Top (2007) 15: 211–216
DOI 10.1007/s11750-007-0029-9

DISCUSSION

# Comments on: Dynamic priority allocation via restless bandit marginal productivity indices

**Richard Weber**

Published online: 27 September 2007


José Niño-Mora has given us a very nice survey of his results for theoretical and algorithm aspects of restless bandit indexation. His paper has inspired me to contribute some new results, which I will report in three parts. Firstly, I present an extension to restless bandits in which there are more than two possible actions in each state. Secondly, I show that condition (ii) in the definition of PCL($\mathcal{F}$)-indexability is actually implied by (i), if we add a mild condition that can always be met by perturbing the data. Thirdly, I describe two interesting classes of restless bandits that are always indexable.

In the classic bandits model there are two actions available in each state: $a = 0$ (passive) and $a = 1$ (active). If the passive action is taken the state does not change. However, in the restless bandits model the passive action can produce a change in state. In practical applications one often has that $a = 0$ changes the state less aggressively than does $a = 1$, and is less costly. Nonetheless, there is nothing fundamentally different about the actions $a = 0$ and $a = 1$. This leads one to ask, why limit ourselves to only two actions? Would not most everything that has been written in the research literature about restless bandits work out just as well if we extended the action space to $a \in \{0, 1, \ldots, k\}$?

Here is an example, based on the model that Niño-Mora's describes for admission to queues in Sect. 3.1. We consider an $M/M/1$ queue with finite buffer $n$. The service rate is $\mu$. Using any of 3 possible actions, $a = \{0, 1, 2\}$, the arrival rate can be set to 2, 1, or 0, respectively. The reader can think of this as controlling an entry gate, at which an arriving stream of customers present themselves as a Poisson process

This comment refers to the invited paper available at: http://dx.doi.org/10.1007/s11750-007-0025-0.

R. Weber (✉)
Center for Mathematical Sciences, Statistical Laboratory, University of Cambridge,
Cambridge CB3 0WB, UK
e-mail: R.R.Weber@statslab.cam.ac.uk

of rate 2. So $a = 0$ means the gate is open, $a = 1$ means it is half-closed, and $a = 2$ means it is fully closed. If the gate is half-open then each arriving customer is rejected with probability 1/2. We require that the gate be closed ($a = 2$) when the queue is full. When there are $i$ customers in the system reward is obtained at a rate $r(i) = n - h(i)$, where $h(i)$ is a convex increasing holding cost. There is a cost $c(a)$ for each epoch of time that the action $a$ is taken, plus a cost of 1 for every customer who is rejected.

Any stationary policy can be described by disjoint sets $S_0$, $S_1$, and $S_2$ where $S_j$ is the set of states in which the action to be taken is $a = j$. Since $S_0 \cup S_1 \cup S_2 = \{0, \ldots, n\}$, it is sufficient to describe a stationary policy by a pair of sets $(S_0, S_1)$. It will be convenient to define $j_i$ as that $j$ such that $i \in S_j$.

Given some initial state distribution and discount factor $\beta$, we let $f^\pi$ be the expected discounted sum of rewards $E_\pi[\sum_{t=0}^{\infty} \beta^t r(x_t)]$. Similarly, $g^\pi = E_\pi[\sum_{t=0}^{\infty} \beta^t c(a(x_t))]$. Consider the $\nu$-wage problem:

$$\max_{\pi \in \Pi} \left[ f^\pi - \nu g^\pi \right].$$

If $\nu$ is very large then $g^\pi$ should be as small as possible, so should take $S_1 = S_2 := 0$.

Let us extend Niño-Mora's marginal productivity index as follows. For $i \in S_0 \cup S_1$, let

$$\nu_i^{S_0,S_1} = \frac{f_i^{j_i+1,S_0,S_1} - f_i^{j_i,S_0,S_1}}{g_i^{j_i+1,S_0,S_1} - g_i^{j_i,S_0,S_1}}.$$

We leave the reader to interpret the meanings of $f_i^{j_i+1,S_0,S_1}$ and other quantities by analogy to Niño-Mora's definitions in Sect. 2.3. Essentially, the numerator is the increase in the expected sum of discounted rewards that occurs when the state is $i$ and the action taken is to close the gate one further notch, from $a = j_i$ to $a = j_i + 1$, but then to continue with policy $(S_0, S_1)$ thereafter. The denominator is a similar measure of the increase in the costs $c(a)$ of holding the gate closed.

