Received: 23 July 2023 | Accepted: 28 July 2023 | Published on: 16 August 2023

DOI: 10.1002/nav.22145

RESEARCH ARTICLE

WILEY

Optimal activation of halting multi-armed bandit models

Wesley Cowan[1] | **Michael N. Katehakis**[2] | **Sheldon M. Ross**[3]

[1]Computer Science Department, Rutgers University, Piscataway, New Jersey, USA
[2]Management Science and Information Systems Department, Rutgers University, Piscataway, New Jersey, USA
[3]Systems Engineering Department, University of Southern California, Los Angeles, California, USA

Correspondence
Michael N. Katehakis, Management Science and Information Systems Department, Rutgers University, 100 Rockafeller Road, Piscataway, NJ, 08854, USA.
Email: mnk@rutgers.edu

Funding information
National Science Foundation, Grant/Award Numbers: CMMI-1662629, CMMI-1662442

Abstract

We study new types of dynamic allocation problems the *Halting Bandit* models. As an application, we obtain new proofs for the classic Gittins index decomposition result compare Gittins (Journal of the Royal Statistical Society, Series B, 1979, 41, 148–177), and recent results of the authors in Cowan and Katehakis (Probability in the Engineering and Informational Sciences, 2015, 29, 51–76).

KEYWORDS

adaptive systems, autonomous reasoning and learning, dynamic data driven systems, machine learning, Markovian decision processes

1 | INTRODUCTION

We investigate a class of *Halting Bandit* models, where at every time step a controller must choose which project out of a fixed collection to activate, and at some (stochastic) time, when sufficient time and effort has been invested in a given project or process, it will be completed or "halt." Additionally, halting may be considered a catastrophic event, such as a project breaking down. These halting events allow bandits to be "singled out"—receiving rewards from successful bandits and paying costs for unsuccessful bandits. This singling out of projects based on state status is novel; prior results focused mainly on maximizing cumulative collective payouts compare model (CCP) of Section 5.

In this article, we consider the following models for maximizing terminal rewards (or minimizing terminal costs): two versions of expected *terminal solo payout*, taken to be a reward dependent on the last (ultimate) or second to last (penultimate) state of the first bandit to halt successfully; the *terminal collective payout* reward, taken to be a reward dependent on the final states of *all* bandits at the first halting; the terminal *non-halting costs*, taken to be a cost incurred by all bandits that *failed* to halt; the terminal *collective profit*, taken to be a reward from the successfully halted bandit less the cost incurred by bandits that failed to halt. After establishing these results, we consider the same model in the framework of cumulative rewards, rather than terminal, when bandits are taken to generate rewards each time they are activated until halting. We use a standard technique to reduce these models to corresponding terminal halting models and in this way, we recover prior results in Cowan and Katehakis (2015) and hence the celebrated Gittins' decomposition compare Gittins (1979).

The central results presented here, the derivation of optimal policies for the terminal solo payout and terminal collective payout models, rests on establishing a correspondence between the two payout models; essentially, the game where every bandit contributes to the total reward can be replaced by an equivalent game where only a single bandit contributes to the terminal reward. This gives further insight into why classical bandit decomposition results work compare Chakravorty and Mahajan (2014), Gittins et al. (2011), Mahajan and Teneketzis (2008), Ishikida and Varaiya (1994), Weber (1992).

For related work we first note that for the finite state Markov chain version of the CCP model of Section 5, Sonin (2008) introduced an equivalent formulation of the indices derived herein in order to derive an efficient algorithm for the calculation of the indices for all states of the Markov chain. The basic idea of this article's *generalized indices* was to use a common Markov decision processes theory interpretation of of the expected discounted total reward with a discount factor β where the state space is complemented by an absorbing state x^* and new transition probabilities that are defined as follows. The probability of entering an absorbing state x^* in one step for any state $y \neq x^*$ ("probability of termination") is equal to $1 - \beta$, and all other initial transition probabilities are multiplied by β. In other words, β is the probability of "survival," or not "halting" herein. Sonin (2008) considered variable probabilities of survival $\beta(x)$ and defined a generalized index $\alpha(x)$ taken to be the maximum ratio of the expected discounted total reward up to the time τ of halting ("termination") per chance of termination at the time τ of halting. He established that for nonconstant discount factors the the equality of the new *generalized index* with the *retirement index* of Whittle (1980) and the restart index of Katehakis and Veinott Jr. (1987), thus he argued that the true meaning of the Gittins index is given by its expression as a ratio of the expected discounted total reward up to the time τ of halting ("termination") per chance of termination at the time τ of halting, and pointed out its relation with the work in Mitten (1960). These results can be extended along the lines of El Karoui and Karatzas (1994) who established the *restart* representation of the Gittins index in a continuous time framework without making further use of it. Additional results connecting the Sonin indices with other problems of stochastic optimization are given in Bank and El Karoui (2004) and in Sonin and Steinberg (2016).

For other related work we refer to Szepesvári (2010), Slivkins (2019), Dumitriu et al. (2003), Katta and Sethuraman (2004), and to Stadje (1995), Pinedo and Rammouz (1988), Pinedo (2012), Glazebrook et al. (2007), Negoescu et al. (2018), Villar et al. (2015) Glazebrook et al. (2014), Denardo et al. (2007), Katehakis and Rothblum (1996), Katehakis and Derman (1986), and Skitsas et al. (2022), Talebi et al. (2021), Cowan et al. (2018).

The article presents a collection of results, organized sequentially to build off each other to the final result. It is worth outlining this explicitly at the start, with a roadmap:

Section 2 gives the underlying mathematical framework of the discussion to follow, to guarantee the necessary processes and control processes are well defined. Ultimately, the key point of these results is this: the relation between the "single payout" model and the "collective payout" model reveals why the contributions of each bandit in the original formulation can be considered individually, by expressing the total game in terms of an equivalent one where only one bandit gives rewards. Section 3 considers a simplified or "solo-payout" model, where only the bandit that halts (or breaks) yields a reward to the controller. These solo payout model bandits have a simple optimal policy. In Section 4, we consider a collective-payout model (rewards from all bandits) and derive equivalent (or bounding) solo-payout models. The optimal solo-payout policy on the equivalent (or bounding) model is then shown to give an equivalent reward to a simple index policy on the collective-payout model, yielding a proof of optimality. In Section 5, a number of alternative payout models are introduced, and all are shown to be equivalent to the solved collective-payout model. The classical Gittins formulation is recovered herein. Some proofs, technical and uninstructive, are relegated to Section 6.

2 | PROBLEM FORMULATION

2.1 | Probability framework

A controller is presented with a finite collection of $N \geq 2$ probability spaces, $(\Omega^i, \mathcal{F}^i, \mathbb{P}^i, \mathbb{F}^i)$, for $1 \leq i \leq N$, representing N environments in which experiments will be performed or rewards collected—the "bandits," or "projects." To each space, we associate an $\mathbb{F}^i$-adapted *reward process* $X^i = \{X^i_t\}_{t\geq 0}$. For $t \in \{0, 1, \dots\}$, we take $X^i_t (= X^i_t(\omega^i)) \in \mathbb{R}$ to represent the reward (or state) attained from the ith bandit on its tth activation. We denote the collection of these processes as $\mathbb{X}$.

Additionally, to each bandit, we associate an $\mathbb{F}^i$-stopping time $\sigma^i > 0$, the "halting time" of the bandit, so that at the σ^ith activation of bandit i, we take the bandit to be stopped, and no longer capable of being activated. Note, σ^i represents the number of times bandit i can be activated, so the last activation of bandit i occurs at bandit-time $\sigma^i - 1$, and at bandit time σ^i, the bandit is permanently stopped. On every activation prior to halting, we assume there is a positive probability of halting. We take the first of any bandit halting to halt the entire decision process (game).

In what follows, we reserve the term "round" to differentiate global controller time (denoted with s), when the controller must decide which bandit to activate, from local bandit times (denoted by t), indicating the current total activations of a given bandit. In each round, the controller activates a bandit, advancing both its local time and the global time by one time step. All bandits begin at local time 0, and advance only on activation, that is, in every round *unactivated bandits remain frozen*. As stated, the game halts upon the first halting of any bandit. The controller needs a *control policy* π, that specifies, at each round s of global time, which bandit to activate.

We embed these bandits in a larger product 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. This "global" probability space is necessary for making sure processes at the controller level (e.g., the policy for bandit activation) are well defined. This construction captures the first key aspect of the model: *the bandits are mutually independent* (e.g., X^i, X^j are independent relative to $\mathbb{P}$ for $i \neq j$). Expectations relative to the local space, that is, bandit i, will be denoted $\mathbb{E}^i$, while expectations relative to the global space are simply $\mathbb{E}$.

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

We make the following assumptions.

Assumption 1. For each bandit i

$$\mathbb{E}^i\left[\sup_{n\geq 0} |X_n^i|\right] < \infty. \tag{1}$$

Assumption 2. For each bandit i the following are true.

$$a) \quad \mathbb{P}^i(\sigma^i < \infty) = 1, \tag{2}$$

$$b) \quad \mathbb{P}^i(\sigma^i = t+1 | \mathcal{F}^i(t)) > 0, \text{ for all } t < \sigma^i, \quad (\mathbb{P}^i, \mathbb{P}\text{-a.e.}). \tag{3}$$

Remark 2. Note, the above assumptions, while technical in statement, have natural interpretations: (2.a) each bandit will halt after finite activations, almost surely; (2.b) at any time prior to halting, there is non-zero probability of halting on the next activation.

A *control policy* π, is a stochastic process on $(\Omega, \mathcal{G}, \mathbb{P})$ that specifies, at each round s of global time, which bandit to activate and collect from, for example, $\pi(s)(= \pi(s, \omega)) = i$ activates bandit i at round s. We restrict attention to the set of policies $\mathcal{P}$ defined to be *non-anticipatory*, that is, the choice of which bandit to activate at round s does not depend on outcomes that have not yet occurred, or information not yet available.

A policy π defines $T_\pi^i(s)$ the *π-local time* of bandit i just prior to the sth round under it, that is, $T_\pi^i(0) = 0$, and for $s > 0$,

$$T_\pi^i(s) = \sum_{s'=0}^{s-1} \mathbb{1}\{\pi(s') = i\}. \tag{4}$$

Note, this gives as a result that at global time s, the sum of all the local times must be s, that is,

$$\begin{aligned} T_\pi^1(s) + T_\pi^2(s) + \dots + T_\pi^N(s) &= \sum_{i=1}^{N}\sum_{s'=0}^{s-1} \mathbb{1}\{\pi(s') = i\} \\ &= \sum_{s'=0}^{s-1}\sum_{i=1}^{N} \mathbb{1}\{\pi(s') = i\} = \sum_{s'=0}^{s-1} 1 = s, \end{aligned} \tag{5}$$

where the inner sum reduces to 1 since exactly one bandit is activated each round.

