Optimal Allocation Strategies in a Discrete-Time Two-Armed Bandit Problem

Audrey Hu[*]
City University of Hong Kong

Liang Zou
University of Amsterdam

February 28, 2024

Abstract. This study addresses a two-armed bandit problem involving a "safe" and a "risky" arm across a countable number of periods. The agent, with one time unit per period, strategically allocates time between these two arms aiming at achieving a "breakthrough." The risky arm's type is unknown, which can be “good” or “bad,” and breakthrough depends on proving it to be good. Breakthrough probability is an exponential function of the allocated time, given the risky arm is good. Departing from the “either-or” binary choices in previous studies, we explore smooth allocation strategies in the $[0, 1]$ range. Our analytical solution reveals that the optimal allocation plan significantly differs from binary strategies, and stopping after any finite periods of unsuccessful trials is suboptimal. A methodological contribution of this study lies in a problem transformation that enhances tractability, going beyond the standard Bellman-equation approach for bandit problems.

Keywords: two-armed bandit; learning; discrete time; exponential distribution.

[*] *Corresponding author.* Address: 9-256, Lau Ming Wai Academic Building, City University of Hong Kong, Hong Kong SAR. Telephone number: +852 34426767. Email: audrey.hu@cityu.edu.hk.

1 Introduction

This paper investigates a discrete-time, two-armed bandit problem, representing a classic dilemma between exploitation and exploration. The scenario involves a choice between a "safe" arm, which guarantees a certain return, and a "risky" arm, which could turn out to be either "good" or "bad". The "good" type offers a higher present value than the safe arm, while the "bad" type is worthless. An economic agent, endowed with one unit of time per period, has to decide how to allocate his time $\alpha_t \in [0,1]$ between the two arms in each period $t = 1, 2, \ldots$. A "breakthrough" occurs when the risky arm is revealed to be good, thereby successfully ending the problem. However, if the type is bad, the agent's exploration time is wasted. This bandit model is representative of a wide range of real-world problems, including R&D, drug trials, mineral site exploration, and the pursuit of proof or evidence for a conjecture.

Despite extensive research, analyses of bandit problems have been largely confined to discrete choice sets or convex/linear payoff functions. Consequently, the optimal strategies are typically characterized by an optimal stopping time combined with an indexing policy—in line with the celebrated index theorem by Gittins and Jones (1974)[1]. However, in many situations, the decision variables of interest are naturally continuous, such as the amount of time, monetary budget, monopoly prices, etc. When these decision variables affect the probability of a breakthrough such that the objective functions are concave, we would expect the optimal allocation policies to be characterized by interior solutions. Indeed, the scarcity of research on smooth solutions to bandit problems is more due to analytical complexity rather than lack of relevance. For instance, in his pioneering work applying bandit theory to stores' learning about consumer demand by changing prices over time, Rothschild (1974) concluded: "I assumed ... that there were only two prices that stores could charge

[1] See, e.g., Bergemann and Valimaki (2010) for a survey of the extensive literature on economic analyses of bandit problems.

. . . However prices are often considered to be naturally continuous variables, and it is not clear that Theorems I and II hold when they are. This seems an open and difficult question [p.200]."

A methodological contribution of our study is to demonstrate that the class of two-armed bandit problems under investigation, i.e., those involving either a "breakthrough" or "nothing" (henceforth "BorN"), can be formulated in tractable forms by a transformation lemma (see Lemma 1). The standard Bellman-equation approach to bandit problems typically involves working on the future posterior beliefs as state variables. When allocation strategies affect the subsequent posterior beliefs, the problem rapidly becomes intractable. Our re-formulation of the problem reveals that, at least for the BorN bandits, tracking the future posteriors is a redundant exercise. The insight is that the continuation payoffs in the future depend on unrealized posteriors, and by the law of iterated expectations, the expected continuation payoffs can be calculated without invoking the posterior conditional probabilities. By transforming the problem, we are able to derive clear analytical solutions and perform comparative statics exercises in a transparent manner. We show that when the conditional probability of a breakthrough is an exponential function of the time allocated to the risky arm, no indexing policy combined with a stopping time can be optimal. The optimal allocation plan in our setting is a gradually declining sequence in $(0, 1]$ that converges to zero as t tends to infinity.

Recently, the economic literature on bandit problems has largely focused on continuous time. As Bolton and Harris (1999) acknowledged in their seminal paper: "We have chosen a continuous-time formulation of the two-armed bandit problem because of its tractability [p. 350]." In a BorN setting with exponential bandits, Malueg and Tsutsui (1997) were the first to characterize optimal allocation strategies by interior solutions, assuming a quadratic cost function. They observed the possibility of no stopping but assumed it away by introducing a fixed cost component in the cost function. Keller, Rady, and Cripps (2005), in their seminal work, formulated the BorN problem in continuous time and derived the optimal solution for exponential bandits. This is the case where, given any positive and constant proportion of time allocated to the risky arm, the probability of a breakthrough has a constant hazard rate that is positive for the good type and zero for the bad type. The solution features a bang-bang strategy: there exists a cut-off posterior belief such that it is optimal to allocate full time to the risky arm when the posterior is above the cut-off, and shift to allocating full time to the safe arm when the posterior drops below the cut-off. The strategy implies an optimal stopping time: given no breakthrough by that time, no resource will be allocated to the risky arm anymore. The foundation of their optimal/efficient result is the fact that the exponential distribution reduces to an instantaneous linear function of the allocation variable. The Keller, Rady, and Cripps' (2005) continuous-time BorN framework for exponential bandits has been since adopted by numerous follow-up studies. The bang-bang strategy remains a common feature of these studies—either by assumption (e.g., Awaya and Krishna 2021; Thomas 2021) or by derivation (e.g., Besanko and Wu 2013; Besanko, Tong, and Wu 2018).

Our analysis of the discrete-time counterpart of the exponential bandit problem reveals a very different story. The bang-bang strategy in discrete time is, in general, suboptimal when the action set is $[0,1]$ rather than $\{0,1\}$. Moreover, for the exponential bandit problem considered, the optimal allocation strategy features a "never give up" attitude toward exploring the risky arm. A rough intuition for this result is that when α_t is sufficiently close to 0, it does no affect the change in the posterior very much. In turn, when the posterior does not change much, it does not cause the subsequent allocation to change much. Consequently, as t tends toward infinity, α_t gradually declines toward 0 but there exists no "last period" in which α_t would drop to 0 completely. Obviously, this cannot happen with the bang-bang strategies. Therefore, conclusions from the present study complement and significantly enrich the state-of-the-art results derived either in continuous time or under linear probabilities.

