Probability in the Engineering and Informational Sciences, **24**, 2010, 349–374.
doi:10.1017/S0269964810000021

INDEXABILITY OF BANDIT PROBLEMS WITH RESPONSE DELAYS

FELIPE CARO AND ONESUN STEVE YOO
UCLA Anderson School of Management
Los Angeles, CA 90095
E-mail: fcaro@anderson.ucla.edu; onesun.yoo.2010@anderson.ucla.edu

This article considers an important class of discrete time restless bandits, given by the discounted multiarmed bandit problems with response delays. The delays in each period are independent random variables, in which the delayed responses do not cross over. For a bandit arm in this class, we use a coupling argument to show that in each state there is a unique subsidy that equates the pulling and nonpulling actions (i.e., the bandit satisfies the indexibility criterion introduced by Whittle (1988). The result allows for infinite or finite horizon and holds for arbitrary delay lengths and infinite state spaces. We compute the resulting marginal productivity indexes (MPI) for the Beta-Bernoulli Bayesian learning model, formulate and compute a tractable upper bound, and compare the suboptimality gap of the MPI policy to those of other heuristics derived from different closed-form indexes. The MPI policy performs near optimally and provides a theoretical justification for the use of the other heuristics.

1. INTRODUCTION

1.1. Motivation

Dynamic allocation of activity under uncertainty is a fundamental decision problem faced by decision makers everyday. In each time period, unable to engage in (pull) all of the existing projects (arms), the decision maker must carefully choose a subset of the projects in which to engage. Once the projects are chosen, corresponding events are set in motion and outcomes are observed, based on which the states of each projects are updated. The objective of the decision maker is to utilize the given information about the projects and choose the subset of projects each period that would maximize the long-term horizon discounted rewards.

However, this widely studied dynamic decision model, a variant of a problem better known as the multiarmed bandit problem, ignores an important dimension

 0269-9648/10 $25.00

of response delays. In practice, a project's outcome is not observed immediately, but only after a *delay* (whose length might be random), during which the decision maker continues to make decisions. Incorporating delays provide a powerful modeling framework, as it can be generalized to aid decision making in many application areas. We illustrate a few examples.

- **Clinical Trials** (Whittle [24]). In this setting, the arms correspond to medical treatments. The state of an arm represents one's state of knowledge on the effectiveness of the corresponding treatment. Pulling an arm corresponds to treating a patient with the corresponding medical treatment. One's state of knowledge on the effectiveness of the treatment will be updated only after observing the patient's treatment outcome.
- **Dynamic Assortment** (Caro and Gallien [7]). In this setting, the set of arms represent the unproduced assortment of fashion items. The state of each project represents one's knowledge on how popular the item will be. The knowledge of each item's popularity will be refined only after observing the sales, which is possible only after incurring production and distribution lead time.
- **Corporate Strategy** (Bernardo and Chowdhry [3]). In this setting, the arms correspond to regions where franchises can be opened. The state of the arms represents the revenue expectations of each region prior to opening a franchise. Once a franchise is opened, the actual sales is observed only after a delay during which the franchise reaches out to the customers. During the delay, the headquarters might decide to open more franchises in the region.
- **Management of Employees**. In this setting, the arms represent employees, and the state of the arms represents the manager's belief about the skill level of each employee. After delegating assignments to different employees, the manager can update his belief on each employee's skill levels based on the outputs, which occurs only after a delay.

Despite their practical relevance, the bandit problems with response delays have received only moderate attention in the literature (see Sect. 1.2 for a review). One reason is because the problem becomes an intractable *restless bandit* problem (Whittle [24]), as the state of an arm that is not pulled (passive) might still change when a backlogged decision is implemented.

Although they are difficult to solve optimally, many restless bandit problems can, nonetheless, be solved near optimally using the marginal productivity index (MPI)-based heuristic (Niño-Mora [19]), provided that the problem satisfies the *indexability criterion* (Whittle [24]). Hence, indexability is a desirable property, as it makes the restless bandit problem practically solvable by employing the MPI-based heuristic.

In this article, we prove that the discrete time bandit problems with stationary random delays satisfy the *indexability criterion* as long as the delayed responses do not cross over. After an overview of the related literatures in Section 1.2, we introduce the multiarmed bandit problem with response delay and describe the basic properties

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

in Section 2. In Section 3 we present the indexability result, and in Section 4 we compute the indexes for the multiarmed bandit with delay for the canonical Beta-Bernoulli learning model and test its performance and compare it to those of other closed-form index heuristics. We conclude in Section 5.

1.2. Literature Review

The literature on restless bandit indexation was created when Whittle [24] first generalized the classic bandit framework (Gittins [13]) by allowing the passive arms to change states and termed it the *restless bandit* problem. The restless bandit problems are computationally intractable to solve optimally; hence, the primary research concerns the development of heuristic policies that can be shown to be near optimal. Whittle formed a Lagrangian dual problem and defined a priority index as the Lagrange multiplier associated with an arm that makes the decision maker indifferent between pulling and not pulling the arm. He showed that this priority index generalizes the Gittins index and devises a priority-index policy that pulls the arms with the highest index values. He further conjectured the asymptotic optimality of the priority-index policy, which Weber and Weiss [22] later largely validated and Weiss [23] showed a special case for which the conjecture holds. However, Whittle stated that for the index to be well defined, the restless bandit problem must first satisfy the *indexability criterion*; that is, the Lagrange multiplier that equates the pulling and nonpulling actions must be unique for every possible state of a given arm. He showed that indexibility cannot be taken for granted by providing counterexamples. Moreover, verifying indexibility itself is nontrivial, and, until recently, sufficient conditions satisfied by a broad subclass of restless bandits were unknown.

Niño-Mora [18] pioneered the field of restless bandit indexation to theoretically provide sufficient conditions for indexability. In particular, Niño-Mora [18] generalized the Whittle priority index by defining the MPI in terms of the more general and economically intuitive reward/work measure and showed that the MPI's interpretation can be applied in an identical manner to many other classic index policies that were shown to be optimal, including the celebrated Gittins index. Using the MPI, he identified classes of restless bandit problems that satisfy the sufficient conditions, mostly under the assumption of a finite state space (see Niño-Mora [19] and references therein). Our work contributes to the literature by expanding the known class of indexable restless bandit problems.

The problem of bandits with response delays has received only moderate attention in the literature. Eick [10] examined the clinical trials setting in which a patient's lifetime is modeled as a geometric random variable and provided the first proof of indexability for a delayed response bandit when the discount factor δ is than $1/2$. Wang and Bickis [21] extended this result to arbitrary lifetime distributions under certain regularity conditions, but those conditions reduce to $\delta < 1/2$ in the discrete time case. In contrast, our result shows indexability for the more applicable discount factors $\delta < 1$. Hardwick, Oehmke, and Stout [15] considered the response delay bandit model in which patients arrive according to a Poisson process with the treatment time

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

having exponential response delays. They identified heuristics that perform well under the objective of minimizing patient loss. However, the heuristics are randomized rules that are not grounded in indexability theory. More recently, Niño-Mora [19] examined a finite queue with a one-period response delay and showed its indexability. However, the model lacks generality in that the state space must be finite and the delay is limited to one period, whereas our model allows for infinite state space and arbitrary delay lengths, which can be stationary random as long as the delayed responses do not cross over. We refer the interested reader to Altman and Stidham [1] and Ehsan and Liu [9] for other queuing applications with delayed information. Finally, Caro and Galien [7] introduced a closed-form index, generalized it to incorporate response delay, and showed that the resulting index policy has near-optimal performance. Our work suggests that their method performs well because their closed-form index is a good approximation of the MPI.

2. PROBLEM DESCRIPTION

2.1. Model Basics

The decision problem is defined in discrete time, where each period is indexed by t, representing t steps to go, and the rewards are discounted by $\delta < 1$ each period. The response delay ℓ is also a discrete quantity. In each time period, with S available arms but only able to pull N ($N < S$), the decision maker must carefully assess the state of each arm s. Once the arms are pulled, the outcomes are observed ℓ periods later, at which point the state of the arm changes. The objective of the decision maker is to pull the N arms each period to maximize the long-term discounted rewards.

Let $x_s \in \Re$ denote the state of arm s and the vector $\mathbf{x} \in \Re^S$ denote the state of all S arms. Let $R_s(x_s)$ denote the reward of arm s that depends on its state. For simplicity, we assume that the reward functions R_s are uniformly bounded, but this assumption can be relaxed (for instance, see Condition B in Gittins [14, p. 17] or the Bayesian formulation given in Burnetas and Katehakis [6]). The decision on arm s each period is represented by $u_s \in \{0, 1\}$, where a value of $u_s = 1$ corresponds to a (*Pull*) decision, whereas a value of $u_s = 0$ corresponds to a (*NotPull*) decision. In each period, it is not possible to pull more than N arms (i.e., $\sum_s^S u_s \leq N$). The vector $\mathbf{u} \in \{0, 1\}^S$ denotes the decision on all S arms, and each of the vectors $(\mathbf{v}^1, \ldots, \mathbf{v}^\ell)$ represent the decisions that had been made in previous periods, with $\mathbf{v}^1$ being the oldest decision that will be implemented this period and the $\mathbf{v}^\ell$ being the most recent decision.

Each arm s follows an independent Markovian process. If $v_s^1 = 1$, the function $f_s(x_s, v_s^1, w_s)$ denotes the state that the arm s transitions to from state x_s given the decision v_s^1 and the random component $w_s(x_s)$, which depends on state x_s; if $v_s^1 = 0$, $f_s(x_s, v_s^1, w_s) = x_s$, signifying that the state of the arm remains unchanged. Letting the vector $\mathbf{w(x)} \in \Re^S$ represent a vector of random variables $w_s(x_s)$, the vector $f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}) \in \Re^S$ represents state that all the arms transitions to state $\mathbf{x}$ given the decision vector $\mathbf{v}^1$ and the random component $\mathbf{w}$.

