Obtaining Experience in a Multi-Armed Bandit

Mikhail Safronov*

November 14, 2014

Abstract

I study a multi-armed bandit problem in which an agent by experimenting with an arm becomes experienced in this arm. The event of becoming experienced is exponentially distributed random variable. When the agent becomes experienced, the chances of getting the prize at the arm increase. I argue that experience makes even an inexperienced agent less willing to experiment, compared to the model with the constant chances of getting the prize. Perhaps interestingly, this effect is strongest for intermediate arrival rates of experience. This implies that among the arms of similar quality the agent will prefer to experiment with the ones that have extreme arrival rates.

I extend the model to a continuum of agents, where the more people pull an arm, the less chance one has to obtain the prize. Agents differ in talent (the arrival rates of experience). I study the distribution of agents’ talents over the arms following an appearance of a new arm. One of the results shows that even if the new arm is ex ante identical to the existing arm, in the long run it will have fewer people with high talent. This effect is driven by congestion and reduced willingness to experiment. In a benchmark case without experimentation, when agents know their private value at the new arm, in the long run there may be fewer or more people with high talent at the new arm.

1 Introduction

Multi-armed bandit model is an important framework for analyzing experimentation. It has been used in a variety of economic applications, such as product search, clinical trials, monopolist learning the market demand, corporate finance. The main idea behind these models is that the agent faces

*Department of Economics, Northwestern University, 2001 Sheridan Rd., Evanston IL 60208-2600; E-mail: mikhailsafronov2014@u.northwestern.edu. I am grateful to Wojciech Olszewski and Bruno Strulovici for insightful conversations and constant support. I would like to thank Jeff Ely and Eddie Dekel, and my peers for insightful discussions and comments. Please visit http://gradstudents.wcas.northwestern.edu/ mvs902/indexjm.html for the most current version.
a trade-off between exploration (trying the new arms to learn about them) and exploitation (trying the best known arm). In this paper a new effect is introduced: by pulling an arm, the agent not only learns about its payoff, but also gets better at pulling that arm, which increases the payoff.

The idea that an agent becomes better, or experienced at pulling the arm, seems a natural addition to the multi-armed bandit model. One example is an agent who has started working as a door-to-door salesman. He learns how much people like to buy the goods he sells. At the same time, he becomes better at advertising his goods. With time, he succeeds more often in selling and his profits increase. The salesman may decide to change the type of goods he sells or begin selling in another neighbourhood, expecting higher profits. However, by switching his current occupation, the agent abandons the experience gained from the past job, and has to learn the specifics of the new market.

Another example of an agent becoming experienced in a multi-armed bandit model is a scientist. She chooses an area of research, for example, she is an economist and decides to focus on game theory. The agent will read the literature, develop projects, learn the questions of an interest in the area. With time, the agent will become a better game theorist and will have greater chances of publishing in top journals. However, the agent may decide to switch to another area, for example, decision theory, in which case she has to become experienced in the new area to achieve the same level of proficiency.

One of the main results of the paper is that adding the possibility of becoming experienced into the multi-armed bandit model makes the agent less willing to experiment. This result is driven by two effects. There is a straightforward ex post effect: if the agent has become experienced at the arm, he is less willing to experiment with the other arms. However, there is also not so obvious ex ante effect: let the agent quickly learn the payoff at the first arm he pulls, without obtaining any experience at the first arm. Then the possibility of becoming experienced makes him less willing to experiment with the other arms. If the agent starts experimenting with the second arm, by the time he learns the payoff he may become experienced at the second arm. It may happen, that the (initial) payoff at the second arm is lower than at the first arm. In this situation, if the agent switches back to the first arm, he abandons experience at the second arm. If the agent stays at the second arm, that arm provides lower (initial) payoff. At the same time, if the agent kept pulling the first arm, he would have become experienced there, and obtained higher payoff. The possibility of becoming experienced at the second arm with the low payoff decreases the agent’s incentives to experiment.

In this paper, I first study a single agent multi-armed bandit model, and later extend it to a continuum of agents. The agent can pull one of several arms, each arm generating the payoff as a
Poisson event. When pulling an arm, the agent may become experienced at that arm; the arrival of experience is exponentially distributed random event. When the agent becomes experienced, the payoff at the arm increases. The agent observes the moment of becoming experienced. The events of learning the payoff and becoming experienced at the arm are independent. In particular, the agent may first become experienced at the arm, and then learn its payoff.

In a single-agent model, I show that the negative effect of experience on the incentives to experiment is non-monotone in the arrival rate of experience. This effect is large for intermediate arrival rates, and small for extreme arrival rates. When the agent chooses from the arms with ex ante equal payoffs, he prefers to experiment with the arms with the extreme arrival rates of experience. The ex ante effect of experience on experimentation vanishes when the agent cannot switch back to the arms he previously pulled.

One way to interpret experience at an arm increasing the payoff, is to assume that when the arm generates a prize, the agent only has occasional chances to receive the prize. Experience increases this chance, and therefore the expected payoff at the arm. Experienced agent “succeeds” at the arm more often. For example, during each day a salesman can visit a certain number of houses, offering his good. Each time he attempts to sell the good, he either succeeds or not. Experience increases the probability of a sale and the expected profit of a salesman, even though he is limited to a certain number of visits per day due to time constraints.

I also study a multi-agent setting in which agents exert negative externalities on each other. Each arm generates a limited number of prizes, and if it is pulled by more agents, the chance that a single agent will obtain the prize becomes lower. If there are more salesmen in the neighbourhood, each of them has less chances to sell the good. If many scientists work in game theory, it is hard to find the unexplored idea. I assume that the experience helps the agent to win the competition, and it increases the probability of the agent receiving the prize, similar to single agent case.

In the model agents differ by the arrival rate of experience, called talent in this setting. Agents with high talent are called talented. The agent is assumed to know her own talent. For example, when studying at school, the agent understands how much time it takes her to learn new skills. Each agent can pull one of two arms: an old arm which has existed for some time, or a new arm which has just arrived. Due to congestion, and incentives to experiment, some of the agents switch to the new arm. One of the results shows that even if both old and new arms are ex ante identical, they may have different long-run distributions over talents of agents pulling them. In case the payoffs at the arms do not vary much for different agents, the old arm will have more talented people. The experience in this setting matters more than the value of the payoff. When experimenting with the
new arm with (relatively) low payoff, the agent may first become experienced and then learn the payoff. In this case the agent will keep pulling the new arm with the low payoff. Talented agents have higher arrival rates of experience, and are more likely to end up pulling the arm with the low payoff, if they experiment with the new arm.

When all prizes give the same known value to every agent, in equilibrium at any time the chance that an (in)experienced agent gets the prize is the same for both arms. Any inexperienced agent, regardless of the talent, will be indifferent between the arms at any time. In case the prizes at each arm give the same known value to every agent, but different arms generating different prizes, in equilibrium the arm with the higher prize will have more talented agents pulling it. In case prizes at each arm give different values to different agents, the long-run distribution over the prize values among the agents pulling the arm, depends on when exactly the prizes arrive at each arm. For each realisation of the times of prize arrival, the final distribution is characterized by the cutoff value of the prize, all agents who value the prize below the cutoff leave the arm. The lower (upper) bound on this cutoff makes experienced agent to have the same incentives to stay at the arm, as the inexperienced agent with unknown (maximal) value of the prize.

Literature Review. The possibility of becoming experienced is captured by the classic multi-armed bandit model. The feature of this model is that the distribution of payoffs from any arm changes when the arm is pulled, and it can mean as well that these payoffs increase with experience. The theorem by Gittins and Jones (1974), one of the most important results in multi-armed bandit models, states that with several arms to pull, each arm can be given an index, and the arm with the highest index is optimal to be pulled. At each arm one finds the optimal stopping time, such that the discounted expected payoff per period is the highest. That payoff value is called the Gittins index, and it depends only on the current state of the arm. An alternative derivation of the Gittins index was made by Whittle (1982), where for any arm $i$ a safe arm is considered. The payoff at the safe arm is such that the agent in the imagined environment with only arm $i$ and the safe arm is indifferent between pulling the safe arm and experimenting with arm $i$ (with a potential of switching to the safe arm). That payoff is the Gittins index. A survey on the literature on multi-armed bandits and Gittins index can be found in the paper by Bergemann and Valimaki (2006).

The possibility of becoming experienced in the model of multi-armed bandit seems to occur in most applications. Yet, to my knowledge, there is not much literature on this topic, with some exceptions. The first one is the paper by Fryer and Harms (2013), where they introduced experience in the experimentation problem. In their paper experience at the arm decreases over time, if that arm is not being pulled. The Gittins index cannot be used in general in this type of model. The
main result of Fryer and Harms is the derivation of an alternative index, that can be used for the choice of the arm. In my model the main attention is paid to incentives of agents to experiment, and to the dynamics of switching the arms due to congestion.

The other papers that consider agents becoming experienced are the papers on labor economics. Yamaguchi (2012) considers the model with workers, who can choose the sector of occupation, and the complexity of task within the sector. The complex tasks help worker to build their skills faster; and, if the worker switches to another sector, he cannot use his accumulated skill at the old sector. Kambourov and Manovskii (2009) show that when workers choose new jobs in the same sector, their wage is higher compared to the ones who switch the type of job. Sanders (2010) develops a model where employees by working in the sector both learn their skills at the sector and increase them at the same time. He estimates the influence of both effects on worker’s mobility. Workers learning their skills makes them change jobs across sectors, while workers increasing their skills makes them choose the jobs with more complex tasks within the sector.

There are papers on multi-armed bandits, which make an assumption similar to experience: they introduce costs of switching between the arms. The experience at the arm endogenously decreases the agent’s incentives to experiment, while the costs do so explicitly. The main question studied in these papers is that one cannot use Gittins index, and the alternative methods of finding the optimal solution in experimentation problem have to be developed. Banks and Sundaram (1994) prove that the optimal solution in the model with the switching costs cannot be indexable. Jun (2004) provides a survey on the bandits with switching costs.

Another related paper on multi-armed bandits is by Thomas (2012), where congestion is introduced in the multi-armed bandit model. There are 2 agents and 3 arms: one risky arm for each agent, and the common safe arm. If one agent occupies the safe arm, the other cannot pull it. The distinctive result is that the agent who is more likely to succeed in own risky arm, nevertheless pulls the safe arm, to make the opponent experiment more and ease the pressure on the safe arm. The current paper differs from the one by Thomas by having the possibility of becoming experienced, and studying a continuum of agents.

There are papers on multi-armed bandits with no experience, but similar in style. Bolton, Harris (1999) and Keller, Rady, Cripps (2005) consider strategic experimentation with N agents and 2 arms. Keller, Rady (2010) study strategic experimentation with Poisson bandits. In the paper by Rosenberg, Solan, and Vieille (2007) the agents observe the actions, but not the payoffs of opponents. Strulovici (2010) considers the voting model with many agents and determines the incentives for collective experimentation. These papers contain similar techniques to the ones used in the current paper.
The paper is organized as follows. Section 2 describes the models with a single agent and a continuum of agents. Section 3 is devoted to the decision problem of a single agent. Part 3.1. describes the effect of experience factor making the agent less willing to experiment. In part 3.2. I show the non-monotonicity of the above effect on the arrival rate of experience. In part 3.3. one considers alternative information structures and how they affect the results. Section 4 describes the case with a continuum of agents and two symmetric arms, with part 4.1. considering the special case with no experimentation. Then, part 4.2. is devoted to the dynamics in case of small relative spread of prizes for different agents at any arm; part 4.3. considers the limit distributions of prize valuations. Part 4.4. discusses the extension for asymmetric arms. Section 5 concludes.

2 Model

2.1 Model for a single agent

An infinitely living agent experiments with \( I \) arms, denoted as \( i \in \{1, 2, ..., I\} \). Time is continuous, and the agent has discount factor \( r \). The agent may pull one arm at a time, and may switch arms at no costs.

Each arm \( i \) generates prizes as a lump sum, with Poisson intensity \( \lambda_i \), known to the agent. If the agent is pulling the arm at the moment of prize arrival, he gets the prize. It will be assumed that the value of the prize is the same for all arrivals, but ex ante unknown to the agent. Therefore by observing the first prize, the agent learns its value for all future prizes as well.

When pulling arm \( i \), the agent may also become experienced at that arm. Experience at arm \( i \) is a binary variable \( X_i \in \{0, 1\} \), with \( X_i = 0 \) meaning that the agent is inexperienced, and \( X_i = 1 \) meaning that the agent is experienced. Experience of the agent at all arms is a vector \( X \in \{0, 1\}^I \). Obtaining experience at arm \( i \) does not affect \( X_j \) for \( j \neq i \), and once the agent becomes experienced, the value of \( X_i \) remains 1 forever. The event of the agent becoming experienced is exponentially distributed with the rate \( \mu_i \), over the time agent spends pulling arm \( i \). That is, if the agent has been pulling arm \( i \) for time \( T \), the probability of agent becoming experienced during that time is \( 1 - e^{-\mu_i T} \). The agent knows \( \mu_i \) and observes the moment of becoming experienced. The process of becoming experienced at an arm is memoryless: if the agent has been pulling the arm for some time, and has not become experienced, the time spent at the arm does not increase the chance of becoming experienced in the future. However, the main result of the model works in case of continuous experience arrival as well.

Experience at the arm increases agent’s payoff as described below. When the prize arrives at
arm $i$ the payoff of an *inexperienced* agent equals $\theta_i$. The payoff $\theta_i$ will be also referred to as the agent’s *type* at arm $i$. Before receiving the first prize, the agent does not know the type $\theta_i$, but has a prior $F_i(\theta_i)$ over the interval $[\theta_i, \bar{\theta}_i]$, where $\theta_i \geq 0$, with density $f_i(\theta_i)$. If the agent is experienced at arm $i$, then he obtains higher payoff $m_i \theta_i$ when the prize arrives. The value $m_i > 1$ is known ex ante, and it shows the gain from experience.

The processes of becoming experienced and prize arrival are independent. In particular, each of the two events of the agent obtaining experience and learning his type at the arm may occur first.

One way to interpret the assumption of experience increasing the payoff at the arm in a multiplicative way is to assume that the agent has only some chance to receive the prize when it arrives, and that chance increases with experience. In this case type $\theta$, the payoff for an inexperienced agent, contains both the value of the prize and the chance of an inexperienced agent to obtain it. This case is motivated by the examples of a scientist publishing in a good journal or a salesman selling good. From time to time these people have a potential opportunity to succeed, but they may not make use of it. Experience increases their chances, and this is reflected in the model. Moreover, in the professions described above the agent may not receive the prize himself, but observe others do so and realize how much he likes it. A young scientist can look at more experienced colleagues, publishing in good journals and receiving grants, and can get an idea how much he would like it. Thus, it is possible for the agent to get information about the prize without actually consuming it.