2 The Model

Time is discrete, with a countable number periods $t = 1, 2, \ldots.$ A decision maker (henceforth, agent), endowed with one unit of a perfectly divisible resource (henceforth, time) per period, faces an identical two-armed bandit problem. One arm is "safe;" the other is "risky." The agent must allocate in each period t a fraction of his time, denoted $a_t \in [0, 1]$, to the risky arm and $1 - a_t$ to the safe arm. In each period t, the safe arm yields a known payoff that is proportional to the fraction of the resource allocated to it, i.e., $(1 - a_t)\ell$. The discount factor is $\delta \in (0, 1)$. Thus, the safe arm offers a present value of $L = \ell/(1-\delta)$ when the agent allocates full time to it indefinitely. The risky arm can be either "good" or "bad." It yields nothing if it is bad, and, if it is good, has a present value $G = \gamma L$ where $\gamma > 1$ measures the attractiveness of the good risky arm relative to the safe arm.[2] If the risky arm is good, the probability of a breakthrough by the agent, e.g., finding solid evidence (or proof) that the arm is good, is $p(a_t) = 1 - e^{-\lambda a_t}$, $\lambda > 0$.

The agent has a prior probability belief $\pi_0 \in [0, 1]$ that the risky arm is good. So, by the Bayes rule, if the agent has chosen allocations $a_1 ..., a_t$ in periods 1 through t without a breakthrough, the updated probability (or posterior belief) that the risky arm is good equals

$$\pi_t = \frac{\pi_0 \prod_{s=1}^{t} (1 - p(a_s))}{1 - \pi_0 + \pi_0 \prod_{s=1}^{t} (1 - p(a_s))} \tag{1}$$

[2] Under risk neutrality, defining returns from the "good" risky arm by a present value G is legitimate and general. For instance, we can write $G = g/(1 - \delta)$ and interpret g as a per period expected return from the risky arm when the agent allocates full time to it. We can also write $G = L + R$ and interpret R as a lump sum reward to each agent for their research breakthrough, etc.

3 The Bandit Problem

The bandit problem boils down to maximizing the agent's expected, discounted payoff. By the principle of optimality for dynamic programming, the agent's optimal conditional payoff V_t^*—given no breakthrough until period $t-1$—can be described by the Bellman equation[3]

$$V_t^*(\pi_{t-1}) = \max_{a_t \in [0,1]} V_t(a_t, \pi_{t-1}) \tag{2}$$

where

$$\begin{aligned} & V_t(a_t, \pi_{t-1}) \\ = \quad & (1-a_t)\ell + \delta\left\{\pi_{t-1}p(a_t)G + (1-\pi_{t-1}p(a_t))V_{t+1}^*(\pi_t)\right\} \end{aligned} \tag{3}$$

This program has the following interpretations. Suppose the risky arm has returned nothing until period $t-1$, and that the agent's past allocations imply the posterior π_{t-1}. If the agent allocates a_t to the risky arms in period t, he receives an immediate return $(1-a_t)\ell$ from the safe arm. By the end of period t, with probability $\pi_{t-1}p(a_t)$ there will be a breakthrough, in which case the exploration is over and the agent enjoys the present value of the risky arm G. With probability $1-\pi_{t-1}p(a_t)$, however, the risky arm yields nothing. Then, the problem continues and the agent's continuation conditional payoff becomes $V_{t+1}^*(\pi_t)$ given updated posterior π_t. If the agent stops the experiment after period t, given that no breakthrough has occurred, we assume all resources will be allocated to the safe arm indefinitely from $t+1$ onwards. Consequently, stopping in period $t+1$ implies $V_{t+s} = L$ for all $s \geq 1$.

Now let $\boldsymbol{\alpha} = (\alpha_t(\pi_{t-1}))_{t=1}^{\infty}$ denote the optimal allocation plan that will be implemented by each agent, such that

$$\alpha_t(\pi_{t-1}) \in \arg\max_{a_t \in [0,1]} V_t(a_t, \pi_{t-1})$$

[3] Our analysis and results may treat t as a variable, hence it is useful to keep the subscript for V_t and a_t.

By standard arguments (e.g., Stokey et al. 1989, Chapter 4), we will establish the existence and uniqueness of $\boldsymbol{\alpha}$. For now, notice that the simple structure of V_t implies that it is differentiable in $(a_t, \pi_{t-1}) \in [0,1]^2$.[4] Thus, for all $a_t = \alpha_t(\pi_{t-1})$ at which (2) has an interior solution, they must satisfy the dynamic first-order condition:

$$\begin{aligned} \frac{\partial}{\partial a_t} V_t(a_t, \pi_{t-1}) &= -\ell + \delta \pi_{t-1} \lambda e^{-\lambda a_t} \left[G - V_{t+1}^*(\pi_t) \right] & (4) \\ &\quad + \delta(1 - \pi_{t-1}\left(1 - e^{-\lambda a_t}\right)) V_{t+1}^{*\prime}(\pi_t) \frac{\partial \pi_t}{\partial a_t} & (5) \\ &= 0 \end{aligned}$$

where, by the envelope theorem,

$$V_{t+1}^{*\prime}(\pi_t) = \frac{\partial V_{t+1}(\alpha_{t+1}(\pi_t), \pi_t)}{\partial \pi_t} \tag{6}$$

The term in (5) captures the learning effect of our bandit problem. Since the agent's payoff increases in his subjective probability that the risky arm is good, implying $V_{t+1}^{*\prime}(\pi_t) \geq 0$ (see Corollary 1) and since

$$\frac{\partial \pi_t}{\partial a_t} = -\frac{(1 - \pi_{t-1})\, \pi_{t-1} \lambda e^{-\lambda a_t}}{\left(1 - \pi_{t-1}\left(1 - e^{-\lambda a_t}\right)\right)^2} < 0,$$

the sign of this learning effect is negative. Therefore, compared to the myopic solution that equates the right-hand side of (4) to zero, for all interior solutions for the program in (2), the optimal allocations under the learning effects are lower than the myopic solutions.

Of course, the bandit problem should be sufficiently interesting so that the risky arm is worthy of experimentations. We invoke the following assumption.

Assumption 1 $\pi_0 > \pi_{\min}$, *where* $0 < \pi_{\min} = \frac{1-\delta}{\delta\lambda(\gamma-1)} < 1$.

