Cripps, Martin W.; Keller, Godfrey; Rady, Sven

Working Paper

Strategic Experimentation: The Case of Poisson Bandits

CESifo Working Paper, No. 737

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Cripps, Martin W.; Keller, Godfrey; Rady, Sven (2002) : Strategic Experimentation: The Case of Poisson Bandits, CESifo Working Paper, No. 737, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/76082

A joint Initiative of Ludwig-Maximilians-Universität and Ifo Institute for Economic Research

Working Papers

STRATEGIC EXPERIMENTATION: THE CASE OF POISSON BANDITS

Martin W. Cripps
Godfrey Keller
Sven Rady*

CESifo Working Paper No. 737 (9)

May 2002

Category 9: Industrial Organisation

Presented at CESifo Area Conference on Industrial Organisation, April 2002

CESifo
Center for Economic Studies & Ifo Institute for Economic Research
Poschingerstr. 5, 81679 Munich, Germany
Phone: +49 (89) 9224-1410 - Fax: +49 (89) 9224-1409
e-mail: office@CESifo.de
ISSN 1617-9595

* Our thanks for very helpful discussions and suggestions are owed to Dirk Bergemann, Antoine Faure-Grimaud, Christian Gollier, Thomas Mariotti, Klaus Schmidt, Jeroen Swinkels and Achim Wambach, and to seminar participants at the IGIER WS in Economic Theory at Erasmus University Rotterdam, the LSE, the Stanford GSB Brown Bag Lunch, the Universities of Oxford, Pennsylvania, Warwick, Wisconsin-Madison, Munich, the 2000 European Summer Symposium in Economic Theory in Gerzensee, the 8th World Congress of the Econometric Society in Seattle, and the DET WS on Learning in Economics at the University of Copenhagen. We would like to thank the Financial Markets Group at the LSE and the Studienzentrum Gerzensee for their hospitality.

CESifo *Working Paper No. 737*

STRATEGIC EXPERIMENTATION: THE CASE OF POISSON BANDITS

Abstract

This paper studies a game of strategic experimentation which the players have access to two-armed bandits where the risky arm distributes lump-sum payoffs according to a Poisson process with unknown intensity. Because of free-riding, there is an inefficiently low level of experimentation in any equilibrium where the players use stationary Markovian strategies. We characterize the unique symmetric Markovian equilibrium of the game, which is in mixed strategies. A variety of asymmetric pure-strategy equilibria is then constructed for the special case where there are two players and the arrival of the first lump-sum fully reveals the quality of the risky arm. Equilibria where players switch finitely often between the roles of experimenter and free-rider all lead to the same pattern of information acquisition; the efficiency of these equilibria depends on the way players share the burden of experimentation among them. We show that at least for relatively pessimistic beliefs, even the worst asymmetric equilibrium is more efficient than the symmetric one. In equilibria where players switch roles infinitely often they can acquire an approximately efficient amount of information, but the rate at which it is acquired still remains inefficient.

JEL Classification: C73, D83, H41, O32.

Keywords: strategic experimentation, two-armed bandit, poisson process, Bayesian learning, Markov perfect equilibrium, public goods.

Martin W. Cripps
Olin School of Business
Washington University
One Brookings Drive
St. Louis MO 63130
U.S.A.

Godfrey Keller
Department of Economics
University of Oxford
Manor R. Building
Oxford OX1 wUQ
United Kingdom

Sven Rady
Department of Economics
University of Munich
Kaulbachstr. 45
80539 Munich
Germany
sven.rady@lrz.uni-muenchen.de

Introduction

When a new restaurant of unknown quality arrives in your neighbourhood you can choose to visit it and risk getting a bad meal yourself; or you can wait until an acquaintance does and then find out about their meal. Furthermore, it may be difficult to determine the quality of the restaurant from one visit alone – it may take many visits to find out whether it is good or bad – so this is a dynamic problem in which the players can perform repeated costly experiments (visit the restaurant) or learn from the experimental observations of others. There are strategic issues in this: first because you can choose to free-ride on the costly information acquisition of your acquaintances (and they can on yours), and second because you can generate information which may encourage others to share the future burden of acquiring information. Such a game of *strategic experimentation* arises in a variety of economic contexts; besides consumer search (as in the restaurant example) or experimental consumption (of a new drug, for instance), firms' research and development activities are a prominent example. Academic researchers pursuing a common research agenda or simply working on a joint paper are also effectively engaged in strategic experimentation.

In the present paper we analyse a game of strategic experimentation based upon two-armed bandits with a safe arm that offers a deterministic flow payoff and a risky arm whose lump-sum payoffs are driven by a Poisson process. This Poisson model is a natural generalization to continuous time of the two-outcome bandit model in Rothschild (1974), the paper that started the economics literature on active Bayesian learning. Nevertheless, Poisson bandits have received little attention so far. Presman (1990) covers the single-agent case. Bergemann and Hege (1998, 2001) study models of financial contracting that embed a Poisson bandit, but the emphasis of their analysis lies on the contractual relationship between a single experimenter and a financier, not on strategic experimentation itself. Malueg and Tsutsui (1997) analyze a model of a patent race with learning where the arrival time of the innovation is exponentially distributed given the stock of knowledge. This leads to the same structure of belief revisions as with Poisson bandits, yet the nature of firms' interaction in their model is entirely different from the situation that we consider. Our motivation for this paper is therefore two-fold: to introduce the notion of a Poisson bandit to a wider audience, and to offer a systematic examination of multi-agent experimentation with such bandits.

With Poisson bandits, news arrives in a 'lumpy' fashion. Examples would be the occasional 'breakthrough' in research and development, failure of an equipment or technology whose reliability is being tested, a completed research paper in a longer-term research agenda, or one of a sequence of crucial proofs in a paper. This should be contrasted with the model of strategic experimentation in Bolton and Harris (1999), which is based upon two-armed bandits where the risky arm yields a flow payoff with Brownian noise. There, both good and bad news arrives continuously, and beliefs are continually adjusted by infinitesimally small increments. The Poisson framework offers an alternative modelling tool for situations where news leads to more drastic revisions of beliefs. For concreteness, we focus on a situation where this news is good. So beliefs jump to a more optimistic level whenever a 'news event' occurs, whereas they gradually

become more pessimistic in between such events.

Another difference between Poisson and Brownian bandits is the greater simplicity of the former. Owing to the technical complexity of the Bolton-Harris model, a complete characterization of its equilibria seems out of reach.[1] In the Poisson model, by contrast, studying the issues associated with strategic experimentation presents much less of a technical challenge. While the Brownian set-up relies on the theory of diffusion processes and associated second-order differential equations (without explicit solutions), the Poisson model involves elementary calculus and first-order differential-difference equations that possess explicit solutions.

As a consequence, we are able to provide a relatively simple and tractable taxonomy of what is possible in our model of strategic experimentation. The results of Bolton and Harris (1999) carry over, among them of course the fundamental inefficiency of information acquisition due to free-riding and, if a single success is not fully revealing, the encouragement effect where one agent's current experimentation may lead to another performing more experimentation in the future (current experiments and future experiments are strategic complements). Exactly as their game, moreover, ours has a unique symmetric Markovian equilibrium, which is in mixed strategies. This is useful of itself because it provides a robustness check of Bolton and Harris's results in a simpler context, and allows a transparent demonstration of the properties and comparative statics of the symmetric equilibrium.

Our main results concern asymmetric (pure-strategy) Markovian equilibria. Here we restrict ourselves to the special case where the arrival of the first lump-sum fully reveals the quality of the risky arm. This case is interesting in its own right. Many examples of strategic experimentation, and especially those involving rare events that carry bad news, will indeed exhibit the feature that a single event is sufficient to determine the optimal decision. Mathematically, the restriction to fully revealing successes simplifies matters in that value functions are (closed-form) solutions to first-order differential equations. For ease of presentation, we focus on the two-player case. All our results generalise to more than two players.

The restriction to fully revealing successes has the important consequence of shutting down the encouragement effect: experimentation at the symmetric equilibrium ceases altogether at the cut-off belief where a single experimenter would stop, and the same holds for any pure-strategy Markovian equilibrium where the players' strategies switch actions a finite number of times only. The reason for this is simple. For the encouragement effect to work, additional experimentation by one player must increase the likelihood that other players will experiment in the future, and this future experimentation must be valuable to the player who acted as a 'volunteer'. To encourage the others, this player needs a success on his risky arm – but in the case of fully revealing successes, he knows everything there is to know from then on, and the additional experimentation by the other players is of no value to him. As we shall see, however, there will be a different sort of encouragement in equilibrium, with players alternating

[1] Bolton and Harris (1999) restrict themselves to studying symmetric equilibria. Park (1999) investigates existence of a particular type of asymmetric equilibrium in the Bolton-Harris model.

between the roles of free-rider and 'lone ranger'.

We show that (at least for relatively pessimistic beliefs) asymmetric (pure-strategy) Markovian equilibria are more efficient than the symmetric equilibrium. The players generate the same *amount* of information at all pure-strategy Markovian equilibria if their strategies switch actions a finite number of times only. This result is driven by backward induction: with finite switching there is a last agent to engage in experimentation and this agent has no incentive to provide more information than would be optimal in a single-agent set-up. Although the amount of information acquired is constant over all pure-strategy Markovian equilibria with finite switching, the *rate* at which the information is acquired does vary. The more equitably the players share the burden of experimentation when it becomes costly (i.e., when ceasing to experiment would yield a higher short-term payoff), the longer they are able to maintain the maximal rate of information acquisition, and the more efficient is the equilibrium. The extreme equilibria where one player bears most of the costs of experimentation are the least efficient.

Casual intuition might lead one to believe that the simplest pure-strategy equilibrium had one player ceasing to experiment when the cost of experimentation became significant and free-riding ever after. In fact no such equilibrium exists. At the simplest pure-strategy equilibrium of the two-player game, one player switches from experimentation to free-riding when beliefs hit a threshold, leaving her opponent to continue experimenting. Then, at a more pessimistic belief threshold, the two players exchange actions – the player who was experimenting free-rides and the player who was free-riding experiments – until all experimentation ceases at the lowest threshold for beliefs. Why do we observe such an equilibrium? In Markovian equilibria the players are not really choosing strategies to affect the amount of information acquired (in aggregate the same amount is always acquired) – but instead they are choosing strategies to adjust the rate at which information is acquired. The last player to experiment is obliged to do this at some cost to herself (and benefit to her opponent). Thus she is not in a hurry to find herself in this role and is willing to delay the time at which this phase of play arrives. Her opponent benefits from this phase of play and so is prepared to experiment in order to accelerate its arrival. Prior to this final phase, therefore, the player who must run the final leg is prepared to defer it by not experimenting herself, whilst the free-rider on the final leg is happy to carry the burden of the experimentation before it. Thus there must be at least two thresholds where actions switch. This simplest equilibrium can be elaborated on by many switches between the role of free-rider and experimenter. We give a complete characterization of when and how this can happen. As the players share the last leg more equally the equilibrium becomes more efficient, because there is less of a temptation for either of the players to free-ride before the last phase.

Our last major result is to show that an approximately efficient amount of information can be acquired in the case of fully revealing successes if we allow the players to use Markovian strategies that switch actions an infinite number of times during a finite time interval. To put this result in perspective, note that in a situation of strategic experimentation with observable actions and outcomes, the players are pro-

viding each other with a public good (information). The provision of this public good is irreversible and ultimately costly (if the experiments are unsuccessful). Recent work on the dynamic provision of public goods has found that efficient provision is possible if the players make smaller and smaller contributions over time and there is no one player who is the last to contribute; see, for example, Admati and Perry (1991), Marx and Matthews (2000), or Lockwood and Thomas (1999). These models use (non-Markovian) trigger strategies to achieve efficiency. If a player deviates from the agreed path of contributions at any point in time, then no other player will make contributions to the public good in the future. Thus the players choose to continue to contribute to the public good because their net gain (of others' future contributions) outweighs their current cost of provision. The absence of a final period is vital here. If there were a last player to provide the public good, she would have no incentive to provide more than the individually rational quantity of the public good and the candidate equilibrium would unravel by backward induction. Although our model is very different – time is continuous rather than discrete – the information transmitted is a natural public good. If there is never a last period of experimentation for any player, each individual can be given an incentive to take turns in providing additional (smaller and smaller) amounts of experimentation. A level of experimentation which is approximately socially efficient can then be induced; the rate at which this information is acquired is, however, socially inefficient. Trigger-strategies are unnecessary here because the beliefs encode the punishment. If a player does not perform an appropriate amount of experimentation, then her opponents' beliefs will not fall sufficiently for them to embark on their round of experimentation, and this hurts the deviating player. In summary, while there is no encouragement effect in the sense of Bolton and Harris (1999) here, players still do encourage each other by taking turns in an incentive-compatible way.

The paper is organised as follows. Section 1 sets up the Poisson bandit model. Section 2 characterizes the optimal strategy for a single player. Section 3 establishes the efficient benchmark where several players coordinate in order to maximize joint expected payoffs. Section 4 introduces the strategic problem and shows that, because of free-riding, any equilibrium of the game leads to inefficiently low levels of experimentation. Section 5 presents the unique symmetric Markov perfect equilibrium, which is in mixed strategies. Section 6 describes pure-strategy, and hence asymmetric, equilibria. Section 7 contains some concluding remarks. Some of the proofs are relegated to the Appendix.

1 Poisson Bandits

The purpose of this section is to introduce continuous-time two-armed bandit problems with Poisson uncertainty. One arm S is 'safe' and yields a known deterministic *flow* payoff whenever it is played; the other arm R is 'risky' and yields a known *lump-sum* reward at random times whenever it is played, the lump-sums arriving according to a Poisson process. The risky arm can be either 'bad' or 'good'. If it is good, the

lump-sums (or 'successes') arrive more frequently than if it is bad.[2] We assume that the agent strictly prefers R, if it is good, to S, and strictly prefers S to R, if it is bad, so she has a motive to experiment with the risky action in the hope of discovering that R is indeed good. The problem she faces, however, is that when she plays R she cannot immediately tell whether it is good or bad, because in either case she initially receives no payoff at all, and the longer she waits without getting a lump-sum, the less optimistic she becomes. Of course, if she eventually receives a lump-sum then she becomes more optimistic again that R is good, but if she waits and waits without the lump-sum arriving then there will come a time when it is optimal for her to cut her losses and switch irrevocably to S.

More formally, time $t \in [0, \infty[$ is continuous, and the discount rate is $r > 0$. The known *flow* payoff of the safe arm is s. The known *lump-sum* payoff of the risky arm is h, the intensity of the Poisson process which determines the arrival of the lump-sums is λ_1 for a good risky arm, and λ_0 for a risky arm, and so the expected payoff from the risky arm is equivalent to a *flow* payoff of $\lambda_1 h$ and $\lambda_0 h$, respectively. We assume that $0 \leq \lambda_0 h < s < \lambda_1 h$.

