UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Institute for New Economic Thinking

Cambridge-INET Institute

Cambridge-INET Working Paper Series No: 2016/24

Cambridge Working Paper Economics: 1667

EXPERIMENTATION AND LEARNING-BY-DOING

Mikhail Safronov
(University of Cambridge)

I consider a multi-armed bandit problem, where by experimenting with any arm an agent not only learns its payoffs, but also due to learning-by-doing becomes experienced at that arm. Experience provides an additional payoffs to the agent. I study the interaction between the processes of experimentation, and learning-by-doing.

The presence of learning-by-doing always reduces the agent's willingness to experiment, regardless of whether the agent is actually experienced at the arm she is currently pulling. Moreover, this effect is non-monotone in the arrival rate of experience, and reaches maximum at intermediate arrival rates. The arms with extreme arrival rate of experience yield the highest payoff to the agent, making her pulling those arms first. This non-monotonicity result is extended to the case of collective experimentation with two agents, where equilibrium payoffs of the agents reach maximum at extreme arrival rates of experience.

If the agent obtains experience by learning certain 'skills' at the arm, then the presence of experimentation effects which skills the agent learns first. If the process of learning-by-doing is deterministic, the agent learns the easier skills first; if the process is stochastic and memoryless, the agent learns the harder skills first.

Experimentation and Learning-by-Doing

Mikhail Safronov*
University of Cambridge

November 24, 2016

Abstract

I consider a multi-armed bandit problem, where by experimenting with any arm an agent not only learns its payoffs, but also due to learning-by-doing becomes experienced at that arm. Experience provides an additional payoff to the agent. I study the interaction between the processes of experimentation, and learning-by-doing.

The presence of learning-by-doing always reduces the agent's willingness to experiment, regardless of whether the agent is actually experienced at the arm she is currently pulling. Moreover, this effect is non-monotone in the arrival rate of experience, and reaches maximum at intermediate arrival rates. The arms with extreme arrival rate of experience yield the highest payoff to the agent, making her pulling those arms first. This non-monotonicity result is extended to the case of collective experimentation with two agents, where equilibrium payoffs of the agents reach maximum at extreme arrival rates of experience.

If the agent obtains experience by learning certain 'skills' at the arm, then the presence of experimentation affects which skills the agent learns first. If the process of learning-by-doing is deterministic, the agent learns the easier skills first; if the process is stochastic and memoryless, the agent learns the harder skills first.

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, job search. The main idea behind these models is that an agent faces a trade-off between exploration (trying the new arms to learn about them) and exploitation (trying the

*E-mail: ms2329@cam.ac.uk. This project started as a chapter in my doctoral dissertation. 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 for helpful discussions and comments. All errors are mine.

best known arm). In this paper I introduce an additional affect of learning-by-doing: 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 arms, seems a natural addition to the multi-armed bandit model. When working, an employee is learning how to perform her duties at lower costs. When using a Mac laptop, an agent becomes faster in accessing its features. If the agent switches arms (job, brand of laptop), she cannot perfectly transfer her experience to a new arm and has to re-engage in learning-by-doing from a lower level of experience. The presence of learning-by-doing therefore affects agent's decision to experiment with arms.

I consider a simple multi-armed bandit model to study the interaction between experimentation and learning-by-doing. When pulling any arm, the payoff for the agent comes from two sources. First, any arm generates prizes as Poisson events with known intensity, each prize giving an ex ante unknown positive payoff to the agent. To simplify the analysis, I assume that each arm generates the same prize every time. When pulling an arm, observing (and consuming) the first prize makes the agent to learn everything about the arm. Second, when pulling an arm, the agent gets an additional flow payoff from her experience at the arm. I assume the agent to be either inexperienced or experienced at any arm. The (stochastic) process of becoming experienced is a Poisson event. The agent knows its intensity and observes the moment of becoming experienced. The two processes of experimentation and learning-by-doing are therefore completely separate. The process of prize generation is stationary over time and does not change with agent's experience. The process of learning-by-doing is known and is not learnt through experimentation.

I assume that the agent cannot transfer experience between arms, and once she becomes experienced at any arm, she retains that experience even if she switches to other arms. The agent's optimal problem can be characterized with the Gittins index. Analyzing how the Gittins index depends on experimentation and learning-by-doing allows to show the interactions between the two processes.

One of the main results of the paper is that adding learning-by-doing 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 an arm, she is less willing to experiment with other arms. However, there is also not so obvious *ex ante* effect: if the agent is still inexperienced at an arm, and has observed its prizes, the agent values potential future experience at this arm more compared to arms with unknown prizes. Indeed, if the agent switches to an unknown (second) arm, she needs to spend some time to learn its prizes and may eventually realize the second arm to

generate prizes with low payoff. In this case if the agent decides to switch from the second arm, she abandons any experience she might have obtained. If the agent keeps pulling the second arm, she gets the low prizes. That is, any decision makes the agent to lose either the payoff from prizes, or the payoff from experience. At the same time, if the agent kept pulling the known arm forever, she would not abandon future experience at the known arm, making it more valuable. In other words, learning-by-doing makes the experimentation and switch between the arms more costly, since any switch might lead to abandoning experience.

The negative effect of learning-by-doing on experimentation is non-monotone in the arrival rate of experience. This effect is large for intermediate arrival rates, and small for extreme arrival rates. If the arrival rate of experience is too slow, the agent essentially never gets any experience and her experimentation problem is not affected by learning-by-doing. If the arrival rate of experience is very fast, the agent essentially becomes experienced at any arm almost immediately. As a result, the agent expects to get an additional payoff from experience at any arm, thus neglecting any value of experience. Therefore, only with intermediate arrival rates the agent expects to become experienced within a reasonable time period, and values that experience. This non-monotonicity has a U-shape form: fixing the prize distribution at the arms, there is a unique arrival rate of experience at which the agent is the least willing to experiment.

This non-monotonicity result can be applied to job search. If agents can be sorted by their general ability to acquire new skills, the agents with the highest and the lowest ability should switch jobs more often. The ability to acquire new skills can be thought to relate to grades in school, or results in job application tests.

The non-monotonicity in arrival rate of experience holds in another comparative statics. At the beginning of experimentation the agent can evaluate the value of future experience at any arm. If the agent never switched from an arm, the value of future experience at that arm would reach its maximum, since the agent would never abandon experience. With several arms and the possibility of switch, the value of future experience is less than maximal. The drop in the value of future experience has a U-shape over its arrival rate. All things equal, the agent prefers to experiment with the arms with extreme arrival rates.

The non-monotonicity of agent's payoff over arrival rate of experience seems to persist even when considering the environment with several agents. With experience arriving too fast, agents get its additional payoff at each arm immediately, and do not account for it in their strategies. With experience arriving too slow, agents finish their collective experimentation before becoming experienced. Therefore, with extreme arrival rates of experience, the process of learning-by-doing does not interact with experimentation, and does not reduce the agents' willingness to experiment.

To support this intuition, I consider an extension with two agents, and show that in some natural equilibrium their payoffs reach maximum at extreme arrival rates of experience.

The effect of learning-by-doing reducing experimentation is one part of interaction between two processes. This effect does not depend on the exact timing of learning-by-doing and would not change if experience arrived gradually, and/or deterministically. In the second part of the paper I extend the model to study how experimentation affects learning-by-doing. For the second part, the timing of experience arrival matters.

I extend the model to allow for an endogenous process of learning-by-doing. An arm is assumed to have two separate skills, and the agent can only be acquiring one skill at a time. Each skill, once acquired, gives the agent an additional payoff, independently of other skill. One can think of employee learning programming and social skills, where experience in either skill leads to additional payoff by reducing costs of job duties. The agent has to decide the sequence in which she is learning these skills. I consider two possible cases of learning-by-doing. In the first case, as before, the moment of acquiring each skill is a Poisson event. In the second case the learning-by-doing process is deterministic: the agent has to spend the exact amount of time on each skill to acquire it.

I study how experimentation, and the possibility of switching arms affects agent's choice of which skill to learn first. To make the problem non-trivial, one skill is assumed to be easier: it requires less time to acquire, though it gives less additional payoff. The result depends on the timing of experience arrival. If it's deterministic, then the possibility of potential arm switch makes the agent more willing to start learning easier skill first. If learning-by-doing is a memoryless, Poisson process then the agent is more willing to start learning the harder skill first. Thus, the timing of experience arrival may lead to different optimal patterns of skill acquisition.

The intuition why the optimal skill accumulation depends on timing of learning-by-doing, is as follows. In deterministic case the agent prefers to start getting the additional payoff from acquired skill as soon as possible, before a potential arm switch. In memoryless case there is no exact time by which the agent is guaranteed to acquire any skill. The agent has to estimate and compare the expected benefits from each skill. Conditional on acquiring any skill the agent's chances to remain at the arm increase, and these chances increase more in case of harder skill. This makes learning harder skill safer, since if succeeded, the agent will abandon the arm with lower probability.

The timing of learning-by-doing is therefore essential to determine which skills the agent should focus on first. Alternatively, the results can be applied to firm's decision on the sequence of employee's training. The firm decision should depend on whether the process of learning-by-doing is rather deterministic or rather stochastic and memoryless. In order to attract more employees, the firm should let them train easier skills first in the former case, and harder skills in the latter.

The rest of the paper is structured as follows. Section 2 is devoted to the related literature. Section 3 introduces the model and shows the effect of learning-by-doing reducing experimentation. Section 4 shows the non-monotonicity effect over the arrival rate of experience. Section 5 considers the reverse effect of experimentation on the process of learning-by-doing. Section 6 concludes.

2 Literature Review

The possibility of becoming experienced as modelled in this paper, makes the state of any arm to be fixed unless the arm is pulled. The theorem by Gittins and Jones (1974), a classical result in multi-armed bandit models, states that each arm can be assigned a number, called Gittins index, and the agent's optimal solution is to always pull the arm with the highest current Gittins index. Importantly, the Gittins index of each arm is estimated independently of other arms, significantly simplifying the estimation procedure. There are several ways to estimate Gittins index. A survey on the literature on multi-armed bandits and Gittins index can be found in Bergemann and Valimaki (2006). In this paper I estimate the Gittins index as a retirement value, and borrow this method from the work by Whittle (1982).

