Four proofs of Gittins' multiarmed bandit theorem

Esther Frostig * Gideon Weiss *†
Department of Statistics
The University of Haifa
Mount Carmel 31905, Israel.
{gweiss}{frostig}@stat.haifa.ac.il

October 2, 2013

Abstract

We study four proofs that the Gittins index priority rule is optimal for alternative bandit processes. These include Gittins' original exchange argument, Weber's prevailing charge argument, Whittle's Lagrangian dual approach, and Bertsimas and Niño-Mora's proof based on the achievable region approach and generalized conservation laws. We extend the achievable region proof to infinite countable state spaces, by using infinite dimensional linear programming theory.

keywords: dynamic programming; bandit problems; Gittins index; linear programming
ams classification: 90B36; 62L15; 90C40; 49L20; 90C05; 90C27; 90C57

1 Introduction

Consider a system consisting of a family of N alternative bandit processes, where at time t the state of the system is given by the vector $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_N(t))$ of the states $Z_n(t)$ of the bandit processes $n = 1, \ldots, N$. We assume that these bandits move on countable state spaces E_n, so $Z_n(t) \in E_n, n = 1, \ldots, N$.

At any point in time, $t = 0, 1, 2, \ldots$, we need to take one of N possible actions, namely choose to activate one of the bandit processes, which will then yield a reward and undergo a Markovian state transition, while all the other bandit processes are passive — they yield no reward, and their states remain frozen. More precisely, if we choose at time t action $n(t) = n$, then bandit n in state $Z_n(t) = i$ will be activated. This action will yield a reward $R_n(i)$, where R_n is the reward function for bandit n, and bandit n will undergo a transition, from state i to state j according to $p_n(i, j) = \mathbb{P}(Z_n(t + 1) = j | Z_n(t) = i)$. For all other bandits, $m \neq n(t)$, there will be no change in state, so $Z_m(t+1) = Z_m(t)$, and no reward, so the reward for period t will be given by $\tilde{R}(t) = R_{n(t)}(Z_{n(t)}(t)) = R_n(i)$.

*Research supported in part by Network of Excellence Euro-NGI
†Research supported in part by Israel Science Foundation Grants 249/02, 454/05, 711/09 and 286/13.

We will assume that $|R_n(i)| \leq C$ uniformly for all states and bandits. The objective is to choose a policy π for activating the bandits so as to maximize total discounted reward

$$V_\pi(\mathbf{i}) = \mathbb{E}_\pi \left\{ \sum_{t=0}^{\infty} \alpha^t \tilde{R}(t) | \mathbf{Z}(0) = \mathbf{i} \right\}, \tag{1.1}$$

where $\mathbf{Z}, \mathbf{i}$ denote the state vector, and $0 < \alpha < 1$ is the discount factor.

This problem, introduced by Bellman [4] as the *multiarmed bandit problem*, is clearly a dynamic programming problem, with a countable state space, a finite action space, bounded rewards and discounted infinite horizon objective. As such, by the theory of dynamic programming [32] it has an optimal solution given by a stationary policy, which can be calculated using various general schemes. However, such a direct approach to the problem is impractical due to the high dimensionality of the state space.

What makes the problem tractable is Gittins' discovery that the problem is solved by a priority policy — one needs only to calculate a priority index for each of the bandits (independent of all the other bandits), and activate the bandit with the highest index. Formally

Theorem 1.1 (Gittins, 1976) *There exist functions, $G_n(Z_n(t)), n = 1, \ldots, N$ such that a policy π^* which will in state $\mathbf{Z}(t)$ activate a bandit process (arm) $n(t) \in \arg\max_{1 \leq m \leq N} G_m(Z_m(t))$ is optimal. The function $G_n(\cdot)$ is calculated from the dynamics of process n alone.*

This is a deep result, and as such it has many different aspects and implications, and can be proven in several very different ways. Proofs of this result have been emerging over several decades, and have motivated much further research. In this paper we discuss 4 different proofs of Gittins' result, introduce some extensions, and study some implications. Our purpose is twofold: We feel that some of the original proofs were never compacted or simplified, and as a result the topic of Gittins index acquired an unjustified reputation of being difficult. Hence we believe that a unified presentation of the four proofs together is useful. More important, when several proofs exist for the same theorem it is difficult to avoid mixtures of ideas or even circular arguments in some of the papers which develop these proofs. We have tried here, using the advantage of hindsight, to present the proofs in a 'pure' form. As far as possible we use complete self contained arguments for each of the proofs, and highlight the differences between them. We believe that our choice to focus on four proofs does cover the main different ideas.

The proofs follow a common structure: They start with the study of a single bandit process and the solution of a one dimensional dynamic programming problem. Next, some properties of the single arm solution are derived. These are then used to study the behavior of the controlled multi-armed system, and to prove that Gittins' policy is optimal. In Section 2 we study the single bandit process, and derive the properties needed for the four proofs. We also derive Klimov's algorithm for the computation of the index when the state space is finite.

The four proofs are presented next in Section 3. They include: Gittins' pairwise interchange argument (Section 3.1), Weber's fair charge argument (Section 3.2), Whittle's dual Lagrangian approach (Section 3.3), and the achievable region proof of Bertsimas and Niño-Mora, (Section 3.4).

The achievable region proof of Bertsimas and Niño-Mora is restricted to the case of a finite state space, while the other proofs are valid for infinite countable state space. In Section 4 we present for the first time an extension of the achievable region proof to the case of infinite countable state space. We obtain a new analog to Klimov's algorithm, for the case of infinite countable state space.

Section 5 contains a discussion of the proofs, highlighting the ideas behind them and comparing them, as well as very brief survey of some additional proofs, as well as a brief discussion of further developments.

For ease of notation and without loss of generality we shall assume that all the bandits move on the same state space E, with a single reward function R and a single transition matrix $\mathcal{P}$. It may be the case that in the original problem the bandits are indeed *i.i.d.*. Otherwise one can artificially introduce $E = \bigcup_{n=1}^{N} E_n$, in which case the Markov chain given by E and $\mathcal{P}$ will be reducible, with noncommunicating classes of states for each of the bandits.

bibliographic note: The Gittins index idea was put forward by Gittins as early as 1972 [12, 13, 14]. It was also indicated in several other papers of the time, notably in Klimov's paper [24], and also in Sevcik [35], Harrison [18], Tcha and Pliska [39], and Meilijson and Weiss [25]. It is pointed out by Sonin [38] that as early as in the 60's a simplified bandit problem has been solved simultaneously by several authors (see Mitten [27]). Many researchers contributed to the further development of the theory, and our discussion here does not attempt to cover all of this work. We are also not including in our discussion the many papers dealing with bandit problems which preceded the Gittins index results. A second edition of Gittins' book, with Weber and Glazebrook [15], contains much new material. For a recent survey, focused mainly on computational methods, see also Chakravorty and Mahajan [6].

Remark Throughout the paper the form of address 'we' is meant conversationally to suggest us and the reader and is no indication of original contribution. Our original contribuitons in this paper, beyond surveying existing results, include a new direct proof that the Gittins index is achieved by a stopping time, an analog to Klimov's algorithm for infinite countable state space, and the extension of the achievable region proof to the infinite countable state space case.

2 Preliminary: Studying the single bandit

We start this section with the study of the Gittins Index (Section 2.1), from first principles and definition of the Gittins order (Section 2.2). Next we present 3 closely related formulations of single arm dynamic programming problems which can be used to calculate the index, and study their solutions (Section 2.3). We continue with definition and some properties of fair charge and prevailing charge, the measurable processes of the index values and their lower envelope, which accompany the bandit process (Section 2.4). We conclude with the derivation of Klimov's algorithm for computing the Gittins index, and its infinite countable state space generalization (Section 2.5). Throughout this section, we consider a single bandit process $Z(t)$.

2.1 The Gittins index

Gittins defined his Dynamic Allocation Index, now known as the Gittins Index, as follows:

$$\nu(i) = \sup_{\sigma>0} \nu(i,\sigma) = \sup_{\sigma>0} \frac{\mathbb{E}\left\{\sum_{t=0}^{\sigma-1} \alpha^t R(Z(t)) | Z(0)=i\right\}}{\mathbb{E}\left\{\sum_{t=0}^{\sigma-1} \alpha^t | Z(0)=i\right\}}. \tag{2.1}$$

Here $\nu(i,\sigma)$ is the expected discounted reward per expected unit of discounted time, when the arm is operated from initial state i, for a duration σ, and σ is a $Z(t)$ positive stopping time. The value $\nu(i)$ is the supremum of $\nu(i,\sigma)$ over all positive stopping times. By the boundedness of the rewards, $\nu(i)$ is well defined for all i, and bounded.

Starting from state i we define the stopping time ($\leq \infty$):

$$\tau(i) = \min\{t : \nu(Z(t)) < \nu(i)\} \tag{2.2}$$

An important property of the Gittins index is that the supremum (2.1) is achieved, and in fact it is achieved by $\tau(i)$. This property is well known since [13]. It can be derived from a dynamic programming formulation of the bandit problem, as we shall see in Section 2.3, by invoking the theory of dynamic programming, and in particular the existence of stationary Markovian optimal policies. It is important however to note that this fundamental connection between (2.1) and (2.2) can be proven directly, without reference to dynamic programming theory.

Theorem 2.1 *The supremum of (2.1) is achieved by (2.2). It is also achieved by any stopping time σ which satisfies:*

$$\sigma \leq \tau \quad \textit{and} \quad \nu(Z(\sigma)) \leq \nu(i) \tag{2.3}$$

We present the 'first principles' proof of Theorem 2.1 in Appendix A.

2.2 The Gittins priority order

The value of the index, $\nu(i)$ introduces a semi-order (each pair of states can be compared, but E is divided into equivalence classes of states with equal indexes). We will consider an arbitrary fixed complete order of the states which is consistent with ν: For any two distinct states $i, j \in E$ either $i \prec j$ or $j \prec i$, and $\nu(i) < \nu(j)$ implies $i \prec j$. We say $i \preceq j$ if $i \prec j$ or $i = j$. We will prove the Gittins optimality theorem for this arbitraty fixed priority policy which gives priority to j over i if $i \prec j$. It is well known from general dynamic programming theory that randomization of optimal policies is also optimal, so this is no restriction of generality.

We use values of ν and the prioirty order relation $\prec$ to define subsets of E ordered by inclusion. For a real number v, let

$$\begin{aligned} S(v) &= \{k \in E,\, \nu(k) \leq v\} \\ S^{-}(v) &= \{k \in E,\, \nu(k) < v\}. \end{aligned} \tag{2.4}$$

For $i \in E$, let

$$\begin{aligned} S_i &= \{k \in E,\, k \preceq i\} \subseteq S(\nu(i)) \\ S_i^- &= \{k \in E,\, k \prec i\} \supseteq S^-(\nu(i)). \end{aligned} \tag{2.5}$$

For $Z(0) = i$ we define the stopping time:

$$\varsigma(i) = \min\{t : t > 0,\, Z(t) \prec i\} = \min\{t : t > 0,\, Z(t) \in S_i^-\} \tag{2.6}$$

Let $i \in E$ be a state and $S \subseteq E$ a subset of states. Consider an arm which is initially in state i, it is played once, and is then played until it reaches a state in S. Let:

$$T_i^S = \min\{t : t > 0, Z(t) \in S | Z(0) = i\}. \tag{2.7}$$

We call T_i^S an i to S first passage time. Let:

$$A_i^S = \mathbb{E}\left\{\sum_{t=0}^{T_i^S - 1} \alpha^t | Z(0) = i\right\}. \tag{2.8}$$

We call A_i^S the i to S expected discounted first passage time. Let:

$$W_i^S = \mathbb{E}\left\{\sum_{t=0}^{T_i^S - 1} R(Z(t))\alpha^t | Z(0) = i\right\}. \tag{2.9}$$

We call W_i^S the i to S expected discounted reward.
Remark.

1. The definitions of $\tau(i)$, $\varsigma(i)$, and T_i^S imply that $\tau(i) = T_i^{S^-(\nu(i))}$, and $\varsigma(i) = T_i^{S_i^-}$.

2. By their definitions,
$$\begin{aligned} &S(\nu(i)) \supseteq S_i \supseteq S_i^- \supseteq S^-(\nu(i)), \\ &T_i^{S(\nu(i))} \leq T_i^{S_i} \leq T_i^{S_i^-} \leq T_i^{S^-(\nu(i))}, \\ &A_i^{S(\nu(i))} \leq A_i^{S_i} \leq A_i^{S_i^-} \leq A_i^{S^-(\nu(i))}. \end{aligned}$$

3. Theorem 2.1 and (2.1) imply that:

$$\nu(i) = \frac{W_i^{S(\nu(i))}}{A_i^{S(\nu(i))}} = \frac{W_i^{S_i}}{A_i^{S_i}} = \frac{W_i^{S_i^-}}{A_i^{S_i^-}} = \frac{W_i^{S^-(\nu(i))}}{A_i^{S^-(\nu(i))}} \tag{2.10}$$

2.3 Dynamic programming for a single arm

Three dynamic programming formulations

We now present 3 equivalent dynamic programming problems for the single bandit process, which can be used to calculate the Gittins index. These three formulations provide some insights, and were used to motivate the various proofs.

Playing against a standard arm (Gittins): Assume that you have a single arm (bandit process) Z, and an alternative standard arm which never changes state and yields a reward γ whenever it is played. Consider this as a multiarmed bandit problem; Gittins referred to this as the $1\frac{1}{2}$ bandits problem. Because the standard arm is fixed, the state of the system is described by the state of the bandit $Z = i$. Denote by $V_s(i, \gamma)$ the optimal reward from state i and standard arm γ (the subscript s stands for "standard arm"). The optimality equations for this problem are:

$$V_s(i, \gamma) = \max \left\{ R(i) + \alpha \sum p(i, j) V_s(j, \gamma), \gamma + \alpha V_s(i, \gamma) \right\} \tag{2.11}$$

The fixed charge problem (Weber): Assume that you have a single arm Z, and at any time t you need to choose whether to play the arm for a fixed charge γ and collect the reward from this play, or not to play at time t but wait for $t+1$. Denote by $V_f(i, \gamma)$ the optimal reward from state i and fixed charge γ (the subscript f stands for "fixed charge"). The optimality equations for this problem are:

$$V_f(i, \gamma) = \max \left\{ R(i) - \gamma + \alpha \sum p(i, j) V_f(j, \gamma), \alpha V_f(i, \gamma) \right\} \tag{2.12}$$

The retirement option problem (Whittle): Assume that you can play the arm for as long as you want, then retire for ever and receive a terminal reward M. Denote by $V_r(i, M)$ the optimal reward from state i and retirement reward M (the subscript r stands for "retirement"). The optimality equations for this problem are:

$$V_r(i, M) = \max \left\{ R(i) + \alpha \sum p(i, j) V_r(j, M), M \right\} \tag{2.13}$$

Once it is optimal to play the standard arm (in the standard arm problem) or not pay the fixed charge (in the fixed charge problem) at some time t, then it is optimal to continue not to play the arm Z forever. This is so because if the arm in state $Z(t)$ is not played then its state is frozen, $Z(t + 1) = Z(t)$. Hence all three problems are actually optimal stopping problems, in which one needs to decide when to stop playing the arm. If we take $M = \frac{\gamma}{1-\alpha}$, then the three problems have the same optimal policy. In the context of retirement we can think of the standard reward γ also as a pension payment per unit time, which is equivalent to the retirement reward of M. The fixed charge problem pays at every time t a reward smaller by γ than that of the standard arm problem. The optimal value functions of the three problems satisfy:

$$V_s(i, \gamma) = V_f(i, \gamma) + \frac{\gamma}{1 - \alpha} = V_r(i, \frac{\gamma}{1 - \alpha})$$

.

Solution of the single arm dynamic programs

By the theory of dynamic programming, the three problems have an optimal solution given by a stationary policy, and we have:

• *Optimal policies:* Let

Strict continuation set	$C_M = \{i : V_r(i,M) > M\}$
Strict stopping set	$S_M = \{i : M > R(i) + \alpha \sum p(i,j) V_r(j,M)\}$
Indifferent states	$\partial_M = \{i : M = R(i) + \alpha \sum p(i,j) V_r(j,M)\}$

then any policy which continues to activate the arm while in C_M, acts arbitrarily in ∂_M and stops in S_M is optimal.

• *Stopping time* $\tau(i,M)$ which is the first passage time from i into S_M.

