Experimentation with Costly Project Maintenance*

Jean Guillaume Forand†

PRELIMINARY AND INCOMPLETE

April 14, 2010

Abstract

In standard multi-armed bandit problems, the act of choosing one arm over another entails only an implicit opportunity cost. I study a three-armed exponential bandit problem in which unused risky arms have explicit maintenance costs and can be irreversibly discarded. If all projects are maintained, the optimal experimentation policy has a Gittins index representation and furthermore has the *stay-with-the-winner* property. I characterise the optimal experimentation policy when arms can be discarded, which does not have an index representation, and show that either (i) maintenance costs are sufficiently high that the losing arm is always discarded immediately, or (ii) the stay-with-the-winner property fails. Maintenance costs alter the optimal experimentation by providing incentives to bring losing arms' option values forward.

1 Introduction

When experimentation is costly, decision-makers face choices about which alternatives to investigate. For example, firms that develop new technologies focus on a few projects at a time and do not fund all competing ideas equally. Investing in multiple technologies simultaneously is costly, so that firms choose a limited set of technologies to develop initially and then decide whether to transfer their resources to other projects when experimentation results are not sufficiently good. Professional sports teams can be thought of as managing their rosters in a similar way. Since the rules of the sport specify that a single player can play at a given position at a time, coaches/managers can infer information about players' abilities only by allowing them to enter the game and replace a teammate.

In standard models of experimentation, the choice of gathering information about one alternative as opposed to another entails only an implicit opportunity cost, the foregone opportunity

*I would like to thank Li Hao, Martin Osborne and Colin Stewart for their supervision, comments and suggestions.

†Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario, M5S 3G7. jgforand@yahoo.ca.

of learning about the inactive alternative. However, retaining the option to investigate a currently shelved alternative often involves explicit maintenance costs. For example, a research and development firm must maintain specialised stocks of knowledge simply to keep open the option of developing the technology that is not currently a priority. This consists of skilled workers or scientists that can be lost to other firms or the upkeep of specialised equipment. In professional sports, the option to develop players of unknown quality is kept open by filling roster spots with 'bench' players, who may seldom get the opportunity to play but command non-negligible salaries.

In this paper, I present a standard model of experimentation, but allow alternatives that are not being investigated to have explicit maintenance costs and to be irreversibly discarded if deemed unpromising. Two risky alternatives can be good or bad and only good alternatives eventually succeed if investigated. An experimenter searches for a success among the alternatives by choosing both which one to investigate now and whether or not to maintain the inactive one for (potential) future research. An alternative's current state is characterised by the experimenter's belief that it is good, and repeated failures make the experimenter more pessimistic about the alternative.

Without the option to discard alternatives, the optimal experimentation policy has a 'stay-with-the-winner' property: the alternative that is more likely to be good is investigated first. I fully characterise the optimal experimentation policy with maintenance costs and show that these lead to significant departures from this standard result. In particular, alternatives that are less likely to succeed are sometimes investigated first. In such cases, 'losing' alternatives are granted a 'last chance' to succeed, after which they are permanently discarded in favour of more promising alternatives.

Maintenance costs alter the optimal order in which alternatives are investigated by providing incentives to bring the option value of less promising alternatives forward. To do so, the experimenter needs to investigate the losing alternative first while paying to maintain the winning alternative. The benefit is that the decision to discard the losing alternative following repeated failures is better informed, and the experimenter can then investigate the winning alternative without having to maintain the losing one.

I model experimentation as a multi-armed bandit problem.[1] This is a well-known class of models in statistical decision theory in which an experimenter (e.g. a firm) learns about the distributions generating the payouts of various arms (e.g. technologies) by 'pulling' (e.g. developing) them sequentially over time. In general, these are complex multi-dimensional dynamic optimisation problems, but in the standard discounted multi-armed bandit problem with independent arms the optimal experimentation policy can be represented by a well-known (Gittins) index policy. To each arm is assigned a number (index) that depends only on the ex ante charac-

[1]See Berry and Fristedt (1985). Bergemann and Välimäki (2006) survey the bandits literature with an eye to applications in economics.

teristics and accumulated observations of that arm. The optimal experimentation policy consists of always selecting an arm among those with maximal indices.

The index representation of the optimal experimentation policy is not robust to perturbations in standard assumptions such as geometric discounting and independent arms. Closer to my paper, Banks and Sundaram (1994) have shown that index policies are not optimal in the presence of switching costs between arms. Intuitively, the reason is that switching costs link the experimenter's decision to move away from an active arm, that is, to incur a switching cost, with the characteristics of outside arms in a way that cannot be captured by a single index that depends only on the characteristics of the active arm. General characterisations of optimal experimentation policies with switching costs have proven difficult to obtain.[2]

Although my results are obtained in a special framework, maintenance costs appear to be more tractable than switching costs. Maintenance costs differ from switching costs in that the latter are attributed to an inactive arm only when experimentation transitions to it and are always accompanied by an observation from that arm. Maintenance costs need to be paid whenever an inactive arm is not pulled and never generate observations from that arm. It is not clear how to even define an index policy in the presence of maintenance costs since experimentation policies need to specify both which arm is pulled and which arms are maintained. Nevertheless, even if this difficulty could be surmounted the bandit problem with maintenance costs would fail to admit a Gittins index representation for the same reason Banks and Sundaram (1994) identified for bandit models with switching costs: the index of a given maintained arm would have to be a function of the maintenance cost, and this relationship would depend nontrivially on the characteristics of outside arms.

The bandit model I consider is a continuous-time exponential bandit due to Keller et al. (2005). The exponential bandit model has proved useful in applications due to its tractability. In particular, the experimenter's optimal payoff functions of the two-armed bandit (optimal stopping) problem are simple to compute, which is not the case for bandits with different probability structures. Keller et al. (2005), following Bolton and Harris (1999), study strategic experimentation and the free-riding incentives of multiple agents facing a single risky arm. Keller and Rady (2009) generalise the model to 'poisson' bandits that allow for arms of the bad type to also generate successes. Klein and Rady (2008) allow for each of two experimenters to have perfectly negatively correlated versions of the same risky arm, and hence the state can still be described by a single-dimensional belief. Strulovici (2009) applies the model in a voting framework. Bergemann and Hege (1998) have introduced a discrete-time version of the model to study the moral hazard problem arising between bankers (principal) and venture capitalists (experimenters). In this vein, recent papers by Bonatti and Hörner (2009) and Hörner and Samuelson (2009) focus on the provision of incentives to experimenting agents.

[2]For details, see Jun (2004). An exception is Bergemann and Välimäki (2001), who exploit results of Banks and Sundaram (1992b) on bandits with a countable numbers of ex ante identical arms to show that an experimenter never switches back to an arm it switched away from earlier.

I present the model in Secton 2. I address how to define strategies (which depend on time in the absence of a success) and Markov strategies (which depend on the state, the belief of the experimenter that each arm is good) and clarify the relationships between them. While Markov strategies along with dynamic programming methods and allow for simple expressions for optimal payoffs, many of the arguments regarding when and why maintained arms should be discarded are naturally established by considering time paths of play. In Section 3, I use the simple structure of the continuous-time exponential bandit problem to derive expressions for the experimenter's optimal payoffs. In Section 4, I characterise optimal experimentation in the benchmark model in which inactive arms have maintenance costs but the experimenter cannot discard risky arms. This is equivalent to a standard three-armed bandit with two risky arms, and hence the solution is known by the Gittins index theorem. I show that the optimal experimentation policy involves the stay-with-the-winner rule. In Section 5, I present the main results of the paper for the model in which inactive arms have maintenance costs and the experimenter can discard arms. First, I show that if the optimal policy ever 'goes-with-the-loser', it will do so in a very specific way. The losing arm will be chosen continuously for a short period, after which, in the absence of a success, it will be discarded. Losing arms are used before winning arms only if they are being granted a 'last chance' to succeed, else they are maintained and not used or simply discarded. Second, I give a complete characterisation of the optimal policy and show that whenever it is not the case that maintenance costs are high enough that the 'losing' arm is always discarded immediately, there exist regions of initial beliefs for which the experimenter starts with the losing arm.

2 Model

Consider a continuous time three-armed bandit problem with two risky arms, A and B, and a safe arm S. An experiment consists of pulling a risky arm for some time interval $[t, t+dt]$. The probabilistic structure of the risky arms is as in Keller et al. (2005). Experiments yield either successes or failures. The type of a risky arm is $\theta \in \{G, B\}$. A risky arm of type θ that is pulled continuously in time interval $[t, t+dt]$ succeeds with probability Gdt if $\theta = G$ and fails for sure if $\theta = B$. The types of risky arms A and B are drawn independently. Let $p_J(0)$ be the ex ante probability that arm J is of type G. A safe arm S yields a flow payoff of 0. A success on either risky arm yields a lump-sum payment of 1 and ends the experimentation process.

Pulling risky arm J continuously in time interval $[t, t+dt]$ entails experimentation costs $\overline{k}dt$. As my only departure from standard multi-armed bandit problems, I introduce explicit costs to maintaining inactive risky arms. That is, a risky arm that is maintained but not pulled in time interval $[t, t+dt]$ entails a cost of $\underline{k}dt$. The experimenter can irreversibly discard risky arms without cost. That is, it can avoid paying for the maintenance of inactive arms but only at the cost permanently abandoning some of its options. There are no costs to the safe arm, which can be interpreted as an option to quit the experimentation process. The experimenter discounts

future payoffs at rate r.

