Probability in the Engineering and Informational Sciences, **29**, 2015, 51–76.
doi:10.1017/S0269964814000217

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.

 0269-9648/15 $25.00

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

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

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 $1 \leqslant i \leqslant N < \infty$, representing N environments in which experiments will be performed or rewards collected – the "bandits". To each space, we associate a *reward process* $X^i = \{X^i_t\}_{t\geqslant 0}$, and a *discount factor sequence* $\beta^i = \{\beta^i_t\}_{t\geqslant 0}$. 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 β^i to be $\mathbb{F}^i$-adapted. We denote the reward process collection as $\mathbb{X}$, and the discount factor sequence collection as $\mathbb{B}$.

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 $\mathbb{E}^i$, while expectations relative to the global space are simply $\mathbb{E}$. Note that Assumption A ensures that X^i, X^j are independent relative to $\mathbb{P}$ for $i \neq j$, so too are the β^i, β^j, though β^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 Ω^i to also be considered as a random variable on the global space Ω, taking $Z(\omega) = Z(\omega^i)$, where $\omega = (\omega^1, \ldots, \omega^N) \in \Omega$. Via this extension, we may take expectations involving a process X^i, or $\mathbb{F}^i$-stopping times, relative to $\mathbb{P}$ 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* $\{\alpha^i_t\}_{t\geqslant 0}$, such that α^i_t represents the total discounting incurred by the first t activations of bandit i. That

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

is, we may take $\alpha_0^i = 1$, and

$$\alpha_t^i = \prod_{t'=0}^{t-1} \beta_{t'}^i. \tag{1}$$

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

Assumption C:

$$\lim_{t\to\infty} \alpha_t^i = \prod_{t=0}^{\infty} \beta_t^i = 0 (\mathbb{P}^i - \text{a.e.}), \tag{2}$$

and

$$\mathbb{E}^i \left[\sum_{t=0}^{\infty} \alpha_t^i |X_t^i| \right] < \infty. \tag{3}$$

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

$$\begin{aligned} S_\pi^i(0) &= \inf\{s \geqslant 0 : \pi(s) = i\}, \\ S_\pi^i(t+1) &= \inf\{s > S_\pi^i(t) : \pi(s) = i\}. \end{aligned} \tag{4}$$

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

$$T_\pi^i(s) = \sum_{s'=0}^{s-1} \mathbb{1}\{\pi(s') = i\}. \tag{5}$$

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 π. This will allow us to define concise global time analogs of several processes. For instance, we define the *global reward process* X_π on $(\Omega, \mathcal{G}, \mathbb{P})$ as $X_\pi(s) = X_{T_\pi(s)}^{\pi(s)}$, giving the reward from collection $\mathbb{X}$ under policy π at round s.

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

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

While α_t^i gives the depreciation on X_t^i due to the activations of bandit i, given a policy π it is also convenient to define the depreciation on X_t^i on its collection, due to activations of any process under that policy. This policy-dependent depreciation is given by $\alpha_\pi^i(t) = A_\pi(S_\pi^i(t))$.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

In what follows, we let $V_\pi(\mathbb{X},\mathbb{B})$ denote the value of a policy, the expected total reward given the reward–discount pair $\mathbb{X},\mathbb{B}$ 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

$$V_\pi(\mathbb{X},\mathbb{B})=\mathbb{E}\left[\sum_{s=0}^{\infty}A_\pi(s)X_\pi(s)\big|\mathcal{G}_0\right], \tag{7}$$

relative to global time, or relative to local time as

$$V_\pi(\mathbb{X},\mathbb{B})=\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=0}^{\infty}\alpha_\pi^i(t)X_t^i\big|\mathcal{G}_0\right]. \tag{8}$$

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

$$V_\pi(\mathbb{X},\mathbb{B})\leqslant V_{\pi^*}(\mathbb{X},\mathbb{B})(\mathbb{P}-\text{a.e.}). \tag{9}$$

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 $\mathbb{F}^i=\{\mathcal{F}^i(t)\}_{t\geqslant 0}$ 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 the process history of X^i. Taking X^i as $\mathbb{F}^i$-adapted as we do, we have $\sigma(X_0^i,X_1^i,\ldots,X_t^i)\subset\mathcal{F}^i(t)$.[1]

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 π until round s, the global information available at round s is given by the σ-algebra $\bigotimes_{i=1}^N\mathcal{F}^i(T_\pi^i(s))$. We may therefore refine the prior definition of non-anticipatory policies to be the set of policies $\mathcal{P}$ such that for each $s\geqslant 0$, $\pi(s)$ 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, e.g., coin flips, are addressed in Section 6.

Additionally, given a policy π, it is necessary to define a set of policy-dependent filtrations in the following way: let $\mathbb{H}_\pi^i=\{\mathcal{H}_\pi^i(t)\}_{t\geqslant 0}$, where $\mathcal{H}_\pi^i(t)=\bigotimes_{j=1}^N\mathcal{F}^j(T_\pi^j(S_\pi^i(t)))$ represents the total information available to the controller about all bandits, prior to the t^{th} 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, 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_t^i 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_t^i to remain uncertain just prior to its collection, that is have X_t^i 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 $\hat{X}_t^i=\mathbb{E}^i\left[X_t^i|\mathcal{F}^i(t)\right]$.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

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 $\mathbb{J}$ (e.g., $\mathbb{J} = \mathbb{F}^i, \mathbb{H}^i_\pi$), 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.

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 π with $S^i_\pi(t') < \infty$, we define the following quantities.*

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

$$\rho^i(t', t'') = \frac{\mathbb{E}^i\left[\sum_{t=t'}^{t''-1} \alpha^i_t X^i_t \middle| \mathcal{F}^i(t')\right]}{\mathbb{E}^i\left[\sum_{t=t'}^{t''-1} \alpha^i_t (1-\beta^i_t) \middle| \mathcal{F}^i(t')\right]}. \tag{10}$$

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

$$\nu^i_\pi(t', t'') = \frac{\mathbb{E}\left[\sum_{t=t'}^{t''-1} \alpha^i_\pi(t) X^i_t \middle| \mathcal{H}^i_\pi(t')\right]}{\mathbb{E}\left[\sum_{t=t'}^{t''-1} \alpha^i_\pi(t) (1-\beta^i_t) \middle| \mathcal{H}^i_\pi(t')\right]}. \tag{11}$$

Note, the above quantities are all measurable with respect to the indicated σ-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

$$\rho^i(t', t'') = \frac{\mathbb{E}^i\left[\sum_{t=t'}^{t''-1} \alpha^i_t X^i_t \middle| \mathcal{F}^i(t')\right]}{\mathbb{E}^i\left[\alpha^i_{t'} - \alpha^i_{t''} \middle| \mathcal{F}^i(t')\right]}. \tag{12}$$

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 ρ^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.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

Note, ν_π^i as in Eq. (11) is the obvious policy-dependent generalization of ρ^i in Eq. (10), rather than ρ^i as in Eq. (12). It has no immediate "restart"-type interpretation in this form. Taking view that the π-block value ν_π^i is the value of a block of X^i when activated under a specific policy π, ρ^i should be viewed as the value of a block with respect to consecutive activation.