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

Let $\mathbf{J}_t^*(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^\ell)$ denote the maximum discounted reward with t steps to go, given the arms' state $\mathbf{x}$ and the decisions of the previous periods $\mathbf{v}^1, \ldots, \mathbf{v}^\ell$. Then the multiarmed bandit problem with delay can be expressed as the following dynamic program:

$$\begin{aligned}(\text{BD}): \quad & \mathbf{J}_t^*(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^\ell) \\ & = \max_{\mathbf{u}} \left\{ \sum_{s=1}^{S} R_s(x_s) v_s^1 + \delta \mathrm{E}_{\mathbf{w}} \mathbf{J}_{t-1}^*(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^\ell, \mathbf{u}) \right\} \\ & \quad \text{s.t.} \quad \mathbf{u} \in \{0,1\}^S, \ \sum_{s=1}^{S} u_s \le N\end{aligned}$$

for $t > \ell$, and for $t \le \ell$,

$$\mathbf{J}_t^*(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^\ell) = \sum_{s=1}^{S} R_s(x_s) v_s^1 + \delta \mathrm{E}_{\mathbf{w}} \mathbf{J}_{t-1}^*(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^\ell)\}, \quad \mathbf{J}_0^*(\cdot) \equiv 0.$$

Because a passive arm's state can change when the delayed (*Pull*) decision is implemented (i.e., $v_s^1 = 1$), the problem is a restless bandit problem (Whittle [24]), which is intractable. A known heuristic that can solve the restless bandit problems near optimally is the MPI-index policy, but this policy is well defined only if the problem satisfies the indexability criteria. We explain this next.

2.2. Indexability Criterion and the Equivalence Relation

Whittle [24] formed the Lagrangian dual of the restless bandit problem, where the dual variable λ has the interpretation of *subsidy* for not pulling the arm and defines the priority index as the value of λ that makes the decision indifferent between pulling and not pulling the arm. If the index λ is to be meaningful however, it must induce a consistent ordering of the arms, in that any arm that is not pulled under a subsidy λ will also be not pulled under a higher subsidy $\lambda' > \lambda$. An equivalent statement in terms of *cost* is that if any arm is pulled under a cost λ, it must also be pulled under a lower cost $\lambda' < \lambda$. The formal definition of indexability for a single independent arm is the following:

DEFINITION (Whittle [24]): *Let $D_s(\lambda)$ be the set of values of x_s for which project s would be rested under a λ-subsidy policy. Then the project is* indexable *if $D_s(\lambda)$ increases monotonically from $\emptyset$ to Υ_s as λ increases from $-\infty$ to $+\infty$, where Υ_s is the full state space for project s.*

A restless bandit problems is indexable if each one of its arms is indexable. The proof given in the next section shows that a single-arm bandit with response delay satisfies Whittle's definition. Therefore, the multiarmed bandit problem with response delays (BD) is indexable.

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

Before showing the main indexibility result, note that there are potentially four possible formulations of our problem, due to the four different ways of accounting for the Lagrange multiplier λ in the rewards. First, from the definition, λ can either have a *subsidy* or *cost* interpretation. Moreover, it might be accounted for when the pull/not pull decision is *made* (before the delay) or when the decision gets *implemented* (after the delay). Accounting for λ before the delay has been more prevalent in the literature. For instance, Wang and Bickis [21] and Caro and Gallien [7] considered the subsidy and cost interpretations, respectively, under that framework. In our proof of indexibility we found it easier to account for λ after the delay. Regardless of this choice, our first proposition shows that the accounting method does not affect the indexability result. Furthermore, because the order of the indexes do not change, the priority indexes policy whose indexes are derived from four different accounting methods would be identical.

PROPOSITION 1: *Suppose one formulation is indexable. Then the other three formulations are also indexable. Moreover, the ranking of the indexes does not change from one formulation to the other.*

PROOF: See Appendix A. ■

In the next section, without a loss of generality, we examine the formulation where λ represents a subsidy for not pulling and rewards (including the subsidy) are accounted when the decisions are implemented after the delay.

3. STRUCTURAL RESULTS

In this section, we establish the indexability of multiarmed bandit problem with (1) constant delay, and then (2) with stationary random delays in which the delayed responses do not cross over. We do so by showing that the single-arm bandit with response delay is indexable.

We point out that the underlying bandit problem is not restless; in other words, the state of arm that is not pulled does not change ℓ periods later. Only after the incorporation of the response delay does the problem become "restless." We exploit this underlying nonrestless structure of the problem in the proof by matching sample paths based on a coupling argument.

Let $J^{\lambda}_{t,s}(z)$ denote the maximum profit-to-go function of an arm s with a subsidy λ for a state z with t periods to go. To show that arm s satisfies Whittle's indexibility definition, for each state z a unique index λ must exist such that the expected discounted profit from pulling is equal to that from not pulling. More formally, if we denote the maximum profit-to-go function of the arm after it is pulled and after it is not pulled respectively as

$$J^{\lambda}_{t,s}(z)^{(Pull)} \quad \text{and} \quad J^{\lambda}_{t,s}(z)^{(NotPull)},$$

we would want to show that there exists a unique λ such that $J^{\lambda}_{t,s}(z)^{(Pull)} = J^{\lambda}_{t,s}(z)^{(NotPull)}$. We can achieve this if we show that $\Delta J^{\lambda}_{t,s}(z) \equiv J^{\lambda}_{t,s}(z)^{(Pull)} -$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

$J^{\lambda}_{t,s}(z)^{(NotPull)}$ is a decreasing function of λ for every state z and then take the limit when $t \to \infty$.

3.1. Indexability of Constant Delay

Consider a single-arm s and a constant response delay of ℓ periods. The maximum profit-to-go function at time t and state x_s with delayed orders $(v_s^1, \ldots, v_s^\ell)$ is given by

$$J^{\lambda}_{t,s}(x_s, v_s^1, \ldots, v_s^\ell) = R_s(x_s)v_s^1 + \lambda(1 - v_s^1) + \delta \max\left\{E_{w_s} J^{\lambda}_{t-1,s}(f_s(x_s, v_s^1, w_s), v_s^2, \ldots, v_s^\ell, 1), E_{w_s} J^{\lambda}_{t-1,s}(f_s(x_s, v_s^1, w_s), v_s^2, \ldots, v_s^\ell, 0)\right\}.$$

The expectation is taken with respect to the random variable w_s, which has an arbitrary distribution that is dependent on the current state x_s. When necessary, we will write $w_s(x_s)$ to make the parameter dependence explicit.

The difference in value at time t between the (*Pull*) and (*NotPull*) decisions has the following expression:

$$\begin{aligned}
&\Delta J^{\lambda}_{t,s}(x_s, v_s^1, \ldots, v_s^\ell) \\
&\quad = R_s(x_s)v_s^1 + \lambda(1 - v_s^1) + \delta E_{w_s} J^{\lambda}_{t-1,s}(f_s(x_s, v_s^1, w_s), v_s^2, \ldots, v_s^\ell, 1) \\
&\qquad - \left\{R_s(x_s)v_s^1 + \lambda(1 - v_s^1) + \delta E_{w_s} J^{\lambda}_{t-1,s}(f_s(x_s, v_s^1, w_s), v_s^2, \ldots, v_s^\ell, 0)\right\} \\
&\quad = \delta E_{w_s}\{J^{\lambda}_{t-1,s}(f_s(x_s, v_s^1, w_s), v_s^2, \ldots, v_s^\ell, 1) - J^{\lambda}_{t-1,s}(f_s(x_s, v_s^1, w_s), v_s^2, \ldots, v_s^\ell, 0)\}.
\end{aligned}$$

Letting $z_s = (x_s, v_s^1, \ldots, v_s^\ell)$ denote the augmented state, we can rewrite the value function as

$$J^{\lambda}_{t-1,s}(z_s) = J^{\lambda}_{t-1,s}(z_s)^{(Pull)} + [\Delta J^{\lambda}_{t-1,s}(z_s)]^- = J^{\lambda}_{t-1,s}(z_s)^{(NotPull)} + [\Delta J^{\lambda}_{t-1,s}(z_s)]^+,$$

where $[r]^+ = \max\{0, r\}$ and $[r]^- = \max\{0, -r\}$.

We now prove the monotonicity result. For notational simplicity, we will omit the subscript s in the proof.

PROPOSITION 2: *For all augmented state z, $\Delta J^{\lambda}_t(z)$ is decreasing in λ.*

PROOF: Using induction, we show that for any $\lambda_1 > \lambda_2$, $\Delta J^{\lambda_1}_t(z) < \Delta J^{\lambda_2}_t(z)$ for all z. The proof is for $\ell \geq 2$. For $\ell = 1$, the notation would have to be slightly different, but the argument is exactly the same.

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

Base Case: $t = \ell + 1$. Here, we make the (*Pull*)/(*NotPull*) decision only once and observe the expected outcome in the remaining ℓ periods. We have

$$\begin{aligned}
&\Delta J^{\lambda}_{\ell+1}(x, v^1, \ldots, v^\ell) \\
&\quad = \delta E_{w_1}\left\{J^{\lambda}_{\ell}(f(x, v^1, w), v^2, \ldots, v^\ell, 1) - J^{\lambda}_{\ell}(f(x, v^1, w), v^2, \ldots, v^\ell, 0)\right\} \\
&\quad = \delta^{\ell} E_{w_1} E_{w_2} \cdots E_{w_l}\left\{J^{\lambda}_{1}(f^{\circ\ell}(x, \underline{v}, \underline{w}), 1) - J^{\lambda}_{1}(f^{\circ\ell}(x, \underline{v}, \underline{w}), 0)\right\} \\
&\quad = \delta^{\ell}\{E_{\underline{w}} R(f^{\circ\ell}(x, \underline{v}, \underline{w})) - \lambda\},
\end{aligned}$$

where the vector $\underline{v}$ represents all the delayed decisions $(v^1, \ldots, v^\ell)$, the vector $\underline{w}$ represents the series of dependent random variables $(w_1, w_2, \ldots, w_\ell)$, and $f^{\circ\ell}(x, \underline{v}, \underline{w})$ is a short-hand notation for $f(f \cdots f(f(f(x, v^1, w_1), v^2, w_2), v^3, w_3), \ldots), v^\ell, w_\ell)$. Each w_i's distribution depends on the sample path of the states, and the expression $E_{\underline{w}}$ represents an ℓ-iterated expectation framework. This expression is clearly decreasing in λ, $\forall z$.

Induction Step: $t > \ell + 1$. Assume that $\forall z = (x, v^1, \ldots, v^\ell)$ and $\lambda_1 > \lambda_2$, $\Delta J^{\lambda_1}_{t-1}(z) < \Delta J^{\lambda_2}_{t-1}(z)$. We will show that $\forall z = (x, v^1, \ldots, v^\ell)$ and $\lambda_1 > \lambda_2$, $\Delta J^{\lambda_1}_{t}(z) < \Delta J^{\lambda_2}_{t}(z)$.

