Social Learning and Experimentation with Heterogeneous Agents

Ilari Paasivirta*

New York University

JOB MARKET PAPER

28th November 2016

Click here for the latest version

Abstract

In our increasingly interconnected society, social learning has become an ever more important element in contexts ranging from voting decisions in elections to consumers’ product choices. This paper studies social learning in a sequential decision-making model without private information, where the individual actions of agents produce information externalities that are aggregated over time. We model individual decisions as continuous time two-armed bandit problems, where players learn about two unknown states: the idiosyncratic type of an agent and the type of the risky arm, which is common across all players. Both types influence observed outcomes, which give rise to non-monotone posterior belief paths, which are absent in one-dimensional settings—an environment that the existing literature largely focuses upon. We derive closed-form solutions for the value functions of the agent’s decision problem and characterise all the possible sequences that arise in the sequential learning environment. We also analyse welfare outcomes and find that, in the absence of successes, welfare has a downward trend but follows a wave-like path, which is absent in the planner’s solution, which maximises social welfare.

Keywords: Social learning, observational learning, herding, two-armed bandit, experimentation, Poisson process, optimal stopping.

JEL Codes: C61, C73, D81, D83

*I am indebted to Ennio Stacchetti for his support and guidance, and to Boyan Jovanovic, Samuel Lee, and Ilan Lobel for continued discussions and advice. For helpful discussions and suggestions, I also thank Ala Avoyan, Laurent Cavenaile, Dániel Csaba, Pierre Germain, Seher Gupta, Siddharth Hari, Pavel Kocourek, Xiaoye Liao, Elliot Lipnowski, Erik Madsen, Laurent Mathevet, Andrew McClellan, Keith O’Hara, Hanna Paasivirta, Juhana Siljander and Chen-Hung Wu. I also thank all the seminar participants at NYU for their feedback and useful comments. I bear responsibility for any remaining errors.

E-mail Address: ilari@nyu.edu
1 Introduction

Consider the following stylised example, which captures the main features of our model. Suppose a restauranteur is considering opening a restaurant at some specific location, where a previous restaurant has just gone out of business. The question of why the preceding restaurant failed is pertinent to the restauranteur’s decision problem. If the previous restaurant failed because the location is simply an objectively bad location for a restaurant, then regardless of its quality, any restaurant is doomed to failure if they open at the location. In contrast, if the previous restaurant failed due to idiosyncratic reasons, such as the quality of its food or poor management, then such a failure should not necessarily discourage the restauranteur from opening at the location. However, the underlying reasons for the failure—objective or idiosyncratic—are inaccessible to the restauranteur. Now suppose that in fact the restauranteur has observed a sequence of different restaurants opening and eventually failing at the location. Intuitively, the restauranteur should become more pessimistic about the probability of success at the location.

In the model of this paper, each agent can choose to learn about a risky arm yielding stochastic lump-sum rewards, or choose a safe arm which yields a known fixed payoff. In the preceding example, the restaurant plays the role of the agent, and the location describes the risky arm. The example captures how the type of the agent influences how informative that agent’s experiences are to others. If the agent is a low type, then he is doomed to failure, which means that the lack of success is uninformative about the probability of success of the subsequent agent. Only when an agent is of a high type, and hence has at least a chance of succeeding, does the lack of success suggest that the arm is bad. If the arm is indeed bad, then all agents would prefer to choose the safe arm. However, the types of the preceding agents and the type of the arm are unknown to all, and can only be inferred from observations of success, or the lack thereof. Learning about the agent’s type influences whether the lack of success should be interpreted as informative or uninformative about the risky arm’s type.
This paper analyses social learning in a model where actions of agents generate public information about unknown states, as is typical in the social learning literature. Yet, unlike typical settings, the novelty here is that we assume that information is generated entirely from the publicly observable learning of the unknown types, rather than each agent being endowed with private information, which gets aggregated through publicly observable actions. Thus, the model contains no private information.

One contribution of the model is that it introduces rich learning dynamics that have non-monotonic posterior belief paths where the information externality from an agent’s action decrease over time. The decreasing information externality influences how social welfare evolves over time. The planner would want to shorten the tenure of any agent, and replace them as fast as possible if there were no entry costs. In the absence of successes along a sequence of agents, in the planner’s solution, welfare would decline monotonically. In contrast, when agents are free to choose their exit times, welfare changes non-monotonically over time following a wave-like pattern with a downward drift.

Because the agent’s type is inextricably linked with the arm type, the learning dynamics will involve two-dimensional beliefs. This introduces technical complications given that models with multidimensional state spaces are typically intractable and must be analysed numerically. However, we are able to solve the continuous time decision problem of the agent in closed form, which makes subsequent analysis of the sequential problem easier. Hence, the technical contribution of this paper is to provide a method for solving a multidimensional continuous time optimal stopping problem where, specifically, the state process has trajectories with persistent jump discontinuities. The persistent jumps reflect the fact that learning never ceases in finite time in our model.

However, in order to introduce the minimum amount of complexity needed to analyse the aforementioned setting, we assume that successes on the risky arm perfectly reveal the agent’s type, and imperfectly reveal the type of the arm. One implication of this is that the posterior belief paths about the risky arm are non-monotonic. These non-monotonocities
are not present in settings where successes are perfectly revealing about both types—a setting which, nevertheless, is a limiting case of the model in this paper.

The assumption of perfectly revealing news about the agent’s type allows us to split the problem into two stages, where we consider the agent’s decision problem before and after observing a success on the risky arm. After a success, the second stage of the problem becomes one-dimensional since only arm type uncertainty remains. We can use the solution to the second stage problem to simplify the first stage problem where there is uncertainty about both types. Essentially, this reduces the problem to solving a partial differential equation (PDE) instead of having to solve a partial functional differential equation (PFDE). Given that the resulting PDE is first-order and linear, the simplified problem proves to be amenable to analytic methods.

A technical extension would be to assume that successes are imperfectly revealing about both types. In this case however, the simplification method used in this paper would not be applicable, and it would be necessary to analyse the PFDE directly, or to use an entirely different approach to solving the problem.

The rest of the paper is organised as follows. Section 2 briefly surveys the related literature. Section 3 sets up the model and section 4 analyses the decision problem of the agent, which forms the core part of the paper. Section 5 focuses on the social learning problem and analyses how the decision problems of all the agents are related. Section 7 discusses possible extensions.

2 Related Literature

This paper primarily stands between two existing bodies of literature. The overall sequential entry-exit structure of agents is related to the social learning literature starting with herding models of Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992), while the
decision problems faced by each agent along this sequence are most closely related to the strategic experimentation literature starting with Bolton and Harris (1999); Keller, Rady, and Cripps (2005). We highlight some of the similarities and key differences in the following subsections.

2.1 Social Learning

In the social learning literature,\(^1\) the basic structure typically describes a sequence of agents who choose an action from some finite choice set. Each receives a private signal about the underlying state of the world, where each agent also observes all the previous actions taken earlier on in the sequence. The flavour of many of the results in the social learning literature is similar to ours, in that we characterize when herding occurs and the subsequent welfare losses.

However, the decision problems faced by each agent are clearly distinct from the standard herding models. Firstly, this paper models the agent’s problem as a continuous time optimal stopping problem, and so the timing of the decisions is endogenous. Although herding papers typically model decisions as one-time static decisions, several studies\(^2\) have considered endogenous timing of decisions in various set-ups.

Secondly, the agent has partial control over the signals, given that the agent chooses whether to undertake experimentation which then generates binary outcomes that are publicly observable. One important difference is that we assume that experimentation is a costly activity and so agents do not automatically receive informative signals like in the herding models. Hence, it is not meaningful to talk of information cascades since signals concerning the risky

\(^1\)This literature has grown very large and has branched into several directions since Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992). However, as this paper only shares some of the more general features the herding literature, we do not provide a detailed review. For a more comprehensive overview of this literature, see for instance Vives (2010). Also, see Smith and Sorensen (2000), which analyses more broadly the specific modelling assumptions that enable herding to arise.

\(^2\)See, for example, Chamley and Gale (1994); Faruk Gul (1995); Gale (1996); Murto and Välimäki (2011)
arm cease entirely as soon as agents herd on the safe arm. Thus, there is no wasted information that somehow fails to get aggregated into public information after herding occurs.

More generally, our model does not feature any private information. Thus, the question of how private information is aggregated into public information, which is the central focus of the herding literature, is mute. Nevertheless, it would be possible to introduce private information into our model by assuming, for instance, that agents are endowed with some private signal when they begin their decision problem. However, since the exit times lie in a continuous space, agents would be able to infer the private signals and reward outcomes of other agents from their exit timing decisions. As a result, this paper abstracts entirely away from private information and focuses on observational learning based solely on public information.

### 2.2 Experimentation

The agent’s problem in this paper can be seen as a multidimensional version of the Poisson bandit models in Keller, Rady, and Cripps (2005) and Keller and Rady (2010). In these models, a finite number of agents simultaneously choose between a risky choice with an unknown stochastic payoff, and a safe choice which yields a fixed known payoff. Agents experiment on the risky arm and update their beliefs in a Bayesian manner based on the reward realisations. The focus in these studies is on the strategic aspects of experimentation and the subsequent free-riding effects that are created by the information externalities of experimentation. Such aspects are not present in this paper given that the sequential structure entails that only one agent faces a decision problem at a time, and so agents cannot benefit from the actions of future agents. Nevertheless, the solution to the planner’s problem in Keller and Rady (2010) constitutes one key component of the agent’s problem in this paper.

The multidimensional nature of our model stems from the fact that each agent has beliefs over both arm and agent types. The dearth of papers that consider environments that allow
for multidimensional types, is perhaps partly due to the considerable technical complications that additional dimensions introduce. However, allowing for a multidimensional state space can give raise to non-monotone posterior beliefs which are not present in one-dimensional models.

In our paper, the absence of reward realisations first causes an outside observer to become more pessimistic with regard to the arm type. Nevertheless, after some time, the lack of rewards will begin to induce increasing optimism concerning the arm’s type. In other words, the absence of rewards is **bad news** at first, but after some point in time the absence of rewards becomes **good news**. This stands in contrast to the existing experimentation literature using exponential bandits, where the absence of signals from the risky arm is assumed to be either good news or bad news—but not both. For instance, in Keller, Rady, and Cripps (2005); Bonatti and Hörner (2011), lack of arrivals from the risky arm is always bad news, since these random events are assumed to be lump-sum rewards similarly to our model. In contrast, in Keller and Rady (2015); Bonatti and Hörner (2016) the lack of arrivals is good news since the random events are assumed to be lump-sum costs. Other papers, such as Simon Board (2013); Che and Hörner (2015); Frick and Ishii (2016), explicitly compare the two news varieties and analyse their effects on incentives in their respective settings.

Some recent papers concerning strategic experimentation allow the risky arms to have type heterogeneity across players by assuming that the types are correlated rather than being identical for all players.\(^3\) This renders the problem multidimensional given that players must now have beliefs over the arms of all players. Ostensibly, it might seem that the model of this paper could be cast as a sequential version of the models with positive correlation. However it should be noted that in our setting, the **reward processes**\(^4\) across players are uncorrelated,

---

\(^3\)See for instance, Klein and Rady 2011; Rosenberg, Salomon, and Vieille 2013

\(^4\)In our model, the Poisson process describing the reward realisations is determined by a single Poisson parameter $\lambda$. In turn, the Poisson parameter for each agent is determined by a product of two random variables $\lambda = \lambda_i \cdot \lambda_c$, where $\lambda_i$ is i.i.d. across players and $\lambda_c$ is identical and hence perfectly correlated across players. Thus the composite random variables $\lambda$ faced by players are uncorrelated but dependent across players.
although they are not independent. This is because nature draws a fixed but unknown arm type at time zero, which influences the reward processes of all players. However, agent types are i.i.d. over the sequence. This means that the beliefs and their dynamics are fundamentally different in our setting and the setting where risky arms are correlated. Also, in our model, beliefs are not over reward rates per se, but over the underlying parameters that determine the reward rate for each arm.

From a methodological point of view, our paper shares similarities with Klein and Rady (2011), which studies a strategic game with negative correlation. Specifically, in the case of imperfect negative correlation\(^5\) when rewards are perfectly revealing about the arm type, the authors construct an equilibrium by solving partial differential equations which yield closed-form solutions up to some integrating constant. In our setting, we allow for rewards to be imperfectly revealing about arm types which means that learning never ceases in finite time. We are able to derive closed-form solutions for the agents’ value functions, as well as explicitly deriving a decision boundary where any agent switches from the risky to the safe arm.

3 Model

We consider a continuous time model of learning, where the baseline setup closely follows Keller, Rady, and Cripps (2005); Keller and Rady (2010). Agents have a common discount rate \(r > 0\), and face a two-armed Poisson bandit problem. Each agent can choose between a safe arm that yields a known flow payoff of zero, or a risky arm which yields rewards at random times according to a homogeneous Poisson point process. The agent has to pay a flow cost of \(c > 0\) in order to use the risky arm.

Departing from the set-ups of Keller, Rady, and Cripps (2005); Keller and Rady (2010), we assume that the reward processes can differ across players. More precisely, we assume that

\(^5\)See section 5.
the parameter governing the reward process of the risky arm is given by

\[ \lambda = \lambda_c \cdot \lambda_i, \] (1)

where \( \lambda_i \in \{0, 1\} \), and \( \lambda_c \in \{\lambda_L, \lambda_H\} \), where \( 0 < \lambda_L < \lambda_H \). While the parameters \( \lambda_i \) and \( \lambda_c \) are unknown and fixed, the supports are common knowledge. Thus, if agent \( i \) plays the risky arm, it will yield a reward at a rate \( \lambda_c \cdot \lambda_i \). The parameters \( \lambda_c \) and \( \lambda_i \) are random variables where the key difference is that \( \lambda_c \) is drawn once and is common to all agents, whilst \( \lambda_i \) is drawn i.i.d. for each agent separately.

If a reward is realised on the risky arm, it is drawn i.i.d. from some fixed distribution with a known mean \( h > 0 \). To avoid trivial cases, we assume that \( \lambda_H h - c > 0 > \lambda_L h - c \). This implies that, conditional on the agent being a high type, the expected flow payoff from the risky arm is strictly higher than the safe arm. Similarly, conditional on the agent being a low type, the safe arm has a strictly higher expected value. For ease of reference, we summarise these key assumptions below.

**Assumption 1.** *(Payoff Parameters)* We assume that \( \lambda_H h - c > 0 > \lambda_L h - c \).

**Assumption 2.** *(Learning Parameters)* Agent types take values \( \lambda_i \in \{0, 1\} \). Arm types take values \( \lambda_c \in \{\lambda_L, \lambda_H\} \), where we assume that \( 0 \leq \lambda_L < \lambda_H \).

All agents share common priors \( p_{1,0} := Pr[\lambda_c = \lambda_H] \) and \( p_{2,0} := Pr[\lambda_i \neq 0] \). All agents update their beliefs in a Bayesian fashion when observing the realisations of the reward process. It turns out to be convenient to model the belief dynamics using an unconditional belief \( p_{1,t} \), and a conditional belief \( p_{2,t} \). More precisely, we use the following beliefs as state variables:

\[ p_{1,t} := Pr[\lambda_i \neq 0 | \mathcal{F}_t] \]

\[ p_{2,t} := Pr[\lambda_c = \lambda_H | \mathcal{F}_t, \lambda_i \neq 0]. \]
Conditional on the information $\mathcal{F}_t$, the expected value of the Poisson parameter is given by

$$E_t \lambda = \lambda(p_{2,t}) \cdot p_{1,t}$$

where we define the function $\lambda(p) := p\lambda_H + (1 - p)\lambda_L$.

At each instance of time $t$, an agent chooses $k_t \in [0, 1]$, which indicates the fraction of resources devoted to the risky arm. Hence, choices are described by $\{k_t\}_{t \geq 0}$, which is measurable with respect to filtration $\mathcal{F}_t$, generated by the publicly observable reward process up to time $t$.