### 2.2 Model for a continuum of agents

In this part I study a market setting with a continuum of agents of total mass 1, each pulling one of $I$ arms. As in the single agent setting, each agent can try one arm at a time, with no costs of switching. I assume that each agent has the same discount factor of $r$. Agents, however, may have different valuations of prizes and rates of becoming experienced. These parameters are independent across agents.

Agents will influence payoffs of one another by congestion. The more agents try the arm, the lower payoff each of them gets. I will consider a specific type of congestion, in which each arm $i$ is characterized by a *capacity* $\alpha_i$. When arm $i$ generates prizes, if the mass of people trying this arm at this moment is smaller than $\alpha_i$, then every person will receive the prize (one prize for one agent). However, if the mass of people exceeds the capacity of the arm, then only $\alpha_i$ of people will receive the prize, others will not. The related environment would be that if there are few salesmen in the neighbourhood, they do not affect each other. However, if there are many salesmen, the individual profit is small.
Each agent at each arm can be either inexperienced or experienced. I will denote as $\beta_i(t)$ and $\gamma_i(t)$ the masses of inexperienced (at arm $i$) and experienced agents pulling arm $i$ at time $t$. Time index $t$ will be omitted, if it will cause no confusion. I assume that experience helps the agent to win the congestion, more specifically, experienced agents will have the priority in obtaining the prizes. That is, when the prizes arrive, the chance that inexperienced agent receives the prize is:

$$
\begin{align*}
&\begin{cases}
1, & \text{if } \beta_i + \gamma_i \leq \alpha_i \\
(\alpha_i - \gamma_i)/\beta_i, & \text{if } \gamma_i \leq \alpha_i < \beta_i + \gamma_i \\
0, & \text{if } \gamma_i > \alpha_i
\end{cases} \\
&\text{and the chance that experienced agent receives the prize, is}
\end{align*}
$$

$$
\begin{align*}
&\begin{cases}
1, & \text{if } \gamma_i \leq \alpha_i \\
\alpha_i/\gamma_i, & \text{if } \gamma_i > \alpha_i
\end{cases}
\end{align*}
$$

This specific type of congestion works differently depending on whether mass of experienced people $\gamma_i$ is larger or smaller than the capacity $\alpha_i$. If $\gamma_i > \alpha_i$, then inexperienced people do not receive any prizes at all. Alternatively, if $\gamma_i \leq \alpha_i$, then each experienced agent receives a prize with probability 1.

With congestion, experience affects the probability of agent receiving the prize, as in the single agent case. In the single agent case, however, the probability of receiving the prize was constant, given the experience. With a continuum of agents, this probability may change over time.

Working with a continuum of agents, I will assume that the rate of becoming experienced for each agent is the same across all arms and it will be denoted by $\mu$. However, $\mu$-s can be different for the different agents. Each agent knows own $\mu$, as well as the distribution across $\mu$-s across agents. The rate $\mu$ will also be referred to as the talent of agent. The more talented agent is, the less time it is expected for her to obtain experience in any activity.

At any moment, each arm will be pulled by some mass of agents. These agents can be distinguished further by their experience values. Conditional on the experience value and the arm being pulled, there will be a distribution over talents $\mu$ across agents. The latter distributions determine the congestion at the moment, as well as the aggregate rate at which agents are becoming experienced at different arms. These distributions are taken into account by all agents, when they maximize their payoff.

Another factor determining agents’ behavior is their private types. A private type of an agent at an arm is how much she values the prize at the arm. Among the agents pulling an arm, there may be ones who already know their private types, and may be those who don’t. When the prize arrives at the arm, any agent who is pulling the arm at the moment, learns her private type. As
in the single agent case, it will be assumed that the value of the prize does not change over time. The configuration of agents with different private types, as well as different talents and values of experience, is evolving with time, as agents are maximizing their payoffs.

In this paper I restrict attention to two arms: an old arm which has existed for some time, and a new arm, which has just appeared. Every agent knows her private type at the old arm, but no one knows private type at the new arm. At the old arm some agents are experienced, others are inexperienced; and no one is experienced at the new arm. I will be interested in the dynamics of agents’ switching to the new arm, including the long-run configuration.

3 The single agent case

In this section an agent experiments with \( I \) arms, denoted by \( i \in \{1,...,I\} \). Each arm \( i \) generates prizes, and conditional on prize arrival, the (inexperienced) agent gets expected payoff \( \theta_i \), which will also be referred to as the agent’s type at arm \( i \). At any moment arm \( i \) is characterized by the type \( \theta_i \) (or prior, if the agent has not learnt \( \theta_i \) yet) and experience value \( X_i \). The variables \( (\theta_i, X_i) \) constitute the state of arm \( i \). The state of arm \( i \) is changed only when arm \( i \) is pulled, and it does so in a way that depends only on the state itself. The results of the paper by Gittins (1979) imply that in this model each arm can be put in correspondence with a Gittins index, and the agent’s optimal behavior in experimentation is to pull the arm with the highest Gittins index. In the current paper the Gittins index will be calculated as follows. For each arm \( i \) one will consider an imaginable environment with arm \( i \) and a safe arm, the payoff of the safe arm making the agent indifferent between pulling the safe arm or experimenting with arm \( i \). The payoff of the safe arm is the Gittins index of arm \( i \).

I will start with the model with two symmetric arms and consider two specifications: with and without the possibility of becoming experienced at the arms. I will demonstrate that adding experience factor in the model reduces agent’s incentives to experiment. The effect of experience on experimentation appears to be non-monotone with the arrival rate of experience. In the next part of the section the influence of experience factor on the Gittins index of the arms will be considered. If the agent has to choose between ex ante identical arms, it will be shown, that it is optimal for the agent to pull the arms with the extreme arrival rates of experience. The last part of the section is devoted to the models with alternative information structures. The effect of experience decreasing the agent’s incentives to experiment appears to be robust.
3.1 Two symmetric arms

In this part I consider the case of two symmetric arms, denoted by $i, j$. Both arms provide the same ex ante distribution of the payoff $\theta$, with density $f$ over the interval $[\underline{\theta}, \overline{\theta}]$, where $\underline{\theta} > 0$. Each arm provides the same arrival rate of experience $\mu$ and the same Poisson intensity $\lambda$ of generating the prize. Becoming experienced increases the expected payoff of the agent, conditional on prize arrival, by the factor $m$ at both arms.

Let us first solve the agent’s problem for the case when there is no experience. Since the arms are symmetric, the agent is initially indifferent between which one to pull. Without loss of generality suppose he pulls arm $i$. The agent pulls arm $i$ until observing the type $\theta_i$. Once he does so, he has to decide between two options. The first option is to keep pulling arm $i$ forever and periodically enjoy consumption of the known payoff at this arm. The second option is to switch to arm $j$, and keep pulling it until observing the payoff of arm $j$ as well. Then, once the agent learns his payoffs at both arms, he will choose the arm with the higher value $\theta$ (or break ties at random if $\theta_i = \theta_j$). The main decision of the agent is whether to experiment with arm $j$, once he has learned the type $\theta_i$ at arm $i$.

It is immediate that the agent’s optimal strategy is to pull arm $j$ if and only if arm $i$’s type $\theta_i$ is below some cutoff $\theta^\ast$. When $\theta_i = \theta^\ast$, agent is indifferent between experimenting with arm $j$ or keeping pulling arm $i$. In this case any choice of the arm will lead to the same value of expected discounted payoff of the agent, or the same continuation payoff of the agent. This indifference condition allows to derive the expression for the cutoff $\theta^\ast$, which is done below.

In case $\theta_i = \theta^\ast$ the agent may choose to keep pulling arm $i$ forever. His continuation payoff, denoted as $V$, satisfies the following condition:

$$V = \lambda dt \theta^\ast + (1 - r dt)V$$  \hspace{1cm} (1)

Expression (1) means that when pulling arm $i$ for an infinitesimal time interval $dt$, the chance of the prize arrival is $\lambda dt$. Conditional on prize arrival, the agent gets the expected payoff of $\theta^\ast$. After $dt$ has passed, the state of arm $i$ does not change, and therefore the agent’s continuation payoff $V$ remains the same. Hence, the value of $V$ at the current moment equals the expected payoff the agent receives during time $dt$, plus the continuation payoff after $dt$ has passed, calculated with the discount $1 - r dt$.

One can subtract $(1 - r dt)V$ from both parts of (1):
and obtain the expression for $V$:

$$V = \frac{\theta^* \lambda}{r} \tag{2}$$

Expression (2) shows the continuation payoff of the agent pulling arm $i$ with known type $\theta^*$. The higher is the intensity $\lambda$ of prize arrival, the more often the agent will receive the prize, and the higher is the continuation payoff.

Let us find now the agent’s continuation payoff if he decides to experiment with arm $j$, conditional on its type being $\theta_j$. Let’s denote that continuation payoff as $V_1$, in case when type $\theta_j$ is bigger than the cutoff $\theta^*$. Once the agent learns $\theta_j$, he will stay at arm $j$. In other words, once the agent start experimenting with arm $j$, he will pull it forever. One already knows from (2), that in this case the continuation payoff of the agent will be

$$V_1 = \frac{\theta_j \lambda}{r} \tag{3}$$

If, however, the type $\theta_j$ is lower than the cutoff $\theta^*$, then once the agent learns $\theta_j$, he will switch back to arm $i$ and stay there. This means that at the moment of prize arrival at arm $j$ the agent gets the expected payoff $\theta_j$ and then switches to arm $i$. Let us denote the continuation payoff of the agent when he starts experimenting with arm $j$ with low type $\theta_j < \theta^*$ as $V_2$. Then one has:

$$V_2 = \lambda dt \theta_j + \lambda dt \left( 1 - \frac{\theta^* \lambda}{r} \right) + (1 - \lambda dt)(1 - rd) V_2 \tag{4}$$

Expression (4) means that when experimenting with arm $j$, within the infinitesimal time interval $dt$ the agent observes the type $\theta_j$ with the probability $\lambda dt$, and gets the expected payoff $\theta_j$. In this case after learning the type $\theta_j$, the continuation payoff of the agent after period $dt$ has passed, is given by (2), and equals $\frac{\theta^* \lambda}{r}$. With probability $(1 - \lambda dt)$, however, the agent does not learn $\theta_j$. His continuation payoff therefore does not change, and remains $V_2$.

Let us subtract the value $V_2$ from both sides of the expression (4). What is left are variables which are either multiplied by $dt$, or by $dt^2$. As the time interval $dt$ can be chosen arbitrarily small, if the expression (4) is the equality, it will remain the equality if one disregards terms with $dt^2$. As a result, one has:

$$0 = \lambda dt \theta_j + \lambda dt \left( \frac{\theta^* \lambda}{r} \right) - (\lambda + r) dt V_2$$
Now one can derive the value of \( V_2 \), and rewrite it as:

\[
V_2 = \frac{\lambda \theta_j + \lambda \left( \frac{\theta^* \lambda}{r} \right)}{\lambda + r} = \theta_j \frac{\lambda}{\lambda + r} + \theta^* \left[ \frac{\lambda}{r} - \frac{\lambda}{\lambda + r} \right]
\]

(5)

The expression (5) means that the low payoff \( \theta_j < \theta^* \) is obtained by the agent with the "weight" \( \frac{\lambda}{\lambda + r} \), and the same weight is taken from the payoff \( \theta^* \) at arm \( i \), compared to (2).

The expression (5) was obtained in a way similar to the expression (2): one evaluates what happens during an infinitesimal interval \( dt \), subtracts the value function, and then limits \( dt \) to 0, disregarding the terms with \( dt^2 \). The same technique will be used later in the paper, and sometimes the derivations will be put in the appendix.

Having calculated the agent’s continuation payoff when he tries arm \( j \), given the private type \( \theta_j \); one derives the ex ante continuation payoff, denoted by \( V_j \), when the type \( \theta_j \) is unknown. One gets \( V_j \) by integrating continuation payoffs \( V_1 \) from (3) and \( V_2 \) from (5) over \( \theta_j \) with the density \( f(\theta_j) \):

\[
V_j = \int_{\theta_j}^{\theta^*} \left[ \frac{\lambda}{r + \lambda} \theta_j + \left( \frac{\lambda}{r} - \frac{\lambda}{r + \lambda} \right) \theta^* \right] f(\theta_j) d\theta_j + \int_{\theta_j}^{\theta^*} \frac{\lambda}{r} \theta_j f(\theta_j) d\theta_j
\]

(6)

If \( \theta_i = \theta^* \) the agent is indifferent between pulling arm \( i \), or experimenting with arm \( j \). In this case the expressions (2) and (6) give the same result. As arms are symmetric, the index \( j \) can be omitted in (6). Thus, one has:

\[
\frac{\lambda}{r} \theta^* = \int_{\theta^*}^{\theta_j} \frac{\lambda}{r} \theta f(\theta) d\theta + \int_{\theta_j}^{\theta^*} \left[ \frac{\lambda}{r + \lambda} \theta + \left( \frac{\lambda}{r} - \frac{\lambda}{r + \lambda} \right) \theta^* \right] f(\theta) d\theta
\]

(7)

The expression (7) shows that the agent is indifferent between pulling arm \( i \) with the cutoff type \( \theta^* \), and experimenting with arm \( j \), in the model with no experience. The next part of this subsection will be devoted to deriving the expression, similar to (7), for the case when there is experience.

In the model with experience factor, at the beginning the agent is indifferent between two arms, as in the case without experience. Without loss, let the agent pull arm \( i \). When pulling arm \( i \), one of the two events will happen first: the agent observing the type \( \theta_i \), or the agent becoming experienced. In case the agent becomes experienced prior to learning the type \( \theta_i \), the agent will continue pulling arm \( i \), as it provides now increased payoffs. Therefore the agent will consider switching to arm \( j \) only when he learns the private type \( \theta_i \).
When the agent learns \( \theta_i \), he may either be inexperienced, or experienced at arm \( i \). His behavior is characterized by the two cutoff values of type \( \theta_i \), one cutoff for each case above. If the agent is experienced at arm \( i \), he is less willing to experiment with arm \( j \), compared to the case with no experience. The reason is at arm \( i \) the agent is already experienced and getting the increased payoff, while if he switches to arm \( j \) he has yet to become experienced. The related cutoff value is therefore lower than \( \theta^* \).