If an agent plays S over a period of time dt then her payoff is $s\,dt$, and if she plays R over this period then her expected payoff is $\lambda h\,dt$, where $\lambda \in \{\lambda_0, \lambda_1\}$ is unknown. Thus, if k indicates her current choice between S ($k = 0$) and R ($k = 1$), then her expected current payoff (conditional on the unknown state λ of the risky arm) is $[(1-k)s + k\lambda h]\ dt$. Starting with a prior belief p_0, her overall objective is to choose a strategy $\{k_t\}_{t\geq 0}$ that maximises

$$\mathrm{E}\left[\int_0^\infty r\,e^{-r\,t}\ [(1-k_t)s + k_t\lambda h]\ dt \,\middle|\, p_0\right],$$

which expresses the payoff in per-period terms. Of course, this choice of strategy is subject to the constraint that the action taken at any time t be measurable with respect to the information available at that time.

Let p_t denote the subjective probability at time t that the agent assigns to the risky arm being good, so that her current expectation of the flow equivalent of playing R is $\lambda(p_t)h$ with

$$\lambda(p) = p\lambda_1 + (1-p)\lambda_0.$$

By the Law of Iterated Expectations, we can rewrite the above payoff as

$$\mathrm{E}\left[\int_0^\infty r\,e^{-r\,t}\ [(1-k_t)s + k_t\lambda(p_t)h]\ dt \,\middle|\, p_0\right].$$

This highlights the potential for beliefs to serve as a state variable.

Were an agent to act myopically over a period of time dt, she would weigh the short-run payoff from playing S, $rs\,dt$, against what she expects from playing R, $r\lambda(p)h\,dt$. So let us define p^m as the belief that makes her indifferent between these choices,

$$p^m = \frac{s - \lambda_0 h}{\Delta\lambda\, h},$$

[2]Presman (1990) calls this set-up, where one arm is of known quality, the Bellman case.

where $\Delta\lambda = \lambda_1 - \lambda_0$. For $p > p^m$ it is myopically optimal to play R; for $p < p^m$ it is myopically optimal to play S. As we shall see below, a forward-looking agent (who values information) continues to play R for some beliefs $p < p^m$, and is said to *experiment*.

We shall consider the cases where there is a single agent, where there are N agents playing as a team, and where there are N players who act strategically but use only Markovian strategies with the state variable being the belief p.

2 The Single-Agent Problem

When S is played over a period of time dt, the belief does not change. When R is played over a period of time dt, the lump-sum h arrives with probability $\lambda_1\,dt$ if the risky arm is good, and with probability $\lambda_0\,dt$ otherwise.[3] If the agent starts with the belief p, plays R over a period of time dt and does not obtain a reward, then the updated belief at the end of that time period is

$$p + dp = \frac{p\,(1-\lambda_1\,dt)}{p\,(1-\lambda_1\,dt) + (1-p)\,(1-\lambda_0\,dt)}$$

by Bayes' rule. Simplifying, we see that the belief changes by

$$dp = -\Delta\lambda p(1-p)\,dt$$

as long as there is no success. Once a lump-sum arrives, on the other hand, the belief *jumps* up to

$$j(p) = \lambda_1 p/\lambda(p).$$

We now derive the agent's Bellman equation. By the Principle of Optimality, the agent's value function satisfies

$$u(p) = \max_{k\in\{0,1\}} \left\{ r\left[(1-k)s + k\lambda(p)h\right]\,dt + e^{-r\,dt}\mathrm{E}\left[u(p+dp)\mid p\right]\right\}$$

where the first term is the expected current payoff and the second term is the discounted expected continuation payoff.

As to the expected continuation payoff, with subjective probability $k\lambda(p)\,dt$ the lump-sum arrives and the agent expects $u(j(p))$; with probability $1-k\lambda(p)\,dt$ no lump-sum arrives and she expects $u(p)+u'(p)dp = u(p) - k\Delta\lambda p(1-p)u'(p)\,dt$.[4]

Using $1 - r\,dt$ to approximate $e^{-r\,dt}$, we see that her discounted expected continuation payoff is

$$(1-r\,dt)\,\left\{u(p) + k[\lambda(p)\,(u(j(p)) - u(p)) - \Delta\lambda p(1-p)u'(p)]\,dt\right\}$$

[3] This is up to terms of the order $o(dt)$, which we can ignore here and in what follows.

[4] Note that infinitesimal changes of the belief are always downward, so strictly speaking only the left-hand derivative of the value function u matters here. While this turns out to be of no relevance to the single-agent and team cases, we will indeed see equilibria of the strategic experimentation game where a player's payoff function is not of class C^1.

and so her expected total payoff is

$$u(p)+r\left\{(1-k)s+k\lambda(p)h+k[\lambda(p)\left(u(j(p))-u(p)\right)-\Delta\lambda p(1-p)u'(p)]/r-u(p)\right\}dt\,.$$

When this is maximised it equals $u(p)$. Simplifying and rearranging, we thus obtain the Bellman equation

$$u(p)=\max_{k\in\{0,1\}}\left\{(1-k)s+k\lambda(p)h+k[\lambda(p)\left(u(j(p))-u(p)\right)-\Delta\lambda p(1-p)u'(p)]/r\right\}\,.$$

Note that the maximand is linear in k, and the equation can be rewritten more succinctly as

$$u(p)=s+\max_{k\in\{0,1\}}k\left\{b(p,u)-c(p)\right\}\,,$$

where

$$c(p)=s-\lambda(p)h$$

and

$$b(p,u)=[\lambda(p)\left(u(j(p))-u(p)\right)-\Delta\lambda\,p(1-p)u'(p)]/r.$$

Clearly, $c(p)$ is the opportunity cost of playing R; the other term, $b(p,u)$, is the (discounted) expected benefit of playing R, and has two parts: first, $\lambda(p)(u(j(p))-u(p))$ is the expected improvement in the overall payoff should a success occur; second, $-\Delta\lambda p(1-p)u'(p)$ is the negative effect on the overall payoff should no success occur. The agent is indifferent between the two options when cost equals expected benefit, each option resulting in $u(p)=s$. Thus she is effectively unrestricted by the discrete nature of her choice; as usual in single-agent decision problems, there is no scope for randomisation.

So, when it is optimal to play S ($k^*=0$), $u(p)=s$ as one would expect; and when it is optimal to play R ($k^*=1$), u satisfies the first-order differential-difference equation

$$\Delta\lambda\,p(1-p)u'(p)+ru(p)-\lambda(p)[u(j(p))-u(p)]=r\lambda(p)h\,. \tag{1}$$

A particular solution to this equation is $u(p)=\lambda(p)h$, the expected payoff from using the risky arm forever. The option value of being able to switch to the safe arm is then captured by the solution to the homogeneous equation, for which we try $u_0(p)=(1-p)\left(\frac{1-p}{p}\right)^{\mu}$ for some $\mu>0$ to be determined.[5]

Now

$$u_0'(p)=-\frac{\mu+p}{p(1-p)}\,u_0(p),\quad\text{and}\quad u_0(j(p))=\frac{\lambda_0}{\lambda(p)}\left(\frac{\lambda_0}{\lambda_1}\right)^{\mu}u_0(p).$$

Inserting these into the homogeneous equation and simplifying leads to the requirement that

$$r+\lambda_0-\mu\Delta\lambda=\lambda_0\left(\frac{\lambda_0}{\lambda_1}\right)^{\mu}. \tag{2}$$

[5]This guess can be obtained by 'extrapolation' from the limiting case where $\lambda_0=0$. In this case, $j(p)=1$ and $u(j(p))=\lambda_1 h$, so (1) becomes a linear differential equation; the above function u_0 is easily seen to solve the homogeneous equation for $\mu=r/\lambda_1$. A more systematic approach relies on a change of the independent variable from p to $\ln\frac{1-p}{p}$. This transforms (1) into a linear differential-difference equation with constant delay to which results from Bellman and Cooke (1963) can be applied.

As a function of μ, the LHS is a negatively sloped straight line which cuts the vertical axis at $r + \lambda_0$. The RHS is a decreasing exponential function which tends to 0 as $\mu \to +\infty$, tends to ∞ as $\mu \to -\infty$, and cuts the vertical axis at λ_0. Thus the above equation in μ has two solutions, one positive and one negative; we write μ_1 for the positive solution. As the LHS of (2) rises with r, we see that μ_1 is increasing in the discount rate.

The solution to the difference-differential equation for the single-agent case is thus

$$V_1(p) = \lambda(p)h + C\,(1-p)\left(\frac{1-p}{p}\right)^{\mu_1} \tag{3}$$

with C being the constant of integration. Economically relevant are solutions with $C > 0$; these are convex in p.

Proposition 2.1 (Single-agent optimum) *In the single-agent problem, there is a cut-off belief p_1^* given by*

$$p_1^* = \frac{\mu_1(s - \lambda_0 h)}{(\mu_1 + 1)(\lambda_1 h - s) + \mu_1(s - \lambda_0 h)} < p^m \tag{4}$$

such that below the cut-off it is optimal to play S and above it is optimal to play R. The value function V_1^* for the single-agent is given by

$$V_1^*(p) = \lambda(p)h + (s - \lambda(p_1^*)h)\left(\frac{1-p}{1-p_1^*}\right)\left(\frac{1-p}{p}\right)^{\mu_1}\left(\frac{p_1^*}{1-p_1^*}\right)^{\mu_1} \tag{5}$$

when $p > p_1^*$, and $V_1^*(p) = s$ otherwise.

PROOF: The expression for p_1^* and the constant of integration in (5) are obtained by imposing $V_1^*(p_1^*) = s$ (value matching) and $(V_1^*)'(p_1^*) = 0$ (smooth pasting). To verify optimality, note that for any function V_1 of the form (3), at any p such that $V_1(p) = s$, it is the case that $V_1'(p) < 0$ if $p < p_1^*$ and that $V_1'(p) > 0$ if $p > p_1^*$. Now, playing S when $p \in [0, p_1^*[$ gives a payoff of s; playing R on any interval to the left of p_1^* would give a payoff less than s and is therefore sub-optimal. On the other hand, playing R when $p \in]p_1^*, 1]$ gives a payoff greater than s; playing S on any interval to the right of p_1^* would give a payoff of s and is therefore also sub-optimal. ■

The value function for a single agent is illustrated in Figure 1 – it is the lower of the two curves. (The solid kinked line is the expected per-period payoff from the myopic strategy; the upper curve is relevant for the next section.) Note that an individual agent can never be forced to accept a worse payoff, since any player can always act unilaterally.

This solution exhibits all of the familiar properties, which were elegantly described in Rothschild (1974): the optimal strategy has a threshold where the experimenter switches irrevocably from R to S; there are occasions where the experimenter makes a mistake by switching from R to S although the risky action is actually better (R is good); the probability of mistakes decreases as the experimenter becomes more patient, and as the reward from the safe action decreases.

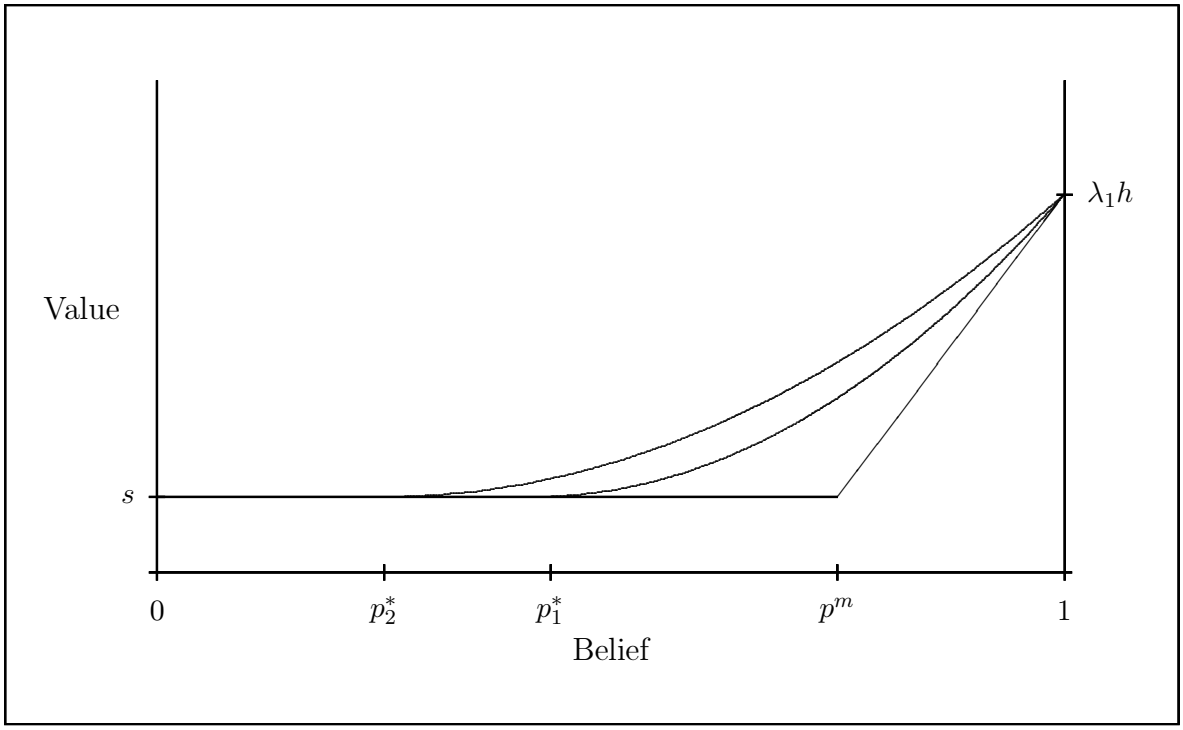

Figure 1: Payoffs for myopic agent, single agent, 2-agent team

3 The N-Agent Team Problem

Now suppose that there are $N \geq 2$ identical agents (same prior belief, same discount rate), each with a replica two-armed bandit (same safe payoff, same lump-sum arriving according to i.i.d. Poisson processes with same parameter), who are working as a team, i.e. who want to maximise the *average* expected payoff. Information is public: the players can observe each other's actions and outcomes, so the players' hold common beliefs throughout time.

If K of them play R over a period of time dt when the risky arm is good, the probability of *none* of them getting a lump-sum is $(1-\lambda_1\,dt)^K = 1-K\lambda_1\,dt$, the probability of *exactly one* of them getting a lump-sum is $K\lambda_1\,dt(1-\lambda_1\,dt)^{K-1} = K\lambda_1\,dt$, and the probability of *more than one* of them getting a lump-sum is negligible.[6] Analogous statements hold in the case of a bad risky arm. If these K players do not obtain a reward, therefore, the belief decays K times as fast as in the single-agent case, $dp = -K\Delta\lambda p(1-p)\,dt$. Once a lump-sum arrives, on the other hand, the belief jumps to the *same* value $j(p)$ as in the single-agent case.

Lemma 3.1 *In the N-agent team problem, it is optimal either for* all *players to play R or for* none *of them to do so.*

[6] Again, we are ignoring terms of order $o(dt)$.

