ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Klein, Nicolas; Rady, Sven

Working Paper

Negatively Correlated Bandits

SFB/TR 15 Discussion Paper, No. 243

Provided in Cooperation with:
Free University of Berlin, Humboldt University of Berlin, University of Bonn, University of Mannheim, University of Munich, Collaborative Research Center Transregio 15: Governance and the Efficiency of Economic Systems

Suggested Citation: Klein, Nicolas; Rady, Sven (2008) : Negatively Correlated Bandits, SFB/TR 15 Discussion Paper, No. 243, Sonderforschungsbereich/Transregio 15 - Governance and the Efficiency of Economic Systems (GESY), München, https://doi.org/10.5282/ubm/epub.13309

This Version is available at:
https://hdl.handle.net/10419/93959

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

GOVERNANCE AND THE EFFICIENCY OF ECONOMIC SYSTEMS
GESY

Discussion Paper No. 243

Negatively Correlated Bandits

Nicolas Klein*
Sven Rady**

* Munich Graduate School of Economics, Kaulbachstr. 45, D-80539 Munich, Germany;
email: kleinnic@yahoo.com.

**Department of Economics, University of Munich, Kaulbachstr. 45, D-80539 Munich, Germany;
email: sven.rady@lrz.uni-muenchen.de

August 2008

Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 is gratefully acknowledged.

Negatively Correlated Bandits*

Nicolas Klein† Sven Rady‡

This version: August 1, 2008

Abstract

We analyze a two-player game of strategic experimentation with two-armed bandits. Each player has to decide in continuous time whether to use a safe arm with a known payoff or a risky arm whose likelihood of delivering payoffs is initially unknown. The quality of the risky arms is perfectly *negatively* correlated between players. In marked contrast to the case where both risky arms are of the same type, we find that learning will be complete in any Markov perfect equilibrium if the stakes exceed a certain threshold, and that all equilibria are in cutoff strategies. For low stakes, the equilibrium is unique, symmetric, and coincides with the planner's solution. For high stakes, the equilibrium is unique, symmetric, and tantamount to myopic behavior. For intermediate stakes, there is a continuum of equilibria.

Keywords: Strategic Experimentation, Two-Armed Bandit, Exponential Distribution, Poisson Process, Bayesian Learning, Markov Perfect Equilibrium.

JEL Classification Numbers: C73, D83, O32.

*We are grateful to Philippe Aghion, Patrick Bolton, Kalyan Chatterjee, Martin Cripps, Matthias Dewatripont, Jan Eeckhout, Florian Englmair, Eduardo Faingold, Philipp Kircher, George Mailath, Timofiy Mylovanov, Stephen Ryan, Klaus Schmidt, Larry Samuelson, as well as seminar participants at Bonn, Munich, UPenn, Yale, HEC Paris, the 2007 SFB/TR 15 Summer School in Bronnbach, the 2007 SFB/TR 15 Workshop for Young Researchers in Bonn, the 2008 SFB/TR 15 Meeting in Gummersbach, the 2008 North American Summer Meetings of the Econometric Society, the 2008 Meeting of the Society for Economic Dynamics and the European Summer Symposium in Economic Theory (ESSET) 2008 for helpful comments and suggestions. We thank the Department of Economics at the University of Bonn, the Business and Public Policy Group at the Wharton School, the UPenn Economics Department and the Studienzentrum Gerzensee for their hospitality. Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 and GRK 801 is gratefully acknowledged.

†Munich Graduate School of Economics, Kaulbachstr. 45, D-80539 Munich, Germany; email: kleinnic@yahoo.com.

‡Department of Economics, University of Munich, Kaulbachstr. 45, D-80539 Munich, Germany; email: sven.rady@lrz.uni-muenchen.de.

1 Introduction

Starting with Rothschild (1974), two-armed bandit models have been used extensively in economics to formalize the trade-off between experimentation and exploitation in dynamic decision problems with learning; see Bergemann and Välimäki (2008) for a survey of this literature. The use of the two-armed bandit framework as a canonical model of strategic experimentation in teams is more recent: Bolton and Harris (1999, 2000) analyze the case of Brownian motion bandits, while Keller, Rady and Cripps (2005) and Keller and Rady (2007) analyze bandits where payoffs are governed by Poisson processes. These papers assume perfect *positive* correlation across bandits; what is good news to any given player is assumed to be good news for everybody else.

There are many situations, however, where one man's boon is the other one's bane. Think of a suit at law, for instance: whatever is good news for one party is bad news for the other. Or consider two firms pursuing research and development under different, incompatible working hypotheses. One pharmaceutical company, for example, might base its drug development strategy on the hypothesis that the cause of a particular disease is a virus, while the other might see a bacterium as the cause. An appropriate model of strategic experimentation in such a situation must assume perfect *negative* correlation across bandits. This we propose to do in the present paper.[1]

We consider two players, each facing a continuous-time exponential bandit as in Keller, Rady and Cripps (2005). One arm is safe, generating a known payoff per unit of time. The other arm is risky, and can be good or bad. If it is good, it generates lump-sum payoffs after exponentially distributed random times; if it is bad, it never generates any payoff. A good risky arm dominates the safe one in terms of average payoff per unit of time, whereas the safe arm dominates a bad risky one. At the start of the game, the players do not know the type of their risky arm, but it is common knowledge that exactly one risky arm is good. Each player's actions and payoffs are perfectly observable to the other player. Starting from a common prior, the players' posterior beliefs thus agree at all times.

The dynamics of these beliefs are easy to describe. If both players play safe, no new information is generated and beliefs stay unchanged. If only one player plays risky and he has no success, the posterior probability that his risky arm is the good one falls gradually over time; if he obtains a lump-sum payoff, all uncertainty is resolved and beliefs become

[1] There is a decision-theoretic literature on correlated bandits which analyzes correlation across different arms of a bandit operated by a single agent; see e.g. Camargo (2007) for a recent contribution to this literature, or Pastorino (2005) for economic applications. Our focus here is quite different, though, in that we are concerned with correlation between different bandits operated by two players who interact strategically.

degenerate at the true state of the world. If both players play risky, finally, and there is no success on either arm, this is again uninformative about the state of the world, so beliefs are constant up to the random time when the first success occurs. It is important to note that a success on one player's risky arm is always bad news for the other player, while lack of success gradually makes the other player more optimistic.

We restrict players to stationary Markov strategies with the common posterior belief as the state variable. As is well known, this restriction is without loss of generality in the decision problem of a single agent experimenting in isolation, whose optimal policy is given by a cutoff strategy, i.e. has him play risky at beliefs more optimistic than some threshold, and safe otherwise. As the same structure prevails in the optimization problem of a utilitarian planner who maximizes the average of the two players' expected discounted payoffs, the Markov restriction is without loss of generality there as well. In the non-cooperative experimentation game, the Markov restriction rules out history-dependent behavior that is familiar from the analysis of infinitely repeated games in discrete time, yet technically quite intricate to formalize in continuous time (Simon and Stinchcombe 1989, Bergin 1992, Bergin and McLeod 1993). Imposing Markov perfection allows us to focus on the experimentation tradeoff that the players face and makes our results directly comparable to those in the previous literature on strategic experimentation in bandits.

After solving the planner's optimization problem, we characterize the Markov perfect equilibria of the experimentation game. We find that all Markov perfect equilibria are in cutoff strategies. This is in stark contrast to Keller, Rady and Cripps (2005), who find that there is *no* such equilibrium when all risky arms are of the same type.

On account of the symmetry of the situation, it is not surprising that there always exists a symmetric equilibrium, where both players use the same cutoff. This symmetric equilibrium is unique. What is more, based upon a characterization of the beliefs at which best responses can change, we are able to determine the parameter values for which there is no other equilibrium besides the symmetric one. This uniqueness result is again in sharp contrast to the multiplicity of equilibria in Keller, Rady and Cripps (2005).

When stakes (as measured by the payoff advantage of a good risky arm over a safe one) are sufficiently low, players' respective single-agent cutoffs are such that a higher than 50–50 chance of having the good risky arm is required for a player to play risky. In this case, the single-agent optimal strategies let at most one player play risky at any given belief and, in the absence of a success, make this player switch to the safe arm at a point where the other player's threshold for playing risky is not yet reached. This means that neither player ever benefits from any experimentation by the other, and so the single-agent optimal strategies are mutually best responses. As experimentation on one bandit can only make

players *incrementally* more optimistic about the other bandit, the same logic applies to the planner's problem, so both players using their single-agent optimal strategies is efficient. Finally, since any equilibrium must give each player at least his single-agent optimal payoff and any payoff higher than that would be incompatible with the upper bound given by the planner's solution, the Markov perfect equilibrium where both players use their single-agent optimal cutoffs is unique.

When stakes are sufficiently high, a lower than 50–50 chance is sufficient for a myopic player, i.e. one who is merely interested in maximizing *current* payoffs, to play risky. In this case, equilibrium is again unique, but inefficient, with both players applying the myopic cutoff strategy. With the stakes high, players are so eager to play risky that there exists a range of beliefs where both are experimenting. As long as no lump-sum arrives, no new information is then made available, and the players are effectively freezing the problem in its current state. This, however, they are only willing to do if the current state is attractive from a myopic perspective.

If the stakes are intermediate in size, there is a continuum of equilibria, each of them characterized by a single belief at which both players change their actions. As the stakes gradually increase and we move from the low to the intermediate case, at first, given *any* initial belief, there still exists an equilibrium that achieves the efficient outcome. As stakes increase further, there then appears a range of initial beliefs for which no equilibrium achieves efficiency. As we move from high stakes down to intermediate stakes, there at first always exists an equilibrium that involves *one* player behaving myopically. To achieve this, the other player has to bear the entire load of experimentation by himself when the uncertainty is greatest. As stakes gradually grow lower, however, the other player will at some point no longer be willing to bear this burden, and the equilibrium disappears.