• *Optimal value function:*

$$V_r(i,M) = \mathbb{E}\left\{ \sum_{t=0}^{\tau(i,M)-1} \alpha^t R(Z(t)) + \alpha^{\tau(i,M)} M | Z(0) = i \right\}, \tag{2.14}$$

where we can also write alternatively $\alpha^{\tau(i,M)} M = \sum_{t=\tau(i,M)}^{\infty} \alpha^t \gamma$.

• Clearly, ∂_M is non-empty only for a discrete set of values M.

• As M increases we have: C_M decreases, S_M increases, and $\tau(i,M)$ decreases. The changes in these sets occur exactly for values of M where ∂_M is non-empty.

• In particular, $C_M = E$ and $\tau(i,M) = \infty$ for $M < \frac{-C}{1-\alpha}$, and $C_M = \emptyset$, $\tau(i,M) = 0$ for $M > \frac{C}{1-\alpha}$ (recall that the objective is bounded, $|R(i)| \leq C$, $i \in E$).

Calculation of the Gittins index from single arm dynamic programs

Define

$$M(i) = \sup\{M : i \in C_M\} = \inf\{M : V_r(i,M) = M\} \tag{2.15}$$

$$\gamma(i) = (1-\alpha) M(i) \tag{2.16}$$

$M(i)$ is the value of M such that $i \in \partial_M$, and in the optimal solution of the single arm dynamic programming problems one is indefferent between playing in state i or stopping.

Proposition 2.2 *The quantity $\gamma(i)$ equals the Gittins index,*

$$\nu(i) = \gamma(i) \tag{2.17}$$

$$\tau(i) = \tau(i, M(i)-) \tag{2.18}$$

The proof is given in Appendix A

Note: by (2.14) it is clear that $\tau(i, M(i)-)$ achieves $\nu(i, \tau(i, M(i)-) = \gamma(i) = \nu(i)$, and therefore we have "for free" a proof of the fact that the supremum in (2.1) is achieved. However this proof is based on deep results from the theory of dynamic programming, as opposed to the direct proof of Theorem 2.1 given in Appendix A.

The optimal value function of the retirement option

We now consider $V_r(i, M)$, the optimal value to the single arm retirement option problem (2.13) for initial state i and terminal reward M. We examine this as a function of M. We already noted that it is bounded. We further state:

Proposition 2.3 *(a) $V_r(i, M) = V_r(i)$, the constant expected reward with no retirement option, for $M \leq -\frac{C}{1-\alpha}$.*

(b) $V_r(i, M) = M$, the reward for immediate retirement, for $M \geq \frac{C}{1-\alpha}$.

(c) $V_r(i, M)$ is nondecreasing and convex in M.

Proof. The only nontrivial part is the convexity. For any fixed policy π let τ_π denote the (possibly infinite) random retirement time. Then:

$$V_r^\pi(i, M) = \mathbb{E}_\pi(\text{reward up to } \tau_\pi + \alpha^{\tau_\pi} M) \tag{2.19}$$

which is linear in M. Hence $V_r(i, M)$, as the supremum of these linear functions over all π is convex. ■

As a convex function $V_r(i, M)$ is differentiable at all but a countable number of points, at which it has subgradients. A glance at (2.14) or (2.19) suggests the form of the derivative.

Proposition 2.4 *Let τ_M denote the optimal retirement time for terminal reward M. Then $\mathbb{E}(\alpha^{\tau_M})$ is a subgradient of $V_r(i, M)$ (the line through $(M, V_r(i, M))$ with slope $\mathbb{E}(\alpha^{\tau_M})$ is below the curve $V_r(i, \cdot)$), and at every M for which $\frac{\partial V_r(i,M)}{\partial M}$ exists,*

$$\frac{\partial V_r(i, M)}{\partial M} = \mathbb{E}(\alpha^{\tau_M} \mid Z(0) = i) \tag{2.20}$$

Proof. Fix M and i, and let $\bar{\pi}$ be an optimal policy for M; let $\epsilon > 0$. Utilizing the policy $\bar{\pi}$ for $M + \epsilon$,

$$V_r^{\bar{\pi}}(i, M + \epsilon) = \mathbb{E}_{\bar{\pi}}(\text{reward up to } \tau_M) + \mathbb{E}(\alpha^{\tau_M})(M + \epsilon)$$

Hence,

$$V_r(i, M + \epsilon) \geq V_r^{\bar{\pi}}(i, M + \epsilon) = V_r(i, M) + \epsilon \mathbb{E}(\alpha^{\tau_M})$$

Similarly,

$$V_r(i, M - \epsilon) \geq V_r(i, M) - \epsilon \mathbb{E}(\alpha^{\tau_M})$$

Hence $\mathbb{E}(\alpha^{\tau_M})$ is a subgradient of $V_r(i, M)$. By definition it is equal to the derivative wherever such exists. ■

2.4 The fair charge and the prevailing charge of a bandit

Consider the fixed charge Problem (2.12). If the arm is in state i and the value of the charge is equal to $\gamma(i)$, then it is optimal to either play or stop, and in either case the expected optimal

revenue (rewards minus charges) is $V_r(i, \gamma(i)) = 0$. Hence we call $\gamma(i)$ the *fair charge* for the arm in state i. Define the *fair charge stochastic process*

$$g(t) = \gamma(Z(t))$$

Note that $g(t)$ is observable (can be calculated for each t from values of the process $Z(\cdot)$ up to time t; more formally, it is measurable with respect to $Z(s), s \leq t$).

As we said, in state i for the fair charge $\gamma(i)$ it is optimal to either play or stop. However, if one does play one needs to continue playing optimally. Let $Z(0) = i$, and the fixed charge be $\gamma(i)$. If one plays the arm at time 0, one needs to continue to play it as long as $g(t) > \gamma(i)$. Consider to the contrary that one starts playing and then stops at a stopping time $\sigma > 0$ such that $P\{g(\sigma) > \gamma(i)\} > 0$. Then the expected revenue up to time σ is < 0. This is clear from the solution of the fixed charge optimality equations (2.12). It is also exactly what is shown in step 1 of the proof of Theorem 2.1, in Appendix A.

In particular it is optimal to play for the duration $\tau(i)$. At the time $\tau(i)$ one has $g(\tau(i)) < \gamma(i) = g(0)$, i.e. the fair charge is less than the fixed charge, and it is optimal to stop. The expected revenue from this play is 0.

Consider now lowering the fixed charge, at the time $\tau(i)$, to the new fair charge value $g(\tau(i))$. Then it will again be optimal to either stop or play, and if one plays one would need to continue to the next appropriate stopping time.

Define the *prevailing charge stochastic process*

$$\underline{g}(t) = \min_{s \leq t} g(s),$$

note that $\underline{g}(t)$ is also observable.

Note also that the fair charge and the prevailing charge processes remain well defined and observable if the bandit process is not played continuously, but is played intermittently, with some idle periods and later continuations.

Assume now that instead of a fixed charge, the charge levied for playing at time t equals the prevailing charge $\underline{g}(t)$. It is then optimal to continue to play forever, and the expected total discounted revenue (rewards minus charges) is 0. On the other hand, at time 0, at the time $\tau(i)$, and in fact at all successive times at which $\underline{g}(t) = g(t)$ it is also optimal to stop. In contrast, it is strictly not optimal to stop when the fair charge exceeds the prevailing charge.

We summarize these results in the following proposition, corollary, and technical result:

Proposition 2.5 *If arm $Z(t)$ is played up to a stopping time σ then:*

$$\mathbb{E}(\sum_{t=0}^{\sigma-1} \alpha^t R(t) | Z(0) = i) \leq \mathbb{E}(\sum_{t=0}^{\sigma-1} \alpha^t \underline{g}(t) | Z(0) = i)$$

Equality holds if and only if $\underline{g}(\sigma) = g(\sigma)$ a.s.

Suppose now that the bandit process is played at a sequence of nonnegative integer times $t(s), s = 1, 2, \ldots$, where $t(s)$ is strictly increasing in s for all s or it increases up to $t(\bar{s})$ and is

infinite for $s > \bar{s}$. Let $Z(t)$ be the state of the bandit at time t. The state is frozen at times $t \notin \{t(s)\}_{s=1}^{\infty}$, and it is changing at times $t \in \{t(s)\}_{s=1}^{\infty}$. Typically $\{t(s)\}_{s=1}^{\infty}$ will be random, but we assume that $t(s)$ is measurable with respect to $Z(t), t \leq t(s)$.

Corollary 2.6

$$\mathbb{E}(\sum_{s=0}^{\infty} \alpha^{t(s)} R(Z(t(s)))|Z(0) = i) \leq \mathbb{E}(\sum_{t=0}^{\infty} \alpha^{t(s)} \underline{g}(t(s))|Z(0) = i)$$

with equality if and only if $\underline{g}(t) = g(t)$ *for all* $t \notin \{t(s)\}_{s=1}^{\infty}$ *a.s.*

We will require a technical point here: Corollary 2.6 remains valid if $t(s)$ are measurable with respect to the cartesian product of the sigma field generated by $Z(t), t \leq t(s)$ with a sigma field Σ which is independent of it.

2.5 The index sample path and Klimov's algorithm

If an arm is played at successive times $t = 0, 1, \ldots$, we now have three stochastic processes: At time t the state of the bandit is $Z(t)$. We then have at each time the process $g(t)$, the fair charge process, with values $g(t) = \gamma(Z(t)) = \nu(Z(t))$ (the last equality is by Proposition 2.2). We refer to it as the index process (we will however continue to call it the fair charge process in Section 3.2, for the second proof). Finally, we have the stochastic process of prevailing charge $\underline{g}(t)$, minimum of all fair charges up to time t, which we now refer to as the lower envelope of the index process. Figure 1 illustrates the index process and its lower envelope.

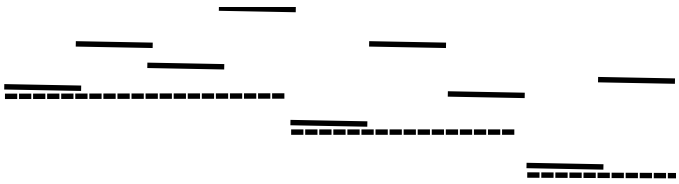

Figure 1: Fair and prevailing charges, the index process and its lower envelope.

We consider the times at which the index (the prevailing charge) and its lower envelope (the fair charge) are equal, and choose out of those a subsequence of times and of states. The subsequence is defined in terms of $Z(t), g(t), \underline{g}(t), t = 0, 1, \ldots$, and the successive intervals $\varsigma(j)$ defined in (2.6). Starting from some arbitrary state $Z(0)$ let:

$$\begin{array}{lll} \mathcal{T}_0 = 0, & & k_0 = Z(0), \\ \mathcal{T}_\ell = \mathcal{T}_{\ell-1} + \varsigma(k_{\ell-1}), & & k_\ell = Z(\mathcal{T}_\ell), \qquad \ell > 0. \end{array} \tag{2.21}$$

For a given sample path, the sequence of times $\mathcal{T}_0 < \mathcal{T}_1 < \ldots < \mathcal{T}_\ell < \ldots$ is the sequence of times at which the priority of the current state is lower than all the states previously encountered. The sequence of states k_ℓ are states of decreasing priority, $k_0 \succ k_1 \succ \cdots \succ k_\ell \succ \cdots$ (hence they are all different), and each k_ℓ is visited for the first time at $\mathcal{T}_\ell$. Note that the sequence of states need not include all the states in E, nor all the states which $Z(t)$ visits.

If a state is visited for the first time only after some other state with lower priority has been visited, then it will not be included as one of the k_ℓ.

For some sample paths we may have a smallest ℓ for which $\varsigma(k_\ell) = \infty$, which means that following $\mathcal{T}_\ell$ the sample path never reaches a state $j \prec k_\ell$. In such a case the sequence of states is finite and terminates at k_ℓ, and all stopping times after $\mathcal{T}_\ell$ are ∞. This will in particular always happen if E is finite.

A countably infinite set is well ordered by an order relation if there is a one to one order preserving mapping of the set into the integers. We note that when E is countably infinite then the Gittins order is not necessarily a well ordering. The importance of $\mathcal{T}_\ell$, k_ℓ, $\ell = 0, 1, 2, \ldots$ is precisely the property that they are an increasing sequence of times and a subset of states which is well ordered by the Gittins priority order. This enables us to use mathematical induction, and perform summations of times and rewards over sample paths of $Z(t)$.

We now use the sequence of times and states $\mathcal{T}_\ell, k_\ell, \ell = 0, 1, \ldots$ to calculate discounted expected first passage times and discounted expected rewards.

The expected discounted first passage times can be expressed as follows:

$$A_j^{S_i} = \mathbb{E}\{\sum_{t=0}^{T_j^{S_i}-1} \alpha^t | Z(0) = j\} = 1 + \mathbb{E}\{\sum_{\ell=1}^{\infty} I(k_\ell \succ i)\alpha^{\mathcal{T}_\ell} \sum_{t=0}^{\varsigma(k_\ell)-1} \alpha^t \mid Z(0) = j\} \tag{2.22}$$

where $I(\cdot)$ is the indicator function. Similarly:

$$A_j^{S_i^-} = 1 + \mathbb{E}\{\sum_{\ell=1}^{\infty} I(k_\ell \succeq i)\alpha^{\mathcal{T}_\ell} \sum_{t=0}^{\varsigma(k_\ell)-1} \alpha^t | Z(0) = j\} \tag{2.23}$$

and by subtracting (2.22) from (2.23) we obtain:

$$A_j^{S_i^-} - A_j^{S_i} = \mathbb{E}\{\sum_{\ell=1}^{\infty} I(k_\ell = i)\alpha^{\mathcal{T}_\ell} \sum_{t=0}^{\varsigma(i)-1} \alpha^t | Z(0) = j\} \tag{2.24}$$

Remark. Note that for every sample path the summation over ℓ in (2.24) is either empty, or it consists of a single term. If a state with priority $\prec i$ is visited by the sample path before the first visit to state i, then $k_\ell \neq i$ for every ℓ and the summation is empty. If on the first visit to i it has the lowest priority of all the states previously visited then $i = k_\ell$ for exactly one ℓ. In that case $T_j^{S_i} = \mathcal{T}_\ell$, and $T_j^{S_i^-} = T_j^{S_i} + \varsigma(i) = \mathcal{T}_{\ell+1}$.

Next we consider $Z(0) = j$ and the first passage stopping time $T_j^{S_j}$, and calculate the total expected revenue up to that time. This gives us an expression for the value of the index $\nu(j)$.

Proposition 2.7 *The Gittins index $\nu(j), j \in E$ as defined in (2.1) satisfies:*

$$\nu(j) = \frac{R(j) + \sum_{\{i:i\succ j\}} (A_j^{S_i^-} - A_j^{S_i})\nu(i)}{A_j^{S_j}} \tag{2.25}$$

Proof. Equations (2.1), (2.10) and (2.24) imply that:

$$\begin{aligned}
\nu(j)A_j^{S_j} &= R(j) + \mathbb{E}\{\sum_{\ell=1}^{\infty} \sum_{\{h:h\succ j\}} I(k_\ell = h)\alpha^{\mathcal{T}_\ell} \sum_{t=0}^{\varsigma(h)-1} R(Z(\mathcal{T}_\ell + t))\alpha^t \;\Big|\; Z(0) = j\} \\
&= R(j) + \mathbb{E}\{\sum_{\ell=1}^{\infty} \sum_{\{h:h\succ j\}} I(k_\ell = h)\alpha^{\mathcal{T}_\ell} \mathbb{E}[\sum_{t=0}^{\varsigma(h)-1} R(Z(\mathcal{T}_\ell + t))\alpha^t \mid \mathcal{T}_\ell, Z(\mathcal{T}_\ell) = h] \;\Big|\; Z(0) = j\} \\
&= R(j) + \mathbb{E}\{\sum_{\ell=1}^{\infty} \sum_{\{h:h\succ j\}} I(k_\ell = h)\alpha^{\mathcal{T}_\ell} \nu(h) \mathbb{E}[\sum_{t=0}^{\varsigma(h)-1} \alpha^t \mid \mathcal{T}_\ell, Z(\mathcal{T}_\ell) = h] \;\Big|\; Z(0) = j\} \\
&= R(j) + \mathbb{E}\{\sum_{\ell=1}^{\infty} \sum_{\{h:h\succ j\}} I(k_\ell = h)\alpha^{\mathcal{T}_\ell} \nu(h) \sum_{t=0}^{\varsigma(h)-1} \alpha^t \;\Big|\; Z(0) = j\} \\
&= R(j) + \sum_{\{h:h\succ j\}} \nu(h) \mathbb{E}\{\sum_{\ell=1}^{\infty} I(k_\ell = h)\alpha^{\mathcal{T}_\ell} \sum_{t=0}^{\varsigma(h)-1} \alpha^t \;\Big|\; Z(0) = j\} \\
&= R(j) + \sum_{\{h:h\succ j\}} (A_j^{S_h^-} - A_j^{S_h})\nu(h) \qquad (2.26)
\end{aligned}$$

The conditional expectation on $\mathcal{T}_\ell, Z(\mathcal{T}_\ell) = h$ allows us to take out $I(k_\ell = h)\alpha^{\mathcal{T}_\ell}$.