PROOF: Let u be the value function for the team problem, expressed as average payoff per team member. When the current belief is p and the current choice is for K agents to play R, the average expected current payoff is $r\left[(1-\frac{K}{N})s+\frac{K}{N}\lambda(p)h\right]\,dt$. Paralleling the calculation for the single-agent problem, we see that the discounted expected continuation payoff is

$$(1-r\,dt)\,\{u(p)+K[\lambda(p)\,(u(j(p))-u(p))-\Delta\lambda p(1-p)u'(p)]\,dt\}$$

and so the average expected total payoff is

$$u(p)+r\left\{(1-\tfrac{K}{N})s+\tfrac{K}{N}\lambda(p)h+K[\lambda(p)\,(u(j(p))-u(p))-\Delta\lambda p(1-p)u'(p)]/r-u(p)\right\}dt\,.$$

Thus the value function satisfies the Bellman equation

$$u(p)=\max_{K\in\{0,1,\ldots,N\}}\left\{(1-\tfrac{K}{N})s+\tfrac{K}{N}\lambda(p)h+K[\lambda(p)\,(u(j(p))-u(p))-\Delta\lambda p(1-p)u'(p)]/r\right\},$$

or equivalently

$$u(p)=s+\max_{K\in\{0,1,\ldots,N\}}K\,\{N\,b(p,u)-c(p)\}\,/N.$$

Once again, the maximand is linear in K, and the team is indifferent between all levels of K when $c(p)$, the opportunity cost of playing R, equals $N\,b(p,u)$, the expected *social* benefit, each of them resulting in $u(p)=s$. Thus at all beliefs $K^*=N$ or $K^*=0$ is optimal. ■

So, when it is optimal for all players to play S, $u(p)=s$ as usual; and when it is optimal for them all to play R, u satisfies

$$N\Delta\lambda\,p(1-p)u'(p)+ru(p)-N\lambda(p)[u(j(p))-u(p)]=r\lambda(p)h\,, \tag{6}$$

which is like equation (1) with λ_0 and λ_1 replaced by $N\lambda_0$ and $N\lambda_1$, respectively (reflecting an N-times faster rate of information acquisition), and h replaced by h/N (reflecting the fact that lump-sum rewards are shared amongst the N team members). Arguing exactly as in the single-agent case, we see that this has the solution

$$V_N(p)=\lambda(p)h+C\,(1-p)\left(\frac{1-p}{p}\right)^{\mu_N} \tag{7}$$

where μ_N is the unique positive solution of the equation

$$\frac{r}{N}+\lambda_0-\mu\Delta\lambda=\lambda_0\left(\frac{\lambda_0}{\lambda_1}\right)^{\mu}. \tag{8}$$

Proposition 3.1 (Team solution) *In the N-agent team problem, there is a cut-off belief p_N^* given by*

$$p_N^*=\frac{\mu_N(s-\lambda_0 h)}{(\mu_N+1)(\lambda_1 h-s)+\mu_N(s-\lambda_0 h)}<p_1^* \tag{9}$$

such that below the cut-off it is optimal for all to play S and above it is optimal for all to play R. The value function V_N^* for the N-agent team is given by

$$V_N^*(p) = \lambda(p)h + (s - \lambda(p_N^*)h)\left(\frac{1-p}{1-p_N^*}\right)\left(\frac{1-p}{p}\right)^{\mu_N}\left(\frac{p_N^*}{1-p_N^*}\right)^{\mu_N} \tag{10}$$

when $p > p_N^*$, and $V_N^*(p) = s$ otherwise.

The proof proceeds exactly like that of Proposition 2.1 and is therefore omitted.

As the LHS of (8) rises with r and falls with N, we see that μ_N and p_N^* are increasing in r and decreasing in N, and it is straightforward to show that each player's payoff $V_N^*(p)$ increases in N over the range of beliefs where playing the risky arm is optimal. Note that the *average* payoff of the players in *any* N-player problem can never be higher than this, since the team can always replicate their strategies. The value function for either member of a two-agent team is illustrated in Figure 1 – it is the upper of the two curves.

The above proposition determines the *efficient* experimentation strategies for N players acting as a team. We can distinguish two aspects of efficiency here. Given a strategy profile $\{(k_{1,t}, \ldots, k_{N,t})\}_{t\geq 0}$ for the team members, the sum $K_t = \sum_{n=1}^N k_{n,t}$ measures how many risky arms are used at a given time t. We will call this number the *intensity* of experimentation. On the other hand, the integral $\int_0^\infty K_t\, dt$ measures how much the risky arms are used overall. We will call this number the *amount* of experimentation that is performed. The efficient intensity of experimentation exhibits a bang-bang feature, being N when the current belief is above p_N^*, and 0 when it is below. Thus, starting from a prior belief $p_0 > p_N^*$, the efficient intensity is maximal as long as successes occur frequently enough, and minimal after a sufficiently long spell without a success. The efficient amount of experimentation depends on the initial belief, the belief at which all experimentation ceases and the arrival times of rewards on the risky arm.

As we shall see next, equilibria of the N-player *strategic* problem are never efficient.

4 The N-Player Strategic Problem

We continue to assume that the players have the same prior belief, the same discount rate, replica two-armed bandits, and that information is public. We consider stationary Markovian pure strategies with the common belief as the state variable.

Let $k_n \in \{0, 1\}$ indicate the current choice of player n between S ($k_n = 0$) and R ($k_n = 1$); let $K = \sum_{n=1}^N k_n$ and $K_{\neg n} = K - k_n$, so that $K_{\neg n}$ summarises the current choices of the other players. Taking into account the information generated if the other players play R, we see that player n's value function satisfies the Bellman equation

$$\begin{aligned} u_n(p) = \max_{k_n \in \{0,1\}} \{&(1-k_n)s + k_n\lambda(p)h \\ &+ (k_n + K_{\neg n})[\lambda(p)(u_n(j(p)) - u_n(p)) - \Delta\lambda\, p(1-p)u_n'(p)]/r\}\ , \end{aligned}$$

where $u_n'(p)$ should be taken to mean the left-hand derivative of the payoff function (see footnote 4 above). In terms of opportunity cost and expected benefit, the Bellman equation reads

$$u_n(p) = s + K_{\neg n}\, b(p, u_n) + \max_{k_n \in \{0,1\}} k_n \left\{ b(p, u_n) - c(p) \right\} .$$

Immediately we see that the best response, $k_n^*(p)$, is determined by comparing the opportunity cost of playing R with the expected *private* benefit:

$$k_n^*(p) \begin{cases} = 0 & \text{if } c(p) > b(p, u_n), \\ \in \{0,1\} & \text{if } c(p) = b(p, u_n), \\ = 1 & \text{if } c(p) < b(p, u_n). \end{cases} \tag{11}$$

If the best response is to play R ($k_n^* = 1$) then player n's value function u_n satisfies

$$K\Delta\lambda\, p(1-p)u'(p) + ru(p) - K\lambda(p)[u(j(p)) - u(p)] = r\lambda(p)h \tag{12}$$

with $K = K_{\neg n} + 1$.[7] If the best response is to free-ride by playing S ($k_n^* = 0$) then u_n satisfies

$$K\Delta\lambda\, p(1-p)u'(p) + ru(p) - K\lambda(p)[u(j(p)) - u(p)] = rs \tag{13}$$

with $K = K_{\neg n}$. Finally, using the indifference condition from (11) to substitute $c(p)$ for $b(p, u_n)$ in the Bellman equation, we see that for $K_{\neg n} > 0$, player n is indifferent if and only if $u_n(p) = s + K_{\neg n}(s - \lambda(p)h)$. Note that

$$\mathcal{D}_K := \{(p, u) \in [0,1] \times I\!R_+ :\ u = s + K\,(s - \lambda(p)h)\}$$

is a diagonal line in the (p, u)-plane which cuts the safe payoff line $u = s$ at $p = p^m$, the myopic switch-point.

We first show that the incentive to free-ride on the experimentation efforts of the other players makes it impossible to reach efficiency.

Proposition 4.1 (Inefficiency) *All Markov perfect equilibria of the N-player strategic game are inefficient.*

Proof: All we need to show is that the efficient strategies from Proposition 3.1 are *not* an equilibrium. Suppose therefore that players $1, \ldots, N-1$ use the risky arm at beliefs above the cut-off p_N^* and the safe arm below. If player N adopts the same

[7] Note that equation (12) for the strategic problem is the same differential-difference equation as that for the team problem with K players; cf. equation (6). To see why, suppose for example that the risky arm is good. Then, whenever K agents play the risky arm, a lump-sum arrives with probability $K\lambda_1\, dt$ over the next instant. In the K-agent team problem, this lump-sum is shared amongst K players, so the expected lump-sum reward over the next instant is $\frac{h}{K} K\lambda_1\, dt = h\lambda_1\, dt$ per player. In the strategic problem, the lump-sum arrives with probability $\lambda_1\, dt$ on player n's arm and with probability $(K-1)\lambda_1\, dt$ on someone else's arm. Since player n keeps her own lump-sum in full and receives no share of someone else's, her expected lump-sum reward is also $h\lambda_1\, dt$. The same argument applies when the risky arm is bad.

strategy, her payoff function is V_N^*. Now, as p approaches p_N^* from above, $b(p, V_N^*)$ tends to $c(p_N^*)/N$. This means that $b(p, V_N^*) < c(p)$ at beliefs just above p_N^*, so using the risky arm is not optimal for player N there. ■

It is obvious that in any Markov perfect equilibrium, at least one player must be using the risky arm at any belief above p_1^*. The interesting question is whether experimentation continues beyond the single-agent optimum, i.e., whether there is an encouragement effect.

Proposition 4.2 (Encouragement effect) *Assume $\lambda_0 > 0$. Then in any Markov perfect equilibrium where at least two players use the risky arm on an interval of beliefs $[j(p_1^*) - \epsilon, j(p_1^*)]$, at least one player experiments at some beliefs below p_1^*. This is the case in* all *Markov perfect equilibria if $j(p_1^*) \geq p^m$, and in particular if $\lambda_0 \leq r$.*

PROOF: Suppose to the contrary that all players play S at all beliefs below p_1^*. Then each player's payoff function satisfies $u_n(p_1^*) = s$, the left-hand derivative $(u_n)'(p_1^*) = 0$ and $b(p_1^*, u_n) \leq c(p_1^*) = b(p_1^*, V_1^*)$, hence $u_n(j(p_1^*)) \leq V_1(j(p_1^*))$, which must in fact hold as an equality since each player can always guarantee herself $V_1(j(p_1^*))$ at the belief $j(p_1^*)$. But each player who uses R at $j(p_1^*)$ must have $u_n(j(p_1^*)) > V_1^*(j(p_1^*))$ because she benefits from the experimentation of at least one other player. This is the desired contradiction.

Next, if $j(p_1^*) \geq p^m$, *all* players use R at least on the interval $[\hat{p}_1, j(p_1^*)]$ where $\hat{p}_1$ is the belief at which the graph of V_1^* intersects the diagonal $\mathcal{D}_{N-1}$. To see that the inequality $\lambda_0 \leq r$ implies that $j(p_1^*) \geq p^m$, we note that with the notation $\Omega(p) = \frac{1-p}{p}$ for the "odds ratio" corresponding to the belief p, we have

$$\Omega(j(p)) = \frac{\lambda_0}{\lambda_1}\,\Omega(p)$$

and

$$\Omega(p_N^*) = \frac{\mu_N + 1}{\mu_N}\,\Omega(p^m).$$

In particular, $\Omega(j(p_1^*)) \leq \Omega(p^m)$ (that is, $j(p_1^*) \geq p^m$) if and only if $(\mu_1+1)/\mu_1 \leq \lambda_1/\lambda_0$, which in turn is equivalent to

$$\mu_1 \geq \frac{\lambda_0}{\Delta\lambda}.$$

This inequality holds if and only if at $\mu = \lambda_0/\Delta\lambda$, the RHS of (2) does not exceed the LHS. Simple algebra shows that this is the case if and only if

$$\lambda_0 \left(\frac{\lambda_0}{\lambda_1}\right)^{\lambda_0/\Delta\lambda} \leq r.$$

Given r and λ_1, this clearly holds for all λ_0 sufficiently close to zero; as $\lambda_0/\lambda_1 < 1$ and $\lambda_0/\Delta\lambda \geq 0$, in fact, it holds whenever $\lambda_0 \leq r$. ■

So the only possibility for the absence of an encouragement effect when $\lambda_0 > 0$ is a situation where only one player experiments at $j(p_1^*)$. A necessary condition for this is that $j(p_1^*) < p^m$, which requires that λ_0 exceed r and be close to λ_1, so that a success of a 'pioneer' who considers experimenting beyond p_1^* would not make other players sufficiently optimistic to engage in further experimentation themselves. For $\lambda_0 = 0$, on the other hand, all the other players would definitely switch to the risky arm after observing the pioneer's success, but this would not help the pioneer because her continuation value has already jumped to $u_n(1) = \lambda_1 h$, the highest possible level.

In the following two sections we turn to a more detailed investigation of Markov perfect equilibria. We shall consider symmetric mixed-strategy equilibria of the N-player game and asymmetric pure-strategy equilibria of the two-player game.

5 Symmetric Equilibria

Since the efficient strategy profile is symmetric and Markovian with the belief as state variable, it is natural to ask what outcomes can be achieved in symmetric Markovian equilibria of the N-player game. We maintain the assumptions of the previous sections, but allow for mixed strategies now. Following Bolton and Harris (1999), we actually consider the time-division game in which agent n allocates a fraction κ_n of the current period $[t, t+dt[$ to R, and the remainder to S; this is isomorphic to the player using the mixed strategy that places probability κ_n on playing R, and the remainder on S.

So, let $\kappa_n \in [0,1]$ indicate the current decision of player n, $K = \sum_{n=1}^{N} \kappa_n$, and $K_{\neg n} = K - \kappa_n$. Once again taking into account the information generated by the other players, we see that player n's value function satisfies the Bellman equation Player n's value function satisfies the Bellman equation

$$\begin{aligned} u_n(p) = \max_{\kappa_n \in [0,1]} \{ & (1-\kappa_n)s + \kappa_n \lambda(p) h \\ & + (\kappa_n + K_{\neg n})[\lambda(p)(u_n(j(p)) - u_n(p)) - \Delta\lambda\, p(1-p)u_n'(p)]/r \} \ , \end{aligned}$$

or alternatively,

$$u_n(p) = s + K_{\neg n}\, b(p, u_n) + \max_{\kappa_n \in [0,1]} \kappa_n \left\{ b(p, u_n) - c(p) \right\} \ .$$

Again the best response, $\kappa_n^*(p)$, is determined by comparing the opportunity cost of experimentation with the expected benefit:

$$\kappa_n^*(p) \begin{cases} = 0 & \text{if } c(p) > b(p, u_n), \\ \in [0,1] & \text{if } c(p) = b(p, u_n), \\ = 1 & \text{if } c(p) < b(p, u_n). \end{cases}$$

In any Markov perfect equilibrium player n's value function will be defined piecewise: when all the time is devoted to S it satisfies equation (13) with $K = K_{\neg n}$; when

all the time is devoted to R it satisfies equation (12) with $K = K_{\neg n} + 1$; and when the time is divided strictly between S and R it satisfies

$$\Delta\lambda\, p(1-p)u'(p) - \lambda(p)[u(j(p)) - u(p)] = r\lambda(p)h - rs. \tag{14}$$

In a *symmetric* equilibrium, the region where all players use the risky arm all the time is separated from the region of strict mixing by the diagonal

$$\mathcal{D}_{N-1} := \{(p,u) \in [0,1] \times I\!R_+ :\ u = s + (N-1)\,(s - \lambda(p)h)\}.$$

Given the post-jump value $u(j(p))$, we have smooth pasting of the solutions to (6) and (14) along this diagonal. To the left of the diagonal, the Bellman equation implies that the players' common strategy $\kappa : [0,1] \to [0,1]$ is given by

$$\kappa(p) = \frac{1}{N-1}\frac{u(p) - s}{s - \lambda(p)h}$$

where u is the common payoff function. As this payoff function is continuous, so is κ.

