

# UNDISCOUNTED BANDIT PROBLEMS\*

Godfrey Keller<sup>†</sup>      Sven Rady<sup>‡</sup>

May 18, 2010

## Abstract

We analyze a game of strategic experimentation with two-armed bandits when there is no discounting. We show that for all specifications of prior beliefs and payoff-generating processes that satisfy some separability condition, all Markov perfect equilibria exhibit striking similarities in the sense that action profiles depend only on the expected current payoff of the risky arm and the expected full-information payoff, given current information. The separability condition holds in a variety of models that have been explored in the literature, all of which assume that the risky arm's expected payoff per unit of time is time-invariant and actual payoffs are generated by a process with independent and stationary increments. The separability condition does not hold in examples where the expected payoff per unit of time is subject to state-switching.

KEYWORDS: Strategic Experimentation, Two-Armed Bandit, Markov-Perfect Equilibrium.

*JEL* CLASSIFICATION NUMBERS: C73, D83.

---

\*Our thanks for helpful discussions and suggestions are owed to workshop and seminar participants at EUI (Florence), SED (2005, Budapest), UCL/Birkbeck and the Universities of Cambridge, Keele, Oxford and Southampton. The first author would like to thank the Center for Economic Studies at the University of Munich for its hospitality. Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 and GRK 801 is gratefully acknowledged.

<sup>†</sup>Department of Economics, University of Oxford, Manor Road Building, Oxford OX1 3UQ, UK.

<sup>‡</sup>Department of Economics, University of Munich, Kaulbachstr. 45, D-80539 Munich, Germany.

# Introduction

We analyze the undiscounted version of a class of continuous-time two-armed bandit models, in which a number of players act non-cooperatively, trying to learn an unknown parameter that governs the risky arm's expected payoff per unit of time. We introduce background information to ensure that the problem is well-posed, and assume a separability condition which restricts the expected infinitesimal change in a players' payoff function to be proportional to the overall intensity of experimentation performed at the given point in time.

Generalizing the findings of Bolton and Harris (2000), we show that under the separability condition, the Markovian equilibria of these undiscounted experimentation games exhibit striking similarities in the sense that action profiles depend only on the expected current payoff of the risky arm and the expected full-information payoff, given current information. For equilibrium strategies, therefore, the specification of the payoff-generating process for the risky arm is irrelevant – what counts is the specification of the agents' prior belief.

We present five examples that satisfy the separability condition. In the first, payoffs are generated by a Brownian motion with unknown drift, and the agents' prior belief about this drift is an arbitrary discrete distribution; this extends the setup of Bolton and Harris (1999, 2000). In the second, payoffs come from a Poisson process with unknown intensity, and the agents' prior belief about this intensity is again a discrete distribution; this generalizes the setup of Keller, Rady and Cripps (2005) and Keller and Rady (2010). These two examples are special cases of a third, explored by Cohen and Solan (2009), in which payoffs are generated by a Lévy process, that is, a continuous-time process with independent and stationary increments. In the fourth example, payoffs stem again from a Brownian motion with unknown drift, but prior beliefs are normally distributed; this is the same specification as in Jovanovic (1979). In the fifth, payoffs are generated by a Poisson process with unknown intensity, but now the agents' prior belief about this intensity is characterized by a Gamma distribution; this specification has been assumed by Moscarini and Squintani (2004).

We also provide an example that violates the separability condition. There, payoffs are generated by a Brownian motion with an unknown drift that is subject to Markovian state-switching as in Keller and Rady (1999, 2003).

These examples suggest that separability hinges on the stationarity of the environment in which the players are learning (the average payoff per unit of time must not change over time) and on the stationarity and independence of the actual payoff increments. In such a situation, experiments performed contemporaneously by different players are indeed perfect substitutes as far as information acquisition is concerned.

The rest of the article is organized as follows. We first set up the general model, introduce the separability condition and present the examples. We then establish the efficient benchmark where players cooperate to maximize joint expected payoffs. Turning to the strategic problem, we provide an inefficiency result and then present a characterization of all Markov perfect equilibria, echoing that from Bolton and Harris (2000). Finally, we offer some concluding remarks.

## 1 Undiscounted Bandits

Time  $t \in [0, \infty)$  is continuous, and there is no discounting. There are  $N \geq 1$  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 she continually has to decide what fraction of the available resource to allocate to each arm.

If a player uses the safe arm  $S$  over an interval  $[t, t + dt)$ , the expected payoff increment is  $s dt$ , where  $s$  is fixed and known to all players; if a player uses the risky arm  $R$  over an interval  $[t, t + dt)$ , the expected payoff increment is  $\mu dt$ , where  $\mu$  is fixed but unknown. So  $s$  and  $\mu$ , common across players, are the expected flow equivalents of the two arms, and if a player allocates the fraction  $k_t \in [0, 1]$  of the resource to  $R$  over an interval of time  $[t, t + dt)$ , and consequently the fraction  $1 - k_t$  to  $S$ , then her expected payoff increment conditional on  $\mu$  is  $[(1 - k_t)s + k_t\mu] dt$ .

Regardless of the players' choices,  $(k_{1,t}, \dots, k_{N,t})$ , all the players observe a background signal which is informationally equivalent to  $k_0 > 0$  units of the resource being allocated to  $R$ . This ensures that the players eventually learn the value of  $\mu$ , even if they all play  $S$  all the time.

The players start with a common prior belief about  $\mu$ , and thereafter they all observe each other's actions and outcomes, so they hold common posterior beliefs throughout time. Assume that at time  $t$  the players believe that  $\mu$  has a cumulative distribution function  $H(\cdot; \pi_t)$ , where  $\pi_t$  is a sufficient statistic for the observations on  $R$  and the background signal up to time  $t$ , and  $H$  represents a conjugate family of distributions. We assume that  $s$  lies between the infimum and supremum of the support of  $H(\cdot; \pi_0)$ , so that, at least initially, each player prefers  $R$ , if it is 'good', to  $S$ , and prefers  $S$  to  $R$ , if it is 'bad'.

