Applied Probability Trust (5 August 2014)

ON THE ASYMPTOTIC OPTIMALITY OF GREEDY INDEX HEURISTICS FOR MULTI-ACTION RESTLESS BANDITS

D. J. HODGE,* *The University of Nottingham*

K. D. GLAZEBROOK,** *Lancaster University*

Abstract

The class of restless bandits as proposed by Whittle [21] have long been known to be intractable. This paper presents an optimality result which extends that of Weber and Weiss [19] for restless bandits to a more general setting in which individual bandits have multiple levels of activation but are subject to an overall resource constraint. The contribution is motivated by recent works of Glazebrook et al. [10, 11] who discuss the performance of index heuristics for resource allocation in such systems. Hitherto, index heuristics have been shown, under a condition of *full indexability*, to be optimal for a natural Lagrangian relaxation of such problems in which resource is purchased rather than constrained. We find that under key assumptions about the nature of solutions to a deterministic differential equation that the index heuristics above are asymptotically optimal in a sense described by Whittle. We then demonstrate that these assumptions always hold for three-state bandits.

Keywords: index heuristic; asymptotic optimality; multi-action restless bandit; stochastic resource allocation

2010 Mathematics Subject Classification: Primary 90C40

Secondary 49L20; 49M20; 93E20

1. Introduction

In what is now regarded as a classical result, Gittins [4] demonstrated the optimality of index policies for a class of reward discounted Markov decision processes called *Multi-*

* Postal address: School of Mathematical Sciences, University Park, Nottingham, NG7 2RD, UK

** Postal address: Lancaster University Management School, Bailrigg, Lancaster, LA1 4YX, UK

Armed Bandits (MABs). MABs are a class of simple models for dynamic resource allocation in which, at each decision epoch $t \in \mathbb{N}$, a choice is made of one from n stochastic projects (or *bandits),* for activation. The decision-maker is aided by the fact that the current *state* of each bandit is always observable. Gittins' solution has the following form: with each bandit is associated an *index,* namely a real-valued function on that bandit's state space. At each epoch it is optimal to activate any bandit whose current index value is maximal. These so-called *Gittins' indices* offer a simple and interpretable calibration of the value of project activation and Gittins' result offers a significant easing of the computational burden involved in developing an optimal policy. A sizeable literature now exists which is devoted to applications and extensions of Gittins' work. The recent book by Gittins et al. [5] gives an introduction to, and an overview of, the area.

It is a feature of Gittins' MABs that bandits which are not activated are frozen, that is their state does not change. This feature delimits the range of application of the model and associated results. In order to address this, Whittle [21] introduced a class of so-called *restless bandits* (RBs) in which the constituent projects have different stochastic dynamics depending on whether they are active or passive. This generalisation is bought at some cost, since unlike MABs, Whittle's RBs are almost certainly intractable, having been shown to be PSPACE-hard by Papadimitriou and Tsitsiklis [15]. Whittle proposed an *index heuristic* for RBs, with Whittle's index emerging from a Lagrangian relaxation of the optimisation problem as a *fair charge* for the activation of a given bandit in a given state. Whittle's indices reduce to those of Gittins in the special MAB case. Whittle's index policy has been shown empirically to perform well in a range of application domains including asset management [3, 11, 14, 17], inventory routing [1], machine maintenance [6] and queueing control [2, 8, 9]. However, the question naturally arises as to what can be said theoretically and in generality about its quality of performance.

This challenge was taken up by Weber and Weiss [19, 20] who explored a conjecture due to Whittle [21]. The setting for these ideas is that a RB features n bandits, m of which may be activated at any time. We consider *average reward per unit time* over an infinite horizon as the criterion by which policies are evaluated. Whittle conjectured that his index heuristic was asymptotically optimal in a limit in which $n, m \to \infty$

in fixed proportion. Weber and Weiss demonstrated that this was indeed the case under a condition in which a particular differential equation's stationary point is a *global attractor*. They further argued that even when that is not the case the degree of suboptimality of the index policy is often very small.

More recently, Glazebrook et al. [10] have discussed a class of models for dynamic resource allocation which extend Whittle's RBs away from the simple *active/passive* dichotomy for bandit treatment to one in which the key resource may be applied at a range of levels to each. In these new scenarios the resource can be concentrated on a few projects which are in urgent need of it or can be spread much more widely. The indices which emerge from this set-up, which is the natural one for the allocation of a single *divisible resource*, are functions not only of bandit state (i) but also of resource level (a). Index $I_i(a)$ can now be viewed as a *fair charge* for raising the resource level in state i to level a (from $a-1$). In Glazebrook et al. [7] the authors present some of the first work on indexability for dual-speed bandits with bounds on suboptimality of index heuristics. The first reference to indexability with multiple action levels comes from [18] in Weber's comments on the paper of Niño Mora [13]. In his comments the author conjectures that the index heuristics used in this paper might have asymptotic properties similar to those in Weber and Weiss [19]. Glazebrook et al. [10] develop this index heuristic and demonstrate empirically its strong performance in the context of models for queueing control and asset management. The goal of our paper is to establish asymptotic optimality of the index heuristic and so extend the result of Weber and Weiss to this more general model class.

In Section 2 we describe the model class of interest in more detail and give a brief account of the key notions of indices and indexability. The statement and proofs of our principal results are contained in Section 3. Section 4 then extends the result of [20] establishing asymptotic optimality for three-state indexable bandits to our new framework. A typical illustrative example is then presented. Section 5 contains some concluding remarks.

2. The model and key ideas

We shall consider a set up in which n finite state continuous time stochastic projects (or bandits) are in competition for a key divisible resource. We shall assume without loss of generality that each bandit has state space $\{1, 2, \ldots, k\}$, that actions $\{0, 1, \ldots, A\}$ are available in each state and that the stochastic dynamics for bandits are Markovian and identical, with $\lambda_{ij}(a)$ the exponential rate that a bandit in state i transitions to j, per unit time, under the application of action a. Given any choice of actions, bandits evolve independently. A decision regarding which action should be applied to each bandit will be made at time zero and at all state transitions of the process. Action a is to be understood here to be the application of a units of resource. In what follows we shall discuss a range of approaches to the way in which resource availability constrains the choice of actions. A reward rate $g(i, a)$ is obtained per unit of time that a bandit spends in state i under action a. Our goal will be the selection of a stationary randomized (see §8.9, [16]) policy σ to maximise the average long-run expected reward rate aggregated across bandits, namely

$$r_\sigma := \lim_{t\to\infty} \frac{1}{t} \int_0^t \sum_{l=1}^{n} \mathbb{E}_\sigma \left[g\left(x_l(s), A_\sigma(l, s)\right) \right] \mathrm{d}s,$$

with $x_l(s)$ the state of bandit l at time s and $A_\sigma(l, s)$ the random action applied to bandit l at time s, under policy σ.

To simplify ergodicity concerns we shall assume here that, under all policies, the underlying Markovian process is irreducible. This is easily achieved by, if necessary, peturbing transition rates to achieve $\lambda_{i,1+(i \bmod k)}(0) \geq \epsilon$ for all i. This could be weakened to demand that the process is irreducible under the optimal policy proposed later; state removal to consider only states visited under an optimal policy would allow a further weakening.

We now describe two problems of interest corresponding to two different approaches to constraining resource availability. Suppose that $0 < \beta < A$. We define

$P_h^n(\beta)$: This is the $n-$bandit reward maximisation problem where the total resource allocated to the bandits at each $t \in \mathbb{N}$ is fixed at $n\beta$, which here needs to be a positive integer. This latter requirement will be relaxed later;

$P_r^n(\beta)$: This is the $n-$bandit reward maximisation problem where the *time average*

amount of resource allocated to the bandits over an infinite horizon is fixed at $n\beta$, which here need not be a positive integer.