Smooth pasting of the payoff function u occurs not only along the diagonal $\mathcal{D}_{N-1}$ but also at the belief where this payoff reaches the level s. In other words, u must be of class C^1. To see this, suppose we had a symmetric equilibrium with a payoff function that hits the level s at the belief $\tilde{p}$ with slope $u'(\tilde{p}+) > 0$. Then we would have $b(p,u) = c(p)$ or

$$\lambda(p)[u(j(p)) - u(p)]/r = c(p) + \Delta\lambda p(1-p)u'(p)/r$$

at beliefs immediately to the right of $\tilde{p}$, implying

$$\lambda(\tilde{p})[u(j(\tilde{p})) - s]/r = c(\tilde{p}) + \Delta\lambda\tilde{p}(1-\tilde{p})u'(\tilde{p}+)/r > c(\tilde{p})$$

by continuity. Immediately to the left of $\tilde{p}$, continuity of $u(j(p))$ and the fact that $u'(p) = 0$ would then imply $b(p,u) = \lambda(p)[u(j(p)) - s]/r > c(p)$, so there would be an incentive to deviate from S to R.

Our next result describes the unique symmetric Markov perfect equilibrium of the strategic experimentation game. To prove existence of a symmetric equilibrium, we first construct a family of candidate payoff functions, that is, solutions to the differential-difference equation

$$\Delta\lambda\, p(1-p)u'(p) - \lambda(p)[u(j(p)) - u(p)] = r\min\left\{\lambda(p)h - s, \frac{\lambda(p)h - u(p)}{N}\right\}, \tag{15}$$

which combines (14) and (6). We then show that there is at least one such solution with zero slope at the belief where it assumes the value s. In a last step, we establish uniqueness.

Proposition 5.1 (Symmetric equilibrium) *The dynamic experimentation game admits a unique symmetric Markov perfect equilibrium, which is necessarily in mixed*

strategies. The corresponding payoff function is the unique function $W_N^* : [0,1] \to [s, \lambda_1 h]$ *of class* C^1 *with the following properties:* $W_N^*(p) = s$ *on an interval* $[0, \tilde{p}_N]$ *with* $0 < \tilde{p}_N < 1$*;* $W_N^*(p) > s$ *on* $]\tilde{p}_N, 1]$*; and* W_N^* *solves the differential-difference equation* (15) *on* $]\tilde{p}_N, 1[$*. The cut-off belief* $\tilde{p}_N$ *satisfies* $p_N^* < \tilde{p}_N < p_1^*$ *if* $\lambda_0 > 0$*, and* $\tilde{p}_N = p_1^*$ *if* $\lambda_0 = 0$*. The equilibrium strategy is given by*

$$\kappa^*(p) = \min\left\{\frac{1}{N-1}\frac{W_N^*(p) - s}{s - \lambda(p)h}, 1\right\},$$

and there is a second cut-off $\hat{p}_N$ *with* $\tilde{p}_N < \hat{p}_N < 1$ *such that* $0 < \kappa^*(p) < 1$ *precisely when* $\tilde{p}_N < p < \hat{p}_N$*.*

PROOF: A solution u to (15) is entirely determined by its point of intersection $(\bar{p}, u(\bar{p}))$ with the diagonal $\mathcal{D}_{N-1}$. To the right of $\mathcal{D}_{N-1}$, we know already that $u = u^{(0)}$ where

$$u^{(0)}(p) = V_N(p) = \lambda(p)h + C\,(1-p)\left(\frac{1-p}{p}\right)^{\mu_N}$$

for some constant C.

We can now rewrite (14) as an ordinary differential equation on the interval $[j^{-1}(\bar{p}), \bar{p}]$:

$$\Delta\lambda\, p(1-p)u'(p) + \lambda(p)u(p) = r\lambda(p)h - rs + \lambda(p)u^{(0)}(j(p)). \tag{16}$$

Standard results imply that this ODE has a unique solution for any initial condition; in particular, there is a unique solution $u^{(1)}$ on $[j^{-1}(\bar{p}), \bar{p}]$ such that $u^{(1)}(\bar{p}) = u^{(0)}(\bar{p})$ and, by construction, $u^{(1)\prime}(\bar{p}) = u^{(0)\prime}(\bar{p})$.

Iterating this step, we construct functions $u^{(i)}$ defined on $[j^{-i}(\bar{p}), j^{-(i-1)}(\bar{p})]$ for $i = 2, 3, \ldots$ by choosing $u^{(i)}$ as the unique solution of the ODE

$$\Delta\lambda\, p(1-p)u'(p) + \lambda(p)u(p) = r\lambda(p)h - rs + \lambda(p)u^{(i-1)}(j(p)) \tag{17}$$

subject to the condition $u^{(i)}(j^{-(i-1)}(\bar{p})) = u^{(i-1)}(j^{-(i-1)}(\bar{p}))$. Setting $u(p) = u^{(i)}(p)$ whenever $j^{-i}(\bar{p}) \le p < j^{-(i-1)}(\bar{p})$, we thus obtain a function u of class C^1 on $]0,1]$ that solves (14) to the left of $\bar{p}$, and (6) to the right of $\bar{p}$.

Standard results imply further that this function depends in a continuous fashion on $\bar{p}$, i.e. on the point of intersection with the diagonal $\mathcal{D}_{N-1}$. In particular, $M(\bar{p})$, the minimum of this function on the interval $[p_N^*, p^m]$, is continuous in $\bar{p}$. Let $\bar{p}_N$ denote the belief where the graph of V_N^* cuts $\mathcal{D}_{N-1}$, and $\bar{p}_{1,N}$ denote the belief where the graph of V_1^* cuts $\mathcal{D}_{N-1}$. We want to show that there exists a $\hat{p}$ between $\bar{p}_N$ and $\bar{p}_{1,N}$ such that $M(\hat{p}) = s$. With $\hat{u}$ denoting the function corresponding to $\hat{p}$, let $\tilde{p}$ be the highest belief where $\hat{u}$ achieves this minimum. We want to show further that $\tilde{p}$ is strictly between p_N^* and p_1^*.

Consider a solution u to (15) which is (strictly) above V_N^* for some belief $p_r \in]p_N^*, 1]$. If u and V_N^* have the same value at some belief $p_\ell \in [p_N^*, p_r[$, then $u - V_N^*$

has a strictly positive maximum at some belief $p \in]p_\ell, 1]$. As $u'(p) = (V_N^*)'(p)$ and $u(j(p)) - V_N^*(j(p)) \leq u(p) - V_N^*(p)$, (6) and (15) imply

$$\lambda(p)h - V_N^*(p) \leq N \min\left\{\lambda(p)h - s, \frac{\lambda(p)h - u(p)}{N}\right\}.$$

So $\lambda(p)h - V_N^*(p) \leq \lambda(p)h - u(p)$ or $u(p) \leq V_N^*(p)$, which is a contradiction. Consequently, u lies strictly above V_N^* on $[p_N^*, p_r]$, and this implies that $M(\bar{p}) > s$ for $\bar{p} < \bar{p}_N$.

Next, consider a solution u to (15) which is (strictly) below V_1^* for some belief $p_r \in]p_1^*, 1]$. If u and V_1^* have the same value at some belief $p_\ell \in [p_1^*, p_r[$, then $V_1^* - u$ has a strictly positive maximum at some belief $p \in]p_\ell, 1]$. As $(V_1^*)'(p) = u'(p)$ and $V_1^*(j(p)) - u(j(p)) \leq V_1^*(p) - u(p)$, (1) and (15) imply

$$\lambda(p)h - V_1^*(p) \geq \min\left\{\lambda(p)h - s, \frac{\lambda(p)h - u(p)}{N}\right\}.$$

As $V_1^*(p) > s$, the minimum on the RHS must be $[\lambda(p)h - u(p)]/N$. But then $NV_1^*(p) \leq (N-1)\lambda(p)h + u(p) < (N-1)\lambda(p)h + V_1^*(p)$ or $V_1^*(p) < \lambda(p)h$, which is a contradiction. Consequently, u lies strictly below V_1^* on $[p_1^*, p_r]$, and this implies that $M(\bar{p}) < s$ for $\bar{p} > \bar{p}_{1,N}$.

Continuity of M together with the two arguments above imply that $M(\bar{p}_N) \geq s$ and $M(\bar{p}_{1,N}) \leq s$, and so there exists a $\hat{p}$ between $\bar{p}_N$ and $\bar{p}_{1,N}$ such that $M(\hat{p}) = s$. Recall that $\hat{u}$ denotes the function corresponding to $\hat{p}$, and $\tilde{p}$ is the highest belief where $\hat{u}(p) = s$. The first argument above implies that $\hat{u} \leq V_N^*$ and so $\tilde{p} \geq p_N^*$, while the second argument above implies that $\hat{u} \geq V_1^*$ and so $\tilde{p} \leq p_1^*$; also, since $(V_N^*)'(p_N^*) = 0$, we see that $u'(\tilde{p}) = 0$.

Note that $\hat{u}$ is the players' common payoff function if they all use the strategy

$$\kappa(p) = \begin{cases} 0 & \text{if} \quad p \leq \tilde{p}, \\ \frac{1}{N-1}\frac{\hat{u}(p)-s}{s-\lambda(p)h} & \text{if} \quad \tilde{p} < p \leq \hat{p}, \\ 1 & \text{if} \quad p > \hat{p}; \end{cases}$$

as $\hat{u} \leq V_N^*$ it stays below $\mathcal{D}_{N-1}$ on $[\tilde{p}, \hat{p}[$ and is indeed a solution to (15). We thus have shown existence of a symmetric equilibrium.

We want to show that the inequalities in $p_N^* \leq \tilde{p} \leq p_1^*$ are strict.

If $\tilde{p} = p_N^*$, then $\hat{u}(\tilde{p}) = s = V_N^*(p_N^*)$ and $\hat{u}'(\tilde{p}) = 0 = (V_N^*)'(p_N^*)$, and now (14) and (6) imply

$$\lambda(p_N^*)[\hat{u}(j(p_N^*)) - s] = rs - r\lambda(p_N^*)h = N\lambda(p_N^*)[V_N^*(j(p_N^*)) - s]$$

and hence $\hat{u}(j(p_N^*)) - s = N[V_N^*(j(p_N^*)) - s]$. So $\hat{u}(j(p_N^*)) = V_N^*(j(p_N^*)) + (N-1)[V_N^*(j(p_N^*)) - s] > V_N^*(j(p_N^*))$, which is a contradiction.

If $\tilde{p} = p_1^*$, then $\hat{u}(\tilde{p}) = s = V_1^*(p_1^*)$ and $\hat{u}'(\tilde{p}) = 0 = (V_1^*)'(p_1^*)$, and now (14) and (1) imply

$$\lambda(p_1^*)[\hat{u}(j(p_1^*)) - s] = rs - r\lambda(p_1^*)h = \lambda(p_1^*)[V_1^*(j(p_1^*)) - s]$$

and hence $\hat{u}(j(p_1^*)) = V_1^*(j(p_1^*))$. So $V_1^* - u$ attains its maximum of 0 at $j(p_1^*)$ as well as at p_1^*. A variant of the second argument above (with $j(p_1^*)$ replacing p and "maximum of 0" replacing "strictly positive maximum") leads to the contradiction $V_1^*(j(p_1^*)) \leq \lambda(j(p_1^*))h$.

When $\lambda_0 = 0$, on the other hand, the differential-difference equation (14) for strict mixing simplifies to the differential equation

$$\lambda_1 p(1-p)u'(p) + \lambda_1 p u(p) = (r+\lambda_1)\lambda_1 ph - rs,$$

while the single-agent solution V_1^* solves

$$\lambda_1 p(1-p)(V_1^*)'(p) + (r+\lambda_1 p)V_1^*(p) = (r+\lambda_1)\lambda_1 ph.$$

As $u(\tilde{p}) = V_1^*(p_1^*) = s$ and $u'(\tilde{p}) = (V_1^*)'(p_1^*) = 0$, we see immediately from these two ODEs that $\tilde{p} = p_1^*$.

Finally, we want to show uniqueness of the symmetric MPE. Suppose therefore that we have two symmetric equilibria with different payoff functions u and $\hat{u}$, respectively. Without loss of generality, let $u - \hat{u}$ assume a strictly positive global maximum at the belief p. Here, $u'(p) = \hat{u}'(p)$ and $u(j(p)) - \hat{u}(j(p)) \leq u(p) - \hat{u}(p)$, so $b(p,u) \leq b(p,\hat{u})$. We cannot have both $u(p)$ and $\hat{u}(p)$ above $\mathcal{D}_{N-1}$ since in this region both u and $\hat{u}$ are of the form V_N and the difference $u - \hat{u}$ is strictly decreasing to the right of $\mathcal{D}_{N-1}$. Further, if $u(p)$ is above $\mathcal{D}_{N-1}$ and $\hat{u}(p)$ is on or below, then $b(p,u) > c(p) = b(p,\hat{u})$ in contradiction to what we derived before. Consequently, we must have both $u(p)$ and $\hat{u}(p)$ on or below $\mathcal{D}_{N-1}$, so $b(p,u) = c(p) = b(p,\hat{u})$. This in turn yields $u(j(p)) - \hat{u}(j(p)) = u(p) - \hat{u}(p)$, so the difference $u - \hat{u}$ is also at its maximum at the belief $j(p)$. Iterating the argument until we get to the right of p^m (and hence to the right of $\mathcal{D}_{N-1}$), we obtain the desired contradiction. This establishes the existence of a unique symmetric equilibrium. ■

The symmetric equilibrium of the Poisson model shares the main features with its counterpart in the Brownian model of Bolton and Harris (1999). First, it clearly shows the fundamental inefficiency of information acquisition due to free-riding. In fact, not only is the amount of experimentation inefficiently low (as can be seen from the lower cut-off $\tilde{p}_N$ being above the team cut-off p_N^*) and the intensity of experimentation inefficiently low (at any belief between p_N^* and $\hat{p}_N$ there is strictly too little use of risky arms), but the acquisition of information is slowed down so severely near the cut-off $\tilde{p}_N$, that the equilibrium amount of experimentation is never performed in finite time – as the following result shows, the players never actually stop allocating at least some of their time to playing the risky arm.[8]

Corollary 5.1 *Starting from a prior belief above the equilibrium cut-off $\tilde{p}_N$, the players' common posterior belief never reaches this cut-off in the symmetric Markov perfect equilibrium.*

[8]To some readers, this phenomenon might be familiar from the production of joint research papers. Once the initial enthusiasm has waned, each co-author might spend less and less time working on the paper, without actually withdrawing completely. And the paper might never be put out of its misery.

PROOF: Close to the right of $\tilde{p}_N$, the dynamics of the belief p given no success are

$$dp = -\Delta\lambda \, \frac{N}{N-1} \, \frac{W_N^*(p) - s}{s - \lambda(p)h} \, p\,(1-p)\, dt.$$

(A success merely causes a delay before the belief decays to near $\tilde{p}_N$ again; when $\lambda_0 = 0$, this 'delay' is itself infinite.) As W_N^* is C^2 to the right of $\tilde{p}_N$ with $W_N^*(\tilde{p}_N) = s$, $(W_N^*)'(\tilde{p}_N) = 0$ and $(W_N^*)''(\tilde{p}_N+) > 0$, we can find a positive constant c such that

$$\Delta\lambda \, \frac{N}{N-1} \, \frac{W_N^*(p) - s}{s - \lambda(p)h} p\,(1-p) < c\,(p - \tilde{p}_N)^2$$

in a neighbourhood of $\tilde{p}_N$.