We restrict our attention to Markovian strategies$^6$ $k_t : [0, 1]^2 \rightarrow [0, 1]$, where we impose standard regularity conditions on the strategies so as to guarantee that the resulting laws of motion for the beliefs are well-defined. We assume that the strategies are left-continuous and piece-wise Lipschitz-continuous, where the state variables are assumed to be the left limiting beliefs $\lim_{\delta \downarrow 0} (p_{1,t-\delta}, p_{2,t-\delta})$. This simply means that at the instance of a payoff realization, strategies depend on the beliefs an instant before time $t$.

The discounted expected payoff to a risk-neutral agent is given by

$$\mathbb{E}\left[\int_0^\infty re^{-rt}k_t [h\lambda(p_{2,t}) \cdot p_{1,t} - c] \, dt\right],$$

where we multiply the integrand by $r$ to express expected payoffs in per-period units. If the agent’s type, $\lambda_i$, is equal to zero, then that agent will never observe a reward. As the type of an agent can only take two values $\{0, 1\}$, it is clear that if an agent observes a reward, then we can be certain that the agent is a high type. Hence, rewards are perfectly revealing about the agent’s type. In contrast, the arm type, $\lambda_c$, is restricted to taking values which are strictly positive. Hence, reward observations yield only imperfect information about the true arm type, and learning never stops in finite time and continues as long as the risky arm

$^6$Given that we do not have to worry about strategic considerations, this assumption is not restrictive.
is being played.

It is useful to analyse separately the cases before and after a reward realisation, since the former case involves beliefs concerning both agent and arm types, whilst the latter case is reduced to solely tracking beliefs concerning the risky arm. We refer to the phase where there is uncertainty about both types as *Stage I*, and when only arm uncertainty remains, we refer to this phase as *Stage II*.

### 3.1 Belief Dynamics: Stage I

If the risky arm does not yield a reward in the time interval \([t, t + dt]\), then each agent will update their beliefs in a Bayesian manner as follows: The updated belief that the risky arm is good, conditional on the agent being a high type, is given by

\[
p_{2,t+dt} = \frac{(1 - k_t \lambda_H dt)p_{2,t}}{(1 - k_t \lambda_H dt)p_{2,t} + (1 - k_t \lambda_L dt)(1 - p_{2,t})}.
\]

The updated unconditional belief that the agent is a high type is given by

\[
p_{1,t+dt} = \frac{(1 - k_t \lambda(p_{2,t})dt)p_{1,t}}{(1 - k_t \lambda(p_{2,t})dt)p_{1,t} + (1 - p_{1,t})}.
\]

Thus, the evolution of the belief can be described in differential form as

\[
dp_{2,t} = -k_t(\lambda_H - \lambda_L)p_{2,t}(1 - p_{2,t})dt \tag{2}
\]

\[
dp_{1,t} = -k_t \lambda(p_{2,t})p_{1,t}(1 - p_{1,t})dt \tag{3}
\]

When the risky arm yields a reward, the updated beliefs cause \(p_{1,t}\) to jump to one, and the
jump in $p_{2,t}$ is given by

$$
\lim_{\Delta t \to 0} p_{2,t+\Delta t} = \frac{k_t p_{2,t} \lambda H \Delta t}{k_t p_{2,t} \cdot \lambda H \Delta t + k_t (1 - p_{2,t}) \cdot \lambda_L \Delta t} = \frac{\lambda_H}{\lambda(p_{2,t})} p_{2,t}
$$

The beliefs $p_{1,t}$ and $p_{2,t}$ serve as the state variables of an agent’s optimisation problem, which we analyse in the next section. The unconditional belief that the risky arm is good is denoted by $p_{2,t}^{UC}$, and it can be expressed in terms of the state variables $p_{1,t}$ and $p_{2,t}$, and the prior belief $p_{2,0}$.

$$
p_{2,t}^{UC} = p_{2,t} p_{1,t} + p_{2,0} (1 - p_{1,t}) \tag{4}
$$

Although abstracting away from time and casting the model solely in terms of beliefs makes solving the model easier, the belief dynamics are best understood by considering their evolution over time. In Figure 1 we plot the unconditional and conditional beliefs that the risky arm is good. If there are no reward observations both $p_{1,t}$ and $p_{2,t}$ will drift down towards zero over time. Thus, $t \to \infty$ implies $p_{2,t}^{UC} \to p_{2,0}$. In other words, if there are no reward realisations on the risky arm, the belief $p_{2,t}^{UC}$ will eventually return back to the prior $p_{2,0}$ in the limit. This captures the intuition that, if no rewards are realised over a long period of time, it becomes more apparent that the agent is a low type, since the rewards should be eventually observed on the risky arm, even if it is a bad arm. As a result, it becomes evident that the lack of reward realisations has been completely uninformative about the arm type, given that a low type agent never receives any rewards regardless of the arm type. Thus, the beliefs concerning the risky arm return back toward the prior $p_{2,0}$, which summarises the information before the current agent started playing the risky arm. We can see these dynamics in figure 1, where we plot the posterior beliefs when nothing has been observed on the risky arm. In particular, we see the non-monotone nature of the belief $p_{2,t}^{UC}$, which starts at the prior $p_{2,0} = 0.6$, and eventually returns back to it, at as $t \to \infty$. 

12
Figure 1: Belief dynamics when there are no reward realisations on the risky arm. Here we assume parametric values $\lambda_H = 2$, $\lambda_L = \frac{1}{2}$, with priors $p_{1,0} = 0.9$ and $p_{2,0} = 0.6$. The unconditional, and conditional probabilities that the risky arm is good are given by $p_{2,UC}^t$, while $p_1$ is the unconditional probability that the agent is a high type.

In the following lemma, we solve the belief paths as functions of time, which allows us to subsequently characterise the belief dynamics more precisely.

**Lemma 1.** Given the prior beliefs $(p_{1,0}, p_{2,0})$, the belief dynamics for the case when there are no reward observations can be expressed as deterministic functions of time as follows:

\[
\begin{align*}
    p_{2,t} &= \left[1 + \left(1 - \frac{p_{2,0}}{p_{2,0}}\right) e^{k(\lambda_H - \lambda_L)t}\right]^{-1} \\
    p_{1,t} &= \left[1 + \frac{\left(1 - p_{1,0}\right)}{p_{2,0}} \frac{e^{k\lambda_H t}}{p_{1,0}} + \left(1 - p_{2,0}\right) e^{k(\lambda_H - \lambda_L)t} \right]^{-1} \\
    p_{2,UC}^t &= \frac{1}{1 + \left(1 - p_{1,0}\right) e^{k\lambda_H t}} + \left(1 - p_{2,0}\right) e^{k(\lambda_H - \lambda_L)t}
\end{align*}
\]

**Proof.** We solve the ordinary differential equations defined in equations 2 and 3. By first solving equation 2 for $p_{2,t}$, and then plugging this solution into 3, we can solve it for $p_{1,t}$.

**Remark 1.** The belief $p_{1,t}$ can also be expressed explicitly as a function of $p_{2,t}$:

\[
p_{1}(p_{2,t}) = \left[1 + \left(1 - \frac{p_{1,0}}{p_{1,0}}\right) \left(1 - \frac{p_{2,t}}{p_{2,0}}\right) e^{k(\lambda_H - \lambda_L)t} \left(\frac{p_{2,t}}{p_{2,0}}\right)^{-\frac{\lambda_H}{\lambda_H - \lambda_L}} \right]^{-1}
\]
The non-monotonic paths of $p_{2,t}^{UC}$ can only arise when $\lambda_c > 0$, which means that there is imperfect learning about the risky arm’s type. If $\lambda_c = 0$, whereby rewards would be perfectly revealing about the arm type, then in the absence of rewards the agent would grow increasingly pessimistic about both the arm and agent type in a monotone fashion.

Proposition 1. Taking as given that assumptions 1 and 2 hold, suppose that the belief dynamics are given in lemma 1, and that there have been no reward observations on the risky arm. We specify the following two properties:

1. $p_{2,t}^{UC}$ has a unique interior global minimum $t^* \in (0, \infty)$, where $t^*$ depends on the prior $p_{1,0}$, but is independent of $p_{2,0}$.

2. $p_{2,t}^{UC}$ is monotone decreasing on interval $[0, t^*)$, and monotone increasing on interval $[t^*, \infty)$.

Then, $p_{2,t}^{UC}$ satisfies properties (1) and (2) below if, and only if, $\lambda_L > 0$.

Proof. The derivative of $p_{2,t}^{UC}$ is given by

$$
\frac{dp_{2,t}^{UC}}{dt} = -kp_{1,0}(1 - p_{2,0})p_{2,0}e^{k(\lambda_H + \lambda_L)t} \left( \lambda_H(1 - p_{1,0})e^{k\lambda_L t} - \lambda_L(1 - p_{1,0})e^{k\lambda_H t} + p_{1,0}(\lambda_H - \lambda_L) \right)
$$

$$
\left( (1 - p_{1,0})e^{k(\lambda_H + \lambda_L)t} + p_{1,0}(1 - p_{2,0})e^{k\lambda_H t} + p_{1,0}p_{2,0}e^{k\lambda_L t} \right)^2
$$

It is easy to see that the first-order condition depends only on the numerator, and it is satisfied when

$$
\lambda_L e^{k\lambda_H t} - \lambda_H e^{k\lambda_L t} = (\lambda_H - \lambda_L) \left( \frac{p_{1,0}}{1 - p_{1,0}} \right)
$$

First observe that the right-hand side is constant. Thus, if $\lambda_L = 0$, the first-order condition is never satisfied and we have $\frac{dp_{2,t}^{UC}}{dt} < 0$ for all $t \geq 0$. Hence, there is no minimum $t^*$.

Now suppose that $0 < \lambda_L < \lambda_H$. Observe that the left-hand side is strictly increasing, where it is negative at $t = 0$, and grows without bound as $t \rightarrow \infty$. This guarantees the existence and uniqueness of an interior $t^*$ that satisfies the first-order condition. Furthermore, note
that the first-order condition 6 is independent of $p_{2,0}$, and is only influenced by $p_{1,0}$. Also, it is clear that the sign of the derivative switches from negative to positive as $t$ crosses $t^*$, which implies that $t^*$ is the global minimum and conditions (1) and (2) are satisfied.

Firstly, it is worth emphasising here that the non-monotonic belief dynamics do not rely on the assumption that rewards are perfectly revealing about one type and imperfectly revealing about the other. The non-monotonicities, do however, rely on having imperfect learning concerning at least one of the types. If rewards gave imperfect information about both types, then non-monotonicities could still arise. Nonetheless, even in this case, only one of the posterior beliefs would exhibit behaviour.

3.2 Belief Dynamics: Stage II

Once a reward has been observed, all uncertainty about the agent’s type is resolved, and the only remaining uncertainty pertains to the risky arm. Given that observing rewards is more likely when the risky arm is good, beliefs jump upward every time a reward is observed. Between reward realisations, beliefs will continuously drift down in a deterministic fashion. If the agent plays the risky arm forever, then the arm type is learned in the limit with probability 1. Figure 3.2 gives an example of how beliefs might evolve if rewards are realised at a rate which is lower than would be expected if the true arm state was of type $\lambda_H$. Hence, beliefs trend downward despite the occasional upward jump at times when reward is observed.

4 Agent’s Decision Problem

Solving the agent’s decision problem is made easier by choosing the state space judiciously. It might seem odd that we choose both a conditional and an unconditional belief as our state variables. Essentially, conditioning on the agent being a high type, makes it easier
Figure 2: An example of a realised posterior belief path concerning the arm type. Here, after the agent has observed a reward, he enters Stage II with the belief $p_2 = 0.8$, and subsequently observes lump-sum rewards at times $t = 1$ and $t = 3$. To capture the value of observing a reward for the first time. At the moment of the first reward realisation, the belief concerning the agent type $p_1$ jumps to one, and the conditional belief $p_2$ concerning the arm type is the only relevant state variable thereafter. This structure allows us to solve the problem by backward induction, where we begin by solving the decision problem after the first success has already been observed. We refer to this as Stage II of the decision problem. We can then use the Stage II solution to solve the problem for the case when there have been no reward realisations. We refer to this as Stage I of the decision problem. This is made possible since the value function of Stage II can be solved in closed form, which allows us to directly simplify the Stage I problem given that we now know the value of reaching Stage II from any state.

### 4.1 Stage II Problem

The solution to the Stage II problem is virtually identical to the cooperative problem in Keller and Rady (2010, Section 3). The only difference is that we have normalised the value of the safe arm to zero, but included a flow cost $c > 0$ to using the risky arm. These are equivalent formulations, but it will ease our later analysis to explicitly include a variable for costs. For completeness, we outline the Stage II problem and its solution in an informal...
manner below, where more details can be found in Appendix A.\footnote{For a more complete analysis, see Keller and Rady (2010)}

The recursive formulation for the Stage II dynamic programming problem is given below

\[
V(p_2) = \max_{k \in [0,1]} \{rk \ [h\lambda(p_2) - c] \ dt + (1 - rdt)E[V(p_2 + dp_2)|p_2]\}
\]

where at each moment in time, the agent chooses the fraction \( k \) of how to allocate his time between the safe and risky arm.

This yields the following Hamilton-Jacobi-Bellman (HJB) equation

\[
V(p_2) = \max_{k \in [0,1]} k \cdot \left\{ h\lambda(p_2) - c + \frac{1}{r} [\lambda(p_2) \cdot [V(j(p_2)) - V(p_2)] - V'(p_2) (\lambda_H - \lambda_L) (1 - p_2)p_2] \right\}
\]

Given that the HJB equation is linear in the intensity of experimentation \( k \), the optimal decision is characterised by a threshold rule, where \( k \) takes binary values \( k \in \{0,1\} \). When the right-hand side of the equation 7 is non-negative, \( k = 1 \), and otherwise \( k = 0 \). Since the value of the outside option is zero, it is clear that \( V(p_2) = 0 \), when \( k = 0 \). If it is optimal to play the risky arm whereby \( k = 1 \), the HJB equation yields the functional difference equation (FDE)\footnote{With a change of variables, the differential equation can be transformed into a delay differential equation (DDE) where the arguments of \( V \) differ at most by a fixed constant. These types of differential equations are also sometimes referred to as retarded differential equations (RDEs) or differential-difference equations (for more details, see for instance Bellman and Cooke (1963), or for a more recent introduction, see Smith (2010)).} given below.

\[
V(p_2) = h\lambda(p_2) - c + \frac{1}{r} [\lambda(p_2) \cdot [V(j(p_2)) - V(p_2)] - V'(p_2) (\lambda_H - \lambda_L) (1 - p_2)p_2] \tag{8}
\]

The jump term \( j(p_2) \) term renders the differential equation non-local, and so standard ODE techniques do not apply. Nevertheless, a guess and verify approach can still be applied to solve the FDE, which is summarised in the proposition below and can be found in Keller and Rady (2010).
Proposition 2. (Keller and Rady, 2010) The single agent solution to the Stage II decision problem is described by a threshold rule, where the agent plays the risky arm with maximum intensity \((k = 1)\) for beliefs \(p_2 \geq p_2^{**}\), and plays the safe arm exclusively for beliefs below this threshold. The threshold belief is given by

\[
p_2^{**} = \frac{\mu (c - h\lambda_L)}{h\lambda_H - c + h\mu(\lambda_H - \lambda_L)}
\]

where \(\mu\) is the positive root of the following equation

\[
r + \lambda_L - \mu (\lambda_H - \lambda_L) = \lambda_L \left( \frac{\lambda_L}{\lambda_H} \right)^\mu
\]

For beliefs \(p_2 \geq p_2^{**}\), the agent’s value function is given by

\[
V(p_2) = (\lambda(p_2)h - c) + (c - \lambda(p_2^{**})h) \left( \frac{1 - p_2}{1 - p_2^{**}} \right) \left( \frac{\Omega(p_2)}{\Omega(p_2^{**})} \right)^\mu
\]

where \(\Omega(p_2) := \frac{1-p_2}{p_2}\). \(V(p_2) = 0\) for beliefs below the threshold \(p_2^{**}\).

