Discussion Paper Series – CRC TR 224

Discussion Paper No. 130
Project B 04

Undiscounted Bandit Games

Godfrey Keller*
Sven Rady**

October 2019

*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

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Collaborative Research Center Transregio 224 - www.crctr224.de
Rheinische Friedrich-Wilhelms-Universität Bonn - Universität Mannheim

Undiscounted Bandit Games*

Godfrey Keller† Sven Rady‡

September 26, 2019

Abstract

We analyze undiscounted continuous-time games of strategic experimentation with two-armed bandits. The risky arm generates payoffs according to a Lévy process with an unknown average payoff per unit of time which nature draws from an arbitrary finite set. Observing all actions and realized payoffs, players use Markov strategies with the common posterior belief about the unknown parameter as the state variable. We show that the unique symmetric Markov perfect equilibrium can be computed in a simple closed form involving only the payoff of the safe arm, the expected current payoff of the risky arm, and the expected full-information payoff, given the current belief. In particular, the equilibrium does not depend on the precise specification of the payoff-generating processes.

Keywords: Strategic Experimentation, Two-Armed Bandit, Strong Long-Run Average Criterion, Markov Perfect Equilibrium, HJB Equation, Viscosity Solution.

JEL Classification Numbers: C73, D83.

*We thank three anonymous referees and an advisory editor for their comments on an earlier version of this work (circulated in 2015 as SFB TR 15 Discussion Paper No. 520). Our thanks for helpful discussions and suggestions are owed to Chris Harris, Alois Kneip, Albert N. Shiryaev, Philipp Strack, Bruno Strulovici, and seminar participants in Austin, Budapest, Cambridge, Florence, Keele, Oxford, Paris, 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 GRK 801, SFB TR 15 and SFB TR 224 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.

1 Introduction

We analyze a class of continuous-time two-armed bandit models in which a number of players act non-cooperatively, trying to learn an unknown state of the world that governs the risky arm's expected payoff per unit of time. Actual payoffs are given by Lévy processes, that is, processes with independent and stationary increments. In addition, players receive free background information in the form of a process of the same type as the payoff processes. Rather than discounting future payoffs, players evaluate their payoff streams according to the strong long-run average criterion.[1] 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, and we look for Markov perfect equilibria.

This setting 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. First, the unknown state of the world can be drawn from an arbitrary finite set, whereas the literature assumes a binary state. Second, the payoff processes can combine continuous with discrete increments, whereas the literature assumes either Brownian or Poisson payoffs. Third, lump-sum payoffs can be good or bad news, whereas the literature assumes that news is of one type only.

The 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 the Brownian model of Bolton and Harris (1999, 2000) and the Poisson models of Keller, Rady and Cripps (2005) and Keller and Rady (2010, 2015), beliefs evolve on the unit interval, which 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. For these settings, the results of Bolton and Harris (2000) yield a characterization of the entire set of undiscounted Markov perfect equilibria. In general, however, one 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. As already noted in Bolton and Harris (2000), the absence of discounting and the presence of background

[1] This criterion is the limit of the standard discounted performance criterion as the discount rate goes to zero, both in terms of value functions and optimal strategies. See Dutta (1991) for the connection between performance criteria with and without discounting in discrete time, and Bolton and Harris (2000) for a detailed treatment of the strong long-run average criterion in a continuous-time Bayesian-learning setting such as ours.

information imply that a player's best response can be computed without knowledge of the 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 payoff of the safe arm, 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 feature carries over to the symmetric Markov perfect equilibrium, where one and the same functional form – the natural generalization of that in Bolton and Harris (2000) – applies across all specifications that we consider. The common equilibrium action is a piecewise linear function of the ratio of two differences: that between the risky arm's expected full-information payoff and the safe payoff, and that between the safe payoff and the risky arm's expected current payoff.

We show that this result extends to two specifications of priors and payoff-generating processes in which the unknown state of the world is drawn from a continuous distribution of unbounded support: Brownian payoffs with normal priors as in Jovanovic (1979), and Poisson payoffs with gamma priors as in Moscarini and Squintani (2010). In either specification, the players' information is captured by a two-dimensional sufficient statistic, which can serve as the state variable for Markov strategies.

Our result hinges on four features of the setting that we study: (i) players receive free background information; (ii) they use the strong long-run average criterion; (iii) the experimentation game is played in continuous time; and (iv) the players' risky payoff processes and the background information are all of the same (unknown) type, hence perfect substitutes with respect to learning. The background information ensures that players learn the true state eventually, no matter what strategy profile they use. This makes it possible to evaluate players' random payoff streams according to the strong long-run average criterion, that is, by computing the expected accumulated shortfall of received payoffs relative to the expected full-information payoff. Under this criterion, the problem of finding a best response to the opponents' Markovian strategy profile has a recursive structure amenable to dynamic-programming techniques. In continuous time, this leads to an HJB equation in which the value function enters only through the expected rate of change of continuation payoffs. When the players' risky payoff processes and the background information are all of the same type, moreover, the rate of change of expected continuation values is linear in the total intensity of experimentation. This makes it possible to eliminate a player's value function completely from the maximization problem in the HJB equation, so best responses can be determined without reference to the value function and the payoff-generating processes.

Each of these four features is indeed crucial. Without background information, the strong long-run average criterion would be ill-defined because the expected accumulated shortfall of received payoffs relative to the expected full-information payoff would always grow infinitely large. With discounting, the HJB equation would necessarily contain a term 'discount rate times current value' that is not multiplied by the total intensity of experimentation, so best responses would depend on current values. As pointed out in Dutta (1991), moreover, alternative undiscounted performance criteria would not permit a recursive representation. If the model were set in discrete time, the expected rate of change of continuation payoffs would not be linear in the total intensity of experimentation. In a discrete-time version of an exponential bandit game à la Keller, Rady and Cripps (2005), for example, the probability of a success in any given round is clearly non-linear in the number of players pulling the risky arm. Linearity would also fail if the type of the risky arm were independent or imperfectly correlated across players, or if the law of the payoff process differed across players.[2]

While the computation of best responses does not involve the specifics of the payoff-generating processes, the evolution of the players' posterior beliefs obviously does depend on how the payoffs are generated, as do the players' equilibrium payoffs. To characterize the latter, and to verify that a certain profile of Markov strategies constitutes an equilibrium, one has to solve a functional equation that involves the infinitesimal generator of the belief process. Our approach here is to show that the player's value function is the unique viscosity solution of the HJB equation subject to the relevant boundary conditions, and that the payoff function for the suggested strategy profile also solves this boundary-value problem, so the two must agree and the player indeed plays a best response.[3]

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 (2017).

The rest of the paper is organized as follows. Section 2 sets up the game and states our assumptions on priors, payoff-generating processes and strategy spaces. Section 3 presents the infinitesimal generator of the process of posterior beliefs. Section 4 constructs the unique symmetric Markov perfect equilibrium. Section 5 presents extensions of our analysis to two settings with a continuously distributed state of world. Section 6 offers some concluding remarks.

[2]Linearity would also fail in a restless bandit model in which the state of the world changed exogenously over time. This would be the case, for example, if payoffs were generated by a Brownian motion with an unknown drift subject to Markovian state-switching between a high and a low level as in Keller and Rady (1999, 2003).

[3]For a recent application of this approach to the verification of optimality in a single-agent learning context with Brownian signals, see Ke and Villas-Boas (2019).

2 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. One arm is safe, the other risky.

The safe arm generates a known constant payoff $s > 0$ per unit of time. The evolution of the payoffs generated by the risky arm depends on a state of the world, ℓ, which nature draws from the set $\{0, 1, \ldots, L\}$ with $L \geq 1$ according to the positive probabilities $\pi_0, \ldots, \pi_L$. Players do not observe the state, but know its distribution. They also know that the payoff process associated with player n's risky arm is of the form

$$X_t^n = \rho\, t + \sigma Z_t^n + Y_t^n,$$

where Z^n is a standard Wiener process and Y^n is a compound Poisson process whose Lévy measure ν is finite and has a finite second moment $\int h^2\, \nu(dh)$.[4] The drift rate ρ, the diffusion coefficient $\sigma > 0$ and the Lévy measure ν are the same for all players. While σ is the same in all states of the world, moreover, ρ and ν vary with the state.[5] Conditionally on ℓ, the processes $Z^1, \ldots, Z^N, Y^1, \ldots, Y^N$ are independent.

We write ρ_ℓ and ν_ℓ for the drift rate and Lévy measure in state ℓ, $\lambda_\ell = \nu_\ell(\mathbb{R}\backslash\{0\})$ for the expected number of jumps per unit of time, and $h_\ell = \int_{\mathbb{R}\backslash\{0\}} h\, \nu_\ell(dh) \,/\, \lambda_\ell$ for the expected jump size. The state-contingent expected risky payoff per unit of time is $\mu_\ell = \rho_\ell + \lambda_\ell\, h_\ell$. We assume that $\mu_0 < \mu_1 < \ldots < \mu_{L-1} < \mu_L$ with $\mu_0 < s < \mu_L$, so that neither arm dominates the other in terms of expected payoffs. Writing π for the vector of probabilities $(\pi_1, \ldots, \pi_L)$, we let $m(\pi)$ denote the expected current (or myopic) payoff from the risky arm, and $f(\pi)$ a player's expected full-information payoff:[6]

$$m(\pi) = \sum_0^L \pi_\ell \mu_\ell, \qquad f(\pi) = \sum_0^L \pi_\ell \max\{s, \mu_\ell\}.$$

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

[4] 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 X^n have finite mean and finite quadratic variation.

