Chapter 4
Learning in a Game of Strategic Experimentation with Three-Armed Exponential Bandits

Nicolas Klein

Abstract The present article provides some additional results for the two-player game of strategic experimentation with three-armed exponential bandits analyzed in Klein (Games Econ Behav 82:636–657, 2013). Players play replica bandits, with one safe arm and two risky arms, which are known to be of opposite types. It is initially unknown, however, which risky arm is good and which is bad. A good risky arm yields lump sums at exponentially distributed times when pulled. A bad risky arm never yields any payoff. In this article, I give a necessary and sufficient condition for the state of the world eventually to be found out with probability 1 in *any* Markov perfect equilibrium in which at least one player's value function is continuously differentiable. Furthermore, I provide closed-form expressions for the players' value function in a symmetric Markov perfect equilibrium for low and intermediate stakes.

4.1 Introduction

Think of a situation in which agents are initially uncertain about some payoff-relevant aspect of their environment. Yet, they can learn about it over time by exploring different options. Thus, a farmer may not know the yield of a new crop before trying it out. Trying it out implies an opportunity cost, however, as using his field to try the new crop means that he cannot use it to plant a traditional crop, whose yield he already knows. The trade-off he faces is thus between optimally using the information he already has (*exploitation*) and investing resources in order to acquire new information, which will potentially be useful to him in the future (*exploration*).

N. Klein (✉)
Université de Montréal and CIREQ, Département de Sciences Économiques, Montréal, QC, Canada

L. A. Petrosyan et al. (eds.), *Frontiers of Dynamic Games*,
Static & Dynamic Game Theory: Foundations & Applications,
https://doi.org/10.1007/978-3-319-92988-0_4

The so-called *multi-armed bandit model* has become canonical in economics to analyze a decision maker's trade-off between exploration and exploitation.[1]

But now suppose that our farmer has a neighbor and that he can observe the kind of crop planted by his neighbor, as well as its yield. Our farmer would of course prefer that his neighbor experiment with the new crop, as this would allow him to get some information about it without having to bear the (opportunity) cost of producing the information himself. Of course, his neighbor faces precisely the same trade-off, and the informational externality leads to a situation of strategic interaction. Such *strategic* bandit problems have been introduced by Bolton and Harris [2, 3], where players choose between a risky option and a safe one. Here, I use the exponential-bandits variant introduced by Keller et al. [6], and, in particular, adopt the three-armed model of Klein [7].

While in [2, 3] and in [6], the risky option was of the same quality for all players, Klein and Rady [8] introduced negative correlation between players: what was good news to one player was bad news to the other. In [7], I have introduced a setting in which two players have access to *two* risky arms of perfectly negatively correlated types. The comparison of the results in [8] and [7] in particular thus allow for the analysis of the impact of delegating project choice to individual agents.

For the case of perfectly positively correlated two-armed bandits, Keller et al. [6] show that players experiment inefficiently little in equilibrium, as compared to the cooperative benchmark. Indeed, the information players produce is a public good; hence they produce too little of it. Indeed, they both give up on finding out the state of the world too soon (i.e., the *amount* of experimentation is too low) and they learn too slowly (i.e., the *intensity* of experimentation will be inefficiently low). By contrast, Klein and Rady [8] find that, with perfectly negatively correlated two-armed bandits, the *amount* of experimentation is always at the efficient level. Furthermore, there exists an efficient equilibrium if and only if the stakes at play are *below* a certain threshold. By contrast, in [7], I show that, when both agents have access to two perfectly negatively correlated risky arms, there exists an efficient equilibrium if and only if the stakes at play *exceed* a certain threshold. In the present article, I provide closed-form expressions for the players' value function in a symmetric Markov perfect equilibrium for the cases in which there does not exist an efficient equilibrium. Furthermore, I give a necessary and sufficient condition for learning to be complete, i.e. for the state of the world to be found out with probability 1, in *any* Markov perfect equilibrium in which at least one player's value function is continuously differentiable.

The rest of this article is organised as follows. Section 4.2 explains the model setup; Sect. 4.3 analyzes conditions under which complete learning will prevail; Sect. 4.4 analyzes equilibrium for low and intermediate stakes, while Sect. 4.5 concludes. Formal proofs are collected in Sect. 4.6.

[1]The multi-armed bandit model was first introduced by Thompson [10] and Robbins [9], and subsequently analyzed, amongst others, by Bradt et al. [4] and Bellman [1]. Gittins and Jones [5] provided the famous Gittins-index characterization of an optimal policy.

4.2 Model Setup

The setup is as in [7]: There are two agents playing a three-armed bandit in continuous time each. One arm is safe in that it yields a known flow payoff of $s > 0$ when pulled; the other two arms, A and B, are risky in that they can be either good or bad. It is known that exactly one of the two risky arms is good and that the same risky arm is good for both players. Which between arms A and B is good and which is bad is initially unknown. The good risky arm yields lump sums $h > 0$ at exponentially distributed times with parameter $\lambda > 0$, when it is pulled. The bad risky arm always yields 0. The parameters λ, s and h, are common knowledge among the players. I assume that $g := \lambda h > s > 0$.

More specifically, either player $i \in \{1, 2\}$ can decide in continuous time how to distribute a unit endowment flow over the three arms of his bandit; i.e., at each instant $t \in \mathscr{R}_+$, he chooses $(k_{i,A}, k_{i,B}) \in \{(a, b) \in [0, 1]^2 : a + b \leq 1\}$, where $k_{i,A}(t)$ ($k_{i,B}(t)$) denotes the fraction of the unit endowment flow player i devotes to arm A (B) at instant t.