We write out the expression for $\Delta J^{\lambda_1}_{t-1}(z)$ and $\Delta J^{\lambda_2}_{t-1}(z)$ as follows:

$$\begin{aligned}
\Delta J^{\lambda_1}_{t}(x, v^1, \ldots, v^\ell) = \delta E_{w_1}&\left\{J^{\lambda_1}_{t-1}(f(x, v^1, w_1), v^2, \ldots, v^\ell, 1)\right. \\
&\left. - J^{\lambda_1}_{t-1}(f(x, v^1, w_1), v^2, \ldots, v^\ell, 0)\right\}, \\
\Delta J^{\lambda_2}_{t}(x, v^1, \ldots, v^\ell) = \delta E_{w'_1}&\left\{J^{\lambda_2}_{t-1}(f(x, v^1, w'_1), v^2, \ldots, v^\ell, 1)\right. \\
&\left. - J^{\lambda_2}_{t-1}(f(x, v^1, w'_1), v^2, \ldots, v^\ell, 0)\right\}.
\end{aligned}$$

The difference between the first and second expressions gives us the following:

$$\begin{aligned}
\text{DIFF} &\equiv \Delta J^{\lambda_1}_{t}(x, v^1, \ldots, v^\ell) - \Delta J^{\lambda_2}_{t}(x, v^1, \ldots, v^\ell) \\
&= \delta E_{w_1}\left\{J^{\lambda_1}_{t-1}(f(x, v^1, w_1), v^2, \ldots, v^\ell, 1)\right. \\
&\quad \left. - J^{\lambda_1}_{t-1}(f(x, v^1, w_1), v^2, \ldots, v^\ell, 0)\right\} \\
&\quad - \left(\delta E_{w'_1}\left\{J^{\lambda_2}_{t-1}(f(x, v^1, w'_1), v^2, \ldots, v^\ell, 1)\right.\right. \\
&\quad \left.\left. - J^{\lambda_2}_{t-1}(f(x, v^1, w'_1), v^2, \ldots, v^\ell, 0)\right\}\right).
\end{aligned}$$

After rewriting each row's expression $J = \max\{J^{(Pull)}, J^{(NotPull)}\}$ in terms of $J = J^{(Pull)} + [J^{(NotPull)} - J^{(Pull)}]^+$ and $J = J^{(NotPull)} + [J^{(NotPull)} - J^{(Pull)}]^-$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

and rearranging the terms, we have

$$\begin{aligned}
\text{DIFF} = {} & \delta E_{w_1}\left\{J_{t-1}^{\lambda_1}(f(x,v^1,w_1),v^2,\ldots,v^\ell,1)^{(NotPull)}\right. \\
& \left. -J_{t-1}^{\lambda_1}(f(x,v^1,w_1),v^2,\ldots,v^\ell,0)^{(Pull)}\right\} \\
& -\delta E_{w_1'}\left\{J_{t-1}^{\lambda_2}(f(x,v^1,w_1'),v^2,\ldots,v^\ell,1)^{(NotPull)}\right. \\
& \left. -J_{t-1}^{\lambda_2}(f(x,v^1,w_1'),v^2,\ldots,v^\ell,0)^{(Pull)}\right\} \\
& +\delta E_{w_1}[\Delta J_{t-1}^{\lambda_1}(f(x,v^1,w_1),v^2,\ldots,v^\ell,1)]^+ \\
& -\delta E_{w_1'}[\Delta J_{t-1}^{\lambda_2}(f(x,v^1,w_1'),v^2,\ldots,v^\ell,1)]^+ \\
& -\delta E_{w_1}[\Delta J_{t-1}^{\lambda_1}(f(x,v^1,w_1),v^2,\ldots,v^\ell,0)]^- \\
& +\delta E_{w_1'}[\Delta J_{t-1}^{\lambda_2}(f(x,v^1,w_1'),v^2,\ldots,v^\ell,0)]^-.
\end{aligned}$$

Each of the last two rows is less than or equal to zero via the induction assumption, and we will denote the sum of the last two rows as $C \le 0$. After evaluating out each term in the first two rows—for example

$$\begin{aligned}
& J_{t-1}^{\lambda_1}(f(x,v^1,w_1),v^2,\ldots,v^\ell,1)^{(NotPull)} \\
& \quad = R(f(x,v^1,w_1))v^2+\lambda_1(1-v^2) \\
& \qquad +\delta E_{w_2} J_{t-2}^{\lambda_1}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,1,0)
\end{aligned}$$

we arrive at the following expression:

$$\begin{aligned}
\text{DIFF} = {} & \delta E_{w_1}\{R(f(x,v^1,w_1))v^2+\lambda_1(1-v^2) \\
& +\delta E_{w_2} J_{t-2}^{\lambda_1}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,1,0)\} \\
& -\delta E_{w_1}\{R(f(x,v^1,w_1))v^2+\lambda_1(1-v^2) \\
& +\delta E_{w_2} J_{t-2}^{\lambda_1}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,0,1)\} \\
& -\delta E_{w_1'}\{R(f(x,v^1,w_1'))v^2+\lambda_2(1-v^2) \\
& +\delta E_{w_2'} J_{t-2}^{\lambda_2}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,1,0)\} \\
& +\delta E_{w_1'}\{R(f(x,v^1,w_1'))v^2+\lambda_2(1-v^2) \\
& +\delta E_{w_2'} J_{t-2}^{\lambda_2}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,0,1)\}+C \\
\le {} & \delta^2 E_{w_1}E_{w_2}\{J_{t-2}^{\lambda_1}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,1,0) \\
& \qquad -J_{t-2}^{\lambda_1}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,0,1)\} \\
& +\delta^2 E_{w_1'}E_{w_2'}\{J_{t-2}^{\lambda_2}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,0,1) \\
& \qquad -J_{t-2}^{\lambda_2}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,1,0)\}.
\end{aligned}$$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

We now introduce the coupling argument. Consider the bandit with subsidy λ_1 starting from two different states. The first, which we refer to as System A, with time $t-2$ to go from state $f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,1,0)$, follows the optimal policy. The second, which we refer to as System B, starts from state $f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,0,1)$, but it implements the same decision as System A in the first ℓ stages, and after that, it follows its own optimal policy. Let π^* denote the optimal policy of System A (which is followed by System B for the first ℓ periods). Note that both System A and System B start from the same (nonaugmented) state $f(f(x,v^1,w_1),v^2,w_2)$ and experience the same number of state transitions within the next ℓ periods. Moreover, these transitions have exactly the same Markovian dynamics, so by defining the two processes on a common probability space, we can assume that the actual transitions are the same.

Let $G^{\lambda_1}_{\pi^*_{t-2}}(z)$ represent the value of being in state z with time $t-2$ to go and following the policy π^*. Then we have

$$J^{\lambda_1}_{t-2}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,1,0) = G^{\lambda_1}_{\pi^*_{t-2}}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,1,0)$$

and

$$J^{\lambda_1}_{t-2}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,0,1) \geq G^{\lambda_1}_{\pi^*_{t-2}}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,0,1).$$

The first equality and the second inequality follow because π^* is optimal for System A but suboptimal for System B. The same coupling argument can be used for the bandit with subsidy λ_2, only that System A would start in state $(\ldots,0,1)$ and System B would start in (. . .,1,0). Denoting π^{**} as the optimal policy of System A, we have

$$J^{\lambda_2}_{t-2}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,0,1) = G^{\lambda_2}_{\pi^{**}_{t-2}}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,0,1)$$

and

$$J^{\lambda_2}_{t-2}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,1,0) \geq G^{\lambda_2}_{\pi^{**}_{t-2}}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,1,0).$$

By subtracting these smaller values of $G^{\lambda_2}_{\pi^{**}_{t-2}}$ and $G^{\lambda_1}_{\pi^*_{t-2}}$, the DIFF can be bounded above as follows:

$$\text{DIFF} \leq \delta^2 E_{w_1} E_{w_2} \left\{ G^{\lambda_1}_{\pi^*_{t-2}}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,1,0) - G^{\lambda_1}_{\pi^*_{t-2}}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,0,1) \right\} \quad \text{(DIFF.1)}$$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

$$+\delta^2 E_{w_1'} E_{w_2'} \Big\{ G^{\lambda_2}_{\pi^{**}_{t-2}}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,0,1) - G^{\lambda_2}_{\pi^{**}_{t-2}}(f(f(x,v^1,w_1'),v^2,w_2'),v^3,\ldots,v^\ell,1,0)\Big\} \quad \text{(DIFF.2)}.$$

We now evaluate expressions (DIFF.1) and (DIFF.2). (DIFF.1):

$$\begin{aligned}
&\delta^2 E_{w_1} E_{w_2} G^{\lambda_1}_{\pi^*_{t-2}}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,1,0)\\
&\quad = \delta^\ell E_{w_1} E_{w_2} \cdots E_{w_\ell} G^{\lambda_1}_{\pi^*_{t-l}}(f^{\circ\ell}(x,\underline{v},\underline{w}),1,0,u_1^*,\ldots,u^*_{\ell-2})\\
&\quad = \delta^\ell E_{w_1} E_{w_2} \cdots E_{w_\ell}\{R(f^{\circ\ell}(x,\underline{v},\underline{w}))\\
&\qquad + \delta E_{w_{\ell+1}} G^{\lambda_1}_{\pi^*_{t-\ell-1}}(f(f^{\circ\ell}(x,\underline{v},\underline{w}),1,w_{\ell+1}),0,u_1^*,\ldots,u^*_{\ell-2},u^*_{\ell-1})\}\\
&\quad = \delta^\ell E_{\underline{w}}\{R(f^{\circ\ell}(x,\underline{v},\underline{w})) + \delta E_{w_{\ell+1}}(\lambda_1 + \delta J^{\lambda_1}_{t-\ell-2}\\
&\qquad \times (f(f^{\circ\ell}(x,\underline{v},\underline{w}),1,w_{\ell+1}),u_1^*,\ldots,u^*_{\ell-1},u^*_\ell))\},
\end{aligned}$$

where u^*_τ means the optimal τth action for System A. Similarly,

$$\begin{aligned}
&\delta^2 E_{w_1} E_{w_2} G^{\lambda_1}_{\pi^*_{t-2}}(f(f(x,v^1,w_1),v^2,w_2),v^3,\ldots,v^\ell,0,1)\\
&\quad = \delta^\ell E_{w_1} E_{w_2} \cdots E_{w_\ell} G^{\lambda_1}_{\pi^*_{t-l}}(f^{\circ\ell}(x,\underline{v},\underline{w}),0,1,u_1^*,\ldots,u^*_{\ell-2})\\
&\quad = \delta^\ell E_{w_1} E_{w_2} \cdots E_{w_\ell}\{\lambda_1 + \delta G^{\lambda_1}_{\pi^*_{t-\ell-1}}(f^{\circ\ell}(x,\underline{v},\underline{w}),1,u_1^*,\ldots,u^*_{\ell-2},u^*_{\ell-1})\}\\
&\quad = \delta^\ell E_{\underline{w}}\{\lambda_1 + \delta(R(f^{\circ\ell}(x,\underline{v},\underline{w}))\\
&\qquad + \delta E_{w_{\ell+2}} J^{\lambda_1}_{t-\ell-2}(f(f^{\circ\ell}(x,\underline{v},\underline{w}),1,w_{\ell+2}),u_1^*,\ldots,u^*_{\ell-1},u^*_\ell))\}.
\end{aligned}$$