Given perfect observability of actions and payoffs, any information that a player garners via experimentation is a public good.[2] In contrast to the case where all risky arms are of the same type, however, the resulting free-rider problem does not cause learning to cease prematurely. In fact, we find complete learning in equilibrium (meaning that beliefs converge to the truth almost surely) for intermediate and high stakes, which is precisely the parameter range where the planner's solution calls for complete learning. The intuition is quite straightforward. If players hold common beliefs and there is perfect negative correlation between the types of the risky arms, it can never be the case that both players are simultaneously very

[2]Murto and Välimäki (2006) and Rosenberg, Solan and Vieille (2007) study strategic experimentation with two-armed bandits where the players' actions are publicly observable, but their payoffs are private information. To simplify the analysis, these authors assume that the decision to stop playing risky is irreversible. In our model, players can freely switch back and forth between the two arms.

pessimistic about their respective prospects; with stakes sufficiently high, this implies that at least one player must be using the risky arm at any time, and so learning never stops.

Thus, whenever society places a lot of emphasis on uncovering the truth, as one may argue is the case with medical research or the justice system, our analysis would suggest an adversarial setup was able to achieve this goal. In this respect, our work is related to Dewatripont and Tirole (1999) who, in a moral hazard setting bearing no resemblance to ours, pose the question whether it is socially better to adjudicate disputes through a centralized system of gathering evidence, which they assimilate to the inquisitorial system of Civil Law countries, or whether the interests of justice may be better served in a decentralized, adversarial system, as it is found in the Common Law countries. They show that, in a centralized system, it is not possible to give adequate incentives to make sure the truth is uncovered, and conclude that the Common Law system of gathering information was therefore superior. Our model provides an alternative framework to ascertain the effectiveness of information-gathering processes in a strategic setting where the parties' expected risky payoffs are perfectly negatively correlated across states of the world.

Chatterjee and Evans (2004) analyze an R&D race with two firms and two projects in which, like in our setup, there is common knowledge that exactly one of these projects will bear fruit if pursued long enough, and actions and payoffs are observable. Their discrete-time model differs from ours in several respects, chief of which is the payoff externality implied by the firms' choices. In our model, there is no payoff rivalry between players – strategic interaction arises out of purely informational concerns. Moreover, Chatterjee and Evans allow firms to change their projects at any time, so that it is possible for them to explore the same project. Our analysis, by contrast, presumes that projects of opposite type have been irrevocably assigned to players at the start of the experimentation game.[3]

The rest of the paper is structured as follows. Section 2 introduces the model. Section 3 solves the planner's problem. Section 4 sets up the non-cooperative game. Section 5 discusses long-run properties of learning in equilibrium. Section 6 characterizes the Markov perfect equilibria of the non-cooperative game. Section 7 concludes. Proofs are provided in the appendix.

[3]In the concluding remarks, we briefly report on an extension of our model in which we partially relax this assumption by giving the players a sequential once-and-for-all choice of bandits prior to the experimentation game.

2 The Model

There are two players, 1 and 2, each of whom faces a two-armed bandit problem in continuous time. Bandits are of the exponential type studied in Keller, Rady and Cripps (2005). One arm is safe in that it yields a known payoff flow of s; the other arm is risky in that it is either good or bad. If it is bad, it never yields any payoff; if it is good, it yields a lump-sum payoff with probability λdt when used over a time of length dt. Let $g\,dt$ denote the corresponding expected payoff increment; thus, g is the product of the arrival rate λ and the average size of a lump-sum payoff. The time-invariant constants $\lambda > 0$ and $g > 0$ are common knowledge. It is also common knowledge that exactly one bandit's risky arm is good. To have an interesting problem, we assume that the expected payoff of a good risky arm exceeds that of the safe arm, whereas the safe arm is better than a bad risky arm, i.e., $g > s > 0$.

Each player chooses actions $\{k_t\}_{t\geq 0}$ such that $k_t \in \{0, 1\}$ is measurable with respect to the information available at time t, with $k_t = 1$ indicating use of the risky arm, and $k_t = 0$ use of the safe arm. The strategic link between the two players' actions is provided by the assumption that players perfectly observe each other's actions and payoffs. Thus, as the bandits are perfectly negatively correlated, any information that is garnered about the quality of the risky arm is a public good. At the outset of the game, players have a common prior about which of the risky arms is good. Since the results of each player's experimentation are public, players share a common posterior at all times. We write p_t for the players' posterior probability assessment at time t that player 1's risky arm is good.

The posterior belief jumps to 1 if there has been a breakthrough on player 1's bandit, and to 0 if there has been a breakthrough on player 2's bandit, where in either case it will remain ever after. If there has been no breakthrough on either bandit by time T given the players' actions $\{k_{1,t}\}_{0\leq t\leq T}$ and $\{k_{2,t}\}_{0\leq t\leq T}$, Bayes' rule yields

$$p_T = \frac{p_0 e^{-\lambda \int_0^T k_{1,t}\,dt}}{p_0 e^{-\lambda \int_0^T k_{1,t}\,dt} + (1-p_0) e^{-\lambda \int_0^T k_{2,t}\,dt}},$$

and so the posterior belief satisfies

$$\dot{p}_t = -(k_{1,t} - k_{2,t})\lambda p_t (1 - p_t)$$

at almost all t up to the first breakthrough on a risky arm.

We restrict players to stationary Markov strategies with the common posterior belief as the state variable. The precise definition of the space of Markov strategies available to each player requires some care. Suppose for example that player 1 plays risky at all beliefs $p \geq \frac{1}{2}$

and safe otherwise, while player 2 plays risky at all beliefs $p < \frac{1}{2}$ and safe otherwise. Then the above expression for the time derivative of the posterior belief translates into $\dot{p}_t = f(p_t)$ where

$$f(p) = \begin{cases} \lambda p(1-p) & \text{for} \quad p < \frac{1}{2}, \\ -\lambda p(1-p) & \text{for} \quad p \geq \frac{1}{2}. \end{cases}$$

There is no continuous function $t \mapsto p_t$ from $[0, \infty[$ to $[0,1]$ with $p_0 = \frac{1}{2}$ that is differentiable almost everywhere and satisfies $\dot{p}_t = f(p_t)$ at almost all t.[4] So the above strategies do not induce a well-defined law of motion for beliefs, which means that there are histories of the game after which these strategies do not pin down the players' actions.[5]

To avoid this problem, we impose specific one-sided continuity requirements on the players' Markov strategies: each player's action is right-continuous with left limits with respect to the posterior probability of that player's risky arm being the good one. More precisely, we define a Markov strategy for player 1 to be a function $k_1 : [0,1] \to \{0,1\}$ such that $k_1(0) = 0$ and $k_1^{-1}(1)$ is the disjoint union of the singleton $\{1\}$ and a finite number of left-closed intervals $[p_i', p_i''[$. Symmetrically, a Markov strategy for player 2 is a function $k_2 : [0,1] \to \{0,1\}$ such that $k_2(1) = 0$ and $k_2^{-1}(1)$ is the disjoint union of the singleton $\{0\}$ and a finite number of right-closed intervals $]p_j', p_i'']$. Fixing the players' actions at the boundaries of the unit interval is innocuous because we are simply imposing the dominant actions under subjective certainty. More importantly, our specification of Markov strategies ensures that for any pair (k_1, k_2) of such strategies, the differential equation

$$\dot{p} = -[k_1(p) - k_2(p)]\lambda p(1-p)$$

has a unique global solution for any initial value in the unit interval, which implies that the law of motion for the posterior belief, the players' actions $k_{n,t} = k_n(p_t)$ and their payoff functions $u_n : [0,1] \to I\!R$ are all well-defined. The latter are given by

$$\begin{aligned} u_1(p) &= \mathrm{E}\left[\int_0^\infty re^{-rt}\left\{k_1(p_t)p_t g + [1-k_1(p_t)]s\right\} dt \middle| p_0 = p\right], \\ u_2(p) &= \mathrm{E}\left[\int_0^\infty re^{-rt}\left\{k_2(p_t)(1-p_t)g + [1-k_2(p_t)]s\right\} dt \middle| p_0 = p\right], \end{aligned}$$

[4]Similar problems have been encountered in the decision-theoretic literature. To guarantee a well-defined law of motion, Presman (1990) allows for simultaneous use of both arms, i.e. for experimentation intensities $k_t \in [0,1]$.

[5]Replacing the differential equation for posterior beliefs by a differential inclusion as in Filippov (1988) does not help. Following this approach, one replaces the function f by a correspondence F with $F(p) = \{f(p)\}$ for $p \neq \frac{1}{2}$ and $F(\frac{1}{2}) \supseteq [-\frac{\lambda}{4}, \frac{\lambda}{4}]$, the convex hull of the left and right limits of f at $p = \frac{1}{2}$. A solution to the differential inclusion $\dot{p}_t \in F(p_t)$ with $p_0 = \frac{1}{2}$ is then easily found in the constant function $p_t \equiv \frac{1}{2}$. However, this solution is not meaningful in the context of our model since it is not compatible with Bayes' rule under the given strategies: if $p_t = \frac{1}{2}$ for all t, the action profile must always be $(k_1, k_2) = (1, 0)$, in which case Bayes' rule would imply a downward trend in beliefs conditional on no success on player 1's bandit.

where $r > 0$ is the players' common discount rate and the expectation is taken with respect to the law of motion of posterior beliefs induced by the strategy pair (k_1, k_2). By standard results, our specification of Markov strategies further implies that the Bellman equations appearing in subsequent sections are necessary and sufficient for optimality.

A profile of Markov strategies (k_1, k_2) is called *symmetric* if $k_1(p) = k_2(1-p)$ for all $p \in [0,1]$. A Markov strategy k_1 for player 1 is called a *cutoff strategy* with cutoff $\hat{p}_1$ if $k_1^{-1}(1) = [\hat{p}_1, 1]$. Analogously, a Markov strategy k_2 for player 2 is a cutoff strategy with cutoff $\hat{p}_2$ if $k_2^{-1}(1) = [0, \hat{p}_2]$. If players were myopic, i.e. merely maximizing current payoffs, player 1 would use the cutoff $p^m = \frac{s}{g}$ and player 2 the cutoff $1 - p^m$. If they were forward-looking but experimenting in isolation, player 1 would optimally use the single-agent cutoff computed in Keller, Rady and Cripps (2005), $p^* = \frac{rs}{(r+\lambda)g - \lambda s} < p^m$, and player 2 the cutoff $1 - p^*$. In both cases, the resulting strategy profile would be symmetric.