It is convenient to define the global time analog, $T_\pi(s) = T_\pi^{\pi(s)}(s)$ to denote the current π-local time of the bandit activated at round s under policy π. This will allow us to define concise global time analogs of several processes. An important example of such a process is the *global reward process* X_π on $(\Omega, \mathcal{G}, \mathbb{P})$ defined as

$$X_\pi(s) = X_{T_\pi(s)}^{\pi(s)},$$

giving the reward available from collection $\mathbb{X}$ under policy π, which is to be received if the game halts at round s.

To be able to translate between global time and local times, when the controller operates according to a policy π, we define the random variables $S_\pi^i(t)$ to represent *the round at which bandit i is activated for the tth time*, that is,

$$\begin{aligned} S_\pi^i(0) &= \inf\{s \geq 0 : \pi(s) = i\}, \\ S_\pi^i(t+1) &= \inf\{s > S_\pi^i(t) : \pi(s) = i\}. \end{aligned} \tag{6}$$

Utilizing this notation, we may define a *global halting time* σ_π, that is, the first round under policy π at which one of the bandits has halted, ending the game:

$$\sigma_\pi = \min_i \{S_\pi^i(\sigma^i - 1)\} + 1. \tag{7}$$

Remark 3. To clarify the above definition, note that $S_\pi^i(0)$ is the time that a policy first activates bandit i, advancing it from local time 0 to local time 1. So $S_\pi^i(\sigma^i - 1)$ is the global time round at which the policy π advances bandit i from local time $\sigma^i - 1$ to local time σ^i, halting that bandit. The expression above for σ_π therefore identifies the first global round at which no further activations will be made, because one bandit has been halted.

In what follows, for a given policy π, we take the final reward the controller receives to be a function of the last rewards of the game, generally a linear combination of $\{X_{T_\pi^i(\sigma_\pi)}^i\}_{1\leq i\leq N}$, or in the penultimate model a function of the second to last rewards. To maximize her expected reward, in every round the controller's decision of which bandit to activate must balance not only the current rewards of each bandit, but also the probability of halting that bandit and in doing so ending the game—losing all potential future rewards.

2.2 | 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 a given time. The following statements are somewhat technical, but necessary for the purpose of making sure all relevant processes are mathematically well-defined, and that our control processes do not depend on information they should not have access to. Ultimately, the optimal policy results of Theorems 1 and 4 (essentially stating the simplicity of the optimal policy) demonstrate that in the optimal policy, any decision to activate a given bandit depends only on information from other bandits individually, thus rendering these filtrations unnecessary under an optimal policy. However, these extended filtrations are a technical necessity for the proof of Theorem 4.

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 σ-algebra $\mathcal{F}^i(t)$ representing the local information available about bandit i at local time t, such as (but not limited to) the process history of X^i. Taking X^i as $\mathbb{F}^i$-adapted as we do, we have $\sigma(X_0^i, X_1^i, \dots, X_t^i) \subset \mathcal{F}^i(t)$.

At round s, all information available to the controller is determined by the state of each bandit at that round, that is, acting under a given policy π until round s, the global information available at round s is given by the σ-algebra $\otimes_{i=1}^N \mathcal{F}^i(T_\pi^i(s))$. 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 σ-algebra, that is, determined by the information available at round s. Weaker definitions of non-anticipatory, such as dependence on random events, for example, coin flips, are addressed in Section 6. It is convenient to define the initial global σ-algebra $\mathcal{G}_0 = \otimes_{i=1}^N \mathcal{F}^i(0)$, representing the initial information available from each bandit, which is independent of policy π.

Additionally, given a policy π, it is necessary to define a set of policy-dependent filtrations in the following way: let $\mathbb{H}_\pi^i = \{\mathcal{H}_\pi^i(t)\}_{t\geq 0}$, where $\mathcal{H}_\pi^i(t) = \otimes_{j=1}^N \mathcal{F}^j(T_\pi^j(S_\pi^i(t)))$ represents the total information available to the controller about all bandits, prior to the tth activation of bandit i under π. 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 $T_\pi^i(S_\pi^i(t)) = t$, $\mathcal{H}_\pi^i(t)$ 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 $\mathbb{H}_\pi^i$-stopping times, compare Remark 1.

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

The following simple example illustrates the random variables we have defined in this section.

Example 1. Take $N = 2$ bandits, independent geometric stopping times σ^i with

$$\mathbb{P}^i(\sigma^i > t) = \beta_i^t, \text{ for } t = 0, 1, \dots$$

for some constants $\beta_i \in (0,1)$, $i = 1, 2$, and consider a cyclic policy $\pi^1(t) = 1$ for $t = 0, 2, \dots$, and $\pi^1(t) = 2$ for $t = 1, 3, \dots$ Under the policy π^1 for any sample path for which $\sigma^1 > 2$ and $\sigma^2 > 2$ we will have:

Thus is easy to see that under π^1 the expected total reward received from the two bandits is

$$V_{\pi^1}(\mathbb{X}) = \mathbb{E}\big[X_0^1 + \beta_1 X_0^2 + \beta_1\beta_2 X_1^1 + \beta_1^2\beta_2 X_1^2 + \beta_1^2\beta_2^2 X_2^1 + \cdots\big].$$

Note also that:
$S_{\pi^1}^1(0) = \inf\{s > 0 : \pi^1(s) = 1\} = 0$, $S_{\pi^1}^1(1) = \inf\{s > S_{\pi^1}^1(0) : \pi^1(s) = 1\} = 2$, $S_{\pi^1}^1(2) = 4$
and
$S_{\pi^1}^2(0) = \inf\{s > 0 : \pi^1(s) = 2\} = 1$, $S_{\pi^1}^2(1) = \inf\{s > S_{\pi^1}^2(0) : \pi^1(s) = 2\} = 3$ and so forth.

Finally note that under policy π^1 on the event $\{\sigma^1 \geq 2, \sigma^2 = 1\}$ bandit 2 causes the game to end at round $s = 2$ that is, the *global halting time* is

$$\sigma_{\pi^1} = \min\{S_{\pi^1}^1(\sigma^1 - 1), S_{\pi^1}^2(0)\} + 1 = 1 + 1 = 2,$$

since $S_{\pi^1}^1(\sigma^1 - 1) \geq S_{\pi^1}^1(1) = 2$.

s	$T_{\pi^1}^1$	$T_{\pi^1}^2$	$\pi^1(s)$	Reward	Probability of not stopping at s
0	0	0	1	X_0^1	$\beta_1 = \mathbb{P}^i(\sigma^1 > 1)$
1	1	0	2	X_0^2	$\beta_1\beta_2 = \mathbb{P}^i(\sigma^1 > 1, \sigma^2 > 1)$
2	1	1	1	X_1^1	$\beta_1^2\beta_2 = \mathbb{P}^i(\sigma^1 > 2, \sigma^2 > 1)$
3	2	1	2	X_1^2	$\beta_1^2\beta_2^2 = \mathbb{P}^i(\sigma^1 > 2, \sigma^2 > 2)$
4	2	2	1	X_2^1	$\beta_1^3\beta_2^2 = \mathbb{P}^i(\sigma^1 > 3, \sigma^2 > 2)$
⋮	⋮	⋮	⋮	⋮	⋮

3 | MAXIMIZING SOLO PAYOUTS: NON-INCREASING REWARDS

In this section, we consider the problem of maximizing the expected *penultimate* reward from the bandit that halts and ends the game. That is, if a bandit is activated and halts, stopping the game, the controller receives the reward that bandit offered prior to its last activation. Additionally, in this section, we assume that the reward processes of each bandit are non-increasing. In fact, under this restriction, we may even maximize the reward *almost surely*. This result, while intuitive, acts as the basis of all future optimality results herein.

We define the *penultimate solo payout* value of a policy π as,

$$\begin{aligned} V_\pi^{PSP}(\mathbb{X}) &= \mathbb{E}[X_\pi(\sigma_\pi - 1)|\mathcal{G}_0] \\ &= \sum_{i=1}^N \mathbb{E}\Big[\mathbb{1}\{i = \pi(\sigma_\pi - 1)\} X^i_{T^i_\pi(\sigma_\pi - 1)} |\mathcal{G}_0\Big]. \end{aligned} \tag{8}$$

Theorem 1 (A greedy, almost-sure result for non-increasing solo payout processes). *Given a collection of reward processes $\mathbb{X}$ such that for each i, X^i is almost surely non-increasing for $t < \sigma^i$, there exists a policy $\pi^* \in \mathcal{P}$ such that for any policy $\pi \in \mathcal{P}$,*

$$X_\pi(\sigma_\pi - 1) \leq X_{\pi^*}(\sigma_{\pi^*} - 1) \quad (\mathbb{P}\text{-a.e.}). \tag{9}$$

In particular, such a π^* is given by the following greedy rule: In each round $s \geq 0$, activate the bandit with the largest current value of X^i, that is,

$$\pi^*(s) = \quad \arg\max_i X^i_{T^i_{\pi^*}(s)}.$$

Proof. The proof proceeds by incremental improvements on an arbitrary policy.

Let $X_0^i = \max_j X_0^j$. Let $\pi \in \mathcal{P}$ be arbitrary, and define $S = S_\pi^i(0)$, the first round bandit i is activated under π. If i is never activated, we take S to be infinite.

From π, we construct a policy $\pi' \in \mathcal{P}$ as follows: π' activates bandits in the same order as π, but it advances the first activation of bandit i from round $s = S$ to round $s = 0$. That is,

$$\pi'(s) = \begin{cases} i & \text{for } s = 0, \\ \pi(s-1) & \text{for } s = 1, 2, \dots S, \\ \pi(s) & \text{for } s \geq S+1. \end{cases} \tag{10}$$

That is, after the initial round policy π' activates the bandit that policy π activated in the previous round, continuing this through the first round that π activates bandit i, then making the same choice in each round as does π. Policy π' is well-defined and in $\mathcal{P}$, as at every round s, the information available under π' about each bandit is greater than or equal to the information available under π at that round.

We next compare the performance of these two policies by cases.