Since experimentation ends after the first success, the only histories after which the experimenter selects an arm to pull are intervals of time in which only failures have been observed. Strategies should properly be defined on histories, however, any such strategy can be redefined to depend solely on time in the absence of a success. A *strategy* is a collection (α, ϕ_A, ϕ_B) for some function $\alpha : \mathbf{R}_+ \to [0,1] \cup \{S\}$ and decreasing functions $\phi_J : \mathbf{R}_+ \to \{0,1\}$ for $J \in \{A, B\}$. The function α is an *assignment rule* and $\int_t^{t+dt} \alpha(t)$ specifies the fraction of time devoted to pulling arms A in time interval $[t, t+dt]$ if the experimenter pulls only risky arms in that interval, while $\alpha(t) = S$ if the experimenter pulls the safe arm at time t. The principal is allowed to share the responsibility for the project between the agents in any interval of time. The assumption that the experimenter cannot share the assignment between all three arms and must decide first whether or not to pull risky arms and then in what ratio is made to simplify the exposition and is in fact without loss of generality for optimal experimentation. Functions ϕ_A and ϕ_B specify *maintenance rules*, with $\phi_J(t) = 1$ if and only if J is maintained at time t. Strategy (α, ϕ_A, ϕ_B) is *admissible* if each component is right-continuous and piecewise Lipschitz continuous. Let $t_J \in [0, \infty) = \sup\{t : \phi_J(t) = 1\}$. Given any initial conditions $(p_A(0), p_B(0)) \in [0,1]^2$, admissible strategy (α, ϕ_A, ϕ_B) induces uniquely defined and continuously differentiable laws of motion for the beliefs $(p_A(t), p_B(t))$ that arms A and B are of type G at time t.[3] These laws of motion are given by

$$\dot{p}_A(t) = \begin{cases} -\alpha(t) H p_A(t)(1 - p_A(t)) \text{ for } t \in [0, t_A), \\ 0 \text{ for } t \geq t_A. \end{cases}$$

$$\dot{p}_B(t) = \begin{cases} -(1 - \alpha(t)) H p_B(t)(1 - p_B(t)) \text{ for } t \in [0, t_B), \\ 0 \text{ for } t \geq t_B. \end{cases}$$

These laws of motion are derived in a straightforward way by requiring that the evolution of beliefs be consistent with α and Bayes' rule, and follows Keller et al. (2005).

For much of the paper, it will be more convenient to work with Markov strategies. These are strategies that are conditioned on the state variable, which is the current belief along with the set of maintained arms. More formally, a *state* consists of $(p_A, p_B, I_A, I_B) \in [0,1]^2 \times \{0,1\}^2$. A *Markov assignment* is a function $\beta : [0,1]^2 \times \{0,1\}^2 \to [0,1] \cup \{S\}$. *Markov maintenance rules* are functions $\varphi_J : [0,1]^2 \times \{0,1\}^2 \to \{0,1\}$ for $J \in \{A, B\}$ such that $\varphi_J(p_A, p_B, I_A, I_B) = 0$ whenever $I_J = 0$.

Imposing admissibility requirements directly on Markov strategies can be cumbersome.[4] A further source of difficulty in this framework is to determine how the monotonicity (irreversibility) requirements on maintenance rules carry over to restrictions on Markov maintenance rules. To

[3] In turn, this ensures that the optimal control problem of finding a payoff-maximising strategy is well-defined.
[4] See Fleming and Rishel (1975), Theorem 6.1.

get around these issues, I rely on the admissibility requirement already stated for strategies. Markov strategy $(\beta, \varphi_A, \varphi_B)$ will be said to be *admissible* if given any state (p_A, p_B, I_A, I_B) and initial beliefs $(p_A(0), p_B(0)) = (p_A, p_B)$, there exists an admissible strategy (α, ϕ_A, ϕ_B) such that for all t

$$\begin{aligned}\alpha(t) &= \beta(p_A(t), p_B(t), \phi_A(t), \phi_B(t)),\\ \phi_A(t) &= \varphi_A(p_A(t), p_B(t), \phi_A(t), \phi_B(t)),\\ \phi_B(t) &= \varphi_B(p_A(t), p_B(t), \phi_A(t), \phi_B(t)).\end{aligned}$$

Henceforth I will not explicitly restrict the experimenter to using admissible Markov strategies, or rather I will assume that the Markov strategies evoked yield well-defined solutions to the differential equations for the evolution of beliefs. However, I will verify that the optimal Markov strategies I derive, as well as the deviating strategies that support various proofs, are admissible.

A Markov strategy (β, φ) is *symmetric* if

$$\beta(p_B, p_A, I_B, I_A) = \begin{cases} 1 - \beta(p_A, p_B, I_A, I_B) & \text{if } \beta(p_A, p_B, I_A, I_B) \neq S,\\ S & \text{if } \beta(p_A, p_B, I_A, I_B) = S,\end{cases}$$

$$\varphi_J(p_B, p_A, I_B, I_A) = \varphi_J(p_B, p_A, I_B, I_A) \quad \text{for } J \in \{A, B\}.$$

To any optimal strategy (β^*, φ^*) will correspond (at least) an optimal symmetric strategy, and hence restricting to symmetric strategies is without loss of generality for the experimenter's payoffs. Given the restriction to symmetric strategies, it is without loss of generality to assume that $p_A \geq p_B$. Henceforth, arm A will always be the 'winning' arm, with arm B the 'losing' arm.

Let $W(\alpha, \phi; t)$ be the experimenter's payoff at time t to strategy (α, ϕ). If a success arrives at time $\tau < \min\{t_A, t_B\}$

$$W(\alpha, \phi; t) = e^{-r\tau} - \int_t^\tau e^{-rs}(\overline{k} + \underline{k})ds,$$

while if a success arrives at $\tau \in [\min\{t_A, t_B\}, \max\{t_A, t_B\}]$

$$W(\alpha, \phi; t) = e^{-r\tau} - \int_t^{\min\{t_A, t_B\}} e^{-rs}(\overline{k} + \underline{k})ds - \int_{\min\{t_A, t_B\}}^\tau e^{-rs}\overline{k}ds,$$

and finally if a success never arrives

$$W(\alpha, \phi; t) = -\int_t^{\min\{t_A, t_B\}} e^{-rs}(\overline{k} + \underline{k})ds - \int_{\min\{t_A, t_B\}}^{\max\{t_A, t_B\}} e^{-rs}\overline{k}ds.$$

The expected payoff to strategy (α, ϕ) given belief $(p_A(0), p_B(0))$ is

$$V(\alpha, \phi; t) \equiv \mathbf{E}W(\alpha, \phi; t),$$

where the expectation is taken over the distribution of stopping times τ determined by (α, ϕ) and $(p_A(s), p_B(s))_t$. Consider an admissible Markov strategy (β, φ), a state (p, I) and the corresponding strategy (α, ϕ). The expected payoff to (β, φ) in state (p, I) is given by

$$v(\beta, \varphi; p, I) \equiv V(\alpha, \phi; 0)$$

The objective of the experimenter is to find a payoff-maximising strategy. To this end, let $U(t) = \max_{(\alpha,\phi)} V(\alpha, \phi; t)$. Similarly, let $u(p, I) = \max_{(\beta,\varphi)} v(\beta, \varphi; p, I)$.

3 Preliminaries: Optimal Payoff Functions

The principal advantage of the continuous-time exponential bandit framework is its tractability, and the fact that simple expressions for optimal value functions exist for the standard two-armed bandit (optimal stopping) problem. In this section, I derive the expressions satisfied by the optimal payoff u that will support the characterisations of Sections 4 and 5. To simplify notation, let the number of beliefs listed in a state implicitly denote the set of maintained arms. Hence (p_A, p_B) can stand for state $(p_A, p_B, 1, 1)$, (p_A) for state $(p_A, p_B, 1, 0)$ given any p_B, and so on.

In any open region of the state space in which both arms are maintained, u must satisfy the following Bellman equation

$$u(p_A, p_B) = \max \Bigg\{ e^{-rdt} u(p_A, p_B),\ u_A(p_A),\ u_B(p_B),\ \max_{\beta \in [0,1]} \Big\{ [\beta p_A G + (1-\beta) p_B G - (\overline{k} + \underline{k})] dt + e^{-rdt} \mathbf{E}[u(p_A + dp_A, p_B + dp_B) | p_A, p_B] \Big\} \Bigg\}. \tag{1}$$

The first term in the brackets of (1) corresponds to the option of employing the safe arm in a time interval of length dt. The second and third terms correspond to the options of discarding arms A and B respectively, where u_J corresponds to the optimal payoff to the two-armed bandit problem with risky arm J and the safe arm. The final term corresponds to the payoffs from maintaining both arms and allocating the experimentation effort optimally. When a risky arm has been discarded, the experimenter faces a standard optimal stopping problem with the remaining risky arm, and payoff u_J solves

$$u_J(p_J) = \max \Big\{ e^{-rdt} u_J(p_J),\ [p_J G - \overline{k}] dt + e^{-rdt} \mathbf{E}[u_J(p_J + dp_J) | p_J] \Big\}.$$

The probability of a success in an interval of length dt is $p_J G dt$, and the payoff to a success is 1. The probability of failure is $1 - p_J G dt$. In case of failure, the payoff to the experimenter is

$u_J(p_J) + u'_J(p_J)dp_J$, which is equal to $u(p_J) - u'_J(p_J)p_J(1-p_J)Gdt$. By rewriting and cancelling dominated terms

$$ru_J(p_J) = \max\left\{0, p_JG - \overline{k} - u'_J(p_J)p_J(1-p_J)G - u_J(p_J)Gp_J\right\}.$$

Hence, in an open region of beliefs in which arm J is used, u_J satisfies the differential equation

$$u_J(p_J)(r + Gp_J) = p_JG - \overline{k} - u'_J(p_J)Gp_J(1-p_J), \tag{2}$$

which can be solved to yield

$$u_J(p_J) = \tilde{C}_J\left(\frac{1-p_J}{p_J}\right)^{\frac{r}{G}}(1-p_J) + p_J\frac{G-\overline{k}}{r+G} - (1-p_J)\frac{\overline{k}}{r}, \tag{3}$$

with the constant of integration $\tilde{C}_J = \left(\frac{\overline{k}}{G-\overline{k}}\right)^{\frac{r}{G}}\frac{G\overline{k}}{r(r+G)}$ and the stopping belief $p_J^* = \frac{\overline{k}}{G}$ determined by value-matching and smooth-pasting conditions

$$u_J(p_J^*) = 0, \text{ and}$$
$$u'_J(p_J^*) = 0.$$

The setup here is slightly different than in Keller et al. (2005), but the expression (3) admits the same interpretation. The term $p_J\frac{G-\overline{k}}{r+G} - (1-p_J)\frac{\overline{k}}{r}$ is the payoff to risky arm J in the absence of the ability to quit experimentation, while the term $\tilde{C}_J\left(\frac{1-p_J}{p_J}\right)^{\frac{r}{G}}(1-p_J)$ captures the option value of the safe arm S, the quitting option.

