Searching for Arms*

Daniel Fershtman† **Alessandro Pavan**‡

July 21, 2019

PRELIMINARY AND INCOMPLETE

Abstract

We study a model of experimentation or sequential learning where the set of alternatives is endogenously shaped by past search decisions. The environment reflects a tradeoff between exploration of existing alternatives and search for additional ones, faced with uncertainty about the set of outside alternatives that may be discovered. Despite the effect of search on the future set of alternatives and the potential correlation between the types of alternatives found over time, we obtain a simple characterization of the optimal policy. We extend our analysis to environments in which an irreversible choice may be made among the endogenous set of alternatives, at any time and given any amount of potentially inconclusive information. As a special case, the analysis yields a generalization of Weitzman's (1979) canonical Pandora's boxes problem, as well as its solution, to an environment where the set of boxes is endogenous.

*For useful comments and suggestions we thank seminar participants at various conferences and workshops where the paper was presented. Fershtman gratefully acknowledges financial support from the German Research Foundation through CRC TR 224.

†Eitan Berglas School of Economics, Tel-Aviv University, and Institute for Microeconomics, University of Bonn. Email: danielfer@tauex.tau.ac.il

‡Department of Economics, Northwestern University. Email: alepavan@northwestern.edu

1 Introduction

Classical search or experimentation problems involve a decision maker exploring a fixed set of options with unknown returns. In many decision problems, however, the set of options is not fixed ex-ante, and the decision maker may choose to seek out additional options as part of the exploration process. This paper studies the question of how to balance exploration among existing options and search for additional options to explore, faced with uncertainty not only about the returns from the options already available, but also about the options that may be found as a result of further search.

We believe this tradeoff is present in many economic environments involving experimentation, sequential learning, or search among alternatives. For example, consider a firm seeking to hire a candidate for a position. The firm begins to interview candidates among an initial pool, but at any point may decide to expand the pool of candidates by searching for additional prospects. Such search is likely to be costly and time consuming, and could delay the final decision, but may yield candidates better suited for the position. A researcher works on a number of ongoing projects with unknown return, but may also exert time and effort searching for additional potential projects to begin pursuing. The tradeoff between the two evolves over time based on the development of existing work and the researcher's ability to generate ideas. A consumer typically balances exploration among different alternatives within her "consideration set" and expansion of this set through search for other products (e.g., by visiting a new store or website). Similarly, a policy maker chooses which policy to promote among a set of available policies with uncertain returns, buy may also search for alternative policies to examine. A platform matches existing buyers and sellers over time, soliciting additional buyers/sellers in response to past outcomes. In all of these examples, the set of options being explored is not fixed, but endogenous to the decision maker's problem. It is a consequence of past deliberate decisions to search for options.

To study this tradeoff, we develop a new model of experimentation or sequential learning among an endogenous set of alternatives. We augment the classical multi-armed bandit problem allowing the set of "arms" to be endogenously determined by past search decisions. In each period, the decision maker can either pull a single arm among the set of existing arms, or search for new arms. As in the classical bandit problem, pulling an arm results in a stochastic reward, a function of the arm's current state, and triggers the transition of the arm to a new state, while the states corresponding to the other arms remain unchanged. However, in contrast to the classical bandit problem, where the set of arms is fixed ex-ante, the set of arms available in each period is endogenously shaped by past search decisions. Search for new arms is costly, and yields a stochastic set of new arms that are added to the pool of arms which can be subsequently pulled. The search technology itself – i.e., the cost and distribution over the set of arms that are discovered – may change over time as a function of past search outcomes.

The framework is flexible and can accommodate various models of experimentation and sequential learning, with an endogenous pool of alternatives. Despite the complication stemming from the fact that search competes with the same arms brought in by past searches, and the potential correlation between the types of arms found over time, we obtain a simple characterization of the optimal policy. At each period, each physical arm is assigned an index identical to the one in the classical bandit problem (Gittins and Jones, 1974), irrespective of the time at which it was found and independently of the information pertaining the state of the search technology and of any other arm. Search is also assigned an independent index, but the latter must account for the fact that (a) the "rewards" from search are not incurred until the arms discovered through search are used, and (b) the fact that the activation of the arms that search will bring must itself be disciplined by an endogenous rule that is part of the index's definition. We show that, despite these complications, the new index for the search technology, which is independent of information about the other existing arms, admits a convenient recursive representation that favors its computation in applications. The optimal policy then consists of choosing in each period the arm with the highest index, including search. In the special case in which search is degenerate (i.e., the set of arms is fixed ex-ante), the model and the optimal policy coincide with the classical bandit problem.

To illustrate why the problem differs from the classical bandit problem, note that problems in which alternatives take the form of "meta arms", i.e., arms that correspond to their own sub-decision problem typically do not admit an index solution, even if each sub-problem is independent from the others, and even if one knows the solution to each independent sub-problem. In the same vein, dependence or correlation between arms typically precludes an index solution. Even if a subset of dependent arms evolves independently of all other arms, and even if one knows the optimal way of choosing among the dependent arms in each subset in isolation from the other arms, generally, it is impossible to assign an index to each subset of dependent arms and utilize an index solution for the "overall" problem. Before describing the optimal policy for our search model, we illustrate these difficulties in an example.

After characterizing the optimal policy, we show how it can be used to shed light on the dynamics that arise as a function of the evolution of the search technology. When the quality of the search technology improves with the number of past searches, or is stationary, dynamics under the optimal policy take the form of "replacement". Whenever search is launched, the decision maker never returns to any of the arms that were available at the time search was launched. The optimality of replacement policies under (weakly) improving technologies, while perhaps anticipated, hinges on the optimal policy taking an index form. In particular, the latter property implies that the *composition* of the set of available arms at the time search is launched is irrelevant for the decision to stop exploring the existing arms and instead search for new arms. This is despite the fact that if the decision maker were to continue with the existing arms, her continuation value would depend on the *entire set's* composition. Importantly, however, if the search technology deteriorates over time (as is the case, for example, when the number or arms that can be found is limited), new arms found through search may be put on hold and returned to at a later stage, after additional searches are conducted. That is, if the search technology is not (weakly) improving, search does not correspond to replacement of existing arms.

We extend our analysis to a class of environments in which, in addition to searching and exploring the available arms, the decision maker must decide if, and when, to *irreversibly commit* to one of the available alternatives. Without further assumptions, the irreversibility of choice may preclude an index solution. However, we identify a condition that guarantees an index policy remains optimal. A special version of this problem that readily satisfies this condition is a generalization of Weitzman's (1979) canonical Pandora's boxes problem to an environment in which the set of boxes is endogenous; that is, besides opening boxes and choosing among them, the agent may also choose to search for additional boxes. Our solution extends Weitzman's and coincides with his solution in the special case in which the set of arms is fixed ex-ante.

We believe many of the economic environments to which the classical bandit problem has been applied – or, similarly, Pandora's boxes problem has been applied – naturally feature the possibility to expand the set of alternatives over time as part of the exploration process. In Section 6, we discuss a number of problems to which our model can be applied, including the design of search engines, solicitation and marketing in dynamic allocation problems, and two-sided markets.

The rest of the paper is organized as follows. The remainder of this section discusses the related literature. Section 2 describes the model. Section 3 presents a cautionary example, characterizes the optimal policy, and discusses its implications for dynamics as a function of the search technology. Section 4 extends the analysis to a class of problems with irreversible choice. Further extensions are discussed in Section 5. Section 6 illustrates several applications, while Section 7 concludes. Some more technical steps of the proofs are relegated to an Appendix at the end of the document.

Related literature. In the classical multi-armed bandit problem (Robbins 1952), a decision maker sequentially pulls a single arm out of an exogenously fixed set, with the goal of maximizing the expected discounted sum of the rewards. The arms' states evolve independently. When an arm is pulled, it yields a reward as a function of its state and its state evolves, while the states of all other arms remain "frozen". In a celebrated result, Gittins and Jones (1974) show that the optimal policy in this problem takes the form of an "index policy", with each arm assigned an index which is a function only of its own state, and with the arm with the highest index pulled at each stage.

First introduced into economics by Rothschild (1974), the multi-armed bandit model has been widely applied to a variety of problems. Versions of it have become workhorse frameworks in the study of experimentation, with applications ranging from labor markets, political economy, dynamic pricing, auctions, corporate finance, and asset pricing (for excellent surveys, see Bergemann and Valimaki (2008) and Horner and Skrzypacz (2017)). Besides experimentation, a growing literature studies optimal sequential learning about several options when attention must focus on one item at a time, before choosing one of the options. Recent contributions include Ke, Shen, and Villas-Boas (2016), Austen-Smith and Martinelli (2018), Ke and Villas-Boas (2019), and Gossner, Steiner and Stewart (2019), each of which assumes different costs and information structures, as well as different assumptions on the termination of learning. In contrast to our work, in all of these papers the set of alternatives is fixed ex-ante.

A few extensions of the canonical bandit problem in which arms arrive over time have been explored in the statistics and operations research literature, but thus far such extensions have received scarce attention in economics. Whittle (1981) shows that the optimality of an index policy extends to a setting in which new arms arrive over time, provided the arms arrive according to an exogenous process with i.i.d arrivals, independently of past decisions (see also Varaiya et al., 1985). More closely related to our framework is Weiss (1988)'s "branching bandits problem," in which an arm, when pulled, yields a reward and is replaced by a new set of arms (see also Keller and Oldale (2003) for a variation of this problem). Our problem differs from this branching problem in two dimensions. First, the state of the search arm is allowed to contain information about the results of past searches and is potentially unbounded, which is essential, for example, to accommodate for the possibility that search may last forever, as is the case for stationary environments. Second, our proof relies on a recursive characterization of the search index, whereas these works establish indexability but do not provide a characterization of the indexes.

As mentioned above, a special case of our results in Section 4 yields an extension of Pandora's boxes problem, first introduced by Weitzman (1979), to an environment in which the set of boxes is endogenous. Weitzman shows that, when the set of boxes is exogenous, the solution takes the following form: Each box is assigned a "reservation price", which is a function of only the distribution over the box's prize and the cost of opening it. Boxes are opened in descending order of these reservation prices. Search stops when the maximum among the realized rewards among the opened boxes exceeds the reservation price of all unopened boxes.[1] Few extensions of Pandora's problem have been considered in the literature, with the exception of Choi and Smith (2016), Olszewski and Weber (2015), and Doval (2018). Doval (2018) studies a version of Pandora's boxes problem in which the decision maker can irreversibly select an unopened box. This problem is related to the extension we consider in Section 4. As mentioned above, the key feature distinguishing our problem from these papers is that the pool of boxes is endogenous in our model and determined by past search decisions.

2 Model

Pulling arms and searching for new ones. Time is discrete and indexed by $t = 0, ..., \infty$. In each period t, the decision maker can either pull one arm among the existing set of arms, denoted by $I_t = \{0, ..., n_t\}$, or search for new arms. Denote by $I_0 = \{0, ..., n_0\}$ the set of arms the decision maker is endowed with. In each period t, given the set I_t of arms already available at period t, the decision to search brings a (stochastic) set of new arms $I_{t+1} \backslash I_t$, which are added to the pool of existing arms. Pulling an existing arm $i \in I_t$ at period t yields a stochastic reward $r_{it} \in \mathbb{R}$, the distribution of which is a function of the arm's state. Arms that are not pulled yield no reward.

