Applied Probability Trust (November 10, 1999)

FOUR PROOFS OF GITTINS' MULTIARMED BANDIT THEOREM

ESTHER FROSTIG,* *The University of Haifa*

GIDEON WEISS,** *The University of Haifa*

Abstract

We survey 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 a proof based on generalized conservation laws and LP duality.

DYNAMIC PROGRAMMING; BANDIT PROBLEMS; GITTINS INDEX

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 90B36;62L15;90C40
SECONDARY 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 $\tilde{R}(t) = R_{n(t)}(Z_{n(t)}(t)) = R_n(i)$.

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

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

* Postal address: Department of Statistics, The University of Haifa, Mount Carmel, 31905, Israel, e-mail *frostig@stat.haifa.ac.il*.

* Postal address: Department of Statistics, The University of Haifa, Mount Carmel, 31905, Israel, e-mail *gweiss@stat.haifa.ac.il*.

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

This problem, introduced by Bellman [2] 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 [13] 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 (Gittins, 1976) There exist functions, $G_n(Z_n(t)), n = 1, \ldots, N$ such that for any state $\mathbf{Z}(t)$ the policy π^* which will activate a bandit process (arm) $n(t) = n$ which satisfies $G_n(Z_n(t)) = \max_{1 \le m \le N} G_m(Z_m(t))$ 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 the past 25 years, and have motivated much further research. In this paper we give 4 different proofs of Gittins' result. Our purpose is twofold: We feel that having these proofs together will be useful to the general reader, in particular in view of the fact 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. 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, from the hindsight advantage of twenty five years, to present the proofs in a 'pure' form, that is use complete self contained arguments for each of the proofs, and highlight the differences between them.

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.

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 a proof based on a more recent approach, of achievable regions and generalized conservation laws, as proposed by Tsoucas, Bertsimas and Niño-Mora, and others (Section 3.4).

We will for ease of notation and without loss of generality 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 a noncommunicating class of states for each of the bandits.

bibliographic note: The Gittins index idea was put forward by Gittins as early as 1972 [5, 6, 7]. It was also indicated in several other papers of the time, notably in Klimov's paper [10], and also in Harrison [8], Sevcik [14] Tcha and Pliska [15] and Meilijson and Weiss [11]. This is a survey paper and our original contribution here is modest — throughout the paper the form of address 'we' is meant conversationally to suggest us and the reader.

2. Preliminary: Studying the single bandit

We start this section with the study of the Gittins index (Section 2.1), from first principles. Next we present 3 closely related formulations of single arm dynamic programming problems (Section 2.2). We then (in Section 2.3) derive properties of the solutions, as are needed for the later proofs.

2.1. *The Gittins index* Gittins defined his dynamic allocation index, now known as the Gittins index, as follows:

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

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 ($\leq \infty$):

$$\tau(i) = \min\{t : \nu(Z(t)) < \nu(i)\} \tag{2}$$

An important property of the Gittins index is that the supremum (1) is achieved, and in fact it is achieved by $\tau(i)$. The following is a 'first principles' proof of this fact — the proof is quite long, but it is instructive in that it is independent of all the ideas which led to the later proofs. We shall see in the next section that Theorem 2 can be derived more easily from the single arm dynamic programs, using the theory of dynamic programming.

Theorem 2 The supremum of (1) is achieved by (2). It is also achieved by any stopping time σ which satisfies:

$$\sigma \leq \tau \quad \textit{and} \quad \nu(Z(\sigma)) \leq \nu(i) \tag{3}$$

Proof. Recall the inequality: For $a, b, c, d > 0$,

$$\frac{a}{c} < \frac{a+b}{c+d} < \frac{b}{d} \Longleftrightarrow \frac{a}{c} < \frac{b}{d} \tag{4}$$

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:

$$\mathbb{P}(Z(\sigma)=j|Z(0)=i)>0 \tag{5}$$

By the definition (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:

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

by (4,5).