We will find it useful below to distinguish three cases depending on the size of the stakes involved, i.e. on the value of information as measured by the ratio $\frac{g}{s}$, and on the parameters λ and r that govern the speed of resolution of uncertainty and the player's impatience, respectively. We speak of *low stakes* if $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$; *intermediate stakes* if $\frac{2r+\lambda}{r+\lambda} \leq \frac{g}{s} < 2$, and *high stakes* if $\frac{g}{s} \geq 2$. These cases are easily distinguished by the positions of the cutoffs p^m and p^*: stakes are low if and only if $p^* > \frac{1}{2}$; intermediate if and only if $p^* \leq \frac{1}{2} < p^m$; and high if and only if $p^m \leq \frac{1}{2}$.

3 The Planner's Problem

In this section, we examine a benevolent utilitarian social planner's behavior in our setup. Proceeding exactly as Keller, Rady and Cripps (2005), we can write the Bellman equation for the maximization of the average payoff from the two bandits as

$$u(p) = s + \max_{(k_1,k_2)\in\{0,1\}^2} \left\{ k_1 \left[B_1(p,u) - \frac{c_1(p)}{2} \right] + k_2 \left[B_2(p,u) - \frac{c_2(p)}{2} \right] \right\},$$

where $B_1(p,u) = \frac{\lambda}{r} p[\frac{g+s}{2} - u(p) - (1-p)u'(p)]$ measures the expected learning benefit of playing risky arm 1, $B_2(p,u) = \frac{\lambda}{r}(1-p)[\frac{g+s}{2} - u(p) + pu'(p)]$ the expected learning benefit of playing risky arm 2, $c_1(p) = s - pg$ the opportunity cost of playing risky arm 1, and $c_2(p) = s - (1-p)g$ the opportunity cost of playing risky arm 2. Thus, the planner's problem is linear in both k_1 and k_2, and he is maximizing separately over k_1 and k_2.

If it is optimal to set $k_1 = k_2 = 0$, the value function works out as $u(p) = s$. If it is optimal to set $k_1 = k_2 = 1$, the Bellman equation reduces to $u(p) = \frac{\lambda}{r}\left[\frac{g+s}{2} - u(p)\right] + \frac{g}{2}$, and so $u(p) = u_{11} = \frac{1}{2}\left(g + \frac{\lambda}{\lambda+r} s\right)$.

If it is optimal to set $k_1 = 0$ and $k_2 = 1$, the Bellman equation amounts to the first-order ODE

$$\lambda p(1-p)u'(p) - [r + \lambda(1-p)]u(p) = -\tfrac{1}{2}\left\{[r + \lambda(1-p)]s + (r+\lambda)(1-p)g\right\}.$$

This has the solution

$$u(p) = \tfrac{1}{2}[s + (1-p)g] + Cp^{\frac{r+\lambda}{\lambda}}(1-p)^{-\frac{r}{\lambda}},$$

where C is some constant of integration.

Finally, if it is optimal to set $k_1 = 1$ and $k_2 = 0$, the Bellman equation is tantamount to the first-order ODE

$$\lambda p(1-p)u_1'(p) + (r + \lambda p)u_1(p) = \tfrac{1}{2}\left\{(r + \lambda p)s + (r+\lambda)pg\right\},$$

which is solved by

$$u(p) = \tfrac{1}{2}(s + pg) + C\,(1-p)^{\frac{r+\lambda}{\lambda}}p^{-\frac{r}{\lambda}}.$$

Note that whenever $k_1 = k_2$, the value function is constant as the planner does not care which arm is good. For the same reason, the problem is symmetric around $p = \frac{1}{2}$. All the planner cares about is the uncertainty that stands in the way of his realizing the upper bound on the value function, $\frac{g+s}{2}$. Hence, intuitively, the planner's value function will admit its global minimum at $p = \frac{1}{2}$, where the uncertainty is starkest.

It is clear that $(k_1, k_2) = (1, 0)$ will be optimal in a neighborhood of $p = 1$, and $(k_1, k_2) = (0, 1)$ in a neighborhood of $p = 0$. What is optimal at beliefs around $p = \frac{1}{2}$ depends on which of the two possible plateaus s and u_{11} is higher. This in turn depends on the size of the stakes involved. In fact, $s > u_{11}$ if and only if stakes are low, i.e. $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$. This is the case we consider first.

Proposition 3.1 (Planner's solution for low stakes) *If $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, and hence $p^* > \frac{1}{2}$, it is optimal for the planner to apply the single-agent cutoffs p^* and $1-p^*$, respectively, that is, to set $(k_1, k_2) = (0, 1)$ on $[0, 1-p^*]$, $k_1 = k_2 = 0$ on $]1-p^*, p^*[$, and $(k_1, k_2) = (1, 0)$ on $[p^*, 1]$. The corresponding value function is*

$$u(p) = \begin{cases} \frac{1}{2}\left\{s + (1-p)g + (s - p^*g)\left(\frac{p}{1-p^*}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{1-p}{p^*}\right)^{-\frac{r}{\lambda}}\right\} & \text{if } \ p \le 1-p^*, \\ s & \text{if } \ 1-p^* \le p \le p^*, \\ \frac{1}{2}\left\{s + pg + (s - p^*g)\left(\frac{1-p}{1-p^*}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{p}{p^*}\right)^{-\frac{r}{\lambda}}\right\} & \text{if } \ p \ge p^*. \end{cases}$$

Figure 1 illustrates the result. Thus, when the value of information, as measured by $\frac{g}{s}$, is so low that the single-agent cutoff p^* exceeds $\frac{1}{2}$, it is optimal for the planner to let the players behave as though they were single players solving two separate, completely unconnected, problems.[6]

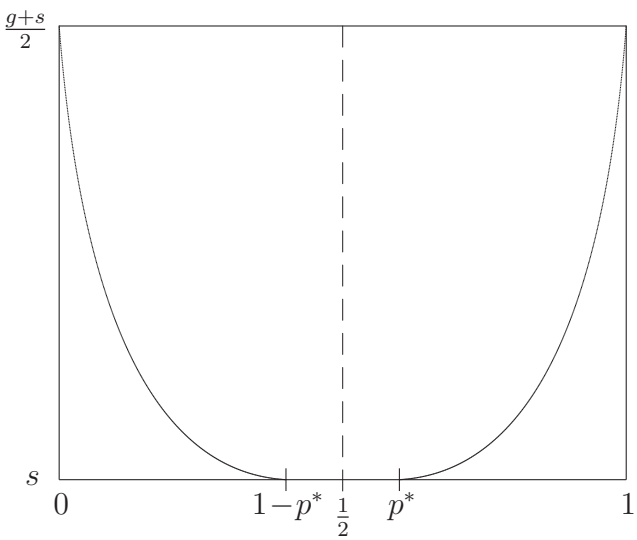

Figure 1: The planner's value function for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$.

Conditional on there not being a breakthrough, the belief will evolve according to

$$\dot{p} = \begin{cases} \lambda p(1-p) & \text{if} \quad p < 1-p^*, \\ 0 & \text{if} \quad 1-p^* \leq p \leq p^*, \\ -\lambda p(1-p) & \text{if} \quad p > p^*. \end{cases}$$

Let us suppose risky arm 1 is good. If the initial belief $p_0 < 1-p^*$, then the posterior belief will converge to $1-p^*$ with probability 1 as there cannot be a breakthrough on risky arm 2. If $1-p^* \leq p_0 \leq p^*$, the belief will remain unchanged at p_0. If $p_0 > p^*$, the belief will converge either to 1 or to p^*. If t^* is the length of time needed for the belief to reach p^* conditional on there not being a breakthrough on risky arm 1, the probability that the belief will converge to p^* is $e^{-\lambda t^*}$. By Bayes' rule, we have $\frac{1-p_t}{p_t} = \frac{1-p_0}{p_0 e^{-\lambda t}}$ in the absence of a breakthrough, and so $e^{-\lambda t^*} = \frac{1-p_0}{p_0}\frac{p^*}{1-p^*}$. The belief will therefore converge to p^* (and learning will remain incomplete) with probability $\frac{1-p_0}{p_0}\frac{p^*}{1-p^*}$, and to 1 (and hence the truth) with the counter-probability. Analogous results hold when risky arm 2 is good.

[6]This would be different if playing the risky arm could also lead to "bad news events" that triggered downward jumps in beliefs. If, starting from p^*, such a jump were large enough to take the belief below $1-p^*$, then letting player 1 play risky at beliefs somewhat below p^* would raise average payoffs if the opportunity cost of doing so were offset by informational gains arising from subsequent experimentation by player 2.

Next, we turn to the case where $u_{11} \geq s$, which is obtained for intermediate and high stakes.

Proposition 3.2 (Planner's solution for intermediate and high stakes) *If $\frac{g}{s} \geq \frac{2r+\lambda}{r+\lambda}$, and hence $p^* \leq \frac{1}{2}$, it is optimal for the planner to apply the cutoffs $\bar{p} = \frac{(r+\lambda)s}{(r+\lambda)g+\lambda s} \in [p^*, \frac{1}{2}]$ and $1-\bar{p}$, respectively, that is, to set $(k_1, k_2) = (0,1)$ on $[0, \bar{p}[$, $k_1 = k_2 = 1$ on $[\bar{p}, 1-\bar{p}]$, and $(k_1, k_2) = (1,0)$ on $]1-\bar{p}, 1]$. The corresponding value function is*

$$u(p) = \begin{cases} \frac{1}{2}\left\{s + (1-p)g + \left[\bar{p}g - \frac{r}{r+\lambda}s\right]\left(\frac{p}{\bar{p}}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{1-p}{1-\bar{p}}\right)^{-\frac{r}{\lambda}}\right\} & \text{if } p \leq \bar{p}, \\ \frac{1}{2}\left(g + \frac{\lambda}{r+\lambda}s\right) & \text{if } \bar{p} \leq p \leq 1-\bar{p}, \\ \frac{1}{2}\left\{s + pg + \left[\bar{p}g - \frac{r}{r+\lambda}s\right]\left(\frac{1-p}{\bar{p}}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{p}{1-\bar{p}}\right)^{-\frac{r}{\lambda}}\right\} & \text{if } p \geq 1-\bar{p}. \end{cases}$$

Figure 2 illustrates this result for $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, which is tantamount to $u_{11} > s$ and easily seen to imply $p^* < \bar{p} < \frac{1}{2}$. To understand why the planner has each player use the risky arm on a smaller interval of beliefs than in the respective single-agent optimum, consider the effect of player 1's action on the average payoff when player 2 is playing risky. If the planner is indifferent between player 1's actions at the belief $\bar{p}$, it must be the case that $B_1(\bar{p}, u) = \frac{c_1(\bar{p})}{2}$. By value matching at the level u_{11} and smooth pasting, this reduces to $\frac{\lambda}{r}\bar{p}[g + s - 2u_{11}] = c_1(\bar{p})$; thus, the possibility of a jump in the sum of the two players' payoffs from $2u_{11}$ to $g + s$ exactly compensates for the opportunity cost of player 1 using the risky arm. For a player 1 experimenting in isolation, the corresponding equation reads $\frac{\lambda}{r}p^*[g - s] = c_1(p^*)$; at the single-agent optimal cutoff, the possibility of a jump in the payoff from s to g exactly compensates for the opportunity cost of player 1 using the risky arm. When $u_{11} > s$, the jump from s to g is larger than the one from $2u_{11}$ to $g + s$, and so we cannot have $\bar{p} = p^*$. That $\bar{p}$ must be greater than p^* follows from the fact that the opportunity cost of using player 1's risky arm is decreasing in p.