A natural extension of the notion of indexability is that Algorithm 1 should produce a set of indices with which it is possible to describe the optimal stationary strategy for solving the $\nu$-wage problem for any $\nu$.

Here $\mathcal{F}$ is the set of partitions of the states into sets $S_0$, $S_1$, and $S_2$ that can be constructed by initially putting all states in $S_0$ and subsequently moving states from $S_0$ to $S_1$, and from $S_1$ to $S_2$. As $\nu \to -\infty$, all states are eventually in $S_2$. The solution to the $\nu$-wage problem can be succinctly described in terms of the $2n$ indices produced by the algorithm. To find the optimal strategy we just put $i$ in $S_2$, $S_1$, $S_0$, as $\nu$ lies in the interval $(-\infty, \nu_{i,1}^*]$, $(\nu_{i,1}^*, \nu_{i,0}^*]$, $(\nu_{i,0}^*, \infty)$, respectively. The queueing control problem above is indexable in just this way when the parameters are $n = 3$, $\mu = 2$, $\beta = 1$ $h(i) = 3 - i$, $c(0) = 0$, $c(1) = 2$, and $c(2) = 5$. Since we work in continuous time, $\beta$ should be thought of as the parameter of the discounting $e^{-\beta t}$. The indices are

$$\begin{aligned}
&\{\nu_{2,1}^*, \nu_{1,1}^*, \nu_{2,2}^*, \nu_{1,2}^*, \nu_{0,1}^*, \nu_{0,2}^*\} \\
&\quad = \left\{ \tfrac{35}{108}, \tfrac{33}{142}, \tfrac{8}{41}, \tfrac{7}{46}, \tfrac{1}{8}, \tfrac{1}{12} \right\} \\
&\quad = \{0.3241, 0.2323, 0.1951, 0.1522, 0.1250, 0.0833\}.
\end{aligned}$$

**Algorithm 1**

**Output**: $\{v_{i,j} : i = 0, \ldots, n-1,\ j = 0, 1\}$.

$S_{0,0} = \{0, 1, \ldots, n\}$, $S_{0,1} = S_{0,2} = 0$.

**for** $k := 1$ **to** $2n$

**pick** $i_k \in \operatorname{argmax}\left\{v_i^{S_{0,k-1}, S_{1,k-1}}, i \in \partial_{\mathcal{F}}^{\text{out}}(S_{0,k-1}, S_{1,k-1})\right\}$

$v^*_{i_k, j_{i_k}} := v_{i_k}^{S_{0,k-1}, S_{1,k-1}}$

**for** $\ell := 0$ **to** 2

$S_{\ell,k} := S_{\ell,k-1}$

**end** {for}

$S_{j_i+1,k} := S_{j_i+1,k-1} \cup \{i_k\}$

$S_{j_i,k} := S_{j_i,k-1} \setminus \{i_k\}$

**end** {for}

**Table 1** Indexability in the control of admission rate to a queue

| $S_0$ | $S_1$ | $S_2$ | Instantaneous cost |
|---|---|---|---|
| {0, 1, 2} | 0 | 0 | $c(0)x_0 + c(0)x_1 + c(0)x_2 + \mu x_2$ |
| {0, 1} | {2} | 0 | $c(0)x_0 + c(0)x_1 + c(1)x_2 + \mu x_2$ |
| {0} | {1, 2} | 0 | $c(0)x_0 + c(1)x_1 + c(1)x_2 + \mu x_2$ |
| {0} | {1} | {2} | $c(0)x_0 + c(1)x_1 + c(2)x_2 + \mu x_2$ |
| {0} | 0 | {1, 2} | $c(0)x_0 + c(2)x_1 + c(2)x_2 + \mu x_2$ |
| 0 | {0} | {1, 2} | $c(1)x_0 + c(2)x_1 + c(2)x_2 + \mu x_2$ |
| 0 | 0 | {0, 1, 2} | $c(2)x_0 + c(2)x_1 + c(2)x_2 + \mu x_2$ |