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 such that:

$$\nu(i,\sigma)>\nu(j) \quad \text{and} \quad \mathbb{P}(\sigma>\sigma'|Z(0)=i)>0 \tag{6}$$

(clearly if $\nu(i,\sigma)\leq\nu(j)$, then σ does not achieve the supremum, hence we need only consider $\nu(i,\sigma)>\nu(j)$). Then:

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

by (4,6).

Steps 1,2 show that the supremum can be taken over stopping times $\sigma>0$ which satisfy (3), and we restrict attention to such stopping times only.

Step 3: The supremum is achieved. Assume that the supremum is not achieved. Consider $\sigma>0$ which satisfies (3), with:

$$\mathbb{P}(\sigma<\tau(i)|Z(0)=i)>0, \quad \nu(i,\sigma)=\nu<\nu(i) \tag{7}$$

Assume that σ stops at a time $<\tau(i)$ when the state is $Z(\sigma)=j$. By (3), $\nu(j)=\nu(i)$. We can then find σ' such that $\nu(j,\sigma')\geq\frac{\nu+\nu(i)}{2}$. 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):

$$\sigma\leq\sigma_1\leq\tau(i), \quad \nu(i,\sigma)<\nu(i,\sigma_1)=\nu_1<\nu(i) \tag{8}$$

We can now construct a sequence of stopping times, with

$$\sigma_{n-1}\leq\sigma_n\leq\tau(i), \quad \nu(i,\sigma_{n-1})<\nu(i,\sigma_n)=\nu_n<\nu(i) \tag{9}$$

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 (3).

Step 4: The supremum is achieved by $\tau(i)$. Start from $Z(0) = i$, and assume that a stopping time σ satisfies (3) and achieves the supremum. Assume

$$\mathbb{P}(\sigma < \tau(i) | Z(0) = i) > 0, \quad \nu(i, \sigma) = \nu(i) \tag{10}$$

and take the event that σ stops at a time $< \tau(i)$ when the state is $Z(\sigma) = j$. By (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:

$$\sigma \leq \sigma_1 \leq \tau(i), \quad \nu(i, \sigma) = \nu(i, \sigma_1) = \nu(i) \tag{11}$$

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 (3). Let σ satisfy (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:

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

This completes the proof.

Theorem 2 is indicated by Gittins [6], but we are not aware of a complete proof which is readily available in the literature.

2.2. *Dynamic programming for a single arm* We now present 3 closely related dynamic programming problems, for a single bandit process. These were posed in some of the papers developing the various proofs.

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

$$V(i) = \max\left\{R(i) + \alpha \sum p(i, j) V(j), \gamma + \alpha V(i)\right\} \tag{12}$$

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 $t+1$. The optimality equations for this problem are:

$$W(i) = \max\left\{R(i) - \gamma + \alpha \sum p(i,j)W(j), \alpha W(i)\right\} \tag{13}$$

The fixed charge problem pays at every time t a reward smaller by γ than that of the standard arm problem. Hence the two problems have the same optimal policy, and $W(i) = V(i) - \frac{\gamma}{1-\alpha}$.

Clearly, 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.

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. The optimality equations for this problem are:

$$V(i,M) = \max\left\{R(i) + \alpha \sum p(i,j)V(j,M), M\right\} \tag{14}$$

We shall take $M = \frac{\gamma}{1-\alpha}$ and then the retirement problem has the same solution as the standard arm problem.

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

- *Optimal policies:* Let

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

then any policy which continues to activates the arm while in C_M, acts arbitrary 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:*

$$V(i,M) = \mathbb{E}\left\{\sum_{t=0}^{\tau(i,M)-1} \alpha^t R(Z(t)) + \alpha^{\tau(i,M)} M | Z(0) = i\right\}, \tag{15}$$

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 M, and as M increases C_M decreases, S_M increases, and $\tau(i,M)$ decreases. 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 $M > \frac{C}{1-\alpha}$.

