MULTI-ARMED BANDITS UNDER GENERAL DEPRECIATION AND COMMITMENT

Wesley Cowan and Michael N. Katehakis

Probability in the Engineering and Informational Sciences / Volume 29 / Issue 01 / January 2015, pp 51 - 76
DOI: 10.1017/S0269964814000217, Published online: 10 October 2014

Link to this article: http://journals.cambridge.org/abstract_S0269964814000217

How to cite this article:

Request Permissions: Click here
MULTI-ARMED BANDITS UNDER GENERAL DEPRECIATION AND COMMITMENT

WESLEY COWAN
Department of Mathematics, Rutgers University
110 Frelinghuysen Road, Piscataway, NJ 08854, USA
E-mail: cwcowan@mah.rutgers.edu

MICHAEL N. KATEHAKIS
Department of Management Science and Information Systems
Rutgers Business School, Newark and New Brunswick
100 Rockafeller Road, Piscataway, NJ 08854, USA
E-mail: mnk@rutgers.edu

Generally, the multi-armed has been studied under the setting that at each time step over an infinite horizon a controller chooses to activate a single process or bandit out of a finite collection of independent processes (statistical experiments, populations, etc.) for a single period, receiving a reward that is a function of the activated process, and in doing so advancing the chosen process. Classically, rewards are discounted by a constant factor $\beta \in (0, 1)$ per round.

In this paper, we present a solution to the problem, with potentially non-Markovian, uncountable state space reward processes, under a framework in which, first, the discount factors may be non-uniform and vary over time, and second, the periods of activation of each bandit may be not be fixed or uniform, subject instead to a possibly stochastic duration of activation before a change to a different bandit is allowed. The solution is based on generalized restart-in-state indices, and it utilizes a view of the problem not as “decisions over state space” but rather “decisions over time”.

1. INTRODUCTION AND SUMMARY

Generally, the multi-armed bandit problem has been described in terms of sequentially allocating effort to one of $N$ independent processes, or bandits, for instance sequentially assigning measurements to one of $N$ possible statistical populations or measurements in clinical trials. In what follows, we discuss the problem in terms of bandit activation. In each period, a controller chooses a single bandit to activate from the $N$ available, basing that decision on all information available about all bandits at that time. The activated bandit yields a reward that depends on its current state, and then moves to a new state according to a probability law of motion that is a function of that bandit’s history. Each bandit is taken to be independent of the others. Inactive bandits in a period yield no rewards, and their states remain frozen for that period.
Central results are the existence and form of index-based policies for certain models that maximize the present value of expected rewards, cf. Gittins et al. [19], Frostig and Weiss [18], Mahajan and Teneketzis [37], Kaspi and Mandelbaum [28], Ishikida and Varaiya [27], El Karoui and Karatzas [14], Gittins [22], Gittins [21] and Gittins et al. [20].

Important extensions of the basic problem were given by Agrawal et al. [3] that considered multiple plays and switching costs, and by Caro and Yoo [9] that considered response delays. Further, in Ouyang and Teneketzis [41] conditions are given under which a myopic policy is optimal for a multi-state channel probing environment, and in Niño-Mora [38] who presents indexability conditions for discrete-state semi-Markov bandits.

Other interesting formulations and applications are discussed in Glazebrook et al. [23], Glazebrook et al. [24], Su et al. [46], Agmon et al. [2], Lai et al. [34], Liu et al. [36] and Aalto et al. [1], Katehakis et al. [30], Weber [51], Weber and Weiss [52] and Chang and Lai [10].

Following Lai et al. [35], alternative treatments have involved minimizing the rate of increase of a regret function, cf. Katehakis and Robbins [32], Burnetas et al. [7], Burnetas and Katehakis [8], Ortner and Auer [40], Oksanen et al. [39]. For other related work we refer to the following: Flint et al. [17], Fernández-Gaucherand et al. [15], Govindarajulu and Katehakis [25], Honda and Takemura [26], Tekin and Liu [47], Tewari and Bartlett [48], Filippi et al. [16], Bertsekas [4], Bubeck and Cesa-Bianchi [5] and Burnetas et al. [6].

In this paper, we consider the following formulation of the problem. In a discrete-time-step model, future rewards from all processes depreciate from period to period according to a possibly stochastic sequence of bandit-dependent discount factors. We show that the optimal policy for the multi-armed under this generalized depreciation model is an index policy, where the indices are propitiously generalized restart in state indices, cf. Katehakis and Veinott [33] and Katehakis et al. [29]; see also Sonin [43], Sonin [44] and Steinberg and Sonin [45]. Furthermore, the overall proof suggests a way of understanding the structure of the reward processes, relative to a “natural” time scale of possibly stochastic intervals of activation or “restart blocks”, rather than steps of unit time.

We note that this problem can be treated to some extent in the classical semi-Markov formulation of the multi-armed, in which the duration a reward process remains in a given state determines the discounting on future rewards. However, the treatment given in this paper is justified by the following reasons:

One, the reward processes and discount factor processes as treated here are defined in considerable generality, as potentially non-Markovian processes over uncountable state spaces. As a result, many classical solutions to this problem, cf. Denardo et al. [12], and Tsitsiklis [49], formulated with finite-state Markov chains, do not apply. The benefit of this increased generality is broader applicability, such as in the case of time-dependent reward processes, or partially observed reward processes (i.e. POMDPs).

Two, many classical treatments of these types of problems treat them as what might be called “decisions over state space”, determining what decision to make in each potential state of the reward processes. The approach taken here might well be described as “decisions over time”, determining when to make what decision and for how long. This can be viewed as a generalization of the approach taken in Kaspi and Mandelbaum [28]. To demonstrate the difference in perspectives, from the first, a simple reward process might be a two-state Markov chain. From the second, a simple reward process would be one characterized simply over time, such as an infinite, monotone process. This perspective leads to a reformulation of the problem, which can be solved simply via a sample-pathwise optimization argument, cf. proof of Theorem 3.

The rest of the paper is organized as follows. In Section 2, we formulate the generalized depreciation model rigorously. Section 3 is devoted to useful notions of the “value” of a set or block of activations of a bandit. In Section 4, we define the generalized restart-in-state
indices and use them to develop the appropriate time scale to understand the structure of reward processes. The necessity of the restart-in-state index is established by example in Section 4.1. In Section 5, we use the generalized restart-in-state indices to construct alternative “summary” reward processes, which are then used to derive the optimal policy in Section 6. Section 7 introduces a model under which activation is subject to periods of commitment, and reduces this model to the previous depreciation model. To close, Section 8 discusses the meaning and implications of a key assumption that allows the techniques presented here to work.

2. FRAMEWORK: A GENERALIZED DEPRECIATION MODEL

A controller is presented with a collection of filtered probability spaces, \((\Omega^i, \mathcal{F}^i, \mathbb{P}^i, \mathbb{F}^i)\), for \(1 \leq i \leq N < \infty\), representing \(N\) environments in which experiments will be performed or rewards collected – the “bandits”. To each space, we associate a reward process \(X^i = \{X^i_t\}_{t \geq 0}\), and a discount factor sequence \(\beta^i = \{\beta^i_t\}_{t \geq 0}\). For \(t \in \{0, 1, \ldots\}\), we take \(X^i_t(=X^i_t(\omega^i)) \in \mathbb{R}\) to represent the reward received from bandit \(i\) on its \(t\)th activation. Additionally, however, we take all rewards collected after the \(t\)th activation of bandit \(i\) to be discounted, reduced by a factor of \(\beta^i_t(=\beta^i_t(\omega^i)) \in (0, 1)\). Following tradition, we take both \(X^i\) and \(\beta^i\) to be \(\mathcal{F}^i\)-adapted. We denote the reward process collection as \(X\), and the discount factor sequence collection as \(B\).

We state the following key assumption that insures that the bandits are mutually independent.

Assumption A: There is a larger “global” probability space \((\Omega, \mathcal{G}, \mathbb{P}) = (\otimes_{i=1}^N \Omega^i, \otimes_{i=1}^N \mathcal{F}^i, \otimes_{i=1}^N \mathbb{P}^i)\), a standard product-space construction, representing the environment of the controller – aware information from all bandits.

Expectations relative to the local space, that is, bandit \(i\), will be denoted \(E^i\), while expectations relative to the global space are simply \(E\). Note that Assumption A ensures that \(X^i, X^j\) are independent relative to \(\mathbb{P}\) for \(i \neq j\), so too are the \(\beta^i, \beta^j\), though \(\beta^i, X^i\) need not be independent.

Remark 1: We adopt the following notational liberty, allowing a random variable \(Z\) defined on a local space \(\Omega^i\) to also be considered as a random variable on the global space \(\Omega\), taking \(Z(\omega) = Z(\omega^i)\), where \(\omega = (\omega^1, \ldots, \omega^N) \in \Omega\). Via this extension, we may take expectations involving a process \(X^i\), or \(\mathcal{F}^i\)-stopping times, relative to \(\mathbb{P}\) instead of \(\mathbb{P}^i\), without additional notational overhead.

In what follows, we reserve the term “round” to differentiate global, controller time, denoted with \(s\), from local bandit times, denoted by \(t\).

The following assumption formally states that in every round a reward is received from the activated bandit, whose state may change, while unactivated bandits remain frozen and yield no rewards.

Assumption B: In each round, the controller selects a bandit \(i\) to activate, receiving its current reward \(X^i_t\) where \(t\) is the current local time for that bandit, and advancing that bandit’s local time one step. All bandits begin at local time 0, and advance only on activation.

For each bandit \(i\), it is convenient to define a total depreciation sequence \(\{\alpha^i_t\}_{t \geq 0}\), such that \(\alpha^i_t\) represents the total discounting incurred by the first \(t\) activations of bandit \(i\). That
is, we may take $\alpha_0^i = 1$, and

$$\alpha_i^t = \prod_{t'=0}^{t-1} \beta_{i}^{t'}.$$  \hfill (1)$$

We additionally make the following assumption stated in terms of restrictions on each bandit $i$:

**Assumption C:**

$$\lim_{t \to \infty} \alpha_i^t = \prod_{t'=0}^{\infty} \beta_{i}^{t'} = 0 (\P^i - \text{a.e.}),$$  \hfill (2)$$

and

$$\mathbb{E}^i \left[ \sum_{t=0}^{\infty} \alpha_i^t | X_i^t \right] < \infty.$$  \hfill (3)

The latter implies immediately that the expected total reward from any bandit is finite.

Aiming to maximize her expected total reward, in every round the controller's decision of which bandit to activate must balance not only which reward to collect in that round, but also the effect of the incurred discounting on all future rewards from all bandits. A **control policy** $\pi$ is a stochastic process on $(\Omega, \mathcal{G}, \P)$ that specifies, at each round $s$ of global time, which bandit to activate and collect from, e.g., $\pi(s) (= \pi(s, \omega)) = i$, specifies to activate bandit $i$ at global time $s$. We restrict ourselves to the set of policies $\mathcal{P}$ defined to be non-anticipatory, that is, polices for which $\pi(s)$ does not depend on outcomes that have not yet occurred, or information not yet available.