[5] Our assumptions on the diffusion coefficient and the Lévy measures ensure that the players cannot infer the true state instantaneously from the continuous and jump part of risky payoffs, respectively.

[6] Given our convention to treat $\pi_1, \ldots, \pi_L$ as the independent variables, π_0 should be viewed as shorthand for $1 - \sum_{\ell=1}^L \pi_\ell$ from now on.

time T is then given by the time-changed process $[T - \tau^n(T)]\, s + X^n_{\tau^n(T)}$ where $\tau^n(T) = \int_0^T k_{n,t}\, dt$ measures the *operational time* that the risky arm has been used. As $X^n_t - \mu t$ is a martingale, the player's expected payoff up to T is

$$\mathbb{E}\left[\int_0^T \{(1 - k_{n,t})s + k_{n,t}\mu\}\, dt\right];$$

here, the expectation is both about the process of allocations $k_{n,t}$ and the unknown expected per-period payoff μ, a random variable with possible values $\mu_0, \ldots, \mu_L$. With s lying in the interior of the range of possible realizations of μ, each player has an incentive to learn the quality of the risky arm.

Players do not discount future payoffs, and are instead assumed to use the strong long-run average criterion.[7] This means that player n chooses allocations $k_{n,t}$ so as to maximize

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

Here, the integrand is the difference between what a player expects to receive at a given point in time and what she would expect to receive were she to be fully informed. Note that this objective function depends on others' actions only through their impact on the player's own choices. In fact, we will soon impose restrictions under which others' actions matter only through their effect on a player's beliefs.

The players start with a common prior belief π_0 about the unknown state ℓ, given by the probabilities with which nature draws this state, namely π. Thereafter, all observe each other's actions and outcomes as well as a common background signal, so they hold common posterior beliefs throughout time. The background signal is generated by the time-changed process $X^0_{\tau^0(t)}$ where X^0 is an independent process of the same law as each player's payoff process from the risky arm, and $\tau^0(t) = k_0 t$ with $k_0 > 0$ exogenously given and arbitrarily small. This signal ensures that the players eventually learn the value of μ even if they all play safe all the time.

Let π_t denote the players' common Bayesian posterior belief about the state given their observations up to time t. With respect to the information filtration generated by these observations, the process of beliefs π_t is a Markov process (in fact, a jump diffusion) and a martingale. The linearity of the functions m and f now implies that $\mathbb{E}[m(\pi_t)] = m(\pi)$ and $\mathbb{E}[f(\pi_t)] = f(\pi)$ for all $t \geq 0$, so we can rewrite the above objective function as

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

[7] For a discussion of this criterion and the role of the background signal introduced in the next paragraph, see Bolton and Harris (2000).

highlighting the potential for the posterior belief to serve as a state variable.

From now on, we restrict players to strategies that are Markovian with respect to this variable, so that the action $k_{n,t}$ chosen at time t is a deterministic function of the posterior π_t only. More precisely, we take the players' common strategy space $\mathcal{K}$ to be the set of all Lipschitz continuous functions from the L-dimensional simplex

$$\Delta_L = \left\{ \pi \in \mathbb{R}^L_+ \colon \sum_{\ell=1}^{L} \pi_\ell \leq 1 \right\}$$

to $[0,1]$. By standard existence and uniqueness results for solutions of stochastic differential equations, any strategy profile $(\kappa_1, \ldots, \kappa_N) \in \mathcal{K}^N$ gives rise to a well-defined process of posterior beliefs,[8] and hence to well-defined payoffs

$$u_n(\pi|\kappa_1, \ldots, \kappa_N) = \mathbb{E}\left[\int_0^\infty \left\{ [1-\kappa_n(\pi_t)]s + \kappa_n(\pi_t)m(\pi_t) - f(\pi_t) \right\} dt \,\middle|\, \pi_0 = \pi \right] \in [-\infty, 0].$$

A player's payoff will indeed be $-\infty$ for certain Markov strategies. If the player always uses the safe arm, for example, and the true state ℓ is such that $\mu_\ell > s$, then by almost sure convergence of posterior beliefs to the truth, the above integrand will converge to $s - \mu_\ell < 0$ as t grows large, implying a diverging integral in that state. Since this occurs with positive prior probability, the expected payoff is $-\infty$, therefore.

We call a strategy in $\kappa_n \in \mathcal{K}$ *reasonable* if each degenerate belief has a neighbourhood in which the strategy prescribes the action that is optimal in the respective state; in particular, $[1-\kappa_n(\pi)]s + \kappa_n(\pi)m(\pi) = \max\{s, m(\pi)\}$ in all these neighbourhoods, and $[1-\kappa_n(\pi)]s + \kappa_n(\pi)m(\pi) - f(\pi) = 0$ in all vertices of the simplex Δ_L. Establishing that posterior beliefs converge exponentially fast to the truth, we show in the appendix that the expected payoff from a reasonable strategy is always finite and, in fact, bounded on the simplex.

Strategy $\kappa_n \in \mathcal{K}$ is a *best response* against $\kappa_{\neg n} = (\kappa_1, \ldots, \kappa_{n-1}, \kappa_{n+1}, \ldots, \kappa_N) \in \mathcal{K}^{N-1}$ if $u_n(\pi|\kappa_n, \kappa_{\neg n}) \geq u_n(\pi|\tilde{\kappa}_n, \kappa_{\neg n})$ for all $\pi \in \Delta^L$ and all $\tilde{\kappa}_n \in \mathcal{K}$. A *Markov perfect equilibrium (MPE)* is a profile of strategies $(\kappa_1, \ldots, \kappa_N) \in \mathcal{K}^N$ that are mutually best responses. Such an equilibrium is *symmetric* if $\kappa_1 = \kappa_2 = \ldots = \kappa_N$. Obviously, each

[8] For $L = 1$ and no discontinuous payoff component, i.e. in the setting 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 Borel measurable Markov strategies implies a unique solution for the belief dynamics; see also Section 5.5 of Karatzas and Shreve (1988). For $L = 1$, no Brownian payoff component, and lump-sum payoffs that are always good news (meaning that $\nu_0(B) \leq \nu_1(B)$ for all Borel sets $B \subseteq \mathbb{R}\backslash\{0\}$), one can proceed as in Keller, Rady and Cripps (2005) and Keller and Rady (2010) and take $\mathcal{K}$ to be the set of functions which are left-continuous and piecewise Lipschitz continuous; as beliefs drift down deterministically in between lump-sums, these properties allow one to construct belief dynamics in a pathwise fashion. Neither approach generalizes to higher dimensions.

player must obtain a finite payoff in any MPE.

3 The Infinitesimal Generator

The evolution of posterior beliefs is driven by up to $N+1$ distinct sources of information: the observations on up to N risky arms plus the background signal. Suppose that only player 1 uses the risky arm, and at full intensity. In other words, consider the time-invariant action profile for which $k_1 = 1$ whereas $k_n = 0$ for all $n > 1$. Write $\mathcal{G}$ for the infinitesimal generator of the corresponding belief process – as the payoff-generating process is the same on every player's risky arm, the identity of the player in question does indeed not matter here.

If we now change player 1's time-invariant intensity to $k_1 < 1$ while keeping all other intensities at zero, the resulting deceleration of the process of observations implies the scaled-down generator $k_1\mathcal{G}$ for the posterior belief; see Dynkin (1965), for example. The same applies to the background signal, of course, if it alone is observed, with associated generator $k_0\mathcal{G}$.

As the processes X^0 and X^1 are independent conditionally on the realized state, Trotter (1959) implies that the infinitesimal generator of posterior beliefs is $(k_0 + k_1)\mathcal{G}$ when both the background signal and player 1's payoffs are observed. By the same token, successively adding the other players with time-invariant allocations $k_2, \ldots, k_N$ leads to the infinitesimal generator $(k_0 + K)\mathcal{G}$ where $K = \sum_{n=1}^N k_n$ measures how much of the N available units of the resource is allocated to risky arms overall. This fact will play a crucial role in our analysis.

The generator $\mathcal{G}$ is that of a jump diffusion. In the interior $\mathring{\Delta}_L$ of the simplex, its action on a C^2 function u is given by

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

where

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

and

$$j_\ell(\pi,h) = \frac{\pi_\ell\,\nu_\ell(dh)}{\nu(\pi)(dh)}$$

is the revised probability of state ℓ after a lump-sum payoff of size h arrives. The first term captures the learning from the continuous part of the payoff-generating process; the second term, the discrete belief revision upon the arrival of a lump-sum payoff; and the third term, the gradual belief revision when no such lump-sum arrives.

