

# Optimal Allocation Strategies in a Discrete-Time Exponential Bandit Problem

Audrey Hu\*

Liang Zou<sup>†</sup>

*City University of Hong Kong*

*University of Amsterdam*

March 14, 2024

---

\**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](mailto:audrey.hu@cityu.edu.hk).

<sup>†</sup>Email: [zou.uva@gmail.com](mailto:zou.uva@gmail.com)

# Optimal Allocation Strategies in a Discrete-Time Exponential Bandit Problem

by Audrey Hu and Liang Zou

**Abstract.** This study addresses a theoretic-bandit problem involving a "safe" and a "risky" arm across countable 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. A methodological contribution of this study lies in a problem transformation that enhances tractability, going beyond the standard Bellman-equation approach for bandit problems. The re-formulation of the problem allows us to obtain clear analytical solutions and comparative statics results. In general, we find that the optimal allocation plan significantly differs from binary strategies, and stopping after any finite periods of unsuccessful trials is suboptimal.

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

# 1 Introduction

This paper examines a discrete-time, two-armed bandit problem, embodying the core economic trade-off between the exploitation of a "safe" arm with a certain return and the exploration of a "risky" arm with an uncertain yield. The risky arm may be categorized as "good", possessing a present value that surpasses the safe arm, or "bad", thereby deeming it devoid of value. In each period, an economic agent, equipped with a single unit of time, must prudently allocate this resource between the two arms. The problem culminates in success if a "breakthrough" transpires—that is, the risky arm is ascertained to be "good". If the arm is "bad", the agent will perpetually fail to achieve a breakthrough, leading to the squandering of all exploration time. This bandit model serves as an apt metaphor for a diverse range of real-world challenges, spanning from R&D, pharmaceutical trials, mineral site exploration, to the pursuit of proof or evidence to substantiate an unproven conjecture.

Despite extensive research, the economic applications of the model have predominantly been constrained to discrete choice sets or convex/linear payoff functions. Consequently, optimal solutions often feature a "bang-bang" strategy: allocating all available time to either the safe or the risky arm, coupled with an optimal stopping time in the absence of a breakthrough. Consistent with the renowned index theorem by Gittins and Jones (1974), these "bang-bang" strategies are relatively straightforward to analyze, eliminating the need to grapple with interior solutions (see, for instance, Bergemann and Valimaki 2010 for a survey). However, in many scenarios, the decision variables of interest are naturally continuous. For example, how much capital should a company allocate to R&D from its retained earnings? How much time should an academic devote to pursuing a highly uncertain but potentially groundbreaking research idea? What should be the trial prices a monopolistic firm might offer for a new product or service to learn the consumer demand function? In these situations, when the objective functions are concave,

we should anticipate the optimal allocation policies to be characterized by interior solutions. Indeed, the scarcity of research on interior solutions to bandit problems is more a reflection of analytical complexity rather than a lack of relevance. For instance, Rothschild (1974), in his seminal work applying bandit theory to stores' learning about consumer demand by changing prices over time, concluded: "I assumed ... that there were only two prices that stores could charge ... 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 key methodological contribution of our study lies in demonstrating 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 conventional 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  $(\alpha_t)_{t=1}^{\infty}$  in  $(0, 1]$  that converges to zero as  $t$  tends to infinity, given no breakthrough has occurred.

The economic literature on bandit problems has largely focused on continuous time in recent years. 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 since been 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).

Bergemann and Hege (1998, 2005) studied the financing of innovation problems in discrete-time BorN bandits. Assuming the probability of breakthrough to be linear in the level of investment in innovation, they derived an optimal bang-bang strategy and focused on the effects of financing decisions on the stopping time. Rosenberg et al. (2007) and Murto and Valimaki (2011) analyzed the two-armed exponential bandit problem in discrete time, assuming the action choice is binary.

Our analysis of the discrete-time counterpart of the exponential bandit prob-

Row Labels	Average of 2020
0	2.4%
Aerospace & Defense	0.5%
Automobiles	1.6%
Biotechnology	30.3%
Capital Markets	0.5%
Chemicals	2.4%
Communications Equipment	16.7%
Electrical Equipment	0.8%
Electronic Equipment, Instruments & Components	8.8%
Energy Equipment & Services	0.6%
Entertainment	9.9%
Equity Real Estate Investment Trusts (REITs)	0.0%
Food Products	0.1%
Health Care Equipment & Supplies	7.4%
Health Care Providers & Services	0.0%
Health Care Technology	13.6%
Hotels, Restaurants & Leisure	0.0%
Household Durables	1.9%
Household Products	0.4%
Independent Power and Renewable Electricity Producers	0.0%
Industrial Conglomerates	2.3%
Interactive Media & Services	18.8%
Internet & Direct Marketing Retail	11.2%
IT Services	1.5%
Leisure Products	4.7%
Life Sciences Tools & Services	7.4%
Machinery	1.8%
Metals & Mining	0.4%
Pharmaceuticals	15.2%
Semiconductors & Semiconductor Equipment	16.7%
Software	19.0%
Technology Hardware, Storage & Peripherals	8.5%

Figure 1: R&D spending in 2020 as a percentage of revenue by industry [S&P500].