If exploring the risky arm is deemed unprofitable, the agents' optimal payoff derives solely from the returns from the safe arm so that $V_t^* \equiv L$ for all $t \geq 1$. To see its implications, substituting $V_2^* = L$ into (4)-(5) to get

[4] At the boundaries of $[0,1]^2$, the (cross-partial) derivatives of V are defined as usual by taking limits.

$$\begin{aligned}\frac{\partial}{\partial a_1}V_1(a_1,\pi_0) &= -\ell + \delta\pi_0\lambda e^{-\lambda a_1}(G-L)\\ \text{and } \frac{\partial^2}{\partial a_1^2}V_1(a_1,\pi_0) &= -\delta\pi_0(\lambda)^2 e^{-\lambda a_1}(G-L) < 0\end{aligned}$$

So, $\pi_0 \leq \pi_{\min}$ implies $\frac{\partial}{\partial a_1}V_1(a_1,\pi_0)|_{a_1=0} \leq 0$ and since the objective function is concave, Assumption 1 provides a necessary condition for the risky arm to be of interest. Given any π_0, the composition of $\pi_{\min}$ suggests that the risky arm is worthy of some investment of time as long as δ, λ, or γ is sufficiently high.

The next assumption ensures congruency for this study.

Assumption 2 $\delta\lambda e^{-\lambda}G - \left(1-\delta e^{-\lambda}\right)L \geq 0.$

The role of this assumption can be illustrated by considering an extreme case with $\pi_0 = 1$ and $T = \infty$. Since π_0 is the agents' *subjective* probability, to realize the present value of the risky arm G they need to provide verifiable evidence, say, to the public or supervision authorities. Suppose this has to be done through experimentation until a breakthrough occurs. Then, Assumption 2 ensures that the team will spend full time exploring the risky arm until achieving a breakthrough, or else never stop. To see this, we replace π_{t-1} and a_t with 1 in (3), and obtain the optimal payoff from the Bellman equation:

$$\begin{aligned}V^*(1) &= \delta\left(1-e^{-\lambda n}\right)G + \delta e^{-\lambda n}V^*(1)\\ &= \frac{\delta\left(1-e^{-\lambda n}\right)G}{1-\delta e^{-\lambda n}}\end{aligned} \tag{7}$$

Since there is no learning in this case, the experimentation plan with $\alpha_t \equiv 1$ must satisfy, substituting (7),

$$\begin{aligned}\frac{\partial}{\partial a_t}V(a_t,1)|_{a_t=1} &= -\ell + \delta n\lambda e^{-\lambda n}\left(G - V^*(1)\right)\\ &= -\ell + \delta n\lambda e^{-\lambda n}G\left(\frac{1-\delta}{1-\delta e^{-\lambda n}}\right) \geq 0\end{aligned}$$

Assumption 2 then implies that the above inequality holds for all $n \geq 1$ (recalling $L = \ell/(1-\delta)$). On the other hand, it is worth observing that when Assumption

2 fails to hold, then, even if the agents believe that the risky arm is good for sure (i.e., $\pi_0 = 1$), they may not be interested in exploring the arm if λ is sufficiently low (i.e., too hard to achieve a breakthrough) or if δ is sufficiently low (i.e., too costly to wait for a breakthrough).

4 Simplifying the Problem

Although the Bellman equation approach in the previous section is quite standard, we present here a re-formulation of the Bellman equation that enhances the tractability of the bandit problem.

Consider a mathematically equivalent problem to (2)-(3) of maximizing

$$\begin{aligned} & V_t(a_t, \pi_{t-1}) - G \\ = \quad & (1 - a_t)\ell - (1 - \delta)G + (1 - \pi_{t-1}p(a_t))\, \delta \left[V^*_{t+1}(\pi_t) - G\right] \end{aligned}$$

To simplify notation, denote $g = (1 - \delta)G$ and define

$$q_t = \pi_{t-1}p(a_t) \text{ and } q^*_t = \pi_{t-1}p(\alpha_t), \text{ for } t = 1, 2, ...$$

Thus, both q_t and q^*_t are the probabilities of a breakthrough in period t. The difference between the two functions is that the former can vary with the control variable a_t while the latter is a fixed quantity with α_t being a point on a given (optimal) path of the allocations. The transformed value function $V_t - G$ can be then expanded *as though* it was the expected sum of a sequence of discounted returns (which are negative here), with an associated probability of receiving the return in each period:

Period	Probability	Discounted payoff
t	1	$-\left[g-(1-a_t)\ell\right]$
$t+1$	$1-q_t$	$-\delta\left[g-(1-\alpha_{t+1})\ell\right]$
...	...	...
$t+s$	$(1-q_t)\prod_{r=1}^{s-1}(1-q^*_{t+r})$	$-\delta^s\left[g-(1-\alpha_{t+s})\ell\right]$
...	...	...

Defining $\Pi^0_{r=1}(\cdot)\equiv 1$ and summing up, we find

$$\begin{aligned}&V_t(a_t,\pi_{t-1}) = G-\left[g-(1-a_t)\ell\right]\\&-\sum_{s=1}^{\infty}\delta^s\left((1-q_t)\prod_{r=1}^{s-1}(1-q^*_{t+r})\right)\left[g-(1-\alpha_{t+s})\ell\right]\end{aligned} \tag{8}$$

Here, given no breakthrough in the previous $t-1$ periods, the term in large brackets is the conditional probability of no breakthrough over the next s periods under the current period allocation a_t and the optimal future allocation plan $\alpha_{t+1},...,\alpha_{t+s-1}$. This transformation allows us to arrive at a much simpler form of the problem, as stated in the following lemma.