Starting from an initial belief $p_0 > \tilde{p}_N$ in this neighbourhood, consider the dynamics

$$dp = -c\,(p - p_1^*)^2\, dt.$$

The solution of these dynamics with initial value p_0 is

$$p_t = \tilde{p}_N + \frac{1}{ct + (p_0 - \tilde{p}_N)^{-1}}.$$

Obviously, this solution does not reach $\tilde{p}_N$ in finite time. Since the modified dynamics have a faster rate of decrease as the original ones, this result carries over to the true evolution of beliefs. ■

A second feature that the symmetric equilibrium shares with that in Bolton and Harris (1999) is the encouragement effect whereby one agent's current experimentation leads to another performing more experimentation in the future. This effect manifests itself in the fact that the cut-off belief $\tilde{p}_N$ where all experimentation stops for good in the symmetric equilibrium is lower than the corresponding single-agent cut-off p_1^*.

Third, the comparative statics with respect to the number of players also play out as in the Brownian set-up. As N increases, the lower cut-off $\tilde{p}_N$ falls, the upper cut-off $\hat{p}_N$ rises, and each player's obtains a higher payoff at all beliefs where the risky arm is used some of the time.

What differentiates the Poisson model from Bolton and Harris (1999) is that the above results can be obtained by elementary methods and constructively. In fact, we can represent the payoff function W_N^* in closed form up to some constants of integration that are implicitly determined by the cut-off $\hat{p}_N$. We use the notation $\Omega(p) = \frac{1-p}{p}$ for the "odds ratio" corresponding to the belief p.

Corollary 5.2 *Define intervals I_i for $i = 0, 1, \ldots$ recursively by setting $I_0 = [\hat{p}, 1]$ and $I_{i+1} = j^{-1}(I_i)$. If $\mu_N \neq \lambda_0/\Delta\lambda$, then*

$$W_N^*(p) \;=\; \left(\lambda_1 h + \frac{r(\lambda_1 h - s)}{\lambda_1}\, i\right) p + \left(\lambda_0 h + \frac{r(\lambda_0 h - s)}{\lambda_0}\, i\right)(1-p)$$

$$+ C^{(0)} \left(\frac{\lambda_0 \left(\lambda_0/\lambda_1\right)^{\mu_N}}{\lambda_0 - \mu_N \Delta\lambda} \right)^i (1-p)\Omega(p)^{\mu_N}$$
$$+ \sum_{n=0}^{i-1} \frac{C^{(i-n)}}{n!} \left(- \frac{\lambda_0 \left(\lambda_0/\lambda_1\right)^{\lambda_0/\Delta\lambda}}{\Delta\lambda} \ln\left[\left(\lambda_0/\lambda_1\right)^{n-1} \Omega(p) \right] \right)^n (1-p)\Omega(p)^{\lambda_0/\Delta\lambda}$$

on $I_i \cap \{p : W_N^*(p) > s\}$ *for some constants* $C^{(i-n)}$ $(n = 0, \ldots i-1)$*, chosen to ensure continuity of* W_N^* *and, by construction, of* $(W_N^*)'$.[9] *The constant* $C^{(0)}$ *that fixes payoffs above* $\mathcal{D}_{N-1}$ *is given by*

$$C^{(0)} = N(s - \lambda_0 h) \left[1 - \frac{\Omega(p^m)}{\Omega(\hat{p})} \right] \Omega(\hat{p})^{-\mu_N}.$$

PROOF: See the Appendix. ■

Proposition 5.1 implies that there is no symmetric MPE in pure strategies. In fact, any candidate for such an equilibrium unravels because of free-riding at lower beliefs. What sort of behaviour can arise in a pure-strategy MPE will be addressed next.

6 Pure-Strategy Equilibria

From now on, we restrict our attention to the special case where a single success reveals the risky arm to be good, i.e. we assume that $\lambda_0 = 0$. This simplifies the construction of equilibria considerably since payoff functions are now characterised by linear first-order differential equations – the post-jump term $u(j(p))$ in equations (12) and (13) is replaced with the constant $\lambda_1 h$.

Accordingly, we simply write λ for λ_1, and then $\Delta\lambda$ reduces to λ and $\lambda(p)$ becomes λp. Thus, if $K_{\neg n}$ other players are using the risky arm and player n's best response is to play R $(k_n^* = 1)$ then her value function u_n satisfies

$$K\lambda p(1-p)u'(p) + (r + K\lambda p)u(p) = (r + K\lambda)\lambda h p \tag{18}$$

with $K = K_{\neg n} + 1$; if the best response is to free-ride by playing S $(k_n^* = 0)$ then u_n satisfies

$$K\lambda p(1-p)u'(p) + (r + K\lambda p)u(p) = rs + K\lambda^2 h p \tag{19}$$

with $K = K_{\neg n}$. Both these ODEs have simple closed-form solutions. The solution to (18) is

$$V_K(p) = \lambda h p + C\,(1-p) \left(\frac{1-p}{p} \right)^{r/K\lambda}, \tag{20}$$

whereas that to (19) is

$$F_K(p) = s + \frac{K\lambda(\lambda h - s)}{r + K\lambda} p + C\,(1-p) \left(\frac{1-p}{p} \right)^{r/K\lambda}. \tag{21}$$

[9] The proof makes it obvious how one has to modify this result in the knife-edge case where $\mu_N = \lambda_0/\Delta\lambda$.

Using these solutions, we will construct two types of asymmetric equilibrium in pure strategies. The first type of MPE consists of strategies where the action of each player switches at finitely many beliefs. As a consequence, there is a last point in time at which any player is willing to experiment. The belief at which this happens (provided no success has been observed) will be the single-player cut-off p_1^*, exactly as in the symmetric MPE with $\lambda_0 = 0$. So a similar inefficiency arises: both the amount and the intensity of experimentation are too low. Nevertheless, these equilibria differ in terms of the time taken to reach the belief where experimentation ceases, and also in terms of aggregate payoffs.

In the second type of MPE, each player's strategy has infinitely many switching points, and although there is a finite time after which no player ever experiments again, no single player has a *last* time for experimentation. That is, immediately prior to reaching a certain cut-off belief, the players switch roles increasingly fast, and infinitely often. We will see that we can take this cut-off belief arbitrarily closely to the efficient cut-off. Still, the equilibrium is inefficient: although an almost efficient amount of experimentation is performed, it is performed with an inefficient intensity.

For ease of exposition, we restrict ourselves to the two-player case from now on. Extending our results to asymmetric equilibria with more than two players poses no conceptual difficulties, but increases the notational burden significantly.[10]

The following result analyses each player's best-response correspondence over the relevant range of pairs of beliefs and continuation payoffs.

Lemma 6.1 *Consider a belief p and a continuation payoff $u \geq V_1^*(p)$ for player i at that belief. Fix an action of player j for all beliefs in an interval $]p, p']$ with $p' > p$. If (p, u) lies on or to the right of the diagonal $\mathcal{D}_1$, then R is the dominant action for player i at all beliefs in $]p, p']$. If (p, u) lies to the left of the diagonal $\mathcal{D}_1$ and $u > s$, then there is an interval $]p, p + \epsilon] \subset]p, p']$ where player i's best response is to play the opposite action to player j's. If $u = s$ and $p < p_1^*$, then S is the dominant action for player i at all beliefs in $]p, p'] \cap]p, p_1^*[$.*

PROOF: See the Appendix. ■

This result is illustrated in Figure 2 where the solid kinked line is the payoff from the myopic strategy, and the solid curve the payoff from the single-agent optimal strategy. From this picture, we can see that a Markov perfect equilibrium has three phases. When the players are optimistic, both play R; when they are pessimistic, both play S; in between, one of them free-rides by playing S while the other is playing R. We shall see that this mid-range of beliefs further splits into two regions: the roles of free-rider and 'lone ranger' are assigned for the whole of the upper region; in the lower region, players can swap roles.

The next proposition first describes the 'simplest' such equilibrium, in which one particular player experiments and the other free-rides throughout the lower region, and

[10] Extending our results to the case $\lambda_0 \neq 0$ is more involved; this is the goal of current research.

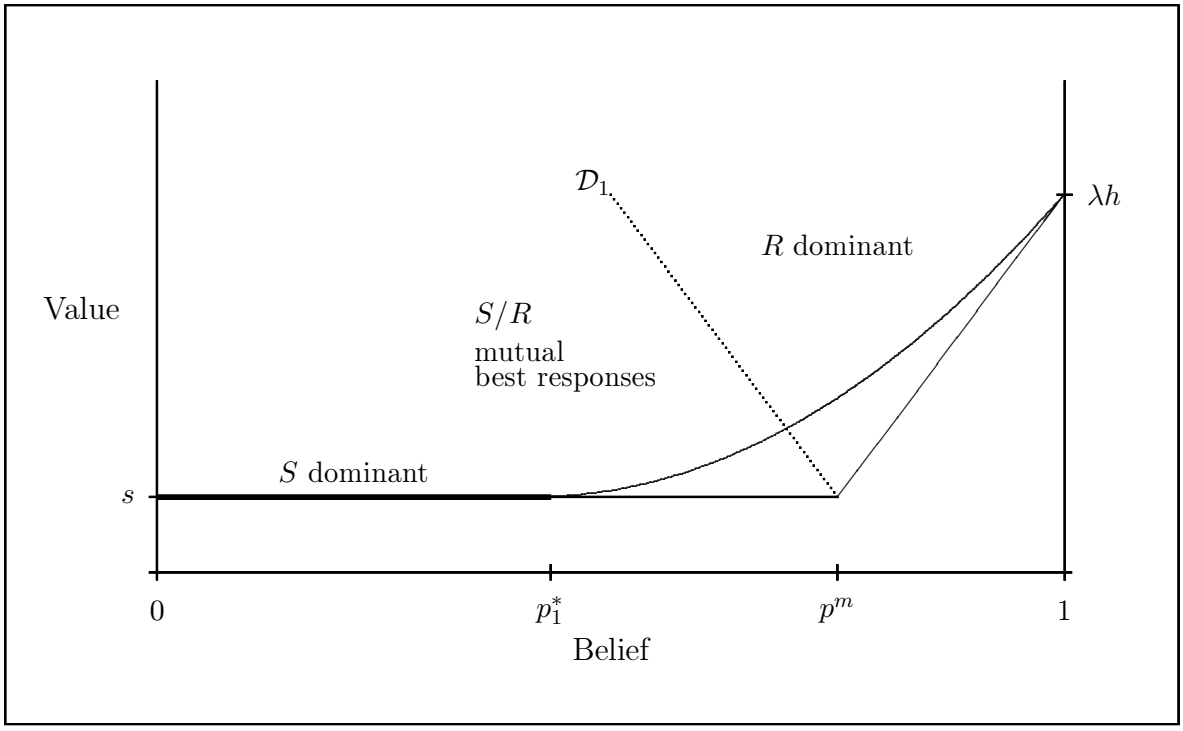

Figure 2: Best responses in the two-player case

then characterizes all pure-strategy MPE where players' actions switch at finitely many beliefs. We use the notation $\Omega(p) = \frac{1-p}{p}$ again.

Proposition 6.1 (Two players, pure strategies, finite number of switches) *In the two-player strategic experimentation problem, there is a pure-strategy Markov perfect equilibrium where the players' actions depend as follows on the common posterior belief. There are three cut-off beliefs $p_1^* < \tilde{p}_\ell < \tilde{p}_r$ such that: on $]\tilde{p}_r, 1]$, both players play R; on $]\tilde{p}_\ell, \tilde{p}_r]$, player 1 plays R and player 2 plays S; on $]p_1^*, \tilde{p}_\ell]$, player 1 plays S and player 2 plays R; on $[0, p_1^*]$, they both play S. The low cut-off, p_1^*, is given in Proposition 2.1; the other two are given by the solution to*

$$\left(\frac{\Omega(\tilde{p}_\ell)}{\Omega(p_1^*)}\right)^{r/\lambda+1} + \frac{r+\lambda}{\lambda}\left[\frac{\Omega(\tilde{p}_\ell)}{\Omega(p^m)} - 1\right] - 1 = 0$$

and the solution to

$$\left\{\frac{(r+\lambda)(2r+\lambda)}{r\lambda}\frac{\Omega(\tilde{p}_\ell)}{\Omega(p^m)} - \frac{r^2+(r+\lambda)(r+2\lambda)}{r\lambda}\right\}\left(\frac{\Omega(\tilde{p}_r)}{\Omega(\tilde{p}_\ell)}\right)^{r/\lambda+1} + \frac{r+\lambda}{\lambda}\left[\frac{\Omega(\tilde{p}_r)}{\Omega(p^m)} - 1\right] - 1 = 0.$$

Moreover, in any pure-strategy MPE with finitely many switching points there are three cut-off beliefs $p_1^* < \bar{p}_\ell \leq \bar{p}_r$, with $\tilde{p}_\ell \leq \bar{p}_\ell$ and $\bar{p}_r \leq \tilde{p}_r$, such that: on $]\bar{p}_r, 1]$, both players play R; throughout $]\bar{p}_\ell, \bar{p}_r]$, one player plays R and the other plays S; on $]p_1^*, \bar{p}_\ell]$, the players share the burden of experimentation by taking turns; on $[0, p_1^*]$, they both play S.

PROOF: Here we just sketch the proof; for details, see the Appendix.

We first note that there must be a last player to experiment since the level $u = s$ can only be reached via the part of the (p, u)-plane where R and S are mutual best responses. This player, say player 2, will necessarily stop experimenting at the single-agent cut-off belief p_1^*.

We can now work backwards (in time) from (p_1^*, s). On an interval to the right of p_1^*, player 2 plays R and his continuation value (as a function of the belief) is a slowly rising convex function. On this interval, player 1 free-rides by playing S and her continuation value is a steeply rising concave function. Thus, at some belief, player 1's value meets $\mathcal{D}_1$ while player 2's value is still below it – this defines $\tilde{p}_\ell$. On an interval to the right of $\tilde{p}_\ell$, player 1 is content to 'go it alone' and play R, while player 2 responds by free-riding with S. At some belief, player 2's value meets $\mathcal{D}_1$ while player 1's value is yet further above it – this defines $\tilde{p}_r$. On the interval to the right of $\tilde{p}_r$, both players optimally play R.

As to other equilibria of this sort, we again work backwards from (p_1^*, s). If the players swap roles (at least once) before the value of either of them has met $\mathcal{D}_1$, then the one with the higher value will be below that of player 1 in the 'simplest' equilibrium sketched above, and the one with the lower value will be above that of player 2. At some belief, the value of one of the players meets $\mathcal{D}_1$ while the other's value is still (weakly) below it – this defines $\bar{p}_\ell > \tilde{p}_\ell$. The one with the higher value plays R to the right of $\bar{p}_\ell$, while the other one free-rides until the value meets $\mathcal{D}_1$ – this defines $\bar{p}_r < \tilde{p}_r$ – and then joins in by playing R. ■

The value functions of the two players in the 'simplest' equilibrium (with cut-offs p_1^*, $\tilde{p}_\ell$ and $\tilde{p}_r$) are illustrated in Figure 3. The faint straight line is $\mathcal{D}_1$. Observe that the lower payoff meets this line at $\tilde{p}_r$ while the higher payoff meets it at $\tilde{p}_\ell$.