Let $x_{jt} \in X_{jt} \equiv \{0, 1\}$ denote the decision to pull arm $j \in \mathbb{N}$ in period t, with $x_{jt} = 1$ if arm j is pulled and $x_{jt} = 0$ otherwise. Denote the period-t history of pulls of arm j by $x_j^t \equiv (x_{js})_{s=0}^t$ and by $X_j^t \equiv \prod_{s=0}^t X_{js}$ the set of such histories. Similarly, denote by $x_t = (x_{jt})_{j=0}^\infty \in X_t \equiv \prod_{j=0}^\infty X_{jt}$ the decision of which arm to pull in period t and by $x^t = (x_s)_{s=0}^t \in X^t \equiv \prod_{s=0}^t X_s$ the period-t history of such pulls. A complete sequence of pulls of arms is denoted by $x = (x_s)_{s=0}^\infty$. At any period t, instead of pulling existing arms, the decision maker can search for additional arms at a cost c_t. Let

[1] Two important assumption implicit in this model are that (i) all uncertainty about a box is resolved at the moment the box is opened, and (ii) a box can only be chosen if it has been opened.

$y_t \in Y_t \equiv \{0, 1\}$, with $y_t = 1$ denoting the period-t decision to search for new arms, and $y_t = 0$ the decision not to search. Denote by $y^t \equiv (y_s)_{s=0}^t \in Y^t \equiv \prod_{s=0}^t Y_s$ the period-t history of search decisions and by $y = (y_s)_{s=0}^\infty$ a complete sequence of search decisions. The description of the reward processes corresponding to the arms and the evolution of the search technology is outlined below.

In order to allow for the possibility of "opting out", we assume that arm zero is a degenerate arm that yields a deterministic reward equal to the outside option (zero) at all periods. The period-t overall decision is summarized by $d_t \equiv (x_t, y_t)$. A sequence of decisions $d = (x, y) \in \Phi \equiv X \times Y$ is *feasible* if for all $t \geq 0$, (i) $x_{jt} = 1$ only if $j \in I_t$, and (ii) $\sum_{j=0}^\infty x_{jt} + y_t = 1$. That is, in each period, the decision maker either pulls an arm from the pool of existing ones, or searches for new arms.

States and arm types. Each arm has a *type* ξ, an element of an arbitrary type space Ξ. An arm's type determines the stochastic process governing the evolution of the arm's state. The process corresponding to each arm of type ξ is Markov and time-homogeneous. With a slight abuse of notation, we denote by $\omega^P = (\xi, \theta) \in \Omega^P = \Xi \times \Theta$ the arm's current *state,* where θ is an element of an arbitrary set Θ. Depending on the application, θ may take different forms. For example, it may contain the history of past rewards, but also additional information the decision maker may have received over time. Importantly, while ξ is fixed, θ evolves whenever the arm is activated. We denote by σ the sigma-algebra associated with Ω^P, and by $H_{\omega^P} \in \Delta(\Omega^P)$ the distribution over Ω^P when the arm's current state is ω^P. The rewards the decision maker receive over time are governed by the same process describing the evolution of the arm's state (that is, they can be represented by deterministic functions of the arm's state). The first time an arm of type ξ is activated, its reward is drawn from the distribution $H_{(\xi,\theta_0)}$ where, without loss of generality, θ_0 can be taken to be the same across all ξ.

The above formulation embodies the following assumptions common to many experimentation problems. The state of an arm evolves only upon being pulled. Furthermore, the distribution H_{ω^P} from which an arm's state is drawn is a function only of the arm's current state and is invariant in calendar time. Importantly, as in the classical bandit problem, the arms' states are drawn independently across arms, conditional on their current realizations. The formulation above is more restrictive than what is required for our results, which extend to more general (and not necessarily Markov) models.

Similarly to the processes corresponding to the "physical" arms, the process governing the cost incurred due to search, the number of new arms found as a result of search, and the types of the newly found arms, is Markov time-homogeneous. The state of the "*search arm*" is summarized by ω^S, which consists of the history $((c_0, E_0), (c_1, E_1), ..., (c_m, E_m))$ of past search costs and of arms' types found. Here $m \in \mathbb{N}$ denotes the number of times search has been carried out in the past, and $E_k = (n_k(\xi) : \xi \in \Xi)$ is a vector representing for each arm's type $\xi \in \Xi$ the number of arms $n_k(\xi) \in \mathbb{N}$ of type ξ found as the result of the k'th search. The first time search is activated the condition of the search arm is (c_0, E_0), where c_0 can be fixed arbitrarily as it plays no role in the analysis (the cost of the first search is c_1, not c_0), and E_0 can be taken to be the description of the types of arms I_0 the decision maker is endowed with at the outset of the decision problem, i.e., at $t = 0$. Denote the set of possible states of the search arm by Ω^S. The distribution over the search cost and the set of new arms found is denoted by H_{ω^S}.

Note that this formulation allows the search technology – the distribution over the search costs and over the new arms found – to depend in flexible ways on the results of previous searches (see the discussion below). The key assumptions are (a) that the search process is time-homogeneous (i.e., invariant in calendar time), and (b) that the outcome of each new search is drawn from H_{ω^S} independently from the idiosyncratic and time-varying component θ of each each physical arm pulled in previous periods.

We define the *state of the system* (for short, the "state") as follows. For each $\omega^P \in \Omega^P$, let $\mathcal{S}^P(\omega^P) \in \mathbb{N}$ denote the number of physical arms in state ω^P. The state of the system is given by the pair $\mathcal{S} \equiv (\omega^S, \mathcal{S}^P)$, where $\mathcal{S}^P : \Omega^P \to \mathbb{N}$ is the function describing, for each state $\omega^P \in \Omega^P$ of the physical arms, the number of physical arms in state ω^P. Next let $\Omega \equiv \Omega^P \cup \Omega^S$ and note that $\Omega^P \cap \Omega^S = \emptyset$. With an abuse of notation, we will sometime find it useful to denote the state as a function $\mathcal{S} : \Omega \to \mathbb{N}$ that specifies, for each $\omega \in \Omega$, including $\omega \in \Omega^S$, the number of arms, including the search arm, in state ω. We will then denote by $\mathcal{S}_t$ the state of the system at the beginning of period t. Clearly in the latter representation, at each period t, there is a unique $\hat{\omega}^s \in \Omega^S$ such that $\mathcal{S}_t(\omega^S) = 1$ if $\omega^s = \hat{\omega}^s$ and $\mathcal{S}_t(\omega^S) = 0$ if $\omega^s \neq \hat{\omega}^s$. Defining the state this way permits us to keep track of all relevant information, while facilitating certain results below.

Policies. A policy χ is a rule governing the decisions in each period – whether to search or pull one of the existing physical arms – based on the available information. More specifically, a process $(\mathcal{S}_t)_{t\geq 0}$, given a sequence of feasible decisions $(d_t)_{t\geq 0}$, generates a natural filtration $\{\mathcal{F}_t\}_{t\geq 0}$. A policy χ is then a $\mathcal{F}_t$-measurable sequence of feasible decisions $\{d_t\}_{t\geq 0}$.[2]

Denote by $u_t \equiv \sum_{j=0}^{\infty} x_{jt} r_{jt} - c_t y_t$ the realized period-t net reward. A policy χ is *optimal* if it maximizes the expected discounted sum of the net rewards

$$\mathbb{E}^{\chi} \left[\sum_{t=0}^{\infty} \delta^t u_t | \mathcal{S}_0 \right],$$

where $\delta \in (0, 1)$ denotes the discount factor. To guarantee that the process of the expected net rewards is well behaved, we assume that for any state of the system $\mathcal{S}$ and policy χ,

$$\lim_{T\to\infty} \delta^T \mathbb{E}^{\chi} \left[\sum_{s=T}^{\infty} \delta^s u_s | \mathcal{S} \right] = 0. \tag{1}$$

The above property guarantees that the solution to the Bellman equation corresponding to the above dynamic program coincides with the true value function. Note that this assumption is always satisfied if the rewards and costs are uniformly bounded, but the latter assumption is stronger than needed for the results. In particular, (1) allows the rewards of the physical arms to be drawn from a sampling process (see the discussion below) with unbounded support.[3]

[2] That is, χ is a policy if and only if the sequence of decisions $(d_t^{\chi})_{t\geq 0}$ under χ is $\{\mathcal{F}_t^{\chi}\}_{t\geq 0}$-adapted, where $\{\mathcal{F}_t^{\chi}\}_{t\geq 0}$ denotes the natural filtration generated by $(\mathcal{S}_t)_{t\geq 0}$ under χ.

[3] For example, rewards may be drawn from a Normal distribution with unknown mean – an environment which has

2.1 Discussion

We now comment on several environments the model captures:

1. *Experimentation.* First, the model may capture the classical setting in which arms are "sampling processes". For each arm, the rewards are drawn from a fixed distribution with an unknown parameter. When an arm is pulled, the agent obtains a reward, and updates her beliefs about this unknown parameter. In this specification, ξ may index elements of the process that are known at the outset (e.g., the variance of the distribution from which the rewards are drawn, as in certain Gaussian models) whereas θ may represent the history of past rewards. The agent's goal is to maximize her discounted sum of rewards.

2. *Sequential learning before choice.* Rewards need not accrue while learning is carried out. The model may capture a problem in which the agent sequentially inspects an endogenous set of alternatives, potentially multiple times as information need not be conclusive, before selecting among them (see Section 4 for the case in which the selection is irreversible). The agent maximizes her discounted reward from the "consumption" of the chosen alternative, net of the inspection costs. In this specification, ξ may represent a set of covariates known to the decision maker, whereas θ may represent the history of signals about an alternative's consumption value received over time.

3. *Beyond learning models.* The model is purposely not restricted to a learning environment. For example, the evolution of the states of the arms, when pulled, may also be triggered by random shocks, and may be driven by a preference for variety, or habit formation.

For all of these environments, the key novelty is that the set of alternatives is determined by past search decisions. The search technology – i.e., the distribution over the search costs and over the new alternatives discovered – is also defined flexibly, and may admit multiple interpretations, including learning about the set of "alternatives" that new searches may bring, or changes in the agent's ability to search. The following example illustrates.

Example 1 (Search technology).

1. A consumer expands her consideration set through search for alternative products. As she searches, she updates her beliefs about the type/number of new alternatives she expects to find from further search.

2. A researcher improves her ability to find new ideas over time (lower search costs, better distribution of new projects). The type of ideas she discovers may be correlated over time.

3. A firm wishes to fill a position. The probability of finding a candidate of given type ξ may decline over time if the number of potential candidates is limited.

3 Optimal policy and dynamics

We now show that the bandit problem with search introduced in the previous section admits a solution in the form of an index policy. That is, each physical arm, as well as the search arm, is

received attention in the literature (see, e.g., Brezzi and Lai, 2000).

attached an index whose value is independent of the information about the other arms. In each period, the decision maker then selects the arm, including search, with the highest index. The dynamics that arise under the optimal policy are discussed in Subsection 3.5 below.

A key feature complicating the analysis is that, in general, forgoing activating currently available arms in favor of search does not imply they will not be pulled in later periods. That is, search typically does not correspond to replacement of the existing arms. This implies that the decision maker must account for the fact that the physical arms that search may bring will subsequently "compete" with the pool of physical arms that past searches brought, as well as with future search decisions. Because the outcomes of search are correlated over time (both in terms of the types of arms found and of the costs of new searches), this introduces novel complications to the canonical problem without search.

A second complication comes from the fact that the results of search are arms, not rewards. The index for the search arm must thus be computed taking into account how the decision maker will activate the arms that search will bring. This is akin to treating search as a "meta arm", i.e., an arm that corresponds to an entire decision problem involving more than the simple decision of when to stop using the arm. As alluded to in the Introduction, in general, bandit problems with "meta arms" do not admit and index solution. The reason is that the optimal rule governing the usage of a meta arm (over and above the decision of when to stop it) typically depends on information that is outside the meta arm (e.g., the state of other arms). The examples below illustrate some of these difficulties.

3.1 Cautionary examples

Consider the following extension of the environment described above. There are $k \in \mathbb{N}$ sets of arms, $K_1, ..., K_k$. Arms from different sets evolve independently of one another, but the state of each arm within a set may depend on the state of other arms from the same set. More generally, suppose that each arm corresponds to a "meta arm", the activation of which involves decisions other than when to stop using it. Each meta arm has its own "meta" process that evolves independently of the other meta arms.