Note that the part of value function (1) in which both arms are maintained is linear in β. Hence, in an open region in which both arms are maintained, the optimal value is attained for $\beta \in \{0,1\}$, and (1) can be rewritten as

$$\begin{aligned} ru(p_A,p_B) = \max\Bigg\{ & p_AG - (\overline{k}+\underline{k}) - \frac{\partial u(p_A,p_B)}{\partial p_A}Gp_A(1-p_A) - u(p_A,p_B)Gp_A, \\ & p_BG - (\overline{k}+\underline{k}) - \frac{\partial u(p_A,p_B)}{\partial p_B}Gp_B(1-p_B) - u(p_A,p_B)Gp_B\Bigg\}. \end{aligned} \tag{4}$$

Contrary to (2), partial differential equation (4) does not have a simple solution, since such a solution to must include an optimal allocation rule.

I approach the solution to (4) by abstracting from allocations in order to reduce the two-dimensional problem (4) to suitably defined single-dimensional problems. First consider an open region of the state space in which arm A is used but both arms are maintained. Then since the optimal Markov strategy is admissible there exists $t' > 0$ and a parametrized path $(p_A(t), p_B)$

such that $U(t) = u(p_A(t), p_B)$ for $t \in (0, t')$. An argument similar to that establishing (2) shows that $U(t)$ satisfies

$$U(t)[r + p_A(t)G] - U'(t) = p_A(t)G - (\overline{k} + \underline{k}). \tag{5}$$

For path $(p_A(t), p_B)$, define $u_A(p_A(t); p_B) \equiv U(t)$. Then $U'(t) = -u'_A(p_A(t); p_B)Gp_A(t)(1 - p_A(t))$, and condition (5) can be rewritten, eliminating the dependence on time, as

$$u_A(p_A; p_B)[r + p_A G] + u'_A(p_A; p_B)Gp_A(1 - p_A) = p_A G - (\overline{k} + \underline{k}). \tag{6}$$

As for (2), (6) can be solved to yield

$$u_A(p_A; p_B) = C_A(p_B)\left(\frac{1 - p_A}{p_A}\right)^{\frac{r}{G}}(1 - p_A) + p_A\frac{G - (\overline{k} + \underline{k})}{r + G} - (1 - p_A)\frac{\overline{k} + \underline{k}}{r}. \tag{7}$$

In (7), the constant of integration will in general depend on p_B, since p_B may affect the payoffs when exiting the A-region. If the parametrised path $(p_A(t), p_B)$ exits the A-region in state (p^*_A, p_B), then p^*_A and $C_A(p_B)$ satisfy the value-matching and smooth-pasting properties

$$\begin{aligned} u_A(p^*_A; C_A(p_B)) &= u(p^*_A, p_B), \text{ and} \\ \frac{\partial}{\partial p_A}u_A(p^*_A; C_A(p_B)) &= \frac{\partial}{\partial p_A}u(p^*_A, p_B). \end{aligned}$$

In general, $u(p^*_A, p_B)$ is endogenous and depends on the experimentation policy once exit from the A-region occurs. If, for example, experimentation exits the A-region into the quitting region at p^*_A, then $u(p^*_A, p_B) = 0$ and $\frac{\partial}{\partial p_A}u(p^*_A, p_B) = 0$, which yields that $p^*_A = \frac{\overline{k}+\underline{k}}{G}$.

Equation (7) shows that when arm A is used in the optimal solution, payoffs evolve essentially as though the experimenter was facing a two-armed bandit problem with cost $\overline{k} + \underline{k}$ for the risky arm. However, the effect of p_B on $u(p_A, p_B)$ is captured by the constant of integration $C_A(p_B)$, which pins down the option value of arm B. It will be useful in the sequel to distinguish a payoff of the form (7) from the optimal payoff $u(p_A, p_B)$ in an A-assignment region. Given (p_A, p_B) and some function $C_A(p_B)$, define the righthand side of (7) as $v_A(p_A; C_A(p_B))$.

4 Optimal Experimentation Without Discarding Arms

In this section, I present the optimal experimentation policy for the benchmark case in which risky arms cannot be discarded individually. That is, the experimenter is restricted to Markov strategies with $\varphi(p, I) = (1, 1)$ for all states (p, I) such that $\beta(p, I) \neq S$. The experimenter can quit the experimentation by moving to the safe arm and discarding both risky arms. This problem is equivalent to the standard three-armed bandit problem with direct costs to experimentation $(\overline{k} + \underline{k})$.

4.1 Stay-with-the-Winner

The next lemma, which is also useful in later sections, deals with the allocation of trials between risky arms conditional on continuing the experimentation. It states that the experimenter should always use the arm with the highest belief, that is, 'stay-with-the-winner'. When beliefs (p_A, p_B) are such that $p_A > p_B$, this means using arm A. When beliefs are such that $p_A = p_B$, then both arms have the highest belief. In the discrete time version of the model, beliefs would jump down following a failure and this would (generically) ensure that there would always exist a 'best' arm, allowing the application of the 'stay-with-the-winner' rule. In continuous time, staying with the winner means sharing experimentation effort equally between both arms when beliefs are on the 45-degree line.[5]

Lemma 1. *Consider $(p_A(0), p_B(0))$ and the belief path $(p_A(t), p_B(t))_t$ under optimal experimentation. If $p_A(0) > p_B(0)$, then $\beta^*(p_A(t), p_B(t)) \in \{1, S\}$ for almost all $t \in [0, \hat{t})$, where $\hat{t}$ is such that $\hat{t} = \min\{\inf\{t : p_A(t) = p_B(t)\}, \infty\}$. If instead $p_A(0) = p_B(0)$, then $\beta^*(p_A(t), p_B(t)) \in \{\frac{1}{2}, S\}$ for almost all t.*

Lemma 1 mimics the Gittins index representation of the optimal experimentation policy, in which an arm's belief is taken to be the index. In fact, its proof is essentially a simplified version of the original 'interchange argument' in Gittins and Jones (1974) and Gittins (1979) that establishes the optimality of the Gittins' index for standard bandit problems.[6] By starting with an assignment in which an arm with a non-maximal Gittins index is chosen before the arm with the maximal index, the argument shows that interchanging the order in which both arms are pulled, keeping continuations following these periods of experimentation fixed, increases the experimenter's payoffs. A special feature of the exponential bandit problems is that the arm with the highest Gittins index is also the arm with the highest belief. Hence, the myopically optimal allocation is also dynamically optimal. Exponential bandits are continuous time versions of Bernoulli bandits in discrete time, for which a result that myopic play is optimal is shown by Berry and Fristedt (1985), who coined the term 'stay-with-the-winner'. Their result was generalised in Banks and Sundaram (1992a), who show that dynamically optimal play is myopic for a class of two-type symmetric bandits.

4.2 Shared Experimentation

Before completing the characterisation of optimal experimentation without discarding arms, I describe the experimenter's payoffs when $p_A = p_B$. According to Lemma 1, if there is no transition to the safe arm, then experimentation is shared and beliefs move down the 45-degree line. The key to obtaining an expression for optimal payoffs is that the partial differential equation (4)

[5] All proofs are in the Appendix.

[6] For a concise presentation of the original proof, see Frostig and Weiss (1999).

can be represented as a differential equation under the assumption that $p_A = p_B = p$ and that $\beta(p,p) = \frac{1}{2}$ for all beliefs p greater than some quiting belief p^*. In that case, the payoff u must satisfy $\frac{\partial}{\partial p_A}u = \frac{\partial}{\partial p_B}u$. Define $u_{AB}(p) \equiv u(p,p)$, then it follows that $u'_{AB}(p) = 2\frac{\partial}{\partial p_A}u(p,p)$ and u_{AB} solves

$$u_{AB}(r+pG) = pG - (\overline{k}+\underline{k}) - \frac{1}{2}Gp(1-p)u'_{AB}.$$

The differential equation (8) has solution

$$u_{AB}(p) = \tilde{C}_{AB}\left(\frac{1-p}{p}\right)^{\frac{2r}{G}}(1-p)^2 + p^2\frac{G-(\overline{k}+\underline{k})}{r+G} + 2p(1-p)\frac{\frac{G}{2}-(\overline{k}+\underline{k})}{r+\frac{G}{2}} - (1-p)^2\frac{\overline{k}+\underline{k}}{r}. \quad (8)$$

If optimal experimentation leads from shared experimentation to the safe arm at belief p^*, the constant of integration $\tilde{C}_{AB}$ and cutoff belief $p^* = \frac{\overline{k}+\underline{k}}{G}$ are determined by value-matching and smooth-pasting conditions.

$$u_{AB}(p^*) = 0, \text{ and}$$
$$u'_{AB}(p^*) = 0.$$

(8) has the same interpretation as (3), where $p^2\frac{G-(\overline{k}+\underline{k})}{r+G} + 2p(1-p)\frac{\frac{G}{2}-(\overline{k}+\underline{k})}{r+\frac{G}{2}} - (1-p)^2\frac{\overline{k}+\underline{k}}{r}$ is the payoff to shared experimentation with belief p for both arms in the absence of the ability to quit experimentation, while $\tilde{C}_{AB}\left(\frac{1-p}{p}\right)^{\frac{2r}{G}}(1-p)^2$ is the option value of the safe arm. Note also that since $p^* = \frac{\overline{k}+\underline{k}}{G}$, the optimal quitting value is the same in the single-arm problem with experimentation cost $\overline{k}+\underline{k}$ as in the two-arm problem with equal beliefs.

4.3 Optimal Policy

Lemma 1 and the previous discussion characterise the optimal experimentation policy without discarding arms. Lemma 1 ensures that off the 45-degree line, the 'best' arm is used, while if on the 45-degree line, shared experimentation must continue until all experimentation ceases. Hence, when off the 45-degree line, two alternatives are available: experiment with arm A until some quitting belief, or experiment with arm A until 45-degree line is reached and then share experimentation. The discussion around (7) and (8) establishes the boundaries between the quitting and experimentation regions.

