Correlated Bandit Games

A. Salomon*†, D. Rosenberg* and N. Vieille*

Strategic experimentation issues are prevalent in most situations of social learning. In such setups, an agent may learn useful information as the result of experimenting by himself, or possibly, from observing other agents. Typical applications include dynamic R&D, competition in an uncertain environment, financial contracting, etc. The trade-off faced by an agent between the exploitation of available information and the acquisition of new information is often modeled by means of a two-arm bandit problem. Each agent chooses to allocate his time/resources between a safe arm, that yields a known payoff flow, and a risky one, which may either be *good* (and then dominate the safe arm), or *bad*. In particular, much attention has been paid to the case of Poisson/exponential bandits, (e.g. Keller, Rady and Cripps [2]).

Within this exponential setup, we study a two-player model, in which the types of the two risky arms are correlated. We analyze both the case in which payoffs are observed, and the case in which only actions are observed. We assume that decisions to switch from the risky to the safe arm are irreversible. We study the joint effect of the correlation and of the information scenario on the equilibrium outcome. Among others, this is related to Murto and Välimäki [1].

Despite differences in the two setups, our main conclusion is the same. When events coming from an other player bring bad news *no* equilibrium exhibits an encouragement effect, in the sense that, when dropping out, the belief held by a player is *equal* to the belief at which he would choose to drop out in a one-player setup. When events instead bring good news, *all* equilibria exhibit an encouragement effect.

1 The model

1.1 The setup

Time is continuous. Each of two players is facing a strategic experimentation problem, modeled as a two-arm bandit problem. At each point in time, each player has to choose which of two arms to pull. One of the arms does not involve any uncertainty, and yields a constant payoff flow with present value s. The other arm's type is *ex ante* unknown, and it may be either *Good* or *Bad*. The decision to switch from the risky arm to the safe one is irreversible. That is, each player has to choose when, if ever, to drop out and stop experimenting.

A risky arm of type *Bad* (B) never yields any payoff. An arm of type *Good* (G) yields a constant payoff flow with present value γ, starting from some random time τ, and no payoff prior to τ. Conditionally to the type being G, the r.v. τ follows an exponential distribution with parameter λ. Of course if the type is B, $\tau = +\infty$.

The players have the same prior over the pair (R_1, R_2) of the types of the risky arms, and share a common discount rate $r > 0$. The decision problems faced by the two players are identical, in that the *ex ante* probability of one's arm being good is the same for both players, and, if good, both arms share the same values of λ and of γ.

Conditional on (R_1, R_2), the two arms behave independently. In particular, if both arms are good, the random times τ_1 and τ_2 associated with the two machines are independent. The strategic interaction between the players comes from the fact that the types R_1 and R_2 of the two arms are correlated. Hence, any information on player j's risky arm is valuable to player i.

The strategic variable of player i is how long to experiment. We will denote by θ_i the time at which player i chooses to drop out.

We will analyze two variants of this model, which differ in the information made available to the players along the play. In both variants, each player i knows at time t whether $\tau_i < t$ or not, and whether (and when) player j already dropped out, or

*HEC Paris
†Université Paris 13

not. Thus, each player knows his own past payoffs and the other player's past action choices. In one variant, no additional information is provided (Observed Actions). In the other variant, player i observes moreover player j's past payoffs – that is, whether $\tau_j < \min(\theta_j, t)$ or not (Observed Payoffs). Our goal is to assess the joint effect of the correlation and of these informational spill-overs. To this end, we measure correlation between the two arms by $\rho := q(R_1 = R_2 = G) - q(R_1 = G)q(R_2 = G)$. The two risky arms are positively (resp. negatively) correlated if $\rho > 0$ (resp. < 0), and this means that an evidence that one's risky arm type is G is also an evidence that the other player's risky arm is G (resp. B).

1.2 Strategies

1.2.1 One-player game

A natural and useful benchmark is the decision problem $\mathcal{P}$ in which player i only observes his own payoffs. The problem $\mathcal{P}$ is a standard one-player decision problem. As long as $\tau_i \geq t$, the probability assigned by player i to his risky arm being good, decreases continuously. The unique optimal policy is to drop out as soon as this belief reaches

$$p_* := \frac{rs}{\lambda(\gamma - s)},$$

which happens (if no payoff is received) at the time T_* defined by $\mathbf{P}(R_i = G|\tau_i \geq T_*) = p_*$. As a function of his initial belief $p \in (0, 1]$, the optimal payoff of the decision maker in $\mathcal{P}$ equals s on $(0, p_*]$ and increases from s to g on $(p_*, 1]$, where $g = \gamma\frac{\lambda}{r+\lambda}$ is the *ex ante* expected payoff of a good risky arm.[1]
For future reference, we also define $\tilde{\mathcal{P}}$ as the decision problem in which player i observes at any time t whether $\tau_i < t$ and/or $\tau_j < t$.[2]

1.2.2 The Observed Payoffs scenario

Consider any time instant t, and let us focus on player i, whom we assume to be active prior to time t. If $\tau_i < t$, player i's risky arm is known to be good, and player i should stick to it. On the either hand, if either $\tau_j < t$ or $\theta_j < t$, player j's future behavior does not prove informative to player i. Thus, at time $\min(\tau_j, \theta_j)$, player i should update his belief, and proceed with the optimal policy in $\mathcal{P}$. Hence, player i's optimal decision at time t is unambiguous unless if $\min(\tau_i, \tau_j, \theta_j) \geq t$. Accordingly, we define a pure strategy of player i to be a time $t \in [0, +\infty]$, with the interpretation that player i drops out at time t if player j has not dropped out, and if no payoff has been received prior to t. A mixed strategy is a probability distribution over $[0, +\infty]$.

