On Optimal Scheduling*

Kfir Eliaz† Daniel Fershtman‡ Alexander Frug§

January 27, 2024

Abstract

We consider a decision-maker sequentially choosing among alternatives when periodic payoffs depend on both chosen and unchosen alternatives in that period. We show that when flow payoffs are the sum or product of payoffs from chosen and unchosen alternatives, the optimal policy is an index policy. We characterize key properties of the optimal dynamics and present an algorithm for computing the indices explicitly. Furthermore, we use the results to generalize Weitzman's (1979) classic Pandora's-boxes problem to allow for complementarities. We illustrate the framework's usefulness through applications, including decision problems with disappearing alternatives, repeated bargaining, dynamic supervision, and dynamic occupational choice.

Keywords: Scheduling, Time management, Multitasking, Task juggling, Multi-armed bandits

*Eliaz acknowledges financial support from the Pinhas Sapir Center for Development and from the Foerder Institute. Frug acknowledges the financial support of the Spanish Ministry of Science and Innovation through the grants: AEI/FEDER, UE - PGC2018-098949-B-I00, SEV-2015-0563, and CEX2019-000915-S.

†Eitan Berglas School of Economics, Tel-Aviv University and David Eccles School of Business, University of Utah. kfire@tauex.tau.ac.il.

‡Eitan Berglas School of Economics, Tel-Aviv University. danielfer@tauex.tau.ac.il.

§Department of Economics and Business, Universitat Pompeu Fabra and Barcelona Graduate School of Economics. alexander.frug@upf.edu.

1 Introduction

As individuals, we are constantly engaged in scheduling tasks and in deciding when to switch from one activity to another. Oftentimes, we are affected by the outcomes of tasks even in periods in which we do not attend to them. For instance, the success of a manager who allocates his time between different teams/departments under his control depends on the overall performance of all of them. While the manager is attending to one of his teams, his overall success continues to depend on the performance of all the other teams. In other words, many scheduling problems have the feature that in each period, the decision-maker receives a payoff from *each* available activity – not just the one he actively engages in. The goal of this paper is to develop a methodology to analyze such decision problems and to illustrate its applicability.

We consider a decision-maker (DM) who faces a finite number of tasks/alternatives, and in each period must decide which task to choose. The per-period payoff from a given task depends on its state and on whether the DM attends to it in the current period, in which case it generates an "active payoff," or not, in which case it generates a "passive payoff." The state of each given task may change only during those periods in which the DM attends to the task, in which case its new state is drawn from a distribution that depends on the task's previous state. The DM's overall flow payoff is a function of *all* of the task-specific per-period payoffs. This captures environments where several activities generate a payoff even when left unattended, as well as environments in which the flow payoff is generated only from the chosen task, but where this payoff depends on the states of the unchosen tasks.

We focus on two canonical cases: the *additive* case, in which the DM's flow payoff is a sum of the payoffs of all tasks – the active payoff from the chosen task and the passive payoffs from the unchosen tasks – and the *multiplicative* case, in which the DM's flow payoff equals the product of the active and passive payoffs. In both cases, the classic Gittins index solution is not applicable since the DM's flow payoff depends on the states of *all* tasks.[1] Nevertheless, the DM's optimal strategy *is* characterized by an index policy. By this we mean that, in both the additive case and the multiplicative case, there exists a function that assigns to each task a score, *which depends only on the characteristics of that particular task,* and a ranking over such scores, such that in each period, given the states of all tasks, it is optimal to pick the highest ranked task. The score of each task, or its index, generalizes the Gittins index underlying the optimal policy in the classic bandit problem. Whereas the Gittins index is the solution to a maximization problem over stopping rules, maximizing the average expected discounted payoff per unit of expected discounted time, the indices in our problem must also account for changes in the "externality" that an alternative imposes on others via its passive payoff. Our analysis goes beyond deriving these indices: we characterize key properties of the dynamics under the optimal policy, and in Appendix A, present an algorithm for calculating the indices *explicitly* in both the additive and multiplicative case.

[1] That is, such problems do not fall into the classic paradigm of the multi-armed bandit problem, in which the DM receives a payoff *only* from the arm he pulls *and only* when he pulls it.

The characterization of the optimal policy in the above two classes of scheduling problems opens the door to analyzing a new rich set of economic applications. We demonstrate this in three types of environments.

First, we generalize the classic Pandora's-boxes problem of Weitzman (1979), and its solution, to allow for complementarities among alternatives. We show that the optimal policy is based on new reservation prizes that govern the order in which boxes are opened, when to stop, and which inspected box to recall upon stopping. We then characterize the effect of a first-order stochastic shift in the distribution of a box's externality on its reservation prize, and on the duration of search in the case of ex-ante identical alternatives. We illustrate the optimal policy and the comparative statics in an example of an entrepreneur searching for investors to implement a new idea, while facing the risk that his idea might be stolen by investors who are not chosen.

Next, we show that in the presence of only two alternatives, our framework enables us to solve scheduling problems that would otherwise be intractable. We illustrate this with two examples, one for the additive case and another for the multiplicative case. In the former case we consider the problem of a firm who needs to decide each period which supplier to choose, where the surplus from the interaction is divided according to a bargaining solution that takes into account the firm's outside option of choosing the competing supplier. For the latter, multiplicative case, we analyze an example where one of the alternatives may cease to be available with some probability.

Finally, we consider an environment where the state of an alternative encodes the number of times that it was chosen. This imposes additional structure on the problem that further simplifies the derivation of the indices. We illustrate this with two examples: supervising agents with stochastic costs of effort in the multiplicative case and career paths and mobility between sectors in the additive case.

All proofs appear in the appendices.

Related literature. The multiplicative case described above was first studied in the statistics literature by Peter Nash (see Nash, 1980), who argued that the optimal policy takes the form of an index policy. Our paper complements Nash's by making the following contributions. First, we introduce the additive case and give a unified proof that covers both cases (Theorem 2). Second, we fully characterize the optimal stopping rules underlying the definition of the indices in the two cases, additive and multiplicative (Theorem 1); Nash asserted that the properties of the optimal stopping rule extend from the standard bandit problem to the multiplicative case, but did not prove this claim. Theorem 1 shows that this is not immediate, and also plays a key role in the analysis that follows. Third, we characterize several key properties which the optimal policy must satisfy (Theorem 3, Corollary 1, Proposition 1, Corollary 2 and Proposition 8), and the implications of such properties for the incentives to backload/frontload decisions given the nature of the externalities alternatives exhibit. Fourth, toward developing a practicable framework, we introduce an intuitive algorithm that allows to explicitly compute the indices in both the additive and multiplicative case, and further elucidates the dynamics under the optimal policy in each of these cases.

One of our main goals is to showcase the new rich set of economic applications that can be analyzed using our framework. For example, we use the results to extend the classic Pandora's-boxes problem (Weitzman, 1979), which plays a central role in the search literature, to allow for complementarities, and illustrate the usefulness of this extension. We also apply the framework to study decision problems with vanishing alternatives, repeated bargaining between a firm and its suppliers, on-the-job-training, dynamic supervision, and dynamic occupational choice.

The problem of allocating time/attention between tasks has been previously studied in the literature under different frameworks. In a series of papers, Coviello et al. (2014, 2015) study the problem of a DM who faces a growing queue of tasks that arrive at an exogenous rate. In their 2014 paper, the authors characterize the production function, which relates the output rate to the effort rate (which governs completion time) and the activation rate (at which tasks are started). Their 2015 paper applies this production function to a dataset of judges' handling of court cases to estimate the effect of increased case load. Bray et al. (2016) model how a judge schedules cases as a classic multi-armed bandit problem, and argue that prioritizing the oldest hearing (case) is optimal when the case completion hazard rate function is decreasing (increasing). Using data on Italian judges, they estimate that a switch from prioritizing the oldest hearing to prioritizing the oldest case greatly decreased average case duration. The present paper complements the work of these authors by considering a different framework where the set of tasks is given, and the DM decides which task to attend to in each period, taking into account how his payoffs depend on both chosen and unchosen tasks. Our main contribution is a characterization of the DM's optimal strategy in the additive and multiplicative cases.

In their classic paper, Radner and Rothschild (1975) study the problem of a DM who needs to allocate a unit of attention in each period among a given set of tasks. The output of an attended (unattended) task increases (decreases) as a function of the amount of effort allocated to it. Instead of deriving the optimal strategy, the authors compare the survival probability of several heuristics (the probability that the outputs of all tasks remain above some threshold).

Our model is naturally related to the multi-armed bandit framework, which has been widely applied in a variety of fields in economics.[2] Due to the dependence of payoffs on the states of all tasks, however, our model does not fall under this classic framework. This feature also distinguishes our analysis relative to the classic literature on "learning by experimentation" (e.g., Keller et al., 2005). The DM's optimal policy in our problem takes the form of an index policy, and in this sense generalizes the key IIA property that has contributed to the applicability of the classic multi-armed bandit framework.

While our model is motivated by environments where learning about alternatives can be linked to the flow payoffs they generate, it is related to a small literature studying the problem of a DM that acquires information about multiple attributes of an object before deciding between the object and an outside option. Since the value of the object depends on all its attributes—whether

[2]See Bergemann and Valimaki (2008) for an excellent survey.

inspected or not—this decision problem resembles the one we study. However, it does not fit into our framework because the final payoff of the DM is the maximum of the object's expected value and the value of the outside option. In particular, it is not known whether the optimal strategy in this environment admits an index policy. Notable examples in this literature include Klabjan et al. (2014), who study a DM inspecting a good's attributes before choosing between the object and an outside option. The DM's payoff from the object is a weighted average of the attributes' values, of which he is initially uninformed. He can inspect attributes at a cost, thereby learning their value, before making a decision. Eliaz and Frug (2018) study a related problem, where a seller decides which attributes of an asset to inspect before proposing a price to a buyer, who does not observe the outcome of the seller's inspections.

Several recent papers have studied the problem of a DM that gradually acquires costly information about a set of options before stopping and choosing one of them. These include Ke et al. (2016), Fudenberg et al. (2018), Che and Mierendorff (2019), Ke and Villas-Boas (2019), and Liang et al. (2021). A key difference between these works and ours is that the DM's payoff in our model is a function of the states of all alternatives, whether chosen or not. Additionally, in contrast to our framework, the optimal strategy in these works does not take the form of an index policy.[3]

Similar to the classic multi-armed bandit framework and the papers discussed above, in the present paper the set of alternatives among which the DM alternates is fixed ex ante. Fershtman and Pavan (2020) analyze a model where, in addition to exploring alternatives already in the DM's consideration set, the DM can choose to search for additional alternatives in response to information gathered about existing ones.

2 Model

There are n alternatives that require the attention of a DM. Each period, $t = 0, 1, 2, ...$, the DM can choose at most a single alternative to attend to. The DM can also decide not to choose any alternative. In each period t, if the DM chooses alternative i, his payoff is $U_t(x_{1,t}, ..., x_{n,t}; i)$, where $x_{i,t} \in X_i$ represents the period-t state of alternative i, and each X_i is an arbitrary state space.

We consider two specifications of the payoff U. The first is the additive case,

$$U_t(x_{1,t}, ..., x_{n,t}; i) = u_i(x_{i,t}) + \sum_{j \neq i} v_j(x_{j,t}), \tag{1}$$

and the second is the multiplicative case,

$$U_t(x_{1,t}, ..., x_{n,t}; i) = u_i(x_{i,t}) \prod_{j \neq i} v_j(x_{j,t}), \tag{2}$$

[3]Gossner et al. (2020) also a study a problem in a which a DM sequentially learns about options before choosing one of them. They assume an exogenous stopping rule, which implies that the optimal learning strategy follows an index policy.

where u_i and v_i are bounded functions. We refer to the former specification as the additive case and to the latter specification as the multiplicative case.[4] In the multiplicative case, u_i and v_i are assumed to be non-negative, whereas in the additive case, u_i and v_i may be positive or negative.

The function u_i represents the *active* payoff from alternative i at the time at which the DM chooses it, while v_i is the *passive* payoff in periods in which the DM chooses another alternative. As we will see, v_i can also capture externalities that unchosen alternatives impose on the chosen alternative in situations where an alternative generates a payoff only when chosen. In the additive case, the possibility of not choosing any alternative can be captured by introducing a fictitious alternative whose state remains constant and for which the functions u and v are constant at zero. Similarly, in the multiplicative case, the possibility of not choosing any alternative can be captured by introducing a fictitious alternative whose state remains constant and for which the functions u and v are constant at 1.

The state of an alternative changes only in a period in which the DM chooses it. Specifically, given $x_{i,t}$, if the DM chooses alternative i in period t, the alternative's next state $x_{i,t+1}$ is drawn from the distribution $F_i(\cdot|x_{i,t})$ defined on X_i, and the states of the other alternatives remain unchanged, $x_{j,t+1} = x_{j,t}$. For example, the change in an alternative's state may capture investment of resources or effort, learning about an unknown characteristic, learning-by-doing, or habit formation.

The additive case fits situations in which each alternative generates some output and the DM cares about the total output. In contrast, the multiplicative case captures situations where the outputs of alternatives are interrelated in the sense that when one alternative generates a very low output, the total output is very low. Put differently, in the additive case, the marginal output obtained from choosing an alternative is independent of the states of the other alternatives, whereas in the multiplicative case, it is dependent. For example, the multiplicative case fits situations in which a manager supervises a team that works on independent components of a single project, such that the project is completed successfully if and only if each component of it is completed successfully. In addition, this case can also address the problem of a firm producing a good using a Cobb–Douglas production function

$$U_t = AL_t^{\beta} K_t^{\alpha}$$

that needs to decide in each period whether to invest in labor (L) or capital (K) with the goal of maximizing its expected discounted production.

A *policy* Γ specifies, given the current state of all alternatives $(x_1, ..., x_n)$, which alternative (if any) the decision maker chooses. The DM wishes to maximize the expected discounted stream of payoffs. An *optimal* policy therefore maximizes

$$\mathbb{E}\left(\sum_{s=0}^{\infty} \delta^s U_s(x_{1,s}, ..., x_{n,s}; i) | x_{1,0}, ..., x_{n,0}\right).$$

[4] As will become clear below, the multiplicative case cannot be converted into the additive case by taking a log transformation.

3 Optimal scheduling

Towards a characterization of the DM's optimal policy, we introduce the following notation. Consider an alternative i that is in state x_i in some unspecified period. With a slight abuse of notation, we denote by $x_i^{+\tau}$ the (possibly random) state of that alternative following τ periods in which it is chosen. We use this notation so that we do not need to specify the particular time period in which the alternative was in the initial state x_i. Using this notation, for any state x_i of alternative i, given a (realization-dependent) stopping rule τ, define

$$a_i(x_i, \tau) \equiv \mathbb{E}\left(\delta^\tau v_i(x_i^{+\tau})|x_i\right) - v_i(x_i). \tag{3}$$

The first component $\mathbb{E}\left(\delta^\tau v_i(x_i^{+\tau})|x_i\right)$ represents the expected discounted *passive* payoff of alternative i after τ consecutive periods in which it is chosen, starting from state x_i. Thus, for a given stopping rule τ, the function $a_i(x_i, \tau)$ captures the expected discounted increase in the passive payoff of alternative i, starting at x_i and stopping according to the rule τ.

Additive case. Given the state $x_{j,t} \in X_j$ of alternative j in period t, define an index

$$I_j(x_{j,t}) \equiv \sup_\tau\{I_j(x_{j,t}, \tau)\} \equiv \sup_\tau \left\{ \frac{(1-\delta)\mathbb{E}\left(\sum_{s=0}^{\tau-1} \delta^s u_j(x_{j,t+s})\right) + a_j(x_{j,t}, \tau)}{\mathbb{E}\left(1-\delta^\tau\right)} \right\}, \tag{4}$$

where the sup is taken over all, possibly stochastic, stopping rules (in particular, the stopping rules are with respect to the natural filtration of $\{x_{j,t}\}$). The index I_j is a function only of the state of alternative j, and is independent of any information about the other alternatives. Note that when $v \equiv 0$ for all the alternatives, the payoff in each period is a function of only the state of the alternative that the decision maker chooses, such that (4) reduces to the classic Gittins index of Gittins and Jones (1974). Given the index defined in (4), we define the following policy.

Definition 1. Γ^* *is a policy choosing in each period the alternative with the highest index.*[5]

Under this policy, the decision in each period boils down to a simple comparison of independent indices.

Multiplicative case. Characterizing the optimal policy for this problem requires a distinction between the states of an alternative in which it is "augmenting," and ones in which it is not.

Definition 2. *Say that a state x_i is* augmenting *if there exists a stopping time τ such that $a_i(x_i, \tau) \equiv \mathbb{E}\left(\delta^\tau v_i(x_i^{+\tau})|x_i\right) - v_i(x_i) > 0$. For each alternative i, denote by $A_i \subseteq X_i$ the set of states that are augmenting.*

That is, a state of an alternative is augmenting if there is a stopping time at which its expected discounted passive payoff increases. This property depends on both the process governing the evolution of the states of an alternative, as well as its current state.

[5] In the case of ties between indices, any tie-breaking rule may be specified.

Given the state $x_{i,t} \in A_i$ of alternative i in period t, define the index

$$J_i(x_{i,t}) \equiv \inf_{\tau: a_i(x_{i,t},\tau)>0} \{J_i(x_{i,t},\tau)\} \equiv \inf_{\tau: a_i(x_{i,t},\tau)>0} \left\{ \frac{\mathbb{E}\left(\sum_{s=0}^{\tau-1} \delta^s u_i(x_{i,t+s})|x_i\right)}{a_i(x_{i,t},\tau)} \right\}. \tag{5}$$

For any state $x_{i,t} \notin A_i$ of alternative i with $a_i(x_{i,t},\tau) < 0$, define the index

$$J_i(x_{i,t}) \equiv \sup_\tau \{J_i(x_{i,t},\tau)\} \equiv \sup_\tau \left\{ \frac{\mathbb{E}\left(\sum_{s=0}^{\tau-1} \delta^s u_i(x_{i,t+s})|x_i\right)}{-a_i(x_{i,t},\tau)} \right\}, \tag{6}$$

and if $a_i(x_{i,t},\tau) = 0$, define $J_i(x_{i,t}) = \infty$.

Note that, as in the additive case, the indices in (5) and (6) are independent of any information about all alternatives $j \neq i$. Furthermore, the denominator of the expression in the curly brackets in (5) is strictly positive when $x_{i,t}$ is augmenting, and for states that are not augmenting, the denominator of (6) is non-negative.

