each state would arrive, we can always operate a project in any state we want. Or equivalently, if we choose a project in some state u at time t, whose state changes to u' at time $t+1$, and we want to play an arm with state u', we can always choose one newly arrived, and treat the project we played at time t as "frozen" so that its state remains u' forever. Therefore, the trade-off between choosing a project in u_1 and another in u_2 at time t is equivalent to taking into account both the difference between saving $-t(u_1)$ and $-t(u_2)$, as well as the difference between the taxes charged from stage $t+1$ to stage ∞ on those two projects with frozen states after the transition. We only consider the transition probability for u since the w process is now irrelevant to our choice, so we reuse the $\nu^u \in \mathscr{N}^u$ notation for the transition probability for u chosen by nature, and denote the nominal transition probability for u by κ^u. Thus, to make the decision between u_1 and u_2, all we need to compare is $h(u_1)$ and $h(u_2)$, where

$$\begin{aligned} h(u) &= -t(u) + \inf_{\nu^u\in\mathscr{N}^u}\left\{\mathbb{E}_{\nu^u}\left[\left(\alpha+\alpha^2+\cdots\right)t(u')\right] + \theta^u I\left(\nu^u \| \kappa^u\right)\right\} \\ &= -t(u) + \inf_{\nu^u\ll\kappa^u}\left\{\mathbb{E}_{\nu^u}\left[\frac{\alpha}{1-\alpha}t(u')\right] + \theta^u I\left(\nu^u \| \kappa^u\right)\right\} \\ &= -t(u) - \theta^u \mathbb{E}_{\kappa^u}\left[\exp\left(-\frac{t(u')}{\theta^u}\frac{\alpha}{1-\alpha}\right)\right], \end{aligned} \tag{5.4}$$

where the last equality because of the variational formula.

Since $h(u)$ is the same at each stage if κ^u and $t(x)$ for all $x \in \mathbb{X}$ are fixed, it is optimal to always play a project in the same state u with the largest index at every stage. □

Remark. *In practical applications the retirement option M in the bandit problems is generally not available. Recall that it was only introduced to facilitate the proof of the policies' optimality, since the case without the retirement option can be deduced by setting $M \to -\infty$. Here we directly consider the normal problem without the retirement option because it is easier to handle.*

5.2 Conjecture

In Kim and Lim [5] there only exists an optimal project-by-project retirement policy and the Gittins policy is not necessarily optimal because the projects are dependent on each other because of the robust structure. In our case, however, when we consider the states instead of the projects themselves, the states are independent to some extent. Therefore, considering the *state-based Gittins index policy* in which the player plays the state with the highest index at each stage, we have the following conjecture.

Conjecture 5.2. *The* state-based Gittins index policy *for the robust arm-acquiring bandit problem is optimal.*

Further research on this problem should be conducted to prove or disprove this conjecture by a counterexample.

References

[1] F. Caro and A.D. Gupta. Robust control of the multi-armed bandit problem. 2013. *Working Paper.*

[2] P. Dupuis and R.S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* Wiley, New York, 1997.

[3] J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[4] J.I. González-Trejo, O. Hernández-Lerma, and L.F. Hoyos-Reyes. Minimax control of discrete-time stochastic systems. *SIAM Journal on Control and Optimization*, 41:1626–1659, 2003.

[5] M.J. Kim and A.E.B. Lim. Robust multi-armed bandit problems. *Management Science*, 2013. *Working Paper*.

[6] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[7] P. Nash. *Optimal Allocation of Resources Between Research Projects*. PhD thesis, University of Cambridge., 1973.

[8] P.P. Varaiya, J.C. Walrand, and C. Buyukkoc. Extensions of the multiarmed bandit problem: The discounted case. *IEEE Transactions on Automatic Control*, 30(5):426–439, 1985.

[9] R. Wheeden, R.L. Wheeden, and A. Zygmund. *Measure and Integral: An Introduction to Real Analysis*. Chapman & Hall/CRC Pure and Applied Mathematics. Taylor & Francis, 1977.

[10] P. Whittle. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):143–149, 1980.

[11] P. Whittle. Arm-acquiring bandits. *The Annals of Probability*, 9(2):284–292, 1981.