Changing the order of summations and expectation is justified since the sums are absolutely convergent and uniformly bounded by $\pm\frac{C}{1-\alpha}$ for all sample paths. ■

From Proposition 2.7 we obtain the next theorem, on which the calculation of the index and the fourth proof depend.

Theorem 2.8 *The Gittins index $\nu(j), j \in E$ satisfies:*

$$\nu(j) = \sup_{k \in S_j} \frac{R(k) + \sum_{\{i:i\succ j\}} (A_k^{S_i^-} - A_k^{S_i})\nu(i)}{A_k^{S_j}} \qquad (2.27)$$

Proof. For every $k \in S_j$:

$$\begin{aligned}
\nu(j) &\geq \frac{R(k) + \sum_{\{i:i\succ k\}} (A_k^{S_i^-} - A_k^{S_i})\nu(i)}{A_k^{S_k}} \\
&\geq \frac{R(k) + \sum_{\{i:i\succ j\}} (A_k^{S_i^-} - A_k^{S_i})\nu(i)}{A_k^{S_j}}
\end{aligned}$$

where the first inequality simply states $\nu(j) \geq \nu(k)$, and the second inquality simply states that $\nu(k) \geq \frac{W_k^{S_j}}{A_k^{S_j}}$. Equality holds since $j \in S_j$. ■

If the state space E is finite, then (2.27) can be used to calculate the Gittins order and index. This is essentialy Klimov's algorithm [24] for calculation of the index. We give two variations of the calculation, the second is from Bertsimas and Niño-Mora [5].

Klimov's Algorithm Version 1

Input $R(i)$, oracle to calculate A_i^S for $i \in S \subseteq E$.

output Gittins order $\varphi(1) \succ \varphi(2) \cdots \succ \varphi(|E|)$, sets $S_{\varphi(1)} \supset S_{\varphi(2)} \supset \ldots \supset S_{\varphi(|E|)}$, values of the index $\nu(\varphi(1)) \geq \nu(\varphi(2)) \geq \cdots \nu(\varphi(|E|))$.

Step 1 Calculate $\max_{i \in E} R(i)$.
Pick $\varphi(1) \in \arg\max$.
Set $\nu(\varphi(1)) =$ the maximal value.
Set: $S_{\varphi(1)} = E$ and $S^-_{\varphi(1)} = S_{\varphi(1)} \backslash \varphi(1)$.

Step k: $(k = 2, \ldots, |E|)$
Calculate: $\max_{i \in S^-_{\varphi(k-1)}} \frac{R(i) + \sum_{j=1}^{k-1} (A_i^{S^-_{\varphi(j)}} - A_i^{S_{\varphi(j)}}) \nu(\varphi(j))}{A_i^{S^-_{\varphi(k-1)}}}$.
Pick $\varphi(k) \in \arg\max$.
Set $\nu(\varphi(k)) =$ the maximal value.
Set: $S_{\varphi(k)} = S^-_{\varphi(k-1)}$ and $S^-_{\varphi(k)} = S_{\varphi(k)} \backslash \varphi(k)$.

Klimov's Algorithm Version 2

Input $R(i)$, oracle to calculate A_i^S for $i \in S \subseteq E$.

output Gittins order $\varphi(1) \succ \varphi(2) \cdots \succ \varphi(|E|)$, sets $S_{\varphi(1)} \supset S_{\varphi(2)} \supset \ldots \supset S_{\varphi(|E|)}$, values of the index $\nu(\varphi(1)) \geq \nu(\varphi(2)) \geq \cdots \nu(\varphi(|E|))$, values of y^S, $S \subseteq E$.

Step 1 Calculate $\max_{i \in E} R(i)$.
Pick $\varphi(1) \in \arg\max$.
Set: $S_{\varphi(1)} = E$.
Set $y^{S_{\varphi(1)}} =$ the maximal value.
Set $\nu(\varphi(1)) = y^{S_{\varphi(1)}}$.
Set: $S^-_{\varphi(1)} = S_{\varphi(1)} \backslash \varphi(1)$.

Step k: $(k = 2, \ldots, |E|)$
Calculate: $\max_{i \in S^-_{\varphi(k-1)}} \frac{R(i) - \sum_{j=1}^{k-1} A_i^{S_{\varphi(j)}} y^{S_{\varphi(j)}}}{A_i^{S^-_{\varphi(k-1)}}}$.
Pick $\varphi(k) \in \arg\max$.
Set: $S_{\varphi(k)} = S^-_{\varphi(k-1)}$.
Set $y^{S_{\varphi(k)}} =$ the maximal value
Set $\nu(\varphi(k)) = \nu(\varphi(k-1)) + y^{S_{\varphi(k)}}$.
Set: $S^-_{\varphi(k)} = S_{\varphi(k)} \backslash \varphi(k)$.

Final step Set $y^S = 0$ for all $S \subset E$, $S \neq S_{\varphi(1)}, \ldots, S_{\varphi(|E|)}$

Note that $y^{S_{\varphi(k)}} = \nu(\varphi(k)) - \nu(\varphi(k-1)) \leq 0$, $k = 2, \ldots, |E|$.

Proposition 2.9 *The two versions of Klimov's algorithm are equivalent*

The proof is in the Appendix. It involves use of Euler summation formula, and is straightforward. Changing the order of summation in the case of infinite countable state space is much more delicate. It is taken up in Section 4.

Algorithms for the calculation of the Gittins index are also discussed in [31] [38] [42] [22].

3 The proofs

3.1 First Proof: Interchange Argument

This proof follows Gittins [12, 13, 14]. We take a fixed priority ordering $\succ$ as defined in Section 2.2. The priority policy will at any time t activate an arm $n^*(t)$ whose state $Z_{n^*(t)}(t) = j^*$ where $j^* \succeq Z_n(t)$, $n = 1, \ldots, N$.

Let π^* denote the priority policy, let n be an arbitrary fixed bandit, and let $\pi^{(0)}$ be the policy which starts at time 0 by activating bandit n and proceeds from time 1 onwards according to the stationary policy π^*. To prove the optimality of π^* it suffices to show that $V_{\pi^*}(\mathbf{i}) \geq V_{\pi^{(0)}}(\mathbf{i})$ for every starting state $\mathbf{i}$. To show this we will define a sequence of additional policies, $\pi^{(s)}, s = 1, 2, \ldots$, such that

$$V_{\pi^{(s)}}(\mathbf{i}) \longrightarrow V_{\pi^*}(\mathbf{i}) \tag{3.1}$$

$$V_{\pi^{(s)}}(\mathbf{i}) \geq V_{\pi^{(s-1)}}(\mathbf{i}) \tag{3.2}$$

We define $\pi^{(s)}$ inductively. For initial state $\mathbf{i}$ let n^*, j^*, ν^*, ς^* be the bandit with the highest priority, the state, the index, and the stopping time $\varsigma^* = \varsigma(j^*)$ as defined in (2.6). Then $\pi^{(s)}$ will activate n^* for duration ς^*, and will then proceed from time ς^* and state $\mathbf{j} = \mathbf{Z}(\varsigma^*)$ as $\pi^{(s-1)}$ would from time 0 and initial state $\mathbf{j}$.

By their construction, $\pi^{(s)}$, $s \geq 1$ and π^* agree for the initial ς^*, $\varsigma^* \geq 1$. Furthermore, they continue to agree after ς^* for as long as $\pi^{(s-1)}$ and π^* agree, from the state reached at ς^*. Hence inductively $\pi^{(s)}$ agrees with π^* for at least the first s time units, hence $\pi^{(s)} \to \pi^*$, and the convergence in (3.1) is proved.

Also, for $s > 1$, $\pi^{(s)}$ and $\pi^{(s-1)}$ agree for the initial ς^* and so

$$V_{\pi^{(s)}}(\mathbf{i}) - V_{\pi^{(s-1)}}(\mathbf{i}) = \mathbb{E}\left\{\alpha^{\varsigma^*}\mathbb{E}\left\{V_{\pi^{(s-1)}}(\mathbf{Z}(\varsigma^*)) - V_{\pi^{(s-2)}}(\mathbf{Z}(\varsigma^*))|\mathbf{Z}(\varsigma^*)\right\}\right\} \tag{3.3}$$

and so to prove (3.2) by induction, and to complete the proof, it remains to show that $V_{\pi^{(1)}}(\mathbf{i}) \geq V_{\pi^{(0)}}(\mathbf{i})$ which is done by the following pairwise interchange argument:

If $n = n^*$ there is nothing to prove since then $\pi^{(1)} = \pi^{(0)} = \pi^*$. Assume then that $n \neq n^*$ for the initial state $\mathbf{i}$. Define the stopping time σ of the bandit process $Z_n(t)$ as the earliest time $t \geq 1$ at which $Z_n(t) \prec j^*$. One sees immediately that $\pi^{(0)}$ will start by activating n for duration σ, since following activation of n at time 0, n remains the bandit with highest priority until $\sigma - 1$. At time σ the highest priority will be the state j^* of bandit n^*, and so $\pi^{(0)}$, which continues according to π^*, will activate n^* for a period ς^*, up to time $\sigma + \varsigma^* - 1$. At time $\sigma + \varsigma^*$ the state will consist of $Z_{n^*}(\varsigma^*)$ for bandit n^*, of $Z_n(\sigma)$ for bandit n, and of $Z_m(0)$ for all other bandits, $m \neq n, n^*$. $\pi^{(0)}$ will proceed according to π^* from then onwards.

Policy $\pi^{(1)}$ will start by activating n^* for a time ς^* then at time ς^* it will activate n, and thereafter it will proceed according to π^*. One sees immediately that $\pi^{(1)}$ will activate n for at least a duration σ from the time ς^* at which it starts to activate n. This is because after n is activated at ς^*, it will remain in a state with priority $\succeq j^*$ for the duration σ, while the state of bandit n^*, following its activation for duration ς^*, is now of priority $\prec j^*$, and all other bandits retain their time 0 states, with priority $\preceq j^*$.

To summarize, $\pi^{(0)}$ activates n for duration σ followed by n^* for duration ς^*, followed by π^*; $\pi^{(1)}$ activates n^* for duration ς^* followed by n for duration σ followed by π^*. The state reached at time $\varsigma^* + \sigma$ by both policies is the same. Note that given n and n^*, the processes $Z_n(t)$ and $Z_{n^*}(t)$ are independent and so the stopping times ς^* and σ are independent. The difference in the expected rewards is:

$$\begin{aligned}
&\mathbb{E}\left\{\sum_{t=0}^{\sigma-1}\alpha^t R(Z_n(t)) + \alpha^\sigma \sum_{t=0}^{\varsigma^*-1}\alpha^t R(Z_{n^*}(t)) + \alpha^{\sigma+\varsigma^*}\sum_{t=0}^{\infty}\alpha^t \tilde{R}(\sigma+\varsigma^*+t) \mid \mathbf{Z}(0)=\mathbf{i}\right\} \\
&-\mathbb{E}\left\{\sum_{t=0}^{\varsigma^*-1}\alpha^t R(Z_{n^*}(t)) + \alpha^{\varsigma^*}\sum_{t=0}^{\sigma-1}\alpha^t R(Z_n(t)) + \alpha^{\sigma+\varsigma^*}\sum_{t=0}^{\infty}\alpha^t \tilde{R}(\sigma+\varsigma^*+t) \mid \mathbf{Z}(0)=\mathbf{i}\right\} \\
&= \mathbb{E}(1-\alpha^{\varsigma^*})\mathbb{E}\left\{\sum_{t=0}^{\sigma-1}\alpha^t R(Z_n(t))\right\} - \mathbb{E}(1-\alpha^\sigma)\mathbb{E}\left\{\sum_{t=0}^{\varsigma^*-1}\alpha^t R(Z_{n^*}(t))\right\} \\
&= \frac{1}{1-\alpha}\mathbb{E}(1-\alpha^{\varsigma^*})\mathbb{E}(1-\alpha^\sigma)\left(\nu(i_n,\sigma)-\nu(i_{n^*})\right) \\
&\leq \frac{1}{1-\alpha}\mathbb{E}(1-\alpha^{\varsigma^*})\mathbb{E}(1-\alpha^\sigma)\left(\nu(i_n)-\nu(i_{n^*})\right) \leq 0. \qquad (3.4)
\end{aligned}$$

This completes the proof of the theorem.

3.2 Second Proof: Interleaving of Prevailing Charges

This proof follows Weber [43]. Similar ideas were also used by Mandelbaum [26] and by Varaiya et al. [42, 19]. We now consider N bandit processes, with initial state $\mathbf{Z}(0) = \mathbf{i}$. We let $t^{(n)}(s), s = 1, 2\ldots$ indicate the times at which bandit n is played, with $t^{(n)}(s)$ strictly increasing in s, or $t^{(n)}(s) = \infty, s > \bar{s}$ if the bandit is only played a finite $\bar{s}$ number of times. Furthermore, $\left\{t^{(n)}(s)\right\}_{s=1}^{\infty}$ are disjoint sets whose union includes all of $\{1, 2\ldots\}$, for any sample path of any policy. We assume $t^{(n)}(s)$ is measurable with respect to $\mathbf{Z}(t), t \leq t^{(n)}(s)$, as it should be for any policy π. We let $g_n(t), \underline{g_n}(t)$ denote the fair and the prevailing charges of bandit n.

By Corollary 2.6, the technical note following it, and the independence of the arms we have:

$$\begin{aligned}
&\text{Expected total discounted reward} = \mathbb{E}\left\{\sum_{n=1}^{N}\sum_{s=1}^{\infty}\alpha^{t^{(n)}(s)}R(Z_n(t^{(n)}(s)))|\mathbf{Z}(0)=\mathbf{i}\right\} \\
&\leq \mathbb{E}\left\{\sum_{n=1}^{N}\sum_{s=1}^{\infty}\alpha^{t^{(n)}(s)}\underline{g_n}(t^{(n)}(s))|\mathbf{Z}(0)=\mathbf{i}\right\} = \mathbb{E}\left\{\sum_{t=0}^{\infty}\alpha^t \underline{\underline{g}}(t)|\mathbf{Z}(0)=\mathbf{i}\right\} \qquad (3.5)
\end{aligned}$$

where we define

$$\underline{\underline{g}}(t) = \underline{g_n}(t) \text{ if } t \in \left\{t^{(n)}(s)\right\}_{s=1}^{\infty}$$

Define now for each sample path

$$\underline{\underline{g}}^*(t) = \text{ The pathwise nonincreasing rearrangement of } \underline{\underline{g}}(t)$$

Note that while both $\underline{\underline{g}}(t)$ and $\underline{\underline{g}}^*(t)$ depend on the sample path of the bandits, the latter does not depend on the policy but only on the sample paths of the individual bandits.

By the Hardy Litlewood Polya inequality:

$$\sum_{t=1}^{\infty} \alpha^t \underline{\underline{g}}(t) \leq \sum_{t=1}^{\infty} \alpha^t \underline{\underline{g}}^*(t) \tag{3.6}$$

with equality holding if and only if $\underline{\underline{g}}(t)$ is nonincreasing.

The proof is now completed by noting the following two points:
(i) Under the Gittins index policy $\underline{\underline{g}}(t)$ is nonincreasing, so (3.6) holds as a pathwise equality.
(ii) Under the Gittins index policy an arm is never left unplayed while its fair charge is greater than the prevailing charge, hence the inequality in (3.5) holds as equality.

3.3 Third Proof: Retirement Option

Following Whittle [47] we consider the multiarmed bandit problem with retirement option. We have N arms in initial state $\mathbf{Z}(0) = \mathbf{i}$ and a retirement reward M. Using the definition (2.15) for arm n we let:

$$M_n(i_n) = \inf\{M : V_n(i_n, M) = M\}$$

Theorem 3.1 (Whittle) *For the multiarmed problem with retirement option the optimal policy is:*

(a) If $M \geq M_n(i_n)$ for all $n = 1, \ldots, N$, retire.

(b) Otherwise activate n^* for which $M_{n^*}(i_{n^*}) = \max_{n=1,\ldots,N}\{M_n(i_n)\}$.