Clearly, the model in Section 2 is a special case of this richer setting. Suppose that, for each set of arms K_i, one can compute the optimal sequence of pulls, independently of the other sets of arms. Equivalently, suppose that for each "meta arm" one can compute the optimal sequence of decisions that define the usage of that arm, independently of the solution to the other meta arms' problems. It is tempting to conjecture that one may then assign an independent index to each set of arms K_i (alternatively, to each "meta arm") and that the optimal policy for the overall problem reduces to an index policy, whereby the meta arm with the highest index is selected in each period.

Perhaps surprisingly, the optimal policy for this enriched problem does not admit an index representation. When arms are not defined as in the canonical multi-armed bandit problem, but rather feature a more complicated internal decision problem (preserving the independence across arms), the optimal policy typically is not an index policy. The following example illustrates.

Example 2. There are two arms. Arm 1 yields a reward of 1000 when it is first pulled. In all subsequent pulls, it yields a reward λ, where λ is initially unknown and may be either 1 or 10, with equal probability. After the first pull of arm 1, λ is perfectly revealed and is fully persistent. Arm 2 is a "meta arm" corresponding to the following decision problem. When the decision maker pulls arm 2 for the first time, she must also choose *how* to pull it. There are two ways to pull this arm, 2(A) and 2(B). If the decision maker selects $2(A)$, the arm yields a reward of 100 for a single period, followed by no rewards thereafter. If, instead, the decision maker selects $2(B)$, the arm yields a reward equal to 11 in each of its subsequent pulls. The choice of which version of arm 2 to use must be made the first time that arm 2 is pulled and can not be reversed.

Assume $\delta = 0.9$. The optimal policy for this problem is the following. In period 1, arm 1 is pulled. If $\lambda = 10$, then arm 2 in version $2(A)$ is then pulled for a single period, followed by arm 1 again in all subsequent periods. If, instead, $\lambda = 1$, arm 2 is then pulled in version $2(B)$ in all subsequent periods. Note that, under the optimal policy, the decision of how to use arm 2 depends on the realization of arm 1's first pull. It is then evident that the optimal policy is not an index policy, no matter how one defines the indices. This is because an index policy requires that both the index of each arm and its utilization (when an arm can be used in different versions, as in the case of "meta arm" 2 in this example) be invariant in the results of the activation of all other arms.

3.2 Optimal policy

For each condition $\omega^P \in \Omega^P$ of the physical arm, let

$$\mathcal{G}^P(\omega^P) \equiv \sup_{\tau>0} \frac{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s r_s | \omega^P\right]}{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s | \omega^P\right]}, \tag{2}$$

be the arm's index, where τ denotes a measurable stopping time.[4] The definition in (2) is equivalent to the standard definition of Gittins (1979).

Given the state $\mathcal{S}$, denote the maximal index among the available physical arms by

$$\mathcal{G}^*(\mathcal{S}^P) = \max_{\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}^P(\hat{\omega}^P)>0\}} \mathcal{G}(\omega^P).$$

Note that $\mathcal{G}^*(\mathcal{S})$ depends on $\mathcal{S}$ only through the state of the physical arms present in $\mathcal{S}$.

We now assign an index to the search arm that is invariant in the information about any of the available physical arms. Analogously to the standard Gittins index, the search index will be defined as the maximal expected average discounted net reward, per unit of expected discounted time, obtained between the current period and an optimal stopping time. Contrary to the standard Gittins index, however, the potential arrival of additional new physical arms as the result of search means the maximization in the definition is not just over stopping times, but also over measurable rules governing the selection among the *new* physical arms and further searches. Denote by τ a

[4]Specifically, τ is a stopping time with respect to the process whose filtration is obtained by pulling the physical arm with initial state ω^P in each period.

measurable stopping time, and by π a measurable rule which prescribes for any period s between the initial one and the stopping time τ which arm is chosen among the new physical arms that arrive over time and the search arm, given the initial state of the search arm, ω^S. Importantly, π selects only among the search arm and the physical arms arriving after the launch of search from state ω^S. That is, it does not select among physical arms present before the launch of search. To make things clear, suppose search is launched in period t and terminated in period $\tau > t$. Then at each period $t < s < \tau$, π selects between search and pulling one of the physical arms available in period s but which were not present at period t. Hence, the period-t search index is independent of any information about the physical arms present at period t.

Formally, given the condition of the search arm $\omega^S \in \Omega^S$, the search index is defined by

$$\mathcal{G}^S(\omega^S) \equiv \sup_{\pi,\tau} \frac{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s (r_s^\pi - c_s^\pi) | \omega^S\right]}{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s | \omega^s\right]}, \tag{3}$$

where r_s^π, c_s^π denote, respectively, the stochastic rewards and costs obtained/incurred between the time search is launched and τ, under the rule π. We are now ready to define the following policy based on comparisons of the indices defined above.

Definition 1. The *index policy* χ^* selects at each period $t \geq 0$, given the state $\mathcal{S}_t$, search if and only if $\mathcal{G}^S(\omega^S) \geq \mathcal{G}^*(\mathcal{S}^\mathcal{P})$, and otherwise an arbitrary physical arm with index $\mathcal{G}^*(\mathcal{S}^P)$.

Ties between the physical arms may be broken arbitrarily. In order to maintain consistency throughout the analysis, we assume that in the case of a tie $\mathcal{G}^S(\omega_t^S) = \mathcal{G}^*(\mathcal{S}_t^P)$, the tie is broken in favor of search.

Theorem 1. *The index policy χ^* is optimal in the bandit problem with search for new arms.*

3.3 Proof of Theorem 1.

The proof consists of three steps. Steps one and two establish certain key properties of the payoff the decision maker obtains under the index policy χ^*. In particular, these steps introduce some auxiliary processes based on the evolution of the indices of the physical arms and of the search arm that favor a certain integral representation of the payoffs under the index policy χ^* and shed light on the dynamics that arise under the index policy. The last step then uses the results in the previous two steps to show that the payoff under the index policy χ^* satisfies the Bellman equation associated with the dynamic program under consideration, which establishes the optimality of χ^*.

Step 1. We first introduce some convenient notation. Define $\mathcal{S}_t^1 \vee \mathcal{S}_t^2 \equiv (\mathcal{S}_t^1(\omega) + \mathcal{S}_t^2(\omega) : \omega \in \Omega)$ and $\mathcal{S}_t^1 \backslash \mathcal{S}_t^2 \equiv (\max\{\mathcal{S}_t^1(\omega) - \mathcal{S}_t^2(\omega), 0\} : \omega \in \Omega)$, and denote by $e_t(\omega)$ the state of the system when only a single arm, physical or search, is present and is in state ω. Recall that any feasible state of the system must contain exactly one search arm (i.e., one state $\hat{\omega}^S$ for which $\mathcal{S}_t(\hat{\omega}^S) = 1$ with $\mathcal{S}_t(\omega^S) = 0$ for all $\omega^S \neq \hat{\omega}^s$). However, it will be convenient to consider also fictitious (unfeasible) states in which there is no search arm, as well as fictitious states with multiple search arms.

Next observe that starting from state $\omega^P \in \Omega^P$, the optimal stopping time $\tau^*(\omega^P)$ in the definition of the index (2) of a physical arm in state ω^P is the first time at which the arm reaches a state in which its index is below its initial value. Formally, for all t, all ω_t^P,

$$\tau^*(\omega_t^P) = \inf\{s > t | \mathcal{G}(\omega_s^P) \leq \mathcal{G}(\omega_t^P)\}.$$

(See Mandelbaum, 1986, for a formal proof of this property). In an environment with search for new arms, the optimal stopping time $\tau^*(\omega_t^S)$ in the RHS of the definition of the index for the search arm, (3), will be the first time at which the new indices corresponding to both the (future) states of the search arm and all *newly found* physical arms are no greater than the index of the search arm (3) at the time of the initial decision to search. Put differently, starting from the state of the system $e(\omega^S)$ at period t, τ^* is the first time in which any remaining arms – including new arms added as the result of search – have an index smaller than the index corresponding to the search arm at period t.

We therefore consider the following process $(\mathcal{S}_t)_{t\geq 0}$ generated by the index policy χ^*. Starting from any state $\mathcal{S}$, denote by $\kappa(v|\mathcal{S}) \in \mathbb{N} \cup \{\infty\}$ the shortest time it takes until (i) the search arm reaches a state in which its index is no greater than v, and (ii) *all* physical arms – regardless of when they have arrived – have an index no greater than v. That is, $\kappa(v|\mathcal{S})$ is the minimal number of periods, starting from the current state $\mathcal{S}$, until all indices of present arms including search are weakly below v. In case this event never occurs, let $\kappa(v|\mathcal{S}) = \infty$. Note that throughout the periods $\kappa(v|\mathcal{S})$, if the search arm is pulled, new arms will arrive and the evolution of their indices must also be taken into account.

Observe that the independence of the processes governing the evolution of the various physical arms, conditional on the arms' types, along with the independence of these processes from the processes governing the evolution of the search technology (again, conditional on the existing arms' types) imply that for any v and states of the system $\mathcal{S}^1$ and $\mathcal{S}^2$,

$$\kappa(v|\mathcal{S}^1 \vee \mathcal{S}^2) = \kappa(v|\mathcal{S}^1) + \kappa(v|\mathcal{S}^2). \tag{4}$$

We shall construct the following stochastic process based on the values of the indices, and their arrival through search, under the index policy χ^*. Starting with the initial state of the system $\mathcal{S}_0$, in which the state of the search arm is ω_0^S, let $v^0 = \max\{\mathcal{G}^*(\mathcal{S}_0^P), \mathcal{G}^S(\omega_0^S)\}$. Consider the first time t^0 in which all indices are strictly below v^0, with $t^0 = \infty$ if this event never occurs. Note that t^0 may differ from $\kappa(v^0|\mathcal{S}_0)$, as t^0 is the first time at which all indices are strictly below v^0, whereas $\kappa(v^0|\mathcal{S}_0)$ is the first time at which all indices are weakly below v^0. Next let $v^1 = \max\{\mathcal{G}^*(\mathcal{S}_{t^0}^P), \mathcal{G}^S(\omega_{t^0}^S)\}$, where $\mathcal{S}_{t^0}^P$ and $\omega_{t^0}^S$ are the states of the physical arms and of the search arm at time t^0. Note that, by construction, $\kappa(v^0|\mathcal{S}_0) = 0$ and $t^0 = \kappa(v^1|\mathcal{S}_0)$. Furthermore, if $t^0 < \infty$ then $v^0 > \mathcal{G}^S(\omega_0^S)$ implies $\omega_{t^0}^S = \omega_0^S$. We can proceed in this manner to obtain a stochastic, *strictly decreasing*, sequence of values $\left(v^i\right)_{i\geq 0}$, with corresponding stochastic times $\left(\kappa(v^i|\mathcal{S}_0)\right)_{i\geq 0}$. Next, for any $i = 0, 1, 2, ...,$ let $\eta(v^i|\mathcal{S}_0) = \sum_{s=\kappa(v^i|\mathcal{S}_0)}^{\kappa(v^{i+1}|\mathcal{S}_0)-1} u_s$ denote the discounted sum of the net rewards between periods $\kappa(v^i|\mathcal{S}_0)$

and $\kappa(v^{i+1}|\mathcal{S}_0) - 1$ and then let $\left(\eta(v^i|\mathcal{S}_0)\right)_{i\geq 0}$ define the corresponding sequence of discounted accumulated net rewards, with $\eta(v^i|\mathcal{S}_0) = 0$ if $\kappa(v^i|\mathcal{S}_0) = \infty$.