For $L = 1$, and hence $\pi = \pi_1$, we obtain the generator computed by Cohen and Solan (2013), with the first term simplifying to

$$\frac{1}{2\sigma^2}(\rho_1 - \rho_0)^2 \pi^2 (1-\pi)^2\, u''(\pi),$$

the expression familiar from Bolton and Harris (1999, 2000). It reflects the fact, established in Liptser and Shiryayev (1977, Theorem 9.1), that when there is no discontinuous payoff component ($\lambda_0 = \lambda_1 = 0$), then the posterior belief π_t of a single agent who allocates his entire resource to the risky arm follows a diffusion process with zero drift and diffusion coefficient $(\rho_1 - \rho_0)\,\sigma^{-1}\pi_t(1-\pi_t)$ relative to the agent's information filtration.[9] For $L > 1$, a generalization of Liptser and Shiryayev (1977, Theorem 9.1) shows that, from the agent's perspective, the corresponding belief process π_t is a driftless L-dimensional diffusion with instantaneous variance-covariance matrix given by

$$\mathrm{Cov}\left[d\pi_{i,t}, d\pi_{\ell,t} \mid \pi_t\right] = \left[\pi_{i,t}\left(\rho_i - \rho(\pi_t)\right)\sigma^{-1}\right]\left[\pi_{\ell,t}\left(\rho_\ell - \rho(\pi_t)\right)\sigma^{-1}\right] dt,$$

hence the structure of the first term in $\mathcal{G}u$.[10]

The second and third terms generalize their counterparts in Cohen and Solan (2013) to $L > 1$ in the obvious way. In the special case that $L = 1$ and the size of lump-sum payoffs is uninformative (meaning that conditional on the arrival of a lump-sum, the distribution of its size does not depend on ℓ), these terms reduce to

$$\lambda(\pi)\left[u\left(\frac{\pi\lambda_1}{(1-\pi)\lambda_0 + \pi\lambda_1}\right) - u(\pi)\right] - (\lambda_1 - \lambda_0)\pi(1-\pi)u'(\pi),$$

as in Keller, Rady and Cripps (2005) and Keller and Rady (2010).

Note that we have not imposed any mutual absolute continuity assumptions on the measures $\nu_0, \ldots, \nu_L$. As a consequence, lump-sum payoffs of a certain size may rule out certain states, so that the posterior belief jumps to a subsimplex of Δ_L of dimension lower than L. Once this happens, Bayesian updating ensures that beliefs remain in this subsimplex.

[9] More precisely, the belief evolves according to $d\pi_t = \sigma^{-1}\,\pi_t[\rho_1 - \rho(\pi_t)]\,d\bar{Z}_t$ where the *innovation* process $\bar{Z}_t$, given by $d\bar{Z}_t = \sigma^{-1}\left([\rho - \rho(\pi_t)]\,dt + \sigma\,dZ_t\right)$, is a Wiener process relative to the agent's information filtration.

[10] This generalization already appears in Veronesi (2000), for example.

4 Symmetric Markov Perfect Equilibrium

Suppose that all players except player n use the strategy $\kappa^\dagger \in \mathcal{K}$, and write $(\kappa_n, \kappa^\dagger_{\neg n})$ for the strategy profile that results when player n uses the strategy $\kappa_n \in \mathcal{K}$.

When choosing κ_n, player n faces a problem of optimal stochastic control of a jump diffusion, and κ_n is a best response if and only if the payoff function $u_n(\cdot|\kappa_n, \kappa^\dagger_{\neg n})$ is the value function for that control problem. This means in particular that a necessary condition for κ_n to be a best response is that the payoff function $u_n(\cdot|\kappa_n, \kappa^\dagger_{\neg n})$ be a viscosity solution of the HJB equation

$$0 = \max_{k\in[0,1]} \left\{(1-k)s + km(\pi) - f(\pi) + [k_0 + (N-1)\kappa^\dagger(\pi) + k]\mathcal{G}u(\pi)\right\} \tag{1}$$

in the interior $\overset{\circ}{\Delta}_L$ of the L-dimensional simplex; see Øksendal and Sulem (2007) or Pham (2009), for example.[11] Conversely, the following conditions are sufficient for κ_n to be a best response: (i) $u_n(\cdot|\kappa_n, \kappa^\dagger_{\neg n})$ is a viscosity solution of the HJB equation in $\overset{\circ}{\Delta}_L$ and satisfies the appropriate boundary condition on $\partial\Delta_L$; (ii) there exists only one such solution.

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

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

which demonstrates that the set of maximizers does not depend on continuation values. Straightforward algebra allows us to further simplify the problem by rewriting the HJB equation so that k appears only in the denominator:

$$0 = \max_{k\in[0,1]} \frac{[k_0 + (N-1)\kappa^\dagger(\pi)][s - m(\pi)] - [f(\pi) - s]}{k_0 + (N-1)\kappa^\dagger(\pi) + k} - [s - m(\pi)] + \mathcal{G}u(\pi). \tag{2}$$

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. When $I(\pi) < k_0 + (N-1)\kappa^\dagger(\pi)$, the numerator in (2) is positive and the maximum is achieved by $k = 0$; when $I(\pi) > k_0 + (N-1)\kappa^\dagger(\pi)$, the numerator is negative and the maximum is achieved by $k = 1$; when $I(\pi) = k_0 + (N-1)\kappa^\dagger(\pi)$, the numerator is zero and the choice of k is inconsequential.

There are three different ways, therefore, in which $k = \kappa^\dagger(\pi)$ can achieve the maximum

[11] A definition of viscosity solutions is given in the appendix.

in the HJB equation: either $\kappa^\dagger(\pi) = 0$ and $I(\pi) \leq k_0$, or $\kappa^\dagger(\pi) = 1$ and $I(\pi) \geq k_0 + N - 1$, or $0 < \kappa^\dagger(\pi) < 1$ and $I(\pi) = k_0 + (N-1)\kappa^\dagger(\pi)$. This pins down $\kappa^\dagger(\pi)$ in terms of the incentive to experiment, $I(\pi)$, the strength of the background signal, k_0, and the number of players, N:

$$\kappa^\dagger(\pi) = \begin{cases} 0 & \text{if } I(\pi) \leq k_0, \\ \frac{I(\pi)-k_0}{N-1} & \text{if } k_0 < I(\pi) < k_0 + N - 1, \\ 1 & \text{if } I(\pi) \geq k_0 + N - 1. \end{cases} \tag{3}$$

As the partial derivatives of the incentive to experiment I are clearly bounded on the compact set $\{\pi \in \Delta_L : k_0 \leq I(\pi) \leq k_0 + N - 1\}$, the function $\kappa^\dagger$ is Lipschitz continuous and hence an element of $\mathcal{K}$. Like the functions m and f, moreover, I and $\kappa^\dagger$ are non-decreasing in π. Finally, it is straightforward to check that $\kappa^\dagger$ prescribes the full-information optimal action in a neighbourhood of each vertex of Δ_L, so the strategy is reasonable.

Proposition. *All players using the strategy $\kappa^\dagger$ constitutes the unique symmetric Markov perfect equilibrium of the experimentation game.*

PROOF: That this strategy profile constitutes an equilibrium is shown in the appendix. Uniqueness follows from the arguments that led us from the HJB equation (1) to the representation (3) for candidate equilibrium actions. ■

Figures 1 and 2 illustrate the case $L = 2$. (In both figures, $\mu_0 = 2$, $\mu_1 = 5$, $\mu_2 = 8$, $N = 4$ and $k_0 = 0.2$; $s = 6$ in Figure 1, and $s = 4$ in Figure 2.) The solid lines are the boundaries of the sets of beliefs at which the equilibrium requires full experimentation ($\kappa^\dagger = 1$) and no experimentation ($\kappa^\dagger = 0$), respectively. The dotted lines are level curves of $\kappa^\dagger$ 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 arm gives the players an increased incentive to experiment.

Note that by equation (3), the set of beliefs for which $\kappa^\dagger(\pi) = 0$ 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 consequence of the linearity of the infinitesimal generator of posterior beliefs in $k_0 + K$: 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.

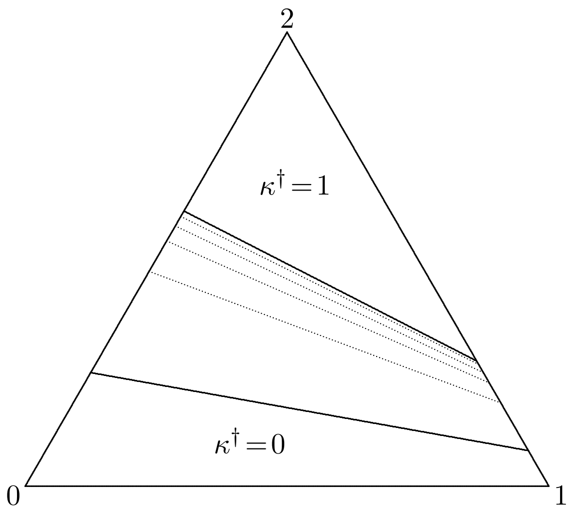

Figure 1: Equilibrium actions for $L = 2$ and $\mu_0 < \mu_1 < s < \mu_2$

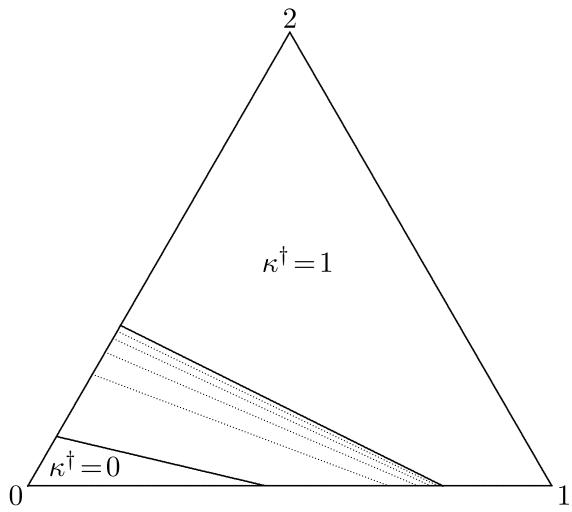

Figure 2: Equilibrium actions for $L = 2$ and $\mu_0 < s < \mu_1 < \mu_2$

Free-riding can also be seen in the fact that $\kappa^\dagger$ is non-increasing in N, and decreasing where it assumes interior values. Figure 3 illustrates this effect. On the horizontal axis