The case when the agent learns private type \( \theta_i \) before becoming experienced at arm \( i \) is less trivial. In particular, it may happen that the agent starts experimenting with arm \( j \) and becomes experienced there. In this case it may be that \( \theta_j < \theta_i \), and agent stays at arm \( j \) nevertheless after learning \( \theta_j \). However, it appears that even with the possibility of becoming experienced at arm \( j \), the agent is less willing to start experimenting with arm \( j \), compared to the case without experience. That is, the agent is less willing to experiment with arm \( j \) even if he is inexperienced at both arms. Let us denote the cutoff of arm \( i \)'s type, at which the inexperienced agent experiments with arm \( j \) as \( \theta^*_1 \). The result is:

**Theorem 1** The cutoff \( \theta^*_1 \) is (weakly) lower than \( \theta^* \): \( \theta^*_1 \leq \theta^* \).

Theorem 1 implies that the agent is less willing to experiment regardless of whether he has become experienced or not, compared to the case without experience.

In order to understand the theorem, let’s look at the formula (7) for the cutoff \( \theta^* \) in the case without experience. The condition (7) means that the agent is indifferent between getting \( \theta_i = \theta^* \) at arm \( i \) forever or experimenting with arm \( j \). The latter choice either leads to the agent getting the payoff \( \theta_j > \theta_i \) from arm \( j \) forever; or the agent getting the payoff \( \theta_j < \theta_i \) once, then switching back to arm \( i \) and getting \( \theta_i \) forever.

When the experience factor is added, that changes the continuation payoff of the agent. Let us look at the option of pulling arm \( i \) forever. When the agent becomes experienced, the payoff from arm \( i \) increases by the factor of \( m \). Therefore, even when the agent is inexperienced, the potential of becoming experienced makes the agent’s continuation payoff higher, compared to the case with no experience. One can calculate that the agent’s continuation payoff increases by a factor of \( \frac{r+mu}{r+\mu} \).

If the agent decides to experiment with arm \( j \) with high type \( \theta_j > \theta_i \), then the agent will keep pulling arm \( j \) forever. Adding experience factor will increase the related continuation payoff by the same factor of \( \frac{r+mu}{r+\mu} \), as in case of staying at arm \( i \). However, if the type of arm \( j \) is low: \( \theta_j < \theta_i \), then adding experience factor will make the continuation payoff of agent’s behavior of "getting the
payoff $\theta_j$ once, then getting $\theta_i$ forever” increase by the factor less than $\frac{r+\mu}{r+\mu}$. This makes the agent less willing to experiment, compared to the case with no experience.

Experience factor makes the continuation payoff of experimenting with arm $j$ in case of $\theta_j < \theta_i$ to increase less than by the factor of $\frac{r+\mu}{r+\mu}$, due to so-called unfortunate case. By this term I mean the situation when the agent tries arm $j$, first becomes experienced there, and then learns that the type there is low: $\theta_j < \theta_i$. Depending on the parameters, the agent may either decide to switch back to arm $i$, and abandon the experience at arm $j$; or keep pulling arm $j$, foregoing the higher type at arm $i$. In both cases the agent has to give up on something. At the same time, if the agent kept pulling arm $i$, he could have become experienced there with the same ex ante chance, and receive increased payoffs at the arm with higher type.

Proof of Theorem 1

The proof repeats the logic in the explanation above and is in the appendix.

Few comments have to be made about theorem 1:

The difference between the cutoffs $\theta^*$ and $\theta^*_1$ due to experience factor arises because of the possibility of unfortunate case. The continuation payoff of the agent trying arm $j$ until he observes the low type $\theta_j < \theta_i$ and then switches back to arm $i$, increased at relatively low rate (smaller than $\frac{r+\mu}{r+\mu}$) when experience factor was added. Therefore, if one would make the decision of switching to arm $j$ irreversible, experience factor will not change the cutoff. More precisely,

**Corollary 1** In the case of irreversible switch, the cutoff does not depend on the experience factor.

Second, the way proof works does not depend on whether the arms have same type distributions of $\theta$ or intensity of prizes arrival $\lambda$, as long as the arm with the known value does not change (i.e., it’s always arm $i$). The arms may even have different values of $\mu$ and $m$. The only parameter that matters is $\frac{r+\mu}{r+\mu}$ - by how much inexperienced agent’s continuation payoff increases, compared to the case with no experience, if he does not switch the arms. Therefore, one has:

**Corollary 2** Theorem 1 continues to hold with the arms having different values of $\mu, m$ if the value $\frac{r+\mu}{r+\mu}$ is the same for both arms.

In the current paper experience is a binary variable. In practice, experience seems rather to accumulate over time. I conjecture that Theorem 1 will hold even in the latter case, provided, that the multiplicator of continuation payoff, (which was $\frac{r+\mu}{r+\mu}$ in the model), for the continuous process
of accumulating experience is the same for both arms. This means, that when the inexperienced agent starts pulling any arm, and learns the type at some point, he will be less willing to experiment with the other arm (comparing to the case with no experience), regardless of the experience accumulated at the arm being pulled.

Another thing to notice is that adding experience factor does not change the agent’s incentives to experiment, if \( \mu = 0 \) or \( \mu = \infty \). The reason is that if the agent gets experience immediately \((\mu = \infty)\) then he thinks of both arms to give him additional gain of factor \( m \) right away and thus, adding multiplicator \( m \) does not change the cutoff in \( \theta \). If the agent never gets experience \((\mu = 0)\) then one has the no experience case and equation (6).

**Corollary 3** Difference between \( \theta^* \) and \( \theta^*_1 \) disappear if either \( \mu = \infty \) or \( \mu = 0 \).

The issue of non-monotonicity of the agent’s incentives to experiment is discussed in the next part, in the general problem with \( I \) arms.

### 3.2 Single agent decision in case of \( I \) arms

In this subsection I consider the general case of the agent who faces \( I \) arms. Each arm is characterized by the Gittins index, which may change over time, as the agent is learning the arm’s type or becoming experienced. The Gittins index of arm \( i \) is the flow payoff from the safe arm, which makes agent indifferent between pulling the safe arm forever, or experimenting with arm \( i \) (with a potential switch back to the safe arm). The Gittins index of arm \( i \) will be denoted as \( s_i \), and as \( s \) for an arbitrary arm. This notation will be used regardless of whether the agent is experienced at the arm. At any moment the agent will be pulling the arm with the highest value of \( s \). The goal of this subsection is to understand the influence of experience factor on the value \( s \). For that, I introduce yet another notation - \( s_{0i} \) (or \( s_0 \) for an arbitrary arm), which would be the Gittins index of arm \( i \), with the same private type \( \theta_i \) (or prior), but without experience factor.

The analysis starts with the following observation:

**Proposition 1** 1. If the agent is experienced, then \( s_i = m_i s_{0i} \)

2. If the agent is inexperienced, then \( s_i \leq \frac{r + m_i \mu_i}{r + \mu_i} s_{0i} \)

**Proof**

Proposition 1 gives some simple results about the influence of the experience factor on the Gittins index, and, respectively, on the agent’s choice. If the agent is already experienced, he
receives payoffs increased by a factor of $m_i$. The only event happening during experimentation is learning, and therefore the agent’s problem is equivalent to the problem with no experience, the payoffs being scaled by $m_i$.

Part 2 of the proposition follows from the theorem 1. Indeed, in the theorem both arms $i$ and $j$ had the same values of $m, \mu$. The Gittins indices of both arms were equal in the case without experience. Adding experience factor multiplied the Gittins index of arm $i$ with the known private value by $\frac{r+\text{\textmu} m}{r+\text{\textmu}}$; and at the same time Gittins index of arm $j$ was increased by lower factor. ■

The possibility of becoming experienced increases $s$ for inexperienced agent. The higher is any of the values $m, \mu$, the higher is the index $s$. However, one may wonder, what would happen with the value of $s$ if one changes both $m$ and $\mu$ in a way that the factor $\frac{r+\text{\textmu} m}{r+\text{\textmu}}$ remains the same. If the agent already knows the private value $\theta$ at the arm, then the above changes of $m, \mu$ would not affect the Gittins index $s$. If the agent does not know private type $\theta$ at the arm, then the Gittins index $s$ will change even though the coefficient $\frac{r+\text{\textmu} m}{r+\text{\textmu}}$ remains the same. One has the following:

**Theorem 2** Let’s change $m$ and $\mu$ in a way that the factor $\frac{r+\text{\textmu} m}{r+\text{\textmu}}$ remains the same. Then, for an inexperienced agent

1. The Gittins index $s$ of the risky arm with unknown type $\theta$ will reach the maximum of $\frac{r+\text{\textmu} m}{r+\text{\textmu}} s_0$ at either $\mu = \infty$ or at $\mu$ limiting to $0$

2. In particular case of type being either high $\bar{\theta}$ or low $\underline{\theta}$, the dependence of the value $s$ on the arrival rate $\mu$ will have a U-shape. The value $s$ will reach its minimum at $\mu^*$, when the experienced agent with low type $\underline{\theta}$ will be indifferent between switching to the safe arm with the value $s$, or not.

The theorem 2 shows that when the agent is facing ex ante identical arms (that is providing the same continuation payoff if the agent cannot switch after he chose the first arm to pull), the agent is willing to pull the arms with the extreme values of $\mu$. The intuition behind this result is as follows: if the arrival rate of experience $\mu$ is high enough ($\mu \sim \infty$), the agent treats the arm as giving the payoffs $m\theta$ at the beginning. From proposition 1 it means that the Gittins index of the arm is increased due to experience factor by $m$, or in case of $\mu = \infty$, by $\frac{r+\text{\textmu} m}{r+\text{\textmu}}$. If the arrival rate of experience $\mu$ is low enough, the chance of unfortunate case to happen is low, and again, the Gittins index increases by the factor of $\frac{r+\text{\textmu} m}{r+\text{\textmu}}$.

**Proof of Theorem 2**

Part 1 follows from the corollary 3. The Gittins index reaches the maximum of $\frac{r+\text{\textmu} m}{r+\text{\textmu}} s_0$ at the extreme values of $\mu$. 

16
In order to prove part 2, let there be two arms: the risky arm, and the safe arm with the value \( s \), such that the agent is indifferent whether to experiment with the risky arm or not. With the two values \( \theta, \tilde{\theta} \) of private types, if the agent experiments with the risky arm in case of high type \( \tilde{\theta} \), he will pull the risky arm forever. If the type of the risky arm is low, and the agent learns it before becoming experienced, he will switch to the safe arm. However, in an unfortunate case (becoming experienced at the risky arm with type \( \theta \), then learning its type) the agent’s decision may vary depending on the parameters. The way to find the optimal behavior, is to assume one particular choice (switch to the safe arm in an unfortunate case or not), calculate the continuation payoff for each choice and then choose whichever gives higher value - this will be the Gittins index \( s \).

Let’s denote the prior probability of the risky arm having the high type \( \theta \) as \( p \), and having the low type \( \theta \) as \( 1 - p \). It is shown in the appendix, that:

**Lemma 1.** If in the unfortunate case the agent has to switch to the safe arm, then the payoff of the safe arm \( s_1 \), which makes the agent indifferent to start experimenting with the risky arm, satisfies the following condition:

\[
s_1 = \frac{\theta \lambda r + m \mu}{r + \mu} + (1 - p) \frac{\theta \lambda (r + \lambda + m \mu)}{r + \lambda + \mu} + \frac{s_1 \lambda}{\lambda + r}
\]

moreover, if the value \( \frac{r + m \mu}{r + \mu} \) is constant, \( s_1 \) increases with \( \mu \);

2. If in the unfortunate case the agent has to keep pulling the risky arm, then the payoff of the safe arm \( s_2 \), which makes the agent indifferent to start experimenting with the risky arm, satisfies the following condition:

\[
s_2 = \frac{\theta \lambda r + m \mu}{r + \mu} + (1 - p) \frac{\theta \lambda (r + \lambda + m \mu)}{r + \lambda + \mu} + \frac{s_2 \lambda}{\lambda + r}
\]

moreover, if the value \( \frac{r + m \mu}{r + \mu} \) is constant, \( s_2 \) decreases with \( \mu \);

The values \( s_1, s_2 \) are calculated at the beginning of experimentation with the risky arm.

Lemma 1 shows that if in the unfortunate case the agent switches (does not switch) to the safe arm, the continuation payoff increases (decreases) with \( \mu \). At \( \mu = \infty \) the agent prefers to switch to the safe arm in the unfortunate case, and at \( \mu \to 0 \) the agent prefers to stay at the risky arm. The optimal behavior of the agent implies that he will choose the option that maximizes the continuation payoff. Therefore the value of Gittins index \( s \) as the maximum over two above choices has the U-shape. Moreover, the minimum is reached at the points when both decisions in the unfortunate case give the same value of \( s \) to the agent, and thus he is indifferent between them. ■
The theorem 2 gives predictions for the career choice of the agent. Let’s assume the agent has to choose between several jobs, each requiring different level of human capital. The jobs are balanced in a way that the higher is the requirement for the human capital at the job, the more rewards it gives. The model predicts that at the beginning of career agent will try the ”extreme” jobs - either the ones where it’s the easiest to become experienced (working in the restaurant) or the ones where the rate of experience accumulation is the slowest (entering the PhD program). The fact that agent may choose jobs with the intermediate values of the experience accumulation can be explained by the correlation of the private types between the jobs. During PhD program the agent may go for an internship. At the start of the career, the agent will choose ”extreme” jobs, and by doing them will also learn his private type at other occupations. Once the agent gets enough information about his private type at the different jobs, he may consider switching from his initial occupation. From ex ante point of view, if the agent has to switch from his first job, the continuation payoff will be high if the first job has the extreme value of $\mu$.

3.3 Alternative information structures

It has been assumed so far that the agent knows the intensity of prize arrival at the arm, and learns the type by pulling the arm. There may be considered different information structures, however. The goal of this part is to consider whether the result from the theorem 1 about the experience factor making the agent less willing to experiment holds under other information structures. I will first consider what happens when the agent learns the intensity of prize arrival, and show that in that case experience factor still makes inexperienced agent less willing to experiment. Then I will give an example where the result does not hold.

The case where theorem 1 holds, is a well known problem of two arms, where one of the arms is safe and gives the agent known flow payoff $s$; while the other arm is the risky one, and can be either good or bad. If the risky arm is good, it gives the prize with the expected payoff $\Delta$ with Poisson intensity $\lambda$, if the risky arm is bad it never gives any prize. It is assumed that the good type of risky arm provides higher payoff in expectation: $\frac{\lambda \Delta}{\gamma} > s$ than the safe arm. If the agent knew the type of the risky arm, he would pull it in case of the good type, or pull the safe arm in case of the bad type. This specification will be referred to as ”no news - bad news”, as when pulling the risky arm and getting no prize, agent’s belief that the arm is of the good type decreases.