Players start out from a common prior $p_0 \in (0, 1)$ that it is their risky arms A that are good. As everyone's action choices, as well as the outcomes of these action choices, are perfectly publicly observable, there is no private information at any time. Thus, players will share a common posterior belief that it is their risky arms A that are good at all times $t \geq 0$. We shall denoted by p_t this belief at instant t. As only a good risky arm can ever yield a lump-sum payoff, $p_\tau = 1$ ($p_\tau = 0$) at all times $\tau > t$ if either player has received a lump sum from arm A (B) at time t. If no such breakthrough has occurred yet by time t, the belief satisfies

$$p_t = \frac{p_0 e^{-\lambda \int_0^t (k_{1,A}(\tau)+k_{2,A}(\tau))\, d\tau}}{p_0 e^{-\lambda \int_0^t (k_{1,A}(\tau)+k_{2,A}(\tau))\, d\tau} + (1 - p_0) e^{-\lambda \int_0^t (k_{1,B}(\tau)+k_{2,B}(\tau))\, d\tau}}. \tag{4.1}$$

Following much of the literature, I focus on Markov perfect equilibria with the common posterior belief p_t as the state variable (which I shall sometimes simply refer to as *equilibrium*). A Markov strategy for player i is a time-invariant, piecewise continuous, function $(k_{i,A}, k_{i,B}) : [0, 1] \to \{(a, b) \in [0, 1]^2 : a + b \leq 1\}$, $p_t \mapsto (k_{i,A}, k_{i,B})(p_t)$. As in [8], a *pair* of Markov strategies is said to be *admissible* if there exists a solution to the corresponding law of motion of beliefs (derived from Bayes' rule) that coincides with the limit of the unique discrete-time solution. An inadmissible strategy pair is assumed to give both players a payoff of $-\infty$.

Players discount payoffs at the common discount rate $r > 0$. An admissible strategy pair $((k_{1,A}, k_{1,B}), (k_{2,A}, k_{2,B}))$ induces a payoff function u_i for players $i \in \{1, 2\}$, which is given by

$$u_i(p) = \mathscr{E}\left[\int_0^\infty re^{-rt} \left\{(k_{i,A}(p_t)p_t + k_{i,B}(p_t)(1 - p_t))g \right.\right.$$
$$\left.\left. + [1 - k_{i,A}(p_t) - k_{i,B}(p_t)]s\right\} dt \,\middle|\, p_0 = p\right], \tag{4.2}$$

where the expectation is taken with respect to the process of beliefs $\{p_t\}_{t\in\mathscr{R}_+}$. Player i's objective is to maximize u_i. As one can see immediately from player i's objective, the other player's actions impact u_i only via the players' common belief process $\{p_t\}_{t\in\mathscr{R}_+}$; i.e., ours is a game of purely informational externalities.

I say that the stakes are *high* if $\frac{g}{s} \geq \frac{4(r+\lambda)}{2r+3\lambda}$; they are *intermediate* if $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$; they are *low* if $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, and *very low* if $\frac{g}{s} < \frac{2(r+\lambda)}{r+2\lambda}$. It is immediate to verify that the stakes are low if and only if $p_1^* := \frac{rs}{(r+\lambda)(g-s)+rs} \geq \frac{1}{2}$; they are very low if and only if $p_2^* := \frac{rs}{(r+2\lambda)(g-s)+rs} \geq \frac{1}{2}$.

Klein [7, Section 4] shows that the utilitarian planner's solution has a bang-bang structure.[2] If the stakes at play are not very low, the planner would always use the risky arm that looks momentarily the most promising; he would never use the safe arm. This means that learning will be complete, i.e. the true state of the world will be found out with probability 1. If the stakes are very low, by contrast, the planner would use the safe arm for all beliefs in $[1-p_2^*, p_2^*]$ and the risky arm that looks momentarily the most promising for all other beliefs. Thus, learning will be incomplete in this case. A single player acting in isolation would optimally pursue the same policy, with p_1^* replacing p_2^*, and "low stakes" replacing "very low stakes," in the previous statements.

4.3 Complete Learning

As already mentioned in the introduction, Keller et al. [6] identified two dimensions of inefficiency in their model: On the one hand, players give up on finding out about the true state of the world too soon, i.e. the experimentation *amount* is inefficiently small. On the other hand, players also learn too slowly, i.e. the experimentation *intensity* is inefficiently low. If one were merely to focus on the long-run properties of learning, only the former effect would be of interest. Keller et al. [6] show that, because of the informational externalities, all experimentation stops at the single-agent cutoff belief in any equilibrium; the efficient cutoff belief would be more pessimistic, though, as it takes into account that the information a player generates benefits the other players also.[3] Furthermore, learning is always incomplete, i.e.

[2] The utilitarian planner maximizes the sum of the players' utilities. The solution to this problem is the policy the players would want to commit to at the outset of the game if they had commitment power. It thus constitutes a natural efficient benchmark against which to compare our equilibria.