A key difference between the indices (5) and (6) is that the optimization in (5) is constrained to stopping times τ for which $a_i(x_i,\tau) > 0$. The reason for this difference is that a state is augmenting if there exists *some* stopping time for which $a_i(x_i,\tau) > 0$, hence in minimizing $J_i(x_{i,t})$ we need to ensure that we remain in the augmenting case. In contrast, a state is non-augmenting if there is *no* stopping time for which $a_i(x_i,\tau) > 0$. Hence, for any τ we pick in the non-augmenting case, $a_i(x_i,\tau) < 0$.

Definition 3. *Define the following order $\succsim$ (with $\succ$ denoting its strict counterpart), according to which the indices of alternatives will be ranked. For any alternatives i, j (including $i = j$):*

1. *If $x_i \in A_i$ and $x_j \notin A_j$, then $J_i(x_i) \succsim J_j(x_j)$.*
2. *If $x_i \in A_i$ and $x_j \in A_j$, then $J_i(x_i) \succsim J_j(x_j)$ if and only if $J_i(x_i) \leq J_j(x_j)$.*
3. *If $x_i \notin A_i$ and $x_j \notin A_j$, then $J_i(x_i) \succsim J_j(x_j)$ if and only if $J_i(x_i) \geq J_j(x_j)$.*

Definition 4. *Denote by Λ^* the policy induced by the order $\succsim$, i.e., the policy that chooses in each period the alternative with the "most preferred" index according to the order $\succsim$.*[6]

As in the case of Γ^*, under the policy Λ^*, the decision of which alternative to choose boils down to a simple comparison of indices across pairs of alternatives, where each alternative's index is a function of its own state only, and independent of any information about other alternatives. For any alternative k, let $\mathcal{I}_k$ denote the index I_k in the additive case and J_k in the multiplicative case. Let $\trianglerighteq$ denote the binary relation $\geq$ in the additive case and $\succsim$ in the multiplicative case. We let $\triangleright$ denote the strict counterpart of $\trianglerighteq$. Finally, let $\mathcal{P}^*$ denote the index policy Γ^* in the additive case, and Λ^* in the multiplicative case.

[6] Ties may be broken according to any prespecified tie-breaking rule.

In order to characterize the optimal policy $\mathcal{P}^*$ we will rely on the following property, which will also be useful for deriving the index in applications. We define the stopping rule τ^* as follows. For any alternative j, beginning at any state x_j, let $\tau_j^*(x_j)$ be the first time at which the index of j becomes *weakly* inferior to $\mathcal{I}_j(x_j)$, according to the order $\trianglerighteq$. That is, $\tau_j^*(x_j)$ is equal to the first period $t > 0$ such that $\mathcal{I}_j(x_j) \trianglerighteq \mathcal{I}_j(x_j^{+t}(x_j))$, where $x_j^{+t}(x_j)$ denotes alternative j's (stochastic) state after being chosen for t periods, starting at state x_j.

The following result shows that the stopping rule τ^* plays a central role in the definition of the indices defined above.

Theorem 1. *For any alternative j and state x_j, $\tau_j^*(x_j)$ attains the value of $\mathcal{I}(x_j)$.*

Nash (1980) did not prove this result, but simply assumed it to be true. As our proof illustrates, establishing the validity of this theorem is not completely obvious. Theorem 1 is a key observation that allows us to characterize the optimal policies for the additive and multiplicative cases, and will also allow us to construct an algorithm for calculating the indices explicitly.

Theorem 2. *The policy $\mathcal{P}^*$ is optimal.*

The proof builds on an interchange argument due to Gittins and Jones (1974). It suffices to show that any policy π^0 that differs from $\mathcal{P}^*$ in period 0 and subsequently coincides with it attains a discounted expected payoff no greater than that of $\mathcal{P}^*$. To show this, starting with any such arbitrary policy π^0, we construct a sequence of modifications of π^0, $(\pi^1, \pi^2, ...)$, such that each modified policy π^k coincides with $\mathcal{P}^*$ for at least the first k periods and attains a weakly higher expected payoff than its predecessor, and furthermore, the expected discounted payoff under π^k converges to the expected discounted payoff under $\mathcal{P}^*$ as $k \to \infty$.

As can be gleaned from the definition of the order $\gtrsim$, the multiplicative case differs from the additive case (as well as the standard bandit problem) in that it requires a classification of states into two classes: augmenting and non-augmenting. The following Theorem highlights the asymmetry between the two classes and reveals general dynamic properties of the optimal policy in the multiplicative case. This is also a new result that was not shown by Nash (1980).

Theorem 3. *Under the optimal policy in the multiplicative case:*

1. *With positive probability, all of the alternatives eventually reach non-augmenting states.*
2. *Once all alternatives are in non-augmenting states, there can be at most a single alternative in an augmenting state thereafter.*

3.1 Intuition and relation to the standard bandit problem

In this subsection we discuss the relation between the indices of the additive and multiplicative case, and their relation to the standard Gittins index. We also provide some economic intuition

for why the optimal strategy takes the form of an index policy, and for the difference between the augmenting and non-augmenting states in the multiplicative case. This discussion and intuition is important for analyzing economic applications. These were not provided in Nash (1980), since he did not consider applications of the index policy in the multiplicative case.

We begin with the additive case, as it is perhaps more intuitive. Although the DM's flow payoff depends on the states of all alternatives—the one he chooses and all the others—Theorem 2 establishes that in each period the DM optimally chooses the alternative with the highest index. This therefore generalizes the IIA property known in the classic bandit framework (that is, the DM prefers to choose alternative i rather than alternative j in a particular period independently of the other alternatives). The indices capture both the direct payoff from choosing an alternative and the indirect effect on payoffs in periods where other alternatives will be selected through v.

The fact that the optimal policy takes such a simple separable form is important for applications. It is useful both for deriving properties of the dynamics and comparative statics under the optimal policy and for computational purposes. Without such separability, the optimal policy, in principle, could still be computed using dynamic programming. However, this would require strong assumptions on the state variables in order to simplify computation due to the "curse of dimensionality."

Rearranging (4) yields the following alternative representation of our index:

$$I_j(x_{j,t}) = \sup_\tau \left\{ \underbrace{\left(\frac{\mathbb{E}\left(\sum_{s=0}^{\tau-1} \delta^s u_j(x_{j,t+s})\right)}{\mathbb{E}(\sum_{s=0}^{\tau-1} \delta^s)} \right)}_{\text{(i)}} - \underbrace{\left(\frac{v_j(x_{j,t}) - \mathbb{E}\left(\delta^\tau v_j(x_{j,t+\tau})\right)}{(1-\delta)\mathbb{E}(\sum_{s=0}^{\tau-1} \delta^s)} \right)}_{\text{(ii)}} \right\}.$$

The indices therefore maximize the difference between two components: (i) the expected discounted payoff per unit of expected discounted time, and (ii) the discounted continuation value of the expected change in the passive payoff per unit of expected discounted time. The first component is the one maximized by the well-known Gittins index. The second component is new, and reflects the fact that when an alternative is not picked, it continues to contribute to the overall payoff, as a function of its state.

Turning to the multiplicative case, the distinction between augmenting and non-augmenting states is not merely a technical one; it is at the heart of the tradeoff between the alternatives. If an alternative is in an augmenting state, it strictly enhances the future benefits from selecting other alternatives—when we take into account discounting—and is therefore currently preferred to alternatives in states that are not augmenting. That is, suppose an alternative i is in a non-augmenting state and generates a higher u than an alternative j, which is in an augmenting state. It is then better to pick i only *after* we choose j, because the increase in the passive payoff of j (which multiplies the payoffs from all the other alternatives) enhances the contribution of alternative i and more than compensates for the delay in choosing it.

Among alternatives in augmenting states, both the direct payoffs that such alternatives are expected to generate (captured by $\mathbb{E}\left(\sum_{s=0}^{\tau-1}\delta^s u_i(x_i^{+s})|x_i\right)$) and the degree to which they are expected to enhance the payoffs from other alternatives in the future (captured by $a_i(x_i,\tau)$) must be weighed. In particular, the index (5) becomes more preferred (that is, its value decreases - note the "inf" in the definition of (5)) the greater is $a_i(x_i,\tau)$ and, perhaps surprisingly, less preferred (that is, its value increases) the greater is $\mathbb{E}\left(\sum_{s=0}^{\tau-1}\delta^s u_i(x_i^{+s})|x_i\right)$. This reflects the fact that, among alternatives in augmenting states, the higher the direct payoffs an alternative generates, the more desirable it is to postpone selecting it in order to allow such payoffs to first be enhanced by choosing alternatives that are augmenting today but generate lower payoffs. Put differently, among alternatives in augmenting states, all things equal, it is desirable (in terms of the ordering of the indices) to *back-load* investment in alternatives with higher direct payoffs.

By contrast, investing in alternatives in non-augmenting states does not enhance the flow payoffs from other alternatives—again, when we take into account discounting. This may be the case simply because $\mathbb{E}\left(v_i(x_i^{+\tau})|x_i\right) - v_i(x_i) < 0$ for any possible stopping time τ. Alternatively, even if an alternative is expected to enhance the future payoffs from investing in the other alternatives—i.e., $\mathbb{E}\left(v_i(x_i^{+\tau})|x_i\right) - v_i(x_i) > 0$ for some τ—it may do so to an extent that is not sufficient to outweigh the opportunity cost of not investing in the other alternatives in the meantime, i.e., $\mathbb{E}\left(\delta^\tau v_i(x_i^{+\tau})|x_i\right) - v_i(x_i) < 0$. Accordingly, among alternatives in non-augmenting states, the index (6) of an alternative is improving (i.e., its value increases) in the direct payoff it generates (captured by $\mathbb{E}\left(\sum_{s=0}^{\tau-1}\delta^s u_i(x_i^{+s})|x_i\right)$) and deteriorating (i.e., its value decreases) in $-a_i(x_i,\tau) = v_i(x_i) - \mathbb{E}\left(\delta^\tau v_i(x_i^{+\tau})|x_i\right) > 0$. In contrast to the intuition for alternatives in augmenting states, among alternatives in non-augmenting states, all things equal, it is desirable (in terms of the ordering of the indices) to *front-load* investment in alternatives with high direct payoffs.

Theorem 3 further highlights the asymmetry between augmenting and non-augmenting states. While there are many natural scheduling problems where all alternatives have only non-augmenting states, there are no environments with alternatives that have only augmenting states. When an alternative is augmenting, the main benefit from choosing it comes from back-loading – enhancing the externality on future payoffs, which will only be collected when *other* alternatives are indeed chosen. Hence, an alternative cannot be augmenting forever. Roughly, this leads to an intuitive two-phase dynamics: in the first phase, the DM *invests in externalities* by selecting augmenting alternatives, and then arrives at an *exploitation phase* during which all alternatives are non-augmenting.

To see how the indices (5)–(6) relate to the index (4), consider the index of an alternative i in

a non-augmenting state $x_{i,t}$. This index satisfies

$$J_i(x_{i,t}) \propto \sup_\tau \left\{ \underbrace{\left(\frac{\mathbb{E}\left(\sum_{s=0}^{\tau-1} \delta^s u_i(x_{i,t+s})\right)}{\mathbb{E}(\sum_{s=0}^{\tau-1} \delta^s)} \right)}_{\text{(i)}} \Big/ \underbrace{\left(\frac{v_i(x_{i,t}) - \mathbb{E}\left(\delta^\tau v_i(x_{i,t+\tau})\right)}{(1-\delta)\mathbb{E}(\sum_{s=0}^{\tau-1} \delta^s)} \right)}_{\text{(ii)}} \right\},$$

maximizing the *ratio* between the two components discussed above: (i) the expected discounted payoff per unit of expected discounted time, and (ii) the net present value of the expected change in the passive payoff, again per unit of expected discounted time. Recall that the first component is the one maximized by the Gittins index, while the second reflects the fact that when an alternative is not chosen, it continues to contribute to the overall payoff, as a function of its state. The index for the augmenting case can be rewritten analogously.

The following Corollary summarizes several special cases of interest in the multiplicative case, which are also useful for applications.[7]

Corollary 1. *The optimal policy $\mathcal{P}^*$ in the multiplicative case satisfies the following properties.*

1. *Suppose there are alternatives for which $u_i \equiv 0$. Then, these alternatives have an index equal to zero. Hence, they must be chosen first when augmenting, and chosen last (if at all) when non-augmenting.*

2. *Suppose there is a "nullifying alternative" i for which $v_i \equiv 0$ (i.e., whenever it is not chosen, the periodic payoff is zero). Then i is always in a non-augmenting state, and its index is infinite. Hence, if there are other alternatives in an augmenting state, they will be chosen before it, but otherwise it is optimal to choose it first.*

3. *Suppose there is a set of alternatives E such that $v_i \equiv \overline{v}$ for all $i \in E$. Then the conditional choice rule among them is specified by their standard Gittins indices.*

We conclude this subsection with the observation that the decision problem in the additive case can be reformulated as a standard bandit problem where each period only the alternative that is chosen generates a payoff, which is *stochastic* as a function of the current state of the alternative (note that in our additive formulation the payoff each period from an alternative is *deterministic* as a function of its current state). Denote the scheduling environment in the additive case described above by $\mathcal{E}$, and consider the fictitious environment $\hat{\mathcal{E}}$ in which, in each period t, the DM's payoff is equal to

$$w_i(x_{i,t}) \equiv u_i(x_{i,t}) - v_i(x_{i,t}) + \frac{\delta}{1-\delta}\left(v_i(x_{i,t+1}) - v_i(x_{i,t})\right) + \sum_{j=1}^{n} v_j(x_{j,0}), \tag{7}$$

[7] These are new observations, which were not shown in Nash (1980).

where i is the alternative selected in period t. Note that in period t the value of $v_i(x_{i,t+1})$ is unknown; that is, $v_i(x_{i,t+1})$ is a random variable (from the perspective of period t), and hence so is $w_i(x_{i,t})$. In this fictitious problem the DM's payoff does not depend on the states of other alternatives (except for the period 0 states, which are known, and which appear in the last summand of (7)). Hence, the fictitious scheduling problem $\hat{\mathcal{E}}$ is a classic bandit problem.

Denote by $\{x_j^\infty\}$ an entire sample path of x_j from its initial state onward, and denote by $\{x_j^\infty\}_{j=1}^n$ the collection of paths for all alternatives.

Proposition 1. *Given any scheduling policy, for any collection of realization paths $\{x_j^\infty\}_{j=1}^n$, the environments $\mathcal{E}$ and $\hat{\mathcal{E}}$ are payoff equivalent.*

The function w_i can be interpreted as follows. Suppose that when a DM selects an alternative he gets the immediate payoff from it and the discounted expected value of the change in the stream of payoffs, assuming the alternative is not picked again. Thus, when the DM selects an alternative, the payoff today from that alternative changes from v_i to u_i and, from the next period on, the per-period payoff changes from $v_i(x_t)$ to $v_i(x_{t+1})$. The reason we can transform the original problem into one where the classic Gittins index solution applies follows from the additive separability across alternatives and across periods. This separability breaks down in the multiplicative case, where it remains an open question whether an analogue of Proposition 1 is true.

In Appendix A, we provide an algorithm for deriving closed-form expressions of the indices in both the additive and multiplicative cases. To the best of our knowledge, this is a new result, which simplifies the analysis of applications (indeed, we rely on this algorithm in analyzing applications in the sections that follow). To better understand how to use the algorithm in applications, we include in the Appendix a simple example that illustrates the different stages of the algorithm. The example, which considers a setting of on-the-job training, also demonstrates a qualitative difference between the additive and multiplicative cases.

4 Weitzman's (1979) problem with complementarities

In the seminal Pandora's-boxes problem of Weitzman (1979), a DM sequentially inspects "boxes" containing unknown prizes. He chooses both the order of inspection and when to stop searching and pick the highest realized prize. Given the central role of this problem in the search literature, we now generalize this classic problem to accommodate complementarities among alternatives. This opens the door to a new variety of applications.

4.1 Pandora's boxes with complementarities

There are n boxes. Each box i is characterized by a pair of values $(\omega_i, \nu_i) \sim F_i$, which are revealed immediately when the box is opened. The DM discounts the future according to δ; we assume that there are no direct costs to opening boxes besides the time cost. The DM can open boxes in any

order he chooses, stopping at any time to take one of the previously inspected boxes. Departing from Weitzman's classic problem, we assume that the boxes' values are complements in the eyes of the DM. In particular, the overall payoff from stopping and taking a previously opened box i is given by

$$U_i \equiv \omega_i \prod_{j \in O \setminus \{i\}} \nu_j,$$

where O is the set of opened boxes. The value ω_i represents the *intrinsic* value from box i and ν_i represents the externality that an inspected box i exerts on the payoff from choosing another box j. In particular, when $\nu_i \equiv 1$ for all i, there are no complementarities, and the problem corresponds to Weitzman's original problem. We refer to the above problem as *Pandora's boxes problem with complementarities*.

One example that fits this environment is an extension of Weitzman's original problem where inspecting a box takes a stochastic number of periods (including the possibility of immediate inspection). In this case $\nu_i = \delta^{\theta_i}$ and $\omega_i = \delta^{\theta_i}\hat{\omega}_i$, where $\hat{\omega}_i$ is box i's unknown prize and θ_i is box i's unknown inspection time. A second example is an environment where past inspections affect the value of the selected alternative. For instance, each alternative may be some project, and the process of learning about a project's value ω_i creates new knowledge that enhances the value of the project that is ultimately selected (this may be captured by $\nu_i > 1$). A third example concerns the problem of an entrepreneur who sequentially approaches potential investors to carry out a venture but faces the risk that each investor who is approached might steal his idea.

4.2 Characterization of the optimal policy

Say that an unopened box i is *augmenting* if $\delta \mathbb{E}(\nu_i) > 1$, where the expectation is taken with respect to the distribution F_i. As will become clear below, this definition is a special case of the more general notion of an augmenting state described above. The following result characterizes the DM's optimal policy.

Theorem 4. *The optimal policy in Pandora's boxes problem with complementarities is the following: At each stage, if there is an augmenting box, open it (the order among augmenting boxes does not matter). If there are no augmenting boxes, then if the highest reservation prize*

$$R = \delta \left(\int_{\omega > R\nu} \omega dF(\omega, \nu) + \int_{\omega < R\nu} R\nu dF(\omega, \nu) \right)$$

among unopened boxes is greater than the highest value of $\frac{\omega}{\nu}$ among the opened boxes, open the box with the highest reservation prize; otherwise, stop and take the prize with the highest $\frac{\omega}{\nu}$ among the open boxes.