Lemma 1 (Transformation) *Given (2)-(3),*

$$V_t(a_t,\pi_{t-1}) = V_t(0,\pi_{t-1})-a_t\ell+\delta\pi_{t-1}p(a_t)H_{t+1} \tag{9}$$

where H_{t+1} is a positive function defined recursively by

$$H_{t+1}=g-(1-\alpha_{t+1})\ell+\delta(1-p(\alpha_{t+1}))H_{t+2} \tag{10}$$

for $t=1,2,....$

Proof. The probability of no breakthrough over the next s periods, starting from period t, can be written alternatively as

$$\begin{aligned}&(1-q_t)\prod_{r=1}^{s-1}(1-q^*_{t+r})\\=\;&1-\pi_{t-1}+\pi_{t-1}\left(1-p(a_t)\right)\prod_{r=1}^{s-1}(1-p(\alpha_{t+r}))\end{aligned} \tag{11}$$

Substitute (11) into (8), and define

$$H_{t+1} = \sum_{s=1}^{\infty} \delta^{s-1} \left(\prod_{r=1}^{s-1} (1 - p(\alpha_{t+r})) \right) \left[g - (1 - \alpha_{t+s})\ell \right]. \tag{12}$$

We obtain by re-arranging the terms that

$$\begin{aligned} V_t(a_t, \pi_{t-1}) &= G - [g - (1 - a_t)\ell] \\ &- (1 - \pi_{t-1}) \sum_{s=1}^{\infty} \delta^s \left[g - (1 - \alpha_{t+s})\ell \right] - \pi_{t-1} (1 - p(a_t))\, \delta H_{t+1} \end{aligned} \tag{13}$$

In particular, choosing $a_t = 0$ gives

$$\begin{aligned} V_t(0, \pi_{t-1}) \;\; &= \;\; G - (g - \ell) \\ &\quad - (1 - \pi_{t-1}) \sum_{s=1}^{\infty} \delta^s \left[g - (1 - \alpha_{t+s})\ell \right] - \pi_{t-1} \delta H_{t+1} \end{aligned}$$

Subtracting this from (13) yields (9).

Since $g > \ell$, then $H_{t+1} > 0$. For $T < \infty$, the exploration stops at $T + 1$ with $\alpha_{T+s} \equiv 0$ for all $s \geq 1$ so that $H_{T+1} = (g - \ell) \sum_{s=1}^{\infty} \delta^{s-1} = G - L$. For $T = \infty$, $\lim_{t \to \infty} H_{t+1} = G - L$ can be seen as a transversality condition consistent with any arbitrarily large, finite T. The rest of the proof follows by straightforward verifications. ■

In this lemma, $V_t(0, \pi_{t-1})$ is the agent's expected payoff from a hypothetical situation of forfeiting the optima planned action α_t, allocating full time to the safe arm, in period t and then follow the previously scheduled optimal allocation path from $t + 1$ onward. Thus, the difference

$$\hat{V}_t(a_t, \pi_{t-1}) = V_t(a_t, \pi_{t-1}) - V_t(0, \pi_{t-1})$$

is the expected benefit of experimenting the risky arm in excess of the safe return in period t. For this reason, we call $\hat{V}_t$ the *risk premium* offered by of the risky arm, and $\hat{V}_t^*(\pi_{t-1}) = \hat{V}_t(\alpha_t(\pi_{t-1}), \pi_{t-1})$ the *optimal risk premium*, in period t. Lemma

1 shows that the solution to the maximization program in (2)-(3) is given by the solution to the program

$$\max_{a_t \in [0,1]} \hat{V}_t(a_t, \pi_{t-1}) \qquad (= -a_t \ell + \delta \pi_{t-1} p(a_t) H_{t+1}) \tag{14}$$

where H_{t+1} depends only on the subsequent planned actions and can be treated as a "given" for each $t \geq 1$. This is a dramatic simplification of the problem.

An example of this simplification is the corollary below.

Corollary 1 *For all $t = 1, 2, ...$, $V_t^{*\prime}(\pi_{t-1}) \geq 0$. The inequality holds strictly unless $\alpha_{t+r} \equiv 0$ for all $r = 0, 1, 2, ...$*

Proof. From (12), we have

$$\begin{aligned} H_{t+1} &= \sum_{s=1}^{\infty} \delta^{s-1} \left(\prod_{r=1}^{s-1} (1 - p(\alpha_{t+r})) \right) [g - (1 - \alpha_{t+s})\ell] \\ &\leq \sum_{s=1}^{\infty} \delta^{s-1} [g - (1 - \alpha_{t+s})\ell] \end{aligned}$$

Therefore, replacing a_t with α_t in (13), and differentiating gives

$$\frac{\partial V_t(\alpha_t, \pi_{t-1})}{\partial \pi_{t-1}} = \sum_{s=1}^{\infty} \delta^s [g - (1 - \alpha_{t+s})\ell] - (1 - p(\alpha_t)) \delta H_{t+1} \geq 0$$

Both the above inequalities hold as an equality only if $\alpha_{t+r} \equiv 0$ for all $r = 0, 1, 2, ...$ Thus, by the envelope theorem, $V_t^{*\prime}(\pi_{t-1}) > 0$ as long as the experimentation of the risky arm continues. ■

This corollary confirms that the learning effect on the continuation payoff in (5), given no breakthrough in the current period, is strictly negative.

The maximization program in (14) has also a very transparent economic interpretation: If $\boldsymbol{\alpha} = (\alpha_t)_{t=1}^{\infty}$ is an optimal allocation plan, then each α_t maximizes the risk premium in period t given the subsequent planned actions. This objective, although equivalent to the one in (2)-(3), seems more relevant in practical terms, as the breakthrough can occur in any period, ending the need for further experimentation.

Proposition 1 *Suppose* $\boldsymbol{\alpha} = (\alpha_t)_{t=1}^{\infty}$ *is an optimal allocation plan. Then, the optimal risk premium in period* t *equals*

$$\hat{V}_t^* = \frac{e^{\alpha_t \lambda} - 1}{\lambda} (\ell + \eta_t) - \alpha_t \ell \tag{15}$$

where $\eta_t = (\geq)\ 0$ *if* $\alpha_t < (=)\ 1$.

Proof. Since V_t in (9) is concave in a_t, so is $\hat{V}_t$. Thus, α_t is characterized by the first-order condition

$$\frac{\partial}{\partial a_t} \hat{V}_t(a_t, \pi_{t-1})|_{a_t = \alpha_t} = -\ell + \delta \pi_{t-1} \lambda e^{-\alpha_t \lambda} H_{t+1} = \eta_t \tag{16}$$

where η_t is the Lagrangian multiplier associated with the constraint $\alpha_t \leq 1$. (The constraint $\alpha_t \geq 0$ is irrelevant because by definition $\hat{V}_t(0, \pi_{t-1}) = 0$.) This gives

$$H_{t+1} = \frac{\ell + \eta_t}{\delta \pi_{t-1} \lambda e^{-\alpha_t \lambda}}$$

Substituting into (9), we obtain

$$\begin{aligned} \hat{V}_t^* &= \hat{V}_t(\alpha_t, \pi_{t-1}) \\ &= \frac{e^{\alpha_t \lambda} - 1}{\lambda} (\ell + \eta_t) - \alpha_t \ell \end{aligned}$$

■

The optimal risk premium given in (15) has a natural interpretation as follows. The term $\alpha_t \ell$ is the opportunity cost of investment of time α_t in the risky arm. The term $\ell + \eta_t$ derives from (16), which equals the *marginal* benefit of investing α_t in the risky arm: $\delta \pi_{t-1} \lambda e^{-\alpha_t \lambda} H_{t+1}$. The overall benefit of investing α_t in the risky arm is then given by the first term in (15), in which we note that $\frac{e^{\alpha_t \lambda} - 1}{\lambda}$ equals the familiar ratio of 1 over the reversed hazard rate: $\frac{p(\alpha_t)}{p'(\alpha_t)}$. It is worth noting that $\hat{V}_t^*$ is a direct function of α_t, depending on the posterior π_{t-1} only indirectly through α_t.