In the case that $\sigma_\pi > S + 1$ $(= S_\pi^i(0) + 1)$, that is when the game halts under π *after* the first activation of bandit i, then there is no difference between the rewards returned by either policy, since both policies perform the same activations after time S (sample path-wise).

Similarly, if $\sigma_\pi = S + 1$, that is π halts *due to* the first activation of bandit i, the reward returned under π is X_0^i, and as bandit i halted on its first activation, the reward returned under π' is also X_0^i.

Finally the only situation in which π and π' differ in their returned rewards is when $\sigma_\pi \leq S$ and $\sigma^i = 1$.

Therefore, it follows from the above cases that:

$$\begin{aligned} X_{\pi'}(\sigma_{\pi'} - 1) - X_\pi(\sigma_\pi - 1) &= (X_{\pi'}(\sigma_{\pi'} - 1) - X_\pi(\sigma_\pi - 1)) \\ &\quad \mathbb{1}_{\{\sigma_\pi \leq S\}} \mathbb{1}_{\{\sigma^i = 1\}} \\ &= (X_0^i - X_\pi(\sigma_\pi - 1)) \\ &\quad \mathbb{1}_{\{\sigma_\pi \leq S\}} \mathbb{1}_{\{\sigma^i = 1\}} \\ &\geq 0 \quad (\mathbb{P}\text{-a.e.}). \end{aligned} \tag{11}$$

The last step follows taking X_0^i as the initial largest reward, and that all bandits are non-increasing.

It follows that advancing the activation of the initial maximal bandit improves or at least does not change the value of a policy. This same argument can be applied at every round that follows, that is, at every round, activation of the current initial maximal bandit is an improvement over (or at least does not change the value) of any other policy. Note, collisions may occur if at a given round two bandits have equal rewards. This may be resolved at the discretion of the controller, such as by always taking the bandit with the smaller index i.

As each bandit halts in a finite time, almost surely, for sufficiently many greedy improvements as outlined above, the resulting improvement of any policy π will return the same value as the completely greedy strategy π^*. Hence,

$$X_{\pi^*}(\sigma_{\pi^*} - 1) \geq X_\pi(\sigma_\pi - 1) \quad (\mathbb{P}\text{-a.e.}). \quad \blacksquare \tag{12}$$

Remark 4. The Necessity of finite σ^i. Note that Assumption 2.a: $\sigma^i < \infty$ almost surely, for each bandit i, is employed to exclude cases such as the following, in which no optimal policy exists.

Consider two bandits, Bandit A offering a potential reward of \$100 in each time step, and Bandit B offering a potential reward of \$50 in each time step. Further, suppose that $\mathbb{P}^A(\sigma^A < \infty) = 0.5$, and $\sigma^B = 1$ almost surely—that is, Bandit B halts after its first activation.

This choice of σ^B implies that any policy on these bandits may be described in the following way: for any a.s. finite $\mathbb{F}^A$-stopping time $\tau \geq 0$, π_τ activates Bandit A until τ, then Bandit B, ending the game. The value of such a policy is given by

$$V_{\pi_\tau}^{PSP}(A, B) = \$100\ \mathbb{P}^A(\sigma^A < \tau) + \$50\ \mathbb{P}^A(\sigma^A \geq \tau) \leq 75. \tag{13}$$

This upper bound may be achieved within an arbitrary amount by choosing a finite, sufficiently large τ—the larger the τ, the closer to achieving the upper bound. However, taking τ to be infinite, the \$100 is only collected with probability 0.5, and Bandit B is never activated at all, yielding a total expected value of $\$100 \times 0.5 = \$50 < \$75$. In this case, there exist ϵ-optimal policies, but no optimal policy. This phenomenon appears in all versions of the problems discussed herein and its investigation is an avenue of interesting additional research.

4 | MAXIMIZING COLLECTIVE PAYOUTS

In this section, we consider a model where rewards are collective, that is, received from all bandits, at the halting of the game. Thus, the expected *collective payout* value of a policy π is

$$V_\pi^{CP}(\mathbb{X}) = \sum_{i=1}^{N} \mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)} | \mathcal{G}_0\right]. \tag{14}$$

In the following subsections, we develop a policy $\pi^* \in \mathcal{P}$ such that for all $\pi \in \mathcal{P}$,

$$V^{CP}_{\pi}(\mathbb{X}) \leq V^{CP}_{\pi^*}(\mathbb{X})(\mathbb{P}\text{-a.e.}). \tag{15}$$

Remark 5. For algebraic convenience in the remainder of this section we take $X^i_0 = 0$ for all i. For a more arbitrary reward processes $\{\hat{X}^i\}$, recall that the initial $\hat{X}^i_0$ are taken to be constant and known at the initial round by assumption. Hence, defining $X^i_t = \hat{X}^i_t - \hat{X}^i_0$, maximizing the total expected reward from the $\{\hat{X}^i\}$ processes is equivalent to maximizing the total expected reward from the $\{X^i\}$ processes.

4.1 | Block values

This section introduces a way of considering the "value" of a set of activations of a bandit. The "true" value of a decision to activate a bandit is not simply the potential reward gained through that decision, but instead it must balance the immediate potential reward with the incurred risk of halting the game through that decision, and the resulting loss of potential future rewards.

For each bandit i, for a given policy π we define τ^i_π to be the first activation of bandit i that *does not occur under* π. That is,

$$\tau^i_\pi = \min\{t \geq 0 : S^i_\pi(t) \geq \sigma_\pi\}. \tag{16}$$

Note, the above makes use in its definition of π "after the halting time σ_π," but we simply mean to observe here that at the global halting time, we can observe what the next activation of each bandit would have been—this is τ^i_π.

With this, we state the following definitions.

Definition 1 (Process blocks and their values). Given times $t' < t''$ with $t' < \sigma^i$, and a policy $\pi \in \mathcal{P}$ with $S^i_\pi(t') < \sigma_\pi$:

1 The *solo-payout value of the* $[t', t'')$*—block of* X^i as:

$$\rho^i(t', t'') = \frac{\mathbb{E}^i\left[X^i_{\sigma^i \wedge t''} - X^i_{t'} \middle| \mathcal{F}^i(t')\right]}{\mathbb{P}^i\left(t' < \sigma^i \leq t'' \middle| \mathcal{F}^i(t')\right)}. \tag{17}$$

2 The π*-value of the* $[t', t'')$*—block of* X^i as:

$$v^i_\pi(t', t'') = \frac{\mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi) \wedge t''} - X^i_{t'} \middle| \mathcal{H}^i_\pi(t')\right]}{\mathbb{P}\left(t' < \sigma^i \leq \tau^i_\pi \wedge t'' \middle| \mathcal{H}^i_\pi(t')\right)}. \tag{18}$$

Remark 6. Due to Equation (3) compare Assumption 2, the denominators of both block values are non-zero. The above quantities are all measurable with respect to the indicated σ-fields, and finite ($\mathbb{P}^i, \mathbb{P}$-a.e.), due to Equation (1) compare Assumption 1.

Notionally, ρ^i can be thought of as the value of a block under consecutive activation, while v^i_π is, correspondingly, the value of a block potentially "diluted" or broken up by activations of other bandits under π. The denominator of v^i_π may be interpreted as the probability that the game halts *due to bandit* i, halting during activation of the $[t', t'')$-block.

Remark 7. The above might be justified as the "value" of a block of activations in the following way: even if the incremental reward gained due to an activation block (the numerators) is small, if the probability of halting due to those activations (the denominators) is sufficiently small, there is very little risk in attempting to gain that increment through that activation. In fact, there might be more to gain in such a case than if the incremental reward were slightly larger, but the probability of halting were also larger. The above values captures this trade-off between risk of halting and reward gained.

The following theorem illustrates the relationship between ρ^i and v^i_π, essentially stating that the value of any block under some policy π is *at most* the value of *some* block activated consecutively.

Theorem 2 (Block value comparison). *For bandit i under policy π, for any time t_0 such that $S^i_\pi(t_0) < \sigma_\pi$, the following holds for any $\mathbb{H}^i_\pi$-stopping time τ with $t_0 < \tau$:*

$$v^i_\pi(t_0, \tau) \leq \operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)} \rho^i(t_0, \hat{\tau})(\mathbb{P}\text{-a.e.}). \tag{19}$$

Proof. Note that it follows from Equations (1), (3) that the essential supremum is finite ($\mathbb{P}$-a.e.).

For each bandit i and any $\pi \in \mathcal{P}$, it can be shown by cases (whether the game does or does not halt due to an activation of i) that $T^i_\pi(\sigma_\pi) = \sigma^i \wedge \tau^i_\pi$.

Therefore, for a given $\tau \in \hat{\mathbb{H}}^i_\pi(t_0)$,

$$\begin{aligned} v^i_\pi(t_0, \tau) &= \frac{\mathbb{E}\left[X^i_{\sigma^i \wedge \tau^i_\pi \wedge \tau} - X^i_{t_0} \middle| \mathcal{H}^i_\pi(t_0)\right]}{\mathbb{P}\left(t_0 < \sigma^i \leq \tau^i_\pi \wedge \tau \middle| \mathcal{H}^i_\pi(t_0)\right)} \\ &= \frac{\mathbb{E}\left[X^i_{\sigma^i \wedge (\tau^i_\pi \wedge \tau)} - X^i_{t_0} \middle| \mathcal{H}^i_\pi(t_0)\right]}{\mathbb{P}\left(t_0 < \sigma^i \leq (\tau^i_\pi \wedge \tau) \middle| \mathcal{H}^i_\pi(t_0)\right)} \\ &\leq \operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{H}}^i_\pi(t_0)} \frac{\mathbb{E}\left[X^i_{\sigma^i \wedge \hat{\tau}} - X^i_{t_0} \middle| \mathcal{H}^i_\pi(t_0)\right]}{\mathbb{P}\left(t_0 < \sigma^i \leq \hat{\tau} \middle| \mathcal{H}^i_\pi(t_0)\right)} (\mathbb{P}\text{-a.e.}). \end{aligned} \tag{20}$$

The last step above follows as, given that τ^i_π and τ are both in $\hat{\mathbb{H}}^i_\pi(t_0)$ by assumption, so too is $\tau^i_\pi \wedge \tau$, and the term on the right hand side is the ess sup over all such stopping times.