The solution to this problem without experience factor is that the agent has some belief $q$ that the risky arm is good. He pulls the risky arm and if he observes prize arrival, he learns that the risky arm is indeed good and keeps pulling it forever. Otherwise, the belief $q$ slowly decreases as
the agent keeps pulling the risky arm; and at some cutoff belief $q^*$ the agent stops pulling the risky arm and switches to the safe one. I will consider the effect of experience factor on the cutoff $q^*$.

It will be assumed that the agent by pulling the risky arm becomes experienced with arrival rate $\mu$, similar to previous section. The experienced agent gets the increased payoff by the factor of $m$. In order to keep payoffs from both arms proportional before and after adding the experience factor, one should have introduced the possibility of becoming experienced at the safe arm as well. However, in that case the safe arm would not be called "safe", as the flow payoff would change over time. Therefore instead I multiply the value of the safe arm $s$ by $\frac{r+mu}{r+\mu}$ - the multiplicator of expected utility that the inexperienced agent gets due to possibility of obtaining experience. Doing so does not change the agent’s continuation payoff if he decides to switch to the safe arm forever; compared to introduction of actual experience possibility.

In the model with experience factor, agent’s behavior will still be to pull the risky arm until the belief $q$ becomes low enough. If the agent while pulling the risky arm becomes experienced, then the resulting cutoff belief will be lower than $q^*$, and the agent would want to experiment longer with the risky arm. The non trivial case is what happens with the cutoff if the agent has not got any experience at the risky arm yet. Let’s denote this cutoff belief as $q^0$ and compare it with $q^*$.

The result is:

**Proposition 2** One has $q^0 > q^*$, and inexperienced agent keeps pulling the risky arm for shorter time.

There is a difference between the case where the agent learns the type (theorem 1) and the case "no news - bad news" when the agent learns the intensity. In the former case if the type of the arm with unknown type is good, the agent will for sure pull it forever. In the case "no news - bad news" even if the risky arm is good, the agent may not learn it before deciding to switch to the safe arm. Adding the possibility of becoming experienced increases the chance of the agent learning that risky arm is of good type, that is, the chance of the agent making correct choice of the arm at the end. This effect increases the agent’s incentives to experiment. However, as in the previous section, if the type of the risky arm is bad, and the agent becomes experienced, then he will be spending more time at the arm with no prizes. This effect decreases the agent’s incentives to pull the risky arm for long time. It appears that the latter effect is always stronger.

**Proof**

The model of "no news - bad news" is similar to Keller, Cripps and Rady (referred to as KCR),
the difference being the experience factor. Using the technique similar to their, one can derive the expression for the cutoff $q^*$ in the benchmark case with no experience:

$$q^* = \frac{sr}{\lambda \Delta (1 + \frac{r}{\lambda}) - s\lambda}$$  \hspace{1cm} (10)

The details are in the appendix.

Now let us see what happens if there exists experience factor. If the belief $q$ is at the cutoff value: $q = q^0$, then the inexperienced agent is indifferent for pulling the risky arm for an infinitesimal time interval $dt$, and then if there is neither good news nor obtaining experience at the risky arm, he switches to the safe arm. Otherwise, he will pull the risky arm forever (in case of observing the prize) or pull it for additional time (in case of becoming experienced). One should compare two variables (I use notation $<>$ to indicate that the sign of the inequality is not known):

$$\frac{sr + m\mu}{r + \mu} <> q\lambda dt[\Delta + \Delta \frac{\lambda r + m\mu}{r + \mu} +$$

$$\mu dt V_m + (1 - q\lambda dt)(1 - r dt)(1 - \mu dt)s \frac{r + m\mu}{r + \mu}$$

Left hand side is the agent’s payoff if he keeps pulling the safe arm. Right hand side shows, that if the agent pulls the risky arm, with probability $q\lambda dt$ he observes the prize and keeps pulling the risky arm. With probability $\mu dt$ the agent becomes experienced at the risky arm and the continuation value becomes $V_m$ (derived in the appendix from the formula in KCR). If nothing happens, the agent switches to the safe arm.

I subtract the value $s$ from both sides, get rid of the terms with $dt^2$ and put all terms with $s$ on the left. The result is:

$$\frac{sr + m\mu}{r + \mu} (r + q\lambda + \mu) <> q\lambda[\Delta + \Delta \frac{\lambda r + m\mu}{r + \mu}] + \mu V_m$$  \hspace{1cm} (11)

If one compares both parts for $q = q^*$, the sign of inequality will show the change in the cutoff, with the ”$>$” indicating that agent pulls the risky arm longer in the case with experience factor. I derive in the appendix, that (11) is equivalent to the following comparison:

$$1 <> \left(\frac{1 - q^*}{q^1 - q^*}\right)^{r/\lambda}$$  \hspace{1cm} (12)

where $q^1$ is the belief cutoff at which the experienced agent switches to the safe arm. One has that $q^1 < q^*$, the inequality in (11) is ”$<$”, and hence one proves proposition 2 ■.

20
Proposition 2 means that the result of theorem 1 holds even in case "no news - bad news", when the agent learns the intensity of prize arrival. It also holds in case, where the risky arm of bad type generates negative payoffs as Poisson events, so called "no news - good news". In the latter case if the type of the risky arm is good, agent pulls it forever, and thus the effect of experience is only to make agent experiment less due to the case of arm being of bad type and experienced agent losing time pulling it. I conjecture therefore, that the result of theorem 1 is robust.

There is a case where theorem 1 does not hold. It happens when the agent receives the prize with occasional chance at the arm being pulled. This case is related to the assumption that when the prize arrives, the agent learns its value even in case of not receiving the prize. If one imposes the agent to learn the private type $\theta$ at the arm only when he obtains the prize, then the experience makes the agent learn the type at the arm faster. Then there is example when adding experience factor makes the agent more willing to experiment. The example is arm $i$ with known type $\theta_i$, and arm $j$ either having $\theta_j > \theta_i$, or 0. This makes the agent switch from arm $j$ in case of it having type of 0, regardless of experience. By adding experience the agent is more willing to experiment with arm $j$, as it will require less time in expectation to learn its type and then make an optimal choice.

4 Continuum of agents

In this section I consider a continuum of agents pulling two arms. When an arm generates prizes, the mass of these prizes equals to the capacity of the arm $\alpha$. If the mass of agents pulling that arm at the moment of prize arrival is larger than $\alpha$, some of them will not get prizes. Experienced agents get priority in obtaining the prize. With time the amount of experienced people at any arm increases, and therefore from individual’s point of view the payoff at the arm will change, even though she may not be pulling it at the moment. One can’t thus use the Gittins index approach to solve the decision problem of the agent.

In this model I name the value of the prize for an agent at an arm as the agent’s private type at the arm. I am interested in the dynamics of agents receiving experience, learning their private types, and switching the arms. Such a dynamics assumes that the arms being pulled have arrived a finite time ago, so that the agents are still learning their types and becoming experienced. The simple model allowing the dynamics is the one with two arms: the old arm and the new arm. The old arm arrives first, and all agents start pulling it. The new arm arrives after the old arm, at period $t = 0$, when the agents start making decisions which arm to pull. Since that moment I assume, that no other arm arrives, and no existing arm disappears.
At any time the decision of the agent to switch the arms depends on the agent’s talent $\mu$ (the arrival rate of experience), the current level of experience $X$, and her private type (or prior) at both arms. I will consider two specific cases of the model, to analyze how the values of $\mu, X$ and private types determine agent’s decision.

First I will assume the case with no experimentation. That is, the old and the new arms will be symmetric, and they will provide the prizes with ex ante known value, same for all agents. In this case (some of) inexperienced agents will switch to the new arm once it arrives. The results show that if the mass of experienced agents at the old arm at $t = 0$ is lower than its capacity $\alpha$, then in any equilibrium the chance of an inexperienced (and experienced) agent getting the prize will be the same across both arms. There are multiple equilibria with the different talent distribution among inexperienced people at both arms. The old arm may have fewer, or larger fraction of talented inexperienced people compared to the new arm.

Second, I will add experimentation in the model: at any arm different agents receive different prizes. The value of the prize at the old arm is assumed to be known at the beginning. It can be so, if before arrival of the new arm at $t = 0$, the old arm has generated prizes at least once. Both arms are ex ante symmetric. With experimentation, I first consider the case with a small relative spread of private types across the arms, and show that if the agents are patient enough, talented agents will be less willing to experiment with the new arm.

Next, I will assume any distribution of private types. The goal of this part is to compare the importance of the agent’s experience and the private type at an arm. The capacity of the old arm will be assumed to be much larger than the capacity of the new arm, and thus the old arm can be thought of as the safe arm, as agents’ flow to the new arm is negligible. One can think of the old arm to be the economy of already existing sectors, with established equilibrium; and the new arm to be the new sector. When the agents switch to the new arm, with the prize arrival they learn their private types there. Some of them will already have become experienced at the moment of prize arrival, some will not. Inexperienced agents with the low private type will go back to the old arm, crowded out by the new agents, who are also inexperienced at the new arm but unaware of their private type. Experienced agents with the low type may or may not switch back to the old arm. The goal is to determine the cutoffs in the private type at which the experienced agents will stay at the new arm forever.

In the extension I will consider small deviation from the assumption of two arms being symmetric and providing the known prize, by introducing different prizes at the arms. The results show that inexperienced agents will be separated between the arms, so that agents with enough talent will all choose the arm with the higher prize (this can be the old arm or the new one), others will
choose the arm with the lower prize. Thus, the small deviation allows to predict one equilibrium out of the many in case of symmetric arms.

In the model, due to zero costs of switching and continuum of agents, one should expect equilibrium not to change, if one agent deviates from the equilibrium behavior; or if a mass of agents deviates for a single moment and then comes back. Therefore, one should define properly what the equilibrium is:

**Definition 1** *Equilibrium in the game with two arms and a continuum of agents will be called a set of functions, each function will describe the density of people trying the same arm $i$, with the same talent $\mu$, same experience $X$ at both arms, and same private types $\theta$ at both arms (whether learnt or not). Such a function will be denoted as $f_{i,\mu,X,\theta}(t)$ - as dependent on $t$. These functions have the following properties:*

1) any such function has to be measurable with respect to Borel sets on $t \in [0, \infty)$;

2) when integrated over any measurable set of talents and set of private types, the result has to indicate the proper mass of agents with such talents, private types and experience;

3) the functions $f_{i,\mu,X,\theta}(t)$ as dependent on time, have to properly indicate the change in the mass of experienced people with time;

4) when the prize arrives at any arm, the functions have to properly reflect the fact that the agents who did not know their type, learn it;

5) for the agent of any type the choice described by the related function, is optimal.

Notice, that the definition 1 implies, that the masses $\beta(t)$ and $\gamma(t)$ of inexperienced and experienced people at any arm have to be measurable with respect to Borel sets on time interval $t \in [0, \infty)$.

### 4.1 Symmetric arms with no private type

In this section I will consider the case when the old arm and the new arm are symmetric. The old arm is indexed as 1, the new arm is indexed as 2. Thus, the mass of experienced people at the old arm is denoted by $\gamma_1$. The capacity of each arm is $\alpha$. The total mass of agents is normalized to 1, and $2\alpha < 1$ so that the congestion does play a role. Prizes arrive with the Poisson intensity $\lambda$, independently across arms. The value $\lambda$ is the same for both arms. When receiving the prize from any arm, each agent gets a payoff of 1 (so there is no experimentation).

At any moment the agents differ by their talent $\mu$ and by the level of experience at the arms. The agents with experience at some arm are called *experienced*, and the agents with no experience at both arms are called *inexperienced*. Then the claim is:
Theorem 3 1. If $\gamma_1 < \alpha$ at the old arm at $t = 0$, then

 a) experienced agents will never switch the arms;
 b) at both arms the mass of experienced people $\gamma$ will reach the capacity $\alpha$ at the same time;
 c) while $\gamma < \alpha$, amount of inexperienced agents at any arm will be proportional to $\alpha - \gamma$;
 d) after $\gamma \geq \alpha$ at both arms, inexperienced people will be split between the arms in a way to keep $\gamma$-s the same for both arms at any moment.

2. If $\gamma_1 > \alpha$ at the old arm at $t = 0$, then

 a) experienced people may switch to the new arm at $t = 0$;
 b) inexperienced people will all switch to the new arm at $t = 0$ and stay there until (if) the values of $\gamma$ will become equal.

In the formulation of the theorem experienced people who move at $t = 0$ can only be the ones with experience at the old arm.

One should notice that the theorem 3 may be not that immediate to hold. One could imagine that in equilibrium the experienced agents at any arm may switch to the other arm, in order to become experienced at both arms and later be able to choose an arm with the lowest congestion. Due to them doing so, values of $\gamma$-s at both arms could be oscillating, at some point one arm being less competitive (small $\gamma_i$), and at other time other arm being less competitive, thus supporting experienced agents’ decision to switch and try other arm. However, this event is impossible in equilibrium.

In order to prove theorem 3, one has to derive several results first. The first one is:

Lemma 2 Let at some point there be $\gamma_j > \gamma_i > \alpha$. Assuming that the experienced people will not switch the arms, the inexperienced people will all try arm $i$ until (if) the masses of experienced people $\gamma$-s become equal at both arms. Afterwards, the inexperienced people will be split between the arms in a way that both $\gamma$-s will be equal at any moment in the future.

Proof.