Proof. The optimality equations for the multiarmed bandit problem with retirement option are:

$$V(\mathbf{i}, M) = \max_{n=1,\ldots,N} \left\{ M, R(i_n) + \alpha \sum_{j \in E} p(i_n, j) V(i_1, \ldots, j, \ldots, i_N, M) \right\} \tag{3.7}$$

If (a) and (b) are followed one can speculate on the form of $V(\mathbf{i}, M)$. Let $\tau_n(i_n, M)$ denote the retirement time (could be infinite) for the single bandit n with terminal reward M. Denote by $T(M)$ the retirement time for the entire multiarmed bandit system. Then (a) and (b) imply

$$T(M) = \sum_{n=1}^{N} \tau_n(i_n, M) \tag{3.8}$$

We now speculate that

$$\frac{\partial}{\partial M} V(\mathbf{i}, M) = \mathbb{E}(\alpha^{T(M)}) = \mathbb{E}(\alpha^{\sum_{n=1}^{N} \tau_n(i_n, M)}) = \prod_{n=1}^{N} \mathbb{E}(\alpha^{\tau_n(i_n, M)}) = \prod_{n=1}^{N} \frac{\partial}{\partial M} V_n(i_n, M). \tag{3.9}$$

Here the second equality is (3.8), the third is true because the random variables $\tau_n(i_n, M)$ are independent, and the fourth is true by (2.20). The first equality would hold if Whittle's policy of following (a) and (b) is optimal, since if this polilcy is optimal we can use the argument of Proposition 2.4 to prove the analogous result to (2.20).

We also have that $V(\mathbf{i}, M) = M$ for $M \geq \frac{C}{1-\alpha}$. Integrating (3.9) we get the following conjectured form for the optimal value function:

$$\hat{V}(\mathbf{i}, M) = \frac{C}{1-\alpha} - \int_M^{\frac{C}{1-\alpha}} \prod_{n=1}^{N} \frac{\partial}{\partial m} V_n(i_n, m)\, dm \tag{3.10}$$

For each n define

$$Q_n(\mathbf{i}, M) = \prod_{n' \neq n} \frac{\partial}{\partial M} V_{n'}(i_{n'}, M) \tag{3.11}$$

By Proposition 2.3, Q_n is nonnegative nondecreasing, ranging from 0 at $M \leq -\frac{C}{1-\alpha}$ to 1 at $M \geq \frac{C}{1-\alpha}$.

Substituting (3.11) in (3.10) and integrating by parts we obtain for each n:

$$\hat{V}(\mathbf{i}, M) = V_n(i_n, M) Q_n(\mathbf{i}, M) + \int_M^{\frac{C}{1-\alpha}} V_n(i_n, m) dQ_n(\mathbf{i}, m) \tag{3.12}$$

We will now show that $\hat{V}$ satisfies the optimality equation (3.7), and hence, by the uniqueness of the solution, $\hat{V} = V$; the proof follows in 3 steps.

Step 1: We show that $\hat{V}(\mathbf{i}, M) \geq M$: From the monotonicity of $V_n(i_n, m)$ in m we obtain, using (3.12)

$$\begin{aligned} \hat{V}(\mathbf{i}, M) &\geq V_n(i_n, M) Q_n(\mathbf{i}, M) + V_n(i_n, M) \int_M^{\frac{C}{1-\alpha}} dQ_n(\mathbf{i}, m) \\ &= Q_n(\mathbf{i}, \frac{C}{1-\alpha}) V_n(i_n, M) = V_n(i_n, M) \geq M \end{aligned} \tag{3.13}$$

Step 2: We show that $\Delta_n \geq 0$ for any n where:

$$\Delta_n = \hat{V}(\mathbf{i}, M) - R(i_n) - \alpha \sum_{j \in E} p(i_n, j) \hat{V}(i_1, \ldots, j, \ldots, i_N, M). \tag{3.14}$$

We note that $Q_n(\mathbf{i}, m)$ does not depend on the value of i_n, i.e.

$$Q_n(\mathbf{i}, m) = Q_n(i_1, \ldots, i_{n-1}, j, i_{n+1}, \ldots, i_N, m) \tag{3.15}$$

Substituting (3.12) twice in (3.14), and using $1 = Q_n(\mathbf{i}, M) + \int_M^{\frac{C}{1-\alpha}} dQ_n(\mathbf{i}, m)$, Δ_n is seen to be:

$$\begin{aligned} \Delta_n &= Q_n(\mathbf{i}, M) \left(V_n(i_n, M) - R(i_n) - \alpha \sum_{j \in E} p(i_n, j) V_n(j, M) \right) \\ &+ \int_M^{\frac{C}{1-\alpha}} \left(V_n(i_n, m) - R(i_n) - \alpha \sum_{j \in E} p(i_n, j) V_n(j, m) \right) dQ_n(\mathbf{i}, m) \\ &\geq 0 \end{aligned} \tag{3.16}$$

where

$$V_n(i_n, m) \geq R(i_n) + \alpha \sum_{j \in E} p(i_n, j) V_n(j, m) \tag{3.17}$$

follows from the optimality equation (2.13) for $V_n(i_n, m)$.

Step 3: Equality holds in (3.13,3.16), exactly under the Whittle policy: Consider $M, m > M_{n^*}(i_{n^*})$, then $Q_n(\mathbf{i}, M) = 1$ and $dQ_n(\mathbf{i}, m) = 0$. Looking at (3.12) we see that (3.13) holds as equality for $M \geq M_{n^*}(i_{n^*})$.

Consider next $M \leq m \leq M_n(i_n)$, for which (3.17) holds as an equality. Substituting this in (3.16), we have that for such M:

$$\Delta_n = \int_{M_n(i_n)}^{\frac{C}{1-\alpha}} \left(V_n(i_n, m) - R(i_n) - \alpha \sum_{j \in E} p(i_n, j) V_n(j, m)\right) dQ_n(\mathbf{i}, m). \tag{3.18}$$

In particular if we take n^*, we have

$$\Delta_{n^*} = \int_{M_{n^*}(i_{n^*})}^{\frac{C}{1-\alpha}} \left(V_{n^*}(i_{n^*}, m) - R(i_{n^*}) - \alpha \sum_{j \in E} p(i_{n^*}, j) V_{n^*}(j, m)\right) dQ_{n^*}(\mathbf{i}, m) = 0 \tag{3.19}$$

since $dQ_{n^*}(\mathbf{i}, m) = 0$ for $m \geq M_{n^*}(i_{n^*})$.

We have shown that $\hat{V}$ satisfies the optimality equations. Furthermore, we have shown that for $M > M_{n^*}(i_{n^*})$ we have $V(\mathbf{i}, M) = M$ so the optimal action is to retire, confirming (a), and for $M < M_{n^*}(i_{n^*})$ we have $V(\mathbf{i}, M) = R(i_{n^*}) - \alpha \sum_{j \in E} p(i_{n^*}, j) V_{n^*}(j, m)$, so the optimal action is to activate n^*, confirming (b). ■

3.4 Fourth Proof: The Achievable Region Approach

This proof is due to Bertsimas and Niño-Mora [5]. Conservation laws for queueing systems and their relation to combinatorial optimization structures such as polymatroids and extended polynatroids are discussed by Federgruen and Groenevelt [11], Shanthikumar and Yao [36], and Tsoucas et al. [41, 3]. Remarkably enough this proof is actually quite close to the pioneering proof of Klimov [24].

3.4.1 Achievable region and generalized conservation laws

Consider the N bandit system with initial vector of states $\mathbf{Z}(0) = \mathbf{i}$, and an arbitrary policy π. Let $I_i^\pi(t)$ be the indicator that policy π plays an arm which is in state i at time t. Define:

$$x_i^\pi = \mathbb{E}\left\{\sum_{t=0}^{\infty} I_i^\pi(t) \alpha^t | \mathbf{Z}(0) = \mathbf{i}\right\},$$

to be the total expected sum of discounted times at which an arm in state i is activated.

The vector $x^\pi = \{x_i^\pi,\, i \in E\}$ is a vector of performance measures of the policy π. For a linear reward function given by $R(i)$ the value of the objective function is obtained from

these performance measures as $\sum_{i\in E} R(i)x_i^\pi$. The set $\{x^\pi,\ \pi$ an admissible policy$\}$ is called the achievable region.

Recall the definitions (2.7, 2.8) of T_i^S, A_i^S, $S \subseteq E$. For initial state $\mathbf{Z}(0) = \{i_1, \ldots, i_N\}$, denote $T^S_{\mathbf{Z}(0)} = \sum_{n:i_n \notin S} T^S_{i_n}$, and let:

$$b(S) = \frac{\mathbb{E}\{\alpha^{T^S_{\mathbf{Z}(0)}}\}}{1-\alpha} \tag{3.20}$$

Theorem 3.2 (Generalized conservation law) *For initial state* $\mathbf{Z}(0)$*, for every policy* π *and every* $S \subseteq E$

$$\sum_{i\in S} A_i^S x_i^\pi \geq b(S). \tag{3.21}$$

Equality for S *holds if and only if* π *gives priority to states in* $E\backslash S$ *over states in* S*.*

Proof. Consider a realization (single sample path) under policy π. Then $T^S_{Z(0)}$ is the total time necessary to get all the arms not initially in S into S. We can divide the time axis according to what we do at each time into three parts:

Let $s_0(1) < \cdots < s_0(T^S_{Z(0)})$, be the times at which we operate on arms which were initially in $E\backslash S$, before they have entered a state in S.

Let $s_i(1) < \cdots < s_i(l) < \cdots$ be the times at which we activate arms in state i, where $i \in S$.

Let $s_{i,l}(1) < \cdots < s_{i,l}(T_i^S(i,l)-1)$ be the times at which, following the l'th activation of an arm in state i (where $i \in S$) we activate that same arm, until it returns to S. Here $T_i^S(i,l)$ counts the number of active steps of the arm which is activated on this lth visit to state i, from i until it returns to S.

It is possible that in the infinite time horizon we will never activate a particular arm again, in which case the corresponding $s_0(l)$, $s_i(l)$, $s_{i,l}(k)$ will be $= \infty$.

Clearly, at any time that we activate any arm we are doing one of the above three things. Hence:

$$\frac{1}{1-\alpha} = \sum_{k=1}^{T^S_{Z(0)}} \alpha^{s_0(k)} + \sum_{i\in S}\sum_{l=1}^{\infty}\left(\alpha^{s_i(l)} + \sum_{k=1}^{T_i^S(i,l)-1} \alpha^{s_{i,l}(k)}\right)$$

which is a conservation law, in that it holds for every sample path, for every policy.

We now obtain an inequality:

$$\sum_{i\in S}\sum_{l=1}^{\infty} \alpha^{s_i(l)}\left(1+\alpha+\cdots+\alpha^{T_i^S(i,l)-1}\right) \geq \sum_{i\in S}\sum_{l=1}^{\infty}\left(\alpha^{s_i(l)} + \sum_{k=1}^{T_i^S(i,l)-1} \alpha^{s_{i,l}(k)}\right) \tag{3.22}$$

$$= \frac{1}{1-\alpha} - \sum_{k=1}^{T^S_{Z(0)}} \alpha^{s_0(k)} \geq \frac{1}{1-\alpha} - \left(1+\alpha+\cdots+\alpha^{T^S_{Z(0)}-1}\right) = \frac{\alpha^{T^S_{Z(0)}}}{1-\alpha}$$

where the first inequality can hold as equality if and only if $s_{i,l}(1), \ldots, s_{i,l}(T_i^S(i,l)-1) = s_i(l)+1, \ldots, s_i(l)+T_i^S(i,l)-1$, for every i,l, and the second inequality can hold as equality if and only if $s_0(1), \ldots, s_0(T^S_{Z(0)}) = 0, 1, \ldots, T^S_{Z(0)}-1$. But that happens exactly whenever π gives priority to states in $E \backslash S$ over states in S.

This proves a pathwise version of the theorem. The theorem now follows by taking expectations on the two sides of the inequalities (3.22). ■

Corollary 3.3

$$\sum_{i \in E} x_i^\pi = \sum_{i \in E} A_i^E x_i^\pi = b(E) = \frac{1}{1-\alpha} \tag{3.23}$$

Proof. The first equality follows from $A_i^E = 1$. The third equality follows from the definition of $b(E)$, since $T^E_{\mathbf{Z}(0)} = 0$. Finally, clearly $\sum_{i \in E} x_i^\pi = \frac{1}{1-\alpha}$. The second equality is consistent with the preivous theorem, in the sense that every policy gives priority to $\emptyset$ over E. ■

According to the generalized conservation law, the following linear programming problem is a relaxation of the multiarmed bandit problem:

$$\begin{aligned} \max \quad & \sum_{i \in E} R(i) x_i \\ s.t. \quad & \sum_{i \in S} A_i^S x_i \geq b(S), \qquad S \subset E, \\ & \sum_{i \in E} x_i = \sum_{i \in E} A_i^E x_i = b(E) = \frac{1}{1-\alpha}, \\ & x_i \geq 0, \quad i \in E. \end{aligned} \tag{3.24}$$

It is a relaxation in the sense that for any policy π the performance measures x_i^π have to satisfy the constraints of the linear program.

3.4.2 The Linear Program

To complete the proof we investigate the linear program (3.24). The rest of the proof is restricted to the case of a finite number of states, $|E| < \infty$. We generalize the proof to infinite countable state spaces in Section 4.

Let $\varphi(1), \ldots, \varphi(|E|)$ be a permutation of the states $1, \ldots, |E|$, and denote by φ the priority policy which uses this permutation order (i.e. $\varphi(1)$ has highest priority, $\varphi(2)$ 2nd highest etc.). Denote $S_{\varphi(i)} = [\varphi(i), \ldots, \varphi(|E|)]$, $i = 1, \ldots, |E|$. We now look at the performance measures $x^\varphi_{\varphi(i)}$.

Consider the upper triangular matrix (which is part of the coefficient matrix of the LP (3.24), consisting of the constraints corresponding to the subsets $S_{\varphi(1)}, \ldots, S_{\varphi(|E|)}$):

$$D = \begin{bmatrix} A_{\varphi(1)}^{S_{\varphi(1)}} & A_{\varphi(2)}^{S_{\varphi(1)}} & \cdots & A_{\varphi(|E|)}^{S_{\varphi(1)}} \\ 0 & A_{\varphi(2)}^{S_{\varphi(2)}} & \cdots & A_{\varphi(|E|)}^{S_{\varphi(2)}} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_{\varphi(|E|)}^{S_{\varphi(|E|)}} \end{bmatrix}.$$

By Theorem 3.2, the performance measures $\mathbf{x}^\varphi$ are the unique solution of the triangular set of equations:

$$D \begin{bmatrix} x^\varphi_{\varphi(1)} \\ x^\varphi_{\varphi(2)} \\ \vdots \\ x^\varphi_{\varphi(|E|)} \end{bmatrix} = \begin{bmatrix} b(S_{\varphi(1)}) \\ b(S_{\varphi(2)}) \\ \vdots \\ b(S_{\varphi(|E|)}) \end{bmatrix}. \tag{3.25}$$

In fact they are a basic feasible solution to the LP (3.24), in which all the remaining inequalities, for all $S \neq S_{\varphi(1)}, \ldots, S_{\varphi(|E|)}$ have non-zero slacks. Thus, the vector of performance measures of each priority policy is an extreme point of the LP, corresponding to a non-degenerate basic solution.

The dual of the linear program (3.24) is

$$\begin{aligned} \min \quad & \sum_{S \subseteq E} b(S) y^S \\ s.t. \quad & \sum_{S: i \in S} A_i^S y^S \geq R(i), \qquad i \in E, \\ & y^S \leq 0, \quad S \subset E. \end{aligned} \tag{3.26}$$

The complementary slack dual solution corresponding to $\mathbf{x}^\varphi$ is of the form $y^S = 0, S \neq S_{\varphi(1)}, \ldots, S_{\varphi(|E|)}$ while the remaining dual variables solve:

$$D' \begin{bmatrix} y^{S_{\varphi(1)}} \\ y^{S_{\varphi(2)}} \\ \vdots \\ y^{S_{\varphi(|E|)}} \end{bmatrix} = \begin{bmatrix} R(\varphi(1)) \\ R(\varphi(2)) \\ \vdots \\ R(\varphi(|E|)) \end{bmatrix} \tag{3.27}$$

which gives recursively, for $S_{\varphi(1)}, \ldots, S_{\varphi(|E|)}$:

$$y^{S_{\varphi(i)}} = \frac{R(\varphi(i)) - \sum_{j=1}^{i-1} A^{S_{\varphi(j)}}_{\varphi(i)} y^{S_{\varphi(j)}}}{A^{S_{\varphi(i)}}_{\varphi(i)}},$$

To show that $\mathbf{x}^\varphi$ is optimal it remains to show that $y^{S_{\varphi(2)}} \leq 0, \ldots, y^{S_{\varphi(|E|)}} \leq 0$ (recall that $S_{\varphi(1)} = E$, and hence y^E is unrestricted in sign).