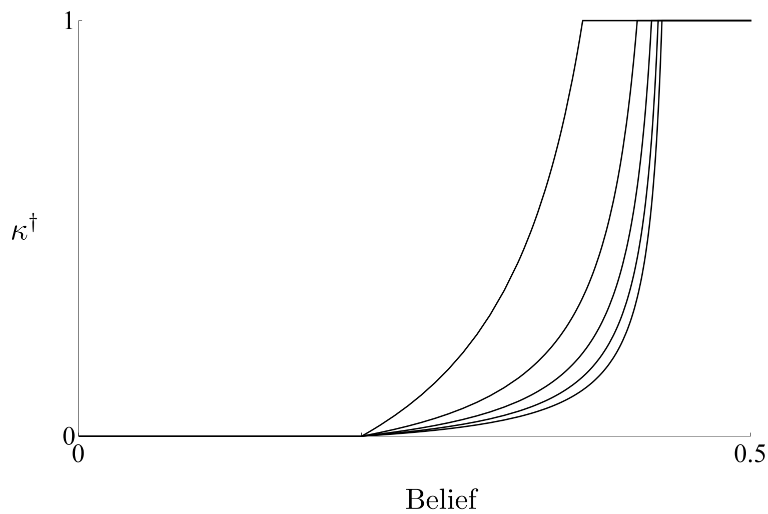

Figure 3: Equilibrium actions for $L = 2$, $\pi_1 = \pi_2$ and $N \in \{2, 4, 6, 8, 10\}$

we set $\pi_1 = \pi_2$ and let that common belief range from 0 to 0.5: so it is a slice through the simplex from the 0-vertex to the midpoint of the opposite edge. (In this figure, the parameters are as in Figure 1 except that N varies from 2 for the leftmost curve to 10 for the rightmost curve.)

The dependence of the overall intensity of experimentation on the number of players is less clear cut: roughly speaking, $N\kappa^\dagger$ increases in N at beliefs where $\kappa^\dagger$ requires exclusive use of the risky arm, but decreases at beliefs where both arms are used simultaneously.

As to the dynamics of beliefs in equilibrium, the present framework permits the analysis of experimentation games in which large payoff increments are *bad* news.[12] 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 lump-sum 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 lump-sum 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.

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 the true μ, and so $\kappa^\dagger(\pi)$ converges to either 0 or 1. As was already said in Section 2, we show in the appendix that the convergence of beliefs is exponentially fast in expectation; this immediately implies that equilibrium actions converge exponentially fast as well.

5 Continuous State Spaces and Sufficient Statistics

This section presents two specifications of priors and payoff-generating processes that fall outside the framework of Section 2 but still permit the same analysis as in Sections 3 and 4. In both settings, the unknown state of the world is drawn from a continuous distribution of unbounded support, with conjugate priors ensuring that the players' information is captured by a two-dimensional sufficient statistic, which can serve as the state variable for Markov strategies.[13] Models in which agents have beliefs and observe stochastic processes like those in Sections 5.1 and 5.2 can be found in Jovanovic (1979) and Moscarini and Squintani (2010), respectively.

[12] In Keller, Rady and Cripps (2005) and Keller and Rady (2010, 2015) lump-sum sizes are completely uninformative, while in Cohen and Solan (2013) lump-sums are informative, but always good news.

[13] The unbounded state space requires adjustments to the proof (via uniqueness of viscosity solutions to the HJB equation) that every player using the strategy $\kappa^\dagger$ constitutes an MPE of the game; we omit the details here.

5.1 Brownian Payoffs, Normal Prior

Suppose that the payoff-generating processes and the background signal are of the form

$$X_t^n = \mu\, t + \sigma Z_t^n,$$

where the Z^n are independent standard Wiener processes and nature draws the unknown drift μ from a normal distribution with mean m_0 and precision $\tau_0 > 0$. This is also the players' common prior. Given the Gaussian process they observe, players then believe at time t that μ is distributed according to a normal distribution with some mean m_t and precision $\tau_t > 0$; see DeGroot (1970, Chapter 9), for example. The pair $\pi_t = (m_t, \tau_t)$ constitutes a sufficient statistic for the updating of beliefs, therefore. Given a generic $\pi = (m, \tau) \in \mathbb{R} \times]0, \infty[$, the corresponding probability density function for μ is $g(\mu; \pi) = \tau^{1/2}\phi\left((\mu - m)\tau^{1/2}\right)$, where ϕ denotes the standard normal density. Let $G(\cdot; \pi)$ denote the associated cumulative distribution function.

As in Section 3, 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.[14] As a result, we see that

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

for any function of class $C^{2,1}$. By the same arguments as in Section 3, moreover, the generator associated with time-invariant intensities $(k_0, k_1, \ldots, k_N) \in [0,1]^{N+1}$ is again $(k_0 + K)\mathcal{G}$.

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 $\mathcal{K}$ to be the set of all functions $\kappa : \Pi \to [0,1]$ such that $\kappa\tau^{-1}$ is Lipschitz continuous on Π. Given a strategy profile $(\kappa_1, \ldots, \kappa_N) \in \mathcal{K}^N$, the sum $K = \sum_{n=1}^N \kappa_n$ also lies in $\mathcal{K}$, 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, as was to be shown.

We can now replicate the arguments of Section 4 in the present setting. As a first step, we compute the expected current payoff $m(\pi)$, the expected full-information payoff

[14] 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 $[\mu - m_t]\, dt$.

$f(\pi)$, and the incentive to experiment $I(\pi)$. 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 $f(\pi) = sG(s;\pi) + \int_s^\infty \mu\, g(\mu;\pi)\, d\mu$. We trivially obtain $G(s;\pi) = \int_{-\infty}^s g(\mu;\pi)\, d\mu = \int_{-\infty}^z \phi(x)\, dx = \Phi(z)$. Since $g(\mu;\pi) \propto \exp\left(-\frac{1}{2}(\mu - m)^2\tau\right)$, moreover, we have $dg(\mu;\pi) = -(\mu - m)\tau\, g(\mu;\pi)\, d\mu$ and so $\mu\, g(\mu;\pi)\, d\mu = m\, g(\mu;\pi)\, d\mu - \tau^{-1}\, dg(\mu;\pi)$, implying

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

The above representation makes it straightforward to verify that f is strictly increasing in m and strictly decreasing in τ.[15] This implies that I and $\kappa^\dagger$ as defined in (3) 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 that $\kappa^\dagger \in \mathcal{K}$ by showing that $I\tau^{-1}$ is Lipschitz continuous on the set $\{\pi \in \Pi : k_0 \leq I(\pi) \leq k_0 + N - 1\}$. This is more involved than in scenarios with a discrete prior because the set in question 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 or no experimentation, and the dashed lines are level curves for $\kappa^\dagger$ 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.

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

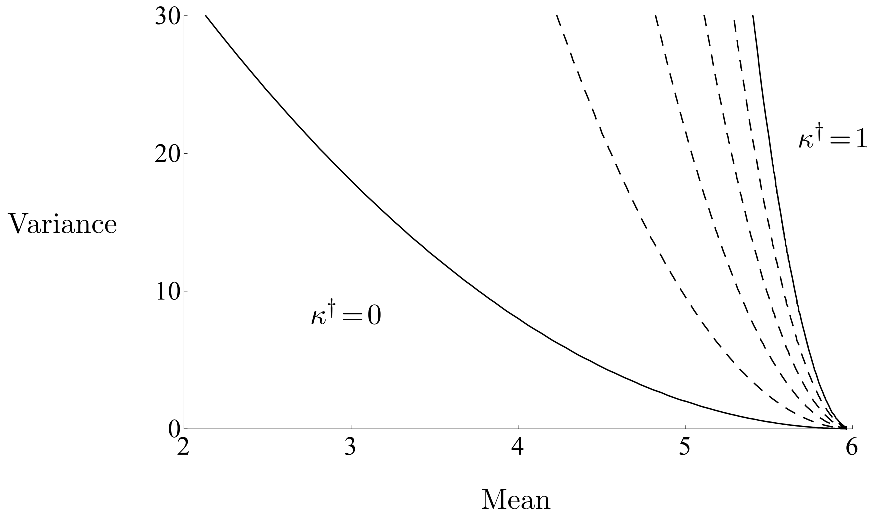

Figure 4: Equilibrium actions for Brownian payoffs and normal prior

5.2 Poisson Payoffs, Gamma Prior

Let $s > 0$ for the safe arm. Suppose that the payoff-generating processes and the background signal are independent Poisson processes whose unknown common intensity μ is drawn from a gamma distribution with parameters $\alpha_0 > 0$ and $\beta_0 > 0$. This is also the players' common prior. Given the processes they observe, players then believe at time t that μ is distributed according to a gamma distribution with some parameters $\alpha_t > 0$ and $\beta_t > 0$, which together constitute a sufficient statistic again; see DeGroot (1970, Chapter 9), for example. Given a generic $\pi = (\alpha, \beta) \in \,]0, \infty[^2$, the probability density function for μ is $g(\mu; \pi) = [\beta^\alpha / \Gamma(\alpha)] \mu^{\alpha-1} e^{-\beta\mu}$; the mean and variance of μ are α/β and α/β^2, respectively. We again write $G(\cdot; \pi)$ for the corresponding cumulative distribution function.