Compared with (5), we can see from (16) that the functional H_{t+1} captures in a way the overall marginal benefit, including the learning effects, of investing time in the risky arm. The simplicity of the proof for Proposition 1 is largely due to the fact that H_{t+1} does not depend directly on the process of the posteriors $\{\pi_t, \pi_{t+1}, ...\}$, as does V^*_{t+1}.

Of course, we still have the burden to solve for the optimal allocation plan $\boldsymbol{\alpha}$.

5 Optimal Allocation Plan

In this section, we analyze the optimal allocation plan. If the agent decides to stop the experiment after period t, given that no breakthrough has occurred, we assume all resources will be allocated to the safe arm indefinitely from $t+1$ onwards. Consequently, stopping in period $t+1$ implies $H_{t+1} = G - L$.

Proposition 2 *Suppose Assumptions 1-2 hold and let $\pi_0 \in (\pi_{\min}, 1)$ be given. (i) The optimal allocation plan $\boldsymbol{\alpha} = (\alpha_t)_{t=1}^{\infty}$ satisfies $\alpha_t > 0$ for all $t \geq 1$, i.e., the experimentation of the risky arm never stops without a breakthrough. (ii) If $\alpha_1 < 1$, then $(\alpha_t)_{t=1}^{\infty}$ is a strictly decreasing sequence that converges to 0 as $t \to \infty$, in which each α_t is determined by*

$$\alpha_t = \frac{1}{\lambda} \ln \frac{\delta\pi_{t-1}\left(1 + \lambda\left(\gamma - 1 + \alpha_{t+1}\right)\right)}{\delta\pi_{t-1} + 1 - \delta} \in (0,1) \tag{17}$$

where the sequence of posteriors $(\pi_t)_{t=0}^{\infty}$,

$$\pi_t = \frac{\pi_{t-1}e^{-\lambda a_t}}{1 - \pi_{t-1}\left(1 - e^{-\lambda a_t}\right)}, \tag{18}$$

converges to $\pi_{\min}$ as $t \to \infty$. (iii) If $\alpha_1 = 1$, there exists $\tau \geq 1$ such that $\alpha_t = 1$ for all $t \leq \tau$ and $\alpha_t < 1$ for all $t > \tau$. Conclusion (ii) regarding the properties of $(\alpha_t)_{t=1}^{\infty}$ holds for $(\alpha_t)_{t=\tau+1}^{\infty}$.

Proof. (i) It suffices to show that $\pi_{t-1} > \pi_{\min}$ implies $\pi_t > \pi_{\min}$ for all $t \geq 1$. We prove this by contradiction. Pick any t such that $\pi_{t-1} > \pi_{\min}$, or $\alpha_t > 0$. Suppose $\pi_t \leq \pi_{\min}$. Then $\alpha_{t+1}(\pi_t) = 0$, implying $H_{t+s} = G - L$ for all $s = 1, 2, ...$ It follows then from (9) of Lemma 1 that α_t, π_{t-1} and π_t must satisfy

$$\begin{aligned} \frac{\partial}{\partial a_t} V_t(\alpha_t, \pi_{t-1}) &= -\ell + \delta\pi_{t-1}\lambda e^{-\lambda\alpha_t}(G - L) \geq 0 \\ \text{and } \frac{\partial}{\partial a_{t+1}} V_t(0, \pi_t) &= -\ell + \delta\pi_t\lambda(G - L) \leq 0 \end{aligned}$$

where the second inequality derives from the hypothesis $\pi_t \leq \pi_{\min}$. Cancelling terms, the above two conditions imply

$$\frac{\pi_{t-1}}{\pi_t} \geq e^{\lambda \alpha_t}. \tag{19}$$

By the Bayes rule, (18) holds and thus $\alpha_t > 0$ implies

$$\frac{\pi_{t-1}}{\pi_t} = e^{\lambda \alpha_t}\left(1 - \pi_{t-1}\right) + \pi_{t-1} < e^{\lambda \alpha_t} \tag{20}$$

The contradiction between (19) and (20) proves $\alpha_t > 0$ for all $t \geq 1$.

(ii) By Lemma 1, differentiating V_t in (9) and applying (10) give us

$$\begin{aligned} \frac{\partial}{\partial a_t} V_t(a_t, \pi_{t-1}) &= -\ell + \delta \pi_{t-1} \lambda e^{-\lambda a_t} H_{t+1} \\ &= -\ell + \delta \pi_{t-1} \lambda e^{-\lambda a_t}\left(g - (1-\alpha_{t+1})\ell + \delta e^{-\lambda \alpha_{t+1}} H_{t+2}\right) \end{aligned} \tag{21}$$

It can be easily argued that $\alpha_t \equiv 1$ for all t cannot be optimal. So, we may assume that $\alpha_{t+1} < 1$, which implies

$$\frac{\partial}{\partial a_{t+1}} V_{t+1}(\alpha_{t+1}, \pi_t) = 0 \text{ or } H_{t+2} = \frac{\ell}{\delta \pi_t \lambda e^{-\lambda \alpha_{t+1}}}$$

Substituting H_{t+2} into (21), replacing π_t with the right-hand side of (18), and rearranging we derive

$$\begin{aligned} & \frac{\partial}{\partial a_t} V_t(a_t, \pi_{t-1}) \\ = \;& -\ell + \delta \pi_{t-1} \lambda e^{-\lambda a_t}\left(g - (1-\alpha_{t+1})\ell + \frac{\ell}{\pi_t \lambda}\right) \\ = \;& \ell\left[e^{-\lambda a_t} \delta \pi_{t-1}\left(1 + \lambda\left(\gamma - 1 + \alpha_{t+1}\right)\right) - \delta \pi_{t-1} - (1-\delta)\right] \end{aligned} \tag{22}$$