Where might such indices be useful in suggesting heuristics for other problems? Suppose we are simultaneously controlling $N$ such queues. We wish to maximize the expected discounted sum of rewards subject to a constraint that instantaneous rate of incurring cost is always equal to some given constant, say $\gamma N$ (which we let scale with $N$). If the sequence of the indices is as above, then this implies an ordering of the stationary policies $(S_0, S_1, S_2)$. Let $x = (x_0, x_1, x_2)$ denote a state in which there are $x_i$ queues with $i$ customers. Our analysis has produced for us an ordering of stationary policies, with progressively increasing instantaneous costs shown in the table below.

Let us assume $c(0)N + \mu N < \gamma N < c(2)N$, i.e., $2 < \gamma < 5$. This ensures that $\gamma$ lies between the values of instantaneous cost that appear in the first and last rows of Table 1. Suppose $x$ is such that $\gamma$ is between the values in the second and third rows:

$$c(0)x_1 + c(0)x_1 + c(1)x_2 + \mu x_2 < \gamma N < c(0)x_1 + c(1)x_1 + c(1)x_2 + \mu x_2.$$

This suggests that we should set $a = 0$ in the $x_0$ queues with 0 customers; set $a = 1$ in the $x_2$ queues with 2 customers, and then set $a = 0$ and $a = 1$ in appropriate numbers

of the $x_1$ queues with 1 customer. We expect this to be better than an alternative way to spend our budget, such as setting $a = 0$ in the queues with 0 and 1 customer, and $a = 1$ and $a = 2$ in appropriate numbers of the $x_2$ queues with 2 customers, since $(S_0, S_1, S_2) = (\{0, 1\}, 0, \{2\})$ does not appear in our table. (For other choices of $c(0), c(1), c(2)$ it does.)

This example is illustrative. A more natural problem to consider would be one in which the are $N$ queues with unlimited buffer spaces, there are rewards for completing customers, and holding costs while customers are in the system. The constraint is that the total work done keeping entry doors closed or half-closed is never more than some $\gamma N$. As Weber and Weiss (1991) have found for classic restless bandits, we might expect this heuristic to be asymptotically optimal or near-optimal, as $N \to \infty$.

Now let me return to the classic restless bandit in which there are just two possible actions: $a \in \{0, 1\}$. I should like to draw attention to the connection between the adaptive-greedy index algorithm $\mathrm{AG}_{\mathcal{F}}$ (as applied to PCL($\mathcal{F}$)-indexable systems) and the policy improvement algorithm. Given $\nu = \nu^*_{i_{k-1}}$, there is some policy $(S_0, S_1)$ that is optimal for the $\nu^*_{i_{k-1}}$-wage problem. Suppose we imagine decreasing $\nu$ from $\nu^*_{i_{k-1}}$ until it reaches a value for which this policy is no longer optimal for the $\nu$-wage problem. This point can be recognized when the policy improvement algorithm tells us that it is possible to make a strictly improving change to the policy. This occurs when for some $i \notin S$ we have

$$f_i^{1,S} - \nu g_i^{1,S} > f_i^{0,S} - \nu g_i^{0,S},$$

or equivalently, when $\nu$ is less than

$$\nu^*_{i_k} = \max_{i \notin S} \left\{ \frac{f_i^{1,S} - f_i^{0,S}}{g_i^{1,S} - g_i^{0,S}} \right\} = \max_{i \notin S} \{ \nu_i^S \}.$$

The assumption that $g_i^{1,S} - g_i^{0,S} > 0$ means that it is impossible for a policy improvement step to take a state $i$ out of $S$ as $\nu$ decreases from $\nu^*_{i_{k-1}}$.

Niño-Mora has rightly remarked that condition (ii) in the definition of PCL($\mathcal{F}$)-indexability can be hard to verify. He proposes alternatives that are easier to apply. We can use the connection with policy improvement to argue that (ii) can be simplified to just (i) if we impose the condition (ii)′ that for every $\nu$

(ii)′: For every $\nu$, $\max_{S \in \mathcal{F}} \left[ f^S - \nu g^S \right]$ is attained by at most two distinct sets $S$ and $S'$.