In keeping with the usage in [19] and [20] we shall use $R_{opt}^{(n)}(\beta)$ for the optimal value of $P_h^n(\beta)$ and $r^{(n)}(\beta)$ for the optimal value of $P_r^n(\beta)$. Subscripts h and r should be understood as 'hard' and 'relaxed' and hence descriptors of the resource constraint in each case, with $P_r^n(\beta)$ the more amenable to solution. With reference to Puterman (Theorem 8.9.6 and Corollary 8.9.7 of [16]), it is clear that for the relaxed problem simultaneously performing a, possibly randomized, stationary optimal relaxed β-constraint strategy at each bandit achieves the optimum for the n bandit problem. One way to see this is to take a solution to the n bandit problem, randomize the bandit labels, then we have an optimal policy which shares expected resource levels equally between the bandits. This policy may be history dependent, in that it may depend on the bandit labellings, however Theorem 8.9.6 ([16]) allows us to transform this into an equally good randomized stationary policy. The performance of this policy at any chosen bandit is bounded above by an optimal policy for that specific bandit with this average resource constraint β. The performance of using these optimal policies acheiving $r^{(1)}(\beta)$ at each bandit is therefore at least as good (and therefore equal to) an optimal policy for $r^{(n)}(\beta)$. In the case $n = 1$, for simplicity we shall denote $r^{(1)}(\beta)$ by $r(\beta)$.

Gittins' MABs have $A = 1$, $n\beta = 1$ and $\lambda_{ij}(0) = 0\ \forall i \neq j$ and his index policies solve $P_h^n(\beta)$ for such cases. Whittle's RBs have $A = 1$ and hence the resource constraint is on the *number* of bandits to be declared active at each time step. He developed an index heuristic for $P_h^n(\beta)$ for *indexable* problems of this type in which the solution to the corresponding $P_r^n(\beta)$ is of index form. The problem with $A > 1$ is considerably more complex, not least in the combinatorial complexity of the partitioning of the available resource to the competing bandits.

In the previous paragraph we observed that $r^{(n)}(\beta) = nr(\beta)$ and hence that a policy solving $P_r^n(\beta)$ is obtained by applying a policy solving $P_r^1(\beta)$ to each bandit in parallel. To develop a solution to the single bandit problem, we first observe that $r(\beta)$ is concave in β. We can see this by considering optimal policies at distinct $\beta_1 < \beta_2$, and constructing the natural, feasible randomized policy at any $\beta = \theta\beta_1 + (1-\theta)\beta_2$

from them. Feasibility of such a randomization establishes that

$$r(\theta\beta_1 + (1-\theta)\beta_2) \geq \theta r(\beta_1) + (1-\theta)r(\beta_2). \tag{1}$$

To develop a the solution to $P_r^1(\beta)$ we pose the *W-charge* problem for a single bandit as follows: abandon any constraint on the resource available to the single bandit and hence allow the decision-maker to choose any action from $\{0, 1, \ldots, A\}$ at each time $t \in \mathbb{N}$. A resource charge of Wa per unit time is made whenever action a is chosen and hence the net reward rate from the application of action a in state i is $g(i,a) - Wa$. We write the average long-run reward rate achieved under policy σ as r_σ^W and r^W for the maximal reward rate. It is easy to show that r^W is decreasing in W by a direct policy comparison contradiction argument. Convexity of r^W follows since r^W is a piece-wise linear hull formed by maximizing (in fact finitely many) affine functions of W with negative gradients. This together with the concavity of $r(\beta)$ allows us to deduce that

$$r(\beta) = \inf_W \left(r^W + W\beta\right), \tag{2}$$

writing $W(\beta)$, say, for any value achieving the infimum in (2). Hence $P_r^1(\beta)$ is solved by a policy which solves the $W(\beta)$-charge problem.

We require that such policies be of *index form*, equivalently, that bandits be *indexable* as follows:

Definition 1. A bandit is indexable if there exists a family $\left\{\sigma^W, -\infty < W < \infty\right\}$ of stationary policies $\sigma^W : \{1, 2, \ldots, k\} \to \{0, 1, \ldots, A\}$ such that $(a)\,\sigma^W$ is optimal for the W-charge problem, and $(b)\,\forall i \in \{1, 2, \ldots, k\}, \sigma^W(i)$ as a function of W is decreasing and onto $\{0, 1, \ldots, A\}$. For an indexable bandit and a choice of $(i, a) \in \{1, 2, \ldots, k\} \times \{1, 2, \ldots, A\}$, the index $I_i(a)$ is given by

$$I_i(a) = \inf_W \left\{W; \sigma^W(i) = a - 1\right\},$$

where we set $I_i(0) = \infty, I_i(A+1) = -\infty$.

When $W = I_i(a)$, the policy σ^W will be indifferent between actions $a-1$ and a in state i. In words, $I_i(a)$ is the *fair charge* for raising the resource level allocated to the bandit from $a-1$ to a when in state i. The solution to the $W-$charge problem is straightforward for indexable bandits: in the system state $\mathbf{i} = (i_l, 1 \leq l \leq n)$ in which

bandit l is in state i_l, allocate resource level a_l to bandit l, where

$$I_{i_l}(a_l+1) \leq W < I_{i_l}(a_l), 1 \leq l \leq n. \tag{3}$$

From (2) it follows that a solution to the relaxed problem $P_r^n(\beta)$ for indexable bandits can be found with an appropriate choice of $W = W(\beta)$.

In words, we accumulate resource at bandit l until the fair charge for allocating further resource falls below resource charge $W(\beta)$. For $W(\beta)$ chosen equal to an index value many solutions exist (through action randomization) to the relaxed problem. However, when later adapting a relaxed policy solution to the hard constraint problem the actual resource level used will be critical. In this scenario we will achieve a given instantaneous resource usage level by choosing $W(\beta)$ equal to some $I_i(a)$, and randomizing between actions a and $a-1$ on bandits in state i.

We develop a *greedy index heuristic* for $P_h^n(\beta)$ when bandits are indexable as follows: in system state $\mathbf{i} = (i_l, 1 \leq l \leq n)$ apply action $\sigma^W(i_l)$ to bandit l, $1 \leq l \leq n$, where W is any value satisfying

$$\sum_{l=1}^{n} \sigma^W(i_l) = n\beta.$$

From (3) we note that at non-index values of W the solution to $P_r^n(\beta)$ is constant in W thus solutions to the above are unlikely. Non-uniqueness of optimal policy at index values, however, allows us to choose W equal to an appropriate index value and if $n\beta \in \mathbb{Z}$ to treat some bandits in the same state differently (as proposed by two or more jointly optimal policies) to achieve the desired resource usage through a deterministic policy.

It is helpful to extend the definition of this greedy heuristic to suitably defined problems $P_h^n(\beta)$ with $n\beta \notin \mathbb{N}$ by use of randomisation. If $n\beta \notin \mathbb{N}$ then in system state $\mathbf{i}$ we find the unique $W(\mathbf{i}) \in \{I_j(a); (j,a) \in \{1,2,\ldots,k\} \times \{1,2,\ldots,A\}\}$ for which

$$\sum_{l=1}^{n} \sigma^{W(\mathbf{i})^+}(i_l) < n\beta, \quad \sum_{l=1}^{n} \sigma^{W(\mathbf{i})^-}(i_l) > n\beta.$$

Suppose that $W(\mathbf{i}) = I_{i_m}(a)$ and hence that $\sigma^{W(\mathbf{i})^+}(i_m) = a-1$. The greedy heuristic will then, in state $\mathbf{i}$, randomize between the actions $\left\{\sigma^{W(\mathbf{i})^+}(i_l), 1 \leq l \leq n\right\}$ and another set of actions, identical save only that the resource allocated to bandit m

is increased by one. The randomization will be designed to ensure that the expected resource level chosen is exactly $n\beta$.