The chance of an inexperienced person to get a prize is 0 at any arms, as $\gamma > \alpha$. The agent may start receiving the prize, only when she becomes experienced. As it is assumed that the experienced people do not switch away from the arm, once person becomes experienced at the arm, she does not switch to the other arm. Therefore, if one denotes as $V_i(t)$ the value function of a person who is experienced at arm $i$ (and pulls it forever) at time $t$, inexperienced person will try the arm with the higher current value of $V(t)$. 
Let’s show that for any inexperienced person it is better to try an arm with currently lower $\gamma$. Suppose the opposite: $\gamma_j > \gamma_i$, and there is an inexperienced person $A$ who is pulling the arm $j$ nevertheless. This means that $V_j(t) \geq V_i(t)$ at the current moment $t$. This inequality, however, will remain after an infinitesimal time interval $dt$ as well, and it will become strict, as it’s shown below:

The value function $V_i(t)$ satisfies the following equation:

$$V_i(t) = \frac{\alpha}{\gamma_i(t)} \lambda dt + (1 - r dt)V_i(t + dt)$$

With the precision of the order $dt$, one can rewrite the equation as the expression for the continuation payoff after $dt$ has passed:

$$V_i(t + dt) = (1 + r dt)V_i(t) - \frac{\alpha}{\gamma_i(t)} \lambda dt$$

and one has:

$$V_i(t + dt) = (1 - r dt)V_i(t) - \frac{\alpha}{\gamma_i(t)} r dt < V_j(t + dt) = (1 - r dt)V_j(t) - \frac{\alpha}{\gamma_j(t)} r dt$$

and therefore as long as $\gamma_j(t) > \gamma_i(t)$, one will always have $V_j > V_i$. All inexperienced people will forever try arm $j$, which is not an equilibrium, as there is less congestion at the arm $i$. Thus, it can’t be that any inexperienced agent tries an arm with higher $\gamma$. This also means that the inequality between $\gamma$-s can’t reverse (that is, for some positive time one has $\gamma_i(t) > \gamma_j(t)$, and then for some positive time one has $\gamma_j(t) > \gamma_i(t)$). The inexperienced agents will first all try the arm $i$, and then behave in a way to keep $\gamma$-s equal.

Lemma 2 shows that if the experienced agents do not switch the arms, then the inexperienced agents will behave in a way to equate $\gamma$-s (given that the inequality $\gamma > \alpha$ holds at both arms). The next result is that if the experienced people can switch the arms, the inequality between $\gamma$-s still can’t be reversed:

**Lemma 3** Let at some point there be $\gamma_j > \gamma_i > \alpha$. If there is a positive mass of people experienced at the arm they are pulling, and who pull the other arm for positive time, then one will not have the reverse inequality $\gamma_i > \gamma_j$ for a positive time.

**Proof**

The proof is in the appendix. The intuition for this lemma is that if the positive mass of experienced people pulls the arm where they have no experience, some of them become experienced in both arms. The people with experience at both arms will always choose the arm with the lowest
value of $\gamma$; they always behave to equalize the $\gamma$-s. Suppose there occurs the reversal in inequality between $\gamma$-s (as described in the lemma). This can happen only if there exists yet another reversal in $\gamma$-s in the future. That is, if one observes reversal from $\gamma_j > \gamma_i$ to $\gamma_i > \gamma_j$, then at some moment afterwards one would again have $\gamma_j > \gamma_i$ (all the inequalities are assumed to hold for a positive time). When the agents switch from arm $j$ to arm $i$ to create the first reversal, they pull the arm $i$ with the higher level of congestion. It’s optimal for them to do so, only if they expect another reversal in inequality, with arm $i$ having lower $\gamma$. By the similar logic, any reversal in inequalities between $\gamma$-s can happen only if there is another one in the future. Therefore, one has to have infinitely many reversals between $\gamma$-s. Each time such a reversal happens, people with experience at one arm pull the other arm, become experienced at both arms, and increase the mass of people experienced at both arms. At some point the mass of people with experience at both arms becomes larger than half of total population. At this point one can’t have strict inequality between $\gamma$-s anymore, as the agents with experience at both arms will equalize the congestion. This contradicts the infinite reversal possibility, proving the lemma.

Lemma 3 implies that the result of Lemma 2 (inexperienced people all going to the arm with the lower $\gamma$) holds even if some experienced people switch the arms. These lemmas show what happens if the amount of experienced people $\gamma$ exceeds the capacity of the arm $\alpha$ at both arms. However, the same statement can be made about the case where only one arm has $\gamma > \alpha$:

**Corollary 4** If one has $\gamma_j \geq \alpha \geq \gamma_i$, then all inexperienced agents will try arm $i$ until $\gamma$-s are equalized.

Next claim is about the equilibrium dynamics, when the amount of experienced people $\gamma$ at any arm does not exceed its capacity $\alpha$:

**Lemma 4** If at both arms $\gamma < \alpha$, then there will be a moment $T$, after which values $\gamma$ will become higher than the capacity $\alpha$, simultaneously at both arms.

**Proof**

Once $\gamma$ reaches $\alpha$ at some arm, all inexperienced people will try the other arm, until values of $\gamma$ are equal. Afterwards, they will be split between the arms, so that $\gamma$-s increase at the same rate. Moreover, experienced people will not switch to the arm with $\gamma = \alpha$ either, as it has higher level of congestion. ■

**Proof of Theorem 3**

Lemmas 2 and 3 imply, that if at $t = 0$ at the old arm one has $\gamma_1 \leq \alpha$ then the chance of an experienced person to get the prize will be the same for both arms at any point. This chance
equals 1 during the process $\gamma$-s at both arms reach $\alpha$, and it is equal between the arms afterwards. This means, that the experienced agents will not switch between the arms at any time, due to equal chances of getting the good while being experienced and the fact that there will always be competition for the prize among inexperienced people (see below). This proves 1a) and 1b).

As the chances of getting prize while being experienced are the same for both arms, inexperienced people will not take this factor into the account when choosing the arm to pull. They will, however, look at the probability of getting the prize while being inexperienced. They will be split between the arms in a way to equate that probability as well. This happens if the amount of inexperienced people is proportional to $\alpha - \gamma$, which proves 1c).

Parts 1d) and 2 of the theorem were proved in Lemmas 2 and 3. ■

Theorem 3 implies that the inexperienced people will behave in a same way, regardless of the talent. If at $t = 0$ one has $\gamma_1 > \alpha$ at the old arm, all of inexperienced agents will pull the new arm, and then once (if) $\gamma$-s equalize, they will be split in any way to keep these $\gamma$-s same in the future. If at $t = 0$ one has $\gamma_1 < \alpha$, the inexperienced agents will be split proportionally to $\alpha - \gamma$, but with any possible distribution of talent in these groups. In this environment, the only case where talent affects the behavior of the agent, is when at the moment of arrival of new arm one has $\gamma_1 > \alpha$ at the old arm. Then experienced people at the old arm may consider to switch to the new arm, and that decision will have a cutoff in $\mu$: only talented agents will consider pulling the new arm.

While being split between two arms, the distributions of talent among the inexperienced agents may not be the same between the arms. The arm with the higher fraction of talented people will have the mass $\alpha - \gamma$ of available prizes for inexperienced people decreasing at higher rate. This will lead into additional flow of inexperienced people from the arm with the higher fraction of talented people to the other arm.

In the next two sections I will consider adding the private type (and experimentation). First specification will be for small relative spread of private types, and the importance of talent $\mu$ on the decision to switch will be discussed. Second specification will be for the old arm having larger capacity than the new arm, which allows to interpret old arm as the safe arm. The long run distribution of types at the new arm will be discussed.

4.2 Symmetric arms with a private type

In this section there will be considered a deviation from the benchmark case of theorem 3. Namely, both arms (old and new) will provide any agent with the prize, which is a private type, from the
same distribution $f(\theta)$ over $(\bar{\theta}, \bar{\theta})$ across arms and agents. At $t = 0$ each agent already knows the private type at the old arm, but no one knows the private type at the new arm. When the new arm arrives, the decision to switch will be a cutoff decision in private type. Agents with different talents will have different cutoffs. The goal is to estimate how the cutoff depends on the talent of the agent, how the ability to become experienced affects agent’s willingness to experiment.

Experience in this model is made much more important than any news about unknown type at the new arm. I assume that the relative spread of the types: $\delta \theta = \frac{\bar{\theta} - \bar{\theta}}{\bar{\theta}}$ is small. The following proposition shows that the dynamics in the model with the private types is similar to the one without private types, if the relative spread of the types is small:

**Lemma 5** 1) there exists relative spread $\delta \theta_0$, such that for any $\delta \theta < \delta \theta_0$ experienced agents do not switch the arms;

2) for any $\epsilon > 0$ there exists $\delta \theta(\epsilon)$ such that for any $\delta \theta < \delta \theta(\epsilon)$, the differences in probability of experienced (inexperienced) agent receiving the prize will differ from the one in theorem 3 by more than $\epsilon$ for times of measure 0.

Lemma 5 implies that if the relative spread of types is small, one can essentially consider the probability of agent receiving the prize in equilibrium equal to the one in the case without private types. The next result is derived, in case when the probability of the agent receiving the prize coincides with the no experimentation case of theorem 3. I will state the result, and then later conjecture that it will hold even without explicit assumption about the probabilities.

**Proposition 3** If one has small relative spread of types, and the probability of getting the prize coincides with the one from the model with no experimentation, then:

1. For any two values of talent $\mu_1 \neq \mu_2$, the comparison of the cutoffs between $\mu_1, \mu_2$, at which inexperienced agents switch to the new arm at $t = 0$ will not depend on the density $f$ of distribution over private types at the new arm;

2. For any pari of talents $\mu_1 < \mu_2$ there exists a discount factor $r_0$ such that for all $r < r_0$, inexperienced agents with the higher talent $\mu_2$ will be less willing to switch to the new arm at $t = 0$, compared to inexperienced agents with $\mu_1$.

Part 1 of the proposition implies that as long as one is close to the case with no experimentation, one is free to choose any type distribution $f(\theta)$; and that will not influence whether talented or not talented inexperienced agents have more incentives to try the new arm. Even though private types matter little, they are the source of the agents of different talents to have different incentives
to switch to the new arm. Part 2 says that if the agents are patient, then the talented agents will be less likely to switch to the new arm.

The intuition for part 2) of proposition 3 is that if the agents are sufficiently patient, they will only care about the prize they end up consuming in the limit. When trying the new arm, talented agents are more likely to first become experienced at the new arm, and only then learn the type. These agents are thus more likely to end up consuming prize of low type from the new arm, and so they are less willing to experiment.

**Proof of proposition 3**

Given this assumption of proposition, one has the same probabilities at both arms for inexperienced (experienced) agents to get the prize, at any time. These probabilities are denoted as $p_0(t)$ and $p_1(t)$, for inexperienced and experienced agents, respectively. Let’s consider an agent with the talent $\mu$, who knows the type $\theta$ at the old arm and is inexperienced. If she decides to stay at the old arm, then one can calculate her expected payoff:

$$\theta \int_0^\infty ([e^{-\mu t}]p_1(t) + [1 - e^{-\mu t}]p_0(t))\lambda re^{-rt}dt$$

At time $t$ the chance of being experienced is $e^{-\mu t}$ and the chance of being inexperienced is $1 - e^{-\mu t}$. Therefore, the expected probability of getting the prize at time $t$ is the expression in round brackets, and the overall integral is the continuation payoff. One can see that the continuation payoff is proportional to $\theta$ and can be written as: $G\theta$, where $G$ does not depend on $\theta$. Value of $G$ can be thought of as expected discounted weight of all prizes agent is expected to get.

Let now agent switch to the new arm. As the probabilities of getting the prize are equal for both arms, if the type at the new arm, denoted as $\theta_{new}$, is higher than $\theta$, agent will keep pulling the new arm forever. As a result, she will receive the expected payoff of $G\theta_{new}$, with the same multiplicator $G$ as in case of staying at the old arm. If $\theta_{new} < \theta$, then the agent may go back to the old arm, but only if she is not experienced at the new arm yet. Otherwise, she stays at the new arm. As $p_1(t)$ and $p_0(t)$ are the same across both arms, and agent will not switch from arm if being experienced, one has that the total weighted sum of prizes when agent tries new arm with $\theta_{new} < \theta$, will still be $G$. That is, if $\theta_{new} < \theta$, then agent will get the prize $\theta_{new}$ with some discounted weight of $F$ (dependent on the agent’s talent $\mu$), and prize $\theta$ with discounted weight of $G - F$.

When making the decision to pull new arm, agent will compare $G\theta$ with
\[
\int_{\theta}^{\theta_{\text{new}}} (F \theta_{\text{new}} + (G - F) \theta) f(\theta_{\text{new}}) d\theta_{\text{new}} + \int_{\theta}^{\infty} (G \theta_{\text{new}}) f(\theta_{\text{new}}) d\theta_{\text{new}}
\]

and coefficients \( F \) and \( G \) do not depend on the distribution \( f(\theta) \). As \( f(\theta) \) is the same for all agents, one can say that the ratio \( \frac{G - F}{G} \) is what determines the agent’s willingness to switch. This ratio determines the chances of consuming \( \theta_{\text{new}} < \theta \) from the new arm with the low type. The lower this ratio is, the higher is the chance of being stuck at the new arm with the low type, and the less willing agent is to experiment. The coefficients \( F, G \) do not depend on the distribution \( f(\theta) \), and this proves part 1.

Values of \( G \) and \( G - F \) are calculated in the appendix:

**Lemma 6** One has:

1. \( G = \int_{0}^{\infty} \left[ \left( e^{-\mu t} p_0(t) \right) + \left( (1 - e^{-\mu t}) p_1(t) \right) \right] \lambda e^{-rt} dt \) (14)

2. \( G - F = \int_{0}^{\infty} \left[ \left( \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t - e^{-\mu t}} p_1(t) + (e^{-\mu t} - e^{-(\lambda + \mu)t}) p_0(t) \right) \right] \lambda e^{-rt} dt \) (15)

Having these formulas, one can see how value \( \frac{G - F}{G} \) depends on \( \mu \); the higher it is, the more willing agent is to experiment. It’s hard to get any analytical comparison, due to complicated dependencies \( p_0(t) \) and \( p_1(t) \), where the former drops to 0 non-exponentially, and the latter equals 1 up to some point, and then decreases non-exponentially. However, one can make a comparison if, fixed all other parameters, \( r \) becomes very small (agent becomes very patient). In this case one should just see what happens when time goes to infinity, as by that time all probabilities would have been close enough to their limits. In particular, \( p_0(t) \) would have become 0, \( p_1(t) \) would have reached limit of \( 2\alpha \). Then in the expression (14) for \( G \) the integrand will be \( 2\alpha\lambda \), and in the expression (15) for \( G - F \) it will be \( 2\alpha\lambda(\frac{\lambda}{\lambda + \mu}) \). The latter diminishes with \( \mu \), and that proves part 2.

\[\blacksquare\]

A possible critique of part 2 in the proposition 3 is that one assumes that agents are patient, and therefore they are willing to experiment, even being experienced. However, in the theorem the main condition is the small relative spread of private types, which dominates the one about the discount factor. If agents are pulling the arms which are identical, except for each agent receiving a small additional private payoff from any arm, the existence of a small payoff purifies the set of equilibria from theorem 3 to the ones with more talented people being less willing to experiment.
I conjecture that the proposition 3 holds even without assuming the equality of probabilities; the small relative spread of types is enough. The part 2 uses the fact of experienced agent not switching the arms, and talented agent consuming the low prize at the end. Lemma 5 is enough for this logic. As for part 1, in the proof of the proposition 3 the ratio $\frac{G-F}{G}$ was used. This ratio does not change much with the small disturbances in the probabilities $p_0, p_1$, and one can approximate the discounted weights of agent getting the good with the values $G, G - F$. The fact that the ratio $\frac{G-F}{G}$ is different for the agents with different values of $\mu$ is the source of having changes in probabilities $p_0, p_1$ compared to the case with no experimentation. Therefore one should expect changes in probabilities to influence agents’ decisions to switch to the new arm at lesser extent, then the discounted weights $G, G - F$.

