Discounted Brownian Bandits with Normal Beliefs

Ege Can Doğaroğlu
3464688

Introduction

Learning through experimentation is typically characterized as a multi-armed bandit problem, where the agents face the decision of choosing among a set of actions, whose consequences might be contingent on the state of the world. In the simplest case with two actions, we refer to the two-armed bandit problem that involves one safe arm with a known expected payoff and a second risky arm with an unknown expected payoff that may be higher or lower than the safe payoff. The problem here is typically portrayed as a trade off between exploiting what we already know (and is safe) and exploring the condition of something that might be possibly better than the safe course of action. As opposed to other types of learning environments, where typically the cost of information is exogenous, in bandit problems the cost of learning is simply the opportunity cost of not exploiting the returns of the safe action. So, one faces the decision of to what degree they should experiment and when to stop experimenting. Typically, the strategies are defined as mappings from the belief space to the allocation of a resource among the actions, where it's assumed that the agent makes her decision only depending on her current subjective belief of the world. The strategies are then chosen in order to maximize total expected payoffs.

Historically, the environment usually involved only a single-agent acting on her own and only observing her own actions and their consequences, whereas in the modern literature the strategic interactions in a social learning environment, allowing agents to learn from each others' actions is the focal point of the analysis. With the recent developments, notoriously intractable nature of the bandit problems are understood to a greater degree. Mainly, we talk about two effects that evolves from the strategic consideration in a social learning environment: the free-riding effect and the encouragement effect. The first is a natural result of the two main properties of the model: ability to learn from each others' actions and the assumption that the experimentation procedure and the state of the world is identical among the agents. This conjunction leads agents to shy away from costly experimentation as long as they believe that there are other people, who are undertaking a sufficient amount of experimentation and (depending on the model) are going to share their experimentation results fully and truthfully with them. Believing this, agents have no incentive to undertake costly experimentation, when they can just as well continue to use the safe arm and be indirectly benefiting from the experimentation at the same time. The latter effect can also be thought of as a result of free-riding. Given that incentives to experiment are problematic and as nobody wants to be the guinea pig of the group, experimentation process can slow down undesirably, even though it's beneficial for everyone. Hence, knowing that a good experimentation outcome today might lead someone to join experimentation tomorrow one has then the incentive to be the leader of the group to motivate social welfare. This effect is only sustained if the good news are inconclusive i.e. when there is still something to learn after a good outcome. Nevertheless, typically the free-riding effect is still persistent enough to cause inefficient levels of experimentation. Although, with recent developments, we see that efficiency is in fact achievable under different equilibrium considerations.

As a technical simplification, the literature typically treats the problem in a setting where the state of the world is considered to be binary: *good* or *bad*. This naturally leads to the conclusion that prior to conducting experimentation, agents have to know exactly what two outcomes they might be facing. In an artificial game where rules are communicated in advance (like a gambling machine) this assumption might be non-restrictive. Although, as especially in the economics literature the setup is thought of as an abstraction of real-world experimentation like drug trials or mineral discoveries; thinking of the state space as higher dimensional (potentially continuous) should allow for an environment that is more suitable for real-world justifications, which may also lead to a more elastic learning environment and higher levels of efficiency.

A potential benefit of achieving tractable analysis under a continuous state space might be that, then, the state of the world can be thought of as the resulting outcome of a completely different non-binary process. For example, the *goodness* of the arm might be time-dependent. This might relate to concepts like time-dependent preferences or external shocks that change the underlying characteristic of payoffs. In the case of drug trials, the binary setup only allows for whether the drug is more effective than a benchmark level or not. In a non-binary state, how effective the drug is can also be characterized in order to potentially compare with other opportunities like simultaneous experimentation. Further, the effectiveness of the drug might be dependent on time-dependent complications with other stimulants and hence the signal that experimentation produces might also be biased depending on at which point in time the experimentation is undertaken. With time dependent tastes, expected benefit of experimentation (and it's relation to the opportunity cost) can be variable at different points in time, which might lead to time variable cutoff beliefs that agents use to decide when to experiment. Such a model might be used to explain periodic risk sensitives depending on factors like self-confidence and liquidity.