Once more, consider a single player allocating his entire resource to the risky arm. He expects to obtain a positive increment between t and $t + dt$ with probability $(\alpha_t/\beta_t)\, dt$, in which case Bayes' rule implies that π_t jumps to $(\alpha_t + 1, \beta_t)$; with probability $1 - (\alpha_t/\beta_t)\, dt$, 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 – again, see DeGroot (1970, Chapter 9). As a consequence, we have

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

Once more, the generator associated with time-invariant intensities $(k_0, k_1, \ldots, k_N) \in [0, 1]^{N+1}$ is $(k_0 + K)\mathcal{G}$.

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 $\Pi = \{\alpha_0 + j : j = 0, 1, 2, \ldots\} \times [\beta_0, \infty[$. For $\mathcal{K}$, we choose the set of all functions $\kappa : \Pi \to [0,1]$ such that $\kappa(\alpha_0+j,\cdot)$ is right-continuous and piecewise Lipschitz continuous for all j. Starting from any $\pi \in \Pi$, any strategy profile $(\kappa_1,\ldots,\kappa_N) \in \mathcal{K}^N$ induces a well-defined and unique law of motion for π_t,

As the unknown intensity μ is also the risky arm's average payoff per unit of time, we see that the expected current payoff is $m(\pi) = \alpha/\beta$. The expected full-information payoff is

$$f(\pi) = s\,G(s;\pi) + \frac{\alpha}{\beta}\,[1 - G(s;\alpha+1,\beta)],$$

with the second term obtained as follows:

$$\begin{aligned}
\int_s^\infty \mu\, g(\mu;\pi)\,d\mu &= \int_s^\infty \mu\,\frac{\beta^\alpha}{\Gamma(\alpha)}\,\mu^{\alpha-1}e^{-\beta\mu}\,d\mu = \frac{\alpha}{\beta}\int_s^\infty \frac{\beta^{\alpha+1}}{\alpha\Gamma(\alpha)}\,\mu^\alpha e^{-\beta\mu}\,d\mu \\
&= \frac{\alpha}{\beta}\int_s^\infty \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)}\,\mu^\alpha e^{-\beta\mu}\,d\mu = \frac{\alpha}{\beta}\int_s^\infty g(\mu;\alpha+1,\beta)\,d\mu \\
&= \frac{\alpha}{\beta}\,[1 - G(s;\alpha+1,\beta)].
\end{aligned}$$

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.[16] Consequently, the incentive to experiment I and the strategy $\kappa^\dagger$ as defined in (3) are non-decreasing in α and non-increasing in β.

For $m(\pi) < s$, we have

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

In the appendix, we verify that $\kappa^\dagger \in \mathcal{K}$ by showing for any fixed α that $I(\alpha,\cdot)$ has a bounded first derivative when $m(\pi) < s$.

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 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 α on the whole real axis, whereas in

[16] Alternatively, for $\alpha' > \alpha''$ the likelihood ratio $g(\alpha;\alpha',\beta)/g(\alpha;\alpha'',\beta)$ is increasing, and for $\beta' > \beta''$ the likelihood ratio $g(\alpha;\alpha,\beta')/g(\alpha;\alpha,\beta'')$ is decreasing. Since the likelihood-ratio ordering implies first-order stochastic dominance, f has the stated monotonicity properties.

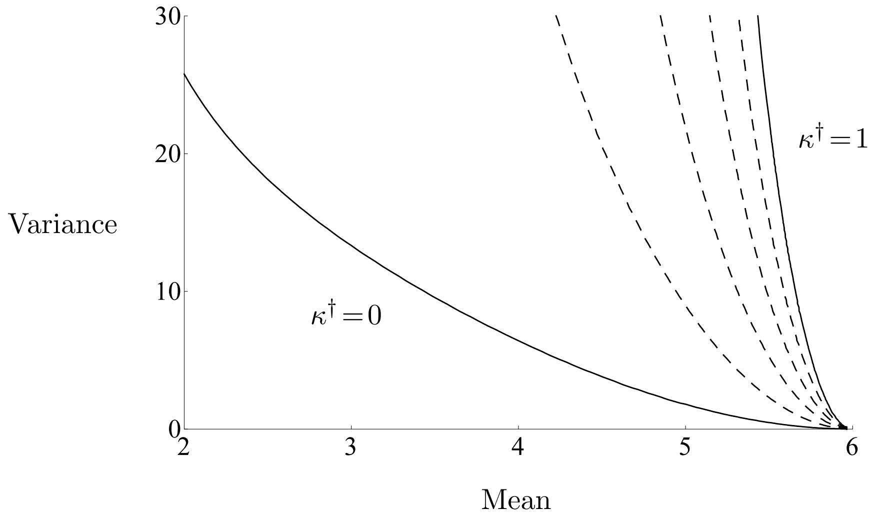

Figure 5: Equilibrium actions for Poisson payoffs and gamma prior

the latter, the mean-preserving spread is concentrated on the positive half-axis and thus raises the option value of experimentation by more.

6 Concluding Remarks

We have seen that when rewards from the risky arm are generated by IID Lévy processes with an unknown average payoff per unit of time, the players' strategy in a symmetric MPE of the undiscounted experimentation game depends only – and in a very simple functional form – on the safe payoff, the expected current payoff of the risky arm, and the expected full-information payoff. Given a finite set from which nature draws the unknown average payoff, the equilibrium strategy is then independent of the actual specification of the payoff-generating processes.

As to the settings with a continuous prior, recall that in the Brownian-normal case the precision of the posterior distribution increases unboundedly with time, as does the inverse of the variance in the Poisson-gamma case. Consequently, the posterior probability density function becomes concentrated on a narrow domain of the support. If we approximated the normal or gamma distribution with a discrete distribution then, over time, the beliefs would become more and more concentrated on the discrete values closest to the true parameter – this suggests that we could take the 'engineering' approach and focus on discrete distributions, with the specification of the payoff-generating processes being irrelevant.[17]

[17] But note that if the two closest neighbours of the true average payoff μ per unit of time are μ_ℓ and $\mu_{\ell+1}$ with $\mu_\ell < \mu < \mu_{\ell+1}$, then, although $m(\pi_T) \simeq \mu$ for large T, we would have $\mathrm{Var}[\mu|\pi_T] \simeq (\mu_{\ell+1}-\mu)(\mu-\mu_\ell)$, which is bounded away from zero.

Letting the discount rate go to zero is going to make the analysis easier in many dynamic settings, but it remains unclear, in general, whether the simplification will be as great as in the present case. Candidates for optimal strategies or best responses may be easier to identify in the undiscounted limit, but there remains the need to obtain a well-defined law of motion, which may again require restrictions such as Lipschitz continuity and could even lead to existence problems. Nevertheless, we believe that the strong long-run average criterion has the potential to prove useful in other contexts, especially since strategies which are optimal under this criterion will shed light on (at least approximately) optimal behaviour for small positive discount rates.

Appendix

Boundedness of Payoffs from Reasonable Strategies

We present the case $L = 1$ only, so that $\ell \in \{0,1\}$, $\pi = \pi_1 \in [0,1]$, $\mu_0 < s < \mu_1$, $m(\pi) = (1-\pi)\mu_0 + \pi\mu_1$ and $f(\pi) = (1-\pi)s + \pi\mu_1$. Suppose first that the Lévy measures ν_0 and ν_1 are non-trivial and equivalent.