Given that we have a closed form solution for \(V\), it is straightforward to show properties of the value function.

**Corollary 1.** For all \(p_2 \geq p_2^{**}\), the value function \(V\) is a non-negative function, which is increasing and convex in \(p_2\).

**Proof.** The second derivative of \(V\) is given by

\[
V''(p_2) = -\frac{\mu(\mu + 1) \left( \frac{1-p_2}{p_2} \right)^\mu \left( \frac{1-p_2^{**}}{p_2^{**}} \right)^{-\mu} (h\lambda(p_2^{**}) - c)}{(1-p_2)p_2^2(1-p_2^{**})}
\]

Given the Stage II threshold belief \(p^{**}\) defined in Equation 9, \(\mu > 0\), and the parametric
restrictions of Assumptions 1 and 2, the following inequality holds

\[ h\lambda(p_2^*) - c = \frac{(h\lambda_H - c)(h\lambda_L - c)}{\mu h(\lambda_H - \lambda_L) + h\lambda_H - c} < 0 \]  

(11)

Thus, \( V'' > 0 \) and hence \( V \) is convex.

By the smooth pasting condition \( V'(p_2^*) = 0 \), and given that \( V \) is convex, we have \( V'(p_2) \geq 0, \forall p_2 \geq p_2^* \). Since \( V'(p_2) \geq 0 \) and by value matching we have \( V(p_2^*) = 0 \), this shows that \( V(p_2) \) for all \( p_2 \geq p_2^* \).

### 4.2 Stage I Problem

The solution to the Stage II problem captures the value of playing a risky arm assuming that the agent is a high type. We can use this value as one component in the Stage I problem where there is uncertainty about both types. One substantial benefit of being able to use the Stage II solution is that this eliminates the need to deal with a partial functional differential equation (PFDE). As a result, the problem is reduced to solving a linear first-order PDE, where standard methods can now be applied. We outline the key steps below, where details can be found in Appendix A.

The recursive formulation of the agent’s dynamic programming problem is given below, where the state space consists of two beliefs.

\[
U(p_1, p_2) = \max_{k \in [0,1]} \{ r k \left[ \lambda(p_2)p_1 h - c \right] dt + (1 - r dt) E \left[ U(p_1 + dp_1, p_2 + dp_2) \mid p_1, p_2 \right] \}
\]

This yields the following HJB equation⁹

⁹For a more detailed derivation of the HJB equation, see Appendix A.

\[
\begin{align*}
U(p_1, p_2) &= \max_{k \in [0,1]} \left\{ \lambda(p_2)p_1 h - c + \frac{1}{r} \left\{ \lambda(p_2)p_1 [U(1, j(p_2)) - U(p_1, p_2)] \ldots \\
& \quad \ldots - \left[ \frac{\partial U(p_1, p_2)}{\partial p_1} \cdot \lambda(p_2)p_1(1 - p_1) + \frac{\partial U(p_1, p_2)}{\partial p_2} \cdot (\lambda_H - \lambda_L)p_2(1 - p_2) \right] \right\}
\end{align*}
\]
Similarly to before, as the HJB equation is linear in $k$, the solution will exhibit a bang-bang property where $k$ takes values from $\{0,1\}$. If the safe arm is optimal, then $k = 0$ and $U(p_1, p_2) = 0$. If the risky arm is optimal, then $k = 1$ and after re-arranging the HJB equation, the problem yields the following partial functional differential equation:

$$
\frac{\partial U(p_1, p_2)}{\partial p_1} \cdot \lambda(p_2)p_1(1 - p_1) + \frac{\partial U(p_1, p_2)}{\partial p_2} \cdot (\lambda_H - \lambda_L)p_2(1 - p_2) + [r + \lambda(p_2) \cdot p_1] U(p_1, p_2) = \\
r [\lambda(p_2) \cdot p_1 h - c] + \lambda(p_2) \cdot p_1 \cdot U(1, j(p_2)) \quad (13)
$$

First note that $U(1, j(p_2))$ corresponds to the value of entering Stage II, with the updated beliefs $(p_1, p_2) = (1, j(p_2))$. Given that we have a closed-form solution for the Stage II problem, we can simply replace $U(1, j(p_2))$ with the known function $V(j(p))$ given in Proposition 2. Note that this eliminates the non-localities from the DE, and as a result the PFDE simplifies into a first-order linear PDE. In Appendix A, we derive the general closed-form solution for $U$ using the Method of Characteristics.\(^\text{10}\) By imposing value matching and smooth pasting conditions, we can find a particular solution for $U$. Note that if the rewards were not perfectly revealing about the agent’s type, we would not have been able to solve the problem using the strategy summarised above. The fact that the agent’s problem becomes one-dimensional after a reward realisation allows us to use the backward induction approach and simplify the DE. If rewards were imperfectly revealing about both states then some other solution method would be necessary.

We summarise the solution to the agent’s decision problem in the proposition below.

**Proposition 3.** The single-agent decision problem is characterised by a threshold strategy

$$
k^*(p_1, p_2) = \begin{cases} 
1 & (p_1, p_2) \in \mathcal{D} \\
0 & \text{otherwise}
\end{cases}
$$

\(^{10}\)For details of this method, see for instance Ch.1. of John (1978).
The continuation region $\mathcal{D}$ is defined by the epigraph

$$
\mathcal{D} := \{(p_1, p_2) | p_1 > p_1^*(p_2), p_2 > p_2^*\}
$$

where the boundary $\partial \mathcal{D}$ is described by a function $p_1^*(p_2)$ given by

$$
p_1^*(p_2) = \frac{rc}{rc + V(p_2)(\lambda(p_2) + r) + (1 - p_2)p_2(\lambda_H - \lambda_L)V'(p_2)} \tag{14}
$$

Here, $V(p_2)$ is the value function of the Stage II problem. The Stage I value function $U$ is given by

$$
U(p_1, p_2) = p_1V(p_2) - c (1 - p_1) + p_1\theta(p_2) f^*(\eta(p_1, p_2))
$$

where $\theta(p_2) := p_2 \left( \frac{1-p_2}{p_2} \right)^{\frac{\lambda_H - \lambda_L}{\lambda_H - \lambda_L}}$. Value matching and smooth pasting conditions determine equations 15 and 18 respectively.

$$
f^*(\eta) = \frac{c(1 - \Gamma(p_2^*(\eta); \eta)) - \Gamma(p_2^*(\eta); \eta)V(p_2^*(\eta))}{\Gamma(p_2^*(\eta))\theta(p_2^*(\eta))} \tag{15}
$$

$$
\Gamma(p_2^*(\eta); \eta) = \left( 1 + e^{-\frac{\eta}{\lambda_H - \lambda_L}p_2^*(\eta)} \left( \frac{1 - p_2^*(\eta)}{p_2^*(\eta)} \right) \right)^{-1} \tag{16}
$$

$$
\eta = \eta(p_1, p_2) = \log \left\{ \left( \frac{(1 - p_2)^{\lambda_H}}{(p_2)^{\lambda_L}} \right) \left( \frac{p_1}{1 - p_1} \right)^{\lambda_H - \lambda_L} \right\} \tag{17}
$$

Here, $p_2^*(\eta(p_1, p_2))$ is defined implicitly as the value $p_2^*$ that, for a given $(p_1, p_2)$, satisfies the equation

$$
\Gamma(p_2^*; \eta(p_1, p_2))(V(p_2^*)(\lambda(p_2^*) + r) + (1 - p_2^*)p_2^*(\lambda_H - \lambda_L)V'(p_2^*)) = rc \left( 1 - \Gamma(p_2^*; \eta(p_1, p_2)) \right) \tag{18}
$$

Proof. See Appendix.

Remark 2. Since the above proposition guarantees existence of $U$ which satisfies the value matching and smooth pasting conditions, we can derive an equivalent expression for the
decision boundary. In Equation 13, if we set $U$ and its partials equal to zero, and solve the equation for $p_1$, we get the following expression for the boundary $\partial \mathcal{D}$ which is equivalent to the one given in Proposition 3:

$$p_1^*(p_2) = \frac{rc}{\lambda(p_2)(rh + V(j(p_2)))} \quad (19)$$

Figure 3: Stage I value function $U(p_1, p_2)$, where the decision boundary $\partial \mathcal{D}$ is marked by the solid convex curve.

The next corollaries describes the relationship between the threshold beliefs in Stage I and II. We see that the beliefs regarding the arm type at the time of choosing the safe arm, will be strictly lower in Stage II when compared with Stage I. The agent is willing to endure higher levels of pessimism regarding the risky arm if he knows that his own (agent) type is high. This is not surprising, given that the reward process is determined by the Poisson parameter $\lambda = \lambda_i \cdot \lambda_c$. In other words, the parameters describing the types enter multiplicatively into the reward process, and so pessimism with respect to one parameter can be compensated with optimism with respect to the other.

**Lemma 2.** Let $\gamma$ denote the following function

$$\gamma(p_2) := V(p_2)(\lambda(p_2) + r) + (1 - p_2)p_2(\lambda_H - \lambda_L)V'(p_2)$$
Then, $\gamma \geq 0$ and $\frac{d\gamma}{dp_2} > 0$ for all $p_2 \geq p_2^{**}$.

**Proof.** See Appendix B \qed

**Corollary 2.** The Stage I threshold belief $p_2^*(\eta)$ is strictly greater than the Stage II threshold belief $p_2^{**}$, for any $\eta \in \mathbb{R}$.

**Proof.** By Proposition 3 the Stage I threshold belief $p_2^*(p_1)$ is characterised implicitly by the equation

$$
\Gamma(p_2; \eta) (V(p_2)(\lambda(p_2) + r) + (1 - p_2)p_2(\lambda_H - \lambda_L)V'(p_2)) = rc (1 - \Gamma(p_2; \eta)) \quad (20)
$$

From Equation 16 it follows that for any constant $\eta \in \mathbb{R}$ and $p_2 \in (0,1)$, $\Gamma$ satisfies $0 < \Gamma(p_2; \eta) < 1$, and $\frac{d\Gamma}{dp_2} > 0$. Hence, at the Stage II threshold belief $p_2^{**}$, the right-hand side of the above equation is strictly lower than the left-hand side, since $V(p_2^{**}) = V'(p_2^{**}) = 0$, and $r, c > 0$. By Lemma 2 and the fact that $\frac{d\Gamma}{dp_2} > 0$, the left-hand side is strictly increasing in $p_2$. It follows that the solution $p_2^*(\eta)$ to Equation 20 must be strictly greater than $p_2^{**}$ for all $\eta \in \mathbb{R}$. \qed

**Corollary 3.** The function $p_1^*(p_2)$ defining the decision boundary $\partial D$ in Proposition 3 is a decreasing function on the interval $p_2 \in (p_2^{**}, 1)$.

**Proof.** By Lemma 2, the denominator of $p_1^*(p_2)$ is increasing in $p_2$ on the domain $[p_2^{**}, 1]$. Thus, $p_1^*(p_2)$ is decreasing in $p_2$. \qed

The decision boundary satisfies most of the comparative statics that we might expect. The boundary shifts towards the origin if we increase $r, h$, or $\lambda_H$ or if we decrease $c$. In other words, if the agent becomes more patient, or if the risky arm becomes more valuable, the agent must become more pessimistic in order for him to be induced to exit. The interesting
Figure 4: Example of a belief path in the absence of reward observations. The agent starts with the prior beliefs $p_1 = 0.9$, $p_2 = 0.9$, after which the beliefs drift towards the decision boundary.

Comparative statics relate to $\lambda_L$—the reward rate of the bad arm. Increasing $\lambda_L$ has two effects: On the one hand, the value of experimentation of the risky arm increases since the agent will experience rewards at a higher rate even if the arm is bad. On the other hand, increasing $\lambda_L$ makes it more difficult for the agent to learn the type of the arm, and so the agent is more likely to spend longer on a bad arm than when $\lambda_L$ is lower. The next result shows that a marginal increase of $\lambda_L$ decreases the agent’s expected utility on at low values of $\lambda_L$. Nevertheless, it is intuitively clear that as we increase $\lambda_L$ closer toward $\lambda_H$, the benefits should eventually dominate the costs generated though slower learning.

**Proposition 4.** If $\frac{c}{h} > r$, then there exists a threshold $\lambda_0 > 0$, such that for any $p_2$

$$
\begin{cases}
\frac{\partial p^*_1(p_2)}{\partial \lambda_L} > 0 & \text{for } \lambda_L < \lambda_0 \\
\frac{\partial p^*_1(p_2)}{\partial \lambda_L} < 0 & \text{for } \lambda_L \geq \lambda_0
\end{cases}
$$
5 Social Learning

We now consider a social learning setting where agents arrive sequentially and face the single agent decision problem described in the previous section. We assume that choosing the safe arm leads to permanent exit of the agent, after which, a new agent arrives and takes his place. Consequently, an agent has a monopoly over the risky arm from the time of entry until the time of his exit. Although agents cannot control their entry time, they have full control over when to exit. Payoff realisations and actions are publicly observable, which entails that beliefs of entering agents are fully determined by the actions and outcomes that preceded the agent’s entry.

5.1 Public Beliefs

The unconditional belief that the risky arm is good is equally relevant for all agents. Thus, we will refer to this belief as the public belief, to emphasise its role over the sequence of entries and exits. In order to see how this belief evolves over the sequence, let $t^*_n$ denote the moment that agent $n \in \mathbb{N}$ enters after the incumbent has exited. Then, the public belief $p_{2,t}^{UC}$ is given by

$$
p_{2,t}^{UC} := Pr[\lambda_c = \lambda_H|\mathcal{F}_t]
= Pr[\lambda_c = \lambda_H|\mathcal{F}_t, \lambda_i = 1] Pr[\lambda_i = 1|\mathcal{F}_t] + Pr[\lambda_c = \lambda_H|\mathcal{F}_{t^*_n}] Pr[\lambda_c = \lambda_L|\mathcal{F}_t]
= p_{2,t}p_{1,t} + p_{2,t^*_n}^{UC}(1 - p_{1,t})
$$

Here, $p_{2,t^*_n}^{UC}$ denotes the public belief at the moment of entry $t^*_n$, where $t \geq t^*_n$. Crucially, $p_{2,t^*_n}^{UC}$ incorporates all the information attained through observational learning. As time passes, the public belief $p_{2,t}^{UC}$ is given by an evolving weighted sum of $p_{2,t}$ and the prior $p_{2,t^*_n}^{UC}$, where the weighting is determined by beliefs about the agent’s type $p_{1,t}$. This captures the fact that if the agent is a low type, then that agent’s experiences are uninformative about the
arm’s type. To see this more clearly, when \( p_{1,t} = 0 \) in equation 21, we see that the public belief \( p^{UC}_{2,t} \) remains constant and equal to the prior \( p^{UC}_{2,t_n} \). In contrast, if a reward observation were to occur whereby \( p_{1,t} = 1 \), all uncertainty regarding the agent’s type disappears and so the conditional and unconditional beliefs about the arm are now equal, \( p^{UC}_{2,t} = p_{2,t} \).

Suppose agent \( n \) exits at some point in time \( t^* > t^*_n \), and at the moment of exit his beliefs were \((p_{1,t^*}, p_{2,t^*})\). Then, agent \( n + 1 \) enters and begins the decision problem with the belief \( p^{UC}_{2,t^*} = p_{2,t^*}p_{1,t^*} + p^{UC}_{2,t_n}(1 - p_{1,t^*}) \), which now captures agent \( n \’)s experiences on the risky arm in addition to all the previous agents’ experiences, which are captured in \( p^{UC}_{2,t_n} \). To keep track of the exit times and exit beliefs of the agents, we let \( t^*_{n+1} \) denote the time of entry of agent \( n + 1 \). Hence, conditional on there being no reward observations, the sequence of prior beliefs for the entering agents along the sequence is described by \( \{p^{UC}_{2,t_n}\}_{n \in I} \), which is a monotonically decreasing sequence with a finite limit point.