The problem described above is one in which the decision problem ends when a box is selected

and the DM receives its prize. The framework described in Section 2, however, has an infinite horizon, and the DM chooses indefinitely among the alternatives, enjoying flow payoffs along the way. Hence, it may not be immediately clear why the framework in Section 2 can accommodate the Pandora's boxes problem with complementarities. To get an intuition, consider the following auxiliary problem, which is a special case of the general environment in Section 2. Suppose that any alternative that has never been used has the same state x_0, and that the payoffs corresponding to an unused box i are given by $u_i(x_0) = 0$ and $v_i(x_0) = 1$. When an alternative is used for the first time, its state transitions from x_0 to a state x_i consisting of the realization of (ω_i, ν_i) drawn from F_i. Using an alternative for the second time, when its state is $x_i = (\omega_i, \nu_i)$, yields a payoff of $u_i(x_i) = \omega_i(1-\delta)$, and the passive payoff corresponding to i is ν_i. Additionally, when an alternative i is chosen for the second time (or any k'th time, where $k \geq 2$), its state does not change (i.e., u_i and v_i remain the same thereafter). Clearly, this problem is just a special case of the environment in Section 2. Importantly, note that in this auxiliary problem – under the optimal policy – once an alternative is chosen for the second time, it will continue to be chosen forever, yielding the same payoff in each subsequent period.

To understand the connection to the Pandora's boxes problem with complementarities, note that using an alternative for the first time in this problem is akin to opening a box, in which case no payoff is received and the intrinsic value and externality of the box are revealed. Using a box for the second time yields a payoff of ω_i multiplied by the externalities of the opened boxes, and therefore corresponds to taking a previously opened box. In fact, the only differences between the two environments is that under the Pandora's boxes problem with complementarities, taking a box ends the decision problem, whereas in the auxiliary problem it does not. However, as noted above, under the optimal policy in the auxiliary problem, the decision and payoffs do not change once an alternative is used for the second time. The fact that the active payoff in the auxiliary problem is multiplied by $(1-\delta)$ accounts for this difference, and the problems become essentially identical. The proof of Theorem 4 in the Appendix formalizes this argument.

Beyond the Pandora's boxes problem with complementarities, it is important to note that the framework in Section 2 permits a great deal of flexibility in how payoffs depend on the state of an alternative. In many classic application of the bandit problem, payoffs are linked to realizations of an alternative. However, as the application in this section illustrates, the framework is far more general, and can also accommodate a variety of learning environments where the DM engages in information acquisition, receiving a positive payoff only when a decision is ultimately made.

In the case with no complementarities where $\nu \equiv 1$, the reservation prize becomes $R = \delta\left(\int_R^\infty \omega dF(\omega, 1) + RF(R, 1)\right)$, which corresponds to Weitzman's reservation prize in his original problem. Furthermore, since boxes are never augmenting, the optimal policy described in Theorem 4 above collapses to Weitzman's solution in the case of $\nu \equiv 1$ in which there are no complementarities.

The interpretation of Weitzman's reservation prize is the value of the hypothetical prize that makes one indifferent between taking it immediately and opening the box to learn its value while

maintaining the option to take the hypothetical prize after observing its realization. In our extension with complementarities, the reservation prize R is the hypothetical prize that makes one indifferent between taking it now and opening the box (to learn the value ω and "activate" its externality ν), maintaining the opportunity to take the hypothetical prize *multiplied by the realized externality* ν, i.e., $R\nu$.

Interestingly, note that upon stopping, in contrast to Weitzman's original problem, it is not the box with the highest intrinsic value ω that is picked, but rather the one with the highest ratio ω/ν.

The characterization of the optimal policy implies the following interesting comparative statics. For any distribution $F(\omega, \nu)$ denote by $F(\omega)$ the marginal over ω, denote by $F_\omega(\nu) \equiv F(\nu|\omega)$ the conditional distribution over ν, and let R_F denote the reservation prize under the distribution F.

Proposition 2. *Let F and G be a pair of distributions on (ω, ν) such that for every ω, the distributions $F_\omega(\nu)$ and $G_\omega(\nu)$ are continuous in ν, $F(\omega) = G(\omega)$, and F_ω FOSD G_ω.*

1. *$R_F > R_G$.*
2. *If all alternatives are ex ante identical, then the expected time of selecting an alternative under G is lower than under F.*

The result above has the following implications. When the distribution of an alternative's externality ν increases in the FOSD sense, the *pre-inspection* ranking of that alternative, as captured by its reservation prize R, improves. However, after it is inspected, its *post-inspection* ranking, captured by ω/ν, deteriorates. The intuition is that when an alternative's externality is "activated" upon its inspection, in order to reap the benefits of this externality, another alternative must be selected. This contrasts with a first-order stochastic increase in the marginal distribution $F(\omega)$, which improves both the pre- and post-inspection ranking of the alternative.

4.3 Application – approaching investors with a new idea

To demonstrate the characterizations from the previous subsection, consider the following example in the spirit of Baccara and Razin (2007) and Luo (2014). An innovator with a new idea wishes to match with an investor (or partner) to help implement the idea. The innovator knows that once he describes his idea to a potential investor he faces the risk that the investor may steal it.

We model the innovator as sequentially approaching potential investors. When an investor is approached, the innovator learns the value that the investor can generate, captured by ω, and the risk that his idea will be stolen in case the investor will not be chosen. We assume that each ω_i is independently drawn from a uniform distribution over $[0, 1]$. We also assume that an investor with a realized value of ω_i will steal the idea (if he is not selected) with probability $\beta_i\omega_i$, where β_i captures the investor's reputation for expropriating others' ideas.

Thus, the value from matching with investor i will be realized only if none of the approached investors end up stealing the idea.[8] It follows that the expected payoff from choosing investor i, having approached a set of investors O is given by

$$U_i = \omega_i \prod_{j \in O \setminus \{i\}} (1 - \beta_j \omega_j).$$

Applying Theorem 4 to this example yields the following characterization.

Proposition 3. *The optimal policy is the following. If the highest reservation prize,*

$$R = \frac{-2 + \beta\delta + \sqrt{4 + 4\beta\delta - \delta^2(4 - \beta^2)}}{2(2\beta - \delta)},$$

among investors that have not been approached is greater than the highest value of $\frac{\omega}{1-\beta\omega}$ among the investors that have been approached, then the investor with the highest reservation prize is approached. Otherwise, stop and sign with the investor for which the highest value of $\frac{\omega}{1-\beta\omega}$ has been observed.

As in Proposition 2 we obtain the following comparative statics. First, an investor with a more dubious reputation is less likely to be approached (i.e., his reservation prize is lower), but once approached, he is more likely to be chosen (higher $\frac{\omega}{1-\beta\omega}$). Moreover, in a symmetric setting where all investors are ex-ante identical, the expected time until an investor is selected decreases in β.

5 Two alternatives

In environments with only two alternatives our framework can be used to solve interesting scheduling problems that would not otherwise be tractable. In particular, in the presence of only two alternatives, the externality function v of a given alternative can be used to encode direct information on the other alternative. In this section we present two such examples. In each example we highlight where the assumption of two alternatives plays a crucial role.

5.1 Repeated bargaining

Consider the dilemma often faced by firms of whether to switch supplier/contractor or remain with the current one. On the one hand, repeatedly using the same supplier enables the supplier to improve over time its process for producing the input needed by the firm. On the other hand, staying with the same supplier may improve the supplier's bargaining position over time, as his level of expertise increases. Firms may differ in their solutions to this dilemma, which raises the

[8] We are implicitly assuming that it takes time to steal an idea, so that an idea can be stolen only after the innovator has completed his search. For simplicity, we assume that a chosen investor has no incentive to steal the idea.

question of what factors favor one decision over the other.[9] Applying the additive case of our framework to a simple stylized model of the firm's decision problem, we show that a firm either sticks with the same supplier or switches suppliers in every period. In particular, we show that if a firm finds it optimal to switch every period, then so will a more patient firm.

Consider a principal P and two identical agents, 1 and 2. In each period there is a project that the principal can assign to one of the agents. When an agent is assigned a project, the surplus he generates is a function of his current productivity. This productivity evolves over time, such that the more projects are assigned to an agent, the higher his productivity (possibly due to learning-by-doing). More specifically, we assume that the surplus from the k-th assigned project is equal to $\sum_{n=0}^{k-1} \theta^n$, where $\theta \in (0, \frac{1}{2})$. The surplus from a project is divided between the principal and the agent to whom the project is assigned through bargaining, as described below. At the start of each period, the principal simultaneously bargains with each of the agents on the division of surplus in case the project is assigned to that agent. Given the bargaining outcome, the principal assigns the project to one of the agents and payoffs are realized. P seeks to maximize the discounted sum of his payoffs with the discount factor δ.

Following Stole and Zwiebel (1996b), we take a reduced-form approach to modeling the bargaining between P and the agents, and assume that agreements are non-binding in the sense that in case of disagreement with agent i, the principal can renegotiate with j, and players anticipate the possibility of such changes following disagreement.[10] The outside option of each agent is normalized to zero. For the principal, the outside option when bargaining with i is the outcome of bargaining with the other agent, j, *in the absence of agent* i.

In our problem, this means that if P assigns the project to agent i in period t, the payoffs to P and i are the following. Let $\beta \in [\frac{1}{2}, 1]$ be P's bargaining power, so that each agent's bargaining power is $1 - \beta$. To guarantee the validity of the bargaining solution (described below), we assume $1 - \theta \geq \beta$. Suppose that the project is assigned to agent i with productivity q_i. If there were no agent i, and P were to bargain only with j, the surplus to be divided would be q_j with outside options of zero for both P and j. This would yield a payoff of βq_j to P and $(1 - \beta)q_j$ to agent j. It follows that in the solution to the bargaining between P and agent i with productivity q_i, the payoff to each side $h \in \{P, i\}$ is defined as follows:

$$[h\text{'s outside option}] + [h\text{'s bargaining power}] \times (\text{surplus} - [P\text{'s outside option}] - [i\text{'s outside option}]) .$$

[9] For instance, Helper and Levine (1992) noted that in the auto industry, contracts with suppliers are typically renegotiated each period, and while Japanese automakers tend to maintain long-term relationships with the same supplier, American automakers often switch between different suppliers.

[10] Stole and Zwiebel (1996b) characterizes a profile of payoffs that is stable in the following sense: prior to production, no individual agent can benefit from renegotiating with the principal, and the principal cannot benefit from renegotiating with the other agent. In all such negotiations, the principal and the agents split the joint surplus from their relationship according to their respective (exogenously given) bargaining powers, and relative to their respective outside options. Stole and Zwiebel (1996a) show that the stable payoff profile coincides with the unique subgame perfect equilibrium outcome of an extensive-form bargaining game.

Hence, P's payoff is $\beta q_i + \beta(1-\beta)q_j$, while i's payoff is $(1-\beta)q_i - \beta(1-\beta)q_j$. Note that a necessary condition for this solution to be valid is that at any history,

$$\frac{q_i}{q_j} \geq \beta. \tag{8}$$

Otherwise, i's payoff is negative and the bargaining solution is invalid. When (8) is met, although the payoff in every period is obtained only from the selected supplier in that period, the functional form of the payoff allows us to apply Theorem 2. Specifically, setting $u_i(q_i) = \beta q_i$ and $v_i(q_i) = \beta(1-\beta)q_i$ translates the problem into our general form. Intuitively, the "passive payoff" $v_j(\cdot)$ captures the externality of supplier j (which depends on j's productivity) in periods when i is chosen.

Note that beyond two suppliers, P's outside option when bargaining with one supplier is a function of only the *most* productive of the remaining suppliers. Thus, when there are more than two suppliers the payoff function is no longer additive in the suppliers' productivity.

As long as P does not have all the bargaining power, he benefits from a relatively high outside option when bargaining with a supplier. On the one hand, this may motivate P to frequently switch between suppliers so as to maintain a high outside option with each. On the other hand, it may create an incentive to stick with one supplier for some time, enabling that supplier to increase his productivity, and then capitalize on this improvement by switching to the other supplier with an improved outside option. The next result shows that P either goes back and forth between the suppliers in each period or remains with the same one throughout, and characterizes which policy is optimal as a function of the parameters.

Proposition 4. *Let $\hat{\delta} = \frac{\beta}{1-\theta(1-\beta)}$. If $\delta > \hat{\delta}$, the unique optimal policy is to assign the project to a different agent in each period. If $\delta < \hat{\delta}$, the unique optimal policy is to assign the project to the same agent in all periods. At the threshold, $\delta = \hat{\delta}$, any allocation policy is optimal.*

Clearly, if P has full bargaining power, it is optimal for him to stick with the same supplier at all periods. However, when $\beta < 1$, the extent to which P can capitalize on the improvements in a supplier's productivity depends on his outside option, creating the incentive to switch between the suppliers. The smaller P's bargaining power (β) or the slower the suppliers' improvement rate (θ), the stronger P's incentive to switch becomes, and therefore the less patient P must be in order for the strategy of continually switching between suppliers each period to become optimal. The result therefore suggests a possible channel through which greater bargaining power on the side of the firm (and/or a sufficiently fast learning-by-doing on the side of the suppliers) gives rise to the emergence of long-term exclusive relationships.

The qualitative features of the above insights extend to a more general surplus function. Let $f(k)$ be the surplus from the k-th assigned project, where f is concave and increasing in k such that condition (8) holds, i.e., that $f(k)/f(k') \geq \beta$ for all k and k'.

First, we show that a relatively impatient P remains with the same agent in all periods.

Proposition 5. *If $\delta < \beta$, the unique optimal policy is to assign the project to the same agent in all periods.*

Next, we show that a sufficiently patient P switches between agents every period when f is "sufficiently concave".

Proposition 6. *If $\delta > \beta$ and for every k,*

$$\frac{f(k+1) - f(k)}{f(k+2) - f(k+1)} > \frac{\delta(1-\beta)}{\delta - \beta}, \tag{9}$$

then the unique optimal policy is to assign the project to a different agent in each period.

5.2 Disappearing alternatives

In many scheduling environments, alternatives may deteriorate or become unavailable independently of whether they were chosen. Such problems do not fit the classic multi-armed bandit paradigm. However, when there are only two alternatives, our framework with complementarities allows to characterize the optimal scheduling policy. We demonstrate this with the following example.

Suppose each alternative $i \in \{A, B\}$ generates a payoff of q^n when it is selected for the $n-$th time, for some $q \in (0,1)$. Alternative A is always available. On the other hand, conditional on being available in a given period, alternative B will disappear with probability $p \in (0,1)$, irrespective of whether it is selected in that period.

To apply our framework, let the state space of A be $X_A = \mathbb{N} \times \{0,1\}$, and the payoff functions be $u_A(n,\chi) = q^n$, and $v_A(n,\chi) = \chi$, for all $n \in \mathbb{N}$ and $\chi \in \{0,1\}$. Suppose that if alternative A is selected at state $(n,0)$, it advances to the state $(n+1,0)$ with probability 1, while if it is selected at state $(n,1)$ it advances to the state $(n+1,1)$ with probability $1-p$ and to the state $(n+1,0)$ with probability p. For alternative B, let $X_B = \mathbb{N} \cup \{\phi\}$, and define the payoff functions $u_B(x_B) = \mathbf{1}_{\{x_B \neq \phi\}} q^{x_B}$ and $v_B(x_B) \equiv 1$, for all $x_B \in X_B$. Suppose that if alternative B is selected at state $x_B \neq \phi$, it advances to the state $x_B + 1$ with probability $1-p$ and to the absorbing state ϕ with probability p.

The idea of the above formulation is the following. Since B can disappear both when B is chosen and when A is chosen, $\phi \in X_B$ encodes the former event, in which case $u_B = 0$, and $\chi = 0$ encodes the latter event in which case $v_A = 0$. In either case, the payoff from selecting B is zero. Note that this formulation relies on there being only two alternatives. The next result characterizes the optimal policy in such a setting.

Proposition 7. *The optimal policy is as follows. Let $c(\delta,p,q) \equiv \frac{\log\left[\frac{1-\delta}{1-\delta q} \cdot \frac{1-\delta(1-p)q}{1-\delta(1-p)}\right]}{\log(q)} > 0$. First select B for $\lceil c(\delta,p,q) \rceil$ periods, then switch to A and alternate every period. When B disappears, choose A forever.*

Since both alternatives have the same productivity which diminishes over time, in principle, the DM would want to alternate between them every period. However, as B can disappear, it receives a head-start which keeps its productivity below that of A throughout. For example, if $q = 0.9,\ \delta = 0.9$, and $p = 0.1$, we get $c(\delta = 0.9, p = 0.1, q = 0.9) \approx 2.72$. Hence, before alternating between the alternatives (so long as B is available), it is optimal to choose B three times in a row.

6 Counting space

There are many interesting economic environments where the state of each alternative encodes the number of times that the alternative was chosen. This state space imposes enough structure that further simplifies the derivation of the indices. Many environments in the multiplicative case have the feature that alternatives impose a diminishing marginal externality on others. This property turns out to have some useful implications that are described in the next proposition.

Proposition 8. *Consider the multiplicative case. Assume that the state space of alternatives is* $\mathbb{N} \cup \{0\}$*, with each state* s *of an alternative representing the number of times that the alternative was chosen. If* v *is concave then:*

1. *The alternative is augmenting in state* s *if and only if* $a(s, 1) = \delta v(s + 1) - v(s) > 0$.

2. *There exists* $\bar{s} \geq 0$ *such that a state* s *is non-augmenting if and only if* $s \geq \bar{s}$.

We demonstrate the usefulness of this result in the following application.

Supervising agents with stochastic costs of effort. The literature on moral hazard has focused on characterizing payment schemes that incentivize agents to exert effort. However, there are many environments in which workers get a fixed wage, and hence cannot be incentivized with monetary transfers that depend on their output (for instance, in the public sector). In these environments, a principal may need to supervise agents while they work on a task, in order to ensure that the task is completed successfully. If the principal is in charge of multiple agents, he must decide in each period which agent to supervise, taking into account the effect of repeated supervision on the agent's willingness to work when left unsupervised. For some agents, supervision can be constructive and helpful, while for others it may be perceived as an unwanted annoyance. The following simple example illustrates how our framework can be applied to characterize the principal's optimal policy in the presence of heterogeneous agents.