The following theorem illustrates the relationship between ρ^i and ν_π^i, essentially stating that the value of any block under some policy π 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* $(\Omega, \mathcal{J}, \mathbb{P})$ *consider a discrete-time process* $\{Z_t\}_{t\geqslant 0}$ *such that* $\mathbb{E}\left[\sum_{t=0}^{\infty}|Z_t|\right] < \infty$. *Let* $\mathbb{J} = \{\mathcal{J}_t\}_{t\geqslant 0}$ *be a filtration, and* $\{\alpha_t\}_{t\geqslant 0}$ *be a* $\mathbb{J}$*-adapted process such that* $\alpha_t \geqslant \alpha_{t+1} \geqslant 0 (\mathbb{P} - a.e.)$. *In such a case the following is true for any* $\tau \in \hat{\mathcal{J}}(0)$*:*

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1}\alpha_t Z_t \big| \mathcal{J}_0\right] \leqslant \alpha_0 \operatorname*{ess\,sup}_{\tau' \in \hat{\mathbb{J}}(0)} \mathbb{E}\left[\sum_{t=0}^{\tau'-1} Z_t \big| \mathcal{J}_0\right] (\mathbb{P} - a.e.). \tag{13}$$

THEOREM 1 (Block Value Comparison): *For bandit* i *under policy* π*, for any time* t_0 *such that* $S_\pi^i(t_0) < \infty$*, the following holds for any* $\mathbb{H}_\pi^i$*-stopping time* τ *with* $t_0 < \tau$*:*

$$\nu_\pi^i(t_0, \tau) \leqslant \operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)} \rho^i(t_0, \hat{\tau}) (\mathbb{P} - \text{a.e.}). \tag{14}$$

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

PROOF: Denote the essential supremum above by ρ, which is $\mathcal{F}^i(t_0)$-measurable. Note, for any $\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)$,

$$\frac{\mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1}\alpha_t^i X_t^i \big| \mathcal{F}^i(t_0)\right]}{\mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1}\alpha_t^i(1-\beta_t^i) \big| \mathcal{F}^i(t_0)\right]} \leqslant \rho \ (\mathbb{P}^i - \text{a.e.}). \tag{15}$$

We may rearrange Eq. (15) to yield

$$\mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1}\alpha_t^i(X_t^i - \rho(1-\beta_t^i)) \big| \mathcal{F}^i(t_0)\right] \leqslant 0 \ (\mathbb{P}^i - \text{a.e.}). \tag{16}$$

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

$$\operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)} \mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1}\alpha_t^i(X_t^i - \rho(1-\beta_t^i)) \big| \mathcal{F}^i(t_0)\right] \leqslant 0 \ (\mathbb{P}^i - \text{a.e.}). \tag{17}$$

To demonstrate that $\nu_\pi^i(t_0, \tau) \leqslant \rho (\mathbb{P} - \text{a.e.})$, it is enough to show that

$$\mathbb{E}\left[\sum_{t=t_0}^{\tau-1}\alpha_\pi^i(t)(X_t^i - \rho(1-\beta_t^i)) \big| \mathcal{H}_\pi^i(t_0)\right] \leqslant 0 \ (\mathbb{P} - \text{a.e.}). \tag{18}$$

It is useful to factor α_π^i into contributions to the discounting from bandit i and contributions from the remaining bandits. That is, we take

$$\alpha_\pi^i(t) = \hat{A}_\pi^i(S_\pi^i(t))\alpha^i(t), \tag{19}$$

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

where

$$\hat{A}^i_\pi(s) = \prod_{j\neq i} \alpha^j_{T^j_\pi(s)}. \tag{20}$$

In the spirit of applying Lemma A.1, note that $\hat{A}^i_\pi(S^i_\pi(t))$ is $\mathbb{H}^i_\pi$-adapted, and

$$1 \geqslant \hat{A}^i_\pi(S^i_\pi(t)) \geqslant \hat{A}^i_\pi(S^i_\pi(t+1)) \geqslant 0 \ (\mathbb{P}-\text{a.e.}).$$

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 $Z^i_t = X^i_t - \rho\,(1-\beta^i_t)$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=t_0}^{\tau-1} \alpha^i_\pi(t) Z^i_t \middle| \mathcal{H}^i_\pi(t_0)\right] &= \mathbb{E}\left[\sum_{t=t_0}^{\tau-1} \hat{A}^i_\pi(S^i_\pi(t))\alpha^i_t Z^i_t \middle| \mathcal{H}^i_\pi(t_0)\right] \\
&\leqslant \hat{A}^i_\pi(S^i_\pi(t_0)) \operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{H}}^i_\pi(t_0)} \mathbb{E}\left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha^i_t Z^i_t \middle| \mathcal{H}^i_\pi(t_0)\right] \\
&\leqslant \hat{A}^i_\pi(S^i_\pi(t_0)) \operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^i(t_0)} \mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha^i_t Z^i_t \middle| \mathcal{F}^i(t_0)\right] \\
&\leqslant 0
\end{aligned} \tag{21}$$

the above relations holding ($\mathbb{P}$-a.e.). The exchange from the essential sup over $\mathbb{H}^i_\pi$-stopping times to $\mathbb{F}^i$-stopping times follows as the α^i_t, Z^i_t 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. ■

The following proposition provides, using ρ^i and ν^i_π, 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.*

$$\mathbb{E}^i\left[\sum_{t=\tau'}^{\tau''-1} \alpha^i_t X^i_t \middle| \mathcal{F}^i(\tau')\right] = \mathbb{E}^i\left[\sum_{t=\tau'}^{\tau''-1} \alpha^i_t \rho^i(\tau',\tau'')(1-\beta^i_t) \middle| \mathcal{F}^i(\tau')\right], \tag{22}$$

$$\mathbb{E}\left[\sum_{t=\tau'}^{\tau''-1} \alpha^i_t X^i_t \middle| \mathcal{H}^i_\pi(\tau')\right] = \mathbb{E}\left[\sum_{t=\tau'}^{\tau''-1} \alpha^i_t \rho^i(\tau',\tau'')(1-\beta^i_t) \middle| \mathcal{H}^i_\pi(\tau')\right], \tag{23}$$

$$\mathbb{E}\left[\sum_{t=\tau'}^{\tau''-1} \alpha^i_\pi(t) X^i_t \middle| \mathcal{H}^i_\pi(\tau')\right] = \mathbb{E}\left[\sum_{t=\tau'}^{\tau''-1} \alpha^i_\pi(t) \nu^i_\pi(\tau',\tau'')(1-\beta^i_t) \middle| \mathcal{H}^i_\pi(\tau')\right]. \tag{24}$$

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 $\nu^i_\pi(\tau',\tau'')(1-\beta^i_t)$ under π, for $\tau' \leqslant t < \tau''$. This will prove to be central to the final proof.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

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*

$$\rho^i(t) = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{F}}^i(t)} \rho^i(t, \tau). \tag{25}$$

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, $\beta_t^i = \beta$ for some β, 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 β_t^i 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 ($\mathbb{P}^i - $ 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 $\{Z_t\}_{t \geqslant 0}$ such that $\mathbb{E}\left[\sum_{t=0}^{\infty} |Z_t|\right] < \infty$. Let $\mathbb{J} = \{\mathcal{J}_t\}_{t \geqslant 0}$ be a filtration such that the Z_t are $\mathbb{J}$-adapted. Then, there exists a (potentially infinite) stopping time $\tau^* \in \hat{\mathbb{J}}(0)$ such that*

$$\mathbb{E}\left[\sum_{t=0}^{\tau^*-1} Z_t \middle| \mathcal{J}_0\right] = \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{J}}(0)} \mathbb{E}\left[\sum_{t=0}^{\tau-1} Z_t \middle| \mathcal{J}_0\right] (\mathbb{P} - \text{a.e.}). \tag{26}$$

In particular, we may take

$$\tau^* = \inf\left\{ n > 0 : \operatorname*{ess\,sup}_{\tau \in \hat{\mathbb{J}}(n)} \mathbb{E}\left[\sum_{t=n}^{\tau-1} Z_t \middle| \mathcal{J}_n\right] < 0 \right\}. \tag{27}$$

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 $\tau \in \hat{\mathbb{F}}^i(t_0)$ such that $\rho^i(t_0) = \rho^i(t_0, \tau)(\mathbb{P}^i - \text{a.e.})$.*