We can now contrast two different learning dynamics that arise in the model. First, note that if experimentation on the risky arm is cheap, agent \( n \) would play the arm for a long time and exit only when becoming very pessimistic. If, even after a long time, there were no reward realisations, \( p_{1,t} \) would approach zero where \( p^{UC}_{2,t} \) would in turn approach agent \( n \’)s prior, \( p^{UC}_{2,t_n} \). This is what gives rise to the non-monotonic beliefs which were apparent in Figure 1. Over time the distance between beliefs \( p^{UC}_{2,t} \) and \( p^{UC}_{2,t_n} \) shrinks, which reflects the fact that it becomes increasingly clear that agent \( n \’)s experiences provide little additional information over what existed before agent \( n \’)s entry, which was captured by \( p^{UC}_{2,t_n} \). Thus, if it becomes apparent that the agent is a low type, then nothing can be learned from observing his experiences on the risky arm.

Contrast this with a setting where experimentation is expensive, and so agents exit quickly unless they see immediate results from their experimentation on the risky arm. In Figure 1, this situation would correspond to a case where exit occurs at a time before \( p^{UC}_{2,t} \) starts to approach back towards the prior belief. As a result, early exit implies that there has not been enough time to determine whether the lack of reward observations is due to the agent being
a low type or the risky arm being bad. Furthermore, this implies that over the sequence of entries and exits, agents grow more pessimistic about the risky arm much faster, since there is not enough time for \( p_{2,t}^{UC} \) to return back close to its prior, which determines the prior for the next entering agent.

In order to investigate the welfare implications of some of these considerations, the next subsections characterise more the paths of play that can arise along the sequence of agents. It is clear that if the risky arm is good, then herding on the safe arm is bad for welfare. Conversely, if the risky arm is bad, then it is desirable that herding on the safe arm happens as fast as possible. These considerations are the same as in the classic exploration versus exploitation trade-off in multi-armed bandit models. Hence, any notion of welfare should capture this trade-off.

As before, we analyse the model following the backward induction approach, and look first at the Stage II setting where a reward has already been realised.

### 5.2 Stage II: Reward Observed

As we saw in Section 4, the agent who experiences a reward enters Stage II of his decision problem, which is considerably simpler as on type uncertainty about the arm remains. This means that now all observations are known to be informative about the risky arm. The next proposition shows that, if an agent nevertheless chooses to eventually exit in Stage II even after having observed a reward, then all subsequent agents herd on the safe arm.

**Proposition 5.** If an agent exits to the safe arm during Stage II of his decision problem, then all subsequent agents choose the safe arm.

**Proof.** First observe that once a reward is observed, the conditional belief \( p_2 \) equals the unconditional belief \( p_2^{UC} \). Also, if an agent exits in Stage II, the exit belief for \( p_2 \) is equal to \( p^{**} \), given in Proposition 2. Consequently, at the time of exit, the public belief \( p_2^{UC} \), which
serves as the prior for the subsequent agent, is equal to \( p^{**} \). By Proposition 2, we know that the threshold belief is always lower in Stage II than in Stage I, i.e. \( p^{**}_2 \leq p^*_2(p_1) \), for all \( p_1 \in (0, 1] \). Since \( p^*_2(p_1) \) is the threshold belief in Stage I, and the prior belief of the entering agent is less than this, exit occurs immediately. Since exit occurs immediately, all beliefs remain constant, which implies that all subsequent decisions are identical given that the beliefs act as the state variables of the agent’s decision problem.

Whether exit happens in Stage II, depends entirely on the frequency of reward realisation at that stage. Every time a reward is observed, beliefs \( p_2 \) experience a discrete upward jump, and drift down in a continuous manner between reward observations. As we saw in the example in Figure 3.2, if the frequency of reward realisations is too low, then eventually beliefs \( p_2 \) will approach zero despite the upward jumps when the occasional reward is observed.

\section*{5.3 Stage I: No Reward Observed}

When any reward realisation is yet to be observed, characterising the paths of plays is more complex. Along the sequence of entering agents, the public belief \( p_{2,t}^\text{UC} \) connects all the individual decision problems since it summarises all the observed information concerning the risky arm. As we already know what happens after a reward realisation, we now focus exclusively on the path of play when no rewards are observed.

To get an idea of the belief dynamics, Figure 5.3 gives an example of how beliefs evolve over a sequence of entering and exiting agents. In the figure, player 1 begins with the prior beliefs \((p_{1,0}, p_{2,0}) = (0.9, 0.9)\), and over time he becomes more pessimistic and so the beliefs \((p_1, p_2)\) drift along the solid line towards the decision boundary. The dotted line depicts how the corresponding public beliefs \( p_{2,t}^\text{UC} \) evolve. Once the boundary is reached, the agent exits. At this moment the vertical arrow marks the belief \( p_{2,t}^\text{UC} \) at the moment of exit, and the following horizontal arrow captures the fact that \( p_{2,t}^\text{UC} \) serves as the prior for the state variable \( p_2 \) for the incoming player—agent 2. The prior \( p_1 \) concerning the agent’s type is identical to the
previous agent’s prior, since agent types are i.i.d.. Once agent 2 begins playing the risky arm, beliefs $p_2$ and $p_2^{UC}$ begin again to diverge, and the agent exits when $(p_1, p_2)$ reaches the boundary. The sequence continues on in a similar fashion, with agents entering and exiting.

Since agent types are i.i.d., each agent begins with the same prior $p_{1,0} = 0.9$, which is captured in the figure by the fact that all the starting points for belief paths are vertically aligned. Secondly, along the sequence of agents, each agent starts with a prior belief $p_2$ which is strictly lower than what the preceding agent started with. In fact, the prior beliefs over the sequence of agents approach the point at which the decision boundary intersects with the vertical line $p_{1,0}$.

![Belief paths for a sequence of agents.](image)

Figure 5: Belief paths for a sequence of agents. The solid lines depict the evolution of the state variables $(p_1, p_2)$. The dashed lines depict the public belief $p_2^{UC}$.

We now show that in the absence of reward realisations, the sequence of entries and exits depicted in figure 5.3 consists of an infinite number of agents. Moreover, the only case where an infinite number of agents play the risky arm is when rewards have never been observed along the sequence.
Proposition 6. Suppose that an agent begins with the prior beliefs \((p_{1,0}, p_{2,0}) \in D\). If no reward realisations are observed along the sequence of agents, then an infinite number of agents will play the risky arm for a strictly positive length of time before exiting to the safe arm. Furthermore, along the sequence the public beliefs satisfy \(p_{2,t_n}^{UC} > p_{2,t_{n+1}}^{UC}\) for all \(n \in \mathbb{N}\), where \(t_n^*\) denotes the exit time of agent \(n\).

Proof. For any priors \((p_{1,0}, p_{2,0}) \in D\), there exists some open ball \(B_\delta((p_{1,0}, p_{2,0})) \subset D\), with radius \(\delta > 0\). Thus, it must take a strictly positive amount of time for agent 1’s beliefs to cross \(B_\delta((p_{1,0}, p_{2,0}))\). Assumptions 1 and 2 imply that beliefs must reach the boundary \(\partial D\) in finite time. The agent’s beliefs at the time of exit \(t^*\) are given by \((p_{1,t^*}, p_{2,t^*})\). Since \(t^* > 0\) and beliefs \((p_{1,t}, p_{2,t})\) have a strictly negative drift, it must be the case that \(p_{1,t^*} < p_{1,0}\) and \(p_{2,t^*} < p_{2,0}\). By Lemma 3 in Appendix B, we also know that \(p_{2,t}^{UC} > p_{2,t}\) for \(t > 0\), with \(p_{2,t}^{UC} = p_{2,t}\) only at the instance of entry. Moreover, since the boundary is reached in finite time, it must be the case that \(p_{2,0} > p_{2,t}\). Following agent 1’s exit, agent 2 enters and begins with the priors \((p_{1,0}, p_{2,t}^{UC})\), where the entry time \(t_2^*\) of agent 2 is equal to the exit time \(t^*\) of agent 1. Thus, by definition \(p_{2,t^*}^{UC} = p_{2,t}^{UC}\). As the exit beliefs of agent 1 satisfy \(p_{1,t^*} < p_{1,0}\) and \(p_{2,t^*} < p_{2,0}\), this implies that \(p_{2,t^*}^{UC} \in \partial D\). And since the prior \(p_{1,0}\) is the same for all agents, we have \(p_{1,0} \in D\). Consequently, the agent 2 begins with beliefs which are strictly in the interior of the continuation region, i.e. \((p_{1,0}, p_{2,t}^{UC}) \in D\). The statement then follows by induction. \(\square\)

The next result shows that the sequence of prior beliefs \(\{p_{2,t_n}^{UC}\}_{n \in \mathbb{N}}\) describing the entry beliefs of agents, approaches a limit point \(p_2^\dagger \in (0, 1)\). However, this limit point is never reached with a sequence consisting of a finite number \(a\) of entries and exits.

Proposition 7. Suppose there are never any reward observations, and consider the sequence of priors \(\{p_{2,t_n}^{UC}\}_{n \in \mathbb{N}}\) for each entering agent. Also assume that the first agent in the sequence has prior beliefs which are in the interior of the continuation region, i.e. \((p_{1,0}, p_{2,0}) \in D \cap (0, 1)^2\). Then, there exists a unique limit point \(p_2^\dagger \in (0, 1)\), such that \(\lim_{n \to \infty} p_{2,t_n}^{UC} = p_2^\dagger\). 30
where the belief $p^1_2$ is the belief which would induce an entering agent to exit immediately. Furthermore, $p^1_2$ is value of $p_2$ that satisfies $p_{1,0} = p^*_1(p_2)$.

**Proof.** See Appendix B

We can see that in the absence of payoff realisations, the sequence of entries and exits is composed of an infinite number of strictly positive experimentation durations on the risky arm. Since no agent $n$ with a finite index ever reaches the limit point $(p_{1,0}, p^1)$, it is not immediately clear whether the sum of experimentation lengths of agents is finite. The next result shows that the length of time between entry and exit shrinks fast enough so that the total sum of the experimentation durations is indeed finite.

**Proposition 8.** In the absence of reward observations on the risky arm, there exists a finite time $T^* > 0$, such that $T^* \geq \sum_{n=1}^{\infty} t^*_n$, where $t^*_n$ denotes the exit time of an arbitrary agent $n$. Consequently, herding on the safe arm must begin by some finite time. Furthermore, the upper bound $T^*$ can be chosen to be arbitrarily close to $\sum_{n=1}^{\infty} t^*_n$.

**Proof.** Given that $\frac{dp^{UC}_2}{dt}$ is continuous with respect to $t$ and the prior $p_{2,0}$, and since $\frac{dp^{UC}_2}{dt} < 0$ for $p_{2,0} \in (0, 1)$, there exists a $\epsilon > 0$, $\tilde{t} > 0$, such that $\frac{dp^{UC}_2}{dt} < 0$, for all $(p_{2,0}, t) \in [p^1_2, p^1_2 + \epsilon] \times [0, \tilde{t}]$, where $p^1 \in (0, 1)$ is defined in Proposition 7. We define

$$\frac{dp^{UC}_2}{dt} := \max_{(p_{2,0}, t) \in [p^1_2, p^1_2 + \epsilon] \times [0, \tilde{t}]} \frac{dp^{UC}_2}{dt}$$

In particular, note that since we are maximising over a closed set $\frac{dp^{UC}_2}{dt} < 0$. For any $\Delta t \in (0, \tilde{t})$, there exists a $N \in \mathbb{N}$ such that $n \geq N$ implies $0 < \frac{p^{UC}_{2,t_n}}{p^1_2} - p^1_2 < \frac{|dp^{UC}_2|}{dt} |\Delta t |$. Consequently, we have $p^{UC}_{2,t_n} + \frac{dp^{UC}_2}{dt} |\Delta t | < 0$. Because the definition of $\frac{dp^{UC}_2}{dt}$ applies to every agent $n \geq N$, and it describes the least steep downward drift of $p^{UC}_{2,t}$ on $(p_{2,0}, t) \in [p^1_2, p^1_2 + \epsilon] \times [0, \tilde{t}]$, it must be the case that all agents are herding on the safe arm by time $t^*_n + \Delta t$. If it were not so, beliefs $p^{UC}_{2,t}$ would be strictly below the threshold belief $p^1_2$. 31
Consequently, in the absence of reward observations, herding on has begun by time $T^*$, which is defined by $T^* := \Delta t + \sum_{i=1}^{n} t^*_i$.

### 5.4 Summary of Possible Sequences

From the results above, we can see that the only case where an infinite number of agents play the risky arm is when there are never any reward realisations. As agents continue to enter and exit after seeing no rewards, they experiment for increasingly shorter lengths of time. Every entering agent has a strictly higher valuation for the risky arm than the exiting agent, which implies that the entering agent will experiment for a strictly positive amount of time. This implies that there will be an infinite number of agents entering and playing the risky arm. Even though there are an infinite number of agents experimenting on the risky arm, the sum total of the lengths of time that players have experimented on the risky arm is bounded. This implies that in the absence of reward realisations, herding must happen by a finite time $T^*$.

In contrast, the first agent who observes a reward enters Stage II. In this stage two things can happen. Firstly, reward realisations can be too infrequent, whereby the agent eventually becomes pessimistic enough to exit to the safe arm. In this case, given that all uncertainty relating to the agent’s type has been eliminated after the first reward realisation, the next agent is now more pessimistic about his own type. This is because the new agent will be a high type with probability $p_{1,0} < 1$, while the preceding agent was a high type with probability 1. As a result, if the most optimistic agent chooses to exit, then all agents who are more pessimistic will also exit.

Secondly, if after entering Stage II, an agent receives rewards at a rate in accordance to the arm being of a good type, then the agent will become increasingly optimistic over time. As time goes on, if the arm is good, then it will become increasingly less likely that the agent
will exit, since by the Law of Large Numbers, the belief $p_{2,t} \to 1$ as $t \to \infty$.

As a result, if any agent ever observes a reward, then that agent will be the last agent in the sequence to ever play the risky arm. Either that agent plays the risky arm for a finite time and exits, or else he plays the risky arm forever.

6 Welfare

Agents do not internalise the information externality of their actions, nor the social cost from monopolising the risky arm. Consider, Figure 1 where if experimentation lasts for long enough and there are no rewards observations, the public belief $p_{t}^{UC}$ starts returning back toward the prior. At this stage, it becomes increasingly likely that the agent is a low type. Hence, with a high probability, the observations made during the tenure of the incumbent agent are of no value to the other players. In order to make this more precise, we adopt the following definition of welfare.

**Definition 1.** The social welfare is defined as follows:

$$W_t \left( p_{1,t}, p_{2,t}, \left\{(t^*_{n}, p_{2,t_n}^{UC})\right\}_{n \in \mathbb{N}} \right) = \begin{cases} U_t(p_{1,t}, p_{2,t}) + \sum_{n} e^{-r(t^*_n - t)}U_t(p_{1,0}, p_{2,t_n}^{UC}) & \text{Stage I: No reward obs.} \\ V(p_{2,t}) & \text{Stage II: No herding} \\ 0 & \text{Stage II: Herding} \end{cases}$$

where $\left\{(t^*_{n}, p_{2,t_n}^{UC})\right\}_{n \in \mathbb{N}}$ is a sequence of pairs consisting of the entry time $t^*_n$ and public belief at the time of entry $p_{2,t_n}^{UC}$, for each agent $n \in \mathbb{N}$.

When herding on the safe arm has taken place, welfare is zero. When an agent has observed a reward and continues to experiment, then the social welfare is identical to the agent’s value of the risky arm. This is due to the results in the previous section where we saw that if an exit ever happens after a reward observation, then all subsequent agents will choose the safe
arm whereby the welfare will be identically zero from that point on. When no rewards have been observed, welfare is given by the discounted sum of the value functions along a sequence with no reward observations.

We assume that the social planner does not care about distributional issues. If the planner knew that all agents were high types, he would not care whether a single agent or many agents played the risky arm. This simplifies the welfare analysis, since the highest welfare is then achieved when a high type agent plays a good arm forever. Hence, our definition of welfare abstracts away from any concerns of how payoffs are distributed over the sequence of agents.