There are two agents who jointly work on a project. Each agent is in charge of a task, and the project is completed successfully if and only if both agents successfully complete their respective task. When an agent is supervised, he successfully completes his task with certainty. When left unsupervised, the agent faces a stochastic cost of effort in completing the task. That is, the agent's motivation for carrying out the task fluctuates (e.g., it may depend on his mood that day, which is affected by factors outside of his control), and his realized motivation determines his perceived

cost of exerting effort. The agent exerts effort if and only if the realized cost does not exceed a threshold. Each agent i starts with an initial threshold c_i, which can change with the number of times in which he is supervised.

Assume that in each period an agent draws a cost from a uniform distribution on $[0, 1]$. Agent 1's cost threshold is zero, and remains constant regardless of the number of times he is supervised. Thus, agent 1 never works when left unsupervised. By contrast, agent 2's threshold as a function of the number of times he was supervised s is $v(s)$, where $v(\cdot)$ is increasing and concave. The interpretation is that supervision helps agent 2 to learn how to perform the task more efficiently, and hence with a lower perceived cost, but the returns to supervision are diminishing. The question we consider is: How should the principal optimally supervise the agents over time?

Let us now show how Corollary 1 and Proposition 8 can be applied to this example. Let the state s of an agent denote the number of times he was supervised. Since $v_1 \equiv 0$, by Corollary 1 (part 3) it follows that whenever agent 2 is in an augmenting state it is optimal to supervise him; otherwise, it is optimal to supervise agent 1. Moreover, from Proposition 8, since $v(s)$ is concave, there exists a threshold before which agent 2 is augmenting and beyond which he is not. This implies the following result.

Proposition 9. *Under the optimal policy, there exists $T \geq 0$ such that the principal supervises agent 2 for T consecutive periods and then switches to supervising agent 1 indefinitely.*

To illustrate the above result, let $v(s) = (\frac{1}{2})^{\frac{1}{s+1}}$. If $\delta > \sqrt{\frac{1}{2}}$, then $a_2(0, 1) > 0$, and the principal will begin supervising agent 2, as he is augmenting. He will continue to do so until a state in which agent 2 is non-augmenting. Agent 2 is non-augmenting at s if $a(s, 1) \leq 0$, i.e., if $\left(\frac{1}{2}\right)^{\frac{1}{(s+1)(s+2)}} \geq \delta$. If $\delta \leq \sqrt{\frac{1}{2}}$, then agent 2 is also non-augmenting and it is optimal to supervise agent 1 only. For example, if $\delta = 0.99$, then the principal supervises agent 2 for six periods, after which he switches to supervise agent 1 in all periods. Hence, from period 7 onward, the expected per-period output is $\left(\frac{1}{2}\right)^{\frac{1}{7}} \approx 0.9$. Thus, if the principal is sufficiently patient, he will first "invest" in lowering the cost of agent 2, even at the expense of no output, and only then switch to supervising agent 1.

We conclude this section with an application that utilizes the additive case of our framework to study the dynamics of occupational choice, allowing human capital in each sector to depend on the experience in both sectors.

Career paths and mobility between sectors. The multi-armed bandit problem has been a natural framework for studying the dynamics of occupational choice. In an influential paper, Jovanovic (1979) uses a multi-armed bandit problem to study a model in which an individual sequentially chooses employment among multiple firms, and learns through specific experience how suited he is to a given job. The individual's optimal policy yields a decreasing hazard: the conditional probability of turnover falls as tenure increases.[11] Intuitively, the more experience the

[11]The relationship between job-specific skills and turnover decisions has been central to the economics of labor mobility since the work of Becker (1962), Mincer (1962), and Oi (1962).

individual accumulates in a particular job, the more precise is his assessment of his competence in this job. Therefore, new information is less likely to affect this assessment and therefore less likely to cause the individual to leave his job. Miller (1984) enriches this model by introducing ex-ante heterogeneity in jobs, using the multi-armed bandit problem to characterize the dynamics of optimal job choice. Since the value of job-specific experience varies across jobs, jobs yielding riskier (but potentially higher) returns are experimented with earlier. Young, inexperienced workers therefore experiment more with such risky jobs.

In the above papers, an individual's expertise in a given job has no bearing on the returns from other jobs. Our framework enables us to extend the classic literature on dynamic occupational choice by allowing transfer of human capital across jobs. The extent to which accumulated human capital is transferable across jobs is relevant for individuals' decisions to switch jobs between sectors and, in particular, between the private and public sectors. For instance, it is fairly common for academics and professionals (e.g., lawyers, accountants, economists, engineers) to switch from private firms to government departments/agencies (which oftentimes involves a salary cut) and then switch back to the private sector.

The following simple example illustrates how career paths that display these movements between sectors are captured by our framework. In each period, an individual can work in one of two sectors, A or B. In each period that he works in a sector, the individual accumulates human capital (measured in monetary units). A fraction of that human capital is transferable to the other sector. The individual's per-period payoff is equal to the accumulated human capital in the current sector plus the transferable portion of the human capital that he accumulated in the other sector.

The initial human capital in sector A is zero and each period of experience in that sector increases the human capital by $r > 0$. After a total of T periods of experience (not necessarily consecutive) in sector A, the total human capital reaches its maximal level of Tr. That is, denoting by $u_A(s)$ the total human capital accumulated in sector A after s periods of experience, we have $u_A(s) = (s+1)r$ in every state $s < T$, and $u_A(s) = Tr$ for all $s \geq T$. Not all of the human capital accumulated in sector A is directly transferable to sector B. Specifically, if an individual with a total experience of s periods in sector A decides to switch to sector B, then only a portion γ of his accumulated human capital is added to his accumulated human capital in the new sector. This transfer of human capital is modeled via the function v_A, such that $v_A(s) = \gamma sr$, where $\gamma \in (0,1)$ is the human capital that is transferred to sector B when the individual switches to that sector after accumulating s periods of experience in sector A. To simplify the analysis, we assume that from the very first period of work in sector B, the human capital in that sector remains constant at $b > 0$, and that a portion β of it is transferable to sector A. Thus, $u_B(s) = b$ for every s, while $v_B(0) = 0$ and $v_B(s) = \beta b$ for $s > 0$ where $\beta \in (0,1)$.

Our objective is to illustrate that an important reason for switching sectors is to accumulate human capital that can be useful in another sector. For instance, a law graduate may enhance his future productivity in a private law firm by first starting out working in a public defender's

office. An alternative career path may begin in a private law firm, followed by a move to the justice department, and then a return to a senior position in a private law firm. To highlight the role of transferable human capital, we assume that there is no uncertainty.[12]

The next result gives necessary and sufficient conditions for each possible optimal career path when the individual is sufficiently patient.

Proposition 10. *Assume that* $\delta \geq \frac{(1-\gamma)T-1}{(1-\gamma)(T-1)}$. *The optimal career paths are characterized by:*

(A) It is optimal to work only in A if and only if $A_1 \equiv (1-\gamma)Tr \geq b\left(\frac{1-\delta(1-\beta)}{1-\delta}\right) \equiv B_0$.

(B) It is optimal to work only in B if and only if $B_1 \equiv (1-\beta)b \geq \frac{r}{1-\delta} - \frac{rT\delta^T(1-\gamma)}{1-\delta^T} \equiv A_0$.

(AB) It is optimal to start a career in A and then move to B and remain there if and only if $A_0 \geq B_0$ *and* $B_1 \geq A_1$.

(BA) It is optimal to start a career in B and then move to A and remain there if and only if $B_0 \geq A_0$ *and* $A_1 \geq B_1$.

(ABA) It is optimal to start a career in A, then move to B, and then return to A and remain there if and only if $A_0 \geq B_0 \geq A_1 \geq B_1$.

(BAB) It is optimal to start a career in B, then move to A, and then return to B and remain there if and only if $B_0 \geq A_0 \geq B_1 \geq A_1$.

Our simple model admits several possible career paths. In particular, it accommodates paths where the individual switches sectors at most twice during his career. The contribution of the proposition is to give the precise conditions on the model's primitives that correspond to each possible path. The condition on the discount factor ensures that the index of sector A is minimal when the individual reaches the maximal human capital in that sector. As a corollary, the above result also offers the following simple characterization of the individual's long-run occupation: in the long run, it is optimal for the individual to work in sector A if and only if $\frac{Tr}{b} \geq \frac{1-\beta}{1-\gamma}$. Hence, under the maintained assumption on δ, the sector where the individual will eventually work is determined by comparing the ratio of the long-term accumulated human capital in the two sectors with the ratio of the fraction of non-transferable human capital in the sectors.

One of the career paths described in the proposition consists of the individual switching to a stint in the public sector only after reaching a senior position in the private sector. This career path is consistent with the finding by Su and Bozeman (2009) that the probability of switching to the public sector is much higher for those who held a managerial position in their previous private sector job than for those who held professional and technical positions.

[12]The model can be extended to allow for the combination of both learning about the quality of matches and transferable human capital.

More generally, the framework allows to extend the classic model of dynamic occupational choice (Jovanovic, 1979; Miller, 1984) to one reflecting the "skill-weights approach" due to Lazear (2009), which views skills as general, but assumes different jobs attach different weights to these skills. In such a model, the dynamics of occupational choice are driven by the combination of learning and the applicability of accumulated experience across jobs.

References

Baccara, M. and R. Razin (2007). Bargaining over new ideas: the distribution of rents and the stability of innovative firms. *Journal of the European Economic Association 5*(6), 1095–1129.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy 70*(5, Part 2), 9–49.

Bergemann, D. and J. Valimaki (2008). Bandit problems. In: The *New Palgrave Dictionary of Economics.* Ed. by Steven N. Durlauf and Lawrence E. Blume. Basingstoke, UK: Palgrave Macmillan.

Bray, R. L., D. Coviello, A. Ichino, and N. Persico (2016). Multitasking, multiarmed bandits, and the Italian judiciary. *Manufacturing & Service Operations Management 18*(4), 545–558.

Che, Y.-K. and K. Mierendorff (2019). Optimal dynamic allocation of attention. *American Economic Review 109*(8), 2993–3029.

Coviello, D., A. Ichino, and N. Persico (2014). Time allocation and task juggling. *American Economic Review 104*(2), 609–623.

Coviello, D., A. Ichino, and N. Persico (2015). The inefficiency of worker time use. *Journal of the European Economic Association 13*(5), 906–947.

Eliaz, K. and A. Frug (2018). Bilateral trade with strategic gradual learning. *Games and Economic Behavior 107*, 380–395.

Fershtman, D. and A. Pavan (2020). Searching for arms: Experimentation with endogenous consideration sets. *Working paper*.

Fudenberg, D., P. Strack, and T. Strzalecki (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review 108*(12), 3651–3684.

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: North-Holland.

Gossner, O., J. Steiner, and C. Stewart (2020). Attention please! *Econometrica*, forthcoming.

Helper, S. and D. I. Levine (1992). Long-term supplier relations and product-market structure. *Journal of Law, Economics, & Organization 8*(3), 561–581.

Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy 87*(5), 972–990.

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., S. Rady, and M. Cripps (2005). Strategic experimentation with exponential bandits. *Econometrica 73*(1), 39–68.

Klabjan, D., W. Olszewski, and A. Wolinsky (2014). Attributes. *Games and Economic Behavior 88*, 190–206.

Lazear, E. P. (2009). Firm-specific human capital: A skill-weights approach. *Journal of Political Economy 117*(5), 914–940.

Liang, A., X. Mu, and V. Syrgkanis (2021). Dynamically aggregating diverse information. *Working paper*.

Luo, H. (2014). When to sell your idea: Theory and evidence from the movie industry. *Management Science 60*(12), 3067–3086.

Miller, R. A. (1984). Job matching and occupational choice. *Journal of Political Economy 92*(6), 1086–1120.

Mincer, J. (1962). On-the-job training: Costs, returns, and some implications. *Journal of Political Economy 70*(5, Part 2), 50–79.

Nash, P. (1980). A generalized bandit problem. *Journal of the Royal Statistical Society: Series B (Methodological) 42*(2), 165–169.

Oi, W. Y. (1962). Labor as a quasi-fixed factor. *Journal of Political Economy 70*(6), 538–555.

Radner, R. and M. Rothschild (1975). On the allocation of effort. *Journal of Economic Theory 10*(3), 358–376.

Stole, L. A. and J. Zwiebel (1996a). Intra-firm bargaining under non-binding contracts. *Review of Economic Studies 63*(3), 375–410.

Stole, L. A. and J. Zwiebel (1996b). Organizational design and technology choice under intrafirm bargaining. *American Economic Review 86*(1), 195–222.

Su, X. and B. Bozeman (2009). Dynamics of sector switching: Hazard models predicting changes from private sector jobs to public and nonprofit sector jobs. *Public Administration Review 69*(6), 1106–1114.

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

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, 641–654.

Appendix A: Calculating the indices explicitly

In this section, we introduce an algorithm for deriving the closed-form expressions of the indices in both the additive and multiplicative cases. Since we calculate the index for any given alternative, we will suppress the subscript referring the identity of the alternative. We focus on the case where the set of an alternative's states $X \equiv \{1, ..., m\}$ is finite (but note that the set of states can differ across alternatives), and the transition between states is Markovian and specified by an $m \times m$ probability matrix $\boldsymbol{P} \equiv [P_{x,y}]$. Denote by $\boldsymbol{u}$ and $\boldsymbol{v}$ the $m \times 1$ vectors such that $\boldsymbol{u}_x = u(x)$ and $\boldsymbol{v}_x = v(x)$.

The relation $\trianglerighteq$ defined above over indices also induces a complete ordering on the set of states X. With slight abuse of notation, we will therefore use $\trianglerighteq$ also as the relation over states. Using this order, for any state $x \in X$, the state space X may be split into two sets, $C(x) \equiv \{y \in X : y \triangleright x\}$ and $S(x) \equiv \{y \in X : x \trianglerighteq y\}$. We refer to these sets as the *continuation set* and the *stopping set*, respectively. The interpretation is that, given Theorem 1, the optimal stopping rule $\tau^*(x)$ continues for all states in $C(x)$ and stops for those in $S(x)$.

We now generalize an algorithm due to Varaiya et al. (1985) – henceforth, VWB – to calculate the exact value of $\mathcal{I}(x)$ for any state $x \in X$. It will be convenient to order the states in X in decreasing order: $\alpha_1 \trianglerighteq \alpha_2 \trianglerighteq ... \trianglerighteq \alpha_m$, where α_i is the i-th ranked state according to $\trianglerighteq$. This ranking is initially unknown and will be derived by the algorithm. The algorithm proceeds in m steps, where each step $k = 1, ..., m$ identifies α_k and calculates its index.

Consider first the multiplicative case: *Step k.* Although in the beginning of step k we do not yet know the identity of α_k, we know that $C(\alpha_k) = \{\alpha_1, ..., \alpha_{k-1}\}$ and $S(\alpha_k) = X - C(\alpha_k)$. Note that in step $k = 1$, $C(\alpha_1) = \emptyset$ and $S(\alpha_1) = X$. Define the $m \times m$ matrix

$$\boldsymbol{Q}^{(k)}_{x,y} \equiv \begin{cases} \boldsymbol{P}_{x,y}, & \text{if } y \in C(\alpha_k) \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Note that in the step $k = 1$, $\boldsymbol{Q}^{(1)}$ is a matrix of zeros. Define the vectors

$$\boldsymbol{d}^{(k)} \equiv [\boldsymbol{I} - \delta \boldsymbol{Q}^{(k)}]^{-1} \boldsymbol{u} \quad \text{and} \quad \boldsymbol{a}^{(k)} \equiv [\boldsymbol{I} - \delta \boldsymbol{Q}^{(k)}]^{-1} [\delta \boldsymbol{P} \boldsymbol{v} - \boldsymbol{v}],$$

where $\boldsymbol{I}$ is the $m \times m$ identity matrix. Let σ be a stopping rule that continues in states $y \in C(\alpha_k)$ and otherwise stops. Then it is easy but tedious to verify that for any state x,

$$\frac{\boldsymbol{d}^{(k)}_x}{|\boldsymbol{a}^{(k)}_x|} = \frac{\mathbb{E}\left(\sum_{s=0}^{\sigma-1} \delta^s u(x^{+s}) | x\right)}{|\mathbb{E}(\delta^\sigma v(x^{+\sigma})) - v(x)|} = J(x, \sigma).$$

To determine α_k, we must distinguish between the case in which it is augmenting and the case in which it is not. If α_k is augmenting, then it must be the case that $a(\alpha_k, \tau^*(\alpha_k)) > 0$, which is equivalent to $\boldsymbol{a}^{(k)}_{\alpha_k} > 0$. Hence, if α_k is augmenting, it must be an element of

$$S^{(+)}(\alpha_k) = \{\alpha_j \in S(\alpha_k) : \boldsymbol{a}^{(k)}_j > 0\}.$$

Furthermore, if α_k is non-augmenting, then $S^{(+)}(\alpha_k)$ must be empty. It follows that

$$\alpha_k = \begin{cases} \arg\min_{\alpha \in S^{(+)}(\alpha_k)} \frac{\boldsymbol{d}_\alpha^{(k)}}{\boldsymbol{a}_\alpha^{(k)}} & if \quad S^{(+)}(\alpha_k) \neq \emptyset \\ \arg\max_{\alpha \in S(\alpha_k)} \frac{\boldsymbol{d}_\alpha^{(k)}}{-\boldsymbol{a}_\alpha^{(k)}} & if \quad S^{(+)}(\alpha_k) = \emptyset \end{cases}$$

and that the index of α_k is

$$J(\alpha_k) = \frac{\boldsymbol{d}_{\alpha_k}^{(k)}}{|\boldsymbol{a}_{\alpha_k}^{(k)}|}.$$

Note that this algorithm implies the following sufficient condition for the non-existence of augmenting states (this condition does not depend on the finiteness of X).

Corollary 2. *If $a(x,1) < 0$ for all $x \in X$, then there is no augmenting state.*[13]

The result follows from the fact that if the condition in the Corollary is satisfied, then α_1 must be non-augmenting, which implies that all states are non-augmenting.