Related Literature

In the modern economics literature, strategic experimentation with bandit problems generally dates to Bolton and Harris (1999), where authors characterize the flow payoffs as increments of a Brownian motion that has an unknown drift coefficient. Uncertainty here is characterized as a binary state of the world and the change in the posterior belief is normally distributed with zero mean and a variance that is dependent on the spread between these two possible states. This is the pioneering work for the analysis of free-riding and encouragement effects, where authors are able to provide a symmetric equilibrium using Markov strategies with beliefs as the state variable. Throughout the paper as value functions, best responses and hence the equilibrium all depend on the belief evolution, the spread between these two variables play a central role in the analysis.

In their benchmark model Keller, Rady, and Cripps (2005), consider the case of *exponential bandits*, where lump-sum payoffs of the arms arrive according to a Poisson process and hence time between each arrival is distributed exponentially. Here, the *goodness* of the arm is reduced to the intensity parameter of the Poisson process i.e. how frequently the lump-sums arrive. In particular, the mean of the lump-sums are identical whether the risky arm is good or bad. This characterization simplifies the problem to only learning about the intensity of the Poisson process and in turn lets the subjective belief on this binary variable move deterministically according to (un)observed arrivals of lump-sums and the level of undertaken experimentation. In this model, arrivals are conclusive, in that when they exist, they reveal the state to be good to all players. In the absence of arrivals, the belief moves continuously in the direction of the state that is known to have the lowest Poisson intensity - the bad risky arm. Similarly to the previous paper, in both states, intensity of the arrivals is exactly known to all players. In the analysis of this problem, authors first consider two extreme benchmark cases. First, the agent is acting myopically i.e. discount future rewards infinitely high and act alone. Second, a social planner coordinates all players to maximize the social welfare and agents act in full cooperation. They provide the unique symmetric equilibrium outcomes as well as some characterization of asymmetric equilibria. It's shown in particular, that the encouragement effect in Bolton and Harris (1999) is not existent in this model, due to the fact that good news on the risky arm are conclusive. This causes naturally for future experimentation on the risky arm to be of no worth after a lump-sum arrival (good news) as the state is fully revealed. It's also shown that all Markov perfect equilibria are inefficient due to the free-riding effect.

In Keller and Rady (2010) the framework is expanded to include the case, where the good news are inconclusive about the state of the risky arm. Here, in the absence of arrivals subjective belief still deteriorates but it's speed is dependent on the difference of intensities in two states i.e. how spread out the payoffs of the risky arm are. In case of good news, belief jumps up independently of experimentation intensity, but dependent on the risk level of the arm. The encouragement effect that was absent in the previous paper exists here due to inconclusive good news. This effect is also sustained when *breakdowns* are considered instead of *breakthroughs*, shown by Keller and Rady (2015) whether the news are conclusive or not. In a more recent paper, Hörner, Klein, and Rady (2022) show that for Lévy bandits where the payoff generating process can have both Brownian or Poisson components, the inefficiency of Markov perfect equilibria can be overcome with relaxing the equilibrium concept to *Strongly Symmetric Equilibrium* that allows for punishing deviant players

in a grim trigger fashion.

In all of these models, it can be seen that much of the results like the evolution of posterior beliefs, value functions, equilibrium strategies depend on the exact values of the binary state of the world whether it leads to a certain Poisson intensity or a drift coefficient. This suggests in particular that in case of experimentation, agents are fully aware of the magnitude of the risk they are taking and update their beliefs or use optimal strategies accordingly. In many cases, risk might be non-binary and the state of the world can be treated as a continuous random variable instead. One such treatment is undertaken in Keller and Rady (2020) where authors primarily consider an arbitrary number of finite states related to uncertainty of a Lévy process. As a result, they consider discrete probability mass functions to characterize the probability measures over the finite and discrete state space and utilize *infinitesimal generator* to characterize the effect of the learning process on the value function, which greatly simplifies the analysis. In a specification, they consider Brownian payoffs with normal prior beliefs and characterize the equilibrium actions in the two dimensional posterior mean-variance space that constitutes a sufficient statistic for posterior beliefs thanks to the normality assumption. Quite intuitively, authors show the trade off between mean and variance, where for low posterior mean one needs a higher variance (option value of learning) in order to continue experimenting, whereas for high mean values continuing is an equilibrium outcome even with low variance levels. The setup I consider mostly adopts the setting in this specification with the fundamental difference that for optimization, they consider the *strong long-run average criterion* that minimizes the expected deviation from the full information payoffs. In contrast, I will follow the previous literature, where the agents optimize to maximize their expected total discounted payoffs. Due to this difference, I expect the existence of encouragement effect, which doesn't exist in Keller and Rady (2020).