2.3. *Properties of the single arm solutions* Define now:

$$M(i) = \sup\{M : i \in C_M\} = \inf\{M : V(i,M) = M\} \tag{16}$$

$$\gamma(i) = (1-\alpha)M(i) \tag{17}$$

2.3.1. *Properties related to the Gittins Index*

Lemma 1 The quantity $\gamma(i)$ equals the Gittins index,

$$\nu(i) = \gamma(i) \tag{18}$$

$$\tau(i) = \tau(i, M(i)-) \tag{19}$$

Proof. step 1: We show that $\nu(i) \leq \gamma(i)$. Consider any $y < \nu(i)$, let $M = \frac{y}{1-\alpha}$. By definition (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:

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

Hence $V(i,M) > M$, 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 $V(i,M)$. Writing (15), and using the fact that for $M < M(i)$ we have $i \in C_M$ and $V(i,M) > M$:

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

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 (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).

Note: by (15) it is clear that $\tau(i, M(i)-)$ achieves $\nu(i, \tau(i, M(i)-) = \gamma(i) = \nu(i)$, and therefore we have a free proof for the fact that the supremum in (1) is achieved, and we can forego the detailed proof of Theorem 2. The key point here is that instead of a first principles proof, we use the theory of dynamic programming to guarantee the existence and identity of $\tau(i, M(i)-)$ as the optimal stopping time for terminal reward M.

2.3.2. *Definition of fair charge and prevailing charge* Consider the fixed charge Problem (13). If the arm is in state i and the value of the charge is $\gamma(i)$ it is optimal to either play or stop, and in either case the expected optimal revenue (rewards minus charges) is $W(i) = 0$. 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 Problem (13). It is also exactly what was shown in step 1 of the proof of Theorem 2.

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 it 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 stoppings 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 revenue 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 lemma:

Lemma 2 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 $t(s), s = 1, 2, \ldots$, and $t(s)$ are strictly increasing in s for all s or increasing up to $t(\bar{s})$ and infinite for $s > \bar{s}$. Let $Z(t)$ be the state (frozen at times $t \notin \{t(s)\}_{s=1}^{\infty}$,

changing at times $t \in \{t(s)\}_{s=1}^{\infty}$). Note that $\{t(s)\}_{s=1}^{\infty}$ will typically be random, but we assume that $t(s)$ is measurable with respect to $Z(t), t \leq t(s)$.

Corollary 1

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

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 1 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.

2.3.3. *Investigating the retirement option* We now consider $V(i, M)$, the optimal value to the single arm retirement option problem (14) 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:

Lemma 3 (a) $V(i, M) = V(i)$ *for* $M \leq -\frac{C}{1-\alpha}$.
(b) $V(i, M) = M$ *for* $M \geq \frac{C}{1-\alpha}$.
(c) $V(i, M)$ *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:

$$V_\pi(i, M) = \mathbb{E}_\pi(\text{reward up to } \tau_\pi + \alpha^{\tau_\pi} M) \tag{20}$$

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

As a convex function $V(i, M)$ is differentiable at all but a countable number of points, at which it has subgradients. A glance at (15) or (20) suggests the form of the derivative.

Lemma 4 Let τ_M *denote the optimal retirement time for terminal reward* M. *Then* $\mathbb{E}\,(\alpha^{\tau_M})$ *is a subgradient of* $V(i, M)$ *(the line through* $(M, V(i, M))$ *with slope* $\mathbb{E}\,(\alpha^{\tau_M})$ *is below the curve* $V(i, \cdot)$*), and at every* M *for which* $\frac{\partial V(i,M)}{\partial M}$ *exists,*

$$\frac{\partial}{\partial M} V(i, M) = \mathbb{E}\,(\alpha^{\tau_M} \mid Z(0) = i) \tag{21}$$

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

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

Hence,

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

Similarly,

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

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

2.3.4. *First passage times* 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:

$$T_i^S = \min\{t : t > 0, Z(t) \in S | Z(0) = i\}, \tag{22}$$

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

$$A_i^S = \mathbb{E}\left\{\sum_{t=1}^{T_i^S - 1} \alpha^t | Z(0) = i\right\}, \tag{23}$$

we call A_i^S the i to S expected discounted first passage time. These quantities are needed in the achievable region proof (Section 3.4).