Consider next the additive case: *Step k.* Define $\boldsymbol{b}^{(k)} \equiv [\boldsymbol{I} - \delta \boldsymbol{Q}^{(k)}]^{-1}\mathbf{1}$, where $\mathbf{1}$ is the $m \times 1$ vector of 1's. One can then verify algebraically that

$$\alpha_k = \arg \max_{\alpha \in S(\alpha_k)} \frac{\boldsymbol{d}_\alpha^{(k)} + \frac{1}{1-\delta}\boldsymbol{a}_\alpha^{(k)}}{\boldsymbol{b}_\alpha^{(k)}}$$

and its index is

$$I(\alpha_k) = \frac{\boldsymbol{d}_{\alpha_k}^{(k)} + \frac{1}{1-\delta}\boldsymbol{a}_{\alpha_k}^{(k)}}{\boldsymbol{b}_{\alpha_k}^{(k)}},$$

where $\boldsymbol{a}^{(k)}$ and $\boldsymbol{d}^{(k)}$ are defined as in the multiplicative case.

Note that the original algorithm of VWB requires knowledge of the payoff of an alternative given its state; in particular, the reward from the selected alternative cannot be a random variable as a function of its state. This means that transforming the additive case into a standard bandit problem using (7), does not permit the use of the VWB-algorithm to compute the indices.

We conclude this section with an example illustrating the workings of the algorithm and the qualitative difference between the additive and multiplicative cases.

Example: On-the-job training. Consider an environment in which two new workers require training by a principal. The workers' productivity is measured by the *success rate* in a given period. Suppose that the workers must receive, in total, two periods of training from the principal before attaining full proficiency. The only difference between the workers is that while the productivity of one remains constant before completing full training, the other already improves after a single period of training. The following table describes the success rates of each of the workers, given their level of training, where $p \in (0,1)$ and $q \in (p, \sqrt{p})$:

[13] The stopping rule 1 refers to the stopping rule that stops immediately, for any realization.

Training	0	1	2
worker A	p	q	1
worker B	p	p	1

We now apply the algorithm to derive indices for each case separately.

Additive. Consider first worker B. In step 1 of the algorithm, $C(\alpha_1) = \emptyset$. Hence, to find α_1 we compare $I_B(x,1)$ across states: $I_B(0,1) = 0$, $I_B(1,1) = \frac{\delta(1-p)}{1-\delta}$, and $I_B(2,1) = 0$. It follows that $\alpha_1 = 1$. Moving on to step 2, we know that $C(\alpha_2) = \{1\}$, and hence, we need to compare $I_B(0,2)$ to $I_B(2,1)$. Since $I_B(0,2) = \frac{\delta^2(1-p)}{1-\delta^2} > 0$, it follows that $\alpha_2 = 0$ and thus, we conclude that $\alpha_3 = 2$. Knowing the stopping rules in each of the states we can write:

$$I_B(0) = I_B(0,2) = \frac{\delta^2(1-p)}{1-\delta^2},\ I_B(1) = I_B(1,1) = \frac{\delta(1-p)}{1-\delta},\ I_B(2) = I_B(2,1) = 0.$$

Using analogous arguments (and noting that $q < \sqrt{p}$ implies that $q - p < 1 - q$) we can show that, for worker A, again, $\alpha_1 = 1$, $\alpha_2 = 0$ and $\alpha_3 = 2$, and that the indices are given by

$$I_A(0) = I_A(0,2) = \tfrac{\delta^2(1-p)}{1-\delta^2} + \tfrac{\delta(q-p)}{1+\delta},\quad I_A(1) = I_A(1,1) = \tfrac{\delta(1-q)}{1-\delta},\quad I_A(2) = I_A(2,1) = 0.$$

It follows that $I_A(1) > I_A(0) > I_B(0)$. It is therefore optimal to train worker A for two consecutive periods and then switch to training worker B for two consecutive periods.

Multiplicative. Consider first worker B. In step 1, $C(\alpha_1) = \emptyset$. Hence, to find α_1 we compare $J_B(x,1)$ across states: $J_B(0,1) = \frac{p}{|\delta p - p|}$, $J_B(1,1) = \frac{p}{|\delta - p|}$, and $J_B(2,1) = \frac{1}{1-\delta}$. Whether or not $a_B(1,1) = \delta - p$ is positive, $\alpha_1 = 1$. In Step 2, we know that $C(\alpha_2) = \{1\}$. Note that whether or not $a_B(0,2) = \delta^2 - p$ is positive, $\alpha_2 = 0$. To see this, note that if $\delta^2 > p$ then $S_B^{(+)}(\alpha_2) = \{0\}$. Otherwise, $J_B(0,2) = \frac{p+\delta p}{p-\delta^2} > \frac{1}{1-\delta}$. In step 3 we conclude that $\alpha_3 = 2$. Hence,

$$J_B(0) = J_B(0,2) = \tfrac{p+\delta p}{|\delta^2 - p|},\quad J_B(1) = J_B(1,1) = \tfrac{p}{|\delta - p|},\quad J_B(2) = J_B(2,1) = \tfrac{1}{1-\delta}.$$

Consider next worker A. In step 1 we compare $J_A(x,1)$ across states: $J_A(0,1) = \frac{p}{|\delta q - p|}$, $J_A(1,1) = \frac{q}{|\delta - q|}$, and $J_A(2,1) = \frac{1}{1-\delta}$. First note that whether δ is above or below $\frac{p}{q}$, $\alpha_1 = 1$. To see this, suppose $\delta > \frac{p}{q}$, then since $q < \sqrt{p}$ we have that $\delta > q$. Hence, $a_A(0,1) > 0$, $a_A(1,1) > 0$ and $a_A(2,1) < 0$. Since $J_A(0,1) = \frac{p/q}{\delta - p/q} > \frac{q}{\delta - q} = J_A(1,1)$ it follows that $\alpha_1 = 1$. Suppose next that $\delta < \frac{p}{q}$. If $\delta > q$, then both $a_A(0,1)$ and $a_A(2,1)$ are negative, while $a_A(1,1)$ is positive. Hence, $\alpha_1 = 1$. If $\delta < q$, then $a_A(\cdot,1) < 0$ for all states. Since in this case $\frac{q}{q-\delta} > \max\{\frac{p/q}{p/q-\delta}, \frac{1}{1-\delta}\}$ it follows that $\alpha_1 = 1$. Turning to step 2, we claim that $\alpha_2 = 0$. To see this, note that if $\delta^2 > p$ then $S_A^{(+)}(\alpha_2) = \{0\}$. Otherwise, $S_A^{(+)}(\alpha_2) = \emptyset$ and $J_A(0,2) = \frac{p+\delta q}{p-\delta^2} > \frac{1}{1-\delta} = J_B(2,1)$ as $q > p$. Finally, $\alpha_3 = 2$. It follows that

$$J_A(0) = J_A(0,2) = \tfrac{p+\delta q}{|\delta^2 - p|},\quad J_A(1) = J_A(1,1) = \tfrac{p/q}{|\delta - p/q|},\quad J_A(2) = J_A(2,1) = \tfrac{1}{1-\delta}$$

To derive the optimal policy, note that both workers are augmenting at 0 if and only if $p < \delta^2$. In this case, $J_A(0) = \frac{p+\delta q}{\delta^2 - p} > \frac{p+\delta p}{\delta^2 - p} = J_B(0)$ and it is optimal to start with B. Since $J_B(1) \succ J_B(0) \succ J_A(0)$ it

follows that it is optimal to train B in the next period as well. Finally, since $J_A(1) \succ J_A(0) \succ J_B(2)$, worker A will be selected in the two subsequent periods.

Suppose next that $p \geq \delta^2$. In this case, both workers are non-augmenting at 0. Since $J_A(0) = \frac{p+\delta q}{p-\delta^2} > \frac{p+\delta p}{p-\delta^2} = J_B(0)$, it is optimal to start with A and continue with A in the next period since $J_A(1) \succ J_A(0) \succ J_B(0)$. Finally, since $J_B(1) \succ J_B(0) \succ J_A(2)$, worker B will be selected in the two subsequent periods. The following summarizes the above discussion.

Proposition 11. *Under the optimal training policy, one of the workers is trained until full proficiency and only then the remaining worker is trained until full proficiency.*

1. *In the additive case, it is optimal to start with A for all $\delta \in (0,1)$.*
2. *In the multiplicative case, if $\delta > \sqrt{p}$, it is optimal to start with B, otherwise, it is optimal to start with A.*

This example illustrates that the training schedule depends on whether the untrained workers are augmenting. In the example, the overall increase in the success rate as a result of full training is identical (from p to 1 in two periods) for both workers. The only difference is that, for A, part of the increase (from p to q) is already attained after one period of training. The question is, when does the principal benefit the most from this intermediate increase?

In the additive case, appropriating the extra gain as early as possible is optimal because of discounting. Hence, for all $\delta \in (0,1)$ the principal trains A first. In the multiplicative case, B's success rate affects the expected benefit from training A in a multiplicative manner—in the same way the discount factor affects the (current) benefit from A's future training. When the untrained worker B is augmenting, the effect of discounting (i.e., delaying A's training) is weaker than that of increasing the success rate of B. Hence, to maximize the *current value* of the benefit from the intermediate increase in A's productivity, the principal first trains B. When an untrained B is non-augmenting, the effect of discounting dominates, and the result is similar to the additive case.

Appendix B: Proofs

Proof of Theorem 1. In what follows, to ease the exposition, we omit the subscript j referring to the alternative, as we are considering a single alternative. We define a stopping rule σ as *feasible* if it satisfies the following. In the additive case, any stopping rule is feasible, and in the multiplicative case, any stopping rule is feasible unless the initial state x is augmenting, in which case σ must satisfy that $a(x,\sigma) > 0$.

Step 1: Any feasible stopping rule that stops at a state y for which $\mathcal{I}(y) \triangleright \mathcal{I}(x)$ does not attain the value of the index. Assume that the initial state is x, and fix a state y such that $\mathcal{I}(y) \triangleright \mathcal{I}(x)$. Consider a feasible stopping rule σ that stops at state y with positive probability. There exists a feasible stopping time σ' such that

$$\mathcal{I}(y,\sigma') \triangleright \frac{\mathcal{I}(y)+\mathcal{I}(x)}{2}. \tag{11}$$

We define the following feasible stopping rule, $\hat{\sigma}$, which coincides with σ unless σ stops at state y, in which case $\hat{\sigma} = \sigma + \sigma'$; that is, whenever σ stops at state y, $\hat{\sigma}$ initiates the stopping rule σ'. We now show that $\mathcal{I}(x, \hat{\sigma}) \triangleright \mathcal{I}(x, \sigma)$.

For the additive case, from (11),

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right) + a(x,\sigma)}{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t|x\right)} = I(x,\sigma) < I(y,\sigma') = \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma'-1} \delta^t u(y^{+t})|y\right) + a(y,\sigma')}{\mathbb{E}\left(\sum_{t=0}^{\sigma'-1} \delta^t|y\right)}.$$

Therefore,

$$\begin{aligned} I(x,\hat{\sigma}) &= \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right) + a(x,\sigma) + \mathbb{E}\left(\sum_{t=\sigma}^{\sigma+\sigma'-1} \delta^t u(y^{+t})|y\right) + \mathbb{E}\left(\delta^\sigma a(y,\sigma')|x\right)}{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t|x\right) + \mathbb{E}\left(\sum_{t=\sigma}^{\sigma+\sigma'-1} \delta^t|y\right)} \\ &= \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right) + a(x,\sigma) + \mathbb{E}\left(\delta^\sigma|x\right)\left(\mathbb{E}\left(\sum_{t=0}^{\sigma'-1} \delta^t u(y^{+t})|y\right) + a(y,\sigma')\right)}{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t|x\right) + \mathbb{E}\left(\delta^\sigma|x\right)\mathbb{E}\left(\sum_{t=0}^{\sigma'-1} \delta^t|y\right)} \\ &> \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right) + a(x,\sigma)}{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t|x\right)} = I(x,\sigma). \end{aligned}$$

Turning to the multiplicative case, suppose first that x is an augmenting state. From (11), y must also be an augmenting state, and

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right)}{a(x,\sigma)} = J(x,\sigma) \geq J(x) > J(y,\sigma') = \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma'-1} \delta^t u(y^{+t})|y\right)}{a(y,\sigma')}.$$

Therefore,

$$\begin{aligned} J(x,\hat{\sigma}) &= \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right) + \mathbb{E}\left(\sum_{t=\sigma}^{\sigma+\sigma'-1} \delta^t u(y^{+t})|y\right)}{a(x,\sigma) + \mathbb{E}\left(\delta^\sigma a(y,\sigma')|x\right)} \\ &= \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right) + \mathbb{E}\left(\delta^\sigma|x\right)\left(\mathbb{E}\left(\sum_{t=0}^{\sigma'-1} \delta^t u(y^{+t})|y\right)\right)}{a(x,\sigma) + \mathbb{E}\left(\delta^\sigma|x\right) a(y,\sigma')} < \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right)}{a(x,\sigma)} = J(x,\sigma). \end{aligned}$$

Suppose next that x is non-augmenting and its index is finite. Then $a(x,\hat{\sigma}) < 0$. From (11), if y is non-augmenting and $a(y,\sigma') < 0$, then

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right)}{-a(x,\sigma)} = J(x,\sigma) < J(y,\sigma') = \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma'-1} \delta^t u(y^{+t})|y\right)}{-a(y,\sigma')}.$$

Otherwise, $a(y,\sigma') \geq 0$. In either case,

$$J(x,\hat{\sigma}) = \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right) + \mathbb{E}\left(\sum_{t=\sigma}^{\sigma+\sigma'-1} \delta^t u(y^{+t})|y\right)}{-a(x,\sigma) - \mathbb{E}\left(\delta^\sigma a(y,\sigma')|x\right)} > \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1} \delta^t u(x^{+t})|x\right)}{-a(x,\sigma)} = J(x,\sigma).$$

Finally, suppose x is non-augmenting and $J(x) = \infty$. If $J(x,\sigma) < \infty$, the previous argument

continues to hold. Suppose $J(x,\sigma)=\infty$. Since $J(y)\succ J(x)$, y must be augmenting. Since $a(x,\hat{\sigma})=a(x,\sigma)+\mathbb{E}\left(\delta^{\sigma}|x\right)a(y,\sigma')$, and since $a(y,\sigma')>0$ from the feasibility of σ', it follows that $a(x,\sigma)>0$, a contradiction.

Therefore, in each of the cases, σ cannot attain the value of the index $\mathcal{I}(x)$.

Step 2. Any feasible stopping rule that continues at a state y for which $\mathcal{I}(x)\triangleright\mathcal{I}(y)$ does not attain the value of the index. Fix a state y such that $\mathcal{I}(x)\triangleright\mathcal{I}(y)$, and define $\sigma'=\min\{t:x^{+t}=y\}$. Consider a feasible stopping rule σ that stops at y with positive probability and satisfies

$$\mathcal{I}(x,\sigma)\triangleright\mathcal{I}(y). \tag{12}$$

Consider the following stopping rule $\hat{\sigma}=\min\{\sigma,\sigma'\}$ that coincides with σ, except when the state is y, in which case $\hat{\sigma}$ stops.

Consider the additive case. From (12),

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1}\delta^{t}u(x^{+t})|x\right)+a(x,\sigma)}{\mathbb{E}\left(\sum_{t=0}^{\sigma-1}\delta^{t}|x\right)}=I(x,\sigma)>I(y)\geq\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^{t}u(y^{+t})|y\right)+a(y,\sigma-\hat{\sigma})}{\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^{t}|y\right)}.$$

Therefore,

$$\begin{aligned}I(x,\sigma)&=\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^{t}u(x^{+t})|x\right)+a(x,\hat{\sigma})+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)\left(\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^{t}u(y^{+t})|y\right)+a(y,\sigma-\hat{\sigma})\right)}{\mathbb{E}\left(\sum_{t=0}^{\sigma-1}\delta^{t}|x\right)+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^{t}|y\right)}\\&<\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^{t}u(x^{+t})|x\right)+a(x,\hat{\sigma})}{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^{t}|x\right)}=I(x,\hat{\sigma}).\end{aligned}$$

Turning to the multiplicative case, suppose that x is non-augmenting. Then, $a(x,\sigma)\leq 0$, and from (12), y is non-augmenting and $J(y)$ is finite. It then follows that $a(x,\sigma)<0$. To see this, note that since $J(y)$ is finite, $a(y,\sigma-\hat{\sigma})<0$, and since x is non-augmenting, $a(x,\hat{\sigma})\leq 0$. Hence, $a(x,\sigma)=a(x,\hat{\sigma})+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)a(y,\sigma-\hat{\sigma})<0$. From (12), we therefore have

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1}\delta^{t}u(x^{+t})|x\right)}{-a(x,\sigma)}=J(x,\sigma)>J(y)\geq J(y,\sigma-\hat{\sigma})=\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^{t}u(y^{+t})|y\right)}{-a(y,\sigma-\hat{\sigma})},$$

where $a(y,\sigma-\hat{\sigma})<0$ since $J(y)$ is inferior to $J(x)$. Therefore,

$$\begin{aligned}J(x,\sigma)&=\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^{t}u(x^{+t})|x\right)+\mathbb{E}\left(\sum_{t=\hat{\sigma}}^{\sigma-1}\delta^{t}u(y^{+t})|y\right)}{-a(x,\hat{\sigma})-\mathbb{E}\left(\delta^{\hat{\sigma}}a(y,\sigma)|x\right)}\\&=\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^{t}u(x^{+t})|x\right)+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)\left(\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^{t}u(y^{+t})|y\right)\right)}{-a(x,\hat{\sigma})-\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)a(y,\sigma-\hat{\sigma})}<\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^{t}u(x^{+t})|x\right)}{-a(x,\hat{\sigma})}=J(x,\hat{\sigma}).\end{aligned}$$

Suppose x is augmenting. Since σ is feasible, $a(x,\sigma)>0$. Suppose $a(y,\sigma-\hat{\sigma})>0$. Then

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-1}\delta^{t}u(x^{+t})|x\right)}{a(x,\sigma)}=J(x,\sigma)<J(y)\leq J(y,\sigma-\hat{\sigma})=\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^{t}u(y^{+t})|y\right)}{a(y,\sigma-\hat{\sigma})}. \tag{13}$$

In this case, $a(x,\hat{\sigma})>0$; that is, $\hat{\sigma}$ is feasible. To see this, note that, from (13),

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^t u(x^{+t})|x\right)+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)\left(\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^t u(y^{+t})|y\right)\right)}{a(x,\hat{\sigma})+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)a(y,\sigma-\hat{\sigma})}<\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^t u(y^{+t})|y\right)}{a(y,\sigma-\hat{\sigma})}.$$

Cross multiplying and rearranging gives

$$0\leq\frac{a(y,\sigma-\hat{\sigma})\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^t u(x^{+t})|x\right)}{\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^t u(y^{+t})|y\right)}<a(x,\hat{\sigma}).$$

Alternatively, suppose $a(y,\sigma-\hat{\sigma})\leq 0$. In this case, $a(x,\sigma)=a(x,\hat{\sigma})+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)a(y,\sigma-\hat{\sigma})>0$ implies $a(x,\hat{\sigma})>0$. In both the cases $a(y,\sigma-\hat{\sigma})>0$ and $a(y,\sigma-\hat{\sigma})\leq 0$,

$$\begin{aligned}J(x,\sigma)&=\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^t u(x^{+t})|x\right)+\mathbb{E}\left(\sum_{t=\hat{\sigma}}^{\sigma-1}\delta^t u(y^{+t})|y\right)}{a(x,\hat{\sigma})+\mathbb{E}\left(\delta^{\hat{\sigma}}a(y,\sigma)|x\right)}\\&=\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^t u(x^{+t})|x\right)+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)\left(\mathbb{E}\left(\sum_{t=0}^{\sigma-\hat{\sigma}-1}\delta^t u(y^{+t})|y\right)\right)}{a(x,\hat{\sigma})+\mathbb{E}\left(\delta^{\hat{\sigma}}|x\right)a(y,\sigma-\hat{\sigma})}>\frac{\mathbb{E}\left(\sum_{t=0}^{\hat{\sigma}-1}\delta^t u(x^{+t})|x\right)}{a(x,\hat{\sigma})}=J(x,\hat{\sigma}).\end{aligned}$$

Therefore, in each of the cases, σ cannot attain the value of the index $\mathcal{I}(x)$.