A treatment of Bayesian belief updating with Brownian payoffs and normal beliefs can be found in Chernoff (1968), while the posterior mean and variance distribution, as well as the converge of the posterior mean to the true value of the state can be found in Keller (2009). A comprehensive survey regarding strategic experimentation with various considerations of motivation, delegation, information disclosure and multiplicity of options is given by Hörner and Skrzypacz (2017).

Model

As mentioned, I largely adopt the Brownian specification in Keller and Rady (2020) with the fundamental difference that the future payoffs are discounted.

Information and Incentives. Time $t \in [0, \infty)$ is continuous. There are $N \geq 1$ identical agents who at every instant of time have to decide how they should allocate a perfectly divisible unit resource among two available actions. One of the actions is considered to be *safe*, producing a known constant payoff $s > 0$ per unit of time. The other action is considered to be *risky* and it produces payoffs according to a Brownian motion with an unknown drift and known diffusion coefficient. The state of the world is characterized by a continuous real-valued random variable θ with unbounded support, which is also the drift coefficient of the payoff process

$$X_t^n = \theta t + \sigma dZ_t^n$$

where Z^n are independent standard Wiener processes and σ is the known diffusion coefficient. It's assumed that the random variable θ is drawn from a normal distribution with mean m_0 and precision $\tau_0 > 0$, which also corresponds to agents' common prior belief.

At each instant of time t, all players decide what portion $k_{n,t} \in [0, 1]$ of their available resource they should allocate to the risky arm. They perfectly observe the decisions taken by other players and the resulting payoffs. Using the safe arm is by construction not informative about the state of the world and the risky arm only conveys noisy information due to the diffusion component. In particular, the risky arm is informative if and only if a positive amount of experimentation is undertaken by some agent. I also assume that $\sigma > 0$, so that there is always noise in the experimentation outcome and hence the risky arm never fully reveals the true state of the world.[1]

By the properties of Brownian motion, the expected payoff increment from using the risky arm over an interval of time $[t, t + dt)$ is θdt, which is random due to the uncertainty regarding the state of the world. Hence, if a player allocates $\{k_t\}_{t\geq 0}$ portion of her resource to the risky arm,

[1] Although, beliefs do converge to the true value eventually. See Keller (2009)

her expected payoff increment over the interval $[t, t+dt)$ is $[(1-k_{n,t})s + k_{n,t}\theta]dt$. Then, players (weakly) prefer the risky arm over the safe arm, whenever they think $\theta \geq s$. As θ has unbounded support, regardless of the value of s, there is always incentive to learn about the state of the world. The total expected discounted payoff, expressed in per-period units, is

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

where the expectation is both over the allocation process $\{k_{n,t}\}$ and the state of the world θ. This is also the objective function that the player n chooses the allocations $\{k_{n,t}\}$ in order to maximize. Other players' actions affect this objective function only through the evolution of posterior beliefs about θ, so it's instructive to develop the evolution of such beliefs in order to analyze strategic interactions.

Evolution of Beliefs. Players start the game with a common prior belief regarding the state of the world and update this belief in a Bayesian fashion according to the payoff realizations from the risky arm. As all actions and the resulting payoffs are perfectly observable, at any point in time, the posterior beliefs that players hold are identical. We assume that the prior belief is normally distributed with $\mu \sim N(m_0, \tau_0^{-1})$. Following Keller and Rady (2020) posterior belief then has the form $\mu_t \sim N(m_t, \tau_t^{-1})$ and hence is completely characterized by the pair $\pi_t = (m_t, \tau_t)$ which evolves according to

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

where $d\bar{Z} = \sigma^{-1}([\mu - m_t]dt + \sigma dZ_t)$ is the *innovation process*. Here, change in the posterior mean is stochastic, being pulled towards the true value of the state, while the change in the posterior precision is deterministic and increases linearly in time. Both increments are proportional to the amount of experimentation by the agents. Further, as the process $\{m_t\}$ corresponds to the subjective expected value of the state of the world at time t, we can write the above objective function as

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

highlighting the importance of the posterior belief as the state variable. Again following Keller and Rady (2020), the infinitesimal generator resulting from the above specified posterior belief process for any twice differentiable function u is

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

which will be used below to conduct dynamic programming analysis.