If a planner had complete control over the timing of entry along the sequence of agents, the planner would replace an agent at each instance in time in Stage I. This is because without payoff observations, both beliefs \((p_{1,t}, p_{2,t})\) have a strict downward drift. Thus, an incoming agent will automatically have a strictly higher expected value since both of his priors \((p_{1,0}, p_{2,t}^{UC})\) are strictly greater than the incumbent’s beliefs \((p_{1,t}, p_{2,t})\) regardless of the exit time of the incumbent. As a result, by replacing an agent at each instant, the planner would effectively be holding the belief concerning the agent’s type fixed at the prior \(p_{1,0}\).

Consequently, because the planner can keep the belief \(p_{1}\) fixed by replacing agents continuously, the only belief of relevance remaining is \(p_{2,t}^{UC}\). This belief will serve as the sole state variable, whose dynamics can be summarised in differential form as follows

\[
dp_{2}^{UC} = -kp_{1,0}(\lambda_H - \lambda_L)(1 - p_{2}^{UC})p_{2}^{UC} dt
\]

The HJB equation associated with the planner’s problem is given by

\[
W(p_{2}^{UC}) = \max_{k \in [0,1]} k \left\{ \lambda(p_{2}^{UC})p_{1,0}h - c + \frac{1}{r} \left\{ \lambda(p_{2}^{UC})p_{1,0} \left[ V(j(p_{2}^{UC})) - W(p_{2}^{UC}) \right] - \frac{\partial W}{\partial p_{2}^{UC}} p_{1,0}(\lambda_H - \lambda_L)p_{2}^{UC}(1 - p_{2}^{UC}) \right\} \right\}
\]
Note that although the form of the above equation is similar to the previous Stage II HJB equation of the agent’s decision problem given in equation 7, they are not identical. The planner’s problem still has two distinct phases. In Stage I, even though the continuous replacement of agents keeps beliefs $p_1$ fixed, at any instance of time there is always uncertainty about the agent’s type, which is given by $p_{1,0}$. Thus, in the above equation, the learning rates are multiplied by $p_{1,0}$. In Phase II the value is equal to $V(j(p_{UC}^2))$ and $p_{1,0}$ no longer plays a role.

Given that we already know $V(j(p_{UC}^2))$, the planner’s problem reduces to solving the following ODE

$$\frac{\partial W}{\partial p_2} + a(p_{UC}^2)W(p_{UC}^2) = b(p_{UC}^2)$$  \hspace{1cm} (22)$$

$$a(p_{UC}^2) = \frac{(r + \lambda(p_{UC}^2)p_{1,0})}{p_{1,0}(\lambda_{H} - \lambda_{L})p_{UC}^2(1 - p_{UC}^2)}$$

$$b(p_{UC}^2) = \frac{r\left(\lambda(p_{UC}^2)p_{1,0}h - c\right) + p_{1,0}\lambda(p_{UC}^2)V(j(p_{UC}^2))}{p_{1,0}(\lambda_{H} - \lambda_{L})p_{UC}^2(1 - p_{UC}^2)}$$

The solution to the DE in 22 is given by:

$$W(p_{UC}^2) = e^{-\int a} \left[ \int b(p_{UC}^2)e^{\int a} dp_{UC}^2 + C \right]$$

The value matching condition $W(p_{\dagger}^2) = 0$ allows us to eliminate the constant $C$ giving us

$$W(p_{UC}^2) = e^{-\int a} \left[ \int b(p_{UC}^2)e^{\int a} dp_{UC}^2 \right]$$

The smooth pasting condition $W'(p_{\dagger}^2) = 0$ is equivalent to requiring that $b(p_{\dagger}^2) = 0$, which
gives us the decision boundary, which in this case is just the scalar value $p_2^\dagger$.

\[
0 = r\left(\lambda(p_2^\dagger)p_{1,0}h - c\right) + p_{1,0}\lambda(p_2^\dagger)V(j(p_2^\dagger)) \\
p_{1,0} = \frac{rc}{\lambda(p_2^\dagger)(rh + V(j(p_2^\dagger)))}
\]

Threshold belief $p_2^\dagger$ is defined implicitly as the solution to the above equation. Note that $p_2^\dagger$ is equivalent to the limit point in Proposition 7.

### 6.1 Policymaker

Suppose that the policymaker cannot control the entry and exit times of agents, but can influence them by varying the flow costs $c$. Is it possible that increasing the costs of experimentation are offset by the gains from reducing socially wasteful experimentation? To investigate this question, we must first understand exactly how $c$ affects the value functions $U$. Moreover, given that $c$ will have an effect on the exit times, the entire entry-exit sequence $\{t_k^*\}$ will change if we change $c$. In order to investigate this, we derive a procedure which allows us to compute the exit time and exit beliefs over a sequence of no reward observations. This procedure is summarised in Algorithm 1.

We can express the evolution of $p_1$ purely in terms of $p_2$, similarly by solving $\frac{dp_1}{dp_2}$ which gives us

\[
p_1(p_2) = \left(1 + e^{-\frac{\eta}{\lambda H - \lambda L}p_2}\left(1 - p_2^\dagger\right)^{\frac{\lambda H}{\lambda H - \lambda L}}\right)^{-1}
\]

The arbitrary constant $\eta$ is pinned down by the initial conditions $(p_{1,0}, p_{2,0})$ and is computed
in the Appendix in equation 32, which we can re-arrange to get

\[ \eta = \log \left( \frac{p_{1,0}}{1 - p_{1,0}} \right)^{\lambda_H - \lambda_L} \left( \frac{1 - p_{2,0}}{p_{2,0}} \right)^{\lambda_L} \]

\[ e^{-\frac{\eta}{\lambda_H - \lambda_L}} = \frac{1 - p_{1,0}}{p_{1,0}} \left( \frac{\lambda_H}{p_{2,0}} \right)^{\frac{\lambda_H}{\lambda_H - \lambda_L}} \left( \frac{1 - p_{2,0}}{p_{2,0}} \right)^{\frac{\lambda_L}{\lambda_H - \lambda_L}} \]

where in the second line follows from the first. This in turn gives us

\[ p_1(p_2) = \left( 1 + \frac{1 - p_{1,0}}{p_{1,0}} \right) \left( 1 - p_{2,t} \right) \left( \frac{\lambda_H}{\lambda_H - \lambda_L} \right) \left( \frac{\lambda_L}{\lambda_H - \lambda_L} \right)^{-1} \]

We can now find the specific belief \( p_2 \), where \( p_1(p_2) \) intersects with the decision boundary \( p_1^*(p_2) \) by equating the two:

\[ p_1^*(p_2) = p_1(p_2) \]

\[ \frac{rc}{(\lambda(p_2)(rh + V(j(p_2))))} = \left( 1 + \frac{1 - p_{1,0}}{p_{1,0}} \right) \left( 1 - p_{2,t} \right) \left( \frac{\lambda_H}{\lambda_H - \lambda_L} \right) \left( \frac{\lambda_L}{\lambda_H - \lambda_L} \right)^{-1} \]

The left-hand side is given in equation 19. The above equation is a transcendental equation, which cannot be solved explicitly for \( p_2 \). However, we can see that for \( p_2 \geq p_2^{**} \), there is a unique solution, since the left-hand side is strictly decreasing in \( p_2 \), and the right-hand side is strictly increasing in \( p_2 \). Denote the solution by \( p_2^\dagger \), whereby the exit beliefs are thus given by \( (p_1^*(p_2^\dagger), p_2^\dagger) \). This in turn allows us to compute the public belief \( p_{2,t}^{UC} \), which serves as the prior for the next agent, who enters at a time which we denote by \( t_1^* \).

\[ p_{2,t_1^*}^{UC} = p_2^\dagger p_1^*(p_2^\dagger) + p_{2,0}(1 - p_1^*(p_2^\dagger)) \]

Once we have solved for \( p_2^\dagger \), we can back out the exit time using the expressions of \( p_{2,t} \) given in Lemma 1 and solving for the time \( t \) at which \( p_{2,t} \) hits \( p_2^\dagger \). In other words we solve the
The following equation for $t$:

\[
p^\dagger_2 = \left[ 1 + \left( \frac{1 - p_{2,0}}{p_{2,0}} \right) e^{k(\lambda_H - \lambda_L)t} \right]^{-1}
\]

Where the unique solution $t^*_1$ is given by

\[
t^*_1 = \frac{1}{\lambda_H - \lambda_L} \log \left[ \left( \frac{1 - p^\dagger_2}{p^\dagger_2} \right) \left( \frac{1 - p_{2,0}}{p_{2,0}} \right)^{-1} \vphantom{\frac{1}{\lambda_H - \lambda_L}} \right]
\]

We summarise these results below, which describe a procedure of how to solve for the entry-exit beliefs and experimentation durations along the sequence.

**Algorithm 1.** In the case when there are no reward observations, we can solve for the sequence $\{t^*_n, (p^*_1, p^*_2)\}$ of entry-exit times and beliefs by the following procedure:

1. **Given priors** $(p_{1,0}, p_{2,0})$, find the value of $p_2$ that solves the following equation and denote it by $p_2^\dagger$:

\[
\frac{r_c}{\lambda(p_2)(rh + V(j(p_2)))} = \left( 1 + \frac{1 - p_{1,0}}{p_{1,0}} \right) \left( \frac{1 - p_2}{1 - p_{2,0}} \right) e^{\lambda_H - \lambda_L} \left( \frac{p_2}{p_{2,0}} \right)^{-1}
\]

(23)

2. $(p^*_1(p_2^\dagger), p^*_2)$ gives the beliefs at which the first agent exits, where $p^*_1(p_2)$ denotes the left-hand side function in 23.

3. Compute the exit time from the following equation:

\[
t^* = \frac{1}{\lambda_H - \lambda_L} \log \left[ \left( \frac{1 - p^\dagger_2}{p^\dagger_2} \right) \left( \frac{1 - p_{2,0}}{p_{2,0}} \right)^{-1} \vphantom{\frac{1}{\lambda_H - \lambda_L}} \right]
\]

3. Compute the public belief $p^\text{UC}_2$ at the time of exit from the following equation:

\[
p^\text{UC}_{2,t^*_1} = p^\dagger_2 p^*_1(p^\dagger_2) + p_{2,0}(1 - p^*_1(p_2^\dagger))
\]
4. Set $p_{2,0} = p_{2,0}^{UC}$, and go back to Step 1 and repeat.

### 6.2 Numerical Example

In this subsection we give a numerical example which shows how the experimentation durations and welfare evolve along the sequence in the absence of reward observations. Figure 6 shows how at the beginning of the sequence agents spend a long time on the risky arm. However, as the risky arm continues to yield no rewards, we can see that the duration of each agent’s experimentation rapidly declines after the first few agents have exited. Eventually, the entering agents will spend only a moment on the risky arm before exiting. This captures the result of Proposition 7, which showed that an infinite number of agents arrive before experimentation on the risky arm ceases completely.

![Experimentation Duration for Each Agent](image)

Figure 6: Experimentation durations on the risky arm along the sequence of entries and exits.

In Figure 7 we see how welfare evolves along the same sequence. Welfare follows a wave-like pattern, since just before the entry of a new agent, it becomes increasingly likely that the
incumbent is a low type. As the moment of the incumbent’s exit draws near, the value of the entering agent will be strictly higher, which is reflected in the upward-sloping curve right before the exit of the incumbent.

![Welfare in Stage I](image)

Figure 7: Welfare in Stage I in the absence of reward observations.

7 Extensions

Perhaps the main advantage of the framework of this paper is its simple structure. The fact that the framework permits closed-form solutions suggests that extending the model into strategic settings should be relatively easy. Firstly, we have derived the posterior belief paths so that they depend on the arm choices of the agents. Secondly, the agent’s decision problem uses the posterior beliefs as state variables. Together these imply that in settings with pure informational externalities,\(^{11}\) we would not have to change the structure of the decision problems of the strategic actors. In equilibrium, actions would influence the belief

\(^{11}\)Where the utility function of an agent is not directly influenced by the actions of other agents.
evolution, which in turn would determine the actions. All of this can be captured without changing the basic structure of the model.

Consider strategic environments similar to those of Bolton and Harris (1999); Keller, Rady, and Cripps (2005), that focus on settings with pure informational externalities that create incentives for free-riding. If we allow for type heterogeneity, this might reduce the free-riding incentive, as one agent’s experimentation would not necessarily be informative to another agent. In some respects we can think of models which are, in some respects, at the other end of the spectrum from the strategic settings. Think of $N$ agents having access to completely independent bandit arms. In this case the equilibrium levels of experimentation would be trivially efficient. Given that all Markov Perfect Equilibria in the aforementioned strategic models are inefficient, one would naturally expect that as we increase the amount of correlation between agent types, any equilibrium becomes increasingly inefficient.

Heterogeneity would also introduce more complex dynamics to any symmetric MPE. Consider a strategic setting of the model of our paper where rewards produced imperfectly revealing good news about the arms. In this setting, agents who had observed nothing would be in the first wave to switch to the safe arm, whilst those that had experienced a reward would switch to the safe arm in a second wave at a later point in time if the rewards were not realised frequently enough. A more complicated case worth considering would be one where we assumed that reward realisations were also imperfectly revealing about agent types. Then, foreseeably, agents would switch to the safe arm based on the number of realisations they had observed and so one might expect to still see exit waves of switching to the safe arm. However, the belief-updating in such a setting would be more complicated, since an agent would need to track the number of reward realisations that each agent experienced in order to be able to accurately interpret what their experiences revealed about the state.

We can think of numerous other environments where multidimensional type heterogeneity might play a crucial role. For instance, consider principal-agent models where there is uncertainty about the quality of a project and about the skill of the worker. If the worker
does not succeed in his project, the principal would want to know whether this was due to a poor match between project and worker. If the match was poor, then was this because the project was bad or because the worker lacked the necessary skills? Knowing the reason for the lack of success has wide variety of implications for decision-making. For instance, if the project were bad, then the principal might simply want to re-allocate the worker to a different project. However, if the worker lacked in skill, then the principal might want to terminate the employment relationship.

More applied settings where heterogeneity plays an important role concerns firm investments into technological development. We can think of firms competing with one another by investing in R&D. On the one hand, it might be highly lucrative to be the first firm to make some technological breakthrough, but on the other hand the probability of achieving such a breakthrough might be low, and R&D is costly. Suppose many firms were investing in a similar type of technology, where the R&D investments yielded profits that were imperfectly revealing about the long-term value of the underlying technology. Then, the success of rival firms would make all firms more optimistic about the potential of the underlying technology. Nonetheless, if R&D expenditures are increased as a result, competition will lower expected future profit flows due to increased future competition. In some respect, free-riding forces similar to Bolton and Harris (1999); Keller, Rady, and Cripps (2005) are present here.