Both $w_{\ell+1}$ and $w_{\ell+2}$ have dependence on the same state $f^{\circ\ell}(x,\underline{v},\underline{w})$ and hence have the same distribution. Therefore, the last terms from the expressions, $\delta E_{w_{\ell+1}} J^{\lambda_1}_{t-\ell-2}(\cdot)$ and $\delta E_{w_{\ell+2}} J^{\lambda_1}_{t-\ell-2}(\cdot)$ cancel and we have

$$\text{(DIFF.1)} = \delta^\ell(1-\delta) E_{\underline{w}} R(f^{\circ\ell}(x,\underline{v},\underline{w})) - \delta^\ell(1-\delta)\lambda_1.$$

Following the same sequence of reasoning, we get the expression for (DIFF.2):

$$\text{(DIFF.2)} = \delta^\ell(1-\delta)\lambda_2 - \delta^\ell(1-\delta) E_{\underline{w'}} R(f^{\circ\ell}(x,\underline{v},\underline{w'})).$$

Summing expressions (DIFF.1) and (DIFF. 2),

$$\begin{aligned}
\text{DIFF} &\le \text{(DIFF.1)} + \text{(DIFF.2)}\\
&= \delta^\ell(1-\delta)\lambda_2 - \delta^\ell(1-\delta)\lambda_1\\
&\quad + \delta^\ell(1-\delta) E_{\underline{w}} R(f^{\circ\ell}(x,\underline{v},\underline{w})) - \delta^\ell(1-\delta) E_{\underline{w'}} R(f^{\circ\ell}(x,\underline{v},\underline{w'})).
\end{aligned}$$

The initial w_1 and w_1' share the same distribution because it is dependent on the original state of the arm x at time t. Also, since $\underline{v}$ are identical, the expression

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

involving the expectations cancel and we have that DIFF is bounded above by

$$\begin{aligned}\text{DIFF} &\leq \delta^{\ell}(1-\delta)\lambda_2 - \delta^{\ell}(1-\delta)\lambda_1 \\ &= (\lambda_2 - \lambda_1)\delta^{\ell}(1-\delta) < 0, \quad \forall \delta < 1.\end{aligned}$$

■

We now present the result that multiarmed bandit problems with constant response delay are indexable.

THEOREM 1: *The multiarmed bandit problem with constant response delay ℓ is indexable.*

PROOF: First, we have $\Delta J_t^{\lambda}(z)$ decreasing in $\lambda, \forall t, z$, and it is easy to see that $\Delta J_t^{0}(z) > 0$ and $\Delta J_t^{\infty}(z) < 0$. Thus, to show that a well-defined λ exists such that $\Delta J_t^{\lambda}(z) = 0$, it suffices to show that $\Delta J_t^{\lambda}(z)$ is continuous in λ. We do this by induction.

When $t = \ell$, we have $\Delta J_{\ell}^{\lambda}(x, v^1, \ldots, v^{\ell}) = E_w(R(x) - \lambda)$ and $J_{\ell}^{\lambda}(x, v^1, \ldots, v^{\ell}) = \max\{E_w(R(x)), \lambda\}$, which are clearly continuous in λ for all z. Suppose that $\Delta J_{t-1}^{\lambda}(z)$ and $J_{t-1}^{\lambda}(z)$ are continuous in λ for all z. Then

$$\begin{aligned}\Delta J_t^{\lambda}(x, v^1, \ldots, v^{\ell}) = \delta E_w\{&J_{t-1}^{\lambda}(f(x, v^1, w), v^2, \ldots, v^{\ell}, 1) \\ &- J_{t-1}^{\lambda}(f(x, v^1, w), v^2, \ldots, v^{\ell}, 0)\}\end{aligned}$$

and

$$\begin{aligned}&J_t^{\lambda}(x, v^1, \ldots, v^{\ell}) \\ &\quad = \max\{E_w J_{t-1}^{\lambda}(f(x, v^1, w), v^2, \ldots, v^{\ell}, 1), E_w J_{t-1}^{\lambda}(f(x, v^1, w), v^2, \ldots, v^{\ell}, 0)\}\end{aligned}$$

are clearly continuous in λ. Moreover, as $R(x)$ is uniformly bounded (by problem assumption), $J_t^{\lambda}(z)$ converges as $t \to \infty$. Hence, there is a well-defined λ such that $\Delta J^{\lambda}(z) = 0$. ■

3.2. Indexability of Stationary Random Delay

In many practical settings, delays might be random. We show that the indexability result can be generalized to bandit problems with stationary random delays, in which the delayed responses do not cross over (i.e., the stochastic delay $\ell \in \{m, m+1\}$ for some fixed integer m). If, however, the randomness in the delay lengths permits the delayed responses to cross over (i.e., $\ell \in \{m, \ldots, m+K\}$, $K > 1$), then the bandit problem is no longer indexable.

THEOREM 2: *The bandit with stationary random delay is indexable if the delayed responses do not crossover. However, indexability need not hold if the delayed responses are allowed to cross over.*

PROOF: See Appendix B. ■

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

This result is analogous to the inventory systems with stochastic lead times. In particular, if the random delay process does not have order crossovers, the base-stock policy is shown to be optimal (e.g., Kaplan [16], Muharremoglu and Yang [17]). However, Robinson, Bradley, and Thomas [20] showed that the base-stock policy is no longer optimal when the order are allowed to cross over.

4. NUMERICAL WORK

In this section, we examine the Beta-Bernoulli learning model in which the prior distribution of the success probability p of the Bernoulli random variable is characterized by a Beta distribution with parameters (α, β). The latter also represents the state of an arm that is updated in a Bayesian manner: to $(\alpha + 1, \beta)$ after observing a success or to $(\alpha, \beta + 1)$ after observing a failure (see Gittins [14]). Since the uniform distribution on $[0, 1]$ can be written as a Beta with parameters $(1, 1)$, a bandit in state (α, β) is statistically equivalent to one that began with its success probability p having an a priori uniform distribution and that has now shown $\alpha - 1$ successes and $\beta - 1$ failures in $\alpha + \beta - 2$ pulls.

We compute the indexes for the multiarmed bandit model with constant delay ℓ. Then, using the indexes, we examine the performance of the resulting MPI policy against an upper bound and compare it to those of other existing closed-form indexes.

4.1. Index Computation

Compared to the classical multiarmed bandit problem (with no delay), the indexes from Theorems 1 and 2 do not have an equivalent representation as an optimal stopping-time problem. Therefore, an approach to compute the indexes based on this property, which Gittins [14] called the *direct approach*, is not available. Instead, we adopt the *calibration approach*, which uses dynamic programming value iteration (see Gittins [14] for further discussion of both approaches).

The indexes for the bandit problem without delay using the Beta-Bernoulli learning model have been computed and tabulated in Gittins [14]. We extend this table by adding the indexes for delays $\ell \in \{1, 2, 3, 4, 5\}$ and discount factors $\delta \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$ and make it available online for public use (see the authors' website). The indexes have been computed using the subsidy/implementation framework, and by Proposition 1, the index values under other reward accounting methods are the same up to a constant factor, which does not affect the actions suggested by the MPI-index policy.

4.2. Numerical Simulation

In this subsection, we examine the performance of the MPI policy for the Beta-Bernoulli learning model (DeGroot [8]) with constant delays. We compute a performance upper bound by solving a relaxed multiarmed bandit problem in which the

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

TABLE 1. Closed-Form Index Formulas.

Name	Closed-Form Index	Index-Specific
Myopic (MYO)	$E[B]$	
Caro–Gallien (CG)	$E[B] + z_{\delta,\ell}\sqrt{V[B]}(V[B]/V[Y])^{c_{\delta,\ell}}$	$z_{\delta,\ell}$, $c_{\delta,\ell}$
Brezzi–Lai (BL)	$E[B] + \sqrt{V[B]}\psi\left(\frac{V[B]}{\ln(1/\delta)V[Y]}\right)$	$\psi(\cdot)$
Ginebra–Clayton (GC)	$E[B] + k_{\delta,\ell}\sqrt{V[B]}$	$k_{\delta,\ell}$

Note: Here B denotes the Beta prior with parameters (α, β) and a *delayed-adjusted* variance equal to $V[B] = \alpha\beta\left((\alpha+\beta)^2(\alpha+\beta+\sum_{\tau=1}^{\ell} v^{\tau}+1)\right)^{-1}$, where v^{τ} is the τth delayed action. Accordingly, Y denotes a Bernoulli random variable with success probability $\alpha(\alpha+\beta)^{-1}$. For index-specific coefficients and functions, refer to the original articles.

TABLE 2. Suboptimality Gap for the MPI Policy and the Closed-Form Benchmark Policies.

ℓ	MYO (%)	CG (%)	BL (%)	GC (%)	MPI (%)	UpperBnd
1	7.96	0.74	0.69	0.53	**0.51**	60.30
2	8.80	1.43	1.66	1.31	**1.53**	59.65
3	8.98	2.26	2.72	2.71	**2.15**	59.01
4	7.23	2.61	2.81	2.80	**2.82**	58.38
5	6.34	3.56	3.51	3.66	**3.63**	57.80

Note: $(\alpha, \beta) = (1, 1)$, $(S, N) = (32, 4)$, $\delta = 0.95$.

constraint that does not allow more than N arms pulled per period is only required to hold on average (see the Appendix C for the upper bound formulation). We use this to gauge the suboptimality of the MPI policy. We then compare its performance with the myopic policy (MYO) that maximizes the single-period reward (see, for instance, Aviv and Pazgal [2]), and the closed-form index policies developed by Caro and Gallien ([7]; denoted CG), Brezzi and Lai ([5]; BL), and Ginebra and Clayton ([12]; GC), in which the respective index formulas (shown in Table 1) have been modified as in Caro and Gallien [7] to account for delays.[1]

The simulation and the upper bound optimization codes are written in Matlab and are available from the authors upon request. Using a discount rate of $\delta = 0.95$, we run a series of simulations for five delay periods $\ell \in \{1, 2, 3, 4, 5\}$ for T periods such that $\sum_{t=T+1}^{\infty} \delta^t < 10^{-6}$ to approximate infinite horizon. We let our initial prior to be the uniform distribution, corresponding to the Beta distribution with parameter $(\alpha, \beta) = (1, 1)$, as it best represents the initial state of knowledge. We did an extensive simulation study and here we show the results for a few representative instances.