[3] By contrast, Bolton and Harris [2] identified an *encouragement effect* in their model. It makes players experiment at beliefs that are more pessimistic than their single-agent cutoffs. This is because they will receive good news with some probability, which will make the other players more optimistic also. This then induces them to provide more experimentation, from which the first player then benefits in turn. With fully revealing breakthroughs as in [6, 8], or this model, however, a player could not care less what others might do after a breakthrough, as there will not be anything left to learn. Therefore, there is no encouragement effect in these models.

there is a positive probability that the truth will never be found out.[4] In [8], however, the *amount* of experimentation is always at the efficient level.[5] This is because both players cannot be exceedingly pessimistic at the same time. Indeed, as soon as players' single-agent cutoffs overlap, at any possible belief at least one of them is loath to give up completely, although players may not be experimenting with the enthusiasm required by efficiency. In particular, learning will be complete in any equilibrium if and only if efficiency so requires.

This section will show that which of these effects prevails here depends on the stakes at play: If stakes are so high that the single-agent cutoffs overlap, players would not be willing ever completely to give up on finding out the true state of the world even if they were by themselves. Yet, since all a player's partner is doing is to provide him some additional information for free, a player should be expected to do at least as well as if he were by himself. Hence, the Klein and Rady [8] effect obtains if players' single-agent cutoffs overlap, and, in any equilibrium (in which at least one player's value function is smooth),[6] the true state of the world will eventually be found out with probability 1 (i.e. learning will be *complete*), as efficiency requires. In the opposite case, however, the informational externality identified by Keller et al. [6] carries the day, and, as we will see in the next section, there exists an equilibrium entailing an inefficiently low amount of experimentation. For some parameters, this implies incomplete equilibrium learning while efficiency calls for complete learning.

To state the next lemma, I write u_1^* for the value function of a single agent operating a bandit with only a safe arm and a risky arm A, while I denote by u_2^* the value function of a single agent operating a bandit with only a safe arm and a risky arm B. It is straightforward to verify that $u_2^*(p) = u_1^*(1-p)$ for all p and that[7]

$$u_1^*(p) = \begin{cases} s & \text{if } p \leq p_1^*, \\ g\left[p + \frac{\lambda p_1^*}{\lambda p_1^* + r}(1-p)\left(\frac{\Omega(p)}{\Omega(p_1^*)}\right)^{\frac{r}{\lambda}}\right] & \text{if } p > p_1^* \end{cases}, \tag{4.3}$$

[4]The efficient solution in [6] also implies incomplete learning.

[5]For perfect negative correlation, this is true in any equilibrium; for general negative correlation, there always exists an equilibrium with this property.

[6]The technical requirement that at least one player's value function be continuously differentiable is needed on account of complications pertaining to the admissibility of strategies. I use it in the proof of Lemma 4.1 to establish that the safe payoff s constitutes a lower bound on the player's equilibrium value. However, by e.g. insisting on playing $(1, 0)$ at a single belief $\hat{p}$ while playing $(0, 0)$ everywhere else in a neighborhood of $\hat{p}$, a player could e.g. force the other player to play $(0, 1)$ at $\hat{p}$ for mere admissibility reasons. Thus, both players' *equilibrium* value functions might be pushed below s at certain beliefs $\hat{p}$. For the purposes of this section, I rule out such implausible behavior by restricting attention to equilibria in which at least one player's value function is smooth.

[7]See Prop.3.1 in [6].

where $\Omega(p) := \frac{1-p}{p}$ denotes the odds ratio. The following lemma tells us that u_1^* and u_2^* are both lower bounds on a player's value in *any* equilibrium, provided his value is smooth.

Lemma 4.1 (Lower Bound on Equilibrium Payoffs) *Let $u \in C^1$ be a player's equilibrium value function. Then, $u(p) \geq \max\{u_1^*(p), u_2^*(p)\}$ for all $p \in [0, 1]$.*

The intuition for this result is very straightforward. Indeed, there are only informational externalities, no payoff externalities, in our model. Thus, intuitively, a player can only benefit from any information his opponent provides him for free; therefore, he should be expected to do at least as well as if he were by himself, forgoing the use of one of his risky arms to boot.

Now, if $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, then $p_1^* < \frac{1}{2} < 1 - p_1^*$, so at any belief p, we have that $u_1^*(p) > s$ or $u_2^*(p) > s$ or both. Thus, there cannot exist a p such that $(k_{1,A}, k_{1,B})(p) = (k_{2,A}, k_{2,B})(p) = (0, 0)$ be mutually best responses as this would mean $u_1(p) = u_2(p) = s$. This proves the following proposition:

Proposition 4.1 (Complete Learning) *If $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$, learning will be complete in any Markov perfect equilibrium in which at least one player's value function is continuously differentiable.*

It is the same threshold $\frac{2r+\lambda}{r+\lambda}$ above which complete learning is efficient, and prevails in any equilibrium, in the perfectly negatively correlated two-armed bandit case.[8] In our setting, however, complete learning is efficient for a larger set of parameters, as we saw in Sect. 4.2. In the following section, I shall proceed to a more thorough analysis of the strategic problem.

4.4 Equilibrium Payoff Functions

In [7], I have shown that there exists an efficient equilibrium in this model if and only if the stakes are high. The purpose of this section is to construct a symmetric equilibrium for those parameter values for which there does not exist an efficient equilibrium. I define symmetry in keeping with [2] as well as [6]:

Definition 4.1 An equilibrium is said to be *symmetric* if equilibrium strategies $((k_{1,A}, k_{1,B}), (k_{2,A}, k_{2,B}))$ satisfy $(k_{1,A}, k_{1,B})(p) = (k_{2,A}, k_{2,B})(p)$ for all $p \in [0, 1]$.

As a matter of course, in any symmetric equilibrium, $u_1(p) = u_2(p)$ for all $p \in [0, 1]$. I shall denote the players' common value function by u. By the same token, I shall write $k_{1,A} = k_{2,A} = k_A$ and $k_{1,B} = k_{2,B} = k_B$.

