GOVERNANCE AND THE EFFICIENCY
OF ECONOMIC SYSTEMS
GESY

Discussion Paper No. 520

Undiscounted Bandit Games

Godfrey Keller*
Sven Rady**

* University of Oxford
** University of Bonn

September 2015

Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 is gratefully acknowledged.

GOVERNANCE AND THE EFFICIENCY OF ECONOMIC SYSTEMS
GESY

Discussion Paper No. 520

Undiscounted Bandit Games

Godfrey Keller*
Sven Rady**

* Cardiff University
** University of Mannheim

September 2015

Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 is gratefully acknowledged.

Undiscounted Bandit Games*

Godfrey Keller† Sven Rady‡

September 14, 2015

Abstract

We analyze continuous-time games 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, the unique symmetric Markov perfect equilibrium can be computed in a simple closed form involving only 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 when 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 Chris Harris, Albert N. Shiryaev, Bruno Strulovici, and seminar participants in Budapest, Cambridge, the European University Institute in Florence, Keele, Oxford, Shanghai, Southampton, and UCL/Birkbeck. We thank the Center for Economic Studies at the University of Munich and the Studienzentrum Gerzensee for their hospitality. Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 and GRK 801 is gratefully acknowledged.

†Department of Economics, University of Oxford, Manor Road Building, Oxford OX1 3UQ, UK.

‡Department of Economics and Hausdorff Center for Mathematics, University of Bonn, Adenauerallee 24-42, D-53113 Bonn, 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. Actual payoffs are given by stochastic processes with stationary and independent increments. Assuming that all actions and payoffs are public information, we restrict players to Markov strategies with the common posterior belief about the unknown parameter as the natural state variable.

In this setting, the expected infinitesimal change in a players' payoff function is proportional to the overall intensity of experimentation performed at the given point in time. As in Bolton and Harris (2000), this *separability condition* implies that best responses can be computed without knowledge of a player's value function. In fact, given the current belief, a player's optimal action depends only on the intensity of experimentation performed by the other players, the expected current payoff of the risky arm, and the expected full-information payoff – it does *not* depend on the precise specification of the payoff-generating process.

This insight allows us to handle a much larger class of priors and payoff-generating processes than the existing literature on bandit-based multi-agent learning in continuous time. More specifically, we present five examples that fit in our general framework. 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) where the prior is a Bernoulli distribution. In the second, payoffs come from a Poisson process with unknown intensity, and the agents' prior belief about this intensity is again an arbitrary discrete distribution; this generalizes the setup of Keller, Rady and Cripps (2005) and Keller and Rady (2010) who also assume a Bernoulli prior. These two examples are special cases of a third in which payoffs are generated by a Lévy process, that is, a continuous-time process with independent and stationary increments; a single-agent version of this setup was explored by Cohen and Solan (2013) under a Bernoulli prior and a specific assumption on the distribution of jumps. In the fourth example, payoffs stem again from a Brownian motion with unknown drift, but the prior belief is a normal distribution; 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 (2010).

This broadening of the class of payoff-generating processes, and the generalization from Bernoulli to arbitrary discrete priors in particular, is not entirely without costs, however. In fact, while the specifics of the payoff-generating process do not affect mutual best responses at a *given* belief, they are highly relevant when it comes to 'synthesizing'

belief-contingent action profiles into a profile of Markov strategies that induces a well-defined law of motion for posterior beliefs. In the models of Bolton and Harris (1999, 2000), Keller, Rady and Cripps (2005) and Keller and Rady (2010), beliefs evolve on the unit interval, and this allows for a space of admissible Markov strategies large enough to accommodate the discontinuities of actions with respect to beliefs which are an immutable feature of asymmetric equilibria. Such a strategy space can also be defined in the Poisson-Gamma case where only one component of the posterior distribution (the shape parameter) evolves stochastically, and in a special case of Lévy payoffs where the size of observed jumps is the only source of information, so posterior beliefs are piecewise constant. In general, however, we must invoke results on the existence and uniqueness of solutions to stochastic differential equations that rely on Lipschitz continuity of coefficients. This rules out asymmetric equilibria but, as our main result shows, the space of Lipschitz continuous strategies is large enough to ensure existence of a unique symmetric Markov perfect equilibrium. The equilibrium strategy has a simple explicit form, moreover.

Besides Bolton and Harris (2000), the undiscounted limit of a continuous-time stochastic game with one-dimensional state space has also been studied in Harris (1988, 1993) and Bergemann and Välimäki (1997, 2002), yielding a much simpler characterization of equilibria than under discounting. More recent applications of this methodology to single-agent experimentation problems can be found in Bonatti (2011) and Peitz, Rady and Trepper (2015). It should be noted, however, that the advantages of considering the undiscounted game hinge on the stationarity of the environment in which the players are learning (meaning in particular that the average payoff per unit of time does not change over time). When payoffs are generated by a Brownian motion with an unknown drift that is subject to Markovian state-switching between a high and a low level as in Keller and Rady (1999, 2003), for example, the separability condition is violated, and the computation of best responses requires knowledge of the value function.

The rest of the paper is organized as follows. Section 1 sets up the game, introduces the separability condition and states our assumptions on priors, payoff-generating processes and strategy spaces. Section 2 proves existence of a unique symmetric Markov perfect equilibrium under these assumptions. Section 3 presents our five examples. Section 4 briefly considers a setting where separability fails because of state switching. Section 5 offers some concluding remarks.

1 The Experimentation Game

Time $t \in [0, \infty)$ is continuous. 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.

Assumption 1 (Payoffs) *There are independent stochastic processes $S^1, \dots, S^N$, R^0, $R^1, \dots, R^N$, a real number s and a real-valued random variable μ such that: (i) the processes $S^n - st$, $n = 1, \dots, N$, are identically distributed martingales independent of μ; (ii) conditional on the realization of μ, the processes $R^n - \mu t$, $n = 0, 1, \dots, N$, are identically distributed martingales.*

We interpret S^n as the payoff-generating process on player n's safe arm, and assume that its expected flow payoff s is commonly known. For $n = 1, \dots, N$, we interpret R^n as the payoff-generating process on player n's risky arm, and assume that its expected flow payoff μ is unknown to the players. The process R^0, finally, provides a background signal, ensuring that the players eventually learn the value of μ even if they all play safe all the time.

Let $k_{n,t} \in [0, 1]$ be the fraction of the available resource that player n allocates to the risky arm at time t; this fraction is required to be measurable with respect to the information that the player possesses at time t. The player's cumulative payoff up to time T is then given by the time-changed process $S^n_{T-\tau^n(T)} + R^n_{\tau^n(T)}$ where $\tau^n(T) = \int_0^T k_{n,t}\, dt$ measures the *operational time* that the risky arm has been used. In view of property (iii), the player's expected payoff (conditional on μ) is $\mathrm{E}\left[\int_0^T \{(1 - k_{n,t})s + k_{n,t}\mu\}\, dt\right]$.

The players start with a common prior belief about μ, and thereafter all observe each other's actions and outcomes as well as the time-changed process $R^0_{\tau^0(t)}$ where $\tau^0(t) = k_0 t$ with $k_0 > 0$ exogenously given and arbitrarily small. So they hold common posterior beliefs throughout time.

Assumption 2 (Beliefs) *At time t the players believe that μ has a cumulative distribution function $H(\cdot; \pi_t)$, where π_t is a sufficient statistic for the observations on $R^0, \dots, R^N$ up to time t, and H represents a conjugate family of distributions. The safe expected flow payoff s lies in the interior of the support of $H(\cdot; \pi_0)$.*

With s lying in the interior of the support of the prior distribution of μ, each player has an incentive to learn the quality of the risky arm.

