Strategic communication in exponential bandit problems

Chantal Marlats*and Lucie Ménager†

January 5, 2011

Abstract

We generalize Keller, Rady and Cripps's [2005] model of strategic experimentation by assuming that transfers of information between players are costly. We introduce costly communication in three different ways. First, we consider the *Paying to exchange information* game: the exchange of information between players occurs if and only if both payed the communication cost. Second, we consider the *Paying to buy information* case, where players pay the cost to observe their opponent's action. Finally, we study the *Paying to give information* case, where players pay the communication cost to display their actions and outcomes. We study the existence and the structure of equilibria in each setting. We show that making communication costly is efficient, in the sense that it decreases free-riding, and increases the speed of learning at equilibrium.

1 Introduction

In many economic situations, agents are trying to optimize their decisions while improving their information at the same time. Consider for instance oil or gas companies, contemplating the exploitation of a new site. The new site can be either very rewarding,

*CORE, UCL, 34 rue du roman pays, Louvain la Neuve, 1348, Belgique.

†LEM Université Paris 2, 5-7 avenue Vavin, 75006 Paris, France

having more reserves than the old one, or can contain no oil at all. Each company has to decide how much of its effort to allocate to the new site, whose reward is unknown, and how much to the old one, whose reward can be considered as certain in the short run.

A large literature in operation research and in game theory has analyzed the decision problem of a single agent who has to choose sequentially between two alternatives whose expected return are uncertain. Multi-armed bandits models[1] where an agent has to decide whether to play a safe arm, offering a known payoff, and a risky arm of unknown payoff, have been used to formalized this tradeoff between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the best payoff).

In particular situations such as those described in the oil company example, the first breakthrough discovery of oil in the new site reveals its superiority to the old site and leads all companies to drill there and to abandon the exploitation of the old site. In such situations, no news is bad news: players gradually become less optimistic as long as no breakthrough happens, and fully informed as soon as it does. This particular exploration versus exploitation trade-off has been studied by Keller, Rady, and Cripps [2005] (KRC hereafter), using a game of strategic experimentation in continuous time where the risky arm generates positive payoffs after exponentially distributed random times if it is good, and never pays out anything if it is bad. In this game, players have to decide what fraction of a given resource to allocate to the risky arm (the new site in the oil company example), and to the safe arm (the old site). Players are said to *experiment* if they allocate some resource to the risky arm while its type is still unknown. Players observe each other's actions (the resource allocated to each arm) and outcomes (the occurrence or absence of a breakthrough), so that information about the type of the risky arm is a public good.

[1]For a review of literature on bandit models and their applications in economics, see Bergemann and Välimäki (2006).

It follows that players *free-ride* on experimentation at equilibrium, in the sense that, at a given belief, they experiment less than what they would have done, were they isolated.

We may think of many situations in which players facing the same exploration versus exploitation trade-off cannot physically observe each other. For instance, the old and the new sites can be so vast that companies cannot observe where their competitors are drilling, and whether they find oil in the new site or not. The fact that players "observe" each other in KRC's model implies that there is some mechanism such that the information about each player's action is public and free (giant screen, oral announcement,...).

In this paper, we generalize KRC's model by assuming that transfers of information between players are costly. We consider two players[2] who have to choose sequentially what fraction of their resource to allocate between a risky and a safe arm. As in KRC, the risky arm generates positive payoffs after exponentially distributed random times if it is good, and never pays out anything if it is bad. At each date, players have the opportunity to pay a cost to "communicate" with their opponent in a particular sense. We introduce costly communication in KRC's model in three different ways. By communication we mean that players can choose to truthfully give or to obtain information, that is the history of his actions and observation, at some cost. First, we consider the *Paying to exchange information* game: the exchange of information between players occurs if and only if both payed the communication cost. Second, we consider the *Paying to buy information* case, where players pay the cost to observe their opponent's action. Finally, we study the *Paying to give information* case, where players pay the communication cost to display their actions and outcomes.

We study the existence and the structure of equilibria in Markov strategies in the different communication settings, and investigate whether making information transfers

[2]Results could be easily obtain in the n-player game, with the appropriate communication structure.

costly reduces free-riding and modifies the speed of learning.

We show that in any setting, there exist equilibria in Markov strategies with individual beliefs as state variable. Their structure strongly depend on the communication setting. 1) When players pay to exchange their information, there exist multiple symmetric equilibria in which players communicate for intermediate beliefs if the communication cost is not too high. Their structure is as follows: when players are very pessimistic, namely when their belief of the risky arm being good is small, they allocate all of their resource to the safe arm and do not communicate. When they are very optimistic, they devote all their resource to the risky arm and don't communicate either. For intermediate beliefs, the expected gain of information is greater that its cost: players communicate, and allocate a positive share of their resource to both arms. We identify the symmetric equilibrium that maximizes players' expected payoff. We also show that if the communication cost is positive, there is no asymmetric equilibrium, whereas it is shown in KRC that there exist several asymmetric equilibria when communication is free. 2) When players pay to buy their opponent's information, there is a unique symmetric equilibrium. This equilibrium is identical to the one with the largest communication interval in the case where players exchange information.However, there exists at least one asymmetric equilibrium, whose structure is as follows. Players have two distinct roles, one being a pioneer (say player 1) and the other one a free-rider (player 2). For very pessimistic beliefs, no player experiments. For optimistic beliefs, both players experiment, buying the other one's information except for very optimistic beliefs where the expected gain of new information is not worth the cost. For intermediate beliefs, only one player experiments, while the other one free-rides in the sense that he plays the safe arm but buys the other player's information. The two players swap the role of pioneer and free-rider on this range of beliefs. 3) When players pay to give information, there is no equilibrium in which players communicate, whether

symmetric or asymmetric. This result partly follows from the absence of *encouragement effect* in the exponential bandits model, first analyzed by Bolton and Harris [1999]. By this effect, the players experiment at some beliefs at which they wouldn't have experimented, were they isolated. As explained in KRC, its absence in the exponential bandits model follows from the fact that a player will experiment more than if he were alone if he thinks that it encourages its opponent to experiment. The only way for this to happen is to have a breakthrough. Yet in this case, all the uncertainty would be resolved, and the additional information that the player would receive would be of no value to him. In our game, a player will display his outcome if he thinks he may receive useful information from it. Yet the only way for him to make his opponent communicates is to display a useful information, that is to have a breakthrough, in which case his opponent's information is of no value to him. The absence of asymetric equilibrium where player exchange or give information is true for all $c > 0$. This implies that the asymetric equilibria found in KRC are not robust to communication cost in some sense.

We show that the amount of experimentation, that is the total quantity of resource allocated by both players to the risky arm over time, increases with the communication cost. This result shows that, quite intuitively, making communication costly reduces free-riding. Indeed, making communication costly tends to reduce the exchange of information at equilibrium, and then reduces the possibility of free-riding. Another important welfare issue is that players make the right decision, that is play R if the risky arm is good, and S otherwise. From this point of view, the relevant criterium to maximize is the speed of learning. We show that there exists an optimal communication cost, for which players learn faster than when they never communicate or when they always communicate.

Utile? We use a model of strategic experimentation that generalizes that in KRC, and whose characteristics are that 1) *many agents* face a bandit problem in *continuous time*,

where the risky arm might yield payoffs after *exponentially* distributed random times, and that 2) agents *do not observe* others' actions and outcomes.

Some works use exponential bandits with public observation: the financial contracting models of Bergemann and Hege [1998,2005], the investment timing model of Décamps and Mariotti [2004]. Other works study bandit problems with many agents in continuous time with a different information structure: Bolton and Harris (1999) with a model where the risky arm yields a flow payoff with Brownian noise, Keller and Rady [2009] with a model where the risky arm distributes lump-sum payoffs according to a Poisson process; some others study bandit models with many agents in discrete time: Bergemann and Valimaki [1996], in which the model is set in discrete time and a general model of uncertainty is considered. In all these works, actions and outcomes of players are publicly observed. A recent literature focuses on the case in which only the actions of the opponents (Rosenberg, Solan and Vieille [2007], Valimaki and Murto [2009]), or only the payoffs of the opponents (Bonnatti and Horner [2010], Horner and Samuelson [2010]) are observed . To the best of our knowledge, strategic communication in a bandit model where actions and outcomes are private information has not been studied.

The rest of the paper is organized as followed. In section 2, we introduce KRC's model of strategic experimentation. In section 3 we present the general game of strategic costly communication we consider. We study the equilibria of the Paying to exchange information, Paying to buy information, and Paying to give information in sections 4, 5, and 6. In section 7, we study the welfare properties of costly communication. We discuss in section 8 of remaining questions and possible extensions.

2 Strategic experimentation with exponential bandits

The aim of this section is to introduce KRC's model of strategic experimentation with exponential bandits. This model corresponds to those studied in this paper for a communication cost zero.