[8] See Proposition 8 in [8].

4.4.1 Low Stakes

Recall that the stakes are low if, and only if, the single-agent cutoffs for the two risky arms do not overlap. It can be shown that in this case there exists an equilibrium that is essentially two copies of the Keller et al. [6] symmetric equilibrium (see their Proposition 5.1), mirrored at the $p = \frac{1}{2}$ axis.

Proposition 4.2 (Symmetric MPE for Low Stakes) *If $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, there exists a symmetric equilibrium where both players exclusively use the safe arm on $[1 - p_1^*, p_1^*]$, the risky arm A above the belief $\hat{p} > p_1^*$, and the risky arm B at beliefs below $1 - \hat{p}$, where $\hat{p}$ is defined implicitly by*

$$\Omega(p^m)^{-1} - \Omega(\hat{p})^{-1} = \frac{r+\lambda}{\lambda}\left[\frac{1}{1-\hat{p}} - \frac{1}{1-p_1^*} - \Omega(p_1^*)^{-1}\ln\left(\frac{\Omega(p_1^*)}{\Omega(\hat{p})}\right)\right]. \tag{4.4}$$

In $[p_1^*, \hat{p}]$, the fraction $k_A(p) = \frac{u(p)-s}{c_A(p)}$ is allocated to risky arm A, while $1 - k_A(p)$ is allocated to the safe arm; in $[1 - \hat{p}, 1 - p_1^*]$, the fraction $k_B(p) = \frac{u(p)-s}{c_B(p)}$ is allocated to risky arm B, while $1 - k_B(p)$ is allocated to the safe arm.

Let $V_h(p) := pg + C_h(1-p)\Omega(p)^{\frac{r}{2\lambda}}$, and $V_l(p) := (1-p)g + C_l p\Omega(p)^{-\frac{r}{2\lambda}}$. Then, the players' value function is continuously differentiable, and given by $u(p) = W(p)$ if $1 - \hat{p} \leq p \leq \hat{p}$, where $W(p)$ is defined by

$$W(p) := \begin{cases} s + \frac{r}{\lambda}s\left[\Omega(p_1^*)^{-1}\left(1 - \frac{p}{p_1^*}\right) - p\ln\left(\frac{\Omega(p)}{\Omega(p_1^*)}\right)\right] & \text{if } 1-\hat{p} < p < 1-p_1^* \\ s \text{ if } 1-p_1^* \leq p \leq p_1^* & \\ s + \frac{r}{\lambda}s\left[\Omega(p_1^*)\left(1 - \frac{1-p}{1-p_1^*}\right) - (1-p)\ln\left(\frac{\Omega(p_1^*)}{\Omega(p)}\right)\right] & \text{if } p_1^* < p < \hat{p} \end{cases}; \tag{4.5}$$

$u(p) = V_h(p)$ if $\hat{p} \leq p$, while $u(p) = V_l(p)$ if $p \leq 1 - \hat{p}$, where the constants of integration C_h and C_l are determined by $V_h(\hat{p}) = W(\hat{p})$ and $V_l(1-\hat{p}) = W(1-\hat{p})$, respectively.

Thus, in this equilibrium, even though either player knows that one of his risky arms is good, whenever the uncertainty is greatest, the safe option is attractive to the point that he cannot be bothered to find out which one it is. When players are relatively certain which risky arm is good, they invest all their resources in that arm. When the uncertainty is of medium intensity, the equilibrium has the flavor of a mixed-strategy equilibrium, with players devoting a uniquely determined fraction of their resources to the risky arm they deem more likely to be good, with the rest being invested in the safe option. As a matter of fact, the experimentation intensity decreases continuously from $k_A(\hat{p}) = 1$ to $k_A(p_1^*) = 0$ (from $k_B(1-\hat{p}) = 1$ to $k_B(1-p_1^*) = 0$). Intuitively, the situation is very much reminiscent of the classical Battle of the Sexes game: If one's partner experiments, one would like to free-ride on his efforts; if one's partner plays safe, though, one would rather do the experimentation oneself than give up on finding out the truth. On the relevant

range of beliefs it is the case that as players become more optimistic, they have to raise their experimentation intensities in order to increase free-riding incentives for their partner. This is necessary to keep their partner indifferent, and hence willing to mix, over both options.

Having seen that for $\frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, there exists an equilibrium with smooth value functions that implies incomplete learning, we are now in a position to strengthen our result on the long-run properties of equilibrium learning:

Corollary 4.1 (Complete Learning) *Learning will be complete in any Markov perfect equilibrium in which at least one player's value function is smooth, if and only if* $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$.

For perfect negative correlation, Klein and Rady [8] found that with the possible exception of the knife-edge case $\frac{g}{s} = \frac{2r+\lambda}{r+\lambda}$, learning was going to be complete in any equilibrium if and only if complete learning was efficient. While the proposition pertains to the exact same parameter set on which complete learning prevails in [8], we here find by contrast that if $\frac{2(r+\lambda)}{r+2\lambda} < \frac{g}{s} \leq \frac{2r+\lambda}{r+\lambda}$, efficiency uniquely calls for complete learning, yet there exists an equilibrium entailing incomplete learning. This is because with three-armed bandits information is more valuable to the utilitarian planner, as in case of a success he gets the full payoff of a good risky arm. With negatively correlated two-armed bandits, however, the planner cannot shift resources between the two types of risky arm; thus, his payoff in case of a success is just $\frac{g+s}{2}$.