Denote by $\mathcal{V}(\mathcal{S}_0)$ the average expected discounted sum of net reward under the index policy χ^*, given the initial state of the system $\mathcal{S}_0$. That is, $\mathcal{V}(\mathcal{S}_0) = (1-\delta)\mathbb{E}^{\chi^*}\left[\sum_{t=0}^{\infty}\delta^t u_t|\mathcal{S}_0\right]$. Below we identify properties of the function $\mathcal{V}(\mathcal{S}_0)$ that permit us to establish the optimality of the index policy χ^*. It should be clear that the same properties apply to $\mathcal{V}(\mathcal{S}_t)$, for any $t \geq 0$ (to see this, it suffices to note that the value of $\mathcal{S}_0 = (\omega_0^S, \mathcal{S}_0^P)$ plays no role in the arguments below). We start with the following result:[5]

Lemma 1. *Given the state of the system $\mathcal{S}_0$, the expected continuation payoff under the index policy χ^* is given by*

$$\mathcal{V}(\mathcal{S}_0) = \int_0^{\infty}\left(1 - \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)}\right)dv. \tag{5}$$

Proof of Lemma 1. First, observe that by the definition of the processes $\left(\kappa(v^i|\mathcal{S}_0)\right)_{i\geq 0}$ and $\left(\eta(v^i|\mathcal{S}_0)\right)_{i\geq 0}$ we have

$$\mathcal{V}(\mathcal{S}_0) = (1-\delta)\mathbb{E}\left[\sum_{i=0}^{\infty}\delta^{\kappa(v^i)}\eta(v^i)|\mathcal{S}_0\right].$$

Next, using the definition of the indices (2) and (3), along with the properties of the optimal stopping times in the definition of the indices, we have

$$v^i = \frac{(1-\delta)\mathbb{E}\left[\eta(v^i)|\mathcal{F}_{\kappa(v^i)}\right]}{\mathbb{E}\left[1-\delta^{\kappa(v^{i+1})-\kappa(v^i)}|\mathcal{F}_{\kappa(v^i)}\right]}, \tag{6}$$

with $\mathcal{F}_0$ corresponding to the initial state $\mathcal{S}_0$. To see why (6) holds, note that at $\kappa(v^i)$, given $\mathcal{F}_{\kappa(v^i)}$, the optimal stopping time defining the index v^i of the arm with the largest index among those available in period $\kappa(v^i)$ is the first time at which the index of that arm, and in case the arm is the search arm, also of all arms found through future searches, drop below v^i.[6]

Rearranging, multiplying both sides of (6) by $\delta^{\kappa(v^i)}$, and using the fact that $\delta^{\kappa(v^i)}$ is known at $\kappa(v^i|\mathcal{S}_0)$, we have that

$$v^i\mathbb{E}\left[\delta^{\kappa(v^i)} - \delta^{\kappa(v^{i+1})}|\mathcal{F}_{\kappa(v^i)}\right] = (1-\delta)\mathbb{E}\left[\delta^{\kappa(v^i)}\eta(v^i)|\mathcal{F}_{\kappa(v^i)}\right].$$

[5] In the proof of Lemma 1 below, as well as in the entire analysis that follows the lemma, when we take expectations over the processes for κ and η, we condition on the initial value of the system $\mathcal{S}_0$. To ease the notation, we drop the conditioning on $\mathcal{S}_0$ from the arguments of κ and η when there is no risk of confusion.

[6] Note that if at period $\kappa(v^i)$ there are multiple arms with index v^i, the average sum $\mathbb{E}\left[\eta(v^i)|\mathcal{F}_{\kappa(v^i)}\right]$ of the discounted net rewards across all arms with index v^i until the indices of all such arms – and in case such arms include search, also of the newly found arms –drop below v^i, per unit of average discounted time, $\mathbb{E}\left[1-\delta^{\kappa(v^{i+1})-\kappa(v^i)}|\mathcal{F}_{\kappa(v^i)}\right]/(1-\delta)$, is the same as the average sum of the discounted net reward of each individual arm with index v^i normalized by the average discounted time until the index of that arm (and again, in case that arm is search, also of its descendants) fall below v^i. This follows from the independence of the processes. Hence, Condition (6) holds irrespectively of whether at $\kappa(v^i)$ there is a single or multiple arms with index v^i.

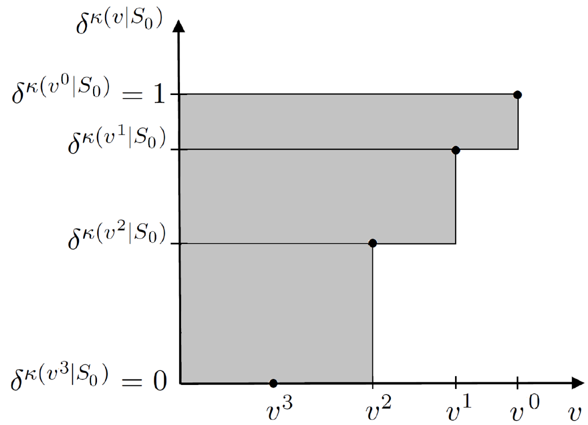

Figure 1: An illustration of the function $\delta^{\kappa(v|S_0)}$ and the region $\sum_{i=0}^{\infty} v^i \left(\delta^{\kappa(v^i|S_0)} - \delta^{\kappa(v^{i+1}|S_0)}\right) = \int_0^\infty v \mathrm{d}\delta^{\kappa(v|S_0)}$, for a particular path with $\kappa(v^3|S_0) = \infty$.

Taking expectations of both sides of the previous equality given the initial state $\mathcal{S}_0$, and using the law of iterated expectations, we have that

$$\mathbb{E}\left[v^i \mathbb{E}\left[\delta^{\kappa(v^i)} - \delta^{\kappa(v^{i+1})}|\mathcal{F}_{\kappa(v^i)}\right]|\mathcal{S}_0\right] = (1-\delta)\mathbb{E}\left[\mathbb{E}\left[\delta^{\kappa(v^i)}\eta(v^i)|\mathcal{F}_{\kappa(v^i)}\right]|\mathcal{S}_0\right],$$

and hence that

$$\mathbb{E}\left[v^i\left(\delta^{\kappa(v^i)} - \delta^{\kappa(v^{i+1})}\right)|\mathcal{S}_0\right] = (1-\delta)\mathbb{E}\left[\delta^{\kappa(v^i)}\eta(v^i)|\mathcal{S}_0\right].$$

If follows that

$$\mathcal{V}(\mathcal{S}_0) = (1-\delta)\mathbb{E}\left[\sum_{i=0}^{\infty}\delta^{\kappa(v^i)}\eta(v^i)|\mathcal{S}_0\right] = \mathbb{E}\left[\sum_{i=0}^{\infty} v^i\left(\delta^{\kappa(v^i)} - \delta^{\kappa(v^{i+1})}\right)|\mathcal{S}_0\right]. \tag{7}$$

Next, note that $\delta^{\kappa(v^i)} = 0$ whenever $\kappa(v^i) = \infty$. Furthermore, for any $i = 0, 1, ..., \kappa(v) = \kappa(v^{i+1})$ for all $v^{i+1} < v < v^i$. It follows that (7) is equivalent to

$$\mathcal{V}(\mathcal{S}_0) = \mathbb{E}\left[\int_0^\infty v\mathrm{d}\delta^{\kappa(v)}|\mathcal{S}_0\right] = \mathbb{E}\left[\int_0^\infty \left(1-\delta^{\kappa(v)}\right)\mathrm{d}v|\mathcal{S}_0\right] = \int_0^\infty \left(1 - \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)}\right)\mathrm{d}v,$$

as claimed. □

Step 2. We now use the result in Lemma 1 to characterize how much the decision maker obtains from following the index policy from the outset rather than being forced to pick a specific arm in the first period and the reverting to the index policy from the next period onward. The characterization will permit to establish in Step 3 the optimality of the index policy through dynamic programming.

Given the state of the system $\mathcal{S}_0$, for any $\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0^P(\hat{\omega}^P) > 0\}$, denote by $\mathbb{E}\left[\tilde{r}|\omega^P\right]$ the immediate expected reward from pulling the physical arm in state ω^P (the expectation is taken under the distribution H_{ω^P}) and by $\tilde{\omega}^P$ the new state of the physical arm. Let

$$V^P(\omega^P|\mathcal{S}_0) \equiv (1-\delta)\mathbb{E}\left[\tilde{r}|\omega^P\right] + \delta\mathbb{E}\left[\mathcal{V}\left(\mathcal{S}_0 \backslash e(\omega^P) \vee e(\tilde{\omega}^P)\right)|\omega^P\right] \tag{8}$$

denote the payoff from starting with a physical arm in state ω^P and then following the index policy from the next period onward. Similarly, let

$$V^S(\omega^S|\mathcal{S}_0) \equiv -(1-\delta)\mathbb{E}\left[\tilde{c}|\omega^S\right] + \delta\mathbb{E}\left[\mathcal{V}\left(S_0 \backslash e(\omega^S) \vee e(\tilde{\omega}^S) \vee W^P(\tilde{\omega}^S)\right)|\omega^S\right], \tag{9}$$

denote the payoff from starting with the search arm in state ω^s and then following the index policy from the next period onward, where $\mathbb{E}\left[\tilde{c}|\omega^S\right]$ is the immediate expected cost incurred from searching, $\tilde{\omega}^S$ is the new state of the search arm after search is carried out, and $W^P(\tilde{\omega}^S)$ is the state of the new physical arms obtained as the result of the search, with $\tilde{c}$ and $W^P(\tilde{\omega}^S)$ jointly drawn from the distribution H_{ω^S}.[7]

Following Whittle (1980, 1981), it is convenient to introduce an additional auxiliary arm which is available at all periods and which yields a constant reward equal to $M < \infty$ whenever pulled. Denote the state corresponding to this arm by ω_M^A, and enlarge Ω^P to include ω_M^A. Similarly, let $e(\omega_M^A)$ denote the state of the system when only the auxiliary arm with fixed reward M is present. Since the reward from the auxiliary arm is constant at M, if $v \geq M$, then $\kappa(v|\mathcal{S}_0 \vee e(\omega_M^A)) = \kappa(v|\mathcal{S}_0)$. If, instead, $v < M$, then clearly $\kappa(v|\mathcal{S}_0 \vee e(\omega_M^A)) = \infty$. Hence, Lemma 1, adapted to the fictitious environment that includes the auxiliary arm, implies that

$$\begin{aligned}\mathcal{V}(\mathcal{S}_0 \vee e(\omega_M^A)) &= \int_0^\infty \left(1 - \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0 \vee e(\omega_M^A))}\right) \mathrm{d}v \\ &= M + \int_M^\infty \left(1 - \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)}\right) \mathrm{d}v \\ &= \mathcal{V}(\mathcal{S}_0) + \int_0^M \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)} \mathrm{d}v. \end{aligned} \tag{10}$$

The definition of the index policy, along with Conditions (8) and (9), also implies the following.