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

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

$$\mathbb{E}^i \left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha_t^i \left(X_t^i - \rho^i(t_0)(1-\beta_t^i) \right) \middle| \mathcal{F}^i(t_0) \right] \leqslant 0 \ (\mathbb{P}^i - \text{a.e.}). \tag{28}$$

Defining

$$\epsilon = -\operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)} \mathbb{E}^i \left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha_t^i \left(X_t^i - \rho^i(t_0)(1-\beta_t^i) \right) \middle| \mathcal{F}^i(t_0) \right], \tag{29}$$

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

$$\begin{aligned} \rho^i(t_0, \hat{\tau}) &\leqslant \rho^i(t_0) - \frac{\epsilon}{\mathbb{E}^i \left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha_t^i (1-\beta_t^i) \middle| \mathcal{F}^i(t_0) \right]} \\ &= \rho^i(t_0) - \frac{\epsilon}{\mathbb{E}^i \left[\alpha_{t_0}^i - \alpha_{\hat{\tau}}^i \middle| \mathcal{F}^i(t_0) \right]} \\ &\leqslant \rho^i(t_0) - \epsilon (\mathbb{P}^i - \text{a.e.}). \end{aligned} \tag{30}$$

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

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

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

$$\operatorname*{ess\,sup}_{\hat{\tau} \in \hat{\mathbb{F}}^i(t_0)} \mathbb{E}^i \left[\sum_{t=t_0}^{\hat{\tau}-1} \alpha_t^i \left(X_t^i - \rho^i(t_0)(1-\beta_t^i) \right) \middle| \mathcal{F}^i(t_0) \right] = 0 \ (\mathbb{P}^i - \text{a.e.}). \tag{32}$$

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 $\tau^* \in \hat{\mathbb{F}}^i(t_0)$ such that

$$\mathbb{E}^i \left[\sum_{t=t_0}^{\tau^*-1} \alpha_t^i \left(X_t^i - \rho^i(t_0)(1-\beta_t^i) \right) \middle| \mathcal{F}^i(t_0) \right] = 0 \ (\mathbb{P}^i - \text{a.e.}), \tag{33}$$

or

$$\rho^i(t_0) = \frac{\mathbb{E}^i \left[\sum_{t=t_0}^{\tau^*-1} \alpha_t^i X_t^i \middle| \mathcal{F}^i(t_0) \right]}{\mathbb{E}^i \left[\sum_{t=t_0}^{\tau^*-1} \alpha_t^i (1-\beta_t^i) \middle| \mathcal{F}^i(t_0) \right]} = \rho^i(t_0, \tau^*) \ (\mathbb{P}^i - \text{a.e.}). \tag{34}$$

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

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

DEFINITION 3 (Restart-in-State Index Times): *Define a sequence of* $\mathbb{F}^i$*-stopping times* $\{\tau_k^i\}_{k\geqslant 0}$ *in the following way, that* $\tau_0^i = 0$*, and for* $k > 0$*,*

$$\tau_{k+1}^i = arg\,\text{ess sup}\{\rho^i(\tau_k^i, \tau) : \tau \in \hat{\mathbb{F}}^i(\tau_k^i)\}. \tag{35}$$

In the case that τ_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 $\rho^i(\tau_k^i) = \rho^i(\tau_k^i, \tau_{k+1}^i)$. 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 τ_{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:*

$$\rho^i(\tau_{k-1}^i) \geqslant \rho^i(\tau_k^i)(\mathbb{P}^i - \text{a.e.}).$$

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 τ_{k-1}^i is realized ($\mathbb{P}^i$ – a.e.) by a τ_k^i such that, via Lemma 2,

$$\underset{\hat{\tau}\in\hat{\mathbb{F}}^i(\tau_k^i)}{\text{ess sup}}\mathbb{E}^i\left[\sum_{t=\tau_k^i}^{\hat{\tau}-1}\alpha_t^i\left(X_t^i - \rho^i(\tau_{k-1}^i)(1-\beta_t^i)\right)\Big|\mathcal{F}^i(\tau_k^i)\right] < 0\ (\mathbb{P}^i - \text{a.e.}). \tag{36}$$

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

$$\frac{\mathbb{E}^i\left[\sum_{t=\tau_k^i}^{\hat{\tau}-1}\alpha_t^i X_t^i\big|\mathcal{F}^i(\tau_k^i)\right]}{\mathbb{E}^i\left[\sum_{t=\tau_k^i}^{\hat{\tau}-1}\alpha_t^i(1-\beta_t^i)\big|\mathcal{F}^i(\tau_k^i)\right]} < \rho^i(\tau_{k-1}^i)(\mathbb{P}^i - \text{a.e.}). \tag{37}$$

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

4.1. Necessity of the Restart in State Indices

We construct the following toy example to demonstrates 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 $V_{\text{DAI}} = 100 + 1b + 2ab$.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

However, consider the alternative strategy of activating X^1 twice first, then activating X^2. This gives a value of $V' = 1 + 2\,a + 100\,a^2$. Comparing the two, we have

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

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* $\mathbb{X} = (X^1, X^2, \ldots, X^N)$, *discount factor sequences* $\mathbb{B} = (\beta^1, \beta^2, \ldots, \beta^N)$, *and* $\{\tau_k^i\}_{k \geqslant 0}$ *as by Definition 3, we define:*

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

$$Y^i(t) = \rho^i(\tau_k^i)(1-\beta_t^i), \quad \text{if } \tau_k^i \leqslant t < \tau_{k+1}^i. \tag{38}$$

2. *For* $\pi \in \mathcal{P}$, *the* π*-equivalent collection* $\mathbb{Y}_\pi^X = (Y_\pi^1, \ldots, Y_\pi^N)$, *by*

$$Y_\pi^i(t) = \nu_\pi^i(\tau_k^i, \tau_{k+1}^i)(1-\beta_t^i), \quad \text{if } \tau_k^i \leqslant t < \tau_{k+1}^i. \tag{39}$$

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 τ_k^i. However, as the $\nu_\pi^i(\tau_k^i, \tau_{k+1}^i)$ depend on the specifics of a policy π, so do the Y_π^i processes; the Y_π^i processes are $\mathbb{H}_\pi^i$-adapted, but not $\mathbb{F}^i$-adapted. Note that, should bandit i be activated only finitely many times under π, Y_π^i will only really be defined up to some τ_{k+1}^i such that $S_\pi^i(\tau_{k+1}^i) = \infty$. For such undefined $Y_\pi^i(t)$, we take $0 * Y_\pi^i(t) = 0$.

The following are simple, but important, properties of the Y^i, Y_π^i processes.

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

$$\mathbb{E}^i\left[\sum_{t=\tau_k^i}^{\tau_{k+1}^i - 1} \alpha_t^i X_t^i \big| \mathcal{F}^i(\tau_k^i)\right] = \mathbb{E}^i\left[\sum_{t=\tau_k^i}^{\tau_{k+1}^i - 1} \alpha_t^i Y^i(t) \big| \mathcal{F}^i(\tau_k^i)\right], \tag{40}$$

$$\mathbb{E}\left[\sum_{t=\tau_k^i}^{\tau_{k+1}^i - 1} \alpha_t^i X_t^i \big| \mathcal{H}_\pi^i(\tau_k^i)\right] = \mathbb{E}\left[\sum_{t=\tau_k^i}^{\tau_{k+1}^i - 1} \alpha_t^i Y^i(t) \big| \mathcal{H}_\pi^i(\tau_k^i)\right], \tag{41}$$

$$\mathbb{E}\left[\sum_{t=\tau_k^i}^{\tau_{k+1}^i - 1} \alpha_\pi^i(t) X_t^i \big| \mathcal{H}_\pi^i(\tau_k^i)\right] = \mathbb{E}\left[\sum_{t=\tau_k^i}^{\tau_{k+1}^i - 1} \alpha_\pi^i(t) Y_\pi^i(t) \big| \mathcal{H}_\pi^i(\tau_k^i)\right]. \tag{42}$$

As with Proposition 1, equality also holds when conditioning with respect to $\mathcal{F}^i(0), \mathcal{G}_0$.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

PROOF: This follows as an application of Proposition 1 and the definitions of Y^i, Y^i_π. ■

The following proposition serves as a justification of the term "equivalent" to describe Y^i, Y^i_π.