We now write $R_{ind}^{(n)}(\beta)$ for the long run average reward rate achieved when the greedy index heuristic just defined is applied to $P_h^n(\beta)$. It follows trivially from the definitions of the quantities concerned that

$$R_{ind}^{(n)}(\beta) \leq R_{opt}^{(n)}(\beta) \leq nr(\beta). \tag{4}$$

When $A = 1$, namely in the restless bandit case, Weber and Weiss were able to show that when bandits are indexable we must have

$$\lim_{n\to\infty} \frac{R_{opt}^{(n)}(\beta)}{n} = r(\beta), \tag{5}$$

and, moreover, that if a particular differential equation's stationary point is a *global attractor* that

$$\lim_{n\to\infty} \frac{R_{ind}^{(n)}(\beta)}{n} = r(\beta). \tag{6}$$

In Section 3 we extend these results to accommodate $A \geq 1$ as is demanded by our general model for the dynamic allocation of a divisible resource. Of these two results, namely those extending (5) and (6) respectively, it is the latter which is more challenging and it is certainly the case that more serious consideration has to be given to understanding the evolution of the n bandits in a scenario in which $A > 1$. Further, with the counterexample provided in [19] to universal asymptotic optimality of a greedy index policy recoverable as a special case (with $A = 1$) we are guaranteed to need conditions at least as strong as are demonstrated necessary there. Our condition for the general model here involves a considerably more complicated differential equation, but one which maintains a similar form which allows for a proof of the asymptotic optimality of the greedy index policy via a convexity argument.

Plainly, courtesy of (4), the result in (5) is implied by that in (6) while the latter has the global attractor condition. For this reason, the verification of the former result provides the important insight that the only reason for possible asymptotic strict inequality of the three quantites in (4) arises from the first of the two inequalities. This also suggests that our relaxation approach is promising for future research since we need only focus on closing the suboptimality gap between policy performance and the value of the easily solved relaxed form of the problem.

We now proceed to an account of our main theoretical results.

3. Results

The following result states that the time-average version, $P_r^n(\beta)$, of the problem $P_h^n(\beta)$ yields an asymptotically tight relaxation.

Theorem 1.

$$\frac{R_{opt}^{(n)}(\beta)}{n} \longrightarrow r(\beta), \qquad as\ n \to \infty. \tag{7}$$

Proof. We start by writing $\boldsymbol{\pi}$ for the equilibrium distribution of a single bandit arising under use of the optimal policy for $P_r^n(\beta)$, i.e. under σ_{rel}. Note that by symmetry this is the same for all bandits. The DP equation for $P_h^n(\beta)$ can be written

$$b(\boldsymbol{x}) + \frac{R_{opt}^{(n)}(\beta)}{n\Lambda} = \max_{\boldsymbol{a}\in\hat{\boldsymbol{A}}} \left\{ \frac{1}{n\Lambda} \sum_{i=1}^{n} g(x_i, a_i) + \mathbb{E}\left[b(\hat{\boldsymbol{x}}) \,|\, \boldsymbol{x}, \boldsymbol{a}\right] \right\}, \tag{8}$$

where Λ is a uniformization parameter, $b(\boldsymbol{x})$ is the usual bias function, $\hat{\boldsymbol{x}}$ denotes the system state after one (uniformized) timestep under action $\boldsymbol{a}$ and $\hat{\boldsymbol{A}}$ is the set of feasible actions for the hard constraint problem $P_h^n(\beta)$. By rescaling time we can assume $\Lambda = 1$, without loss of generality.

We begin by assuming that all uniformized transition probabilities are rational (which implies that all equilibrium probabilities are rational) and choose an appropriate large n so that $n\pi_i \in \mathbb{N}$ for all i. We then consider starting the n-bandit system from a state, $\boldsymbol{x}$, mirroring $\boldsymbol{\pi}$, by having $n\pi_i$ of the n bandits starting in state i, for each i. Then we can write $n_i(\boldsymbol{x})$ for the integer-valued number of projects in state i where $n_i(\boldsymbol{x}) = n\pi_i$. We now consider the action of the relaxed optimal policy, σ_{rel}, which will use a total activation resource of exactly $n\beta$ units in state $\boldsymbol{x}$, since we are in a scaled up version of the single bandit equilibrium.

We now consider the (suboptimal) policy for $P_h^n(\beta)$ defined by performing the initially non-dynamic action specified by the relaxed policy σ_{rel} in our starting state *for a fixed time* δ before switching to an optimal policy for $P_h^n(\beta)$ thereafter. Suppose now that the action defined by this policy is $\boldsymbol{a}^* = \{a_1^*, a_2^*, \ldots, a_n^*\}$ with each $a_i^* \in [0, \ldots, A]$. Then the expected number of state changes (including null events as a result of uniformization) over this time period of length δ is $n\delta$. Over the same time the expected

reward obtained will be at least $\delta n r(\beta) - n\delta^2 G$, where $G = 2\max|g(i,a)|$, since only rewards incurred after a transition could be being incurred suboptimally (not at rate $r(\beta)$) and for each such state the loss in reward is bounded by twice the maximal reward. Since our δ-policy is suboptimal we can bound its performance above by

$$\delta R_{opt}^{(n)}(\beta) + b(\boldsymbol{x}) \geq \delta r(\beta) n - n\delta^2 G + \mathbb{E}_{\boldsymbol{a}^*} b(\boldsymbol{X}^\delta), \tag{9}$$

where $\boldsymbol{X}^\delta$ is the random system state after time δ. We desire to show that $|b(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{a}^*} b(\boldsymbol{X}^\delta)|$ is $o(\delta)$ and $o(n)$.

To do this we introduce the distance $d(\cdot,\cdot)$ defined by $d(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\sum|n_i(\boldsymbol{x}) - n_i(\boldsymbol{y})|$ – the minimal number of distinct state components between $\boldsymbol{x}$ and $\boldsymbol{y}$ after any permitted reordering of $\boldsymbol{y}$. We can write the number of state i bandits at time δ as $n_i(\boldsymbol{X}^\delta) = Y_1 + Y_2 + \ldots + Y_k$ where Y_j is the number of time-δ bandits which begin in state j at time 0 and end in state i at time δ. Writing $P_{ji}(\boldsymbol{a}^*, \delta)$ for the probability that a bandit in state j at time 0 is in state i at time δ under constant action level $\boldsymbol{a}^*$ then $Y_j \sim \mathrm{Bin}(n_j(\boldsymbol{x}), P_{ji}(\boldsymbol{a}^*, \delta))$.