This can always be made true by perturbation of the data. Let

$$\phi(\nu) = \max_{\pi} \left[ f^\pi - \nu g^\pi \right].$$

Note that $\phi(\nu)$ is piecewise linear and a convex function of $\nu$. Let us denote by $S = S(\nu)$ the minimal active set such that $\phi(\nu)$ is achieved by the stationary $S$-active set policy. For $\nu$ large enough $S = 0$. Now suppose the $S$-active set policy is optimal for some $\nu$, and we then decrease $\nu$ until it reaches some $\bar{\nu}$, where $S$ is no longer optimal, in the sense that for some $i$, the $S'$-active set policy obtained by adding or

subtracting $i$ from $S$ (which we denote as $S'$) is better than the $S$-active set policy when $\nu = \bar{\nu} - \epsilon$, and $\epsilon > 0$ is arbitrarily small. This point can be recognized because the policy improvement algorithm will tell us that we can improve things by adding or subtracting a state from $S(\nu)$. By assumption (ii)$'$ there is just one such state and it is

$$i: \quad R_i^{S'} - \nu Q_i^{S'} + \beta \sum_j p_{ij}^{S'} \big(f_j^S - \nu g_j^S\big) \geq R_i^S - \nu Q_i^S + \beta \sum_j p_{ij}^S \big(f_j^S - \nu g_j^S\big),$$
$$\text{for all } \nu \leq \bar{\nu}, \text{ and where there is equality for } \nu = \bar{\nu},$$

where $p^S$ is the probability transition matrix under the $S$-active set policy. We also must have $f^S - \bar{\nu} g^S = f^{S'} - \bar{\nu} g^{S'}$. In fact, we cannot have $S' = S \setminus \{i\}$. If this were so, we could supposedly improve the value function $f^S - (\nu - \epsilon) g^S$ by making a change in policy from $S$ to $S'$. However, positive marginal work implies that $g^{S'} < g^S$ (by policy improvement-type arguments applied to $g^{S'}$), which contradicts $f^S - (\bar{\nu} - \epsilon) g^S < f^{S'} - (\bar{\nu} - \epsilon) g^{S'}$. So we may assume that $S'$ is formed by adding a state $i$ to $S$. Note that its marginal productivity index is $\nu_i^S = \bar{\nu}$.

We now conduct policy improvement by changing the action from passive to active in state $i$. The $S'$-active set policy is optimal in an interval where $\nu$ is less than $\bar{\nu}$. If $|S| = k-1$, we now have $i_k = i$ and $\nu_{i_k} = \bar{\nu}$. The assumption that $\mathcal{F}_0 \subseteq \mathcal{F}$ is important to ensure that $S' \in \mathcal{F}$ and so our procedure gives the same output as algorithm $\mathrm{AG}_{\mathcal{F}}$ and finds $\mathcal{F}_0$. Thus with the addition of (ii)$'$, the assumptions of $\mathcal{F}_0 \subseteq \mathcal{F}$ and positive marginal work imply that $S(\nu)$ increases monotonically as $\nu$ decreases, taking in the states one by one. The indices produced by $\mathrm{AG}_{\mathcal{F}}$ are necessarily strictly decreasing.

We can almost remove (ii) entirely. The only problem occurs if $\mathcal{F}$ is something like

$$\big\{\{1\}, \{1,2\}, \{1,3\}, \{1,2,4\}, \{1,3,5\}, \{1,2,3,4\}, \ldots\big\},$$

we have $\nu_2 = \nu_3 = \nu_4 = \bar{\nu}$, and the optimal active set changes from $\{1\}$ to $\{1,2,3,4\}$ as $\nu$ passes though $\bar{\nu}$. In following $\mathrm{AG}_{\mathcal{F}}$ we might have $S = \{1\}$, and first choose to add 3 to $S$. There is now no way to reach $\{1,2,3,4\}$. The algorithm will produce indices that violate (ii). Another way to ensure that $\mathrm{AG}_{\mathcal{F}}$ works is to replace (ii) by a condition which ensures that $\mathcal{F}$ has some sufficient connectedness, so that examples like the above cannot arise. This is what Niño-Mora does in Sect. 4.2 with his definition of a monotonically connected set system and LP($\mathcal{F}$)-indexability. We note that where Niño-Mora's results are based upon parametric linear programming, the methodology of policy improvement is essentially equivalent, but can provide a second helpful viewpoint.