Defining a "global" π-analog of ρ^i,

$$\rho^i_\pi(t',t'') = \frac{\mathbb{E}\left[X^i_{\sigma^i\wedge t''} - X^i_{t'}\middle|\mathcal{H}^i_\pi(t')\right]}{\mathbb{P}\left(t' < \sigma^i \le t''\middle|\mathcal{H}^i_\pi(t')\right)}, \tag{21}$$

we have the following relations:

$$v^i_\pi(t_0,\tau) \le \operatorname*{ess\,sup}_{\hat\tau\in\hat{\mathbb{H}}^i_\pi(t_0)} \rho^i_\pi(t_0,\hat\tau) \le \operatorname*{ess\,sup}_{\hat\tau\in\hat{\mathbb{F}}^i(t_0)} \rho^i(t_0,\hat\tau)(\mathbb{P}\text{-a.e.}). \tag{22}$$

The first inequality above is simply a restatement of Equation (20). The second inequality, the exchange from $\mathbb{H}^i_\pi$-stopping times to $\mathbb{F}^i$-stopping times, is intuitive: as the X^i process and σ^i are independent of the non-i bandits, information about those independent bandits (through the $\mathbb{H}^i_\pi$-stopping times) cannot assist in maximizing the quotient. Rigorously, this amounts to integrating out the independent bandits; this is done in detail as Proposition 1. ■

Proposition 1. *For bandit i under policy π, for any time t_0 such that $S^i_\pi(t_0) < \sigma_\pi$, the following holds:*

$$\operatorname*{ess\,sup}_{\hat\tau\in\hat{\mathbb{H}}^i_\pi(t_0)} \rho^i_\pi(t_0,\hat\tau) \le \operatorname*{ess\,sup}_{\hat\tau\in\hat{\mathbb{F}}^i(t_0)} \rho^i(t_0,\hat\tau)(\mathbb{P}\text{-a.e.}). \tag{23}$$

See Section 6 for its proof.

The following proposition provides, using ρ^i and v^i_π, expressions for the incremental reward gained through consecutive or under π activation of a block.

Proposition 2. *For each bandit i, the following relations hold for any $\mathbb{F}^i$-stopping times $\tau' < \tau''$ where the quantities are well defined. Equality also holds when conditioning with respect to the initial information, $\mathcal{F}^i(0)$, $\mathcal{G}_0$ respectively via the tower property.*

$$\mathbb{E}^i\left[X^i_{\sigma^i\wedge\tau''} - X^i_{\tau'}\middle|\mathcal{F}^i(\tau')\right] = \mathbb{E}^i\left[\sum_{t=\tau'}^{\tau''-1}\rho^i(\tau',\tau'')\mathbb{1}_{\{\sigma^i=t+1\}}\middle|\mathcal{F}^i(\tau')\right], \tag{24}$$

$$\mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)\wedge\tau''} - X^i_{\tau'}\middle|\mathcal{H}^i_\pi(\tau')\right] = \mathbb{E}\left[\sum_{t=\tau'}^{\tau''-1}v^i_\pi(\tau',\tau'')\mathbb{1}_{\{\sigma_\pi=S^i_\pi(t)+1\}}\middle|\mathcal{H}^i_\pi(\tau')\right]. \tag{25}$$

Proof. The above equations follow directly from Equations (17), (18), observing the following relations:

$$\mathbb{P}^i\left(t' < \sigma^i \le t''\middle|\mathcal{F}^i(t')\right) = \mathbb{E}^i\left[\sum_{t=t'}^{t''-1}\mathbb{1}_{\{\sigma^i=t+1\}}\middle|\mathcal{F}^i(t')\right],$$

$$\mathbb{P}\left(t' < \sigma^i \le \tau^i_\pi \wedge t''\middle|\mathcal{H}^i_\pi(t')\right) = \mathbb{E}\left[\sum_{t=t'}^{t''-1}\mathbb{1}_{\{\sigma_\pi=S^i_\pi(t)+1\}}\middle|\mathcal{H}^i_\pi(t')\right]. \tag{26}$$

■

4.2 | Solo payout indices and times

Theorem 2 indicates the significance of the following quantity.

Definition 2 (The Solo-Payout Index). For any $t < \sigma^i$, the incremental *Solo-Payout Index at t* is defined to be

$$\rho^i(t) = \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{F}}^i(t)} \rho^i(t,\tau). \tag{27}$$

This index can be interpreted as the maximal quotient of "incremental reward" over "probability of termination/halting" as in Equation (17). Sonin (2011) defined this index for the case of finite state Markov chain reward processes, in order to provide an efficient computation of the Gittins indices of all states.

The following result demonstrates that $\rho^i(t)$ is realized as the value of *some* block from time t, that is, for some $\tau > t$, $\rho^i(t) = \rho^i(t,\tau)(\mathbb{P}^i\text{-a.e.})$. As such, $\rho^i(t)$ represents the *maximal block value achievable from process i from time t.*

Proposition 3. *For any time $t_0 < \sigma^i$, there exists a $\tau \in \hat{\mathbb{F}}^i(t_0)$ such that $\rho^i(t_0) = \rho^i(t_0,\tau)(\mathbb{P}^i\text{-a.e.})$.*

The proof is relegated to Section 6, since it specialized and not the focus of this article.

The solo-payout indices and their realizing blocks provide a natural time scale with which to view a process, in terms of a sequence of blocks. In particular, we define the following sequence:

Definition 3 (Solo-Payout Index Times). Define a sequence of $\mathbb{F}^i$-stopping times $\{\tau^i_k\}_{k\ge0}$ in the following way, that $\tau^i_0 = 0$, and for $k > 0$,

$$\tau^i_{k+1} = \arg\ \operatorname{ess\,sup}\{\rho^i(\tau^i_k,\tau) : \tau \in \hat{\mathbb{F}}^i(\tau^i_k)\}. \tag{28}$$

In the case that $\tau^i_k = \sigma^i$ for some k, then $\tau^i_{k'}$ is taken to be infinite for all $k' > k$. In the case that $\tau^i_k < \sigma^i$, we have that $\rho^i(\tau^i_k) = \rho^i(\tau^i_k,\tau^i_{k+1})$. The question of whether the "arg ess sup" exists is resolved in the positive by Proposition 3; if there is more than one stopping time that attains the "arg ess sup," we take τ^i_{k+1} to be the one demonstrated by the application of Lemma 1 in the proof of Proposition 3.

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

$$\mathbb{N}^i = [0,\tau^i_1) \cup [\tau^i_1,\tau^i_2) \cup [\tau^i_2,\tau^i_3) \cup \dots .$$

One important property of this partition is the following:

Proposition 4 (Solo-Payout Indices are Non-Increasing over Index Times). *For any $k > 0$ such that $\tau^i_k < \sigma^i$, the following is true:* $\rho^i(\tau^i_{k-1}) \geq \rho^i(\tau^i_k)(\mathbb{P}^i\text{-a.e.})$.

For intuition, recall the $\{\tau^i_k\}_k$ are meant to realize successively the maximal indices of the process $\{X^i_t\}_t$. If $\rho^i(\tau^i_{k-1}) = \rho^i(\tau^i_{k-1}, \tau^i_k) < \rho^i(\tau^i_k)$, the index from τ^i_{k-1} may be increased by taking a block that extends from τ^i_{k-1} *past* τ^i_k. This contradicts the idea of the $\{\tau^i_k\}_k$ as realizing the maximal indices. The proof is relegated to Section 6, as technical, and not the focus of this article.

4.3 | Equivalent solo payout processes

For each bandit, we have developed a partition of local time into blocks of activations via the solo payout index stopping times. With Proposition 2 in mind, we use these blocks to define a set of reward equivalent penultimate solo payout processes, and π-equivalent solo payout processes.

Definition 4. Given the collection of reward processes $\mathbb{X} = (X^1, \dots, X^N)$, and $\{\tau^i_k\}_{k\geq 0}$ for each i as in Definition 3, we define:

1. The reward-equivalent solo payout collection $\mathbb{Y}^X = (Y^1, \dots, Y^N)$ by
$$Y^i(t) = \rho^i(\tau^i_k), \quad \text{if } \tau^i_k \leq t < \tau^i_{k+1}. \tag{29}$$
2. For $\pi \in \mathcal{P}$, the *π-equivalent solo payout collection* $\mathbb{Y}^X_\pi = (Y^1_\pi, \dots, Y^N_\pi)$, by
$$Y^i_\pi(t) = \nu^i_\pi(\tau^i_k, \tau^i_{k+1}), \quad \text{if } \tau^i_k \leq t < \tau^i_{k+1}. \tag{30}$$

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 the $\rho^i(\tau^i_k)$ is defined by the information available locally at time τ^i_k. However, as the $\nu^i_\pi(\tau^i_k, \tau^i_{k+1})$ depend on the specifics of policy π, so do the Y^i_π processes; the Y^i_π processes are $\mathbb{H}^i_\pi$-adapted, but not $\mathbb{F}^i$-adapted. Note, Y^i is only really defined for $t < \sigma^i$, and Y^i_π is only defined for t such that $S^i_\pi(t) < \sigma_\pi$. However, since no rewards are collected from bandit i after these times, this lack of definition is of no consequence.

The following are simple, but important properties of the $\mathbb{Y}^X, \mathbb{Y}^X_\pi$ processes.

Proposition 5. *For $\pi \in \mathcal{P}$, for each i, and any k where the following quantities are defined,*

$$\mathbb{E}^i\left[X^i_{\sigma^i \wedge \tau^i_{k+1}} - X^i_{\tau^i_k}\middle|\mathcal{F}^i(\tau^i_k)\right] = \mathbb{E}^i\left[\sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} Y^i(t)\mathbb{1}_{\{\sigma^i=t+1\}}\middle|\mathcal{F}^i(\tau^i_k)\right], \tag{31}$$

$$\mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)\wedge\tau^i_{k+1}} - X^i_{\tau^i_k}\middle|\mathcal{H}^i_\pi(\tau^i_k)\right] = \mathbb{E}\left[\sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} Y^i_\pi(t)\mathbb{1}_{\{\sigma_\pi=S^i_\pi(t)+1\}}\middle|\mathcal{H}^i_\pi(\tau^i_k)\right]. \tag{32}$$

As with Proposition 2, equality also holds when conditioning with respect to $\mathcal{F}^i(0), \mathcal{G}_0$.

Proof. This follows as an application of Proposition 2 and the definitions of Y^i, Y^i_π. ■

The following proposition serves as justification of the term "equivalent" in describing the $\mathbb{Y}^X, \mathbb{Y}^X_\pi$ collections.