Source: <https://einvestingforbeginners.com/rd-spending-as-a-percentage-of-revenue-by-industry/>

lem 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  $\alpha_t$  is sufficiently close to 0, it does not 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,  $\alpha_t$  gradually declines toward 0 but there exists no "last period" in which  $\alpha_t$  would drop to 0 completely. Obviously, this cannot happen with the bang-bang strategies.

In the framework of binary decisions, since allocations are a constant number (0 or 1) in each per period, comparative statics analyses typically center around effects on the optimal stopping time, or on the *total* amount of resources invested in

exploring the risky arm. One appeal of our study is that it offers comparative statics predictions directly related to the *level* of resources allocated to the exploration of the unknown, thereby enabling us to adequately address empirical questions such as "What determines the cross-industry differences in R&D spending as a percentage of revenue?" For instance (see Figure 1), among the S&P 500 companies in 2020, the biotechnology industry tops the list of having invested an average of \$30.3% of their revenue on R&D, while some other industries virtually invested nothing (Sather, 2021). In short, the analysis and results presented in this study complement and significantly enrich the state-of-the-art knowledge about the BorN exponential bandit problems.

## 2 The Model

Time is discrete, with countable periods  $t = 1, 2, \dots$ . 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.<sup>1</sup> If the risky arm is good, the probability

---

<sup>1</sup>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.

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  $a_1, \dots, a_t$  in periods 1 through  $t$  without a breakthrough, the updated probability (or posterior belief) that the risky arm is good equals

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

### 3 The Bandit Problem

Let  $\pi_0 \in (\pi_{\min}, 1)$  be given, and let  $\boldsymbol{\alpha} = (\alpha_t)_{t=1}^{\infty}$  denote a feasible allocation plan such that each  $\alpha_t \in [0, 1]$  is measurable with respect to the information available at the start of period  $t$ . Let  $t^*$ , if any, denote the period in which a breakthrough occurs such that  $\alpha_t \equiv 1$  for all  $t > t^*$ . Let  $t_s$ , if any,

denote the last period of the experimentation in the absence of a breakthrough such that  $\alpha_t \equiv 0$  for all  $t > t_s$ . Then, the bandit problem boils down to maximizing the agent's expected, discounted payoff  $E(\bar{V}(\boldsymbol{\alpha}))$  in which

$$\bar{V}(\boldsymbol{\alpha}) = \begin{cases} \sum_{t=1}^{t^*} \delta^{t-1} (1 - \alpha_t) \ell + \delta^{t^*} G & \text{if } t^* \leq t_s \\ \sum_{t=1}^{t_s} \delta^{t-1} (1 - \alpha_t) \ell + \delta^{t_s} L & \text{if } t_s < t^* \end{cases}$$

and the expectation  $E(\cdot)$  is over the type of the risky bandit, and the stochastic processes  $(\alpha_t)_{t=1}^{\infty}$  and  $(\pi_t)_{t=0}^{\infty}$ .

Now let  $V_t^*$  denote the agent's optimal, conditional expected payoff in period  $t$ —given no breakthrough until period  $t - 1$  and the updated belief  $\pi_{t-1}$ . By the principle of optimality for dynamic programming, if  $\boldsymbol{\alpha}$  is optimal it must satisfy the



Bellman equation<sup>2</sup>

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

$$\text{s.t. } \pi_t = \frac{\pi_{t-1} e^{-\lambda a_t}}{1 - \pi_{t-1} (1 - e^{-\lambda a_t})} \quad (3)$$

where

$$\begin{aligned} & V_t(a_t, \pi_{t-1}) \\ = & (1 - a_t)\ell + \delta \{ \pi_{t-1} p(a_t) G + (1 - \pi_{t-1} p(a_t)) V_{t+1}^*(\pi_t) \} \end{aligned} \quad (4)$$

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  $\pi_{t-1}$ . 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 expected payoff becomes  $V_{t+1}^*(\pi_t)$  given updated posterior  $\pi_t$ . If the agent stops the experiment after period  $t_s$  without any breakthrough, all future resources will be allocated to the safe arm from  $t_s + 1$  onwards, implying  $V_t = L$  for all  $t > t_s$ .

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

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

We will establish the existence and uniqueness of  $\boldsymbol{\alpha}$  in Proposition 3. For now, notice that the simple structure of  $V_t$  implies that it is differentiable in  $(a_t, \pi_{t-1})$  on

---

<sup>2</sup>Our analysis and results may treat  $t$  as a variable, hence it is useful to keep the subscript for  $V_t$  and  $a_t$ .