Given the current belief  $H(\cdot; \pi)$ , let  $m(\pi)$  denote the expected current (or myopic) payoff from  $R$ , and let  $f(\pi)$  denote the expected full-information payoff:

$$m(\pi) = \int \mu dH(\mu; \pi), \quad f(\pi) = \int (s \vee \mu) dH(\mu; \pi).$$

As  $m(\pi_t)$  and  $f(\pi_t)$  are conditional expectations given all the information available at time

$t$ , the Law of Iterated Expectations implies that  $E_t[m(\pi_T)] = m(\pi_t)$  and  $E_t[f(\pi_T)] = f(\pi_t)$  for all  $T > t$ , i.e. both  $m(\pi_t)$  and  $f(\pi_t)$  are martingales with respect to the players' information sets.

Each player chooses actions  $\{k_t\}_{t \geq 0}$  such that  $k_t$  is measurable with respect to the information available at time  $t$ . The objective is to maximize

$$E \left[ \int_0^\infty [(1 - k_t)s + k_t m(\pi_t) - f(\pi_t)] dt \right],$$

where the expectation is over the stochastic processes  $\{k_t\}$  and  $\{\pi_t\}$  – that is, players use the catching-up criterion.<sup>1</sup> This objective highlights the potential for the sufficient statistic to serve as a state variable. It also shows that a player's payoff depends on others' actions only through their impact on the evolution of the sufficient statistic.

Suppose players choose the actions  $\{(k_{1,t}, \dots, k_{N,t})\}_{t \geq 0}$ . Let  $K_t = \sum_{n=1}^N k_{n,t}$ . This sum measures how much of the  $N$  units of the resource is allocated to risky arms at time  $t$  – we call it the *intensity of experimentation*. For  $n = 1, \dots, N$ , let  $u_n(\pi)$  denote the value of player  $n$ 's objective function when  $\pi_0 = \pi$  and the above strategies are used. The assumption that the following separability condition holds is crucial to our analysis:

$$E[u_n(\pi_{t+dt}) \mid \pi_t, K_t] = u_n(\pi_t) + (K_t + k_0) D(\pi_t, u_n) dt \quad (1)$$

for some functional operator  $D(\cdot, \cdot)$  whose domain contains all possible pairs of realizations of the sufficient statistic and payoff functions.

## 2 Examples

This section presents five specifications of priors and payoff-generating processes that satisfy the above separability condition, and one that does not. For more details of Example 2.1, see Bolton and Harris (1999, 2000), and for the discounted version of Example 2.2, see Keller, Rady and Cripps (2005) and Keller and Rady (2010). The discounted single-agent version of Example 2.3 with a two-point prior is solved in Cohen and Solan (2009). Models in which agents observe stochastic processes and have beliefs like those in Example 2.4 and 2.5 can be found in Jovanovic (1979) and Moscarini and Squintani (2004), respectively. The state-switching specification in Example 2.6 is the same as in Keller and Rady (1999, 2003).

In this section we write  $K_t^\ddagger$  for  $K_t + k_0$ ,  $d\pi_t$  for  $\pi_{t+dt} - \pi_t$ , and ignore terms of order higher than  $dt$ .

---

<sup>1</sup>For a discussion of this objective, see Bolton and Harris (2000). Essentially, the integrand is the difference between what a player expects to receive and what she would expect to receive were she to be fully informed.

## 2.1 Brownian payoffs, discrete prior

We start with the case of a two-point prior. If player  $n$  allocates a fraction  $k_n$  of her unit resource to the risky arm over a time interval  $[t, t + dt)$ , her payoff increment from the risky arm is  $k_n \mu dt + \sqrt{k_n} \sigma dZ_{n,t}$  where  $(Z_1, \dots, Z_N)$  is an  $N$ -dimensional Wiener process,  $\mu \in \{\mu_0, \mu_1\}$  and  $\mu_0 < s < \mu_1$ . The background signal that all players receive is of the form  $k_0 \mu dt + \sqrt{k_0} \sigma dZ_{0,t}$  where  $Z_0$  is a Wiener process orthogonal to  $Z_1, \dots, Z_N$ .

Let  $\pi_t$  denote the probability that the players assign to the event  $\mu = \mu_1$  given their observations up to time  $t$ . This is an obvious sufficient statistic for the problem at hand. We have

$$m(\pi) = (1 - \pi)\mu_0 + \pi\mu_1, \quad f(\pi) = (1 - \pi)s + \pi\mu_1.$$

Moreover, it follows from Liptser and Shiriyayev (1977, Theorem 9.1) that

$$\mathbb{E}[d\pi_t | \pi_t, K_t] = 0, \quad \text{Var}[d\pi_t | \pi_t, K_t] = K_t^\dagger \left[ \pi_t(1 - \pi_t) \Delta\mu \sigma^{-1} \right]^2 dt,$$

where  $\Delta\mu = \mu_1 - \mu_0$ . Applying Itô's lemma and taking expectations, we find

$$\mathbb{E}[u_n(\pi_{t+dt}) | \pi_t, K_t] = u_n(\pi_t) + K_t^\dagger \left\{ \frac{1}{2} \left[ \pi_t(1 - \pi_t) \Delta\mu \sigma^{-1} \right]^2 u_n''(\pi_t) \right\} dt,$$

so (1) holds with

$$D(\pi, u) = \frac{1}{2} \left[ \pi(1 - \pi) \Delta\mu \sigma^{-1} \right]^2 u''(\pi).$$

There is a straightforward generalization to the case where  $\mu$  can take any one of  $L+1$  possible values, with  $L$  finite or countably infinite. Let  $\mu \in \{\mu_0, \mu_1, \dots, \mu_L\}$  and  $\mu_0 < \dots < s < \dots < \mu_L$ . Players' beliefs become the  $L$ -vector  $\pi = (\pi_1, \dots, \pi_L)$ ,  $H(\mu; \pi)$  is the obvious step function and, with  $\pi_0 = 1 - \sum_{\ell \geq 1} \pi_\ell$ ,

$$m(\pi) = \sum_{\ell \geq 0} \pi_\ell \mu_\ell, \quad f(\pi) = \sum_{\ell \geq 0} \pi_\ell (s \vee \mu_\ell).$$