PROPOSITION 5: *For each bandit i,*

$$\mathbb{E}^i\left[\sum_{t=0}^{\infty}\alpha_t^i X_t^i\big|\mathcal{F}^i(0)\right]=\mathbb{E}^i\left[\sum_{t=0}^{\infty}\alpha_t^i Y^i(t)\big|\mathcal{F}^i(0)\right], \tag{43}$$

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\alpha_\pi^i(t) X_t^i\big|\mathcal{G}_0\right]=\mathbb{E}\left[\sum_{t=0}^{\infty}\alpha_\pi^i(t) Y_\pi^i(t)\big|\mathcal{G}_0\right]. \tag{44}$$

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:*

$$\alpha_\pi^i(t)Y_\pi^i(t)\leqslant\alpha_\pi^i(t)Y^i(t)\quad(\mathbb{P}-\text{a.e.}). \tag{45}$$

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

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

$$Y_\pi^i(t)=\nu_\pi^i(\tau_k^i,\tau_{k+1}^i)(1-\beta_t^i)\leqslant\rho^i(\tau_k^i)(1-\beta_t^i)=Y^i(t)\ (\mathbb{P}-\text{a.e.}). \tag{46}$$

■

6. A GREEDY RESULT AND THE OPTIMAL CONTROL POLICY

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

$$\sum_{s=0}^{\infty}A_\pi(s)Y_\pi^X(s)\leqslant\sum_{s=0}^{\infty}A_{\pi^*}(s)Y_{\pi^*}^X(s)(\mathbb{P}-\text{a.e.}). \tag{47}$$

In particular, such a π^* is given by the following rule: activate bandit i corresponding to the largest current value of $\rho^i(\tau_k^i)$, for a duration corresponding to the realizing $[\tau_k^i,\tau_{k+1}^i)$-block, repeating this ad infinitum.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

PROOF: Without loss of generality, let $\rho^1(0) \geqslant \rho^i(0)$ for all i. Since the $\rho^i(\tau_k^i)$ are decreasing with k, we have for all i, k that

$$\rho^1(0) \geqslant \rho^i(\tau_k^i)(\mathbb{P} - \text{a.e.}). \tag{48}$$

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

We may express the total reward under π sample path wise as

$$\begin{aligned} R_\pi &= \sum_{s=0}^{\infty} A_\pi(s) Y_\pi^X(s) \\ &= \sum_{s=0}^{S-1} A_\pi(s) Y_\pi^X(s) + A_\pi(S) Y^1(0) + \sum_{s=S+1}^{\infty} A_\pi(s) Y_\pi^X(s). \end{aligned} \tag{49}$$

From π, we construct a policy $\pi' \in \mathcal{P}$ in the following way: π' is identical to π 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,

$$\pi'(s) = \begin{cases} 1 & \text{for } s = 0, \\ \pi(s-1) & \text{for } s = 1, 2, \ldots S, \\ \pi(s) & \text{for } s \geqslant S+1. \end{cases} \tag{50}$$

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

Using this relation, we may express the sample path wise reward under policy π', relative to π, as

$$\begin{aligned} R_{\pi'} &= \sum_{s=0}^{\infty} A_{\pi'}(s) Y_{\pi'}^X(s) \\ &= Y^1(0) + \sum_{s=1}^{S} A_{\pi'}(s) Y_{\pi'}^X(s) + \sum_{s=S+1}^{\infty} A_{\pi'}(s) Y_{\pi'}^X(s) \\ &= Y^1(0) + \sum_{s=1}^{S} \beta_0^1 A_\pi(s-1) Y_\pi^X(s-1) + \sum_{s=S+1}^{\infty} A_\pi(s) Y_\pi^X(s) \\ &= Y^1(0) + \beta_0^1 \sum_{s=0}^{S-1} A_\pi(s) Y_\pi^X(s) + \sum_{s=S+1}^{\infty} A_\pi(s) Y_\pi^X(s). \end{aligned} \tag{51}$$

Comparing the two, we have

$$\begin{aligned} R_{\pi'} - R_\pi &= \left(Y^1(0) + \beta_0^1 \sum_{s=0}^{S-1} A_\pi(s) Y_\pi^X(s) \right) - \left(\sum_{s=0}^{S-1} A_\pi(s) Y_\pi^X(s) + A_\pi(S) Y^1(0) \right) \\ &= Y^1(0)(1 - A_\pi(S)) - (1 - \beta_0^1) \sum_{s=0}^{S-1} A_\pi(s) Y_\pi^X(s) \\ &= Y^1(0)(1 - A_\pi(S)) - \sum_{s=0}^{S-1} A_\pi(s)(1 - \beta_0^1) Y_\pi^X(s). \end{aligned} \tag{52}$$

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

Defining $\beta_\pi(s) = \beta^{\pi(s)}_{T_\pi(s)}$, we have by Eq. (48) that $(1-\beta^1_0)Y^X_\pi(s) \leqslant (1-\beta_\pi(s))Y^1(0)$. Hence,

$$\begin{aligned} R_{\pi'} - R_\pi &\geqslant Y^1(0)(1-A_\pi(S)) - \sum_{s=0}^{S-1} A_\pi(s)(1-\beta_\pi(s))Y^1(0) \\ &= Y^1(0)(1-A_\pi(S)) - Y^1(0)\sum_{s=0}^{S-1}(A_\pi(s) - A_\pi(s)\beta_\pi(s)) \\ &= Y^1(0)(1-A_\pi(S)) - Y^1(0)\sum_{s=0}^{S-1}(A_\pi(s) - A_\pi(s+1)) \\ &= Y^1(0)(1-A_\pi(S)) - Y^1(0)\,(1-A_\pi(S)) \\ &= 0\ (\mathbb{P}-\text{a.e.}). \end{aligned} \tag{53}$$

Immediately, for any policy π, advancing the first activation of the bandit with the largest current ρ^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 ρ-greedy advancements almost surely improves or does not change the value of the policy. Collisions, when two bandits have the same current ρ-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 ρ-greedy strategy as in the theorem. Let $\pi^* \in \mathcal{P}$ be the completely ρ-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 π after N-many ρ-greedy advancements. Notice, π^* and π_N agree for the first N rounds. Let $\tilde{\tau}^i_N = T^i_{\pi^*}(N)$. Recalling the definition of $\hat{A}^i_\pi(s)$, cf. Eq. ¡eqref¿eq:Ahat¡/eqref¿, as the discounting due to non-i bandits at round s, we have the following bound:

$$\begin{aligned} |R_{\pi^*} - R_{\pi_N}| &\leqslant \sum_{i=1}^{N}\sum_{t=\tilde{\tau}^i_N}^{\infty} |\alpha^i_{\pi^*}(t) - \alpha^i_{\pi_N}(t)||Y^i(t)| \\ &\leqslant \sum_{i=1}^{N}\sum_{t=\tilde{\tau}^i_N}^{\infty} \alpha^i_t \hat{A}^i_{\pi^*}(S^i_{\pi^*}(\tilde{\tau}^i_N))|Y^i(t)| \\ &\leqslant \sum_{i=1}^{N} \hat{A}^i_{\pi^*}(S^i_{\pi^*}(\tilde{\tau}^i_N)) \left(\sum_{t=\tilde{\tau}^i_N}^{\infty} \alpha^i_t |Y^i(t)|\right). \end{aligned} \tag{54}$$

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

For a given bandit i under π^*, 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 π^*, then $\tilde{\tau}^i_N$ increases without bound with N. As such, $\sum_{t=\tilde{\tau}^i_N}^{\infty} \alpha^i_t |Y^i(t)|$ converges to 0 ($\mathbb{P}$ – 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 π^*, then some bandit $j \neq i$ is activated infinitely many times under π^*. Thus, for some finite N, $S^i_{\pi^*}(\tilde{\tau}^i_N) = \infty$, and for the infinitely activated bandit $j \neq i$ we have $\hat{A}^i_{\pi^*}(S^i_{\pi^*}(\tilde{\tau}^i_N)) \leqslant \alpha^j_\infty = 0 (\mathbb{P} - \text{a.e.})$, by Eqs (2) and (20). In

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

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

The above imply the following:

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

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

Remark 4: An important feature of the optimal policy π^* for the $\mathbb{Y}^X$ processes is that it *preserves restart in state index blocks*. Because it is ρ-greedy, we have explicitly, for $\tau_k^i \leqslant \tau_k^i + t < \tau_{k+1}^i$, that $\pi^*(S_{\pi^*}^i(\tau_k^i)+t) = i$, ($\mathbb{P}$-a.e.).