Over the time interval of length δ only 'no change', or 'a single state change' are the events with probabilities not $o(\delta)$ — in particular, the probabilities are $e^{-\delta} = 1-\delta+o(\delta)$ and $\delta e^{-\delta} = \delta + o(\delta)$ respectively. If we introduce the notation $p_{ji}(a)$ for the jump-chain probability under action a, conditioned on there being one 'event' in $[0,\delta]$ (note that $p_{jj}(a)$ may be strictly positive), then we can also find expressions for the expectation and variance of $n_i(\boldsymbol{X}^\delta)$ by repeated use of the identity

$$P_{ji}(\boldsymbol{a}^*, \delta) = \mathbb{I}(j=i)(1-\delta) + \delta p_{ji}(a_j^*) + o(\delta). \tag{10}$$

We have

$$\begin{aligned}\mathbb{E}(n_i(\boldsymbol{X}^\delta)) &= n_i(\boldsymbol{x}) P_{ii}(\boldsymbol{a}^*, \delta) + \sum_{j\neq i} n_j(\boldsymbol{x}) P_{ji}(\boldsymbol{a}^*, \delta) \\ &= n_i(\boldsymbol{x})(1-\delta) + \delta n_i(\boldsymbol{x}) p_{ii}(a_i^*) + \sum_{j\neq i} n_j(\boldsymbol{x})\delta p_{ji}(a_j^*) + no(\delta).\end{aligned}$$

But $\boldsymbol{x}$ is an equilibrium state satisfying the detailed balance equations and so

$$\sum_{j\neq i} n_j(\boldsymbol{x}) p_{ji}(a_j^*) = n_i(\boldsymbol{x})(1 - p_{ii}(a_i^*)), \tag{11}$$

from which it easily follows that

$$\mathbb{E}(n_i(\boldsymbol{X}^\delta)) = n_i(\boldsymbol{x}) + no(\delta).$$

Similarly, simplification allows us to write,

$$\mathrm{Var}(n_i(\boldsymbol{X}^\delta)) = \sum_i \mathrm{Var}(Y_i) = 2\delta n_i(\boldsymbol{x})\left[1 - p_{ii}(a_i^*)\right] + no(\delta),$$

with use of (11). These allow us to deduce

$$\begin{aligned}
\mathbb{E}(d(\boldsymbol{x}, \boldsymbol{X}^\delta)) &= \frac{1}{2}\sum_i \mathbb{E}|n_i(\boldsymbol{x}) - n_i(\boldsymbol{X}^\delta)|, \\
&\leq \frac{1}{2}\sum_i \sqrt{\mathrm{Var}(n_i(\boldsymbol{X}^\delta))} + no(\delta) = \frac{1}{2}\sum_i \sqrt{2\delta n_i(\boldsymbol{x})\left[1 - p_{ii}(a_i^*)\right]} + no(\delta), \\
&\leq B\sqrt{\delta n} + no(\delta), \qquad (12)
\end{aligned}$$

for some constant B. We have therefore obtained a bound on $\mathbb{E}(d(\boldsymbol{x}, \boldsymbol{X}^\delta))$ which is of order $\sqrt{n}$.

Consider now $\boldsymbol{x}, \boldsymbol{y}$ with $d(\boldsymbol{x}, \boldsymbol{y}) = 1$ and suppose $x_l \neq y_l$. Further introduce the notation σ_{opt}^h for an optimal policy for $P_h^n(\beta)$. Now apply σ_{opt}^h, starting in state $\boldsymbol{x}$, and couple the actions with acting on state $\boldsymbol{y}$. After the first transition (real or virtual) of bandit l, the interpretation of the bias function as the difference in long-term rewards between different starting states yields

$$b(\boldsymbol{x}) - b(\boldsymbol{y}) \geq -G + \mathbb{E}_{\sigma_{opt}^h}\left[b(\hat{\boldsymbol{x}}) - b(\hat{\boldsymbol{y}})\right], \qquad (13)$$

where $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are the states after the first transition. Since we coupled the processes we know that $d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \leq 1$ a.s. We would like that $\mathbb{P}(\hat{x}_l = \hat{y}_l) > 0$ but with one transition this may not be true. Assumed irreducibility ensures that $\exists\ r \in \mathbb{Z}^+$ such that after r transitions there is a path of positive probability from x_l to y_l. Thus if we denote by $\hat{\boldsymbol{x}}^{(r)}$ the system state after r iterations (starting from state $\boldsymbol{x}$) of potential state l changes then we can find a constant $0 < \omega = \mathbb{P}(\hat{x}_l^{(r)} = \hat{y}_l^{(r)})$. Indeed uniformization, and therefore null transitions, ensure that for r we can choose $\omega = \min_{\boldsymbol{x},\boldsymbol{y}:d(\boldsymbol{x},\boldsymbol{y})=1} \mathbb{P}(\hat{x}_l^{(r)} = \hat{y}_l^{(r)})$. Also, the r-step version of (13) is

$$b(\boldsymbol{x}) - b(\boldsymbol{y}) \geq -rG + \mathbb{E}_{\sigma_{opt}^h}\left[b(\hat{\boldsymbol{x}}^{(r)}) - b(\hat{\boldsymbol{y}}^{(r)})\right].$$

By considering the pair $(\boldsymbol{x}, \boldsymbol{y})$ which minimise the LHS of this equation, we deduce that

$$\min_{\boldsymbol{x},\boldsymbol{y}:d(\boldsymbol{x},\boldsymbol{y})=1} \Big\{ b(\boldsymbol{x}) - b(\boldsymbol{y}) \Big\} \geq -rG/\omega.$$

To extend to more general $(\boldsymbol{x}, \boldsymbol{y})$ with $d(\boldsymbol{x}, \boldsymbol{y}) = d > 1$ we merely construct a sequence of vectors $\boldsymbol{x} = \boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{d-1}, \boldsymbol{v}_d = \boldsymbol{y}$ with $d(\boldsymbol{v}_i, \boldsymbol{v}_{i+1}) = 1$ for all i. Then expand $b(\boldsymbol{x}) - b(\boldsymbol{y})$ as $b(\boldsymbol{v}_0) - b(\boldsymbol{v}_1) + b(\boldsymbol{v}_1) - b(\boldsymbol{v}_2) + \ldots - b(\boldsymbol{v}_d)$ to yield

$$b(\boldsymbol{x}) - b(\boldsymbol{y}) \geq -rGd(\boldsymbol{x}, \boldsymbol{y})/\omega. \tag{14}$$

Combining (9), (12) and (14), and dividing through by $n\delta$ we deduce that

$$\begin{aligned} \frac{R^{(n)}_{opt}(\beta)}{n} \geq r(\beta) - \delta G + \frac{\mathbb{E}_{\boldsymbol{a}^*}\Big[b(\boldsymbol{X}^\delta) - b(\boldsymbol{x})\Big]}{n\delta} &\geq r(\beta) - \delta G - \frac{rG}{\omega}\frac{1}{n\delta}\mathbb{E}_{\boldsymbol{a}^*}\Big[d(\boldsymbol{X}^\delta, \boldsymbol{x})\Big] \\ &= r(\beta) - \delta G - \frac{rG}{\omega}\left(\frac{o(\delta)}{\delta} + \frac{B}{\sqrt{n\delta}}\right). \end{aligned}$$

Thus choosing small δ and then letting $n \to \infty$ (through appropriate values) we obtain $\lim \frac{R^{(n)}_{opt}(\beta)}{n} \geq r(\beta)$, from which we deduce equality since we already know that $R^{(n)}_{opt}(\beta) \leq nr(\beta)$.

Remark 1: One real difference of note from the account of [19] is the correction that two states $\boldsymbol{x}$ and $\boldsymbol{y}$ with $d(\boldsymbol{x}, \boldsymbol{y}) = 1$ need not have a positive probability of single-step intercommunication. This cannot be remedied by a perturbation of the 1-step transition matrix (replacing all zero entries by ε) since then ω^{-1} may be unbounded.