Again from Liptser and Shiriyayev (1977, Theorem 9.1), for  $i, \ell \geq 1$  we have<sup>2</sup>

$$\begin{aligned} \mathbb{E}[d\pi_{\ell,t} | \pi_t, K_t] &= 0, \quad \text{Var}[d\pi_{\ell,t} | \pi_t, K_t] = K_t^\dagger \left[ \pi_{\ell,t}(\mu_\ell - m(\pi_t)) \sigma^{-1} \right]^2 dt, \\ \text{and } \text{Cov}[d\pi_{i,t}, d\pi_{\ell,t} | \pi_t, K_t] &= K_t^\dagger \left[ \pi_{i,t}(\mu_i - m(\pi_t)) \sigma^{-1} \right] \left[ \pi_{\ell,t}(\mu_\ell - m(\pi_t)) \sigma^{-1} \right] dt. \end{aligned}$$

Using Itô's lemma and taking expectations now leads to (1) holding with

$$D(\pi, u) = \frac{1}{2} \sum_{i \geq 1} \sum_{\ell \geq 1} \pi_i \pi_\ell (\mu_i - m(\pi))(\mu_\ell - m(\pi)) \sigma^{-2} \frac{\partial^2 u(\pi)}{\partial \pi_i \partial \pi_\ell}.$$

---

<sup>2</sup>They show that for a single agent playing  $R$ , and no background information, the belief evolves according to  $d\pi_{\ell,t} = \sigma^{-1} \pi_{\ell,t} [\mu_\ell - m(\pi_t)] d\bar{z}_t$ , where  $d\bar{z}_t = \sigma^{-1} ([\mu - m(\pi_t)] dt + \sigma dZ_t)$  is the *innovation* process.

## 2.2 Poisson payoffs, discrete prior

As in Example 2.1, we start with the case of a two-point prior. If player  $n$  allocates a fraction  $k_n$  of her unit resource to the risky arm over a time interval  $[t, t + dt)$ , she receives lump-sums from the risky arm corresponding to the increments of a Poisson process with parameter  $k_n \mu$  where  $\mu \in \{\mu_0, \mu_1\}$  and  $\mu_0 < s < \mu_1$ . These processes are independent across players. The background signal that all players observe is the increment of a Poisson process with parameter  $k_0 \mu$  which is independent of the processes that generate players' payoffs.

We can again take  $\pi_t$ , the posterior probability that  $\mu = \mu_1$ , as the sufficient statistic. In particular,  $m(\pi) = (1 - \pi)\mu_0 + \pi\mu_1$  and  $f(\pi) = (1 - \pi)s + \pi\mu_1$  are the same as in Example 2.1.

Now, with probability  $K_t^\ddagger m(\pi_t) dt$ , there is a positive increment on one of the risky arms or the background signal between  $t$  and  $t + dt$ , and Bayes' rule implies that  $\pi_t$  jumps to  $j(\pi_t)$  given by

$$\pi_{t+dt} = \pi_t \mu_1 / m(\pi_t).$$

With probability  $1 - K_t^\ddagger m(\pi_t) dt$ , there is no such increment and Bayes' rule yields

$$d\pi_t = -K_t^\ddagger \pi_t (1 - \pi_t) \Delta\mu dt$$

with  $\Delta\mu = \mu_1 - \mu_0$ . So, we have

$$\begin{aligned} & \mathbb{E}[u_n(\pi_{t+dt}) \mid \pi_t, K_t] \\ &= K_t^\ddagger m(\pi_t) u_n(j(\pi_t)) dt + (1 - K_t^\ddagger m(\pi_t) dt) (u_n(\pi_t) - K_t^\ddagger \pi_t (1 - \pi_t) \Delta\mu u'_n(\pi_t) dt) \\ &= u_n(\pi_t) + K_t^\ddagger \left\{ m(\pi_t) [u_n(j(\pi_t)) - u_n(\pi_t)] - \pi_t (1 - \pi_t) \Delta\mu u'_n(\pi_t) \right\} dt, \end{aligned}$$

and (1) holds with

$$D(\pi, u) = m(\pi) [u(j(\pi)) - u(\pi)] - \pi(1 - \pi) \Delta\mu u'(\pi).$$

As in Example 2.1, there is a simple generalization to the case where  $\mu$  can take any one of  $L + 1$  possible values, with  $L$  finite or countably infinite. Just as there, players' beliefs become an  $L$ -vector, and  $m(\pi)$  and  $f(\pi)$  are the same as given earlier. After a positive increment, the belief jumps to  $j(\pi_t)$  given by the  $L$ -vector

$$\pi_{t+dt} = (\pi_{1,t} \mu_1 / m(\pi_t), \dots, \pi_{\ell,t} \mu_\ell / m(\pi_t), \dots, \pi_{L,t} \mu_L / m(\pi_t));$$

if no increment arrives, for  $\ell \geq 1$  beliefs adjust infinitesimally by

$$d\pi_{\ell,t} = -K_t^\ddagger \pi_{\ell,t} (\mu_\ell - m(\pi_t)) dt.$$

This leads to (1) holding with

$$D(\pi, u) = m(\pi) [u(j(\pi)) - u(\pi)] - \sum_{\ell \geq 1} \pi_\ell (\mu_\ell - m(\pi)) \frac{\partial u(\pi)}{\partial \pi_\ell}.$$

### 2.3 Lévy payoffs, discrete prior

Examples 2.1 and 2.2 are special cases of a specification where payoffs are generated by a Lévy process, that is, a continuous-time process with independent and stationary increments. If player  $n$  allocates a fraction  $k_n$  of her unit resource to the risky arm over a time interval  $[t, t + dt)$ , her payoff from the risky arm is the increment of the process

$$k_n \lambda t + \sqrt{k_n} \sigma Z_{n,t} + k_n Y_{n,t}$$

where  $(Z_1, \dots, Z_N)$  is again an  $N$ -dimensional Wiener process, and  $Y_n$  is a compound Poisson process with finite Lévy measure  $\nu$ , that is,  $\nu(A)$  is the expected number of jumps per unit of time whose size is in the Borel set  $A \subseteq \mathbb{R} \setminus \{0\}$ ; the compound Poisson processes are independent across players. The background signal is informationally equivalent to an amount  $k_0$  of the resource being devoted to the risky arm.