Proposition 1. *When arms cannot be discarded, the following admissible Markov strategy is*

optimal

$$\beta^*(p_A,p_B)=\begin{cases}1 & \text{if } p_A>p_B \text{ and } p_A>\frac{\overline{k}+k}{G},\\ \frac{1}{2} & \text{if } p_A=p_B>\frac{\overline{k}+k}{G},\\ S & \text{if } p_A\leq\frac{\overline{k}+k}{G}.\end{cases}$$

$$\phi^*(p_A,p_B)=\begin{cases}(1,1) & \text{if } p_A>\frac{\overline{k}+k}{G}.\\ (0,0) & \text{if } p_A\leq\frac{\overline{k}+k}{G}.\end{cases}$$

Figure 1 illustrates belief paths consistent with the optimal experimentation policy when arms cannot be discarded. From belief (p_A,p_B) with $p_A>p_B>\frac{\overline{k}+k}{G}$, the use of arm A followed by shared experimentation until belief $(\frac{\overline{k}+k}{G},\frac{\overline{k}+k}{G})$ is optimal. From belief (p'_A,p'_B) with $p'_A>\frac{\overline{k}+k}{G}>p'_B$, only arm A will ever be used, until belief $(\frac{\overline{k}+k}{G},p_B)$.

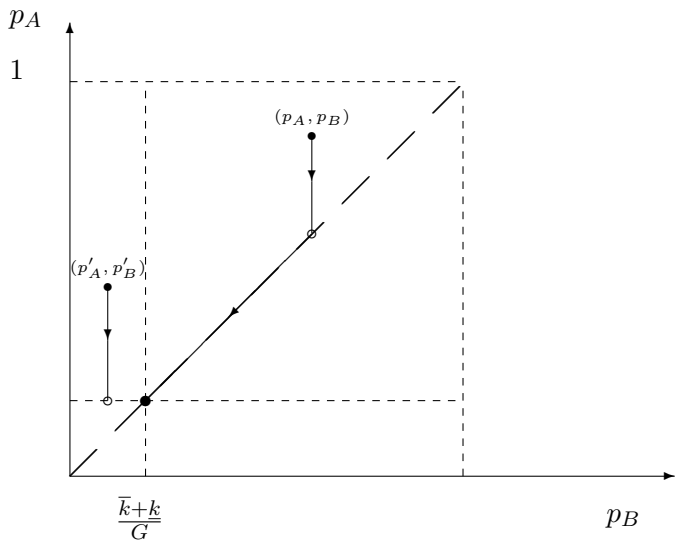

Figure 1: Optimal Experimentation Without Discarding of Arms.

5 Optimal Experimentation with Discarding of Arms

This section returns to the problem in which risky arms can be discarded and characterises optimal experimentation. When arms can be discarded, the experimenter can avoid accumulating maintenance costs on inactive arms, yet may hesitate to irreversibly abandon the option of

exploiting them. The experimenter has an incentive to bring the option value of inactive arms forward. How this incentive is resolved in the optimal experimentation policy is the subject of the following sections.

5.1 When to 'Go-with-the-Loser'

I first provide necessary conditions which deal with which arms can be discarded, and when. These establish that the patterns of optimal experimentation when arms can be discarded are very simple. The first result is quite natural and states that if an arm is ever discarded it is the losing arm B, and it is discarded as soon as the experimenter no longer intends to use it. That is, there is no value in 'stringing' B along without pulling it, only to discard is later.

Lemma 2. *Consider $(p_A(0), p_B(0))$ and the belief path $(p_A(t), p_B(t))_t$ under optimal experimentation.*

i. Suppose there exist $\hat{t}$ and $\epsilon > 0$ such that $\varphi^*(p_A(t), p_B(t)) \neq (1,1)$ for all $t \in [\hat{t}, \hat{t} + \epsilon)$ and $\beta^*(p_A(t), p_B(t)) \neq S$ for almost all $t \in [\hat{t}, \hat{t} + \epsilon)$. Then, without loss of generality, $\varphi^*(p_A(t), p_B(t)) = (1,0)$ for all $t \in [\hat{t}, \hat{t} + \epsilon)$.

ii. Suppose there exists $\hat{t}$ such that $\beta^*(p_A(t), p_B(t)) = 1$ for all $t \in [0, \hat{t})$ and that there exists $t' < \hat{t}$ such that $\varphi^*(p_A(t), p_B(t)) = (1,1)$ for almost all $t \in [0, t')$. Then $\varphi^*(p_A(t), p_B(t)) = (1,1)$ for all $t \in [t', \hat{t})$.

The proof of Lemma 2 is simple. If the better arm A were discarded before B, and B was used after having discarded A, then inverting the roles of arms A and B would increase the experimenter's payoff. If, on the other hand, arm B were maintained but never used again, were arm B to be discarded immediately, discoveries would occur with the same probability and maintenance costs would be avoided for a random time of positive expected length.

The next result derives the precise conditions under which the losing arm B can be used under optimal experimentation. It shows that whenever arm A is strictly better than arm B but is not used, then it must be that arm B is used exclusively for some time and then discarded.

Lemma 3. *Suppose that $p_A(0) > p_B(0)$ and consider the belief path $(p_A(t), p_B(t))_t$ under optimal experimentation. Suppose that there exists $\hat{t} > 0$ such that $\beta^*(p_A(t), p_B(t)) \neq 1$ for almost all $t \in [0, \hat{t})$ and $\varphi^*(p_A(t), p_B(t)) = (1,1)$ for all $t \in [0, \hat{t})$. Then there exists t^* such that $\hat{t} \leq t^*$, $\beta^*(p_A(t), p_B(t)) = 0$ for almost all $t \in [0, t^*)$, $\varphi^*(p_A(t), p_B(t)) = (1,1)$ for all $t \in [0, t^*)$ and $\varphi^*(p_A(t^*), p_B(t^*)) = (1,0)$.*

Lemma 3 answers the question at the head of the section regarding how the experimenter manages to bring the option value of inactive arms forward to avoid maintenance costs. Lemma 3 relies on Lemma 1, which shows that if both arms are maintained, optimal experimentation

requires that the better arm be used. Hence, any period of experimentation in which arm B is used and arm A is maintained must end by arm B being discarded. That is, whenever arm B is used, it is given a 'last chance' before it is culled. By Lemma 1, the losing arm B 'should' be the inactive arm. To avoid both maintaining arm B while using arm A and sacrificing its option value by discarding it, the experimenter must inefficiently use arm B for a short period while maintaining the better arm A. However, this is done with an eye to discarding arm B quickly in the absence of success. Hence to bring the option value of arm B forward, the experimenter faces a trade-off between inefficient assignment and the maintenance cost savings of experimenting with the better arm A in the absence of arm B.

5.2 Discarding Boundary

Together, Lemmas 1, 2 and 3 imply that given $p_A(0) > p_B(0)$, either *i*) arm B is discarded immediately, *ii*) arm B is used until it is discarded in favour of arm A, *iii*) arm A is used for some time, a switch to B occurs and B is used until it is discarded, or *iv*) arm A is used until the 45-degree line is reached, followed by shared experimentation. From shared experimentation, either one or both arms are discarded, or arm B is used exclusively until it is discarded.

This section focuses on the experimenter's discarding decision. To this end, suppose that $p_A > p_B$ and that (p_A, p_B) lies in an open region of beliefs in which arm B is used. Then, by Lemma 3, arm B will be used until it is discarded at some belief p_B^*, and as was shown in Section 3, the experimenter's payoff satisfies

$$u(p_A, p_B) = v_B(p_B; C_B(p_A)),$$

for some constant of integration $C_B(p_A)$. The experimenter's payoff at belief (p_A, p_B^*) once arm B has been discarded is given by $u_A(p_A)$ and is independent of p_B. Recall that $u_A(p_A)$ is the optimal payoff to a single risky arm with cost $\overline{k}$ and belief p_A. Hence value-matching and smooth-pasting conditions at the discarding belief (p_A, p_B^*) yield

$$v_B(p_B^*; C_B(p_A)) = u_A(p_A), \text{ and} \tag{9}$$

$$\begin{aligned} \frac{\partial}{\partial p_B} v_B(p_B^*; C_B(p_A)) &= \frac{\partial}{\partial p_B} u_A(p_A) \\ &= 0. \end{aligned} \tag{10}$$

Rearranging (10) yields

$$C_B(p_A) \left(\frac{1 - p_B^*}{p_B^*} \right)^{\frac{r}{G}} = \frac{G(r + \overline{k} + \underline{k})}{r(r+G)} \frac{p_B^* G}{p_B^* G + r}. \tag{11}$$

(11), along with (9), yield that p_B^* solves

$$u_A(p_A) = \frac{p_B^* G - (\overline{k} + \underline{k})}{p_B^* G + r} \tag{12}$$

Note that the right-hand side in (12) is the payoff to an arm that is known to be of type G but has a success rate $p_B^* G$ and associated experimentation cost $\overline{k} + \underline{k}$, since the flow payoff of such an arm is $p_B^* G - (\overline{k} + \underline{k})$, while the expected wait until a success is $p_B^* G$ and hence the effective discount is $p_B^* G + r$. Hence (12) states that at a cutoff belief (p_A, p_B^*) at which arm B is discarded, the experimenter is indifferent between its payoff to arm A in the absence of arm B and a riskless arm with a payoff equal to arm B's flow payoff at the belief p_B^* at which it is discarded. Note that (12) also implies that given arm A with belief p_A, there is a unique candidate cutoff state (p_A, p_B^*) at which arm B is discarded.