The simulation results where the decision maker pulls 4 arms out of a total of 32 arms (i.e., $(S, N) = (32, 4)$) are shown in Table 2. The first observation is that MPI index policy is near optimal since the suboptimality gap is very small. In general, it

[1] The index-specific coefficients of the CG and GC formulas were obtained through least squares using a small sample of exact MPI values.

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

TABLE 3. Asymptotic Suboptimality Gap.

	(S,N):	(32,4)	(S,N):	(160, 20)	(S,N):	(320,40)
ℓ	MYO (%)	MPI (%)	MYO (%)	MPI (%)	MYO (%)	MPI (%)
1	7.96	**0.51**	8.65	**0.64**	8.78	**0.07**
2	8.80	**1.53**	7.64	**0.65**	7.72	**0.28**
3	8.98	**2.15**	4.75	**0.47**	4.62	**0.45**
4	7.23	**2.82**	3.97	**0.87**	3.74	**0.68**
5	6.34	**3.63**	4.75	**1.60**	4.77	**1.17**

Note: $(\alpha, \beta) = (1, 1)$, $\delta = 0.95$.

was less than 4% in all, of the simulations we ran, and in most cases, it was actually less than 2%. The gap has a slight tendency to increase with the length of the delay ℓ. This could suggest that the MPI policy becomes slightly worse. However, it could also be that the upper bound deteriorates with longer delays.

We also note that all the delay-incorporated closed-form index policies perform very close to the MPI policy and that the differences are not statistically significant. We attribute the performance similarity to the fact that all the values of the modified closed-form indexes provide good approximations of the MPIs. We do, however, find that the myopic policy performs significantly worse than all other policies. This is to be expected because the myopic policy ignores the delayed actions as well as the future benefits from learning.

Computing a large table of necessary MPIs often requires a high level of computational complexity. Our finding suggests that, in such cases, one should adjust the existing closed-form indexes and use the policy as a substitute for the MPI policy and attain comparable results.

Furthermore, we find that as the number of projects S and the number of allowable pulls N increase while maintaining a constant ratio N/S, the suboptimality gap of the MPI policy approaches zero. The suboptimality gaps for $(S, N) = (32, 4)$, $(S, N) = (160, 20)$, and $(S, N) = (320, 40)$ are shown in Table 3. Whittle [24] initially conjectured that the MPI index policy is asymptotically optimal. This was largely validated by Weber and Weiss [22] for finite-state restless bandits. Our results support the conjecture for infinite-state bandits with response delay.

5. CONCLUSION

In this article, we prove the indexability of the multiarmed bandit problem with response delay, where the delays are of arbitrary length and are allowed to be stationary random as long as the delayed responses do no cross over. We show that, under the stationarity assumption, the problem is not indexable if the order is allowed to cross over. The MPI policy performs near-optimally, and the closed-form index policies when adjusted for delay represent good estimations of the MPI and perform well.

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

Further refinements of these policies are worth studying. For example, Kaplan [16] formulated a stochastic lead-time process in which the delays for each period are identically distributed but statistically dependent random variables so that orders do not cross over. It would be worthwhile to examine whether our results hold for nonstationary random delays. Another interesting variation is to make the pulls *irrevocable*; that is, once an arm stops being pulled, it can never be pulled again. This can be a desirable property from a practical standpoint and the results available for the classical bandit problem show a high performance that might extend to the case with response delays (see Farias and Madan [11]).

Acknowledgment

The authors would like to thank Steve Lippman and Charles Corbett for their helpful comments. We would also like to thank Georgios Georgiadis for organizing the index tables. Finally, we would like to thank one anonymous referee for his/her constructive suggestions, which helped improve the quality of the article considerably.

References

1. Altman, E. & Stidham, S., (1995). Optimality of monotonic policies for two-action Markovian decision processes, with applications to control of queues with delayed information source. *Queueing Systems* 21(3-4): 267–291.
2. Aviv, Y., Pazgal, A. (2002). Pricing of short life-cycle products through active learning. Working paper, Washington University, St. Louis, MO.
3. Bernardo, A. Chowdhry, B. (2002). Resources, real options, and corporate strategy. *Journal of Financial Economics* 63: 211–234.
4. Bertsimas, D. & Mersereau, A.J. (2007). A learning approach for interactive marketing to a customer segment. *Operations Reserach* 55(6): 1120–1135.
5. Brezzi, M. & Lai, T.L. (2002). Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control* 27: 87–108.
6. Burnetas, A.N. & Katehakis, M.N. (2003). Asymptotic Bayes analysis for the finite-horizon one-armed-bandit problem. *Probability in the Engineering and Informational Science* 17:53–83.
7. Caro, F. & Gallien, J. (2007). Dynamic sssortment with demand learning for seasonal consumer goods. *Management Science*, 53(2): 276–292.
8. DeGroot, M.H. (1970). *Optimal statistical decisions.* New York: McGraw-Hill.
9. Ehsan, N. & Liu, M. (2004). On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services. In *Proceedings INFOCOM 2004*, Vol. 3, pp. 1974–1983.
10. Eick, S.G. (1988). Gittins procedures for bandits with delayed responses. *Journal of the Royal Statistics Society B* 50(1): 125–132.
11. Farias, V. & Madan, R. (2008). Irrevocable multi-armed bandit policies. Working paper. MIT Sloan School of Management.
12. Ginebra, J. & Clayton, M. K. (1995). Response surface bandits. *Journal of the Royal Statistical Society B* 57: 771–784.
13. Gittins, J.C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society B* 14: 148–167.
14. Gittins, J.C. (1989). Multi-armed bandit allocation indices. Chichester, UK: John Wiley.
15. Hardwick, J., Oehmke, R., Stout, Q.F. (2006). New adaptive designs for delayed response models. *Journal Sequential Planning Inference* 136: 1940–1955.
16. Kaplan, R.S. (1970). A dynamic inventory model with stochastic lead times. *Management Science* 16(7): 491–507.

17. Muharremoglu, A. & Yang, N. (2008). Inventory management with an exogenous supply process. Working Paper, Columbia University New York.
18. Niño-Mora, J. (2006). Dynamic priority allocation via restless bandit marginal productivity indices. *Top* 15: 161–198.
19. Niño-Mora, J. (2007). Marginal productivity index policies for scheduling multiclass delay-/loss-sensitive traffic with delayed state observation. In *NGI 2007, Proceedings of the 3rd EuroNGI conference on next generation Internet networks: design and engineering for heterogeneity*. Piscataway, NJ:IEEE, pp. 209–217.
20. Robinson, L.W., Bradley, J.R., & Thomas, L.J. (2001). Consequences of order crossover under order-up-to inventory policies. *Manufacturing & Service Operations Management* 3(3): 175–188.
21. Wang, X. & Bickis, M. (2003). One-armed bandit models with continuous and delayed responses. *Mathematical Methods of Operations Research* 58: 209–219.
22. Weber, R.R. & Weiss, G. (1990). On an index policy for restless bandits. *Journal of Applied Probability* 27: 637–648.
23. Weiss, G. (1992). Turnpike optimality of Smith's rule in parallel machines stochastic scheduling. *Mathematics of Operations Research* 17(2): 255–270.
24. Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability* 25A: 287–298.

APPENDIX A.

Equivalence Relation

There are potentially four different ways of accounting for the Lagrange multiplier λ, which are given below and are summarized in Table A.1

$$\widehat{J}_t^{\lambda}(x,v^1,\ldots,v^{\ell}) = R(x)v^1 + \max\left\{\delta E_w \widehat{J}_{t-1}^{\lambda}(f(x,v^1,w),v^2,\ldots,v^{\ell},1),\right.$$
$$\left.\lambda + \delta E_w \widehat{J}_{t-1}^{\lambda}(f(x,v^1,w),v^2,\ldots,v^{\ell},0)\right\}.$$

$$J_t^{\lambda}(x,v^1,\ldots,v^{\ell}) = R(x)v^1 + \lambda(1-v^1) + \max\left\{\delta E_w J_{t-1}^{\lambda}(f(x,v^1,w),v^2,\ldots,v^{\ell},1),\right.$$
$$\left.\delta E_w J_{t-1}^{\lambda}(f(x,v^1,w),v^2,\ldots,v^{\ell},0)\right\}.$$

$$\widehat{H}_t^{\lambda}(x,v^1,\ldots,v^{\ell}) = R(x)v^1 + \max\left\{-\lambda + \delta E_w \widehat{H}_{t-1}^{\lambda}(f(x,v^1,w),v^2,\ldots,v^{\ell},1),\right.$$
$$\left.\delta E_w \widehat{H}_{t-1}^{\lambda}(f(x,v^1,w),v^2,\ldots,v^{\ell},0)\right\}.$$

$$H_t^{\lambda}(x,v^1,\ldots,v^{\ell}) = (R(x)-\lambda)v^1 + \max\left\{\delta E_w H_{t-1}^{\lambda}(f(x,v^1,w),v^2,\ldots,v^{\ell},1),\right.$$
$$\left.\delta E_w H_{t-1}^{\lambda}(f(x,v^1,w),v^2,\ldots,v^{\ell},0)\right\}.$$

PROOF OF PROPOSITION 1: Through induction on t it can be shown that $\Delta J_t^{\lambda}(z) = \Delta H_t^{\lambda}(z) = \Delta \widehat{J}_t^{\delta^{\ell}\lambda}(z) = \Delta \widehat{H}_t^{\delta^{\ell}\lambda}(z)$ for every (augmented) state z, where the operator Δ denotes the difference between the expected profits under the *Pull* and *NotPull* decisions. Clearly, if the difference is decreasing in λ for any of the formulations, then it is also decreasing for the other formulations. The proposition follows from this observation. ■

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

TABLE A.1. Four Different Representations

Lagrange Multiplier λ Accounted:	As Subsidy for Not Pulling	As Cost for Pulling
When decision made	$\widehat{J}_t^{\lambda}$	$\widehat{H}_t^{\lambda}$
When decision implemented	J_t^{λ}	H_t^{λ}

APPENDIX B.