Let  $\lambda \in \{\lambda_0, \lambda_1, \dots, \lambda_L\}$  and  $\nu \in \{\nu_0, \nu_1, \dots, \nu_L\}$ ; define  $\bar{\nu}_\ell = \nu_\ell(\mathbb{R} \setminus \{0\})$  as the expected number of jumps per unit of time if  $\nu = \nu_\ell$ , and  $g_\ell = \int_{\mathbb{R} \setminus \{0\}} g \nu_\ell(dg) / \bar{\nu}_\ell$  as the expected jump size. Finally, let  $\mu_\ell = \lambda_\ell + g_\ell \bar{\nu}_\ell$ .

With Lévy payoffs and a discrete prior, (1) holds with  $D(\pi, u)$  being given by a combination of expressions that generalize those in Examples 2.1 and 2.2, namely

$$\begin{aligned} D(\pi, u) &= \frac{1}{2} \sum_{i \geq 1} \sum_{\ell \geq 1} \pi_i \pi_\ell (\lambda_i - \lambda(\pi)) (\lambda_\ell - \lambda(\pi)) \sigma^{-2} \frac{\partial^2 u(\pi)}{\partial \pi_i \partial \pi_\ell} \\ &\quad + \int_{\mathbb{R} \setminus \{0\}} [u(j(\pi, g)) - u(\pi)] \nu(\pi)(dg) - \sum_{\ell \geq 1} \pi_\ell (\bar{\nu}_\ell - \bar{\nu}(\pi)) \frac{\partial u(\pi)}{\partial \pi_\ell}, \end{aligned}$$

where

$$\lambda(\pi) = \sum_{\ell \geq 0} \pi_\ell \lambda_\ell, \quad \nu(\pi) = \sum_{\ell \geq 0} \pi_\ell \nu_\ell, \quad \bar{\nu}(\pi) = \sum_{\ell \geq 0} \pi_\ell \bar{\nu}_\ell,$$

and  $j_\ell(\pi, g) = \pi_\ell \nu_\ell(dg) / \nu(\pi)(dg)$  is the revised belief after a jump of size  $g$  arrives.

### 2.4 Brownian payoffs, normal prior

The risky arms and background signal are specified as in Example 2.1 except for the assumption that  $\mu$  can now take any real value. At time  $t$ , players believe that  $\mu$  is distributed according to a normal distribution with mean  $m_t$  and precision  $\tau_t > 0$ , and so

we take  $\pi = (m, \tau)$ .

For future reference, note that (i) since  $s \vee \mu$  is increasing in  $\mu$ , a first-order stochastic dominance argument can be used to establish that  $\partial f(\pi)/\partial m > 0$ , and (ii) since  $s \vee \mu$  is convex in  $\mu$ , a second-order stochastic dominance argument can be used to establish that  $\partial f(\pi)/\partial \tau < 0$ . Consequently,  $f$  has the same monotonicity properties as  $m$ .

Now, following Chernoff (1968, Lemma 4.1), or Liptser and Shiriyayev (1977, Theorem 10.1), we have<sup>3</sup>

$$\mathbb{E}[dm_t | \pi_t, K_t] = 0, \quad \text{Var}[dm_t | \pi_t, K_t] = K_t^\dagger \tau_t^{-2} \sigma^{-2} dt, \quad \mathbb{E}[d\tau_t | \pi_t, K_t] = K_t^\dagger \sigma^{-2} dt.$$

Applying Itô's lemma and taking expectations, we see that

$$\mathbb{E}[u_n(\pi_{t+dt}) | \pi_t, K_t] = u_n(\pi_t) + K_t^\dagger \left\{ \sigma^{-2} \left[ \frac{1}{2} \tau_t^{-2} \partial^2 u_n(\pi_t) / \partial m^2 + \partial u_n(\pi_t) / \partial \tau \right] \right\} dt,$$

so (1) holds with

$$D(\pi, u) = \sigma^{-2} \left[ \frac{1}{2} \tau^{-2} \frac{\partial^2 u(\pi)}{\partial m^2} + \frac{\partial u(\pi)}{\partial \tau} \right].$$

## 2.5 Poisson payoffs, Gamma prior

The risky arms and background signal are specified as in Example 2.2 except for the assumption that  $\mu$  can now take any non-negative value. Let  $s > 0$  for the safe arm.

At time  $t$ , players believe that  $\mu$  is distributed according to the Gamma distribution  $\text{Ga}(\alpha_t, \beta_t)$  with parameters  $\alpha_t > 0$  and  $\beta_t > 0$ . With  $\pi = (\alpha, \beta)$ , the probability density function for  $\mu$  is  $h(\mu; \pi) = [\beta^\alpha / \Gamma(\alpha)] \mu^{\alpha-1} e^{-\mu\beta}$ , and we have

$$m(\pi) = \alpha/\beta, \quad f(\pi) = \int_0^\infty (s \vee \mu) h(\mu; \pi) d\mu.$$

(The corresponding variance of  $\mu$  is  $\alpha/\beta^2$ .)

For future reference, note that for  $\alpha' > \alpha''$  the likelihood ratio  $h(\mu; \alpha', \beta)/h(\mu; \alpha'', \beta)$  is increasing, and for  $\beta' > \beta''$  the likelihood ratio  $h(\mu; \alpha, \beta')/h(\mu; \alpha, \beta'')$  is decreasing. Since the likelihood ratio ordering implies first-order stochastic dominance,  $f$  has the same monotonicity properties as  $m$ .

Now, with probability  $K_t^\dagger m(\pi_t) dt$ , there is a positive increment on one of the risky

---

<sup>3</sup>It can be shown that for a single agent playing  $R$ , and no background information, the mean and precision evolve according to  $dm_t = \sigma^{-1} \tau_t^{-1} d\bar{z}_t$  and  $d\tau_t = \sigma^{-2} dt$ , where, now, the *innovation* process is  $d\bar{z}_t = \sigma^{-1} ([\mu - m_t] dt + \sigma dZ_t)$ .