Proposition 6. *For each i, for any policy $\pi \in \mathcal{P}$,*

$$\mathbb{E}^i\left[X^i_{\sigma^i}\middle|\mathcal{F}^i(0)\right] = \mathbb{E}^i\left[Y^i(\sigma^i - 1)\middle|\mathcal{F}^i(0)\right], \tag{33}$$

$$\mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)}\middle|\mathcal{G}_0\right] = \mathbb{E}\left[\mathbb{1}_{\{i=\pi(\sigma_\pi-1)\}}Y^i_\pi(T^i_\pi(\sigma_\pi - 1))\middle|\mathcal{G}_0\right]. \tag{34}$$

Proof. Each follows from the corresponding equation in Proposition 5, summing over k and taking expectations from the initial time, via the tower property. On the right hand sides, the X^i terms telescope in the sum, and X^i_0 is taken to be 0. On the left hand sides, the sums over Y may be expressed as single terms, due to the indicators. ■

Proposition 7. *For each i, and any time $t > 0$ such that $Y^i(t)$ is well defined,*

$$Y^i(t-1) \geq Y^i(t)(\mathbb{P}^i\text{-a.e.}). \tag{35}$$

Proof. This follows immediately from Proposition 4, and Definition 4.1. ■

Theorem 3 (Comparison of Equivalent, π-Equivalent Solo Payout Processes). *For any $\pi \in \mathcal{P}$, for each i and all time t where both are defined, we have:*

$$Y^i_\pi(t) \leq Y^i(t) \quad (\mathbb{P}\text{-a.e.}). \tag{36}$$

Proof. For such a t, we have for some k that $\tau^i_k \leq t < \tau^i_{k+1}$, and as an application of Theorem 1,

$$Y^i_\pi(t) = \nu^i_\pi(\tau^i_k, \tau^i_{k+1}) \leq \operatorname*{ess\,sup}_{\tau' \in \hat{\mathbb{H}}^i_\pi(\tau^i_k)} \nu^i_\pi(\tau^i_k, \tau') \leq \operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(\tau^i_k)} \rho^i(\tau^i_k, \hat{\tau}) = \rho^i(\tau^i_k) = Y^i(t) \quad (\mathbb{P}\text{-a.e.}). \tag{37}$$

Note in the above that the first and the last relations are just definitions, the second follows naturally by comparing one instance of the function to an ess sup of the same function, the third is due to Theorem 2, the fourth is due to the definition of the ρ^i function. ■

4.4 | The optimal policy

The derivation of the optimal control policy for an arbitrary collection of reward processes $\mathbb{X}$ under a collective reward structure is all but immediate now.

Theorem 4 (The Optimal Collective Payout Control Policy). *For a collection of reward processes* $\mathbb{X} = (X^1, X^2, \dots, X^N)$, *and the associated stopping times* $\{\sigma^i\}_{i=1,\dots,N}$, *there exists a strategy* $\pi^* \in \mathcal{P}$ *such that for all* $\pi \in \mathcal{P}$,

$$V_\pi^{CP}(\mathbb{X}) \le V_{\pi^*}^{CP}(\mathbb{X})(\mathbb{P}\text{-a.e.}). \tag{38}$$

In particular, such an optimal policy π^* *can be described in the following way: successively activate the bandit with the largest current solo payout index,*

$$\rho^i(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[X^i_{\sigma^i \wedge \tau} - X^i_t \middle| \mathcal{F}^i(t)\right]}{\mathbb{P}^i\left(t < \sigma^i \le \tau \middle| \mathcal{F}^i(t)\right)}, \tag{39}$$

for the duration of the corresponding index block.

Before giving the proof of this theorem, we give a corollary, which gives a useful alternative characterization of the policy π^*.

Corollary 1. *An alternative characterization of the policy* π^* *in Theorem 4 is the following: at every round, activate the bandit with the largest current solo payout index.*

Proof. From Theorem 4, it follows that the optimal first activation is to activate a bandit with the largest current solo payout index. If that activation does not halt the bandit and end the game, the controller is faced with a structurally identical decision problem. It follows that again, the optimal activation is to activate a bandit with the largest current solo payout index. This argument may be iterated until halting, which will occur in finite time by assumption on the $\{\sigma^i\}$. ■

Proof of Theorem 4. For an arbitrary policy π, and π^* as indicated above, we establish the following relations:

$$V_\pi^{CP}(\mathbb{X}) = V_\pi^{PSP}(\mathbb{Y}_\pi^X) \le V_\pi^{PSP}(\mathbb{Y}^X) \le V_{\pi^*}^{PSP}(\mathbb{Y}^X) = V_{\pi^*}^{CP}(\mathbb{X})(\mathbb{P}\text{-a.e.}), \tag{40}$$

that is, for any policy π, we have that $V_\pi^{CP}(\mathbb{X}) \le V_{\pi^*}^{CP}(\mathbb{X})(\mathbb{P}\text{-a.e.})$ and therefore π^* is an optimal policy.

In the following steps we prove relations (40).

Step 1: $V_\pi^{CP}(\mathbb{X}) = V_\pi^{PSP}(\mathbb{Y}_\pi^X)$, ($\mathbb{P}$-a.e.).

We have, via Proposition 6, Equation (34),

$$\begin{aligned} V_\pi^{CP}(\mathbb{X}) &= \sum_{i=1}^N \mathbb{E}\left[X^i_{T^i_\pi(\sigma_\pi)} \middle| \mathcal{G}_0\right] \\ &= \sum_{i=1}^N \mathbb{E}\left[\mathbb{1}_{\{i=\pi(\sigma_\pi-1)\}} Y^i_\pi(T^i_\pi(\sigma_\pi-1)) \middle| \mathcal{G}_0\right] \\ &= V_\pi^{PSP}(\mathbb{Y}_\pi^X). \end{aligned}$$

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

Step 2: $V_\pi^{PSP}(\mathbb{Y}_\pi^X) \le V_\pi^{PSP}(\mathbb{Y}^X)(\mathbb{P}\text{-a.e.})$.

This follows from the point-wise inequality of Theorem 3, $Y^i_\pi(t) \le Y^i(t)$ for all t. Note that for any t where $Y^i_\pi(t)$ is not defined, the tth activation of i does not occur under π, and no comparison is necessary.

Step 3: $V_\pi^{PSP}(\mathbb{Y}^X) \le V_{\pi^*}^{PSP}(\mathbb{Y}^X)(\mathbb{P}\text{-a.e.})$.

This follows simply from Theorem 1 as, by construction, the terms of each Y^i process are equal to the solo payout indices of X^i, piecewise constant over blocks, and non-increasing.

Step 4: $V_{\pi^*}^{PSP}(\mathbb{Y}^X) = V_{\pi^*}^{CP}(\mathbb{X})(\mathbb{P}\text{-a.e.})$.

Note that π^* activates bandits consecutively over the duration of their index blocks. For a given i, define

$$k_i^* = \min_{k \ge 0}\{S^i_{\pi^*}(\tau^i_k) \ge \sigma_\pi\}, \tag{41}$$

the first block of i that is *not* activated under π^*. Note then that for each i, we have the following relation

$$T^i_{\pi^*}(\sigma_{\pi^*}) = \sigma^i \wedge \tau^i_{k_i^*}. \tag{42}$$

Expressing the value of policy π^* relative to activations over blocks, and utilizing the tower property, we have the following equivalences:

$$\begin{aligned} V_{\pi^*}^{PSP}(\mathbb{Y}^X) &= \sum_{i=1}^N \sum_{k=0}^\infty \mathbb{E}\left[\mathbb{1}_{\{k_i^*>k\}} \sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} Y^i(t)\mathbb{1}_{\{\sigma^i=t+1\}} \middle| \mathcal{G}_0\right] \\ &= \sum_{i=1}^N \sum_{k=0}^\infty \\ &\quad \mathbb{E}\left[\mathbb{1}_{\{k_i^*>k\}} \mathbb{E}\left[\sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} Y^i(t)\mathbb{1}_{\{\sigma^i=t+1\}} \middle| \mathcal{H}^i_\pi(\tau^i_k)\right] \middle| \mathcal{G}_0\right] \\ &= \sum_{i=1}^N \sum_{k=0}^\infty \\ &\quad \mathbb{E}\left[\mathbb{1}_{\{k_i^*>k\}} \mathbb{E}\left[X^i_{\sigma^i \wedge \tau^i_{k+1}} - X^i_{\tau^i_k} \middle| \mathcal{H}^i_\pi(\tau^i_k)\right] \middle| \mathcal{G}_0\right] \\ &= \sum_{i=1}^N \mathbb{E}\left[X^i_{\sigma^i \wedge \tau^i_{k_i^*}} - X^i_0 \middle| \mathcal{G}_0\right] \\ &= \sum_{i=1}^N \mathbb{E}\left[X^i_{T^i_{\pi^*}(\sigma_{\pi^*})} \middle| \mathcal{G}_0\right] = V_{\pi^*}^{CP}(\mathbb{X}). \end{aligned} \tag{43}$$

Note the exchange over blocks of the Y^i rewards for the X^i rewards is due to Proposition 5, Equation (31), taking the extension to $\mathcal{H}^i_{\pi^*}(\tau^i_k)$ in place of $\mathcal{F}^i(\tau^i_k)$. ■

Remark 8. 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 2.2, 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, for example, coin flips. However, such a construction simply amounts to placing a distribution on $\mathcal{P}$. Since π^* is $\mathbb{P}$-a.e. superior to any $\pi \in \mathcal{P}$, π^* would be similarly superior to any policy sampled randomly from $\mathcal{P}$.

The structure of the proof of Theorem 4 above is based on deriving an optimality result for the collective payout model by reducing it to an instance of the a solo payout model. It suggests an interesting correspondence between the two. Under the collective payout model, in any period the controller wishes to achieve via bandit activation high collective rewards of all bandits on halting. Under a solo payout model, in any period the controller wishes to achieve via bandit activation high rewards of a given bandit on halting. However, since (under either model) the controller can only activate one bandit at a time, under the collective payout model the controller essentially seeks in every period to maximize the change in collective reward due to a single bandit *should that bandit halt*, or equivalently to maximize the change in reward *of that single bandit* should that bandit halt. The collective payout model can therefore be cast as a penultimate solo payout model, where the payout on halting is based on the *change* in reward of the activated bandit rather than the final collective rewards of all bandits. This can be further seen in the following section, where optimal index policies for the general (penultimate and ultimate) solo payout model are given.