The preceding discussion does not say whether arm B is actually ever used when $p_A > p_B$, just when it should be discarded were it to be used. Define mapping $p_B^* : [0,1] \to [0,1]$ such that $p_B^*(p_A)$ is the solution to (12) if it exists, and is equal to p_A otherwise. Clearly, a necessary condition for arm B to be used before arm A is that there exists belief p_A such that $p_B^*(p_A) < p_A$. This occurs whenever, for fixed p_A, there exits p_B such that $u_A(p_A) < \frac{p_B^* G - (\overline{k} + \underline{k})}{p_B^* G + r}$. To this end, consider the mapping $p_B \mapsto \frac{p_B G - (\overline{k} + \underline{k})}{p_B G + r}$. It is straightforward to verify that this mapping is increasing and concave. Hence, for fixed p_A, the inequality $u_A(p_A) \leq \frac{p_B G - (\overline{k} + \underline{k})}{p_B G + r}$ is easiest to satisfy for $p_B = p_A$. Note that

$$\begin{aligned} \lim_{p_A \nearrow 1} \left[u_A(p_A) - \frac{p_A G - (\overline{k} + \underline{k})}{p_A G + r} \right] &= \frac{G - \overline{k}}{G + r} - \frac{G - (\overline{k} + \underline{k})}{G + r} \\ &> 0. \end{aligned} \tag{13}$$

That is, as the probability that arm A is of type G approaches 1, the payoff to a single risky arm with cost $\overline{k}$ approaches the payoff to an arm known to be of type G with cost $\overline{k}$ and success rate G. This dominates the payoff to an arm known to be of type G with cost $\overline{k} + \underline{k}$ and success rate G. Furthermore,

$$\begin{aligned} \lim_{p_A \searrow \frac{\overline{k} + \underline{k}}{G}} \left[u_A(p_A) - \frac{p_A G - (\overline{k} + \underline{k})}{p_A G + r} \right] &= v_A(\frac{\overline{k} + \underline{k}}{G}; \tilde{C}_A) \\ &> 0. \end{aligned} \tag{14}$$

That is, as p_A approaches the quitting belief $\frac{\overline{k} + \underline{k}}{G}$ (for a risky arm with cost $\overline{k} + \underline{k}$), the payoff to an arm known to be of type G with cost $\overline{k} + \underline{k}$ and success rate $p_A G$ approaches 0, while the payoff to a risky arm with cost $\overline{k}$ is strictly positive, since its own quitting belief is $\frac{\overline{k}}{G}$. This implies that when arms can be discarded, contrary to the results of Proposition 1, the experimenter will never reach the quitting belief $\frac{\overline{k} + \underline{k}}{G}$ with both arms maintained. (13) and (14) show that discarding arm B is always best either when arm A is almost sure to be good or when the belief in arm A approaches the quitting belief for the case in which arms cannot be discarded. Thus, if arm B is ever used, beliefs must lie 'between' $(\frac{\overline{k} + \underline{k}}{G}, \frac{\overline{k} + \underline{k}}{G})$ and $(1, 1)$.

The following lemma collects some of the previous discussion to give a necessary and sufficient condition for the existence of a set of beliefs with positive Lebesgue measure in which arm B is used before arm A.

Lemma 4. *One of the two following cases must obtain. Either*

i. $u_A(p_A) > \frac{p_A G - (\overline{k}+\underline{k})}{p_A G + r} > 0$ *for all* p_A*, and for almost all* (p_A, p_B)*,* $\varphi^*(p_A, p_B) = (1, 0)$*, or*

ii. there exist $\overline{p}_A > \underline{p}_A$ *such that* $u_A(p_A) \leq \frac{p_A G - (\overline{k}+\underline{k})}{p_A G + r}$ *if and only if* $p_A \in [\underline{p}_A, \overline{p}_A]$*. Then for almost all* (p_A, p_B) *with* $\varphi^*(p_A, p_B) = (1, 1)$*,* $\beta^*(p_A, p_B) = 0$ *only if* $p_A \in [\underline{p}_A, \overline{p}_A]$ *and* $p_B \in [p_B^*(p_A), p_A]$*. Furthermore, the set* $\{(p_A, p_B) : \varphi^*(p_A, p_B) = (1, 1)$ *and* $\beta^*(p_A, p_B) = 0\}$ *has positive Lebesgue measure.*

Figure 2 illustrates the discarding boundary when the condition of part *ii* of Lemma 4 obtains. Define

$$\mathcal{P}_M = \{(p_A, p_B) : p_A \geq p_B, p_B \geq p_B^*(p_A)\},$$

which is the set of beliefs which is inside the discarding boundary. That is, $\mathcal{P}_M$ is the maintenance region, the set of beliefs inside which arm B is never discarded. Further define

$$\mathcal{P}_D = \{(p_A, p_B) : p_A \geq p_B\} \setminus \mathcal{P}_M,$$

which is the set of beliefs outside the discarding boundary. This is the discarding region, in which arm B can possibly be maintained but never used. It is easily verified that the boundary separating $\mathcal{P}_M$ from $\mathcal{P}_D$ is concave. From state (p_A, p_B), if it is optimal to use arm B, then B must be used until $(p_A, p_B^*(p_A))$, after which B is discarded and A must be used until $p_A^* = \frac{\overline{k}}{G}$, the quitting belief with a single risky arm.

Part *ii* of Lemma 4 states that the set of beliefs for which arm B us used before arm A is nonempty whenever arm B is not always immediately discarded. The set $\mathcal{P}_B$ in Figure 2 illustrates the beliefs for which the argument in the proof applies, which are those beliefs close to $(\underline{p}_A, \underline{p}_A)$ and $(\overline{p}_A, \overline{p}_A)$. For any beliefs (p_A, p_B), the payoff to using arm A (or to shared experimentation) and maintaining B is at most the payoff to using an arm known to be of type G with success rate $p_A G$ and experimentation cost $\overline{k} + \underline{k}$. However, near $(\overline{p}_A, \overline{p}_A)$, discarding arm B yields a payoff close to the payoff to an arm known to be of type G with success rate $\overline{p}_A G$ but reduced experimentation cost $\overline{k}$. Hence near $(\overline{p}_A, \overline{p}_A)$, discarding arm B yields strictly higher payoffs than either using arm A (or shared experimentation). Yet, for beliefs inside $\mathcal{P}_M$, using arm B until the discarding boundary yields strictly higher payoffs than discarding it. By continuity, using arm B yields strictly higher payoffs than using arm A (or shared experimentation). The same argument applies around $(\underline{p}_A, \underline{p}_A)$. Intuitively, around $(\overline{p}_A, \overline{p}_A)$ and $(\underline{p}_A, \underline{p}_A)$, the experimenter has already decided that it no longer wishes to maintain both arms in the long term. Yet arm B may still be considered to be of some value, which the experimenter may want to exploit before discarding it.

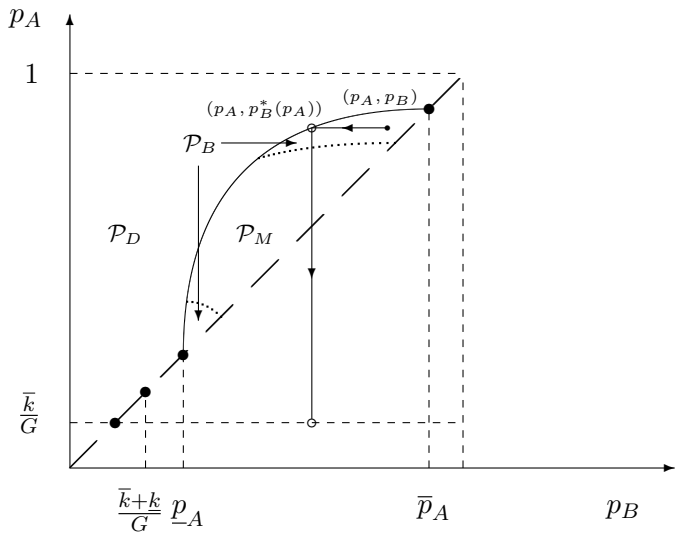

Figure 2: Discarding Boundary.

5.3 Optimal Policy

The previous sections derived the possible patterns of optimal experimentation when arms can be discarded. In this final section, I complete the characterisation of optimal experimentation by showing when it entails these various patterns. A useful starting point is to focus on beliefs on the 45-degree line. Lemma 3 states that if $p_A = p_B$, then either there is shared experimentation or arm B is used until it is discarded. By Lemma 4, shared experimentation always ends with the selection of arm B (in order to discard it). The next lemma addresses the question of how many exit points from shared experimentation to arm B there can be on the 45-degree line, and shows that only a single belief $(\underline{p}, \underline{p})$ can satisfy both the value-matching and smooth-pasting conditions associated to such an exit.

Lemma 5. *Suppose there exists $p' < p''$ such that $\beta^*(p,p) = \frac{1}{2}$ for almost all $p \in [p', p'']$. Then there exists $\underline{p}$ and $\overline{p}$ such that $\underline{p} \leq p'$, $p'' \leq \overline{p}$ and $\beta^*(p,p) = \frac{1}{2}$ for almost all $p \in P$ if and only if $P \subset [\underline{p}, \overline{p}]$.*

Lemma 5 implies that if the belief $\underline{p}$, derived explicitely in the Appendix is such that $\underline{p} \in (\underline{p}_A, \overline{p}_A)$ and if $v_{AB}(p; C_{AB}(\underline{p})) > v_B(p; C_B(p))$ for a set of beliefs (p,p) such that $p \in (\underline{p}, \underline{p} + \epsilon]$ for some $\epsilon > 0$, then there exists $\overline{p} \in (\underline{p}, \overline{p}_A)$ such that $\beta(p,p) = \frac{1}{2}$ for almost all $p \in (\underline{p}, \overline{p})$ and $\beta(p,p) = 0$ for almost all $p \in [\underline{p}_A, \underline{p}] \cup [\overline{p}, \overline{p}_A]$. That is, optimal experimentation calls for shared experimentation only for those beliefs (p,p) with $p \in (\underline{p}, \overline{p})$.

To complete the characterisation of the optimal experimentation policy, I will define two sets of beliefs, $\mathcal{P}_B \subset \mathcal{P}_M$ and $\mathcal{P}_A$ via a backwards induction argument. Intuitively, these sets will correspond the the regions of the state space in which arms A and B are used under optimal experimentation. In the following, assume that the conditions of Lemma 5 are met and that there exists a (unique) portion of the 45-degree line $(\underline{p}, \overline{p})$ for which shared experimentation is optimal. The arguments that follow apply in a straightforward way when this is not the case.

First, let

$$\mathcal{P}_B^1 = \left\{(p_A, p_B) \in \mathcal{P}_M : p_A \in [\underline{p}_A, \underline{p}] \cup [\overline{p}, \overline{p}_A]\right\}.$$

By Lemma 3, it must be that given $p \in [\underline{p}_A, \underline{p}] \cup [\overline{p}, \overline{p}_A]$, $\beta^*(p, p_B) = 0$ for almost all $p_B \in (p, p_B^*(p))$. That is, if arm B is used from the 45-degree line, it cannot be that a transition to arm A occurs before arm B is discarded.