8 Conclusion

We have presented a tractable continuous time model to study learning about a multidimensional state. The agent’s decision problem yields simple closed-form solutions, which facilitates the subsequent analysis and opens up the possibility of extending the basic framework in other directions. This paper has studied how social learning is influenced by heterogeneity of agents. This dimension is important since it shows how learning about one dimension can interact with learning about another dimension. This can give rise to non-monotonic
posterior belief dynamics which are not present in the existing literature. Furthermore, we characterise the sequence of play when agents enter and exit, and somewhat surprisingly find that the only situation when an infinite number of agents will play the risky arm is when the risky arm has never yielded any rewards. The welfare analysis shows that in the absence of any fixed entry costs, the social planner would want to replace the agents infinitely fast. This is driven by the fact that the absence of rewards implies strictly increasing pessimism about the agent’s type, and so a new agent will always have a higher value for the risky arm.
A Appendix: Agent’s Decision Problem

A.1 Stage II Decision Problem

A.1.1 Hamilton-Jacobi-Bellman Equation

We provide a heuristic derivation of the HJB equation corresponding to the agent’s dynamic programming problem in Stage II:

\[
V(p_2) = \max_{k \in [0,1]} \left\{ r k \left[ h \lambda(p_2) - c \right] dt + (1 - r dt) E \left[ V(p_2 + dp_2) \mid p_2 \right] \right\}
\]

\[
V(p_2) = \max_{k \in [0,1]} \left\{ r k \left[ h \lambda(p_2) - c \right] dt + (1 - r dt) \left[ \lambda(p_2) k dt \cdot V(j(p_2)) \right. \right.
\]
\[
\left. \left. + (1 - \lambda(p_2) k dt)V(p_2 + dp_2) \right] \right\}
\]

\[
V(p_2) [1 - (1 - r dt)] = \max_{k t} \left\{ r k \left[ h \lambda(p_2) - c \right] dt + (1 - r dt) \left[ \lambda(p_2) k dt \cdot [V(j(p_2)) - V(p_2)] \right. \right.
\]
\[
\left. \left. \left. \left. + (1 - \lambda(p_2) k dt)V'(p_2)dp_2 \right] \right] \right\}
\]

Let \( dt \to 0 \),

\[
V(p_2) = \max_{k \epsilon} k \cdot \left\{ h \lambda(p_2) - c + \frac{1}{r} [\lambda(p_2) \cdot [V(j(p_2)) - V(p_2)] - V'(p_2) (\lambda_H - \lambda_L) (1 - p_2)p_2] \right\}
\]

A.1.2 Solution to ODDE: Overview

We follow Keller and Rady (2010) and solve the ordinary differential-difference equation (ODDE) in 8 in a manner analogous to solving regular ODEs:

1. Find a particular solution \( V_p \) to the DE given in 8.

2. Find the general solution \( V_c \) to the homogeneous (complementary) solution to the DE given in 8.
3. A general solution to equation 8 is given by $V(p_2) = V_p(p_2) + V_c(p_2)$

After having solved equation 8, the value function $V^*$ of the single-agent decision problem can be found by solving for the cut-off belief $p_2^{**}$ where the agent switches from the risky arm to playing the safe arm. The cut-off belief can be solved with standard techniques by imposing value matching $V(p_2^{**}) = 0$ and smooth pasting $V'(p_2^{**}) = 0$.

A.1.3 Solving the ODDE: A Particular Solution

We guess that a particular solution to equation 8 is given by $V(p_2) = \lambda(p_2)h - c$, and verify it below:

$$V(p_2) = h\lambda(p_2) - c + \frac{1}{r}[\lambda(p_2) \cdot [V(j(p_2)) - V(p_2)] - V'(p_2) (\lambda_H - \lambda_L) (1 - p_2)p_2]$$

$$\lambda(p_2)h - c = h\lambda(p_2) - c + \frac{1}{r}[\lambda(p_2) \cdot \lambda(j(p_2))h - \lambda(p_2)h] - \lambda'(p_2)h (\lambda_H - \lambda_L) (1 - p_2)p_2]$$

$$0 = \lambda(p_2) \cdot \left[\lambda_H (\frac{\lambda_H}{\lambda(p_2)}) p_2 + \lambda_L \left(1 - \frac{\lambda_H}{\lambda(p_2)} \right) p_2\right] - (\lambda_H - \lambda_L)^2 (1 - p_2)p_2$$

$$0 = \lambda^2 p_2 + \lambda_L (\lambda(p_2) - \lambda_H p_2) - (\lambda(p_2))^2 - (\lambda_H - \lambda_L)^2 (1 - p_2)p_2$$

$$0 = (\lambda_H - \lambda_L)^2 p_2 \left(1 - p_2\right) - (\lambda_H - \lambda_L)^2 \left(1 - p_2\right)p_2$$

This verifies $V(p_2) = \lambda(p_2)h - c$ as a particular solution.

A.1.4 Solving the ODDE: Solution to the Homogeneous Equation

The homogeneous version of equation 8 is given below

$$V(p_2) = \frac{1}{r}[\lambda(p_2) \cdot [V(j(p_2)) - V(p_2)] - V'(p_2) (\lambda_H - \lambda_L) (1 - p_2)p_2]$$

(24)

Note that this is essentially the same as in Keller and Rady (2010), and so we can follow the method of solution of the authors.
Guess that the solution takes the general form

\[ V(p_2) = (1 - p_2) [\Omega(p_2)]^\mu \]

\[ \Omega(p_2) := \frac{1 - p_2}{p_2} \]

Firstly, note that

\[ V'(p_2) = -[\Omega(p_2)]^\mu + \mu (1 - p_2) [\Omega(p_2)]^{\mu-1} \Omega'(p_2) \]

\[ = -[\Omega(p_2)]^\mu \left(1 + \frac{\mu}{p_2}\right) \]

Secondly, note that

\[ V(j(p_2)) = \left(1 - \frac{\lambda_H}{\lambda(p_2)} p_2\right) \left[\Omega\left(\left(\frac{\lambda_H}{\lambda(p_2)}\right) p_2\right)\right]^\mu \]

\[ = \left(\frac{\lambda_L(1 - p_2)}{\lambda(p_2)}\right) \left(\frac{\lambda_L}{\lambda_H}\right)^\mu [\Omega(p_2)]^\mu \]

Plugging these into equation 24 gives us

\[ (1 - p_2) [\Omega(p_2)]^\mu = \frac{1}{r} \left\{\lambda(p_2) \cdot \left[\left(\frac{\lambda_L(1 - p_2)}{\lambda(p_2)}\right) \left(\frac{\lambda_L}{\lambda_H}\right)^\mu - (1 - p_2)\right] [\Omega(p_2)]^\mu \right\} \]

\[ + [\Omega(p_2)]^\mu \left(1 + \frac{\mu}{p_2}\right) (\lambda_H - \lambda_L) (1 - p_2) p_2 \]

\[ (1 - p_2)r = \left[\lambda_L(1 - p_2) \left(\frac{\lambda_L}{\lambda_H}\right)^\mu - \lambda(p_2)(1 - p_2)\right] + \left(1 + \frac{\mu}{p_2}\right) (\lambda_H - \lambda_L) (1 - p_2) p_2 \]

\[ r = \lambda_L \left(\frac{\lambda_L}{\lambda_H}\right)^\mu - \lambda_L + \mu (\lambda_H - \lambda_L) \] (25)

Re-arranging the equation, we get

\[ r + \lambda_L - \mu (\lambda_H - \lambda_L) = \lambda_L \left(\frac{\lambda_L}{\lambda_H}\right)^\mu \] (26)
Since this an exponential polynomial equation, it cannot be solved explicitly for $\mu$. However, as both sides of the equation are continuous in $\mu$, it is not hard to show the existence two roots $\mu$, which satisfy the equation—one positive and one negative root. Also, considering that with $\mu < 0$, we would have $\lim_{p \to 1} V(p_2) = \infty$, we choose the positive root of the equation, and discard the negative one.

**A.1.5 Solving the ODDE: Value Matching and Smooth Pasting**

Putting the particular solution and the solution to the homogeneous equation together, the general solution to equation 8 is given by

$$V(p_2) = \lambda(p_2)h - c + C_0(1 - p_2)[\Omega(p_2)]^\mu \quad C_0 \in \mathbb{R} \quad (27)$$

Note that $V(p_2)$ in 27 is only the value of the risky arm. To solve the agent's decision problem which takes into account the value of the outside option, we need to solve for the cut-off belief $p_2^{**}$ where the agent exits and switches to the safe arm forever and gets $V = 0$. The existence of such a threshold belief is guaranteed by Assumption 1. This guarantees that the agent will choose the safe arm for some range of beliefs, and the risky arm for some other range. We can get an expression for the cut-off belief $p_2^{**}$ and the constant $C_0$ by imposing value matching $V(p_2^{**}) = 0$ and smooth pasting $V'(p_2^{**}) = 0$.

By value matching we have

$$0 = V(p_2^{**}) = \lambda(p_2^{**})h - c + C_0(1 - p_2^{**})[\Omega(p_2^{**})]^\mu$$

$$C_0 = \frac{c - \lambda(p_2^{**})h}{(1 - p_2^{**})[\Omega(p_2^{**})]^\mu}$$
Thus,
\[
V(p_2) = (\lambda(p_2)h - c) + (c - \lambda(p_2^{**})h) \left(\frac{1 - p_2}{1 - p_2^{**}}\right) \left(\frac{\Omega(p_2)}{\Omega(p_2^{**})}\right)^\mu
\]

Smooth pasting gives us
\[
V'(p_2) = (\lambda_H - \lambda_L)h - (c - \lambda(p_2^{**})h) \left(\frac{1}{1 - p_2^{**}}\right) \left(\frac{\Omega(p_2)}{\Omega(p_2^{**})}\right)^\mu
+ (c - \lambda(p_2^{**})h) \left(\frac{1 - p_2}{1 - p_2^{**}}\right) (-\mu) \left(\frac{\Omega(p_2)}{\Omega(p_2^{**})}\right)^{\mu-1} \left(-1/p_2\right)
\]
Setting \(V'(p_2^{**}) = 0\) we get
\[
0 = (\lambda_H - \lambda_L)h - (c - \lambda(p_2^{**})h) \left(\frac{1}{1 - p_2^{**}}\right) \left(\frac{\Omega(p)}{\Omega(p_2^{**})}\right)^\mu
+ (c - \lambda(p_2^{**})h) \left(\frac{1 - p}{1 - p_2^{**}}\right) (-\mu) \left(\frac{\Omega(p)}{\Omega(p_2^{**})}\right)^{\mu-1} \left(-1/p_2^{**}\right)
\]
This yields the cut-off belief
\[
p_2^{**} = \frac{\mu \left(c - h\lambda_L\right)}{h\lambda_H - c + h\mu(\lambda_H - \lambda_L)} \tag{28}
\]
Hence, we get the final form for the value function for the Stage II problem:
\[
V(p_2) = (\lambda(p_2)h - c) + (c - \lambda(p_2^{**})h) \left(\frac{1 - p_2}{1 - p_2^{**}}\right) \left(\frac{\Omega(p_2)}{\Omega(p_2^{**})}\right)^\mu
\]
where \(p_2^{**}\) is given in equation 28 and \(\mu\) is given implicitly in equation 26.
A.2 Stage I Decision Problem

A.2.1 Hamilton-Jacobi-Bellman Equation

We provide a heuristic derivation of the HJB equation corresponding to the agent’s dynamic programming problem in Stage I:

\[
U(p_1, p_2) = \max_{k \in [0,1]} \{rk [\lambda(p_2) \cdot p_1 h - c] dt + (1 - r dt) E [U(p_1 + dp_1, p_2 + dp_2)|p_1, p_2]\}
\]

\[
= \max_{k \in [0,1]} \left\{ rk [\lambda(p_2) \cdot p_1 h - c] dt + (1 - r dt) \left[ k \lambda(p_2) \cdot p_1 dt \cdot U(1, j(p_2)) + (1 - k \lambda(p_2) \cdot p_1 dt) \left[ U(p_1, p_2) + \frac{\partial U(p_1, p_2)}{\partial p_1} \cdot dp_1 + \frac{\partial U(p_1, p_2)}{\partial p_2} \cdot dp_2 + o(||(dp_1, dp_2)||) \right] \right] \right\}
\]

Re-arranging and subtracting \((1 - r dt)U(p_1, p_2)\) from both sides.

\[
U(p_1, p_2)r dt = \max_{k \in [0,1]} \left\{ rk [\lambda(p_2) \cdot p_1 h - c] dt + (1 - r dt) \left[ k \lambda(p_2) \cdot p_1 dt \cdot [U(1, j(p_2)) - U(p_1, p_2)] \right] \right\}
\]

\[
... + (1 - k \lambda(p_2) \cdot p_1 dt) \left[ \frac{\partial U(p_1, p_2)}{\partial p_1} \cdot dp_1 + \frac{\partial U(p_1, p_2)}{\partial p_2} \cdot dp_2 + o(||(dp_1, dp_2)||) \right] \right\}
\]

Ignoring the higher-order terms and dividing by \(r dt\), gives us

\[
U(p_1, p_2) = \max_{k \in [0,1]} \left\{ k [\lambda(p_2) \cdot p_1 h - c] + \frac{1}{r} \left[ k \lambda(p_2) \cdot p_1 \cdot [U(1, j(p_2)) - U(p_1, p_2)] \right] \right\}
\]

\[
... + (1 - k \lambda(p_2) \cdot p_1 dt) \left[ \frac{\partial U(p_1, p_2)}{\partial p_1} \cdot \frac{dp_1}{dt} + \frac{\partial U(p_1, p_2)}{\partial p_2} \cdot \frac{dp_2}{dt} \right] \right\}
\]

Substituting in the laws of motion \(dp_1\) and \(dp_2\), and letting \(dt \to 0\), gives

\[
U(p_1, p_2) = \max_{k \in [0,1]} \left\{ k [\lambda(p_2) \cdot p_1 h - c] + \frac{k}{r} \left\{ \lambda(p_2) \cdot p_1 \cdot [U(1, j(p_2)) - U(p_1, p_2)] \right\} \right\}
\]

\[
... - \left[ \frac{\partial U(p_1, p_2)}{\partial p_1} \cdot \lambda(p_2)p_1(1 - p_1) + \frac{\partial U(p_1, p_2)}{\partial p_2} \cdot (\lambda_H - \lambda_L)p_2(1 - p_2) \right] \right\}
\]

49
Since the right-hand side above is linear in \( k \), the maximising agent will either set \( k = 0 \) and get \( U = 0 \), or set \( k = 1 \), which leads to a differential equation. In the Stage II problem, solving the HJB equation led to solving an ODDE. In the Stage I problem, solving the above HJB equation leads us to a partial functional differential equation (PFDE), which we solve in the next subsection.

A.2.2 Solving the PFDE

In the continuation region \( D \) where the agent chooses \( k = 1 \), the HJB equation yields the following PFDE:

\[
\frac{\partial U(p_1, p_2)}{\partial p_1} \cdot \lambda(p_2)p_1(1 - p_1) + \frac{\partial U(p_1, p_2)}{\partial p_2} \cdot (\lambda_H - \lambda_L)p_2(1 - p_2) + [r + \lambda(p_2)p_1] U(p_1, p_2) = \\
r \left[ \lambda(p_2) \cdot p_1 h - c \right] + \lambda(p_2)p_1 U(1, j(p_2))
\]

(29)

The problematic term is \( U(1, j(p_2)) \), which is an unknown function evaluated at a point \( (1, j(p_2)) \), while the other functions \( U \) and its partials are evaluated at \( (p_1, p_2) \). This renders the differential equation non-local. We can nevertheless simplify the problem by noticing that \( U(1, j(p_2)) \) is the value an instance after observing the first reward. Thus, \( U(1, j(p_2)) \) is equal to the Stage II value \( V(j(p_2)) \). Since we have already solved the Stage II problem, we can replace the problematic term \( U(1, j(p_2)) \) with \( V^*(j(p_2)) \) in the differential equation. Thus the PFDE simplifies into a standard first-order linear partial differential equation (PDE).

\[
\frac{\partial U(p_1, p_2)}{\partial p_1} \cdot \lambda(p_2)p_1(1 - p_1) + \frac{\partial U(p_1, p_2)}{\partial p_2} \cdot (\lambda_H - \lambda_L)p_2(1 - p_2) + [r + \lambda(p_2) \cdot p_1] U(p_1, p_2) = \\
r \left[ \lambda(p_2) \cdot p_1 h - c \right] + \lambda(p_2)p_1 V^*(j(p_2))
\]

(30)

To solve this PDE, we use the Method of Characteristics,\(^{12}\) which essentially reduces the PDE into a parametrised ODE by using a conveniently chosen change of variables. In essence, this

\(^{12}\)See, for instance, John (1978).
enables us to analyse the resulting ODE on parametrised one-dimensional curves\textsuperscript{13} in the belief domain $(0, 1)^2$, where the ODE can be solved with standard methods. The solution can then be extended over the whole family of parametrised one-dimensional curves which cover the belief space in $(0, 1)^2$. Finally, reversing the change of variables gives the solution to the PDE.