Consider then the permutation defined by the Gittins priority order. Then the values of the index calculated by Klimov's algorithm are $\nu(\varphi(1)) \geq \cdots \geq \nu(\varphi(|E|))$. Furthermore, the values of $y^{S_{\varphi(i)}}$ of the dual solution are exctly those calculated by the second version of Klimov's algorithm, and they satisfy:

$$y^{S_{\varphi(1)}} = \nu(\varphi(1)) \tag{3.28}$$

$$y^{S_{\varphi(j)}} = \nu(\varphi(j)) - \nu(\varphi(j-1)) \leq 0, \quad j = 2, \ldots, |E|, \tag{3.29}$$

Hence the Gittins index priority policy is optimal.

3.4.3 Extended polymatroids

Theorem 3.4 *The achievable region is a bounded convex polytope which coincides with the feasible solutions of the linear program (3.24). It has $|E|!$ distinct extreme points which are the performance vectors of the priority policies given by all the permutations of the states.*

Proof. If we let R vary over all possible $|E|$ vectors, the solutions of (3.24) vary over all the extreme points of the feasible polyhedron of the LP. But for each such R Klimov's algorithm finds an optimal permutation priority policy which has that extreme point as its performance vector. Hence: The achievable performance region contains all the solutions of the LP, and so it coincides with it. Further more, its extreme points coincide with the performance vectors of the priority policies, which are defined by the $|E|!$ permutations of the states. ■

Theorem 3.4 establishes that the achievable region of a multiarmed bandit problem is an extended polymatroid, as we explain now.

Polyhedral sets of the form:

$$\mathcal{M} = \left\{ \mathbf{x} \in \mathbb{R}_+^{|E|} : \sum_{i \in S} x_i \geq b(S), S \subseteq E \right\}$$

where b is a supermodular function, i.e. $b(S_1) + b(S_2) \leq b(S_1 \cup S_2) + b(S_1 \cap S_2)$, are called *polymatroids* [10], and are of great importance in combinatorial optimization because of the following property: Let $\varphi(1), \ldots, \varphi(|E|)$ be a permutation of $1, \ldots, |E|$, and let $S_{\varphi(1)}, \ldots, S_{\varphi(|E|)}$ be the nested subsets $S_{\varphi(i)} = \{\varphi(i), \ldots, \varphi|E|)\}$. Then $\mathcal{M}$ has exactly $|E|!$ extreme points given by the solutions of:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \mathbf{x}^{\varphi} = \begin{bmatrix} b(S_{\varphi(1)}) \\ b(S_{\varphi(2)}) \\ \vdots \\ b(S_{\varphi(|E|)}) \end{bmatrix}.$$

This property implies that the optimization of any objective function linear in $\mathbf{x}$ is achieved by a greedy solution: for reward function R the optimal solution is given by the $\mathbf{x}^{\varphi}$ of the permutation $R(\varphi(1)) \geq R(\varphi(2)) \geq \cdots \geq R(\varphi(|E|))$.

Relations between polymatroids and conservation laws of stochastic systems are explored by Federgruen and Gronevelt [11], and by Shanthikumar and Yao [36].

Tsoucas [3, 41] and Bertsimas and Niño-Mora [5] define extended polymatroids as a generalization to polymatroids. An extended polymatroid with coefficients $a_i^S > 0, i \in S \subseteq E$ is a polyhedral set:

$$\mathcal{EM} = \left\{ \mathbf{x} \in \mathbb{R}_+^{|E|} : \sum_{i \in S} a_i^S x_i \geq b(S), S \subseteq E \right\}$$

which satisfies the following property (rather than require b supermodular): For every permutation and nested sets as above, the solution to

$$\begin{bmatrix} a_{\varphi(1)}^{S_{\varphi(1)}} & a_{\varphi(2)}^{S_{\varphi(1)}} & \cdots & a_{\varphi(|E|)}^{S_{\varphi(1)}} \\ 0 & a_{\varphi(2)}^{S_{\varphi(2)}} & \cdots & a_{\varphi(|E|)}^{S_{\varphi(2)}} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{\varphi(|E|)}^{S_{\varphi(|E|)}} \end{bmatrix} \mathbf{x}^{\varphi} = \begin{bmatrix} b(S_{\varphi(1)}) \\ b(S_{\varphi(2)}) \\ \vdots \\ b(S_{\varphi(|E|)}) \end{bmatrix},$$

is in $\mathcal{EM}$.

Our discussion here shows that extended polymatroids share the property of polymatroids: The above solutions are the $|E|!$ extreme points of the polyhedral set, and the optimization of any objective function linear in $\mathbf{x}$ is achieved by a greedy solution, which constructs the optimal permutation. The proof of the generalized conservation laws in Section 3.4.1 shows that the achievable region of the multiarmed bandit problem is an extended polymatroid.

4 Extension of achievable region proof to infinite countable state spaces

4.1 Linear programs in general vector spaces

In this section we summarize the general structure of a linear programming problem and its dual as defined on general vector spaces; for a more detailed reference see Anderson and Nash's book [1], see also Shapiro [37] and Barvinok [2].

The primal space X is a linear vector space over the real numbers $\mathbb{R}$. The positive cone P_X is a convex cone (set of vectors in X which is closed under positive linear combinations). X is partially ordered by P_X, where we say that x is greater or equal to y if $x - y \in P_X$. The algebraic dual of X is the vector space X^* of all the linear functionals from X to $\mathbb{R}$. The image of $x \in X$ under $x^* \in X^*$ is denoted $\langle x, x^* \rangle$. The positive cone P_{X^*} is defined by:

$$P_{X^*} = \{x^* \in X^* : \langle x, x^* \rangle \geq 0 \text{ for all } x \in P_X\}.$$

Similarly for the dual problem we have a linear space Y, a positive cone P_Y, and their algebraic duals Y^* and P_{Y^*}.

Let A be a linear map from X to Y, let $r \in X^*$ and $b \in Y$ be given. The primal linear program is to find $x \in X$ so as to:

$$\begin{array}{lll} LP: & \text{maximize} & \langle x, r \rangle \\ & \text{subject to} & Ax - b \in P_Y \\ & & x \in P_X \end{array}$$

Let A^* be the dual linear map from Y^* to X^*, defined by

$$\langle x, A^* y^* \rangle = \langle Ax, y^* \rangle, \qquad \text{for all } x \in X, y^* \in Y^*. \tag{4.1}$$

The dual linear program is to find $y^* \in Y^*$ so as to:

$$\begin{array}{lll} AD: & \text{minimize} & \langle b, y^* \rangle \\ & \text{subject to} & A^* y^* - r \in P_{X^*} \\ & & -y^* \in P_{Y^*} \end{array}$$

The following is the weak duality theorem for linear programs in general vector spaces. It establishes sufficient conditions for x and y^* to be optimal solutions for the primal and the dual problems. We provide the proof here for the sake of completeness.

Theorem 4.1 *If x is feasible for LP and y^* is feasible for AD then*

$$\langle b, y^* \rangle \geq \langle x, r \rangle.$$

Equality holds if and only if

$$\langle b - Ax, y^* \rangle = 0, \qquad \langle x, A^* y^* - r \rangle = 0,$$

in which case we say that x, y^* are complementary slack.

If x, y^* are feasible and complementary slack then x and y^* are optimal solutions to LP and AD respectively.

Proof. Assume that x is feasible for LP and y^* is feasible for AD.

$Ax - b \in P_Y$ and $-y^* \in P_{Y^*}$ imply $\langle Ax - b, -y^* \rangle \geq 0$ or equialently $\langle b - Ax, y^* \rangle \geq 0$.

$x \in P_X$ and $A^* y^* - r \in P_{X^*}$ imply $\langle x, A^* y^* - r \rangle \geq 0$.

Hence:

$$\langle b, y^* \rangle \geq \langle Ax, y^* \rangle = \langle x, A^* y^* \rangle \geq \langle x, r \rangle$$

Clearly if this inequality holds as an equality for some feasible x, y^* then x and y^* are optimal solutions to LP and AD respectively.

Equality can hold if and only if $\langle b - Ax, y^* \rangle = 0$ and $\langle x, A^* y^* - r \rangle = 0$. ■

4.2 Formulation of the multi-armed bandit problem as general LP

We have shown in Section 3.4.1 that the achievable region of the bandit problem under all policies must obey conservation laws which define the LP problem (3.24). We now reformulate the LP problem and its dual (3.26) in terms of Section 4.1.

We take as our primal space X the space of all sequences $x = \{x_i\}_{i \in E}$, and $x_i \in \mathbb{R}$ such that $\sum_{i \in E} |x_i| < \infty$. X is a linear vector space. We take as the positive cone the non-negative sequences, $P_X = \{x \in X, x_i \geq 0, i \in E\}$. The reward function $\mathbf{R}$: $\mathbf{R} = \{R(i)\}_{i \in E}$, where $|R(i)| \leq C$, is a linear functional on X, that is $\mathbf{R} \in X^*$, with $\langle x, \mathbf{R} \rangle = \sum_{i \in E} x_i R(i)$.

Remark Summation over $i \in E$ is well defined even if we do not have a well ordering of the countable set E, whenever the sum is absolutely convergent. This is always the case here, by the boundedness of $\mathbf{R}, b(S), A_i^S$, and absolute convergence of $x \in X$.

We take as our dual space Y a subspace of all the functions mapping the subsets of E to the real line. We include in Y the function $\mathbf{b} : 2^E \to \mathbb{R}$ given by $b(S)$ as defined in (3.20), for

every $S \subseteq E$. We also include in Y all the images of X under the linear map A defined as follows: The function Ax for $x \in X$ is given by

$$Ax(S) = \sum_{j \in S} A_j^S x_j. \tag{4.2}$$

Recall that $b(S), A_j^S$ are bounded below and above by 0 and $\frac{1}{1-\alpha}$, for all $S \subseteq E$. We let Y be the linear vector space spanned by $\mathbf{b}$ and $Ax, x \in X$, so that every $y \in Y$ is of the form $y = Ax + a\mathbf{b}$ for some $a \in \mathbb{R}$ and some $x \in X$. We take as the positive cone in Y the set $P_Y = \{y \in Y : y(S) \geq 0, S \subset E, y(E) = 0\}$.

For any $y \in Y$ we will be particularly interested in the values of $y(S)$ for subsets of the form $S(v), S^-(v), S_i, S_i^-$ as defined in (2.4, 2.5). For $y \in Y$ we define a real function $g_y : \mathbb{R} \to \mathbb{R}$ as:

$$g_y(v) = y(S(v)). \tag{4.3}$$

We need the following Proposition, which derives the properties of g_y, $y \in Y$. Property (iv) below which allows the use of integration by parts is the crucial element in our extension of the proof to infinite countable state spaces.

Proposition 4.2 *Let $y \in Y$. Then:*

(i) $y(\emptyset) = 0$.

(ii) g_y *is of bounded variation.*

(iii) *The function $g_y(v)$ is continuous from the right with left limits, given by:*

$$g_y(v^-) = \lim_{v_n \nearrow v} g_y(v_n) = y(S^-(v)), \qquad g_y(v^+) = \lim_{v_n \searrow v} g_y(v_n) = y(S(v)) = g_y(v). \tag{4.4}$$

(iv) *The jumps of $g_y(v)$ are given by:*

$$g_y(v) - g_y(v^-) = \sum_{i:\nu(i)=v} y(S_i) - y(S_i^-), \tag{4.5}$$

The proof of the Proposition is quite technical, and is based on a detailed study of the properties of the functions $A_j^S, b(S)$. In the proof of the crucial part (iv) we need to resort back to sample path arguments based on the sequence of stopping time and states $\mathcal{T}_\ell$, k_ℓ. We present the complete proof in the Appendix.

4.3 The solution of the infinite LP

Let $x^G \in X$ be the performance measure under the Gittins index prioirty policy, that is x_i^G is the expected discounted time that an arm in state i is activated under the policy.

Let $\nu^* \in Y^*$ be the linear functional on Y defined by the Gittins index, as

$$\langle y, \nu^* \rangle = \sum_{i \in E} \nu(i)[y(S_i) - y(S_i^-)]. \tag{4.6}$$

We now show that x^G and ν^* solve the infinite linear programs LP and AD respectively. To do so we need to show, by the weak duality theorem 4.1:

The feasibility of x^G:

(a) $Ax^G - \mathbf{b} \in P_Y$

(b) $x^G \in P_X$

The feasibility of ν^*:

(c) $A^*\nu^* - \mathbf{R} \in P_{X^*}$

(d) $-\nu^* \in P_{Y^*}$

Complementary slackness

(e) $\langle Ax^G - \mathbf{b}, \nu^* \rangle = 0$

(f) $\langle x^G, A^*\nu^* - \mathbf{R} \rangle = 0$

We know that (a) and (b) hold and x^G is feasible for LP, because by the conservation theorem 3.2 and corollary 3.3, the performance measure x^π for every policy π satisfies (3.21, 3.23), which is (a), and $x_i^\pi \geq 0$ which is (b).

We will show in next Proposition 4.3 that $A^*\nu^* - \mathbf{R} = 0$, which will prove (c) and (f). Propositions 4.4, 4.5 will prove (e) and (d) respectively.

Proposition 4.3 *The linear functional ν^* satisfies*

$$A^*\nu^* = \mathbf{R}$$

Proof. To show that $A^*\nu^*$ is in fact the linear functional $\mathbf{R}$ we calculate $\langle x, A^*\nu^* \rangle$ and show that it is equal to $\langle x, \mathbf{R} \rangle$ for all $x \in X$. This is where we use the result of Proposition 2.7.

$$\begin{aligned}
\langle x, A^*\nu^* \rangle &= \langle Ax, \nu^* \rangle \\
&= \sum_{i \in E} \nu(i)[Ax(S_i) - Ax(S_i^-)] \\
&= \sum_{i \in E} \nu(i)[\sum_{j \in S_i} A_j^{S_i} x_j - \sum_{j \in S_i^-} A_j^{S_i^-} x_j] \\
&= \sum_{i \in E} \nu(i) A_i^{S_i} x_i - \sum_{i \in E} \nu(i) \sum_{j : j \prec i} (A_j^{S_i^-} - A_j^{S_i}) x_j \\
&= \sum_{i \in E} \nu(i) A_i^{S_i} x_i - \sum_{j \in E} x_j \sum_{i : i \succ j} \nu(i)(A_j^{S_i^-} - A_j^{S_i}) \\
&= \sum_{j \in E} x_j [\nu(j) A_j^{S_j} - \sum_{i : i \succ j} \nu(i)(A_j^{S_i^-} - A_j^{S_i})] \\
&= \sum_{j \in E} x_j R(j) \\
&= \langle x, \mathbf{R} \rangle
\end{aligned}$$

The first equality is the definition of A^* in (4.1). The second follows from the definition of ν^* in (4.6). The third follows from the definition of the map Ax in (4.2). In the fourth we separate $S_i = S_i^- \cup i$. In the fifth we change the order of summation, which is justified by $\nu(i), A_j^S$ bounded and $x \in X$ absolutely convergent. In the sixth we relabel i to j in the first summation and take out x_j before the brackets. The seventh follows from the statement (2.25) of Proposition 2.7. The last is the definition of $\mathbf{R}$. ■

Proposition 4.4 *The primal solution x^G and the dual solution ν^* satisfy:*

$$\langle Ax^G, \nu^* \rangle = \langle \mathbf{b}, \nu^* \rangle$$

Proof.

$$\begin{aligned}
\langle Ax^G, \nu^* \rangle &= \sum_{i \in E} \nu(i)[Ax^G(S_i) - Ax^G(S_i^-)] \\
&= \sum_{i \in E} \nu(i)[\sum_{j \in S_i} A_j^{S_i} x_j^G - \sum_{j \in S_i^-} A_j^{S_i^-} x_j^G] \\
&= \sum_{i \in E} \nu(i)[b(S_i) - b(S_i^-)] \\
&= \langle \mathbf{b}, \nu^* \rangle
\end{aligned}$$

Where the first, the second, and the last equalities follow by definition, and the third equality holds by the conservation law of theorem 3.2, because the policy G gives prioirty to $E \backslash S_i$ over S_i, and to $E \backslash S_i^-$ over S_i^-. ■

Proposition 4.5 *The linear functional ν^* is non-positive, that is $-\nu^* \in P_{Y^*}$.*

Proof. We need to show that for all $y \in P_Y$, $\langle y, -\nu^* \rangle \geq 0$, equivalently that $\langle y, \nu^* \rangle \leq 0$. Let

$$\mathcal{E} = \{v : \exists j \in E, \nu(j) = v\}$$

denote the countable set of values of ν. The steps of the proof are:

$$\begin{aligned}
\langle y, \nu^* \rangle &= \sum_{i \in E} \nu(i)[y(S_i) - y(S_i^-)] \\
&= \sum_{v \in \mathcal{E}} v \sum_{i:\, \nu(i) = v} [y(S_i) - y(S_i^-)] \\
&= \sum_{v \in \mathcal{E}} v[g_y(v) - g_y(v^-)] \\
&= \int_{-C}^{C} v dg_y(v) \\
&= Cg_y(C) + Cg_y(-C) - \int_{-C}^{C} g_y(v) dv \\
&= Cy(E) + Cy(\emptyset) - \int_{-C}^{C} y(S(v)) dv \\
&\leq 0
\end{aligned}$$

The first equality follows from the definition of ν^* in (4.6). In the second we write the sum over E as a sum over sets with equal values of the index. The third equality is the crucial result (4.5) of part (iv) of Proposition 4.2.