Indexability of Random Delay Without Order Crossover

We first establish the following monotonicity result using the coupling argument as was done previously for Proposition 2.

PROPOSITION B.1: *For all states z, $\Delta J_t^{\lambda}(z)$ is monotonically decreasing in λ for stationary random delay ℓ if the orders do not cross over (i.e., $\ell \in \{m, m+1\}$).*

PROOF: We show that if $\ell \in \{m, m+1\}$, then the problem is indexable.

We show that for any $\lambda_1 > \lambda_2$, $\Delta J_t^{\lambda_1}(z) < \Delta J_t^{\lambda_2}(z)$ for all z, via induction. For simplicity of illustration, we will assume that $m = 0$—in other words, that the delay is uncertain between no delay and a delay of period 1. The structure of the proof remains identical for $m > 0$.

Base Case: $t = 1$: In the final decision period $t = 1$, there will be zero delay with probability p_0, and a delay of one period with probability p_1. We have

$$J_1^{\lambda}(x,1) = R(x) + \max\{p_0 E_{w_1} R(f(x,1,w_1)) + p_1 0, p_0\lambda + p_1 0\},$$
$$J_1^{\lambda}(x,0) = \lambda + \max\{p_0 R(x) + p_1 0, p_0\lambda + p_1 0\},$$
$$J_1^{\lambda}(x,\emptyset) = 0 + \max\{p_0 R(x) + p_1 0, p_0\lambda + p_1 0\},$$

and

$$\Delta J_1^{\lambda}(x,1) = p_0\{E_{w_1} R(f(x,1,w_1)) - \lambda\},$$
$$\Delta J_1^{\lambda}(x,0) = \Delta J_1^{\lambda}(x,\emptyset) = p_0\{R(x) - \lambda\}.$$

All are clearly decreasing in λ.

Induction Step: Suppose that $\forall z$, $\lambda_1 > \lambda_2$, $\Delta J_{t-1}^{\lambda_1}(z) < \Delta J_{t-1}^{\lambda_2}(z)$. We will show that $\forall z$, $\lambda_1 > \lambda_2$, $\Delta J_t^{\lambda_1}(z) < \Delta J_t^{\lambda_2}(z)$.

Again, there will be no delay with probability p_0 and a delay of one period with probability p_1. We have

$$\begin{aligned} J_t^{\lambda}(x,1) = R(x) + \max\{&p_0(E_{w_1} R(f(x,1,w_1)) + \delta E_{w_2} J_{t-1}^{\lambda}(f(f(x,1,w_1),1,w_2),\emptyset)) \\ &+ p_1(\delta J_{t-1}^{\lambda}(f(x,1,w_1),1)), p_0(\lambda + \delta E_{w_1} J_{t-1}^{\lambda}(f(x,1,w_1),\emptyset)) \\ &+ p_1(\delta E_{w_1} J_{t-1}^{\lambda}(f(x,1,w_1),0))\}, \end{aligned}$$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

$$J_t^\lambda(x,0) = \lambda + \max\{p_0(R(x) + \delta E_{w_1} J_{t-1}^\lambda(f(x,1,w_1),\emptyset)) + p_1(\delta J_{t-1}^\lambda(x,1)),$$
$$p_0(\lambda + \delta J_{t-1}^\lambda(x,\emptyset)) + p_1(\delta J_{t-1}^\lambda(x,0))\},$$
$$J_t^\lambda(x,\emptyset) = 0 + \max\{p_0(R(x) + \delta E_{w_1} J_{t-1}^\lambda(f(x,1,w_1),\emptyset)) + p_1(\delta J_{t-1}^\lambda(x,1)),$$
$$p_0(\lambda + \delta J_{t-1}^\lambda(x,\emptyset)) + p_1(\delta J_{t-1}^\lambda(x,0))\},$$

and

$$\Delta J_t^\lambda(x,1) = p_0(E_{w_1} R(f(x,1,w_1)) - \lambda) + p_0(\delta E_{w_2} J_{t-1}^\lambda(f(f(x,1,w_1),1,w_2),\emptyset)$$
$$- \delta E_{w_1} J_{t-1}^\lambda(f(x,1,w_1),\emptyset)) + p_1(\delta J_{t-1}^\lambda(f(x,1,w_1),1)$$
$$- \delta E_{w_1} J_{t-1}^\lambda(f(x,1,w_1),0)),$$
$$\Delta J_t^\lambda(x,0) = \Delta J_t^\lambda(x,\emptyset)$$
$$= p_0(R(x) - \lambda) + p_0(\delta E_{w_1} J_{t-1}^\lambda(f(x,1,w_1),\emptyset) - \delta J_{t-1}^\lambda(x,\emptyset)),$$
$$+ p_1(\delta J_{t-1}^\lambda(x,1) - \delta J_{t-1}^\lambda(x,0)).$$

For simplicity, we will examine the indexability for state $z = (x,0)$, but the identical argument holds for other states. We have

$$\mathrm{DIFF} \equiv \Delta J_t^{\lambda_1}(x,0) - \Delta J_t^{\lambda_2}(x,0)$$
$$= p_0 \left\{ (\lambda_2 - \lambda_1) + \delta E_{w_1} J_{t-1}^{\lambda_1}(f(x,1,w_1),\emptyset) - \delta J_{t-1}^{\lambda_1}(x,\emptyset) \right.$$
$$\left. -\delta E_{w_1'} J_{t-1}^{\lambda_2}(f(x,1,w_1'),\emptyset) + \delta J_{t-1}^{\lambda_2}(x,\emptyset) \right\}$$
$$+ p_1 \left\{ \delta J_{t-1}^{\lambda_1}(x,1) - \delta J_{t-1}^{\lambda_1}(x,0) - \delta J_{t-1}^{\lambda_2}(x,1) + \delta J_{t-1}^{\lambda_2}(x,0) \right\}.$$

After rewriting the expression and rearranging the terms, we have

$$\mathrm{DIFF} = p_0 \left\{ (\lambda_2 - \lambda_1) + \delta E_{w_1} J_{t-1}^{\lambda_1}(f(x,1,w_1),\emptyset)^{(NotPull)} - \delta J_{t-1}^{\lambda_1}(x,\emptyset)^{(Pull)} \right.$$
$$\left. -\delta E_{w_1'} J_{t-1}^{\lambda_2}(f(x,1,w_1'),\emptyset)^{(NotPull)} + \delta J_{t-1}^{\lambda_2}(x,\emptyset)^{(Pull)} \right\}$$
$$+ p_1 \left\{ \delta J_{t-1}^{\lambda_1}(x,1)^{(NotPull)} - \delta J_{t-1}^{\lambda_1}(x,0)^{(Pull)} \right.$$
$$\left. - \delta J_{t-1}^{\lambda_2}(x,1)^{(NotPull)} + \delta J_{t-1}^{\lambda_2}(x,0)^{(Pull)} \right\}$$
$$+ p_0 \left\{ \delta E_{w_1}[\Delta J_{t-1}^{\lambda_1}(f(x,1,w_1),\emptyset)]^+ - \delta E_{w_1'}[\Delta J_{t-1}^{\lambda_2}(f(x,1,w_1'),\emptyset)]^+ \right.$$
$$\left. -\delta[\Delta J_{t-1}^{\lambda_1}(x,\emptyset)]^- + \delta[\Delta J_{t-1}^{\lambda_2}(x,\emptyset)]^- \right\}$$
$$+ p_1 \left\{ \delta[\Delta J_{t-1}^{\lambda_1}(x,1)]^+ - \delta[\Delta J_{t-1}^{\lambda_2}(x,\emptyset)]^+ \right.$$
$$\left. - \delta[\Delta J_{t-1}^{\lambda_1}(x,0)]^- + \delta[\Delta J_{t-1}^{\lambda_2}(x,0)]^- \right\}.$$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

Eliminating the bottom two expressions, which are both nonpositive by the induction assumption, we have

$$
\begin{aligned}
\text{DIFF} \le\ & p_0 \Big\{(\lambda_2-\lambda_1)+\delta E_{w_1} J_{t-1}^{\lambda_1}(f(x,1,w_1),\emptyset)^{(NotPull)} - \delta J_{t-1}^{\lambda_1}(x,\emptyset)^{(Pull)} \\
& \quad -\delta E_{w_1'} J_{t-1}^{\lambda_2}(f(x,1,w_1'),\emptyset)^{(NotPull)} + \delta J_{t-1}^{\lambda_2}(x,\emptyset)^{(Pull)}\Big\} \\
& + p_1 \Big\{\delta J_{t-1}^{\lambda_1}(x,1)^{(NotPull)} - \delta J_{t-1}^{\lambda_1}(x,0)^{(Pull)} \\
& \quad - \delta J_{t-1}^{\lambda_2}(x,1)^{(NotPull)} + \delta J_{t-1}^{\lambda_2}(x,0)^{(Pull)}\Big\} \\
=\ & p_0 \Big\{(\lambda_2-\lambda_1)+p_0\Big\{\delta E_{w'}(\lambda_1+\delta J_{t-2}^{\lambda_1}(f(x,1,w_1),\emptyset)) - \delta(R(x) \\
& \quad + \delta E_{w_1} J_{t-2}^{\lambda_1}(f(x,1,w_1),\emptyset)) - \delta E_{w_1'}(\lambda_2+\delta J_{t-2}^{\lambda_2}(f(x,1,w_1'),\emptyset)) + \delta(R(x) \\
& \quad + \delta E_{w_1'} J_{t-2}^{\lambda_2}(f(x,1,w_1'),\emptyset)) + p_1\Big\{\delta^2 E_{w_1} J_{t-2}^{\lambda_1}(f(x,1,w_1),0) - \delta^2 J_{t-2}^{\lambda_1}(x,1) \\
& \quad -\delta^2 E_{w_1'} J_{t-2}^{\lambda_2}(f(x,1,w_1'),0) + \delta^2 J_{t-2}^{\lambda_2}(x,1)\Big\}\Big\}\Big\} \\
& + p_1\Big\{p_0\Big\{\delta(R(x)+\lambda_1+\delta E_{w_1} J_{t-2}^{\lambda_1}(f(x,1,w_1),\emptyset)) - \delta(\lambda_1+R(x) \\
& \quad + E_{w_1} J_{t-2}^{\lambda_1}(f(x,1,w_1),\emptyset)) - \delta(R(x)+\lambda_2+\delta E_{w_1'} J_{t-2}^{\lambda_2}(f(x,1,w_1'),\emptyset)) \\
& \quad + \delta(\lambda_2+R(x)+E_{w_1'} J_{t-2}^{\lambda_2}(f(x,1,w_1'),\emptyset)) \\
& + p_1\Big\{\delta(R(x)+\delta E_{w_1} J_{t-2}^{\lambda_1}(f(x,1,w_1),0)) - \delta(\lambda_1+\delta J_{t-2}^{\lambda_1}(x,1)) \\
& \quad -\delta(R(x)+\delta E_{w_1'} J_{t-2}^{\lambda_2}(f(x,1,w_1'),0)) + \delta(\lambda_2+\delta J_{t-2}^{\lambda_2}(x,1))\Big\}\Big\}\Big\} \\
=\ & p_0(\lambda_2-\lambda_1)+\delta(\lambda_1-\lambda_2)(p_0^2-p_1^2) \\
& + p_1\Big\{\delta^2 E_{w_1} J_{t-2}^{\lambda_1}(f(x,1,w_1),0) - \delta^2 J_{t-2}^{\lambda_1}(x,1) \\
& \quad - \delta^2 E_{w_1'} J_{t-2}^{\lambda_2}(f(x,1,w_1'),0) + \delta^2 J_{t-2}^{\lambda_2}(x,1)\Big\}.
\end{aligned}
$$