The possibility of learning-by-doing in the process of experimentation seems natural. Yet, to my knowledge, there is not much theoretical literature combining the two effects together. The paper by Fryer and Harms (2015) considers the learning-by-doing effect in the multi-armed bandit model, where the experience is decreasing over time, if the 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 index, which similar to Gittins index and which can be used to characterize the agent's optimal choice. The current paper differs from Fryer and Harms by having much simpler model, which allows to achieve a closed-form solution, and have comparative statics.

The process of learning-by-doing introduces an implicit cost of switching arms. When switching, the agent abandons any accumulated experience. An alternative assumption, made in several papers, would be to make the switching costs explicit. This assumption would not allow to directly use the Gittins index. The main goal of the papers with explicit switching costs is development of alternative solution methods for optimal experimentation. Jun (2004) provides a survey on such models. One of the main results, by Banks and Sundaram (1994), shows that the optimal solution cannot be indexable.

The current paper is related to constantly growing literature in labor economics on human capital accumulation and learning the job match quality, and their influence on job search and wages. There are recent papers studying the interaction between the two effects. Yamaguchi

(2012) considers the model, where workers may choose the speed of learning-by-doing by varying their effort. Kambourov and Manovskii (2009) show that when workers switch to new jobs in the same sector, their wage drop is milder compared to the workers who switch to a job in the new sector. Sanders (2016) develops a model where employees by working in the sector simultaneously learn their skills at the sector and increase them. He estimates empirically the influence of both effects on worker's wages and finds learning-by-doing to be dominant. In comparison, the current paper studies the interaction between experimentation and learning-by-doing more theoretically. I believe there is not yet sufficient understanding of this interaction, and view my paper as providing a simple intuitive model of this interaction.

An important extension of the multi-armed bandit model is to consider the collective experimentation by several agents. Assuming arms are providing the same payoffs for the agents, there are informational externalities the agents exert on each other. In papers by Keller, Rady, Cripps (2005), and Keller, Rady (2010), there is a free rider effect, reducing the equilibrium amount of experimentation below efficient level. In the paper by Bolton, Harris (1999) there is an additional encouragement effect, where the agents' motivation to experiment may be increased, since the new information may encourage other agents to experiment more in the future. The current paper relates to collective experimentation by showing that the process of learning-by-doing reduces the equilibrium amount of experimentation, and the size of this reduction becomes negligible with extreme arrival rates of experience.

Other works contain techniques and insight, used in this paper. In 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. Thomas (2012) considers the model with two agents and congestion: when pulling an arm, the agent prevents his opponent from pulling the same arm.

3 Model

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. If the agent is pulling arm i, it generates prizes with Poisson intensity λ_i, the intensity is ex ante known to the agent. I assume that each arm i has the same payoff value $\theta_i > 0$ of the prize at each arrival. Ex ante the value of θ_i is unknown to the agent, and is distributed over the interval $(\underline{\theta_i}, \overline{\theta_i})$, $0 < \underline{\theta_i} \leq \overline{\theta_i} < \infty$, with density $f(\theta_i)$ independently of other arms. When the agent gets the prize for the first time, she receives the payoff θ_i and learns its value for all future prizes. This learning specification determines the optimal experimentation scheme for the agent:

she pulls arm i until she gets the prize θ_i for the first time, and obtains full information about arm i. Afterwards the agent has to decide whether to keep pulling arm i forever or experiment with another arm.

When pulling arm i, the agent not only learns about its prize, but also may become experienced at arm i. Experience at arm i is a binary variable $X_i \in \{0, 1\}$, with $X_i = 1$ meaning that the agent is experienced at arm i. I assume that a) pulling any arm i does not change the experience at other arms. and b) once the agent becomes experienced at arm i, the value of X_i remains 1 forever. In this part of the paper I consider the moment of experience arrival to be a Poisson event, with arrival rate μ_i for each arm i, and independent of other arms and of prize arrival at arm i. The agent knows μ_i ex ante and observes the moment of becoming experienced. This process of becoming experienced is memoryless: if the agent has been pulling any 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.

An alternative way of modelling the learning-by-doing is to assume a deterministic process: each arm i is characterized by time T_i, known to the agent ex ante. The agent obtains experience at arm i at the moment the aggregate time of her having pulled arm i, equals T_i. In the first part of the paper, devoted to how learning-by-doing affects experimentation, the results will not significantly depend on the timing process of learning-by-doing, so I will focus purely on memoryless process. In the second part, devoted to how experimentation affects learning-by-doing, I will compare both deterministic and memoryless processes.

Experience at any arm i increases agent's payoff: in addition to (occasionally) receiving the prize θ_i, the experienced agent gets a flow payoff of rm_i, when pulling arm i. The value of m_i is ex ante known to the agent, and if the experienced agent keeps pulling arm i forever, the experience will add a discounted total payoff of m_i. The overall benefit of pulling arm i therefore comes from enjoying the prizes at Poisson events and the flow payoff from experience.

At any moment arm i is characterized by the prize value θ_i (or prior, if the agent has not learnt θ_i yet) and experience value X_i. Those variables are changed only when arm i is pulled, and they do so in a Markov way that depends only on their current values. 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 at any moment is to pull the arm with the highest current Gittins index. I use the following notations for the Gittins index of arm i:

DEFINITION 1 *1. $G_i(\theta_i)$, and G_i are Gittins indices of arm i if there is no process of learning-by-doing, and the agent has, respectively, observed or not observed prize θ_i at arm i;*

2. $G_i^0(\theta_i)$, and G_i^0 are Gittins indices of arm i if there is learning-by-doing, the agent is yet inexperienced at arm i and has, respectively, observed or not observed prize θ_i at arm i;

3. $G_i^1(\theta_i)$, and G_i^1 are Gittins indices of arm i if there is learning-by-doing, the agent is experienced at arm i, and has, respectively, observed or not observed prize θ_i at arm i.

Comparing these indices allows to analyze the agent's willingness to experiment in different situations.

3.1 Learning-by-doing decreases experimentation

In this section I consider how the process of learning-by-doing affects the agent's willingness to experiment. That is, I compare two cases: a benchmark case where there is no process of learning-by-doing, and the case with learning-by-doing. Both cases have the same processes of generating prizes for all arms, and the same distribution of prize value θ-s. The main result is that learning-by-doing unambiguously reduces experimentation. I first show the result in a special case of two arms, and then formulate it rigorously.

Let's consider the case with only two arms: i, j, which generate prizes with the same Poisson intensity, and have the same parameters of learning-by-doing, μ, m. The agent has already pulled arm i and learned θ_i. She has to decide whether to keep pulling arm i forever, or to experiment with arm j.

Let's assume that in case with no learning-by-doing (or, alternatively, $\mu = 0$), the agent would be indifferent between keeping pulling arm i forever or experimenting with arm j. In that case, the presence of learning-by-doing ($\mu > 0$) makes the agent to strictly prefer pulling arm i forever. The result is straightforward if the agent is already experienced at arm i. Indeed, arm i provides the agent with an additional flow payoff rm, while at arm j the agent has to spend some time to start receiving that flow payoff of rm.

The more subtle case is when in the presence of learning-by-doing the agent is currently inexperienced at arm i. That is, the agent is inexperienced at any arm and can become experienced at any arm. One could think the presence of learning-by-doing essentially cancels out and has no effect on experimentation. However, even though the agent gets the same arrival rate of experience at each arm, arm j with unknown value θ_j depreciates the value of experience by being risky. If the agent learns θ_j to be low, she may decide to switch back to arm i, potentially abandoning her experience at arm j. At the same time, when pulling arm i, the agent would not interrupt the learning-by-doing process, and enjoy the full benefit of experience.

To better illustrate the logic, let's assume the prize at arm i $\theta_i = 20$, and arm j having either the prize $\theta_j = 0$, or the prize $\theta_j = 30$. If the agent keeps pulling arm i forever, whenever the

agent obtains a prize, it will always be 20, making the sequence of agent's payoffs from prizes to be $20, 20, 20, ...$ If the agent experiments with arm j, she does so until learning θ_j, and then chooses the arm with higher prize. That is, if $\theta_j = 30$, the agent pulls arm j forever and enjoys the sequence of prizes $30, 30, 30, ...$ If $\theta_j = 0$, then the agent enjoys the sequence of prizes $0, 20, 20, ...,$ since she switches back to arm i.

The agent is indifferent whether to experiment in the absence of learning-by-doing. This means that the agent is indifferent between enjoying the prize sequence $20, 20, 20, ...$; and the lottery over the sequences $30, 30, 30, ...$ and $0, 20, 20, ...$ Now let's add learning-by-doing and look at how the agent's payoff changes. If the agent keeps pulling arm i forever, then she gets the prize sequence $20, 20, 20, ...$; plus in addition she becomes experienced at some point and enjoys the flow payoff of mr. From the point of inexperienced agent, the learning-by-doing process adds an additional value, denoted by Δ, which the agent enjoys if she keeps pulling arm i forever.

If instead the agent experiments with arm j, then if $\theta_j = 30$, the agent keeps pulling arm j forever. The agent gets the prize sequence $30, 30, 30, ...$ and the additional payoff of Δ from learning-by-doing at arm j. The presence of learning-by-doing increases the agent's continuation payoff by Δ in both cases of pulling arm i, and pulling arm j with $\theta_j = 30$. However, in case of $\theta_j = 0$, learning-by-doing increases the continuation payoff of the agent by less than Δ. When experimenting with arm j with $\theta_j = 0$, the agent might first become experienced and then learn $\theta_j = 0$. The agent has now two options. If she switches back to arm i, then she abandons the experience at arm j; she gets the same prize sequence $0, 20, 20, ...,$ but the additional payoff from experience will be lower than Δ. Alternatively, if the agent keeps pulling arm j forever, she will get a payoff Δ from experience, but a worse prize sequence $0, 0, 0....$ That is, regardless of agent's choice, the payoff from pulling arm j when $\theta_j = 0$ gives her less than a prize sequence of $0, 20, 20, ...$ plus Δ, making her less willing to experiment.

The effect of learning-by-doing reducing experimentation works in general case of I arms. Adding the the possibility of obtaining the additional payoff from experience increases the continuation payoff of pulling any arm. However, this increase is the highest at any arm with known prize, since if the agent keeps pulling that arm forever she gets the full benefit of experience. The agent is less willing to pull the arms with unknown payoffs, since the additional value of learning-by-doing is lower.