For each $y \in Y$, by parts (ii,iii) of Proposition 4.2, the function g_y is of bounded variation and right continuous, so it can be written as the difference of two monotone non-decreasing right continuous functions. Hence there exists a unique signed Borel measure $\mu(a,b] = g_y(b) - g_y(a)$ (see Royden [33] Proposition 12 page 262), and the Lebesgue Stielties integral $\int_{-C}^{C} v dg_y(v)$ is well defined as the Lesbegue integral of the function v over the interval (-C,C) with respect to the measure μ. From the definition of g_y the measure μ is a discrete measure, concentrated on the set $\mathcal{E}$. Recall that ν is bounded (same bound as R), $|\nu(i)| < C$, hence all the values in $\mathcal{E}$ are bounded between $-C, C$. Hence the fourth equality holds.

Furthermore, because v is continuous and $v, g_y(v)$ are of bounded variation, the Riemann-Stieltjes integral $\int_{-C}^{C} v dg_y(v)$ is well defined and the Lebesgue-Stieltjes integral agrees with the Riemann-Stieltjes integral (see [33] page 262, and Rudin [34] theorem 6.8), and for the Rieman-Stieltjes integral we can use integration by parts (see [34] theorem 6.30) to obtain the fifth equality. The sixth equality follows from the definition of g_y where we note that $S(C) = E$, and $S(-C) = \emptyset$.

By part (i) of Proposition 4.2 $y(\emptyset) = 0$ for all $y \in Y$. Furthermore, by definition of P_Y, $y(E) = 0$, and $y(S) \geq 0$ for all $y \in P_Y$. Hence the last inequality of ≤ 0. This completes the proof. ■

5 Discussion

In this section we discuss the various proofs, and their uses in the develpment of further theory and results. We start in Section 5.1 with a brief analysis of the main features of each proof, and how they relate to each other. We also include a brief discussion of some additional proof ideas in Section 5.2. We remark briefly on some more recent results that go beyond optimality of the index in Section 5.3.

5.1 General observations about the proofs

We have surveyed four proofs each based on a different idea. There is a certain structure to the proofs; the first two are essentially primal proofs while the last two are dual proofs. The proofs provide successively more detailed information about the model and its solution.

5.1.1 Comments on the interchange argument

The interchange argument proof of Section 3.1 is the simplest, most direct proof. If alternative actions provide different rewards per unit time, then doing the one with highest reward per unit time earlier is preferable, because the rewards are discounted. Changing the order of two such action one obtains immediate proof. This is essentially the idea of the proof. However, to evaluate the reward per unit time for a bandit process one should choose a duration to

maximize this reward per unit time, and the maximum is over all random positive stopping times.

Notice that the times which are switched around in the interchange argument are maximizing stopping times for the highest index arm n^* and an arbitrary arm n. Hence the proof makes very direct use of the definition of the Gittins.

While the proof is very direct and simple, it tells us nothing about the features of the optimal solution.

5.1.2 Comments on interleaving of prevailing charges

The interleaving of prevailing charges proof of Section 3.2 does a little more towards studying the model and its behavior under general policies and under the Gittins index prioirty policy.

One observes that the sample path of each arm during the periods in which it is active is independent of other arms and of the policy. The fair charge and prevailing charge are two stochastic processes which accompany the states of the arm, and share this property. Thus a policy is simply a way of interleaving the active times, and creating an interleaved sequence of the activation times, of the states, of the fair charges, and of the prevailing charges from all the N arms. The conservation laws of the fourth proof are also based on this observation.

The proof then is quite simple: Interleaving of the prevailing charge process and adding up the discounted prevailing charges produces a quantity whose expectation is an upper bound on the expeceted sum of the discounted interleaved rewards of the arms.

This expected sum of discounted interleaved prevailing charges is maximized when the prevailing charges are rearranged to form a decreasing sequence. The maximization follows by Hardy Littelwood Polya argument, or equivalently by an interchange argument.

The next observation is that one can indeed interleave the prevailing charges so that they form a decreasing sequence, and this is exactly what is produced by the Gittins index priority policy. Furthermore, in this particular case the upper bound is sharp: The expected sum of discounted interleaved prevailing charges and the expected sum of discounted rewards are equal under the Gittins priority policy.

Note that a property of the Gittins priority policy is that an arm which is activated is kept active for the duration of the Gittins stopping time. In terms of fair and prevailing charges this says that arms can only be switched when prevailing and fair charges are equal.

To maximize the discounted sum of rewards it is clearly optimal to try and catch the big rewards early. This of course cannot be realized on every sample path. However, the sequence of prevailing charges under the Gittins priority policy is indeed non-increasing for every sample path. It is in fact equal to a conditional expected reward per unit time at all times, where the conditioning is on the state of all the inactive arms and on the state of the active arm at the time of activation.

This second proof is based on the existence of the prevailing charge, and the fact that the Gittins priority policy produces the interleaving which is the non-increasing rearrangement of the prevailing charges. The proof does not use the explicit form of the prevailing charges. The next proofs study the form of the optimal solution in more details.

5.1.3 Comments on the retirement option approach

In the retirement option proof we add to the initial state of the system, given by $i_1, \ldots, i_N$ another parameter M which is a reward for retiring, and consider the optimal expected reward $V(\mathbf{i}, M)$ as a function of M. For a single arm one would retire at time $\tau(i, M)$ for a reward $\sum_{t=0}^{\tau(i,M)-1} \alpha^t R(Z(t)) + M\alpha^{\tau(i,M)}$, and one has in fact (see (2.20)) that $\frac{\partial}{\partial M} V(i, M) = \mathbb{E}(\alpha^{\tau(i,M)})$. For the multiarmed bandit with retirement option one wishes to prove that the optimal policy is to follow the Gittins priority policy, but never activate again an arm n which has reached its $\tau(i_n, M)$. Taking this as a conjecture implies that $\frac{\partial}{\partial M} V(\mathbf{i}, M) = \prod_{n=1}^{N} \mathbb{E}(\alpha^{\tau_n(i_n,M)}) = \prod_{n=1}^{N} \frac{\partial}{\partial M} V_n(i_n, M)$, from which a more explicit form (3.12) is derived. The proof is completed by a technical part showing that this conjectured form satisfies the optimality equations.

As we have seen, the Gittins index $\nu(i)$ is equal to the fair charge $\gamma(i)$ or the retirement reward per unit discounted time $(1-\alpha)M(i)$, which indicates that ν, γ, M can all be interpreted as shadow costs for expected active time. This economic interpretation leads to a Lagrangean formulation and a dual problem, which are the basis for the retirement option proof. We describe this Lagrangean dual approach now.

The solution of the retirement option problem consists of activating arms up to some time τ and retirement for the remaining time. In terms of discounted expected time, one is active for a total of $\mathbb{E}\left(\frac{1-\alpha^\tau}{1-\alpha}\right)$, and retired for the remaining $\mathbb{E}\left(\frac{\alpha^\tau}{1-\alpha}\right)$.

Consider then a constrained optimization problem: Maximize the expected rewards from the N arms, while using a limited amount of expected discounted active time:

$$\text{Primal Problem:} \qquad \begin{array}{ll} \max_\pi & \mathbb{E}_\pi \left\{ \sum_{t=0}^{\infty} \alpha^t R(t) | \mathbf{Z}(0) = \mathbf{i} \right\} \\ \text{subject to} & \mathbb{E}_\pi \left\{ \sum_{t=0}^{\infty} \alpha^t \, 1_{\text{retired}}(t) | \mathbf{Z}(0) = \mathbf{i} \right\} = b \end{array} \tag{5.1}$$

We can dualize the work constraint by a Lagrange multiplier M, to write the Lagrangean:

$$\mathcal{L}(\pi, b, M) = \mathbb{E}_\pi \sum_{t=0}^{\infty} \alpha^t R(t) + \frac{M}{1-\alpha} \left(\mathbb{E}_\pi \sum_{t=0}^{\infty} \alpha^t \, 1_{\text{retired}}(t) - b \right).$$

Maximizing this for given M, b over all policies is equivalent to solving Whittle's retirement option problem:

$$V(\mathbf{i}, M) = \max_\pi \mathbb{E}_\pi \left\{ \sum_{t=0}^{\infty} \alpha^t R(t) + \frac{M}{1-\alpha} \sum_{t=0}^{\infty} \alpha^t \, 1_{\text{retired}}(t) | \mathbf{Z}(0) = \mathbf{i} \right\}.$$

The dual to (5.1) is:

$$\text{Dual Problem:} \quad \min_M \max_\pi \mathcal{L}(\pi, b, M) = \min_M \left\{ V(\mathbf{i}, M) - \tfrac{M}{1-\alpha} \, b \right\}. \tag{5.2}$$

For any two policies π', π'' we can randomize and choose π' with probability θ, π'' with probability $1-\theta$. Hence our problem has convex domain and concave objective, and the optimal values of the primal (5.1) and the dual (5.2) are equal.

5.1.4 Comments on the achievable region approach

The idea of achievable regions is useful in many optimization problems. Instead of optimizing the objective function over decision variables $\theta \in \Theta$ one maps Θ onto an achievable region $\mathbb{X}$, and considers the optimization of the objective function over the acievable values $\chi \in \mathbb{X}$, where the χ is some performance measure of θ, often the value of some constrained quantity. In the context of dynamic programming, one can map the policies π (Markovian randomized policies) onto the simpler space of $\chi_s^a, s \in S, a \in A$ of the expected discounted time spent in state s taking action a, where S is the state space, A the action space. In the general context of dynamic programming we get an LP formulation based on this achievable region. The objective is the linear $\sum_{s,a} R(s,a)\chi_s^a$, and the achievable region consists of the solution to the balance equations (Poisson equation), so that the LP formulation is

$$\begin{aligned} \max \quad & \sum_{s,a} R(s,a)\chi_s^a \\ s.t. \quad & \sum_a (I - \alpha P'^a)\chi^a = P_0 \end{aligned}$$

where P^a is the transition matrix under action a, and P_0 the initial distribution of the states.

In the case of bandit problems the full state space consists of the states $i \in E$ of the N arms, given by $(i_1, \ldots, i_N)$, or more compactly by the counts $n_1, \ldots, n_{|E|}$ of the number of bandit arms in state $i \in E$. The size of the state space $\kappa =$ the number of partitions of N arms into $|E|$ states, and the LP has therefore $|E| \times \kappa$ variables and κ constraints. Note that often though not always $\kappa > 2^{|E|}$.

In the achievable region approach of Section 3.4 we are not using the full state description with full performance measures $\chi_s^a, s \in S, a \in A$. Instead, because at any time only one arm is chosen to be activated, the measure used is $x_i, i \in E$, the expected discounted time that the active arm is an arm in state i. For every policy π there will be a full set of $\chi_s^a, s \in S, a \in A$, however we will only require the values of $x_i^\pi, i \in E$ the values of the x_i under π.

The generalized conservation laws are the outcome of the same idea on which the interleaving of prevailing charges is based: The path of each arm is independent of the policy. Taking expectations, the generalized conservation laws become a set of linear constraints on the achievable $x_i^\pi, i \in E$, with one constraint for each subset $S \subseteq E$, and the resulting LP (3.24) has $|E|$ variables and $2^{|E|}$ constraints.

The proof of the Gittins index theorem is now as follows: The dual problem to (3.24) is (3.26), and the primal and dual problems (3.24, 3.26) are then solved by Klimov's algorithm, which calculates the Gittins priority order and the values of the Gittins index.

Apart from providing an algorithm to claculate the Gittins index, the achievable region and generalized conservation laws approach also provides us with a very informative description of the achievable region - as we discussed in Section 3.4.3 the achievable region is en extended polymatroid.

5.2 Some additional proofs

Several additional proofs, using various deep insights from the theory of dynamic programming have been proposed. We survey a few of those in this section. For some additional methods and ideas for computation of the index see recent survey by Chakravorty and Mahajan [6]. Our presentation of these proofs is limited to the basic insights on which they are based, because as far as we understand them these proofs are based to a lesser or greater extent on the four proofs of Section 3. An interesting connection with risk sensitive control (see Whittle [50]) is presented in Katehakis and Rothblum [23] and in Denardo, Park and Rothblum [8]. See also Denardo, Feinberg and Rothblum [9].

5.2.1 Induction on the number of states

A short proof by Tsitsiklis [40] is based on induction on the number of states, hence this proof covers only the finite state space case. It is immediate to see that the state with highest Gittins index is the state $1 \in \arg\max\{R(i) : i = 1, \ldots, L\}$. It is now shown that it is optimal always to activate arm in state 1, if such an arm is available — this is done by a pairwise interchange argument, similar to that employed in the general interchange proof of Section 3.1. Next one observes that we can add periods of activating state 1 as part of the stopping times for states $2, \ldots, L$, to obtain a new (semi-Markov) multiarmed bandit problem with $L-1$ states, whose indices are the same as in the original problem. This provides the induction step.

5.2.2 Epsilon optimality

Ketehakis and Veinott [22] use essentially the interchange argument of 3.1 to prove that if at all times one chooses to activate arms with index which is no more than ϵ away from the highest index, then the total objective is no more than $\frac{\epsilon}{1-\alpha}$ away from the optimum. To evaluate the sub-optimality they use the following property of dynamic programming: the optimal policy and the optimal value function of a discounted problem with infinite time horizon is identical to that of an undiscounted problem with a finite random time horizon, where the finite time horizon is given by a properly chosen stopping time.

5.2.3 Restart at i

In the same paper Katehakis and Veinott [22] also point out another interpretation of the Gittins index, in addition to those listed in 2.3. See also [21]

Restart at i problem (Katehakis–Veinott): Assume that you have a single arm Z, and at any time t you can choose to activate the arm in its current state, or to restart with the arm in state i. The optimality equations for this problem are:

$$V^i(k) = \max\left\{ R(k) + \alpha \sum_j p(k,j) V^i(j), R(i) + \alpha \sum_j p(i,j) V^i(j) \right\} \tag{5.3}$$

We note that $V^i(i)$ is the retirement reward under which in state i one is indifferent between activation and retirement, and in fact $V^i(i) = \frac{\nu(i)}{1-\alpha}$.

5.3 Extensions of Gittins index optimality

Many extensions of the Gittins index and of the Gittins index optimality theorem are possible. These include the following:

- non-discounted positive dynamic programming, negative dynamic programming, and average cost dynamic programming extensions
- Semi Markov decision moments and continuous time bandit problems [20].
- Arm acquiring and branching bandit problems [48, 46].
- Tax problem formulation [42].

All four proof methods can be adapted to these problems.

5.4 Beyond Gittins optimality

The Gittins index policy fails to be optimal if more than one arm is activated at each decision moment. It also fails if the remaining arms are not frozen. Two more recent developments address these problems.

When the choice is to activate M of the N arms, while the rest of the arms remain passive and their states are frozen, we have the model of *parallel bandits*. The Gittins index policy may be approximately optimal for this model. This is shown via approximate conservation laws which extend the polymatroid structure of the achievable region proof [7, 16, 17].

When the choice is to activate M of the N arms, with the remaining arms passive, but passive arms are not frozen, so that we have an active reward $R_n(i)$, and active as well as passive transition probabilities, given by $p_n^a(i,j)$ and $p_n^p(i,j)$ respectively, we have the model of *restless bandits*. Here an idea that extends Whittle's retirement proof leads to the Whittle index: Assume there is a subsidy γ paid for each time that an arm is passive, and solve the single arm dynamic programming average cost problem. This will partition all the states of the arm into a passive subset of states $S_p(\gamma) \subseteq E$, and its complement of active states. If $S_p(\gamma)$ is monotonically increasing with γ the restless bandit is indexable, and the value of $\gamma(i)$ at which state i is indifferent between passive or active is the Whittle index of state i. Using the Whittle index priority policy for the choice of the M active arms while not optimal, may under certain circumstances be close to optimal [49, 44, 45]. Indexability has been further investigated by Niño-Mora [28, 29, 30].