5 | ADDITIONAL PAYOUT SCHEMES

Utilizing the results of the previous section, we next provide index policies for optimizing the rewards/costs from a number of additional payout models, by reducing them to the collective payout model compare Equation (14) of the previous section, and utilizing Theorem 4. We construct the models below, specified by different ways in which rewards are received and/or costs are paid. We note that analogous results can be obtained for the penultimate solo payout model without the monotonicity restriction of Section 3 on the underlying reward processes. They are omitted for brevity.

1. The *Ultimate Solo Payout* model (SP). In this model, the controller aims to maximize the expected final reward from the bandit that halts the game, that is, the value of a policy π is defined as,

$$V^{SP}_\pi(\mathbb{X}) = \mathbb{E}[X_\pi(\sigma_\pi)|\mathcal{G}_0] = \sum_{i=1}^{N} \mathbb{E}\Big[\mathbb{1}_{\{i=\pi(\sigma_\pi-1)\}} X^i_{T^i_\pi(\sigma_\pi)} |\mathcal{G}_0\Big]. \quad (44)$$

2. The *non-halting cost* model (NH). In this model, the controller *pays a cost* based on the bandits that did not halt the game, and wishes to minimize this expected cost. The *halting cost* of a policy π is

$$V^{NH}_\pi(\mathbb{X}) = \mathbb{E}\left[\sum_{i\neq\pi(\sigma_\pi-1)} X^i_{T^i_\pi(\sigma_\pi)} |\mathcal{G}_0\right] = \sum_{i=1}^{N} \mathbb{E}\Big[\mathbb{1}_{\{i\neq\pi(\sigma_\pi-1)\}} X^i_{T^i_\pi(\sigma_\pi)} |\mathcal{G}_0\Big]. \quad (45)$$

3. The *total profit* model (TP). In this model to each bandit i we associate a reward process $\{R^i_t\}_{t\geq 0}$ and a cost process $\{C^i_t\}_{t\geq 0}$. The controller gains a reward from the bandit that halts the game, and pays a cost for each bandit that does not halt. The controller wishes to maximize her expected total profit, that is, the value of a policy π is now defined as,

$$V^{TP}_\pi(\mathbb{R},\mathbb{C}) = \sum_{i=1}^{N} \mathbb{E}\Big[\mathbb{1}_{\{i=\pi(\sigma_\pi-1)\}} R^i_{T^i_\pi(\sigma_\pi)} - \mathbb{1}_{\{i\neq\pi(\sigma_\pi-1)\}} C^i_{T^i_\pi(\sigma_\pi)} |\mathcal{G}_0\Big]. \quad (46)$$

4. The *cumulative collective payout* (CCP) model and the Gittins index. In this model, the controller gains a bandit's current reward each time that bandit is chosen to be activated. Bandits that are never activated give no rewards. The controller wishes to maximize her expected total payout, that is, the value of a policy π is now defined as,

$$V^{CCP}_\pi(\mathbb{X}) = \sum_{i=1}^{N} \mathbb{E}\left[\sum_{t=0}^{T^i_\pi(\sigma_\pi)-1} X^i_t \middle| \mathcal{G}_0\right]. \quad (47)$$

Note, in the above expressions we take empty sums to be 0.

For all these models, we will provide an index policy to maximize the corresponding value function as follows.

1. For the *SP model*, define a collection of reward processes $\mathbb{Z} = \{Z^i\}_{1\leq i\leq N}$ by for each i, each $t \geq 0$,

$$Z^i_t = \mathbb{1}_{\{\sigma^i=t\}} X^i_t. \quad (48)$$

Notice that at round σ_π, $Z^i_t = 0$ for all bandits that did not halt the game, and $Z^i_t = X^i_{\sigma^i}$ for the bandit that did halt the game. Hence the collective payout under $\mathbb{Z}$ is equal to the solo payout under $\mathbb{X}$, $V_\pi^{CP}(\mathbb{Z}) = V_\pi^{SP}(\mathbb{X})$. Applying Theorem 4, the optimal policy for the collective payout under $\mathbb{Z}$ yields an optimal policy for the solo payout under $\mathbb{X}$, and it is given by a policy that always activates bandits according to the maximum *solo payout index*:

$$\rho_{SP}^i(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\mathbb{1}_{\{\tau \geq \sigma^i\}} X_{\sigma^i}^i \middle| \mathcal{F}^i(t)\right]}{\mathbb{P}^i\left(t < \sigma^i \leq \tau \middle| \mathcal{F}^i(t)\right)}. \quad (49)$$

It is interesting to observe that the policy based on the above index has a very natural interpretation, viewing the index as the maximal conditional expected payout of a bandit on its halting, that is, the policy always activates the bandit with the largest potential payout—should it pay out. Additionally, comparing the above index to the optimal index for the collective payout model, it is clear that the collective payout index emphasizes the "change in reward on halting" of a single bandit, while the solo payout index emphasizes only the final reward of a single bandit on halting. This again highlights the correspondence between these two models, as discussed at the end of Section 4.4.

2. We reduce the *non-halting cost* model, to the collective payout model in the following way. Define a collection of reward processes $\mathbb{Z} = \{Z^i\}_{1 \leq i \leq N}$ by for each i, each $t \geq 0$,

$$Z_t^i = -\mathbb{1}_{\{\sigma^i \neq t\}} X_t^i. \quad (50)$$

Notice that at round σ_π, if bandit i was activated to halt the game (i.e., $\pi(\sigma_\pi - 1) = i$), Equation (50) implies that $Z_t^i = 0$ and $Z_t^j = -X_t^j$, for $j \neq i$. Hence, the collective payout under $\mathbb{Z}$ is equal to the negative of the halting cost under $\mathbb{X}$: $V_\pi^{CP}(\mathbb{Z}) = -V_\pi^{NH}(\mathbb{X})$; it follows that maximizing the collective payout under $\mathbb{Z}$ minimizes the halting cost under $\mathbb{X}$. Applying Theorem 4, the optimal policy for the collective payout under $\mathbb{Z}$ yields an optimal policy for the non-halting cost model under $\mathbb{X}$, and it is given by a policy that always activates bandits according to the minimum *non-halting cost index*:

$$\rho_{NH}^i(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\mathbb{1}_{\{\sigma^i > \tau\}} X_\tau^i - X_t^i \middle| \mathcal{F}^i(t)\right]}{\mathbb{P}^i\left(t < \sigma^i \leq \tau \middle| \mathcal{F}^i(t)\right)}. \quad (51)$$

3. For the *total profit* model (TP) model, in order to provide an index policy to maximize its value function, we reduce it to the collective payout model in the following way. Define a collection of reward processes $\mathbb{Z} = \{Z^i\}_{1 \leq i \leq N}$ by for each i, each $t \geq 0$,

$$Z_t^i = \mathbb{1}_{\{\sigma^i = t\}} R_t^i - \mathbb{1}_{\{\sigma^i \neq t\}} C_t^i. \quad (52)$$

Notice that at round σ_π, $Z_t^i = -C_t^i$ for all bandits that did not halt the game, and $Z_t^i = R_t^i$ for the bandit that did halt the game. Hence the collective payout under $\mathbb{Z}$ is equal to the collective profit solo payout under $(\mathbb{R}, \mathbb{C})$, $V_\pi^{CP}(\mathbb{Z}) = V_\pi^{TP}(\mathbb{R}, \mathbb{C})$, compare Equation (14). Thus, as before, the optimal policy for the collective payout under $\mathbb{Z}$ yields an optimal policy for the total profit under $(\mathbb{R}, \mathbb{C})$, given by a policy that always activates bandits according to the maximum *total profit index*:

$$\rho_{TP}^i(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\mathbb{1}_{\{\sigma^i \leq \tau\}} R_{\sigma^i}^i - \mathbb{1}_{\{\sigma^i > \tau\}} C_\tau^i + C_t^i \middle| \mathcal{F}^i(t)\right]}{\mathbb{P}^i\left(t < \sigma^i \leq \tau \middle| \mathcal{F}^i(t)\right)}. \quad (53)$$

4. For the CCP model the controller gains a bandit's current reward each time that a bandit is chosen to be activated. Bandits that are not activated give no rewards. To provide an index policy to maximize this value function, we reduce it to the collective payout model, in the following way. Define a collection of reward processes $\mathbb{Z} = \{Z^i\}_{1 \leq i \leq N}$ by

$$Z_t^i = \sum_{t'=0}^{t-1} X_{t'}^i, \text{ for each } i, \text{each} t \geq 0. \quad (54)$$

It follows easily that the collective payout model value under $\mathbb{Z}$ is equal to the collective cumulative payout under $\mathbb{X}$, that is, $V_\pi^{CP}(\mathbb{Z}) = V_\pi^{CCP}(\mathbb{X})$. Thus, applying Theorem 4, the optimal policy for the collective payout under $\mathbb{Z}$ yields an optimal policy for the collective cumulative payout under $\mathbb{X}$, given by a policy that always activates bandits according to the maximum *collective cumulative payout index*:

$$\rho_{CCP}^i(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \frac{\mathbb{E}^i\left[\sum_{t'=t}^{\sigma^i \wedge \tau - 1} X_{t'}^i \middle| \mathcal{F}^i(t)\right]}{\mathbb{P}^i\left(t < \sigma^i \leq \tau \middle| \mathcal{F}^i(t)\right)}. \quad (55)$$

This extension of the collective payout model is interesting in its own right, because it allows us to readily recover and provide new simple proofs for the classic result of Gittins (1979) and the recent results in Cowan and Katehakis (2015).

Indeed, consider the case in which each time the controller activates a bandit, all future expected rewards are effectively discounted by a factor equal to the probability of that decision not halting the game.

In the special case that each halting time $\sigma^i > 0$ is a geometric random variable with a constant parameter $0 < \beta < 1$, independent of the reward processes $\mathbb{X}$, that is, $\mathbb{P}^i(\sigma^i = t + 1 | \mathcal{F}^i(t)) = 1 - \beta$. This results in every activation discounting all future rewards by a factor of β.