The evolution of the sufficient statistic over time is driven by $N + 1$ distinct sources of information, namely the observations on $R^0, R^1, \dots R^N$. The following assumption specifies how beliefs evolve when only one of these sources is observed, and at full intensity.

Assumption 3 (Generator) *Fix a player n and consider the time-invariant action profile for which $k_n = 1$ whereas $k_j = 0$ for all $j \in \{0, \dots, N\} \setminus \{n\}$. Then the corresponding process π is a time-homogeneous Markov process with infinitesimal generator $\mathcal{G}^n$.*

By property (ii) in Assumption 1, we have $\mathcal{G}^1 = \mathcal{G}^2 = \ldots = \mathcal{G}^N$, for which we simply write $\mathcal{G}$.

If we change player n's time-invariant intensity to $k_n \in [0,1]$ while keeping all other intensities at zero, the resulting deceleration of the process of observations implies the scaled-down generator $k_n\mathcal{G}$ for the sufficient statistic; see Dynkin (1965), for example. The same applies to the background signal, of course, with associated generator $k_0\mathcal{G}$. Finally, a repeated application of Trotter (1959) invoking the conditional independence of the processes $R^0, \ldots, R^N$ establishes that the generator associated with time-invariant intensities $(k_0, k_1, \ldots, k_N) \in [0,1]^{N+1}$ is $(k_0 + K)\mathcal{G}$ where $K = \sum_{n=1}^N k_n$ measures how much of the N available units of the resource is allocated to risky arms – we call it the *intensity of experimentation*.

The fact that the infinitesimal generator of the sufficient statistic π is linear in $k_0 + K$ will play a crucial role in our analysis. In more heuristic fashion, we can rewrite this *separability condition* as

$$\mathrm{E}\left[u(\pi_{t+dt}) \mid \pi_t, k_{1,t}, \ldots, k_{N,t}\right] = u(\pi_t) + (k_0 + K_t)\,\mathcal{G}u(\pi_t)\,dt, \tag{1}$$

where $u : [0,1] \to I\!R$ is any function in the domain of $\mathcal{G}$, and $K_t = \sum_{n=1}^N k_{n,t}$.

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:

$$m(\pi) = \int \mu \; dH(\mu;\pi), \quad f(\pi) = \int (s \vee \mu) \; dH(\mu;\pi) = sH(s;\pi) + \int_s^\infty \mu \; dH(\mu;\pi).$$

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 $\mathrm{E}_t[m(\pi_T)] = m(\pi_t)$ and $\mathrm{E}_t[f(\pi_T)] = f(\pi_t)$ for all $T > t$, i.e. both $m(\pi_t)$ and $f(\pi_t)$ are martingales with respect to the players' information sets.

Players do not discount future payoffs, and are instead assumed to use the catching-up criterion.[1] This means that player n chooses allocations $k_{n,t}$ so as to maximize

$$\mathrm{E}\left[\int_0^\infty \left\{(1 - k_{n,t})s + k_{n,t}m(\pi_t) - f(\pi_t)\right\} \, dt\right].$$

Here, 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. Note that a player's payoff depends on others' actions only through their impact on the evolution of the sufficient statistic.

[1] For a discussion of this objective and the role of the background signal, see Bolton and Harris (2000).

The above objective highlights the potential for the sufficient statistic to serve as a state variable; from now on, we shall restrict players to strategies that are Markovian with respect to this variable. More precisely, the players' common strategy space is a non-empty set $\mathcal{A}$ of functions from the state space Π (consisting of all possible realizations of the sufficient statistic) to $[0,1]$.

Assumption 4 (Strategies) *The common strategy space $\mathcal{A}$ has the property that, starting from any $\pi \in \Pi$, any strategy profile $(k_1, \ldots, k_N) \in \mathcal{A}^N$ induces a well-defined and unique law of motion for π_t.*

This assumption implies that any strategy profile $(k_1, \ldots, k_N) \in \mathcal{A}^N$ gives rise to well-defined payoff functions

$$u_n(\pi|k_1, \ldots, k_N) = \mathrm{E}\left[\int_0^\infty \left\{(1 - k_n(\pi_t)s + k_n(\pi_t)m(\pi_t) - f(\pi_t)\right\}\, dt \;\middle|\; \pi_0 = \pi\right].$$

Strategy $k_n \in \mathcal{A}$ is a *best response* against $k_{\neg n} = (k_1, \ldots, k_{n-1}, k_{n+1}, \ldots, k_N) \in \mathcal{A}^{N-1}$ if $u_n(\pi|k_n, k_{\neg n}) \geq u_n(\pi|\tilde{k}_n, k_{\neg n})$ for all $\pi \in \Pi$ and all $\tilde{k}_n \in \mathcal{A}$. A *Markov perfect equilibrium (MPE)* is a profile of strategies $(k_1, \ldots, k_N) \in \mathcal{A}^N$ that are mutually best responses. Such an equilibrium is *symmetric* if $k_1 = k_2 = \ldots = k_N$.

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 ∞ otherwise. Note that when the functions m and f are co-monotonic, I inherits their monotonicity property.

Assumption 5 (Regularity) *For any positive real numbers $a < b$, the function $k : \Pi \to [0,1]$ defined by*

$$k(\pi) = \begin{cases} 0 & \text{if } I(\pi) \leq a, \\ \frac{I(\pi)-a}{b-a} & \text{if } a < I(\pi) < b, \\ 1 & \text{if } I(\pi) \geq b \end{cases}$$

is an element of $\mathcal{A}$.

The type of strategy considered in Assumption 5 arises in symmetric Markov perfect equilibria of the experimentation game, to which we turn now.

2 Symmetric Markov Perfect Equilibrium

Suppose that all players except player n use the strategy $k^\dagger \in \mathcal{A}$. By using the separability condition in equation (1), the Bellman equation for player n's best response then becomes

$$0 = \max_{k_n \in [0,1]} \left\{ s - f(\pi) + k_n[m(\pi) - s] + [k_0 + (N-1)k^\dagger(\pi) + k_n]\, \mathcal{G}u_n(\pi) \right\}.$$

As the left-hand side is zero (a consequence of no discounting) and $k_0 + (N-1)k^\dagger(\pi) + k_n$ is positive (because of the background signal), the Bellman equation can be rearranged as

$$0 = \max_{k_n \in [0,1]} \left\{ \frac{s - f(\pi) + k_n[m(\pi) - s]}{k_0 + (N-1)k^\dagger(\pi) + k_n} \right\} + \mathcal{G}u_n(\pi),$$

which demonstrates that the optimal k_n does not depend on continuation values. Straightforward algebra allows us to further simplify the problem by rewriting the Bellman equation so that k_n appears only in the denominator:

$$0 = \max_{k_n \in [0,1]} \left\{ \frac{[k_0 + (N-1)k^\dagger(\pi)][s - m(\pi)] - [f(\pi) - s]}{k_0 + (N-1)k^\dagger(\pi) + k_n} \right\} - [s - m(\pi)] + \mathcal{G}u_n(\pi).$$

When $I(\pi) < k_0 + (N-1)k^\dagger(\pi)$, the numerator in the reworked Bellman equation is positive and it is optimal to minimize the denominator by choosing $k_n = 0$; when $I(\pi) > k_0 + (N-1)k^\dagger(\pi)$, the numerator is negative and it is optimal to maximize the denominator by choosing $k_n = 1$; when $I(\pi) = k_0 + (N-1)k^\dagger(\pi)$, the numerator is zero and all choices of k_n are optimal.