Given a policy $\pi$, it is convenient to be able to translate between global time and local time. Define $S_i^\pi(t)$ to be the round at which process $i$ is activated for the $t$th time when the controller operates according to policy $\pi$. This may be expressed as

$$S_i^\pi(0) = \inf \{ s \geq 0 : \pi(s) = i \},$$

$$S_i^\pi(t + 1) = \inf \{ s > S_i^\pi(t) : \pi(s) = i \}. \hfill (4)$$

We may also define $T_i^\pi(s)$ to denote the local time of bandit $i$ just prior to the $s$th round under a policy $\pi$, that is, $T_i^\pi(0) = 0$, and for $s > 0$, and

$$T_i^\pi(s) = \sum_{s'=0}^{s-1} 1 \{ \pi(s') = i \}. \hfill (5)$$

It is convenient to define the global time analog, $T_\pi(s) = T_\pi^\pi(s) (s)$ to denote the current local time of the bandit activated at round $s$ under policy $\pi$. This will allow us to define concise global time analogs of several processes. For instance, we define the global reward process $X_\pi$ on $(\Omega, \mathcal{G}, \P)$ as $X_\pi(s) = X_{T_\pi(s)}^\pi$, giving the reward from collection $X$ under policy $\pi$ at round $s$.

Given a policy $\pi \in \mathcal{P}$, the reward collected at round $s$ under $\pi$ is discounted by a factor of

$$A_\pi(s) = \prod_{i=1}^{N} \alpha_i^{T_\pi(s)}. \hfill (6)$$

While $\alpha_i^t$ gives the depreciation on $X_i^t$ due to the activations of bandit $i$, given a policy $\pi$ it is also convenient to define the depreciation on $X_i^t$ on its collection, due to activations of any process under that policy. This policy-dependent depreciation is given by $A_\pi(t) = \alpha_i^{T_\pi(s)}(T_\pi(s))$. 
In what follows, we let $V_\pi(X, \mathcal{B})$ denote the value of a policy, the expected total reward given the reward–discount pair $X, \mathcal{B}$ under policy $\pi \in \mathcal{P}$. Taking $\mathcal{G}_0 = \otimes_{i=1}^N \mathcal{F}_i(0)$ as the initial global information available to the controller, we may express the value of a policy as

$$V_\pi(X, \mathcal{B}) = \mathbb{E}\left[ \sum_{s=0}^{\infty} A_\pi(s) X_\pi(s) \mid \mathcal{G}_0 \right],$$

relative to global time, or relative to local time as

$$V_\pi(X, \mathcal{B}) = \sum_{i=1}^{N} \mathbb{E}\left[ \sum_{t=0}^{\infty} \alpha_\pi^i(t) X_i^t \mid \mathcal{G}_0 \right].$$

The problem the controller faces is to determine a policy $\pi^* \in \mathcal{P}$ that is optimal in the sense that for any other $\pi \in \mathcal{P}$,

$$V_\pi(X, \mathcal{B}) \leq V_{\pi^*}(X, \mathcal{B}) (\mathbb{P} - \text{a.e.}).$$

In the remainder of the paper, we construct just such an optimal policy.

### 2.1. Global Information Versus Local Information

One of the intricacies of the results to follow is in properly distinguishing and determining what information is available to the controller to act on at any given time. For each bandit $i$, the filtration $\mathbb{F}^i = \{ \mathcal{F}^i(t) \}_{t \geq 0}$ represents the progression of information available about that bandit – the $\sigma$-algebra $\mathcal{F}^i(t)$ representing the local information available about bandit $i$ at local time $t$, such as the process history of $X^i$. Taking $X^i$ as $\mathbb{F}^i$-adapted as we do, we have $\sigma(X^i_0, X^i_1, \ldots, X^i_t) \subset \mathcal{F}^i(t)$.

At round $s$, the total, global information available to the controller is determined by the state of each bandit at that round, that is, acting under a given policy $\pi$ until round $s$, the global information available at round $s$ is given by the $\sigma$-algebra $\otimes_{i=1}^{N} \mathcal{F}^i(T^i_\pi(s))$. We may therefore refine the prior definition of non-anticipatory policies to be the set of policies $\mathcal{P}$ such that for each $s \geq 0$, $\pi(s)$ is measurable with respect to the prior $\sigma$-algebra, that is, determined by the information available at round $s$. Weaker definitions of non-anticipatory, such as dependence on random events, e.g., coin flips, are addressed in Section 6.

Additionally, given a policy $\pi$, it is necessary to define a set of policy-dependent filtrations in the following way: let $\mathbb{H}^i_\pi = \{ \mathcal{H}^i_\pi(t) \}_{t \geq 0}$, where $\mathcal{H}^i_\pi(t) = \otimes_{j=1}^{N} \mathcal{F}^j(T^j_\pi(S^j_\pi(t)))$ represents the total information available to the controller about all bandits, prior to the $t^{th}$ activation of bandit $i$ under $\pi$. It is indexed by the local time of bandit $i$, but at each time $t$ gives the current state of information of each bandit. Note that, since $T^i_\pi(S^i_\pi(t)) = t$, $\mathcal{H}^i_\pi(t)$ contains the information available in $\mathcal{F}^i(t)$. This filtration is necessary for expressing local stopping times, that is, concerning $X^i$, from the perspective of the controller – $\mathbb{F}^i$-stopping times no longer suffice, since the controller has access to information from all the other processes as well. Note though, $\mathbb{F}^i$-stopping times may be viewed as $\mathbb{H}^i_\pi$-stopping times, cf. Remark 1. Ultimately, the optimal policy result demonstrates that any decision about a given bandit depends only on information from that bandit, thus rendering these

---

1 This means the value of $X^i_t$ is revealed prior to its collection by the controller, determined by the information available up to time $t$. A more general model might consider $X^i_t$ to remain uncertain just prior to its collection, that is have $X^i_t$ be measurable with respect to $\mathcal{F}^i(t + 1)$, but not $\mathcal{F}^i(t)$. However, this may be reduced to the case we present here, taking the reward process to be given by $\hat{X}^i_t = \mathbb{E}^i \left[ X^i_t \mid \mathcal{F}^i(t) \right]$. 

---

MULTI-ARMED BANDITS UNDER GENERAL DEPRECIATION AND COMMITMENT
filtrations unnecessary in practice. However, they are a technical necessity for the proof of that result.

When discussing stopping times, we will utilizing the following notation: For a general filtration $\mathcal{J}$ (e.g., $\mathcal{J} = \mathcal{F}_i, \mathcal{H}_i$), we denote by $\tilde{\mathcal{J}}(t)$ the set of all $\mathcal{J}$-stopping times strictly greater than $t$ ($\mathbb{P}^i, \mathbb{P}$-a.e.). For a $\mathcal{J}$-stopping time $\tau$, $\tilde{\mathcal{J}}(\tau)$ is similarly defined.

3. BLOCK VALUES

This section introduces a way of considering the “value” of a set of activations of a bandit. As noted previously, the “true” value of a decision to activate a bandit is not simply the reward gained through that decision, but instead must balance the immediate reward with the effect on all future rewards of the discounting incurred through that decision.

We start with the following definitions.

**Definition 1 (Process Blocks and their Values):** Given times $t' < t''$, and a policy $\pi$ with $S^i_\pi(t') < \infty$, we define the following quantities.

1. The restart value of the $[t', t'')$ - block of $X^i$ as:

$$
\rho^i(t', t'') = \frac{\mathbb{E}^i \left[ \sum_{t=t'}^{t''-1} \alpha^i_t X^i_t \mid \mathcal{F}^i(t') \right]}{\mathbb{E}^i \left[ \sum_{t=t'}^{t''-1} \alpha^i_t (1 - \beta^i_t) \mid \mathcal{F}^i(t') \right]}.
$$

2. The $[t', t'')$ - $\pi$-block value of $X^i$ as:

$$
\nu^i_\pi(t', t'') = \frac{\mathbb{E}^i \left[ \sum_{t=t'}^{t''-1} \alpha^i_\pi(t) X^i_t \mid \mathcal{H}^i_\pi(t') \right]}{\mathbb{E}^i \left[ \sum_{t=t'}^{t''-1} \alpha^i_\pi(t) (1 - \beta^i_t) \mid \mathcal{H}^i_\pi(t') \right]}.
$$

Note, the above quantities are all measurable with respect to the indicated $\sigma$-fields, and finite, due to the assumptions of Eq. (3) and that $\beta^i_t < 1$ for all $i, t, (\mathbb{P}^i, \mathbb{P}$-a.e.).

**Remark 2:** We might offer the following justification of the above block “values”. Noting that $\alpha^i_{t+1} = \beta^i_t \alpha^i_t$, the denominator of $\rho^i(t', t'')$ becomes telescoping, and we may equivalently express it as

$$
\rho^i(t', t'') = \frac{\mathbb{E}^i \left[ \sum_{t=t'}^{t''-1} \alpha^i_t X^i_t \mid \mathcal{F}^i(t') \right]}{\mathbb{E}^i \left[ \alpha^i_t - \alpha^i_{t'} \mid \mathcal{F}^i(t') \right]}.
$$

In this form, it can be shown that $\rho^i(t', t'')$ represents the total reward accrued from bandit $i$ starting at time $t'$ if the controller could, at time $t''$, restart the block at time $t'$ and continue to collect rewards, repeating this on an infinite-time horizon.

From this perspective, it can be seen that the less depreciation a block incurs over its duration, the higher the corresponding value of $\rho^i$. As such, a block that yields a small reward but very little depreciation might in fact have a higher value than a block yielding high reward but incurring serious depreciation. This seems to capture the balance the controller must strike, between reward and depreciation – and indeed does so, as the optimal policy will demonstrate.
Note, $\nu^i_\pi$ as in Eq. (11) is the obvious policy-dependent generalization of $\rho^i$ in Eq. (10), rather than $\rho^i$ as in Eq. (12). It has no immediate “restart”-type interpretation in this form. Taking view that the $\pi$-block value $\nu^i_\pi$ is the value of a block of $X^i$ when activated under a specific policy $\pi$, $\rho^i$ should be viewed as the value of a block with respect to consecutive activation.

The following theorem illustrates the relationship between $\rho^i$ and $\nu^i_\pi$, essentially stating that the value of any block under some policy is at most the value of some block activated consecutively. It depends on the following lemma concerning stochastic control, inspired by a lemma in Varaiya et al. [50]. We present its proof in Appendix A.1.