I will now formulate rigorously that the benefit of future experience is the highest at the arm with known prize. First, I estimate some continuation payoffs for the agent (equal in this case to Gittins index) in case she does not switch arms:

LEMMA 1 *1. In case of the agent pulling arm i with known value θ_i, and no learning-by-doing*

$\mu_i = 0$, *her continuation payoff equals* $G_i(\theta_i) = \frac{\lambda_i}{r}\theta_i$;

2. In case of pulling arm i*, which generates no prizes, the continuation payoff of an experienced agent equals to* $G_i^1 = m_i$*, and the continuation payoff of an inexperienced agent equals* $G_i^0 = \Delta_i \equiv \frac{m_i\mu_i}{r+\mu_i}$.

The proof of Lemma 1 is in the Appendix. The variable $\Delta_i \equiv \frac{m_i\mu_i}{r+\mu_i}$, referred later to *ex ante value of experience*, represents the maximal expected benefit of learning-by-doing to currently inexperienced agent, in case of the agent never switching from arm i.

The learning-by-doing reduces experimentation, as shown in:

THEOREM 1 *1. If the agent is inexperienced at arm* i *with known* θ_i*, then the Gittins index of arm* i *equals* $G_i^0(\theta_i) = \frac{\lambda_i}{r}\theta_i + \Delta_i$;

2. If the agent is experienced at arm i *with unknown* θ_i*, then the Gittins index equals* $G_i^1 = G_i + m_i$;

3. If the agent is inexperienced at arm i *with unknown* θ_i*, then the Gittins index equals* $G_i^0 \leq G_i + \Delta_i$.

The proof of Theorem 1 is in the Appendix, and follows the intuition from the two-arm case. With only one process of experimentation or learning-by-doing in parts 1, 2, the Gittins index equals sum of Gittins indices from the two processes. With both experimentation and learning-by-doing in part 3, the overall Gittins index is weakly lower than the sum of Gittins indices from the two processes. This is due to the fact, that the agent may become experienced at arm i and then learn θ_i to be low. The agent then has to decide whether to give up experience, or keep pulling arm i. There is therefore negative interaction between the two processes. Comparing expressions in parts 1 and 3, one can see that the additional value of learning-by-doing for inexperienced agent equals ex ante value Δ_i if the agent already knows prize value θ_i, and weakly lower otherwise. This difference makes the agent less willing to experiment.

4 Non-monotonicity over the rate of learning-by-doing

The learning-by-doing process reduces the agent's willingness to experiment, since the value of experience is the highest at an arm with known prize. The size of this reduction depends on the rate of experience arrival, μ. Assuming all arms are characterized by the same learning-by-doing parameters μ, m, with extreme values of μ the agent would experiment the most. If the value of μ would be close to zero, then the agent would not become experienced at all. On the other hand, with extremely large values of μ, the agent would immediately become experienced and start

receiving the flow payoff of rm at any arm. The agent would not value the additional value of experience since she would obtain it with no delay. Only with intermediate rates μ of experience arrival, the learning-by-doing reduces experimentation: the agent expects to become experienced within a reasonable time period, and at the same time values the experience.

The non-monotonicity result is formally stated as follows:

PROPOSITION 1 *The effect of learning-by-doing reducing experimentation at arm i, $G_i + \Delta_i - G_i^0$, has a U-shape as a function of μ_i, and reaches zero at either $\mu_i = 0$ or $\mu_i \longrightarrow \infty$.*

The proof resembles the intuition above and is in the Appendix. It can be interpreted as follows: let the agent know the prize value of θ_i at arm i and be inexperienced at any arm. The decision of the agent to experiment with other arms is a cutoff decision over θ_i. If all arms had the same parameters of learning-by-doing, μ, m, then the cutoff would have a U-shape over μ and reach its maximum at $\mu \in \{0, \infty\}$. With extreme arrival rates the agent would be most willing to switch from arm i and experiment with unknown arms.

REMARK 1 *Proposition 1 holds for the deterministic process of experience arrival as well. The non-monotonicity holds over the time period T_i required to become experienced.*

This non-monotonicity property has certain predictions for job search. Assuming each agent can be characterized by 'talent' - how fast she learns new skills at any job, the least and the most talented agents would switch the jobs the most. Observing any agent to learn slow on job means the agent does not have enough accumulated human capital to prevent her from leaving. Observing any agent to learn fast on job, does not necessarily mean that the agent will be most willing to stay at the current job. If that agent expects her fast learning to persist at other jobs, she would not care much about her success at current job. Only the agents with intermediate levels of talent would be the ones who will switch their jobs the least. Any such agent will obtain significant value of experience and will value it.

4.1 Value of future experience

The previous section on non-monotonicity considered the willingness to experiment across several agents, dependent on arrival rate μ of experience for each agent. That is, the arrival rate μ was varied, while the payoff m from experience remained the same. This section focuses on a single agent and does somewhat different exercise. The values of m, μ are both varied in a way that the ex ante value of experience $\Delta = \frac{m\mu}{r+\mu}$ remains the same. In other words, if the agent had only one arm to pull, those variations would be irrelevant. However, with the possibility of switching

arms, the agent's preferences for pulling the arm do depend on the arrival rate μ, and change non-monotonically. That is, all things equal, the agent prefers to experiment with arms with extreme arrival rates of experience:

THEOREM 2 *Consider arm i and vary the values μ_i, m_i while fixing the ex ante value of experience $\Delta_i = \frac{m_i \mu_i}{r+\mu_i}$. Then the Gittins index of inexperienced agent G_i^0 has a U-shape as a function of μ, reaching maximum of $G_i + \Delta_i$ at $\mu = 0$ and limits to $G_i + \Delta_i$ when μ converges to ∞.*

The proof is in the Appendix and has the same intuition of non-monotonicity. When experience arrives too fast, the agent starts getting payoff m_i immediately at arm i, increasing Gittins index accordingly. When experience arrives too slow, then the agent would almost surely learn θ_i before becoming experienced, making the negative interaction between experimentation and learning-by-doing negligible.

Unlike previous results, Theorem 2 is different for the deterministic process of learning-by-doing, where the agent becomes experienced after time T_i of pulling arm i. One can define the ex ante value of experience for deterministic process as $\Delta_i = e^{-rT_i} m_i$, which adds to the continuation payoff if the agent never switches from arm i. The result is formulated as:

PROPOSITION 2 *With fixed Δ_i, for deterministic process of learning-by-doing the Gittins index decreases with time T_i of obtaining experience, and reaches maximum of $G_i + \Delta_i$ if $T_i = 0$.*

The difference from Theorem 2 is that with deterministic process there is a predetermined time of experience arrival. Any switch from arm i means the agent is abandoning the ex post accumulated time for becoming experienced. With deterministic process, therefore, the agent always prefers the arrival time to be be as low as possible. At the same time, with memoryless process the agent rather cares about the processes of experimentation and learning-by-doing to not interact with each other much. With extreme arrival rates of experience, either process will be finished much before the other, negating any interaction.

4.2 Two-agent case

This paper is devoted to analyze the interaction between experimentation and learning-by-doing. A straightforward extension of the model would be to have several agents, simultaneously experimenting and becoming experienced. There are several ways the agents might affect each other. The agents may share the experience accumulated by any of them, like the employees advising each other on efficient methods to do their work. The agents may share their values of prizes generated by each arm, like the employees share the working conditions, or Mac users share the build-in features of their PCs.

The interaction between experimentation and learning-by-doing is more complicated with several agents. Nevertheless, I conjecture that the non-monotonicity result holds for the case of several agents. Intuitively, if the process of learning-by-doing is much faster or much slower than the rate of experimentation, then one of the two processes should end much before the other, as in case with a single agent. One might still expect the continuation payoffs of agents to depend non-monotonically on the arrival rate of experience, reaching the maximum at extreme rates. In this section I show this non-monotonicity in one specific case with two agents.

I assume two agents have access to two arms. One arm is safe, generating the constant flow payoff. The other arm is risky, generating prizes as Poisson events. The values of prizes are assumed to be same for both agents. In addition, when pulling the risky arm, each agent may become experienced, independently of the other agent. One could think of two firms exploiting a new technology, the profitability of which is the same for both firms. In the process, one of the firms may happen to optimize own production and implement new technology at lower costs.

I consider the following model. Both agents share the same discount factor r. At any moment t each agent chooses action $\alpha_t \in [0,1]$ as how to split time between both arms, with α_t being fraction devoted to the risky arm. The safe arm generates a constant flow payoff of sr, and has no learning-by-doing. Given α_t, the agent gets the flow payoff of $(1-\alpha_t)sr$ from the safe arm.

At the risky arm the agent obtains a prize with payoff θ as a Poisson event with arrival rate $\alpha_t\lambda$. The value $\theta > 0$ is fixed over time, and is the same for both agents. The value θ is ex ante unknown and is distributed over the interval $(\underline{\theta}, \overline{\theta})$ with $-\infty < \underline{\theta} \leq \overline{\theta} < \infty$, with density $f(\theta)$. The prize arrivals are assumed to be independent across the agents. If the agent is pulling the risky arm being inexperienced, she becomes experienced as a Poisson event with arrival rate $\alpha_t\mu$, independently of the other agent. The experienced agent gets an additional flow payoff of $\alpha_t mr$ from the risky arm. I assume that there is no congestion at the arms: the agent's payoff does not directly depend on the other agent's action.

The values of s, m, μ, λ are the same for both agents and ex ante known. I assume that the action of each agent, as well as the moment of becoming experienced are not observable by the other agent. However, when any agent gets the prize from the risky arm for the first time, the value θ becomes publicly known.

The optimal action of each agent depends on whether she is experienced at the risky arm, and on the strategy of the opponent. I focus on a particular symmetric equilibrium, where from each agent's point of view, the expected fraction of time her opponent devotes to pulling the risky arm, is constant over time and denoted as γ. In this equilibrium the continuation payoff of each agent,

as well as her incentives to pull the risky arm, depend only on her experience at the risky arm, and not on the time passed.