Suitable characteristic curves are given by the (locally deterministic) belief paths that can be found explicitly by solving for one belief in terms of the other. We can express $p_1$ in terms of $p_2$ by solving the following ODE, which is the ratio of the differentials given in equations 2 and 3.

$$\frac{dp_1}{dp_2} = \frac{\lambda(p_2)}{(\lambda_H - \lambda_L)} \cdot \frac{p_1(1 - p_1)}{p_2(1 - p_2)}$$  \hspace{1cm} (31)

This is a separable differential equation, which we can solve by direct integration.

$$\int \frac{\lambda(p_2)}{p_2(1 - p_2)} dp_2 = \int \frac{(\lambda_H - \lambda_L)}{p_1(1 - p_1)} dp_1$$

This yields the following solution:

$$\lambda_H log(1 - p_2) - \lambda_L log(p_2) + (\lambda_H - \lambda_L)(-log(1 - p_1) + log(p_1)) = C$$

Using this solution we can introduce the following change of variables, which in effect allows us to eliminate one of the state variables in equation 30, thereby simplifying the PDE into an ODE.

\textsuperscript{13}These are often referred to as characteristic curves.
\[
\begin{align*}
    \xi(p_1, p_2) &= p_2 \\
    \eta(p_1, p_2) &= \lambda_H \log(1 - p_2) - \lambda_L \log(p_2) + (\lambda_H - \lambda_L)(-\log(1 - p_1) + \log(p_1))
\end{align*}
\] (32)

To make sure that the mapping between the original and new coordinate systems is bijective, note that

\[
    \det(J) = \det \begin{bmatrix} \xi_{p_1} & \xi_{p_2} \\ \eta_{p_1} & \eta_{p_2} \end{bmatrix} = \det \begin{bmatrix} 0 & 1 \\ \frac{\lambda_H - \lambda_L}{p_1(1 - p_1)} & -\frac{\lambda(p_2)}{p_2(1 - p_2)} \end{bmatrix} = -\frac{\lambda_H - \lambda_L}{p_1(1 - p_1)} < 0
\]

We can solve the second line in Equation 32 for \( p_1 \) expressing it in terms of \( \eta \) and \( \xi \). We denote the resulting solution by \( \Gamma \), which is given by

\[
    \Gamma(\xi; \eta) := p_1 = \left(1 + e^{-\frac{\eta}{\lambda_H - \lambda_L} \xi(1 - \xi) \frac{\lambda_H}{\lambda_H - \lambda_L} \xi - \frac{\lambda_H}{\lambda_H - \lambda_L}}\right)^{-1}
\] (33)

This allows us to substitute out \( p_1 \) with \( \Gamma(\xi, \eta) \), and \( p_2 \) with \( \xi \) in equation 30. We treat \( \eta \) as a constant, which effectively restricts the beliefs to a locus of points that lie on the same characteristic curve. The variable \( \xi \) specifies a point on the characteristic curve. Choosing a different value of \( \eta \) is equivalent to choosing a different characteristic curve.

Treating \( \eta \) as a constant also implies that one of the partial derivatives is equal to zero, and so we are left with a first-order linear ODE in \( \xi \) given by

\[
    -\frac{\partial U(\xi; \eta)}{\partial \xi} \cdot (\lambda_H - \lambda_L)\xi(1 - \xi) = \\
    \left[ r + \lambda(\xi) \cdot \Gamma(\xi; \eta) \right] U(\xi; \eta) - r \left[ \lambda(\xi) \cdot \Gamma(\xi; \eta) h - c \right] - \lambda(\xi) \cdot \Gamma(\xi; \eta) \cdot V^*(j(\xi))
\] (34)
We can express this more concisely as

\[ U_\xi + a(\xi; \eta) U = b(\xi; \eta) \]

where

\[
a(\xi; \eta) = \frac{r + \lambda(\xi) \cdot \Gamma(\xi; \eta)}{(\lambda_H - \lambda_L)\xi(1 - \xi)}
\]

\[
b(\xi; \eta) = \frac{r [\lambda(\xi) \cdot \Gamma(\xi; \eta)h - c]}{(\lambda_H - \lambda_L)\xi(1 - \xi)} + \frac{\lambda(\xi) \cdot \Gamma(\xi; \eta) \cdot V^*(j(\xi))}{(\lambda_H - \lambda_L)\xi(1 - \xi)}
\]

The solution to 34 is given by

\[
U(\xi; \eta) = \left(e^{-\int a(\xi; \eta)d\xi}\right) \int \left(b(\eta, \xi)e^{\int a(\xi; \eta)d\xi}\right)d\xi + f(\eta) \left(e^{-\int a(\xi; \eta)d\xi}\right)
\]

where \( f(\eta) \) is the constant of integration and by direct integration it can be shown that

\[
e^{-\int a(\xi; \eta)d\xi} = \left(\xi^{\lambda_H+r}_{\lambda_H-\lambda_L} (1 - \xi)^{-\lambda_H+r}_{\lambda_H-\lambda_L} + e^{-\lambda_H+r}_{\lambda_H-\lambda_L} (1 - \xi)^{-\lambda_H+r}_{\lambda_H-\lambda_L} \xi^{\lambda_H+r}_{\lambda_H-\lambda_L}\right)^{-1}
\]

\[
= (1 - \xi)^{\lambda_H+r}_{\lambda_H-\lambda_L} \xi^{-\lambda_H+r}_{\lambda_H-\lambda_L} \Gamma(\xi; \eta)
\]

and

\[
\int \left(b(\eta, \xi)e^{\int a(\xi; \eta)d\xi}\right)d\xi = -ce^{-\lambda_H+r}_{\lambda_H-\lambda_L} (1 - \xi)^{-\lambda_H+r}_{\lambda_H-\lambda_L} \xi^{\lambda_H+r}_{\lambda_H-\lambda_L}
\]

\[
+ (1 - \xi)^{-\lambda_H+r}_{\lambda_H-\lambda_L} \xi^{\lambda_H+r}_{\lambda_H-\lambda_L} V(\xi)
\]

After re-arranging and substituting in \( \Gamma(\xi; \eta) \), we get

\[
\left(e^{-\int a(\xi; \eta)d\xi}\right) \int \left(b(\xi; \eta)e^{\int a(\xi; \eta)d\xi}\right)d\xi = \Gamma(\eta; \xi)V(\xi) - c(1 - \Gamma(\eta; \xi))
\]
Combining all the parts gives us the general solution to the ODE

\[ U(\xi; \eta) = \Gamma(\xi; \eta) V(\xi) - c (1 - \Gamma(\xi; \eta)) + \Gamma(\xi; \eta) (1 - \xi) \frac{\lambda_{H+r}}{\lambda_{H-r}} \xi - \frac{\lambda_{L+r}}{\lambda_{L-r}} \xi \]

\[ + \Gamma(\xi; \eta) (1 - \xi) \theta(\xi) f(\eta) \quad (35) \]

### A.3 Value Matching and Smooth Pasting

So far, we have only found the general solution to the parametrised ODE that corresponds to the agent’s decision problem. In contrast to standard solutions to ODE, the general solution contains an unknown function \( f(\eta) \), rather than just an unknown constant. Nevertheless, for any value of \( \eta \), we can pin down the value of \( f(\eta) \), by imposing value matching and smooth pasting conditions. As a result, this defines \( f \) as a function of \( \eta \).

#### A.3.1 Value Matching and Smooth Pasting

For conciseness, define \( \theta(\xi) := (1 - \xi) \frac{\lambda_{H+r}}{\lambda_{H-r}} \xi - \frac{\lambda_{L+r}}{\lambda_{L-r}} \xi \). The general solution in equation 35 can then be expressed as

\[ U(\xi; \eta) = \Gamma(\xi; \eta) V(\xi) - c (1 - \Gamma(\xi; \eta)) + \Gamma(\xi; \eta) \theta(\xi) f(\eta) \]

We can solve for the unknown function \( f \) by imposing value matching \( U(\xi^*; \eta) = 0 \). Here, \( \xi^* \) is a belief where the agent is indifferent between the safe and risky arm. This gives us the following condition:

\[ 0 = \Gamma(\xi^*; \eta) V(\xi^*) - c (1 - \Gamma(\xi^*; \eta)) + \Gamma(\xi^*; \eta) \theta(\xi^*) f(\eta) \]

Because the value of \( \xi^* \) that solves the above equation depends on the value of the constant \( \eta \), strictly speaking, we should denote this threshold value as \( \xi^*(\eta) \). Solving the equation for
the unknown function \( f(\eta) \) gives

\[
f(\xi(\eta), \eta) = \frac{c(1 - \Gamma(\xi(\eta); \eta)) - \Gamma(\xi(\eta); \eta)V(\xi(\eta))}{\Gamma(\xi(\eta); \eta)\theta(\xi(\eta))}
\]

Substituting this back into \( U \) gives

\[
U(\xi; \eta) = \Gamma(\xi; \eta)V(\xi) - c(1 - \Gamma(\xi; \eta)) + \frac{\Gamma(\xi; \eta)\theta(\xi)}{\Gamma(\xi(\eta); \eta)\theta(\xi(\eta))} [c(1 - \Gamma(\xi(\eta); \eta)) - \Gamma(\xi(\eta); \eta)V(\xi(\eta))]\]

Next, we use the smooth pasting condition to find the cut-off belief \( \xi^* \), which satisfied

\[
U'(\xi^*; \eta) = 0.
\]

We evaluate the derivative at \( \xi^*(\eta) \), where we suppress the argument for brevity. Imposing smooth pasting \( U' = 0 \), gives

\[
0 = \Gamma'(\xi^*; \eta)V(\xi^*) + \Gamma(\xi^*; \eta)V'(\xi^*) + c\Gamma'(\xi^*; \eta) + \Gamma'(\xi^*; \eta)\theta(\xi^*) + [\Gamma'(\xi^*; \eta)\theta(\xi^*) + \Gamma(\xi^*; \eta)\theta'(\xi^*)] f(\xi^*, \eta)
\]

Substituting in \( f(\xi^*, \eta) \) and re-arranging, and then dividing by \( \Gamma(\xi^*; \eta) \) gives

\[
0 = (c + V(\xi^*)) \frac{\Gamma'(\xi^*; \eta)}{\Gamma(\xi^*; \eta)} + V'(\xi^*) + \left( \frac{\Gamma'(\xi^*; \eta)}{\Gamma(\xi^*; \eta)} + \frac{\theta'(\xi^*)}{\theta(\xi^*)} \right) \left[ c \frac{1 - \Gamma(\xi^*; \eta)}{\Gamma(\xi^*; \eta)} - V(\xi^*) \right]
\]

Since

\[
\frac{d\Gamma}{d\xi} = \frac{\lambda(\xi)\Gamma(\eta, \xi) (1 - \Gamma(\eta, \xi))}{\xi(1 - \xi)(\lambda_H - \lambda_L)}
\]

\[
\frac{d\theta}{d\xi} = \frac{\theta(\xi) (\lambda(\xi) + r)}{\xi(1 - \xi)(\lambda_H - \lambda_L)}
\]
the equation simplifies to

$$\Gamma(\xi^*; \eta) \left[ V(\xi^*)(\lambda(\xi^*) + r) + (1 - \xi^*)\xi^*(\lambda_H - \lambda_L)V'(\xi^*) \right] = rc \left( 1 - \Gamma(\xi^*; \eta) \right)$$ (36)

This equation determines the threshold value $\xi^*$ for any given value of $\eta$. To show the existence and uniqueness of $\xi^*$ on the domain $\xi \in [p_2^{**}, 1]$, first note that by Lemma 2 the expression in the square brackets on the left-hand side is strictly increasing in $\xi$. Also recall that $\Gamma$ satisfies properties $\Gamma'(\xi; \eta) > 0$, $\Gamma(0; \eta) = 0$ and $\lim_{\xi \to 1} \Gamma(\xi; \eta) = 1$. At $\xi = p_2^{**}$, the left-hand side of Equation 36 is strictly smaller than the right-hand side, while as $\xi \to 1$ the left-hand side is strictly greater than the right-hand side. Since all the terms are continuous, by the Intermediate Value Theorem there exists a solution $\xi^*$ to Equation 36 for each $\eta$.

Uniqueness follows from the fact that on the domain $[p_2^{**}, 1]$, the right-hand side of Equation 36 is strictly decreasing in $\xi$ since $\Gamma' > 0$. By Lemma 2 the left-hand side is strictly increasing in $\xi$. Thus there must exist a unique value $\xi^*$ which satisfies the equation.

Since $\eta$ fully determines $\xi^*$, we write $f^*(\eta) = f(\xi^*(\eta), \eta)$. Thus, the unique solution of the ODE that satisfies value matching and smooth pasting is given by

$$U(\xi; \eta) = \Gamma(\xi; \eta)V(\xi) - c \left( 1 - \Gamma(\xi; \eta) \right) + \frac{\Gamma(\xi; \eta)\theta(\xi)}{\Gamma(*) \theta(*) \eta} \left[ c \left( 1 - \Gamma(*) \eta \right) - \Gamma(*) \eta V(*) \eta \right]$$ (37)

More concisely, we can express this as

$$U(\xi; \eta) = \Gamma(\xi; \eta)V(\xi) - c \left( 1 - \Gamma(\xi; \eta) \right) + \Gamma(\xi; \eta)\theta(\xi)f^*(\eta)$$

A.3.2 Solution in the Belief Domain

To express the solution in equation 37 in terms of $p_1$ and $p_2$, we substitute $\eta = \eta(p_1, p_2)$ back into $U(\xi; \eta)$. In other words, while before we held $\eta$ fixed, we now let $p_1$ and $p_2$ vary freely and so these beliefs will now determine the value of $\eta$ according to 32. Next, we can substitute
out $\Gamma(\xi; \eta)$ which is given in equation 33, whereby $p_1 \equiv \Gamma(p_2; \eta(p_1, p_2))$. Nevertheless, note that $\Gamma(\xi^*(\eta(p_1, p_2)); \eta(p_1, p_2))$ is not identically equal to $p_1$, since $\xi^*(\eta(p_1, p_2))$ varies in both $p_1$ and $p_2$, and so $\xi^*$ is not equal to $p_2$ in general. With this in mind, equation 37 can be expressed as

$$U(p_1, p_2) = p_1 V(p_2) - c (1 - p_1) + p_1 \theta(p_2) f^*(\eta(p_1, p_2))$$

where

$$f^*(\eta(p_1, p_2)) := \frac{c (1 - \Gamma(\xi^*(\eta(p_1, p_2)); \eta(p_1, p_2))) - \Gamma(\xi^*(\eta(p_1, p_2)); \eta(p_1, p_2)) V(\xi^*(\eta(p_1, p_2)))}{\Gamma(\xi^*(\eta(p_1, p_2)); \eta(p_1, p_2)) \theta(\xi^*(\eta(p_1, p_2)))}$$

### A.3.3 Decision Boundary in Terms of $p_1$

Recall that we derived $\xi^*(p_1, p_2)$ from the smooth pasting equation

$$\Gamma(\xi^*; \eta) (V(\xi^*)(\lambda(\xi^*) + r) + (1 - \xi^*)\xi^*(\lambda_H - \lambda_L)V'(\xi^*)) = rc (1 - \Gamma(\xi^*; \eta))$$

On the boundary, we have $\Gamma(\xi^*; \eta(p_1, \xi^*)) = p_1$, whereby

$$p_1 (V(\xi^*)(\lambda(\xi^*) + r) + (1 - \xi^*)\xi^*(\lambda_H - \lambda_L)V'(\xi^*)) = cr (1 - p_1)$$

For a fixed $p_1$ the above equation gives the threshold value of belief $p_2$ where agents exit. This is denoted by $\xi^*$. Although we cannot write $\xi^*$ explicitly in terms of $p_1$ and the parameters, we can express the boundary values $p_1$ in terms of $p_2$. Thus, an explicit form for the decision boundary can be obtained from the smooth pasting equation which gives us

$$p_1^* = \frac{rc}{rc + V(p_2)(\lambda(p_2) + r) + (1 - p_2)p_2(\lambda_H - \lambda_L)V'(p_2)}$$

$\Box$
B Appendix

B.1 Risky Arm Beliefs

**Lemma 3.** In the absence of reward observations, the beliefs $p_{1,t}$ and $p_{2,t}$ are monotonically decreasing over time. Beliefs concerning the risky arm satisfy $p_{2,t} \leq p_{2,t}^{UC} \leq p_{2,0}$, for all $t$, where $p_{2,0}$ denotes the prior belief. The inequalities are strict for all $t > 0$ and hold with equality only at $t = 0$, and in the limit as $t \to \infty$.