1.2.3 The Observed Actions scenario

Similarly, a player who is still active prior to time t will play according to the optimal strategy in $\mathcal{P}$ unless if $\min(\tau_i, \theta_j) \geq t$. Again, we let a pure strategy of player i be a time $t \in [0, +\infty]$, with the interpretation that player i drops out at t if $\tau_i \geq t$ and $\theta_j \geq t$. A mixed strategy is a probability distribution over $[0, +\infty]$.

2 Results

We let dates T_ϕ and T_ψ be defined by

$$\mathbf{P}(R_i = G| \min(\tau_j, \tau_i) \geq T_\phi) = p_* \text{ and } \mathbf{P}(R_i = G|R_j = G, \tau_i) \geq T_\psi) = p_*.$$

T_ϕ (resp. T_ψ) can be interpreted as the date when the belief held by a player reaches the cutoff p_* in the OP scenario when no payoff occurs (resp. when the opponent receives a payoff from his risky arm).

[1] To avoid trivialities, we assume $g > s > 0$.
[2] Thus, $\tilde{\mathcal{P}}$ is the best-reply problem faced by player i when he is able to watch player's j payoffs and when player j plans never to drop out.

2.1 Negative informational shocks

We here cover the case where events from the opponent bring bad news. This is the case when either (i) $\rho < 0$ in the Observed Payoffs scenario, or (ii) $\rho > 0$ in the Observed Actions scenario. Indeed in (i) the possible event is the other player receiving a payoff from his risky arm, which is evidence that its type is G, and (statistical) evidence that one's risky arm type is B when $\rho < 0$. In (ii) the possible event is the opponent dropping out. This shows that he did not get any payoff from his risky arm, and that its type tends to be B. This is also (statistical) evidence that one's risky arm type is B when $\rho > 0$.

Theorem 1 (OP scenario, $\rho < 0$) *The equilibria are the profiles (T_ϕ, σ) and (σ, T_ϕ), where σ is any distribution such that $\sigma([T_\phi, +\infty]) = 1$. In particular, there is a unique symmetric equilibrium, (T_ϕ, T_ϕ).*

Theorem 2 (OA scenario, $\rho > 0$) *There is a unique symmetric equilibrium, (σ_*, σ_*). The distribution σ_* has a support $[T_*, T_\psi]$, and a density, which is positive and continuous on $[T_*, T_\psi]$.*
One has $\mathbf{P}(R_i = G | \theta_j \geq t, \tau_i \geq t) = p_*$ for each $t \in [T_*, T_\psi]$.

The main conclusion to be drawn is that, at equilibrium, no player is willing to keep experimenting with a belief below p_*, the optimal cut-off in $\mathcal{P}$. Put otherwise, while the informational spill-over/externality clearly affects the speed at which beliefs change, the marginal option value of observing the other player is equal to *zero*. Indeed, when player i's belief reaches p_*, he finds it optimal to exit, *whether or not* he may benefit from observing the other player's action/payoffs in the future. This is partially explained by the fact that informational events from the other player only bring bad news.

2.2 Positive informational shocks

We discuss the case where events bring good news. This is the case when either (i) $\rho > 0$ in the OP scenario or (ii) $\rho < 0$ in the OA scenario.
In the observed payoffs scenario, the game is a game of pure coordination. Player i is willing to experiment beyond T_ϕ only to the extent that player j also experiments beyond T_ϕ. On the other hand, no player will ever be willing to experiment beyond T_{**}, defined as the (unique) optimal policy in the decision problem $\tilde{\mathcal{P}}$. If $\rho < 0$, former results show that $T_{**} = T_\phi$, and the uniqueness of the equilibrium outcome follows. When $\rho > 0$, one has $T_{**} > T_\phi$ and, not surprisingly, a continuum of equilibrium outcomes.

Theorem 3 (OP scenario, $\rho > 0$) *One has $T_{**} > T_\phi$. All symmetric equilibria are pure, and all pure equilibria are symmetric. The symmetric equilibria are the pairs (T, T), with $T \in [T_\phi, T_{**}]$. The asymmetric equilibria are the pairs (T_{**}, σ) and (σ, T_{**}), where σ assigns probability 1 to $[T_{**}, +\infty]$.*

Conclusions are much sharper in the observed actions scenario. Basically, there can not be atom at any time t in an equilibrium strategy. Indeed the other player could benefit from waiting time t and then dropping out immediately if no good news arise. Then there is a unique equilibrium strategy, which is only dictated by the necessity to make the other player indifferent between staying active and dropping out.

Theorem 4 (OA scenario, $\rho < 0$) *There is a unique equilibrium, which is a symmetric and non-atomic equilibrium (σ, σ). The support of the distribution σ is equal to $\left[T_*, \hat{T}\right]$, where $\hat{T} < T_\phi$.*

Observe that in all equilibria, with the exception of (T_ϕ, T_ϕ) in the OP scenario, players are able to remain active with beliefs below p_*.

References

[1] P. Murto and J. Välimäki (2008): Learning and Information Aggregation in an Exit Game,*Helsinki Center of Economic Research, Discussion Papers*, Discussion Paper No 235 / September 2008, ISSN 1795-0562.

[2] G. Keller, S. Rady and M. Cripps (2005): Strategic Experimentation with exponential bandits, *Econometrica*, Vol. 73, No. 1, 39-68.