**Lemma 1:** In an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$ consider a discrete-time process \{$Z_t\}_{t \geq 0}$ such that $\mathbb{E}[\sum_{t=0}^{\infty} |Z_t|] < \infty$. Let $\mathcal{J} = \{\mathcal{J}_t\}_{t \geq 0}$ be a filtration, and $\{\alpha_t\}_{t \geq 0}$ be a $\mathcal{J}$-adapted process such that $\alpha_t \geq 0(t_0 - a.e.).$ In such a case the following is true for any $\tau \in \mathcal{F}(0)$:

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha_t Z_t | \mathcal{J}_0\right] \leq \alpha_0 \text{ess sup}_{\tau' \in \mathcal{J}(0)} \mathbb{E}\left[\sum_{t=0}^{\tau'-1} Z_t | \mathcal{J}_0\right] (\mathbb{P} - a.e.).$$

(13)

**Theorem 1 (Block Value Comparison):** For bandit $i$ under policy $\pi$, for any time $t_0$ such that $S^i(t_0) < \infty$, the following holds for any $\mathcal{H}^i_{\tau}$-stopping time $\tau$ with $t_0 < \tau$:

$$\nu^i_\pi(t_0, \tau) \leq \text{ess sup}_{\hat{\tau} \in \mathcal{F}(t_0)} \rho^i(t_0, \hat{\tau}) (\mathbb{P} - a.e.).$$

(14)

Note that by Eq. (3), the essential supremum is finite ($\mathbb{P}$-a.e).

**Proof:** Denote the essential supremum above by $\rho$, which is $\mathcal{F}(t_0)$-measurable. Note, for any $\hat{\tau} \in \mathcal{F}(t_0)$,

$$\mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha^i_t X^i_t | \mathcal{F}(t_0)\right] \leq \rho (\mathbb{P}^i - a.e.).$$

(15)

We may rearrange Eq. (15) to yield

$$\mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha^i_t (X^i_t - \rho (1 - \beta_i^t)) | \mathcal{F}(t_0)\right] \leq 0 (\mathbb{P}^i - a.e.).$$

(16)

Since this holds for all such $\hat{\tau}$, we have:

$$\text{ess sup}_{\hat{\tau} \in \mathcal{F}(t_0)} \mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha^i_t (X^i_t - \rho (1 - \beta_i^t)) | \mathcal{F}(t_0)\right] \leq 0 (\mathbb{P}^i - a.e.).$$

(17)

To demonstrate that $\nu^i_\pi(t_0, \tau) \leq \rho(\mathbb{P} - a.e.)$, it is enough to show that

$$\mathbb{E}\left[\sum_{t=t_0}^{\tau-1} \alpha^i_t (X^i_t - \rho (1 - \beta_i^t)) | \mathcal{H}^i_{\tau}(t)\right] \leq 0 (\mathbb{P} - a.e.).$$

(18)

It is useful to factor $\alpha^i_\pi$ into contributions to the discounting from bandit $i$ and contributions from the remaining bandits. That is, we take

$$\alpha^i_\pi(t) = \hat{A}^i_\pi(S^i(t)) \alpha^i(t),$$

(19)
where
\[
\hat{A}^i_n(s) = \prod_{j \neq i} \alpha^j T^i_n(s),
\]
(20)

In the spirit of applying Lemma A.1, note that \( \hat{A}^i_n(S^i_n(t)) \) is \( \mathbb{H}^i_n \)-adapted, and
\[
1 \geq \hat{A}^i_n(S^i_n(t)) \geq \hat{A}^i_n(S^i_n(t + 1)) \geq 0 \text{ (P-a.e.).}
\]

Noting that the integrability conditions follow from Eq. (3), Lemma A.1 may then be applied to demonstrate Eq. (18) in the following way. Defining \( Z^i_t = X^i_t - \rho^i \alpha^i \),
\[
\mathbb{E} \left[ \sum_{t=t_0}^{\tau-1} \alpha^i(t) Z^i_t \mid \mathcal{H}^i_n(t_0) \right] = \mathbb{E} \left[ \sum_{t=t_0}^{\tau-1} \hat{A}^i_n(S^i_n(t)) \alpha^i Z^i_t \mid \mathcal{H}^i_n(t_0) \right]
\leq \hat{A}^i_n(S^i_n(t_0)) \text{ ess sup} \mathbb{E} \left[ \sum_{t=t_0}^{\hat{\tau}-1} \alpha^i Z^i_t \mid \mathcal{H}^i_n(t_0) \right]
\leq \hat{A}^i_n(S^i_n(t_0)) \text{ ess sup} \mathbb{E} \left[ \sum_{t=t_0}^{\hat{\tau}-1} \alpha^i Z^i_t \mid \mathcal{F}^i(t_0) \right]
\leq 0
\]
(21)

the above relations holding (P-a.e.). The exchange from the essential sup over \( \mathbb{H}^i_n \)-stopping times to \( \mathbb{F}^i \)-stopping times follows as the \( \alpha^i, Z^i \) depend only on bandit \( i \), and independent information about non-\( i \) bandits cannot help in maximizing that sum. This exchange is proven in more detail in Appendix B.1, but the proof simply amounts to integrating out the independent bandits. The last step follows from Eq. (17).

This gives Eq. (18), and completes the proof. \( \blacksquare \)

The following proposition provides, using \( \rho^i \) and \( \nu^i \), alternative expressions for the reward accrued through the activation of a block. They follow immediately from Eqs (10) and (11).

**Proposition 1:** The following hold for any \( \mathbb{F}^i \)-stopping times \( \tau' < \tau'' \). Equality also holds when conditioning with respect to the initial information \( \mathcal{F}^i(0), \mathcal{G}_0 \) respectively via the tower property.
\[
\mathbb{E}^i \left[ \sum_{t=\tau'}^{\tau''-1} \alpha^i(t) X^i_t \mid \mathcal{F}^i(\tau') \right] = \mathbb{E}^i \left[ \sum_{t=\tau'}^{\tau''-1} \alpha^i(t) \rho^i(\tau', \tau'')(1 - \beta^i_t) \mid \mathcal{F}^i(\tau') \right],
\]
(22)
\[
\mathbb{E} \left[ \sum_{t=\tau'}^{\tau''-1} \alpha^i(t) X^i_t \mid \mathcal{H}^i_n(\tau') \right] = \mathbb{E} \left[ \sum_{t=\tau'}^{\tau''-1} \alpha^i(t) \rho^i(\tau', \tau'')(1 - \beta^i_t) \mid \mathcal{H}^i_n(\tau') \right],
\]
(23)
\[
\mathbb{E} \left[ \sum_{t=\tau'}^{\tau''-1} \alpha^i_n(t) X^i_t \mid \mathcal{H}^i_n(\tau') \right] = \mathbb{E} \left[ \sum_{t=\tau'}^{\tau''-1} \alpha^i_n(t) \nu^i_n(\tau', \tau'')(1 - \beta^i_t) \mid \mathcal{H}^i_n(\tau') \right].
\]
(24)

**Remark 3:** Note the relationship the above suggests between \( X^i_t \) and \( \rho^i(\tau', \tau'')(1 - \beta^i_t) \), and between \( X^i_t \) and \( \nu^i_n(\tau', \tau'')(1 - \beta^i_t) \) under \( \pi \), for \( \tau' < t < \tau'' \). This will prove to be central to the final proof.
4. THE RESTART-IN-STATE INDEX: DEFINITION AND PROPERTIES

Theorem 1 indicates the significance of the following quantity.

**Definition 2 (The Restart-in-State Index):** For any $t < \infty$, the Restart-in-State Index at $t$ is defined to be

$$\rho^i(t) = \text{ess sup}_{\tau \in F^i(t)} \rho^i(t, \tau).$$

(25)

This form of the index, based on the quotient in Eq. (10), was anticipated by Sonin [44], who defined it on Markov chain reward processes as a generalization of the classical Gittins dynamic allocation index, and as an extension of the restart-in-state index of Katehakis and Veinott [33]. Note that if all the discount factors are equal across bandits, that is, $\beta^i_t = \beta$ for some $\beta$, for all $i$, and $t$, then this index differs from the classical Gittins index by merely a factor of $1/(1 - \beta)$, and index policies based on either index will be equivalent. However, taking $\beta^i_t$ in its full generality, no such relationship exists between the indices and therefore between policies.

The necessity of the restart-in-state index, within the discrete-time model framework, is established by example in Section 4.1.

Noting that $\rho^i(t, \tau)$ is the value of the $(t, \tau)$ – block, we may interpret $\rho^i(t)$ as the maximal block value achievable from bandit $i$ from time $t$. The use of the terms maximal and achievable is justified here as $\rho^i(t)$ is realized $(P^i - \text{a.e.})$ as the value of some block starting at $t$. To show this requires the following technical lemma, which follows as a special case of classic results of Snell [42] and others, cf. the Optional Stopping Lemma of Derman and Sacks [13] and its discussion in Katehakis et al. [31].

**Lemma 2:** In an arbitrary probability space, consider a discrete-time process $\{Z_t\}_{t \geq 0}$ such that $E[\sum_{t=0}^{\infty} |Z_t|] < \infty$. Let $\mathcal{J} = \{J_t\}_{t \geq 0}$ be a filtration such that the $Z_t$ are $\mathcal{J}$-adapted. Then, there exists a (potentially infinite) stopping time $\tau^* \in \hat{\mathcal{J}}(0)$ such that

$$E\left[\sum_{t=0}^{\tau^*-1} Z_t \big| J_0\right] = \text{ess sup}_{\tau \in \hat{\mathcal{J}}(0)} E\left[\sum_{t=0}^{\tau-1} Z_t \big| J_0\right] (P - \text{a.e.}).$$

(26)

In particular, we may take

$$\tau^* = \inf \left\{ n > 0 : \text{ess sup}_{\tau \in \hat{\mathcal{J}}(n)} E\left[\sum_{t=n}^{\tau-1} Z_t \big| J_n\right] < 0 \right\}.$$  

(27)

Note, we allow infinite stopping times, as the sum in Eq. (27) is well defined by assumption.

Utilizing this lemma, we have the following result.

**Proposition 2:** For any time $t_0 < \infty$, there exists a $\tau \in \hat{F}^i(t_0)$ such that $\rho^i(t_0) = \rho^i(t_0, \tau) (P^i - \text{a.e.})$.  