There are three different ways, therefore, in which $k_n = k^\dagger(\pi)$ can be an optimal choice for player n: either $k^\dagger(\pi) = 0$ and $I(\pi) \leq k_0$, or $k^\dagger(\pi) = 1$ and $I(\pi) \geq k_0 + N - 1$, or $0 < k^\dagger(\pi) < 1$ and $I(\pi) = k_0 + (N-1)k^\dagger(\pi)$. This pins down $k^\dagger(\pi)$ in terms of the incentive to experiment, $I(\pi)$, the strength of the background signal, k_0, and the number of players, N:

$$k^\dagger(\pi) = \begin{cases} 0 & \text{if } I(\pi) \leq k_0, \\ \frac{I(\pi) - k_0}{N-1} & \text{if } k_0 < I(\pi) < k_0 + N - 1, \\ 1 & \text{if } I(\pi) \geq k_0 + N - 1. \end{cases}$$

Proposition. *Under Assumptions 1–5, all players using the strategy $k^\dagger$ constitutes the unique symmetric Markov perfect equilibrium of the experimentation game.*

Proof: By Assumption 5, the strategy $k^\dagger$ is an element of $\mathcal{A}$; by Assumption 4, all players using this strategy gives rise to a well-defined common payoff function $u^\dagger$. By standard results, this function satisfies

$$0 = s - f(\pi) + k^\dagger(\pi)[m(\pi) - s] + [k_0 + Nk^\dagger(\pi)]\, \mathcal{G}u^\dagger(\pi),$$

and the above arguments imply that

$$0 \geq s - f(\pi) + k(\pi)[m(\pi) - s] + [k_0 + (N-1)k^\dagger(\pi) + k(\pi)]\,\mathcal{G}u^\dagger(\pi)$$

for any strategy $k \in \mathcal{A}$. The standard verification argument now shows that all players using the strategy $k^\dagger$ indeed constitutes an MPE. Uniqueness (as usual, up to changes on a null set of states) follows from the fact that, in view of the above arguments, the common strategy in any symmetric MPE must agree with $k^\dagger$ almost everywhere. ■

Note that the set of beliefs for which $k^\dagger(\pi) = 0$ is independent of the number of players and actually the same as for a single agent experimenting in isolation. This is a stark manifestation of the incentive to free-ride on information generated by others. In the terminology coined by Bolton and Harris (1999), it means that there is no "encouragement effect": the prospect of subsequent experimentation by other players provides a player *no* incentive to increase the current intensity of experimentation and thereby shorten the time at which the information generated by the other players arrives. Intuitively, this simply reflects our assumption that players do not discount future payoffs and hence are indifferent as to their timing. Formally, the absence of the encouragement effect is a direct consequence of the separability condition: as the value of future experimentation by other players is captured by a player's equilibrium continuation values, yet best responses are independent of those continuation values, there is no channel for future experimentation by others to impact current actions.

Free-riding can also be seen in the fact that $k^\dagger$ is non-increasing in N, and decreasing where it assumes interior values. (See Figure 3 at the end of Section 3.1 for an illustration of these two points.) The dependence of the overall intensity of experimentation on the number of players is less clear cut: roughly speaking, $Nk^\dagger$ increases in N at beliefs where $k^\dagger$ requires exclusive use of the risky arm, but decreases at beliefs where both arms are used simultaneously. Further, any weak monotonicity of I in a component of π is inherited by $k^\dagger$.

Finally, by the martingale convergence theorem, beliefs converge almost surely to the degenerate distribution concentrated on the true value of μ; therefore $f(\pi)$ converges to either s or μ, and so $k^\dagger(\pi)$ converges to either 0 or 1.

3 Examples

This section presents five specifications of priors, payoff-generating processes and strategy spaces that satisfy Assumptions 1–5. For more details of Example 3.1, see Bolton and Harris (1999, 2000), and for the discounted version of Example 3.2, see Keller, Rady

and Cripps (2005) and Keller and Rady (2010).[2] A discounted single-agent version of Example 3.3 with a two-point prior is solved in Cohen and Solan (2013). Models in which agents observe stochastic processes and have beliefs like those in Examples 3.4 and 3.5 can be found in Jovanovic (1979) and Moscarini and Squintani (2010), respectively.

3.1 Brownian payoffs, discrete prior

We start with the case of a two-point prior. For $n = 0, 1, \ldots, N$, we let $R^n = \mu t + \sigma Z^n$ where $(Z^0, Z^1, \ldots, Z^N)$ is an $(N+1)$-dimensional Wiener process, $\sigma > 0$ and $\mu \in \{\mu_0, \mu_1\}$ with $\mu_0 < s < \mu_1$.[3]

Let π_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, and 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 Shiryayev (1977, Theorem 9.1) that for a single player who allocates his entire resource to the risky arm, π_t is a diffusion process with zero drift and diffusion coefficient $(\mu_1 - \mu_0)\,\sigma^{-1}\pi_t(1-\pi_t)$ relative to the player's information filtration.[4] This implies

$$\mathcal{G}u(\pi) = \frac{1}{2\sigma^2}(\mu_1 - \mu_0)^2\pi^2(1-\pi)^2\,u''(\pi).$$

There is a straightforward generalization to the case where μ can take any one of $L+1$ possible values $\mu_0 < \mu_1 < \ldots < \mu_{L-1} < \mu_L$ with $\mu_0 < s < \mu_L$. Players' beliefs now become an L-vector $\pi = (\pi_1, \ldots, \pi_L)$ where π_ℓ is the probability assigned to $\mu = \mu_\ell$. The state space is $\Pi = \Delta^L$, the standard L-simplex, and, with $\pi_0 = 1 - \sum_{\ell=1}^L \pi_\ell$,

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

An extension of Liptser and Shiryayev (1977, Theorem 9.1) shows that for a single player allocating his entire resource to the risky arm, π_t is a driftless L-dimensional diffusion

[2] Keller and Rady (2015) consider a 'bad news' variant of this example in which the processes S^n and R^n represent the cumulative *cost* of using an arm rather than the cumulative payoff.

[3] Note that the same parameter σ applies in both states of the world. If this were not the case, the players could infer the true state in an instant from the quadratic variation of risky payoffs.

[4] More precisely, they show that the belief evolves according to $d\pi_t = \sigma^{-1}\,\pi_t[\mu_1 - m(\pi_t)]\,d\bar{Z}_t$ where the *innovation* process $\bar{Z}_t$, given by $d\bar{Z}_t = \sigma^{-1}\left([\mu - m(\pi_t)]\,dt + \sigma\,dZ_t\right)$, is a Wiener process relative to the player's information filtration.

process with instantaneous variance-covariance matrix given by

$$\mathrm{Cov}\left[d\pi_{i,t}, d\pi_{\ell,t} \mid \pi_t\right] = \left[\pi_{i,t}\left(\mu_i - m(\pi_t)\right)\sigma^{-1}\right]\left[\pi_{\ell,t}\left(\mu_\ell - m(\pi_t)\right)\sigma^{-1}\right]dt.$$

Thus,

$$\mathcal{G}u(\pi) = \frac{1}{2\sigma^2}\sum_{i=1}^{L}\sum_{\ell=1}^{L}\pi_i\,\pi_\ell\left(\mu_i - m(\pi)\right)\left(\mu_\ell - m(\pi)\right)\frac{\partial^2 u(\pi)}{\partial\pi_i\,\partial\pi_\ell}\,.$$

For $L = 1$, the case analyzed in Bolton and Harris (2000), the presence of background information allows one to invoke a result of Engelbert and Schmidt (1984) whereby any profile of Markov strategies in $\mathcal{M}([0,1],[0,1])$, the set of Borel measurable functions from the unit interval to itself, implies a unique solution for the belief dynamics.[5] As this result does not generalize to higher dimensions, the set of admissible strategies for $L \geq 2$ will necessarily be a strict subset of $\mathcal{M}(\Delta^L,[0,1])$. In view of standard existence and uniqueness results for solutions of stochastic differential equations, a natural choice is $\mathcal{A} = \mathcal{L}(\Delta^L,[0,1])$, the set of all Lipschitz continuous functions from the simplex to the unit interval.