If $\alpha_t = 1$, we define $\tau = t$. In this case (22) ≥ 0 at $a_t = 1$, implying

$$e^{-\lambda} \delta \pi_{\tau-1}\left(1 + \lambda \gamma\right) - \delta \pi_{t-1} - (1-\delta) > 0$$

Assumption 2 implies that the term above is nondecreasing in $\pi_{\tau-1}$. Therefore, $\alpha_s = 1$ for all $1 \leq s < \tau$ because

$$\begin{aligned} \frac{\partial}{\partial a_s} V_s(1, \pi_{s-1}) &= e^{-\lambda} \delta \pi_{s-1}\left(1 + \lambda \gamma\right) - \delta \pi_{t-1} - (1-\delta) \\ &\geq e^{-\lambda} \delta \pi_{\tau-1}\left(1 + \lambda \gamma\right) - \delta \pi_{\tau-1} - (1-\delta) > 0 \end{aligned}$$

If $\alpha_t < 1$, then (21) $= 0$ yields

$$\alpha_t = \frac{1}{\lambda} \ln \frac{\delta \pi_{t-1} \left(1 + \lambda \left(\gamma - 1 + \alpha_{t+1}\right)\right)}{\delta \pi_{t-1} + 1 - \delta}, \tag{23}$$

confirming (17). What remains is to show that $(\alpha_t)_{t=1}^{\infty}$ is a strictly decreasing sequence that converges to 0 as $t \to \infty$.

Given $\alpha_t > 0$ for all $t \geq 1$, (18) implies that $(\pi_t)_{t=0}^{\infty}$ is a strictly decreasing sequence bounded from below by $\pi_{\min}$. Thus, by the monotone convergence theorem, π_t tends to a limit $\pi_\infty \geq \pi_{\min}$. By backward induction, let us "guess" that $\alpha_{T+1} \geq \alpha_{T+2}$ for some arbitrarily large T. Then, given that α_t increases in π_{t-1} and α_{t+1}, as can be easily checked from (23), for $t = T - 1$ we have

$$\alpha_t > \frac{1}{\lambda} \ln \frac{\delta \pi_t \left(1 + \lambda \left(\gamma - 1 + \alpha_{t+2}\right)\right)}{\delta \pi_t + 1 - \delta} = \alpha_{t+1}$$

This shows that $(\alpha_t)_{t=1}^{T}$ is a strictly decreasing sequence. Since T is arbitrary, taking limit as $T \to \infty$ verifies that $\alpha_t > \alpha_{t+1}$ for all $t \geq 1$. Finally, since $\alpha_t > 0$ implies $\pi_{t-1} > \pi_t$, $\lim_{t\to\infty} \pi_t = \pi_\infty$ implies $\lim_{t\to\infty} \alpha_t = 0$, which, in turn, implies $\pi_\infty = \pi_{\min}$.

Conclusion (iii) is now a direct consequence of (ii). ■

The "none-stop" result of this proposition might seem surprising. A rough intuition for this result is that when α_t is sufficiently close to 0, it does no affect the change in the posterior very much. In turn, when the posterior does not change much, it does not cause the subsequent allocation to change much. Consequently, as t tends toward infinity, α_t gradually declines toward 0 but there exists no "last period" in which α_t would drop to 0 completely. The proof of this result reveals a more fundamental reason: For stopping to be optimal at certain time $t + 1$, two conditions must be simultaneously met. One of these (see (19)) derives from the first-order conditions for the optimality of the allocation plan, and the other (see (19)) from the Bayes rule. The former requires the relative change in the posterior to be larger than a certain level, while the latter requires the change to be smaller than that level. These conflicting conditions imply the impossibility of finding a finite time horizon to optimally stop the experimentation completely, no matter how long the horizon is.

Proposition 2 provides a useful characterization of the optimal allocation plan, where α_t is determined by the current posterior belief π_{t-1} and, recursively, the next-period allocation plan α_{t+1}. In the next proposition, we solve $(\alpha_t)_{t=1}^{\infty}$ in which α_t is determined solely by the posterior π_{t-1} and other exogenous variables. We then conduct a comparative statics analysis of the properties of α_t.

Define

$$\pi_{\max} = \frac{1-\delta}{\delta\left(e^{-\lambda}\left(\lambda\gamma+1\right)-1\right)},$$

which can be done under Assumption 2 and satisfies $\pi_{\min} < \pi_{\max} < 1$.

Proposition 3 *Suppose Assumptions 1-2 hold. Given* δ, γ, λ, *assume* $\pi_0 \in [\pi_{\min}, \pi_{\max}]$. *Then, for all* $1 < T \leq \infty$, *where* T *is an exogenously scheduled stopping time for the bandit exploration, there exists a unique optimal allocation plan* $(\alpha_t)_{t=1}^{T}$, *in which each* $\alpha_t = \alpha_t(\pi_{t-1}, \delta, \gamma, \lambda) \in (0, 1)$ *is a continuously differentiable function such that*

$$\textit{(i)}\ \frac{\partial \alpha_t}{\partial \pi_{t-1}} > 0,\ \textit{(ii)}\ \frac{\partial \alpha_t}{\partial \delta} > 0,\ \textit{(iii)}\ \frac{\partial \alpha_t}{\partial \gamma} > 0,$$

and (iv) there exists $\hat{\pi}_t \in [\pi_{\min}, \pi_{\max}]$ *such that* $\frac{\partial \alpha_t}{\partial \lambda} > (\leq)\ 0$ *if* $\pi_{t-1} < (\geq)\ \hat{\pi}_t$.

Proof. We first show that $\pi_0 \leq \pi_{\max}$ implies $\alpha_t < 1$ for all $t \geq 1$. Inspecting (22) in the proof of Proposition 2, if $\alpha_t = 1$, then (22) ≥ 0 at $a_t = 1$. Consequently,

$$e^{-\lambda}\delta\pi_{t-1}\left(1+\lambda\gamma\right) + \delta\pi_{t-1} + 1 - \delta > 0$$

which implies $\pi_{t-1} > \pi_{\max}$. But this violates the assumption $\pi_0 \leq \pi_{\max}$. This shows $\alpha_t \in (0, 1)$ for all $t \geq 1$.