PROOF: We have that for all $\hat{\tau} \in \hat{\mathcal{F}}^i(t_0)$, $\rho^i(t_0, \hat{\tau}) \leq \rho^i(t_0)$ ($\mathbb{P}^i$ - a.e.), or in parallel with Eq. (16),

$$\mathbb{E}^i \left[ \sum_{t=t_0}^{\hat{\tau} - 1} \alpha^i_t (X^i_t - \rho^i(t_0)(1 - \beta^i_t)) \mid \mathcal{F}^i(t_0) \right] \leq 0 \ (\mathbb{P}^i - \text{a.e.}). \quad (28)$$

Defining

$$\epsilon = - \text{ess sup}_{\hat{\tau} \in \hat{\mathcal{F}}^i(t_0)} \mathbb{E}^i \left[ \sum_{t=t_0}^{\hat{\tau} - 1} \alpha^i_t (X^i_t - \rho^i(t_0)(1 - \beta^i_t)) \mid \mathcal{F}^i(t_0) \right],$$

we have that $\epsilon \geq 0 (\mathbb{P}^i - \text{a.e.})$. We may use $-\epsilon$ as an improved upper bound in Eq. (28). This may be rearranged to yield

$$\rho^i(t_0, \hat{\tau}) \leq \rho^i(t_0) - \frac{\epsilon}{\mathbb{E}^i \left[ \sum_{t=t_0}^{\hat{\tau} - 1} \alpha^i_t (1 - \beta^i_t) \mid \mathcal{F}^i(t_0) \right]} \rho^i(t_0) - \frac{\epsilon}{\mathbb{E}^i [\alpha^i_{t_0} - \alpha^i_{t_0} \mid \mathcal{F}^i(t_0)]} \leq \rho^i(t_0) - \epsilon (\mathbb{P}^i - \text{a.e.}). \quad (30)$$

Since the above property holds for all such $\hat{\tau}$, it extends to the essential supremum, yielding

$$\rho^i(t_0) \leq \rho^i(t_0) - \epsilon (\mathbb{P}^i - \text{a.e.}), \quad (31)$$

or equivalently that $\epsilon \leq 0 (\mathbb{P}^i - \text{a.e.})$. In conjunction with the first observation, that $\epsilon \geq 0 (\mathbb{P}^i - \text{a.e.})$, we have $\epsilon = 0 (\mathbb{P}^i - \text{a.e.})$, that is,

$$\text{ess sup}_{\hat{\tau} \in \hat{\mathcal{F}}^i(t_0)} \mathbb{E}^i \left[ \sum_{t=t_0}^{\hat{\tau} - 1} \alpha^i_t (X^i_t - \rho^i(t_0)(1 - \beta^i_t)) \mid \mathcal{F}^i(t_0) \right] = 0 (\mathbb{P}^i - \text{a.e.}). \quad (32)$$

To satisfy the hypotheses of Lemma 2, note that the integrability stems by the assumption of (3). We may apply Lemma 2 in this instance to yield a stopping time $\tau^* \in \hat{\mathcal{F}}^i(t_0)$ such that

$$\mathbb{E}^i \left[ \sum_{t=t_0}^{\tau^* - 1} \alpha^i_t (X^i_t - \rho^i(t_0)(1 - \beta^i_t)) \mid \mathcal{F}^i(t_0) \right] = 0 (\mathbb{P}^i - \text{a.e.}), \quad (33)$$

or

$$\rho^i(t_0) = \frac{\mathbb{E}^i \left[ \sum_{t=t_0}^{\tau^* - 1} \alpha^i_t X^i_t \mid \mathcal{F}^i(t_0) \right]}{\mathbb{E}^i \left[ \sum_{t=t_0}^{\tau^* - 1} \alpha^i_t (1 - \beta^i_t) \mid \mathcal{F}^i(t_0) \right]} = \rho^i(t_0, \tau^*) (\mathbb{P}^i - \text{a.e.}). \quad (34)$$

Hence, the restart-in-state index $\rho^i(t_0)$ is realized ($\mathbb{P}^i$ - a.e.) for some $\hat{\mathcal{F}}^i$-stopping time $\tau^* > t_0$. In particular, the block value of $X^i$, starting at $t_0$, is maximized ($\mathbb{P}^i$ - a.e.) by the $[t_0, \tau^*)$ block.

The restart-in-state indices and their realizing blocks provide a natural time scale to view the bandits in, in terms of a sequence of blocks. In particular, we define the following sequence.
Suppose the controller is given two deterministic, finite processes, \( X^1 \) and \( X^2 \). The dynamic allocation (Gittins) indices, within the discrete-time model framework, specify to activate \( X^1 \) at the first, then \( X^2 \) twice. This gives a value of \( V_{\text{DAI}} = 100 + 1b + 2ab \).

\[ \tau_{k+1}^i = \text{arg} \sup \{ \rho^i(\tau_k^i, \tau) : \tau \in \hat{\mathcal{F}}^i(\tau_k^i) \} \]  

(35)

In the case that \( \tau_k^i \) is infinite for some \( k \), then \( \tau_k^i \) is taken to be infinite for all larger \( k' \). In the case that \( \tau_k^i < \infty \), we have that \( \rho^i(\tau_k^i) = \rho^i(\tau_k^i, \tau_{k+1}) \). The question of whether the “arg ess sup” exists in this case is resolved in the positive by Proposition 3; if there is more than one stopping time that attains the “arg ess sup” above we take \( \tau_{k+1}^i \) to be the one demonstrated by the application of Lemma 2.

Using this sequence of stopping times, we partition the local process times \( \mathbb{N}^i = \{0, 1, 2, \ldots \} \) into

\[ \mathbb{N}^i = [0, \tau_1^i) \cup [\tau_1^i, \tau_2^i) \cup [\tau_2^i, \tau_3^i) \cup \ldots. \]

One important property of this partition is the following.

**Proposition 3 (Restart-in-State Indices Non-Increasing over Index Times):** For any \( k > 0 \) such that \( \tau_k^i < \infty \), the following is true:

\[ \rho^i(\tau_{k-1}^i) \geq \rho^i(\tau_k^i)(\mathbb{P}^i - \text{a.e.}) \]

**Proof:** For \( k > 0 \), if \( \tau_k^i < \infty \) and therefore \( \tau_{k-1}^i < \infty \), we have that the restart-in-state index from time \( \tau_{k-1}^i \) is realized \((\mathbb{P}^i - \text{a.e.})\) by \( \tau_k^i \) such that, via Lemma 2,

\[ \text{ess sup} \mathbb{E}^i [ \frac{\sum_{t=\tau_k^i}^{\tau_{k-1}^i-1} \alpha_t^i (X_t^i - \rho^i(\tau_{k-1}^i)(1 - \beta_t^i)) | \mathcal{F}^i(\tau_k^i) ] < 0 (\mathbb{P}^i - \text{a.e.}) \]  

(36)

From this, for any \( \hat{\tau} \in \hat{\mathcal{F}}^i(\tau_k^i) \), we have

\[ \mathbb{E}^i \left[ \frac{\sum_{t=\tau_k^i}^{\hat{\tau}+1} \alpha_t^i X_t^i | \mathcal{F}^i(\tau_k^i) }{\sum_{t=\tau_k^i}^{\hat{\tau}+1} \alpha_t^i(1 - \beta_t^i) | \mathcal{F}^i(\tau_k^i) } < \rho^i(\tau_{k-1}^i)(\mathbb{P}^i - \text{a.e.}) \right. \]

(37)

Taking the essential supremum over such \( \hat{\tau} \) (and noting that the essential sup will again be realized) establishes that \( \rho^i(\tau_k^i) \leq \rho^i(\tau_{k-1}^i) \) \((\mathbb{P}^i - \text{a.e.})\). \( \blacksquare \)

### 4.1. Necessity of the Restart in State Indices

We construct the following toy example to demonstrate the inapplicability of the classical form of the dynamic allocation (Gittins) indices, within the discrete-time model framework. Suppose the controller is given two deterministic, finite processes, \( X^1 = \{1, 2\} \) and \( X^2 = \{100\} \); they can also be thought of as infinite processes with infinite trailing 0’s. Further, each activation of \( X^1 \) incurs a discount factor of \( a \) (i.e. \( \beta_0^1 = \beta_1^1 = a \)) and each activation of \( X^2 \) incurs a discount factor of \( b \) (i.e. \( \beta_0^2 = b \)) with \( a, b \in (0, 1) \).

The classic Gittins (1979) DAI dynamic allocation indices of \( X^1 \) and \( X^2 \) at \( s = 0 \) are:

\[ \gamma_{X^1} = (1 + 2a)/(1 + a) = \max\{1/1, (1 + 2a)/(1 + a)\} \] and \( \gamma_{X^2} = 100 \). From this, it is clear that for any value of \( a \in (0, 1) \), \( \gamma_{X^1} < \gamma_{X^2} \). Hence, the DAI based dynamic allocation policy specifies to activate \( X^2 \) first, then \( X^1 \) twice. This gives a value of \( V_{\text{DAI}} = 100 + 1b + 2ab \).
However, consider the alternative strategy of activating $X^1$ twice first, then activating $X^2$. This gives a value of $V' = 1 + 2a + 100a^2$. Comparing the two, we have

$$\frac{V' - V_{DAI}}{(1 - b)(1 - a^2)} = \frac{1 + 2a}{1 - a^2} - \frac{100}{1 - b}.$$ 

It is clear from the above that for $a$ especially large and $b$ small, the difference in values is positive, and the policy determined by the restart dynamic allocation indices is in fact superior in value to the policy determined by the DAI indices, that is, the traditional DAI policy may specify a non-optimal policy. On the other hand, it is easy to see that the policy based on the restart in state indices of $X^1$ and $X^2$ which at $s = 0$ are: $\rho_{X^1} = (1 + 2a)/(1 - a^2) = \max\{1/(1 - a), (1 + 2a)/(1 - a^2)\}$, $\rho_{X^2} = 100/(1 - b)$ is always optimal.

5. BANDIT AND POLICY EQUIVALENT REWARD PROCESSES

For each bandit, we have developed a partition of local time for into blocks of activations, via the restart-in-state index stopping times. We extend on the remarks at the end of Section 3, by defining the following alternative reward processes.

**Definition 4:** Given the collection of reward processes $X = (X^1, X^2, \ldots, X^N)$, discount factor sequences $B = (\beta^1, \beta^2, \ldots, \beta^N)$, and $\{\tau_k^i\}_{k \geq 0}$ as by Definition 3, we define:

1. The reward-equivalent collection $Y^X = (Y^1, \ldots, Y^N)$ by

   $$Y^i(t) = \rho^i(\tau_k^i)(1 - \beta^i_k), \quad \text{if } \tau_k^i \leq t < \tau_{k+1}^i.$$

2. For $\pi \in P$, the $\pi$-equivalent collection $Y^X_\pi = (Y^1_\pi, \ldots, Y^N_\pi)$, by

   $$Y^i_\pi(t) = \nu^i_\pi(\tau_k^i, \tau_{k+1}^i)(1 - \beta^i_k), \quad \text{if } \tau_k^i \leq t < \tau_{k+1}^i.$$

Like $X^i$, the process $Y^i$ is defined on $(\Omega^i, \mathcal{F}^i, \mathbb{P}^i, \mathbb{F}^i)$ and is $\mathbb{F}^i$-adapted, as $\rho^i(\tau_k^i)$ is defined by the information available locally at time $\tau_k^i$. However, as the $\nu^i_\pi(\tau_k^i, \tau_{k+1}^i)$ depend on the specifics of a policy $\pi$, so do the $Y^i_\pi$ processes; the $Y^i_\pi$ processes are $\mathbb{F}^i_\pi$-adapted, but not $\mathbb{F}^i$-adapted. Note that, should bandit $i$ be activated only finitely many times under $\pi$, $Y^i_\pi$ will only really be defined up to some $\tau_{k+1}^i$ such that $S^i_\pi(\tau_{k+1}^i) = \infty$. For such undefined $Y^i_\pi(t)$, we take $0 \ast Y^i_\pi(t) = 0$.

The following are simple, but important, properties of the $Y^i, Y^i_\pi$ processes.