The equilibrium evolves over time as follows. With $\gamma = 1$, each agent devotes all the time to pull the risky arm, regardless of experience. With $\gamma \in (0, 1)$, the inexperienced agent is always indifferent on whether to pull the risky arm, and chooses a mixed time allocation for the arms. The experienced agent, having more incentives to pull the risky arm, devotes all time to it. Each agent is constantly updating her belief on whether the opponent has become experienced. This belief increases over time due to learning-by-doing. On the other hand, conditional on θ not being revealed, the belief in the opponent being of high type decreases. In the limit, conditional on θ not being revealed, the belief in the opponent being high type converges to some fixed value in $(0, 1)$. Respectively, over time, the fraction which inexperienced agent is devoting pulling the risky arm, decreases, so that the expected value γ remains the same.

Similar to Theorem 2 in one-agent case, I consider the variations of arrival rate of experience μ and additional payoff m at the risky arm, such that the ex ante value of experience $\Delta = \frac{m\mu}{r+\mu}$ remains the same. Then, the continuation payoff of an inexperienced agent reaches maximum at extreme arrival rates $\mu \in \{0, \infty\}$:

THEOREM 3 *In equilibrium the values of γ and the continuation payoff of an inexperienced agent reach maximum at either $\mu = 0$ or μ limiting to infinity.*

The proof of Theorem 3 is in the Appendix. It no longer guarantees that the continuation payoff of each agent depends on μ as a U-shape. There is an endogenous parameter γ from opponent's strategy, which affects the agent's continuation payoff, making the overall dependence on μ complicated. Nevertheless, the result about the extreme values of μ remains, and has the same intuition as in one-agent case. With μ too high, both agents start getting an additional payoff of m immediately, and it does not affect their interaction. With μ too low, the agents would finish experimentation before becoming experienced. With no interaction between the two processes of experimentation and learning-by-doing, the agents' continuation payoffs reach their maximum.

5 Experimentation affects learning-by-doing

The model has focused so far on the effect of learning-by-doing on experimentation: there is an exogenous process of experience arrival at each arm, and the agent has to make a decision regarding which arm to pull. In this section I consider the process of experience arrival to be endogenous. If the learning-by-doing is seen as the process of the agent accumulating certain skills, then each arm may have *several* skills to acquire, each increasing the flow payoff at the arm. Assuming the

agent can only acquire one skill at a time, she has to decide which skill to acquire first. The choice of skill will depend on the experimentation, on the possibility of abandoning the arm.

As before, I consider the arm, which payoff comes from prizes generated as Poisson events, and from experience. When pulling the arm it generates prizes with known Poisson intensity λ. Each time the arm generates the prize, it gives the same payoff θ to the agent, the value of θ is ex ante unknown. In addition to prizes, the arm has two skills k, l for the agent to acquire. Once acquired, each skill gives an additional flow payoff to the agent, independent of the other skill. Each skill k, l is characterized by the additional flow payoff rm_k, rm_l it provides to the agent.

I consider two different time processes of acquiring the skills. In memoryless, Poisson case the event of acquiring skill has a Poisson distribution with arrival rates μ_k, μ_l. In deterministic case the agent has to spend certain time T_k or T_l to acquire, respectively, skill k or skill l.

The agent may only be acquiring one skill at a time and has to decide which skill to acquire first. As a benchmark case, I make the following

ASSUMPTION 1 *If the agent had to pull the arm forever, she would be indifferent which skill to acquire first.*

That is, if the agent had no outside option, any sequence of skill acquisition gives the same continuation payoff to the agent. This means that one of the skills, say skill k, has lower payoff rm_k, but is acquired faster, with higher arrival rate μ_k in Poisson case, or with lower time T_k in deterministic case. With no experimentation the agent is indifferent between which skill to acquire first.

With the presence of outside option, the agent may expect to switch from the arm, dependent on the realized value of θ. The possibility of switching makes the agent to prefer one of two sequences of acquiring skills:

PROPOSITION 3 *1. In deterministic case, the agent prefers to acquire first the skill with lower value T, which requires less time to learn.*

2. In memoryless case, the agent prefers to acquire first the skill, which has lower arrival rate μ.

The proof of Proposition 3 is in the Appendix. The Proposition states that the timing of learning-by-doing process determines the agent's preferences over which skill to acquire first. In deterministic case the presence of potential switch to the safe arm makes the agent to acquire the easier skill first and start having an additional payoff earlier. In memoryless case, however, the solution is reverse. The agent does not have a certain time for acquiring the skills. Conditional on acquiring the harder skill, rather than the easier skill, the agent will be less likely to abandon

the arm. The payoff from skills is therefore more secure in case of the agent starting acquiring the harder skill first, inducing the agent to choose that learning sequence.

Proposition 3 provides insight on the efficient on-job training. The acquisition of new skills is in general a somewhat stochastic process, with agent's competence gradually increasing over time. There is usually a trade-off: the easier a skill is learnt, the lower payoff it gives. The optimal solution to this trade-off depends on whether the training is rather deterministic or rather stochastic and memoryless. In case of skills being a routine and acquired deterministically through courses, the easier skills should be acquired first. On the contrary, in case of skills being acquired somewhat randomly through actual learning-by-doing, the harder skills might be a better choice to start with.

6 Conclusion

In this paper I consider the interaction between experimentation and learning-by-doing. Learning-by-doing unambiguously reduces the agent's willingness to experiment, since the experience is always more valued at the known arm. This effect is non-monotone in arrival rate of experience, inducing the agent to pull the arms with extreme arrival rates of experience. Intuitively, with extreme arrival rates, one of the processes of experimentation or learning-by-doing is finished much before the other process, and therefore reducing the negative interaction between them. The same idea of non-monotonicity is extended for a two-agent case.

Experimentation affects learning-by-doing by changing the agent's choice between quick and cheap experience or slow and valuable experience. The quick experience allows the agent to enjoy its benefits faster, and is preferred with deterministic process of learning-by-doing. The slow experience, conditional on its acquisition, secures a greater payoff for the agent, and is preferred with stochastic process of learning-by-doing.

I believe the model is simple enough to be used in a variety of extensions. One can generalize the model with many agents, who may affect each other in both the processes of experimentation and learning-by-doing. The model can be adapted to a delegation problem of a principal and an agent, or to a matching environment with the unknown match value increasing over time. These extensions are a potential for future work.

7 References

References

Banks, J.S., and Sundaram, R.K. (1994) "Switching Costs and The Gittins Index," *Econometrica*, Vol. 62, No. 3, pp. 687-694.

Bergemann, D., and Välimäki, J. (2006) "Bandit Problems,", ISSN 1795-0562, Discussion Paper No. 93.

Bolton, P., and Harris, C. (1999) "Strategic Experimentation," *Econometrica*, Vol. 67, No. 2, pp. 349-374.

Fryer, R.G., Harms, P. (2015) "Two-Armed Restless Bandits with Imperfect Information: Stochastic Control and Indexability,", working paper.

Gittins, J.C., and Jones, D.M. (1974) ""A Dynamic Allocation Index for the Sequential Allocation of Experiments," *Progress in Statistics*, ed. by J. Gani et al. Amsterdam: North-Holland, pp. 241-266.

Gittins, J.C. (2010) "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 41, No. 2, pp. 148-177.

Jun, T. (2004) "A Survey on Bandit Problem with Switching Cost," *De Economist 152*, No. 4, pp. 513-541.

Kambourov, G., and Manovskii, I. (2009) "Occupation Specifity on Human Capital," *International Economic Review*, Vol. 50, No.1, pp. 63-115.

Keller, G., Rady, S., and Cripps, M. (2005) "Strategic Experimentation with Exponential Bandits," *Econometrica*, Vol. 73, No. 1, pp. 39-68.

Keller, G., and Rady, S. (2010) "Strategic Experimentation with Poisson Bandits," *Theoretical Economics*, 5, pp. 275-311.

Rosebberg, D., Solan, E., and Vieille, N. (2007) "Social Learning in One-Arm Bandit Problems," *Econometrica*, Vol. 75, No. 6, pp. 1591-1611.

Sanders, C. (2016) "Skill Uncertainty, Skill Accumulation, and Occupational Choice,", Working paper.

Strulovici, B. (2010) "Learning While Voting: Determinants of Collective Experimentation," *Econometrica*, Vol. 78, No.3, pp. 933-971.

Thomac, C.D. (2012) "Strategic Experimentation with Congestion,", avaialble at https://webspace.utexas.edu/ ct23744.

Whittle, P. (1982) *Optimization Over Time*, vol. 1, Wiley, Chichester.

Yamaguchi, S. (2012) "Tasks and Heterogeneous Human Capital," *Journal of Labor Economics*, Vol. 30, No. 1, pp. 1-53.

8 Appendix

Proof of Lemma 1

1. Let the agent pull arm i with known value of θ_i. Then the continuation payoff of the agent, equal to Gittins index $G_i(\theta_i)$, can be written as:

$$G_i(\theta_i) = \theta_i \lambda_i dt + (1 - rdt) G_i(\theta_i)$$

Indeed, within an infinitesimal period dt the agent gets a payoff θ_i with probability $\lambda_i dt$. After period dt, the agent will get the same continuation payoff $G_i(\theta_i)$, discounted with a factor of $(1 - rdt)$. This equation is equivalent to point 1 of Lemma 1

2. Let the experienced agent pull arm i, which gives her a constant payoff of rm_i. Then the continuation payoff of the experienced agent, G_i^1, satisfies the following:

$$G_i^1 = m_i r dt + (1 - rdt)G_i^1$$

which yields $G_i^1 = m_i$.

Let the inexperienced agent pull arm i, Then the continuation payoff of the inexperienced agent, G_i^0, satisfies the following:

$$G_i^0 = \mu_i dt * G_i^1 + (1 - rdt)(1 - \mu_i dt)G_i^0$$

Within an infinitesimal time period dt, the agent becomes experienced with probability μ_i, and her continuation payoff changes to G_i^1. Otherwise, with probability $1 - \mu_i dt$, the agent's continuation payoff remains the same. Getting rid of terms of order dt^2, and recalling that $G_i^1 = m_i$, yields $G_i^0 = \frac{m_i \mu_i}{r+\mu_i}$.

Proof of Theorem 1

The Gittins index of any arm is a retirement value, such that if the agent can always take this retirement value and stop experimentation, she is indifferent at the moment between taking the retirement value or experimenting with the arm.