### 4.3 A new arm in a stable environment

In the previous parts the results were focused on the talent of the agent, and how it affects experimentation problem. In this section I study the relative importance of the agent’s experience value and the private type at the arm. I assume the environment where the old arm is the safe arm. More specifically, I assume that every agent is experienced at the old arm, and that the capacity of the new arm is much lower than the capacity of the old arm, and the total mass of agents. Thus, agents’ switching to the new arm will have almost no effect for the agents at the old arm. In this scenario the agents who switch to the new arm, will have the highest possible value of $\mu$ (so that they have the highest expected utility at the new arm) and the lowest type $\theta$ at the old arm. Therefore, all agents switching to the new arm have the same value of $\mu$, and for any of them the old arm is the safe arm, giving the same flow payoff $s$.

Once the new arm arrives, agents will start pulling it. I assume that

$$\lambda \theta > s$$  \hspace{1cm} (16)

Expression (16) means that without the congestion even the agent with the lowest private type wants to pull the new arm. The mass of people at the new arm at $t = 0$ will be such that any person will be indifferent between pulling the new arm or receiving the payoff from the safe arm.

As time goes, some agents at the new arm will become experienced, and when the prize arrives, everyone pulling the new arm will learn the private type. Then there will be inexperienced agents with low private type, who will be crowded out by completely new agents from the old arm, as the latter ones have more incentives to try the new arm. When the next prize arrives, there may be yet another inflow of new agents. The inexperienced people at the arm are thus divided into two
groups, depending on whether they have learnt their private type. The agents from these groups pulling the new arm is the source of increasing the mass of experienced people $\gamma$. Comparing the incentives of these agents to stay at the new arm with experienced agents allows to estimate the cutoffs on private types at which the experienced people will not be crowded out by inexperienced agents.

The main result of this section is:

**Theorem 4 1.** In the limit the private types of the agents pulling the new arm will be above the cutoff value, which will lie in the interval $(\theta_0, \theta_1)$, given by the equations below:

a. The lowest possible type $\theta_0$ to pull the new arm in the limit satisfies the equation

$$\theta_0 = \int_\theta^{\overline{\theta}} \left[ \frac{\lambda}{r + \lambda + \mu} \max\{\theta_0, \frac{\mu}{r + \mu} \theta \} + \frac{\mu}{r + \lambda + \mu} \left( \frac{r}{r + \lambda} \theta + \frac{\lambda}{r + \lambda} \max\{\theta_0, \theta\} \right) \right] f(\theta) \, d\theta \quad (17)$$

b. The lowest type $\theta_1$, which will for sure be present at the new arm in the limit, satisfies the equation:

$$\theta_1 = \frac{\mu}{r + \mu} \overline{\theta} \quad (18)$$

The proof is in the appendix.

The intuition behind the theorem 4 is as follows. When the new arm arrives, some agents switch to it. On aggregate level, there is no uncertainty in agents becoming experienced at the arm. However, the time of prize arrival, and respectively, of agents learning their types, is random. That’s why the distribution of the types in the limit may vary, and one has the uncertainty about the cutoff.

At the moment of prize arrival some of the agents are already experienced. Among these experienced agents the distribution of the private types is ex-ante distribution with density $f(\theta)$ over $[\underline{\theta}, \overline{\theta}]$. Thus, agents of any private type with experience are present at the new arm. When more agents become experienced, the congestion level increases, and some of the agents get crowded out.

The equation (17) gives the condition on the private type $\theta_0$ under which the experienced agent with private type $\theta < \theta_0$ at the new arm is crowded out by the completely new agents. Expression (17) means the following: let’s fix the congestion value $\gamma > \alpha$ over time and find the cutoff level $\gamma^*$ under which completely new agents are indifferent between pulling the new arm or not. Then, given $\gamma^*$, one can find the type $\theta_0$ of experienced agents who are indifferent between staying at the
new arm or switching to the safe arm. Thus, equation (17) means that experienced agents with the type $\theta_0$ are willing to stay at the arm if and only if completely new agents are willing to try the arm. If in the limit there are experienced agents with the type below $\theta_0$, this means that the arm is still attractive for the completely new agents, and therefore the congestion will still increase over time.

The equation (18) gives the condition on the private type $\theta_1$ under which the experienced agent is not crowded by inexperienced agents with the maximal private type $\bar{\theta}$. The expression (18) means that, fixed some level of congestion $\gamma_0 > \alpha$, inexperienced people with the highest possible type are indifferent between keeping pulling the new arm or switching to the safe arm. At this level $\gamma_0$ experienced people with the type $\gamma_1$ are also indifferent between pulling the new arm or switching back to the safe option. If the type of experienced agent is at least $\theta_1$, she will not be crowded out by increased congestion, as at that moment all inexperienced people would have left the arm.

One may wonder what happens if the capacity of the old arm is comparable to the capacity of the new arm. When the prizes arrive at the new arm, some mass of agents switch back to the old arm after learning their low private type; and some mass of agents switch from the old arm to the new arm. With time, the agents who switch to the new arm, will have higher utility at the old arm compared to the ones who switched to the new arm before. Thus, the expression (17) will not hold (or more precisely, has to be adjusted for different utilities at the old arm): it was obtained assuming that experienced agent with $\theta_0$ at the new arm has the same utility from the old arm, as inexperienced agents with unknown private type. The latter agents arrive at the new arm later, and therefore have now higher utility at the old arm, and have higher incentives to switch back. By the same logic, the expression (18) still holds.

4.4 Extension: asymmetric arms with no private type

The equilibria described in theorem 3 are of large variety due to free switching costs. In those equilibria inexperienced agents behave in the same way regardless of their talent. This result is very sensitive to small disturbances though. In this section arms provide slightly different prizes, but each arm gives the same value of prize to any agent, thus no experimentation.

As the result of this section does not depend on which arm (old or new) has higher prize, let’s denote arm with higher prize as arm $H$, and it has the prize $\theta_H$, which is greater than $\theta_L$ provided by the other arm, denoted as $L$. Except for the prizes, arms are otherwise identical. One will focus
on the case when at $t = 0$ at the old arm one has $\gamma_1 < \alpha$ and therefore inexperienced people will be split between the arms; while experienced people will not change their arm.

The main result is:

**Proposition 4** Let the value $\theta_H$ limit to $\theta_L$ from above. In the limit one has:

1) For the period when $\gamma < \alpha$ at both arms there is a cutoff $\mu^*(t)$ such that any inexperienced agent with a talent $\mu < \mu^*(t)$ at time $t$ will be pulling arm $L$, and any inexperienced person with a talent $\mu > \mu^*(t)$ will be pulling arm $H$;

2) during the time when $\gamma < \alpha$ at both arms, there will be flow of people from arm $H$ to arm $L$; making the cutoff $\mu^*(t)$ increasing over time;

3) amount of experienced people $\gamma$ will hit $\alpha$ sooner at arm $H$;

4) Once $\gamma$ hits $\alpha$ at arm $L$, ratio between $\gamma$-s at both arms will be proportional to ratio over $\theta$-s:

$$\frac{\gamma_H}{\gamma_L} = \frac{\theta_H}{\theta_L}$$

and this ratio will remain constant at any moment afterwards.

**Proof**

The proof of Proposition 4 will be similar to theorem 3. First, if at both arms $\gamma > \alpha$, then the expected payoff of experienced person at any arm will be proportional to $\frac{\theta_0}{\gamma}$. Using the same logic as in theorem 3 one can show that the inexperienced people will behave in a way that keeps the ratio $\frac{\theta}{\gamma}$ same for both arms.

In order to see that at arm $H$ $\gamma$ hits $\alpha$ earlier, let’s suppose that $\gamma$ hits $\alpha$ at the same time $T$ at both arms. Then at the moment just before $T$ it is strictly profitable for an inexperienced person to deviate from arm $L$ to arm $H$: the expected chance for the inexperienced agent to get the prize is close to 0 at both arms, but at arm $H$ the prize $\theta_H$ is larger. Therefore all inexperienced people will go to arm $H$, and at arm $L$ there will be no people becoming experienced, and therefore $\gamma$ will not hit $\alpha$ at time $T$ at arm $L$. At arm $L$ $\gamma$ will hit $\alpha$ only when the expected payoff of the experienced person is equal at both arms. That is, $\gamma_H = \frac{\gamma_L \theta_H}{\theta_L} = \frac{\alpha \theta_H}{\theta_L}$ and $\gamma_L = \alpha$. That proves 3) and 4).

Let’s now look at how behavior of inexperienced people differs with the talent. Due to previous discussion, at the beginning expected flow payoff of experienced person is higher at arm $H$ than at arm $L$, and this inequality is preserved until $\gamma = \alpha$ at arm $L$, afterwards it becomes equality and stays so forever. Inexperienced person, when choosing an arm, will look at two values: the flow payoff for her as an inexperienced person (which depends on amount of prizes $\alpha - \gamma$ over mass of inexperienced people), and what continuation payoff she gets when she becomes experienced. Talented people value latter term more, as they are more likely to become experienced.
If at arm $H$ both the continuation payoff of experienced and the expected flow payoff of inexperienced person were higher, every inexperienced person will go to arm $H$. That is not possible, however: if the difference between the prizes $\theta$ is not high, all inexperienced people being at arm $H$ will create competition, while at arm $L$ one could get the prize with higher probability. Therefore, at arm $H$ one has greater continuation payoff for experienced person, but lower flow payoff for the inexperienced person. Given that, talented people have higher incentives to go to arm $H$. That proves 1).

Finally, if the difference between prizes $\theta$ limits to zero, then one will limit to the circumstances of the theorem 3. The probability of getting the prize should be almost the same for both arms, as otherwise inexperienced agents will all pull the arm with the significantly higher probability. Thus when $\gamma < \alpha$, the mass of inexperienced people at any arm is (almost) proportional to $\alpha - \gamma$. Therefore the rate at which the mass of prizes for inexperienced people $\alpha - \gamma$ will deteriorate, will be higher at arm $H$ with more talented agents, and one will need flow of people from arm $H$ to arm $L$, in order to recover indifference condition. That proves 2). ■

Proposition 4 was proven regardless of which arm provides higher prize. If it is the old arm, then the inexperienced people with low talent will switch to the new arm at $t = 0$, and later more people will go from the old arm to the new one. Agents with low talent will switch to the new arm first, followed by more and more talented as time goes. The (old) arm with the higher prize will be the place of high competition, and agents pull it hope to become experienced. The ones who become experienced, stay at that arm, others will be switching to the arm with the lower prize.

Proposition 4 was proven under the conditions very close to the benchmark case of the theorem 3. However, the result that talented people go to the arm with the higher prize, seems to hold under more general conditions. The reason is that talented inexperienced people value the continuation payoff of experienced person more than the flow payoff of inexperienced (compared to not talented). Therefore they are ready to face higher competition among inexperienced people (as long as there is no satiation: $\gamma < \alpha$). After amount of experienced people becomes greater than the capacity at both arms, every inexperienced parson will be indifferent between which arm to pull.

5 Conclusion

In this paper I consider the model with a multi-armed bandit, where an agent both learns the payoff and becomes experienced at the arm. Experience factor in the model reduces agent’s willingness to experiment. This effect induces the agent in case of all arms being ex ante identical, to pull the
ones with the extreme arrival rates of experience. The results can be used in modelling occupation choices of agents. At the start of career the agents are predicted to either choose jobs which are easy to become proficient but with low added value of experience; or the jobs where experience accumulation is slow, but it provides high benefits. When estimating empirically the influence of experience on the mobility between the jobs, one should keep in mind that the potential to become experienced may prevent the agents from switching the jobs, as well as experience itself.

The second part of the paper is devoted to a continuum of agents. It studies the dynamics of the agents switching to the newly arrived arm. The incentives of talented agents to switch to the new arm may be the same, or even lower, compared to the agents with low talent. For example, when there is a new topic in the science, it is not immediate that the most talented young scientists will work there. When there is a new sector in economics, it’s not always the case that it will attract more talented employees, compared to already existing sectors. In order to attract talented agents to the new area, one should make the salaries there high, rather than create more jobs in that area.

6 References

References


7 Appendix

7.1 Proofs for the model with the single agent

Proof of Theorem 1

First, I claim that:

**Proposition 5** Let us assume that the agent has types \( \theta_i, \theta_j \) at arms \( i, j \), and he is inexperienced at both arms. Then one has the following continuation payoffs for the different choices of the agent:

1. If the agent keeps pulling arm \( i \) forever, his continuation payoff is:

\[
\frac{\theta_i \lambda r + m \mu}{r + \mu}.
\]

2. If the agent pulls arm \( j \), observes the value \( \theta_j \), and then switches to arm \( i \) if and only if he has not become experienced at arm \( j \), then his continuation payoff is:

\[
\left[ (\frac{\lambda}{r} \frac{r + \mu}{r + \lambda + \mu}) \theta_j + (\frac{\lambda}{r} \frac{r + \mu}{r + \lambda}) \theta_i \right] \frac{r + m \mu}{r + \mu}.
\]

3. If the agent pulls arm \( j \), observes the value \( \theta_j \), and then switches to arm \( i \) regardless of whether he has become experienced at arm \( j \), then his continuation payoff is:

\[
\frac{\lambda \theta_j}{r} \frac{r + \lambda + m \mu}{r + \mu} + (\frac{\lambda}{r} \frac{r + \lambda}{r + \mu}) \theta_i \frac{r + m \mu}{r + \mu}.
\]
The proof is after the theorem.

If the agent is pulling arm $j$, and its value $\theta_j$ is higher than the cutoff $\theta^*_j$, then agent will pull arm $j$ forever. If the value $\theta_j < \theta^*_j$, then if agent learns the value $\theta_j$ prior to becoming experienced, he will switch back to arm $i$. However, there may happen the unfortunate case, when agent becomes experienced at arm $j$, and then learns its low value $\theta_j$. In this case the agent compares two continuations payoffs: the value of $\frac{\theta_j r + m \mu}{r + \mu}$ of switching back to arm $i$ and $m \frac{\theta_j r}{r + \mu}$ of staying at arm $j$. Thus, there exists a value $\theta_0 = \frac{r + m \mu}{r + \mu} \theta^*_j < \theta^*_j$ such that if $\theta_j < \theta_0$, then the agent will switch back to arm $i$ once he learns the value $\theta_j$, regardless of experience. Otherwise, if $\theta_0 < \theta_j < \theta^*_j$, then when the agent learns the value $\theta_j$, he will switch back to arm $i$ only if he has not so far become experienced at arm $j$.