Lemma 2. *For any state of the search technology ω^S, state of an arm ω^P, and reward M from the auxiliary arm:*

$$\mathcal{V}(e(\omega^S) \vee e(\omega_M^A)) = \begin{cases} V^S(\omega^S|e(\omega^S) \vee e(\omega_M^A)) & \text{if } M \leq \mathcal{G}^S(\omega^S) \\ M > V^S(\omega^S|e(\omega^S) \vee e(\omega_M^A)) & \text{if } M > \mathcal{G}^S(\omega^S) \end{cases} \tag{11}$$

[7] Note that $W^P(\tilde{\omega}^S)$ is a deterministic function of the new state $\tilde{\omega}^S = ((c_0, E_0), ..., (c_m, E_m))$ of the search arm. To see this, recall that $E_m = (n_m(\xi) : \xi \in \Xi)$ is the list of the number of arms of each type found at the last search. The state $W^P(\tilde{\omega}^S) : \Omega \to \mathbb{N}$ is then obtained from E_m by letting $W^P(\tilde{\omega}^S)(\omega) = 0$ for all $\omega \in \Omega^S \cup \{(\xi, r^s) \in \Omega^P : s > 0 \text{ or } E_m(\xi) = 0\}$ and $W^P(\tilde{\omega}^S)(\omega) = E_m(\xi)$ for all states $(\xi, r^0) \in \Omega^P$ such that $E_m(\xi) > 0$.

and similarly

$$\mathcal{V}(e(\omega^P) \vee e(\omega^A_M)) = \begin{cases} V^P(\omega|e(\omega^P) \vee e(\omega^A_M)) & \textit{if } M \leq \mathcal{G}^P(\omega^P) \\ M > V^P(\omega|e(\omega^P) \vee e(\omega^A_M)) & \textit{if } M > \mathcal{G}^P(\omega^P). \end{cases} \tag{12}$$

Proof of Lemma 2. See the Appendix.

Next, for any initial state of the system $\mathcal{S}_0$, and any state $\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0(\hat{\omega}^P) > 0\}$ of the physical arms present in $\mathcal{S}_0$, let $D^P(\omega^P|\mathcal{S}_0) \equiv \mathcal{V}(\mathcal{S}_0) - V^P(\omega^P|\mathcal{S}_0)$ denote the payoff differential between starting with the index policy right away and starting with a physical arm in state ω^P and then following the index policy from the following period onward. Similarly, let $D^S(\omega^S|\mathcal{S}_0) \equiv \mathcal{V}(\mathcal{S}_0) - V^S(\omega^S|\mathcal{S}_0)$ denote the payoff differential between starting with the index policy right away and starting with the search arm in state ω^S and then following the index policy from the following period onward, where ω^S is the state of the search arm, as specified by $\mathcal{S}_0$.[8] We are now ready to establish the following lemma.[9]

Lemma 3. *Let $\mathcal{S}_0$ be the initial state of the system, with $\omega^S \in \Omega^S$ being the state of the search technology, as specified in $\mathcal{S}_0$. Then*

$$D^S(\omega^S|\mathcal{S}_0) = \int_0^{\mathcal{G}^*(\mathcal{S}_0^P)} D^S(\omega^S|e(\omega^S) \vee e(\omega^A_v))d\mathbb{E}\delta^{\kappa(v|\mathcal{S}_0 \setminus e(\omega^S))}. \tag{13}$$

Similarly, for any physical arm of type $\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0^P(\hat{\omega}^P) > 0\}$,

$$D^P(\omega^P|\mathcal{S}_0) = \int_0^{max\{\mathcal{G}^*(\mathcal{S}_0^P \setminus e(\omega^P)), \mathcal{G}^S(\omega^S)\}} D^P(\omega^P|e(\omega^P) \vee e(\omega^A_v))d\mathbb{E}\delta^{\kappa(v|\mathcal{S}_0 \setminus e(\omega^P))}. \tag{14}$$

Proof of Lemma 3. See the Appendix.

Step 3. Using Lemma 3, we can now directly verify that the average payoff under an index policy satisfies the dynamic programming equation, implying the optimality of the index policy. Let

$$\mathcal{V}^*(\mathcal{S}_0) \equiv (1-\delta)\text{sup}_{\chi \in \mathcal{X}} \mathbb{E}^{\chi}\left[\sum_{t=0}^{\infty} \delta^t u_t | \mathcal{S}_0\right] \tag{15}$$

denote the value function for the dynamic programming problem under consideration. We then have the following result:

Lemma 4. *For any state of the system $\mathcal{S}_0$, with ω^S denoting the state of the search technology as specified under $\mathcal{S}_0$,*

1. *$\mathcal{V}(\mathcal{S}_0) \geq V^S(\omega^S|\mathcal{S}_0)$, with the inequality holding as an equality if and only if $\mathcal{G}^S(\omega^S) \geq \mathcal{G}^*(\mathcal{S}_0^P)$;*

[8]That is, ω^S is the unique state in Ω^S such that $\mathcal{S}_0(\omega^s) > 0$.

[9]In the statement of the lemma, $\mathcal{S}_0 \setminus e(\omega^S)$ is the state of a fictitious system in which the search arm is absent, whereas $\mathcal{S}_0^P \setminus e(\omega^P)$ is the state of the physical arms obtained from $\mathcal{S}_0^P$ by subtracting one physical arm in state ω^P.

2. *for any $\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0(\hat{\omega}^P) > 0\}$, $\mathcal{V}(\mathcal{S}_0) \geq V^P(\omega^P|\mathcal{S}_0)$ with the inequality holding as an equality if and only if $\mathcal{G}^P(\omega^P) = \mathcal{G}^*(\mathcal{S}_0^P) \geq \mathcal{G}^S(\omega^S)$.*

Hence, for any $\mathcal{S}_0$, $\mathcal{V}(\mathcal{S}_0) = \mathcal{V}^*(\mathcal{S}_0)$, and the index policy χ^* is optimal.

Proof of Lemma 4. First, use (11) to note that the integrand in (13) is non-negative for all $0 \leq v \leq \mathcal{G}^*(\mathcal{S}_0^P)$, and that the entire integral in (13) is equal to zero if and only $\mathcal{G}^*(\mathcal{S}_0^P) \leq \mathcal{G}^S(\omega^S)$. This establishes Condition 1 in the lemma. Similarly, use (12) to observe that for any $\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0^P(\hat{\omega}^P) > 0\}$, the integrand in (14) is non-negative for any $0 \leq v \leq \max\{\mathcal{G}^*(\mathcal{S}_0^P \backslash e(\omega^P)), \mathcal{G}^S(\omega^S)\}$, and that the entire integral in (14) is equal to zero if and only if $\mathcal{G}^P(\omega^P) \geq \max\{\mathcal{G}^*(\mathcal{S}_0^P \backslash e(\omega^P)), \mathcal{G}^S(\omega^S)\}$, which is the case if and only if $\mathcal{G}^P(\omega^P) = \mathcal{G}^*(\mathcal{S}_0^P) \geq \mathcal{G}^S(\omega^S)$. This establishes Condition 2 of the lemma.

Next, note that, jointly, Conditions 1 and 2 in the lemma imply that

$$\mathcal{V}(\mathcal{S}_0) = \max\left\{V^S(\omega^S|\mathcal{S}_0), \max_{\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0^P(\hat{\omega}^P) > 0\}} V^P(\omega^P|\mathcal{S}_0)\right\}.$$

Hence the payoff under the index policy $\mathcal{V}$ solves the Bellman equation corresponding to the dynamic program under consideration. Assumption (1) then guarantees that $\mathcal{V}$ must coincide with the value function, i.e., $\mathcal{V}(\mathcal{S}_0) = \mathcal{V}^*(\mathcal{S}_0)$, and hence that the index policy χ^* is optimal. □

The result in the last lemma completes the proof of the theorem. ■

3.4 Recursive representation of the indices

An important feature of the optimal policy χ^* is its recursive structure. Note that the index for search as defined in (3) does not provide much intuition as to what the policy π and stopping time τ attaining the supremum may be. Interestingly, the intuition developed in the previous section implies that the policy π attaining the supremum is in fact the index policy χ^* itself.

Corollary 1. *The index $\mathcal{G}^S(\omega^S)$ of the search arm can be re-written as*

$$\mathcal{G}^S(\omega^S) = \frac{\mathbb{E}^{\chi^*}\left[\sum_{s=0}^{\tau^*-1} \delta^s (r_s - c_s)|\omega^S\right]}{\mathbb{E}^{\chi^*}\left[\sum_{s=0}^{\tau^*-1} \delta^s |\omega^s\right]}, \tag{16}$$

where the expectation is taken with respect to the process induced by the index policy χ^*, as defined in Definition 1, and where τ^* is the first time $s \geq 1$ at which the indices of the search arm and of all physical arms brought in by search fall below the value of the search arm (16) at the time search was launched (i.e., at $s = 0$).

3.5 Dynamics and the search technology

A special case in which the dynamics under the optimal policy χ^* takes a particularly simple form is when the search technology is constant, i.e., the distribution H_{ω^S} from which (c, E) is drawn is invariant in the state of the search arm. In this case, an immediate implication of Theorem 1 is that search corresponds to replacement of the existing set of arms and, as a result, at each time in which search is carried out the continuation value is exactly the same.

Proposition 1. *Assume the search technology is constant: $H_{\omega^S} = H$ for all $\omega^s \in \Omega^S$. For any two states of the system $\mathcal{S}$, $\mathcal{S}'$ for which search takes place under χ^*, $\mathcal{V}(\mathcal{S}) = \mathcal{V}(\mathcal{S}')$. Furthermore, without loss of optimality, all physical arms present at state $\mathcal{S}$ are subsequently never pulled.*

Proof. From Theorem 1, the decision to search is optimally determined under χ^* based on a comparison of the index (3) corresponding to search – which is defined independently of the state of the existing physical arms and is constant under a fixed search technology – and the maximal index among the existing physical arms. Since the state of a physical arm changes only when pulled, if in period t, $\mathcal{G}^S \geq \mathcal{G}^*(\mathcal{S}^P)$, this must remain true in all subsequent periods. ■

The result above continues to hold if the search technology improves over time. For example, a researcher may grow better at finding promising research projects over time. Formally, recall that $E_k = (n_k(\xi) : \xi \in \Xi)$ is a vector representing for each arm's type $\xi \in \Xi$ the number of arms $n_k(\xi) \in \mathbb{N}$ of type ξ found as the result of the k'th search. We say that the search technology is improving if $(-c_k, E_k)$ increases in k, in the sense of first-order stochastic dominance.[10] In this case, physical arms are required to pass a higher and higher threshold in order to be pulled, while search becomes more and more desirable.

On the other hand, in some circumstances it is instead more reasonable to suppose the search technology deteriorates over time (defined analogously to an improving technology). This is the case, for example, if the number of potential arms that may be found is finite, in which case the more alternatives have been found through search, the fewer alternatives remain to be found. In the latter case, the simple dynamics of Proposition 1 no longer apply. Search is instead front-loaded, with the pool of arms increased early on with the expectation of being pulled in the future, as the threshold the index of physical arms must pass in order to be pulled decreases.

4 Search, experimentation, and irreversible choice

There are many environments in which in addition to learning about existing alternatives and searching for new ones the decision maker has the option to choose *irreversibly* one of the alternatives, triggering a discrete change in payoffs. An example of such a setting is Weitzman's (1979) Pandora's problem, where uncertainty about the reward from each box is resolved at the moment the box is opened, and a box can be irreversibly chosen only if it was previously opened. These assumptions guarantee that the irreversibly of the agent's choice has no bite: Even if the choice was reversible, since no additional information about an opened box is available, there is no reason for the decision maker to change her decision. Doval (2018) studies a generalization of this problem in

[10] That is, for any k and for any upper set $Z \in \mathbb{R} \times \mathbb{R}^{|\Xi|}$, $\Pr((-c_{k+1}, E_{k+1}) \in Z) \geq \Pr((-c_k, E_k) \in Z)$. This definition of an improving search technology is quite strong. In more specific environments in which there is an order on the set of types Ξ, weaker definitions can be introduced.

which a box may be selected even if it has not been previously opened, showing that the optimal solution takes the form of an index policy only under certain conditions, and studying the case in which the index policy need not be optimal.

In this section, we extend our analysis to a general model of search, experimentation, and irreversible choice in which, at each period, the decision maker can (i) experiment among the alternatives already discovered, (ii) search for new ones, or (iii) irreversibly select an alternative among those discovered through past searches, based on *possibly partial* information about the profitability of the available alternatives.