Step 3. There exists a feasible stopping rule that satisfies the necessary conditions of steps 1 and 2 and that attains the value of the index $\mathcal{I}(x)$. Suppose the value of the index is not attained by any stopping rule. Let $\overline{\tau}$ be a stopping rule defined as follows. For any x, $\overline{\tau}(x)=\min\{t:\mathcal{I}(x)\triangleright\mathcal{I}(x^{+t})\}$. That is, $\overline{\tau}(x)$ is the first time at which the value of the index becomes strictly inferior to $\mathcal{I}(x)$. Consider a stopping rule σ that satisfies the necessary conditions implied by steps 1 and 2 that stops before $\hat{\tau}(x)$ with positive probability and satisfies $\mathcal{I}(x)\triangleright\mathcal{I}(x,\sigma)\equiv\mathcal{I}$. Assume that σ stops at a time before $\overline{\tau}(x)$ when the state is y. By steps 1 and 2, it must be that $\mathcal{I}(x)=\mathcal{I}(y)$. We can then find a feasible rule σ' such that $\mathcal{I}(x,\sigma')\triangleright\frac{\mathcal{I}+\mathcal{I}(x)}{2}$. Define σ' accordingly for the states in which σ stops before $\overline{\tau}(x)$, and let $\sigma'=0$ whenever $\sigma=\overline{\tau}(x)$. Let $\sigma_1=\sigma+\sigma'$. Repeating the argument of step 1, we have that $\sigma\leq\sigma_1\leq\overline{\tau}$ and $\mathcal{I}(x)\triangleright\mathcal{I}(x,\sigma_1)\equiv\mathcal{I}_1\triangleright\mathcal{I}(x,\sigma)$ (note that σ_1 is feasible as a concatenation of σ and σ', which are both feasible). We can continue to construct a sequence of stopping rules such that $\sigma\leq\sigma_1\leq...\leq\sigma_n\leq\overline{\tau}(x)$ and $\mathcal{I}(x)\triangleright\mathcal{I}(x,\sigma_n)\equiv\mathcal{I}_n\triangleright\mathcal{I}(x,\sigma_{n-1})$. This sequence either continues indefinitely, or there exists some n_0 such that $\sigma_{n_0}=\overline{\tau}(x)$, in which case we let $\sigma_n=\overline{\tau}(x)$ for all $n>n_0$. By construction, in each step σ_n either coincides with $\bar{\tau}(x)$ or increases (for any state at which $\sigma_n<\bar{\tau}(x)$) by at least a period. Hence, $\sigma_n\to\bar{\tau}(x)$ a.s. From the monotone convergence theorem, it follows that $\mathcal{I}(x,\sigma_n)\to\mathcal{I}(x,\bar{\tau}(x))$. But note that this implies that $\mathcal{I}(x,\overline{\tau}(x))\triangleright\mathcal{I}(x,\sigma)$. Hence, $\overline{\tau}(x)$ attaines the value of the index $\mathcal{I}(x)$, a contradiction.

Step 4. The value of the index $\mathcal{I}(x)$ is attained by $\overline{\tau}(x)$. Suppose that a feasible stopping rule σ satisfies the necessary conditions of steps 1 and 2, and attains the value of the index $\mathcal{I}(x)$. Assume that σ stops before $\overline{\tau}$ with positive probability and that $\mathcal{I}(x,\sigma)=\mathcal{I}(x)$. Consider the event that σ stops before $\overline{\tau}(x)$ when the state is y. By steps 1 and 2, it must be that $\mathcal{I}(x)=\mathcal{I}(y)$. By step 3, we can find a feasible σ' such that $\mathcal{I}(y,\sigma')=\mathcal{I}(y)=\mathcal{I}(x)$. Define σ' accordingly for the states in which σ stops before $\overline{\tau}(x)$, and let $\sigma'=0$ whenever $\sigma=\overline{\tau}(x)$. Let $\sigma_1=\sigma+\sigma'$. Then, repeating the

argument in step 1, $\sigma \leq \sigma_1 \leq \overline{\tau}(x)$ and $\mathcal{I}(x) = \mathcal{I}(x, \sigma_1) = \mathcal{I}(x, \sigma)$. We can continue to consrtuct an increasing sequence of stopping times such that $\sigma_n \to \bar{\tau}(x)$ a.s., with $\mathcal{I}(x, \sigma_n) = \mathcal{I}(x)$. Hence, $\mathcal{I}(x, \overline{\tau}(x)) = \mathcal{I}(x)$.

Step 5. $\tau^*(x)$ attains the value of the index $\mathcal{I}(x)$. Whenever $\tau^*(x) < \overline{\tau}(x)$, we have $\overline{\tau}(x) - \tau^*(x) = \overline{\tau}(x^{+\tau^*})$, where $x^{+\tau^*}$ will denote the state at time $\tau^*(x)$, and from the previous steps, $\mathcal{I}(x^{+\tau^*}, \overline{\tau}(x) - \tau^*(x)) = \mathcal{I}(x^{+\tau^*}, \overline{\tau}(x^{+\tau^*})) = \mathcal{I}(x)$. We now show that $\mathcal{I}(x) = \mathcal{I}(x, \tau^*(x))$. Suppose $\mathcal{I}(x)$ is finite. In this case, we can denote the ratio $\mathcal{I}(x, \tau^*(x)) = \frac{\alpha}{\beta}$, and $\mathcal{I}(x^{+\tau^*}, \overline{\tau}(x) - \tau^*(x)) = \frac{\alpha'}{\beta'}$. By considerations similar to those in step 2, we can write $\mathcal{I}(x) = \mathcal{I}(x, \overline{\tau}(x)) = \frac{\alpha + \mathbb{E}\left(\delta^{\tau^*(x)}|x\right)\alpha'}{\beta + \mathbb{E}\left(\delta^{\tau^*(x)}|x\right)\beta'}$, which implies $\frac{\alpha'}{\beta'} = \frac{\alpha}{\beta}$. Finally, if $\mathcal{I}(x) = \infty$ (which occurs if and only $a(x, \bar{\tau}(x)) = 0$ in the multiplicative case), then $\mathcal{I}(x^{+\tau^*}, \overline{\tau}(x) - \tau^*(x)) = \infty$, and hence, $a(x^{+\tau^*}, \overline{\tau}(x) - \tau^*(x)) = 0$. Since $a(x, \bar{\tau}(x)) = a(x, \tau^*(x)) + \mathbb{E}\left(\delta^{\tau^*(x)}|x\right) a(x^{+\tau^*}, \overline{\tau}(x) - \tau^*(x))$, it follows that $a(x, \tau^*(x)) = 0$ and hence $\mathcal{I}(x, \tau^*(x)) = \infty$. ■

Proof of Theorem 2. We give a unified proof of the optimal policy for both the additive and multiplicative cases. Whenever the details of the arguments depend on the relevant case, we give a separate argument for each of them. Let π^0 be a policy that chooses some alternative i in period 0 and then proceeds according to the policy $\mathcal{P}^*$ from period 1 onward. In order to prove the optimality of $\mathcal{P}^*$, it is enough to show that the expected discounted payoff under the policy π^0 is no greater than the expected discounted payoff under $\mathcal{P}^*$, given any initial state $(x_{1,0}, ..., x_{n,0})$ of the DM.[14]

Let $(x_{1,0}, ..., x_{n,0})$ be the initial state of the DM's problem. Consider the policy π^0. If π^0 chooses in period 0 the same alternative $\mathcal{P}^*$ would have chosen, the two policies coincide in all periods. Therefore, suppose that π^0 chooses alternative i in period 0, while $\mathcal{P}^*$ would have chosen alternative $j \neq i$ in period 0. Note that this means that $\mathcal{I}_j(x_{j,0}) \trianglerighteq \mathcal{I}_i(x_{i,0})$. Also note that despite the fact that π^0 proceeds according to $\mathcal{P}^*$ from period 1 onward, π^0 need not choose j in period 1, since the state of alternative i may change as a result of being chosen in period 0. Define $\tau_k^*(x_k) = \min\{t > 0 : \mathcal{I}_k(x_k) \trianglerighteq \mathcal{I}_k(x_k^{+t}(x_k))\}$, where $x_k^{+t}(x_k)$ denotes alternative i's (stochastic) state after t periods of being chosen, starting in state x_k. In other words, beginning in state x_k, $\tau_k^*(x_k)$ is the first time at which the index $\mathcal{I}_k$ becomes weakly worse than $\mathcal{I}_k(x_k)$ according to $\trianglerighteq$.[15]

Denote by σ_1 the stochastic time at which an alternative other than i is chosen under π^0. Without loss of optimality, we can assume that this will be alternative j, and as j has not been chosen yet, its state in period σ_1 is equal to that of period 0. Let $\tau_j^*(x_{j,0})$ be the optimal stopping time in the definition of the index of j given state $x_{j,0}$. Setting $\sigma_2 = \tau_j^*(x_{j,0})$, π^0 will therefore choose alternative j from period σ_1 until (at least) period $\sigma_1 + \sigma_2 - 1$. At time $\sigma_1 + \sigma_2$, the index of alternative i will be $\mathcal{I}_i(x_i^{+\sigma_1}(x_{i,0}))$, the index of alternative j will be $\mathcal{I}_j(x_j^{+\sigma_2}(x_{j,0}))$, and the index of all other alternatives will be $\mathcal{I}_k(x_{k,0})$. Define the policy π^1 that initially picks alternative j during periods $0, ..., \sigma_2 - 1$, then picks alternative i during periods $\sigma_2, ..., \sigma_2 + \sigma_1 - 1$, and then coincides with $\mathcal{P}^*$ thereafter.

The final step of the proof will rely on the following Lemma. The details of its proof depend

[14] This follows from standard results in the literature on Markov decision processes.

[15] By Theorem 1, in the additive case, $\tau_k^*(x_k)$ attains the supremum in (4), and similarly, in the multiplicative case, $\tau_k^*(x_k)$ attains the infimum of (5) or the supremum of (6). We use this repeatedly throughout the proof.

on whether we are in the additive or multiplicative case. In order not to disrupt the flow of the proof of the Theorem, we give the separate proof of each case after the final step of the proof of Theorem 2.

Lemma 1. *The expected payoff under π^1 is weakly greater than that under π^0.*

If π^1 coincides with $\mathcal{P}^*$ during the periods $\sigma_2, ..., \sigma_2 + \sigma_1 - 1$, then π^1 and $\mathcal{P}^*$ are identical and the proof is complete. Otherwise, we can modify π^1 to a new policy π^2, repeating the argument in the preceding paragraphs.[16] We can proceed inductively and construct a sequence of policies $(\pi^0, \pi^1, \pi^2, ...)$, such that: (i) given the initial state $(x_{1,0}, ..., x_{n,0})$, π^{s+1} yields an expected discounted payoff no smaller than π^s, and (ii) the expected discounted payoff under π^s converges to the expected discounted payoff under $\mathcal{P}^*$ as $s \to \infty$ (to see this, note that π^s coincides with $\mathcal{P}^*$ for at least the first s periods). It follows that the expected discounted payoff under π^0 is no greater than under $\mathcal{P}^*$, which completes the proof. ■

Proof of Lemma 1. The policies π^0 and π^1 coincide from period $\sigma_1 + \sigma_2$ onward. Therefore we focus on periods $0, ..., \sigma_1 + \sigma_2 - 1$.

Additive case. Under π^1, the expected discounted payoff at periods $0, ..., \sigma_1 + \sigma_2 - 1$ is

$$\mathbb{E}\left(\sum_{k\neq i,j} \delta^{\sigma_1+\sigma_2-1} v_k(x_{k,0})\right) + \mathbb{E}\left(\sum_{t=0}^{\sigma_2-1} \delta^t u_j(x_{j,t})\right) + \mathbb{E}\left(\sum_{t=0}^{\sigma_2-1} \delta^t v_i(x_{i,0})\right)$$
$$+ \mathbb{E}\left(\delta^{\sigma_2} \sum_{t=0}^{\sigma_1-1} \delta^t v_j(x_{j,\sigma_2})\right) + \mathbb{E}\left(\delta^{\sigma_2} \sum_{t=0}^{\sigma_1-1} \delta^t u_i(x_{i,t})\right).$$

Similarly, the expected payoff during periods $0, ..., \sigma_1 + \sigma_2 - 1$ under π^0 is equal to

$$\mathbb{E}\left(\sum_{k\neq i,j} \delta^{\sigma_1+\sigma_2-1} v_k(x_{k,0})\right) + \mathbb{E}\left(\sum_{t=0}^{\sigma_1-1} \delta^t u_i(x_{i,t})\right) + \mathbb{E}\left(\sum_{t=0}^{\sigma_1-1} \delta^t v_j(x_{j,0})\right)$$
$$+ \mathbb{E}\left(\delta^{\sigma_1} \sum_{t=0}^{\sigma_2-1} \delta^t v_i(x_{i,\sigma_1})\right) + \mathbb{E}\left(\delta^{\sigma_1} \sum_{t=0}^{\sigma_2-1} \delta^t u_j(x_{j,t})\right).$$

Subtracting the two and rearranging, we have

$$\mathbb{E}\left(\sum_{t=0}^{\sigma_2-1} \delta^t u_j(x_{j,t})\right)(1 - \mathbb{E}(\delta^{\sigma_1})) + \frac{v_i(x_{i,0})\mathbb{E}(1-\delta^{\sigma_2})}{1-\delta} + \mathbb{E}(\delta^{\sigma_2} v_j(x_{j,\sigma_2}))\,\mathbb{E}\left(\frac{1-\delta^{\sigma_1}}{1-\delta}\right)$$
$$- \mathbb{E}\left(\sum_{t=0}^{\sigma_1-1} \delta^t u_i(x_{i,t})\right)(1 - \mathbb{E}(\delta^{\sigma_2})) - \frac{v_j(x_{j,0})\mathbb{E}(1-\delta^{\sigma_1})}{1-\delta} - \mathbb{E}(\delta^{\sigma_1} v_i(x_{i,\sigma_1}))\,\mathbb{E}\left(\frac{1-\delta^{\sigma_2}}{1-\delta}\right).$$

[16] In particular, consider the vector of states of all of the alternatives, starting from $(x_{1,0}, ..., x_{n,0})$ and having followed $\mathcal{P}^*$ in periods $0, ..., \sigma_2 - 1$. Suppose $\mathcal{P}^*$ would proceed to choose alternative $k \neq i$ at this stage (it may or may not be the case that $k = j$). Let $\tau_k^*(x_{k,\sigma_2})$ be the optimal stopping time in the definition of the index of k given state x_{k,σ_2}. The policy π^1 therefore chooses alternative j during periods $0, ..., \sigma_2 - 1$, then alternative i during $\sigma_2, ..., \sigma_2 + \sigma_1 - 1$, and then alternative k during (at least) $\sigma_2 + \sigma_1 + \tau_k^*(x_{k,\sigma_2}) - 1$. Denoting $\sigma_3 = \sigma_2 + \tau_k^*(x_{k,\sigma_2})$, define π^2 as follows. First, π^2 follows $\mathcal{P}^*$ during periods $0, ..., \sigma_3 - 1$, then it chooses alternative i during $\sigma_3, ..., \sigma_3 + \sigma_1 - 1$, and from then on it proceeds according to $\mathcal{P}^*$. Following precisely the same argument as above, π^2 yields an expected discounted payoff no smaller than π^1.

To see that this expression is non-negative note that multiplying by $\frac{1-\delta}{\mathbb{E}(1-\delta^{\sigma_1})\mathbb{E}(1-\delta^{\sigma_2})}$ and rearranging, the difference can be written as

$$\frac{(1-\delta)\mathbb{E}\left(\sum_{t=0}^{\sigma_2-1}\delta^t u_j(x_{j,t})\right)+\mathbb{E}\left(\delta^{\sigma_2}v_j(x_{j,\sigma_2})\right)-v_j(x_{j,0})}{\mathbb{E}\left(1-\delta^{\sigma_2}\right)}$$
$$-\left(\frac{(1-\delta)\mathbb{E}\left(\sum_{t=0}^{\sigma_1-1}\delta^t u_i(x_{i,t})\right)+\mathbb{E}\left(\delta^{\sigma_1}v_i(x_{i,\sigma_1})\right)-v_i(x_{i,0})}{\mathbb{E}\left(1-\delta^{\sigma_1}\right)}\right).$$

This difference is non-negative since the first summand is equal to $I_j(x_{j,0})$ (as σ_2 is an optimal stopping time) and the second summand is at most $I_i(x_{i,0})$ (as σ_1 is some stopping time), and $I_j(x_{j,0}) \geq I_i(x_{i,0})$. This completes the proof for the additive case.