We then have the following result.

Theorem 4 (The Optimal Control Policy for Generalized Deprecation): *For a collection of reward processes* $\mathbb{X} = (X^1, X^2, \ldots, X^N)$, *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* $\pi \in \mathcal{P}$,

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

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

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

$$V_\pi(\mathbb{X}, \mathbb{B}) = V_\pi(\mathbb{Y}_\pi^X, \mathbb{B}) \leqslant V_\pi(\mathbb{Y}^X, \mathbb{B}) \leqslant V_{\pi^*}(\mathbb{Y}^X, \mathbb{B}) = V_{\pi^*}(\mathbb{X}, \mathbb{B})(\mathbb{P}-\text{a.e.}), \tag{57}$$

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

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

We have, by Proposition 5,

$$V_\pi(\mathbb{X}, \mathbb{B}) = \sum_{i=1}^N \mathbb{E}\left[\sum_{t=0}^\infty \alpha_\pi^i(t) X_t^i \big| \mathcal{G}_0\right] = \sum_{i=1}^N \mathbb{E}\left[\sum_{t=0}^\infty \alpha_\pi^i(t) Y_\pi^i(t) \big| \mathcal{G}_0\right] = V_\pi(\mathbb{Y}_\pi^X, \mathbb{B}).$$

Note, because the Y_π^i 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 $V_\pi(\mathbb{Y}_\pi^X)$ is well defined via the above equation.

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

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

$$V_\pi(\mathbb{Y}_\pi^X, \mathbb{B}) = \sum_{i=1}^N \mathbb{E}\left[\sum_{t=0}^\infty \alpha_\pi^i(t) Y_\pi^i(t) \big| \mathcal{G}_0\right] \leqslant \sum_{i=1}^N \mathbb{E}\left[\sum_{t=0}^\infty \alpha_\pi^i(t) Y^i(t) \big| \mathcal{G}_0\right] = V_\pi(\mathbb{Y}^X, \mathbb{B}). \tag{58}$$

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

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

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

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

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

$$\begin{aligned} V_{\pi^*}(\mathbb{Y}^X, \mathbb{B}) &= \sum_{i=1}^{N} \sum_{k=0}^{\infty} \mathbb{E}\left[\hat{A}^i_\pi(S^i_\pi(\tau^i_k)) \mathbb{E}\left[\sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} \alpha^i_t Y^i(t) \middle| \mathcal{H}^i_\pi(\tau^i_k) \right] \middle| \mathcal{G}_0 \right] \\ &= \sum_{i=1}^{N} \sum_{k=0}^{\infty} \mathbb{E}\left[\hat{A}^i_\pi(S^i_\pi(\tau^i_k)) \mathbb{E}\left[\sum_{t=\tau^i_k}^{\tau^i_{k+1}-1} \alpha^i_t X^i_t \middle| \mathcal{H}^i_\pi(\tau^i_k) \right] \middle| \mathcal{G}_0 \right] \\ &= V_{\pi^*}(\mathbb{X}, \mathbb{B})(\mathbb{P} - \text{a.e.}). \end{aligned}$$

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 π^* is $\mathbb{P}$ − a.e. superior to any $\pi \in \mathcal{P}$, π^* would be similarly superior to any policy sampled randomly from $\mathcal{P}$.

7. A MODEL OF COMMITMENTS

One way of interpreting the discount factor β^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 $1 \leqslant i \leqslant N < \infty$, each indexed in continuous time, and satisfying the usual conditions. To each space, we associate a *continuous time reward rate process* $X^i = \{X^i_t\}_{t \geqslant 0}$. For $t \in [0, \infty)$, we take $X^i_t (= X^i_t(\omega^i)) \in \mathbb{R}$ 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 to be $\mathbb{F}^i$-adapted, and denote the collection of reward processes as $\mathbb{X}$.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

Additionally, to each bandit we associate a *commitment process* $c^i = \{c_k^i\}_{k\geqslant 0}$. 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 $c_k^i(\omega^i) \in \mathbb{R}^+$ before the controller is again free to activate from the N bandits as she chooses. We will refer to c_k^i as the duration of the k^{th} commitment epoch of bandit i. We denote the collection of commitment processes as $\mathbb{C}$.

For each bandit, it is convenient to define a sequence $\{C_k^i\}_{k\geqslant 0}$ of *process epoch times* by $C_0^i = 0$, and $C_k^i = \sum_{k'=0}^{k-1} c_{k'}^i$. Note that C_k^i gives the "local" time at which the k^{th} activation epoch of bandit i begins. In addition to the assumption that X^i is $\mathbb{F}^i$-adapted, we take each c_k^i as $\mathcal{F}^i(C_k^i)$-measurable, that is, the duration of a given epoch is determined by all the results prior to the start of that epoch. Note then that the C_k^i represent $\mathbb{F}^i$-stopping times.

In addition to Assumptions A and B, we take the following additional restrictions on each bandit i.

Assumption C′.

$$\lim_{k\to\infty} C_k^i = \sum_{k=0}^{\infty} c_k^i = \infty \ (\mathbb{P}^i - \text{a.e.}), \tag{59}$$

and

$$\mathbb{E}^i\left[\int_0^\infty e^{-rt}|X_t^i|dt\right] < \infty. \tag{60}$$

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 π 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 π, we may define continuous time versions of $T_\pi^i(s)$ and $S_\pi^i(t)$ as

$$T_\pi^i(s) = \int_0^s \mathbb{1}_{\{\pi(s')=i\}} ds', \tag{61}$$

and, utilizing the above,

$$S_\pi^i(t) = \inf\{s \geqslant 0 : T_\pi^i(s) \geqslant t\}. \tag{62}$$

Extending the ideas of Section 2.1, we define a policy π to be non-anticipatory if for $s \geqslant 0$, $\pi(s)$ is measurable with respect to $\bigotimes_{i=1}^N \mathcal{F}^i(T_\pi^i(s))$. However, we are only interested in the set $\mathcal{P}_C$ of non-anticipatory policies π that satisfy the commitment constraints, that is, for each bandit i, for each $k \geqslant 0$,

$$\begin{aligned} S_\pi^i(C_k^i + t) &= S_\pi^i(C_k^i) + t, \quad \text{for } 0 \leqslant t < c_k^i, \\ \pi(S_\pi^i(C_k^i + t)) &= i, \qquad\qquad\quad \text{for all } 0 \leqslant t < c_k^i. \end{aligned} \tag{63}$$

In parallel with the previous sections, we let $V_\pi^C(\mathbb{X}, \mathbb{C})$ denote the value of a policy, the expected total reward given the reward–commitment pair $(\mathbb{X}, \mathbb{C})$ under policy $\pi \in \mathcal{P}_C$. We