Environment with irreversible choice. Formally, we modify the environment as follows. In addition to the actions x_t and y_t defined above, we introduce an additional action, $z_{jt} \in \{0, 1\}$, representing the irreversible choice of selecting arm j in period t, with $z_{jt} = 1$ denoting the decision to irreversibly commit to arm j, and $z_{jt} = 0$ the decision to not commit to that arm in period t. The period-t complete decision is then summarized by $d_t \equiv (x_t, y_t, z_t)$, with $z_t = (z_{jt})_{j=0}^{\infty}$. A sequence of decisions d is *feasible* if, for all $t \geq 0$, (i) $x_{jt} = 1$, or $z_{jt} = 1$, only if $j \in I_t$, (ii) $\sum_{j=1}^{\infty} x_{jt} + y_t + z_{jt} = 1$, and (iii) $z_{jt} = 1$ if $z_{js} = 1$ for some $s < t$. Together, the last two assumptions imply that, once an arm is irreversibly chosen, there are no further decisions to be made.

To allow for the possibility that arms have to be explored a few times before they can be irreversibly selected, we assume that each arm of type ξ must be pulled at least $M_\xi \geq 0$ times before it can be irreversibly chosen (the case where $M_\xi = \infty$ corresponds to the baseline model with no irreversible choice studied in the previous sections). Pulling a physical arm without committing to it (formally captured by $x_{jt} = 1$) triggers a change in the arm's state and brings a flow reward, as in the baseline model. Irreversibly committing to arm j of type ξ in period t (formally captured by the decision $z_{jt} = 1$), instead, yields a flow reward to the decision maker from that moment onward, the value of which may be only imperfectly known to the decision maker at the time the irreversible decision is made. We denote by $R(\omega^P)$ the expected flow value of committing to an arm of type ξ in state $\omega^P = (\xi, \theta)$, where ω^P is the arm's state at the moment the irreversible decision is made (i.e., at the first t at which $z_{jt} = 1$). Note that since choice is irreversible, $R(\omega^P)$ admits two equivalent interpretations: (i) If the the arm is chosen, an immediate expected reward equal to $R(\omega^P)/(1 - \delta)$ is obtained and there are no further rewards; (ii) rewards continue to accrue at all subsequent periods after the irreversible choice is made, with each expected reward equal to $R(\omega^P)$. Under the latter, $R(\omega^P)/(1 - \delta)$ denotes the discounted expected payoff from the stream of future rewards.

Now suppose that each arm's states can be partially ordered, based on the number of times that an arm has been activated. Formally, suppose that the set Θ takes the form $\Theta = \Lambda \times \mathbb{N}$, with $m \in \mathbb{N}$ denoting the number of times the arm has been activated. For any $\omega^P = (\xi, (\lambda, m))$, say that $\hat{\omega}^P$ "follows" ω^P if and only if $\hat{\omega}^P = (\xi, (\lambda, \hat{m}))$ for some $\hat{m} \geq m$. We denote this relation by $\hat{\omega}^P \succeq \omega^P$.

Condition 1. An arm of type ξ has *the* **"better-later-than-sooner"** property if, for any $\omega^P = (\xi, (\lambda, m))$, with $m \geq M_\xi$, and any $\hat{\omega}^P \succeq \omega^P$, either $R(\hat{\omega}^P) \geq R(\omega^P)$, or $R(\hat{\omega}^P), R(\omega^P) \leq 0$.

As explained above, the model in the previous sections corresponds to the special case where

$M_\xi = \infty$, for all $\xi \in \Xi$. Weitzman's (1979) Pandora's boxes problem corresponds to the special case where (i) $M_\xi = 1$ for all $\xi \in \Xi$, (ii) the set of boxes is exogenous (which can be captured by assuming that H_{ω^S} is a Dirac), (iii) the reward from opening a box for the first time is equal to the cost of opening the box, the reward from opening a box for the k-th time, with $k > 1$, is equal to a large negative number, and (iv) the reward $R(\omega^P)$ from choosing a box irreversibly is invariant in the number of times the box has been opened and is equal to the box's prize. Clearly, this problem satisfies the "better-later-than-sooner" property.

Another environment that is a special case of the model above is the following. Each time an arm is pulled (without commitment), a reward r is obtained and the decision maker updates her belief about the value R of committing to the arm through the observation of a new signal. In particular, H_{ω^P} could be a Dirac measure assigning probability one to a set of states in which $r = 0$, in which case the (reversible) pull of an arm corresponds to the acquisition of new information about the arm's "retirement" value (i.e., its reward R upon committing to it). Condition 1 then has the following interpretation: at the M_ξ-th pull, the arm is either revealed unprofitable (meaning that its retirement value is revealed to be negative), or at each subsequent pull, favorable news about the arm's retirement value arrives. Note that, under Condition 1, before the decision maker commits to an arm, she may swing between different arms for arbitrarily long horizons.

Theorem 2. *Suppose Condition 1 is satisfied for all $\xi \in \Xi$. Then the index policy χ^* is optimal in the extended model with irreversible choice.*

Proof. To ease the notation, assume that $I_0 = \emptyset$. That is, there are initially no arms, so that all arms arrive through search. It will be evident from the arguments below that the optimality of an index policy does not hinge on this simplifying assumption.

Consider first an environment in which $M_\xi = 0$ for all ξ; we will show below that the result easily extends to an environment in which $M_\xi > 0$ for some ξ. Consider the following *auxiliary environment*, where all choices are *reversible*. Suppose that whenever an arm of type ξ is discovered through search, an additional *auxiliary* arm is also discovered. Whenever pulled, this auxiliary arm yields a fixed flow reward equal to $R(\xi, (\lambda_0, 0))$ where λ_0 is an arbitrary element of a set Λ. Next, suppose that, whenever a physical non-auxiliary arm in state ω^P is pulled for the m-th time, a new auxiliary arm is immediately "discovered". Whenever pulled, this auxiliary arm yields a fixed flow reward equal to $R(\omega^P)$. We say that an auxiliary arm *corresponds to arm j* if it has been discovered as a result of arm j being discovered (through search) or pulled in some prior period. Define the index for the search arm as in (3). Note that this definition is now based on the modified search technology which includes the discovery of auxiliary arms. For each physical arm, given its state ω^P, define the arm's new index as

$$\hat{\mathcal{G}}^P(\omega^P) \equiv \sup_{\pi,\tau} \frac{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s r_s | \omega^P\right]}{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s | \omega^P\right]}, \tag{17}$$

where τ is again a measurable stopping time and where π is a measurable rule selecting between pulling the physical arm or any of the *new* auxiliary arms discovered as the result of *future* pulls of the same physical arm (note that the selection excludes any auxiliary arms discovered by pulling the same arm in previous periods).

It should be clear that the same steps as in the proof of Theorem 1 guarantee the index policy based on the above new indices is optimal in this auxiliary environment.[11] We now interpret pulling an auxiliary arm corresponding to arm j as irreversibly choosing arm j. Note that in this auxiliary environment, once an auxiliary arm is pulled, it will continue to be pulled in all subsequent periods, for its index, which is equal to $R(\omega^P)$, is invariant in subsequent pulls.

For the auxiliary environment to be formally equivalent to the primitive one (in the sense of generating the same dynamics and yielding the decision maker the same payoff) the following property must be true. For any arm j and any period $t > 0$, if an auxiliary arm corresponding to arm j is discovered in period t, then under the index policy in the auxiliary environment the decision maker will never pull any other auxiliary arm corresponding to arm j discovered prior to period t. That is, for each given arm, the "newest" auxiliary arm corresponding to it must have the highest (per-period) expected retirement value among all auxiliary arms corresponding to it. Condition 1 guarantees that this is the case.

Finally, note the the proof immediately extends to an environment in which $M_\xi > 0$ for some ξ by assuming that in the fictitious auxiliary environment described above, an auxiliary arm is discovered only when an arm of type ξ has been pulled at least M_ξ times. ■

5 Other extensions

Below we discuss how the results can accommodate several other extensions.

Relative length of search. In order to accommodate for the possibility that search may create frictions, we assume it competes with the pulling of the existing arms. That is, in each period in which the decision maker searches for new arms, she cannot pull existing ones. In reality, the length of time that search occupies relative to experimentation among the existing arms may differ. For example, the online search for alternative providers of a given service may take seconds, but searching for a potentially suitable candidate for a given position make take longer than an interview.

The assumption that the length of time search occupies is identical to the length of time pulling an arm occupies is innocuous. The results immediately extend to a setting in which the number of periods a pull of an arm, or search, occupy before a different choice can be made differs as a function of its state; in particular, the length of time may differ not only between search and the physical arms, but also across physical arms, and even over time.[12] Because the length of time a pull of a physical arm occupies can be made arbitrary large, by rescaling the payoffs and adjusting the discount factor appropriately, one can make the relative length of time in which the pulling of

[11]The proof must be adjusted to allow auxiliary arms to arrive as the result of pulls of physical arms. Since all the steps are virtually the same, however, the proof is omitted.

[12]The length of time search, or a pull, takes may also be stochastic, and need not be finite. More generally, all of the results may be extended to a semi-Markov environment, where time is not slotted.

physical arms is interrupted to permit search for new arms arbitrarily small. The result thus also apply to problems in which search and experimentation occur "almost" in parallel.

Multiple search possibilities. As illustrated in Example 2, if there are multiple options for search for which the outcome is correlated, an index policy cannot be guaranteed to be optimal.[13] However, the analysis readily extends to an environment in which there are multiple search possibilities with independent outcomes, by allowing for the possibility of multiple "search arms". For example, a researcher may choose in which field to search for a new project. A department with a single new faculty position may choose in which field to search for candidates. In an application we discuss in Section 6, a platform matching agents from two sides of a market may choose to solicit additional agents from either side. The analysis can also be extended to allow the results of search to include not just physical arms but also new search possibilities.

No discounting. The proofs of Theorems 1 and 2 rely on the assumption that $\delta < 1$. As mentioned in Section 4, an important special case of our analysis is an extension of Weitzman's Pandora's boxes problem with search for new boxes. Most applications of Weitzman's framework assume no discounting (i.e., $\delta = 1$). Our results extend to this case. The reason is that, as noted in Olszewski and Weber (2015), the Pandora's boxes problem with $\delta = 1$ is a special case of a setting with undiscounted "target processes", in which upon an arm reaching a certain (target) state, rewards no longer accrue. A well known result for such problems is that the finiteness they impose allows to take the limit as $\delta \to 1$ (see, e.g., Dimitriu, Tetali, and Winkler, 2003).

We note, however, that, as for the classical bandit problem with fixed arms, outside of this class of target processes, the optimality of an index policy does not generally extend to the case of $\delta = 1$.

6 Applications

6.1 Search engine design

Sponsored search advertising, where a set of sponsored links is displayed accompanying the results of consumers' search queries online, accounts for a large fraction of Internet advertising revenues (see, e.g., Edelman, Ostrovsky and Schwarz (2007)). Despite the importance that sponsored search occupies in modern business activities, the few models of online consumer search that have been developed remain quite restrictive. For example, the pertinent literature has typically assumed that consumers click ads sequentially in the order they are displayed, and that click-through-rates depend on positions but not on the ads displayed at the various positions. Such assumptions, however, do not appear to square well with empirical observations.[14]

An important feature of online consumer search is that the consumer is presented with a potentially huge amount of search results, which are displayed in a sequence across multiple pages.

[13] For example, such correlation arises naturally in an environment in which the DM can choose how much to invest in search, with different levels of investment corresponding to different "intensities" of search.

[14] For example, in an empirical analysis of consumer search in online advertising markets using data from Microsoft Live, Jeziorski and Segal (2015) find that almost half of the users who click on a link do not click in sequential order of positions, and that CTRs do depend on the identity of competing ads.