*Proof.* The monotonic downward drift of $p_{1,t}$ and $p_{2,t}$ is immediate from the laws of motion 2 and 3. These downward drifts in turn imply that the prior $p_{2,0}$ acts as an upper bound for $p_{2,t}$ and $p_{2,t}^{UC}$, which is apparent from the functional form of equation 4. From equation 4 we also see that $p_{2,t} \leq p_{2,t}^{UC}$. The strict downward drift of $p_{1,t}$ and $p_{2,t}$, implies that $p_{2,t} < p_{2,0}$ for all $t > 0$. Given that $p_{2,t}^{UC}$ is a convex combination of $p_{2,t}$ and $p_{2,0}$, and that due to the downward drift we have $p_{2,t} \leq p_{2,0}$, it is clear that $p_{2,t} \leq p_{2,t}^{UC}$ for all $t \geq 0$. □

**Lemma 4.** Let $p_{2}^{\dagger}$ denote the value of $p_{2}$ that satisfies $p_{1,0} = p_{1}^{*}(p_{2})$, where $p_{1}^{*}(p_{2})$ is the function describing the decision boundary given in equation 14, and $p_{2}^{**}$ denotes the Stage II threshold belief. For all $t \geq 0$, the beliefs $p_{1,t}$, $p_{2,t}$ and $p_{2,t}^{UC}$ satisfy the following inequalities in the continuation region $\mathcal{D}$

$$
\begin{align*}
    p_{2}^{**} \leq p_{2}^{\dagger} \leq p_{2,t} \leq p_{2,t}^{UC} \\
    p_{1}^{*}(p_{2,t}^{UC}) \leq p_{1}^{*}(p_{2,t}) \leq p_{1,t} \leq p_{1,0}
\end{align*}
$$

All inequalities are strict for all times other than the moment of entry or exit.

*Proof.* WLOG, we assume that an agent enters at time $t = 0$ with priors $(p_{1,0}, p_{2,0}) \in \mathcal{D}$. In the first line, $p_{2,t} \leq p_{2,t}^{UC}$ follows from Lemma 3 in Appendix B. Given that $p_{1}^{*}(p_{2})$ is continuous, we know that there must exist some $p_{2}^{\dagger}$ such that $p_{1,0} = p_{1}^{*}(p_{2}^{\dagger})$. Otherwise, given
that \( p_1^*(p_2^{**}) = 1 \), it would then have to be the case that \( p_{1,0} < p_1^*(p_2) \) for all \( p_2 \), contradicting the assumption that \( p_{1,0} \in D \). Since by definition, \( p_2^\dagger \) lies on the boundary \( \partial D \), we get \( p_2^\dagger \leq p_{2,t} \), and by Corollary 2 we get \( p_2^{**} \leq p_2^\dagger \).

In the second line, given that \( p_1^*(p_2) \) is a strictly decreasing function, \( p_{2,t} \leq p_{2,t}^{UC} \) implies \( p_1^*(p_{2,t}^{UC}) \leq p_1^*(p_{2,t}) \). As long as \( p_{1,t} \in D \), we have \( p_1^*(p_{2,t}) \leq p_{1,t} \), and given the downward drift of \( p_{1,t} \), we have \( p_{1,t} \leq p_{1,0} \). 

\[ \square \]

**B.2 Proof of Lemma 2**

By Corollary 1 we have \( V, V' \geq 0 \) for all \( \xi \geq p_2^{**} \). This implies that \( \gamma(\xi) \geq 0 \) for all \( \xi \geq p_2^{**} \).

Next, we show that \( \frac{d^2 \gamma}{d\xi^2} > 0 \).

\[
\frac{d\gamma}{d\xi} = \frac{\left(\frac{1-\xi}{\xi}\right)^\mu (\mu + \xi)(h\lambda_H - c)(r + \lambda_L - (\lambda_H - \lambda_L)\mu) \left((\mu + 1)(h\lambda_H - c)\right) - \mu - 1}{\mu \xi} 
+ (\lambda_H - \lambda_L)(h(\lambda_H + \lambda_L + r) - c)
\]

To show that \( \frac{d\gamma}{d\xi} > 0 \), first note that when we evaluate the derivative at \( p_2^{**} \), we have

\[
\left. \frac{d\gamma}{d\xi} \right|_{\xi = p_2^{**}} = \mu h(\lambda_H - \lambda_L)^2 + (\lambda_H - \lambda_L)(h\lambda_H - c) > 0
\]

Next, we show that \( \gamma(\xi) \) is convex.

\[
\frac{d^2 \gamma}{d\xi^2} = \frac{\mu \left(\frac{1-\xi}{\xi}\right)^\mu (c - h\lambda_L)(-\lambda_H \mu + \lambda_L \mu + \lambda_L + r) \left((\mu + 1)(h\lambda_H - c)\right) - \mu}{(1 - \xi)\xi^2}
\]

Recall that in Stage II, \( \mu \) was defined as the positive root of the equation

\[
-\mu (\lambda_H - \lambda_L) + r + \lambda_L = \lambda_L \left(\frac{\lambda_L}{\lambda_H}\right)\mu
\]
Substituting the above expression into \( \frac{d^2 \gamma}{d \xi^2} \) gives
\[
\frac{d^2 \gamma}{d \xi^2} = \frac{\mu \left( \frac{1-\xi}{\xi} \right)^\mu (c - h \lambda_L) \lambda_L \left( \frac{\lambda_L}{\mu h} \right)^\mu \left( \frac{(\mu+1)(h \lambda_H - c)}{\mu (\mu - h \lambda_L)} \right)^{-\mu}}{(1 - \xi) \xi^2}
\]

Given Assumptions 2 and 1, imply that \( \frac{d^2 \gamma}{d \xi^2} > 0 \). Together, \( \frac{d \gamma}{d \xi} \bigg|_{\xi = \bar{p}_2^*} > 0 \) and \( \frac{d^2 \gamma}{d \xi^2} > 0 \) imply that \( \frac{d \gamma}{d \xi} > 0 \) for all \( \xi \geq p_2^{**} \).

\[\square\]

### B.3 Proof of Proposition 7

In order to reach a contradiction, suppose that \( \exists \bar{p}_2 > \bar{p}_2^* \), such that \( \lim_{n \to \infty} p_{2,t_n}^{UC} = \bar{p}_2 \). This implies that \( p_{1,0} - p_{1}^*(\bar{p}_2) > 0 \), given that \( p_{1}^* \) is decreasing in its argument. Thus, there exists \( \delta_1 > 0 \), such that the open disk \( B_\delta((p_{1,0}, \bar{p}_2)) \cap (0,1)^2 \subset \mathcal{D} \), for any \( \delta \in (0, \delta_1) \). Fix a \( \delta \in (0, \delta_1) \). Since by Proposition 6 \( p_{2,t_n}^{UC} > p_{2,t_{n+1}}^{UC} \), and because \( \bar{p}_2 \) is a limit point, there exists a \( N_0 \in \mathbb{N} \) s.t. \( p_{2,t_n}^{UC} \in (\bar{p}_2, \bar{p}_2 + \delta) \) for all \( n > N_0 \). For simplicity, let \( p_{2,0} \) denote the first element along the sequence \( \{ p_{2,t_n}^{UC} \}_{n=1}^\infty \) s.t. \( p_{2,t_n}^{UC} \in (\bar{p}_2, \bar{p}_2 + \delta) \).

Using equations 2 and 3, we can express the law of motion of \( p_{2,t}^{UC} \) in differential form as
\[
dp_{2,t} = -p_{1,t} [(\lambda_H - \lambda_L)p_{2,t}(1 - p_{2,t}) - (p_{2,t} - p_{2,0}) \lambda(p_{2,t})(1 - p_{1,t})] \quad (38)
\]

Recall from Proposition 1, that \( \exists ! t^* > 0 \) that \( dp_{2,t}^{UC} < 0 \), \( \forall t \in [0, t^*) \), and \( dp_{2,t}^{UC} > 0 \), \( \forall t \in (t^*, \infty) \). Furthermore, \( t^* \) is independent of \( p_{2,0} \) and only depends on the fixed prior of the agent’s type \( p_{1,0} \).

**Case 1:** Fix an arbitrary pair of priors \( (p_{1,0}, p_{2,0}) \), and suppose that exit happens at some time \( t_n^* \in [0, t^*) \). As \( (p_{1,0}, p_{2,0}) \in \mathcal{D} \cap (0,1)^2 \) holds by assumption, it is clear from equation 38 that at time \( t = 0 \), \( dp_{2,0}^{UC} < 0 \) for all \( p_{2,0} \in [\bar{p}, \bar{p} + \delta] \). Define the following bounds on the
time derivatives of the beliefs:

\[
\frac{\dot{p}_2^\text{UC}}{dt} := \max_{p_2, 0 \in [\tilde{p}_2, \bar{p}_2 + \delta], t \in [0, T]} \frac{dp_{2,t}^\text{UC}}{dt}
\]

\[
\frac{\dot{p}_2}{dt} := \min_{p_2, 0 \in [\tilde{p}_2, \bar{p}_2 + \delta], t \in [0, T]} \frac{dp_{2,t}}{dt}
\]

Since the max is taken over a closed set, \( \frac{\dot{p}_2^\text{UC}}{dt} \) and \( \frac{\dot{p}_2}{dt} \) exist and \( \frac{\dot{p}_2^\text{UC}}{dt} < 0 \). Because it is always the case that \( \frac{dp_{2,t}}{dt} \leq \frac{dp_{2,t}^\text{UC}}{dt} \), we have

\[
\frac{\dot{p}_2}{dt} \leq \frac{dp_{2,0}}{dt} \leq \frac{dp_{2,t}^\text{UC}}{dt} \leq \frac{dp_{2,t}^\text{UC}}{dt}
\]

Also note that \( \exists \alpha \in (0, 1) \) which guarantees that

\[
p_2,0 + \alpha dp_2 \in D \quad \forall dp_2 \in \left[ \frac{\dot{p}_2}{dt}, \frac{\dot{p}_2^\text{UC}}{dt} \right], \forall p_{2,0} \in [\tilde{p}_2, \bar{p} + \delta]
\]

Now, since \( p_{2,t_n^*}^\text{UC} \to \tilde{p}, \exists N_1 > N_0, \text{ s.t } 0 < |\bar{p} - p_{2,t_n^*}^\text{UC}| < \frac{\alpha \frac{dp_{2,t}^\text{UC}}{dt}}{2} \) for all \( n > N_1 \). For simplicity, we re-label the first time \( t_n^* \) that an incoming agent satisfies the inequality, by \( t = 0 \). Given that \( p_{2,0}^\text{UC} + \alpha dp_2 \in D \), we know that \( p_{2,t}^\text{UC} \) decreases by an amount, which is at least \( \frac{dp_{2,t}^\text{UC}}{dt} < 0 \) while still remaining in the continuation region \( D \). Since \( dp_{2,t}^\text{UC} \leq 0 \) for all \( t \in [0, t^*] \), we know that if the agent exits at some point in time contained in the interval \([0, t^*]\), it must be the case that at the time of exit \( t^*_n > 0 \), it is the case that \( p_{2,t_n^*}^\text{UC} < \tilde{p} \). Contradiction.

Case 2: Fix an arbitrary pair of priors \((p_1,0, p_2,0)\), and suppose that exit happens at some time \( t_n^* > t^* \), when \( dp_{2}^\text{UC} \geq 0 \). Since exit occurs in finite time, Proposition 6 guarantees that \( p_{2,t_n^*}^\text{UC} < p_{2,0} \).

Now note that

\[
p_{1,t} \geq \tilde{p}_1 := \lim_{p_2 \to 1} p_{1,t}^\text{UC}(p_2) = \frac{rc}{rc + (h\lambda_H - c)(\lambda_H + r)} \in (0, 1) \quad \forall t \geq 0
\]
Thus, when \( t > t^* \) we have

\[
p_{2,t}^{UC} = p_{2,t}p_{1,t} + p_{2,0}(1 - p_{1,t})
\]

\[
\leq p_{2,t}\hat{p}_1 + p_{2,0}(1 - \hat{p}_1)
\]

\[
< p_{2,0}
\]

Given that \( p_{2,t} \) is strictly decreasing and \( \hat{p}_1 \) is a global lower bound, we have

\[
p_{2,0} - p_{2,t_n}^{UC} \geq (p_{2,0} - p_{2,t_n}^{*})\hat{p}_1
\]

\[
\geq (p_{2,0} - p_{2,t^*})\hat{p}_1
\]

since by \( t_n^* > t^* \) we have \( p_{2,t^*} > p_{2,t_n^*} \).

Using the expression for \( p_{2,t} \) from Lemma 1, we can define a minimum bound for \( (p_{2,0} - p_{2,t^*}) \) as follows

\[
\hat{p}_2 := \min_{\tilde{p}_{2,0} \in [p_{2,t_n^*}^{*}, p_{2,0}]} \left\{ \tilde{p}_{2,0} - \left[ 1 + \left( 1 - \frac{\tilde{p}_{2,0}}{p_{2,0}} \right) e^{k(\lambda_H - \lambda_L)t^*} \right]^{-1} \right\}
\]

Recall that the threshold belief of Stage II always satisfies \( p_{2,t_n}^{**} \leq p_{2,t} \). Let \( \tilde{p}_{2,0} \) denote the prior of the first agent in the sequence. By assumption \( \tilde{p}_{2,0} \in (0, 1), \) and given that \( p_{2,t_n}^{UC} > p_{2,t_{n+1}}^{UC} \) holds for all \( n \), we also have \( \tilde{p}_{2,0} > p_{2,t_{n+1}}^{UC} \) for all \( n \). It is clear that \( \tilde{p}_2 \in (0, 1) \). As a result, we know that if there is an exit at time \( t_n^* > t^* \), then, between any two consecutive agents, the common belief must have decreased by an amount which is at least \( (p_{2,0} - \tilde{p}_2)\hat{p}_1 > 0 \).

Let \( \delta := (p_{2,0} - \tilde{p}_2)\hat{p}_1 \). Then \( \exists N \in \mathbb{N}, \text{ s.t. } n \geq N \) implies \( |p_{2,t_n}^{UC} - \tilde{p}| < \frac{\delta}{2} \). If agent \( n \) enters with the prior \( p_{2,t_n}^{UC} \), then agent \( n + 1 \) will enter with a prior \( p_{2,t_{n+1}}^{UC} \) such that \( p_{2,t_{n+1}}^{UC} < \tilde{p} < p_{2,t_n}^{UC} \). Contradiction.

Thus, we have showed that there cannot exit a limit point \( \tilde{p} \) such that \( \tilde{p}_2 > \tilde{p}^t \). Similarly, there cannot exit a limit point \( \tilde{p} \) such that \( \tilde{p} < \tilde{p}^t \). This follows by definition, as we defined
$p^\dagger$ as the belief that satisfies $p_{1,0} = p_{1}^\dagger(p_2^\dagger)$. This simply says that the point $(p_{1,0}, p_2^\dagger)$ lies on $\partial C$ and hence triggers immediate exit. By Proposition 6, we know that there is an infinite sequence of priors $\{p_{2,t_n}^{UC}\}_{n \in \mathbb{N}}$ where agents play the risky arm for a positive length of time. Thus it must be the case that $p_{2,t_n}^{UC} > p^\dagger_2$ for all $n$. Thus, it can never be the case that $\lim_{n \to \infty} (p_{2,t_n}^{UC}) = \bar{p}_2 < p^\dagger_2$.

Uniqueness follows from the above arguments. This completes the proof that $\lim_{n \to \infty} (p_{1,0}, p_{2,t_n}^{UC}) = p^\dagger_2$. 

\[\square\]

63
References