2.1 The model

Bandit problem

Time t is continuous. There are two players, each of them endowed with one unit of a perfectly divisible resource per unit of time. Each player faces a two-armed bandit problem where he continually has to decide what fraction of the resource to allocate to each arm. One arm, denoted S, is safe and yields a deterministic payoff $s > 0$ per unit of resource allocated to it. The other arm, denoted R, is risky and can be either "bad" or "good". If it is bad, then it always yield a payoff 0. If it is good, then it yields random payoffs of mean $h > 0$ at random times, the arrival rate of these payoffs being a constant λ per unit of resource allocated to the risky arm. The average payoff per unit of resource allocated to the risky arm over time is denoted by $g := \lambda h$. Furthermore, the arrival of lump-sums is independent across players. The term *exponential bandits* used by KRC comes from the fact that the time of arrival of the first lump-sum would be exponentially distributed if players were to use a time-invariant allocation.

Formally, if a player allocates the fraction $k_t \in [0, 1]$ of the resource to the risky arm over an interval of time $[t, t + dt)$, and consequently the fraction $1 - k_t$ to the safe arm, then he receives the payoff $(1 - k_t)sdt$ from the arm S, the payoff 0 from the risky arm if it is bad, and the payoff $k_t g$ from the risky arm if it is good. The payoffs are supposed to be such that $0 < s < g$, so that each player strictly prefers R to S if the risky arm is good, and strictly prefers S to R if it is bad.

Beliefs

At the beginning of the game, players do not know the state of the risky arm but have a common prior belief about it. In KRC's model, players observe each other's actions and outcomes at any time, and therefore will hold common posterior beliefs throughout time. We will depart from this setting by assuming that players decide whether to show or not their actions and outcomes. Let p_t denote the players' probability at date t that the risky arm is good. Since a bad risky arm always yields a payoff 0, the first arrival of a lump-sum payoff, called a breakthrough, reveals to all players that the risky arm is good. In other words, the arrival of a breakthrough resolves the players' uncertainty about the type of the risky arm. Players are said to *experiment* when they use R while its type is still unknown. As long as players experiment, the probability that R is good decreases.

Payoffs

The belief p_t depends on the arrival of a breakthrough, and is then a random variable. The actions of players depend on their beliefs and are then also random variables. Let k_t^i be the fraction of the unit resource allocated by player i to the risky arm over the interval $[t, t+dt)$, and $\{k_t^i\}_{t\geq 0}$ the stochastic process of player i's actions, such that k_t^i is measurable with respect to the information available to player i at time t. Player i's total expected discounted payoff is

$$E_{\{k_t^i\},\{p_t\}}\left[\int_0^\infty re^{-rt}[(1-k_t^i)s + k_t^i g p_t]dt\right]$$

A player's payoff depends on others' actions only through their impact on the evolution of his beliefs.

Evolution of beliefs

Let $K_t := k_t^1 + k_t^2$ be the total amount of resource allocated to the risky arm over the interval $[t, t+dt)$. If the risky arm is good, the probability of none of the players achieving

a breakthrough is $\prod_i (1 - k_t^i \lambda dt)$ which is equal to $1 - K_t \lambda dt$ up to terms to the order $o(dt)$ which will be ignored in the rest of the paper; if the risky arm is bad, the probability of none of them achieving a breakthrough is 1. Therefore, if players start with the common belief p_t at time t and don't achieve a breakthrough in $[t, t + dt)$, the updated belief at the end of the period is

$$p_{t+dt} = \frac{p_t(1 - K_t \lambda dt)}{1 - p_t + p_t(1 - K_t \lambda dt)}$$

by Bayes'rule.

Therefore, as long as there is no breakthrough, the belief changes by[3]

$$dp_t = -K_t \lambda p_t (1 - p_t) dt$$

Once there is a breakthrough, the posterior belief jumps to 1.

2.2 Strategic experimentation

Myopic behavior

A myopic agent would simply maximize the expected short-run payoff $(1 - k_t)s + k_t g p_t$. Therefore, for $p_t > p^m := \frac{s}{g}$, it is myopically optimal to allocate the resource only to R; for $p_t < p^m$, it is myopically optimal to allocate the resource only to S;

Farsighted behavior

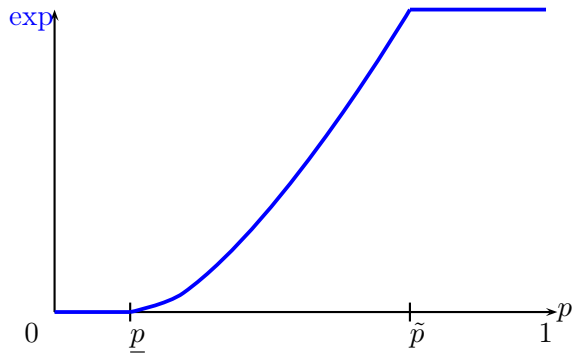
Since $0 < s < g$, agents strictly prefer the risky arm if it is good. Farsighted agents anticipate that they may receive more in the future by playing R if it is good. Therefore, there exists a belief threshold $\underline{p}$ such that it is optimal for foresighted agents to devote some resource to R for $p_t \in [\underline{p}, p^m]$.

In the one-agent model, the information of the agent only comes from his own experi-

[3] $\frac{dp_t}{dt} = p_t' = \lim_{dt \to 0} \frac{p_{t+dt} - p_t}{dt} = \lim_{dt \to 0} \frac{-(1-p_t)p_t K_t \lambda}{1 - p_t + p_t(1 - K_t \lambda dt}} = -K_t \lambda p_t (1 - p_t)$

mentation. It has been shown[4] that it is optimal for the agent to allocate all of his resource to S if his belief is below a threshold $\underline{p}$, and to allocate all of his resource to R otherwise. $\underline{p}$ will therefore be called the single agent cut-off belief.

In the multi-player model of KRC however, the information of a player comes from the experimentation of all of them. This implies that at the unique symmetric Markovian equilibrium, players experiment less than in the one-agent case, in the sense that they allocate only a fraction of the resource to the risky arm for beliefs at which they would have allocated all the resource if they were isolated. More precisely, KRC show that the equilibrium strategies of the players at the symmetric equilibrium are as follows. When a player is very confident that the risky arm is good, that is when p_t is above a threshold $\overline{p} < p^m$, then it is optimal for him to play R with probability 1. When his belief of R being good is under $\underline{p}$, then it is optimal for him to play S with probability one. However, for intermediate beliefs, that is for $p_t \in [\underline{p}, \overline{p}]$, it is optimal for players to allocate an increasing fraction of the resource to the risky arm. This equilibrium behavior features *free-riding* in the following sense. For $p_t \in [\underline{p}^*, \overline{p}^*]$, it would be optimal for an isolated agent to devote all of his resource to R. When the player benefits from the information gained by observing others' actions and outcomes, he is better off by insuring himself in allocating a part of his resource to S and letting the other player experimenting for him.



KRC also show that the encouragement effect analyzed by Bolton and Harris (1999)

[4] See KRC

doesn't exist. By this effect, the presence of other players encourages at least one of them to continue experimenting at beliefs more pessimistic than the single-agent cut-off belief. It rests on two conditions: the 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 acts as a pioneer. With exponential bandits, the likelihood that others will experiment decreases unless a breakthrough happens. But since a breakthrough is fully revealing, the additional experimentation by other players after the breakthrough is of no value to the pioneer.

3 Strategic communication

We generalize KRC's model by assuming that transfers of information between players are costly. We introduce costly information in three different ways: first, players may pay a cost $c > 0$ to exchange information, in the sense that both players observe their opponent's action if and only if they both payed the cost c. Second, they may pay the cost c to get information about others. Finally, they may pay the cost c to inform their opponent of their own actions and outcomes.

For each setting, we describe the game of strategic acquisition of information, then we characterize players's best responses and we study equilibria.

At each date t, players decide whether to pay the communication cost $c > 0$, or to pay nothing. What they expect to receive from paying c depends on the setting considered. At each time t, players also decide what quantity of the resource to be allocated to the risky arm, and whether they communicate or not. Players' strategies have then two components: an experimentation strategy $k_t \in [0, 1]$, and an information acquisition strategy $q_t \in \{0, 1\}$, where $q_t = 1$ if players pay the cost c, and $q_t = 0$ otherwise. We will call *communication*

the action of paying c. More precisely players pay c, a flow cost to observe (or reveal) the histories of actions and observation that takes place while the flow is bieng paid.

As in KRC, we will consider stationary Markov strategies, namely strategies that depend only on individual beliefs.

Fix a belief p and consider $k_i \in [0,1]$ and $q_i \in \{0,1\}$ player i's experimentation and communication decisions for this belief. Let $K := k_i + k_j$ be the total amount of experimentation.

If player i's actions are k_i, q_i, then i gets $(1-k_i)s$ from the safe arm, $k_i g$ from the risky arm if it is good, which is an event of probability p, and pays c if he communicates, that is if $q_i = 1$. Therefore, i's expected current payoff is $(1-k_i)s + k_i gp - q_i c$. By the principle of optimality, player i's value function satisfies the following Bellman equation:

$$u(p) = \max_{k_i, q_i} \left\{ r((1-k_i)s + k_i gp - q_i c)dt + e^{-rdt} E[u(p+dp) \mid p, k_i, k_j, q_i, q_j] \right\}$$

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

The expected continuation payoff $u(p+dp)$ is g if a breakthrough occurs, and $u(p) + u'(p)dp$ otherwise. If individual actions k_i and k_j are known to player i, his probability of a breakthrough is $pK\lambda dt$ and his belief evolutes following $dp = -K\lambda p(1-p)dt$. If player i doesn't know his opponent's action, then his subjective probability of a breakthrough is $pk_i\lambda dt$, and his belief changes following $dp = -k_i\lambda p(1-p)dt$.