Remark 2: We thus know that the solutions to $P^n_h(\beta)$ and $P^n_r(\beta)$ agree, on a 'per bandit' average reward scale. The important outstanding question, and the harder question, asks whether implementing a greedy index based heuristic, when the bandits are indexable, achieves the same asymptotic performance of average reward per bandit per unit time.

Notationally, since we shall treat all bandits of identical state identically we change our state variable from $\boldsymbol{x} \in \{1, \ldots, k\}^n$ to $\boldsymbol{z}^{(n)} \in [0, 1]^k$ where $z^{(n)}_i$ is the proportion of the n bandits which are in state i. The action space can appropriately be rewritten too in terms of actions taken as functions of the bandit state i, not bandit number l.

For a quick example, suppose that $\beta = 0.45$, $k = 4$, and $A = 2$ with $I_4(1) > I_4(2) > I_3(1) > I_2(1) > I_3(2) > I_1(1) > I_2(2) > I_1(2)$. Suppose the system is in

state $\boldsymbol{z}^{(n)} = \{0.3, 0.4, 0.2, 0.1\}$, then the basic action under the greedy index policy is $\{0, 0, 1, 2\}$ since $0.1 + 0.1 + 0.2 < 0.45 < 0.1 + 0.1 + 0.2 + 0.4$ and a randomly chosen bandit in state 2 would randomize between action 1 and action 0. The exact nature of the extra action would depend upon n, but would involve some bandits in state 2 taking action $a = 0$ and some taking action $a = 1$ and possibly one randomizing between the two actions.

We now introduce a result of Mitra & Weiss [12], which we shall use later in the exposition, but appears now to contextualize the approach. This is a result bounding the time-average deviation from an equilibrium distribution of a family of continuous-time Markov chains. The family of processes is that which arises by speeding up time and scaling down jump sizes accordingly (speed up time by factor n – the number of bandits – and scale down jump-sizes to $1/n$) from a single chain. We consider the evolution of the processes $\left(\boldsymbol{z}^{(n)}\right)$ described above as $n \to \infty$ under application of the greedy index policy, using this result. For fixed n, and each system state $\boldsymbol{z}^{(n)}$, we introduce the notation $\Psi_{ij}(\boldsymbol{z}^{(n)})$ for the individual rate at which one of the n bandits transitions from $i \mapsto j$ under the actions taken by greedy index policy σ_{ind} in state $\boldsymbol{z}^{(n)}$. We further write $\boldsymbol{\Psi}(\boldsymbol{z}^{(n)})$ for the matrix of values $\left\{\Psi_{ij}(\boldsymbol{z}^{(n)})\right\}$. We shall show that the fluid limit of the $\boldsymbol{z}^{(n)}$ processes exists and satisfies the equation $\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = \Psi(\boldsymbol{z})\boldsymbol{z}$, given explicitly in (18).

Theorem 2. (Mitra & Weiss, 1988.) *Suppose that there exists a probability distribution $\boldsymbol{\zeta}$ such that for every initial probability distribution $\boldsymbol{z}(0)$ the fluid approximation $\frac{\mathrm{d}z}{\mathrm{d}t} = \boldsymbol{\Psi}(\boldsymbol{z})\boldsymbol{z}$ has $\boldsymbol{z}(t) \to \boldsymbol{\zeta}$, and the transition rates $\Psi_{ij}(\boldsymbol{z})$ are bounded and Lipschitz-continuous. Then for every $\epsilon > 0$ there exist positive constants c_1 and c_2 such that for any initial state $\boldsymbol{z}(0)$*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P\left(\left\|\boldsymbol{z}^{(n)}(u) - \boldsymbol{\zeta}\right\|_2 > \epsilon \,\middle|\, \boldsymbol{z}(0)\right) du \leq c_1 \exp(-nc_2). \tag{15}$$

We will use this to show that, if all paths in the fluid limit process of the system evolving under the greedy index policy tend to a unique equilibrium then asymptotically the average reward per bandit under σ_{ind} approaches the average reward in the fluid limit problem. Finally we can then observe that the fluid limit problem has $r(\beta)$ as its limiting average reward.

Theorem 3. *Let $\boldsymbol{\pi}$ be the equilibrium distribution of a single bandit operated under the relaxed-constraint optimal policy σ_{rel}. Suppose that the σ_{ind} fluid limit differential equation* (18) *has a global attractor. Then if the n bandits are indexable,* (18) *has the unique fixed point $\boldsymbol{\pi}$ and $\boldsymbol{z}(t) \to \boldsymbol{\pi}$ for all $\boldsymbol{z}(0)$. Furthermore, we have $R^{(n)}_{ind}(\beta)/n \to r(\beta)$, as $n \to \infty$, $\beta \in (0, A)$.*

In contrast to Theorem 2 of [19] we have $\beta \in (0, A)$ as opposed to $(0, 1)$, but more importantly the natural greedy index policy is not just a decision of which bandits to activate.

Given an ordering on indices, such as the one above, we define a sequence of linear functions $f^a_i(\boldsymbol{z})$ as follows:

$$f^a_i(\boldsymbol{z}) = \sum_{j=1}^{k} c(i, a, j) z_j,$$

where

$$c(i, a, j) = \sum_{\hat{a}=1}^{A} \mathbb{I}\left[I_j(\hat{a}) > I_i(a)\right], \text{ and } \mathbb{I} \text{ is an indicator.}$$

In words, $c(i, a, j)$ is the number of state j indices which are greater than $I_i(a)$. Thus in the example above, $f^1_4(\boldsymbol{z}) = 0$, $f^2_4(\boldsymbol{z}) = z_4$, and $f^2_2(\boldsymbol{z}) = 2z_4 + 2z_3 + z_2 + z_1$. If we write $p^a_i(\boldsymbol{z})$ for the probability that a randomly chosen bandit in state i receives an activation level of at least a, in system state $\boldsymbol{z}$ then

$$p^a_i(\boldsymbol{z}) = \min\left\{z_i, \max\left[0, \beta - f^a_i(\boldsymbol{z})\right]\right\} \Big/ z_i, \quad \text{for } z_i > 0, \tag{16}$$

setting $p^a_i(\boldsymbol{z}) = 0$ when $z_i = 0$ for completeness. Note that $p^a_i(\boldsymbol{z})$ is decreasing in a. Since $f^{a+1}_i(\boldsymbol{z}) - f^a_i(\boldsymbol{z})$ is a non-zero polynomial in z_i with non-negative coefficients, if $z_i > 0$ then $p^a_i(\boldsymbol{z}) \in [0, 1) \Rightarrow p^{a+1}_i(\boldsymbol{z}) = 0$.

The expected activation level applied to a state i bandit under σ_{ind}, and therefore to all bandits in z_i is then given by

$$u_i(\boldsymbol{z}) := \sum_{a=1}^{A} p^a_i(\boldsymbol{z}).$$

For a vector $\boldsymbol{h} \in \mathbb{R}^A$ we now define

$$\phi_i(\boldsymbol{z}, \boldsymbol{h}) = \sum_{a=1}^{A} \left[p^a_i(\boldsymbol{z}) - p^{a+1}_i(\boldsymbol{z})\right] h_a, \tag{17}$$

then $\sum_i \phi_i(\boldsymbol{z}^{(n)}(t), \boldsymbol{g}(i,\cdot))$ is the instantaneous reward rate at time t, under the greedy index heuristic σ_{ind}.