As the partial derivatives of the incentive to experiment I are clearly bounded on the compact set $\Pi(a,b) = \{\pi \in \Pi : a \leq I(\pi) \leq b\}$, Assumption 5 holds trivially. Like the functions m and f, moreover, I and the equilibrium strategy $k^\dagger$ are non-decreasing in π.

Figures 1 and 2 illustrate the case where $L = 2$. (In all the figures, $k_0 = 0.2$; when μ

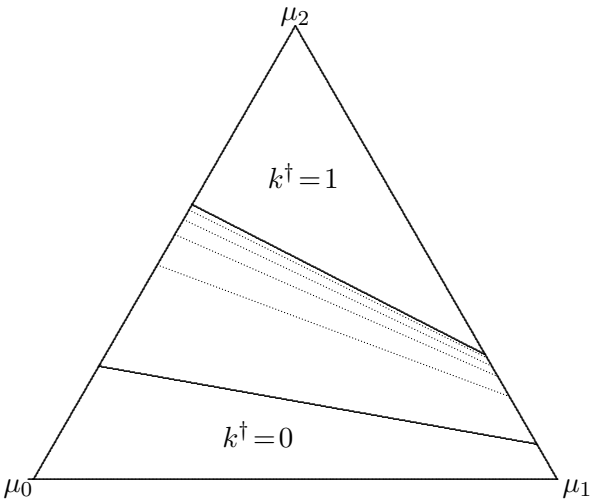

Figure 1: Equilibrium actions for $L = 2$ and $\mu_0 < \mu_1 < s < \mu_2$

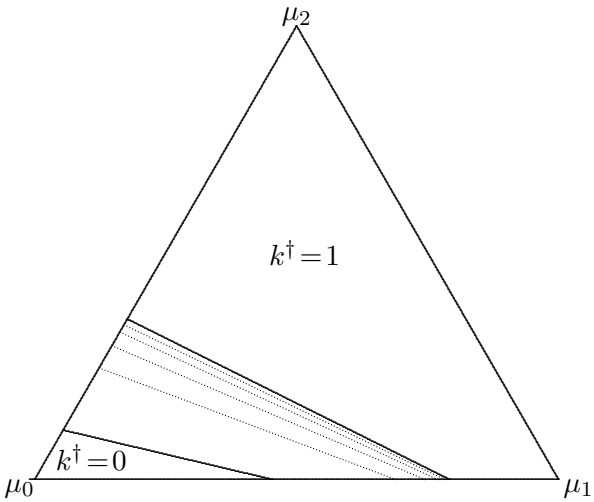

Figure 2: Equilibrium actions for $L = 2$ and $\mu_0 < s < \mu_1 < \mu_2$

has a discrete distribution, we use parameter values $\mu_0 = 2$, $\mu_1 = 5$, $\mu_2 = 8$; in this pair of figures, $s = 6$ on the left, $s = 4$ on the right, and in both cases $N = 4$.) The solid lines are

[5] See also Section 5.5 of Karatzas and Shreve (1988).

the boundaries of the sets of beliefs at which the equilibrium requires full experimentation ($k^\dagger = 1$) and no experimentation ($k^\dagger = 0$), respectively. The dotted lines are level curves of $k^\dagger$ for the experimentation intensities 0.2, 0.4, 0.6 and 0.8. A comparison of the two figures exhibits the familiar property that a decrease in the reward from the safe action gives the players an increased incentive to experiment.

Figure 3 illustrates the effect that increasing the number of players has on the equilibrium actions. On the horizontal axis we set $\pi_1 = \pi_2$ and let that common belief range

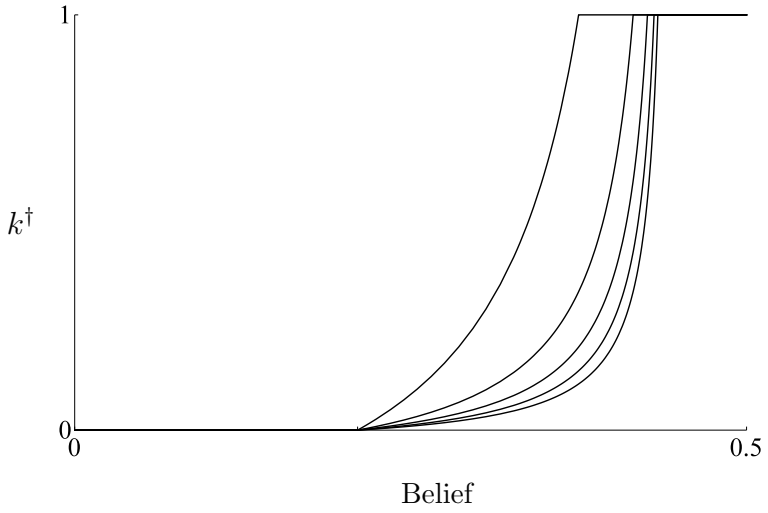

Figure 3: Equilibrium actions for $L = 2$, $\pi_1 = \pi_2$ and $N \in \{2, 4, 6, 8, 10\}$

from 0 to 0.5: so it is a slice through the simplex from the μ_0-vertex to the midpoint of the opposite edge. (In this figure, $s = 6$, and N varies from 2 for the leftmost curve to 10 for the rightmost curve; so the second curve from the left has the same parameter values as in Figure 1.) As can clearly be seen, the set of beliefs where players use the safe arm exclusively is independent of the number of players, reflecting the absence of the encouragement effect. Free-riding is also evident: as the number of players, N, increases, $k^\dagger$ decreases at any belief where $0 < k^\dagger < 1$. In this range of beliefs, the curves become more convex as N increases – less steep for low beliefs, and much steeper for high beliefs.

3.2 Poisson payoffs, discrete prior

As in Example 3.1, we start with the case of a two-point prior and let $R^0, R^1, \ldots, R^N$ be independent Poisson processes with common intensity μ where $\mu \in \{\mu_0, \mu_1\}$ and $\mu_0 < s < \mu_1$. We can again take π_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 3.1.

Now, consider a single player allocating his entire resource to the risky arm. With

probability $m(\pi_t)\,dt$, he will observe a positive increment between t and $t+dt$, and his belief jumps to

$$j(\pi_t) = \frac{\pi_t \mu_1}{m(\pi_t)}$$

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

$$d\pi_t = -(\mu_1 - \mu_0)\,\pi_t(1-\pi_t)\,dt.$$

So, we have

$$\begin{aligned}
&\mathrm{E}\left[u(\pi_{t+dt}) \mid \pi_t\right] \\
&\quad = \; m(\pi_t)\,u(j(\pi_t))\,dt + \Big(1 - m(\pi_t)\,dt\Big)\Big(u(\pi_t) - (\mu_1-\mu_0)\,\pi_t(1-\pi_t)\,u'(\pi_t)\,dt\Big) \\
&\quad = \; u(\pi_t) + \Big\{m(\pi_t)\left[u(j(\pi_t)) - u(\pi_t)\right] - (\mu_1-\mu_0)\,\pi_t(1-\pi_t)\,u'(\pi_t)\Big\}\,dt,
\end{aligned}$$

or

$$\mathcal{G}u(\pi) = m(\pi)\left[u(j(\pi)) - u(\pi)\right] - (\mu_1-\mu_0)\,\pi(1-\pi)\,u'(\pi).$$