In fact, most of the options are initially invisible, and require the consumer to incur the (time, or mental) cost of moving to the next page or scrolling further down the list to view additional alternatives. Clearly, it is unrealistic to assume that a consumer reads all of the results. Instead, the set of search results that a consumer reads (and considers clicking on) is endogenous.[15]

Our model can be applied to such an environment as a first step toward a better understanding of consumer search in these markets. We apply it as follows. When a consumer enters a query, a list of results is presented in a sequence. Reading the text displayed in a result is costly, and adds the result to the pool of alternatives under consideration. The consumer may also click on one of the results already read (and hence in the consideration set) to access the corresponding page (also at a cost) in order to learn her valuation for the corresponding good. At any point in time, the consumer can then stop and make a purchase among those results she has clicked on. To recap, at each stage the consumer can either read the information displayed in response to the search query (this amounts to searching), click on one of the texts she read to be redirected to the corresponding vendor's webpage (this corresponds to exploring an alternative), or purchase a product, in which case the decision problem ends (this corresponds to an irreversible choice, as examined in Section 4).[16]

This formulation therefore corresponds to an extension of Pandora's boxes problem where the set of boxes (the search results) is endogenous, and is determined by the combination of what the platform displays and of the consumer's search (i.e., reading) behavior. In particular, it is a special case of the model in Section 4. Condition 1 is satisfied in this environment ($M_\xi = 1$ for all $\xi \in \Xi$ and all uncertainty about each product is resolved after the first click). As a result, Theorem 2 applies and the consumer's optimal policy (reading/clicking/purchasing) takes the form of an index policy.

Once the consumer's behavior is described, the model can be used to endogenize the click-through-rates, as well as the probability the consumer selects the various products. The analysis may also be used to explain why and when higher positions imply higher CTRs, a property that has been exogenously assumed in existing models, but which has not been formally micro-founded. Finally, the analysis can be used to endogenize the various externalities across positions, by allowing the click-through-rates (and the eventual purchasing probabilities) to depend on the identities of the ads displayed in the various positions. Once at hand, these results can potentially be used for search-engine and auction design.

6.2 Dynamic allocation problems

In many dynamic allocation problems, the set of agents competing for an indivisible good is endogenous. A firm that repeatedly procures a service from a set of possible suppliers may choose to expand its supplier base in response to new needs or past outcomes. A seller of a house may

[15] For example, approximately 70 percent of Amazon users do not visit pages other than the first one.

[16] Note that this formulation implicitly assumes the consumer does not click on a result without having read its text first, and that a purchase cannot be made without having clicked on the link directing the consumer to the vendor's webpage. Both assumptions seem quite natural in this context.

solicit additional bidders in response to past offers. Such problems offer another application for our model, with each arm corresponding to a strategic agent and with search corresponding to the principal's decision to enlarge the set of available agents. The type of arm ξ may then capture intrinsic characteristics of an agent, while the time-varying component θ may capture signals the principal received about her payoff from transacting with the agent (such signals can also be endogenous as in the case of bids submitted by the agents).

For example, consider the following problem of a seller of an indivisible, durable, good to one of countably infinitely many potential buyers. The seller's payoff from selling to buyer $i \in \mathbb{N}$ at price p_i in period t is equal to $\delta^t(p_i - \gamma_i)$, where δ is the common discount factor, and where γ_i is a buyer-specific cost. Buyer i's payoff from purchasing the good at price p_i, $s \geq 0$ periods after becoming aware of the seller's product, is $\delta^s(v_i - p_i)$, where v_i denotes buyer i's gross value for the good. Buyers become aware of the seller's product stochastically over time, according to the Poisson process described below. Suppose each (v_i, γ_i) is drawn from a distribution $F_i(\cdot)$, independently across buyers. At each period, the seller can (a) inspect one of the buyers who has made an offer and has not been inspected yet, (b) accept one of the inspected offers, or (c) search for new buyers. Inspecting buyer i's offer costs the seller σ_i, and perfectly reveals γ_i to the seller.[17]

In this environment, searching for new buyers can be interpreted as the decision to invest in marketing or other activities that raise the awareness of potential buyers about the seller' product. For example, each search may come at a cost c, and may bring n new buyers, drawn from a Poisson distribution with rate λ, whose magnitude may depend on past search outcomes. In this setting, it is natural to assume that when a buyer becomes aware of the seller's product (e.g., by being reached by an ad), she does not know how many periods the seller has been in the market and/or how many buyers are also aware of the seller's product, as in Abreu and Brunnermeier (2003).

The results above can then be used to shed light on the joint determination of the principal's allocation and solicitation policy, and of the agents' bidding behavior.

6.3 Two-sided markets with side-specific solicitation

Another interesting application is the analysis of pricing and marketing decisions in markets where platforms mediate the interactions among agents from various sides of the market. Our results may shed light on platforms' strategies, especially in the early stages when building a user base is critical.

The possibility of bringing in new agents from various sides, however, poses interesting challenges. Consider a platform mediating interactions between sellers (from side A) and buyers (from side B). To capture the relevant capacity constraints in the simplest possible way, suppose a single match between a buyer and a seller can be accommodated in each period. In this application, an arm corresponds to a potential match between a buyer and a seller who are "on board", i.e., in the platform's consideration set. Over time, in response to the evolution of the match values, the platform may decide to invest to bring additional buyers or sellers on board. For simplicity, suppose a single buyer or a single seller may be introduced at each time.

[17] That is, the seller observes the offers made by the buyers for free, but must inspect them to learn her idiosyncratic cost of serving them.

The possibility of adding either a buyer or a seller requires allowing for two distinct search arms. While independence of search outcomes across the two sides may be natural, the number of new potential matches introduced by adding a buyer (alternatively, a seller) is equal to the number of sellers (alternatively, buyers) already on board. Thus, the two search technologies are clearly not independent. For example, suppose that at some time t there are n_A sellers and n_B buyers on board. If the platform brings a new seller on board in period t then, in period $t+1$, soliciting a new buyer brings $n_A + 1$ new potential matches. Thus, the search technology corresponding to the solicitation of buyers changes state from period t to period $t+1$, even when no buyers are solicited in period t. Using arguments similar to those in Theorem 2, however, one can show that this difficulty is resolved under the condition that the value of each new potential match is non-negative in expectation. Under such an assumption, more new matches are better than fewer. This property plays a role similar to the "better-later-than-sooner" property in Section 4, ensuring the platform never wishes to utilize search in one of its "previous states", which in turn guarantees the optimality of an index policy.

7 Concluding remarks

The paper provides a characterization of the optimal policy for a new class of problems in which the set of alternatives to explore changes endogenously over time as the result of search. The solution takes the form of an index policy where each alternative is assigned an index equal to Gittins (1979)'s, whereas search is assigned a new index that accounts for the way the decision maker selects among the new alternatives that search brings and future searches. The new index is equal to the expected sum of the rewards, net of the search costs, under a rule that selects among newly discovered alternatives and search, stopping when a state is reached in which (a) the new index of search (defined in a recursive manner) and (b) the indices of all new alternatives discovered after search was initiated drop below the value of the search index at the time search was initiated. Importantly, the index accounts for the fact that between the time search is launched and the stopping time in the definition of the index, search may alternate between the exploration of the newly found alternatives and additional search, in a way that maximizes the average sum of the net rewards, per unit of average discounted time.

Equipped with this characterization, we then show that when the search technology is (weakly) improving, search implies "replacement" of all alternatives available at the time search is launched. When, instead, the search technology deteriorates over time (as it is the case when there are finitely many alternatives that search can bring), the existing alternatives are put on hold and the decision maker can come back to them at later periods.

We also show that the optimal policy continues to be an index policy in certain enriched settings where, in addition to exploring the existing alternatives and searching for new ones, the decision maker can commit irreversibly to one of the alternatives, triggering a discrete change in payoffs. This setting admits, as a special case, a generalization of Weitzman (1979)'s Pandora boxes problem in which the set of boxes changes endogenously over time as the result of search.

The model is quite flexible and can be used in a variety of economic applications in which search, experimentation, or sequential learning are important features of the problem under examination.

References

Abreu, D. and M. Brunnermeier (2003). Bubbles and Crashes, *Econometrica*, 71(1), 173-204.

Austen-Smith, D. and C. Martinelli (2018). Optimal exploration. Working paper.

Bergemann, D., and J., Välimäki (2008). Bandit problems. In: Durlauf, S., Blume, L. (Eds.), The New Palgrave Dictionary of Economics, 2nd Edition. Vol. 1. Palgrave Macmillan, New York, pp. 336–340.

Brezzi, M., and T. L. Lai (2000). Incomplete learning from endogenous data in dynamic allocation. *Econometrica*, 68(6), 1511-1516.

Choi, M., and L. Smith (2016). Optimal sequential search among alternatives. Working paper.

Dumitriu, I., Tetali, P., and P. Winkler (2003). On playing golf with two balls. *SIAM Journal on Discrete Mathematics*, 16(4), 604-615.

Doval, L. (2018). Whether or not to open Pandora's box. *Journal of Economic Theory* 175, 127–158.

Edelman, B., Ostrovsky, M., and M. Schwarz (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, 97(1), 242-259.

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

Gittins, J. and D. Jones (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani (Ed.), *Progress in Statistics*, pp. 241-266. Amsterdam, NL: North- Holland.

Gossner, O., Steiner, J., and C. Stewart (2019). Attention, please!. Working paper.

Hörner, J., and A. Skrzypacz (2017). Learning, Experimentation, and Information Design. *Advances in Economics and Econometrics*, 1, 63-98.

Jeziorski, P., and I. Segal (2015). What makes them click: Empirical analysis of consumer demand for search advertising. *American Economic Journal: Microeconomics*, 7(3), 24-53.

Ke, T. T., Z.-J. M. Shen, and J. M. Villas-Boas (2016). Search for information on multiple products. *Management Science* 62 (12), 3576-3603.

Ke, T. T. and J. M. Villas-Boas (2019). Optimal learning before choice. *Journal of Economic Theory* 180, 383-437.

Keller, G., and A. Oldale (2003). Branching bandits: a sequential search process with correlated pay-offs. *Journal of Economic Theory*, 113(2), 302-315.

Mandelbaum, A. (1986). Discrete multi-armed bandits and multi-parameter processes. *Probability Theory and Related Fields*, 71(1), 129-147.

Olszewski, W., and R. Weber (2015). A more general Pandora rule? *Journal of Economic Theory* 160, 429–437.

Pandey, S., Chakrabarti, D., and D. Agarwal (2007). Multi-armed bandit problems with dependent arms. *In Proceedings of the 24th international conference on Machine learning, ACM*, pp. 721-728.

Robbins, H. (1952). Some aspects of the sequential design of experiments." *Bulletin of the American Mathematical Society*, 58(5): 527–35.

Rothschild, M. (1974). A two-armed bandit theory of market pricing, *Journal of Economic Theory* 9 (1974) 185–202.

Varaiya, P., Walrand, J., and C. Buyukkoc (1985). Extensions of the multiarmed bandit problem: The discounted case. *IEEE transactions on automatic control*, 30(5), 426-439.

Weiss, G. (1988). Branching bandit processes. *Probability in the Engineering and Informational Sciences*, 2(3), 269-278.

Weitzman, M. (1979). Optimal search for the best alternative. *Econometrica* 47 (3), 641–654.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2), 143-149.

——— (1981). Arm-acquiring bandits. *The Annals of Probability*, 9(2), 284-292.