The discounted expected continuation payoff $E[u(p+dp) \mid p, k_i, k_j, q_i, q_j]$ depends on the communication setting we consider.

1. In the setting where players pay to exchange their information, i knows j's action if and only if $q_i = q_j = 1$. His expected continuation payoff is then

$$E[u(p+dp) \quad | \quad p, k_i, k_j, q_i, q_j] = q_i q_j \left[gpK\lambda dt + (1 - pK\lambda dt)(u(p) - K\lambda p(1-p)u'(p)dt) \right]$$
$$+(1 - q_i q_j) \left[gpk_i\lambda dt + (1 - pk_i\lambda dt)(u(p) - k_i\lambda p(1-p)u'(p)dt) \right]$$

2. In the setting where players pay to get the information, i knows j's action if and only if $q_i = 1$. His expected continuation payoff is then

$$E[u(p+dp) \quad | \quad p, k_i, k_j, q_i, q_j] = q_i \left[gpK\lambda dt + (1 - pK\lambda dt)(u(p) - K\lambda p(1-p)u'(p)dt) \right]$$
$$+(1 - q_i) \left[gpk_i\lambda dt + (1 - pk_i\lambda dt)(u(p) - k_i\lambda p(1-p)u'(p)dt) \right]$$

3. In the setting where players pay to display their information, i knows j's action if and only if $q_j = 1$. His expected continuation payoff is then

$$E[u(p+dp) \quad | \quad p, k_i, k_j, q_i, q_j] = q_j \left[gpK\lambda dt + (1 - pK\lambda dt)(u(p) - K\lambda p(1-p)u'(p)dt) \right]$$
$$+(1 - q_j) \left[gpk_i\lambda dt + (1 - pk_i\lambda dt)(u(p) - k_i\lambda p(1-p)u'(p)dt) \right]$$

In the rest of the paper, we will use the following notation: $c(p) := s - gp$ and $b(p,u) := \frac{\lambda}{r} p(g - u(p) - (1-p)u'(p))$. $c(p)$ is the opportunity cost of playing R, and $b(p,u)$ is the discounted expected private benefit of playing R, and has two parts: $\lambda p(g - u(p))$ is the expected value of the jump to $u(p) = 1$ should a breakthrough occur, and $-\lambda p(1-p)u'(p)$ is the negative effect on the overall payoff should no breakthrough occur.

4 Paying to exchange the information (PTEI)

We make the assumption that the exchange of information between players occurs only if both decided to pay the communication cost. The kind of communication we consider is then that of a *club*: to communicate with each other, two agents have to undertake a

costly action. In other words, both have to go to the club to be able to talk with each other.

Using $1-rdt$ as an approximation to e^{-rdt}, and neglecting terms of the order $o(dt)$, we can rewrite player i's payoff $u(p) = s + \max_{k_i \in [0,1], q_i \in \{0,1\}} \Big\{ k_i(gp - s + \frac{\lambda}{r} p(g - u(p) - (1-p)u'(p))) + q_i(-c + q_j k_j \frac{\lambda}{r} p(g - u(p) - (1-p)u'(p))) \Big\}$

player i's payoff rewrites

$$u(p) = s + \max_{k_i \in [0,1], q_i \in \{0,1\}} \{k_i(b(p,u) - c(p)) + q_i(q_j k_j b(p,u) - c)\}$$

where: - c is the communication cost

- $q_2 k_2 b(p,u)$ is the discounted expected private benefit of communicating, that is the benefit to player i of the information generated by player j, and has also two parts: $q_2 k_2 \lambda p(g - u(p))$ is the expected value of the jump to $u(p) = 1$ should a breakthrough occur for player j, and $-\lambda p(1-p)u'(p)$ the negative effect on the overall payoff should no breakthrough occur for player j.

4.1 Best responses

Players' best-responses are determined by comparing $c(p)$ and $b(p,u)$, namely the opportunity cost with the expected private benefit of playing R for the experimentation decision, and by comparing c and $k_j q_j b(p,u)$, namely the instantaneous cost of communication with the expected benefit of the information gained from player j for the communication decision.

Let us first point out that no player will communicate alone at equilibrium. Indeed, if $q_j = 0$, then $u_i(p) = s + \max_{k_i \in [0,1], q_i \in \{0,1\}} \{k_i(b(p,u) - c(p)) + q_i(-c)\}$. If i were to communicate, he would pay the communication cost without gaining anything from it, then $q_i = 0$.

Suppose now that player j communicates ($q_j = 1$), and let us determine player i's best responses to k_j. When $q_j = 1$, player i's continuation payoff is

$$u_i(p) = s + \max_{k_i \in [0,1], q_i \in \{0,1\}} \{k_i(-c(p) + b(p,u))) + q_i(-c + k_j b(p,u))\}$$

• $q_i = 0$ and $k_i = 0$ if $k_j b(p,u) - c < 0$ and $b(p,u) - c(p) < 0$. In this case, $u_i(p) = s$, so $b(p,u) = \frac{p}{\mu}(g - s)$ and these are best-response if $p \leq \min\{\underline{p}, \frac{\mu c}{k_j(g-s)}\}$, where $\underline{p} := \frac{\mu s}{(1+\mu)(g-s)+\mu s}$ is the single-agent cut-off.

• $q_i = 0$ and $k_i = 1$ if $k_j b(p,u) - c < 0$ and $b(p,u) - c(p) > 0$. In this case, $u_i(p) = s + b(p,u) - c(p)$ so these are best-response if $u_i(p) \in [s, s + \frac{c}{k_j} - c(p)]$.

• $q_i = 0$ and $k_i \in [0,1]$ if $k_j b(p,u) - c < 0$ and $b(p,u) - c(p) = 0$. In this case, $u_i(p) = s$, so these are best-responses only if $p = \underline{p}$.

• $q_i = 1$ and $k_i = 0$ if $k_j b(p,u) - c > 0$ and $b(p,u) - c(p) < 0$. In this case, $u_i(p) = s + k_j b(p,u) - c$ so these are best-responses if $u_i(p) \in [s, s + k_j c(p) - c]$.

• $q_i = 1$ and $k_i \in [0,1]$ if $k_j b(p,u) - c > 0$ and $b(p,u) - c(p) = 0$. So these are best-responses if $u_i(p) = s + k_j c(p) - c$.

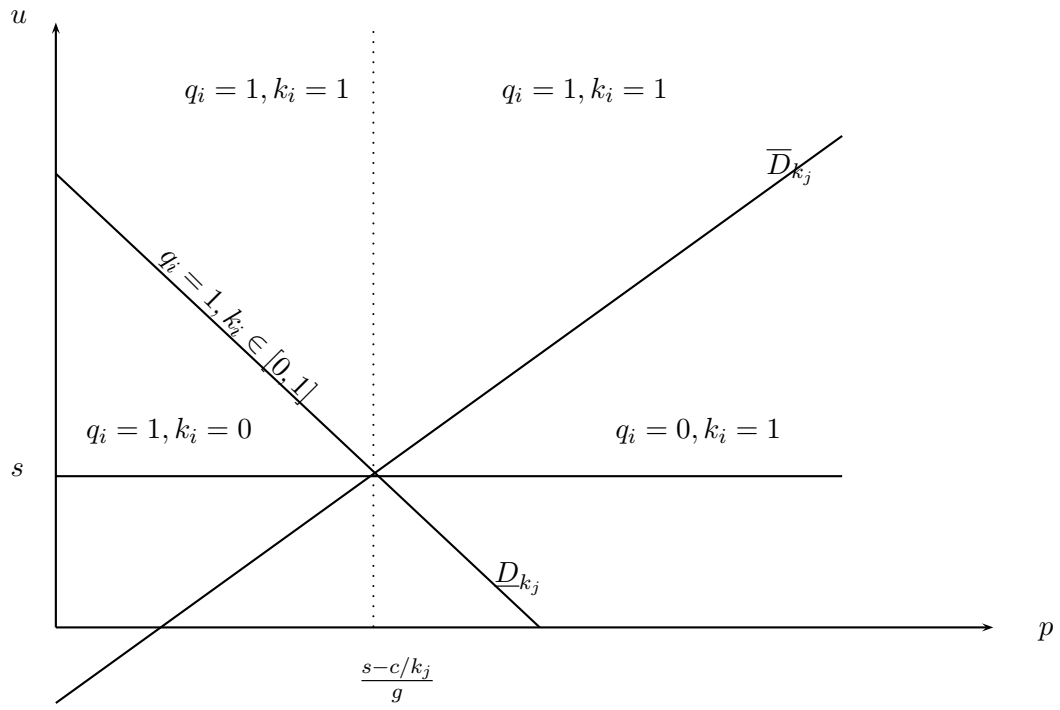
• $q_i = 1$ and $k_i = 1$ if $k_j b(p,u) - c > 0$ and $b(p,u) - c(p) > 0$. In this case, $u_i(p) = s - c + k_j b(p,u) + b(p,u) - c(p)$, so these are best-responses if $u_i(p) > s - c - c(p) + (1 + k_j) \max\{\frac{c}{k_j}, c(p)\}$.