The dynamics of beliefs conditional on there not being a breakthrough are now given by

$$\dot{p} = \begin{cases} \lambda p(1-p) & \text{if } p < \bar{p}, \\ 0 & \text{if } \bar{p} \leq p \leq 1-\bar{p}, \\ -\lambda p(1-p) & \text{if } p > 1-\bar{p}. \end{cases}$$

Whenever the stakes are intermediate or high, therefore, the planner shuts down *incremental* learning on $[\bar{p}, 1-\bar{p}]$. Yet he still learns the truth with probability 1 in the long run because this interval is absorbing for the posterior belief process in the absence of a breakthrough, and once it is reached, the planner uses both risky arms until all uncertainty is resolved.

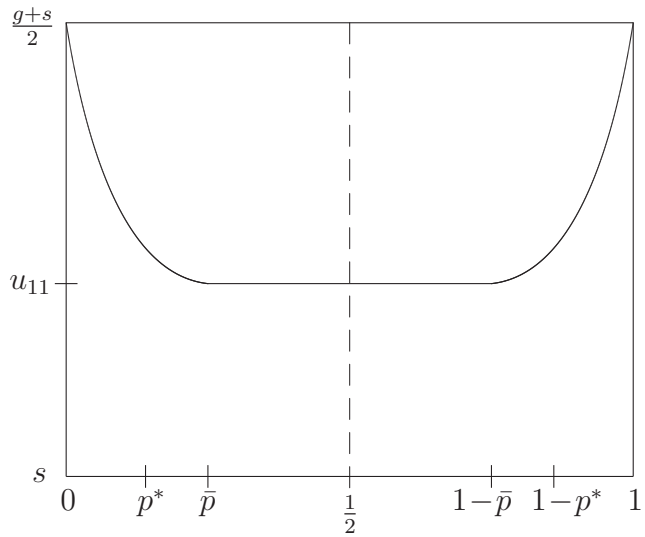

Figure 2: The planner's value function for $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$.

In summary, when stakes are intermediate or high, efficiency calls for complete learning, i.e., almost sure convergence of the posterior belief p_t to the truth. When stakes are low, however, efficient learning can be incomplete.

4 The Strategic Problem

Our solution concept is Markov perfect equilibrium, with the players' strategies as defined in Section 2 above.

Again proceeding as in Keller, Rady and Cripps (2005), we see that the following Bellman equation characterizes player 1's best responses against his opponent's strategy k_2:

$$u_1(p) = s + k_2(p)\beta_1(p, u_1) + \max_{k_1 \in \{0,1\}} k_1[b_1(p, u_1) - c_1(p)],$$

where $c_1(p) = s - pg$ is the opportunity cost player 1 has to bear when he plays risky, $b_1(p, u_1) = \frac{\lambda}{r}p[g - u_1(p) - (1-p)u_1'(p)]$ is the learning benefit accruing to player 1 when he plays risky, and $\beta_1(p, u_1) = \frac{\lambda}{r}(1-p)[s - u_1(p) + pu_1'(p)]$ is his learning benefit from player 2's playing risky.[7]

[7]By standard results, player 1's payoff function from playing a best response against k_2 is once continuously differentiable on any open interval of beliefs where player 2's action is constant. At a belief where k_2 is discontinuous, $u_1'(p)$ must be understood as the one-sided derivative of u_1 in the direction implied by the law of motion of beliefs.

Analogously, the Bellman equation for player 2 is

$$u_2(p) = s + k_1(p)\beta_2(p, u_2) + \max_{k_2 \in \{0,1\}} k_2[b_2(p, u_2) - c_2(p)],$$

where $c_2(p) = s - (1-p)g$ is the opportunity cost player 2 has to bear when he plays risky, $b_2(p, u_2) = \frac{\lambda}{r}(1-p)[g - u_2(p) + pu_2'(p)]$ is the learning benefit accruing to player 2 when he plays risky, and $\beta_2(p, u_2) = \frac{\lambda}{r}p[s - u_2(p) - (1-p)u_2'(p)]$ is his learning benefit from player 1's playing risky.

It is straightforward to obtain closed-form solutions for the payoff functions. If $k_1(p) = k_2(p) = 0$, the players' payoffs are $u_1(p) = u_2(p) = s$. If $k_1(p) = k_2(p) = 1$, the Bellman equations yield $u_1(p) = pg + \frac{\lambda}{\lambda+r}(1-p)s$ and $u_2(p) = u_1(1-p)$. On any interval where $k_1(p) = 1$ and $k_2(p) = 0$, u_1 and u_2 satisfy the ODEs

$$\begin{aligned} \lambda p(1-p)u_1'(p) + (r + \lambda p)u_1(p) &= (r+\lambda)pg, \\ \lambda p(1-p)u_2'(p) + (r + \lambda p)u_2(p) &= (r+\lambda p)s, \end{aligned}$$

which have the solutions $u_1(p) = pg + C_1(1-p)^{\frac{r+\lambda}{\lambda}}p^{-\frac{r}{\lambda}}$ and $u_2(p) = s + C_2(1-p)^{\frac{r+\lambda}{\lambda}}p^{-\frac{r}{\lambda}}$ with constants of integration C_1 and C_2, respectively. Finally, on any interval where $k_1(p) = 0$ and $k_2(p) = 1$, u_1 and u_2 solve

$$\begin{aligned} \lambda p(1-p)u_1'(p) - [r + \lambda(1-p)]u_1(p) &= -[r + \lambda(1-p)]s, \\ \lambda p(1-p)u_2'(p) - [r + \lambda(1-p)]u_2(p) &= -(r+\lambda)(1-p)g, \end{aligned}$$

which implies $u_1(p) = s + C_1 p^{\frac{r+\lambda}{\lambda}}(1-p)^{-\frac{r}{\lambda}}$ and $u_2(p) = (1-p)g + C_2 p^{\frac{r+\lambda}{\lambda}}(1-p)^{-\frac{r}{\lambda}}$.

Note that each of the above closed-form solutions is the sum of one term that expresses the expected payoff from committing to a particular action and another term that captures the option value of being able to switch to the other action.

5 Complete Learning

In this section, we shall show that whenever the planner's solution leads to complete learning, so will any Markov perfect equilibrium of the experimentation game. To this end, we first establish a lower bound on equilibrium payoffs.

From Keller, Rady and Cripps (2005), the optimal payoffs of player 1 and 2, if they were experimenting in isolation and hence applying the cutoffs p^* and $1 - p^*$, respectively, would be

$$u_1^*(p) = \begin{cases} s & \text{if} \quad p \le p^*, \\ pg + (s - p^*g)\left(\frac{1-p}{1-p^*}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{p}{p^*}\right)^{-\frac{r}{\lambda}} & \text{if} \quad p \ge p^* \end{cases}$$

and $u_2^*(p) = u_1^*(1-p)$. Since each player in the experimentation game always has the option to act as though he were a single player by just ignoring the additional signal he gets from the other player, it is quite intuitive that he cannot possibly do worse with the other player around than if he were by himself. The following lemma confirms this intuition.

Lemma 5.1 *The value function of the respective single-agent problem constitutes a lower bound on each player's equilibrium value function in any Markov perfect equilibrium.*

Now, if $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, then $p^* < \frac{1}{2} < 1-p^*$, so at any belief p, Lemma 5.1 implies $u_1^*(p) > s$ or $u_2^*(p) > s$ or both. Thus, there cannot exist a p such that $k_1(p) = k_2(p) = 0$ be mutually best responses as this would mean $u_1(p) = u_2(p) = s$. This proves the following proposition:

Proposition 5.2 (Complete learning) *If $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, learning will be complete in any Markov perfect equilibrium.*

Proposition 6.5 below will show that in the knife-edge case where $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$, the unique Markov perfect equilibrium also entails complete learning. Whenever efficiency calls for complete learning, therefore, learning will be complete in equilibrium. This result is in stark contrast to the benchmark problem of perfect positive correlation in Bolton and Harris (1999) and Keller, Rady and Cripps (2005), where any MPE entails an inefficiently large probability of incomplete learning.

6 Markov Perfect Equilibria

Our next aim is to the characterize the Markov perfect equilibria of the experimentation game.

The profile of actions (k_1, k_2) must be $(0,0)$, $(0,1)$, $(1,0)$ or $(1,1)$ at any belief. For $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, the profile $(1,1)$ cannot occur in equilibrium since it would imply an average payoff of $u_{11} < s$ at the relevant belief, giving at least one player a payoff below s, and hence below his single-agent optimum. For $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, on the other hand, the profile $(0,0)$ cannot occur since it would imply incomplete learning.

We say that the *transition* (k_1^-, k_2^-)—(k_1^+, k_2^-)—(k_1^+, k_2^+) occurs at the belief $\hat{p} \in]0,1[$ if $\lim_{p\uparrow\hat{p}} k_1(p) = k_1^-$, $(k_1(\hat{p}), k_2(\hat{p})) = (k_1^+, k_2^-)$, $\lim_{p\downarrow\hat{p}} k_2(p) = k_2^+$, and at least one of the sets $\{k_1^-, k_1^+\}$ and $\{k_2^-, k_2^+\}$ contains more than one element. Given our definition of strategies, each equilibrium has a finite number of transitions.