Using the proposition 5, one can get the expression for the cutoff $\theta^*_j$. I will assume that the value $\theta_0$, which determines agent’s decision to switch back to arm $i$ in the unfortunate case, is higher than $\theta_4$, that is, $\theta_0$ is a feasible value. If it were not the case, the expression for the cutoff $\theta^*_j$ would be simpler, and the same logic would work for the proof.

The expression for the cutoff $\theta^*_j$ determines inexperienced agent’s indifference between keeping pulling the arm $i$ and experimenting with the arm $j$:

$$
\frac{\lambda \theta^*_j r + m \mu}{r + \mu} = \int_{\theta_0}^{\theta^*_j} \left[ \frac{\lambda}{\lambda + r} \frac{r + \lambda + m \mu}{r + \lambda + \mu} + \frac{\lambda}{r + \mu} \right] d\theta +
\int_{\theta_0}^{\theta^*_j} \left[ \frac{\lambda}{r(r + \lambda + \mu)} \frac{r + \lambda + m \mu}{r + \mu} + \frac{\lambda}{r} \right] d\theta +
\int_{\theta_0}^{\theta^*_j} \left[ \lambda \frac{r + m \mu}{r + \mu} \right] d\theta
$$

(22)

The expression (22) means that if the agent learns private type at arm $i$ to be $\theta^*_j$, while being inexperienced, then he is indifferent between pulling arm $i$ forever, or experimenting with arm $j$. If agent experiments with arm $j$, then, depending on the type $\theta_j$, he will either switch back to arm $i$ once he learns $\theta_j$ regardless of experience (first term on the right hand side), switch back to arm $i$ if he learns $\theta_j$ while being inexperienced at arm $j$ (second term), or never switch back to arm $i$ (third term).

Let’s compare part by part expressions (7) and (22) to obtain the result of the theorem. Left hand side of (22) equals left hand side of (7) times $\frac{r + m \mu}{r + \mu}$. This means when agent does not change the arms, the coefficient $\frac{r + m \mu}{r + \mu}$ shows the increase in continuation payoff of the agent due to experience. The integrand of the third term of right hand side in (22) equals the integrand of the second term in (7) times the same value of $\frac{r + m \mu}{r + \mu}$. However, integrands in both first and second terms of right hand side of (22) are less than the integrand in first term in right hand side in (7) times $\frac{r + m \mu}{r + \mu}$. For the second term in right hand side of (22) this is true due to increased coefficient $\frac{\lambda(r + \mu)}{r(r + \lambda + \mu)} > \frac{\lambda}{\lambda + r}$ of low value of $\theta_j < \theta^*_j$ compared to (7). This reflects the fact, that in an unfortunate case agent will prefer to stay at the arm $j$ with the lower private value, and thus agent gets low payoff with higher chance. For the first term, the "experience" coefficient
\( \frac{r + \lambda + m\mu}{r + \lambda + \mu} \) of the type \( \theta_j \) is lower than \( \frac{r + m\mu}{r + \mu} \). This happens due to the agent abandoning the experience he obtained at arm \( j \).

I have shown, that the integrand of right hand side in case when \( \theta_j \) is below the corresponding cutoff is (relatively) lower for the expression (10), for the model with the experience factor. This means, that the cutoff \( \theta^*_j \) is lower than the benchmark one \( \theta^* \), and proves the theorem.■

Proof of Proposition 5

1. Let agent pull arm \( i \) with the known type \( \theta_i \) forever, and denote as \( V_{0i} \) the continuation payoff when the agent is inexperienced, and as \( V_{1i} \) the continuation payoff if the agent is experienced. Then one has:

\[
V_{0i} = \lambda dt \theta_i + \mu dt (1 - r dt) V_{1i} + (1 - r dt)(1 - \mu dt)V_{0i}
\]

The equation above tells that within infinitesimal time interval \( dt \), with probability \( \lambda dt \) the agent gets the prize \( \theta_i \). With probability \( \mu dt \) the agent becomes experienced at arm \( i \), and the continuation payoff after \( dt \) has passed becomes \( V_{1i} \). With the probability \( 1 - \mu dt \) agent does not become experienced, and therefore the continuation payoff remains the same, \( V_{0i} \).

Subtracting the value \( V_{0i} \) from both sides, and getting rid of terms with \( dt^2 \), one gets:

\[
0 = \lambda dt \theta_i + \mu dt V_{1i} - (r + \mu) dt V_{0i}
\]

and therefore,

\[
V_{0i} = \frac{\lambda \theta_i}{r + \mu} + \frac{\mu}{r + \mu} V_{1i}
\]

(23)

The continuation payoff \( V_{1i} \) is found similar to expression (2):

\[
V_{1i} = \frac{m \lambda \theta_i}{r}
\]

Putting it into (23) yields the equation (19).

2. Let the agent now pull arm \( j \) and switch back to arm \( i \) only once the prize arrives at arm \( j \), and it happens while the agent is still inexperienced at arm \( j \). Let’s denote the continuation payoff of the agent when he starts pulling arm \( j \) as \( V_{0j} \), the continuation payoff if he becomes experienced at arm \( j \) before learning \( \theta_j \) as \( V_{1j} \). When one writes the expression for the continuation payoff \( V_{0j} \), two events may happen: with probability \( \lambda dt \) agent learns the type at arm \( j \), with probability \( \mu dt \) he becomes experienced at arm \( j \). With the product of these probabilities agent both becomes experienced and learns the type at arm \( j \), but this event is of order \( dt^2 \), and therefore is omitted. Thus, one has:

\[
V_{0j} = \lambda dt \theta_j + \lambda dt (1 - r dt)V_{0i} + \mu dt (1 - r dt)V_{1j} + (1 - r dt)(1 - \lambda dt)(1 - \mu dt)V_{0j}
\]

The expression above tells that when pulling the arm \( j \), with probability \( \lambda dt \) the agents learns the type \( \theta_j \) and gets the expected payoff of \( \theta_j \). If the agent learns the type \( \theta_j \), he switches back to arm \( i \), which gives the continuation payoff \( V_{0i} \), calculated in the previous part and given by (19). With probability \( \mu dt \) the agent becomes experienced, and his continuation payoff changes to \( V_{1j} \). Otherwise, if the agent neither learns \( \theta_j \), nor becomes experienced, the continuation payoff remains the same, \( V_{0j} \).
Subtracting the value function $V_{0j}$ and getting rid of terms with $dt^2$ one gets:

$$0 = \lambda dt \theta_j + \lambda dt V_{0i} + \mu dt V_{1j} - (r + \lambda + \mu) dt V_{0j}$$

or,

$$V_{0j} = \frac{\lambda}{r + \lambda + \mu} (\theta_j + V_{0i}) + \frac{\mu}{r + \lambda + \mu} V_{1j}$$  \hspace{1cm} (24)

The value $V_{1j}$ is the continuation payoff of the agent who is experienced at arm $j$ and pulls it forever. It equals $m \lambda \theta_j$. Putting this expression into (24) together with the expression (19) for the continuation payoff $V_{0i}$ yields the equation (20).

3. Now let the agent pull arm $j$ and switch to arm $i$ as soon as he learns the value $\theta_j$. Let’s denote the continuation payoff of the agent when he starts pulling arm $j$ as $V_{0j}'$, and the continuation payoff of the agent when he has become experienced at arm $j$ before learning the type $\theta_j$ as $V_{1j}'$. One writes the expression for $V_{0j}$ when the agent pulls arm $j$ for an infinitesimal period $dt$, and disregards the chance of agent both learning $\theta_j$ and becoming experienced at arm $j$ during period $dt$, as this chance is of order $dt^2$:

$$V_{0j}' = \lambda dt \theta_j + \lambda dt(1 - r dt) V_{0i} + \mu dt(1 - r dt) V_{1j}' + (1 - r dt)(1 - \lambda dt)(1 - \mu dt) V_{0j}'$$

The equation above tells that when pulling the arm $j$ for period $dt$, with probability $\lambda dt$ the agent learns the type $\theta_j$, gets expected payoff of $\theta_j$ and afterwards switches to arm $i$, which gives the continuation payoff $V_{0i}$. With the probability $\mu dt$ the agent becomes experienced at arm $j$, and the continuation payoff changes to $V_{1j}'$. Otherwise the continuation payoff remains the same, $V_{0j}'$. Getting rid of the terms $dt^2$, the expression above is reduced to:

$$V_{0j}' = \frac{\lambda}{r + \lambda + \mu} (\theta_j + V_{0i}) + \frac{\mu}{r + \lambda + \mu} V_{1j}'$$  \hspace{1cm} (25)

Now let’s write the expression for the continuation payoff $V_{1j}'$:

$$V_{1j}' = \lambda dt m \theta_j + \lambda dt(1 - r dt) V_{0i} + (1 - \lambda dt)(1 - \mu dt) V_{1j}'$$

which means that as soon as the agent learns the type $\theta_j$ at arm $j$, he switches to arm $i$. When learning the type $\theta_j$, the agent gets the expected payoff of $m \theta_j$ due to being experienced at arm $j$. This expression is equivalent to:

$$V_{1j}' = \frac{\lambda}{r + \lambda} (m \theta_j + V_{0i})$$  \hspace{1cm} (26)

Uniting the expressions (25), (26), and the equation (19) for the continuation payoff $V_{0i}$, one obtains the expression (21). ■

Proof of Lemma 1

1. Let the risky arm have the high type $\theta$ with probability $p$, and otherwise have the low type $\bar{\theta}$. One assumes that the agent in case of pulling the risky arm of low type, becoming experienced and then learning
its type, switches back to the safe arm. Let’s denote the continuation payoff of pulling the risky arm of a high type as \( V_{1H} \), and the continuation payoff of pulling the risky arm of low type as \( V_{1L} \). At the beginning, the agent is indifferent between experimenting with the risky arm, or pulling the safe arm and getting payoff \( s_1 \).

If the type of the risky arm is high, agent will keep pulling it forever. Therefore, the value \( V_{1H} \) can be found from the equation (19), of the inexperienced agent’s value function when he pulls the same arm forever:

\[
V_{1H} = \frac{\theta \lambda r + m \mu}{r + \mu}\]  
(27)

If the type of the risky arm is low, the continuation payoff \( V_{1L} \) of pulling it can be found from the expression (22), by substituting \( \theta_j \) for \( \theta \), and \( \frac{\lambda \theta}{r} \frac{r + m \mu}{r + \mu} \) for \( s_1 \). The result it:

\[
V_{1L} = \frac{\theta \lambda r + \lambda + m \mu}{r + \lambda + \mu} + \frac{s_1 \lambda}{r + \lambda}\]  
(28)

Recalling that \( s_1 \) is the ex ante continuation payoff of pulling the risky arm, it has to be that \( s_1 = p V_{1H} + (1 - p) V_{1L} \). Putting expressions (27), (28) yields the equation (8).

If one denotes coefficient \( z = \frac{r + m \mu}{r + \mu} > 1 \), then (8) becomes:

\[
s_1 = \frac{p \theta \lambda}{r} z + (1 - p) \left( \frac{\theta \lambda}{r} \frac{z(r + \lambda + \mu)}{r + \lambda + \mu} + \frac{s_1 \lambda}{r + \lambda} \right)
\]

The coefficient \( \frac{z(r + \lambda + \mu)}{r + \lambda + \mu} = z - (z - 1) \frac{\lambda}{r + \lambda + \mu} \) increases with \( \mu \), and so does the value \( s_1 \).

2. If in the unfortunate case the agent has to keep pulling the risky arm, then the Gittins index is denoted as \( s_2 \). Let’s also denote the continuation payoff of pulling the risky arm of high type as \( V_{2H} \), and of low type as \( V_{2L} \).

One has that the value \( V_{2H} \) satisfies the same formula (27) as in the previous case, with the agent switching to the safe arm in an unfortunate case. As for the value of \( V_{2L} \), it can be derived from the expression (20), by substituting \( \theta_j \) into \( \theta \), and \( \frac{\lambda \theta}{r} \frac{r + m \mu}{r + \mu} \) for \( s_2 \). The result it:

\[
V_{2L} = \frac{\theta \lambda}{r} \frac{r + m \mu}{r + \lambda + \mu} + \frac{\lambda}{r + \lambda + \mu} s_2
\]  
(29)

Recalling that \( s_2 = p V_{2H} + (1 - p) V_{2L} \), and using formulas (27), (29), one obtains the expression (9).

If one denotes coefficient \( z = \frac{r + m \mu}{r + \mu} > 1 \), then (9) becomes:

\[
s_2 = \frac{p \theta \lambda}{r} z + (1 - p) \frac{\theta \lambda}{r} \frac{z(r + \mu)}{r + \lambda + \mu} + \frac{s_2 \lambda}{r + \lambda + \mu}
\]

or,

\[
s_2 = \frac{p \theta \lambda}{r} z + (1 - p) \frac{\theta \lambda}{r + \lambda + \mu} \frac{z(r + \mu)}{r + \lambda + \mu} = \frac{z \lambda p \theta}{r} + (1 - p) \frac{\theta \lambda}{r + \lambda + \mu}
\]
and the last expression decreases with \( \mu \). □

Proof of Proposition 2
1. Derivation of expression (10)

Let’s find cutoff \( q^* \). At the belief \( q^* \) agent is indifferent between pulling risky arm for a infinitesimal period \( dt \) or stick with the safe arm. One has:

\[
s = q^* \lambda dt [\Delta + \frac{\lambda}{r} \Delta] + (1 - q^* \lambda dt)(1 - r dt) s
\]

Left hand side is the agent’s payoff from the safe arm. Right hand side shows that if the agent tries the risky arm for an infinitesimal period of \( dt \), then with probability \( q^* \lambda \) he receives good news (\( \Delta \)) and keeps with the good arm forever. Otherwise, he switches to the safe arm.

Subtracting \( s \) from both sides and disregarding terms of \( dt^2 \), one gets:

\[
0 = q^* \lambda dt [\Delta + \frac{\lambda}{r} \Delta] - s(q^* \lambda + r) dt
\]

which is equivalent to (10).