3. The proofs

3.1. *First Proof: Interchange Argument* This proof follows Gittins [5, 6, 7]. We wish to show that any priority policy which activates at time t an arm

$$n(t) \in \arg\max\{\nu(Z_n(t)) : n = 1, \ldots, N\}$$

is optimal. To avoid trivial complications we will assume that there are never any ties. Any problem can be changed into one in which ties never occur by arbitrarily small perturbations, so this assumption can be made. Once the unique theorem is proved for all perturbed problems, we can let the perturbations approach zero, and obtain the optimality of arbitrary tie breaking.

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

$$V_{\pi^{(s)}}(\mathbf{i}) \longrightarrow V_{\pi^*}(\mathbf{i}) \tag{24}$$

$$V_{\pi^{(s)}}(\mathbf{i}) \geq V_{\pi^{(s-1)}}(\mathbf{i}) \tag{25}$$

We define $\pi^{(s)}$ inductively. For initial state $\mathbf{i}$ let n^*, ν^*, τ^* be the bandit with the highest index, the index, and the stopping time achieving the index. Then $\pi^{(s)}$ will

activate n^* for duration τ^*, and will then proceed from time τ^* and state $\mathbf{j} = \mathbf{Z}(\tau^*)$ as $\pi^{(s-1)}$ would from time 0 and initial state $\mathbf{j}$.

By their construction, $\pi^{(s)}$ and π^* agree for the initial τ^*, $\tau^* \geq 1$. Furthermore, they continue to agree after τ^* for as long as $\pi^{(s-1)}$ and π^* agree, from the state reached at τ^*. Hence inductively $\pi^{(s)}$ agrees with π^* for at least the first s time units, hence $\pi^{(s)} \to \pi^*$, and the convergence in (24) is proved.

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

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

and so to prove (25) 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 $\nu(Z_n(t)) < \nu^*$. 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 index until $\sigma - 1$. At time σ the highest index is ν^* of bandit n^*, and so $\pi^{(0)}$, which continues according to π^*, will activate n^* for a period τ^*, up to time $\sigma + \tau^* - 1$. At time $\sigma + \tau^*$ the state will consist of $Z_{n^*}(\tau^*)$ 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 τ^* then at time τ^* 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 τ^* at which it starts to activate n. This is because after n is activated at τ^*, it will retain an index higher or equal to ν^* for the duration σ, while the index of bandit n^*, following its activation for duration τ^*, is now $\leq \nu^*$, and all other bandits retain their time 0 states, with indexes $\leq \nu^*$.

To summarize, $\pi^{(0)}$ activates n for duration σ followed by n^* for duration τ^*, followed by π^*; $\pi^{(1)}$ activates n^* for duration τ^* followed by n for duration σ followed by π^*. The state reached at time $\tau^* + \sigma$ 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 τ^* and σ are independent. The difference in the expected rewards is:

$$\begin{aligned}
&\mathbb{E}\left\{\sum_{t=0}^{\sigma-1}\alpha^t R(Z_n(t)) + \alpha^\sigma\sum_{t=0}^{\tau^*-1}\alpha^t R(Z_{n^*}(t)) + \alpha^{\sigma+\tau^*}\sum_{t=0}^{\infty}\alpha^t\tilde{R}(\sigma+\tau^*+t)\mid\mathbf{Z}(0)=\mathbf{i}\right\}\\
&-\mathbb{E}\left\{\sum_{t=0}^{\tau^*-1}\alpha^t R(Z_{n^*}(t)) + \alpha^{\tau^*}\sum_{t=0}^{\sigma-1}\alpha^t R(Z_n(t)) + \alpha^{\sigma+\tau^*}\sum_{t=0}^{\infty}\alpha^t\tilde{R}(\sigma+\tau^*+t)\mid\mathbf{Z}(0)=\mathbf{i}\right\}\\
&= \mathbb{E}\left(1-\alpha^{\tau^*}\right)\mathbb{E}\left\{\sum_{t=0}^{\sigma-1}\alpha^t R(Z_n(t))\right\} - \mathbb{E}\left(1-\alpha^{\sigma}\right)\mathbb{E}\left\{\sum_{t=0}^{\tau^*-1}\alpha^t R(Z_{n^*}(t))\right\}\\
&= \frac{1}{1-\alpha}\mathbb{E}\left(1-\alpha^{\tau^*}\right)\mathbb{E}\left(1-\alpha^{\sigma}\right)\left(\nu(i_n,\sigma)-\nu(i_{n^*})\right)\\
&\leq \frac{1}{1-\alpha}\mathbb{E}\left(1-\alpha^{\tau^*}\right)\mathbb{E}\left(1-\alpha^{\sigma}\right)\left(\nu(i_n)-\nu(i_{n^*})\right)\leq 0. \qquad (26)
\end{aligned}$$

This completes the proof of the theorem.

3.2. *Second Proof: Interleaving of Prevailing Charges* This proof follows Weber [18]. Similar ideas were also used by Mandelbaum [12] and by Varaiya et al. [17, 9]. We now consider N bandit processes, with initial state $\mathbf{Z}(0)=\mathbf{i}$. We let $t^{(n)}(s), s=1,2\dots$ 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 $\{1,2\dots\}$, as it should be under any sample path of any policy. We assume $t^{(n)}(s)$ is measurable with respect to $\mathbf{Z}(t), t\le t^{(n)}(s)$. We let $g_n(t), \underline{g}_n(t)$ denote the fair and the prevailing charges of bandit n.

By Corollary 1, the technical note following it, and the independence of the arms we have:

$$\text{Expected total discounted reward}=\mathbb{E}\left\{\sum_{n=1}^{N}\sum_{s=1}^{\infty}\alpha^{t^{(n)}(s)}R(Z_n(t^{(n)}(s)))|\mathbf{Z}(0)=\mathbf{i}\right\}$$

$$\le \quad \mathbb{E}\left\{\sum_{n=1}^{N}\sum_{s=1}^{\infty}\alpha^{t^{(n)}(s)}\underline{g}_n(t^{(n)}(s))|\mathbf{Z}(0)=\mathbf{i}\right\}=\mathbb{E}\left\{\sum_{t=0}^{\infty}\alpha^{t}\underline{\underline{g}}(t)|\mathbf{Z}(0)=\mathbf{i}\right\} \tag{27}$$

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:

$$\sum_{t=1}^{\infty}\alpha^{t}\underline{\underline{g}}(t)\le\sum_{t=1}^{\infty}\alpha^{t}\underline{\underline{g}}^{*}(t) \tag{28}$$

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 (28) 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 (27) holds as equality.

3.3. *Third Proof: Retirement Option* Following Whittle [19] 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 (16) for arm n we let:

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

Theorem 3 (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:

$$V(\mathbf{i}, M) = \max_{n=1,\ldots,N} \left\{ M, R(i_n) + \alpha \sum_{j \in E} p(i_n, j) V(i_1, \ldots, j, \ldots, i_N, M) \right\} \tag{29}$$

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

$$T(M) = \sum_{n=1}^{N} \tau_n(i_n, M) \tag{30}$$

We now speculate that

(31)
$$\frac{\partial}{\partial M} V(\mathbf{i}, M) = \mathbb{E}\left(\alpha^{T(M)}\right) = \mathbb{E}\left(\alpha^{\sum_{n=1}^{N} \tau_n(i_n, M)}\right) = \prod_{n=1}^{N} \mathbb{E}\left(\alpha^{\tau_n(i_n, M)}\right) = \prod_{n=1}^{N} \frac{\partial}{\partial M} V_n(i_n, M)$$

where the first equality might hold by analogy to (21), the second is (30), the third is true because the random variables $\tau_n(i_n, M)$ are independent, and the fourth is true by (21).