We first consider transitions where one player's action stays fixed. Invoking the standard principles of value matching and smooth pasting, we obtain the following result.

Lemma 6.1 *In any Markov perfect equilibrium, a transition* $(k_1^-, 0)$—$(k_1^+, 0)$—$(k_1^+, 0)$ *can only occur at the belief* p^*, $(0, k_2^-)$—$(0, k_2^-)$—$(0, k_2^+)$ *only at* $1 - p^*$, $(k_1^-, 1)$—$(k_1^+, 1)$—$(k_1^+, 1)$ *only at* p^m, *and* $(1, k_2^-)$—$(1, k_2^-)$—$(1, k_2^+)$ *only at* $1 - p^m$.

While it is intuitive that a player would apply the single-agent cutoff rule against an opponent who plays safe and thus provides no information, it is surprising that the myopic cutoff determines equilibrium behavior against an opponent who plays risky. Technically, this result is due to the fact that along player 1's payoff function for $k_1 = k_2 = 1$, his learning benefit from playing risky vanishes:

$$b_1(p, u_1) = \frac{\lambda}{r} p \left[g - \left(pg + \frac{\lambda}{\lambda + r}(1 - p)s \right) - (1 - p) \left(g - \frac{\lambda}{\lambda + r} s \right) \right] = 0,$$

and so $k_1 = 1$ is optimal against $k_2 = 1$ if and only if $c_1(p) \leq 0$, that is, $p \geq p^m$. This is best understood by recalling the law of motion of beliefs in the absence of a success on either arm, $\dot{p} = -(k_1 - k_2)\lambda p(1 - p)$, which tells us that if both players are playing risky, the state variable does not budge until the first success occurs and all uncertainty is resolved. In other words, all a player does by chiming in in his opponent's experimentation is to keep the belief, his action and his continuation value constant and wait for the resolution of uncertainty. But this can only be optimal if he reaps maximal current payoffs while waiting. So his playing the risky arm must be myopically optimal.

In the following lemma, we consider the transitions where both players change action.

Lemma 6.2 *The following statements hold for all Markov perfect equilibria. (i) The transition* $(1, 0)$—$(0, 0)$—$(0, 1)$ *can only occur if* $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$ *and only at beliefs in* $[1 - p^*, p^*]$. *(ii) The transition* $(0, 1)$—$(1, 1)$—$(1, 0)$ *can only occur if* $\frac{2r+\lambda}{r+\lambda} \leq \frac{g}{s} \leq 2$ *and only at beliefs in* $[1 - p^m, p^m]$.

The structure of Markov perfect equilibria depends on the relative position of the possible transition points, which in turn depends on the stakes involved, i.e. on the ratio $\frac{g}{s}$. For expositional reasons, we shall first analyze the cases of low and high stakes.

6.1 Low Stakes

Recall that the low-stakes case is defined by the inequality $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$. In this case, $1 - p^m < 1 - p^* < \frac{1}{2} < p^* < p^m$.

Proposition 6.3 (Markov perfect equilibrium for low stakes) *When $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$, the unique Markov perfect equilibrium is symmetric and coincides with the planner's solution. That is, player 1 plays risky on $[p^*, 1]$ and safe on $[0, p^*[$, while player 2 plays risky on $[0, 1-p^*]$ and safe on $]1-p^*, 1]$. The pertaining value functions are those of the respective single-agent problems, u_1^* and u_2^*.*

Figure 3 illustrates this result.[8] The players' average payoff function coincides with the planner's value function as stated in Proposition 3.1.

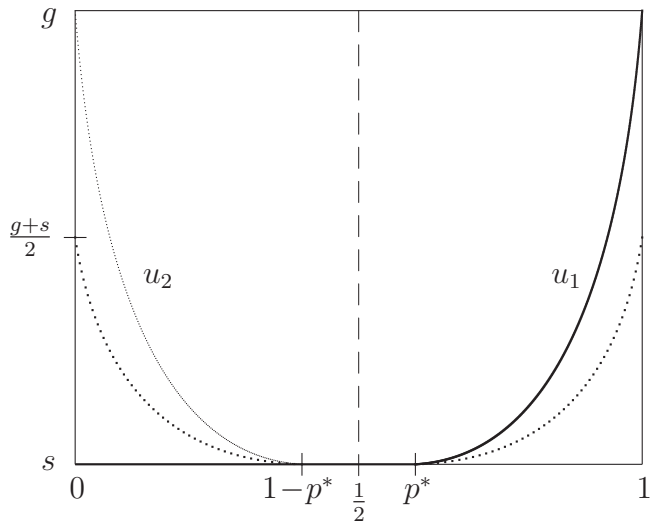

Figure 3: The equilibrium value functions for $\frac{g}{s} < \frac{2r+\lambda}{r+\lambda}$.

Why we should have efficiency in this case is intuitively quite clear, as the planner lets players behave as though they were single players. As $p^* > \frac{1}{2}$, there is no spillover from a player behaving like a single agent on the other player's optimization problem. Hence the latter's best response calls for behaving like a single player as well. Thus, there is no conflict between social and private incentives.[9]

The law of motion for the belief and the probability of the players' eventually finding out the true state of the world are thus the same as in the planner's solution for low stakes.

[8] In this and all subsequent figures, the thick solid line depicts the value function of player 1, the thin solid line that of player 2, and the dotted line the players' average payoff function.

[9] As already discussed in the introduction, the uniqueness part of Proposition 6.3 is obvious given the lower bound on equilibrium payoffs in Lemma 5.1 and the planner's solution in Proposition 3.1. The proof of uniqueness that we give in the appendix does not rely on knowledge of the planner's solution. This method of proof has the advantage that it carries over to intermediate and high stakes.

6.2 High Stakes

The high-stakes case is defined by the inequality $\frac{g}{s} \geq 2$. In this case, $p^* < p^m \leq \frac{1}{2} \leq 1-p^m < 1-p^*$.

Proposition 6.4 (Markov perfect equilibrium for high stakes) *When $\frac{g}{s} \geq 2$, the unique Markov perfect equilibrium is symmetric and has both players behave myopically. That is, player 1 plays risky on $[p^m, 1]$ and safe on $[0, p^m[$, while player 2 plays risky on $[0, 1-p^m]$ and safe on $]1-p^m, 1]$. The pertaining value functions are*

$$u_1(p) = \begin{cases} s + \frac{\lambda}{\lambda+r}(1-p^m)s\left(\frac{p}{p^m}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{1-p}{1-p^m}\right)^{-\frac{r}{\lambda}} & \text{if } \ p \leq p^m, \\ pg + \frac{\lambda}{\lambda+r}(1-p)s & \text{if } \ p^m \leq p \leq 1-p^m, \\ pg + \frac{\lambda}{\lambda+r}p^m s\left(\frac{1-p}{p^m}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{p}{1-p^m}\right)^{-\frac{r}{\lambda}} & \text{if } \ p \geq 1-p^m \end{cases}$$

and $u_2(p) = u_1(1-p)$.

When the stakes are high, the unique equilibrium calls for both players' behaving myopically. This is best understood by recalling from our discussion above that individual optimality calls for myopic behavior whenever one's opponent is playing risky. When the stakes are high, players' myopic cutoff beliefs are more pessimistic than $p = \frac{1}{2}$, so the relevant intervals overlap.

Figure 4 illustrates this result. Player 1's value function has a kink at $1-p^m$, where player 2 changes action. Symmetrically, player 2's value function has a kink at p^m, where player 1 changes action. As a consequence, the average payoff function has a kink both at p^m and at $1-p^m$. That it dips below the level u_{11} close to these kinks is evidence of the inefficiency of equilibrium. We will return to this point in Section 6.4 below.

Arguing exactly as after Proposition 3.2, it is straightforward to see that learning will be complete, as predicted by Proposition 5.2.

6.3 Intermediate Stakes

This case is defined by the condition that $\frac{2r+\lambda}{r+\lambda} \leq \frac{g}{s} < 2$. In this case, $p^* \leq \frac{1}{2} < p^m$.

When the stakes are intermediate in size, equilibrium is not unique; rather there is a continuum of equilibria, as the following proposition shows.

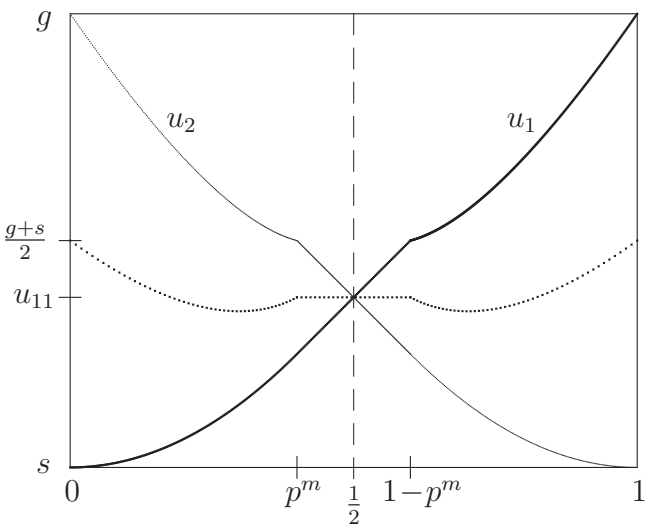

Figure 4: The equilibrium value functions for $\frac{g}{s} > 2$.