Second, consider

$$\mathcal{P}_M \setminus \mathcal{P}_B^1 = \left\{(p_A, p_B) \in \mathcal{P}_M : p_A \in [\underline{p}, \overline{p}], p_B \in [p_B^*(\underline{p}), \overline{p}]\right\},$$

the set of beliefs in the maintenance region that have not been attributed to $\mathcal{P}_B^1$. By Lemma 3, from such beliefs, an optimal policy will either use arm B immediately until it is discarded, or use arm A either until beliefs reach the 45-degree or until a switch to arm B occurs. Define $v_A(p_A; C_A(p_B; p'_A))$ to be the payoff to the experimenter in state $(p_A, p_B) \in \mathcal{P}_M \setminus \mathcal{P}_B^1$ were it to pull arm A until belief $p'_A \in [\max\{p_B, \underline{p}\}, p_A)$, and then switch to arm B until discarding belief $p_B^*(p'_A)$. Hence the constant of integration $C_A(p_B; p'_A)$ depends on the belief p_B and on the switching belief p'_A, but not on p_A. Similarly, if $p_B > \underline{p}$, define $v_A(p_A; C_A^{45}(p_B))$ to be the payoff to the experimenter in state (p_A, p_B) were it to pull arm A until it reaches the 45-degree line, after which it shares experimentation until joint belief $\underline{p}$. If $p_B \leq \underline{p}$, then define $v_A(p_A; C_A^{45}(p_B)) = v_A(p_A; C_A(p_B; \underline{p}))$. Note that $v_A(p_A; C_A(p_B; p'_A)) \geq v_A(p_A; C_A(p_B; p''_A))$ if and only if $C_A(p_B; p'_A) \geq C_A(p_B; p''_A)$ and $v_A(p_A; C_A(p_B; p'_A)) \geq v_A(p_A; C_A^{45}(p_B))$ if and only if $C_A(p_B; p'_A) \geq C_A^{45}(p_B)$. Hence if $C_A(p_B; p_A) = \max_{\{p'_A \in [\max\{p_B, \underline{p}\}, p_A]\}} C_A(p_B; p'_A)$, then the experimenter has no incentive to use arm A. If, on the other hand, there exists a p'_A such that $C_A(p_B; p'_A) > C_A(p_B; p_A)$, then the experimenter gains by staying with arm A until belief p'_A. Let

$$\mathcal{P}_B^2 = \left\{(p_A, p_B) \in \mathcal{P}_M \setminus \mathcal{P}_B^1 : \max\left\{\max_{p' \in [\max\{p_B, \underline{p}\}, p_A]} C_A(p_B; p'), C_A^{45}(p_B))\right\} \leq C_A(p_B; p_A)\right\},$$

and let $\mathcal{P}_B = \mathcal{P}_B^1 \cup \mathcal{P}_B^2$. Finally, let $\mathcal{P}_A^1 = \mathcal{P}_M \setminus \mathcal{P}_B$. Hence, all the beliefs in $\mathcal{P}_M$ have been attributed either to $\mathcal{P}_B$ or to $\mathcal{P}_A^1$.

Third, consider the beliefs in $\mathcal{P}_D$, those outside the discarding boundary. By Lemma 2, it must be that $\varphi^*(p_A, p_B) = (1, 0)$ for all $(p_A, p_B) \in \mathcal{P}_D$ that are not in the set $\{(p_A, p_B) :$

$\beta^*(p_A, p_B(p_A)) = 1\}$. That is, if arm B is maintained, it must be that it will not be discarded once beliefs reach the discarding boundary. Let

$$\mathcal{Q} = \Big\{(p_A, p_B) \in \mathcal{P}_D : (p_A, p_B^*(p_A)) \in \mathcal{P}_A^1\Big\}.$$

That is, $\mathcal{Q}$ is the set of beliefs in the discarding region such that were A to be used and B maintained until the discarding bound, B would also be maintained when beliefs cross into $\mathcal{P}_M$. For any $(p_A, p_B) \in \mathcal{Q}$, define $p_A^{**}(p_B) = \sup\{p_A : (p_A, p_B) \in \mathcal{P}_A^1\}$. Furthermore, define

$$\mathcal{P}_A^2 = \Big\{(p_A, p_B) \in \mathcal{Q} : v_A(p_A; C_A(p_B; p_A^{**}(p_B)) > v_A(p_A; \tilde{C}_A)\Big\}.$$

Finally, let $\mathcal{P}_A = \mathcal{P}_A^1 \cup \mathcal{P}_A^2$. The next proposition collects these various results into an optimal Markov assignment.

Proposition 2. *When arms can be discarded, the following admissible Markov strategy is optimal.*

$$\beta^*(p_A, p_B) = \begin{cases} 0 & \text{if } (p_A, p_B) \in \mathcal{P}_B, \\ 1 & \text{if } (p_A, p_B) \in \mathcal{P}_A, \\ \frac{1}{2} & \text{if } p_A = p_B = p \text{ and } p \in (\underline{p}, \overline{p}), \\ S & \text{otherwise.} \end{cases}$$

$$\beta^*(p_A) = \begin{cases} 1 & \text{if } p_A \geq \frac{\overline{k}}{G} \\ S & \text{otherwise.} \end{cases} \qquad \beta^*(p_B) = \begin{cases} 0 & \text{if } p_B \geq \frac{\overline{k}}{G} \\ S & \text{otherwise.} \end{cases}$$

$$\varphi^*(p_A, p_B) = \begin{cases} (1,1) & \text{if } (p_A, p_B) \in \mathcal{P}_A \cup \mathcal{P}_B, \\ (1,0) & \text{otherwise.} \end{cases}$$

$$\varphi^*(p_A)) = \begin{cases} 1 & \text{if } p_A \geq \frac{\overline{k}}{G}, \\ 0 & \text{otherwise.} \end{cases} \qquad \varphi^*(p_B) = \begin{cases} 1 & \text{if } p_B \geq \frac{\overline{k}}{G}, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3 provided an illustration of the sets $\mathcal{P}_A$ and $\mathcal{P}_B$. The figure as drawn assumes that $\mathcal{P}_A$ is convex, which need not necessarily be the case. However, note that the boundary between sets $\mathcal{P}_A$ and $\mathcal{P}_B$ must always be downward-sloping, else this would violate Lemma 3.

When shared experimentation occurs on the 45-degree line, which by continuity implies that $\mathcal{P}_A$ is nonempty, the reversal of the 'go-with-the-winner' property exhibits a noteworthy non-monotonicity. For fixed p_B, arm B is used first in state (p_A, p_B) only for intermediate values of p_A. Suppose that arm A is believed to be of type G with very high probability, and hence to succeed quickly. In that case, it is best for the experimenter to discard project B immediately and exploit project A. Suppose now that p_A is substantially higher than p_B, yet arm B is still believed to be of type G with relatively high probability. The optimal experimentation rule in the absence

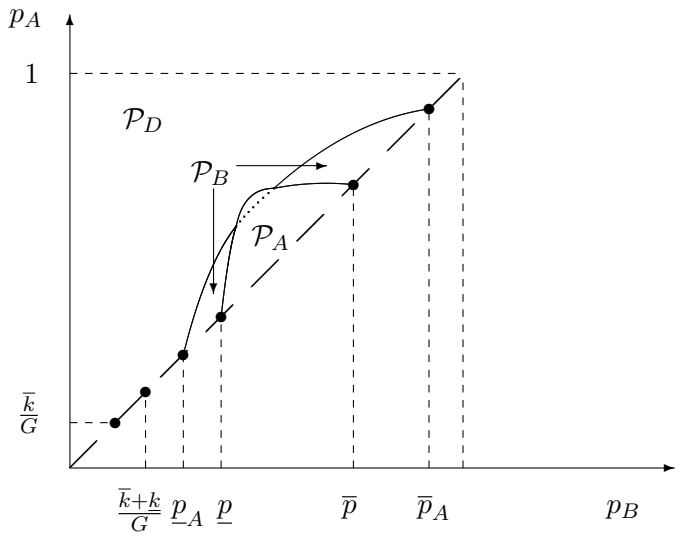

Figure 3: Optimal Experimentation Policy.

of maintenance costs would use arm A until its belief dropped to p_B, after which experimentation would be shared. However, since arm A is thought fairly likely to succeed, maintaining the option value of arm B is very costly, since this value can be realised only after arm A has failed for a long time. In this case, the optimal policy with maintenance costs starts by using arm B, discards it after a short period of failure, and moves to the clearly more promising arm A. If instead beliefs p_A and p_B are close to each other, it may still be optimal to develop projects as in the optimal policy without maintenance costs. In this case, discarding arm B immediately or giving it its 'last chance' is more costly, since the realisation of its option value is not so far away and arm A is not the clear-cut superior arm.

6 Conclusion

The standard approach to experimentation has been to assume that when currently occupied by other projects, keeping the option of researching various alternatives at later dates is costless. That keeping options open can involve maintenance costs is natural in many settings. This paper shows that such costs generate new trade-offs for experimenters by giving them incentives to manage the timing of the realisation of inactive alternatives' option values and have important implications for optimal experimentation policies.

While I have focused on the simple and tractable exponential bandit problem, it is not unrea-

sonable to expect that my main arguments extend to more general bandit settings. If they do, this would show that maintenance costs are indeed much more tractable than switching costs. Note that more generally, the arguments used in the paper are based on finite backwards induction, where the recursion is on the set of maintained arms. At each step of the recursion, the arguments rely on maintained arms' Gittins indices. This is made clear by the common structure of Lemmas 1 and 3 and the original 'interchange argument' of Gittins (1979) that establishes the optimality of index policies in standard bandit problems.

References

Banks, J. and R. Sundaram (1992a). A class of bandit problems yielding myopic optimal strategies. *Journal of Applied Probability 29*(3), 625–632.

Banks, J. and R. Sundaram (1992b). Denumerable-armed bandits. *Econometrica 60*(5), 1071–1096.

Banks, J. and R. Sundaram (1994). Switching costs and the Gittins index. *Econometrica 62*(3), 687–694.

Bergemann, D. and U. Hege (1998). Venture capital financing, moral hazard, and learning. *Journal of Banking & Finance 22*(6-8), 703–735.