As is customary in the literature, I will continue with the case when agents act alone in the single-agent problem. The corresponding value function will also determine the lower-bound off payoffs one can achieve, as it's assumed that players can always choose to not listen to each others' observations. Later, I will focus on the cooperative solution to inspect the optimal outcome if the decision was given to a social planner and each agent perfectly followed their instructions. This scenario corresponds to the case where the free-riding effect does not exist and hence any deviation in the experimentation level or expected payoff from the cooperative solution will be seen as *inefficient*. Lastly, I will look at the strategic considerations in which, each agent plays their best response to the situation at hand in order to maximize their own expected payoff. In particular, I will consider symmetric Markov perfect equilibrium, where the strategies are deterministic functions of the current belief, which are identical across all agents. This, in turn, implies that the actions taken are also identical since the posterior beliefs at any point in time is same across agents due to perfect observation.

Single-Agent Solution. Given the above objective function if the agent acts alone and does not observe the realizations of other players' actions, her continuation rewards only depend on her current action. Then, by the Principle of Optimality, the value function that arises from the objective maximization must satisfy the HJB equation

$$ru(\pi) = \max_{k\in[0,1]} \left(r[(1-k)s + km] + kGu(\pi)\right)$$

as the infinitesimal generator can be linearly scaled. Further treatment of this function is given in the cooperative case as both equations are identical up to the level of total experimentation.

Myopic Solution. When the discount rate r is infinitely large, players may think of maximizing their returns over the next instant only. Then, they act on their current beliefs where the optimal action has the particularly easy *bang-bang* solution, in which it's optimal to play risky as long as one believes that the state of the world is larger than the known underlying payoff of the safe arm. As the posterior mean m_t equals the subjective expected value of the state, best responses are given by

$$k^* = \begin{cases} 1, & \text{for } m_t \geq s \\ 0, & \text{otherwise} \end{cases}$$

Cooperative Solution. For the efficiency benchmark we can compute the optimal allocation when agents act cooperatively and choose their actions $(k_1, ..., k_N)_{t\geq 0}$ in order to maximize the average total expected payoff. Letting $K = \sum_{n=1}^{N} k_n$ denote the total amount of experimentation and given the subjective expected value m of θ, the average expected payoff increment is

$$\left[\left(1 - \frac{K}{N}\right) s + \frac{K}{N} m\right] dt$$

with the corresponding HJB equation

$$ru(\pi) = \max_{K\in[0,N]} \left\{ r\left[\left(1 - \frac{K}{N}\right) s + \frac{K}{N} m\right] + K\mathcal{G}u(\pi) \right\}$$

which following the literature can be rewritten as

$$u(\pi) = s + \max_{K\in[0,N]} K\left[b(\pi, u) - \frac{c(\pi)}{N}\right]$$

where $b(\pi, u) = \frac{1}{r}\mathcal{G}u(\pi)$ is seen as the expected informational benefit of using the risky arm and $c(\pi) = s - m_t$ is seen as the per player opportunity cost of using the risky arm. Due to linearity, it's easy to see that the maximum is achieved at $K = 1$ if $b(\pi, u) > c(\pi)$ and at $K = 0$ if $b(\pi, u) < c(\pi)$. When $b(\pi, u) = c(\pi)$ any choice of $K \in [0, N]$ can be a solution to the maximization problem. This intuition suggest that there is a cutoff point in beliefs that determines whether continuing to explore or stopping experimentation and choosing the safe action will be a best response. Then, we face the reformulation of the problem as the one of optimal stopping with the HJB equation

$$u(\pi) = \max\left\{ s, m + \frac{N}{r} Gu(\pi) \right\}$$

hence the value function equals to either the stopping payoff or the continuation payoff with full experimentation intensity depending on the current state of the world. By standard arguments, along the stopping boundary that separates these *stopping* and *continuation* regions we must have *value matching* and *smooth pasting* properties so that at the boundary π_t^*

$$\begin{aligned} u(\pi_t^*) &= s \\ \frac{\partial u(\pi_t^*)}{\partial m_t} &= 0 \\ \frac{\partial u(\pi_t^*)}{\partial \tau_t} &= 0 \end{aligned}$$

must hold. Further, in the continuation region, the value function has to satisfy the parabolic second-order linear partial differential equation (PDE) with

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

which is essentially similar to *Reaction-diffusion systems* like the *Kolmogorov-Petrovsky-Piskunov equation* with the fundamental difference that the curvature has a variable coefficient. A quick

analysis suggests that, if the smooth pasting property is to be satisfied, at the boundary, the terms in the square brackets have to be equal to 0. Hence, the boundary posterior mean must be such that $m^* = s$ in order to satisfy the value matching property. This is also the cutoff belief of the myopic single agent. Further, for a function to satisfy the value matching property with the boundary mean $m^* = s$ it is sufficient that it satisfies the smooth pasting properties. So we can deduce the problem of verifying both properties to verifying smooth pasting for the two variables only.