This analysis shows that player i's best-response depends on whether in the (p,u)-plane, the point $(p, u_i(p))$ lies below, on or above the line $\underline{\mathcal{D}}_{k_j}$ and below or above the line $\overline{\mathcal{D}}_{k_j}$, where

$$\underline{\mathcal{D}}_{k_j} := \{(p,u) \in [0,1] \times \mathbb{R}_+ \mid u = s - c + k_j c(p)\}$$

$$\overline{\mathcal{D}}_{k_j} := \{(p,u) \in [0,1] \times \mathbb{R}_+ \mid u = s + \frac{c}{k_j} - c(p)\}$$

For $k_j > 0$, $\underline{\mathcal{D}}_{k_j}$ and $\overline{\mathcal{D}}_{k_j}$ are respectively a downward and an upward sloping diagonal, which both cross the safe payoff line $u(p) = s$ at $p = \frac{s-c/k_j}{g}$. For $k_j = 0$, $\underline{\mathcal{D}}_{k_j}$ coincides with the safe payoff line, and $\overline{\mathcal{D}}_{k_j}$ "tends" to $u(p) = \infty$, so that the area where $q_i = 1$ best-responses is empty. The following graph gives the area of best responses when the opponent communicate and spends a proportion of time k_j playing the risky arm in a given interval of time.



4.2 Equilibria

We now study the Markovian equilibria of the game. We first show that there is no asymmetric equilibrium if the communication cost is positive. Then we show that there is a multiplicity of symmetric Markovian equilibria, with all the same structure.

Proposition 1 (Asymmetric equilibrium)**.** If $c > 0$, there is no asymmetric equilibrium in Markov strategies.

Proof of Proposition 1. By best-responses analysis, we know that $q_j = 0 \Rightarrow q_i = 0$. Therefore, either $q_i = q_j = 0$, or $q_i = q_j = 1$. There can be no asymmetry in communication strategies. If $q_i(p) = q_j(p) = 0$, then there is no asymmetry in experimentation strategies since both players face the single-agent problem. Suppose now that $q_i(p) = q_j(p) = 1$. By best-response analysis, we know that $k_j = 0 \Rightarrow q_i = 0$. Therefore, $q_i = q_j = 1 \Rightarrow k_i > 0$ and $k_j > 0$. Let us show that $k_i \in]0,1[\Rightarrow k_j \in]0,1[$. If $k_i \in]0,1[$, then $b(p,u) = c(p)$ and

$u = s - c + k_j c(p)$. If $k_j = 1$, then $k_i \in]0,1[$ only when i's continuation payoff crosses the line $u = s - c + c(p)$, which happens only for some belief $\widetilde{p}$. For $p > \widetilde{p}$, the continuation payoff u is above the line $s-c+c(p)$, and is then in the area where $k_i = 1$ is a best-response to $k_j = 1$. Thus there is no range of beliefs such that $k_i \in]0,1[$ is a best-response to $k_j = 1$.

Therefore, there is no equilibrium in which both players communicate, with only one of them allocating all the resource to R. □

It is noteworthy that there exist asymmetric equilibria when $c = 0$. Indeed, KRC show that there exist several types of asymmetric equilibria. In one of these types for instance, when the players are optimistic, they play R; when they are pessimistic, they play S; in between, there are two regions in which one of them free-rides by playing S while the other one plays R, players swaping roles of pioneer and free-rider between the two regions. Formally, there are two cut-offs p_1 and p_2, and one switchpoint p_s, such that both players play S when the common belief p is below p_1, both play R when $p > p_2$; on $(p_1, p_s]$, player 1 plays R and player 2 plays S, and on $(p_s, p_2]$, player 1 plays S and player 2 plays R. When $c = 0$ there exists also an equilibrium profile where players do not communicate. Thus in $c = 0$ the set of asymmetric equilibria is greater than in KRC. Note that the set of equilibrium payoffs is not lower hemi continuous in c. Moreover the following proposition means that if communication is costly then there is no symmetric equilibria, even is c is arbitrarily small. This means that asymmetric equilibrium are not robust to communication cost.

In the PTEI setting, the fact that communication is costly implies that either both players communicate, or both don't. If they don't communicate, they both play the single-agent optimal strategy. Since a player will not communicate if his opponent doesn't experiment, there can be no equilibrium in which for some beliefs, both players communicate, one of

them experimenting while the other one free-rides.

We now show that the PTEI game has infinitely many equilibria, all with the same structure.

Proposition 2 (Structure of symmetric equilibria)**.** Let $c \geq 0$ be a communication cost. The PTEI game has a infinitely many equilibria in Markovian strategies with the common posterior belief as the state variable.

In these equilibria, the equilibrium allocation of the resource and the communication strategies are defined as follows. There exist $\underline{p}$, $\overline{p}(c)$, $\underline{a}(c)$, and $\overline{a}(c)$ such that $\underline{p} \leq \underline{a}(c) \leq \overline{p}(c) \leq \overline{a}(c)$ and such that for all $\overline{p}(c) \leq x \leq y \leq \overline{a}(c)$, the following strategy is an equilibrium strategy:

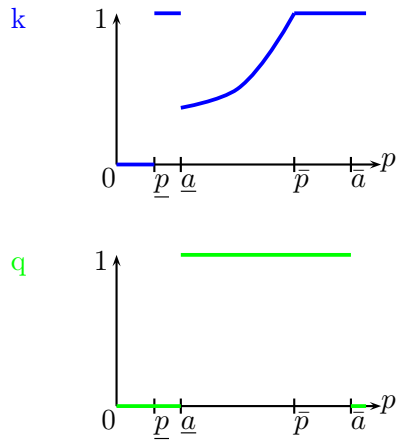
- the safe arm is used exclusively at beliefs below the single-player cut-off $\underline{p}$;
- the risky arm is used exclusively on $[\underline{p}, \underline{a}(c)]$ and on $[\overline{p}(c), 1]$.
- for beliefs on $[\underline{a}(c), \overline{p}(c)]$, players communicate and allocate an increasing fraction of the resource to R up to the belief cut-off $\overline{p}(c)$.
- for beliefs above $\overline{p}(c)$, they use R exclusively and communicate only on $[x, y]$.

There exists $c_{max} > 0$ such that $\forall\ c < c_{max},\ \underline{p} < \underline{a}(c) < \overline{p}(c) < \overline{a}(c)$.

Let us first remark that for $c = 0$, this equilibrium is KRC's equilibrium.

Furthermore, if $c \geq c_{max}$, then $\underline{a}(c_{max}) = \overline{a}(c_{max})$, and then players do not communicate at equilibrium.

Finally, the equilibrium where $x = \underline{a}(c)$ and $y = \overline{a}(c)$, namely the equilibrium in which the range of beliefs for which players communicate is the biggest, is the one that procures the maximal payoffs to players. The following graph gives a rough shape of this equilibrium:



Proof of Proposition 2. We first show that the strategy where $x = \overline{p}(c)$ and $y = \overline{a}(c)$ is the strategy of a symmetric equilibrium. Then we show that any strategy with $\overline{p} < x < \overline{a}(c)$ is also a strategy of a symmetric equilibrium.

- Let us first list player i's best-response at symmetric equilibrium:
- $(q_i = 0, k_i = 0)$ if $p \leq \underline{p}$;
- $(q_i = 0, k_i = 1)$ if $p > \underline{p}$;
- $(q_i = 1, k_i = 0)$ if $u \in [s, s - c] = \emptyset$;
- $(q_i = 1, k_i \in [0, 1])$ if $u = s + k(p)c(p) - c$ and $k(p)c(p) > c$;
- $(q_i = 1, k_i = 1)$ if $u > s - c - c(p) + 2\max\{c, c(p)\} = (s + c - c(p))_{\ p > \frac{s-c}{g}} + (s + c(p) - c)_{\ p < \frac{s-c}{g}}$.

Lemma 1. If $\underline{p} \geq \frac{s-c}{g}$, then $\underline{a}(c) = \overline{p}(c) = \overline{a}(c)$. It follows that $q(p) = 0$ for all p, and $k(p) = 0$ on $p \leq \underline{p}$, and $k(p) = 1$ on $p > \underline{p}$.

Proof: On $p \leq \underline{p}$, $k = 0$. The diagonal $\underline{D}_{k_j}$ crosses the safe payoff line $u(p) = s$ in $\frac{s-c/k_j}{g} < \underline{p}$ for any value of k_j. Since $u(\underline{p}) = s$, players' payoff cannot cross $\underline{D}_{k_j}$, and always remains in the area where $q = 0$ and $k = 1$ are best-response. □

Proof: □

Lemma 2. If $c > 0$, then $\underline{p} < \underline{a}(c)$.