**Proposition 4:** For each bandit $i$, the following hold any $k \geq 0$:

$$\mathbb{E}^i \left[ \sum_{t = \tau_k^i}^{\tau_{k+1}^i-1} \alpha^i_{t} X^i_{\pi}(t) | \mathcal{F}^i(\tau_k^i) \right] = \mathbb{E}^i \left[ \sum_{t = \tau_k^i}^{\tau_{k+1}^i-1} \alpha^i_{t} Y^i_\pi(t) | \mathcal{F}^i(\tau_k^i) \right],$$

$$\mathbb{E} \left[ \sum_{t = \tau_k^i}^{\tau_{k+1}^i-1} \alpha^i_{t} X^i_{\pi}(t) | \mathcal{H}^i_\pi(\tau_k^i) \right] = \mathbb{E} \left[ \sum_{t = \tau_k^i}^{\tau_{k+1}^i-1} \alpha^i_{t} Y^i_\pi(t) | \mathcal{H}^i_\pi(\tau_k^i) \right],$$

$$\mathbb{E} \left[ \sum_{t = \tau_k^i}^{\tau_{k+1}^i-1} \alpha^i_{t}(t) X^i_{\pi}(t) | \mathcal{H}^i_\pi(\tau_k^i) \right] = \mathbb{E} \left[ \sum_{t = \tau_k^i}^{\tau_{k+1}^i-1} \alpha^i_{t}(t) Y^i_\pi(t) | \mathcal{H}^i_\pi(\tau_k^i) \right].$$

As with Proposition 1, equality also holds when conditioning with respect to $\mathcal{F}^i(0), G_0$. 


Proof: This follows as an application of Proposition 1 and the definitions of $Y^i, Y^i_\pi$. ■

The following proposition serves as a justification of the term “equivalent” to describe $Y^i, Y^i_\pi$.

**Proposition 5:** For each bandit $i$,

$$
\mathbb{E}^i \left[ \sum_{t=0}^{\infty} \alpha^i_t X^i_t | F^i(0) \right] = \mathbb{E}^i \left[ \sum_{t=0}^{\infty} \alpha^i_t Y^i_t | F^i(0) \right],
$$

(43)

$$
\mathbb{E} \left[ \sum_{t=0}^{\infty} \alpha^i_\pi(t) X^i_t | G_0 \right] = \mathbb{E} \left[ \sum_{t=0}^{\infty} \alpha^i_\pi(t) Y^i_\pi(t) | G_0 \right].
$$

(44)

Proof: Each follows from the corresponding equation in Prop. 4, summing over $k$ and taking expectations from the initial time, via the tower property. ■

**Theorem 2 (Comparison of the Equivalent Processes):** For each $i$ and all time $t$, we have:

$$
\alpha^i_\pi(t) Y^i_\pi(t) \leq \alpha^i_\pi(t) Y^i(t) \quad (P - \text{a.e.}).
$$

(45)

Proof: If the controller does not activate bandit $i$ at least $t$ times under $\pi$, ever (i.e. $S^i_\pi(t) = \infty$), then some bandit $j \neq i$ must be activated infinitely many times under $\pi$. As such, $\alpha^i_\pi(t) \leq \alpha^j_\infty = 0$, by Eq. (2). Hence, $\alpha^i_\pi(t) = 0$, and the inequality holds.

If the controller does activate bandit $i$ at least $t$ times under $\pi$, then $\alpha^i_\pi(t)$ is non-zero and can be ignored on both sides of the inequality. In that case, we have for some $k$ that $\tau^i_k \leq t < \tau^i_{k+1}$, and as an application of Theorem 1,

$$
Y^i_\pi(t) = \nu^i_\pi(\tau^i_k, \tau^i_{k+1}) (1 - \beta^i_t) \leq \rho^i(\tau^i_k)(1 - \beta^i_t) = Y^i(t) \quad (P - \text{a.e.}).
$$

(46)

6. A GREEDY RESULT AND THE OPTIMAL CONTROL POLICY

The importance of the $\mathcal{Y}^X$ collection is that the optimal policy for these reward processes may be derived with relative ease. In fact, for these processes, not only may the total reward be maximized in expectation, but a policy exists which maximizes the reward almost surely. That is, we have the following theorem.

**Theorem 3 (Pointwise Optimization on $\mathcal{Y}^X$ Processes):** Given a reward - discount pair $(X, \mathbb{B})$, let $\mathcal{Y}^X$ be the collection of reward-equivalent processes. There exists a policy $\pi^* \in \mathcal{P}$ such that for any other policy $\pi \in \mathcal{P}$, $\pi^*$ yields a greater total reward, almost surely. That is, defining $Y^X_\pi(s) = Y^\pi(s)(T_\pi(s))$,

$$
\sum_{s=0}^{\infty} A_\pi(s) Y^X_\pi(s) \leq \sum_{s=0}^{\infty} A_{\pi^*}(s) Y^X_{\pi^*}(s) \quad (P - \text{a.e.}).
$$

(47)

In particular, such a $\pi^*$ is given by the following rule: activate bandit $i$ corresponding to the largest current value of $\rho^i(\tau^i_k)$, for a duration corresponding to the realizing $[\tau^i_k, \tau^i_{k+1}]$-block, repeating this ad infinitum.
PROOF: Without loss of generality, let $\rho^1(0) \geq \rho^i(0)$ for all $i$. Since the $\rho^i(\tau_k^i)$ are decreasing with $k$, we have for all $i, k$ that
\[
\rho^1(0) \geq \rho^i(\tau_k^i)(\mathbb{P} - \text{a.e.}). \tag{48}
\]

Let $\pi$ be an arbitrary policy in $\mathcal{P}$, and define $S = S^1_\pi(0)$, the first time bandit 1 is activated under $\pi$. Note, if bandit 1 is never activated, we take $S$ to be infinite.

We may express the total reward under policy $\pi$ sample path wise as
\[
R_\pi = \sum_{s=0}^{\infty} A_\pi(s) Y^{X}_{\pi}(s)
\]
\[
= \sum_{s=0}^{S-1} A_\pi(s) Y^{X}_{\pi}(s) + A_\pi(S) Y^{1}(0) + \sum_{s=S+1}^{\infty} A_\pi(s) Y^{X}_{\pi}(s). \tag{49}
\]

From $\pi$, we construct a policy $\pi' \in \mathcal{P}$ in the following way: $\pi'$ is identical to $\pi$ in that it activates bandits in the same order, but it advances the first activation of bandit $i$ from round $s = S$ to round $s = 0$. That is,
\[
\pi'(s) = \begin{cases} 
1 & \text{for } s = 0, \\
\pi(s-1) & \text{for } s = 1, 2, \ldots, S, \\
\pi(s) & \text{for } s \geq S + 1. 
\end{cases} \tag{50}
\]

It is important to observe that $\pi'$ is in $\mathcal{P}$, as at every round $s$, the information available under $\pi'$ is greater than or equal to the information available under $\pi$.

Using this relation, we may express the sample path wise reward under policy $\pi'$, relative to $\pi$, as
\[
R_{\pi'} = \sum_{s=0}^{\infty} A_{\pi'}(s) Y^{X}_{\pi'}(s)
\]
\[
= Y^{1}(0) + \sum_{s=1}^{S} A_{\pi'}(s) Y^{X}_{\pi'}(s) + \sum_{s=S+1}^{\infty} A_{\pi'}(s) Y^{X}_{\pi'}(s)
\]
\[
= Y^{1}(0) + \sum_{s=1}^{S} \beta_1 A_\pi(s-1) Y^{X}_{\pi}(s-1) + \sum_{s=S+1}^{\infty} A_\pi(s) Y^{X}_{\pi}(s)
\]
\[
= Y^{1}(0) + \beta_1 \sum_{s=0}^{S-1} A_\pi(s) Y^{X}_{\pi}(s) + \sum_{s=S+1}^{\infty} A_\pi(s) Y^{X}_{\pi}(s). \tag{51}
\]

Comparing the two, we have
\[
R_{\pi'} - R_\pi = \left( Y^{1}(0) + \beta_1 \sum_{s=0}^{S-1} A_\pi(s) Y^{X}_{\pi}(s) \right) - \left( \sum_{s=0}^{S-1} A_\pi(s) Y^{X}_{\pi}(s) + A_\pi(S) Y^{1}(0) \right)
\]
\[
= Y^{1}(0)(1 - A_\pi(S)) - (1 - \beta_1) \sum_{s=0}^{S-1} A_\pi(s) Y^{X}_{\pi}(s) \tag{52}
\]
\[
= Y^{1}(0)(1 - A_\pi(S)) - \sum_{s=0}^{S-1} A_\pi(s)(1 - \beta_1) Y^{X}_{\pi}(s).
\]
Defining $\beta_\pi(s) = \beta_{T_\pi(s)}$, we have by Eq. (48) that $(1 - \beta_0^i)Y^X_\pi(s) \leq (1 - \beta_\pi(s))Y^1(0)$. Hence,

$$R_{\pi^*} - R_\pi \geq Y^1(0)(1 - A_\pi(S)) - \sum_{s=0}^{S-1} A_\pi(s)(1 - \beta_\pi(s))Y^1(0)$$

$$= Y^1(0)(1 - A_\pi(S)) - Y^1(0) \sum_{s=0}^{S-1} (A_\pi(s) - A_\pi(s)\beta_\pi(s))$$

$$= Y^1(0)(1 - A_\pi(S)) - Y^1(0) \sum_{s=0}^{S-1} (A_\pi(s) - A_\pi(s + 1))$$

$$= Y^1(0)(1 - A_\pi(S)) - Y^1(0)(1 - A_\pi(S))$$

$$= 0 \ (\mathbb{P} - \text{a.e.}).$$

Immediately, for any policy $\pi$, advancing the first activation of the bandit with the largest current $\rho^i$ value almost surely increases, or does not change, the value of the policy. This argument can be extended, via a forward induction-type argument, to show that any finite number of these $\rho$-greedy advancements almost surely improves or does not change the value of the policy. Collisions, when two bandits have the same current $\rho$-value, are left to the discretion of the controller, but may be resolved with a simple rule such as always picking the lowest numbered bandit.

It remains to compare these improved strategies to the completely $\rho$-greedy strategy as in the theorem. Let $\pi^* \in \mathcal{P}$ be the completely $\rho$-greedy strategy, described in the theorem. For a given $\pi \in \mathcal{P}$, let $\pi_N \in \mathcal{P}$ be the policy that results from $\pi$ after $N$-many $\rho$-greedy advancements. Notice, $\pi^*$ and $\pi_N$ agree for the first $N$ rounds. Let $\tilde{T}_N^i = T^i_{\pi^*}(N)$. Recalling the definition of $\tilde{A}_N^i(s)$, cf. Eq. eqrefAhat, as the discounting due to non-$i$ bandits at round $s$, we have the following bound:

$$|R_{\pi^*} - R_{\pi_N}| \leq \sum_{i=1}^{N} \sum_{t=\tilde{T}_N^i}^{\infty} |\alpha_{i,\pi^*}(t) - \alpha_{i,\pi_N}(t)||Y^i(t)|$$

$$\leq \sum_{i=1}^{N} \sum_{t=\tilde{T}_N^i}^{\infty} \alpha_{i,\pi_N}(t) A_{\pi^*}(i, (S_{\pi^*}^i(\tilde{T}_N^i)))|Y^i(t)|$$

$$\leq \sum_{i=1}^{N} \tilde{A}_{\pi^*}(i, (S_{\pi^*}^i(\tilde{T}_N^i))) \left( \sum_{t=\tilde{T}_N^i}^{\infty} \alpha_{i,\pi_N}(t)|Y^i(t)| \right).$$

Note that it follows from the definition of $Y^i$ and Eq. (3) that $\sum_{t=0}^{\infty} \alpha_{i,\pi_N}(t)|Y^i(t)| < \infty (\mathbb{P} - \text{a.e.}).$

For a given bandit $i$ under $\pi^*$, two things may happen: either $i$ is activated infinitely many times, or $i$ is activated finitely many times.