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

may therefore express the value of a policy as

$$V_\pi^C(\mathbb{X},\mathbb{C}) = \mathbb{E}\left[\int_0^\infty e^{-rs}X_\pi(s)ds\middle|\mathcal{G}_0\right], \tag{64}$$

relative to global time, or relative to local time as

$$V_\pi^C(\mathbb{X},\mathbb{C}) = \sum_{i=1}^N \mathbb{E}\left[\int_0^\infty e^{-rS_\pi^i(t)}X_t^i dt\middle|\mathcal{G}_0\right]. \tag{65}$$

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 $\pi \in \mathcal{P}_C$,

$$V_\pi^C(\mathbb{X},\mathbb{C}) \leqslant V_{\pi^*}^C(\mathbb{X},\mathbb{C})\ (\mathbb{P}-\text{a.e.}). \tag{66}$$

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

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

$$V_\pi^C(\mathbb{X},\mathbb{C}) = V_{\hat{\pi}}(\hat{\mathbb{X}},\hat{\mathbb{B}})\ (\mathbb{P}-\text{a.e.}), \tag{67}$$

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:

$$\begin{aligned}\hat{X}_k^i &:= \mathbb{E}^i\left[\int_0^{c_k^i} e^{-rt}X^i_{C_k^i+t}dt\middle|\mathcal{F}^i(C_k^i)\right],\\ \hat{\beta}_k^i &:= e^{-rc_k^i},\\ \hat{\mathcal{F}}^i(k) &:= \mathcal{F}^i(C_k^i).\end{aligned} \tag{68}$$

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 $\hat{X}^i = \{\hat{X}_k^i\}_{k\geqslant 0}$, a discount factor sequence $\hat{\beta}^i = \{\hat{\beta}_k^i\}_{k\geqslant 0}$, and a discrete-time filtration $\hat{\mathbb{F}}^i = \{\hat{\mathcal{F}}^i(k)\}_{k\geqslant 0}$ to which they are both adapted. We may then define the total depreciation sequence $\hat{\alpha}^i = \{\hat{\alpha}_k^i\}_{k\geqslant 0}$ by Eq. (1)

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

using $\hat{\beta}^i$, which yields $\hat{\alpha}^i_k = e^{-rC^i_k}$. Observe then that we have the following:

$$\lim_{k\to\infty} \hat{\alpha}^i_k = \lim_{k\to\infty} e^{-rC^i_k} = 0 \ (\mathbb{P}^i - \text{a.e.}), \tag{69}$$

and

$$\begin{aligned} \mathbb{E}^i\left[\sum_{k=0}^{\infty} \hat{\alpha}^i_k |\hat{X}^i_k|\right] &\leqslant \mathbb{E}^i\left[\sum_{k=0}^{\infty} \int_{C^i_k}^{C^i_{k+1}} e^{-rt}|X^i_t|dt\right] \\ &= \mathbb{E}^i\left[\int_0^{\infty} e^{-rt}|X^i_t|dt\right] < \infty. \end{aligned} \tag{70}$$

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

Taking a policy $\pi \in \mathcal{P}_C$ on $(\mathbb{X}, \mathbb{C})$, we may translate it into a discrete-time policy $\hat{\pi} \in \mathcal{P}$ on $(\hat{\mathbb{X}}, \hat{\mathbb{B}})$ in the following way: if the h^{th} decision made under policy π 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 π 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 $\hat{\alpha}^i_{\hat{\pi}}(k) = e^{-rS^i_{\pi}(C^i_k)}$, 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,

$$\begin{aligned} V^C_{\pi}(\mathbb{X}, \mathbb{C}) &= \sum_{i=1}^{N} \mathbb{E}\left[\int_0^{\infty} e^{-rS^i_{\pi}(t)} X^i_t dt \big| \mathcal{G}_0\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[\sum_{k=0}^{\infty} \int_{C^i_k}^{C^i_{k+1}} e^{-rS^i_{\pi}(t)} X^i_t dt \big| \mathcal{G}_0\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[\sum_{k=0}^{\infty} \int_0^{c^i_k} e^{-rS^i_{\pi}(C^i_k+t)} X^i_{C^i_k+t} dt \big| \mathcal{G}_0\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[\sum_{k=0}^{\infty} \int_0^{c^i_k} e^{-r(S^i_{\pi}(C^i_k)+t)} X^i_{C^i_k+t} dt \big| \mathcal{G}_0\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[\sum_{k=0}^{\infty} e^{-rS^i_{\pi}(C^i_k)} \int_0^{c^i_k} e^{-rt} X^i_{C^i_k+t} dt \big| \mathcal{G}_0\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[\sum_{k=0}^{\infty} \hat{\alpha}^i_{\hat{\pi}}(k) \hat{X}^i_k \big| \mathcal{G}_0\right] \\ &= V_{\hat{\pi}}(\hat{\mathbb{X}}, \hat{\mathbb{B}})(\mathbb{P} - \text{a.e.}). \end{aligned} \tag{71}$$

■

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

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

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, $c_k^i \in \mathbb{N}$. Everything done in the continuous case goes through for the discrete with $\hat{X}^i(k)$ and $\hat{\beta}_k^i$ defined by the following as an extension of Eq. (68):

$$\hat{X}^i(k) = \mathbb{E}^i\left[\sum_{t=0}^{c_k^i-1} \beta^t X^i(C_k^i + t) \big| \mathcal{F}^i(C_k^i)\right], \qquad \hat{\beta}_k^i = \beta^{c_k^i}. \tag{72}$$

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): $\prod_{t=0}^{\infty} \beta_t^i = 0$, $\sum_{k=0}^{\infty} c_k^i = \infty$ ($\mathbb{P}^i$ – 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 ρ-greedy policy improvements converged to the value of the infinite ρ-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. (2).

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, $\beta_t^i = \beta$, 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.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