Proof: As soon as $c > 0$, $(0,0)$ followed by $(0,1)$. $c > 0 \Rightarrow k(p) = 0$ for $p < \underline{p}$ and $u(p) = s$. At $p = \underline{p}$, the continuation payoff enters in the area where $q = 0$ and $k = 1$ are best-responses with $k = 0$. At symmetric equilibrium, it means that $k_j = 0$ and that the line $s - c + k_j c(p)$ is $s - c$. Therefore, i's continuation payoff cannot cross the line $s - c + k_j c(p)$ at $p = \underline{p}$. Therefore, it will cross the line for some $\widetilde{p} > \underline{p}$, and i will play $q_i = 0$ and $k_i = 1$ for p in between. □

Lemma 3. If $\underline{p} < \frac{s-c}{g}$, and if there exists $\widetilde{p} \in [\underline{p}, \frac{s-c}{g}$ such that $s - c + k(\widetilde{p})c(\widetilde{p}) = s$, then $\underline{a}(c) < \overline{p}(c) < \overline{a}(c)$.

Proof If $p \leq \underline{p}$, then $k = 0, q = 0$, and $u(p) = s$. The continuation payoff enters in the area where it could cross the line $\underline{D}_{k_j}$ with $k(p) = 0$. At symmetric equilibrium, k_j will then be equal to 0, and $\underline{D}_0 = s - c$ is strictly above the safe payoff line. Therefore, there exists $\varepsilon > 0$ such that $u(p)$ cannot cross $D_{k(p)}$ on $[\underline{p}, \underline{p} + \varepsilon]$. If $k(p) = 1$ on $[\underline{p}, \underline{p} + \varepsilon]$, then $s < s - c + c(p)$ so $q_i = 0$ is a best response to any strategy of j. Therefore, $k = 1$ and $q = 0$ on $[\underline{p}, \underline{p} + \varepsilon]$. Players payoff is then the single agent payoff $V_0(p) = gp + (1-p)K_0\Omega(p)^\mu)$, with $\Omega(p) = \frac{1-p}{p}$ and $\mu = \frac{\lambda}{r}$, obtained by solving $V = s + b(p, V(p)) - c(p)$ up to a constant of integration K_0. The constant is determined by the usual smooth-pasting condition $V(\underline{p}) = s$.

On $[\underline{p}, \underline{p} + \varepsilon]$, $k(p) = 1$, so D_1 is the line of equation $u = s - c + c(p)$. Since there exists $\widetilde{p}$ such that $s - c + k(\widetilde{p}) = s$, the diagonal $\underline{D}_{k_j}$ belongs to the area between $u = s$ and $u = s - c + c(p)$. Since furthermore players' payoff $V(p)$ is increasing, it will cross $\underline{D}_1$ for some value $\underline{a}(c)$, determined by

$$V(\underline{a}(c)) = s - c + c(\underline{a}(c))$$

There exists $\varepsilon >$ such that for p on $[\underline{a}(c), \underline{a}(c) + \varepsilon]$, players play $q = 1$ and $k \in [0,1]$. Let $W(p)$ the payoff they obtain in that case. $q = 1$ and $k \in [0,1]$ are optimal if $b(p, W) = c(p)$,

that is if $W(p) = s + (1+\mu)(g-s) + \mu s(1-p)\ln\Omega(p) + K_W(1-p)$, with K_W a constant of integration. By the smooth-pasting condition, K_w is determined by $W(\underline{a}(c)) = V(\underline{a}(c)$. Players' payoff in that case is also determined by $u(p) = s - c + k(p)c(p)$. This give the optimal fraction of resource allocated to the risky arm: $k(p) = \dfrac{W(p) - s + c}{c(p)}$. $k(p)$ is increasing. It is a best-response as long as $k(p) \leq 1$. Let $\overline{p}(c)$ be the cut-off such that $k(\overline{p}(c)) = 1$, namely such that

$$k(\overline{p}(c)) = \frac{W(\overline{p}(c)) - s + c}{c(\overline{p}(c))}$$

There exists $\varepsilon > 0$ such that on $[\overline{p}(c), \overline{p}(c) + \varepsilon]$, $k(p) = 1$, and players'payoff is above the line $\underline{D}_1$, and above $\overline{D}_1$. Therefore, it is in the area where $q = 1$ and $k = 1$ are best-responses. Let us denote by V_1 the continuation payoff in that area. $V_1(p)$ is obtained by resolving $V_1(p) = s - c + 2b(p, V_1) - c(p)$, and is $V_1(p) = gp + (1-p)K_1\Omega(p)^{\frac{\mu}{2}} - c(1 - p\frac{2}{2+\mu})$, with K_1 a constant of integration determined by $W(\overline{p}(c)) = V_1(\overline{p}(c))$. Since $V_1(1) = g - c\frac{\mu}{2+\mu} < s - c - c(1) = g$ and $V_1(\overline{p}(c)) > s - c + c(\overline{p}(c))$, there exists $\overline{a}(c)$ such that

$$V_1(\overline{a}(c)) = s - c + c(\overline{a}(c))$$

For $p > \overline{a}(c)$, players' payoff is in the area where $q = 0, k = 1$ is dominant. Players' payoff in this area is $V_2(p) = gp + (1-p)K_2\Omega(p)^{\mu})$, the constant K_2 being determined by $V_1(\overline{a}(c)) = V_2(\overline{a}(c))$. □

Lemma 4. There exists c_{max} such $c \leq c_{max} \Rightarrow \widetilde{p} \geq \underline{p}$.

Proof: $\widetilde{p}$ is defined by $k(\widetilde{p})c(\widetilde{p}) = c \Leftrightarrow W(\widetilde{p}) = s$. If $c = 0$, $\widetilde{p} = \frac{s}{g}$ the myopic cut-off, so $\frac{s}{g} > \underline{p}$ is a sufficient condition for $\widetilde{p} > \underline{p}$ if $c = 0$. If $c = s$, then $\widetilde{p} < \underline{p}$. Furthermore, differentiating with respect to c, we find that $K'_W(1 - \widetilde{p}) = \widetilde{p}'(\mu s \ln(\Omega(\widetilde{p})) + \frac{\mu s}{\widetilde{p}} + K_W)$. Easy calculations show that $K'_W < 0$, which implies that $\widetilde{p}' < 0$. Since $\frac{s}{g} > \underline{p}$ is always true, there exists $c_{max} > 0$ such that for all $c \geq c_{max}$, $\widetilde{p} \geq \underline{p}$. □

□

5 Paying to buy information (PTBI)

We now consider the case where players simply pay to get information about their opponent's actions and outcomes. As in the previous section, we use $1-rdt$ as an approximation to e^{-rdt}, and neglect terms of the order $o(dt)$, so that player i's payoff rewrites

$$u(p) = s + \max_{k_i \in [0,1], q_i \in \{0,1\}} \{k_i(b(p,u) - c(p)) + q_i(k_j b(p,u) - c)\}$$

5.1 Best-responses

The analysis of best-responses is the same as that of the Paying to exchange information case except that player i may purchase information ($q_i = 1$) even if player j doesn't ($q_j = 0$).

Let k_j be player j's experimentation decision. Player i's continuation payoff is

$$u(p) = s + \max_{k_i \in [0,1], q_i \in \{0,1\}} \{k_i(-c(p) + b(p,u))) + q_i(-c + k_j b(p,u))\}$$

• $q_i = 0$ and $k_i = 0$ if $k_j b(p,u) - c < 0$ and $b(p,u) - c(p) < 0$. In this case, $u = s$, so these are best-response if $p \leq \min\{\underline{p}, \frac{\mu c}{k_j(g-s)}\}$.

• $q_i = 0$ and $k_i = 1$ if $k_j b(p,u) - c < 0$ and $b(p,u) - c(p) > 0$. In this case, $u = s + b(p,u) - c(p)$ so these are best-response if $u \in [s, s + \frac{c}{k_j} - c(p)]$.

• $q_i = 0$ and $k_i \in [0,1]$ if $k_j b(p,u) - c < 0$ and $b(p,u) - c(p) = 0$, so these are best-responses only if $p = \underline{p}$.

• $q_i = 1$ and $k_i = 0$ if $k_j b(p,u) - c > 0$ and $b(p,u) - c(p) < 0$. In this case, $u = s + k_j b(p,u) - c$ so these are best-responses if $u \in [s, s + k_j c(p) - c]$.

• $q_i = 1$ and $k_i \in [0,1]$ if $k_j b(p,u) - c > 0$ and $b(p,u) - c(p) = 0$. In this case, $u = s + k_j c(p) - c$, so these are best-responses if $k_j c(p) - c > 0$.

• $q_i = 1$ and $k_i = 1$ if $k_j b(p,u) - c > 0$ and $b(p,u) - c(p) > 0$. In this case, $u = s - c + k_j b(p,u) + b(p,u) - c(p)$, so these are best-responses if $u > s - c - c(p) + (1+k_j)\max\{\frac{c}{k_j}, c(p)\}$.

5.2 Equilibria

At symmetric equilibrium, the situation where one player purchases information while the other doesn't will not occur. Therefore, there is a unique symmetric equilibrium, which is the same as the one in the PTBI game whose communication interval is the largest.

Proposition 3 (Symmetric equilibrium)**.** There is unique symetric equilibrium in which players play the equilibrium strategy profile with the largest communication interval in the PTEI game.