Note that the expression equivalent to that for  $dm_t$  to be found in equation (9) of Jovanovic (1979) omits the term  $[\mu - m_t]$ .

arms or the background signal between  $t$  and  $t + dt$ , and  $\pi_{t+dt} = (\alpha_t + 1, \beta_t + K_t^\ddagger dt)$ ; with probability  $1 - K_t^\ddagger m(\pi_t) dt$ , there is no such increment, and  $\pi_{t+dt} = (\alpha_t, \beta_t + K_t^\ddagger dt)$ . (Essentially,  $\alpha$  counts arrivals of the lump-sums and  $\beta$  measures the effective time that  $R$  has been used – see, for example, DeGroot, 1970, Chapter 9.) So, we find

$$\begin{aligned} \mathbb{E}[u_n(\pi_{t+dt}) \mid \pi_t, K_t] &= K_t^\ddagger m(\pi_t) u_n(\alpha_t + 1, \beta_t) dt + (1 - K_t^\ddagger m(\pi_t) dt) \left( u_n(\pi_t) + K_t^\ddagger \frac{\partial u_n(\pi_t)}{\partial \beta_t} dt \right) \\ &= u_n(\pi_t) + K_t^\ddagger \left\{ m(\pi_t) [u_n(\alpha_t + 1, \beta_t) - u_n(\pi_t)] + \partial u_n(\pi_t) / \partial \beta_t \right\} dt. \end{aligned}$$

Thus, (1) holds with

$$D(\pi, u) = m(\pi) [u(\alpha + 1, \beta) - u(\pi)] + \frac{\partial u(\pi)}{\partial \beta}.$$

## 2.6 Brownian payoffs, state-switching

We modify Example 2.1 by introducing state-switching: the unknown drift at time  $t$ ,  $\mu_t$ , switches between  $\mu_0$  and  $\mu_1$  according to a continuous-time Markov process with transition probabilities

$$\Pr(\mu_{t+dt} = \mu_1 \mid \mu_t = \mu_0) = p_0 dt + o(dt), \quad \Pr(\mu_{t+dt} = \mu_0 \mid \mu_t = \mu_1) = p_1 dt + o(dt),$$

where  $p_\ell \geq 0$  ( $\ell = 0, 1$ ).

Given the belief  $\pi_t$  that  $\mu_t = \mu_1$ , the players assign probability  $(1 - \pi_t)p_0 dt$  to a transition from  $\mu_t = \mu_0$  to  $\mu_{t+dt} = \mu_1$ ; similarly, they assign probability  $\pi_t p_1 dt$  to a transition from  $\mu_t = \mu_1$  to  $\mu_{t+dt} = \mu_0$ . The former increases  $\pi_t$  whereas the latter decreases it, and the combined effect leads to

$$\mathbb{E}[d\pi_t \mid \pi_t, K_t] = [(1 - \pi_t)p_0 - \pi_t p_1] dt.$$

As a consequence, we have

$$\begin{aligned} \mathbb{E}[u_n(\pi_{t+dt}) \mid \pi_t, K_t] &= u_n(\pi_t) + \left\{ [(1 - \pi_t)p_0 - \pi_t p_1] u'_n(\pi_t) + \frac{1}{2} K_t^\ddagger [\pi(1 - \pi) \Delta \mu \sigma^{-1}]^2 u''(\pi) \right\} dt, \end{aligned}$$

so condition (1) does not hold for this specification when at least one of the transition intensities  $p_0, p_1$  is non-zero.

### 3 Efficient Benchmark and Markov Equilibria

In this section, we characterize efficient strategies as well as Markov perfect equilibria, borrowing some insights from Bolton and Harris (2000). We assume throughout that the separability condition (1) holds.

#### 3.1 The cooperative problem

Suppose that the  $N$  players work cooperatively, i.e. maximize the *average* expected payoff by jointly choosing the action profiles  $\{(k_{1,t}, \dots, k_{N,t})\}_{t \geq 0}$ . This is a dynamic programming problem with the current value of  $\pi$  as the state variable.

If current actions are  $(k_1, \dots, k_N)$ , the average expected payoff increment is given by  $\left[ \left(1 - \frac{K}{N}\right)s + \frac{K}{N}m(\pi) \right] dt$  with  $K = \sum_{n=1}^N k_n$ . As the expected continuation value is of the form  $u(\pi) + (K + k_0) D(\pi, u) dt$ , the cooperative's problem reduces to choosing the optimal intensity of experimentation  $K$  given the current state  $\pi$ .

The Bellman equation for the value function of the cooperative, expressed as average payoff per agent, is

$$0 = \max_{K \in [0, N]} \left\{ s - f(\pi) + \frac{K}{N}(m(\pi) - s) + (K + k_0) D(\pi, u) \right\}.$$

Since the maximand in the Bellman equation is an affine function of  $K$ , it is immediate that it is always optimal to choose either  $K = 0$  (all agents use  $S$  exclusively), or  $K = N$  (all agents use  $R$  exclusively). As the left-hand side is zero (a consequence of no discounting), and since  $K + k_0 > 0$  (because of the background signal), the Bellman equation can be rearranged as

$$0 = \max_{K \in [0, N]} \left\{ \frac{s - f(\pi) + \frac{K}{N}(m(\pi) - s)}{K + k_0} \right\} + D(\pi, u)$$

thereby demonstrating that the current choice does not depend on the continuation value. Simple algebra allows us to further simplify the problem by rewriting the Bellman equation so that the choice variable,  $K$ , appears only in the denominator:

$$0 = \max_{K \in [0, N]} \left\{ \frac{\frac{k_0}{N}(s - m(\pi)) - (f(\pi) - s)}{K + k_0} \right\} - \frac{1}{N}(s - m(\pi)) + D(\pi, u).$$

Following Bolton and Harris (2000), we define the *incentive to experiment* by

$$I(\pi) = \frac{f(\pi) - s}{s - m(\pi)}$$

when  $m(\pi) < s$ , and  $\infty$  otherwise. Note that when  $m(\cdot)$  and  $f(\cdot)$  are co-monotonic,  $I(\cdot)$  inherits their monotonicity properties.

When  $I(\pi) < k_0/N$ , the numerator in the reworked Bellman equation is positive and it is optimal to minimize the denominator by choosing  $K = 0$ ; when  $I(\pi) > k_0/N$ , the numerator is negative and it is optimal to maximize the denominator by choosing  $K = N$ . Values of  $\pi$  such that  $I(\pi) = k_0/N$  make the numerator zero and render all choices of  $K$  optimal. Thus, the state space can be divided into two regions such that in one region it is efficient for all to play  $S$  exclusively and in the other it is efficient for all to play  $R$  exclusively. The boundary between these two regions is given by the set  $\Pi_N^*$  of all  $\pi$  satisfying  $I(\pi) = k_0/N$ . Further, when  $I(\cdot)$  is monotonic, each region is simply-connected.