We now introduce the coupling argument. Consider the bandit with subsidy λ_1 starting from two different states. The first, which we refer to System A, starts from the augmented state $(f(x,1,w_1),0)$ at time $t-2$. The second, which we refer to as System B, starts from state $(x,1)$ but implements the same decisions as System A in the first stage, and after that, it follows its own optimal policy. Let π^* denote the optimal policy of System A.

Let $G_{\pi_{t-2}^*}^{\lambda_1}(z)$ represent the value of being in state z for λ_1 at time $t-2$ and following the policy π^*. Then we have

$$
J_{t-2}^{\lambda_1}(f(x,1,w_1),0) = G_{\pi_{t-2}^*}^{\lambda_1}(f(x,1,w_1),0) \quad \text{and} \quad J_{t-2}^{\lambda_1}(x,1) \ge G_{\pi_{t-2}^*}^{\lambda_1}(x,1).
$$

The same coupling argument can be used for the bandit with subsidy λ_2, only that System A would start in state $(x,1)$ and System B would start in $(f(x,1,w_1'),0)$. Denoting

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

π^{**} as the optimal policy of System A, we have

$$J^{\lambda_2}_{t-2}(x,1) = G^{\lambda_2}_{\pi^{**}_{t-2}}(x,1) \quad \text{and} \quad J^{\lambda_2}_{t-2}(f(x,1,w_1'),0) \geq G^{\lambda_2}_{\pi^{**}_{t-2}}(f(x,1,w_1'),0).$$

By subtracting these smaller values of $G^{\lambda_1}_{\pi^{*}_{t-2}}$ and $G^{\lambda_2}_{\pi^{**}_{t-2}}$, the right-hand side of the inequality, and therefore DIFF, is bounded above by

$$\begin{aligned}
\text{DIFF} \leq\ & p_0(\lambda_2-\lambda_1) + \delta(\lambda_1-\lambda_2)(p_0^2-p_1^2) \\
& + p_1\Big\{\delta^2 E_{w_1} G^{\lambda_1}_{\pi^{*}_{t-2}}(f(x,1,w_1),0) - \delta^2 G^{\lambda_1}_{\pi^{*}_{t-2}}(x,1) \qquad \text{(DIFF.1)} \\
& \quad + \delta^2 G^{\lambda_2}_{\pi^{**}_{t-2}}(x,1) - \delta^2 E_{w_1'} G^{\lambda_2}_{\pi^{**}_{t-2}}(f(x,1,w_1'),0)\Big\}. \qquad \text{(DIFF.2)}
\end{aligned}$$

We now elaborate the expressions (DIFF.1) and (DIFF.2):

$$\begin{aligned}
\text{(DIFF.1)} = p_1\Big\{&\delta^2 E_{w_1}\Big\{p_0(\lambda_1 + R(f(x,1,w_1)))u^* + \lambda_1(1-u^*) \\
&\quad + \delta E_{w_2} G^{\lambda_1}_{\pi^{*}_{t-3}}(f(f(x,1,w_1),u^*,w_2),\emptyset)) \\
&\quad + p_1(\lambda_1 + \delta E_{w_1} G^{\lambda_1}_{\pi^{*}_{t-3}}(f(x,1,w_1),u^*))\Big\} \\
&- \delta^2\Big\{p_0(R(x) + E_{w_1}R(f(x,1,w_1)))u^* + \lambda_1(1-u^*) \\
&\quad + \delta E_{w_2} G^{\lambda_1}_{\pi^{*}_{t-3}}(f(f(x,1,w_1),u^*,w_2),\emptyset)) \\
&\quad + p_1(R(x) + E_{w_1} G^{\lambda_1}_{\pi^{*}_{t-3}}(f(x,1,w_1),u^*))\Big\}\Big\} \\
= p_1\{&\delta^2(\lambda_1 - R(x))\}.
\end{aligned}$$

Following the same argument, we have

$$\text{(DIFF.2)} = p_1\{\delta^2(R(x)-\lambda_2)\}.$$

Thus, after summing the expressions, we have

$$\begin{aligned}
\text{DIFF} &= p_0(\lambda_2-\lambda_1) + \delta(\lambda_1-\lambda_2)(p_0^2-p_1^2) + p_1\delta^2(\lambda_1-\lambda_2) \\
&= (\lambda_2-\lambda_1)(p_0 - \delta p_0^2 + \delta p_1^2 - \delta^2 p_1) \\
&= (\lambda_2-\lambda_1)(p_0(1-\delta p_0) + \delta p_1(p_1-\delta)) < 0.
\end{aligned}$$

Notice that if $p_0 = 1$ and $p_1 = 0$, or $p_0 = 0$ and $p_1 = 1$, the above inequality reduces to respectively

$$(\lambda_2-\lambda_1)(1-\delta) < 0 \quad \text{and} \quad (\lambda_2-\lambda_1)\delta(1-\delta) < 0,$$

which is consistent with the result of Proposition 2. ■

PROPOSITION B.2: *If the delayed responses are allowed to cross over, then* $\Delta J^{\lambda}_t(z)$ *is not necessarily monotonically decreasing.*

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

PROOF: We provide an example of a range of λ's in which $\Delta J_t^{\lambda}(z)$ is increasing when the delayed responses are allowed to cross over. In particular, consider $\ell \in \{0, 2\}$, $z = (x, 0, \emptyset)$ at time $t = 4$, and $J_0^{\lambda}(\cdot) = 0$. We have

$$J_4^{\lambda}(x, 0, \emptyset) = \lambda + \max\{\delta J_3^{\lambda}(x, \emptyset, 1), \delta J_3^{\lambda}(x, \emptyset, 0)\},$$
$$\Delta J_4^{\lambda}(x, 0, \emptyset) = \delta\{J_3^{\lambda}(x, \emptyset, 1) - J_3^{\lambda}(x, \emptyset, 0)\}.$$

We elaborate the necessary $J_3^{\lambda}(\cdot)$'s and $J_2(\cdot)'s$:

$$J_3^{\lambda}(x, \emptyset, 1) = \max\{R(x) + \delta E_w J_2^{\lambda}(f(x, 1, w), 1, \emptyset), \lambda + \delta J_2^{\lambda}(x, 1, \emptyset)\},$$
$$J_3^{\lambda}(x, \emptyset, 0) = \max\{R(x) + \delta E_w J_2^{\lambda}(f(x, 1, w), 0, \emptyset), \lambda + \delta J_2^{\lambda}(x, 0, \emptyset)\},$$
$$\delta E_w J_2^{\lambda}(f(x, 1, w), 1, \emptyset) = \delta E_w R(f(x, 1, w)) + \delta^2 E_w E_{w'} \max\{R(f(f(x, 1, w), 1, w')), \lambda\},$$
$$\delta J_2^{\lambda}(x, 1, \emptyset) = \delta R(x) + \delta^2 E_w \max\{R(f(x, 1, w)), \lambda\},$$
$$\delta E_w J_2^{\lambda}(f(x, 1, w), 0, \emptyset) = \delta\lambda + \delta^2 E_w \max\{R(f(x, 1, w)), \lambda\},$$
$$\delta J_2^{\lambda}(x, 0, \emptyset) = \delta\lambda + \delta^2 \max\{R(x), \lambda\}.$$

We consider the following independent binary random process as shown in Figure A.1. Let $R(x) = x$, and for simplicity, let us take $x = 1$ and let $\lambda_1 = 1.67$, and $\lambda_2 = 1.55$, with $\delta = 0.9$. Substituting these values into the above expression, we have

$$J_3^{\lambda_1}(x, \emptyset, 1) = 4.597, \quad J_3^{\lambda_1}(x, \emptyset, 0) = 4.530, \quad J_3^{\lambda_2}(x, \emptyset, 1) = 4.390, \quad J_3^{\lambda_2}(x, \emptyset, 0) = 4.335,$$

giving us

$$\Delta J_4^{\lambda_1}(x, 0, \emptyset) = \delta\left\{J_3^{\lambda_1}(x, \emptyset, 1) - J_3^{\lambda_1}(x, \emptyset, 0)\right\} = 0.9(0.067) = 0.0603,$$
$$\Delta J_4^{\lambda_2}(x, 0, \emptyset) = \delta\left\{J_3^{\lambda_2}(x, \emptyset, 1) - J_3^{\lambda_2}(x, \emptyset, 0)\right\} = 0.9(0.055) = 0.0495.$$

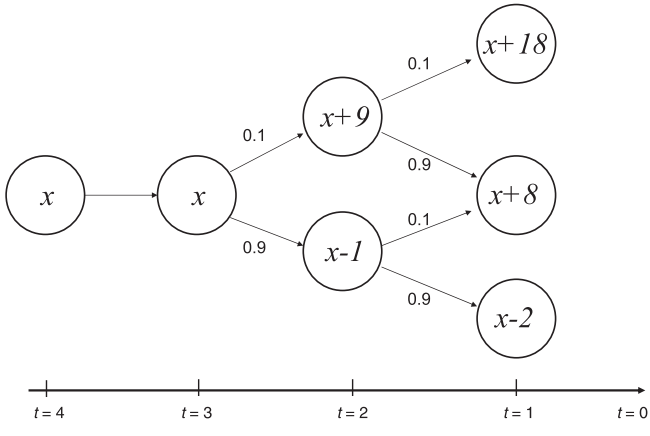

FIGURE A.1. Independent binary random process.