Acknowledgement

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

References

1. Aalto, S., Ayesta, U. & Righter, R. (2011). Properties of the Gittins index with application to optimal scheduling. *Probability in the Engineering and Informational Sciences* 25: 269–288.
2. Agmon, N., Kraus, S. & Kaminka, G.A. (2008). Multi-robot perimeter patrol in adversarial settings. In *2008 IEEE International Conference on Robotics and Automation (ICRA 2008)*, pp. 2339–2345, Pasadena, CA: IEEE.
3. Agrawal, R, Hegde, M. & Teneketzis, D. (1990). Multi-armed bandit problems with multiple plays and switching cost. *Stochastics and Stochastic Reports* 29: 437–459.
4. Bertsekas, D.P. (2011). *Dynamic programming and optimal control*, vol. II, 3rd ed., Belmont, MA: Athena Scientific.
5. Bubeck, S. & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv:1204.5721.*
6. Burnetas, A.N. & Katehakis, M.N. (1997). Optimal adaptive policies for Markovian decision processes. *Mathematics of Operations Research* 22: 222–255.
7. Burnetas, A.N. & Katehakis, M.N. (2002). Asymptotic Bayes analysis for the finite horizon one armed bandit problem. *Probability in the Engineering and Informational Science* 17: 157–161.
8. Burnetas, A.N. & Katehakis, M.N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* 17: 122–142.
9. Caro, F. & Yoo, O.S. (2010). Indexability of bandit problems with response delays. *Probability in the Engineering and Informational Sciences* 24: 349–374.
10. Chang, F. & Lai, T.L. (1987). Optimal stopping and dynamic allocation. *Advances in Applied Probability* 19: 829–53.
11. Chung, K.L. (1982). *Lectures from Markov processes to Brownian motion*. Berlin: Springer-Verlag.
12. Denardo, E.V., Feinberg, E.A. & Rothblum, U.G. (2013). The multi-armed bandit, with constraints. In M.N. Katehakis, S.M. Ross, and J. Yang, (eds.), *Cyrus Derman Memorial Volume I: Optimization under Uncertainty: Costs, Risks and Revenues*, Annals of Operations Research, New York, NY: Springer.
13. Derman, C. & Sacks, J. (1960). Replacement of periodically inspected equipment (an optimal optional stopping rule). *Naval Research Logistics Quarterly* 7: 597–607.
14. El Karoui, N. & Karatzas, I. (1993). General Gittins index processes in discrete time. *Proceedings of the National Academy of Sciences* 90: 1232–1236.
15. Fernández-Gaucherand, E., Arapostathis, A. & Marcus, S.I. (1993). Analysis of an adaptive control scheme for a partially observed controlled Markov chain. *IEEE Transactions on Automatic Control* 38: 987–993.
16. Filippi, S., Cappé, O. & Garivier, A. (2010). Optimism in reinforcement learning and Kullback–Leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing*, pp. 115–122, Monticello, IL: IEEE.
17. Flint, M., Fernández, E. & Kelton, W.D. (2009). Simulation analysis for uav search algorithm design using approximate dynamic programming. *Military Operations Research* 14: 41–50.
18. Frostig, E. & Weiss, G. (2014). Four proofs of Gittins' multiarmed bandit theorem. In M.N. Katehakis, S.M. Ross, and J. Yang, (eds.), *Cyrus Derman Memorial Volume II: Optimization under Uncertainty: Costs, Risks and Revenues*, Annals of Operations Research, New York, NY: Springer.
19. Gittins, J.C., Glazebrook, K.D. & Weber, R.R. (2011). *Multi-armed bandit allocation indices*. West Sussex, UK: Wiley.
20. Gittins, J.C. & Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani, (ed.), *Progress in statistics*, pp. 241–66, Amsterdam, NL: North-Holland. Read at the 1972 European Meeting of Statisticians, Budapest.
21. Gittins, J.C. (1979). Bandit processes and dynamic allocation indices (with discussion). *Journal of Royal Statistics Society, Series B* 41: 335–340.
22. Gittins, J.C. (1989). *Multi-armed bandit allocation indices*. Chichester: Wiley.
23. Glazebrook, K.D., Hodge, D.J. & Kirkbride, C. (2011). General notions of indexability for queueing control and asset management. *The Annals of Applied Probability* 21: 876–907.
24. 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.

25. Govindarajulu, Z. & Katehakis, M.N. (1991). Dynamic allocation in survey sampling. *American Journal of Mathematical and Management Sciences* 11: 199–199.
26. Honda, J. & Takemura, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pp. 67–79.
27. Ishikida, T. & Varaiya, P. (1994). Multi-armed bandit problem revisited. *Journal of Optimization Theory and Applications* 83: 113–154.
28. Kaspi, H. & Mandelbaum, A. (1998). Multi-armed bandits in discrete and continuous time. *The Annals of Applied Probability* 8: 1270–1290.
29. Katehakis, M.N. & Derman, C. (1986). Computing optimal sequential allocation rules in clinical trials. *Lecture Notes-Monograph Series*, 8: 29–39.
30. 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.
31. Katehakis, M.N., Olkin, I., Ross, S.M. & Yang, J. (2013). On the life and work of Cyrus Derman. *Annals of Operations Research*, 208: 1–22.
32. Katehakis, M.N. & Robbins, H. (1995). Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America* 92: 8584–8585.
33. Katehakis, M.N. & Veinott, A.F. (1987). The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research* 12: 262–268.
34. Lai, L., El Gamal, H., Jiang, H. & Poor, V.H. (2008). Optimal medium access protocols for cognitive radio networks. In *6th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops*.
35. Lai, T.L. & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6: 4–22.
36. Liu, K., Zhao, Q. & Krishnamachari, B. (2010). Dynamic multichannel access with imperfect channel state detection. *IEEE Transactions on Signal Processing* 58: 2795–2808.
37. Mahajan, A. & Teneketzis, D. (2008). Multi-armed bandit problems. In A. O. Hero III, D. A. Castanon, D. Cochran, and K. Kastella (eds.), *Foundations and Applications of Sensor Management*, pp. 121–151, New York, NY: Springer.
38. Niño-Mora, J. (2006). Restless bandit marginal productivity indices, diminishing returns, and optimal control of make-to-order/make-to-stock $M/G/1$ queues. *Mathematical Methods of Operational Research* 31: 50–84.
39. Oksanen, J., Koivunen, V. & Poor, H. V. (2012). A sensing policy based on confidence bounds and a restless multi-armed bandit model. In *2012 Conference Record of the Forty-Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 318–323, Pacific Grove, CA: IEEE.
40. Ortner, P. & Auer, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. In *Proceedings of the 2006 Conference on Advances in Neural Information Processing Systems 19*, volume 19, p. 49, Vancouver, BC: MIT Press.
41. Ouyang, Y. & Teneketzis, D. (2013). On the optimality of myopic sensing in multi-state channels. *arXiv:1305.6993*.
42. Snell, J.L. (1952) Applications of martingale system theorems. *Transactions of the American Mathematical Society* 73: 293–312.
43. Sonin, I.M. (2008). A generalized Gittins index for a Markov chain and its recursive calculation. *Statistics and Probability Letters* 78: 1526–1533.
44. Sonin, I.M. (2011). Optimal stopping of Markov chains and three abstract optimization problems. *Stochastics* 83: 405–414.
45. Steinberg, C. & Sonin, I. (2014). Continue, quit, restart probability model. In M.N. Katehakis, S.M. Ross, and J. Yang (eds.), *Cyrus Derman Memorial Volume II: Optimization under Uncertainty: Costs, Risks and Revenues*, Annals of Operations Research, Springer.
46. Su, H., Qiu, M. & Wang, H. (2012). Secure wireless communication system for smart grid with rechargeable electric vehicles. *IEEE Communications Magazine* 50: 62–68.
47. Tekin, C. & Liu, M. (2011). Optimal adaptive learning in uncontrolled restless bandit problems. *arXiv:1107.4042*.
48. Tewari, A. & Bartlett, P.L. (2007). Optimistic linear programming gives logarithmic regret for irreducible MDPs. In J.C. Platt, D. Koller, Y. Singer and S.T. Roweis (eds.), *Advances in Neural Information Processing Systems*, pp. 1505–1512.
49. Tsitsiklis, J.N. (1994). A short proof of the Gittins index theorem. *The Annals of Applied Probability*, 27: 194–199.

50. Varaiya, P., Walrand, J. & Buyukkoc, C. (1985). Extensions of the multiarmed bandit problem: the discounted case. *IEEE Transactions on Automatic Control,* 30: 426–439.
51. Weber, R.R. (1992). On the Gittins index for multiarmed bandits. *The Annals of Applied Probability* 1024–1033.
52. Weber, R.R. & Weiss, G. (1990). On an index policy for restless bandits. *Journal of Applied Probability* 637–648.