To identify the fluid limit model, we note that transitions in $\boldsymbol{z}^{(n)}$ are all of the form $\boldsymbol{z}^{(n)} \mapsto \boldsymbol{z}^{(n)} + (1/n)\boldsymbol{e}_{ij}$ for some bandit states i and j, and where $\boldsymbol{e}_{ij}$ is a vector with -1 in the ith component and $+1$ in the jth component. If transitions occur at rate $\lambda_{ij}(a)$ for a bandit in state i under action a then the rate of $\boldsymbol{e}_{ij}$ transitions under σ_{ind} will be

$$nz_i^{(n)} \sum_a \left[p_i^a(\boldsymbol{z}) - p_i^{a+1}(\boldsymbol{z})\right] \lambda_{ij}(a) = nz_i^{(n)} \phi_i(\boldsymbol{z}^{(n)}, \boldsymbol{\lambda_{ij}}(\cdot)),$$

writing $\boldsymbol{\lambda_{ij}}(\cdot)$ for the vector $\{\lambda_{ij}(1), \lambda_{ij}(2), \ldots, \lambda_{ij}(A)\}$. Since the n-dependence in all rates is purely linear, the relationship between $\boldsymbol{z}^{(n)}$ and $\boldsymbol{z}^{(2n)}$, for example, is that the latter has jumps at twice the rate of the former, but jumps are half the size. The natural limiting process, of the sequence of $\boldsymbol{z}^{(n)}(t)$ is then determined by the differential equation

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = \sum_{i,j} z_i \phi_i(\boldsymbol{z}, \boldsymbol{\lambda}_{ij}(\cdot))\boldsymbol{e}_{ij}. \tag{18}$$

We note in passing, with a slight abuse of notation, that starting from $\boldsymbol{z} := \boldsymbol{\pi}$, $\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = 0$ is just the detailed balance equations of the *relaxed* policy on *a single* bandit since from $\boldsymbol{\pi}$ the relaxed policy uses an expected β resources per bandit at all times. So $\boldsymbol{\pi}$ is a stationary point of (18).

We shall see that the constraint that a policy uses $n\beta$ resource on average per unit of time uniquely identifies an index policy implementation. It will then follow that there can only exist one stationary point of (18) and hence that we have a unique equilibrium distribution for the optimal index policy for the relaxed problem. By being a zero of the right-hand side of (18) we necessarily have a solution to the detailed balance equations of the relaxed policy with $n\beta$ constraint, and thus an equilibrium distribution for the relaxed policy. Any such zero of (18), is therefore an optimal equilibrium of a W-charge problem, in particular a W for which $n\beta$ is the resulting resource usage per unit time. Recall that the gradient of the W-charge problem solution (in W) is the resource used, and as such is decreasing in β since $r(\beta)$ is concave in β and our randomizations induce resource usages strictly decreasing in β. The W-charge problem decomposes as n single bandits, each evolving under an optimal W-charge policy. Here we may use the assumption that under the optimal index policy the induced chain is irreducible to

deduce we have a unique such equilibrium.

Now we may use our assumption that the limit set is a global attractor to invoke Theorem 2, taking $\boldsymbol{\zeta} = \boldsymbol{\pi}$. We start by observing that the term in the integrand $\sum_{i=1}^{k} \phi_i(\boldsymbol{z}^{(n)}(s), \boldsymbol{g}(i,\cdot))$ below yields the instantaneous σ_{ind} rewards, which are piecewise-linear with only finitely many changes in gradient and thus clearly satisfy the continuity conditions of Theorem 2. Furthermore the entire integrand below is bounded, since rewards themselves are bounded. We now consider

$$\frac{R_{ind}^{(n)}(\beta)}{n} - r(\beta) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \sum_{i=1}^{k} \Big[\phi_i(\boldsymbol{z}^{(n)}(s), \boldsymbol{g}(i,\cdot)) - \phi_i(\boldsymbol{\pi}, \boldsymbol{g}(i,\cdot))\Big] \,\mathrm{d}s.$$

Now given $\eta > 0$, the continuity of ϕ means we can find an $\epsilon > 0$ such that

$$\sup_{\boldsymbol{z}:\|\boldsymbol{z}-\boldsymbol{\pi}\|<\epsilon} \sum_{i=1}^{k} |\phi_i(\boldsymbol{z}, \boldsymbol{g}) - \phi_i(\boldsymbol{\pi}, \boldsymbol{g})| < \eta/2,$$

and for this $\epsilon > 0$ Theorem 2 gives that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P\left(\left\|\boldsymbol{z}^{(n)}(u) - \boldsymbol{\pi}\right\|_2 > \epsilon \,\middle|\, \boldsymbol{z}(0)\right) \mathrm{d}u \le c_1 \exp(-nc_2), \tag{19}$$

for some positive constants c_1, c_2. Thus the time-averaged proportion of time that a $\boldsymbol{z}^{(n)}$ path spends outside a small region around $\boldsymbol{\pi}$ decays exponentially with n. Then we just choose n_0 large enough that $Gc_1 \exp(-n_0 c_2) < \eta/2$, and obtain

$$\left|\frac{R_{ind}^{(n)}(\beta)}{n} - r(\beta)\right| < \eta \quad \forall n > n_0. \tag{20}$$

We have thus shown $\lim_{n\to\infty} R_{ind}^{(n)}(\beta)/n = r(\beta)$. This concludes the proof of Theorem 3.

In conclusion, we have seen that, under the condition that the differential equation (18) has a single-point limit set, and assuming the bandits are indexable, the asymptotic performance of the greedy index policy agrees with both the original hard constraint problem and the relaxed problem:

$$\lim_{n\to\infty} \frac{R_{ind}^{(n)}(\beta)}{n} = \lim_{n\to\infty} \frac{R_{opt}^{(n)}(\beta)}{n} = r(\beta) \tag{21}$$

4. Optimality in three dimensions

Having established Theorem 3, we are now able to present a concrete set of problems for which asymptotic optimality of index policies can be demonstrated to hold. We shall

do this by way of a study of the resulting fluid limit differential equation, establishing that the unique stationary point is indeed a global attractor. The beginning of this section is thus an extension of [20] to our multi-action bandits.

We consider bandits evolving on just $k = 3$ states. We know from [19] that when $k = 4$, even for $A = 1$ there are examples where asymptotic optimality does not hold with index policies and indexable bandits, so there is no hope of proving universal optimality of index policies for $k > 3$. Although we restrict ourselves to $k = 3$ we will, however, permit any number of possible activity levels A.

What follows is therefore an extention of Weber and Weiss [20], in which the authors prove that for three-state, two-action bandits, indexability is a sufficient condition for asymptotic optimality of the greedy index heuristic. We do this by establishing the globally attractive nature of the unique solution to the fluid limit equation, and therefore are able to invoke Theorem 3. The irreducibility assumption is still required, however reducible three-state bandits are clearly uninteresting in this problem. We show for any value of A, and any resource constraint value $\beta \in (0, A)$, that in the case of indexable bandits the greedy index heuristic achieves the asymptotically optimal reward.

Lemma 1. *With n bandits in a normalized system state $\boldsymbol{z}^{(n)} = \{z_1^{(n)}, z_2^{(n)}, z_3^{(n)}\}$ the net rate of flow into (or out of) state 1 is an affine function of $z_1^{(n)}$, $z_2^{(n)}$, and $z_3^{(n)}$. Further, after substituting $z_2^{(n)} = 1 - z_1^{(n)} - z_3^{(n)}$, the coefficient of the $z_1^{(n)}$ term is non-positive.*

Proof. For a given state $\boldsymbol{z}^{(n)}$, the rate of flow out of state 1 is

$$z_1^{(n)}\phi_1(\boldsymbol{z}^{(n)}, \boldsymbol{\lambda_{12}}(\cdot)) + z_1^{(n)}\phi_1(\boldsymbol{z}^{(n)}, \boldsymbol{\lambda_{13}}(\cdot)), \tag{22}$$

and the rate of flow into state 1 is

$$z_2^{(n)}\phi_2(\boldsymbol{z}^{(n)}, \boldsymbol{\lambda_{21}}(\cdot)) + z_3^{(n)}\phi_3(\boldsymbol{z}^{(n)}, \boldsymbol{\lambda_{31}}(\cdot)). \tag{23}$$

The form of ϕ_i, given in (16) and (17), is such that for any vector $\boldsymbol{z}$, $z_i\phi_i(\boldsymbol{z}, \cdot)$ is affine in the components of $\boldsymbol{z}$. This establishes the first claim of the lemma.