Proposition 6.5 (Markov perfect equilibria for intermediate stakes) *When $\frac{2r+\lambda}{r+\lambda} \leq \frac{g}{s} < 2$, there is a continuum of Markov perfect equilibria. Each of them is characterized by a unique belief $\hat{p} \in [\max\{1-p^m, p^*\}, \min\{p^m, 1-p^*\}]$ such that player 1 plays risky if and only if $p \geq \hat{p}$, and player 2 if and only if $p \leq \hat{p}$. The pertaining value functions are given by*

$$u_1(p) = \begin{cases} s + \left[\hat{p}g + \frac{\lambda}{\lambda+r}(1-\hat{p})s - s\right]\left(\frac{p}{\hat{p}}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{1-p}{1-\hat{p}}\right)^{-\frac{r}{\lambda}} & \text{if } p \leq \hat{p} \\ pg + \frac{\lambda}{\lambda+r}(1-\hat{p})s\left(\frac{1-p}{1-\hat{p}}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{p}{\hat{p}}\right)^{-\frac{r}{\lambda}} & \text{if } p \geq \hat{p} \end{cases}$$

for player 1, and

$$u_2(p) = \begin{cases} (1-p)g + \frac{\lambda}{\lambda+r}\hat{p}s\left(\frac{p}{\hat{p}}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{1-p}{1-\hat{p}}\right)^{-\frac{r}{\lambda}} & \text{if } p \leq \hat{p} \\ s + \left[(1-\hat{p})g + \frac{\lambda}{\lambda+r}\hat{p}s - s\right]\left(\frac{1-p}{1-\hat{p}}\right)^{\frac{r+\lambda}{\lambda}}\left(\frac{p}{\hat{p}}\right)^{-\frac{r}{\lambda}} & \text{if } p \geq \hat{p} \end{cases}$$

for player 2.

Amongst the continuum of equilibria characterized in Proposition 6.5, there is a unique symmetric one, given by $\hat{p} = \frac{1}{2}$. Figure 5 illustrates this equilibrium. Both players' value functions and their average are kinked at $p = \frac{1}{2}$, where both players change action. At any belief except $p = \frac{1}{2}$, the average payoff function is below the planner's solution; if the initial belief is $p_0 = \frac{1}{2}$, however, the efficient average payoff u_{11} is achieved.

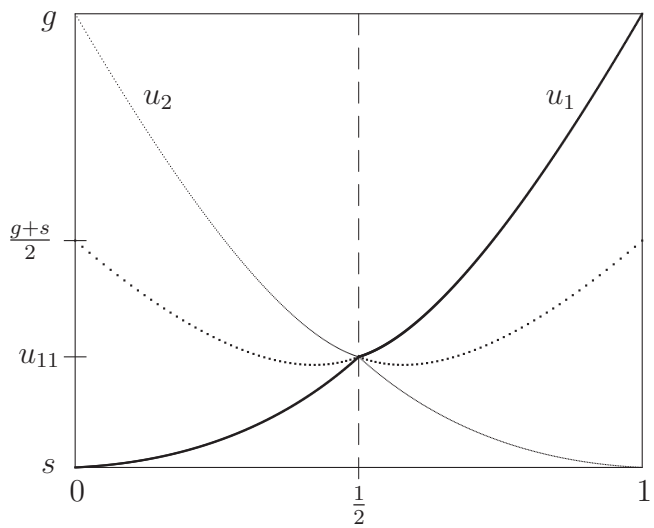

Figure 5: The value functions in the unique symmetric equilibrium for $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < 2$.

For arbitrary $\hat{p}$, the dynamics of beliefs in the absence of a breakthrough are given by

$$\dot{p} = \begin{cases} \lambda p(1-p) & \text{if} \quad p < \hat{p}, \\ 0 & \text{if} \quad p = \hat{p}, \\ -\lambda p(1-p) & \text{if} \quad p > \hat{p}. \end{cases}$$

As predicted by Proposition 5.2, learning is complete in all these equilibria.

6.4 Efficiency vs. Myopia

As we have pointed out already, when the stakes are low, players do not interfere with each other's optimization problem and behave as though they were all by themselves. We have seen that this kind of behavior is also efficient.

If stakes are high, however, we have seen that players behave myopically. This implies that in the unique MPE, experimentation is at efficient levels except on $[\bar{p}, p^m] \cup [1-p^m, 1-\bar{p}]$, the union of two non-empty and non-degenerate intervals, where experimentation is inefficiently low. Put differently, there is a region of beliefs where one player free-rides on the other player's experimentation, which is inefficient from a social point of view.

In the case of intermediate stakes, equilibrium behavior changes gradually from efficiency to myopia. Indeed, as is easily verified, if $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} \leq \frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}}$, then the lower bound on the equilibrium cutoff $\hat{p}$ satisfies $\max\{p^*, 1-p^m\} \leq \bar{p}$. Now, if the players' initial belief is $p_0 < \bar{p}$, the equilibrium with $\hat{p} = \bar{p}$ achieves efficiency as the only beliefs that

are reached with positive probability under the equilibrium strategies are given by the set $\{0,1\} \cup [p_0, \bar{p}]$, and the equilibrium strategies prescribe the efficient actions at all of these beliefs. Similarly, for $p_0 > 1 - \bar{p}$, efficiency is achieved by the equilibrium with $\hat{p} = 1 - \bar{p}$. Finally, if $\bar{p} \leq p_0 \leq 1 - \bar{p}$, efficiency is achieved by the equilibrium with $\hat{p} = p_0$, since this ensures that only beliefs in $\{0, p_0, 1\}$ are reached with positive probability.

If $\frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}} < \frac{g}{s} < 2$, then $p^* < \bar{p} < 1 - p^m$. Now, suppose $\bar{p} \leq p_0 < 1 - p^m$. Equilibrium uniquely calls for $(k_1, k_2)(p) = (0,1)$ for all $p \leq 1 - p^m$, whereas efficiency would require $(k_1, k_2)(p) = (1,1)$ whenever $\bar{p} < p \leq 1 - \bar{p}$. Thus, equilibrium implies inefficient play on the interval $]\bar{p}, 1 - p^m[$ which is reached with positive probability given the initial belief p_0.

Combined with our results for low and high stakes, these arguments establish the following proposition.

Proposition 6.6 (Efficiency) *If $\frac{g}{s} \leq \frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}}$, then for each initial belief, there exists a Markov perfect equilibrium that achieves the efficient outcome. If $\frac{g}{s} > \frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}}$, there are initial beliefs under which the efficient outcome cannot be reached in equilibrium.*

If $1 + \sqrt{\frac{r}{r+\lambda}} \leq \frac{g}{s} < 2$, then $p^m \leq 1 - p^*$. In this situation, setting $\hat{p} = p^m$ ($\hat{p} = 1 - p^m$) yields an equilibrium where only player 1 (player 2) behaves myopically, while the other player bears the entire burden of experimentation by himself, something he is only willing to do provided the stakes involved exceed the threshold of $1 + \sqrt{\frac{r}{r+\lambda}}$. In view of our findings for low and high stakes, this establishes the following result.

Proposition 6.7 (Myopic behavior) *If $\frac{g}{s} \geq 1 + \sqrt{\frac{r}{r+\lambda}}$, there exists a Markov perfect equilibrium where at least one of the players behaves myopically. If $\frac{g}{s} < 1 + \sqrt{\frac{r}{r+\lambda}}$, no player behaves myopically in equilibrium.*

Note that for certain parameter values, namely if $1 + \sqrt{\frac{r}{r+\lambda}} \leq \frac{g}{s} \leq \frac{2r+\lambda}{2(r+\lambda)} + \sqrt{\frac{(2r+\lambda)^2}{4(r+\lambda)^2} + \frac{\lambda}{r+\lambda}}$, equilibria where one player behaves myopically co-exist with equilibria that achieve efficiency given the initial belief.

7 Concluding Remarks

We have analyzed a game of strategic experimentation in continuous time where players' expected risky payoffs are perfectly negatively correlated across states of the world. We have

found that, in sharp contrast to the case of perfectly positive correlation, all the equilibria are of the cutoff type, and that for a large subset of parameters, equilibrium is unique. When the stakes are low, equilibrium behavior is efficient, whereas for high stakes players behave myopically.

In order to ensure a well-defined state variable for all initial values, we have restricted players' strategies to be continuous in the direction of more optimistic beliefs. Although this restriction rules out more transitions than would be necessary to guarantee a well-defined solution to the law of motion for every initial belief, it turns out to be innocuous in the sense that no further equilibria emerge when we rule out only those transitions that are incompatible with a well-defined law of motion.[10]

Furthermore, we have restricted attention to what in the literature have been termed "pure strategy equilibria" (by Bolton and Harris, 1999 and 2000) or "'simple equilibria" (by Keller, Rady and Cripps, 2005, and Keller and Rady, 2007). Our results on efficiency and complete learning are robust to an extension of the strategy space where players are allowed to choose experimentation intensities from the entire unit interval.

Our results on efficiency and complete learning would not change either if, as in Keller and Rady (2007), a bad risky arm also had a non-zero arrival rate of lump-sum payoffs, so that the first success on a risky arm no longer revealed the true state of the world.

Our analysis naturally raises the question under what circumstances players would *choose* to play a strategic experimentation game with perfectly negatively, rather than positively, correlated bandits. To analyze this question, we can extend our model by letting players first decide sequentially whether they want to experiment with risky arm 1, whose prior probability of being good is p_0, or with risky arm 2, whose corresponding probability is $1-p_0$. They then play the strategic experimentation game with either perfectly positively or negatively correlated bandits, as the case might be. Using the fact that in any equilibrium of the experimentation game, no player can obtain a payoff higher than twice the planner's solution minus the single-agent solution, it is straightforward to derive a condition on the model parameters under which equilibrium of the extended game uniquely predicts that players choose different risky arms for all priors p_0 in a neighborhood of $\frac{1}{2}$.[11] It is easy to find parameter combinations that satisfy this condition; for instance, $\frac{r}{\lambda} = 2$ and $\frac{g}{s} = 3$ will do. Given r and λ, moreover, the condition will always be fulfilled if the stakes $\frac{g}{s}$ are

[10]We have chosen the *a priori* more restrictive course because using the alternative method would have entailed the undesirable feature that a player's strategy space depended on his opponent's strategy. The treatment of this case, which, as noted, leads to the exact same set of equilibria, is available from the authors upon request.

[11]Details are available from the authors upon request.

large enough. This shows that for high uncertainty (p_0 close to $\frac{1}{2}$) and sufficiently high stakes, players prefer going their own separate ways on mutually exclusive hypotheses over investigating the same hypothesis together – even when, as we assume here, there are perfect spillovers and no risk of preemption.

While this extension of our model merely allows for one irreversible project selection, we plan to explore repeated selection in a variant of our setup where, akin to Chatterjee and Evans (2004), each player has access to both risky arms and can choose between them at will. This requires each player to solve a three-armed bandit problem with a safe arm and two risky arms that are known to be of opposite types.