Note that the efficient intensity of experimentation exhibits a bang-bang feature, being  $N$  in one region of the state space, and  $0$  in the other. Thus, the efficient intensity is maximal when the incentive to experiment is high, and minimal when it is low.

For a given set of parameters  $\mu_\ell$ , Examples 2.1–2.3 have the same functions  $m(\cdot)$  and  $f(\cdot)$ , so they admit the same solution to the cooperative problem. With discrete distributions for the unknown average payoff per unit of time, the efficient strategies do not depend on whether the payoff-generating processes are Brownian motions, Poisson processes or more general Lévy processes. As we shall see later, this insight carries over to Markov perfect equilibria of the non-cooperative experimentation game.

### 3.2 The strategic problem

Now assume that there are  $N \geq 2$  players acting non-cooperatively and consider Markovian strategies with  $\pi$  as the state variable.

To characterize the best response correspondence, fix a state  $\pi$ . With  $k_n \in [0, 1]$  indicating player  $n$ 's action at that state and  $K = \sum_{n=1}^N k_n$ , let  $K_{-n} = K - k_n$ , which summarizes the actions of the other players. Proceeding in the same way as in the previous subsection, we find that the rearranged Bellman equation for player  $n$  is

$$0 = \max_{k_n \in [0,1]} \left\{ \frac{(K_{-n} + k_0)(s - m(\pi)) - (f(\pi) - s)}{k_n + K_{-n} + k_0} \right\} - (s - m(\pi)) + D(\pi, u_n).$$

Player  $n$ 's best response,  $k_n^*$ , is determined by looking at the sign of the numerator of the maximand:

$$k_n^* \begin{cases} = 0 & \text{if } I(\pi) < k_0 + K_{-n}, \\ \in [0, 1] & \text{if } I(\pi) = k_0 + K_{-n}, \\ = 1 & \text{if } I(\pi) > k_0 + K_{-n}. \end{cases} \quad (2)$$

If the players use symmetric strategies in equilibrium, then, whenever a positive frac-

tion of the resource is allocated to each arm, that fraction,  $k_N^\dagger$ , is calculated from the indifference condition in (2) together with  $K_{-n} = (N - 1)k_N^\dagger$ , i.e.

$$k_N^\dagger(\pi) = \frac{I(\pi) - k_0}{N - 1}.$$

Note that  $k_N^\dagger(\pi) = 0$  for  $\pi \in \Pi_1^*$ , i.e. all the players stop using  $R$  when a single agent would do so. Further, it is clear that whenever  $I(\pi)$  is monotonic, so is  $k_N^\dagger(\pi)$ .

If the players use asymmetric strategies in equilibrium, then, whenever all the other players are using  $S$  exclusively, so  $K_{-n} = 0$ , player  $n$ 's decision is the same as in the single-agent case, so she too would switch from  $R$  to  $S$  when the state is in  $\Pi_1^*$ . Further, as we will see in the next subsection, when fewer than  $N - 1$  players are using  $S$  exclusively, in equilibrium the remaining players use symmetric actions near to  $\Pi_1^*$  and again stop using  $R$  on  $\Pi_1^*$ .

Since the region of the state space where an  $N$ -agent cooperative plays  $R$  increases with  $N$ , and any Markov perfect equilibrium of the  $N$ -player experimentation game has players using  $R$  only where a single agent would, all these equilibria are inefficient. Further, close to the single-agent boundary the equilibrium intensity of experimentation is at most 1, whereas the intensity of experimentation for the cooperative is  $N$ .

### 3.3 Markov perfect equilibria

Sets of mutual best responses are given below and the resulting total experimentation schedule for  $N = 3$  is illustrated in Figure 1, which shows the intensity of experimentation as a function of the incentive to experiment.<sup>4</sup>

The equilibrium actions fall into three broad categories. For any particular level of  $I(\pi)$ : (1) all  $N$  players use either  $S$  exclusively or  $R$  exclusively; (2) exactly  $N - 1$  players use either  $S$  exclusively or  $R$  exclusively; (3) fewer than  $N - 1$  players use either  $S$  exclusively or  $R$  exclusively.

**Case 1** (leading to the horizontal sections in the figure)

Consider  $I(\pi) < k_0$ . Regardless of the choices of the others,  $I(\pi) < k_0 + K_{-n}$ , so  $k_n^*(\pi) = 0$  and it is dominant for all players to choose  $S$  exclusively. (If  $I(\pi) = k_0$  and  $N - 1$  players are choosing  $S$  exclusively, then it is still weakly dominant for the remaining player to choose  $S$  exclusively.)

Now, consider  $I(\pi) > k_0 + N - 1$ . Regardless of the choices of the others,  $I(\pi) > k_0 + K_{-n}$ , so  $k_n^*(\pi) = 1$  and it is dominant for all players to choose  $R$  exclusively. (If

---

<sup>4</sup>This is essentially Figure 4.1 from Bolton and Harris (2000).