1. If the inexperienced agent experiments with arm i with known value θ_i, then, since the payoff from arm i will only increase with experience, the agent will keep pulling arm i forever. The Gittins index $G^0(\theta_i)$ equals to agent's continuation payoff and satisfies:

$$G^0(\theta_i) = \theta_i \lambda_i dt + \mu dt * G^1(\theta_i) + (1 - rdt)(1 - \mu_i dt)G^0(\theta_i)$$

where $G^1(\theta_i)$ is the continuation payoff of experienced agent at arm i, and satisfies:

$$G^1(\theta_i) = \theta_i \lambda_i dt + m_i r dt + (1 - rdt)G^1(\theta_i)$$

Getting rid of terms with dt^2, one gets: $G^0(\theta_i) = \frac{\lambda_i}{r}\theta_i + \frac{m_i \mu_i}{r+\mu_i} = \frac{\lambda_i}{r}\theta_i + \Delta_i$.

2. Without learning-by-doing, the agent keeps pulling arm i until she learns θ_i. Then, if the continuation payoff of pulling arm i, $\frac{\lambda_i}{r}\theta_i$, is lower than the retirement value, G_i, the agent stops pulling arm i and gets G_i, otherwise the agent keeps pulling arm i forever. There is a cutoff $\theta^* = \frac{rG_i}{\lambda_i}$, such that the agent will stop pulling arm i if $\theta_i < \theta^*$.

The expression for continuation payoff of the agent, when she decides to pull arm i, depends on whether the actual value θ_i is higher than θ^*. Let's denote this payoff as $V(\theta_i)$. Conditional on $\theta_i > \theta^*$, one has $V(\theta_i) = \frac{\lambda_i}{r}\theta_i$, as in Lemma 1. Conditional on $\theta_i < \theta^*$, the continuation payoff satisfies:

$$V(\theta_i) = \lambda_i dt(\theta_i + G_i) + (1 - \lambda_i dt)(1 - rdt)V(\theta_i)$$

That is, with probability $\lambda_i dt$ the agent gets prize θ_i, and then immediately quits with retirement value G_i; otherwise she keeps pulling arm i forever. The continuation payoff $V(\theta_i)$ equals to:

$$V(\theta_i) = \frac{\lambda_i \theta_i + G_i}{r + \lambda_i}$$

Since the agent is indifferent at the beginning between pulling arm i or getting G_i, one gets the value G_i equal to expected value of $V(\theta_i)$ over θ_i:

$$G_i = \int_{-\infty}^{\frac{rG_i}{\lambda_i}} \frac{\lambda_i(\theta_i + G_i)}{r + \lambda_i} f(\theta_i) d\theta_i + \int_{\frac{rG_i}{\lambda_i}}^{\infty} \frac{\lambda_i}{r} \theta_i f(\theta_i) d\theta_i \tag{1}$$

With experience, the agent gets an additional flow payoff of rm_i from arm i. In order to keep the agent indifferent between pulling arm i or immediately retire, the retirement value has to increase by m_i. In this case one can think of retirement value to give agent G_i as a lump sum payoff, and also a constant flow payoff of rm_i. Since now the agent gets a flow payoff of rm_i regardless of whether she is pulling arm i, her incentives do not change compared to the case with no learning-by-doing. The value $G_i + m_i$ is the Gittins index for experienced agent.

3. Let the agent be inexperienced at arm i with unknown value θ_i, and Gittins index of arm i equal G_i^0. If the agent experiments with arm i, then she will do it until observing the value θ_i. Afterwards her decision on whether to quit and get G_i^0, depends on θ_i and whether she has become experienced. There are therefore two cutoff values, determining the agent's decision. The cutoff value $\theta_0^* = \frac{r}{\lambda_i}(G_i^0 - m_i)$ satisfies $\frac{\lambda_i}{r}\theta_0^* + m_i = G_i^0$, and makes the experienced agent to be indifferent whether to keep pulling arm i or retire and get G_i^0. The cutoff value $\theta_1^* = \frac{r}{\lambda_i}(G_i^0 - \frac{m_i \mu_i}{r+\mu_i})$ makes the inexperienced agent to be indifferent whether to keep pulling arm i or retire.

If pulling arm i, the continuation payoff of the agent is expressed differently, dependent on actual value of θ_i. Let's denote the continuation payoff of inexperienced agent as $V^0(\theta_i)$, and of experienced agent as $V^1(\theta_i)$. If $\theta_i > \theta_1^*$, the agent will keep pulling arm i forever, getting a continuation payoff of $V^0(\theta_i) = \frac{\lambda_i}{r}\theta_i + \frac{m_i \mu_i}{r+\mu_i}$.

If $\theta_0^* < \theta_i < \theta_1^*$, then the agent will keep pulling arm i only if she has become experienced before learning θ_i. The continuation payoff satisfies:

$V^0(\theta_i) = \lambda_i dt(\theta_i + G_i^0) + \mu_i dt(\frac{\lambda_i}{r}\theta_i + m_i) + (1 - rdt)(1 - \lambda_i dt) + (1 - \mu_i dt)V^0(\theta_i)$

With probability λdt the inexperienced agent receives θ_i, and then immediately retires. With probability $\mu_i dt$ the agent becomes experienced and keeps pulling arm i forever. Otherwise, the continuation payoff does not change and remains $V^0(\theta_i)$. The continuation payoff $V^0(\theta_i)$ equals:

$$V^0(\theta_i) = \frac{\lambda_i(\theta_i + G_i^0) + \mu_i(\frac{\lambda_i}{r}\theta_i + m_i)}{r + \lambda_i + \mu_i}$$

If $\theta_i < \theta_0^*$, then the agent will stop pulling arm i once she learns θ_i. The continuation payoff, if being already experienced, satisfies:

$$V^1(\theta_i) = m_i r dt + \lambda_i dt(\theta_i + G_i^0) + (1 - rdt)(1 - \lambda_i dt)V^1(\theta_i)$$

The continuation payoff of inexperienced agent satisfies:

$$V^0(\theta_i) = \lambda_i dt(\theta_i + G_i^0) + \mu_i dt * V^1(\theta_i) + (1 - rdt)(1 - \lambda_i dt)(1 - \mu_i dt)V^0(\theta_i)$$

which yields:

$$V^0(\theta_i) = \frac{\lambda_i}{r + \lambda_i}(\theta_i + G_i^0) + \frac{r\mu_i m_i}{(r + \lambda_i)(r + \lambda_i + \mu_i)}$$

Combining all the expressions for value function $V^0(\theta_i)$, one gets the Gittins index G_i^0 as expected value over $V^0(\theta_i)$:

$$\begin{aligned} G_i^0 = &\int_{-\infty}^{\frac{r}{\lambda_i}(G_i^0 - m_i)} \frac{\lambda_i}{r + \lambda_i}(\theta_i + G_i^0) + \frac{r\mu_i m_i}{(r + \lambda_i)(r + \lambda_i + \mu_i)} f(\theta_i)d\theta_i + \\ &+ \int_{\frac{r}{\lambda_i}(G_i^0 - m_i)}^{\frac{r}{\lambda_i}(G_i^0 - \frac{m_i\mu_i}{r+\mu_i})} \frac{\lambda_i(\theta_i + G_i^0) + \mu_i(\frac{\lambda_i}{r}\theta_i + m_i)}{r + \lambda_i + \mu_i} f(\theta_i)d\theta_i + \\ &+ \int_{\frac{r}{\lambda_i}(G_i^0 - \frac{m_i\mu_i}{r+\mu_i})}^{\infty} (\frac{\lambda_i}{r}\theta_i + \frac{m_i\mu_i}{r + \mu_i}) f(\theta_i)d\theta_i \end{aligned} \quad (2)$$

Now let's compare two expressions for Gittins indices (1) and (2). Let's show that if one substitutes $G_i^0 = G_i + \Delta_i = G_i + \frac{m_i\mu_i}{r+\mu_i}$ into expression (2), then the left-hand side would become bigger than the right-hand side. That would prove Theorem 1. In other words, let's substitute $G_i^0 = G_i + \frac{m_i\mu_i}{r+\mu_i}$ into expression (2), and show that the right hand side of (2) would not exceed the right-hand side of (1) plus $\frac{m_i\mu_i}{r+\mu_i}$.

The last integrals in both expressions (1) and (2) would coincide if $G_i^0 = G_i + \frac{m_i\mu_i}{r+\mu_i}$. The difference between the integrands in first terms between expressions (2) and (1) would equal

$$\frac{\lambda_i}{r + \lambda_i}(\theta_i + G_i + \frac{m_i\mu_i}{r + \mu_i}) + \frac{r\mu_i m_i}{(r + \lambda_i)(r + \lambda_i + \mu_i)} - \frac{\lambda_i(\theta_i + G_i)}{r + \lambda_i} = \frac{m_i\mu_i}{r + \mu_i} - \frac{r\lambda_i m_i\mu_i}{(r + \lambda_i)(r + \mu_i)(r + \lambda_i + \mu_i)}$$

and is smaller than $\frac{m_i\mu_i}{r+\mu_i}$, reflecting the fact that if agent becomes experienced and then learns θ_i, she prefers to leave arm i and abandon the experience.

The difference between the integrand in the second term of expression (2) and the first term of expression (1) equals

$$\frac{\lambda_i(\theta_i + G_i + \frac{m_i\mu_i}{r+\mu_i}) + \mu_i(\frac{\lambda_i}{r}\theta_i + m_i)}{r + \lambda_i + \mu_i} - \frac{\lambda_i(\theta_i + G_i)}{r + \lambda_i} = \frac{m_i\mu_i}{r + \mu_i} + \frac{\lambda_i\mu_i(\frac{\lambda_i}{r}\theta_i - G_i)}{(r + \lambda_i)(r + \lambda_i + \mu_i)}$$

which is lower than $\frac{m_i\mu_i}{r+\mu_i}$, since this is case for $\theta_i < \frac{r}{\lambda_i}(G_i^0 - \frac{m_i\mu_i}{r+\mu_i}) = \frac{r}{\lambda_i}G_i$. This reflects the fact that the experienced agent does not want to quit pulling arm i and therefore enjoys worse prizes, compared to case without learning-by-doing.