As in Example 3.1, there is a simple generalization to the case where μ can take any one of $L+1$ possible values. 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, beliefs jump to

$$j(\pi_t) = \frac{1}{m(\pi_t)}\left(\pi_{1,t}\,\mu_1, \ldots, \pi_{\ell,t}\,\mu_\ell, \ldots, \pi_{L,t}\,\mu_L\right);$$

if no increment arrives, beliefs adjust infinitesimally by

$$d\pi_{\ell,t} = -\pi_{\ell,t}\left(\mu_\ell - m(\pi_t)\right)dt.$$

This leads to

$$\mathcal{G}u(\pi) = m(\pi)\left[u(j(\pi)) - u(\pi)\right] - \sum_{\ell=1}^{L} \pi_\ell\left(\mu_\ell - m(\pi)\right)\frac{\partial u(\pi)}{\partial \pi_\ell}\,.$$

For $L = 1$, the case analyzed in Keller, Rady and Cripps (2005) and Keller and Rady (2010), one can take $\mathcal{A}$ to be the set of functions from the unit interval to itself which are left-continuous and piecewise Lipschitz continuous; as beliefs drift down deterministically in between Poisson events, these properties allow one to construct belief dynamics in a pathwise fashion. When $L \geq 2$, $\mathcal{A} = \mathcal{L}(\Delta^L, [0,1])$ is again a natural choice.

The functions I and $k^\dagger$ are the same as in Example 3.1, and so Assumption 5 also holds here.

3.3 Lévy payoffs, discrete prior

Examples 3.1 and 3.2 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. For simplicity, we restrict ourselves in the following to Lévy processes with a finite expected number of jumps per unit of time; the jump component of any such process is a compound Poisson process.

Let $R^n = \rho\, t + \sigma\, Z^n + Y^n$, therefore, where $(Z^0, Z^1, \ldots, Z^N)$ is again an $(N+1)$-dimensional Wiener process and $Y^0, Y^1, \ldots, Y^N$ are independent compound Poisson processes whose common Lévy measure ν has a finite second moment $\int g^2\, \nu(dg)$.[6] While $\sigma > 0$ is the same in all states of the world, the drift rate ρ and the Lévy measure ν vary with the state. We write (ρ_ℓ, ν_ℓ) for their realization in state $\ell = 0, 1, \ldots, L$, $\lambda_\ell = \nu_\ell(\mathbb{R}\backslash\{0\})$ for the expected number of jumps per unit of time, and $g_\ell = \int_{\mathbb{R}\backslash\{0\}} g\, \nu_\ell(dg) \,/\, \lambda_\ell$ for the expected jump size.[7] The expected risky payoff per unit of time in state ℓ is $\mu_\ell = \rho_\ell + \lambda_\ell\, g_\ell$. Once more, the functions m and f are the same as in Example 3.1.

With these payoffs, the generator $\mathcal{G}$ is that of a jump-diffusion, given by a combination of expressions that generalize those in Examples 3.1 and 3.2, namely

$$\begin{aligned}\mathcal{G}u(\pi) \;&=\; \frac{1}{2\sigma^2} \sum_{i=1}^{L} \sum_{\ell=1}^{L} \pi_i\, \pi_\ell\, (\rho_i - \rho(\pi))(\rho_\ell - \rho(\pi))\, \frac{\partial^2 u(\pi)}{\partial \pi_i\, \partial \pi_\ell} \\ &\quad + \int_{\mathbb{R}\backslash\{0\}} [u(j(\pi, g)) - u(\pi)]\; \nu(\pi)(dg) \;-\; \sum_{\ell=1}^{L} \pi_\ell\, (\lambda_\ell - \lambda(\pi))\, \frac{\partial u(\pi)}{\partial \pi_\ell}\,,\end{aligned}$$

where

$$\rho(\pi) = \textstyle\sum_{\ell=0}^{L}\; \pi_\ell\, \rho_\ell, \quad \nu(\pi) = \sum_{\ell=0}^{L}\; \pi_\ell\, \nu_\ell, \quad \lambda(\pi) = \sum_{\ell=0}^{L}\; \pi_\ell\, \lambda_\ell,$$

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

Once more, we can take $\mathcal{A} = \mathcal{L}(\Delta^L, [0,1])$, and the functions I and $k^\dagger$ are the same as before. In the special case where $\rho_0 = \ldots = \rho_L$ (so there is nothing to learn about the drift rate) and $\lambda_0 = \ldots = \lambda_L$ (so jumps occur at the same rate in all states of the world), the process of posterior beliefs is piecewise constant, and we can take $\mathcal{A} = \mathcal{M}(\Delta^L, [0,1])$.

The framework with Lévy payoff processes and discrete priors permits the analysis of experimentation games in which the size of a jump in cumulative payoffs is informative, and – in the special case just described – even the only source of information. Moreover,

[6] Here, $\nu(B) < \infty$ is the expected number of jumps per unit of time whose size is in the Borel set $B \subseteq \mathbb{R}\backslash\{0\}$. The finite second moment ensures that the processes R^n have finite mean and finite quadratic variation.

[7] Our assumptions on the Lévy measures ensure that the players cannot infer the true state instantaneously from the jump part of risky payoffs.

it is straightforward to model situations in which large payoff increments are *bad* news.[8]

For example, let $L = 1$ for simplicity, with $\rho_0 = \rho_1$ and $\lambda_0 = \lambda_1$. Assume that the payoff increments are in the set $\{s-10, s-5, s+5, s+10\}$. For the 'good' arm, the associated probabilities of a jump of that size are $\{0.1, 0.3, 0.5, 0.1\}$, so the expected increment is $s+1$; for the 'bad' arm, the associated probabilities of a jump of that size are $\{0.5, 0.1, 0.1, 0.3\}$, and the expected increment is $s-2$. When a payoff increment occurs, the belief jumps – up if the increment is moderate ($s-5$ and $s+5$ are relatively more likely if the arm is 'good'), and down if the increment is extreme ($s-10$ and $s+10$ are relatively more likely if the arm is 'bad'). So, in this stripped-down illustration, an arrival of the largest possible payoff increment is bad news, and may well cause the players to stop experimenting.

3.4 Brownian payoffs, normal prior

In this specification, the risky arms and background signal are as in Example 3.1 except for the assumption that μ can now take any real value. At time t, players believe that μ is distributed according to a normal distribution with mean m_t and precision $\tau_t > 0$. Given $\pi = (m, \tau) \in \mathbb{R} \times]0, \infty[$, the probability density function for μ is thus $h(\mu; \pi) = \tau^{1/2}\phi\left((\mu - m)\tau^{1/2}\right)$, where ϕ denotes the standard normal density.

Again, consider a single player allocating his entire resource to the risky arm. Following Chernoff (1968, Lemma 4.1), or Liptser and Shiryayev (1977, Theorem 10.1), τ_t increases deterministically at the rate σ^{-2} and m_t is a driftless diffusion process with diffusion coefficient $\sigma^{-1}\,\tau_t^{-1}$ relative to the player's information filtration.[9] Applying Itô's lemma and taking expectations, we see that

$$\mathrm{E}\left[u(\pi_{t+dt}) \mid \pi_t\right] = u(\pi_t) + \sigma^{-2}\left[\tfrac{1}{2}\,\tau_t^{-2}\,\frac{\partial^2 u(\pi_t)}{\partial m^2} + \frac{\partial u(\pi_t)}{\partial \tau}\right] dt$$

or

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

Since the precision τ_t increases over time, the relevant state space is the half-plane $\Pi = \mathbb{R} \times [\tau_0, \infty[$. As to admissible strategies, we take $\mathcal{A}$ to be the set of all functions $k\colon \Pi \to [0,1]$ such that $k\tau^{-1}$ is Lipschitz continuous on Π. We show that this is sufficient for Assumption 4 to be satisfied, i.e. there is a well-defined and unique law of motion for

[8] In Keller *et al.* (2005) and Keller and Rady (2010, 2015) jump sizes are completely uninformative, while in Cohen and Solan (2013) jumps are informative, but always good news.