For the description of the evolution of beliefs, it is convenient to work with the log odds ratio

$$\omega_t = \ln \frac{\pi_t}{1-\pi_t}\,,$$

so that

$$\pi_t = \frac{e^{\omega_t}}{1+e^{\omega_t}} \quad \text{and} \quad 1-\pi_t = \frac{e^{-\omega_t}}{1+e^{-\omega_t}}\,.$$

Lemma A.1 *There exists a constant $C > 0$ such that for all $x, y \in \mathbb{R}$,*

$$\frac{e^{x+y}}{1+e^{x+y}} \leq \frac{e^x}{1+e^x} + \frac{e^x}{(1+e^x)^2}\,y + C\,\frac{e^x}{(1+e^x)^3}\,y^2.$$

Proof: For

$$f(x,y) = \frac{e^{x+y}}{1+e^{x+y}}\,,$$

we compute the partial derivatives

$$f_y(x,y) = \frac{e^{x+y}}{(1+e^{x+y})^2}\,, \quad f_{yy}(x,y) = \frac{e^{x+y}(1-e^{x+y})}{(1+e^{x+y})^3}\,.$$

For fixed x, the function $f(x,\cdot)$ thus has the following second-order Taylor approximation around $y_0 = 0$:

$$f(x,y) \approx \frac{e^x}{1+e^x} + \frac{e^x}{(1+e^x)^2}\,y + \frac{1}{2}\frac{e^x(1-e^x)}{(1+e^x)^3}\,y^2.$$

As $1 - e^x \leq 1$, we have the local (with respect to the second variable) upper bound

$$f(x,y) \leq \frac{e^x}{1+e^x} + \frac{e^x}{(1+e^x)^2}\,y + \frac{1}{2}\frac{e^x}{(1+e^x)^3}\,y^2.$$

Replacing the factor $\frac{1}{2}$ in the last term by a sufficiently large constant C ensures a global upper bound.[18] ■

Suppose now that starting from $\pi_0 = \pi$ (and corresponding $\omega_0 = \omega$), the players use the strategy profile $(\kappa_1, \ldots, \kappa_N) \in \mathcal{K}^N$. By an extension of the results in Cohen and Solan (2013, Section 3.2) to more than one agent, the log odds ratio at time $t > 0$ can be written as

$$\omega_t = \omega + \eta_\ell \left[k_0 t + \sum_{n=1}^{N} \int_0^t \kappa_n(\pi_{s-})\,ds \right] + M_t^\ell,$$

[18] Numerical computations suggest that $C = 2$ is large enough.

where

$$\eta_\ell = (-1)^{\ell+1} \frac{(\rho_1 - \rho_0)^2}{2\sigma^2} - (\lambda_1 - \lambda_0) + \int_{\mathbb{R}\setminus\{0\}} \ln \frac{\nu_1}{\nu_0}(h)\, \nu_\ell(dh),$$

$\frac{\nu_1}{\nu_0}$ is the Radon-Nikodym derivative of ν_1 with respect to ν_0, and M^ℓ is a martingale under the probability measure $\mathbb{P}_\ell$ associated with state ℓ. The expectation and variance of M^ℓ under this measure, moreover, satisfy $\mathbb{E}_\ell[M_t^\ell] = 0$ and $\mathrm{Var}_\ell[M_t^\ell] \leq C_\ell\, t$ for all t and a positive constant C_ℓ.[19]

As $\ln x < x - 1$ for all positive $x \neq 1$, and $\frac{\nu_1}{\nu_0} = (\frac{\nu_0}{\nu_1})^{-1}$, one sees that

$$\int_{\mathbb{R}\setminus\{0\}} \ln \frac{\nu_1}{\nu_0}(h)\, \nu_0(dh) < \lambda_1 - \lambda_0 < \int_{\mathbb{R}\setminus\{0\}} \ln \frac{\nu_1}{\nu_0}(h)\, \nu_1(dh)$$

unless $\nu_1 = \nu_0$, in which case the inequality $\mu_1 > \mu_0$ implies $\rho_1 > \rho_0$. So $\eta_0 < 0 < \eta_1$. As $\kappa_n \geq 0$ for all n, this in turn implies

$$\omega + \eta_0 k_0 t + M_t^0 \geq \omega_t \geq \omega + \eta_1 k_0 t + M_t^1.$$

By Lemma A.1,

$$\pi_t \leq \frac{e^{\omega + \eta_0 k_0 t + M_t^0}}{1 + e^{\omega + \eta_0 k_0 t + M_t^0}} \leq \frac{e^{\omega + \eta_0 k_0 t}}{1 + e^{\omega + \eta_0 k_0 t}} + \frac{e^{\omega + \eta_0 k_0 t}}{(1 + e^{\omega + \eta_0 k_0 t})^2}\, M_t^0 + C\, \frac{e^{\omega + \eta_0 k_0 t}}{(1 + e^{\omega + \eta_0 k_0 t})^3}\, (M_t^0)^2$$

and

$$1 - \pi_t \leq \frac{e^{-\omega - \eta_1 k_0 t - M_t^1}}{1 + e^{-\omega - \eta_1 k_0 t - M_t^1}} \leq \frac{e^{-\omega - \eta_1 k_0 t}}{1 + e^{-\omega - \eta_1 k_0 t}} - \frac{e^{-\omega - \eta_1 k_0 t}}{(1 + e^{-\omega - \eta_1 k_0 t})^2}\, M_t^1 + C\, \frac{e^{-\omega - \eta_1 k_0 t}}{(1 + e^{-\omega - \eta_1 k_0 t})^3}\, (M_t^1)^2.$$

Writing $C'_\ell = C C_\ell$, we thus have

$$\mathbb{E}_0[\pi_t] \leq e^{\omega + \eta_0 k_0 t} \left(1 + C \mathrm{Var}_0[M_t^0]\right) = e^{\omega + \eta_0 k_0 t}(1 + C'_0 t) = \frac{\pi}{1 - \pi}\, e^{\eta_0 k_0 t}(1 + C'_0 t)$$

and

$$\mathbb{E}_1[1 - \pi_t] \leq e^{-\omega - \eta_1 k_0 t} \left(1 + C \mathrm{Var}_1[M_t^1]\right) = e^{-\omega - \eta_1 k_0 t}(1 + C'_1 t) = \frac{1 - \pi}{\pi} e^{-\eta_1 k_0 t}(1 + C'_1 t).$$

Now let player n use a reasonable strategy. Then there is a constant $C_2 > 0$ such that

$$[1 - \kappa_n(\pi)]s + \kappa_n(\pi) m(\pi) - f(\pi) \geq C_2 \left[\max\{s, m(\pi)\} - f(\pi)\right]$$

for all π. Note that $\max\{s, m(\pi)\} - f(\pi)$ is bounded below by $s - f(\pi) = \pi\,(s - \mu_1)$ and by $m(\pi) - f(\pi) = (1 - \pi)\,(\mu_0 - s)$.

Given the prior belief $\pi_0 = \pi$, the player uses the expectation operator $\mathbb{E}_\pi = (1 - \pi)\mathbb{E}_0 + \pi \mathbb{E}_1$ to compute her objective function. Thus,

[19]For any fixed action profile, M^ℓ has stationary increments, so its variance grows linearly with time. C_ℓ can be chosen as the rate at which the variance grows when all players use the risky arm exclusively.

$$
\begin{aligned}
&u_n(\pi|\kappa_1,\dots,\kappa_N)\\
&\quad \geq \; (1-\pi)C_2(s-\mu_1)\mathbb{E}_0\Big[\int_0^\infty \pi_t\,dt\Big] + \pi C_2(\mu_0 - s)\mathbb{E}_1\Big[\int_0^\infty (1-\pi_t)\,dt\Big]\\
&\quad = \; (1-\pi)C_2(s-\mu_1)\int_0^\infty \mathbb{E}_0[\pi_t]\,dt + \pi C_2(\mu_0 - s)\int_0^\infty \mathbb{E}_1[1-\pi_t]\,dt\\
&\quad \geq \; \pi C_2(s-\mu_1)\int_0^\infty e^{\eta_0 k_0 t}(1+C_0' t)\,dt + (1-\pi)C_2(\mu_0 - s)\int_0^\infty e^{-\eta_1 k_0 t}(1+C_1' t)\,dt\\
&\quad = \; \pi C_2(s-\mu_1)\frac{C_0' - \eta_0 k_0}{\eta_0^2 k_0^2} + (1-\pi)C_2(\mu_0 - s)\frac{C_1' + \eta_1 k_0}{\eta_1^2 k_0^2}\,.
\end{aligned}
$$

This is the desired result.

Next, suppose that the Lévy measure ν_1, say, is not absolutely continuous with respect to ν_0. Take a ν_0-null set $B \subseteq \mathbb{R}\backslash\{0\}$ with $\nu_1(B) > 0$. In state $\ell = 1$, we then have $\mathbb{P}_1[\pi_t = 1] \geq 1 - e^{-\nu_1(B)t}$, so that

$$
\mathbb{E}_1[1-\pi_t] = \mathbb{P}_1[\pi_t = 1]\cdot 0 + \mathbb{P}_1[\pi_t < 1]\cdot \mathbb{E}_1[1-\pi_t|\pi_t < 1] \leq \mathbb{P}_1[\pi_t < 1] \leq e^{-\nu_1(B)t}.
$$

This exponential convergence again allows us to compute an upper bound for $\int_0^\infty \mathbb{E}_1[1-\pi_t]\,dt$.

Finally, if both Lévy measures are trivial, the inequality $\eta_0 < 0 < \eta_1$ holds trivially, and the result follows as above.

Viscosity Solutions of the HJB Equation

Consider a nonempty, open, connected and bounded set $\Omega \subset \mathbb{R}^L$. Denote the set of all symmetric $L \times L$ matrices by $\mathbb{S}^L$. Let $H \in C(\Omega \times \mathbb{R}^L \times \mathbb{S}^L \times \mathbb{R})$ satisfy

$$
H(x,p,X+Y,d) \geq H(x,p,X,d+q)
$$

for all $(x,p,X,d) \in \Omega \times \mathbb{R}^L \times \mathbb{S}^L \times \mathbb{R}$, all positive semidefinite $Y \in \mathbb{S}^N$ and all $q \geq 0$.[20]

We are interested in solutions $u\colon \overline{\Omega} \to \mathbb{R}$ of boundary value problems of the form

$$
H(x, Du, D^2u, u - Mu) = 0 \qquad \text{in } \Omega, \tag{A.1}
$$

$$
u = v \qquad \text{on } \partial\Omega, \tag{A.2}
$$

where Du and D^2u are the gradient and the Hessian matrix of u, respectively, M is an operator mapping $C(\overline{\Omega})$ into itself, and $v \in C(\overline{\Omega})$.