Bergemann, D. and J. Välimäki (2001). Stationary multi-choice bandit problems. *Journal of Economic Dynamics and Control 25*(10), 1585–1594.

Bergemann, D. and J. Välimäki (2006). Bandit Problems. *Cowles Foundation Discussion Papers*.

Berry, D. and B. Fristedt (1985). *Bandit problems*. London: Chapman and Hall London.

Bolton, P. and C. Harris (1999). Strategic experimentation. *Econometrica*, 349–374.

Bonatti, A. and J. Hörner (2009). Collaborating.

Fleming, W. and R. Rishel (1975). *Deterministic and stochastic optimal control*. Berlin: Springer-Verlag.

Frostig, E. and G. Weiss (1999). Four proofs of Gittins' multiarmed bandit theorem. *Applied Probability Trust*, 1–20.

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

Gittins, J. and D. Jones (1974). A dynamic allocation index for the sequential design of experiments. *Progress in statistics 241*, 266.

Hörner, J. and L. Samuelson (2009). Incentives for Experimenting Agents. *Cowles Foundation Discussion Papers*.

Jun, T. (2004). A survey on the bandit problem with switching costs. *de Economist 152*(4), 513–541.

Keller, G. and S. Rady (2009). Strategic Experimentation with Poisson Bandits. *Discussion Papers in Economics*.

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

Klein, N. and S. Rady (2008). Negatively Correlated Bandits. *Discussion Papers in Economics*.

Strulovici, B. (2009). Learning While Voting: Determinants of Collective Experimentation. *Working paper*.

A Appendix

Proof of Lemma 1. Suppose that $p_A(0) > p_B(0)$, and consider the belief path under optimal experimentation $(p_A(t), p_B(t))_t$. The first step is to show that if there exists $\hat{t} > 0$ and $\hat{T} \subset [0, \hat{t})$ such that $\hat{T}$ has positive Lebesgue measure and $\beta^*(p_A(t), p_B(t)) \neq 1$ for all $t \in \hat{T}$, then $\beta^*(p_A(t), p_B(t)) = 0$ for almost all $t \in \hat{T}$. Suppose instead that $\beta^*(p_A(t), p_B(t)) = \alpha(t) \in (0, 1)$ for all $t \in \hat{T}$. Let $T_A = \int_0^{\hat{t}} \alpha(t)dt$. By assumption, $T_A \in (0, \hat{t})$.

Given allocation $\alpha(t)$ for $t \in [0, \hat{t})$, and initial beliefs $(p_A(0), p_B(0))$, solving the differential equation for the evolution of beliefs yields that

$$p_A(t) = \frac{1}{1 + \frac{1-p_A(0)}{p_A(0)} e^{H \int_0^t \alpha(s)ds}}, \text{ and}$$
$$p_B(t) = \frac{1}{1 + \frac{1-p_B(0)}{p_B(0)} e^{H \int_0^t (1-\alpha(s))ds}}.$$

Belief $p_A(t)$ depends only on the cumulative experimentation on arm A up to time t, $\int_0^t \alpha(s)ds$, and not on when this experimentation occurred within the interval $[0, t]$.

Consider an alternative admissible Markov assignment $\hat{\beta}$ such that

$$\hat{\beta}(p_A(t), p_B(t)) = \begin{cases} 1 & \text{for all } t \in (0, \hat{t} - T_A], \\ 0 & \text{for all } t \in (\hat{t} - T_A, \hat{t}), \end{cases}$$

with $\hat{\beta} = \beta^*$ otherwise. Then $(\hat{p}_A(\hat{t}), \hat{p}_B(\hat{t})) = (p_A(\hat{t}), p_B(\hat{t}))$, where $(\hat{p}_A(t), \hat{p}_B(t))_t$ is the belief path associated with $\hat{\beta}$. Hence, the payoffs following $\hat{t}$ are the same under both assignments.

That is, $v(\hat{\beta}, \varphi^*; p_A(\hat{t}), p_B(\hat{t})) = u(p_A(\hat{t}), p_B(\hat{t}))$. Furthermore, conditional on $(p_A(0), p_B(0))$, the probability that no success occurs until $\hat{t}$ is the same under β and $\hat{\beta}$.

Let τ_{β^*} (respectively $\tau_{\hat{\beta}}$) be the random arrival time of a success under assignment β^* (respectively $\hat{\beta}$) in time interval $[0, \hat{t}]$. Then $Pr[\tau_{\hat{\beta}} \leq t | p_A(0), p_B(0)] > Pr[\tau_{\beta^*} \leq t | p_A(0), p_B(0)]$ for all $t \in (0, \hat{t})$, that is, $\tau_{\hat{\beta}}$ is higher than τ_{β^*} in the sense of first order stochastic dominance. By discounting, the experimenter's payoff is decreasing in the arrival time of a success, and hence $\hat{\beta}$ yields a strictly higher expected payoff than β^* in $[0, \hat{t}]$, or

$$\int_0^{\hat{t}} [1 - (\overline{k} + \underline{k})] e^{-r\tau_{\hat{\beta}}} \hat{\mu}(d\tau_{\hat{\beta}}) > \int_0^{\hat{t}} [1 - (\overline{k} + \underline{k})] e^{-r\tau_{\beta^*}} \mu^*(d\tau_{\beta^*}),$$

where $\hat{\mu}$ and μ^* are the distributions of $\tau_{\hat{\beta}}$ and τ_{β^*}, respectively. Hence,

$$\begin{aligned} v(\hat{\beta}; p_A(\hat{t}), p_B(\hat{t})) &= \int_0^{\hat{t}} [1 - (\overline{k} + \underline{k})] e^{-r\tau_{\hat{\beta}}} \hat{\mu}(d\tau_{\hat{\beta}}) + Pr[\tau_{\hat{\beta}} > \hat{t} | p_A(0), p_B(0)) u(p_A(\hat{t}), p_B(\hat{t})] \\ &> \int_0^{\hat{t}} [1 - (\overline{k} + \underline{k})] e^{-r\tau_{\beta^*}} \mu^*(d\tau_{\beta^*}) + Pr[\tau_{\beta^*} > \hat{t} | p_A(0), p_B(0)) u(p_A(\hat{t}), p_B(\hat{t})] \\ &= u(p_A(0), p_B(0)), \end{aligned}$$

a contradiction. Hence it must be that $\alpha(t) = 0$ for almost all $t \in [0, \hat{t}]$.

That is, the previous argument shows that if arm A is not used, it must be that arm B is used exclusively. Since in that case $p_A(t) > p_B(t)$ for all $t > 0$, the previous argument also ensures that arm B is used until experimentation ceases, which must occur at time t^* such that $p_B(t^*) = \frac{\overline{k}+\underline{k}}{G}$. By mimicking this strategy with arm A instead of B, that is, using arm A until belief $p_A^* = \frac{\overline{k}+\underline{k}}{G}$ and then moving permanently to S, the experimenter's payoff at time 0 would be higher. That is, consider alternative strategy $\hat{\beta}$ such that

$$\hat{\beta}(p_A, p_B) = \begin{cases} 1 & \text{if } p_A > \frac{\overline{k}+\underline{k}}{G} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} v(\beta; p_A(0), p_B(0)) &= v_A(p_A(0); C_A) \\ &\geq v_B(p_B(0, C_A)) \\ &= u(p_A(0), p_B(0)), \end{aligned}$$

a contradiction. C_A is the constant of integration for the optimal stopping problem with a single risky arm and direct cost $\overline{k} + \underline{k}$. Hence, it must be that $\alpha(t) = 1$ for all t such that $p_A(t) > p_B(t)$.

The same argument can be applied if $p_A(0) = p_B(0)$ to show that experimentation is shared until it ceases, i.e., $\beta^*(p_A(t), p_B(t)) = \frac{1}{2}$ for all t such that $\varphi(p_A(t), p_B(t))$.

□

Proof of Proposition 1. Most of the result was proved in the text. All that remains is to show that the assignment β^* is admissible. Clearly, given any t^*, t^{**} and t' such that $0 \leq t^* \leq t^{**}$ and $t' \geq 0$, any strategy (α, ϕ) of the form

$$\alpha(t) = \begin{cases} 1 & \text{if } t < t^*, \\ \frac{1}{2} & \text{if } t \in [t^*, t^{**}), \\ S & \text{if } t \geq t^{**}, \end{cases}$$
$$\phi(t) = \begin{cases} (1,1) & \text{if } t < t', \\ (0,0) & \text{if } t \geq t', \end{cases} \tag{15}$$

is admissible. Furthermore, given any (p_A, p_B), there exist t^*, t^{**} and t' such that $0 \leq t^* \leq t^{**}$ and $t' \geq 0$ such that a strategy (α, ϕ) defined as in (15) is such that

$$\begin{aligned} \alpha(t) &= \beta^*(p_A(t), p_B(t), \phi_A(t), \phi_B(t)), \\ \phi_A(t) &= \varphi_A^*(p_A(t), p_B(t), \phi_A(t), \phi_B(t)), \\ \phi_B(t) &= \varphi_B^*(p_A(t), p_B(t), \phi_A(t), \phi_B(t)), \end{aligned}$$

and hence Markov strategy (β^*, φ^*) is admissible.