It is easy to see that

$$V_{\pi}^{CCP}(\mathbb{X})=\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=0}^{T_{\pi}^{i}(\sigma_{\pi})-1}X_{t}^{i}\middle|\mathcal{G}_{0}\right]=\mathbb{E}\left[\sum_{s=0}^{\infty}\beta^{s}X_{\pi}(s)\middle|\mathcal{G}_{0}\right]. \tag{56}$$

It follows from Equation (56), that maximizing the $V_{\pi}^{CCP}(\mathbb{X})$ under this model (with $\mathbb{P}^i(\sigma^i = t+1|\mathcal{F}^i(t)) = 1-\beta$, for all t and all i) is then equivalent precisely the framework outlined by Gittins (1979), that is, total expected discounted reward of $\mathbb{X}$ for a constant discount factor β. In this case, the collective cumulative payout index reduces to

$$\begin{aligned}\rho_{CCP}^{i}(t)=&\operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{F}}^{i}(t)}\frac{\mathbb{E}^{i}\left[\sum_{t'=t}^{\tau-1}\beta^{t'-t}X_{t'}^{i}\middle|\mathcal{F}^{i}(t)\right]}{\mathbb{E}^{i}\left[1-\beta^{\tau-t}\middle|\mathcal{F}^{i}(t)\right]}\\ =&\frac{1}{1-\beta}\operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{F}}^{i}(t)}\frac{\mathbb{E}^{i}\left[\sum_{t'=t}^{\tau-1}\beta^{t'}X_{t'}^{i}\middle|\mathcal{F}^{i}(t)\right]}{\mathbb{E}^{i}\left[\sum_{t'=t}^{\tau-1}\beta^{t'}\middle|\mathcal{F}^{i}(t)\right]},\end{aligned} \tag{57}$$

where the *essential sup* on the right is precisely the Gittins index for bandit i. As $1/(1-\beta)$ is a constant, positive factor, activating according to the maximal collective cumulative payout index and activating according to the maximal Gittins index result in equivalent, optimal policies. We also note that in this $\rho_{CCP}^{i}(t)$ is the restart index compare Katehakis and Veinott Jr. (1987) and the generalized index of Sonin (2008).

The above is a well known interpretation of the Gittins index problem in terms of halting bandits, but its treatment herein provides a new interesting implication. In its classical form, it is not intuitively clear why the decision problem decomposes into indices that treat each bandit separately. However, framing it as a collective payout halting problem, we may make use of the previously described correspondence with the solo payout model. Reducing the Gittins model to a solo payout model, where in every period the controller wishes to realize the largest change in value of a single bandit on halting, provides additional insight into why the decomposition of the decision process into treating each bandit independently holds.

We additionally note that the above arguments can be extended to generalized sequences of discount factors, for which $\mathbb{P}^i(\sigma^i = t+1|\mathcal{F}^i(t)) = 1-\beta_t^i$, and thus recover the main results of Cowan and Katehakis (2015).

6 | PROOFS OF AUXILIARY PROPOSITIONS

We start with the following.

Proof of Proposition 1. Without loss of generality, we may take $t_0 = 0$. Recall the definition of ρ_{π}^{i}, ρ^{i}:

$$\begin{aligned}\rho_{\pi}^{i}(t',t'')&=\frac{\mathbb{E}\left[X_{\sigma^i\wedge t''}^{i}-X_{t'}^{i}\middle|\mathcal{H}_{\pi}^{i}(t')\right]}{\mathbb{P}\left(t'<\sigma^i\leq t''\middle|\mathcal{H}_{\pi}^{i}(t')\right)},\\ \rho^{i}(t',t'')&=\frac{\mathbb{E}^{i}\left[X_{\sigma^i\wedge t''}^{i}-X_{t'}^{i}\middle|\mathcal{F}^{i}(t')\right]}{\mathbb{P}^{i}\left(t'<\sigma^i\leq t''\middle|\mathcal{F}^{i}(t')\right)}.\end{aligned} \tag{58}$$

Letting R denote the R.H.S. of Equation (23), observe (by the definition of the ess sup) that for any $\hat{\tau}\in\hat{\mathbb{F}}^{i}(0)$,

$$\mathbb{E}\left[X_{\sigma^i\wedge\hat{\tau}}^{i}-X_{0}^{i}-R\mathbb{1}\{0<\sigma^i\leq\hat{\tau}\}\middle|\mathcal{F}^{i}(0)\right]\leq 0(\mathbb{P}\text{-a.e.}). \tag{59}$$

To prove the proposition, it suffices to show that for any $\hat{\tau}\in\hat{\mathbb{H}}_{\pi}^{i}(0)$,

$$\mathbb{E}\left[X_{\sigma^i\wedge\hat{\tau}}^{i}-X_{0}^{i}-R\mathbb{1}\{0<\sigma^i\leq\hat{\tau}\}\middle|\mathcal{H}_{\pi}^{i}(0)\right]\leq 0(\mathbb{P}\text{-a.e.}). \tag{60}$$

For compactness of argument, we take $N=2$ and $i=1$, though the following argument generalizes to arbitrary bandits in the obvious way. For notational compactness, we define $W_t^i = X_{\sigma^i\wedge t}^{i}-X_0^i - R\mathbb{1}\{0<\sigma^i\leq t\}$.

Note that for any set $A\in\mathcal{H}_{\pi}^{1}(0)$, and any $\tau\in\hat{\mathbb{H}}_{\pi}^{1}(0)$,

$$\mathbb{E}\left[\mathbb{1}_A\mathbb{E}\left[W_{\tau}^{1}\middle|\mathcal{H}_{\pi}^{1}(0)\right]\right]=\mathbb{E}\left[\mathbb{1}_A W_{\tau}^{1}\right]. \tag{61}$$

Taking A as a rectangle in $\mathcal{H}_{\pi}^{1}(0)$, $A=A_1\times A_2$, observe that $A_1 \in \mathcal{F}^1(0)$. The indicator may be decomposed as $\mathbb{1}_A(\omega) = \mathbb{1}_{A_1}(\omega^1)\mathbb{1}_{A_2}(\omega^2)$. It follows as a result of the initial integrability assumptions on the bandits, Equations (1), (3), that we may exchange the expectation over the product space for an iterated expectation:

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_A W_{\tau}^{1}\right]&=\mathbb{E}^2\left[\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{1}_{A_2}W_{\tau}^{1}\right]\right]\\ &=\mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}W_{\tau}^{1}\right]\right]\\ &=\mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{E}^1\left[W_{\tau}^{1}\middle|\mathcal{F}^1(0)\right]\right]\right].\end{aligned} \tag{62}$$

Observe that, while τ (begin an $\mathbb{H}_{\pi}^{1}$-stopping time) may have a dependence on Ω^2, inside the iterated integral with the dependence on Ω^2 fixed, it is an $\mathbb{F}^i$-stopping time. Hence, as an application of Equation (59), we have the bound

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_A W_{\tau}^{1}\right]&=\mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{E}^1\left[W_{\tau}^{1}\middle|\mathcal{F}^1(0)\right]\right]\right]\\ &\leq\mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}0\right]\right]=0.\end{aligned} \tag{63}$$

Hence, for all rectangles $A \in \mathcal{H}_{\pi}^{1}(0)$, $\mathbb{E}\left[\mathbb{1}_A\mathbb{E}\left[W_{\tau}^{1}\middle|\mathcal{H}_{\pi}^{1}(0)\right]\right] \leq 0$. This extends via the usual monotone-class type argument to *all* $A \in \mathcal{H}_{\pi}^{1}(0)$. Hence, it follows that for all $\tau\in\hat{\mathbb{H}}_{\pi}^{1}(0)$,

$$\mathbb{E}\left[W_{\tau}^{1}\middle|\mathcal{H}_{\pi}^{1}(0)\right]\leq 0(\mathbb{P}\text{-a.e.}). \tag{64}$$

This establishes the result. ∎

The proof of Proposition 3 below requires the following technical lemma. Its proof follows along the lines of the proofs of Theorems 4.1–4.3 in Snell (1952), see also the optimal optional stopping lemma in Derman and Sacks (1960).

Lemma 1. *In an arbitrary probability space with a filtration $\mathbb{J} = \{\mathcal{J}_t\}_{t\geq 0}$, consider an adapted discrete-time process $\{Z_t\}_{t\geq 0}$ such that $\mathbb{E}\left[\sup_{\mathbb{N}} |Z_t| \middle| \mathcal{J}_0\right] < \infty$. If the $\mathbb{J}$-stopping time $\tau^* \in \hat{\mathbb{J}}(0)$ defined by*

$$\tau^* = \inf\{n > 0 : \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(n)} \mathbb{E}\left[Z_\tau \middle| \mathcal{J}_n\right] \leq Z_n\}, \tag{65}$$

is almost surely finite, then

$$\mathbb{E}\left[Z_{\tau^*} \middle| \mathcal{J}_0\right] = \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(0)} \mathbb{E}\left[Z_\tau \middle| \mathcal{J}_0\right] (\mathbb{P}\text{-a.e.}). \tag{66}$$

Proof of Proposition 3. Recall that we need to show that for any time $t_0 < \sigma^i$, there exists a $\tau \in \hat{\mathbb{F}}^i(t_0)$ such that $\rho^i(t_0) = \rho^i(t_0, \tau)(\mathbb{P}^i\text{-a.e.})$.

We have that for all $\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)$, $\rho^i(t_0, \hat{\tau}) \leq \rho^i(t_0)(\mathbb{P}^i\text{-a.e.})$. Taking

$$\mathbb{P}^i(t_0 < \sigma^i \leq \hat{\tau} \middle| \mathcal{F}^i(t_0)) = \mathbb{E}^i\left[\mathbb{1}_{\{t_0<\sigma^i\leq\hat{\tau}\}} \middle| \mathcal{F}^i(t_0)\right],$$

we have in parallel with Equation (24),

$$\mathbb{E}^i\left[X^i_{\sigma^i\wedge\hat{\tau}} - X^i_{t_0} - \rho^i(t_0)\mathbb{1}_{\{t_0<\sigma^i\leq\hat{\tau}\}} \middle| \mathcal{F}^i(t_0)\right] \leq 0(\mathbb{P}^i\text{-a.e.}). \tag{67}$$

Defining

$$\epsilon = -\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^i(t_0)} \mathbb{E}^i\left[X^i_{\sigma^i\wedge\hat{\tau}} - X^i_{t_0} - \rho^i(t_0)\mathbb{1}_{\{t_0<\sigma^i\leq\hat{\tau}\}} \middle| \mathcal{F}^i(t_0)\right], \tag{68}$$

we have that $\epsilon \geq 0(\mathbb{P}^i\text{-a.e.})$. We may use $-\epsilon$ as an improved upper bound in Equation (67). This may be rearranged to yield

$$\begin{aligned}\rho^i(t_0, \hat{\tau}) &\leq \rho^i(t_0) - \frac{\epsilon}{\mathbb{E}^i\left[\mathbb{1}_{\{t_0<\sigma^i\leq\hat{\tau}\}} \middle| \mathcal{F}^i(t_0)\right]} \\ &\leq \rho^i(t_0) - \epsilon(\mathbb{P}^i\text{-a.e.}).\end{aligned} \tag{69}$$