A function $u \in C(\overline{\Omega})$ is called a *viscosity subsolution* of (A.1) if for every $\phi \in C^2(\overline{\Omega})$ and every $x_0 \in \Omega$ such that $\phi \geq u$ on $\overline{\Omega}$ and $\phi(x_0) = u(x_0)$,

$$
H(x_0, D\phi(x_0), D^2\phi(x_0), \phi(x_0) - M\phi(x_0)) \geq 0.
$$

Analogously, a function $u \in C(\overline{\Omega})$ is called a *viscosity supersolution* of (A.1) if for every $\phi \in$

[20] Note that the variables X and Y just introduced are unrelated to the objects for which we use these symbols in the main text.

$C^2(\overline{\Omega})$ and every $x_0 \in \Omega$ such that $\phi \leq u$ on $\overline{\Omega}$ and $\phi(x_0) = u(x_0)$,

$$H(x_0, D\phi(x_0), D^2\phi(x_0), \phi(x_0) - M\phi(x_0)) \leq 0.$$

Finally, $u \in C(\overline{\Omega})$ is called a *viscosity solution* of (A.1) if it is a viscosity sub- and supersolution of (A.1).

The HJB equation (1) and its reformulation (2) are both of the form (A.1) with $\Omega = \overset{\circ}{\Delta}_L$, the operator in question being

$$Mu(\pi) = \frac{1}{\lambda(\pi)} \int_{\mathbb{R}\setminus\{0\}} u(j(\pi, h))\, \nu(\pi)(dh).$$

By the arguments that led us from (1) to (2) in Section 4, these equations have the same viscosity solutions. We will refer to either equation as the HJB equation in what follows.

Suppose that all players except player n use the strategy $\kappa^\dagger$ defined in (3). Let $u^*(\cdot|\kappa^\dagger_{\neg n})$ denote the value function of the control problem that player n faces when choosing a best response, and $u(\cdot|\kappa^\dagger, \kappa^\dagger_{\neg n})$ the player's payoff function when she also uses strategy $\kappa^\dagger$, that is,

$$u(\pi|\kappa^\dagger, \kappa^\dagger_{\neg n}) = \mathbb{E}^{(\kappa^\dagger, \kappa^\dagger_{\neg n})} \left[\int_0^\infty \left\{ [1 - \kappa^\dagger(\pi_t)]s + \kappa^\dagger(\pi_t) m(\pi_t) - f(\pi_t) \right\} dt \,\middle|\, \pi_0 = \pi \right].$$

By definition, $u^*(\cdot|\kappa^\dagger_{\neg n}) \geq u(\cdot|\kappa^\dagger, \kappa^\dagger_{\neg n})$. We shall establish the converse inequality via a comparison result for viscosity sub- and supersolutions.

We know that both functions are bounded. Assume for now that they are actually continuous on Δ_L; we will justify this assumption later.

Lemma A.2 *The value function $u^*(\cdot|\kappa^\dagger_{\neg n})$ is a viscosity subsolution of the HJB equation.*

PROOF: We simplify the notation by writing u instead of $u^*(\cdot|\kappa^\dagger_{\neg n})$.

Consider $\phi \in C^2(\Delta_L)$ and $\pi_0 \in \overset{\circ}{\Delta}_L$ such that $u - \phi \leq 0 = u(\pi_0) - \phi(\pi_0)$. To establish that u is a viscosity subsolution of (1), we must show that

$$\max_{k \in [0,1]} \left\{ (1-k)s + km(\pi_0) - f(\pi_0) + [k_0 + (N-1)\kappa^\dagger(\pi_0) + k]\mathcal{G}\phi(\pi_0) \right\} \geq 0.$$

Suppose that this is not the case, so that

$$(1-k)s + km(\pi_0) - f(\pi_0) + [k_0 + (N-1)\kappa^\dagger(\pi_0) + k]\mathcal{G}\phi(\pi_0) < 0$$

for all $k \in [0,1]$. For $\varepsilon > 0$, define $\psi \in C^2(\Delta_L)$ by

$$\psi(\pi) = \phi(\pi) + \varepsilon \|\pi - \pi_0\|^4$$

and note that $\psi \to \phi$ uniformly as $\varepsilon \to 0$. For $\delta > 0$, let $B_\delta(\pi_0) \subset \mathbb{R}^L$ be the open ball of radius δ centered at π_0. By continuity, we can find $\varepsilon, \delta > 0$ such that $B_\delta(\pi_0) \subset \overset{\circ}{\Delta}_L$ and

$$(1-k)s + km(\pi) - f(\pi) + [k_0 + (N-1)\kappa^\dagger(\pi) + k]\mathcal{G}\psi(\pi) < 0$$

for all $k \in [0,1]$ and all $\pi \in B_\delta(\pi_0)$. As π_0 is a strict maximizer of $u - \psi$, moreover, there exists $\gamma > 0$ such that $u(\pi) - \psi(\pi) \leq -\gamma$ for $\pi \in \Delta_L \setminus B_\delta(\pi_0)$. Suppose now that player n uses the strategy $\kappa \in \mathcal{K}$ against the other players' common strategy $\kappa^\dagger$. Define $\tau = \inf\{t > 0\colon \|\pi_t - \pi_0\| > \delta\}$. As $k_0 > 0$, we have $\mathbb{E}^{(\kappa,\kappa^\dagger_{\neg n})}[\tau] < \infty$ and

$$\begin{aligned}
&\mathbb{E}^{(\kappa,\kappa^\dagger_{\neg n})}\left[\int_0^\tau \{[1-\kappa(\pi_t)]s + \kappa(\pi_t)m(\pi_t) - f(\pi_t)\}\,dt + u(\pi_\tau)\right] - u(\pi_0) \\
\leq\ & \mathbb{E}^{(\kappa,\kappa^\dagger_{\neg n})}\left[\int_0^\tau \{[1-\kappa(\pi_t)]s + \kappa(\pi_t)m(\pi_t) - f(\pi_t)\}\,dt + \psi(\pi_\tau)\right] - \psi(\pi_0) - \gamma \\
=\ & \mathbb{E}^{(\kappa,\kappa^\dagger_{\neg n})}\left[\int_0^\tau \left\{[1-\kappa(\pi_t)]s + \kappa(\pi_t)m(\pi_t) - f(\pi_t) + [k_0 + (N-1)\kappa^\dagger(\pi) + \kappa(\pi_t)]\mathcal{G}\psi(\pi_t)\right\}\,dt\right] - \gamma \\
<\ & -\gamma,
\end{aligned}$$

where the equality in the third line follows from Dynkin's formula. But this contradicts the dynamic programming principle, which states that

$$u(\pi_0) = \sup_{\kappa \in \mathcal{K}} \mathbb{E}^{(\kappa,\kappa^\dagger_{\neg n})}\left[\int_0^\tau \{[1-\kappa(\pi_t)]s + \kappa(\pi_t)m(\pi_t) - f(\pi_t)\}\,dt + u(\pi_\tau)\right].$$

■

Lemma A.3 *The payoff function $u(\cdot|\kappa^\dagger, \kappa^\dagger_{\neg n})$ is a viscosity supersolution of the HJB equation.*

Proof: We simplify the notation by writing u instead of $u(\cdot|\kappa^\dagger, \kappa^\dagger_{\neg n})$.

Consider $\phi \in C^2(\Delta_L)$ and $\pi_0 \in \mathring{\Delta}_L$ such that $u - \phi \geq 0 = u(\pi_0) - \phi(\pi_0)$. For any deterministic time $\tau > 0$,

$$\begin{aligned}
0 &= \mathbb{E}^{(\kappa^\dagger,\kappa^\dagger_{\neg n})}\left[\int_0^\tau \left\{[1-\kappa^\dagger(\pi_t)]s + \kappa^\dagger(\pi_t)m(\pi_t) - f(\pi_t)\right\}\,dt + u(\pi_\tau)\right] - u(\pi_0) \\
&\geq \mathbb{E}^{(\kappa^\dagger,\kappa^\dagger_{\neg n})}\left[\int_0^\tau \left\{[1-\kappa^\dagger(\pi_t)]s + \kappa^\dagger(\pi_t)m(\pi_t) - f(\pi_t)\right\}\,dt + \phi(\pi_\tau)\right] - \phi(\pi_0) \\
&= \mathbb{E}^{(\kappa^\dagger,\kappa^\dagger_{\neg n})}\left[\int_0^\tau \left\{[1-\kappa^\dagger(\pi_t)]s + \kappa^\dagger(\pi_t)m(\pi_t) - f(\pi_t) + [k_0 + N\kappa^\dagger(\pi_t)]\mathcal{G}\phi(\pi_t)\right\}\,dt\right]
\end{aligned}$$

by Dynkin's formula. Dividing through by τ and letting $\tau \to 0$, we get

$$[1-\kappa^\dagger(\pi_0)]s + \kappa^\dagger(\pi_0)m(\pi_0) - f(\pi_0) + [k_0 + N\kappa^\dagger(\pi_0)]\mathcal{G}\phi(\pi_0) \leq 0,$$

which is equivalent to

$$\frac{[k_0 + (N-1)\kappa^\dagger(\pi_0)][s - m(\pi_0)] - [f(\pi_0) - s]}{k_0 + N\kappa^\dagger(\pi_0)} - [s - m(\pi_0)] + \mathcal{G}\phi(\pi_0) \leq 0.$$