4.4.2 Intermediate Stakes

For intermediate stakes, the equilibrium I construct is essentially of the same structure as the previous one: It is symmetric and it requires players to mix on some interval of beliefs. However, there does not exist an interval where both players play safe, so that players will always eventually find out the true state of the world, even though they do so inefficiently slowly.

Proposition 4.3 (Symmetric MPE for Intermediate Stakes) *If* $\frac{2r+\lambda}{r+\lambda} < \frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$, *there exists a symmetric equilibrium. Let* $\check{p} := \frac{\lambda+r}{\lambda}(2p^m - 1)$, *and* $\mathscr{W}(p)$ *be defined by*

$$\mathscr{W}(p) := \begin{cases} s + \frac{r+\lambda}{\lambda}(g-s) - \frac{r}{\lambda} ps\,(2 + \ln(\Omega(p))) & \text{if } p \leq \frac{1}{2} \\ s + \frac{r+\lambda}{\lambda}(g-s) - \frac{r}{\lambda}(1-p)s\,(2 - \ln(\Omega(p))) & \text{if } p \geq \frac{1}{2} \end{cases} \tag{4.6}$$

Now, let $p_1^\dagger > \frac{1}{2}$ *and* $p_2^\dagger > \frac{1}{2}$ *be defined by* $\mathscr{W}(p_1^\dagger) = \frac{\lambda+r(1-p_1^\dagger)}{\lambda+r} g$ *and* $\mathscr{W}(p_2^\dagger) = 2s - p_2^\dagger g$, *respectively. Then, let* $p^\dagger := p_1^\dagger$ *if* $p_1^\dagger \geq \check{p}$; *otherwise, let* $p^\dagger := p_2^\dagger$.

In equilibrium, both players will exclusively use their risky arm A in $[p^\dagger, 1]$, *and their risky arm B in* $[0, 1-p^\dagger]$. *In* $]\frac{1}{2}, p^\dagger]$, *the fraction* $k_A(p) = \frac{\mathscr{W}(p)-s}{c_A(p)}$ *is allocated*

to risky arm A, while $1 - k_A(p)$ *is allocated to the safe arm; in* $[p^\dagger, \frac{1}{2}[$*, the fraction* $k_B(p) = \frac{\mathscr{W}(p)-s}{c_B(p)}$ *is allocated to risky arm B, while* $1 - k_B(p)$ *is allocated to the safe arm. At* $p = \frac{1}{2}$*, a fraction of* $k_A(\frac{1}{2}) = k_B(\frac{1}{2}) = \frac{(\lambda+r)g-(2r+\lambda)s}{\lambda(2s-g)}$ *is allocated to either risky arm, with the rest being allocated to the safe arm.*

Let $V_h(p) := pg + C_h(1-p)\Omega(p)^{\frac{r}{2\lambda}}$*, and* $V_l(p) := (1-p)g + C_l p\Omega(p)^{-\frac{r}{2\lambda}}$*. Then, the players' value function is continuously differentiable, and given by* $u(p) = \mathscr{W}(p)$ *in* $[1-p^\dagger, p^\dagger]$*, by* $u(p) = V_h(p)$ *in* $[p^\dagger, 1]$*, and* $u(p) = V_l(p)$ *in* $[0, 1-p^\dagger]$*, with the constants of integration* C_h *and* C_l *being determined by* $V_h(p^\dagger) = \mathscr{W}(p^\dagger)$ *and* $V_l(1-p^\dagger) = \mathscr{W}(1-p^\dagger)$*.*

Thus, no matter what initial prior belief players start out from, there is a positive probability that beliefs will end up at $p = \frac{1}{2}$, and hence they will try the risky project that looked initially less auspicious. Therefore, in contrast to the equilibrium for low stakes, there is a positive value attached to the option of having access to the second risky project.

4.5 Conclusion

I have analyzed a game of strategic experimentation with three-armed bandits, where the two risky arms are perfectly negatively correlated. In [7], I have shown that there exists an efficient equilibrium if and only if the stakes are high. Here, we have seen that any equilibrium in which at least one player's value is smooth involves complete learning if stakes are not low. If stakes are intermediate in size, all equilibria are inefficient, though they involve complete learning (provided both players' value functions are not kinked), as required by efficiency. If the stakes are low, all equilibria are inefficient, and there exists an equilibrium implying an inefficiently low amount of experimentation. In particular, if the stakes are low but not very low, there exists an equilibrium that involves incomplete learning while efficiency requires complete learning; if the stakes are very low, the efficient solution also implies incomplete learning.

From an economic point of view, the reason for the prevalence of free-riding in Markov perfect equilibrium when the types of the risky arms are perfectly positively correlated is as follows. If a player deviates by providing less effort than he is supposed to, the other players will be more optimistic than they should be as a result, and hence more willing to pick up the deviating player's slack. This makes players more inclined to free-ride. However, if players' risky arms are negatively correlated as in [8], it is impossible for both of them to be very pessimistic about their respective projects at the same time, and free-riding only appears if the players' respective single-agent cut-offs overlap. Otherwise, i.e., if the stakes are low, there exists an efficient equilibrium. By contrast, in our setting, there exists an efficient equilibrium if and only if the stakes are high [7], i.e. if and only if both players are always sufficiently optimistic about one of their projects. Otherwise, the positive correlation between players makes incentives for free-riding reappear.