Multiplicative case. The expected payoff during periods $0, ..., \sigma_1 + \sigma_2 - 1$ under π^1 is

$$\prod_{k\neq i,j} v_k(x_{k,0})\left\{v_i(x_{i,0})\mathbb{E}\left(\sum_{t=0}^{\sigma_2-1}\delta^t u_j(x_{j,t})\right)+\mathbb{E}\left(\delta^{\sigma_2}v_j(x_{j,\sigma_2})\right)\mathbb{E}\left(\sum_{t=0}^{\sigma_1-1}\delta^t u_i(x_{i,t})\right)\right\}. \tag{14}$$

Similarly, under π^0, the expected payoff during these periods is equal to

$$\prod_{k\neq i,j} v_k(x_{k,0})\left\{v_j(x_{j,0})\mathbb{E}\left(\sum_{t=0}^{\sigma_1-1}\delta^t u_i(x_{i,t})\right)+\mathbb{E}\left(\delta^{\sigma_1}v_i(x_{i,\sigma_1})\right)\mathbb{E}\left(\sum_{t=0}^{\sigma_2-1}\delta^t u_j(x_{j,t})\right)\right\}. \tag{15}$$

Denote by $\Delta(\pi^1, \pi^0)$ the difference between the expected discounted payoff under π^1 and its counterpart under π^0. Subtracting (15) from (14) and rearranging, we have that $\Delta(\pi^1, \pi^0)$ is

$$\prod_{k\neq i,j} v_k(x_{k,0})\left\{\mathbb{E}\left(\sum_{t=0}^{\sigma_2-1}\delta^t u_j(x_{j,t})\right)\left(v_i(x_{i,0})-\mathbb{E}\left(\delta^{\sigma_1}v_i(x_{i,\sigma_1})\right)\right)\right. \tag{16}$$
$$\left.-\mathbb{E}\left(\sum_{t=0}^{\sigma_1-1}\delta^t u_i(x_{i,t})|x_{i,0}\right)\left(v_j(x_{j,0})-\mathbb{E}\left(\delta^{\sigma_2}v_j(x_{j,\sigma_2})\right)\right)\right\}.$$

We now verify that $\Delta(\pi^1, \pi^0) \geq 0$. Recall that $J_j(x_{j,0}) \gtrsim J_i(x_{i,0})$. There are three cases.

Case 1. Suppose that $x_{j,0}$ is an augmenting state and $x_{i,0}$ is non-augmenting. Then, by the definition of $\sigma_2 = \tau_j^*(x_{j,0})$, $v_j(x_{j,0}) - \mathbb{E}\left(\delta^{\sigma_2}v_j(x_{j,\sigma_2})\right) < 0$, and by the definition of $J_i(x_{i,0})$, $v_i(x_{i,0}) - \mathbb{E}\left(\delta^{\sigma_1}v_i(x_{i,\sigma_1})\right) \geq 0$. This guarantees that $\Delta(\pi^1, \pi^0) \geq 0$.

Case 2. Suppose that $x_{j,0}$ and $x_{i,0}$ are both augmenting. Then $J_j(x_j) \gtrsim J_i(x_i)$ implies that $J_i(x_i) \geq J_j(x_j)$. Furthermore, $v_j(x_{j,0}) - \mathbb{E}\left(\delta^{\sigma_2}v_j(x_{j,\sigma_2})\right) < 0$ by the definition of σ_2. Suppose first that $v_i(x_{i,0}) - \mathbb{E}\left(\delta^{\sigma_1}v_i(x_{i,\sigma_1})\right) < 0$. Then

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma_1-1}\delta^t u_i(x_{i,t})\right)}{\mathbb{E}\left(\delta^{\sigma_1}v_i(x_{i,\sigma_1})\right)-v_i(x_{i,0})} \geq J_i(x_{i,0}) \geq J_j(x_{j,0}) = \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma_2-1}\delta^t u_j(x_{j,t})\right)}{\mathbb{E}\left(\delta^{\sigma_2}v_j(x_{j,\sigma_2})\right)-v_j(x_{j,0})}.$$

The first inequality follows from (5), while the equality follows from the fact that $\sigma_2 = \tau_j^*(x_{j,0})$ is the optimal stopping time in the definition of the index $J_j(x_{j,0})$. Rearranging and multiplying both sides by $\prod_{k\neq i,j} v_k(x_{k,0})$, we have that $\Delta(\pi^1, \pi^0) \geq 0$.

Now suppose that $v_i(x_{i,0}) - \mathbb{E}\left(\delta^{\sigma_1} v_i(x_{i,\sigma_1})\right) > 0$. Then

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma_1-1} \delta^t u_i(x_{i,t})\right)}{\mathbb{E}\left(\delta^{\sigma_1} v_i(x_{i,\sigma_1})\right) - v_i(x_{i,0})} \leq \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma_2-1} \delta^t u_j(x_{j,t})\right)}{\mathbb{E}\left(\delta^{\sigma_2} v_j(x_{j,\sigma_2})\right) - v_j(x_{j,0})},$$

and again $\Delta(\pi^1, \pi^0) \geq 0$. Finally, if $v_i(x_{i,0}) - \mathbb{E}\left(\delta^{\sigma_1} v_i(x_{i,\sigma_1})\right) = 0$ then clearly $\Delta(\pi^1, \pi^0) \geq 0$.

Case 3. Suppose that $x_{j,0}$ and $x_{i,0}$ are both non-augmenting. Then $J_j(x_j) \succsim J_i(x_i)$ implies that $J_i(x_i) \leq J_j(x_j)$. Furthermore, $v_i(x_{i,0}) - \mathbb{E}\left(\delta^{\sigma_1} v_i(x_{i,\sigma_1})\right) \geq 0$ and $v_j(x_{j,0}) - \mathbb{E}\left(\delta^{\sigma_2} v_j(x_{j,\sigma_2})\right) \geq 0$, and by (6),

$$\frac{\mathbb{E}\left(\sum_{t=0}^{\sigma_1-1} \delta^t u_i(x_{i,t})\right)}{v_i(x_{i,0}) - \mathbb{E}\left(\delta^{\sigma_1} v_i(x_{i,\sigma_1})\right)} \leq J_i(x_{i,0}) \leq J_j(x_{j,0}) = \frac{\mathbb{E}\left(\sum_{t=0}^{\sigma_2-1} \delta^t u_j(x_{j,t})\right)}{v_j(x_{j,0}) - \mathbb{E}\left(\delta^{\sigma_2} v_j(x_{j,\sigma_2})\right)}.$$

Rearranging and multiplying by $\prod_{k \neq i,j} v_k(x_{k,0})$, we have that $\Delta(\pi^1, \pi^0) \geq 0$, which completes the proof for the multiplicative case. ■

Proof of Theorem 3. *Part 1.* The proof is by induction on the number k of alternatives in an augmenting state. Suppose first that initially there is a single alternative i in an augmenting state. By Theorem 2, alternative i will be chosen until it becomes non-augmenting. Therefore, it suffices to show that choosing this alternative repeatedly will eventually bring it to a non-augmenting state with positive probability. Denote by $\hat{A}_i \subseteq A_i$ the set of augmenting states of i such that, starting from such states, the alternative remains augmenting forever with probability 1. We will show that $\hat{A}_i = \emptyset$. Assume by contradiction otherwise, and let $x_i \in \hat{A}_i$. As x_i is augmenting, there exists a stopping rule τ such that $a_i(x_i, \tau) = \mathbb{E}(\delta^\tau v_i(x_i^{+\tau})) - v_i(x_i) > 0$. Note that since $v_i(\cdot)$ is bounded, the stopping rule τ must stop at a finite time with positive probability.

Lemma 2. *Starting at $x_i \in \hat{A}_i$, the probability that τ stops at finite time outside $\hat{A}_i$ is zero.*

Proof of Lemma 2. Assume by contradiction that, starting from x_i, τ stops at finite time in $\hat{A}_i^C$ with probability $\mu > 0$.[17] Let $\hat{A}_i^C(q)$ denote the set of states of i such that, starting from such states, the alternative eventually reaches a non-augmenting state with probability greater than $q \geq 0$. Observe that $\hat{A}_i^C = \cup_{q>0} \hat{A}_i^C(q)$. Next, for any $q \geq 0$, denote by $E_{x_i,\tau}(q)$ the event that, starting from x_i, the rule τ stops at finite time in $\hat{A}_i^C(q)$. Note that $E_{x_i,\tau}(0)$ is the event that, starting from x_i, the rule τ stops at finite time in $\hat{A}_i^C$, and hence $\Pr(E_{x_i,\tau}(0)) = \mu$.

We now argue that $\Pr(E_{x_i,\tau}(q)) > \mu/2$ for some $q > 0$. Otherwise, there would exist a decreasing sequence $\{q_m\}_{m=1}^\infty \searrow 0$ such that $E_{x_i,\tau}(q_1) \subset E_{x_i,\tau}(q_2) \subset ...$ and $\Pr(E_{x_i,\tau}(q_m)) \leq \mu/2$ for all m. By Kelly's Nested Sets Theorem, $\lim_{m\to\infty} \Pr(E_{x_i,\tau}(q_m)) = \Pr(\cup_{m=1}^\infty E_{x_i,\tau}(q_m)) \leq \mu/2$. But note that by definition of μ, $\Pr(\cup_{m=1}^\infty E_{x_i,\tau}(q_m)) = \Pr(E_{x_i,\tau}(0)) = \mu$. We have thus shown that $\Pr(E_{x_i,\tau}(q)) > \mu/2$ for some $q > 0$. Hence, for such q, starting from x_i, the alternative reaches a non-augmenting state with probability at least $q\mu/2$, contradicting the assumption that $x_i \in \hat{A}_i$. □

The fact that $a_i(x_i, \tau) > 0$, together with the lemma above, implies that there exists a state $x_i' \in \hat{A}_i$ such that $v(x_i') > v(x_i)/\delta$. Applying the entire argument above to x_i', proves the existence

[17] For any set Z, we use Z^C to denote the complement of Z.

of a state $x_i'' \in \hat{A}_i$ such that $v(x_i'') > v(x_i')/\delta$. Applying this argument repeatedly therefore contradicts the assumption that $v(\cdot)$ is bounded. Therefore, $\hat{A}_i = \emptyset$. That is, choosing alternative i repeatedly will eventually bring it to a non-augmenting state with positive probability. We have therefore established the claim for $k = 1$. Suppose we have shown that when there are at most $k \in \{1, ..., n-1\}$ alternatives in augmenting states, eventually, with positive probability, all of them become non-augmenting. To conclude the proof of part 1, we show that this holds also if we start with $k+1$ alternatives in augmenting states. Recall that under the opitmal policy, in each period, whenever there are alternatives in an augmenting state, one of them is selected, and only this alternative can become non-augmenting at that period. Pick one of the $k+1$ alternatives that are initially in an augmenting state and denote it by i. The optimal policy must either choose i infinitely often, or choose among the remaining k alternatives infinitely often (or both). In either case, by the induction hypothesis, eventually, with positive probability, one of the $k+1$ alternatives becomes non-augmenting with positive probability. At that stage, the claim follows by the induction hypotehsis. Part 1 of the Theorem then follows.

Part 2. Suppose that all alternatives are at non-augmenting states. By the optimal policy, whenever a selected alternative becomes augmenting, it will be selected thereafter until it becomes non-augmenting again. ■

Proof of Proposition 1. Let $t_i(s)$ be the calendar time at which the alternative i is chosen for the s'th time, where $s \geq 1$. Given the collection of realization paths, denote the realized value of $w_i(x_{i,s-1})$ by $w_i(x_{i,s-1}, x_{i,s})$ (for example, the payoff from choosing i for the first time in the fictitious environment $\hat{\mathcal{E}}$ is $w_i(x_{i,0}, x_{i,1})$). We can write the payoff in $\hat{\mathcal{E}}$ given the assumed policy and collection of realization paths as follows:

$$
\begin{aligned}
&\sum_{i=1}^{n}\left(\sum_{s=1}^{\infty}\delta^{t_i(s)}w_i(x_{i,s-1},x_{i,s})\right)\\
&=\sum_{i=1}^{n}\left[\sum_{s=1}^{\infty}\delta^{t_i(s)}\left(u_i(x_{i,s-1})-v_i(x_{i,s-1})+\sum_{r=1}^{\infty}\delta^{r}\left(v_i(x_{i,s})-v_i(x_{i,s-1})\right)\right)+\frac{v_i(x_{i,0})}{1-\delta}\right]\\
&=\sum_{i=1}^{n}\left[\sum_{s=1}^{\infty}\delta^{t_i(s)}\left(u_i(x_{i,s-1})-v_i(x_{i,s-1})\right)+\sum_{s=1}^{\infty}\sum_{r=1}^{\infty}\delta^{t_i(s)+r}\left(v_i(x_{i,s})-v_i(x_{i,s-1})\right)+\frac{v_i(x_{i,0})}{1-\delta}\right]\\
&=\sum_{i=1}^{n}\left[\sum_{s=1}^{\infty}\delta^{t_i(s)}\left(u_i(x_{i,s-1})-v_i(x_{i,s-1})\right)+\sum_{t=1}^{\infty}\sum_{\{s\geq 1:t_i(s)<t\}}\left(\delta^{t}\left(v_i(x_{i,s})-v_i(x_{i,s-1})\right)\right)+\frac{v_i(x_{i,0})}{1-\delta}\right]\\
&=\sum_{i=1}^{n}\left[\sum_{s=1}^{\infty}\delta^{t_i(s)}\left(u_i(x_{i,s-1})-v_i(x_{i,s-1})\right)\right]+\sum_{i=1}^{n}\left[\sum_{t=0}^{\infty}\delta^{t}v_i(x_{i,s_i(t)})\right],
\end{aligned}
$$

where $s_i(t)$ is the largest $s \geq 1$ such that $t_i(s) < t$, and if such s does not exist, then $s_i(t) = 0$. The last expression is precisely the payoff in the original environment $\mathcal{E}$ of the additive case under the assumed policy and collection of realization paths. This is because, for each alternative i, when it is chosen for the s'th time at time $t_i(s)$, the DM receives the active payoff $u_i(x_{i,s-1})$, and in all periods $t_i(s) < t < t_i(s+1)$ receives the passive payoff $v_i(x_{i,s})$ from the alternative. ■

Proof of Theorem 4. *Step 1 - from a search problem to an experimentation problem.* The

problem in Section 4.1 is one where the decision problem ends when a box is selected. The framework described in Section 2, however, has an infinite horizon, and the DM chooses indefinitely among the alternatives. We map the Pandora's boxes problem with complementarities into an auxiliary problem that fits in our setting. This auxiliary setting is defined as follows.

Auxiliary problem. Denote by x_0 the initial state of any unopened box. For any unopened box i, let the payoff $u_i(x_0)$ be 0, and let $v_i(x_0) = 1$. Opening a box corresponds to choosing it for the first time. When a box is chosen for the first time, its state transitions from x_0 to the box's realization of (ω_i, ν_i), $x_i = (\omega_i, \nu_i)$. Choosing a box for the second time, when its state is $x_i = (\omega_i, \nu_i)$, yields a payoff of $u_i(x_i) = \omega_i(1-\delta)$. The passive payoff corresponding the such a box is ν_i. When a box i is chosen for the k'th time, $k \geq 2$, its state remains unchanged (and hence u_i and v_i remain the same).

The above auxiliary problem is a special case of the general setting in Section 2, and therefore the optimal policy for the auxiliary problem is a special case of $\mathcal{P}^*$. The crucial point is the following – since the states of all boxes (and hence their indices) do not change after the box is chosen for the second time, if the DM finds it optimal to choose a box for the second time in the auxiliary problem, they will continue to find it optimal to do so in all subsequent periods, yielding a continuation payoff of $\frac{u_i \prod_{j \neq i} v_j}{1-\delta} = \frac{\omega_i(1-\delta)\prod_{j \neq i} v_j}{1-\delta} = \omega_i \prod_{j \in O/\{i\}} \nu_j$. Choosing a box for the second time – under the optimal policy of the auxiliary problem – is therefore the same as stopping and taking the box in Pandora's problem with complementarities and receiving a one time payoff of $\omega_i \prod_{j \neq i} \nu_j$. In other words, the optimal policy for the auxiliary problem immediately delivers the optimal policy in Pandora's boxes problem with complementarities: choosing a box for the first time corresponds to opening it, and choosing it for the second time corresponds to stopping and taking it.

Step 2 – Optimal policy for the auxiliary problem. Consider the auxiliary problem described in the previous step. Let us first consider when a box is in an augmenting state. Note that a box can be in an augmenting state only if it has not been opened, as $v(x_0) = 1$ but $v_i = \nu_i$ at all periods after the box has been opened. Furthermore, since the state of a box changes only after it is opened for the first time and thereafter remains the same, it is enough to consider the stopping time $\tau \equiv 1$, and thus, in particular, whether $\delta\nu_i - 1 > 0$. A box i is therefore augmenting if and only if $\mathbb{E}[\nu_i] > \frac{1}{\delta}$. Note that any augmenting state of a box i has the same index, 0, since $u_i(x_0) = 0$. Next, consider a box that has already been opened. Such a box can only be in a non-augmenting state, and the optimal stopping time in (6) can be taken to be $\tau = 1$ by Theorem 1 (as the state of the box, and therefore its index, no longer changes). The index of a box that has been opened is therefore $J_i(x_i) = \frac{u_i(\omega_i)}{\nu_i - \delta\nu_i} = \frac{\omega_i}{\nu_i}$.

It remains to derive the initial index $J_i(x_0)$ of an unopened box that is in a non-augmenting state. By Theorem 1, the optimal stopping rule τ^* attaining $J_i(x_0)$ must take the following form: τ^* stops if the index $J_i(x_i)$ given the realization $x_i = (\omega_i, \nu_i)$ is no greater than the initial index of the box, $J_i(x_0)$ – in this case, the index $J_i(x_i)$ will be equal to 0, as $u_i(x_0) = 0$. If the index given the realization of (ω_i, ν_i) is strictly greater than the initial index $J_i(x_0)$, however, then τ^* continues forever. The index $J_i(x_0)$ therefore takes the value 0 for realizations (ω_i, ν_i) such that

$J_i(\omega_i,\nu_i) \leq J_i(x_0)$, and $\frac{\delta \mathbb{E}(u_i(\omega_i))/(1-\delta)}{v_i(x_0)-\mathbb{E}(\delta^\infty \nu_i)} = \frac{\delta \mathbb{E}(u_i(\omega_i))}{1-\delta}$ otherwise. Hence,

$$\begin{aligned} J_i(x_0) &= \frac{\delta \int_{J_i(\omega_i,\nu_i)>J_i(x_0)} \frac{u_i(\omega_i)}{1-\delta} dF_i(\omega_i,\nu_i)}{1-\left(\int_{J_i(\omega_i,\nu_i)\leq J_i(x_0)} \delta\nu_i dF_i(\omega_i,\nu_i) + \int_{J_i(\omega_i,\nu_i)>J_i(x_0)} \delta^\infty \nu_i dF_i(\omega_i,\nu_i)\right)} \\ &= \frac{\delta \int_{\omega_i > v_i J_i(x_0)} \omega_i dF_i(\omega_i,\nu_i)}{1-\delta \int_{\omega_i \leq v_i J_i(x_0)} \nu_i dF_i(\omega_i,\nu_i)}, \end{aligned}$$

where the last equality follows from the fact that given a realization (ω_i,ν_i), when the state is non-augmenting, $J_i(\omega_i,\nu_i) = \frac{\omega_i}{\nu_i}$. Rearranning, we have

$$J_i(x_0) = \delta\left(\int_{\omega_i > v_i J_i(x_0)} \omega_i dF_i(\omega_i,\nu_i) + \int_{\omega_i \leq v_i J_i(x_0)} J_i(x_0)\nu_i dF_i(\omega_i,\nu_i)\right).$$

The latter corresponds to the reservation prize R in the statement of the theorem, and extends Weitzman's notion of a reservation prize to the case of complementarities.