If $i$ is activated infinitely many times under $\pi^*$, then $\tilde{T}_N^i$ increases without bound with $N$. As such, $\sum_{t=\tilde{T}_N^i}^{\infty} \alpha_{i,\pi_N}(t)|Y^i(t)|$ converges to 0 ($\mathbb{P} - \text{a.e.}$). As the depreciation due to the non-$i$ bandits is at most 1, the contribution of bandit $i$ in the above sum converges to 0 with $N$.

If $i$ is activated finitely many times under $\pi^*$, then some bandit $j \neq i$ is activated infinitely many times under $\pi^*$. Thus, for some finite $N$, $S_{\pi_N^i}(\tilde{T}_N^i) = \infty$, and for the infinitely activated bandit $j \neq i$ we have $\tilde{A}_{\pi_N^i}(S_{\pi_N^i}(\tilde{T}_N^i)) \leq \alpha_{\infty}^j = 0 (\mathbb{P} - \text{a.e.})$, by Eqs (2) and (20). In...
other words, the depreciation incurred at infinity is at most the depreciation incurred from j at infinity, which is 0. Since the total remaining reward from bandit i is finite almost surely, the contribution of bandit i to the above sum converges to 0 with N (P − a.e.).

The above imply the following:

$$\lim_{N \to \infty} |R_{\pi^*} - R_{\pi^N}| = 0 \text{ (P − a.e.)} \tag{55}$$

Since the value of any policy may be improved by a finite number of $\rho$-greedy advancements, and the value of a policy under $N$ $\rho$-greedy advancements converges to the value of the total $\rho$-greedy policy $\pi^*$, it follows that for any policy $\pi \in \mathcal{P}$ we have (P − a.e.) that $R_{\pi} \leq R_{\pi^*}$, verifying the theorem.

Remark 4: An important feature of the optimal policy $\pi^*$ for the $Y^X$ processes is that it preserves restart in state index blocks. Because it is $\rho$-greedy, we have explicitly, for $\tau^i_k \leq \tau^i_{k+1}$, that $\pi^*(S^i_{\pi^*}(\tau^i_k) + t) = i$, (P-a.e.).

We then have the following result.

**Theorem 4 (The Optimal Control Policy for Generalized Deprecation):** For a collection of reward processes $X = (X^1, X^2, \ldots, X^N)$, and discount factor sequences $\mathbb{B} = (\beta^1, \beta^2, \ldots, \beta^N)$, there exists a strategy $\pi^* \in \mathcal{P}$ such that for all $\pi \in \mathcal{P}$,

$$V_{\pi}(X, \mathbb{B}) \leq V_{\pi^*}(X, \mathbb{B})(P − a.e.). \tag{56}$$

In particular, such an optimal $\pi^*$ can be described in the following way: Successively, activate the bandit with the largest restart in state index $\rho^i$, for the duration of the corresponding index block.

**Proof:** For an arbitrary policy $\pi$, and $\pi^*$ as indicated above, we establish the following relations:

$$V_{\pi}(X, \mathbb{B}) = V_{\pi}(Y^X_\pi, \mathbb{B}) \leq V_{\pi}(Y^X, \mathbb{B}) \leq V_{\pi^*}(Y^X, \mathbb{B}) = V_{\pi^*}(X, \mathbb{B})(P − a.e.), \tag{57}$$

that is, for any policy $\pi$, we have that $\pi(\mathcal{X}, \mathbb{B}) \leq \pi^*(\mathcal{X}, \mathbb{B})(P \text{ − a.e.})$ and therefore $\pi^*$ is an optimal policy. In the following steps, we prove relations (57).

**Step 1:** $V_{\pi}(X, \mathbb{B}) = V_{\pi}(Y^X_\pi, \mathbb{B})(P \text{ − a.e.})$.

We have, by Proposition 5,

$$V_{\pi}(X, \mathbb{B}) = \sum_{i=1}^{N} \mathbb{E} \left[ \sum_{t=0}^{\infty} \alpha^i_{\pi}(t)X^i_{\pi}(t) \bigg| \mathcal{G}_0 \right] = \sum_{i=1}^{N} \mathbb{E} \left[ \sum_{t=0}^{\infty} \alpha^i_{\pi}(t)Y^i_{\pi}(t) \bigg| \mathcal{G}_0 \right] = V_{\pi}(Y^X_\pi, \mathbb{B}).$$

Note, because the $Y^i_{\pi}$ processes are defined in terms of $\pi$, they are not $\mathcal{F}^i$-adapted, and cannot be utilized under any other policy. However, the value $V_{\pi}(Y^X_\pi)$ is well defined via the above equation.

**Step 2:** $V_{\pi}(Y^X_\pi, \mathbb{B}) \leq V_{\pi}(Y^X, \mathbb{B})(P − a.e.)$.

This follows simply from the almost sure comparison of Theorem 2, giving

$$V_{\pi}(Y^X_\pi, \mathbb{B}) = \sum_{i=1}^{N} \mathbb{E} \left[ \sum_{t=0}^{\infty} \alpha^i_{\pi}(t)Y^i_{\pi}(t) \bigg| \mathcal{G}_0 \right] \leq \sum_{i=1}^{N} \mathbb{E} \left[ \sum_{t=0}^{\infty} \alpha^i_{\pi}(t)Y^i(t) \bigg| \mathcal{G}_0 \right] = V_{\pi}(Y^X, \mathbb{B}). \tag{58}$$

**Step 3:** $V_{\pi}(Y^X, \mathbb{B}) \leq V_{\pi^*}(Y^X, \mathbb{B})(P − a.e.)$. 

This is simply an application of Theorem 3, as \( V_\pi(\mathcal{Y}^X, \mathcal{B}) = \mathbb{E} [R_\pi | \mathcal{G}_0] \), with \( R_\pi \) as in the theorem.

**Step 4:** \( V_\pi^*(\mathcal{Y}^X, \mathcal{B}) = V_\pi^*(\mathcal{X}, \mathcal{B})(\mathbb{P} - \text{a.e.}) \).

By the factorization of \( \alpha_i^\pi_t(t) \) as in Eq. (19), and Remark 4, we may express the total reward under \( \pi^* \) relative to the \( [\tau^*_k, \tau^*_{k+1}] \)-blocks. In this way, we may apply Proposition 4 to yield the following equivalence:

\[
V_{\pi^*}(\mathcal{Y}^X, \mathcal{B}) = \sum_{i=1}^N \sum_{k=0}^{\infty} \mathbb{E} \left[ \hat{A}_i^\pi(S^\pi_{\tau^*_k}) \mathbb{E} \left[ \sum_{t=\tau^*_k}^{\tau^*_{k+1}-1} \alpha_i^\pi Y^i(t) | \mathcal{H}_{\tau^*_k}^\pi \right] | \mathcal{G}_0 \right] \\
= \sum_{i=1}^N \sum_{k=0}^{\infty} \mathbb{E} \left[ \hat{A}_i^\pi(S^\pi_{\tau^*_k}) \mathbb{E} \left[ \sum_{t=\tau^*_k}^{\tau^*_{k+1}-1} \alpha_i^\pi X^i_t | \mathcal{H}_{\tau^*_k}^\pi \right] | \mathcal{G}_0 \right] \\
= V_{\pi^*}(\mathcal{X}, \mathcal{B})(\mathbb{P} - \text{a.e.}).
\]

This completes the proof.

**Remark 5:** The above theorem demonstrates a policy \( \pi^* \in \mathcal{P} \) that is \( \mathbb{P} \)-a.e. superior (or equivalent) to every other policy \( \pi \in \mathcal{P} \). However, the set of non-anticipatory policies \( \mathcal{P} \) was defined in a fairly restrictive sense in Section 2.1, so that the decision in any round was completely determined by the results of the past. This might be weakened to allow for randomized policies, so that the decision in a given round might depend on the results of independent events, e.g., coin flips. However, such a construction simply amounts to placing a distribution on \( \mathcal{P} \). Since \( \pi^* \) is \( \mathbb{P} - \text{a.e.} \) superior to any \( \pi \in \mathcal{P} \), \( \pi^* \) would be similarly superior to any policy sampled randomly from \( \mathcal{P} \).

### 7. A MODEL OF COMMITMENTS

One way of interpreting the discount factor \( \beta^i_t \) in the previous model is as related to the duration of the \( t^{th} \) activation of bandit \( i \)-durations that may be non-uniform across bandits, and across time. This section lays out this relationship in detail: we define a continuous time model, in which rewards are discounted continuously at a fixed rate, and in which activation of a bandit is subject to a (potentially stochastic) period of commitment that must pass before the controller is again free to select a bandit to activate. Such commitments might arise as the product of contractual obligations, or properties of the bandits such as operational speeds, or the controller may simply require a certain condition be met before switching.

The key assumptions are again that the bandits are independent, and unactivated bandits remain frozen. The controller’s goal remains the maximization her expected total reward. As such, each decision of which bandit to activate must balance not only the immediate reward collected, but potential delay in collection and additional discounting of all future rewards, due to commitment.

A controller is presented with a collection of filtered probability spaces, \( (\Omega^i, \mathcal{F}^i, \mathbb{P}^i, \mathbb{F}^i) \), for \( 1 \leq i \leq N < \infty \), each indexed in continuous time, and satisfying the usual conditions. To each space, we associate a continuous time reward rate process \( X^i = \{X^i_t\}_{t \geq 0} \). For \( t \in [0, \infty) \), we take \( X^i_t(\omega^i) \in \mathbb{R} \) to represent the reward rate received from bandit \( i \) at time \( t \) during its activation. Rewards are discounted at a fixed rate \( r > 0 \), compounded continuously. We take \( X^i_t \to \mathbb{F}^i \)-adapted, and denote the collection of reward processes as \( \mathcal{X} \).
Additionally, to each bandit we associate a commitment process \( c^i = \{c^i_k\}_{k \geq 0} \). Commitment processes function in the following way: when bandit \( i \) is activated for the \( k \)th time, it must be continuously activated for a duration of \( c^i_k(\omega^i) \in \mathbb{R}^+ \) before the controller is again free to activate from the \( N \) bandits as she chooses. We will refer to \( c^i_k \) as the duration of the \( k \)th commitment epoch of bandit \( i \). We denote the collection of commitment processes as \( C \).