APPENDIX A

A.1. Proof of Lemma A.1

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

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

PROOF: If the result can be shown for infinite τ, the case for arbitrary $\tau > 0$ follows simply, defining a new sequence $\hat{\alpha}_t = \alpha_t \mathbb{1}_{\{\tau>t\}}$. 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,

$$\mathbb{E}\left[\alpha_0 Z_0 + \alpha_1 Z_1\Big|\mathcal{J}_0\right] \tag{A.2}$$

is maximized taking α_1 to be the $\mathcal{J}_1$-measurable random variable defined by

$$\alpha_1 = \begin{cases} \alpha_0 & \text{if } \mathbb{E}\left[Z_1\Big|\mathcal{J}_1\right] > 0, \\ 0 & \text{if } \mathbb{E}\left[Z_1\Big|\mathcal{J}_1\right] \leqslant 0. \end{cases} \tag{A.3}$$

In short, if the remaining contribution is positive, make α_1 as large as possible (i.e. equal to α_0), else make α_1 as small as possible (i.e. equal to 0). Either way, this factors simply, giving in all cases,

$$\mathbb{E}\left[\alpha_0 Z_0 + \alpha_1 Z_1\Big|\mathcal{J}_0\right] \leqslant \alpha_0 \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(0),\tau\leqslant 2}\mathbb{E}\left[\sum_{t=0}^{\tau-1} Z_t\Big|\mathcal{J}_0\right](\mathbb{P}-\text{a.e.}). \tag{A.4}$$

We may apply this result inductively in the following way. Allowing $N < \infty$ terms,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{N-1}\alpha_t Z_t\Big|\mathcal{J}_0\right] &= \mathbb{E}\left[\alpha_0 Z_0 + \mathbb{E}\left[\sum_{t=1}^{N-1}\alpha_t Z_t\Big|\mathcal{J}_1\right]\Big|\mathcal{J}_0\right] \\ &\leqslant \mathbb{E}\left[\alpha_0 Z_0 + \alpha_1 \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(1),\tau\leqslant N}\mathbb{E}\left[\sum_{t=1}^{\tau-1} Z_t\Big|\mathcal{J}_1\right]\Big|\mathcal{J}_0\right].\end{aligned} \tag{A.5}$$

Again, the optimal choice of α_1 is the $\mathcal{J}_1$-measurable random variable given by α_0 if the essential sup is positive and 0 if it is negative. In either case, we have the following,

$$\mathbb{E}\left[\sum_{t=0}^{N-1}\alpha_t Z_t\Big|\mathcal{J}_0\right] \leqslant \alpha_0 \operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(0),\tau\leqslant N}\mathbb{E}\left[\sum_{t=0}^{\tau-1} Z_t\Big|\mathcal{J}_0\right](\mathbb{P}-\text{a.e.}). \tag{A.6}$$

This extends immediately to all finite, $\mathcal{J}_0$-measurable N. We next extend the finite result to the infinite.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

The initial assumption $\mathbb{E}\left[\sum_{t=0}^{\infty}|Z_t|\right]<\infty$ implies that $\mathbb{E}\left[\sum_{t=0}^{\infty}|Z_t||\mathcal{J}_0\right]<\infty$ ($\mathbb{P}-$a.e.). Thus for any fixed $\epsilon>0$, we may find a $\mathcal{J}_0$-measurable and finite $N>0$ ($\mathbb{P}-$a.e.) such that

$$\mathbb{E}\left[\sum_{t=N}^{\infty}|Z_t|\big|\mathcal{J}_0\right]\leqslant\epsilon\ (\mathbb{P}-\text{a.e.}). \tag{A.7}$$

Then,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{\infty}\alpha_t Z_t\big|\mathcal{J}_0\right]&=\mathbb{E}\left[\sum_{t=0}^{N-1}\alpha_t Z_t\big|\mathcal{J}_0\right]+\mathbb{E}\left[\sum_{t=N}^{\infty}\alpha_t Z_t\big|\mathcal{J}_0\right]\\&\leqslant\mathbb{E}\left[\sum_{t=0}^{N-1}\alpha_t Z_t\big|\mathcal{J}_0\right]+\mathbb{E}\left[\sum_{t=N}^{\infty}|Z_t|\big|\mathcal{J}_0\right]\\&\leqslant\alpha_0\operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(0),\tau\leqslant N}\mathbb{E}\left[\sum_{t=0}^{\tau-1}Z_t\big|\mathcal{J}_0\right]+\epsilon(\mathbb{P}-\text{a.e.})\\&\leqslant\alpha_0\operatorname*{ess\,sup}_{\tau\in\hat{\mathbb{J}}(0)}\mathbb{E}\left[\sum_{t=0}^{\tau-1}Z_t\big|\mathcal{J}_0\right]+\epsilon(\mathbb{P}-\text{a.e.}).\end{aligned} \tag{A.8}$$

The last step follows simply, extending the set of τ in question from stopping times at most N, to all possible stopping times in $\hat{\mathbb{J}}(0)$. 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. ■

APPENDIX B

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 $Z_t^i=X_t^i-\rho\,(1-\beta_t^i)$,

$$\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{H}}_\pi^i(t_0)}\mathbb{E}\left[\sum_{t=t_0}^{\hat{\tau}-1}\alpha_t^i Z_t^i\big|\mathcal{H}_\pi^i(t_0)\right]\leqslant\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^i(t_0)}\mathbb{E}^i\left[\sum_{t=t_0}^{\hat{\tau}-1}\alpha_t^i Z_t^i\big|\mathcal{F}^i(t_0)\right](\mathbb{P}-\text{a.e.}). \tag{B.1}$$

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 $W_{t'}^i=\sum_{t=t_0}^{t'-1}\alpha_t^i Z_t^i$.

Note that for any set $A\in\mathcal{H}_\pi^1(t_0)$, and any $\tau\in\hat{\mathbb{H}}_\pi^i(t_0)$,

$$\mathbb{E}\left[\mathbb{1}_A\mathbb{E}\left[W_\tau^1|\mathcal{H}_\pi^1(t_0)\right]\right]=\mathbb{E}\left[\mathbb{1}_A W_\tau^1\right]. \tag{B.2}$$

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

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_A W_\tau^1\right]&=\mathbb{E}^2\left[\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{1}_{A_2}W_\tau^1\right]\right]\\&=\mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}W_\tau^1\right]\right]\\&=\mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{E}^1\left[W_\tau^1|\mathcal{F}^1(t_0)\right]\right]\right].\end{aligned} \tag{B.3}$$

Observe that, while τ (being 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 a $\mathbb{F}^i$-stopping time. Hence, we have the

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press

bound

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_A W_\tau^1\right] &\leqslant \mathbb{E}^2\left[\mathbb{1}_{A_2}\mathbb{E}^1\left[\mathbb{1}_{A_1}\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^1(t_0)}\mathbb{E}^1\left[W_{\hat{\tau}}^1\middle|\mathcal{F}^1(t_0)\right]\right]\right] \\ &= \mathbb{E}^2\left[\mathbb{E}^1\left[\mathbb{1}_{A_1}\mathbb{1}_{A_2}\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^1(t_0)}\mathbb{E}^1\left[W_{\hat{\tau}}^1\middle|\mathcal{F}^1(t_0)\right]\right]\right] \qquad \textbf{(B.4)} \\ &= \mathbb{E}\left[\mathbb{1}_A\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^1(t_0)}\mathbb{E}^1\left[W_{\hat{\tau}}^1\middle|\mathcal{F}^1(t_0)\right]\right].\end{aligned}$$

Hence, for all rectangles $A \in \mathcal{H}_\pi^1(t_0)$,

$$\mathbb{E}\left[\mathbb{1}_A\mathbb{E}\left[W_\tau^1\middle|\mathcal{H}_\pi^1(t_0)\right]\right] \leqslant \mathbb{E}\left[\mathbb{1}_A\operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^1(t_0)}\mathbb{E}^1\left[W_{\hat{\tau}}^1\middle|\mathcal{F}^1(t_0)\right]\right]. \qquad \textbf{(B.5)}$$

This extends via a monotone-class- type argument, cf. Chung [11], to *all* $A \in \mathcal{H}_\pi^1(t_0)$. Hence, for all $\tau \in \hat{\mathbb{H}}_\pi^1(t_0)$,

$$\mathbb{E}\left[W_\tau^1\middle|\mathcal{H}_\pi^1(t_0)\right] \leqslant \operatorname*{ess\,sup}_{\hat{\tau}\in\hat{\mathbb{F}}^1(t_0)}\mathbb{E}^1\left[W_{\hat{\tau}}^1\middle|\mathcal{F}^1(t_0)\right] (\mathbb{P}-\text{a.e.}). \qquad \textbf{(B.6)}$$

This establishes the result.

https://doi.org/10.1017/S0269964814000217 Published online by Cambridge University Press