A preliminary research of the stochastic processes and functional analysis literature unfortunately returned no results regarding the exact treatment of such partial differential equations. This is largely due to the difficulty that the corresponding equation is of second-order, inhomogenous, partial with two variables and that it includes variable coefficients. I have come up with some attempts to simplify some dimension of the difficulty by imposing certain assumptions on the value function that in turn lead to perturbations of the PDE. While these assumptions are quite restrictive and difficult to relax, they allow for closed form solutions for the differential equation and may help with the understanding of the original function in a broader sense. Throughout, I will abstract away from any model parametrization in the PDE in order to simplify the process as much as possible. In many cases, parametrization leads to further infeasibilites regarding the closed form solutions.

No Curvature. The simplest case to consider when trying to find a solution to PDE is to lower it's order by assuming that highest order term is equal to zero. Namely, this leads to the property that $\frac{\partial^2 u(m,\tau)}{\partial m^2} = 0$. Then, the equation can be rewritten as the first-order PDE

$$u(\pi) = m + \frac{\partial u(\pi)}{\partial \tau}$$

with the particular solution

$$u(m,\tau) = m + e^{\frac{\tau r \sigma^2}{N}}$$

It's interesting to see that this function has similar contour lines as the equilibrium actions portrayed in Keller and Rady (2020). Nevertheless, it doesn't satisfy either the smooth pasting or the value matching properties as for any value of τ we must have that $u > s$, which implies that the value function must have curvature in m.

Harmonic Functions. A well-behaved class of functions are of the harmonic types which involve assumptions regarding their curvatures. This condition is common in stochastic processes and it requires that $\frac{\partial^2 u}{\partial m^2} + \frac{\partial^2 u}{\partial \tau^2} = 0$, leading to a rephrasing of the PDE as

$$u(m,\tau) = m + \frac{\partial u}{\partial \tau} - \frac{\partial^2 u}{\partial \tau^2}$$

with the closed form solution

$$u(m,\tau) = m + e^{\tau/2}\left(sin\left(\frac{\sqrt{3}\tau}{2}\right)c_1(m) + cos\left(\frac{\sqrt{3}\tau}{2}\right)c_2(m)\right)$$

where $c_1(\cdot)$ and $c_2(\cdot)$ are arbitrary functions of m. For the boundary $\pi^* = (s, \tau^* = \frac{5\pi}{6\sqrt{3}})$ this function satisfies the smooth pasting property for τ, when $c_1 = c_2$; and for m, when $\frac{c_2'}{c_1'} = \frac{1+\sqrt{3}+2\sqrt{2}e^{-(5\pi)/12}}{1-\sqrt{3}}$. It's unclear whether both properties are satisfied under the same parametrization. Further, scaling the second derivative in the functional equation with τ^{-2} i.e. imposing variable coefficient leads to complications regarding the closed form solution. In addition, if u is both harmonic and bounded, it must be a constant function, which doesn't fit to our case.

Impact Substitution. A perhaps intuitively intriguing property of the value function that one might consider could be that the impact of posterior mean and precision might be substitutes of each other. For a belief that the posterior mean has a high impact on the value function, precision of that belief may have a lower impact and *vice versa*. Usually, substitution effects are presented with the sub-modularity condition that involves cross-partial derivatives of the function, but since

in our differential equation no such term exists, one might impose a much more restrictive, linear, relationship among the first derivatives like $\frac{\partial u}{\partial m} + \frac{\partial d}{\partial \tau} = c$ for some constant c. Then the differential equal can be expressed as

$$u(m,\tau) = m + \left(c - \frac{\partial u(m,\tau)}{\partial m}\right) - \frac{1}{\tau^2}\frac{\partial^2 u(m,\tau)}{\partial m^2}$$

with the particular solution

$$u(m,\tau) = m + c - 1 + e^{1/2m\left(-\tau^2 - \tau\sqrt{\tau^2-4}\right)} + e^{1/2m\left(-\tau^2 + \tau\sqrt{\tau^2-4}\right)}$$