[9] More precisely, it can be shown that $dm_t = \sigma^{-1}\,\tau_t^{-1}\,d\bar{Z}_t$ and $d\tau_t = \sigma^{-2}\,dt$ where, now, the innovation process is $d\bar{Z}_t = \sigma^{-1}\left([\mu - m_t]\,dt + \sigma\,dZ_t\right)$. Note that the expression equivalent to that for dm_t to be found in equation (9) of Jovanovic (1979) omits the term $[\mu - m_t]\,dt$.

the state.

Given a strategy profile $(k_1, \ldots, k_N) \in \mathcal{A}^N$, the corresponding intensity of experimentation $K = \sum_{n=1}^N k_n$ also lies in $\mathcal{A}$, and the system we need to solve is

$$dm = K(m,\tau)\,\tau^{-1}\sigma^{-1}d\bar{Z}, \quad d\tau = K(m,\tau)\,\sigma^{-2}dt.$$

The change of variable $\eta = \ln\tau$ transforms this into $dm = K(m,e^\eta)\,e^{-\eta}\sigma^{-1}d\bar{Z}$ and $d\eta = K(m,e^\eta)\,e^{-\eta}\sigma^{-2}dt$; as $K(m,e^\eta)\,e^{-\eta}$ is Lipschitz continuous in (m,η) on $\mathbb{R} \times [\ln\tau_0, \infty[$, this system has a unique solution, verifying Assumption 4.

In preparation for showing that Assumption 5 is satisfied by this example, we derive $m(\pi)$, $f(\pi)$, and $I(\pi)$. The expected current payoff $m(\pi)$ is simply the projection of π on its first component. For the expected full-information payoff, we have

$$f(\pi) = s\,\Phi(z) + m\,[1-\Phi(z)] + \tau^{-1/2}\phi(z),$$

where $z = (s-m)\tau^{1/2}$ and Φ denotes the standard normal cumulative distribution function. To see this, note first that we trivially obtain $H(s;\pi) = \int_{-\infty}^s h(\mu;\pi)\,d\mu = \int_{-\infty}^z \phi(x)\,dx = \Phi(z)$. Since $h(\mu;\pi) \propto \exp\left(-\frac{1}{2}(\mu-m)^2\tau\right)$, moreover, we have $dh(\mu;\pi) = -(\mu-m)\tau\,h(\mu;\pi)\,d\mu$ and so $\mu\,h(\mu;\pi)\,d\mu = m\,h(\mu;\pi)\,d\mu - \tau^{-1}\,dh(\mu;\pi)$, implying

$$\begin{aligned}\int_s^\infty \mu\,dH(\mu;\pi) &= \int_s^\infty \mu\,h(\mu;\pi)\,d\mu = \int_s^\infty m\,h(\mu;\pi)\,d\mu - \int_s^\infty \tau^{-1}\,dh(\mu;\pi)\\ &= m\,[1-H(s;\pi)] + \tau^{-1}\,h(s;\pi) = m\,[1-\Phi(z)] + \tau^{-1/2}\,\phi(z).\end{aligned}$$

The above representation makes it straightforward to verify that f is strictly increasing in m and strictly decreasing in τ.[10] By what was said above, I and $k^\dagger$ are non-decreasing in m and non-increasing in τ.

When $m < s$ we have

$$I(\pi) = \frac{s\,\Phi(z) + m\,[1-\Phi(z)] + \tau^{-1/2}\phi(z) - s}{s-m} = \Phi(z) - 1 + z^{-1}\phi(z).$$

In the appendix we verify Assumption 5 for this example by showing that $I\tau^{-1}$ is Lipschitz continuous on $\Pi(a,b) = \{\pi \in \Pi : a \le I(\pi) \le b\}$. This is more involved than in the examples with a discrete prior because the set $\Pi(a,b)$ is unbounded.

Figure 4 illustrates equilibrium actions as a function of the posterior mean m and variance τ^{-1}. (In this figure, $s = 6$ and $N = 4$.) As in Figures 1–2, the solid curves are the boundaries of the sets of beliefs at which the equilibrium requires full experimentation

[10] Alternatively, since $s \vee \mu$ is increasing in μ, a first-order stochastic dominance argument can be used to establish that $\partial f(\pi)/\partial m > 0$, and since $s \vee \mu$ is convex in μ, a second-order stochastic dominance argument can be used to establish that $\partial f(\pi)/\partial \tau < 0$.

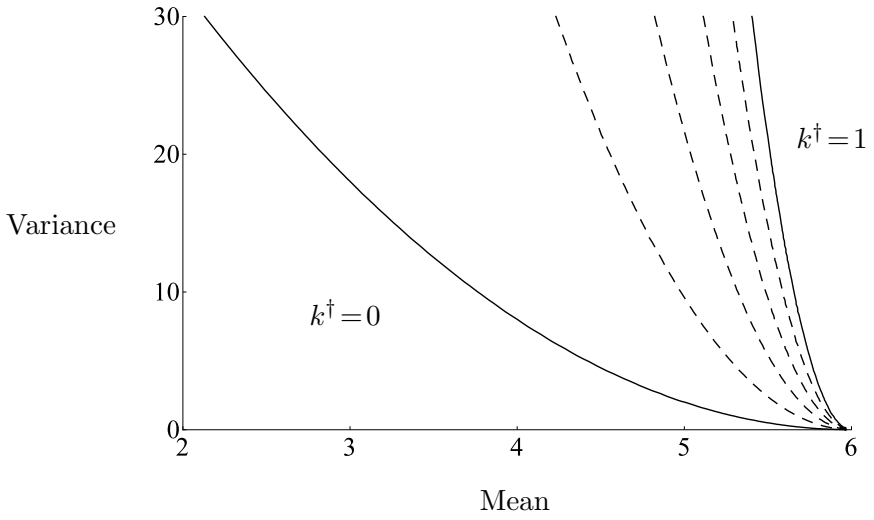

Figure 4: Equilibrium actions for Brownian payoffs and normal prior

or no experimentation, and the dashed lines are level curves for $k^\dagger$ equal to 0.2, 0.4, 0.6 and 0.8. All these curves are downward sloping; as one would expect, there is a trade-off between mean and variance with the latter capturing the "option value" of experimentation. In particular, a very high variance is needed to induce a high intensity of experimentation at low means. As the mean approaches the safe flow payoff, the level curves become steeper and steeper so that the posterior variance has a diminishing impact on the intensity with which the players explore the risky arm.

3.5 Poisson payoffs, gamma prior

The risky arms and background signal are specified as in Example 3.2 except for the assumption that μ can now take any non-negative value. Let $s > 0$ for the safe arm.

At time t, players believe that μ is distributed according to the gamma distribution $\mathrm{Ga}(\alpha_t, \beta_t)$ with parameters $\alpha_t > 0$ and $\beta_t > 0$. Given $\pi = (\alpha, \beta) \in]0, \infty[^2$, the probability density function for μ is $h(\mu; \pi) = [\beta^\alpha / \Gamma(\alpha)] \mu^{\alpha-1} e^{-\beta\mu}$, and we have $m(\pi) = \alpha/\beta$. (The corresponding variance of μ is α/β^2.)