Now assume $1 < T < \infty$, so that $H_{T+r} = G - L$ for $r = 1, 2, \ldots$. Let $\boldsymbol{\alpha} = (\alpha_t)_{t=1}^{T}$ denote an allocation plan and $(\pi_t)_{t=0}^{T-1}$ the sequence of the posteriors associated with α with $\pi_0 \in (\pi_{\min}, 1)$. By Lemma 1,

$$\frac{\partial}{\partial a_T} V_T(\alpha_T, \pi_{T-1}) = -\ell + \delta\pi_{T-1}\lambda e^{-\lambda\alpha_T}(G-L) = 0$$

yields a unique solution

$$\alpha_T(\pi_{T-1}, \delta, \gamma, \lambda) = \frac{1}{\lambda} \ln \frac{\delta \pi_{T-1} \lambda (\gamma - 1)}{1 - \delta}$$

The function α_T has continuous, positive derivative on $[\pi_{\min}, 1]$:

$$\begin{aligned}
\frac{\partial}{\partial \pi_{T-1}} \alpha_T &= \frac{1}{\pi_{T-1} \lambda} > 0 \\
\frac{\partial}{\partial \delta} \alpha_T &= \frac{1}{\lambda \delta (1 - \delta)} > 0 \\
\frac{\partial}{\partial \gamma} \alpha_T &= \frac{1}{\lambda (\gamma - 1)} > 0 \\
\frac{\partial}{\partial \lambda} \alpha_T &= \frac{1}{\lambda^2} \left(1 - \ln \left(\pi_{T-1} \lambda \delta \frac{\gamma - 1}{1 - \delta} \right) \right) \\
&= \frac{1}{\lambda^2} \left(1 - \ln \frac{\pi_{T-1}}{\pi_{\min}} \right) \begin{cases} > 0 \text{ if } \pi_{T-1} < \pi_{\min} e \\ \leq 0 \text{ if } \pi_{T-1} \geq \pi_{\min} e \end{cases}
\end{aligned}$$

This shows that α_T satisfies the properties (i)-(iv) as stated.

Now by backward induction, for some $t+1 \leq T$, suppose the function $\alpha_{t+1}(\pi_t, \delta, \gamma, \lambda)$ is well defined and continuously differentiable, satisfying properties (i)-(iv).

Moving backward in time, let $h(\alpha_{t+1}, \pi_{t-1}, \delta, \gamma, \lambda)$ denote the function on the right-hand side of 17 in Proposition 2:

$$h = \frac{1}{\lambda} \ln \frac{\delta \pi_{t-1} (1 + \lambda (\gamma - 1 + \alpha_{t+1}))}{\delta \pi_{t-1} + 1 - \delta}$$

It has continuous partial derivatives

$$\begin{aligned}
\frac{\partial h}{\partial \alpha_{t+1}} &= \frac{1}{1 + \lambda (\gamma - 1 + \alpha_{t+1})} > 0 \\
\frac{\partial h}{\partial \pi_{t-1}} &= \frac{1}{\pi_{t-1} \lambda} \frac{1 - \delta}{\pi_{t-1} \delta - \delta + 1} > 0 \\
\frac{\partial h}{\partial \delta} &= \frac{1}{\lambda \delta (\delta \pi_{t-1} - \delta + 1)} > 0 \\
\frac{\partial h}{\partial \gamma} &= \frac{1}{h \lambda - \lambda + \lambda \gamma + 1} > 0 \\
\frac{\partial h}{\partial \lambda} &= -\frac{1}{\lambda^2} \ln \frac{\delta \pi_{t-1} (1 + \lambda (\gamma - 1 + \alpha_{t+1}))}{\delta \pi_{t-1} + 1 - \delta} + \frac{\gamma - 1 + \alpha_{t+1}}{\lambda (1 + \lambda (\gamma - 1 + \alpha_{t+1}))}
\end{aligned}$$

In (18), let function $g(a_t, \pi_{t-1}, \lambda)$ denote π_t :

$$g = \frac{\pi_{t-1} e^{-\lambda a_t}}{1 - \pi_{t-1}\left(1 - e^{-\lambda a_t}\right)}$$

It is continuously differentiable, with

$$\begin{aligned}
\frac{\partial g}{\partial a_t} &= -\lambda e^{-\lambda a_t} \frac{\pi_{t-1}\left(1 - \pi_{t-1}\right)}{\left(1 - \pi_{t-1}\left(1 - e^{-\lambda a_t}\right)\right)^2} < 0 \\
\frac{\partial g}{\partial \pi_{t-1}} &= \frac{e^{-\lambda a_t}}{\left(1 - \pi_{t-1}\left(1 - e^{-\lambda a_t}\right)\right)^2} > 0 \\
\frac{\partial g}{\partial \lambda} &= -a_t e^{-\lambda a_t} \frac{\pi_{t-1}\left(1 - \pi_{t-1}\right)}{\left(1 - \pi_{t-1}\left(1 - e^{-\lambda a_t}\right)\right)^2} < 0
\end{aligned}$$

Now define function $J_t(a_t, \pi_{t-1}, \delta, \gamma, \lambda)$ by

$$J_t = a_t - h(\alpha_{t+1}\left(g(a_t, \pi_{t-1}, \lambda), \delta, \gamma, \lambda\right), \pi_{t-1}, \delta, \gamma, \lambda)$$

Since h and g are continuously differentiable, by the induction hypothesis J_t is continuously differentiable in all variables. Importantly, the partial derivative of J_t with respect to a_t is everywhere strictly positive:

$$\frac{\partial J_t}{\partial a_t} = 1 - \frac{\partial h}{\partial \alpha_{t+1}} \frac{\partial \alpha_{t+1}}{\partial g} \frac{\partial g}{\partial a_t} > 0$$