4.6 Proofs

This section collects the proofs of our results. We note that player i's Bellman equation is given by (see [7])

$$u_i(p) = s + k_{j,A} B_A(p, u_i) + k_{j,B} B_B(p, u_i) + \max_{\{(k_{i,A}, k_{i,B}) \in [0,1]^2 : k_{i,A} + k_{i,B} \leq 1\}} \left\{ k_{i,A} \left[B_A(p, u_i) - c_A(p) \right] + k_{i,B} \left[B_B(p, u_i) - c_B(p) \right] \right\}, \tag{4.7}$$

where $\{j\} = \{1, 2\} \setminus \{i\}$, $B_A(p, u) := \frac{\lambda}{r} p[g - u(p) - (1-p)u'(p)]$ and $B_B(p, u) := \frac{\lambda}{r}(1-p)[g - u(p) - pu'(p)]$ measure the learning benefit from playing arm A and arm B, respectively, while $c_A(p) := s - pg$ and $c_B(p) := s - (1-p)g$ measure the appertaining myopic opportunity cost of doing so. A myopic player (i.e. a player whose discount rate $r \to \infty$) would use risky arm A (B) if and only if $c_A(p) > 0$ ($c_B(p) > 0$), i.e., if and only if $p > p^m := \frac{s}{g}$ ($p < 1 - p^m$).

Furthermore, we note for future reference (see Appendix A in [7]) that, on any open interval of beliefs on which $((1, 0), (1, 0))$ is played, both players' value functions satisfy the ODE

$$2\lambda p(1-p)u'(p) + (2\lambda p + r)u(p) = (2\lambda + r)pg. \tag{4.8}$$

On any open interval of beliefs at which a player is indifferent between his safe arm and his risky arm A, his value function satisfies the ODE

$$\lambda p(1-p)u'(p) + \lambda p u(p) = (\lambda + r)pg - rs. \tag{4.9}$$

4.6.1 Proof of Lemma 4.1

In a first step, I show that s is a lower bound on u. Assume to the contrary that there exists a belief $p^\dagger \in]0, 1[$ such that $u(p^\dagger) < s$. Then, since u is continuously differentiable and $u(0) = u(1) = g > s$, there exists a belief $\tilde{p} \in]0, 1[$ such that $u(\tilde{p}) < s$ and $u'(\tilde{p}) = 0$. I write B_A and B_B for $B_A(p, u)$ and $B_B(p, u)$, respectively, suppressing arguments whenever this is convenient. Moreover, I define $\hat{B}_A(p) := \frac{\lambda}{r} p(g - s) > 0$ and $\hat{B}_B(p) := \frac{\lambda}{r}(1-p)(g-s) > 0$, while denoting by $(k_{j,A}, k_{j,B})$ the other player's action at $\tilde{p}$ in the equilibrium underlying the value function u. Now, at $\tilde{p}$, $u < s$ immediately implies $B_A = \frac{\lambda}{r}\tilde{p}(g-u) > \hat{B}_A$ and $B_B = \frac{\lambda}{r}(1-\tilde{p})(g-u) > \hat{B}_B$, and we have that

$$u - s \geq k_{j,A}(B_A - \hat{B}_A) + k_{j,B}(B_B - \hat{B}_B) = (k_{j,A}\tilde{p} + k_{j,B}(1-\tilde{p}))(s-u) \geq 0, \tag{4.10}$$

a contradiction to $u < s$.[9] Thus, we have already shown that u_1^* bounds u from below at all beliefs $p \leq p_1^*$.

Now, suppose there exists a belief $p > p_1^*$ at which $u < u_1^*$. I now write $B_A^* := \frac{\lambda}{r} p[g - u_1^* - (1-p)(u_1^*)'(p)] = u_1^* - pg$ and $B_B^* := \frac{\lambda}{r}(1-p)[g - u_1^* + p(u_1^*)'(p)]$. Since $B_A^* + B_B^* = \frac{\lambda}{r}(g - u_1^*)$, and hence $B_B^* = \frac{\lambda}{r}(g - u_1^*) - (u_1^* - pg)$, we have that $B_B^* \geq 0$ if and only if $u_1^* \leq \frac{\lambda + rp}{\lambda + r} g =: w_1(p)$. Let $\tilde{p}$ be defined by $w_1(\tilde{p}) = s$; it is straightforward to show that $\tilde{p} < p_1^*$. Noting furthermore that $u_1^*(p_1^*) = s$, $w_1(1) = u_1^*(1) = g$, and that w_1 is linear whereas u_1^* is strictly convex in p, we conclude that $u_1^* < w_1$ and hence $B_B^* > 0$ on $[p_1^*, 1[$. Moreover, since $B_A^* \geq 0$ (see [6]), we have $u_1^* = pg + B_A^* \leq pg + k_{j,B} B_B^* + (1 + k_{j,A}) B_A^*$ on $[p^*, 1]$, for any $(k_{j,A}, k_{j,B})$.