Proof of Proposition 3. At symmetric equilibrium, either $q_i = q_i = 0$ or $q_i = q_j = 1$. Therefore, even if best-responses are not the same in the PTBI game, best-responses at symmetric equilibrium are the same than the one with the largest interval. Recall that in the previous case multiplicity was a consequence that if a player communicate in ain $(\underline{a}, \bar{a})$ it was best response to do so (otherwise the player would pay c and receive no information). This argument dos not hold any more in the present case. □

However, since $q_i = 1$ may be a best-response to $q_j = 0$, there may exist asymmetric equilibria, contrary to the previous case. Indeed, we show that there exists at least one asymmetric equilibrium, in which players have two distinct roles, one being a pioneer (say player 1) and the other one a free-rider (player 2). For very pessimistic beliefs, no player experiments. For optimistic beliefs, both players experiment, buying the other one's information except for very optimistic beliefs where the expected gain of new information

is not worth the cost. For intermediate beliefs, only one player experiments, while the other one free-rides in the sense that he plays the safe arm but buys the other one's information. The two player swaps the role of pioneer and free-rider on this range of beliefs.

Proposition 4 (Asymmetric equilibrium)**.** If $\frac{s-c}{g} > \underline{p}$, there is an asymmetric Markovian equilibrium in the "Paying to buy information" game, where the players's actions depend as follows on the common belief.

There are five cut-offs, $\underline{p}$, p_1, p_2, p_3 and p_4 such that:

- player 1 buys player 2's information on $[p_1, p_3]$. He plays S on $[0, \underline{p}] \cup [p_1, p_2]$, and R otherwise.

- player 2 buys player 1's information on $[\underline{p}, p_1] \cup [p_2, p_4]$. He plays S on $[0, p_1]$, and R otherwise.

Proof of Proposition 4. If $\frac{s-c}{g} \leq \underline{p}$, we know from the proof of Proposition 2 that player i's payoff cannot cross the diagonal $\underline{D}_1$, whatever k_i and q_i for $p > \underline{p}$, and then $q_i = 0$ is a dominant strategy for both players. Players then face the single-agent problem and play the same experimentation strategy.

Suppose now that $\frac{s-c}{g} > \underline{p}$. Let player 2 be the free-rider, and player 1 the pioneer.

• For $p < \underline{p}$, both players play $q_i = 0, k_i = 0$.

• Suppose that $k_1 = 1$ for $p \in [\underline{p}, \underline{p} + \varepsilon]$, and let us study player 2's best-responses. If $q_2 = 0$, then $k_2 = 1$ and $u = s + b(p,u) - c(p)$. Yet $q_2 = 0$ if $-c + k_1 b(p,u) < 0 \Rightarrow b(p,u) < c$, so $q_2 = 0$ if $u < s + c - c(p) = c + gp$. Since $\underline{p} < \frac{s-c}{g}$, $u(p) \geq s = u(\underline{p}) > c + g\underline{p}$. So $q_2 = 0$ cannot be a best-response and $q_2 = 1$.

Given $q_2 = 1$, $k_2 = 0$ is optimal if and only if $u < s - c + c(p)$. Since $\underline{p} < \frac{s-c}{g}$, there exists $p_1 > \underline{p}$ such that $k_2(p) = 0$ for all $p \in [\underline{p}, p_1]$. In this case, 2's payoff is defined by $u_2(p) = s - c + b(p, u_2)$ with $u_2(\underline{p}) = s$. p_1 is determined by $u_2(p_1) = s - c + c(p_1)$.

If $k_2 = 0$ for $p \in [\underline{p}, p_1]$, then it is optimal for player 1 to play $k_1 = 1$ since $p > \underline{p}$.

For $p > p_1$ it becomes dominant for player 2 to play $k_2 = 1$, whatever k_1. Indeed, 2's payoff is in the area where $q_i = 1, k_i = 1$ is a best-response against $q_i = 1, k_i = 1$ since $u > s - c + c(p)$, and also in the area where $(q_i = 0, k_i = 1)$ is a best-response against $k_i = 0$. Therefore, $k_2 = 1$. For $p \leq p_1$, player 1's payoff is $u_1(p) = V(p) < u_2(p)$. Therefore, there exists $p_2 > p_1$ such that $u_1(p) < s - c + c(p)$ for all $p \in [p_1, p_2]$. Then it is optimal for player 1 to free-ride in turn and to play $q_1 = 1$ and $k_1 = 0$. On this range of beliefs, 1's payoff is defined by $u_1(p) = s - c + b(p, u_1)$ with $u_1(p_1) = V(p_1)$, and p_2 is determined by $u_2(p_2) = s - c + c(p_2)$.

For $p > p_2$, for the same reasons as for player 2, it becomes dominant for player 1 to play $k_1 = 1$. Since both payoffs are in the area where $(q_i = 1, k_i = 1)$ is a best-response against $(q_i = 1, k_i = 1)$, both players experiment and communicate, as long as their payoffs are above the line $s + c - c(p)$. In this area, individual payoffs are defined by $u_i(p) = s - c + 2b(p, u_i) - c(p)$, with continuity of payoffs in p_2. Since $u_2(p) > u_1(p)$, there exists p_3 and p_4, $p_3 < p_4$, such that $u_1(p)$ and $u_2(p)$ cross the line $s + c - c(p)$ respectively in p_3 and p_4.

□

6 Paying to give information (PTGI)

We now consider the case where players pay to give their information to their opponent. This is the classical kind of "postal" communication, where for instance player 1 sends by e-mail the description of his actions and outcomes to player 2. As in the previous sections, using $1 - rdt$ as an approximation to e^{-rdt} and neglecting terms of the order $o(dt)$, we

rewrite player i's payoff

$$u(p) = s + q_j k_j b(p,u) + \max_{k_i, q_i} \{-q_i c + k_i(b(p,u) - c(p)\}$$

Clearly, it is dominant for i to play $q_i = 0$. It follows that when players use Markov strategies, there is no asymmetric equilibrium, and there is a unique symmetric equilibrium in which players do not communicate and play the single-agent solution.

Proposition 5 (Symmetric equilibrium)**.** The "Paying to give information" game has a unique symmetric equilibrium in Markov strategies, in which players do not communicate and experiment like the single-agent:

For any p, $q^*(p) = 0$ and $k^*(p) = 0$ for $p \geq \underline{p}$ and $k^*(p) = 1$ for $p > \underline{p}$.

Proof of Proposition 5. For any q_j, k_j, $u(p) = s + q_j k_j b(p,u) + \max_{k_i,q_i}\{-q_i c + k_i(b(p,u) - c(p)\} = s + q_j k_j b(p,u) + \max_{k_i}\{k_i(b(p,u) - c(p)\}$. Therefore, at any equilibrium, whether symmetric or asymmetric, i's continuation payoff is $u(p) = s + \max_{k_i}\{k_i(b(p,u) - c(p)\}$. i faces the single-agent problem, and plays $k^*_{sa}(p)$. □

Therefore, there is no equilibrium in Markov strategies, namely when individual actions only depend on individual beliefs in which players communicate. However, we know from the analysis of the two previous settings, PTEI and PTBI, that players would be better off for some intermediate beliefs if they could receive information from the other at a cost c. We then study the possibility for players to coordinate on some "communication phases" by using strategies that depend on their belief *and* on time, that we may call *non-stationary Markov strategies*.

A simple backward induction argument proves that, even in non-stationary Markov strategies, there can be no equilibrium, whether symmetric or asymmetric, in which players communicate.

Proposition 6. There is no equilibrium in non-stationary Markov strategies in which players communicate.

Proof of Proposition 6. Suppose that players know that they stop communicating at some date t. So at $t-dt$, player 1 will not communicate since he earns c by doing so. Consequently player 2 stops also communicating at $t - dt$. By continuing backward, it is easy to show that both players never communicate. □

The crucial difference with the PTEI game is that at symmetric equilibrium, player 1 communicates today not for having player 2's information today, but for having it tomorrow. Therefore, a player will communicate at some date t only if it might provide him with valuable information at date $t + dt$. Suppose that at equilibrium, players communicate for beliefs in $[p_1, p_2]$. Therefore, the impossibility of communication at equilibrium follows from the absence of *encouragement effect* in the exponential bandit model. By this effect, first analyzed by Bolton and Harris [1999], the presence of a player encourages the other to experiment at beliefs more pessimistic than the single-agent cut-off belief $\underline{p}$. We now that for $p < \underline{p}$, players will not experiment, and consequently will not communicate. Then there exists some cut-off $\widetilde{p} \geq \underline{p}$ such that players will stop communicating for $p \leq \widetilde{p}$. Suppose that i is the last player to communicate. He will do so if it might encourage his opponent to experiment and communicate about this experimentation, and if this additional information is valuable to him. Yet the only way to make his opponent experiment more is to have a breakthrough. But in this case, there is no more uncertainty and the additional information is of no value to player i.

7 Welfare results

In this section we study the welfare properties of costly communication following three different approaches: first in terms of amount of experimentation, then in terms of intensity of experimentation, and finally in terms of payoffs. We give the results for games in which players communicate for some beliefs at equilibrium, namely the PTEI and PTBY games.

7.1 Amount of experimentation

The first question we address is that of the impact of communication in terms of amount of experimentation, defined as $\hat{K} = \int_0^\infty K_t dt$. The amount of information measures how much of the resource is allocated to risky arms overall up to time ∞. The next proposition states that the higher the communication cost c, the higher the amount of experimentation at equilibrium.