Finally, we plan to explore the strategic experimentation problem with imperfect correlation between bandit types, which necessarily involves a state space with more than two elements and beliefs that evolve in a simplex of dimension higher than one.

Appendix

Proof of Proposition 3.1

The policy (k_1, k_2) implies a well-defined law of motion for the posterior belief. The function u satisfies value matching and smooth pasting at p^* and $1-p^*$, hence is of class C^1. It is strictly decreasing on $[0, 1-p^*]$ and strictly increasing on $[p^*, 1]$. Moreover, $u = s + B_2 - \frac{c_2}{2}$ on $[0, 1-p^*]$, $u = s$ on $[1-p^*, p^*]$, and $u = s + B_1 - \frac{c_1}{2}$ on $[p^*, 1]$ (we drop the arguments for simplicity), which shows that u is indeed the planner's payoff function from (k_1, k_2).

To show that u and this policy (k_1, k_2) solve the planner's Bellman equation, and hence that (k_1, k_2) is optimal, it is enough to establish that $B_1 < \frac{c_1}{2}$ and $B_2 > \frac{c_2}{2}$ on $]0, 1-p^*[$, $B_1 < \frac{c_1}{2}$ and $B_2 < \frac{c_2}{2}$ on $]1-p^*, p^*[$, and $B_1 > \frac{c_1}{2}$ and $B_2 < \frac{c_2}{2}$ on $]p^*, 1[$. Consider this last interval. There, $u = s + B_1 - \frac{c_1}{2}$ and $u > s$ (by monotonicity of u) immediately imply $B_1 > \frac{c_1}{2}$. Next, $B_2 = \frac{\lambda}{r}[\frac{g+s}{2} - u] - B_1 = \frac{\lambda}{r}[\frac{g+s}{2} - u] - u + s - \frac{c_1}{2}$; this is smaller than $\frac{c_2}{2}$ if and only if $u > u_{11}$, which holds here since $u > s$ and $s > u_{11}$. The other two intervals are treated in a similar way. ∎

Proof of Proposition 3.2

It is straightforward to check that $p^* \leq \bar{p} \leq \frac{1}{2}$ if $\frac{g}{s} \geq \frac{2r+\lambda}{r+\lambda}$. The rest of the proof proceeds along the same lines as the previous one and is therefore omitted. ∎

Definitions and an Auxiliary Result

For $p \in [0, 1]$, we define

$$w_1(p) = pg + \frac{\lambda}{r+\lambda}(1-p)s \quad \text{and} \quad w_2(p) = (1-p)g + \frac{\lambda}{r+\lambda}ps = w_1(1-p).$$

We recall from Section 4 that these are the players' payoff functions when both are playing risky. Furthermore, we define the players' expected full-information payoffs:

$$\overline{u}_1(p) = pg + (1-p)s \quad \text{and} \quad \overline{u}_2(p) = (1-p)g + ps = \overline{u}_1(1-p).$$

The following lemma will be useful in the proofs of Lemma 6.2 and Propositions 6.4–6.5.

Lemma A.1 *At any belief where the payoff function of player n satisfies $u_n(p) = s + \beta_n(p, u_n)$, the sign of $b_n(p, u_n) - c_n(p)$ coincides with the sign of $w_n(p) - u_n(p)$.*

Proof: We first note that $b_n(p, u_n) = \frac{\lambda}{r}[\overline{u}_n(p) - u_n(p)] - \beta_n(p, u_n)$. As $\beta_n(p, u_n) = u_n(p) - s$, this implies $b_n(p, u_n) - c_n(p) = \frac{\lambda}{r}[\overline{u}_n(p) - u_n(p)] - u_n(p) + s - c_n(p) = \frac{r+\lambda}{r}[w_n(p) - u_n(p)]$. ∎

Proof of Lemma 5.1

Let u_1 be player 1's equilibrium value function in some MPE with equilibrium strategies (k_1, k_2). Write $b_1^*(p) = b_1(p, u_1^*)$, and $\beta_1^*(p) = \beta_1(p, u_1^*)$. Henceforth, we shall suppress arguments whenever

this is convenient. Since p^* is the single-agent cutoff belief for player 1, we have $u_1^* = s$ for $p \leq p^*$ and $u_1^* = s + b_1^* - c_1 = pg + b_1^*$ for $p > p^*$. Thus, if $p \leq p^*$, the claim obviously holds as s is a lower bound on u_1.

Now, let $p > p^*$. Then, noting that $b_1^* = u_1^* - pg$, we have $\beta_1^* = \frac{\lambda}{r}[\overline{u}_1 - u_1^*] - (u_1^* - gp)$. Thus, $\beta_1^* > 0$ if and only if $u_1^* < pg + \frac{\lambda}{r+\lambda}(1-p)s = w_1$. Noting that $w_1(p^*) = u_1^*(p^*) = s$, $w_1(1) = u_1^*(1) = g$, and that w_1 is linear whereas u_1^* is strictly convex in p, we conclude that $u_1^* < w_1$ and hence $\beta_1^* > 0$ on $]p^*, 1[$. As a consequence, we have $u_1^* = pg + b_1^* \leq gp + k_2\beta_1^* + b_1^*$ on $[p^*, 1]$.

Now, suppose $u_1 < u_1^*$ at some belief. Since s is a lower bound on u_1, this implies existence of a belief strictly greater than p^* where $u_1 < u_1^*$ and $u_1' \leq (u_1^*)'$. This immediately yields $b_1 > b_1^* > c_1$, so that we must have $k_1 = 1$ and $u_1 = pg + k_2\beta_1 + b_1$ at the belief in question. But now,

$$u_1 - u_1^* \geq pg + k_2\beta_1 + b_1 - (pg + k_2\beta_1^* + b_1^*) = (1-k_2)(b_1 - b_1^*) + k_2\left[\frac{\lambda}{r}(u_1^* - u_1)\right] > 0,$$

a contradiction.

An analogous argument applies for player 2's equilibrium value function u_2. ∎

Proof of Lemma 6.1

At each of these transitions, we must have value matching and smooth pasting for the player who changes his action. For example, suppose that there is a transition $(0,0)$—$(1,0)$—$(1,0)$ at the belief $\hat{p}$. Then the value function of player 1 must satisfy $u_1(\hat{p}) = s$, $u_1'(\hat{p}) = 0$ and $\lambda\hat{p}(1-\hat{p})u_1'(\hat{p}) + (r + \lambda\hat{p})u_1(\hat{p}) = (r+\lambda)\hat{p}g$ by the ODE for $(k_1, k_2) = (1, 0)$. Substituting for $u_1(\hat{p})$ and $u_1'(\hat{p})$ and solving yields $\hat{p} = \frac{rs}{(r+\lambda)g - \lambda s} = p^*$. The other transitions are dealt with in the same way. ∎

Proof of Lemma 6.2

Suppose the transition $(1,0)$—$(0,0)$—$(0,1)$ occurs at belief $\hat{p}$. This implies $u_1(\hat{p}) = u_2(\hat{p}) = s$. Now, player 2's value function solves the ODE for $k_1 = 0$ and $k_2 = 1$ to the right of $\hat{p}$, which, by continuity of u_2, implies $\lambda\hat{p}(1-\hat{p})u_2'(\hat{p}+) = [r + \lambda(1-\hat{p})]s - (r+\lambda)(1-\hat{p})g$, where $u_2'(\hat{p}+) := \lim_{p\downarrow\hat{p}} u_2'(p)$, and so we find $u_2'(\hat{p}+) < 0$ whenever $\hat{p} < 1 - p^*$. So we must have $\hat{p} \geq 1 - p^*$. Now, player 1's value function solves the ODE for $k_1 = 1$ and $k_2 = 0$ to the left of $\hat{p}$, which implies $\lambda\hat{p}(1-\hat{p})u_1'(\hat{p}-) = (r+\lambda)\hat{p}g - (r+\lambda\hat{p})s$, where $u_1'(\hat{p}-) = \lim_{p\uparrow\hat{p}} u_1'(p)$; so we have $u_1'(\hat{p}-) > 0$ whenever $\hat{p} > p^*$. Thus, we must have $\hat{p} \in [1-p^*, p^*]$, which requires $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$. This proves statement (i).

Suppose now that the transition $(0,1)$—$(1,1)$—$(1,0)$ occurs at belief $\hat{p}$. This implies $u_1(\hat{p}) = w_1(\hat{p})$ and $u_2(\hat{p}) = w_2(\hat{p})$. To the right of $\hat{p}$, player 2' value function solves the ODE for $k_1 = 1$ and $k_2 = 0$, which implies

$$u_2'(\hat{p}+) = \frac{r + \lambda\hat{p}}{\lambda\hat{p}(1-\hat{p})}\left[\frac{r + \lambda(1-\hat{p})}{r+\lambda}s - (1-\hat{p})g\right].$$

Now, if $\hat{p} < 1 - p^m$, then $u_2'(\hat{p}+) < w_2'(\hat{p})$ and so $u_2 < w_2$ to the immediate right of $\hat{p}$, implying by Lemma A.1 that $k_2 = 0$ is *not* a best response to $k_1 = 1$ there – a contradiction. Thus, we must

have $\hat{p} \geq 1 - p^m$. To the left of $\hat{p}$, player 1's value function solves the ODE for $k_1 = 0$ and $k_2 = 1$, which implies

$$u_1'(\hat{p}-) = \frac{r + \lambda(1-\hat{p})}{\lambda\hat{p}(1-\hat{p})}\left[\hat{p}g - \frac{r+\lambda\hat{p}}{r+\lambda}s\right].$$

If $\hat{p} > p^m$, then $u_1'(\hat{p}-) > w_1'(\hat{p})$ and so $u_1 < w_1$ to the immediate left of $\hat{p}$ – another contradiction by Lemma A.1. So we must have $\hat{p} \in [1 - p^m, p^m]$, which requires $\frac{g}{s} \leq 2$. Furthermore, we note that the existence of a belief $\hat{p}$ such that $k_1(\hat{p}) = k_2(\hat{p}) = 1$ requires $u_{11} \geq s$ and hence $\frac{g}{s} \geq \frac{2r+\lambda}{r+\lambda}$. This proves statement (ii). ∎

Proof of Proposition 6.3

The functions u_1 and u_2 are of class C^1 with u_2 strictly decreasing on $[0, 1 - p^*]$ and u_1 strictly increasing on $[p^*, 1]$. As $u_2 = s + b_2 - c_2$ on $[0, 1 - p^*]$ and $u_1 = s + b_1 - c_1$ on $[p^*, 1]$ (we drop the arguments for simplicity), u_1 and u_2 are indeed the players' payoff functions for (k_1, k_2).