The integrands in first and second terms of expression (2) are lower than in the first term of (1) plus $\frac{m_i\mu_i}{r+\mu_i}$, reflecting the fact that when the experienced agent learns about low θ, she has to either sacrifice experience, or stay at lower θ. Therefore, the value $G_i^0 \leq G_i + \frac{m_i\mu_i}{r+\mu_i}$, proving Theorem 1.

Proof of Proposition 1

Let's consider a variable $Y \equiv G_i^0 - \frac{m_i\mu_i}{r+\mu_u}$ The expression (2) then can be rewritten as:

$$Y = \int_{-\infty}^{\frac{r}{\lambda_i}(Y-\frac{rm_i}{r+\mu_i})} \left(\frac{\lambda_i}{r+\lambda_i}(\theta_i + Y) - \frac{r\lambda_i\mu_i m_i}{(r+\lambda_i)(r+\mu_i)(r+\lambda_i+\mu_i)} \right) f(\theta_i)d\theta_i +$$
$$+ \int_{\frac{r}{\lambda_i}(Y-\frac{rm_i}{r+\mu_i})}^{\frac{r}{\lambda_i}Y} \left(\frac{\lambda_i(\theta_i+Y)}{r+\lambda_i} + \frac{\lambda_i\mu_i(\frac{\lambda_i}{r}\theta_i - Y)}{(r+\lambda_i)(r+\lambda_i+\mu_i)} \right) f(\theta_i)d\theta_i + \int_{\frac{r}{\lambda_i}Y}^{\infty} \frac{\lambda_i}{r}\theta_i f(\theta_i)d\theta_i$$

First let's notice that if μ_i equals to either zero or ∞, then the expression for Y coincides with the expression (1), and therefore $Y = G_i$. Next, let's consider a derivative Y'_{μ_i}, and show that it is negative below some $\mu_i^* > 0$, and otherwise is positive. This will show the U-shape from Proposition 1, since G_i does not depend on μ_i.

The derivative of Y'_{μ_i} satisfies the following:

$$Y'_{\mu_i} = \int_{-\infty}^{\frac{r}{\lambda_i}(Y-\frac{rm_i}{r+\mu_i})} \left(\frac{\lambda_i}{r+\lambda_i}Y'_{\mu_i} - \frac{r\lambda_i m_i}{r+\lambda_i}\frac{r^2+r\lambda_i-\mu_i^2}{(r+\mu_i)^2(r+\lambda_i+\mu_i)^2} \right) f(\theta_i)d\theta_i +$$
$$+ \int_{\frac{r}{\lambda_i}(Y-\frac{rm_i}{r+\mu_i})}^{\frac{r}{\lambda_i}Y} \left(\frac{\lambda_i Y'_{\mu_i}}{r+\lambda_i} + \frac{\lambda_i(\frac{\lambda_i}{r}\theta_i - Y) - \lambda_i\mu_i Y'_{\mu_i}}{(r+\lambda_i)(r+\lambda_i+\mu_i)} - \frac{\lambda_i\mu_i(\frac{\lambda_i}{r}\theta_i - Y)}{(r+\lambda_i)(r+\lambda_i+\mu_i)^2} \right) f(\theta_i)d\theta_i$$

The coefficient at Y'_{μ_i} is higher on left-hand side, then on right-hand side. Therefore the sign of Y'_{μ_i} coincides with the sign of right-hand side, after removing all coefficients with Y'_{μ_i}:

$$\int_{-\infty}^{\frac{r}{\lambda_i}(Y-\frac{rm_i}{r+\mu_i})} \left(\frac{r\lambda_i m_i}{r+\lambda_i}\frac{-r^2-r\lambda_i+\mu_i^2}{(r+\mu_i)^2(r+\lambda_i+\mu_i)^2} \right) f(\theta_i)d\theta_i +$$
$$+ \int_{\frac{r}{\lambda_i}(Y-\frac{rm_i}{r+\mu_i})}^{\frac{r}{\lambda_i}Y} \left(\frac{\lambda_i(\frac{\lambda_i}{r}\theta_i - Y)}{(r+\lambda_i)(r+\lambda_i+\mu_i)} - \frac{\lambda_i\mu_i(\frac{\lambda_i}{r}\theta_i - Y)}{(r+\lambda_i)(r+\lambda_i+\mu_i)^2} \right) f(\theta_i)d\theta_i$$

or, equivalently, if multiplying by $(r+\lambda_i+\mu_i)^2 > 0$, the sign of

$$\int_{-\infty}^{\frac{r}{\lambda_i}(Y-\frac{rm_i}{r+\mu_i})} \frac{r\lambda_i m_i}{r+\lambda_i}\frac{-r^2-r\lambda_i+\mu_i^2}{(r+\mu_i)^2} f(\theta_i)d\theta_i + \int_{\frac{r}{\lambda_i}(Y-\frac{rm}{r+\mu_i})}^{\frac{r}{\lambda_i}Y} \lambda_i(\frac{\lambda_i}{r}\theta_i - Y)f(\theta_i)d\theta_i$$

The first term increases with μ_i, being negative at $\mu_i = 0$, and positive at $\mu_i = \infty$. Moreover, both the integrand, $\frac{r\lambda_i m_i}{r+\lambda_i}\frac{-r^2-r\lambda_i+\mu_i^2}{(r+\mu_i)^2}$, and the interval over the integration $(-\infty, \frac{r}{\lambda_i}(Y-\frac{rm_i}{r+\mu_i}))$ increase with μ_i. The second term is negative. Its integrand, $\frac{\lambda_i}{r}\theta_i - Y$ does not depend on μ_i, and the interval $(\frac{r}{\lambda_i}(Y-\frac{rm}{r+\mu_i}), \frac{r}{\lambda_i}Y)$ over integration is decreasing with μ_i. This means that the whole expression is negative if μ_i is low enough, and positive otherwise. Therefore, the value Y has a U-shape over the arrival rate μ_i, and reaches maximum at extreme values $\mu_i \in \{0, \infty\}$.

Proof of Theorem 2

Let's consider the expression (2) for Gittins index G_i^0, and substitute $m_i = \frac{\Delta_i(r+\mu_i)}{\mu_i}$ to get:

$$G_i^0 = \int_{-\infty}^{\frac{r}{\lambda_i}(G_i^0 - \frac{\Delta_i(r+\mu_i)}{\mu_i})} \left(\frac{\lambda_i}{r+\lambda_i}(\theta_i + G_i^0) + \frac{r\Delta_i(r+\mu_i)}{(r+\lambda_i)(r+\lambda_i+\mu_i)} \right) f(\theta_i)d\theta_i +$$
$$+ \int_{\frac{r}{\lambda_i}(G_i^0 - \frac{\Delta_i(r+\mu_i)}{\mu_i})}^{\frac{r}{\lambda_i}(G_i^0-\Delta_i)} \left(\frac{\lambda_i(\theta_i + G_i^0) + (\Delta_i(r+\mu_i) + \mu_i\frac{\lambda_i}{r})}{r+\lambda_i+\mu_i} \right) f(\theta_i)d\theta_i +$$

$$+\int_{\frac{r}{\lambda_i}(G_i^0-\Delta_i)}^{\infty}\left(\frac{\lambda_i}{r}\theta_i+\Delta_i\right)f(\theta_i)d\theta_i \tag{3}$$

With μ_i limiting to zero, first term of (3) disappears, and the value of G_i^0 satisfies:

$$G_i^0=\int_{-\infty}^{\frac{r}{\lambda_i}(G_i^0-\Delta_i)}\left(\frac{\lambda_i(\theta_i+G_i^0)+(\Delta_i r)}{r+\lambda_i}\right)f(\theta_i)d\theta_i+\int_{\frac{r}{\lambda_i}(G_i^0-\Delta_i)}^{\infty}\left(\frac{\lambda_i}{r}\theta_i+\Delta_i\right)f(\theta_i)d\theta_i \tag{4}$$

and, if compared to expression (1), one can see that $G_i^0=G_i+\Delta_i$. Similar, with $\mu=\infty$, the second term of (3) disappears, and the value G_i^0 satisfies (4).

Now let's show that the value G_i^0 has a U-shape over μ_i: its derivative $(G_i^0)'_{\mu_i}$ is negative if μ_i is low enough, and positive otherwise. When taking the derivative of expression (3) over μ_i, the left-hand side would give a higher coefficient on the derivative $(G_i^0)'_{\mu_i}$, then the right-hand side. Therefore, the sign of $(G_i^0)'_{\mu_i}$ would be determined by the derivative of right-hand side of (3) over μ_i, without including terms $(G_i^0)'_{\mu_i}$. The sign of $(G_i^0)'_{\mu_i}$ coincides with:

$$\int_{-\infty}^{\frac{r}{\lambda_i}(G_i^0-\frac{\Delta_i(r+\mu_i)}{\mu_i})}\frac{r\Delta_i}{r+\lambda_i}\frac{1}{(r+\lambda_i+\mu_i)^2}f(\theta_i)d\theta_i+\int_{\frac{r}{\lambda_i}(G_i^0-\frac{\Delta_i(r+\mu_i)}{\mu_i})}^{\frac{r}{\lambda_i}(G_i^0-\Delta_i)}\frac{\lambda_i(\frac{\lambda_i}{r}\theta_i+\Delta_i-G^0)}{(r+\lambda_i+\mu_i)^2}f(\theta_i)d\theta_i$$

or, if multiplying by $(r+\lambda_i+\mu_i)^2>0$,

$$\int_{-\infty}^{\frac{r}{\lambda_i}(G_i^0-\frac{\Delta_i(r+\mu_i)}{\mu_i})}\frac{r\Delta_i}{r+\lambda_i}f(\theta_i)d\theta_i+\int_{\frac{r}{\lambda_i}(G_i^0-\frac{\Delta_i(r+\mu_i)}{\mu_i})}^{\frac{r}{\lambda_i}(G_i^0-\Delta_i)}\lambda_i(\frac{\lambda_i}{r}\theta_i+\Delta_i-G^0)f(\theta_i)d\theta_i$$

where the first term is positive, and second is negative. When μ_i increases, the value $\frac{r}{\lambda_i}(G_i^0-\frac{\Delta_i(r+\mu_i)}{\mu_i})$ increases, and both the first term and the second term increase. Therefore, there is a cutoff value of μ_i, such that with low values of μ_i the derivative $(G_i^0)'_{\mu_i}$ is negative; and otherwise it's positive. The value G_i^0 has a U-shape over μ_i.