Proposition 7. The amount of experimentation in the equilibrium of the game increases with c, and is maximal for $c \geq c_{max}$.

Proof of Proposition 7. From the dynamics of beliefs, we know that $K_t = -\frac{dp}{\lambda(1-p)p}dt$ if players communicate, and that $K_t = -\frac{2dp}{\lambda(1-p)p}dt$ if they don't. If $c < c_{max}$ and $p_0 > \overline{a}(c)$, then the amount of experimentation at equilibrium is then:

$$\begin{aligned}
\widehat{K} &= \int_0^\infty K_t dt \\
&= \int_{p_0}^{\overline{a}(c)} -\frac{dp}{\lambda(1-p)p} + \int_{\overline{a}(c)}^{\underline{a}(c)} -\frac{2dp}{\lambda(1-p)p} + \int_{\underline{a}(c)}^{\underline{p}^*} -\frac{dp}{\lambda(1-p)p} \\
&= \tfrac{1}{\lambda}\left(\ln(\underline{a}(c)) - \ln(\overline{a}(c)) + \ln(\underline{p}^*) - \ln(p_0)\right) \\
&> \tfrac{1}{\lambda}\left(\ln(p_1) - \ln(p_0)\right)
\end{aligned}$$

$\ln(\underline{a}(c)) - \ln(\overline{a}(c))$ increases with c as long as $c < c_{max}$, and is equal to $\ln(\underline{p}^*) - \ln(p_0)$ for all $c \geq c_{max}$. Therefore, the amount of experimentation is maximal if players do not communicate, that is for $c \geq c_{max}$. □

This result shows that, quite intuitively, making communication costly reduces free-riding. Indeed, free-riding comes from the fact that players may learn information from the experimentation of others, through communication. Obviously, making communication costly tends to reduce the exchange of information at equilibrium, and then reduces the possibility of free-riding. Therefore, if the objective is to fight free-riding behaviors, an extreme and efficient way is to impose a communication cost high enough to deter communication.

However, why would a social planner want to maximize the amount of experimentation? A somehow more important welfare issue for a social planner is that players make the right decision, that is play R if the risky arm is good, and S otherwise. From this point of view, the amount of information is not the relevant criterium to maximize. What matters is that players learn fast, so that they stop to experiment quickly if the risky arm is bad. We now study how the speed of learning depends on the communication cost, and we show that deterring communication to take place is not efficient.

7.2 Speed of learning

Suppose that the risky arm is bad. The best action for players is then to play the safe arm. Starting with a prior belief p_0 above the single-agent cut-off $\underline{p}$, they will start the game in experimenting. Since R is bad, they will never observe a breakthrough, and their belief will continuously decrease with time, until it reaches the belief threshold $\underline{p}$ under which they will stop experimenting. Obviously, in this situation, the sooner they stop experimenting, the better. Let us call $T(p_0)$ the time at which the common belief reaches $\underline{p}$ starting at p_0. The speed of learning is measured by $T(p_0)$, the smaller being $T(p_0)$, the faster the learning. KRC show that if the prior belief p_0 is above $\underline{p}$, then players' common posterior belief never reaches $\underline{p}$ at equilibrium. In other words, if the risky arm is bad, and

if players start with a prior belief high enough to make them experimenting, they will never stop making the wrong action. We first show that if no breakthrough happens, which is the case if the risky arm is bad, then for any $c > 0$, individual beliefs reach $\underline{p}$ in finite time.

Proposition 8. Let $p_0 > \underline{p}$. $T(p_0) < \infty$ except if a breakthrough appears iff $c > 0$.

Proof of Proposition 8. If $c = 0$, then the setting is that of KRC, and $T(p_0) = \infty$. Suppose now that $c > 0$. At equilibrium, the dynamics of beliefs is given by

$$dp = \begin{cases} 0 & if\ p \leq \underline{p} \\ -\lambda p(1-p)dt & if\ p \in [\underline{p}, \underline{a}(c)] \\ -\lambda 2k(p)p(1-p)dt & if\ p \in [\underline{a}(c), \overline{p}(c)] \\ -\lambda 2p(1-p)dt & if\ p \in [\overline{p}(c), \overline{a}(c)] \\ -\lambda p(1-p)dt & if\ p > \overline{a}(c)] \end{cases}$$

with $k(p) = \frac{W(p)-s+c}{s-gp}$.

Let us show that for any prior belief p_0, p_t reaches $\underline{p}$ in finite time if no breakthrough occurs.

- Suppose that $p_0 > \overline{a}(c)$ and let us show that p_t reaches $\overline{a}(c)$ in finite time. The dynamics of beliefs is $dp = -2\lambda p(1-p)dt$, thus $p_t = \frac{1}{1+\Omega(p_0)e^{2\lambda t}}$. Then $p_t = \overline{p}(c) \Leftrightarrow t = \frac{1}{2\lambda} \ln\left(\frac{\Omega(\overline{p}(c))}{\Omega(p_0)}\right) < \infty$. By the same argument, if $p_0 \in [\underline{p}, \underline{a}(c)]$, the dynamics of beliefs is $dp = -\lambda p(1-p)dt$, thus $p_t = \underline{p} \Leftrightarrow t = \frac{1}{\lambda} \ln\left(\frac{\Omega(\underline{p})}{\Omega(p_0)}\right) < \infty$.

- Suppose that $p_0 \in [\overline{p}(c), \overline{a}(c)]$ and let us show that p_t reaches $\overline{p}$ in finite time. The dynamics of beliefs is $dp = -2\lambda p(1-p)dt$, thus $p_t = \frac{1}{1+\Omega(p_0)e^{2\lambda t}}$. Then $p_t = \overline{a}(c) \Leftrightarrow t = \frac{1}{2\lambda} \ln\left(\frac{\Omega(\overline{p}(c))}{\Omega(p_0)}\right) < \infty$.

- Suppose now that $p_0 \in [\underline{a}(c), \overline{p}(c)]$ and let us show that p_t reaches $\underline{a}(c)$ in finite time. The dynamics of beliefs is $dp = -2\lambda h(p)dt$, with $h(p) := k(p)p(1-p)$. We

have $h'(p) = W'(p)\frac{p(1-p)}{s-gp} + (W(p) - s + c)\frac{s-2ps+p^2g}{(s-pg)^2}$ and $h''(p) = W''(p)\frac{p(1-p)}{s-gp} + 2W'(p)\frac{s-2ps+p^2g}{(s-pg)^2} + (W(p) - s + c)\frac{2s(g-s)}{(s-pg)^3}$. Thus $h''(p) > 0$ and $h'(\underline{a}(c)) > 0$. Then there exists a line that cuts the graph of $h(p)$ at $p = \underline{a}(c)$ and which is strictly below for $p > \underline{a}(c)$. Formally, there exists $m > 0$ such that $h(p) > m(p - \underline{a}(c)) + h(\underline{a}(c))$ for any $p > \underline{a}(c)$.

Consider the dynamics $dp = -2\lambda(m(p - \underline{a}(c)) + h(\underline{a}(c)))dt$. We show easily that $p_t = p_0 e^{-2\lambda mt} + (\underline{a}(c) + \frac{h(\underline{a}(c))}{2\lambda m})(1 - e^{-2\lambda m})$. Therefore, the date t at which p_t reaches $\underline{a}(c)$ is given by $e^{2\lambda mt} = (\underline{a}(c) - p_0)\frac{2\lambda m}{h(\underline{a}(c))} + 1$ and is finite. Since this dynamics has a lower rate of decrease than the true dynamics of beliefs, it is proved that the dynamics of beliefs reaches $\underline{a}(c)$ in finite time.

□

The intuition of this result is the following. In both cases, free and costly information, players allocate a decreasing fraction of the resource to the risky arm as their belief decreases. The difference is that if information is free (KRC's setting), this fraction goes to 0 as the belief goes to $\underline{p}$ with smooth pasting, that is the speed at which the fraction goes to 0 decreases and tends to 0 as the belief tends to $\underline{p}$. If information is costly however, even for a very small cost, the speed at which individual beliefs decreases never tends to 0.

The next results states that there exists a communication cost $c^* \in]0, c_{max}[$ that maximizes the speed of learning.

Let $c_{1/2}$ be the communication cost for which $k(\underline{a}(c)) = \frac{1}{2}$. If the prior belief is below the prior under which players mix between the two arms for $c = c_{1/2}$, $\underline{p}(c_{1/2})$, then $c_{1/2}$ dominates $c = 0$ and $c = c_{max}$ with respect to the speed of learning.

Proposition 9. If $p_0 < \overline{p}(c_{1/2})$, then $T(p_0, c_{1/2}) < T(p_0, 0)$ and $T(p_0, c_{1/2}) < T(p_0, c_{max})$.

Proof of Proposition 9. Let $T_1(p_0, c)$ be the expected time spent in $[\underline{p}, \underline{a}(c_{1/2})]$, $T_2(p_0, c)$ in $[\underline{a}(c_{1/2}), \bar{p}(c_{1/2})]$, $T_3(p_0, c)$ in $[\bar{p}(c_{1/2}), \bar{p}(0)]$ and $T_4(p_0, c)$ in $[\bar{p}(c_{1/2}), p_0]$. Notice that $T(p_0, c) = \sum_{n=1}^{n=4} T_n(p_0, c)$.