and smooth pasting conditions

$$\begin{aligned}
\frac{\partial u(s,\tau^*)}{\partial m} &= \frac{1}{2}\left(-\tau^2 - \tau\sqrt{\tau^2-4}\right) e^{1/2s(-\tau^2-\tau\sqrt{\tau^2-4})} \\
&\quad + \frac{1}{2}\left(-\tau^2 + \tau\sqrt{\tau^2-4}\right) e^{1/2s(-\tau^2+\tau\sqrt{\tau^2-4})} + 1 = 0 \\
\frac{\partial u(s,\tau^*)}{\partial \tau} &= \frac{1}{2}s\left(-\frac{\tau^2}{\sqrt{\tau^2-4}} - \sqrt{\tau^2-4} - 2\tau\right) e^{1/2s(-\tau^2-\tau\sqrt{\tau^2-4})} \\
&\quad + \frac{1}{2}s\left(\frac{\tau^2}{\sqrt{\tau^2-4}} + \sqrt{\tau^2-4} - 2\tau\right) e^{1/2s(-\tau^2+\tau\sqrt{\tau^2-4})} = 0
\end{aligned}$$

where the * sign is omitted due to notational convenience. This formulation has the advantage that it allows for the variable coefficient but prohibits a complex relationship between the partial first derivatives of the function. Parametrization also leads to additional complications.

Viscosity Solutions. A more tractable analysis is given by Keller and Rady (2020) through the one of viscosity solutions. This analysis involves establishing whether a function is a viscosity *sub(super)solution* of a partial differential equation in order to establish properties like existence, uniqueness and stability. Such solutions are considered to be *weak solutions* in that derivatives of the function may not exist in certain regions.

Strategic Solution. Here, we consider the case where agents act non-cooperatively to fulfill their self-interests. In order to establish a symmetric equilibrium, it's instructive to think about a case where each agent uses the same Markov perfect strategy $\kappa(\pi)$ except for one agent to study the unilateral deviations. In such a case, if the best response of the single agent is also to use the same strategy $\kappa(\pi)$, we obtain a symmetric equilibrium. For a strategy κ to be a best response, it has to maximize the objective function; whose value function then has to satisfy the HJB equation

$$ru(\pi) = \max_{k\in[0,1]} \{r\left[(1-k)\,s + km\right] + \left[(N-1)\kappa(\pi) + k\right]\mathcal{G}u(\pi)\}$$

where the opponents strategies and the expected future benefits of experimentation are deterministic functions of the current belief and we used again the linear scalability property of the infinitesimal generator. Rewriting this equation similar to the case in the cooperative solution we get

$$u(\pi) = s + (N-1)\kappa(\pi)b(\pi,u) + \max_{k\in[0,1]} k\left[b(\pi,u) - c(\pi)\right]$$

where the second term in the right hand side of the equation corresponds to the benefit of the player from opponents' experimentation that she acquires for free i.e. the free riding effect. Similar to before, the maximand is linear in k so it will be optimal to fully experiment when the expected benefit of experimentation is higher than the opportunity cost and to not experiment at, all *vice versa.* When the expected benefit of experimentation equals to the cost of experimentation, any choice of k will be optimal.

References

Bolton, P., & Harris, C. (1999). Strategic experimentation. *Econometrica*, *67*(2), 349–374.

Chernoff, H. (1968). Optimal stochastic control. *Sankhyā: The Indian Journal of Statistics, Series A*, 221–252.

Hörner, J., Klein, N., & Rady, S. (2022). Overcoming free-riding in bandit games. *The Review of Economic Studies*, *89*(4), 1948–1992.

Hörner, J., & Skrzypacz, A. (2017). Learning, experimentation and information design. *Advances in Economics and Econometrics*, *1*, 63–98.

Keller, G. (2009). *Brownian motion and normally distributed beliefs* (Tech. Rep.). Working Paper, University of Oxford December.

Keller, G., & Rady, S. (2010). Strategic experimentation with poisson bandits. *Theoretical Economics*, *5*(2), 275–311.

Keller, G., & Rady, S. (2015). Breakdowns. *Theoretical Economics*, *10*(1), 175–202.

Keller, G., & Rady, S. (2020). Undiscounted bandit games. *Games and Economic Behavior*, *124*, 43–61.

Keller, G., Rady, S., & Cripps, M. (2005). Strategic experimentation with exponential bandits. *Econometrica*, *73*(1), 39–68.