Proof of Theorem 3

Let's denote as $E\theta$ the expected value of prize payoff θ at the risky arm. Let's also assume that:

$$sr>\lambda E\theta+\Delta r \tag{5}$$

Otherwise, the expected payoff at the risky arm would be greater than at the safe arm, making each agent to only pull the risky arm, independently of other agent's action.

Given that θ has not been learnt, the continuation payoff of each agent depends on whether the agent is experienced at the risky arm. Let's denote these continuation payoffs as V^1, V^0 for, respectively, experienced and inexperienced agent. When θ has been learnt, each agent independently decides whether to keep pulling the risky arm. Let's denote the expected continuation payoff of the agent, conditional on θ being learnt, as $M^0=max\{s,\frac{\lambda}{r}\theta+\frac{m\mu}{\mu+r}\}$ for inexperienced agent and $M^1=max\{s,\frac{\lambda}{r}\theta+m\}$ for an experienced agent. That is, when the value of θ is learnt, each agent gets a continuation payoff of either M^0 or M^1, plus if she was the one who obtained prize θ, she receives a lump-sum payoff of θ as well.

At any time, if the experienced agent is pulling the arm for a small interval dt, she gets the following expected payoff:

$$mrdt + \lambda dt(E\theta + M^1) + \gamma\lambda dt M^1 + (1 - (r + \lambda + \gamma\lambda)dt)V^1 \tag{6}$$

Indeed, the experienced agent gets a flow payoff of mr. With probability λdt the experienced agent receives θ, and gets a continuation payoff of M^1. With probability $\gamma\lambda dt$ the other agent learns θ, and the experienced agent gets a continuation payoff of M^1. Otherwise, the continuation payoff retains the same value of V^1.

Similarly, if the experienced agent does not pull the risky arm, her payoff is:

$$srdt + \gamma\lambda dt M^1 + (1 - (r + \gamma\lambda)dt)V^1 \tag{7}$$

The experienced agent prefers to pull the risky arm, if

$$sr \leq mr + \lambda(E\theta + M^1 - V^1) \tag{8}$$

If the inexperienced agent is pulling the risky arm for a small time dt, she gets a payoff:

$$\lambda dt(E\theta + M^0) + \gamma\lambda dt M^0 + \mu dt(V^1 - V^0) + (1 - (r + \lambda + \gamma\lambda)dt)V^0$$

and if pulling the safe arm, she gets a payoff

$$srdt + \gamma\lambda dt M^0 + (1 - (r + \gamma\lambda)dt)V^0$$

The inexperienced agent prefers to pull the risky arm, if

$$sr \leq \lambda(E\theta + M^0 - V^0) + \mu(V^1 - V^0) \tag{9}$$

In equilibrium, the incentives to experiment are strictly higher for the experienced agent (this can be directly checked from the expressions below). Assuming the initial value $\gamma > 0$, both types of agents (weakly) prefer pulling the risky arm, as otherwise at the beginning of the game both agents, being inexperienced at the risky arm, would only pull the safe arm. This yields the following expressions for continuation payoffs:

$$V^1 = \frac{mr + \lambda(E\theta + M^1) + \gamma\lambda M^1}{r + \lambda + \gamma\lambda} \tag{10}$$

$$V^0 = \frac{\lambda(E\theta + M^0) + \gamma\lambda M^0 + \mu V^1}{r + \lambda + \gamma\lambda + \mu} \tag{11}$$

If the value γ lies in $(0, 1)$, then the inexperienced agent is indifferent whether to pull the risky arm. Combining the expressions (9), (10), (11), and substituting the value m for $m = \frac{\Delta(r+\mu)}{\mu}$, one gets the following condition for γ:

$$\Delta(r + \mu)r(r + \gamma\lambda) + \lambda\mu(r + \gamma\lambda)(E\theta + M^1) + \gamma\lambda\mu(r + \gamma\lambda)M^1 =$$

$$= sr(\lambda+\mu)(r+\lambda+\gamma\lambda) + \gamma\lambda(\lambda+\mu)(r+\lambda+\gamma\lambda)M^0 + (sr - \lambda E\theta - \lambda M^0)(r+\gamma\lambda)(r+\lambda+\gamma\lambda) \quad (12)$$

Let's analyze the expression (12) for extreme values of μ. If $\mu = 0$, then the expression $\mu M^1 = max\{\mu s, \frac{\mu\lambda}{r}\theta + m\mu\} = m\mu = \Delta r$, and one gets:

$$\Delta r^2(r+\gamma\lambda) + \lambda(r+\gamma\lambda)\Delta r + \gamma\lambda(r+\gamma\lambda)\Delta r =$$
$$= sr\lambda(r+\lambda+\gamma\lambda) + \gamma\lambda^2(r+\lambda+\gamma\lambda)M^0 + (sr - \lambda E\theta - \lambda M^0)(r+\gamma\lambda)(r+\lambda+\gamma\lambda)$$

or, dividing by $r+\lambda+\gamma\lambda$ and rearranging terms,

$$\Delta r(r+\gamma\lambda) - sr\lambda - (sr - \lambda E\theta)(r+\gamma\lambda) + r\lambda M^0 = 0 \quad (13)$$

If $\mu = \infty$, then the value M^1 equals M^0. Leaving only terms proportional to μ in (12), one gets:

$$\Delta r(r+\gamma\lambda) + \lambda(r+\gamma\lambda)(E\theta + M^0) + \gamma\lambda(r+\gamma\lambda)M^0 = sr(r+\lambda+\gamma\lambda) + \gamma\lambda(r+\lambda+\gamma\lambda)M^0$$

which coincides with the expression (13). Due to inequality 5, the expression (13) decreases with γ. This means there may exist at most one value of $\gamma^* \in (0,1)$, which satisfies (13). If such γ does not exist, the expression (13) has to hold as an inequality, and the value of γ^* equals either 0 or 1. This means that both the cases when $\mu = 0$ and μ limiting to ∞ have the same equilibrium value of γ^*.

Let's now analyze the expression (12) for intermediate values of μ. First, let's assume there exists γ^* such that (13) holds as equality. If one substitutes the expression (13) into the expression (12), the residual term will equal:

$$(r+\gamma\lambda)(\lambda+\gamma\lambda)(\mu[M^1 - M^0] - \Delta r) \quad (14)$$

and by definition of M^0, M^1, is non-positive, since $\mu[M^1 - M^0] - \Delta r \leq 0$. Since the expression (12) is a weighted sum of expressions (13) and (14), with the latter being non-positive, the way to make (12) zero is to make (13) positive. This is achieved by making $\gamma < \gamma^*$.

Let's assume now that the equilibrium value of γ^* at extreme values of μ is zero, and the expression (13) is non-positive. This means that the overall expression (12) is non-positive as well, making the equilibrium value of γ for all intermediate value of μ equal zero. Finally, with $\gamma^* = 1$ the equilibrium value of γ at any intermediate value of μ cannot exceed γ^*. This means that the value γ^* for extreme values of μ always exceeds the equilibrium value of γ for intermediate values.

Having shown that the equilibrium value of γ reaches the maximum at extreme values of μ, let's show the same result for a continuation payoff V^0. First, if at some intermediate value of μ in equilibrium one has $\gamma = 0$, then both agents only pull the safe arm. Both the agents get a continuation payoff of $V^0 = s$, which is the minimal possible value and cannot be larger compared to extreme values of μ. If in equilibrium $\gamma > 0$, then from expressions (10) and (11) the value V^0 equals:

$$V^0 = \frac{\lambda(\theta + M^0) + \gamma\lambda M^0 + \mu\frac{mr+\lambda(\theta+M^1)+\gamma\lambda M^1}{r+\lambda+\gamma\lambda}}{r + \lambda + \gamma\lambda + \mu}$$

Taking the partial derivative $\partial V^0/\partial\mu$ yields its sign to coincide with expression:

$$\Delta r(\lambda + \gamma\lambda) + (\lambda + \gamma\lambda)(\mu[\partial(\mu M^1)/\partial\mu] - \mu M^1) + (\lambda + \gamma\lambda)(r + \lambda + \gamma\lambda)([\partial(\mu M^1)/\partial\mu] - M^0)$$

where the values $\mu[\partial(\mu M^1)/\partial\mu] - \mu M^1$, and $[\partial(\mu M^1)/\partial\mu] - M^0$ are negative, and increase with μ. Moreover, the value V^0 reaches

$$V^0 = \frac{\Delta r + \lambda\theta + (\lambda + \gamma\lambda)M^0}{r + \lambda + \gamma\lambda}$$

at both $\mu = 0$ and $\mu = \infty$. This means that the value V^0 (with fixed γ) has a U-shape as a function of μ and reaches maximum at $\mu = 0$ and ∞.

Since the value of γ reaches the maximum at $\mu = 0$ or μ limiting to infinity, and the value of γ increases the value V^0, the overall value V^0 reaches maximum at extreme values of μ.

Proof of Proposition 3

1. Deterministic case. Let there be two skills, denoted as k, l. The agent can only learn one skill at a time, and has to spend time T_k on skill k and $T_l > T_k$ on skill l. When learned, skills k, l give flow payoffs of rm_k or $rm_l > rm_k$, independently of whether the other skill is learned.

If the agent is learning skill k first, then she gets a flow payoff of rm_k after time T_k and gets a flow payoff of $r(m_k + m_l)$ after time $T_k + T_l$. One can estimate the ex ante value $W^0_{k,l}(t)$ of discounted payoff from skills, accumulated up to time t:

$$W^0_{k,l}(t) = \int_0^t rm_{k,l}(s)e^{-rs}ds$$

where $rm_{k,l}(s)$ is the flow payoff for the agent at time s, which equals 0 if $s < T_k$, equals rm_k if $T_k \leq s < T_k + T_l$, and equals $r(m_k + m_l)$, if $T_k + T_l \leq s$.

Similar, if the agent learns skill l first, and skill k second, the value $W^0_{l,k}(t)$ of the payoff from skills, which the agent has accumulated up to time t, and estimated at $t = 0$, equals:

$$W^0_{l,k}(t) = \int_0^t rm_{l,k}(s)e^{-rs}ds$$

where $rm_{l,k}(s)$ equals 0 if $s < T_l$, equals rm_l if $T_l \leq s < T_k + T_l$, and equals $r(m_k + m_l)$, if $T_k + T_l \leq s$.