2. Derivation of the continuation payoff \( V_m \)

Value \( V_m \) is the continuation payoff for the case when agent receives payoff \( m \frac{\lambda \Delta}{r} \) from the risky arm of a good type, and \( s \frac{r + m \mu}{r + \mu} \) from the safe arm. This problem has a cutoff value of \( q^1 \), that can be obtained from the paper by Keller, Cripps, Rady. Given the current belief \( q \) of risky arm having the good type, one can use the formulas (3)-(4) from Keller, Cripps, Rady, for the case \( N = 1 \) of one agent and claim:

\[
V_m = q(m \frac{\lambda \Delta}{r}) + (s \frac{r + m \mu}{r + \mu} - m \frac{\lambda \Delta}{r} q^1) \frac{1 - q}{1 - q^1} (\frac{1 - q}{q^1})^{r/\lambda}
\]

where

\[
q^1 = \frac{r s \frac{r + m \mu}{r + \mu}}{\lambda \Delta m (1 + \frac{r}{\lambda}) - \lambda s \frac{r + m \mu}{r + \mu}}
\]

One gets the expressions (30), (31) from the formulas (3), (4) in the paper by Keller, Cripps and Rady by making substitutions of the parameters. The parameter \( g \) as the payoff of the good type of the risky arm is substituted by \( m \frac{\lambda \Delta}{r} \). The parameter \( s \) as the payoff from the safe arm is substituted by \( s \frac{r + m \mu}{r + \mu} \). The parameter \( \mu \) is substituted by \( \frac{r}{\lambda} \). The beliefs \( p, p^*_N \) are substituted by \( q, q^1 \). Finally the function \( \Omega(p) \) equals \( \frac{1 - p}{p} \). □

3. Putting formulas (30), (31) into (11), after algebra yields the expression (12)
Proof of Lemma 3

Let’s denote agents with experience only at arm $i$ $(j)$ as $i$-experienced $(j$-experienced). If the positive mass of $i$-experienced agents pulls arm $j$ for a positive time, then some of them become experience in both arms (similar effect for $j$-experienced agents). This means that there is a positive mass of agents with experienced at both arms. These agents will always choose the arm with the lower value of $\gamma$.

Let’s suppose that there is a reverse in inequality between $\gamma$-s: one has $\gamma_j > \gamma_i$ for a positive time, and then one has $\gamma_i > \gamma_j$ for a positive time. Let’s show that

**Lemma 7** The reversal $\gamma_j > \gamma_i \rightarrow \gamma_i > \gamma_j$ implies that there will be yet another reversal in the future; that is, there will be another time when $\gamma_j > \gamma_i$.

Proof.

The reversal in inequalities $\gamma_j > \gamma_i \rightarrow \gamma_i > \gamma_j$ can happen due to one of the following reasons:

1) $j$-experienced agents switch to arm $i$, reducing $\gamma_j$
2) inexperienced agent become experienced at arm $i$
3) $i$-experienced agents switch to arm $i$, increasing $\gamma_i$

This reversal can’t happen from behavior of people experienced in both arms, as they always choose the arm with the lower $\gamma$.

1) If $j$-experienced agents switch to arm $i$, and make it $\gamma_i > \gamma_j$, then each of them loses the current probability of getting the prize. For the case $\gamma_i > \alpha$ the probability of these agents getting the prize at arm $j$ is $\max\{\frac{\gamma_j}{\gamma_i}, 1\}$. The probability of them getting the prize at arm $i$ is at most $\frac{\gamma_i}{\gamma_j}$, if they also were experienced at the arm $i$. The latter expression is smaller than the former. For the case $\gamma_i \leq \alpha$ $j$-experienced agents get the prize with probability $1$ at arm $j$. However, if they switch to arm $i$, they will be inexperienced there, and the total mass of agents at arm $i$ is greater than $\alpha$ (otherwise one would not have $\gamma_i > \gamma_j$). Therefore, at arm $i$ the $j$-experienced agents receive the prize with the probability strictly smaller than $1$. Thus, they lose the current probability of getting the prize. In order for their choice to be optimal, it has to be that the inequality between $\gamma$-s reverses again in the future, so that there is a potential reward for $j$-experienced agents to be at arm $i$ and have a chance to become experienced in both arms.

2) If inexperienced agents are pulling arm $i$ at the regime $\gamma_i > \gamma_j$, they can only do so if they expect in the future another reversal of $\gamma$-s

3) Let it be now $i$-experienced agents who switch from arm $j$ to arm $i$, causing the reversal $\gamma_j > \gamma_i \rightarrow \gamma_i > \gamma_j$. The $i$-experienced agent were at arm $j$, losing the current probability of getting the prize; and then they switch back to arm $i$. Let’s show that this event can happen, only if a positive mass of $j$-experienced
agents who switch to arm $i$ at the time when $\gamma_i > \gamma_j$, as in this case one requires another reversal in $\gamma$-s: $\gamma_i > \gamma_j \rightarrow \gamma_j > \gamma_i$.

If all $j$-experienced agents stay at arm $j$ at the time when $\gamma_i > \gamma_j$, this means that when the reversal $\gamma_j > \gamma_i \rightarrow \gamma_i > \gamma_j$ happens, the value of $\gamma_j$ does not decrease. During that reversal inexperienced agents could have become experienced at the arm $j$, and the agents with experience at both arms switch to arm $j$ during this reversal. As $\gamma_j$ does not decrease during the reversal, $\gamma_i$ strictly increases.

Let’s consider the deviation of $i$-experienced agent where he pulls arm $j$ at the very end of the regime $\gamma_j > \gamma_i$ for time $dt$, and then pulls arm $i$ at the very beginning of the regime $\gamma_i > \gamma_j$. By doing this $i$-experienced agent does not change the chance of being experienced at both arms, except for the time of deviation: end of regime $\gamma_j > \gamma_i$ - start of regime $\gamma_i > \gamma_j$. The last effect leads to change in the payoff of order $dt^2$, which can be disregarded (as seen later). As after the reversal $\gamma_i$ increases, the deviation above gives a strictly positive increase in the probability of getting the prize for $i$-experienced agent, which in turn leads to the increase of the continuation payoff of order $dt$. Although $\gamma_j$ did not decrease during the switch, one can still argue that if $i$-experienced (i.e., inexperienced at arm $j$) agent pulls arm $j$ at the start of the regime $\gamma_i > \gamma_j$, the expected chance to get the prize will be higher than at the end of the regime $\gamma_j > \gamma_i$.

There are two types of agents with no experience at the arm $j$: $i$-experienced agents and the inexperienced agents. When $i$-experienced agents switch back to arm $i$, they reduce the competition at arm $j$. If the switch $\gamma_j > \gamma_i \rightarrow \gamma_i > \gamma_j$ were the last one, the inexperienced agents will all switch to arm $j$ some positive time before the end of the regime $\gamma_j > \gamma_i$. Therefore, within $dt$ the competition among the inexperienced agents at arm $j$ decreases when the reversal occurs. Thus, the deviation for $dt$ period gives $i$-experienced agent strictly positive payoff. This contradicts the optimality of behavior of $i$-experienced agent and proves that one needs yet another reversal between $\gamma$-s. ■

Lemma 7 implies that if there is a reversal between $\gamma$-s, there has to be infinite amount of them. Moreover, there is no stopping time $T^*$, after which all reversal cease. Were it the case, one can argue that when the moment $T^*$ almost occurs, the knowledge of the regime $\gamma_j \geq \gamma_i$ for all the future means that $i$-experienced agents will not pull the arm $j$, and that $\gamma_i > \alpha$. This means that $j$-experienced agents can not cause the switch $\gamma_i > \gamma_j \rightarrow \gamma_j > \gamma_i$, as the $dt$-deviation described in the lemma 7 will give higher payoff. Finally, inexperienced agents will not behave to reverse the $\gamma$-s similar to the lemma 2, as none of experienced people does so.

Thus, there are infinite reversals between $\gamma$-s, which happen indefinitely. At some point the mass of inexperienced people will be smaller than the half of mass of people with experience at both arms. Therefore, such reversals can only occur if either $i$- or $j$-experienced people switch the arms. Moreover, the mass of people switching the arms must be at least half of the mass of people with experience at both arms, otherwise there will be no reversal in $\gamma$-s. This means that there will be agents who become experienced at both arms, and the mass of them increases. Once the mass of the agents with experience at both arms becomes half of the total population, there will be no more strict inequality between $\gamma$-s. Therefore, there will be no more reversals from lemma 7, which contradicts the assumption that the first such a reversal occurred, and proves the lemma 3 ■
Proof of Lemma 6

1. The equation (14) means that at any time \( t \) the agent with the talent \( \mu \) is experienced at the arm with the probability \( 1 - e^{-\mu t} \), and is inexperienced with the probability \( e^{-\mu t} \). Thus, the integrand of (14) shows the ex ante chance of the agent to receive the prize.

2. Let agent experiment with the new arm with the low type \( \theta_{\text{new}} < \theta \). She switches back to the old arm when and if she learns the type \( \theta_{\text{new}} \) before becoming experienced at the new arm. The probability \( P(t) \) that the agent has neither learnt the type \( \theta_{\text{new}} \) or become experienced at the new arm is changed over time as follows:

   \[
P(t) = \lambda dt P(t) + \mu dt P(t) + P(t + dt)
   \]

   with the (ex-ante) probability \( \lambda dt P(t) \) the agent learns the type at the new arm during the period \([t, t + dt]\), with probability \( \mu dt P(t) \) the agent becomes experienced. Thus, one gets: \( P(t) = e^{-(\lambda + \mu)t} \), and the ex-ante probability \( P_{\text{old}}(t) \) that the agent learns the value \( \theta_{\text{new}} \) at time \( t \) prior to becoming experienced is found by integration of the change \( \lambda dt P(t) \), and is equal to

   \[
P_{\text{old}}(t) = \frac{\lambda}{\lambda + \mu}(1 - e^{-(\lambda + \mu)t})
   \]

   The ex-ante probability \( P_{\text{old,0}}(t) \) that the agent has switched to the old arm and is pulling it while being inexperienced, is found by:

   \[
P_{\text{old,0}}(t) = \int_0^t (dP_{\text{old}}(s)) e^{-\mu(t-s)} ds = \int_0^t (\lambda e^{-(\lambda + \mu)s} e^{-\mu(t-s)} ds = e^{-\mu t} - e^{-(\lambda + \mu)t}
   \]

   which is the coefficient of the probability \( p_0(t) \) in the integrand of \( G - F \). If one subtracts the expression (34) from the expression (33), one gets the coefficient of the probability \( p_1(t) \) in the integrand \( G - F \). The latter coefficient shows the ex ante probability of agent being at the old arm at time \( t \) and being experienced at the old arm. Thus, the integrand in the expression (15) correctly shows the ex-ante probability of agent consuming the prize \( \theta \) at the old arm at time \( t \).

Proof of Theorem 4

Let’s suppose that the amount of experienced people \( \gamma > \alpha \) is constant over time and find the value of \( \gamma \) at which the new agent, who is inexperienced and is unaware of the private type \( \theta \) at the new arm, is indifferent between experimenting with the new arm or staying at the old arm.

The value function \( V \) of the new agent who decides to pull the new arm, satisfies:

\[
V = \lambda dt V_1 + \mu dt V_2 + (1 - r dt)(1 - \lambda dt)(1 - \mu dt)V
\]

where \( V_1 \) is the continuation payoff of the agent who has learnt his type, \( V_2 \) is the continuation payoff of the agent who has become experienced at the new arm. One has:
\[ V = \frac{\lambda}{r + \lambda + \mu} V_1 + \frac{\mu}{r + \lambda + \mu} V_2 \]  \hspace{1cm} (35)

If the agent learns the type \( \theta \) at the new arm and keeps pulling it forever, the continuation payoff \( V'_1 \) is:

\[ V'_1 = \mu dt \frac{\lambda \theta}{r} \frac{\alpha}{\gamma} + (1 - r dt)(1 - \mu dt) V'_2 \]

and therefore, \( V'_1 = \frac{\mu}{r + \mu} \frac{\lambda \theta}{r} \frac{\alpha}{\gamma} \). The continuation payoff \( V_2 \) is the maximum over \( V'_1 \) of the agent’s choice of staying at the new arm and \( u \) of switching back to the old arm:

\[ V_1 = \max\{ \frac{\mu}{r + \mu} \frac{\lambda \theta}{r} \frac{\alpha}{\gamma}, u \} \]  \hspace{1cm} (36)

Similar, the continuation payoff \( V_2 \) of the experienced agent who has not learnt the type \( \theta \) yet, is:

\[ V_2 = \lambda dt \frac{\alpha}{\gamma} + \lambda dt \max\{ \frac{\lambda \theta}{r} \frac{\alpha}{\gamma}, u \} + (1 - r dt)(1 - \lambda dt) V_2 \]

and therefore,

\[ V_2 = \frac{\lambda}{\lambda + r} (\frac{\alpha}{\gamma} + \max\{ \frac{\lambda \theta}{r} \frac{\alpha}{\gamma}, u \}) \]  \hspace{1cm} (37)

The formulas (35)- (37) provide the expression for the value function \( V \) of completely new agent who is experimenting with the new arm. If he is indifferent whether to experiment, it has to be:

\[ s = \int_0^\theta \left[ \frac{\lambda}{r + \lambda + \mu} \max\{ \frac{\mu}{r + \mu} \frac{\lambda \theta}{r} \frac{\alpha}{\gamma}, u \} + \frac{\mu}{r + \lambda + \mu} \left( \frac{\lambda}{\lambda + \mu} \left( \frac{\lambda \theta}{r} \frac{\alpha}{\gamma} + \max\{ \frac{\lambda \theta}{r} \frac{\alpha}{\gamma}, u \} \right) \right) \right] f(\theta) d\theta \]  \hspace{1cm} (38)

At the level \( \gamma \) when the new agent is indifferent whether to experiment with the new arm, there is a type \( \theta_0 \) such that experienced agent with this type is indifferent whether to switch to the old arm. Type \( \theta_0 \) has to satisfy:

\[ \frac{\lambda \theta_0}{r} \frac{\alpha}{\gamma} = u \]  \hspace{1cm} (39)

Putting (39) into (38) yields the expression (17).

2. The expression (18) provides the condition for the experienced agent with the type \( \theta_1 \) to be indifferent whether to switch to the old arm if and only if the inexperienced agent with the maximal type \( \overline{\theta} \) is also indifferent. If the inexperienced agent decides to stay at the new arm, he has yet to become experienced, hence the coefficient \( \frac{\mu}{r + \mu} \). This is derived in the same way as the first maximum functional in the integrand of the expression (17).