To conclude the proof of the theorem, note that from step 2, the optimal policy in the auxiliary problem is precisely the one in the statement of the theorem, except that – as explained in step 1 – whereas in the auxiliary problem once a box is chosen for the second time it will continue to be chosen indefinitely, in Pandora's problem with complementarities this second time choosing a box corresponds to stopping and taking it. The result therefore follows from steps 1 and 2 above ■

Proof of Proposition 2. To prove (1), note that

$$R_F = \delta \int_\omega \left(\int_\nu \max\{\omega, R\nu\} dF_\omega(\nu)\right) dF(\omega) > \delta \int_\omega \left(\int_\nu \max\{\omega, R\nu\} dG_\omega(\nu)\right) dF(\omega) = R_G.$$

To prove (2), fix a realization (ω,ν) for which $\omega > \nu R_F$ (this corresponds to selecting the alternative immediately). Associate this realization with the realization (ω,ν') where $\nu'(\nu) = G_\omega^{-1}[F_\omega(\nu)]$. By FOSD, $\nu' < \nu$. By part (1), $R_G < R_F$. Hence $\omega > \nu' R_G$. ■

Proof of Proposition 3. The reservation prize corresponding to an investor that has yet to be approached is given by

$$\begin{aligned} R &= \delta\left(\int_{\omega > R\nu(\omega)} \omega dF(\omega) + \int_{\omega < R\nu(\omega)} R\nu(\omega) dF(\omega)\right) \\ &= \delta\left(\int_{\omega > \frac{R}{1+R\beta}} \omega dF(\omega) + \int_{\omega < \frac{R}{1+R\beta}} R(1-\beta\omega) dF(\omega)\right) \\ &= \delta\left(\int_{\frac{R}{1+R\beta}}^1 \omega d\omega + \int_0^{\frac{R}{1+R\beta}} R(1-\beta\omega) d\omega\right) = \frac{1}{2}\delta\left(1 + \frac{R^2}{1+R\beta}\right), \end{aligned}$$

where the last equality follows after some algebra. The reservation prize R therefore solves the following equation, $R^2\left(\frac{1}{2}\delta - \beta\right) - R(1-\frac{1}{2}\delta\beta) + \frac{1}{2}\delta = 0$, and it can be verified that the relevant solution to this equation is $R = \frac{-2+\beta\delta+\sqrt{4+4\beta\delta-\delta^2(4-\beta^2)}}{2(2\beta-\delta)}$. The result then follows immediately from Theorem 4. ■

Proof of Proposition 4. We say that an agent is currently in state k if $k-1$ projects were

assigned to him previously. For each $k \in \mathbb{N}$, let us compare the fractions in (4) for $\tau = 1$,

$$I(k,1) = \frac{(1-\delta)\beta\frac{1-\theta^k}{1-\theta} + \delta\beta(1-\beta)\frac{1-\theta^{k+1}}{1-\theta} - \beta(1-\beta)\frac{1-\theta^k}{1-\theta}}{1-\delta},$$

and $\tau = 2$,

$$I(k,2) = \frac{(1-\delta)\beta\left(\frac{1-\theta^k}{1-\theta} + \delta\frac{1-\theta^{k+1}}{1-\theta}\right) + \delta^2\beta(1-\beta)\frac{1-\theta^{k+2}}{1-\theta} - \beta(1-\beta)\frac{1-\theta^k}{1-\theta}}{1-\delta^2}.$$

Combining and rearranging terms yields $I(k,2) - I(k,1) = \frac{\delta\theta^k\beta}{1-\delta^2}\left(\beta - \delta + \theta\delta - \theta\beta\delta\right)$. Note that $\operatorname{sgn}\left(I(k,2) - I(k,1)\right) = \operatorname{sgn}\left(\beta - \delta + \theta\delta - \theta\beta\delta\right)$ does not depend on k.

Consider $\delta < \hat{\delta}$. Then $\beta - \delta + \theta\delta - \theta\beta\delta > 0$ and $I(k,2) > I(k,1)$ for all k. We now show that $I(k) = I(k, \tau^* = \infty)$ for all $k \in \mathbb{N}$. Assume by contradiction that there is κ for which $I(\kappa) = (\kappa, \tau^*)$ for some finite τ^*. In this case, by arguments analogous to those given in Theorem 1, $I(\kappa+\tau^*-1) \geq I(\kappa) \geq I(\kappa+\tau^*)$. But then we have $I(\kappa+\tau^*-1) = I(\kappa+\tau^*-1, 1) \geq I(\kappa+\tau^*-1, 2)$, contradiction. It follows that if $\delta < \hat{\delta}$, the unique optimal policy is to allocate the project to the same agent in all periods.

Now suppose $\delta > \hat{\delta}$. Then $\beta - \delta + \theta\delta - \theta\beta\delta < 0$ and $I(k,2) < I(k,1)$ for all k. We show that the stopping rule $\tau^* = 1$ is uniquely optimal for all k. Suppose there is κ such that $I(\kappa) = (\kappa, \tau^*)$ for some finite $\tau^* > 1$. In this case, $I(\kappa + \tau^*) \leq I(\kappa) \leq I(\kappa + 1)$. Therefore, there exists $k' \in \{\kappa + 1, ..., \kappa + \tau^* - 1\}$ such that $I(k') \geq I(k'')$ for all $k'' \in \{\kappa, ..., \kappa + \tau^* - 1\}$. Hence, $I(k') = I(k', 1)$ and thus $I(k'-1, 2) \geq I(k'-1, 1)$, a contradiction. It is left to compare $I(k,1)$ with $I(k,\infty)$. Simple algebra yields

$$I(k,1) - I(k, \tau^* = \infty) = \frac{\delta\theta^k\beta}{(1-\delta)(\theta\delta - 1)}\left(\beta - \delta + \theta\delta - \theta\beta\delta\right) > 0,$$

where the inequality follows from $(\beta - \delta + \theta\delta - \theta\beta\delta) < 0$ and the fact that $\theta\delta < 1$. We have therefore shown that $I(k) = I(k,1) > I(k,\tau)$ for any $\tau \neq 1$, whenever $\delta > \hat{\delta}$. Thus, in this case, alternating suppliers every period is uniquely optimal.

Finally, any policy is optimal if $\delta = \hat{\delta}$ since in this case, $\beta - \delta + \theta\delta - \theta\beta\delta = 0$ and thus, for all k and τ, $I(k) = I(k,\tau)$. ■

Proof of Proposition 5. As in the proof of Proposition 5, we begin by computing the difference $I(k,2) - I(k,1)$, which reduces to

$$(\beta - \delta)\left[f(k+1) - f(k)\right] + \delta(1-\beta)\left[f(k+2) - f(k+1)\right]. \tag{17}$$

If $\delta < \beta$, then this expression is positive for all k. Hence, when $\beta > \delta$ we have that $I(k,2) > I(k,1)$ for all k. Using the same arguments as in the proof of Proposition 5, we conclude that $I(k) = I(k, \tau^* = \infty)$ for all k, and the unique optimal policy is to allocate the project to the same agent in all periods. ■

Proof of Proposition 6. Note that if (9) holds for all k, then $I(k,1) > I(k+1,1)$. We now show by induction that for all k, and all $n \geq 1$, $I(k,1) > I(k,n)$. Suppose we established this inequality

for n and consider $I(k, n+1)$. It is straightforward to show that

$$I(k,n+1) = \frac{I(k,1) + (\sum_{s=1}^{n} \delta^s) I(k+1,n)}{\sum_{s=0}^{n} \delta^s} < \frac{I(k,1) + (\sum_{s=1}^{n} \delta^s) I(k+1,1)}{\sum_{s=0}^{n} \delta^s}$$
$$< \frac{I(k,1) + (\sum_{s=1}^{n} \delta^s) I(k,1)}{\sum_{s=0}^{n} \delta^s} = I(k,1).$$

To see why, note first that $I(k, n+1)$ is the average per-period gain from $n+1$ consecutive choices beginning from state k. The same gain can be represented as the average of one choice at state k (i.e., $I(k,1)$) followed by discounted n consecutive choices at state $k+1$ beginning in the next period. This yields the first equality. The next inequalities follow from the induction hypothesis and the fact that $I(k,1) > I(k+1,1)$. This establishes that $I(k,1) > I(k,n)$ for all k and $n > 1$.

Next, note that

$$I(k,n) = \frac{\beta(1-\delta)\sum_{t=0}^{n-1} \delta^t f(k+t) - \beta(1-\beta) f(k)}{1-\delta^n}$$

and hence,

$$\lim_{n\to\infty} I(k,n) = \beta(1-\delta)\sum_{t=0}^{\infty} \delta^t f(k+t) - \beta(1-\beta) f(k) = I(k,\infty).$$

It then follows that $I(k,1) > I(k,\infty)$. ■

Proof of Proposition 7. *Indices.* First note that since $v_B(\cdot)$ is constant and $v_A(\cdot)$ may only decrease, $a_i(x_i, \tau) \leq 0$ for any stopping rule τ, and hence there are no augmenting states for the two alternatives. Next, observe that $J_B(\phi) = 0 < J_A((n,\chi))$ for any $(n,\chi) \in S_A$ and so A is selected whenever $x_B = \phi$. Finally, note that $a_A(n,0) = 0$ and so $J_A((n,0)) = \infty$ while $J_B(x_B)$ is finite. Hence, A is selected whenever $\chi = 0$.

Consider alternative A. We now show that the index at every state $(n,1)$ is given by the following stopping rule which we denote τ': *Select A once; stop if the resulting state is $(n+1,1)$, and continue forever if the state is $(n+1,0)$.*

Suppose the initial state is such that $\chi = 1$. Due to the state transition process, any stopping rule τ can be written as $(\kappa, \xi(k))$ where $\kappa \in \mathbb{N} \cup \{\infty\}$ and $\xi : \{1, .., \kappa\} \to \mathbb{N} \cup \{0, \infty\}$, such that if the state changes to one where $\chi = 0$ in period $k \in \{1, ..., \kappa\}$, the stopping time is $k + \xi(k)$; otherwise, the stopping time is κ. Consider two initial states $(n_1, 1)$, $(n_2, 1)$ such that $n_1 < n_2$. Note that for any stopping rule τ, $J_A((n_1,1),\tau) > J_A((n_2,1),\tau)$, as the denominators of the corresponding ratios are identical but the numerator of $J_A((n_1,1),\tau)$ is larger than that of $J_A((n_2,1),\tau)$. Since the inequality holds for any τ, we can conclude that $J_A((n_1,1)) > J_A((n_2,1))$. By Theorem 1 and the observation that $J_A((n,0)) = \infty$ for all $n \in \mathbb{N}$ (as $a((n,0),\tau) = 0$ for any τ), we can conclude that for each $n \in \mathbb{N}$, $J_A((n,1)) = J_A((n,1),\tau')$. Note that Theorem 1 also implies that $J_A((n,0)) = J_A((n,0),1)$. Direct calculation yields that $J_A((n_A,\chi)) = \frac{q^{n_A}\left(\frac{\delta pq}{1-\delta q}+1\right)}{1-(1-p)\delta}$ when $\chi \neq 0$ and $J_A((n_A,\chi)) = \infty$ when $\chi = 0$.

An analogous argument shows that $J_B(n_1) > J_B(n_2)$ for $n_1 < n_2$. By Theorem 1, for each $n \in \mathbb{N}$, $J_B(n) = J_B(n,1)$. In addition, note that Theorem 1 also implies that $J_B(\phi) = J_B(\phi,1)$. Hence, the index of B is given by: $J_B(x_B) = \frac{q^{n_B}}{1-\delta} \cdot \mathbf{1}_{\{x_B \neq \phi\}}$.

Optimal Policy. Using the indices above we can infer the optimal dynamics. As mentioned

earlier, when x_A is such that $\chi = 0$ or $x_B = \phi$, it is optimal to select A. Hence, suppose that $x_A = (n_A, 1)$ and $x_B = n_B \in \mathbb{N}$. Consider $J_A((n_A, 1))$ and $J_B(n_B)$ as functions of continuous variables n_A and n_B, respectively, and note that $J_A((n_A, 1))$ and $J_B(n_B)$ are continuous and decreasing in these variables. Solving for $J_A((n_A, 1)) = J_B(n_B)$ yields $n_A = n_B - c(\delta, p, q)$. Direct inspection shows that $c(\delta, p, q) > 0$ for all $(\delta, p, q) \in (0, 1)^3$. ■

Proof of Proposition 8. *Part 1.* Assume $a(s, \tau + 1) = \delta^{\tau+1} v(s + \tau + 1) - v(s) > 0$. Since

$$\begin{aligned}
&\delta^{\tau+1} v(s+\tau+1) - v(s) \\
&= \delta^{\tau} [\delta v(s+\tau+1) - v(s+\tau)] + \delta^{\tau-1} [\delta v(s+\tau) - v(s+\tau-1)] + .. + [\delta v(s+1) - v(s)] \\
&< \delta^{\tau} [\delta v(s+1) - v(s)] + \delta^{\tau-1} [\delta v(s+1) - v(s)] + .. + [\delta v(s+1) - v(s)] \\
&= (\delta^{\tau} + ... + 1) [\delta v(s+1) - v(s)],
\end{aligned}$$

we have that $\delta v(s+1) - v(s) > 0$.

Part 2. Assume an alternative is augmenting in state s. By part 1, $\delta v(s+1) - v(s) > 0$. By concavity, for any $s' < s$, $\delta v(s'+1) - v(s') > 0$. By Theorem 3, there exists some state in which the alternative is non-augmenting. ■

Proof of Proposition 10. We begin by deriving the indices for the two sectors. Consider first Sector A. It can be modeled as an alternative whose initial state is 0 and whenever A is selected its state increases by one until the absorbing state T is reached. We derive the indices by applying the algorithm described in Appendix A.

Step 1: To find α_1, recall that $C(\alpha_1) = \emptyset$. Hence, for all states s we compare

$$I_A(s, 1) = \begin{cases} (1-\gamma)(s+1)r + \frac{\gamma r}{1-\delta} & , \quad s < T \\ (1-\gamma)Tr & , \quad s = T. \end{cases}$$

$I_A(s, 1)$ is increasing in s for $s < T$, and $I_A(T-1, 1) > I_A(T, 1)$. Hence $\alpha_1 = T - 1$.

Step 2: To find α_2, note that $C(\alpha_2) = \{T-1\}$ and $S(\alpha_2) = \{0, ..., T\} - C(\alpha_2)$. Observe that selecting A once at any state $s \neq T-2$ will bring us to a state in $S(\alpha_2)$. Therefore, to find α_2 we must compare $I_A(s, 1)$ for $s \neq T-2$ and $I_A(T-2, 2)$. However, since $I_A(T-2, 2) > I_A(T-2, 1) > I_A(s, 1)$ for all $s < T-2$, $\alpha_2 \in \{T-2, T\}$. Our assumption on δ ensures that $\alpha_2 = T-2$. To see this, note that the expression $\frac{(1-\gamma)T-1}{(1-\gamma)(T-1)}$ in our assumption on δ is increasing in T and that $I_A(T-2, 1) \geq I_A(T, 1)$ holds whenever $\delta \geq \frac{(1-\gamma)(T-(T-2))-1}{(1-\gamma)((T-(T-2))-1)}$.

Continuing in the same manner, in Step $k \leq T$ we have $C(\alpha_k) = \{T-j\}_{j=1}^{k-1}$ and $S(\alpha_k) = \{0, ..., T\} - C(\alpha_k)$. Again, by comparing $I_A(s, 1)$ for all states $s \in \{0, ..., T-k\}$ and noting that $I_A(T-k, 1) < I_A(T-k, k)$, we conclude that $\alpha_k \in \{T-k, T\}$. As before, $I_A(T-k, 1) > I_A(T, 1)$ holds whenever $\delta \geq \frac{(1-\gamma)(T-(T-k))-1}{(1-\gamma)((T-(T-k))-1)}$, and thus, our assumption on δ implies that $\alpha_k = T-k$ and that $I_A(T-k) = I_A(T-k, k)$. Finally, in Step $T+1$, we conclude that $\alpha_{T+1} = T$ and, since the state T is absorbing, $I_A(T) = I_A(T, 1)$. To summarize, the indices of A are given by

$$I_A(s) = I_A(s, T - s + \mathbf{1}_{\{s=T\}}) = \begin{cases} \frac{r}{1-\delta} + rs(1-\gamma) - \frac{(1-\gamma)r(T-s)\delta^{(T-s)}}{1-\delta^{(T-s)}} & , \quad s < T \\ (1-\gamma)Tr & , \quad s = T. \end{cases}$$

Sector B can be modeled in a similar way with only two states. The initial state 0 changes to the

absorbing state 1 when B is selected. By Theorem 1, $I_B(1) = I_B(1,1) = (1-\beta)b$. It is immediate to verify that $I_B(0,1) > I_B(1)$, hence $I_B(0) = b\left[\frac{1-(1-\beta)\delta}{1-\delta}\right]$ and $I_B(1) = (1-\beta)b$.

To conclude the proof, note that $A_0 = I_A(0) = I_A(0,T)$, which implies that if the individual starts working in A, he will remain there for T consecutive periods. In addition, $A_1 = I_A(T)$, $B_0 = I_B(0)$, and $B_1 = I_B(1)$. Finally, $B_0 > B_1$ and $A_0 > A_1$. ■