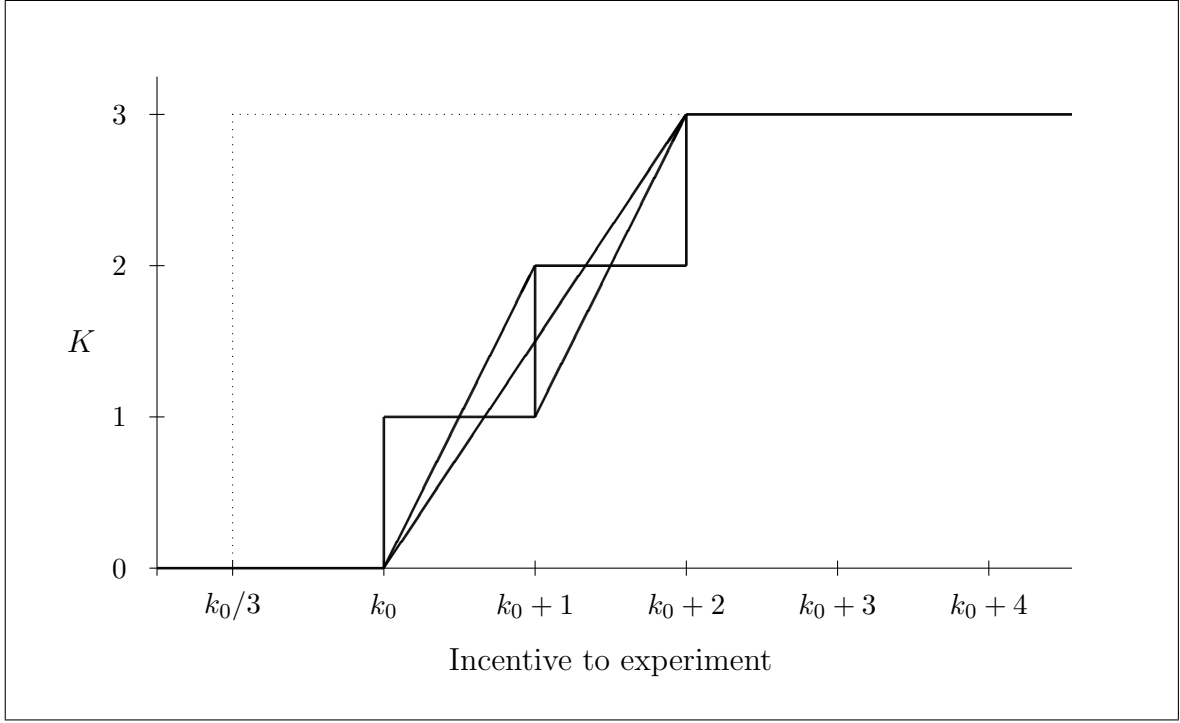


Figure 1: Intensity of experimentation in three-player equilibria

$I(\pi) = k_0 + N - 1$  and  $N - 1$  players are choosing  $R$  exclusively, then it is still weakly dominant for the remaining player to choose  $R$  exclusively.)

Last, consider  $k_0 + \ell - 1 < I(\pi) < k_0 + \ell$  for some  $\ell \in \{1, \dots, N - 1\}$ . If  $\ell$  others are playing  $R$  and  $N - \ell - 1$  others are playing  $S$ , then  $K_{-n} = \ell$ , so  $I(\pi) < k_0 + K_{-n}$  and  $k_n^*(\pi) = 0$ . (Again, if  $I(\pi) = k_0 + \ell$  so that  $I(\pi) = k_0 + K_{-n}$  then  $k_n^*(\pi) = 0$  is still a best response, but not unique.) If  $\ell - 1$  others are playing  $R$  and  $N - \ell$  others are playing  $S$ , then  $K_{-n} = \ell - 1$ , so  $I(\pi) > k_0 + K_{-n}$  and  $k_n^*(\pi) = 1$ . (Once again, if  $I(\pi) = k_0 + \ell - 1$  so that  $I(\pi) = k_0 + K_{-n}$  then  $k_n^*(\pi) = 1$  is still a best response, but not unique.) So, for  $I(\pi)$  in the stated interval, we have  $\ell$  players choosing  $R$  and the rest choosing  $S$ .

**Case 2** (leading to the vertical sections in the figure)

From the parenthetical remarks made when considering Case 1, it is clear that we can only have exactly one player indifferent between  $S$  and  $R$  when  $I(\pi) = k_0 + \ell$  for some  $\ell \in \{0, \dots, N - 1\}$ .

So, consider  $I(\pi) = k_0 + \ell$ ,  $\ell$  players choosing  $R$ ,  $N - \ell - 1$  choosing  $S$ , and the remaining player choosing *any*  $k \in (0, 1)$ . (If  $k = 0$  or  $1$  then we would be in Case 1.)

If player  $n$  is choosing  $R$ , then  $K_{-n} = \ell - 1 + k$ , so  $I(\pi) > k_0 + K_{-n}$  and  $k_n^*(\pi) = 1$  is indeed her best response. Similarly, if player  $n$  is choosing  $S$ , then  $K_{-n} = \ell + k$ , so  $I(\pi) < k_0 + K_{-n}$  and now  $k_n^*(\pi) = 0$  is her best response. Last, if player  $n$  is choosing neither  $R$  nor  $S$  exclusively, then  $K_{-n} = \ell$ , so  $I(\pi) = k_0 + K_{-n}$  and her choice of *any*

$k \in (0, 1)$  is indeed a best response. Thus the choices described in the previous paragraph are mutual best responses at the stated level of  $I(\pi)$ .

**Case 3** (leading to the sloping sections in the figure)

Note that in an equilibrium where more than one player is choosing neither  $R$  nor  $S$  exclusively, each of those players must be facing the *same* indifference condition, implying that each of those players must be choosing the *same*  $k \in (0, 1)$ . Also, we saw in Case 1 that when  $I(\pi) < k_0$  the dominant choice is  $S$ , and when  $I(\pi) > k_0 + N - 1$  it is dominant to choose  $R$ , so we can restrict attention to  $k_0 \leq I(\pi) \leq k_0 + N - 1$ .

Consider  $I(\pi) = k_0 + \ell + [N - L - 1]k$  for some  $L \in \{0, \dots, N - 2\}$ , some  $\ell \in \{0, \dots, L\}$ , and some  $0 < k < 1$ , with  $\ell$  players choosing  $R$ ,  $L - \ell$  choosing  $S$ , and the remaining  $N - L$  players choosing the same  $k \in (0, 1)$ . (Again, if  $k = 0$  or  $1$  then we would be in Case 1.)

If player  $n$  is choosing  $R$ , then  $K_{-n} = \ell - 1 + [N - L]k$ , so  $I(\pi) > k_0 + K_{-n}$  and  $k_n^*(\pi) = 1$  is indeed her best response. Similarly, if player  $n$  is choosing  $S$ , then  $K_{-n} = \ell + [N - L]k$ , so  $I(\pi) < k_0 + K_{-n}$  and now  $k_n^*(\pi) = 0$  is her best response. Last, if player  $n$  is choosing neither  $R$  nor  $S$  exclusively, and the others like her are choosing the same  $k \in (0, 1)$  as each other, then  $K_{-n} = \ell + [N - L - 1]k$ , so  $I(\pi) = k_0 + K_{-n}$  and her choice of that same  $k \in (0, 1)$  as them is indeed a best response. Once again, the choices described in the previous paragraph are mutual best responses at the stated level of  $I(\pi)$ .