Appendix

Proof of Lemma 2. First note that the index corresponding to the auxiliary arm is equal to M. Hence, if $M \leq \mathcal{G}^S(\omega^S)$, the index policy will start by selecting search. If, instead, $M > \mathcal{G}^S(\omega^S)$, the index policy will select the auxiliary arm forever, yielding a discounted expected net reward equal to M. To see why in this case $M > V^S(\omega^S|e(\omega^S) \vee e(\omega^A_M))$, observe that the payoff $V^S(\omega^S|e(\omega^S) \vee e(\omega^A_M))$ from starting with search and then following an index policy in each subsequent period is

$$V^S(\omega|e(\omega^S) \vee e(\omega^A_M)) = \mathbb{E}\left[\sum_{s=0}^{\hat{\tau}-1} \delta^s (r_s^{\hat{\pi}} - c_s^{\hat{\pi}}) + \frac{\delta^{\hat{\tau}}}{1-\delta} M | \omega^S \right],$$

for some stopping and selection rules $\hat{\tau}, \hat{\pi}$, since once the auxiliary arm is selected under the index policy, it will be selected in all subsequent periods. By definition of $\mathcal{G}^S(\omega^S)$,

$$M > \mathcal{G}^S(\omega^S) = \sup_{\pi,\tau} \frac{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s (r_s^{\pi} - c_s^{\pi}) | \omega^S\right]}{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s | \omega^s\right]} \geq \frac{\mathbb{E}\left[\sum_{s=0}^{\hat{\tau}-1} \delta^s (r_s^{\hat{\pi}} - c_s^{\hat{\pi}}) | \omega^S\right]}{\mathbb{E}\left[\sum_{s=0}^{\hat{\tau}-1} \delta^s | \omega^s\right]},$$

and rearranging,

$$M\mathbb{E}\left[\sum_{s=0}^{\hat{\tau}-1} \delta^s | \omega^s\right] > \mathbb{E}\left[\sum_{s=0}^{\hat{\tau}-1} \delta^s (r_s^{\hat{\pi}} - c_s^{\hat{\pi}}) | \omega^S\right].$$

Therefore,

$$V^S(\omega|e(\omega^S)\vee e(\omega^A_M)) = \mathbb{E}\left[\sum_{s=0}^{\hat{\tau}-1}\delta^s(r_s^{\hat{\pi}}-c_s^{\hat{\pi}})+\frac{\delta^{\hat{\tau}}M}{1-\delta}|\omega^S\right] < \mathbb{E}\left[M\sum_{s=0}^{\hat{\tau}-1}\delta^s+\frac{\delta^{\hat{\tau}}M}{1-\delta}|\omega^S\right] = M.$$

Similar arguments establish Condition (12). □

Proof of Lemma 3. Given the state of the system $\mathcal{S}_0 \vee e(\omega^A_M)$ and fixing $\omega^S \in \Omega^S$ to be the state of the search arm, as specified in $\mathcal{S}_0$, we have that

$$\begin{aligned}
&D^S(\omega^S|\mathcal{S}_0\vee e(\omega^A_M)) = \\
&\quad \mathcal{V}(\mathcal{S}_0)+\int_0^M \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)}\mathrm{d}v+\mathbb{E}\left[\tilde{c}|\omega^S\right] \\
&\qquad -\delta\mathbb{E}\left[\mathcal{V}(S_0\backslash e(\omega^S)\vee e(\tilde{\omega}^S)\vee W^P(\tilde{\omega}^S))+\int_0^M \mathbb{E}\delta^{\kappa(v|S_0\backslash e(\omega^S)\vee e(\tilde{\omega}^S)\vee W^P(\tilde{\omega}^S))}\mathrm{d}v|\omega^S\right],
\end{aligned} \tag{18}$$

where the equality follows from combining (9) with (10). Similarly,

$$\begin{aligned}
D^S(\omega^S|e(\omega^S)\vee e(\omega^A_M)) = \mathcal{V}(e(\omega^S)) &+ \int_0^M \mathbb{E}\delta^{\kappa(v|e(\omega^S))}\mathrm{d}v+\mathbb{E}\left[\tilde{c}|\omega^S\right] \\
&-\delta\mathbb{E}\left[\mathcal{V}(e(\tilde{\omega}^S)\vee W^P(\tilde{\omega}^S)))+\int_0^M \mathbb{E}\delta^{\kappa(v|e(\tilde{\omega}^S)\vee W^P(\tilde{\omega}^S))}\mathrm{d}v|\omega^S\right].
\end{aligned} \tag{19}$$

Differentiating (18) and (19) with respect to M, using the independence across the physical and search arms, and the property that $\kappa(v|\mathcal{S}^1\vee\mathcal{S}^2)=\kappa(v|\mathcal{S}^1)+\kappa(v|\mathcal{S}^2)$, it is easily verified that

$$\frac{\partial D^S(\omega^S|\mathcal{S}_0\vee e(\omega^A_M))}{\partial M} = \mathbb{E}\delta^{\kappa(M|S_0\backslash e(\omega^S))}\frac{\partial D^S(\omega^S|e(\omega^S)\vee e(\omega^A_M))}{\partial M}. \tag{20}$$

That is, the improvement in $D^S(\omega^S|\mathcal{S}_0\vee e(\omega^A_M))$ as a result of a slight increase in the reward M from the auxiliary arm is the same as in a setting with only search and the auxiliary arm, $D^S(\omega^S|e(\omega^S)\vee e(\omega^A_M))$, discounted by the expected time it takes under the index policy till there are no indices with value strictly higher than M, in an environment with only physical arms in state $\mathcal{S}^P_0$ where $\mathcal{S}^P_0$ is the same state of the physical arms as in $S_0\backslash e(\omega^S)$. Similar arguments imply that, for any $\omega^P\in\{\hat{\omega}^P\in\Omega^P:\mathcal{S}_0(\hat{\omega}^P)>0\}$,

$$\frac{\partial D^P(\omega^P|\mathcal{S}_0\vee e(\omega^A_M))}{\partial M} = \mathbb{E}\delta^{\kappa(M|\mathcal{S}_0\backslash e(\omega^P))}\frac{\partial D^P(\omega^P|e(\omega^P)\vee e(\omega^A_M))}{\partial M}. \tag{21}$$

Let $M^*\equiv\max\{\mathcal{G}^*(\mathcal{S}^P_0),\mathcal{G}^S(\omega^S)\}$. Integrating (20) over the region $(0,M^*)$ of rewards of the auxiliary arm and rearranging, we have that

$$
\begin{aligned}
D^S(\omega^S|\mathcal{S}_0 \vee e(\omega_0^A)) &= D^S(\omega^S|\mathcal{S}_0 \vee e(\omega_{M^*}^A)) - \int_0^{M^*} \mathbb{E}\delta^{\kappa(v|S_0\backslash e(\omega^S))} \frac{\partial D^S(\omega^S|e(\omega^S) \vee e(\omega_v^A))}{\partial v} \mathrm{d}v \\
&= D^S(\omega^S|\mathcal{S}_0 \vee e(\omega_{M^*}^A)) - \left[\mathbb{E}\delta^{\kappa(v|S_0\backslash e(\omega^S))} D^S(\omega^S|e(\omega^S) \vee e(\omega_v^A))\right]_0^{M^*} \\
&+ \int_0^{M^*} D^S(\omega^S|e(\omega^S) \vee e(\omega_v^A)) \mathrm{d}\mathbb{E}\delta^{\kappa(v|S_0\backslash e(\omega^S))} \\
&= D^S(\omega^S|\mathcal{S}_0 \vee e(\omega_{M^*}^A)) - D^S(\omega^S|e(\omega^S) \vee e(\omega_{M^*}^A)) \\
&+ \int_0^{M^*} D^S(\omega^S|e(\omega^S) \vee e(\omega_v^A)) \mathrm{d}\mathbb{E}\delta^{\kappa(v|S_0\backslash e(\omega^S))}
\end{aligned}
$$

where the second equality follows from integration by parts and the third from the fact that $\kappa(M^*|S_0\backslash e(\omega^S)) = 0$ along with the fact that, because an arm with fixed reward equal to zero (arm 0) is always present, $\kappa(0|S_0\backslash e(\omega^S)) = \infty$. The existence of such an "opt-out" arm also implies that $D^S(\omega^S|\mathcal{S}_0 \vee e(\omega_0^A)) = D^S(\omega^S|\mathcal{S}_0)$. It can also easily be verified that $D^S(\omega^S|\mathcal{S}_0 \vee e(\omega_{M^*}^A)) = D^S(\omega^S|e(\omega^S) \vee e(\omega_{M^*}^A))$.[18] Therefore, we have

$$
D^S(\omega^S|\mathcal{S}_0) = \int_0^{M^*} D^S(\omega^S|e(\omega^S) \vee e(\omega_v^A)) \mathrm{d}\mathbb{E}\delta^{\kappa(v|S_0\backslash e(\omega^S))}. \tag{22}
$$

Similar arguments yield

$$
D^P(\omega^P|\mathcal{S}_0) = \int_0^{M^*} D^P(\omega^P|e(\omega^P) \vee e(\omega_v^A)) \mathrm{d}\mathbb{E}\delta^{\kappa(v|S_0\backslash e(\omega^P))}. \tag{23}
$$

To complete the proof of Lemma 3, consider first the case where the index policy at $\mathcal{S}_0$ specifies starting by pulling a physical arm; i.e., $M^* \neq \mathcal{G}^S(\omega^S)$. Then Condition (13) in the lemma follows directly from (22) by noting that $M^* = \mathcal{G}^*(\mathcal{S}_0^P)$. Also observe that, for any state $\omega^P \in \Omega^P$, if $M^* > \max\{\mathcal{G}^*(\mathcal{S}_0^P\backslash e(\omega^P)), \mathcal{G}^S(\omega^S)\}$ then $M^* = \mathcal{G}^P(\omega^P)$, in which case the integrand $D^P(\omega^P|e(\omega^P) \vee e(\omega_v^A))$ in (23) is equal to zero over the entire region $[0, \mathcal{G}(\omega^P)]$ and hence over $[0, \max\{\mathcal{G}^*(\mathcal{S}_0^P\backslash e(\omega^P)), \mathcal{G}^S(\omega^S)\}]$. That is, in this case, Condition (14) in the lemma clearly holds. Next, pick any state $\omega^P \in \Omega^P$ such that $\mathcal{G}^P(\omega^P) < M^*$. Condition (14) then follows directly from (23) by noting that, in this case, $M^* = \max\{\mathcal{G}^*(\mathcal{S}_0^P\backslash e(\omega^P)), \mathcal{G}^S(\omega^S)\}$.

Next, consider the case where the index policy at $\mathcal{S}_0$ specifies starting by pulling the search arm; i.e., $M^* = \mathcal{G}^S(\omega^S)$. Then, for any $\omega^P \in \Omega^P$, $\max\{\mathcal{G}^*(\mathcal{S}_0^P\backslash e(\omega^P)), \mathcal{G}^S(\omega^S)\} = M^*$, in which case Condition (14) in the lemma follows directly from (23). That Condition (13) in the lemma also holds, follows from the fact that, in this case, $D^S(\omega^S|\mathcal{S}_0) = 0$ and the integrand $D^S(\omega^S|e(\omega^S) \vee e(\omega_v^A))$ in (22) is equal to zero over the entire region $[0, \mathcal{G}^S(\omega^S)]$. □

[18] This follows immediately from the observation that $\mathcal{V}(\mathcal{S}_0 \vee e(\omega_{M^*}^A)) = \mathcal{V}(e(\omega^S) \vee e(\omega_{M^*}^A)) = M^*$, and similarly $\mathbb{E}\left[\mathcal{V}\left(\mathcal{S}_0\backslash e(\omega^S) \vee e(\tilde{\omega}^S) \vee W^P(\tilde{\omega}^S) \vee e(\omega_{M^*}^A)\right)|\omega^S\right] = \mathbb{E}\left[\mathcal{V}\left(e(\tilde{\omega}^S) \vee W^P(\tilde{\omega}^S) \vee e(\omega_{M^*}^A)\right)|\omega^S\right]$. Intuitively, any physical arm with index strictly lower than M^* will never be chosen given the presence of the auxiliary arm with reward M^*.