We also have that $V(\mathbf{i}, M) = M$ for $M \geq \frac{C}{1-\alpha}$. Integrating (31) we get the following conjectured form for the optimal value function:

$$\hat{V}(\mathbf{i}, M) = \frac{C}{1-\alpha} - \int_{M}^{\frac{C}{1-\alpha}} \prod_{n=1}^{N} \frac{\partial}{\partial m} V_n(i_n, m)\, dm \tag{32}$$

For each n define

$$Q_n(\mathbf{i}, M) = \prod_{n' \neq n} \frac{\partial}{\partial M} V_{n'}(i_{n'}, M) \tag{33}$$

By Lemma 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 (33) in (32) and integrating by parts we obtain for each n:

$$\hat{V}(\mathbf{i}, M) = V_n(i_n, M) Q_n(\mathbf{i}, M) + \int_{M}^{\frac{C}{1-\alpha}} V_n(i_n, m) dQ_n(\mathbf{i}, m) \tag{34}$$

To complete the proof we need to show that $\hat{V} = V$ by showing that $\hat{V}$ satisfies the optimality equation (29), which we do 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 (34)

$$\begin{aligned} \hat{V}(\mathbf{i}, M) &\geq V_n(i_n, M)Q_n(\mathbf{i}, M) + V_n(i_n, M)\int_M^{\frac{C}{1-\alpha}} dQ_n(\mathbf{i}, m) \\ &= Q_n(\mathbf{i}, \frac{C}{1-\alpha})V_n(i_n, M) = V_n(i_n, M) \geq M \end{aligned} \tag{35}$$

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

$$\Delta_n = \hat{V}(\mathbf{i}, M) - R(i_n) - \alpha \sum_{j \in E} p(i_n, j)\hat{V}(i_1, \dots, j, \dots, i_N, M). \tag{36}$$

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

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

Substituting (34) in (36), Δ_n is seen to be:

$$\begin{aligned} \Delta_n &= Q_n(\mathbf{i}, M)\left(V_n(i_n, M) - R(i_n) - \alpha \sum_{j \in E} P(i_n, j)V_n(j, M)\right) \\ &+ \int_M^{\frac{C}{1-\alpha}} \left(V_n(i_n, m) - R(i_n) - \alpha \sum_{j \in E} P(i_n, j)V_n(j, m)\right) dQ_n(\mathbf{i}, m) \\ &\geq 0 \end{aligned} \tag{37}$$

where

$$V_n(i_n, m) \geq R(i_n) + \alpha \sum_{j \in E} P(i_n, j)V_n(j, m) \tag{38}$$

by the optimality equation (14) for $V_n(i_n, m)$.

Step 3: Equality holds in (35,37), 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 (34) we see that (35) holds as equality for $M \geq M_{n^*}(i_{n^*})$.

Consider next $M \leq m \leq M_n(i_n)$, for which (38) holds as an equality. Substituting this in (37), we have that for such M:

$$\Delta_n = \int_{M_n(i_n)}^{\frac{C}{1-\alpha}} \left(V_n(i_n, m) - R(i_n) - \alpha \sum_{j \in E} p(i_n, j)V_n(j, m)\right) dQ_n(\mathbf{i}, m). \tag{39}$$

In particular if we take n^*, we have

$$\Delta_{n^*} = \int_{M_{n^*}(i_{n^*})}^{\frac{C}{1-\alpha}} \left(V_{n^*}(i_{n^*}, m) - R(i_{n^*}) - \alpha \sum_{j \in E} P(i_{n^*}, y)V_{n^*}(j, m)\right) dQ_{n^*}(\mathbf{i}, m) = 0 \tag{40}$$

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

3.4. *Fourth Proof: The Achievable Region Approach* This proof follows the ideas of Tsoucas et al. [16, 1] and of Bertsimas and Niño-Mora [3]. Remarkably enough it is actually quite close to the pioneering proof of Klimov [10].