By (17) of Proposition 2, $J_t = 0$ has a solution. Therefore, by the implicit function theorem, $J_t = 0$ defines a continuously differentiable function $\alpha_t(\pi_{t-1}, \delta, \gamma, \lambda)$ satisfying

$$\begin{aligned}
\frac{\partial \alpha_t}{\partial \pi_{t-1}} &= \frac{1}{\partial J_t / \partial a_t} \left(\frac{\partial h}{\partial \pi_{t-1}} + \frac{\partial h}{\partial \alpha_{t+1}} \frac{\partial \alpha_{t+1}}{\partial g} \frac{\partial g}{\partial \pi_{t-1}} \right) > 0 \\
\frac{\partial \alpha_t}{\partial \delta} &= \frac{1}{\partial J_t / \partial a_t} \left(\frac{\partial h}{\partial \delta} + \frac{\partial h}{\partial \alpha_{t+1}} \frac{\partial \alpha_{t+1}}{\partial \delta} \right) > 0 \\
\frac{\partial \alpha_t}{\partial \gamma} &= \frac{1}{\partial J_t / \partial a_t} \left(\frac{\partial h}{\partial \gamma} + \frac{\partial h}{\partial \alpha_{t+1}} \frac{\partial \alpha_{t+1}}{\partial \gamma} \right) > 0 \\
\frac{\partial \alpha_t}{\partial \lambda} &= \frac{1}{\partial J_t / \partial a_t} \left(\frac{\partial h}{\partial \lambda} + \frac{\partial h}{\partial \alpha_{t+1}} \left(\frac{\partial \alpha_{t+1}}{\partial \lambda} + \frac{\partial \alpha_{t+1}}{\partial g} \frac{\partial g}{\partial \lambda} \right) \right)
\end{aligned}$$

Although tedious, it can be readily verified that $\frac{\partial \alpha_t}{\partial \lambda}$ has the property of (iv). By induction, and the principle of dynamic programming, we have thus obtained a unique optimal allocation plan $(\alpha_t)_{t=1}^{T}$, with each α_t satisfying properties (i)-(iv). If there is no exogenous stopping constraint, letting $T \to \infty$ and noting from Proposition 2 that $\alpha_t \to 0$ implies $H_{t+1} \to G - L$ as $t \to \infty$, the optimal plan $(\alpha_t)_{t=1}^{\infty}$ is obtained in the limit. ■

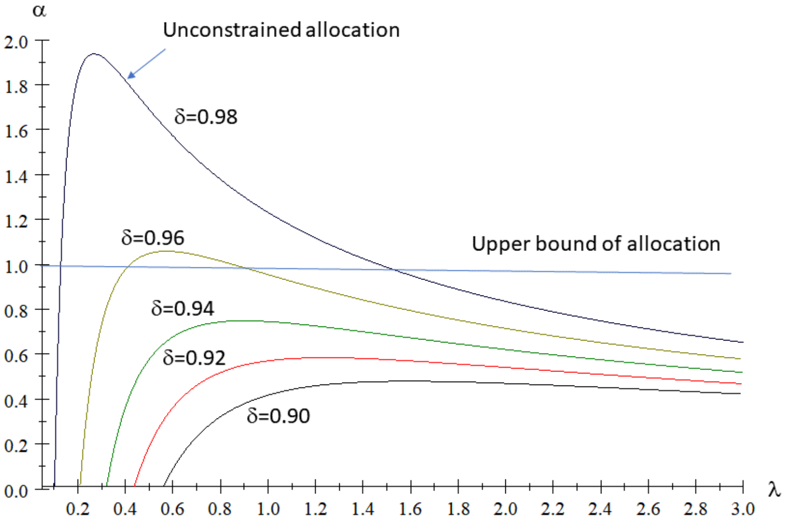

Figure 1: Behavior of α_t as a function of δ and λ, assuming $\pi_{t-1} = 0.1$ and $\gamma = 3$.

The comparative statics results in this proposition are intuitive: increasing the posterior, the discount factor, and the attractiveness of the breakthrough make the exploration of the risky arm more profitable, and therefore warrant more investment of time. Regarding the effect of increasing the hazard rate of the probability p, it is interesting to observe that when the posterior is sufficiently high the effect is negative on the allocation. The intuition is that, with sufficiently high confidence that the breakthrough will occur, the agent does not need to worry too much and a higher λ makes the breakthrough even easier to achieve. Then, the agent may have the incentive to reduce the allocation of time given the opportunity cost of it. On the other hand, when the posterior is sufficiently low, increasing λ encourages the agent to work harder on achieving a breakthrough. Figure 1 shows a graphical example of the behavior of α_t as δ and λ change.

Regarding the optimal risk premium given in Proposition 1, an immediate corollary follows.

Corollary 2 *Under the assumptions of Proposition 3, for all* $t \geq 1$ *the optimal risk premium* $\hat{V}_t$ *(i) increases in* $\pi_{t-1}, \delta, \gamma$ *and (ii) increases in* ℓ *holding* γ *fixed.*

Proof. (i) is a direct consequence of Proposition 3, as $\hat{V}_t$ is an increasing function of α_t, depending on $(\pi_{t-1}, \delta, \gamma)$ solely through α_t. (ii) is a straightforward consequence of $\hat{V}_t$ being an increasing function of ℓ. ■

6 References

Awaya, Y., & Krishna, V. (2021). Startups and upstarts: disadvantageous information in R&D. Journal of Political Economy, 129(2), 534-569. The University of Chicago Press Chicago, IL.

Bergemann, D., & Hege, U. (2005). The financing of innovation: Learning and stopping. RAND Journal of Economics, 719-752. JSTOR.

Bergemann, D., & Hege, U. (1998). Venture capital financing, moral hazard, and learning. Journal of Banking & Finance, 22(6-8), 703-735. Elsevier.

Bergemann, D., & Vimi, J. (2010). The dynamic pivot mechanism. Econometrica, 78(2), 771-789. Wiley Online Library.

Besanko, D., & Wu, J. (2013). The impact of market structure and learning on the tradeoff between R&D competition and cooperation. The Journal of Industrial Economics, 61(1), 166-201. Wiley Online Library.

Besanko, D., Tong, J., & Wu, J. J. (2018). Subsidizing research programs with “if” and “when” uncertainty in the face of severe informational constraints. The RAND Journal of Economics, 49(2), 285-310. Wiley Online Library.

Bolton, P., & Harris, C. (1999). Strategic experimentation. Econometrica, 67(2), 349-374. Wiley Online Library.

Gittins, J. C., & Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. Progress in Statistics, 241-266.

Keller, G., Rady, S., & Cripps, M. (2005). Strategic experimentation with exponential bandits. Econometrica, 73(1), 39-68. Wiley Online Library.

Malueg, D. A., & Tsutsui, S. O. (1997). Dynamic R&D competition with learning. The RAND Journal of Economics, 751-772. JSTOR.

Rothschild, M. (1974). A two-armed bandit theory of market pricing. Journal of Economic Theory, 9(2), 185-202. Elsevier.

Stokey, N. L. (1989). Recursive Methods in Economic Dynamics. Harvard University Press.

Thomas, D. C. (2021). Strategic experimentation with congestion. American Economic Journal: Microeconomics, 13(1), 1-82. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203-2425.