I am intrigued by Table 2, where José has found that amongst $10^7$ randomly generated problems indexability so frequently occurs. Practically interesting problems probably do not have rewards, costs and transition matrices that are similar to those that can be produced sampling $U[0,1]$ random variables. Nonetheless, one would like to understand Table 2. It is easy to appreciate why indexability is more prevalent as $\beta$ decreases, since in that direction myopic policies are optimality. It is less clear why the prevalence of indexability should increase with the number of states.

What might be some sufficient conditions for a classic restless bandit to be indexable? Glazebrook, Niño-Mora and Ansell have described an interesting special case

of a dual-speed discrete-time restless bandit that is always indexable (Index policies for a class of discounted restless bandits, *Adv. Appl. Probab.* 34, 754–774, 2002). We use insights from policy improvement, to provide a quick proof of this result in continuous time. Suppose that the effect of the passive action in state $i$ is to change the speed of transition out of state $i$ by a factor $\rho_i > 0$. Thus $q_{ij}^0 = \rho_i q_{ij}^1$. Rewards are obtained at rates $r_i^0$ and $r_i^1$ using the passive or active action, respectively, (so we have classic bandits for $r^0 = 0$, $\rho_i \to 0$).

Consider the discounted case, with discount rate $\beta$. Suppose $u_i$ is 0 or 1 as the action taken in state $i$ is passive or active. We adopt Whittle's model of a restless bandit, in which in state $i$ reward accrues at rate $u_i r_i^1 + (1-u_i)(r_i^0 + \nu)$, where $\nu$ is now a subsidy for taking the passive action. We make no assumption about the ordering of $r_i^0$ and $r_i^1$, or of $q_i^0$ and $q_i^1$. Under a fixed stationary policy, the expected discounted reward takes the form $V_i = a_i + b_i \nu$, where

$$b_i = \frac{1}{q_i^{u_i} + \beta}\left(1 - u_i + q_i^{u_i} \sum_{j \neq i} p_{ij} b_j\right),$$

$q_i^{u_i}$ is the rate of transitions out of state $i$, and $p_{ij}$ is the jumping chain. Note that since $u_i$ only changes the rate of transition out of state $i$, the jumping chain is the same for $u_i = 0$ and $u_i = 1$. Since the optimal value function is convex in $\nu$, we may argue, as above, that a policy improvement step could take $u_i = 0$ to $u_i = 1$ as $\nu$ increases only if

$$\frac{q_i^1}{q_i^1 + \beta} \sum_{j \neq i} p_{ij} b_j > \frac{1}{q_i^0 + \beta}\left(1 + q_i^0 \sum_{j \neq i} p_{ij} b_j\right).$$

This is equivalent to

$$-\beta - q_i^1 + \beta\left(q_i^1 - q_i^0\right) \sum_j p_{ij} b_j > 0.$$

However, the right hand side is nonpositive, either because $q_i^1 - q_i^0 \leq 0$, or because $q_i^1 - q_i^0 \geq 0$ and $\sum_j p_{ij} b_j \leq 1/\beta$. Thus, we conclude that this type of restless bandit must be indexable.

A second interesting type of restless bandit that is indexable is one in which the transition rates satisfy $q_{ij}^0 = \epsilon_j$ for all $i \neq j$. That is, when the passive action is taken in state $i$ the transition rate to state $j$ is some $\epsilon_j$, independent of $i$. The proof that such restless bandits are indexable involves noting that $b_i = b$, for some constant $b$, for all states $i$ in which the passive action is taken. Also, $b_i < b$ for all states $i$ in which the active action is taken. Thus for a $\nu$ and a state $i$ in which $u_i = 0$ is optimal,

$$\frac{1}{q_i^1 + \beta} \sum_{j \neq i} q_{ij}^1 b_j < b = \frac{1}{q_i^0 + \beta}\left(1 + \sum_{j \neq i} \epsilon_j b_j\right),$$

and so as $\nu$ increases it is impossible for a policy improvement step to take $u_i = 0$ to $u_i = 1$.