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

$$\rho^i(t_0) \leq \rho^i(t_0) - \epsilon(\mathbb{P}^i\text{-a.e.}), \tag{70}$$

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,

$$\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^i(t_0)} \mathbb{E}^i\left[X^i_{\sigma^i\wedge\hat{\tau}} - X^i_{t_0} - \rho^i(t_0)\mathbb{1}_{\{t_0<\sigma^i\leq\hat{\tau}\}} \middle| \mathcal{F}^i(t_0)\right] = 0(\mathbb{P}^i\text{-a.e.}). \tag{71}$$

Define $Z^i_t = X^i_{\sigma^i\wedge t} - X^i_{t_0} - \rho^i(t_0)\mathbb{1}_{\{t_0<\sigma^i\leq t\}}$. Note that the integrability condition of Lemma 1 is satisfied due to Equation (1). For $t \geq \sigma^i$, Z^i_t is constant, hence $\tau^* \leq \sigma^i < \infty$ almost surely. Hence we may apply Lemma 1 here to yield a stopping time $\tau^* \in \hat{\mathbb{F}}^i(t_0)$ such that

$$\mathbb{E}^i\left[X^i_{\sigma^i\wedge\tau^*} - X^i_{t_0} - \rho^i(t_0)\mathbb{1}_{\{t_0<\sigma^i\leq\tau^*\}} \middle| \mathcal{F}^i(t_0)\right] = 0(\mathbb{P}^i\text{-a.e.}), \tag{72}$$

or

$$\rho^i(t_0) = \frac{\mathbb{E}^i\left[X^i_{\sigma^i\wedge\tau^*} - X^i_{t_0} \middle| \mathcal{F}^i(t_0)\right]}{\mathbb{P}^i\left(t_0 < \sigma^i \leq \tau^* \middle| \mathcal{F}^i(t_0)\right)} = \rho^i(t_0, \tau^*)(\mathbb{P}^i\text{-a.e.}). \tag{73}$$

Hence, the solo-payout index $\rho^i(t_0)$ is realized ($\mathbb{P}^i$-a.e.) for some $\mathbb{F}^i$-stopping time $\tau^* > t_0$. ■

Proof of Proposition 4. For $k > 0$, let $\tau^i_k < \sigma^i$, and therefore $\tau^i_{k-1} < \sigma^i$. Defining

$$Z^i_t = X^i_{\sigma^i\wedge t} - X^i_{\tau^i_{k-1}} - \rho^i(\tau^i_{k-1})\mathbb{1}_{\{\tau^i_{k-1}<\sigma^i\leq t\}},$$

note that for $t > \tau^i_k$: $Z^i_t - Z^i_{\tau^i_k} = X^i_{\sigma^i\wedge t} - X^i_{\tau^i_k} - \rho^i(\tau^i_{k-1})\mathbb{1}_{\{\tau^i_k<\sigma^i\leq t\}}$.

It follows from the proof of Proposition 3 that the solo-payout index from time τ^i_{k-1} is realized by a τ^i_k such that

$$\operatorname*{ess\,sup}_{\tau'\in\hat{\mathbb{F}}^i(\tau^i_k)} \mathbb{E}^i\left[Z^i_{\tau'} \middle| \mathcal{F}^i(\tau^i_k)\right] \leq Z^i_{\tau^i_k}(\mathbb{P}^i\text{-a.e.}), \tag{74}$$

or

$$\operatorname*{ess\,sup}_{\tau'\in\hat{\mathbb{F}}^i(\tau^i_k)} \mathbb{E}^i\left[X^i_{\sigma^i\wedge\tau'} - X^i_{\tau^i_k} - \rho^i(\tau^i_{k-1})\mathbb{1}_{\{\tau^i_k<\sigma^i\leq\tau'\}} \middle| \mathcal{F}^i(\tau^i_k)\right] \leq 0(\mathbb{P}^i\text{-a.e.}). \tag{75}$$

From the above, for any $\tau' \in \hat{\mathbb{F}}^i(\tau^i_k)$, we have

$$\frac{\mathbb{E}^i\left[X^i_{\sigma^i\wedge\tau'} - X^i_{\tau^i_k} \middle| \mathcal{F}^i(\tau^i_k)\right]}{\mathbb{P}^i\left(\tau^i_k < \sigma^i \leq \tau' \middle| \mathcal{F}^i(\tau^i_k)\right)} \leq \rho^i(\tau^i_{k-1})(\mathbb{P}^i\text{-a.e.}). \tag{76}$$

Taking the essential supremum over such τ' establishes that $\rho^i(\tau^i_k) \leq \rho^i(\tau^i_{k-1})$, ($\mathbb{P}^i$-a.e.). ■

ACKNOWLEDGMENTS

We acknowledge support for this work from the National Science Foundation, NSF Grants: CMMI-1662629 and CMMI-1662442.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Michael N. Katehakis https://orcid.org/0000-0002-1511-7098

REFERENCES

Bank, P., & El Karoui, N. (2004). A stochastic representation theorem with applications to optimization and obstacle problems. *The Annals of Probability*, *32*, 1030–1067.

Chakravorty, J., & Mahajan, A. (2014). Multi-armed bandits, Gittins index, and its calculation. *Methods and Applications of Statistics in Clinical Trials*, *2*, 1–20.

Cowan, W., & Katehakis, M. N. (2015). Multi-armed bandits under general depreciation and commitment. *Probability in the Engineering and Informational Sciences*, *29*, 51–76.

Cowan, W., Honda, J., & Katehakis, M. N. (2018). Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, *18*, 1–28. http://jmlr.org/papers/v18/15-154.html

Denardo, E. V., Park, H., & Rothblum, U. G. (2007). Risk-sensitive and risk-neutral multiarmed bandits. *Mathematics of Operations Research*, *32*, 374–394.

Derman, C., & Sacks, J. (1960). Replacement of periodically inspected equipment.(an optimal optional stopping rule). *Naval Research Logistics Quarterly*, *7*, 597–607.

Dumitriu, I., Tetali, P., & Winkler, P. (2003). On playing golf with two balls. *SIAM Journal on Discrete Mathematics*, *16*, 604–615.

El Karoui, N., & Karatzas, I. (1994). Dynamic allocation problems in continuous time. *The Annals of Applied Probability*, *4*, 255–286.

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

Gittins, J. C., Glazebrook, K., & Weber, R. R. (2011). Multi-armed bandit allocation indices. John Wiley & Sons.

Glazebrook, K. D., Hodge, D. J., Kirkbride, C., & Minty, R. J. (2014). Stochastic scheduling: A short history of index policies and new approaches to index generation for dynamic resource allocation. *Journal of Scheduling*, *17*, 407–425.

Glazebrook, K. D., Kirkbride, C., Mitchell, H. M., Gaver, D. P., & Jacobs, P. A. (2007). Index policies for shooting problems. *Operations Research*, *55*, 769–781.

Ishikida, T., & Varaiya, P. (1994). Multi-armed bandit problem revisited. *Journal of Optimization Theory and Applications*, *83*, 113–154.

Katehakis, M. N., & Derman, C. (1986). Computing optimal sequential allocation rules in clinical trials. *Lecture Notes-Monograph Series*, *8*, 29–39.

Katehakis, M. N., & Rothblum, U. G. (1996). Finite state multi-armed bandit problems: Sensitive-discount, average-reward and average-overtaking optimality. *The Annals of Applied Probability*, *6*, 1024–1034.

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

Katta, A.-K., & Sethuraman, J. (2004). A note on bandits with a twist. *SIAM Journal on Discrete Mathematics*, *18*, 110–113.

Mahajan, A., & Teneketzis, D. (2008). *Multi-armed bandit problems*. In A. O. Hero, III, D. A. Castanon, D. Cochran, & K. Kastella (Eds.), Foundations and applications of sensor management (pp. 121–151). Springer.

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

Negoescu, D. M., Bimpikis, K., Brandeau, M. L., & Iancu, D. A. (2018). Dynamic learning of patient response types: An application to treating chronic diseases. *Management Science*, *64*, 3469–3970.

Pinedo, M., & Rammouz, E. (1988). A note on stochastic scheduling on a single machine subject to breakdown and repair. *Probability in the Engineering and Informational Sciences*, *2*, 41–49.

Pinedo, M. L. (2012). Scheduling: Theory, algorithms, and systems. Springer Verlag.

Skitsas, K., Papageorgiou, I. G., Talebi, M. S., Kantere, V., Katehakis, M. N., & Karras, P. (2022). Sifter: Space-efficient value iteration for finite-horizon MDPs. *Proceedings of the VLDB Endowment*, *16*, 90–98.

Slivkins, A. (2019). Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, *12*, 1–286.

Snell, J. L. (1952). Applications of martingale system theorems. *Transactions of the American Mathematical Society*, *73*, 293–312.

Sonin, I. M. (2008). A generalized Gittins index for a Markov chain and its recursive calculation. *Statistics & Probability Letters*, *78*, 1526–1533.

Sonin, I. M. (2011). Optimal stopping of Markov chains and three abstract optimization problems. *Stochastics an International Journal of Probability and Stochastic Processes*, *83*, 405–414.

Sonin, I. M., & Steinberg, C. (2016). Continue, quit, restart probability model. *Annals of Operations Research*, *241*, 295–318.

Stadje, W. (1995). Selecting jobs for scheduling on a machine subject to failure. *Discrete Applied Mathematics*, *63*, 257–265.

Szepesvári, C. (2010). Algorithms for reinforcement learning *Synthesis Lectures on Artificial Intelligence and Machine Learning* (Vol. *4*, pp. 1–103). Morgan & Claypool Publishers.

Talebi, M. S., Jonsson, A., & Maillard, O. (2021). *Improved exploration in factored average-reward MDPs*. In International conference on Artificial Intelligence and Statistics (pp. 3988–3996). PMLR.

Villar, S. S., Bowden, J., & Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, *30*, 199–215.

Weber, R. (1992). On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, *2*, 1024–1033.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*, 143–149.

How to cite this article: Cowan, W., Katehakis, M. N., & Ross, S. M. (2023). Optimal activation of halting multi-armed bandit models. *Naval Research Logistics (NRL)*, *70*(7), 639–652. https://doi.org/10.1002/nav.22145