The more dp decreases, the smaller is $T_n(p_0, c)$. On $[\bar{p}(c), \bar{p}(0)]$, when $c = 0$ and $c = c_{1/2}$, players follow the same strategy. So $T_4(p_0, c) = T_4(p_0, 0)$. When players experiment alone, that is when $c = c_{\max}$, dp decreases less rapidly. So $T_4(p_0, c) < T_4(p_0, c_{\max})$. On $[\bar{p}(c_{1/2}), \bar{p}(0)]$, if player $c = 0$ then $k_i = 0$ whereas $k = 1$ when $c = c_{1/2}$. So dp decreases faster when $c = c_{1/2}$. So $T_3(p_0, c_{1/2}) < T_3(p_0, 0)$. As $k = 1$ if $c = c_{1/2}$ and players exchange information then $T_3(p_0, c_{1/2}) < T_3(p_0, c_{\max})$. On $[\underline{a}(c_{1/2}), \bar{p}(c_{1/2})]$ players communicate if $c = 0$ or $c = c_{1/2}$ and $k_{c_{1/2}}(p) > k_0(p)$. So $T_2(p_0, c_{1/2}) < T_2(p_0, 0)$. If $c = c_{1/2}$, $k_i + k_j > 1$ so $T_2(p_0, c_{1/2}) < T_2(p_0, c_{\max})$. On $[p, \underline{a}(c_{1/2})]$, if $c = 0$ then $k_i + k_j > 1$ and if $c = c_{1/2}$ then $k_i = k_j = 1$. So $T_1(p_0, c_{1/2}) < T_1(p_0, 0)$. Finally when $c = c_{1/2}$on this interval players face a situation similar to the case when player experiment alone. So $T_4(p_0, c_{1/2}) = T_4(p_0, 0)$. □

This proposition shows that there exists some cost $c^* \in]0, c_{max}[$ for which the speed of learning is maximal. Let $p_0 > \overline{a}(c)$ and let us denote t_1, t_2, t_3, and t_4 the time spent to go from p_0 to $\overline{a}(c)$, from $\overline{a}$ to p_2, from $\overline{p}(c)$ to $\underline{a}(c)$, and from $\underline{a}$ to $\underline{p}$. The average speed of learning is $V(c) := \frac{p_0 - \underline{p}}{t_1 + t_2 + t_3 + t_4}$.

The optimal cost c^* maximizes $V(c)$. Since t_3 cannot be computed explicitly, we have not been able for now to compute c^*

Two types of mistakes can be made by players: experimenting when the risky arm is bad, and stopping experimentation when the risky arm is good. We showed that making communication costly reduces the occurrence of the first type of mistake. It may seem that it increases the occurrence of the second type. However, if beliefs never reach $\underline{p}$ when $c = 0$, it's basically because players tend to almost not experiment, so in fact players may

commit the second type of mistake even if $c = 0$.

7.3 Payoffs

We now turn to the question of the efficiency of communication in terms of individual payoffs.

Proposition 10. Individual expected payoffs decrease with c as long as $c < c_{max}$.

Proof of Proposition 10. Fix $c \in]0, c_{max}[$. For $p \in [\underline{p}, \underline{a}(c)]$, players get an expected payoff $V(p) = gp + (1-p)K_V\Omega(p)^\mu$ whereas they would get a payoff $W_0(p) = s + (1+\mu)(g-s) + \mu s(1-p)\ln(\Omega(p)) + C(1-p)$ if c were 0. Using the fact that K_V is defined by $V(\underline{p}) = s$ and $W(\underline{p}) = s$, we show that $V(p) \leq W_0(p)$ for all $p \geq \underline{p}$ showing that $W_0''(\underline{p}) > V''(\underline{p})$ and the convexity of W_0 and V.

Since $k(p)$ increases with c, $\overline{p}(c) < \overline{p}(0)$. For $p \in [\underline{a}(c), \overline{p}(c)]$, players get $W(p) = s + (1+\mu)(g-s) + \mu s(1-p)\ln(\Omega(p)) + K_W(1-p)$ and would get a payoff $W_0(p) = s + (1+\mu)(g-s) + \mu s(1-p)\ln(\Omega(p)) + C(1-p)$ if c were 0. Since $W_0(\underline{p}) = s$, we have $W(p) \leq W_0(p)$ for all $p \in [\underline{a}(c), \overline{p}(c)]$.

For $p \in [\overline{p}(c)), \overline{p}(0)]$, players get a payoff $V_1(p)$ which is smaller than $W(p)$ and then smaller than $W_0(p)$ that players would get if c were 0.

For $p \in [\overline{p}(0), \overline{a}(c)]$, players get a payoff $V_1(p) = gp + K_1(1-p)\Omega(p)^{\mu/2} - c(1 - \frac{2p}{2+\mu})$ and would get $gp + C(1-p)\Omega(p)^\mu 2$ if c were 0. They gain from more information without paying the cost.

For $p > \overline{a}(c)$, players get $V(p) = gp + (1-p)K_V\Omega(p)^\mu$, and would get $gp + C(1-p)\Omega(p)^\mu 2$ if c were 0. □

This means that free-riding doesn't affect individual payoffs.

8 Discussion

Information structure Exponential bandit model are somehow specific: in many situations learning does happen through a breakthrough but rather through a gradual process. For instance, suppose that a farmer tests a new pesticide whose efficiency is unknown. An increase in the quantity or quality of his production does not imply that the pesticide is efficient. Other factors, like weather or human effort, may partly explain the observed growth of the production. The observation of a higher quality just means that it is more likely that the pesticide is efficient, and the farmer can learn more about the degree of efficiency of the pesticide in repeating experimentations. Keller and Rady (2010) analyze a bandit that takes into account such situations: they assume that the lump sums, which follow a Poisson distribution, are more frequent when the arm is good. The equilibria of this paper share many qualitative features with KRC. For instance, in the symmetric equilibrium players experiment for high beliefs, take the safe arm for low beliefs and mix for the intermediate beliefs. A huge difference is that encouragement effects occurs.

What would be the consequences of introducing costly information transfers in Keller and Rady (2010)? We may expect that costly transfers reduce the encouragements effect. Indeed, costly transfers of information reduce the benefit generated by the information externalities. Indeed experimenting in order to induce the other to do so is more costly, since the experimenter has to pay to send information.

Irreversibility Irreversibility of the safe action is a relatively frequent assumption in the bandit literature (c.f. Rosenberg, Solan and Vieille (2007) and Murto and Välimäki (2009)). What would be the impact of this assumption in a setting of costly information transfers? Irreversibility of communication would not change our results: in all the equilibria we exhibit, communication strategies are threshold strategies. In other words players

play as if communication were irreversible. Irreversibility of experimentation may increase the amount of experimentation.

Homogeneity of communication costs We assume that the communication cost is the same across players. Does a difference in costs modify the results? Clearly in the case PTEI and PTGI there is no hope for asymmetric equilibrium to emerge. The structure of symmetric equilibria may not be affected by an heterogeneity in communication costs. It is plausible that the equilibria would be those of the homogeneous case with a communication cost equal to the highest cost of both players.

Type of messages An important assumption is that players can not lie on the information they send. Either they show their actions and outcomes, or they don't. An alternative way of modelling communication could be to assume that players can send *messages* to each other chosen in a set of arbitrarily many messages (as in cheap-talk models, but at some cost c). What would be the structure of information at equilibrium? Would it be dominant for all players to send the same message as a player who would have receive a breakthrough? Would a perfectly revealing equilibrium exist, where players' messages reveal their type, namely the history of their actions and outcomes? This kind of communication introduces a new difficulty since players' beliefs may differ at some dates. We let this work for latter research.

References

[1] Bergemann D., Välimäki J., (1996), Learning and strategic pricing, *Econometrica*, **64**, 1125-49.

[2] Bergemann D., Välimäki J., (1997), Market diffusion with two-sided learning, *RAND J. Econo.*, 28, 773-795.

[3] Bergemann D., Välimäki J., (2000), Experimentation in markets, *Review of Economic Studies*, **67**, 213-234.

[4] Bergemann D., Välimäki J., (2006), Bandit problems, *Cowles Foundation Discussion Paper n 1551.*

[5] Bolton P., Harris C., (1999), Strategic experimentation, *Econometrica*, **67**, 349-374.

[6] Bonatti A., Hörner J., (2010), Collaborating, *Cowles Foundation Paper 1695*, forthcoming in *American Economic Review.*

[7] Hörner J., Samuelson L., (2010), Incentives for experimenting agents, *Cowles Foundation Discussion Paper n 1726.*

[8] Keller G., Rady S., Cripps M., (2005), Strategic experimentation with exponential bandits, *Econometrica*, **73**, 39-68.

[9] Keller G., Rady S., (2010), Strategic experimentation with Poisson bandits, *Theoretical Economics, Econometric Society*, **5**, 275-311

[10] Murto P., Välimäki J., (2009), Delay and information aggregation in stopping games with private information.

[11] Murto P., Välimäki J., (2009), Learning and information aggregation in an exit game.

[12] Rosenberg D., Solan E., Vieille N., (2007), Social learning in one-arm Bandit problems, *Econometrica*, **75**, 1591-1611.