For the second part, suppose that in state $\boldsymbol{z}$ the greedy index action is $\boldsymbol{a} =$

$\{a_1, a_2, a_3\}$ then net flow into state 1 is:

$$\begin{aligned}&\big(\lambda_{31}(a_3)z_3 + \lambda_{21}(a_2)z_2\big) - \big(\lambda_{12}(a_1)z_1 + \lambda_{13}(a_1)z_1\big)\\ &= \left(\lambda_{31}(a_3) - \lambda_{21}(a_2)\right) z_3 + \lambda_{21}(a_2) - \lambda_{21}(a_2)z_1 - \lambda_{12}(a_1)z_1 - \lambda_{13}(a_1)z_1.\end{aligned}$$

As all z_1 terms have non-positive coefficients, this proves the result.

The z_1 coefficient of flow into state 1 could be zero even when $z_1 \neq 0$, if $\lambda_{21}(a_2) = \lambda_{12}(a_1) = \lambda_{13}(a_1) = 0$. However, the z_1 coefficient out of state 1 and the z_3 coefficient out of state 3 cannot simultaneously be zero as then we would have zero rate of flow out of state 2 too (i.e. $\lambda_{21}(a_2) = \lambda_{23}(a_2) = 0$) and we would be in a state where all transition rates were zero, contradicting our irreducibility assumption.

Theorem 4. *Assume that we have n copies of an indexable bandit on three states ($k = 3$), with any fixed $\beta \in (0, A)$. Then the fluid limit approximation for $\boldsymbol{z}$, (18), has a globally attractive fixed point. Therefore the proposed greedy index policy is asymptotically optimal as $n \to \infty$.*

Proof. We shall consider the solutions to the fluid limit differential equation (18), noting that $\boldsymbol{z}(t) = \{z_1(t), z_2(t), z_3(t)\}$ with $z_1(t)+z_2(t)+z_3(t) = 1$, so we can eliminate $z_2(t)$ and regard $\boldsymbol{z}(t)$ as a two-dimensional vector. Having assumed an indexable bandit we can order the $3A$ indices, $\mathcal{I}_1 \geq \mathcal{I}_2 \geq \ldots \mathcal{I}_{3A}$, where each $\mathcal{I}_n = I_i(a)$ for some $i \in \{1, 2, 3\}$ and some $a \in \{1, 2, \ldots, A\}$. The form of ϕ (see (16) and (17)) is such that there are $3A$ regions of distinct affine forms taken by the derivatives of z_1 and z_3, determined by which $\mathcal{I}_n$ lie above or below the service charge W which describes a particular greedy action. The form of (18) is such that on each of these $3A$ regions we have a constant 2-vector $\boldsymbol{c}_m$ and constant 2x2-matrix $\boldsymbol{A}_m$, with $m = \{1, 2, \ldots, 3A\}$, satisfying

$$\begin{pmatrix}\dot{z_1}\\ \dot{z_3}\end{pmatrix} = \boldsymbol{c}_m + \boldsymbol{A}_m \begin{pmatrix}z_1\\ z_3\end{pmatrix}. \tag{24}$$

Now, by Lemma 1, $(A_m)_{11} \leq 0$, and by the same argument we also have $(A_m)_{22} \leq 0$. The comment after the lemma allows us to invoke Bendixson's criterion which says that if $\dot{\boldsymbol{z}} = (f_1(z_1, z_3), f_2(z_1, z_3))$ (for continuously differentiable f) and $\nabla \cdot \boldsymbol{f} < 0$ then there exist no periodic solutions. The observation that net flows on the boundary are always strictly towards the interior of the state-space guarantees that no trajectories

ever leave it. An application of the Poincaré-Bendixson theorem in the plane (which classifies solution trajectories) together with the fact that there are no limit cycles, excludes every case except that the already identified unique stationary point must be the limit of all trajectories. Thus the asymptotic performance under the greedy index policy is optimal, by Theorem 3.

4.1. Two worked examples

As a demonstration of these results we now present two examples of three-state irreducible and indexable bandits for which the greedy index heuristic is now known to be optimal. Consider a collection of n identical bandits on states $i \in \{1, 2, 3\}$. Each bandit has three levels of action, named $a \in \{0, 1, 2\}$. Thus if all bandits receive maximal actions we would use $2n$ units of resource per unit time. In these examples we shall use $\beta = \frac{5}{4}$. The only particularly observable impact of a choice of β is that together with the index ordering it determines which of the $3A$ potential regions of differing behaviour (identified in (24)) have full rank and which are empty. This setup models, for example, a scenario where bandit states can represent current tasks of three potential priority types: high, medium and low. Furthermore, resources can be prioritized to work at differing rates on the different bandits, perhaps to prioritize finishing high priority jobs earlier.

Example 1 For each bandit we need to describe two functions: the transition rates between all pairs of states as functions of actions, and the reward rates for each state and action. We shall use

$$\lambda_{ij}(a) = \alpha(a)v_j + \frac{i}{3} \tag{25}$$

where $\boldsymbol{v} \equiv \{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\}$ and $\{\alpha(0), \alpha(1), \alpha(2)\} \equiv \{1, 2, 5\}$. The $\alpha(\cdot)$ function respresents the rate at which those actions complete work at the bandits, then after service is complete the vector $\boldsymbol{v}$ respresents the differing probabilities with which the bandit transitions. The additional term on the end of (25) could be seen as a state dependent abandonment rate.

Secondly, we need to define the reward rates which we take to be $g(i, a) = g_i\alpha(a)$, where $\boldsymbol{g} \equiv \{1, 2, 5\}$.

Under these choices we find the following index ordering:

$$I_3(1) > I_2(1) > I_3(2) > I_1(1) > I_2(1) > I_1(2), \tag{26}$$

in particular we have index values given in Table 1. The table also gives the optimal

Index name	W-value	Policy σ^W
	$-\infty$	2,2,2
$I_1(2)$	0.965	1,2,2
$I_2(2)$	1.083	1,1,2
$I_1(1)$	1.857	0,1,2
$I_3(2)$	1.951	0,1,1
$I_2(1)$	2.159	0,0,1
$I_3(1)$	4.288	0,0,0
	∞	

TABLE 1: Index values and optimal policies for the W-charge problem.

policies σ^W for the W-charge problem for the ranges of W determined by successive index values in Table 1.

There are a number of ways to calculate the index values in Table 1. For an example as small as this it is feasible for the reader to enumerate the 27 stationary deterministic policies $\pi_1,\ldots,\pi_{27}$ for a single bandit and for a given W value find the optimal policy from

$$\max_{1\le j\le 27} \big[R(\pi_j) + WA(\pi_j)\big], \tag{27}$$

where $R(\pi)$ is the expected long-run reward under π and $A(\pi_j)$ is the long-run average action level used under π.