Since s is a lower bound on u, by continuity, $u(p) < u_1^*(p)$ implies the existence of a belief strictly greater than p_1^* where $u < u_1^*$ and $u_1' \leq (u_1^*)'$. This immediately yields $B_A > B_A^* > c_A$, as well as

$$u - u_1^* \geq pg + k_{j,B} B_B + (1 + k_{j,A}) B_A - [pg + (1 + k_{j,A}) B_A^* + k_{j,B} B_B^*] \quad (4.11)$$

$$= k_{j,B}(B_A + B_B - B_A^* - B_B^*) + (1 + k_{j,A} - k_{j,B})(B_A - B_A^*) \quad (4.12)$$

$$= k_{j,B} \frac{\lambda}{r}(u_1^* - u_1) + (1 + k_{j,A} - k_{j,B})(B_A - B_A^*) > 0, \quad (4.13)$$

a contradiction.[10]

An analogous argument applies for u_2^*. □

4.6.2 *Proof of Proposition 4.2*

First, I show that $\hat{p}$ as defined in the proposition indeed exists and is unique in $]p_1^*, 1[$. It is immediate to verify that the left-hand side of the defining equation is decreasing, while the right-hand side is increasing in $\hat{p}$. Moreover, for $\hat{p} = p_1^*$, the left-hand side is strictly positive, while the right-hand side is zero. Now, for $\hat{p} \uparrow 1$, the left-hand side tends to $-\infty$, while the right-hand side is positive. The claim thus follows by continuity.

[9] Strictly speaking, the first inequality relies on the admissibility of the action $(0, 0)$ at $\tilde{p}$. However, even if $(0, 0)$ should not be admissible at $\tilde{p}$, my definition of strategies still guarantees the existence of a neighborhood of $\tilde{p}$ in which $(0, 0)$ is admissible everywhere except at $\tilde{p}$. Hence, by continuous differentiability of u, there exists a belief $\tilde{\tilde{p}}$ in this neighborhood at which the same contradiction can be derived.

[10] Again, strictly speaking, the first inequality relies on the admissibility of the action $(1, 0)$ at the belief in question, and my previous remark applies.

The proposed policies imply a well-defined law of motion for the posterior belief. It is immediate to verify that the function u satisfies value matching and smooth pasting at p_1^* and $1-p_1^*$. To show that it is continuously differentiable, it remains to be shown that smooth pasting is satisfied at $\hat{p}$ and $1-\hat{p}$. From the appertaining ODEs, we have that

$$\lambda\hat{p}(1-\hat{p})u'(\hat{p}-)+\lambda\hat{p}u(\hat{p})=(\lambda+r)\hat{p}g-rs \tag{4.14}$$

and

$$2\lambda\hat{p}(1-\hat{p})u'(\hat{p}+)+(2\lambda\hat{p}+r)u(\hat{p})=(2\lambda+r)\hat{p}g, \tag{4.15}$$

where I write $u'(\hat{p}-) := \lim_{p\uparrow\hat{p}} u'(p)$ and $u'(\hat{p}+) := \lim_{p\downarrow\hat{p}} u'(p)$. Now, $u'(\hat{p}-)=u'(\hat{p}+)$ if and only if $u(\hat{p})=2s-\hat{p}g$. Now, algebra shows that indeed $W(\hat{p})=2s-\hat{p}g$. By symmetry, we can thus conclude that $W(1-\hat{p})=2s-(1-\hat{p})g$ and that u is continuously differentiable. Furthermore, it is strictly decreasing on $]0,1-p_1^*[$ and strictly increasing on $]p_1^*,1[$. Moreover, $u=s+2B_B-c_B$ on $[0,1-\hat{p}]$, $u=s+k_BB_B$ on $[1-\hat{p},1-p_1^*]$, $u=s$ on $[1-p_1^*,p_1^*]$, $u=s+k_AB_A$ on $[p_1^*,\hat{p}]$ and $u=s+2B_A-c_A$ on $[\hat{p},1]$, which shows that u is indeed the players' payoff function from $((k_A,k_B),(k_A,k_B))$.

Consider first the interval $]1-p_1^*,p_1^*[$. It has to be shown that $B_A-c_A<0$ and $B_B-c_B<0$. On $]1-p_1^*,p_1^*[$, we have that $u=s$ and $u'=0$, and therefore $B_A-c_A=\frac{\lambda+r}{r}pg-\frac{\lambda p+r}{r}s$. This is strictly negative if and only if $p<p_1^*$. By the same token, $B_B-c_B=\frac{\lambda+r}{r}(1-p)g-\frac{\lambda(1-p)+r}{r}s$. This is strictly negative if and only if $p>1-p_1^*$.

Now, consider the interval $]p_1^*,\hat{p}[$. Here, $B_A=c_A$ by construction, as k_A is determined by the indifference condition and symmetry. It remains to be shown that $B_B\le c_B$ here. Using the relevant differential equation, I find that $B_B=\frac{\lambda}{r}(g-u)+pg-s$. This is less than $c_B=s-(1-p)g$ if and only if $u\ge\frac{\lambda+r}{\lambda}g-\frac{2r}{\lambda}s$. Yet, $\frac{\lambda+r}{\lambda}g-\frac{2r}{\lambda}s\le s$ if and only if $\frac{g}{s}\le\frac{2r+\lambda}{r+\lambda}$, so that the relevant inequality is satisfied. The interval $]1-\hat{p},1-p_1^*[$ is treated in an analogous way.