By assumption of Proposition 3, $W^0_{k,l}(t = \infty) = W^0_{l,k}(t = \infty)$. Moreover, if the agent learns skill k first, she starts getting a positive flow payoff of rm_k faster, compared to her learning skill l first. After time $T_k + T_l$, regardless of which skill is learnt first, the agent obtains both skills and gets a flow payoff of $r(m_k + m_l)$. This means, that $W^0_{k,l}(t) \geq W^0_{l,k}(t)$, with strict inequality for $T_k < t < T_k + T_l$.

Let the agent now engage in experimentation and be able to switch from the arm with two skills. If the agent starts learning skill k first, then her expected continuation payoff is weakly higher compared to her learning skill l first. Indeed, consider any choice of the agent to whether to pull the arm with two skills as dependent on only what she has learnt about the arm, and not the skills. Since $W^0_{k,l}(t) \geq W^0_{l,k}(t)$, with the same choice of whether to pull the arm, the additional ex ante payoff from skills is always weakly greater in case of learning skill k first. Thus, the agent prefers to start learning skill k.

2. Let now the learning-by-doing process be Poisson. The agent has a choice between two skills: k, l, with parameters of Poisson learning rate $\mu_k > \mu_l$, and flow payoffs $rm_k < rm_l$. If the agent is learning skill k until succeeded, and then is learning skill l, the ex ante continuation payoff from the skills equals:

$$W^0_{k,l} = m_k \frac{\mu_k}{r+\mu_k} + m_l \frac{\mu_k}{r+\mu_k} \frac{\mu_l}{r+\mu_l}$$

That is, the agent gets payoff m_k from skill k, discounted by the value $\frac{\mu_k}{r+\mu_k}$ of expected time of learning skill k; and gets payoff m_l from skill l, discounted by values $\frac{\mu_k}{r+\mu_k}$, $\frac{\mu_l}{r+\mu_l}$ of expected time to learn both skills.

If the agent is first learning skill l, and then skill k, her ex ante continuation payoff from the skills equals:

$$W^0_{l,k} = m_l \frac{\mu_l}{r+\mu_l} + m_k \frac{\mu_l}{r+\mu_l} \frac{\mu_k}{r+\mu_k}$$

By assumption of the Proposition 3, $W^0_{k,l} = W^0_{l,k}$. If the agent has no choice but to pull the arm with two skills, she is indifferent between which skill to learn first.

Let the agent now have a choice of switching away from the arm. In addition to skills, the arm is generating prizes θ with a Poisson rate λ, with θ being distributed with density $f(\theta)$. The agent has to decide on which skill to learn first. To find the solution, one can compare two Gittins indices for the case when the (exogenous) sequence is to learn skill k first, and for the case when the sequence is to learn skill l first. Let's denote those indices as, respectively, $G^0_{k,l}$, $G^0_{l,k}$, and show that $G^0_{l,k} \geq G^0_{k,l}$, proving Proposition 3.

Let's consider the case when the agent learns skill k first. The agent will pull the arm until observing θ, and then decide whether pull the arm forever or switch. The decision will depend on how many skills the agent has acquired before observing θ, and will be a cutoff solution. If the agent has acquired no skills, she will switch the arm if and only if $\theta < \theta^2_{k,l}$, where $\theta^2_{k,l}$ satisfies:

$$G^0_{k,l} = \frac{\lambda}{r}\theta^2_{k,l} + W^0_{k,l}$$

If the agent has acquired skill k, but not l, she will switch from the arm if and only if $\theta < \theta^1_{k,l}$, where $\theta^1_{k,l}$ satisfies:

$$G^0_{k,l} = \frac{\lambda}{r}\theta^1_{k,l} + m_k + \frac{m_l \mu_l}{r+\mu_l}$$

If the agent has acquired both skills k, l, she will switch the arm if and only if $\theta < \theta^0_{k,l}$, where $\theta^0_{k,l}$ satisfies:

$$G^0_{k,l} = \frac{\lambda}{r}\theta^0_{k,l} + m_k + m_l$$

Conditional on θ, the agent's ex ante continuation payoff, denoted as $V^0_{k,l}(\theta)$, depends on which interval $(-\infty, \theta^0_{k,l})$, $(\theta^0_{k,l}, \theta^1_{k,l})$, $(\theta^1_{k,l}, \theta^2_{k,l})$, $(\theta^2_{k,l}, \infty)$ the value θ belongs to. These continuation payoffs can be calculated similar to expression (2), and are given below.

If $\theta < \theta^0_{k,l}$, then

$$V^0_{k,l}(\theta) = (\theta + G^0_{k,l})\frac{\lambda}{r+\lambda} + m_k\frac{r}{r+\lambda}\frac{\mu_k}{r+\lambda+\mu_k} + m_l\frac{r}{r+\lambda}\frac{\mu_k}{r+\lambda+\mu_k}\frac{\mu_l}{r+\lambda+\mu_l} \tag{15}$$

If $\theta^0_{k,l} < \theta < \theta^1_{k,l}$, then

$$V^0_{k,l}(\theta) = (\theta + G^0_{k,l})\frac{\lambda}{r+\lambda} + m_k\frac{\mu_k}{r+\lambda+\mu_k}\frac{r+\mu_l}{r+\lambda+\mu_l} + (m_l + \theta\frac{\lambda}{r})\frac{\mu_k}{r+\lambda+\mu_k}\frac{\mu_l}{r+\lambda+\mu_l} \tag{16}$$

If $\theta^1_{k,l} < \theta < \theta^2_{k,l}$, then

$$V^0_{k,l}(\theta) = (\theta + G^0_{k,l})\frac{\lambda}{r+\lambda+\mu_k} + (m_k + m_l\frac{\mu_l}{r+\mu_l} + \theta\frac{\lambda}{r})\frac{\mu_k}{r+\lambda+\mu_k} \tag{17}$$

Finally, if $\theta^2_{k,l} < \theta$, then

$$V^0_{k,l}(\theta) = \theta\frac{\lambda}{r} + W^0_{k,l} \tag{18}$$

The Gittins index $G^0_{k,l}$ for the arm with skill k being learnt before skill l, can be represented as the expected value of ex ante continuation payoff $V^0_{k,l}(\theta)$:

$$G^0_{k,l} = \int_{-\infty}^{\theta^0_{k,l}} \left((\theta + G^0_{k,l})\frac{\lambda}{r+\lambda} + m_k\frac{r}{r+\lambda}\frac{\mu_k}{r+\lambda+\mu_k} + m_l\frac{r}{r+\lambda}\frac{\mu_k}{r+\lambda+\mu_k}\frac{\mu_l}{r+\lambda+\mu_l}\right) f(\theta)d\theta + \tag{19}$$

$$+ \int_{\theta^0_{k,l}}^{\theta^1_{k,l}} \left((\theta + G^0_{k,l})\frac{\lambda}{r+\lambda} + m_k\frac{\mu_k}{r+\lambda+\mu_k}\frac{r+\mu_l}{r+\lambda+\mu_l} + (m_l + \theta\frac{\lambda}{r})\frac{\mu_k}{r+\lambda+\mu_k}\frac{\mu_l}{r+\lambda+\mu_l}\right) f(\theta)d\theta +$$

$$+ \int_{\theta^1_{k,l}}^{\theta^2_{k,l}} \left((\theta + G^0_{k,l})\frac{\lambda}{r+\lambda+\mu_k} + (m_k + m_l\frac{\mu_l}{r+\mu_l} + \theta\frac{\lambda}{r})\frac{\mu_k}{r+\lambda+\mu_k}\right) f(\theta)d\theta + \int_{\theta^2_{k,l}}^{\infty} \left(\theta\frac{\lambda}{r} + W^0_{k,l}\right) f(\theta)d\theta$$

If the player has instead to start with skill l, she will expect to get the similar expression, but with interchanging k, l. That is, there are two expressions (19): one is for the case when skill k is learnt first (as above), and the alternative, denoted as (19)' is when skill l is learnt first, with interchanged k, l. The optimal decision on which skill depends on which value of Gittins index is higher: $G^0_{k,l}$, or $G^0_{l,k}$.

To show that $G^0_{l,k} \geq G^0_{k,l}$, let's observe that as in (19) (and (19)'), each of the two Gittins indices equals expected value for the related ex ante continuation payoff ($V^0_{k,l}(\theta)$ in case of skill k learnt first). Let's set $G^0_{l,k} = G^0_{k,l} = G^0$, and show that for all θ, the value $V^0_{k,l}(\theta)$ would weakly increase if instead skill l is learnt first, or, respectively, if indices k, l would be interchanged.

The continuation payoff, $V^0_{k,l}(\theta)$, in cases of (15) and (18) does not change with interchanging k, l (and fixing G^0). Similarly, the cutoff values $\theta^0_{k,l}, \theta^2_{k,l}$ do not change if interchanging k, l. The derivative of

continuation payoff $V^0_{k,l}(\theta)$ over θ in expression (16) does not change if skill l is learns first; and the same derivative $V^0_{k,l}(\theta)$ over θ in expression (17) is weakly lower in case of skill l learnt first. Moreover, the cutoff value $\theta^1_{k,l}$ is higher in case when skill l is learnt first.

This means that the continuation payoff $V^0_{k,l}(\theta)$ does not change with interchanging k, l, on the intervals $\theta \in (-\infty, \theta^0_{k,l}) \cup (\theta^2_{k,l}, \infty)$. The continuation payoff $V^0_{k,l}(\theta)$ weakly increases if skill l is learnt first, on the interval $\theta \in (\theta^0_{k,l}, \theta^2_{k,l})$. This interval corresponds to the case in which if the agent learns skill l, she stays at the arm; and the agent might not stay at the arm if instead she only learns skill k. Therefore, learning skill l first increases the agent's chances to stay at the arm, and increases the expected continuation payoff.

The continuation payoff $V^0_{k,l}(\theta)$ is therefore weakly higher for all θ, in case of skill l learnt first. The Gittins index $G^0_{l,k} \geq G^0_{k,l}$, showing Proposition 3.