Under these circumstances we solve the fluid limit differential equation (in Octave/Matlab), over 4 distinct regions. The constraint $\beta = 5/4$ ensures there is always greedily allocated resource for the first unit in state 1 and state 2 bandits, but then there are two further potential regions of different behaviour of the fluid limit differential equation.

Keeping the same notation of (z_1, z_2, z_3) for the proportions of bandits in each of the three states, and reducing to two states by using $z_2 = 1 - z_1 - z_3$, the resulting

phase plane diagram of solutions in the (z_1, z_3) plane appears in Figure 1, with a single global attractor.

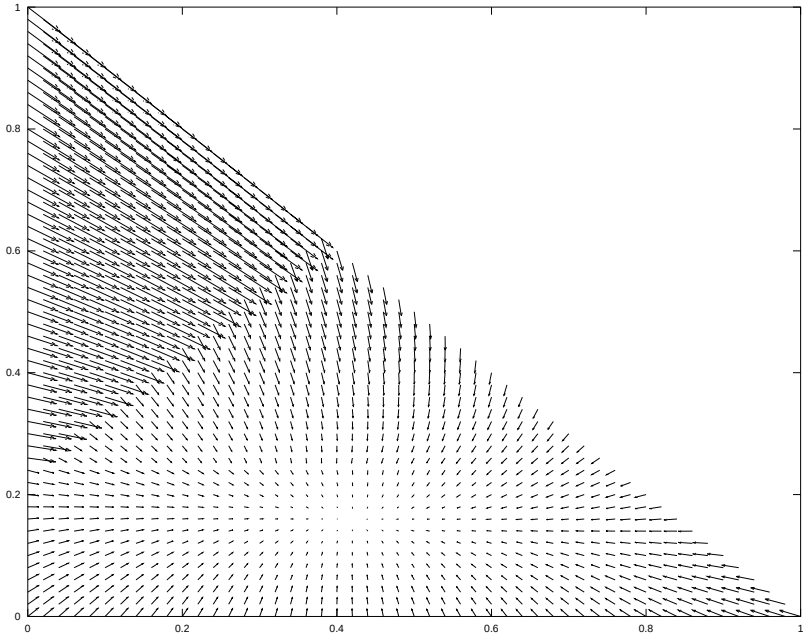

FIGURE 1: Example 1 – phase plane diagram of (z_1, z_3) with fixed point around (0.422,0.158)

Example 2 As a second example we modify the transition rates to make the abandonment rate multiplicative with the natural rate $\boldsymbol{v}$ associated with the states. We therefore use

$$\lambda_{ij}(a) = (\alpha(a) + i)\, v_j \tag{28}$$

instead, with the parameters as defined in Example 1. This modification leads to no change in the relative ordering of the six indices, though it does, of course, give rise to a new (but similar) phase plane diagram. Again by Theorem 4 the observed stationary point is a global attractor and the greedy index policy is asymptotically optimal.

5. Conclusion

We have taken the approaches set forth in [19], and extended them to a much wider range of problems. Specifically we show that in problems for which multiple activation levels are permitted at any bandit, *indexability* (in the new extended multi-action sense) is a critically useful property in finding asymptotically optimal policies. One question we have addressed is the stability of the unique solution to the proposed fluid limit differential equation. We know it is not always a global attractor, but we also know that finding counterexamples even in the single-action cases considered by Weber & Weiss is challenging. We have, however, extended their later result [20] to the multi-action case. So we do know that on three-state bandits indexability suffices to establish asymptotic optimality of the greedy index heuristic.

Of those works which have drawn on [19] since its publication as support for their approaches, a significant number make only a cursory reference because of the required simplification of their own problem to one with an '*active* or *passive*' action set. In expanding the scope of the result to a wider class of restless bandit problems we present a theoretical grounding for observed excellent performances of index-based heuristics in a large variety of problems observed in the literature. This is particularly the case in recent work of [10] on multi-action restless bandits where performance of greedy index heuristics has been seen to at times be astonishingly strong even for small numbers of bandits.

A final note on an extension to our main result, is that our assumption of identical bandits is not necessary. If our n (as $n \to \infty$) bandits are drawn from a *finite* (d, say) collection of different bandit types, then it is apparent that we could instead define $\boldsymbol{z}^{(n)}$ not as $(z_1, z_2, \ldots, z_k)$ but instead take a vector in $[0,1]^{kd}$ keeping track of proportions of the n bandits of each type in each state. Since the optimal solution to the relaxed problem, when the bandits are all indexable, is still of the same greedy index form and the form of our greedy index heuristic is unchanged, we obtain the same results where we effectively just have $\boldsymbol{z}^{(n)}$ defined on a state space of dimension kd instead of k and we have A indices identical for all bandits represented in each of these kd components. The only requirement we need make is that as $n \to \infty$ then the proportions of the n-bandits of each type remain constant, so that the $\boldsymbol{z}^{(n)}$ processes can indeed be seen

as sped up copies of each other with smaller jumps – all under our greedy index policy.

References

[1] T. Archibald, D. Black, and K. D. Glazebrook. Indexability and index heuristics for a simple class of inventory routing problems. *Operations Research*, 57(2):314–326, 2009.

[2] U. Ayesta, P. Jacko, and V. Novak. A nearly-optimal index rule for scheduling of users with abandonment. *Working paper from BCAM*, 2010.

[3] F. Caro and J. Gallien. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292, 2007.

[4] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[5] J. C. Gittins, K. D. Glazebrook, and R. R. Weber. *Multi-Armed Bandit Allocation Indices*. Wiley-Blackwell, London, 2011.

[6] K. Glazebrook, H. Mitchell, and P. Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research*, 165(1):267–284, 2005.

[7] K. D. Glazebrook, J. Niño Mora, and P. Ansell. Index policies for a class of discounted restless bandits. *Advances in Applied Probability*, 34(4):754–774, 2002.

[8] K. D. Glazebrook, P. S. Ansell, R. T. Dunn, and R. R. Lumley. On the optimal allocation of service to impatient tasks. *Journal of Applied Probability*, 41(1): 51–72, 2004.

[9] K. D. Glazebrook, C. Kirkbride, and J. Ouenniche. Index policies for the admission control and routing of impatient customers to heterogeneous service stations. *Operations Research*, 57(4):975–989, 2009.

[10] K. D. Glazebrook, D. J. Hodge, and C. Kirkbride. General notions of indexability for queueing control and asset management. *The Annals of Applied Probability*, 21(3):876–907, 2011.

[11] D. J. Hodge and K. D. Glazebrook. Dynamic resource allocation in a multi-product make-to-stock production system. *Queueing Systems*, 67(4):333–364, 2011.

[12] D. Mitra and A. Weiss. A closed network with a discriminatory processor sharing server. *Performance Evaluation Review*, 17(1):200–208, 1989.

[13] J. Niño Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15(2):161–198, 2007.

[14] M. Opp, K. D. Glazebrook, and V. G. Kulkarni. Outsourcing warranty repairs: Dynamic allocation. *Naval Research Logistics*, 52(5):381–398, 2005.

[15] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

[16] M. L. Puterman. *Markov Decision Processes – Discrete Stochastic Dynamic Programming*. Wiley, New York, 2005.

[17] M. H. Veatch and L. M. Wein. Scheduling a make-to-stock queue: Index policies and hedging points. *Operations Research*, 44(4):634–647, 1996.

[18] R. R. Weber. Comments on: Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15(2):211–216, 2007.

[19] R. R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

[20] R. R. Weber and G. Weiss. Addendum to: On an index policy for restless bandits. *Advances in Applied Probability*, 23(2):429–430, 1991.

[21] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.