For each bandit, it is convenient to define a sequence \( \{C^i_k\}_{k \geq 0} \) of process epoch times by
\[
C^i_0 = 0, \quad C^i_k = \sum_{k' = 0}^{k-1} c^i_{k'},
\]
and
\[
\mathbb{E}^i \left[ \int_0^\infty e^{-rt} |X^i_t| dt \right] < \infty.
\]

The former restriction implies the total time of each bandit is infinite, and the latter again implies that the expected total reward from any one bandit is finite.

A control policy \( \pi \) now becomes a right-continuous stochastic process on the global space, such that \( \pi(s) = i \) means the controller is activating bandit \( i \) at global time \( s \). Note, in continuous time, the use of “round” to describe a unit of global time is no longer sensible.

Given a policy \( \pi \), we may define continuous time versions of \( T^i_\pi(s) \) and \( S^i_\pi(t) \) as
\[
T^i_\pi(s) = \int_0^s 1_{\{\pi(s') = i\}} ds',
\]
and, utilizing the above,
\[
S^i_\pi(t) = \inf\{s \geq 0 : T^i_\pi(s) \geq t\}.
\]

Extending the ideas of Section 2.1, we define a policy \( \pi \) to be non-anticipatory if for \( s \geq 0, \pi(s) \) is measurable with respect to \( \bigotimes_{i=1}^N \mathcal{F}(T^i_\pi(s)) \). However, we are only interested in the set \( \mathcal{P}_C \) of non-anticipatory policies \( \pi \) that satisfy the commitment constraints, that is, for each bandit \( i \), for each \( k \geq 0 \),
\[
S^i_\pi(C^i_k + t) = S^i_\pi(C^i_k) + t, \quad \text{for } 0 \leq t < c^i_k, \\
\pi(S^i_\pi(C^i_k + t)) = i, \quad \text{for all } 0 \leq t < c^i_k.
\]

In parallel with the previous sections, we let \( V^C_\pi(X, C) \) denote the value of a policy, the expected total reward given the reward-commitment pair \((X, C)\) under policy \( \pi \in \mathcal{P}_C \). We
may therefore express the value of a policy as

\[ V^C_\pi(X, C) = \mathbb{E} \left[ \int_0^\infty e^{-rs} X_\pi(s) ds | \mathcal{G}_0 \right] , \]  

(64)

relative to global time, or relative to local time as

\[ V^C_\pi(X, C) = \sum_{i=1}^N \mathbb{E} \left[ \int_0^\infty e^{-rS^i(t)} X^i_t dt | \mathcal{G}_0 \right] . \]  

(65)

The problem the controller faces is to determine a policy \( \pi^* \in \mathcal{P}_C \) that is optimal in the sense that, for any other \( \pi \in \mathcal{P}_C \),

\[ V^C_\pi(X, C) \leq V^C_\pi^*(X, C) (\mathbb{P} - \text{a.e.}) . \]  

(66)

We resolve this by reducing this commitment model to the previous depreciation model. That is, we have the following result.

**Theorem 5:** Given a collection of bandits \( \{(\Omega^i, \mathcal{F}^i, \mathbb{P}^i, \hat{\mathbb{P}}^i)\}_{1 \leq i \leq N} \) in continuous time, with the reward-commitment pair \((X, C)\), there exists a collection of bandits \( \{(\hat{\Omega}^i, \hat{\mathcal{F}}^i, \hat{\mathbb{P}}^i, \hat{\mathbb{P}}^i)\}_{1 \leq i \leq N} \) in discrete time, with the reward-discount pair \((\hat{X}, \hat{B})\), such that for any policy \( \pi \in \mathcal{P}_C \), there exists a policy \( \hat{\pi} \in \mathcal{P} \) such that

\[ V^C_\pi(X, C) = V^C_{\hat{\pi}}(\hat{X}, \hat{B}) (\mathbb{P} - \text{a.e.}) , \]  

(67)

and vice versa, such a \( \pi \in \mathcal{P}_C \) exists for any \( \hat{\pi} \in \mathcal{P} \).

**Proof:** The construction is fairly natural, amounting to a translation from the continuous “commitment time” to a discrete “decision time”, in which one unit of time indicates a single decision by the controller. To the \( k \)th activation of bandit \( i \), we associate the following quantities:

\[ \hat{X}^i_k := \mathbb{E}^i \left[ \int_0^{c^i_k} e^{-rt} X^i_{C^i_k+t} dt | \mathcal{F}^i(C^i_k) \right] , \]

\[ \hat{\beta}^i_k := e^{-rc^i_k}, \]

\[ \hat{\mathcal{F}}^i(k) := \mathcal{F}^i(C^i_k). \]  

(68)

These represent, respectively, the expected reward collected during the \( k \)th commitment period, the depreciation incurred on all future rewards due to the \( k \)th commitment period, and the information available about bandit \( i \) prior to the \( k \)th decision to activate it. These define, therefore, a discrete-time reward process \( \hat{X}^i = \{\hat{X}^i_k\}_{k \geq 0} \), a discount factor sequence \( \hat{\beta}^i = \{\hat{\beta}^i_k\}_{k \geq 0} \), and a discrete-time filtration \( \hat{\mathbb{P}}^i = \{\hat{\mathcal{F}}^i(k)\}_{k \geq 0} \) to which they are both adapted. We may then define the total depreciation sequence \( \hat{\alpha}^i = \{\hat{\alpha}^i_k\}_{k \geq 0} \) by Eq. (1)
using $\dot{\beta}^i$, which yields $\dot{\alpha}^i_k = e^{-rC^i_k}$. Observe then that we have the following:

$$
\lim_{k \to \infty} \dot{\alpha}^i_k = \lim_{k \to \infty} e^{-rC^i_k} = 0 \, (\mathbb{P}^i - \text{a.e.}),
$$

(69)

and

$$
\mathbb{E}^i \left[ \sum_{k=0}^{\infty} \dot{\alpha}^i_k | \hat{X}^i_k \right] \leq \mathbb{E}^i \left[ \sum_{k=0}^{\infty} \int_{C^i_k}^{C^i_{k+1}} e^{-rt} X^i_t dt \right] = \mathbb{E}^i \left[ \int_0^\infty e^{-rt} X^i_t dt \right] < \infty.
$$

(70)

This gives us an instance $(\hat{X}, \hat{B})$ of the depreciation model on the bandits $\{(\Omega^i, \mathcal{F}^i, \mathbb{P}^i, \hat{B}^i)\}_{1 \leq i \leq N}$.

Taking a policy $\pi \in \mathcal{P}_C$ on $(X, \mathbb{C})$, we may translate it into a discrete-time policy $\hat{\pi} \in \mathcal{P}$ on $(\hat{X}, \hat{B})$ in the following way: if the $h^{th}$ decision made under policy $\pi$ is to activate bandit $i$, we have $\hat{\pi}(h) = i$. Similarly, given a $\hat{\pi} \in \mathcal{P}$, it may be translated into a policy $\pi \in \mathcal{P}_C$ by taking $\pi$ to be the continuous-time extension of $\hat{\pi}$ that satisfies the commitment periods. In this way, we may go back and forth between models. Note then, we may define the a policy-dependent depreciation sequence by $\hat{\alpha}^i_\pi(k) = e^{-rS^i_k(C^i_k)}$, the total depreciation incurred on the $k^{th}$ commitment period by all previous activation.

Hence, given $\pi \in \mathcal{P}_C$ and the corresponding $\hat{\pi} \in \mathcal{P}$, or vice versa,

$$
V^C_\pi(X, \mathbb{C}) = \sum_{i=1}^{N} \mathbb{E} \left[ \int_0^\infty e^{-rS^i_t(t)} X^i_t dt | \mathcal{G}_0 \right] = \sum_{i=1}^{N} \mathbb{E} \left[ \sum_{k=0}^{C^i_{k+1}} e^{-rS^i_k(t)} X^i_t dt | \mathcal{G}_0 \right] = \sum_{i=1}^{N} \mathbb{E} \left[ \sum_{k=0}^{c^i_k} e^{-rS^i_k(C^i_k + t)} X^i_{C^i_k + t} dt | \mathcal{G}_0 \right] = \sum_{i=1}^{N} \mathbb{E} \left[ \sum_{k=0}^{c^i_k} e^{-rS^i_k(C^i_k + t)} X^i_{C^i_k + t} dt | \mathcal{G}_0 \right] = \sum_{i=1}^{N} \mathbb{E} \left[ \sum_{k=0}^{c^i_k} \hat{\alpha}^i_\pi(k) \hat{X}^i_k | \mathcal{G}_0 \right] = V_{\hat{\pi}}(\hat{X}, \hat{B})(\mathbb{P} - \text{a.e.}).
$$

(71)

This connection between the commitment model and the depreciation model can also be used to construct a continuous-time commitment model that is equivalent of a given depreciation model, essentially by inverting the relationships in Eq. (68), for example, taking commitment durations as $c^i_k = -\ln(\beta^i_k)/r$ (for a context specific choice of $r > 0$). In this
way, it can be shown that the two models are equivalent. This construction will not be explored further here.

Theorem 5 implies that any instance of the commitment model may be solved by translating it into the corresponding depreciation model and solving it there, via the restart-in-state index.

The model presented above was taken to be in continuous time. A discrete-time model is also possible, in which case future rewards are discounted by some fixed factor $\beta \in (0, 1)$ per round, and each commitment duration refers to a fixed number of activations, $c_k \in \mathbb{N}$. Everything done in the continuous case goes through for the discrete with $\hat{\mathcal{X}}(k)$ and $\hat{\beta}_k^i$ defined by the following as an extension of Eq. (68):

$$
\hat{X}^i(k) = \mathbb{E}[\beta^k \mathcal{X}(C_k^i + k) | \mathcal{F}(C_k^i)],
$$

$$
\hat{\beta}_k^i = \beta \cdot c_k^i.
$$

8. ASSUMPTIONS C AND C’, AND ZENO’S HYPOTHESIS

In both of the models presented in this paper, the individual bandits were restricted in two ways, by Assumptions C and C’. First we assumed the integrability conditions of Eqs (3) and (60), effectively requiring that the total expected reward from each bandit be finite. This requirement was taken simply to ensure a degree of realism in the model.

Additionally the restrictions of Eqs (2) and (59):

$$
\prod_{t=0}^{\infty} \beta^t_i = 0, \sum_{k=0}^{\infty} c_k^i = \infty (\mathbb{P}^i - \text{a.e.}) \forall i,
$$

were respectively placed on each model as a matter of both necessity and convenience. Through the relationships between the models defined in Eq. (68), these assumptions were shown to be equivalent in Eq. (69).

The mathematical necessity of these assumptions arose in the proof of Theorem 3 in Eq. (54), in proving that the values of finite-step $\rho$-greedy policy improvements converged to the value of the infinite $\rho$-greedy policy. The key point was that if a bandit were only activated finitely many times under a policy, the remaining rewards from that bandit were discounted to 0 by the assumption of Eq. (59) that the total commitment time must be infinite. The controller would then be faced with a Zeno-type problem of making the “next” activation decision after an infinite sequence of activation decisions. Taking the assumption of Eq. (59) that the total commitment time must be infinite sidesteps any potential entanglements of this type, guaranteeing that only finitely many decisions can be made in finite time.