As

$$\kappa^\dagger(\pi_0) \in \arg\max_{k \in [0,1]} \frac{[k_0 + (N-1)\kappa^\dagger(\pi_0)][s - m(\pi_0)] - [f(\pi_0) - s]}{k_0 + (N-1)\kappa^\dagger(\pi_0) + k},$$

u is thus a viscosity supersolution of (2). ■

The comparison result that yields the inequality $u^*(\cdot|\kappa^\dagger_{\neg n}) \leq u(\cdot|\kappa^\dagger, \kappa^\dagger_{\neg n})$ is due to Ishii and Yamada (1993). These authors consider functional equations $F(x, u, Du, D^2u, u - Mu) = 0$ such that

$$F(x, r, p, X + Y, d) \leq F(x, r, p, X, d + q)$$

for all $(x, r, p, X, d) \in \Omega \times \mathbb{R} \times \mathbb{R}^L \times \mathbb{S}^L \times \mathbb{R}$, all positive semidefinite $Y \in \mathbb{S}^N$ and all $q \geq 0$. This means that F corresponds to $-H$ here.[21] As a consequence, the inequalities defining sub- and supersolutions in terms of F are the opposite of those in terms of H.

There is a second, more substantive difference between the definitions of Ishii and Yamada (1993) and ours. Translated back into our setting, a function $u \in C(\overline{\Omega})$ is a viscosity subsolution of (A.1) in their sense if for every $\phi \in C^2(\Omega)$ and every $x_0 \in \Omega$ such that $\phi - u$ has a local minimum in x_0,

$$H(x_0, D\phi(x_0), D^2\phi(x_0), u(x_0) - Mu(x_0)) \geq 0.$$

Analogously, a function $u \in C(\overline{\Omega})$ is a viscosity supersolution of (A.1) in their sense if for every $\phi \in C^2(\Omega)$ and every $x_0 \in \Omega$ such that $\phi - u$ has a local maximum at x_0,

$$H(x_0, D\phi(x_0), D^2\phi(x_0), u(x_0) - Mu(x_0)) \leq 0.$$

In these alternative definitions, therefore, u is replaced by ϕ only as far as the gradient and Hessian are concerned, but not in the nonlocal term. When M is an integral operator of the type considered here, however, an argument in Alvarez and Tourin (1996, p. 300) implies that these definitions are in fact equivalent to ours.[22]

Lemma A.4 *Let a function $v \in C(\partial\Delta_L)$ be given. Suppose that $\underline{u}$ is a viscosity subsolution of the HJB equation, $\overline{u}$ a viscosity supersolution, and $\underline{u} \leq v \leq \overline{u}$ on $\partial\Delta_L$. Then $\underline{u} \leq \overline{u}$ on Δ_L.*

PROOF: Equation (2) takes the form assumed in Ishii and Yamada (1993) with the domain $\Omega = \overset{\circ}{\Delta}_L$, the function

$$F(x, p, X, d) = -\frac{1}{2\sigma^2} R(x)' X R(x) + L(x)' p + \lambda(x) d - c(x)$$

where

$$R(x) = \begin{pmatrix} x_1[\rho_1 - \rho(x)] \\ \vdots \\ x_L[\rho_L - \rho(x)] \end{pmatrix}, \qquad L(x) = \begin{pmatrix} x_1[\lambda_1 - \lambda(x)] \\ \vdots \\ x_L[\lambda_L - \lambda(x)] \end{pmatrix}$$

and

$$\begin{aligned} c(x) &= \max_{k \in [0,1]} \frac{[k_0 + (N-1)\kappa^\dagger(x)][s - m(x)] - [f(x) - s]}{k_0 + (N-1)\kappa^\dagger(x) + k} - [s - m(x)] \\ &= \frac{[k_0 + (N-1)\kappa^\dagger(x)][s - m(x)] - [f(x) - s]}{k_0 + N\kappa^\dagger(x)} - [s - m(x)], \end{aligned}$$

[21]Note that Ishii and Yamada (1993) allow the value of the solution to enter as a separate variable besides its difference with the nonlocal operator. Because of the absence of discounting, this generality is not needed here, so H has one argument fewer.

[22]See Azimzadeh, Bayraktar and Labahn (2017) for a related discussion.

and the operator

$$Mu(x) = \frac{1}{\lambda(x)} \int_{\mathbb{R}\setminus\{0\}} u(j(x,h))\, \nu(x)(dh).$$

It is straightforward to check that F, M and the function $B(x,u) = u - v(x)$ defined on $\partial\Delta_L \times \mathbb{R}$ satisfy all the conditions imposed by Ishii and Yamada (1993). The result thus follows from their Theorem 3.1. ■

Corollary A.1 $u^*(\cdot|\kappa^\dagger_{\neg n}) = u(\cdot|\kappa^\dagger, \kappa^\dagger_{\neg n})$.

Proof: The proof is by induction over the dimension of the faces of the simplex. The 0-faces (vertices) correspond to degenerate beliefs that assign probability 1 to one of the states; at all these vertices, both functions assume the value 0. An application of Lemma A.4 for $L = 1$ now yields $u^*(\cdot|\kappa^\dagger_{\neg n}) = u(\cdot|\kappa^\dagger, \kappa^\dagger_{\neg n})$ along any 1-face (edge) of the simplex. Applying the lemma for $L = 2$ then proves this identity for all 2-faces (facets), and so on until the entire simplex is covered. ■

As a by-product, this confirms that the value function is indeed the unique viscosity solution of the HJB equation.

It remains to justify our assumption that the functions $u^*(\cdot|\kappa^\dagger_{\neg n})$ and $u(\cdot|\kappa^\dagger, \kappa^\dagger_{\neg n})$ are continuous. In fact, using upper semicontinuous and lower semicontinuous envelopes, Ishii and Yamada (1993) define the notion of viscosity sub- and supersolution for functions that are merely locally bounded. Lemmas A.2 and A.3 still hold then, and Lemma A.4 generalizes in a way that ensures that any viscosity solution satisfying a continuous boundary condition must be continuous overall; see Ishii and Yamada (1993, Corollary 3.3). Continuity of the functions in question follows from an iterative application of this result as in the proof of Corollary A.1.

Verification that $\kappa^\dagger \in \mathcal{K}$ for Brownian Payoffs and Normal Prior

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 $\Pi = \mathbb{R} \times [\tau_0, \infty[$, 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 that $\kappa^\dagger \in \mathcal{K}$, it suffices to show that $I\tau^{-1}$ is Lipschitz continuous on $\Pi(a,b) = \{\pi \in \Pi : a \leq I(\pi) \leq b\}$ 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 that $\kappa^\dagger \in \mathcal{K}$ for Poisson Payoffs and Gamma Prior

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

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

We fix α as well as positive real numbers $a < b$. To verify that $\kappa^\dagger \in \mathcal{K}$, 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

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

For $\beta = \alpha/s$ and $\mu < s$, the term in square brackets under the integral is positive, so we have $G(s;\alpha,\frac{\alpha}{s}) - G(s;\alpha+1,\frac{\alpha}{s}) > 0$. For $\beta \searrow \frac{\alpha}{s}$, therefore, the numerator $s\, G(s;\alpha,\beta) - \frac{\alpha}{\beta}\, G(s;\alpha+1,\beta)$ 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

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

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

References

Alvarez, O. and A. Tourin (1996): "Viscosity Solutions of Nonlinear Integro-Differential Equations," *Annales de l'Institut Henri Poincaré*, **13**, 293–317.

Azimzadeh, P., E. Bayraktar and G. Labahn (2017): "Convergence of Approximation Schemes for Weakly Nonlocal Second Order Equations," arXiv:1705.02922v4.

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.

Dutta, P.K. (1991): "What Do Discounted Optima Converge to?: A Theory of Discount Rate Asymptotics in Economic Models," *Journal of Economic Theory*, **55**, 64–94.

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.

Ishii, K. and N. Yamada (1993): "Viscosity Solutions of Nonlinear Second Order Elliptic PDEs Involving Nonlocal Operators," *Osaka Journal of Mathematics*, **30**, 439–455.

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.

Ke, T.T. and J.M. Villas-Boas (2019): "Optimal Learning Before Choice," *Journal of Economic Theory*, **180**, 383–437.

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.

Øksendal, B. and A. Sulem (2007): *Applied Stochastic Control of Jump Diffusions (2^{nd} Edition).* New York: Springer-Verlag.

Peitz, M., S. Rady and P. Trepper (2017): "Experimentation in Two-Sided Markets," *Journal of the European Economic Association*, **15**, 128–172.

Pham, H. (2009): *Continuous-time Stochastic Control and Optimization with Financial Applications.* New York: Springer-Verlag.

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

Veronesi, P. (2000): "How Does Information Quality Affect Stock Returns?," *Journal of Finance*, **55**, 807–837.