Note that with finitely many beliefs at which a player changes his action, the threshold belief at which all experimentation stops is again the single-agent cut-off p_1^*; in particular, it is the same for all equilibria of this type (and thus they all exhibit the same amount of experimentation, whereas the higher threshold beliefs are determined endogenously by how the burden of experimentation is shared at beliefs to the right of p_1^* (and hence the intensity of experimentation will vary across these equilibria).

The 'simplest' equilibrium of Proposition 6.1 is also the 'worst' from an efficiency perspective. This is because it gives the player who experiments last the lowest possible payoff function, which in turn implies that the part of the state space where both players experiment is smallest – the threshold belief at which the intensity of experimentation drops from 2 to 1 (that is, the belief at which the lower payoff function crosses $\mathcal{D}_1$) is as high as it can be, namely equal to $\tilde{p}_r$. The 'simplest' equilibrium therefore exhibits the *slowest* experimentation. In an MPE where the threshold belief $\bar{p}_r$ is lower, the maximal intensity of experimentation is maintained for longer, so the same overall amount of information is acquired faster. As the following proposition shows, such an equilbrium is more efficient.

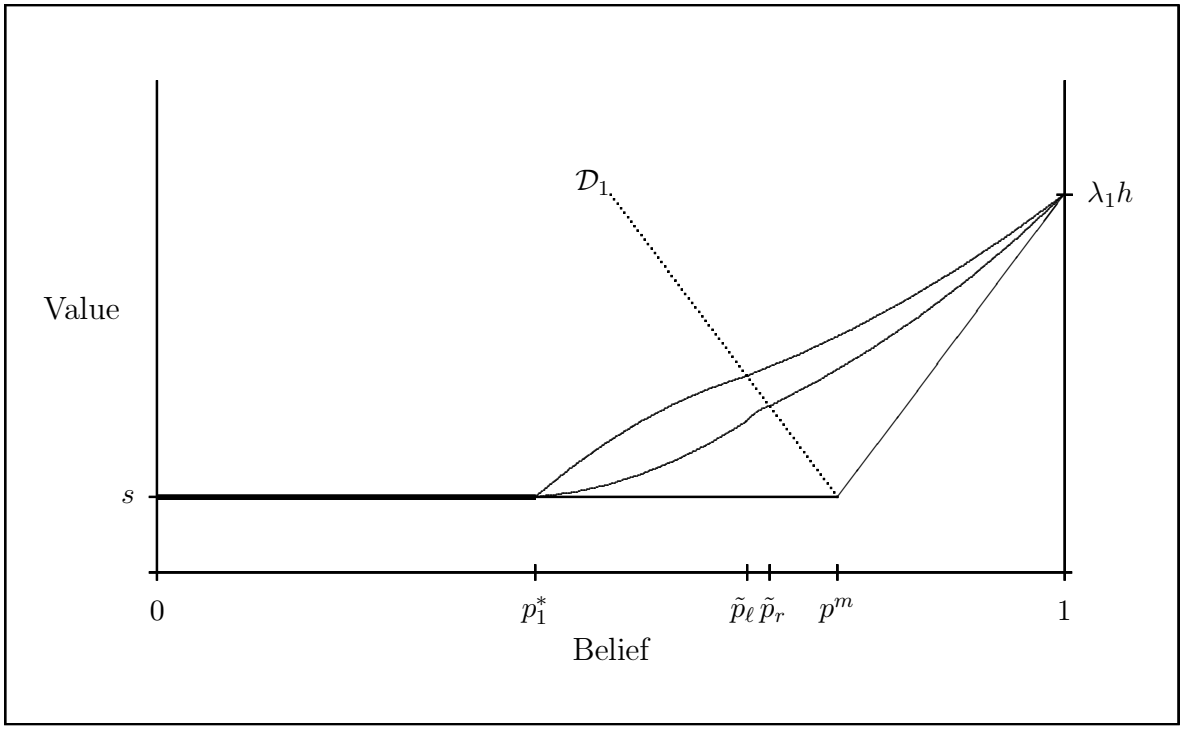

Figure 3: Payoffs in a two-player asymmetric equilibrium.

Proposition 6.2 (Welfare ranking) *The sum of the players' payoffs in the pure-strategy Markov perfect equilibria of Proposition 6.1 is decreasing in the cut-off belief $\bar{p}_r$ where one player switches to the safe arm for the first time, and strictly decreasing in $\bar{p}_r$ at all beliefs where both players use the risky arm.*

PROOF: Let $\bar{u}$ denote the solution to the ODE $u(p) = s + \frac{1}{2}\{2b(p,u) - c(p)\}$ (which corresponds to exactly one member of a two-player team experimenting) with $\bar{u}(p_1^*) = s$. In an equilibrium with right cut-off $\bar{p}_r$, the two players' average payoff function is $\bar{u}$ on $[p_1^*, \bar{p}_r]$; above $\bar{p}_r$, it is of the form V_2 as in (20) with the constant of integration determined by the condition $V_2(\bar{p}_r) = \bar{u}(\bar{p}_r)$. It is straightforward to verify that $V_2'(\bar{p}_r) > \bar{u}'(\bar{p}_r)$, which in turn implies that the V_2 part of the average payoff function is the higher, the lower is $\bar{p}_r$. ■

The way to achieve a more efficient equilibrium is to raise the lower of the two payoff functions by sharing the burden of experimentation more equally, that is, by switching roles more often. The lowest upper bound on aggregate payoffs is then given by a situation of payoff symmetry where each player obtains exactly half the payoff of the team strategy that has one player experiment to the left of the diagonal $\mathcal{D}_1$, and both players to the right of it.[11] This is the same payoff as if each player allocated

[11] This lowest upper bound on a player's payoff function is easy to calculate. To the left of $\mathcal{D}_1$, it solves the ODE $u(p) = s + \frac{1}{2}\{2b(p,u) - c(p)\}$ (which corresponds to exactly one member of a two-

exactly half of his time to the risky arm on the entire region below $\mathcal{D}_1$, hence clearly different from the payoff in the symmetric equilibrium of Proposition 5.1 where the fraction of time allocated to the risky arm falls gradually from 1 to 0 over this region.

In particular, there is a region of beliefs close to the single-agent cut-off where the intensity of experimentation in the symmetric equilibrium is lower than even in the 'worst' asymmetric one. By the logic of the last proposition, this ought to mean that welfare in the symmetric equilibrium should be lower at those beliefs than in any asymmetric equilibrium. The following proposition confirms this.

Proposition 6.3 (Welfare comparison with symmetric MPE) *For beliefs in the interval $]p_1^*, \tilde{p}_\ell]$ the sum of the players' payoffs in the pure-strategy asymmetric equilibria of Proposition 6.1 is strictly greater than the sum of players' payoffs in the mixed-strategy symmetric equilibrium of Proposition 5.1.*

PROOF: See the Appendix. ■

The intuition for this result is that, at each belief in the stated range, players are engaged in a coordination game like the Battle of the Sexes. There are two asymmetric pure equilibria of the type 'free-rider, lone ranger' where one player gets a high payoff and the other gets a low payoff. The players have different preferences over these equilibria (they would both prefer to free-ride) and it is this that presents them with a coordination problem. If the coordination problem is solved by mixing, the players do worse in aggregate.[12]

Propositions 6.2 and 6.3 show that alternating between the roles of free-rider and lone ranger as the belief changes is an effective (and incentive-compatible) way of increasing players' payoffs. Players can do even better if we allow them to switch between actions at *infinitely* many beliefs. In that case, they can take turns experimenting in such a way that no player ever has a last time (or lowest belief) at which he is supposed to use the risky arm. Surprisingly, it is then possible to reach cut-off beliefs below p_1^* in equilibrium. In fact, it is possible to (almost) attain the efficient cut-off p_2^*, but it is still reached too slowly.

Proposition 6.4 (Two players, pure strategies, infinite number of switches) *For each $\epsilon > 0$, there is a strictly decreasing sequence of beliefs $\{p_i^\dagger\}_{i=-2}^{\infty}$ with $p_2^* < p_\infty^\dagger := \lim_{i\to\infty} p_i^\dagger < p_2^* + \epsilon$ such that the following pure strategies constitute a Markov perfect equilibrium of the two-player strategic experimentation game: on $]p_{-2}^\dagger, 1]$, both players play R; on $]p_{i+1}^\dagger, p_i^\dagger]$, player 1 plays R and player 2 plays S if i is even, whereas player 1 plays S and player 2 plays R if i is odd; on $[0, p_\infty^\dagger]$, they both play S.*

player team experimenting) subject to the condition $u(p_1^*) = s$. The intersection of this solution with $\mathcal{D}_1$ determines the lowest possible realisation of the threshold $\bar{p}_r$. To the right of $\mathcal{D}_1$, we then have a function V_2 as in (20).

[12]Note that the symmetric equilibrium could exhibit a higher intensity of experimentation than the asymmetric ones at beliefs close to the cut-off $\hat{p}_2$. The proposition does not rule out that because of this, the mixed equilibrium could be more efficient at beliefs above $\tilde{p}_\ell$. Numerically, however, we find that the asymmetric equilibria are more efficient on the entire interval $]p_1^*, 1[$.

PROOF: See the Appendix. ■

Note that $p^{\dagger}_{-2}$ and $p^{\dagger}_{-1}$ are playing the roles of $\tilde{p}_r$ and $\tilde{p}_\ell$ from Proposition 6.1. Also note that as ϵ tends to zero the amount of experimentation performed in this equilibrium approaches the efficient amount. However, the intensity of experimentation is efficient only at times before $p^{\dagger}_{-2}$ and after $p^{\dagger}_{\infty}$ is reached; at times in between it is 1, and therefore first too low then too high relative to the efficient benchmark.

7 Concluding Remarks

There are some generalisations of our results that follow with no or relatively little additional work. First, all our results apply to bandit problems where the known arm generates a stationary non-deterministic stream of payoffs – we can simply reinterpret s as the expected flow payoff. Second, the construction of asymmetric equilibria for the case of fully revealing successes (Section 6) generalises to more than two players, and the corresponding results carry over. Third, the model is easily adapted to situations where news events carry bad news.

It is more cumbersome to examine asymmetric pure-strategy MPE in the general case where both Poisson intensities are strictly positive. We naturally expect that the results of Section 6 generalise, but the analysis becomes harder because we have to 'paste together' solutions to various differential-difference equations, keeping track of the precise region into which the posterior belief jumps after a success. We are investigating such equilibria in current work.

In the case of fully revealing successes, the model can be reinterpreted as a model of innovation and learning similar to Malueg and Tsutsui (1997). In contrast to these authors, we obtain closed-form solutions in our set-up. It would be interesting to vary the degree to which the post-innovation prize is shared, e.g. by introducing an advantage for whoever is first to experience a success.

Another extension that we intend to pursue is the introduction of asymmetries between players, for example regarding the discount rate or the ability to generate information from their experimentation effort. This may reduce the multiplicity of asymmetric equilibria that we have found for symmetric players. It may also allow us to investigate the question as to with whom a given agent would choose to play the strategic experimentation game.

More generally, we hope that Poisson bandits will prove useful as building blocks for models with a richer structure. Interesting extension in this direction could include rewards that depend on action profiles, unobservable outcomes, or costly communication.

Appendix

Proof of Corollary 5.2

We consider the equation

$$\Delta\lambda\, p(1-p)u'(p) + \lambda(p)u(p) = r\lambda(p)h - rs + \lambda(p)u(j(p)). \tag{A.1}$$

Let $\alpha = \lambda_0/\Delta\lambda$, and, for $i \geq 0$, define

$$g^{(i)}(p) = d_1^{(i)}p + d_0^{(i)}(1-p) + m^{(i)}(1-p)\Omega(p)^\mu + (1-p)\Omega(p)^\alpha \sum_{n=0}^{i-1} l^{(i-n)} \left(\ln\left[(\lambda_0/\lambda_1)^{n-1}\,\Omega(p)\right]\right)^n$$

where i is an iteration counter, and $d_1^{(i)}, d_0^{(i)}, m^{(i)}, l^{(i-n)}$ are constants (which depend on i). We are interested in the situation where $u(j(p)) = g^{(i)}(j(p))$:

$$\begin{aligned} g^{(i)}(j(p)) &= d_1^{(i)}\frac{\lambda_1}{\lambda(p)}\,p + d_0^{(i)}\frac{\lambda_0}{\lambda(p)}\,(1-p) + m^{(i)}\frac{\lambda_0}{\lambda(p)}\left(\frac{\lambda_0}{\lambda_1}\right)^\mu (1-p)\Omega(p)^\mu \\ &\quad + \frac{\lambda_0}{\lambda(p)}\left(\frac{\lambda_0}{\lambda_1}\right)^\alpha (1-p)\Omega(p)^\alpha \sum_{n=0}^{i-1} l^{(i-n)}\left(\ln\left[(\lambda_0/\lambda_1)^n\,\Omega(p)\right]\right)^n \end{aligned}$$

in which case the RHS of (A.1) becomes:

$$G^{(i)}(p) = D_1^{(i)}p + D_0^{(i)}(1-p) + M^{(i)}(1-p)\Omega(p)^\mu + (1-p)\Omega(p)^\alpha \sum_{n=0}^{i-1} L^{(i-n)}\left(\ln\left[(\lambda_0/\lambda_1)^n\,\Omega(p)\right]\right)^n$$

where

$$D_1^{(i)} = d_1^{(i)}\lambda_1 + r(\lambda_1 h - s), \quad D_0^{(i)} = d_0^{(i)}\lambda_0 + r(\lambda_0 h - s)$$

and

$$M^{(i)} = m^{(i)}\lambda_0\,(\lambda_0/\lambda_1)^\mu\,, \quad L^{(i-n)} = l^{(i-n)}\lambda_0\,(\lambda_0/\lambda_1)^\alpha\,.$$

The homogeneous equation has the solution

$$u_0(p) = (1-p)\Omega(p)^\alpha.$$

Using the method of variation of constants, we now write $u(p) = a(p)u_0(p)$ so that

$$\Delta\lambda\, p(1-p)u'(p) + \lambda(p)u(p) = \Delta\lambda\, p(1-p)u_0(p)a'(p).$$

The ODE thus transforms into the following equation for the first derivative of the unknown function a:

$$\begin{aligned} \Delta\lambda\, a'(p) &= \frac{G^{(i)}(p)}{p(1-p)\,u_0(p)} \\ &= D_1^{(i)}\,\Omega(p)^{-\alpha}(1-p)^{-2} + D_0^{(i)}\,\Omega(p)^{-\alpha+1}(1-p)^{-2} + M^{(i)}\,\Omega(p)^{\mu-\alpha+1}(1-p)^{-2} \\ &\quad + \Omega(p)(1-p)^{-2}\sum_{n=0}^{i-1} L^{(i-n)}\left(\ln\left[(\lambda_0/\lambda_1)^n\,\Omega(p)\right]\right)^n. \end{aligned}$$

Make the substitution $\omega = \Omega(p)$ and define $A(\omega) = a(p)$, so $a'(p) = -A'(\omega)/p^2$. Then

$$-\Delta\lambda\, A'(\omega) = D_1^{(i)}\,\omega^{-\alpha-2} + D_0^{(i)}\,\omega^{-\alpha-1} + M^{(i)}\,\omega^{\mu-\alpha-1} + \omega^{-1}\sum_{n=0}^{i-1} L^{(i-n)}\left(\ln\left[(\lambda_0/\lambda_1)^n\,\omega\right]\right)^n,$$

so

$$\begin{aligned} A(\omega) &= \frac{D_1^{(i)}}{\lambda_1}\,\omega^{-\alpha-1} + \frac{D_0^{(i)}}{\lambda_0}\,\omega^{-\alpha} + \frac{M^{(i)}}{\lambda_0-\mu\Delta\lambda}\,\omega^{\mu-\alpha} - \sum_{n=0}^{i-1}\frac{L^{(i-n)}}{(n+1)\Delta\lambda}\left(\ln\left[(\lambda_0/\lambda_1)^n\,\omega\right]\right)^{n+1} \\ &\quad + C^{(i+1)}, \end{aligned}$$

where $C^{(i+1)}$ is a constant of integration. Multiplying by $u_0(p) = (1-p)\,\omega^\alpha$ and substituting $\omega = \Omega(p)$ leads to

$$\begin{aligned} u(p) &= \frac{D_1^{(i)}}{\lambda_1}\,p + \frac{D_0^{(i)}}{\lambda_0}\,(1-p) + \frac{M^{(i)}}{\lambda_0-\mu\Delta\lambda}\,(1-p)\Omega(p)^\mu \\ &\quad + (1-p)\Omega(p)^\alpha\sum_{n=1}^{i}\frac{-L^{(i+1-n)}}{n\,\Delta\lambda}\left(\ln\left[(\lambda_0/\lambda_1)^{n-1}\,\Omega(p)\right]\right)^n + (1-p)\Omega(p)^\alpha C^{(i+1)}. \end{aligned}$$

This completes one iteration.