Finally, consider the interval $]\hat{p},1[$. Plugging in the relevant differential equation yields $B_A-B_B=u-pg-\frac{\lambda}{r}(g-u)$. This exceeds $c_A-c_B=(1-2p)g$ if and only if $u\ge\frac{\lambda+r(1-p)}{\lambda+r}g$. At $\hat{p}$, the indifference condition gives us $k_A(\hat{p})=1$, which implies $u(\hat{p})=2s-\hat{p}g$. Since $p\mapsto\frac{\lambda+r(1-p)}{\lambda+r}g$ is decreasing and u is increasing, it is sufficient for us to show that $u(\hat{p})\ge\frac{\lambda+r(1-\hat{p})}{\lambda+r}g$, which is equivalent to $\hat{p}\le\frac{\lambda+r}{\lambda}(2p^m-1)$. From the indifference condition for the experimentation intensity $\tilde{k}_A(p):=\frac{u(p)-s}{c_A(p)}$, we see that $\tilde{k}_A$ is strictly increasing on $]p_1^*,p^m[$, and that $\lim_{p\uparrow p^m}\tilde{k}_A(p)=+\infty$; hence $\hat{p}<p^m$. Therefore, it is sufficient to show that $p^m\le\frac{\lambda+r}{\lambda}(2p^m-1)$, which is equivalent to $\frac{g}{s}\le\frac{2r+\lambda}{r+\lambda}$. □

4.6.3 *Proof of Proposition 4.3*

The proposed policies imply a well-defined law of motion for the posterior belief. The function u is strictly decreasing on $]0, \frac{1}{2}[$ and strictly increasing on $]\frac{1}{2}, 1[$. Furthermore, as $\lim_{p\uparrow\frac{1}{2}} u'(p) = \lim_{p\downarrow\frac{1}{2}} u'(p) = 0$, the function u is continuously differentiable. Moreover, $u = s + 2B_B - c_B$ on $[0, 1 - p^\dagger]$, $u = s + k_B B_B$ on $[1 - p^\dagger, \frac{1}{2}]$, $u = s + k_A B_A$ on $[\frac{1}{2}, p^\dagger]$ and $u = s + 2B_A - c_A$ on $[p^\dagger, 1]$, which shows that u is indeed the players' payoff function from $((k_A, k_B), (k_A, k_B))$.

To establish existence and uniqueness of $p^\dagger$, note that $p \mapsto \frac{\lambda+r(1-p)}{\lambda+r} g$ and $p \mapsto 2s - pg$ are strictly decreasing in p, whereas $\mathscr{W}$ is strictly increasing in p on $]\frac{1}{2}, 1[$. Now, $\mathscr{W}(\frac{1}{2}) = \frac{r+\lambda}{\lambda} g - \frac{2r}{\lambda} s$. This is strictly less than $\frac{\lambda+\frac{r}{2}}{\lambda+r} g$ and $2s - \frac{g}{2}$ whenever $\frac{g}{s} < \frac{4(r+\lambda)}{2r+3\lambda}$. Moreover, $\mathscr{W}(\frac{1}{2})$ strictly exceeds $\frac{\lambda+r(1-p^m)}{\lambda+r} g = g - \frac{r}{r+\lambda} s$ and $2s - p^m g = s$ whenever $\frac{g}{s} > \frac{2r+\lambda}{r+\lambda}$. Thus, I have established uniqueness and existence of $p^\dagger$ and that $p^\dagger \in]\frac{1}{2}, p^m[$.

By construction, $u > \max\{\frac{\lambda+r(1-p)}{\lambda+r} g, 2s - pg\}$ in $]p^\dagger, 1]$, which, by Lemma A.1 in [7], implies that $((1, 0), (1, 0))$ are mutually best responses in this region; by the same token, $u > \max\{\frac{\lambda+rp}{\lambda+r} g, 2s - (1 - p)g\}$ in $[0, 1 - p^\dagger[$, which, by Lemma A.1 in [7], implies that $((0, 1), (0, 1))$ are mutually best responses in that region.

Now, consider the interval $]\frac{1}{2}, p^\dagger]$. Here, $B_A = c_A$ by construction, so all that remains to be shown is $B_B \leq c_B$. By plugging in the indifference condition for u', I get $B_B = \frac{\lambda}{r}(g - u) + pg - s$. This is less than $c_B = s - (1 - p)g$ if and only if $u \geq \frac{\lambda+r}{\lambda} g - \frac{2r}{\lambda} s = \mathscr{W}(\frac{1}{2}) = u(\frac{1}{2})$, which is satisfied by the monotonicity properties of u. An analogous argument establishes $B_A \leq c_A$ on $[1 - p^\dagger, \frac{1}{2}[$. □

References

1. Bellman, R.: A problem in the sequential design of experiments. Sankhya Indian J. Stat. (1933–1960) **16**(3/4), 221–229 (1956)
2. Bolton, P., Harris, C.: Strategic experimentation. Econometrica **67**, 349–374 (1999)
3. Bolton, P., Harris, C.: Strategic experimentation: the Undiscounted case. In: Hammond, P.J., Myles, G.D. (eds.) Incentives, Organizations and Public Economics – Papers in Honour of Sir James Mirrlees, pp. 53–68. Oxford University Press, Oxford (2000)
4. Bradt, R., Johnson, S., Karlin, S.: On sequential designs for maximizing the sum of n observations. Ann. Math. Stat. **27**, 1060–1074 (1956)
5. Gittins, J., Jones, D.: A dynamic allocation index for the sequential design of experiments. In: Progress in Statistics, European Meeting of Statisticians, 1972, vol. 1, pp. 241–266. North-Holland, Amsterdam (1974)
6. Keller G., Rady, S., Cripps, M.: Strategic experimentation with exponential bandits. Econometrica **73**, 39–68 (2005)
7. Klein, N.: Strategic learning in teams. Games Econ. Behav. **82**, 636–657 (2013)
8. Klein, N., Rady, S.: Negatively correlated bandits. Rev. Econ. Stud. **78**, 693–732 (2011)

9. Robbins, H.: Some aspects of the sequential design of experiments. Bull. Am. Math. Soc. **58**, 527–535 (1952)
10. Thompson, W.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**, 285–294 (1933)