Once more, consider a single player allocating his entire resource to the risky arm. With probability $m(\pi_t)\, dt$, he obtains a positive increment between t and $t + dt$, in which case Bayes' rule implies that π_t jumps to $(\alpha_t + 1, \beta_t)$; with probability $1 - m(\pi_t)\, dt$, there is no such increment and $d\pi_t = (d\alpha_t, d\beta_t) = (0, dt)$. Thus, α counts arrivals of increments and β measures the time that has elapsed – see, for example, DeGroot (1970, Chapter 9). We obtain

$$\mathrm{E}\left[u(\pi_{t+dt}) \mid \pi_t\right] \;=\; m(\pi_t)\, u(\alpha_t + 1, \beta_t)\, dt + \Big(1 - m(\pi_t)\, dt\Big) \left(u(\pi_t) + \frac{\partial u(\pi_t)}{\partial \beta}\, dt\right)$$

$$= \quad u(\pi_t) + \left\{ m(\pi_t) \left[u(\alpha_t + 1, \beta_t) - u(\pi_t) \right] + \frac{\partial u(\pi_t)}{\partial \beta} \right\} dt,$$

hence

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

Given that α_t and β_t increase over time, and α_t can only do so in unit increments, the relevant state space is $\Pi = \{\alpha_0 + \ell : \ell = 0, 1, 2, \ldots\} \times [\beta_0, \infty[$. For $\mathcal{A}$, we choose the set of all functions $k : \Pi \to [0, 1]$ such that $k(\alpha_0 + \ell, \cdot)$ is right-continuous and piecewise Lipschitz continuous for all ℓ, so Assumption 4 is satisfied.

Again, in preparation for showing that Assumption 5 is also satisfied, we restate $m(\pi)$, then derive $f(\pi)$ and $I(\pi)$.

$$m(\pi) = \tfrac{\alpha}{\beta}, \quad f(\pi) = s\, H(s; \alpha, \beta) + \tfrac{\alpha}{\beta} \left[1 - H(s; \alpha + 1, \beta) \right].$$

The second term in the expression for f is obtained as follows:

$$\begin{aligned} \int_s^\infty \mu \, dH(\mu; \pi) &= \int_s^\infty \mu \left[\beta^\alpha / \Gamma(\alpha) \right] \mu^{\alpha-1} e^{-\beta\mu} \, d\mu = \tfrac{\alpha}{\beta} \int_s^\infty \left[\beta^{\alpha+1} / \alpha\Gamma(\alpha) \right] \mu^\alpha e^{-\beta\mu} \, d\mu \\ &= \tfrac{\alpha}{\beta} \int_s^\infty \left[\beta^{\alpha+1} / \Gamma(\alpha + 1) \right] \mu^\alpha e^{-\beta\mu} \, d\mu = \tfrac{\alpha}{\beta} \int_s^\infty h(\mu; \alpha + 1, \beta) \, d\mu \\ &= \tfrac{\alpha}{\beta} \left[1 - H(s; \alpha + 1, \beta) \right]. \end{aligned}$$

The formula for f makes it straightforward to verify that, exactly like m, this function is strictly increasing in α and strictly decreasing in β.[11] Consequently, I and $k^\dagger$ are non-decreasing in α and non-increasing in β.

For $m(\pi) = \alpha/\beta < s$, we have

$$I(\pi) = \frac{s\, H(s; \alpha, \beta) + \frac{\alpha}{\beta} \left[1 - H(s; \alpha + 1, \beta) \right] - s}{s - \frac{\alpha}{\beta}} = \frac{s\, H(s; \alpha, \beta) - \frac{\alpha}{\beta} H(s; \alpha + 1, \beta)}{s - \frac{\alpha}{\beta}} - 1.$$

In the appendix we show that $I(\alpha, \cdot)$ has a bounded first derivative when $m(\pi) < s$ for any fixed α, i.e. on $B(a, b) = \left\{ \beta \in \left] \tfrac{\alpha}{s}, \infty \right[: a \leq I(\pi) \leq b \right\}$, thus verifying Assumption 5 for this example.

Figure 5 illustrates the mean-variance trade-off in equilibrium actions for Poisson payoffs and gamma prior. (Here, as in the example with Brownian payoffs and normal prior, $s = 6$ and $N = 4$; the curves shown are thus the exact counterparts of those in Figure 4.) To compute the level curves, one uses the fact that the shape parameter α

[11] Alternatively, 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 stated monotonicity properties.

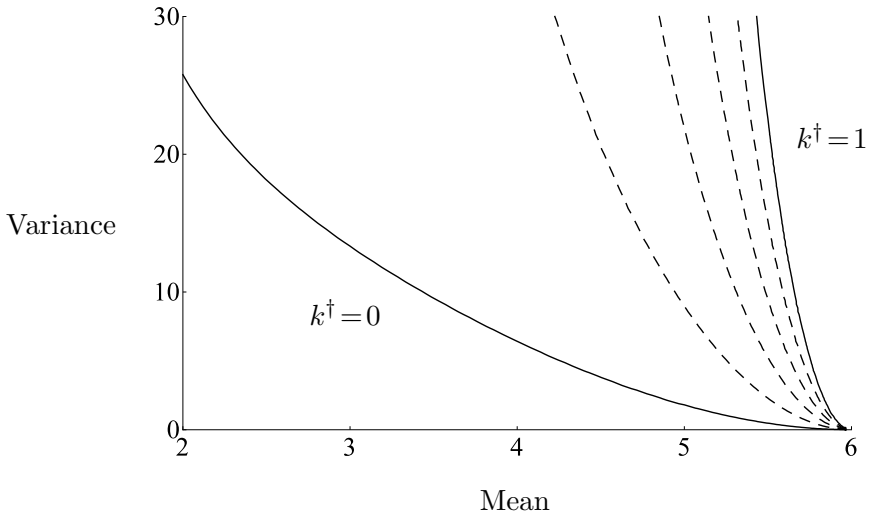

Figure 5: Equilibrium actions for Poisson payoffs and gamma prior

equals the squared mean of the gamma distribution divided by its variance, and β is α divided by the mean. The similarity to Figure 4 is striking; a closer comparison reveals that the level curves in the Brownian-normal case are somewhat steeper than those in the Poisson-gamma case. This is because in the former, an increase in the variance induces a mean-preserving spread for the random variable μ on the whole real axis, whereas in the latter, the mean-preserving spread is concentrated on the positive half-axis and thus raises the option value of experimentation by more.

4 An Example Where Separability Fails

The aim of this section is to present a specification of beliefs and payoffs that violates separability. To this end, we modify Example 3.1 by introducing state-switching: the unknown drift of the Brownian motion switches between levels μ_0 and μ_1 according to a continuous-time Markov process with transition probabilities

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

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

Given the belief π_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 induces a positive drift for π_t, the latter a negative drift, and the combined effect leads to

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

adding a term to the infinitesimal generator of π that is *not* linear in $k_0 + K_t$. In fact, we have

$$\begin{aligned}&\mathrm{E}\left[u(\pi_{t+dt}) \mid \pi_t, k_{1,t}, \ldots, k_{N,t}\right] \\ &\quad = \; u(\pi_t) + \left\{\left[(1-\pi_t)p_0 - \pi_t p_1\right] u'(\pi_t) + \tfrac{1}{2}\left(k_0 + K_t\right)\left[\pi_t(1-\pi_t)\,\Delta\mu\,\sigma^{-1}\right]^2 u''(\pi_t)\right\} dt,\end{aligned}$$

so condition (1) does not hold for this specification. Separability fails because the speed of mean reversion introduced by state switching is completely unaffected by the intensity with which the players sample their payoff generating processes.