All the MPE are just combinations of these three cases. In Figure 1 above, the dotted line is the efficient outcome, and we can see the equilibrium experimentation that approaches this the closest is the upper envelope consisting of alternating horizontal and sloping solid lines, together with a jump at  $k_0$  – this is the equilibrium that maximizes total experimentation at any given belief, and, as such, maximizes aggregate payoffs as we now show (cf. Bolton and Harris, 2000).

### 3.4 Constrained efficiency

We consider a planner who wants to maximize the average payoff of the agents such that the actions assigned to the agents constitute a MPE of the non-cooperative experimentation game. So define  $\mathcal{E}(\pi)$  as the set of all  $(k_1, \dots, k_N)$  such that  $k_n$  is optimal for player  $n$  given  $\pi$  and  $K_{-n}$  according to the best responses in (2).

Paralleling the analysis in section 3.1, we can write the Bellman equation for the value function of this problem as

$$0 = \max_{(k_1, \dots, k_N) \in \mathcal{E}(\pi)} \left\{ \frac{\frac{k_0}{N}(s - m(\pi)) - (f(\pi) - s)}{K + k_0} \right\} - \frac{1}{N}(s - m(\pi)) + D(\pi, u)$$

and again the optimal choice depends on whether or not  $I(\pi) < k_0/N$ . If it is, the planner minimizes the denominator of the maximand by choosing  $K = 0$ , which is incentive compatible since  $(0, \dots, 0) \in \mathcal{E}(\pi)$  when  $I(\pi) < k_0/N < k_0$ . Otherwise, the planner wants to maximize the denominator, and achieves this by choosing  $(k_1, \dots, k_N) \in \mathcal{E}(\pi)$  such that  $K$  maximizes the intensity of experimentation at  $\pi$ . That is, the allocation of actions to agents results in the total experimentation schedule illustrated by the upper envelope in Figure 1.

## 4 Concluding Remarks

We have seen that under the separability condition, the players' best responses depend only on the expected current payoff from the risky arm and the expected full-information payoff. Once a discrete prior distribution for the unknown average payoff per unit of time has been specified, these two expected payoffs are fully determined – the set of Markov perfect equilibria is then invariant to the specification of the payoff-generating process.

As to the examples with a continuous prior distribution, we note that in Example 2.4 (Brownian noise, normal prior) the precision of the posterior distribution increases unboundedly with time, as does the denominator of the variance in Example 2.5 (Poisson noise, Gamma prior) – consequently the posterior probability density function becomes concentrated on a narrow domain of the support. If we approximate the normal or Gamma distribution with a discrete distribution (Example 2.1 or 2.2) then, over time, the beliefs become more and more concentrated on the discrete values closest to the true parameter  $\mu$  – this suggests that we could take the ‘engineering’ approach and focus on discrete distributions, with the specification of the payoff-generating processes being irrelevant.<sup>5</sup>

Of course, the evolution of the agents' posterior belief does depend on how the payoffs are generated, as do the players' equilibrium payoffs. To calculate the latter, one has to solve a differential equation that depends on the functional operator  $D(\cdot, \cdot)$ : if the agents' prior belief is a two-point distribution this will be an ordinary differential equation; otherwise it will be a partial differential equation. In the examples with Brownian payoffs the differential equation is second-order, whereas in the examples with Poisson payoffs it is first-order but with a coupled difference (or ‘jump’) term.

## References

BOLTON, P. AND C. HARRIS (1999): “Strategic Experimentation,” *Econometrica*, **67**, 349–374.

---

<sup>5</sup>But note that if for  $T$  very large the two closest neighbours of  $\mu$  in the support of  $H(\cdot; \pi_T)$  are  $\mu_i$  and  $\mu_\ell$  with  $\mu_i < \mu < \mu_\ell$ , then, although  $m(\pi_T) \simeq \mu$ , we would have  $\text{Var}[\mu | \pi_T] \simeq (\mu_\ell - \mu)(\mu - \mu_i) \gg 0$ .

- BOLTON, P. AND C. HARRIS (2000): “Strategic Experimentation: the Undiscounted Case,” in *Incentives, Organizations and Public Economics – Papers in Honour of Sir James Mirrlees*, ed. by P.J. Hammond and G.D. Myles. Oxford: Oxford University Press, 53–68.
- CHERNOFF, H. (1968): “Optimal Stochastic Control,” *Sankhyā*, **30**, 221–252.
- COHEN, A. AND E. SOLAN (2009): “Bandit Problems with Lévy Payoff Processes,” working paper, Tel Aviv University; archived at <http://arxiv.org/abs/0906.0835v1>.
- DEGROOT, M. (1970): *Optimal Statistical Decisions*. New York: McGraw Hill.
- JOVANOVIĆ, B. (1979): “Job Matching and the Theory of Turnover,” *Journal of Political Economy*, **87**, 972–990.
- KELLER, G. AND S. RADY (1999): “Optimal Experimentation in a Changing Environment,” *Review of Economic Studies*, **66**, 475–507.
- KELLER, G. AND S. RADY (2003): “Price Dispersion and Learning in a Dynamic Differentiated-Goods Duopoly,” *RAND Journal of Economics*, **34**, 138–165.
- KELLER, G. AND S. RADY (2010): “Strategic Experimentation with Poisson Bandits,” *Theoretical Economics*, **5**, 275–311.
- KELLER, G., S. RADY AND M. CRIPPS (2005): “Strategic Experimentation with Exponential Bandits,” *Econometrica*, **73**, 39–68.
- LIPTSER, R.S. AND A.N. SHIRYAYEV (1977): *Statistics of Random Processes I*. New York: Springer-Verlag.
- MOSCARINI, G. AND F. SQUINTANI (2004): “Competitive Experimentation with Private Information,” Cowles Foundation Discussion Paper 1489.