3.4.1. *Generalized conservation laws* Consider the N bandit system with initial state $\mathbf{Z}(0) = \mathbf{i}$, 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 the activated arm is in state i.
Note: The value of the objective function for π is $\sum_{i\in E} R(i)x_i^\pi$.

Recall the definitions (22,23) of T_i^S, A_i^S, $S \subseteq E$. For initial state $\mathbf{Z}(0) = \{i_1, \ldots, i_N\}$, denote $T_{Z(0)}^S = \sum_{n:i_n \notin S} T_{i_n}^S$, and let:

$$b(S) = \frac{\mathbb{E}\{\alpha^{T_{Z(0)}^S}\}}{1-\alpha}. \tag{41}$$

We also use the following notation: If a policy π gives priority to states outside S over states in S, we write: $\pi : S^c \to S$.

Theorem 4 For initial state $\mathbf{Z}(0)$*, for every policy* π *and every* $S \subseteq E$

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

Equality holds if and only if $\pi : S^c \to S$.

Proof. Consider a realization (single sample path) under policy π. Then $T_{Z(0)}^S$ 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 $s_0(1) < \cdots < s_0(T_{Z(0)}^S)$, be the times at which we operate on arms which were initially in S^c, before they have entered a state in S. We take these times in increasing order, and it is possible that $s_0(l) = \infty$ from some point onwards.
- Let $s_i(1) < \cdots < s_i(l) < \cdots$ be the times at which we activate arms in state i, where $i \in S$. Again we write these in increasing order, and it is possible that $s_i(l) = \infty$ from some point onwards.
- 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. We let $T_i^S(i,l)$ denote the number of steps from when an arm was in i on that l's occasion, until following some plays it returns to S. Again we write the times in increasing order, and it is possible that $s_{i,l}(k) = \infty$ from some point onwards.

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

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

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

We now obtain an inequality:

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

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

where the first inequality can hold as equality if and only if $s_{i,l}(1),\ldots,s_{i,l}(T^S_i(i,l)-1) = s_i(l)+1,\ldots,s_i(l)+T^S_i(i,l)-1$, 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 $\pi : S^c \to S$.

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

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

$$\begin{aligned} \max \quad & \sum_{i\in E} R(i)x_i \\ s.t. \quad & \sum_{i\in S} A^S_i x_i \geq b(S), \qquad S\subset E, \\ & \sum_{i\in E} A^E_i x_i = b(E) = \frac{1}{1-\alpha}, \\ & x_i \geq 0, \quad i\in E. \end{aligned} \tag{42}$$

It is a relaxation in the sense that any performance measure given by a policy π has to satisfy the constraints of the linear program.

3.4.2. *The Linear Program* To complete the proof we investigate the linear program (42). For the continuation of the proof we need to restrict ourselves to the case of a finite number of states, $|E|$.

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 $S_i = [\varphi(i),\ldots,\varphi(|E|)]$, $i = 1,\ldots,|E|$. Then $\varphi : S^c_i \to S_i$. Consider the upper triangular matrix (which is part of the coefficient matrix of the

LP (42)):

$$D = \begin{bmatrix} A^{S_1}_{\varphi(1)} & A^{S_1}_{\varphi(2)} & \cdots & A^{S_1}_{\varphi(|E|)} \\ 0 & A^{S_2}_{\varphi(2)} & \cdots & A^{S_2}_{\varphi(|E|)} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A^{S_{|E|}}_{\varphi(|E|)} \end{bmatrix}.$$

By Theorem 4, the performance measures $\mathbf{x}^\varphi$ are the unique solution of the triangular set of equations:

$$D \begin{bmatrix} x^\varphi_1 \\ x^\varphi_2 \\ \vdots \\ x^\varphi_{|E|} \end{bmatrix} = \begin{bmatrix} b(S_1) \\ b(S_2) \\ \vdots \\ b(S_{|E|}) \end{bmatrix}.$$

and are a basic feasible solution to the LP (42). Thus, the vector of performance measures of each priority policy is an extreme point of the LP.