References

[1] Anderson, E. and Nash, P. (1987) *Linear programming in infinite dimensional spaces. Theory and application* Wiley-Interscience, Chichester.

[2] Barvinok, A. (2002) *A Course in Convexity.* AMS Graduate Studies in Mathematics, Vol 54.

[3] Battacharya, P., Georgiadis, L, and Tsoucas, P. (1992). Extended polymatroids, properties and optimization. In E. Balas, G. Cornnéjols and R. Kannan, edrs. *Integer Programming and Combinatorial Optimization, IPCO2.* Carnegie-Mellon University, pp 298-315.

[4] Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhia* **16**, 221–229.

[5] Bertsimas, D. and Niño-Mora, J. (1996). Conservation laws, extended polymatroids and multi-armed bandit problems. *Mathematics of Operations Research* **21,** 257–306.

[6] Chakravorty, J., Mahajan, A. (2013) Multi-armed bandits, Gittins index, and its calculation. `http://www.ece.mcgill.ca/ amahaj1/projects/bandits/book/2013-bandit -computations.pdf`

[7] Dacre, M., Glazebrook, K. and Niño-Mora, J. (1999). The achievable region approach to the optimal control of stochastic systems. With discussion. *Journal of the Royal Statistical Society, Series B, Methodological*, 61:747-791.

[8] Denardo, E. V., Park, H., Rothblum, U. G. (2007). Risk-sensitive and risk-neutral multi-armed bandits. *Mathematics of Operations Research*, **32**(2), 374-394.

[9] Denardo, E. V., Feinberg, E. A., Rothblum, U. G. (2013). The multi-armed bandit, with constraints. *Annals of Operations Research*, **208**:37–62, (Volume 1 of this publication).

[10] Edmonds, J. (1970). Submodular functions, matroids and certain polyhedra. in *Proceedings of Calgary International Conference on Combinatorial Structures and their Applications*, R. Guy, H. Hanani, N. Sauer, and J. Schönheim, eds., Gordon and Breach, New York, pp 69–87.

[11] Federgruen, A. and Groenevelt, H. (1988) Characterization and optimization of achievable performances in general queueing systems *Operations Research* 36:733–741.

[12] Gittins, J.C. and Jones, D.M. (1974). A dynamic allocation indices for the sequential design of experiments. In J. Gani, K. Sarkadi and I. Vince (eds.) *Progress in Statistics, European Meeting of Statisticians 1972, Vol 1* Amsterdam: North Holland, pp 241–266.

[13] Gittins, J.C. (1979). Bandit Processes and Dynamic Allocation Indices. *J Royal Statistical Society Series B* **14**, 148 – 167.

[14] Gittins, J.C. (1989). *Multiarmed Bandits Allocation Indices.* Wiley, New York.

[15] Gittins, J.C., Glazebrook, K., Weber, R.R (2011) *Multiarmed Bandits Allocation Indices. 2nd Edition* Wiley, New York.

[16] Glazebrook, K. D., R. Garbe. (1999). Almost optimal policies for stochastic systems which almost satisfy conservation laws. *Annals of Operations Research* 92:1943.

[17] Glazebrook, K. and Niño-Mora, J. (2001). Parallel scheduling of multiclass M/M/m queues: approximate and heavy-traffic optimization of achievable performance. *Operations Research* 49:609-623.

[18] Harrison, J.M. (1975). Dynamic scheduling of a multiclass queue, discount optimality. *Operations Research* **23**, 270–282.

[19] Ishikada A, T. and Varaiya, P. (1994). Multi-Armed Bandit Problem Revisited. *J. of optimization theory and applications* **83,** 113-154.

[20] Kaspi, H. and Mandelbaum A. (1998) Multi-armed bandits in discrete and continuous time *Annals of Applied Probability* **8**, 1270–1290.

[21] Katehakis M. N. and C. Derman (1986). Computing optimal sequential allocation rules in clinical trials. *Adaptive Statistical Procedures and Related Topics (J. Van Ryzin ed.) I.M.S. Lecture Notes-Monograph Series*, **8**: 29–39

[22] Katehakis, M.N. and Veinott, A.F. (1987). The multi-armed bandit problem: decomposition and computation *Mathematics of Operations Research* **12**, 262–268.

[23] Katehakis M. N. and U. Rothblum (1996). Finite state multi-armed bandit sensitive–discount, average-reward and average-overtaking optimality. *Annals of Applied Probability*, **6**(3):1024–1034.

[24] Klimov, G.P. (1974). Time sharing service systems I. *Theory of Probability and Applications* **19,** 532–551.

[25] Meilijson, I. and Weiss, G. (1977). Multiple feedback at a single server station. *Stochastic Processes and their Applications* **5**, 195–205.

[26] Mandelbaum, A. 1986. Discrete Multi-Armed Bandits and Multiparameter Processes. *Probability Theory and Related Fields* **71,** 129-147.

[27] Mitten, L. G. (1960). An analytic solution to the least cost testing sequence problem. *Journal of Industrial Engineering*, **11**(1), 17.

[28] Niño-Mora, J. 2001. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33:76–98.

[29] Niño-Mora, J. 2002. Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. *Mathematical Programming, Series A* 93:361–413.

[30] Niño-Mora, J. 2006. Restless bandit marginal productivity indices, diminishing returns and optimal control of make-to-order/make-to-stock M/G/1 queues. *Mathematics of Operations Research* 31:50–84.

[31] Niño-Mora, J. (2006). A (2/3)n3 fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov chain. Accepted in *INFORMS Journal on Computing* To appear

[32] Ross, S.M. (1983). *Introduction to Stochastic Dynamic Programming.* Academic Press, New York.

[33] Royden, H. L. (1971). *Real Analysis* Macmillan, New York.

[34] Rudin, W. (1964). *Principles of Mathematical Analysis* Mcgraw-hill, New York.

[35] Sevcik, K.C. (1974). Scheduling for minimum total loss using service time distributions. *J. of the Association for Computing Machinery* **21** 66–75.

[36] Shanthikumar, J.G. and Yao, D.D. (1992) Multiclass queueing systems: polymatroidal structure and optimal scheduling control *Operations Research* 40:S293–299.

[37] Shapiro, A. (2001) On duality theory of conic linear problems. Chapter 7 in *Semi-Infinite Programming*, Editors: M.A. Goberna and M.A. Lopez, Kluwer, Netherlands, 135–165.

[38] Sonin, I. M. (2008). A generalized Gittins index for a Markov chain and its recursive calculation. *Statistics and Probability Letters*, **78**(12), 1526-1533.

[39] Tcha, D. and Pliska, S.R. 1975. Optimal control of single server queueing networks and multi-class M/G/1 queues with feedback. *Operations Research* **25,** 248-258.

[40] Tsitsiklis, J.N. (1994) A short proof of the Gittins index theorem *Annals of Applied Probability* **4**, 194–199

[41] Tsoucas, P. (1991) The region of achievable performance in a model of Klimov. Research Report RC16543, IBM T.J. Watson Research Center, Yorktown Heights, New York.

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

[43] Weber, R. R. (1992) On the Gittins index for multiarmed bandits. *Annals of Probability* **2,**1024-1033.

[44] Weber, R. R, Weiss, G. 1990. On an index policy for restless bandits. in *J Applied Probability* 27:637–648.

[45] Weber, R. R, Weiss, G. 1991. Addendum to 'on an index policy for restless bandits'. in *Advances Applied Probability* 23:429–430.

[46] Weiss G 1988. Branching bandit processes *Probability Engineering Informational Sciences* **2**, 269–278.

[47] Whittle, P. 1980. Multi-armed bandits and the Gittins index. *J. Royal Statistical Society Series B* **42,** 143–149.

[48] Whittle, P. (1981). Arm acquiring bandits *Annals of Probability* **9,** 284–292.

[49] Whittle, P. (1988). Restless bandits: activity allocation in a changing world. in *A Celebration of Applied Probability* ed J. Gani, *J Applied Probability* 25A:287–298.

[50] Whittle, P. (1990). *Risk-sensitive optimal control.* New York: Wiley.

A Proofs of index properties

In this appendix we present the proofs of some results postponed in the paper.

Proof of Theorem 2.1. The direct proof is quite straightforward. Note that in (2.1) the value of $\nu(i,\sigma)$ for each stopping time σ is a ratio of sums over consecutive times. Hence to compate these we use the simple ineaulity:

$$\frac{a}{c} < \frac{a+b}{c+d} \Longleftrightarrow \frac{a+b}{c+d} < \frac{b}{d} \Longleftrightarrow \frac{a}{c} < \frac{b}{d} \tag{1.1}$$

We show that if we start from i and wish to maximize $\nu(i,\sigma)$ it is never optimal to stop in state j if $\nu(j) > \nu(i)$ or to continue in state j if $\nu(j) < \nu(i)$, which allows us to consider only stopping time of the form (2.3). We then assume that no stopping time achieves the supremum, and we construct an increasing sequence of stopping times with increasing ratio, which converges to $\tau(i)$ which therefore must achieve the supremum, and from this contradiction we deduce that the supremum is achieved. Finally we show that since the supremum is achieved, it is acieved by $\tau(i)$ as well as by all stopping times of the form (2.3).

Step 1: Any stopping time which stops while the ratio is $> \nu(Z(0))$ does not achieve the supremum. Assume that $Z(0) = i$, fix j such that $\nu(j) > \nu(i)$, and consider a stopping time σ such that:

$$\mathbb{P}(Z(\sigma) = j | Z(0) = i) > 0. \tag{1.2}$$

By the definition (2.1) there exists a stopping time σ' such that $\nu(j,\sigma') > \frac{\nu(j)+\nu(i)}{2}$. Define $\sigma' = 0$ for all initial values $\neq j$. Then:

$$\nu(i,\sigma+\sigma') = \frac{\mathbb{E}\{\sum_{t=0}^{\sigma-1}\alpha^t R(Z(t))|Z(0)=i\}+\mathbb{E}\{\sum_{t=\sigma}^{\sigma+\sigma'-1}\alpha^t R(Z(t))|Z(0)=i\}}{\mathbb{E}\{\sum_{t=0}^{\sigma-1}\alpha^t|Z(0)=i\}+\mathbb{E}\{\sum_{t=\sigma}^{\sigma+\sigma'-1}\alpha^t|Z(0)=i\}} > \nu(i,\sigma),$$

by (1.1), (1.2).

Step 2: Any stopping time which continues when the ratio is $< \nu(Z(0))$ does not achieve the supremum. Assume that $Z(0) = i$, fix j such that $\nu(j) < \nu(i)$, and let $\sigma' = \min\{t : Z(t) = j\}$. Consider any stopping time σ which does not always stop when it reaches state j, and assume that:

$$\nu(i,\sigma) > \nu(j) \text{ and } \mathbb{P}(\sigma > \sigma' | Z(0) = i) > 0. \tag{1.3}$$

Then:

$$\nu(i,\sigma) = \frac{\mathbb{E}\{\sum_{t=0}^{\min(\sigma,\sigma')-1}\alpha^t R(Z(t))|Z(0)=i\}+\mathbb{E}\{\sum_{t=\sigma'}^{\sigma-1}\alpha^t R(Z(t))|Z(\sigma')=j\}}{\mathbb{E}\{\sum_{t=0}^{\min(\sigma,\sigma')-1}\alpha^t|Z(0)=i\}+\mathbb{E}\{\sum_{t=\sigma'}^{\sigma-1}\alpha^t|Z(\sigma')=j\}} < \nu(i,\min(\sigma,\sigma')),$$

by (1.1), (1.3).

Steps 1,2 show that the supremum can be taken over stopping times $\sigma > 0$ which satisfy (2.3), and we restrict attention to such stopping times only.

Step 3: The supremum is achieved. If $\tau(i)$ is the unique stopping time which satisfies (2.3) then it achieves the supremum and there is nothing more to prove. Assume that the supremum is not achieved. We now consider a fixed stopping time $\sigma > 0$ which satisfies (2.3) and:

$$\mathbb{P}(\sigma < \tau(i)|Z(0) = i) > 0, \quad \nu(i, \sigma) = \nu_0 < \nu(i). \tag{1.4}$$

This is possible, since τ is not unique, and since the supremum is not achieved. Assume that σ stops at a time $< \tau(i)$ when the state is $Z(\sigma) = j$. By (2.3), $\nu(j) = \nu(i)$. We can then find σ' such that $\nu(j, \sigma') \geq \frac{\nu_0 + \nu(i)}{2}$. Define σ' accordingly for the value of $Z(\sigma)$ whenever $\sigma < \tau(i)$, and let $\sigma' = 0$ if $\sigma = \tau(i)$. Let $\sigma_1 = \sigma + \sigma'$. Clearly we have (repeat the argument of step 1):

$$\sigma \leq \sigma_1 \leq \tau(i), \quad \nu(i, \sigma) < \nu(i, \sigma_1) = \nu_1 < \nu(i) \tag{1.5}$$

We can now construct a sequence of stopping times, with

$$\sigma_{n-1} \leq \sigma_n \leq \tau(i), \quad \nu(i, \sigma_{n-1}) < \nu(i, \sigma_n) = \nu_n < \nu(i) \tag{1.6}$$

which will continue indefinitely, or will reach $\mathbb{P}(\sigma_{n_0} = \tau(i)) = 1$, in which case we define $\sigma_n = \tau(i), n > n_0$.

It is easy to see that $\min(n, \sigma_n) = \min(n, \tau(i))$, hence $\sigma_n \nearrow \tau(i)$ a.s. It is then easy to see (use dominated or monotone convergence) that $\nu(i, \sigma_n) \nearrow \nu(i, \tau(i))$. But this implies that $\nu(i, \sigma) < \nu(i, \tau(i))$. Hence the assumption that the supremum is not achieved implies that the supremum is achieved by $\tau(i)$, which is a contradiction. Hence, for any initial state $Z(0) = i$ the supremum is achieved by some stopping time, which satisfies (2.3).

Step 4: The supremum is achieved by $\tau(i)$. Start from $Z(0) = i$, and assume that a stopping time σ satisfies (2.3) and achieves the supremum. Assume

$$\mathbb{P}(\sigma < \tau(i)|Z(0) = i) > 0, \quad \nu(i, \sigma) = \nu(i) \tag{1.7}$$

and take the event that σ stops at a time $< \tau(i)$ when the state is $Z(\sigma) = j$. By (2.3) $\nu(j) = \nu(i)$. We can then find σ' which achieves the supremum, $\nu(j, \sigma') = \nu(j) = \nu(i)$. Define σ' accordingly for the value of $Z(\sigma)$ whenever $\sigma < \tau(i)$, and let $\sigma' = 0$ if $\sigma = \tau(i)$. Let $\sigma_1 = \sigma + \sigma'$. Clearly we have:

$$\sigma \leq \sigma_1 \leq \tau(i), \quad \nu(i, \sigma) = \nu(i, \sigma_1) = \nu(i) \tag{1.8}$$

We can now construct an increasing sequence of stopping times, $\sigma_n \nearrow \tau(i)$ a.s., and all achieving $\nu(i, \sigma_n) = \nu(i)$. Hence (again use dominated or monotone convergence) $\nu(i, \tau(i)) = \nu(i)$.

Step 5: The supremum is achieved by any stopping time which satisfies (2.3). Let σ satisfy (2.3). Whenever $\sigma < \tau(i)$ and $Z(\sigma) = j$, we will have $\tau(i) - \sigma = \tau(j)$, and $\nu(j, \tau(i) - \sigma) = \nu(i)$. Hence:

$$\nu(i) = \nu(i, \tau(i)) = \frac{\mathbb{E}\{\sum_{t=0}^{\sigma-1} \alpha^t R(Z(t))|Z(0)=i\} + \mathbb{E}\{\sum_{t=\sigma}^{\tau(i)-1} \alpha^t R(Z(t))|Z(0)=i\}}{\mathbb{E}\{\sum_{t=0}^{\sigma-1} \alpha^t|Z(0)=i\} + \mathbb{E}\{\sum_{t=\sigma}^{\tau(i)-1} \alpha^t|Z(0)=i\}} = \nu(i, \sigma).$$

This completes the proof. ■

Proof of Proposition 2.2. *step 1: We show that* $\nu(i) \leq \gamma(i)$. Consider any $y < \nu(i)$, let $M = \frac{y}{1-\alpha}$. By definition (2.1) there exists a stopping time τ for which $\nu(i,\tau) > y$.