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

In other words, although $\lambda_1 > \lambda_2$, we have $\Delta J_4^{\lambda_1}(x, 0, \emptyset) > \Delta J_4^{\lambda_2}(x, 0, \emptyset)$, which implies that $\Delta J_4^{\lambda}(x, 0, \emptyset)$ is increasing in this interval. ■

PROOF OF THEOREM 2: The theorem is clear by following the outline of the proof of Theorem 1 and using the results of Proposition B.1 and B.2. ■

APPENDIX C.

Upper-Bound Formulation

So far, we have focused on formulating a heuristic because the dynamic programming formulation that defines the optimal policy is intractable. In this section, we formulate a tractable Lagrangian upper bound of the problem by decoupling the dynamic program into S independent arms. The upper bound enables us to provide a suboptimality guarantee of the resulting index policy.

PROPOSITION C.1: *Define the following function:*

$$L_t^{\lambda}(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{\ell}) = N\lambda + \max_{\boldsymbol{u}} \left\{ \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 + \delta E_{\boldsymbol{w}} L_{t-1}^{\lambda}(f(\boldsymbol{x}, \boldsymbol{v}^1, \boldsymbol{w}), \boldsymbol{v}^2, \ldots, \boldsymbol{v}^{\ell}, \boldsymbol{u}) \right\}$$

$$\text{s.t.} \quad \boldsymbol{u} \in \{0, 1\}^S,$$

for $t > \ell$, and for $t \leq \ell$,

$$L_t^{\lambda}(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{\ell}) = N\lambda + \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 + \delta E_{\boldsymbol{w}} L_{t-1}^{\lambda}(f(\boldsymbol{x}, \boldsymbol{v}^1, \boldsymbol{w}), \boldsymbol{v}^2, \ldots, \boldsymbol{v}^{\ell})\}, \quad L_0^{\lambda}(\cdot) = 0,$$

where the λ represents the cost when the arm is actually pulled*, and*

$$L_t^*(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{\ell}) = \min_{\lambda} L_t^{\lambda}(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{\ell}).$$

Then

$$J_t^*(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{\ell}) \leq L_t^*(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{\ell}) \leq L_t^{\lambda}(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{\ell}).$$

PROOF: We prove by induction. For $t \leq l$, it is clear that the following holds given that $\sum_{s=1}^{S} v^t \leq N, \forall t \leq \ell$.

$$\begin{aligned} J_1^*(\mathbf{x}, \mathbf{v}^1) &= \sum_{s=1}^{S} R_s(x_s) v_s^1 \\ &\leq N\lambda + \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 \\ &= L_1^{\lambda}(\mathbf{x}, \mathbf{v}^1) \end{aligned}$$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

and

$$J_t^*(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^t) = \sum_{s=1}^{S} R_s(x_s) v_s^1 + \delta \mathrm{E}_{\mathbf{w}} J_{t-1}^*(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^t)$$
$$\leq N\lambda + \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 + \delta \mathrm{E}_{\mathbf{w}} L_{t-1}^{\lambda}(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^t)$$
$$= L_t^{\lambda}(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^t).$$

For $t \geq \ell + 1$, suppose $J_{t-1}^* < H_{t-1}^{\lambda}$. Then

$$J_t^*(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^\ell) = \max_{\mathbf{u} \in \{0,1\}^S : \sum_{s=1}^{S} u_s \leq N} \left\{ \sum_{s=1}^{S} R_s(x_s) v_s^1 + \delta \mathrm{E}_{\mathbf{w}} J_{t-1}^*(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^\ell, \mathbf{u}) \right\}$$
$$\leq \max_{\mathbf{u} \in \{0,1\}^S : \sum_{s=1}^{S} u_s \leq N} \left\{ N\lambda + \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 \right.$$
$$\left. + \delta \mathrm{E}_{\mathbf{w}} J_{t-1}^*(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^\ell, \mathbf{u}) \right\}$$
$$\leq \max_{\mathbf{u} \in \{0,1\}^S : \sum_{s=1}^{S} u_s \leq N} \left\{ N\lambda + \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 \right.$$
$$\left. + \delta \mathrm{E}_{\mathbf{w}} L_{t-1}^{\lambda}(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^\ell, \mathbf{u}) \right\}$$
$$\leq \max_{\mathbf{u} \in \{0,1\}^S} \left\{ N\lambda + \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 + \delta \mathrm{E}_{\mathbf{w}} L_{t-1}^{\lambda}(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^\ell, \mathbf{u}) \right\}$$
$$= L_t^{\lambda}(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^\ell).$$

The first inequality follows because $\sum_{s=1}^{S} v_s^1 \leq N$, and the second inequality holds because of the induction assumption. The final inequality is because it is an optimization problem defined over a larger set. ■

We next show that the above expression for $L_t^{\lambda}(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^\ell)$ can be formulated more simply in terms of single-arm problems. Such similar decomposition has been shown previously without delay (see Caro and Gallien [7] and Bertsimas and Mersereau [4]).

PROPOSITION C.2:

(i) *For* $t \leq \ell$,

$$L_t^{\lambda}(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^t) = N\lambda \sum_{\tau=1}^{t} \delta^{\tau-1} + \sum_{s=1}^{S} L_{t,s}^{\lambda}(x_s, v_s^1, v_s^2, \ldots, v_s^t),$$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

where

$$L^{\lambda}_{t,s}(x_s, v^1_s, \ldots, v^t_s) = (R(x_s) - \lambda)v^1_s + \delta E_w L^{\lambda}_{t-1,s}(f(x, v^1, w), v^2, \ldots, v^t).$$

(ii) *For* $t \geq \ell + 1$,

$$L^{\lambda}_t(\boldsymbol{x}, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^t) = N\lambda \sum_{\tau=1}^{t} \delta^{\tau-1} + \sum_{s=1}^{S} L^{\lambda}_{t,s}(x_s, v^1_s, \ldots, v^t_s),$$

where

$$L^{\lambda}_{t,s}(x_s, v^1_s, \ldots, v^t_s) = \max\Big\{(R(x_s) - \lambda)v^1_s + \delta E_w L^{\lambda}_{t-1,s}(f(x, v^1, w), v^2, \ldots, v^t, 1),$$
$$(R(x_s) - \lambda)v^1_s + \delta E_w L^{\lambda}_{t-1,s}(f(x, v^1, w), v^2, \ldots, v^t, 0)\Big\}.$$

PROOF: We prove by induction.

(i) For delay of ℓ, $t = \ell + 1$ is where we will make the final decision. Hence, $L^{\lambda}_{\ell}(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^{\ell})$ can be considered a constant. where the decisions $(\mathbf{v}^1, \ldots, \mathbf{v}^{\ell})$ get carried out. We first evaluate the quantity for $t \leq \ell$. $L^{\lambda}_0(\cdot) = 0$. First by Proposition C.1, we have

$$L^{\lambda}_1(\mathbf{x}, \mathbf{v}^1) = N\lambda + \sum_{s=1}^{S}(R_s(x_s) - \lambda)v^1_s + \delta E_w\{L^{\lambda}_0(\cdot)\}$$
$$= N\lambda + \sum_{s=1}^{S} L^{\lambda}_{1,s}(x_s, v^1_s).$$

Then if we assume the expression holds for L^{λ}_{t-1}, we have

$$L^{\lambda}_t(\mathbf{x}, \mathbf{v}^1, \ldots, \mathbf{v}^t) = N\lambda + \sum_{s=1}^{S}(R_s(x_s) - \lambda)v^1_s + \delta E_{\mathbf{w}}\{L^{\lambda}_{t-1}(f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2)\}$$
$$= N\lambda + \sum_{s=1}^{S}(R_s(x_s) - \lambda)v^1_s + \delta E_{\mathbf{w}}\left\{N\lambda \sum_{\tau=1}^{t-1} \delta^{\tau-1} + \sum_{s=1}^{S} L^{\lambda}_{t-1,s}(f(x_s, v^1_s, w_s), v^2_s, \ldots, v^t_s)\right\}$$
$$= N\lambda \sum_{\tau=1}^{t} \delta^{\tau-1} + \sum_{s=1}^{S}\{(R_s(x_s) - \lambda)v^1_s + \delta E_{w_s} L^{\lambda}_{t-1,s}(f(x_s, v^1_s, w_s), v^2_s, \ldots, v^t_s)\}$$
$$= N\lambda \sum_{\tau=1}^{t} \delta^{\tau-1} + \sum_{s=1}^{S} L^{\lambda}_{t,s}(x_s, v^1_s, \ldots, v^t_s).$$

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press

(ii) Now suppose $t \geq \ell + 1$, and the expression holds for $t-1$. Then, again from Proposition C.1, we have the following expression:

$$\begin{aligned} L_t^{\lambda}(\mathbf{x}, \mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}_s^t) &= N\lambda + \max_{\mathbf{u}} \left\{ \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 + \delta E_{\mathbf{w}} L_{t-1}^{\lambda} \right. \\ &\qquad \left. \times \ (f(\mathbf{x}, \mathbf{v}^1, \mathbf{w}), \mathbf{v}^2, \ldots, \mathbf{v}^l, \mathbf{u}) \right\} \\ &= N\lambda + \max_{\mathbf{u}} \left\{ \sum_{s=1}^{S} (R_s(x_s) - \lambda) v_s^1 \right. \\ &\quad \left. + \delta E_{\mathbf{w}} \left\{ N\lambda \sum_{\tau=1}^{t-1} \delta^{\tau-1} + \sum_{s=1}^{S} H_{t-1,s}^{\lambda}(f(x_s, v_s^1, w), v_s^2, \ldots, v_s^t, u_s) \right\} \right\} \\ &= N\lambda \sum_{\tau=1}^{t} \delta^{\tau-1} + \sum_{s=1}^{S} \max_{\hat{u}} \left\{ (R_s(x_s) - \lambda) v_s^1 \right. \\ &\qquad \left. + \delta E_{\mathbf{w}} L_{t-1,s}^{\lambda}(f(x_s, v_s^1, w), v_s^2, \ldots, v_s^t) \right\} \\ &= N\lambda \sum_{\tau=1}^{t} \delta^{\tau-1} + \sum_{s=1}^{S} L_{t,s}^{\lambda}(x_s, v_s^1, v_s^2, \ldots, v_s^t). \end{aligned}$$

■

https://doi.org/10.1017/S0269964810000021 Published online by Cambridge University Press