Consider now the complementary slack dual solution corresponding to $\mathbf{x}^\varphi$. It is of the form $y^S = 0, S \neq S_1, \ldots, S_{|E|}$ while the remaining dual variables solve:

$$D' \begin{bmatrix} y^{S_1} \\ y^{S_2} \\ \vdots \\ y^{S_{|E|}} \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(|E|) \end{bmatrix}$$

which gives recursively, for $S_1, \ldots, S_{|E|}$:

$$y^{S_i} = \frac{R(i) - \sum_{j=1}^{i-1} A_i^{S_j} y^{S_j}}{A_i^{S_i}},$$

For $\mathbf{x}^\varphi$ to be optimal it is necessary that $y^{S_2} \leq 0, \ldots, y^{S_{|E|}} \leq 0$. We can now use $R(1), \ldots, R(|E|)$ to construct such a permutation φ. The following algorithm, known as Klimov's algorithm will do it. Starting from $S_1 = E$, calculate recursively for $i = 1, \ldots, |E|$,

$$y^{S_i} = \max_{k \in S_i} \frac{R(k) - \sum_{j=1}^{i-1} A_k^{S_j} y^{S_j}}{A_k^{S_i}}, \quad \varphi(i) = \arg\max_{k \in S_i} \frac{R(k) - \sum_{j=1}^{i-1} A_k^{S_j} y^{S_j}}{A_k^{S_i}}$$
$$S_{i+1} = S_i \backslash \varphi(i).$$

Because $\mathbf{x}^\varphi$ is optimal, the priority policy based on this permutation is optimal. It is easy to see that this coincides with the Gittins policy, and in fact the index is $\nu(\varphi(i)) = \sum_{j=1}^{i} y^{S_j}$. This completes the proof.

Note: If we let R vary over all possible $|E|$ vectors, the solutions of (42) vary over all the extreme points of the feasible polyhedron of the LP. But for each such R the above algorithm finds an optimal permutation priority policy which has that extreme point as its performance vector. Hence: The achievable performance region coincides with the feasible region of the LP, and its extreme points coincide with the performance vectors of the priority policies.

3.4.3. *Extended polymatroides* 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* [4], 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 $S_1, \ldots, S_{|E|}$ be the nested subsets $S_i = \{\varphi(i), \ldots, \varphi|E|)\}$. Then $\mathcal{M}$ has exactly $|E|!$ extreme points given by the the solutions of:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \mathbf{x}^{\varphi} = \begin{bmatrix} b(S_1) \\ b(S_2) \\ \vdots \\ b(S_{|E|}) \end{bmatrix}.$$

This property implies that the optimization of any objective function linear in $\mathbf{x}$ is achieved by a greedy solution.

Tsoucas [1, 16] and Bertsimas and Niño-Mora [3] define extended polymatroids as a generalization to polymatroids as 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: For every permutation and nested sets as above, the solution to

$$\begin{bmatrix} a_{\varphi(1)}^{S_1} & a_{\varphi(2)}^{S_1} & \cdots & a_{\varphi(|E|)}^{S_1} \\ 0 & a_{\varphi(2)}^{S_2} & \cdots & a_{\varphi(|E|)}^{S_2} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{\varphi(|E|)}^{S_{|E|}} \end{bmatrix} \mathbf{x}^{\varphi} = \begin{bmatrix} b(S_1) \\ b(S_2) \\ \vdots \\ b(S_{|E|}) \end{bmatrix},$$

is in $\mathcal{EM}$.

The discussion in Section 3.4.2 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.

Acknowledgement

This research was supported by GIF — German Israeli Foundation grant number I-564-246-06/97.

References

[1] 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.

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

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

[4] 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.

[5] 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.

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

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

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

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

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

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

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

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

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

[15] 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.

[16] 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.

[17] 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.

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

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