Hence, a policy π which from state i will play up to time τ and then stop and collect the reward M, will have:

$$\begin{aligned} V_r^{\pi}(i,M) &= \mathbb{E}\left\{\sum_{t=0}^{\tau-1}\alpha^t R(Z(t)) + \sum_{t=\tau}^{\infty}\alpha^t y | Z(0)=i\right\} \\ &> \mathbb{E}\left\{\sum_{t=0}^{\tau-1}\alpha^t y + \sum_{t=\tau}^{\infty}\alpha^t y | Z(0)=i\right\} = \frac{y}{1-\alpha} = M. \end{aligned}$$

Hence $V_r(i,M) > M$, and i belongs to the continuation set, for standard arm reward y, (or fixed charge y, or terminal reward M). Hence, $M(i) \geq M$, and $\gamma(i) \geq y$. But $y < \nu(i)$ was arbitrary. Hence, $\gamma(i) \geq \nu(i)$.

step 2: We show that $\nu(i) \geq \gamma(i)$. Consider any $y < \gamma(i)$. Let $M = \frac{y}{1-\alpha}$, and consider $\tau(i,M)$ and $V_r(i,M)$. Writing (2.14), and using the fact that for $M < M(i)$ we have $i \in C_M$ and $V_r(i,M) > M$:

$$\begin{aligned} V_r(i,M) &= \mathbb{E}\left\{\sum_{t=0}^{\tau(i,M)-1}\alpha^t R(Z(t)) + \sum_{t=\tau(i,M)}^{\infty}\alpha^t y | Z(0)=i\right\} \\ &> \frac{y}{1-\alpha}. \end{aligned}$$

But this means that $\nu(i,\tau(i,M)) > y$. Hence, $\nu(i) > y$. But $y < \gamma(i)$ was arbitrary. Hence, $\nu(i) \geq \gamma(i)$.

step 3: Identification of $\tau(i,M(i)-)$ *as achieving the supremum in (2.1).* Clearly, starting from state i, $\tau(i,M(i)-)$ will continue for the continuation set of $C_{M(i)-}$ which includes all j with $\gamma(j) \geq \gamma(i)$. But we have shown that $\gamma(i) = \nu(i)$, hence clearly $\tau(i,M(i)-)$ is identical to $\tau(i)$ as defined in (2.2). ■

Proof of Proposition 2.9. The equivalence of the two algorithms is easily seen. Step 1 is identical. Assume that the two algorithms are the same for steps $1,\ldots,k-1$. We then have in step k, for any $i \in S^-_{\varphi(k-1)}$ that:

$$\frac{R(i) - \sum_{j=1}^{k-1} A_i^{S_{\varphi(j)}} y^{S_{\varphi(j)}}}{A_i^{S^-_{\varphi(k-1)}}} =$$

$$\frac{R(i) - A_i^{S_{\varphi(1)}}\nu(\varphi(1)) - \sum_{j=2}^{k-1} A_i^{S_{\varphi(j)}}(\nu(\varphi(j)) - \nu(\varphi(j-1)))}{A_i^{S^-_{\varphi(k-1)}}} =$$

$$\frac{R(i)-\sum_{j=1}^{k-1}A_i^{S_{\varphi(j)}}\nu(\varphi(j))+\sum_{j=1}^{k-2}A_i^{S_{\varphi(j+1)}}\nu(\varphi(j))}{A_i^{S^-_{\varphi(k-1)}}}=$$

$$\frac{R(i)+\sum_{j=1}^{k-1}(A_i^{S^-_{\varphi(j)}}-A_i^{S_{\varphi(j)}})\nu(\varphi(j))-A_i^{S^-_{\varphi(k-1)}}\nu(\varphi(k-1))}{A_i^{S^-_{\varphi(k-1)}}}=$$

$$\frac{R(i)+\sum_{j=1}^{k-1}(A_i^{S^-_{\varphi(j)}}-A_i^{S_{\varphi(j)}})\nu(\varphi(j))}{A_i^{S^-_{\varphi(k-1)}}}-\nu(\varphi(k-1))$$

and so the supremum in step k is achieved by the same $\varphi(k)$ in both versions of the algorithm. The quantities y^S appear in the 4th proof, Section 3.4. ■

Proof of Proposition 4.2. **(i)** The definition of $Ax(S)$ implies that $Ax(\emptyset)=0$ because it is an empty sum. Trivially $T_j^{\emptyset}=\infty$, so $\alpha^{T_j^{\emptyset}}=0$, hence the definition (3.20) of $b(S)$ implies $b(\emptyset)=0$. Thus (i) holds.

(ii) First we show that if $x\in X$, and $y(S)=Ax(S)$, then g_y is of bounded variation. Let $x^+=\{x_i^+=\max(x_i,0)\}_{i\in E}$, and $x^-=\{x_i^-=\max(-x_i,0)\}_{i\in E}$. Clearly $x^+,x^-\in X$ and $g_y(v)=Ax^+(S(v))-Ax^-(S(v))$. We have:

$$\begin{aligned} Ax^+(S(v)) &= \sum_{j\in S(v)} A_j^{S(v)}x_j^+ \\ &= \sum_{j\in E} A_j^{S(v)}x_j^+ - \sum_{j\in E\setminus S(v)} A_j^{S(v)}x_j^+ \end{aligned} \tag{1.9}$$

$S(v)$ is increasing in v, as a result $A_j^{S(v)}$ is decreasing in v. Also $E\setminus S(v)$ is decreasing in v. Hence both terms in (1.9) decrease in v. Thus $Ax^+(S(v))$, as the difference of two decreasing functions of v is of bounded variation. Similarly $Ax^-(S(v))$, is of bounded variation. Hence, g_y is of bounded variation.

Next we take $y=\mathbf{b}$, and consider g_y. $g_y(v)=b(S(v))$ increases in v thus it is of bounded variation. This proves (ii).

(iii) To calculate the limits from the left we need to prove some continuity results: consider an increasing sequence $\{v_n\}$, such that $\lim_{n\to\infty}v_n=v$, and $v_n<v$. Consider first $S(v_n)$ and $S^-(v)$. Then $S(v_n)$ are increasing and $\lim_{n\to\infty}S(v_n)=\bigcup_{n=0}^{\infty}S(v_n)=S^-(v)$. To see this, note the if $i\in S^-(v)$ then $\nu(i)<v$, hence for some n_0 we have $\nu(i)\le v_n$ for all $n\ge n_0$, hence $i\in S(v_n), n\ge n_0$.

Next consider $T_j^{S(v_n)}$ and $T_j^{S^-(v)}$ for some given sample path. If $T_j^{S^-(v)}=\infty$ then $T_j^{S(v_n)}=\infty$ for all n. If $T_j^{S^-(v)}=t<\infty$ then we have $Z(t)=i\in S^-(v)$, but in that case $i\in S(v_n), n>n_0$, so $T_j^{S(v_n)}=t$ for all $N\ge n_0$. Thus we have that $T_j^{S(v_n)}$ is non-increasing in n and converges to $T_j^{S^-(v)}$ for this sample path. Hence $T_j^{S(v_n)}\searrow_{a.s.}T_j^{S^-(v)}$.

We now have that $\sum_{t=0}^{T_j^{S(v_n)}-1} \alpha^t \searrow_{a.s.} \sum_{t=0}^{T_j^{S^-(v)}-1} \alpha^t$, and because $\sum_t \alpha^t$ are uniformly bounded, $A_j^{S(v_n)} \searrow A_j^{S^-(v)}$.

Consider now $Ax(S(v_n))$ and $Ax(S^-(v))$. We need to show that as $n \to \infty$,

$$\sum_{j\in S(v_n)} A_j^{S(v_n)} x_j - \sum_{j\in S^-(v)} A_j^{S^-(v)} x_j = \sum_{j\in S(v_n)} \left(A_j^{S(v_n)} - A_j^{S^-(v)}\right) x_j - \sum_{j\in S^-(v)\setminus S(v_n)} A_j^{S^-(v)} x_j \to 0$$

which follows from A_i^S bounded by $\frac{1}{1-\alpha}$, x_i absolutely convergent, and $S(v_n) \nearrow S^-(v)$. To explain this a little further: Since x_j are absolutely convergent, for every $\epsilon > 0$ we can find a finite subset of states E_0 such that $\frac{2}{1-\alpha}\sum_{j\in E\setminus E_0} |x_j| < \frac{1}{2}\epsilon$. If we now examine the sums only over $j \in E_0$, clearly the first sum can be made arbitrarily small as $n \to \infty$, and the second sum becomes empty as $n \to \infty$.

Finally for any given $\mathbf{Z}(0)$, $\mathbb{E}(\alpha^{T_{Z_k(0)}^{S(v_n)}}) \nearrow \mathbb{E}(\alpha^{T_{Z_k(0)}^{S^-(v)}}), k = 1, \ldots, N$ and hence by definition (3.20) $b(S(v_n)) \nearrow b(S^-(v))$. This completes the proof that $g_y(v_n) \to y(S^-(v))$.

To show continuity from the right, consider a decreasing sequence $\{v_n\}$, such that $\lim_{n\to\infty} v_n = v$, and $v_n > v$. Consider the sequence of sets $S(v_n)$ and the set $S(v)$. Then $S(v_n)$ are decreasing and $\lim_{n\to\infty} S(v_n) = \bigcap_{n=0}^{\infty} S(v_n) = S(v)$. To see this, note the if $i \notin S(v)$ then $\nu(i) > v$, hence for some n_0 we have $\nu(i) > v_n$ for all $n \geq n_0$, hence $i \notin S(v_n), n \geq n_0$.

The remaining steps of the proof are as for the limit from the left: one shows that $T_j^{S(v_n)} \nearrow_{a.s.} T_j^{S(v)}$, and so on. This completes the proof of (iii)

(iv) We wish to show

$$y(S(v)) - y(S^-(v)) = \sum_{i:\nu(i)=v} y(S_i) - y(S_i^-),$$

Clearly, if $\{i : \nu(i) = v\} = \emptyset$ then both sides are 0, and if $\{i : \nu(i) = v\}$ consists of a single state $\{i\}$ then $S(v) = S_i, S^-(v) = S_i^-$ and there is nothing to prove. If $\{i : \nu(i) = v\}$ consists of a finite set of states with $i_1 \prec i_2 \ldots \prec i_M$, then $S(i_k) = S^-(i_{k+1})$ and the summation over i_k is a collapsing sum. If $\{i : \nu(i) = v\}$ is infinite countable but well ordered, then we can order $\{i : \nu(i) = v\}$ as $i_1 \prec i_2 \prec \ldots$ (ordinal type ω), and the infinite sum on the right is a collapsing sum, which converges to the right hand side.

The difficulty here is that in the general case $\{i : \nu(i) = v\}$ may not be well ordered by $\prec$, and it is this general case which we wish to prove. We do so by a sample path argument, using the fact that the sequence of activation times, $t = 1, 2, \ldots$ is well ordered. In our sample path argument we make use of the sequence of stopping times and states $\mathcal{T}_\ell$ and k_ℓ, defined in (2.21). Recall that this is the sequence of stopping times and states at which the index sample path 'loses priority height', in that at time $\mathcal{T}_\ell$ it reaches a state k_ℓ which is of lower priority than all the states encountered in $0 < t < \mathcal{T}_\ell$.

Fix a value of v, and consider a sample path starting from $Z(0) = j$. Then for this sample path we will have integers $0 \leq \underline{L} < \overline{L} \leq \infty$ such that:

$$\nu(k_\ell) = \begin{cases} > v & \ell \leq \underline{L} \\ = v & \underline{L} < \ell < \overline{L} \\ < v & \ell \geq \overline{L} \end{cases}$$

It is possible that $\underline{L} = \infty$ because the sample path never reaches $S(v)$. It is also possible that $\overline{L} = \underline{L} + 1$, either because $\{i : \nu(i) = v\} = \emptyset$, or because the first visit of the sample path to $S(v)$ is directly to a state with index $< v$; in each of these cases $T_j^{S(v)} = T_j^{S^-(v)}$. Otherwise, if $\overline{L} - \underline{L} > 1$, then $\mathcal{T}_{\underline{L}+1} = T_j^{S_{k_{\underline{L}+1}}} = T_j^{S(v)}$ will be the first visit of the process in $S(v)$, and $\mathcal{T}_{\overline{L}} = T_j^{S^-_{k_{\overline{L}-1}}} = T_j^{S^-(v)}$ will be the first visit of the process in $S^-(v)$.

We can then write:

$$\sum_{t=0}^{T_j^{S^-(v)}-1} \alpha^t - \sum_{t=0}^{T_j^{S(v)}-1} \alpha^t = \sum_{\underline{L}<\ell<\overline{L}} \left(\sum_{t=0}^{T_j^{S^-_{k_\ell}}-1} \alpha^t - \sum_{t=0}^{T_j^{S_{k_\ell}}-1} \alpha^t \right) = \sum_{i:\nu(i)=v} \left(\sum_{t=0}^{T_j^{S_i^-}-1} \alpha^t - \sum_{t=0}^{T_j^{S_i}-1} \alpha^t \right) \tag{1.10}$$

where the second equality follows from $T_j^{S_i^-} = T_j^{S_i}$ for all $i : \nu(i) = v$ except for $k_\ell, \underline{L} < \ell < \overline{L}$.

By taking ecpectations we now get that:

$$A_j^{S^-(v)} - A_j^{S(v)} = \sum_{i:\nu(i)=v} A_j^{S_i^-} - A_j^{S_i}$$

and also, for all $j \in S(v) \backslash S^-(v)$:

$$A_j^{S_j^-} - A_j^{S(v)} = \sum_{i:\nu(i)=v, i \succeq j} \left(A_j^{S_i^-} - A_j^{S_i} \right).$$

Next we note that

$$b(S) = \mathbb{E}(\frac{\alpha^{T^S_{\mathbf{Z}(0)}}}{1-\alpha}) = \mathbb{E}(\sum_{t=T^S_{\mathbf{Z}(0)}}^{\infty} \alpha^t)$$

where $T^S_{\mathbf{Z}(0)} = \sum_{n=1}^N T^S_{Z_n(0)}$. Hence, using the same sample path result (1.10) for each of $Z_n(0)$ we get

$$b(S(v)) - b(S^-(v)) = \mathbb{E}(\sum_{t=T^{S(v)}_{\mathbf{Z}(0)}}^{T^{S^-(v)}_{\mathbf{Z}(0)}-1} \alpha^t) = \mathbb{E}(\sum_{i:\nu(i)=v} \sum_{t=T^{S_i}_{\mathbf{Z}(0)}}^{T^{S_i^-}_{\mathbf{Z}(0)}-1} \alpha^t) = \sum_{i:\nu(i)=v} \big(b(S_i) - b(S_i^-)\big)$$

We now come to evaluate $Ax(S^-(v)) - Ax(S(v))$. We need to show that for all $x \in X$:

$$\begin{aligned} Ax(S^-(v)) - Ax(S(v)) &= \sum_{j \in S^-(v)} x_j A_j^{S^-(v)} - \sum_{j \in S(v)} x_j A_j^{S(v)} \\ &= \sum_{i:\nu(i)=v} \left(\sum_{j \in S_i^-} x_j A_j^{S_i^-} - \sum_{j \in S_i} x_j A_j^{S_i} \right) = \sum_{i:\nu(i)=v} \big(Ax(S_i^-) - Ax(S_i)\big) \end{aligned} \tag{1.11}$$

Since this has to hold for all x, we need to show for every $j \in E$ that the coefficients of x_j satisfy the equalities individually.

For $j \in S^-(v)$ clearly $j \in S(v), S_i, S_i^-, i : \nu(i) = v$, we need to check that:

$$A_j^{S^-(v)} - A_j^{S(v)} = \sum_{i:\nu(i)=v} \left(A_j^{S_i^-} - A_j^{S_i} \right)$$

which we have just shown.

For $j \in S(v) \backslash S^-(v)$ we have: $j \in S(v)$, and for $i : \nu(i) = v$: $j \in S_i$ if $i \succeq j$, and $j \in S_i^-$ if $i \succ j$. Hence, the coefficients of j which we need to compare are:

$$-A_j^{S(v)} = \sum_{i:\nu(i)=v, i \succ j} A_j^{S_i^-} - \sum_{i:\nu(i)=v, i \succeq j} A_j^{S_i}$$

Which amounts to:

$$A_j^{S_j^-} - A_j^{S(v)} = \sum_{i:\nu(i)=v, i \succeq j} \left(A_j^{S_i^-} - A_j^{S_i} \right)$$

which we have also shown. ■