□

Proof of Lemma 2. For part i, suppose there exists $\hat{t}$ and $\epsilon > 0$ such that $\varphi^*(p_A(t), p_B(t)) \neq (1,1)$ and $\beta^*(p_A(t), p_B(t)) \neq S$ for almost all $t \in [\hat{t}, \hat{t}+\epsilon)$. Then one arm is discarded on the equilibrium path. Let $t^* = \inf\{t < \hat{t} : \varphi^*(p_A(t), p_B(t)) \neq (1,1)\}$. If $\varphi^*(p_A(t^*), p_B(t^*)) = (0,1)$, then since $\beta^*(p_A(t^*), p_B(t^*)) \neq S$ for almost all $t \in [\hat{t}, \hat{t}+\epsilon)$, it must be that $\beta^*(p_A(t^*), p_B(t^*)) = 0$ for almost all $t \in [\hat{t}, \hat{t}+\epsilon)$. Consider a Markov strategy (β', φ') such that

$$\begin{aligned} \varphi'(p_A, p_B) &= (1,0) && \text{for all } (p_A, p_B) \text{ such that } \varphi^*(p_A, p_B) = (0,1), \\ \beta'(p_A(t), p_B(t)) &= 1 && \text{for all } t > t^* \text{ for which } \beta^*(p_A(t), p_B(t)) = 0, \end{aligned}$$

with $(\beta', \varphi') = (\beta^*, \varphi^*)$ otherwise. Under (β', φ'), $p_A'(t) \geq p_B(t)$ for all $t > t^*$ by the assumption of symmetric strategies, and hence for all $t > t^*$ such that $\beta^*(p_A(t), p_B(t)) = 0$,

$$\begin{aligned} v(\beta', \varphi'; p_A(t), p_B(t)) &= v_A(p_A(t); \tilde{C}_A) \\ &\geq v_B(p_B(t); \tilde{C}_A) \\ &= v_B(p_B(t); \tilde{C}_B) \\ &= u(p_A(t), p_B(t)). \end{aligned}$$

If the inequality is strict, this yields the required contradiction, while if it holds with equality, it is without loss of generality to discard arm B instead of arm A.

For part *ii*, suppose that there exists $\hat{t}$ such that $\beta^*(p_A(t), p_B(t)) = 1$ for almost all $t \in [0, \hat{t})$ and that there exists $t' < \hat{t}$ such that $\varphi^*(p_A(t), p_B(t)) = (1,1)$ for almost all $t \in [0, t')$, but that $\varphi^*(p_A(t''), p_B(t'')) \neq (1,1)$ for some $t'' \in (t', \hat{t})$. By part *i*, $\varphi^*(p_A(t''), p_B(t'')) = (1,0)$. Consider Markov strategy (β', φ') such that $\varphi'(p_A(0), p_B(0)) = (1,0)$, with $(\beta', \varphi') = (\beta^*, \varphi^*)$ otherwise. Then we can write

$$v(\beta', \varphi'; p_A(0), p_B(0)) = v_A(p_A(0); C'_A) + p_A(0)\frac{\underline{k}}{r+G} + (1 - p_A(0))\frac{\underline{k}}{r},$$

and

$$u(p_A(t), p_B(t)) = v_A(p_A(0); C_A),$$

where the constants of integration C_A and C'_A are determined at beliefs $(p_A(t''), p_B(t''))$ at which

$$v_A(p_A(t''); C_A) = v_A(p_A(t''); C'_A) = u(p_A(t''), p_B(t'')).$$

Hence

$$C'_A = C_A - \frac{p_A(t'')\frac{\underline{k}}{r+G} + (1 - p_A(t''))\frac{\underline{k}}{r}}{\left(\frac{1-p_A(t'')}{p_A(t'')}\right)^{\frac{r}{G}}(1 - p_A(t''))},$$

and

$$\begin{aligned} v(\beta', \varphi'; p_A(0), p_B(0)) &= v_A(p_A(0); C_A) \\ &\quad - \frac{\left(\frac{1-p_A(0)}{p_A(0)}\right)^{\frac{r}{G}}(1 - p_A(0))}{\left(\frac{1-p_A(t'')}{p_A(t'')}\right)^{\frac{r}{G}}(1 - p_A(t''))}\left[p_A(t'')\frac{\underline{k}}{r+G} + (1 - p_A(t''))\frac{\underline{k}}{r}\right] \\ &\quad + p_A(0)\frac{\underline{k}}{r+G} + (1 - p_A(0))\frac{\underline{k}}{r} \\ &> v_A(p_A(0); C_A). \end{aligned}$$

The inequality follows since $p_A(t'') < p_A(0)$.

□

Proof of Lemma 3. Lemma 1 and part *i* of Lemma 2 imply that if there exists $\hat{t} > 0$ such that $\beta^*(p_A(t), p_B(t)) \neq 1$ for almost all $t \in [0, \hat{t})$ and $\varphi^*(p_A(t), p_B(t)) = (1,1)$ for all $t \in [0, \hat{t})$, then it must be that $\beta^*(p_A(t), p_B(t)) = 0$ for almost all $t \in [0, \hat{t})$ and that if there exists $t^* \geq \hat{t}$ such that $\beta^*(p_A(t^*), p_B(t^*)) > 0$, then by part *i* of Lemma 2 it must be that $\varphi^*(p_A(t^*), p_B(t^*)) \in \{(1,0), (0,0)\}$. That is, if arm A is not pulled it must be that arm B is, and the experimenter cannot go back to arm A without discarding arm B. Since the experimenter must eventually discard B if p_B gets close to 0, it only remains to be shown that $\varphi^*(p_A(t^*), p_B(t^*)) = (1,0)$, that is, that the experimenter will discard B in favour of A at t^*. This follows by part *ii* of Lemma 2.

□

Proof of Lemma 4. First, $u_A(p_A)$ is increasing and convex in p_A. Also, since the mapping $p_A \mapsto \frac{p_A G-(\overline{k}+\underline{k})}{p_A G+r}$ is increasing and concave, then by (13) and (14) either $u_A(p_A) > \frac{p_A G-(\overline{k}+\underline{k})}{p_A G+r} > 0$ for all p_A or there exist $\overline{p}_A > \underline{p}_A$ such that $u_A(p_A) \leq \frac{p_A G-(\overline{k}+\underline{k})}{p_A G+r}$ if and only if $p_A \in [\underline{p}_A, \overline{p}_A]$, where $\underline{p}_A$ and $\overline{p}_A$ are the only two solutions to $u_A(p_A) = \frac{p_A G-(\overline{k}+\underline{k})}{p_A G+r}$.

Now suppose that the conditions of part *ii* obtain. A first claim is that at $(\overline{p}_A, \overline{p}_A)$, discarding arm B is strictly preferred to shared experimentation. By Lemma 3, if B is not discarded then $\beta^*(\overline{p}_A, \overline{p}_A) = \frac{1}{2}$ and the beliefs go down the 45-degree line until some belief (p^*, p^*), and hence the experimenter's payoffs satisfy $u(p_A, p_B) = v_{AB}(\overline{p}_A; C_{AB}(p^*))$. v_{AB} itself satisfies

$$\begin{aligned} v_{AB}(\overline{p}_A; C_{AB}(p^*)) &= \frac{\overline{p}_A G - (\overline{k}+\underline{k})}{(r+\overline{p}_A G)} - \frac{G\overline{p}_A(1-\overline{p}_A)}{2(r+\overline{p}_A G)} v'_{AB}(\overline{p}_A; C_{AB}(p^*)) \\ &< \frac{\overline{p}_A G - (\overline{k}+\underline{k})}{(r+\overline{p}_A G)} \\ &= v_A(\overline{p}_A; \tilde{C}_A). \end{aligned}$$

Hence, by continuity, for states (p, p) with $p < \overline{p}_A$ sufficiently close to $\overline{p}_A$, discarding B is strictly preferred to shared experimentation. A very similar argument shows that discarding B is strictly preferred to using arm A for an open set of states of positive Lebesgue measure (p_A, p_B) with $p_A > p_B$ sufficiently close to $(\overline{p}_A, \overline{p}_A)$. However, within the discarding boundary using arm B (until the boundary) is preferred to discarding it and hence there exists an open region of positive Lebesgue measure around $(\overline{p}_A, \overline{p}_A)$ in which using arm B is optimal. A very similar argument demonstrates the same result for a region around $(\underline{p}_A, \underline{p}_A)$.

□

Proof of Lemma 5. By Lemma 3, once the experimenter quits shared experimentation, arm B is used, then discarded and replaced with arm A. Also, by Lemma 4, there exists a belief $\hat{p} > \underline{p}_A$ such that $\beta^*(p, p) = 0$ for almost all $p \in [\underline{p}_A, \hat{p}]$. Suppose there exists $p' > p''$ such that $\beta^*(p, p) = \frac{1}{2}$ for almost all $p \in [p', p'']$, and that the experimenter switches from shared experimentation to arm B at belief $p^* < p''$. Hence the smooth-pasting condition at belief p^* is

$$\begin{aligned} \frac{\partial}{\partial p_B} u(p^*, p^*) &= \frac{1}{2} \cdot \frac{\partial}{\partial p} v_{AB}(p^*; C_{AB}(p^*)) \\ &= \frac{\partial}{\partial p_B} v_B(p^*, C_B(p^*)), \end{aligned}$$

which, with manipulations, yields that

$$C_{AB} = \frac{C_B(p)}{\left(\frac{1-p}{p}\right)^{\frac{r}{G}}(1-p)} + \frac{p\left[\frac{G-(\overline{k}+\underline{k})}{r+G} - \frac{\frac{G}{2}-(\overline{k}+\underline{k})}{r+\frac{G}{2}}\right] + (1-p)\left[\frac{\frac{G}{2}-(\overline{k}+\underline{k})}{r+\frac{G}{2}} - \frac{\overline{k}+\underline{k}}{r}\right] - \frac{G(r+\overline{k}+\underline{k})}{r(r+G)}}{\left(\frac{1-p}{p}\right)^{\frac{2r}{G}+1}\frac{Gp+r}{G}}.$$

Meanwhile, the value matching condition is

$$v_{AB}(p^*; C_{AB}(p^*)) = v_B(p^*; C_B(p^*)),$$

which, with manipulations, yields that

$$C_{AB} = \frac{C_B(p)}{\left(\frac{1-p}{p}\right)^{\frac{r}{G}}(1-p)} - \frac{\frac{G^2}{2}(r+\overline{k}+\underline{k})}{\left(\frac{1-p}{p}\right)^{\frac{2r}{G}+1} r(r+G)(r+\frac{G}{2})}.$$

Together, these yield that

$$p^* = \frac{2(\overline{k}+\underline{k})(r+G)(r+\frac{G}{2})}{2(\overline{k}+\underline{k})(r+G)(r+\frac{G}{2}) + \frac{G^2}{2}(\overline{k}+\underline{k}+r)} \tag{16}$$

Clearly, $p^* \in [0,1]$ is unique. Define $\underline{p}$ to be the unique solution to (16).

□

Proof of Proposition 2. Most of the result was proved in the text. All that remains is to show that the assignment (β^*, φ^*) is admissible, which follows from an argument very similar to that for Proposition 1. □