From the solution to the team problem:

$$d_1^{(0)} = \lambda_1 h, \quad d_0^{(0)} = \lambda_0 h, \quad \text{and} \quad m^{(0)} = C^{(0)},$$

where $C^{(0)}$ is the constant that fixes payoffs above the diagonal. The final (summed) term is vacuous for $i = 0$.

The above iterative step shows that, in general,

$$d_1^{(i)} = \lambda_1 h + \frac{r(\lambda_1 h - s)}{\lambda_1}\,i, \quad d_0^{(i)} = \lambda_0 h + \frac{r(\lambda_0 h - s)}{\lambda_0}\,i, \quad \text{and} \quad m^{(i)} = C^{(0)}\left(\frac{\lambda_0\,(\lambda_0/\lambda_1)^\mu}{\lambda_0-\mu\Delta\lambda}\right)^i.$$

After a little algebra, we find that the constants in the summation are given by:

$$l^{(i-n)} = \frac{C^{(i-n)}}{n!}\left(-\frac{\lambda_0\,(\lambda_0/\lambda_1)^\alpha}{\Delta\lambda}\right)^n \quad \text{for } n = 0,\ldots,i-1.$$

The constants $C^{(i-n)}$ $(n = 0,\ldots i-1)$ are chosen to ensure continuity. In particular, writing $\hat{\jmath}^{-i}$ for $j^{-i}(\hat{p})$, $C^{(i+1)}$ is chosen such that $u^{(i+1)}(\hat{\jmath}^{-i}) = u^{(i)}(\hat{\jmath}^{-i})$ for $i \geq 0$, and satisfies:

$$\begin{aligned} &C^{(i+1)}\,(1-\hat{\jmath}^{-i})\Omega(\hat{\jmath}^{-i})^\alpha \\ &= -\frac{r(\lambda_1 h - s)}{\lambda_1}\hat{\jmath}^{-i} - \frac{r(\lambda_0 h - s)}{\lambda_0}(1-\hat{\jmath}^{-i}) \\ &\quad + C^{(0)}\left(1 - \frac{\lambda_0\,(\lambda_0/\lambda_1)^\mu}{\lambda_0-\mu\Delta\lambda}\right)\left(\frac{\lambda_0\,(\lambda_0/\lambda_1)^\mu}{\lambda_0-\mu\Delta\lambda}\right)^i(1-\hat{\jmath}^{-i})\Omega(\hat{\jmath}^{-i})^\mu \\ &\quad + \left\{\sum_{n=0}^{i-1} C^{(i-n)}\left[\frac{1}{n!}\left(-\frac{\lambda_0\,(\lambda_0/\lambda_1)^\alpha}{\Delta\lambda}\ln\left[(\lambda_0/\lambda_1)^{n-1}\,\Omega(\hat{\jmath}^{-i})\right]\right)^n\right.\right. \\ &\qquad\qquad \left.\left. - \frac{1}{(n+1)!}\left(-\frac{\lambda_0\,(\lambda_0/\lambda_1)^\alpha}{\Delta\lambda}\ln\left[(\lambda_0/\lambda_1)^n\,\Omega(\hat{\jmath}^{-i})\right]\right)^{n+1}\right]\right\}(1-\hat{\jmath}^{-i})\Omega(\hat{\jmath}^{-i})^\alpha. \end{aligned}$$

■

Proof of Lemma 6.1

First note that each player's value function is continuous as a function of p and takes the value λh at $p = 1$ and s at $p = 0$; moreover it is differentiable *wherever he/she chooses optimally to switch* (from playing R to playing S, or *vice versa*) and the other player does not switch – if the right derivative is smaller, the player should switch at a larger p; if the right derivative is larger, the player should switch at a smaller p.

Our aim is to show that the region bounded below by the myopic payoff in the (p, u)-plane contains three regions, as in the picture in the main text. In one region (when the players are optimistic) it is dominant for each of them to play R and in another region (when the players are pessimistic and $u = s$) it is dominant for each of them to play S; in between, S and R are mutual best responses.

Assume that the continuation value of player n is given by $u_n(p)$, for $n = A, B$.

• Assume that player A (she) is playing R when the belief is in some interval $[p_\ell, p_r]$, and consider the best response of player B (he) on $[p_\ell, p_c] \subseteq [p_\ell, p_r]$. If it is also R then his value function on $[p_\ell, p_c]$ is given by V_2 from equation (20) with $V_2(p_\ell) = u_B(p_\ell)$; if his best response is S then his value function on $[p_\ell, p_c]$ is given by F_1 from equation (21) with $F_1(p_\ell) = u_B(p_\ell)$. Now, if $V_2(p) = F_1(p) = u$, say, then $V_2'(p) > F_1'(p)$ if $u > 2s - \lambda hp$, and $V_2'(p) < F_1'(p)$ if $u < 2s - \lambda hp$. Thus, if $u_B(p_\ell) > 2s - \lambda hp_\ell$, then his best response to R is to "join in" by playing R on $[p_\ell, p_r]$; if $u_B(p_\ell) < 2s - \lambda hp_\ell$, then his best response to R is to free-ride by playing S on $[p_\ell, p_c]$ for any p_c such that $F_1(p_c) < 2s - \lambda hp_c$; and he can only switch optimally at a belief $p_c \in [p_\ell, p_r]$ where $(p_c, u_B(p_c)) \in \mathcal{D}_1$.

• Now, assume that player A (she) is playing S when the belief is in some interval $[p_\ell, p_r]$, and consider the best response of player B (he) on $[p_\ell, p_c] \subseteq [p_\ell, p_r]$. If it is R then his value function on $[p_\ell, p_c]$ is given by V_1 from equation (3) with $V_1(p_\ell) = u_B(p_\ell)$; if his best response is also S then the belief no longer changes, so it must be the case that $u_B(p_\ell) = s$ and his value function on $[p_\ell, p_c]$ is simply s. Now, if $V_1(p) = s$, then $V_1'(p) > 0$ if $p > p_1^*$, and $V_1'(p) < 0$ if $p < p_1^*$. Thus, if $u_B(p_\ell) = s$, then his best response to S is to act unilaterally: if $p_\ell > p_1^*$ then play R on $[p_\ell, p_r]$; if $p_\ell < p_1^*$ then play S on $[p_\ell, p_c]$ for any p_c such that $p_c < p_1^*$; and he can only switch optimally at the belief p_1^*. However, if $u_B(p_\ell) > s$, then his best response to S must be to play R on $[p_\ell, p_r]$ (but note that $V_1'(p) < 0$ if $(r + \lambda p)V_1(p) > (r + \lambda)\lambda hp$). ■

Proof of Proposition 6.1

Let $\bar{p}_r$ denote the smallest belief where each player's continuation value is (weakly) above $\mathcal{D}_1$, and let $\bar{p}_\ell$ denote the largest belief where each player's continuation value is (weakly) below $\mathcal{D}_1$; necessarily, $p_1^* < \bar{p}_\ell \leq \bar{p}_r < p^m$.

For a belief in a neighbourhood of 1, specifically $p \in (\bar{p}_r, 1]$, R is the dominant strategy; and for a belief in a neighbourhood of 0, specifically $p \in [0, p_1^*]$, S is the dominant strategy. (We know that $u_n(0) = s$, and so S is a dominant response on any interval $[0, p_c] \subseteq [0, p_1^*]$). For beliefs $p \in (p_1^*, \bar{p}_\ell]$, the best response to S is to play R (act unilaterally), and the best response to R is to play S (free-ride). Now consider beliefs $p \in (\bar{p}_\ell, \bar{p}_r]$; let A be the player whose continuation value crosses $\mathcal{D}_1$ at $\bar{p}_\ell$ and let B be the player whose continuation value crosses $\mathcal{D}_1$ at $\bar{p}_r$. If B plays S, then A's best response is to play R (act unilaterally), and if B plays R, then A's best response is to play R ("join in"); thus R is the dominant response

for A. So, given A plays R, B's best response is to play S (free-ride). To summarise:

Belief p	0	p_1^*	$\bar{p}_\ell$	$\bar{p}_r$	1
A's strategy	S	S/R	R	R	
B's strategy	S	R/S	S	R	
A's continuation value	s	$F_{1,A}/V_{1,A}$	$V_{1,A}$	$V_{2,A}$	
B's continuation value	s	$V_{1,B}/F_{1,B}$	$F_{1,B}$	$V_{2,B}$	

and the strategies on $(p_1^*, \bar{p}_\ell]$ determine $\bar{p}_\ell$ endogenously, which player plays R and which player plays S on $(\bar{p}_\ell, \bar{p}_r]$, and $\bar{p}_r$ endogenously. If the players have the above continuation values, then the above strategies are best responses to each other; and if the players are using the above strategies, then the continuation values are indeed those given above. Thus the above strategies constitute an equilibrium with the equilibrium value functions given by the continuation values.

The 'simplest' equilibrium is where one player, say player 1, plays S on $(p_1^*, \tilde{p}_\ell]$, and the other player, player 2, plays R on this interval. Then player 1's value function F_1 satisfies equation (21) and player 2's value function V_1 satisfies equation (3), with $F_1(p_1^*) = V_1(p_1^*) = s$. So $F_1'(p_1^*) > V_1'(p_1^*)$, since whenever $F_1(p) = V_1(p) = u$, say, $F_1'(p) > V_1'(p)$ iff $\lambda h p < s$, i.e. iff $p < p^m$. Furthermore, it can be shown that F_1 is concave and V_1 is convex[13] and so if F_1 and V_1 take the same value again, say at $p_c > p_1^*$, then $F_1'(p_c) \leq V_1'(p_c)$, which implies that $p_c \geq p^m$. This shows that F_1 meets $\mathcal{D}_1$ at a *smaller* belief than does V_1, and that $F_1 > V_1$ on $(p_1^*, \tilde{p}_\ell]$; that is, player 1 must be A and switch from playing R on $(\tilde{p}_\ell, \tilde{p}_r]$, and player 2 must be B and switch from playing S on $(\tilde{p}_\ell, \tilde{p}_r]$. This equilibrium is thus given by:

Belief p	0	p_1^*	$\tilde{p}_\ell$	$\tilde{p}_r$	1
A's strategy	S	S	R	R	
B's strategy	S	R	S	R	
A's value function	s	$F_{1,A}$	$V_{1,A}$	$V_{2,A}$	
B's value function	s	$V_{1,B}$	$F_{1,B}$	$V_{2,B}$	

and the components of the value functions, and the switch-points, are determined as follows:

(1) C in $F_{1,A}$ from $F_{1,A}(p_1^*) = s$

(2) C in $V_{1,B}$ from $V_{1,B}(p_1^*) = s$

(3) $\tilde{p}_\ell$ from $F_{1,A}(\tilde{p}_\ell) = 2s - \lambda h \tilde{p}_\ell$

(4) C in $V_{1,A}$ from $V_{1,A}(\tilde{p}_\ell) = F_{1,A}(\tilde{p}_\ell) = 2s - \lambda h \tilde{p}_\ell$

(5) C in $F_{1,B}$ from $F_{1,B}(\tilde{p}_\ell) = V_{1,B}(\tilde{p}_\ell)$

(6) $\tilde{p}_r$ from $F_{1,B}(\tilde{p}_r) = 2s - \lambda h \tilde{p}_r$

(7) C in $V_{2,A}$ from $V_{2,A}(\tilde{p}_r) = V_{1,A}(\tilde{p}_r)$

(8) C in $V_{2,B}$ from $V_{2,B}(\tilde{p}_r) = F_{1,B}(\tilde{p}_r) = 2s - \lambda h \tilde{p}_r$

[13]It transpires that the second derivative of the functions F_1, V_1 and V_2 has the same sign as the constant of integration (in (21), (3) and (20) respectively) and thus the convexity/concavity of the solution is determined by that sign.

Note that the boundary condition at $p = 1$ is automatically satisfied because $V_{2,A}(1) = V_{2,B}(1) = \lambda h$ regardless of the constants of integration.

Noting that when $V_2(p) = V_1(p) = u$, say, $V_2'(p) > V_1'(p)$ iff $u > \lambda hp$ (the payoff from always playing R), we see that

- $0 < F_{1,A}'(p_1^*), \quad F_{1,A}'(\tilde{p}_\ell) > V_{1,A}'(\tilde{p}_\ell), \quad V_{1,A}'(\tilde{p}_r) < V_{2,A}'(\tilde{p}_r)$;
- $0 = V_{1,B}'(p_1^*), \quad V_{1,B}'(\tilde{p}_\ell) < F_{1,B}'(\tilde{p}_\ell), \quad F_{1,B}'(\tilde{p}_r) = V_{2,B}'(\tilde{p}_r)$.

Thus, as the common belief decays, B switches smoothly from R to S against R at $\tilde{p}_r$ (where A has a kink), both A and B switch at $\tilde{p}_\ell$ (each with a kink), and B switches smoothly again from R to S against S at p_1^* (where A again has a kink).

Following steps (1) and (3) determines the equation for $\tilde{p}_\ell$ given in the statement of the proposition; following steps (2), (5) and (6) determines the equation for $\tilde{p}_r$ given in the statement of the proposition; the remaining steps are for completeness only.[14]

Other equilibria for the two-player strategic problem

Any finite partition of the interval to the right of p_1^* can be used to construct a pure strategy equilibrium of the two-player strategic problem.