The convenience of these assumptions can be seen most directly in the continuous-time commitment model. If the commitment durations of a given bandit summed to some finite value, a situation may arise in which the controller may make infinitely many decisions in finite time. For instance taking $c_k^i = 2^{-k}$ for some $i$, the decisions to activate each commitment period of bandit $i$ would only take 2 units of time total. The controller would then be faced with a Zeno-type problem of making the “next” activation decision after an infinite sequence of activation decisions. Taking the assumption of Eq. (59) that the total commitment time must be infinite sidesteps any potential entanglements of this type, guaranteeing that only finitely many decisions can be made in finite time.

From this perspective, Assumptions C and C’ can be interpreted similarly for each model: each ensures that after an infinite sequence of decisions, the value of uncollected rewards is zero – either through the infinite delay of the commitment model, or the discounting to 0 of the depreciation model. Hence, any decision made “after” an infinite sequence of decisions contributes nothing to the total and can be ignored.

We note that in the case of a constant discount factor for all bandits and times, that is, $\beta^i = \beta$, and for commitment times bounded from below, that is, $c_k^i > \delta$ for some $\delta > 0$, the hypotheses of Assumption C and C’ are automatically satisfied.
Acknowledgement

Research partially supported by the National Science Foundation, grant CMMI 14-50743 and the Rutgers Business School Research Resources Committee.

References

Again, the optimal choice of $\alpha$ is maximized taking $\alpha$ such that
\[ \sum_{t=0}^{\tau-1} \alpha_t Z_t | J_0 \leq \alpha_0 \text{ess sup}_{\tau' \in \hat{J}(0)} \left[ \sum_{t=0}^{\tau'-1} Z_t | J_0 \right] (\mathbb{P} \text{ - a.e.}). \]  

(A.1)

Proof: If the result can be shown for infinite $\tau$, the case for arbitrary $\tau > 0$ follows simply, defining a new sequence $\hat{\alpha}_t = \alpha_t 1_{\{\tau > t\}}$. Hence it suffices to take $\tau = \infty$.

This result follows straightforwardly in the case the $\{Z_t\}$ sequence has finitely many terms. The result is trivial if there is a single term. In the case of two terms,
\[ \alpha_1 = \begin{cases} \alpha_0 & \text{if } E \left[ Z_1 | J_1 \right] > 0, \\ 0 & \text{if } E \left[ Z_1 | J_1 \right] \leq 0. \end{cases} \]  

(A.3)

In short, if the remaining contribution is positive, make $\alpha_1$ as large as possible (i.e. equal to $\alpha_0$), else make $\alpha_1$ as small as possible (i.e. equal to 0). Either way, this factors simply, giving in all cases,
\[ E \left[ \alpha_0 Z_0 + \alpha_1 Z_1 | J_0 \right] \leq \alpha_0 \text{ ess sup}_{\tau' \in \hat{J}(0), \tau \leq 2} E \left[ \sum_{t=0}^{\tau-1} Z_t | J_0 \right] (\mathbb{P} \text{ - a.e.}). \]  

(A.4)

We may apply this result inductively in the following way. Allowing $N < \infty$ terms,
\[ E \left[ \sum_{t=0}^{N-1} \alpha_t Z_t | J_0 \right] = E \left[ \alpha_0 Z_0 + \sum_{t=1}^{N-1} \alpha_t Z_t | J_1 \right] | J_0 \]  

\[ \leq E \left[ \alpha_0 Z_0 + \alpha_1 \text{ ess sup}_{\tau \in \hat{J}(1), \tau \leq N} E \left[ \sum_{t=1}^{\tau-1} Z_t | J_1 \right] | J_0 \right]. \]  

(A.5)

Again, the optimal choice of $\alpha_1$ is the $J_1$-measurable random variable given by $\alpha_0$ if the essential sup is positive and 0 if it is negative. In either case, we have the following,
\[ E \left[ \sum_{t=0}^{N-1} \alpha_t Z_t | J_0 \right] \leq \alpha_0 \text{ ess sup}_{\tau' \in \hat{J}(0), \tau \leq N} E \left[ \sum_{t=0}^{\tau-1} Z_t | J_0 \right] (\mathbb{P} \text{ - a.e.}). \]  

(A.6)

This extends immediately to all finite, $J_0$-measurable $N$. We next extend the finite result to the infinite.
The initial assumption \( \mathbb{E} \left[ \sum_{t=0}^{\infty} |Z_t| \right] < \infty \) implies that \( \mathbb{E} \left[ \sum_{t=0}^{\infty} |Z_t| \right] < \infty \) \((\mathbb{P} - \text{a.e.})\). Thus for any fixed \( \epsilon > 0 \), we may find a \( \mathcal{F}_0 \)-measurable and finite \( N > 0 \) \((\mathbb{P} - \text{a.e.})\) such that

\[
\mathbb{E} \left[ \sum_{t=N}^{\infty} |Z_t| \right] \leq \epsilon \quad (\mathbb{P} - \text{a.e.}).
\]

(A.7)

Then,

\[
\mathbb{E} \left[ \sum_{t=0}^{\infty} \alpha_t Z_t \mid \mathcal{F}_0 \right] = \mathbb{E} \left[ \sum_{t=0}^{N-1} \alpha_t Z_t \mid \mathcal{F}_0 \right] + \mathbb{E} \left[ \sum_{t=N}^{\infty} \alpha_t Z_t \mid \mathcal{F}_0 \right] \\
\leq \mathbb{E} \left[ \sum_{t=0}^{N-1} \alpha_t Z_t \mid \mathcal{F}_0 \right] + \mathbb{E} \left[ \sum_{t=N}^{\infty} |Z_t| \right] \\
\leq \alpha_0 \text{ ess sup}_{\tau \in \mathbb{S}(0), \tau \in \mathbb{S}(N)} \mathbb{E} \left[ \sum_{t=0}^{\tau-1} Z_t \mid \mathcal{F}_0 \right] + \epsilon (\mathbb{P} - \text{a.e.})
\]

(A.8)

The last step follows simply, extending the set of \( \tau \) in question from stopping times at most \( N \), to all possible stopping times in \( \mathbb{S}(0) \). Since the above holds for all \( \epsilon > 0 \), the result is immediate. By the previous remarks, \( \tau = \infty \) is sufficient, and the proof is complete.

\section*{APPENDIX B}

\subsection*{B.1. Integration Exchange in Theorem 1}

In the proof of Theorem 1, it remains to rigorously demonstrate the exchange of essential suprema in the third step of relations (21), that is, that for \( Z_i = X_i - \rho (1 - \beta_i) \),

\[
\text{ess sup}_{\tau \in \mathbb{H}_1^1(t_0)} \mathbb{E} \left[ \sum_{t=t_0}^{\tau-1} \alpha_i^1 Z_i \mid \mathcal{H}_1^1(t_0) \right] \leq \text{ess sup}_{\tau \in \mathbb{F}^1(t_0)} \mathbb{E} \left[ \sum_{t=t_0}^{\tau-1} \alpha_i^1 Z_i \mid \mathcal{F}_1^1(t_0) \right] \quad (\mathbb{P} - \text{a.e.}).
\]

(B.1)

This essentially amounts to integrating out the independent bandits, since \( Z_i \) depends only on bandit \( i \). For compactness of argument, we take \( N = 2, i = 1 \), though the following argument generalizes to arbitrary bandits in the obvious way.

For notational compactness, we define \( W_i^1 = \sum_{t=t_0}^{t-1} \alpha_i^1 Z_i \).

Note that for any set \( A \in \mathcal{H}_1^1(t_0) \), and any \( \tau \in \mathbb{H}_1^1(t_0) \),

\[
\mathbb{E} \left[ 1_A \mathbb{E} \left[ W_i^1 \mid \mathcal{H}_1^1(t_0) \right] \right] = \mathbb{E} \left[ 1_A W_i^1 \right].
\]

(B.2)

Taking \( A \) as a rectangle in \( \mathcal{H}_1^1(t_0) \), \( A = A_1 \times A_2 \), observe that \( A_1 \in \mathcal{F}_1^1(t_0) \). The indicator may be decomposed as \( 1_A(\omega) = 1_{A_1}(\omega^1) 1_{A_2}(\omega^2) \). As \( \sum_{t=0}^{\infty} \alpha_i^1 Z_t < \infty \) almost surely, we may exchange the expectation over the product space for an iterated expectation.

\[
\mathbb{E} \left[ 1_A W_i^1 \right] = \mathbb{E}^2 \left[ 1_{\mathcal{A}_1} 1_{\mathcal{A}_2} W_i^1 \right] \\
= \mathbb{E}^2 \left[ 1_{\mathcal{A}_2} \mathbb{E}^1 \left[ 1_{\mathcal{A}_1} W_i^1 \right] \right] \\
= \mathbb{E}^2 \left[ 1_{\mathcal{A}_2} \mathbb{E}^1 \left[ 1_{\mathcal{A}_1} W_i^1 \mid \mathcal{F}_1^1(t_0) \right] \right].
\]

(B.3)

Observe that, while \( \tau \) (being an \( \mathbb{H}_1 \)-stopping time) may have a dependence on \( \Omega^2 \), inside the iterated integral with the dependence on \( \Omega^2 \) fixed, it is a \( \mathcal{F}_1 \)-stopping time. Hence, we have the
\[ E \left[ 1_A W^1_\tau \right] \leq E^2 \left[ 1_{A_1} E^1 \left[ 1_{A_2} \text{ess sup}_{\hat{\tau} \in \hat{F}^1(t_0)} \left[ W^1_\tau | \mathcal{F}^1(t_0) \right] \right] \right] \]
\[ = E^2 \left[ E^1 \left[ 1_{A_1} 1_{A_2} \text{ess sup}_{\hat{\tau} \in \hat{F}^1(t_0)} \left[ W^1_\tau | \mathcal{F}^1(t_0) \right] \right] \right] \]
\[ = E \left[ 1_A \text{ess sup}_{\hat{\tau} \in \hat{F}^1(t_0)} \left[ W^1_\tau | \mathcal{F}^1(t_0) \right] \right]. \quad (B.4) \]

Hence, for all rectangles \( A \in \mathcal{H}^1_{\pi}(t_0) \),
\[ E \left[ 1_A E \left[ W^1_\tau | \mathcal{H}^1_{\pi}(t_0) \right] \right] \leq E \left[ 1_A \text{ess sup}_{\hat{\tau} \in \hat{F}^1(t_0)} \left[ W^1_\tau | \mathcal{F}^1(t_0) \right] \right]. \quad (B.5) \]

This extends via a monotone-class-type argument, cf. Chung [11], to all \( A \in \mathcal{H}^1_{\pi}(t_0) \). Hence, for all \( \tau \in \hat{H}^1_{\pi}(t_0) \),
\[ E \left[ W^1_\tau | \mathcal{H}^1_{\pi}(t_0) \right] \leq \text{ess sup}_{\hat{\tau} \in \hat{F}^1(t_0)} \left[ W^1_\tau | \mathcal{F}^1(t_0) \right] (\mathbb{P} - \text{a.e.}). \quad (B.6) \]

This establishes the result.