The Bellman equation for player n against all other players using a strategy $k^\dagger$ now becomes

$$\begin{aligned}0 \;\; = \;\; \max_{k_n \in [0,1]} \Big\{ & s - \theta_n^* + k_n[m(\pi) - s] + \left[(1-\pi)p_0 - \pi p_1\right] u_n'(\pi) \\ & + \tfrac{1}{2}\left[k_0 + (N-1)k^\dagger(\pi) + k_n\right]\left[\pi(1-\pi)\,\Delta\mu\,\sigma^{-1}\right]^2 u_n''(\pi)\Big\},\end{aligned}$$

with θ_n^* denoting the highest achievable long-run average payoff. It is clearly impossible to rewrite this equation so as to separate all terms involving u_n' or u_n'' from the choice variable k_n. In other words, Markovian best responses can no longer be computed without knowledge of the value function u_n.

5 Concluding Remarks

We have seen that under the separability condition, the players' strategies in a symmetric MPE of the undiscounted experimentation game depend only on the expected current payoff from the risky arm and the expected full-information payoff. Under a discrete prior distribution for the unknown average payoff per unit of time, these two expected payoffs are fully determined – the equilibrium strategy is then invariant to the specification of the payoff-generating process.

As to the examples with a continuous prior distribution, recall that in Example 3.4 (Brownian noise, normal prior) the precision of the posterior distribution increases unboundedly with time, as does the inverse of the variance in Example 3.5 (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 3.1 or 3.2) then, over time, the beliefs become more and more concentrated on the discrete values closest to the true parameter μ – this suggests that we could take the 'engineering' approach and focus on discrete

distributions, with the specification of the payoff-generating processes being irrelevant.[12]

Of course, the evolution of the agents' posterior belief *does* depend on how the payoffs are generated, as do the players' equilibrium payoffs; and to calculate the latter, one has to solve a functional equation that involves the operator $\mathcal{G}$, which encodes the evolution of beliefs.

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

Appendix

Verification of Assumption 5 in Example 3.4

From the main body of the text, for $m < s$ we have

$$I(\pi) = \Phi(z) - 1 + z^{-1}\phi(z)$$

where $z = (s-m)\tau^{1/2}$.

The function $F(z) = \Phi(z) - 1 + z^{-1}\phi(z)$ is a strictly decreasing bijection from $]0, \infty[$ to itself with first derivative $F'(z) = -z^{-2}\phi(z)$. For any positive real number c, therefore, we have $I(\pi) = c$ if and only if $(s-m)\tau^{1/2} = F^{-1}(c)$. At any such (m, τ) in the half-plane $\Pi = I\!R \times [\tau_0, \infty[$, we have $\partial I/\partial m = -F'(F^{-1}(c))\,\tau^{1/2}$ and $\partial I/\partial \tau = \frac{1}{2}F'(F^{-1}(c))F^{-1}(c)\,\tau^{-1}$.

To verify Assumption 5, it suffices to show that $I\tau^{-1}$ is Lipschitz continuous on $\Pi(a,b) = \{\pi \in \Pi : a \le I(\pi) \le b\}$ for any positive real numbers $a < b$. For $I(\pi) = c$, we have $\partial(I\tau^{-1})/\partial m = -F'(F^{-1}(c))\,\tau^{-1/2}$ and $\partial(I\tau^{-1})/\partial \tau = \left(\frac{1}{2}F'(F^{-1}(c))F^{-1}(c) - c\right)\tau^{-2}$. This establishes that both partial derivatives of $I\tau^{-1}$ are bounded along any level curve $I(\pi) = c$ in Π. Letting c range from a to b shows that they are bounded on the whole of $\Pi(a,b)$, so $I\tau^{-1}$ is indeed Lipschitz continuous there.

Verification of Assumption 5 in Example 3.5

Again from the main body of the text, for $m(\pi) = \alpha/\beta < s$ we have

$$I(\pi) = \frac{s\,H(s;\alpha,\beta) - \frac{\alpha}{\beta}\,H(s;\alpha+1,\beta)}{s - \frac{\alpha}{\beta}} - 1.$$

We fix α as well as positive real numbers $a < b$. To verify Assumption 5, it suffices to show that $I(\alpha,\cdot)$ is Lipschitz continuous on the set $B(a,b) = \left\{\beta \in \left]\frac{\alpha}{s}, \infty\right[: a \le I(\pi) \le b\right\}$. To this end, we note first that

$$H(s;\alpha,\beta) - H(s;\alpha+1,\beta) = \int_0^s \frac{\beta^\alpha}{\Gamma(\alpha)}\,x^{\alpha-1}e^{-\beta\mu}\left[1 - \frac{\beta\mu}{\alpha}\right]\,d\mu.$$

For $\beta = \alpha/s$ and $\mu < s$, the term in square brackets under the integral is positive, so we have $H(s;\alpha,\frac{\alpha}{s}) - H(s;\alpha+1,\frac{\alpha}{s}) > 0$. For $\beta \searrow \frac{\alpha}{s}$, therefore, the numerator $s\,H(s;\alpha,\beta) - \frac{\alpha}{\beta}\,H(s;\alpha+1,\beta)$ in the above expression for $I(\pi)$ tends to a positive limit. Given that $I(\pi)$ is finite for $\beta \in B(a,b)$, this implies that the denominator in the above expression must be bounded away from 0, i.e. β must be bounded away from α/s on $B(a,b)$. Using the fact that

$$\frac{\partial H(s;\alpha,\beta)}{\partial \beta} = \frac{\alpha}{\beta}\left[H(s;\alpha,\beta) - H(s;\alpha+1,\beta)\right],$$

it is now straightforward to verify that that $I(\alpha,\cdot)$ has a bounded first derivative on $B(a,b)$.

References

Bergemann, D. and J. Välimäki (1997): "Market Diffusion with Two-sided Learning," *RAND Journal of Economics*, **28**, 773–795.

Bergemann, D. and J. Välimäki (2002): "Entry and Vertical Differentiation," *Journal of Economic Theory*, **106**, 91–125.

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

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.

Bonatti, A. (2011): "Menu Pricing and Learning," *American Economic Journal: Microeconomics*, **3**, 124–163.

Chernoff, H. (1968): "Optimal Stochastic Control," *Sankhyā*, **30**, 221–252.

Cohen, A. and E. Solan (2013): "Bandit Problems with Lévy Payoff Processes," *Mathematics of Operations Research*, **38**, 92–107.

DeGroot, M. (1970): *Optimal Statistical Decisions*. New York: McGraw Hill.

Dynkin, E.B. (1965): *Markov Processes Vol. I*. Berlin: Springer.

Engelbert, H.J. and W. Schmidt (1984): "On One-Dimensional Stochastic Differential Equations with Generalized Drift," *Lecture Notes in Control and Information Sciences*, **69**, Berlin: Springer-Verlag, 143–155.

Harris, C. (1988): "Dynamic Competition for Market Share: An Undiscounted Model," Discussion Paper No. 30, Nuffield College, Oxford.

Harris, C. (1993): "Generalized Solutions to Stochastic Differential Games in One Dimension," Industry Studies Program Discussion Paper No. 44, Boston University.

Jovanovic, B. (1979): "Job Matching and the Theory of Turnover," *Journal of Political Economy*, **87**, 972–990.

Karatzas, I. and S.E. Shreve (1988): *Brownian Motion and Stochastic Calculus*. New York: Springer-Verlag.

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. and S. Rady (2015): "Breakdowns," *Theoretical Economics*, **10**, 175–202.

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 (2010): "Competitive Experimentation with Private

Information: The Survivor's Curse," *Journal of Economic Theory*, **145**, 639–660.

Peitz, M., S. Rady and P. Trepper (2015): "Experimentation in Two-Sided Markets," CESifo Working Paper No. 5346, available at http://www.cesifo-group.de/DocDL/cesifo1_wp5346.pdf; to appear in *Journal of the European Economic Association*.

Trotter, H.F. (1959): "On the Product of Semi-Groups of Operators," *Proceedings of the American Mathematical Society*, **10**, 545–551.