Take any finite (measurable) partition of $(p_1^*, p^m]$ and divide this into two subsets I_n, $n = 1, 2$. Build the continuous functions X_n on $[p_1^*, p^m]$ as follows: $X_n(p_1^*) = s$, X_n satisfies equation (21) on I_n (free-rider), X_n satisfies equation (3) on $I_{\neg n}$ (lone ranger).

Define $\bar{p}_\ell = \min\{p \in [p_1^*, p^m] : X_1(p) \vee X_2(p) = 2s - \lambda hp\}$. If $X_n(\bar{p}_\ell) \geq X_{\neg n}(\bar{p}_\ell)$ then $A = n$, else $A = \neg n$; $B = \neg A$.

Define $\bar{p}_r$ by $X_B(\bar{p}_r) = 2s - \lambda h\bar{p}_r$, so $\bar{p}_\ell \leq \bar{p}_r$.

Now take the partition $J_1 \cup J_2$ of $(p_1^*, \bar{p}_\ell]$, where $J_n = \{p \leq \bar{p}_\ell : p \in I_n\}$, i.e. J_n and I_n agree on $(p_1^*, \bar{p}_\ell]$.

Let A's strategy be as follows:
play S on $[0, p_1^*]$; play S on J_A and R on J_B; play R on $(\bar{p}_\ell, \bar{p}_r]$; play R on $(\bar{p}_r, 1]$.

Let B's strategy be as follows:
play S on $[0, p_1^*]$; play R on J_A and S on J_B; play S on $(\bar{p}_\ell, \bar{p}_r]$; play R on $(\bar{p}_r, 1]$.

Build the continuous functions Y_n on $[0, 1]$ as follows:

$Y_A(p) = s$ on $[0, p_1^*]$; Y_A satisfies equation (21) on J_A (free-rider) and satisfies equation (3) on J_B (lone ranger); Y_A satisfies equation (3) on $(\bar{p}_\ell, \bar{p}_r]$ (lone ranger); Y_A satisfies equation (20) on $(\bar{p}_r, 1]$.

$Y_B(p) = s$ on $[0, p_1^*]$; Y_B satisfies equation (3) on J_A (lone ranger) and satisfies equation (21) on J_B (free-rider); Y_B satisfies equation (21) on $(\bar{p}_\ell, \bar{p}_r]$ (free-rider); Y_B satisfies equation (20) on $(\bar{p}_r, 1]$.

If the continuation values are given by Y_n, then the above strategies are best responses to each other; and if the players are using the above strategies, then the continuation values are indeed given by Y_n. Thus the above strategies constitute an equilibrium with the equilibrium value functions given by Y_n.

Y_A and Y_B lie between $F_{1,A}$ and $V_{1,B} \cup F_{1,B}$ below and to the left of $\mathcal{D}_1$. Thus $\tilde{p}_\ell \leq \bar{p}_\ell \leq \bar{p}_r \leq \tilde{p}_r$, and so the 'simplest' equilibrium exhibits the least experimentation. ■

[14] Details are available from the authors on request.

Proof of Proposition 6.3

It is straightforward to check that at the symmetric equilibrium, each player's payoff function W satisfies

$$\frac{W(p)-s}{s(1-p)} = \frac{r}{\lambda}\left(R(p)^{-\lambda/r}-1\right) + \ln R(p)$$

on $]p_1^*, \hat{p}_2]$, where $\hat{p}_2$ is the upper cut-off from Proposition 5.1 with $N=2$ and

$$R(p) = \left(\frac{\Omega(p)}{\Omega(p_1^*)}\right)^{r/\lambda}$$

is decreasing in p and smaller than 1 for $p > p_1^*$. Aggregate payoff at this equilibrium is twice W on $]p_1^*, \hat{p}_2]$.

On $]p_1^*, \tilde{p}_\ell]$ the payoff function of player 2 (the last experimenter) at the 'simplest' asymmetric equilibrium described in Proposition 6.1 is of the type V_1 with the boundary condition $V_1(p_1^*) = s$. By (20) this satisfies

$$\frac{V_1(p)-s}{s(1-p)} = \frac{r}{r+\lambda}R(p)^{-\lambda/r} - 1 + \frac{\lambda}{r+\lambda}R(p),$$

where we have used the fact that $\Omega(p^m)/\Omega(p_1^*) = r/(r+\lambda)$. On $]p_1^*, \tilde{p}_\ell]$ the payoff function of player 1 (the last free-rider) at this same equilibrium is of the type F_1 with the boundary condition $F_1(p_1^*) = s$. By (21) this satisfies

$$\frac{F_1(p)-s}{s(1-p)} = \frac{r\lambda}{(r+\lambda)^2}R(p)^{-\lambda/r} - \frac{r\lambda}{(r+\lambda)^2}R(p),$$

again using $\Omega(p^m)/\Omega(p_1^*) = r/(r+\lambda)$. Aggregate payoff at this equilibrium thus satisfies

$$\frac{V_1(p)+F_1(p)-2s}{s(1-p)} = \frac{r(r+2\lambda)}{(r+\lambda)^2}R(p)^{-\lambda/r} - 1 + \frac{\lambda^2}{(r+\lambda)^2}R(p).$$

With the notation $\mu = r/\lambda$, a simple calculation now gives

$$\frac{V_1+F_1-2W^*}{s(1-p)} = -\frac{\mu^2(2\mu+3)}{(\mu+1)^2}R^{-1/\mu} + \frac{1}{(\mu+1)^2}R - 2\ln R - 1 + 2\mu,$$

where we have suppressed the dependence of V_1, F_1, W and R on p. We want to show that the right-hand side is positive on the interval $]p_1^*, p^m]$. To this end, we consider the right-hand side as a function $f(R)$ on the interval $[R(p^m), 1]$. As $f(1) = 0$, $f'(1) < 0$ and $f'' < 0$ on this interval, it suffices to show that $f(R(p^m)) > 0$. Now, $R(p^m) = [\mu/(\mu+1)]^\mu$, so

$$f(R(p^m)) = -\frac{2\mu+1}{\mu+1} + \frac{1}{(\mu+1)^2}\left(\frac{\mu}{\mu+1}\right)^\mu - 2\mu\ln\frac{\mu}{\mu+1}\,.$$

As a function of μ on the positive half-axis, this is quasi-concave with limit zero as μ tends to 0 or $+\infty$, hence positive throughout.

For $p \in]p_1^*, \min\{\hat{p}_2, \tilde{p}_\ell\}]$, therefore, the sum of payoffs at the symmetric equilibrium lies strictly below the sum of payoffs at the asymmetric equilibrium. This implies that the payoff of player 1 (the last free-rider) at the asymmetric equilibrium lies above W. Hence, the belief at which player 1's value function intersects $\mathcal{D}_1$ must be strictly lower than the belief at which W intersects $\mathcal{D}_1$, or $\tilde{p}_\ell < \hat{p}_2$. ■

Proof of Proposition 6.4

Our aim is to build an MPE where the players make an infinite number of switches between R and S in finite time. We find equilibria where the beliefs fall arbitrarily close to the team cut-off before the players stop using R for good. This means that the amount of experimentation performed can get arbitrarily close to the efficient level. The intensity of experimentation, however, will be inefficiently low.

The intuition for these equilibria is that for all beliefs above the team cut-off level there is a Pareto gain from performing more experiments, so provided any player's immediate contributions are sufficiently small relative to the long-run Pareto gain, performing experiments in turn can be sustained as an equilibrium.

The equilibrium constructed below is such that a player's payoff before embarking on a round of single-handed experimentation equals s. (At all other beliefs, each player has an expected payoff strictly exceeding s.) While pinning down equilibrium payoffs this way simplifies the construction, other choices would work as well.

Fix a belief $p_0^\dagger$ strictly between the two-player team cut-off p_2^* and the single-agent optimal cut-off p_1^*. Given this starting point, we will define a strictly decreasing sequence of beliefs $\{p_i^\dagger\}_{i=0}^\infty$ bounded below by p_2^* such that the following Markovian pure strategies constitute an equilibrium at beliefs $p \leq p_0^\dagger$: player 1 uses R on any interval $[p_i^\dagger, p_{i+1}^\dagger[$ for even i and S otherwise; player 2 uses R on any interval $[p_i^\dagger, p_{i+1}^\dagger[$ for odd i and S otherwise. In particular, both players use S at beliefs $p \leq p_\infty^\dagger = \lim_{i\to\infty} p_i^\dagger$.

Assuming for the moment that we have already constructed such a sequence of beliefs, let X_i be player 1's expected payoff at the start of the interval $[p_i^\dagger, p_{i+1}^\dagger[$ with even i when the players use the above strategies. Similarly, let Y_i be player 2's expected payoff at the start of the interval $[p_i^\dagger, p_{i+1}^\dagger[$ with odd i when the players use those strategies.

Let i be even. Player 1 uses R on the interval $[p_i^\dagger, p_{i+1}^\dagger[$, so her value function satisfies the single-agent differential equation (1) there. If $u(p_{i+1}^\dagger)$ is her expected payoff once the belief has hit $p_{i+1}^\dagger$, we obtain

$$X_i = \lambda h p_i^\dagger + [u(p_{i+1}^\dagger) - \lambda h p_{i+1}^\dagger] \frac{1-p_i^\dagger}{1-p_{i+1}^\dagger} \left(\frac{x_i}{x_{i+1}}\right)^{r/\lambda}$$

or

$$\frac{X(p_i^\dagger) - s}{s\,(1-p_i^\dagger)} = \frac{1}{x_i} - 1 + \frac{u(p_{i+1}^\dagger) - \lambda h p_{i+1}^\dagger}{s(1-p_i^\dagger)} \left(\frac{x_i}{x_{i+1}}\right)^{r/\lambda}$$

where $x_i = \Omega(p_i^\dagger)/\Omega(p^m)$.

On the interval $[p_{i+1}^\dagger, p_{i+2}^\dagger[$, player 1 watches while player 2 uses R, so her value function satisfies the differential equation (11) with $K = 1$ and the terminal condition $u(p_{i+2}^\dagger) = X_{i+2}$. Solving this gives

$$\frac{u(p_{i+1}^\dagger) - s}{s(1-p_i^\dagger)} = \frac{\lambda}{r+\lambda}\frac{1}{x_{i+1}} + \left(\frac{x_{i+1}}{x_{i+2}}\right)^{r/\lambda} \left[\frac{X_{i+2}-s}{s(1-p_{i+2}^\dagger)} - \frac{\lambda}{r+\lambda}\frac{1}{x_{i+2}}\right].$$

Substituting this into the above equation for X_i, assuming that

$$X_i = s \quad \text{for} \quad i = 0, 2, 4, \ldots,$$

and re-arranging, we get the second-order difference equation

$$0 = \frac{\lambda}{r+\lambda}\left[x_i^{-\frac{r+\lambda}{\lambda}} - x_{i+2}^{-\frac{r+\lambda}{\lambda}}\right] + \left[\frac{r}{r+\lambda}\frac{1}{x_i} - 1\right] x_i^{-\frac{r}{\lambda}} - \left[\frac{r}{r+\lambda}\frac{1}{x_{i+1}} - 1\right] x_{i+1}^{-\frac{r}{\lambda}}$$

for even i. Going through the same steps for player 2 under the assumption that

$$Y_i = s \quad \text{for} \quad i = 1, 3, 5, \ldots,$$

we see that this difference equation holds for *all* i.

With the new variable

$$z_i = \frac{x_{i+1} - x_i}{x_i},$$

this yields the two-dimensional first-order system

$$\begin{aligned} x_{i+1} &= x_i\,(1 + z_i), \\ \frac{\lambda}{r+\lambda}(1+z_{i+1})^{-r/\lambda-1} &= (1+z_i)^{r/\lambda+1} + x_i\,(1+z_i) - x_i\,(1+z_i)^{r/\lambda+1} - \frac{r}{r+\lambda}. \end{aligned}$$

The fact that $p_2^* < p_0^\dagger < p_1^*$ translates into $\frac{r+\lambda}{r} < x_0 < \frac{r+2\lambda}{r}$. Via the above system, each choice of $z_0 > 0$ determines a strictly increasing sequence $\{x_i\}$ (and hence also a strictly decreasing sequence $\{p_i^\dagger\}$.) We are done if we can show that there is a choice of z_0 such that the sequence $\{x_i\}$ is bounded above by $\frac{r+2\lambda}{r}$.

Fix $\delta > 0$ such that $x_0 + \delta < \frac{r+2\lambda}{r}$. Since

$$\frac{\partial z_{i+1}}{\partial z_i}(x_i, 0) = \frac{r}{\lambda} x_i - \frac{r+\lambda}{\lambda},$$

there is a $\gamma > 0$ and a β strictly between 0 and 1 such that for all (x_i, z_i) with $x_0 \le x_i \le x_0 + \delta$ and $0 \le z_i \le \gamma$ the partial derivative of z_{i+1} with respect to z_i satisfies

$$0 < \frac{\partial z_{i+1}}{\partial z_i} \le \beta.$$

Now let

$$z_0 = \min\left\{\gamma, (1-\beta)\ln\frac{x_0+\delta}{x_0}\right\}.$$

A simple induction argument then shows that

$$\ln\frac{x_i}{x_0} = \sum_{j=0}^{i-1} \ln(1+z_j) \le \sum_{j=0}^{i-1} z_j \le \sum_{j=0}^{i-1} \beta^j z_0 = \frac{z_0}{1-\beta} \le \ln\frac{x_0+\delta}{x_0}$$

for all i. This implies that $x_i \le x_0 + \delta < \frac{r+2\lambda}{r}$ for all i, as desired.

Note that by taking x_0 closer and closer to $\frac{r+2\lambda}{r}$ (which corresponds to taking $p_0^\dagger$ closer and closer to p_2^*), we can insure that the limit of the x_i gets arbitrarily close to $\frac{r+2\lambda}{r}$, and so the distance between the limit belief $p_\infty^\dagger$ and the efficient cut-off p_2^* becomes smaller than any given positive ϵ. To complete the construction of the equilibrium, we now only have to move back from $p_0^\dagger$ to higher beliefs and assign actions to the two players in the way we did for the pure-strategy equilibria with a finite number of switches. ■

References

Admati, A.R. and M. Perry (1991): "Joint Projects without Commitment", *Review of Economic Studies*, **58**, 259–276.

Bellman, R. and K.L. Cooke (1963): *Differential-Difference Equations* (New York: Academic Press).

Bergemann, D. and U. Hege (1998): "Venture Capital Financing, Moral Hazard and Learning", *Journal of Banking and Finance*, **22**, 703–735.

Bergemann, D. and U. Hege (2001): "The Financing of Innovation: Learning and Stopping" (CEPR Discussion Paper No. 2763).

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

Lockwood, B. and J.P. Thomas (1999): "Gradual Cooperation in Repeated Games with Reversibilities" (working paper, University of Warwick and University of St. Andrews).

Malueg, D.A. and S.O. Tsutsui (1997): "Dynamic R&D Competition with Learning", *RAND Journal of Economics*, **28**, 751–772.

Marx, L. and S. Matthews (2000): "Dynamic Voluntary Contribution to a Public Project", *Review of Economic Studies*, **67**, 327–358.

Park, K. (1999): "Leader-Follower Model of Strategic Experimentation" (working paper, UCLA).

Presman, E.L. (1990): "Poisson Version of the Two-Armed Bandit Problem with Discounting", *Theory of Probability and its Applications*, **35**, 307–317.

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