To show that u_1 and the policy k_1 solve player 1's Bellman equation given player 2's strategy k_2, and hence that k_1 is a best response to k_2, it is enough to establish that $b_1 < c_1$ on $]0, p^*[$ and $b_1 > c_1$ on $]p^*, 1[$. On this last interval, $u = s + b_1 - c_1$ and $u_1 > s$ (by monotonicity of u_1) immediately imply $b_1 > c_1$. On $]0, p^*[$, we have $u_1 = s$ and $u_1' = 0$, hence $b_1 - c_1 = \frac{\lambda}{r}p(g-s) - (s - pg) = \frac{(r+\lambda)g-\lambda s}{r}p - s < 0$. As $u_2(p) = u_1(1-p)$ and $k_2(p) = k_1(1-p)$, the previous steps also imply $b_2 > c_2$ on $]0, 1 - p^*[$ and $b_2 < c_2$ on $]1 - p^*, 1[$, which completes the proof that (k_1, k_2) constitutes an equilibrium.

For uniqueness, recall that, as $u_{11} < s$, the action profile $(k_1, k_2) = (1, 1)$ cannot be part of an MPE since this would involve a payoff strictly below s for at least one player at some belief. Of the transitions considered in Lemma 6.2, only $(1, 0)$—$(0, 0)$—$(0, 1)$ could happen in this case, and it could only occur at some belief $\hat{p} \in [1 - p^*, p^*]$. It thus follows from Lemma 6.1 that in any MPE, players can only transition out of $(0, 1) = (k_1(0), k_2(0))$ at belief $1 - p^*$, and have to move into $(0, 0)$ to the immediate right of $1 - p^*$. As $(k_1(1), k_2(1)) = (1, 0)$, players cannot transition back into $(0, 1)$ to the right of $1 - p^*$. ∎

Proof of Proposition 6.4

The functions u_1 and u_2 are of class C^1 except at $1 - p^m$ and p^m, respectively, where their first derivative jumps downward; u_1 is strictly increasing, u_2 strictly decreasing. Moreover, $u_1 = s + \beta_1$ and $u_2 = s + b_2 - c_2$ on $[0, p^m]$, $u_1 = s + \beta_1 + b_1 - c_1$ and $u_2 = s + \beta_2 + b_2 - c_2$ on $[p^m, 1 - p^m]$, and $u_1 = s + b_1 - c_1$ and $u_2 = s + \beta_2$ on $[1 - p^m, 1]$. So u_1 and u_2 are indeed the players' payoff functions for (k_1, k_2).

To show that u_1 and the policy k_1 solve player 1's Bellman equation given player 2's strategy k_2, and hence that k_1 is a best response to k_2, it is enough to establish that $b_1 < c_1$ on $]0, p^m[$ and $b_1 > c_1$ on $]p^m, 1[$. On $]1-p^m, 1[$, $u_1 = s+b_1-c_1$ and $u_1 > s$ (by monotonicity of u_1) immediately imply $b_1 > c_1$. On $]p^m, 1-p^m[$, we have $b_1 = 0 > c_1$. On $]0, p^m[$, it is easily verified that $u_1 > w_1$, so Lemma A.1 implies $b_1 < c_1$. As $u_2(p) = u_1(1-p)$ and $k_2(p) = k_1(1-p)$, the previous steps also

imply $b_2 > c_2$ on $]0, 1 - p^m[$ and $b_2 < c_2$ on $]1 - p^m, 1[$, which completes the proof that (k_1, k_2) constitutes an equilibrium.

For uniqueness, we recall that the action profile $(k_1, k_2) = (0, 0)$ cannot be part of an MPE since it would imply incomplete learning. It thus follows immediately from Lemmas 6.1 and 6.2 that the only way for players to transition out of $(0, 1) = (k_1(0), k_2(0))$ is for them to switch to $(1, 1)$ at p^m. Thus, players cannot transition back into $(0, 1)$ to the right of p^m. Therefore, again using Lemma 6.1, the only way to transition out of $(1, 1)$ is to switch to $(1, 0)$ at $1 - p^m$. Hence, players cannot transition back to $(1, 1)$ or $(0, 1)$ to the right of $1 - p^m$. ∎

Proof of Proposition 6.5

The functions u_1 and u_2 are of class C^1 except at $\hat{p}$, where their first derivatives jump; u_1 is strictly increasing, u_2 strictly decreasing. Moreover, $u_1 = s + \beta_1$ and $u_2 = s + b_2 - c_2$ on $[0, \hat{p}[$, u_1 and u_2 coincide with w_1 and w_2, respectively, at $\hat{p}$, and $u_1 = s + b_1 - c_1$ and $u_2 = s + \beta_2$ on $]\hat{p}, 1]$. So u_1 and u_2 are indeed the players' payoff functions for (k_1, k_2).

As $u_1 > w_1$ and $u_2 > s$ on $[0, \hat{p}[$, we have $b_1 < c_1$ (by Lemma A.1) and $b_2 > c_2$ on this interval. Similarly, as $u_1 > s$ and $u_2 > w_2$ on $]\hat{p}, 1]$, we have $b_1 > c_1$ and $b_2 < c_2$ there. The players' respective actions at $\hat{p}$ are fixed by the continuity requirements imposed by our definition of strategies.

To see that there are no other equilibria, note that, by Lemmas 6.1 and 6.2, there might potentially be two ways of transitioning out of $(0, 1) = (k_1(0), k_2(0))$, namely either via $(0, 1)$—$(1, 1)$—$(1, 0)$, which can only happen at points in the interval $[1 - p^m, p^m]$, or via $(0, 1)$—$(1, 1)$—$(1, 1)$, which can only happen at p^m. Now, suppose that there exists an MPE where players transition from $(0, 1)$ into $(1, 1)$ at p^m. To the right of p^m, players cannot transition back into $(0, 1)$ as, to the right of p^m, there is no way for them to transition out of $(0, 1)$ again. Moreover, they can only transition from $(1, 1)$ to $(1, 0)$ via $(1, 1)$—$(1, 1)$—$(1, 0)$, which can only happen at $1 - p^m < p^m$. Thus, in such an MPE, we must have $(k_1(1), k_2(1)) = (1, 1)$ – a contradiction.

Therefore, in any MPE, there exists a belief $\hat{p} \in [1 - p^m, p^m]$ at which a transition of the form $(0, 1)$—$(1, 1)$—$(1, 0)$ occurs. Now, there is no way for the players to transition out of $(1, 0)$ again to the right of $\hat{p}$, as, by Lemma 6.1, $(1, 0)$—$(1, 0)$—$(1, 1)$ can only occur at $1 - p^m$, which already implies the uniqueness of $\hat{p}$.

Thus, we have shown that there is exactly one transition in any MPE, occurring at a belief $\hat{p} \in [1 - p^m, p^m]$. For the case where $p^m > 1 - p^*$, we shall now show that in fact $\hat{p} \in [p^*, 1 - p^*]$. Indeed, suppose that $\hat{p} < p^*$. Then, $\hat{p}g + \frac{\lambda}{r+\lambda}(1 - \hat{p})s < s$ and, by the explicit expression for player 1's value function, $u_1 < s$ on $]0, \hat{p}[$, which is incompatible with player 1 playing a best response. By an analogous argument, we can rule out $\hat{p} > 1 - p^*$. ∎

References

Bergemann, D. and J. Välimäki (2008): "Bandit Problems," in: *The New Palgrave Dictionary of Economics*, 2nd edition, ed. by S. Durlauf and L. Blume. Basingstoke and New York: Palgrave Macmillan Ltd, forthcoming.

Bergin, J. (1992): "A Model of Strategic Behavior in Repeated Games," *Journal of Mathematical Economics*, 21, 113–153.

Bergin, J. and W.B. MacLeod (1993): "Continuous Time Repeated Games," *International Economic Review*, 34, 21–37.

Bolton, P. and C. Harris (1999): "Strategic Experimentation," *Econometrica*, 67, 349–374.

Bolton, P. and C. Harris (2000): "Strategic Experimentation: the Undiscounted Case," in: *Incentives, Organizations and Public Economics – Papers in Honour of Sir James Mirrlees*, ed. by P.J. Hammond and G.D. Myles. Oxford: Oxford University Press, 53–68.

Camargo, B. (2007): "Good News and Bad News in Two–Armed Bandits," *Journal of Economic Theory*, 135, 558–566.

Chatterjee, K. and R. Evans (2004): "Rivals' Search for Buried Treasure: Competition and Duplication in R&D," *RAND Journal of Economics*, 35, 160–183.

Dewatripont, M. and J. Tirole (1999): "Advocates," *Journal of Political Economy*, 107, 1–39.

Filippov A.F. (1988): *Differential Equations with Discontinuous Righthand Sides*. Dordrecht: Kluwer.

Keller, G. and S. Rady (2007): "Strategic Experimentation with Poisson Bandits," working paper, University of Oxford and University of Munich.

Keller, G., S. Rady and M. Cripps (2005): "Strategic Experimentation with Exponential Bandits," *Econometrica*, 73, 39–68.

Murto, P. and J. Välimäki (2006): "Learning in a Model of Exit," Helsinki Center of Economic Research Working Paper No. 110.

Pastorino, E. (2005): "Essays on Careers in Firms," Ph.D. Dissertation, University of Pennsylvania.

Presman, E.L. (1990): "Poisson Version of the Two-Armed Bandit Problem with Dis-

counting," *Theory of Probability and its Applications*, 35, 307–317.

Rosenberg, D., E. Solan and N. Vieille (2007): "Social Learning in One–Armed Bandit Problems," *Econometrica*, 75, 1591–1611.

Rothschild, M. (1974): "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 9, 185–202.

Simon, L.K. and M.B. Stinchcombe (1989): "Extensive Form Games in Continuous Time: Pure Strategies," *Econometrica*, 57, 1171–1214.