$[0, 1]^2$ .<sup>3</sup> 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:

$$\frac{\partial}{\partial a_t} V_t(a_t, \pi_{t-1}) = -\ell + \delta \pi_{t-1} \lambda e^{-\lambda a_t} [G - V_{t+1}^*(\pi_t)] \quad (5)$$

$$\begin{aligned} & + \delta (1 - \pi_{t-1} (1 - e^{-\lambda a_t})) V_{t+1}^{*'}(\pi_t) \frac{\partial \pi_t}{\partial a_t} \\ & = 0 \end{aligned} \quad (6)$$

where, by the envelope theorem,

$$V_{t+1}^*(\pi_t) = V_{t+1}(\alpha_{t+1}(\pi_t), \pi_t) \implies V_{t+1}^{*'}(\pi_t) = \frac{\partial V_{t+1}(\alpha_{t+1}(\pi_t), \pi_t)}{\partial \pi_t} \quad (7)$$

The term in (6) 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  $V_{t+1}^{*'}(\pi_t) > 0$  (see Lemma 2) and since

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

the sign of this learning effect is negative. Therefore, compared to the myopic solutions that neglect the term in (6), all interior solutions to (2) imply that 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 experimentation. 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 (5)-(6) to get

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

---

<sup>3</sup>At the boundaries of  $[0, 1]^2$ , the (cross-partial) derivatives of  $V$  are defined as usual by taking limits.

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  $\pi_0$ , the composition of  $\pi_{\min}$  suggests that the risky arm is worthy of exploration as long as  $\delta$ ,  $\lambda$ , or  $\gamma$  is sufficiently high.

The next assumption ensures congruency of the analysis for our study.

**Assumption 2**  $\delta e^{-\lambda}(\lambda\gamma + 1) - 1 \geq 0$ .

The role of this assumption can be illustrated by considering an extreme case with  $\pi_0 = 1$ . Since  $\pi_0$  is the agent's *subjective* probability, to realize the present value of the risky arm,  $G$ , he needs 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 agent will spend full time exploring the risky arm until achieving a breakthrough, or else never stop. To see this, we replace  $\pi_{t-1}$  and  $a_t$  with 1 in (4), and obtain the optimal payoff from the Bellman equation:

$$\begin{aligned} V^*(1) &= \delta(1 - e^{-\lambda})G + \delta e^{-\lambda}V^*(1) \\ &= \frac{\delta(1 - e^{-\lambda})G}{1 - \delta e^{-\lambda}} \end{aligned} \tag{8}$$

Since there is no learning in this case, the experimentation plan with  $\alpha_t \equiv 1$  must satisfy, substituting (8),

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

Assumption 2 then implies that the above inequality holds. On the other hand, it is worth observing that when Assumption 2 fails to hold, then, even if the agent believes that the risky arm is good for sure (i.e.,  $\pi_0 = 1$ ), he may not be interested in exploring the arm if  $\lambda$  is sufficiently low (i.e., too hard to achieve a breakthrough) or if  $\delta$  is sufficiently low (i.e., too costly to wait for a breakthrough).

Assumption 2 allows us to define

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

which can be easily checked to satisfy  $\pi_{\min} < \pi_{\max} < 1$ . As will be shown later,  $\pi_{t-1} \leq \pi_{\max}$  implies  $\alpha_t < 1$ , ensuring that the first-order conditions can be applied legitimately.

## 4 Simplifying the Problem

Although the Bellman equation approach in the previous section is quite standard, it is not easily tractable. The main reason is that all future posteriors depend on  $\pi_t$ , and therefore a marginal change of  $a_t$  may have effects that propagate to all future continuation payoffs.

We present here a re-formulation of the Bellman equation that enhances the tractability of the bandit problem. Fix  $\alpha$  (except  $a_t$ ), and consider a mathematically equivalent problem to (2)-(4) of maximizing

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

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

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

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  $\alpha_t$  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	$-[g - (1 - a_t)\ell]$
$t + 1$	$1 - q_t$	$-\delta [g - (1 - \alpha_{t+1})\ell]$
$\dots$	$\dots$	$\dots$
$t + s$	$(1 - q_t) \prod_{r=1}^{s-1} (1 - q_{t+r}^*)$	$-\delta^s [g - (1 - \alpha_{t+s})\ell]$
$\dots$	$\dots$	$\dots$

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

$$V_t(a_t, \pi_{t-1}) = G - [g - (1 - a_t)\ell] - E_t \left\{ \sum_{s=1}^{\infty} \delta^s \left( (1 - q_t) \prod_{r=1}^{s-1} (1 - q_{t+r}^*) \right) [g - (1 - \alpha_{t+s})\ell] \right\} \quad (9)$$

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  $\alpha_{t+1}, \dots, \alpha_{t+s-1}$ . 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)-(4),*

$$V_t(a_t, \pi_{t-1}) = V_t(0, \pi_{t-1}) - a_t \ell + \delta \pi_{t-1} p(a_t) H_{t+1} \quad (10)$$

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

$$H_{t+1} = E_t \{ g - (1 - \alpha_{t+1})\ell + \delta(1 - p(\alpha_{t+1}))H_{t+2} \} \quad (11)$$

for  $t = 1, 2, \dots$

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

$$\begin{aligned} & (1 - q_t) \prod_{r=1}^{s-1} (1 - q_{t+r}^*) \\ &= 1 - \pi_{t-1} + \pi_{t-1} (1 - p(a_t)) \prod_{r=1}^{s-1} (1 - p(\alpha_{t+r})) \end{aligned} \quad (12)$$

Substitute (12) into (9), and define

$$H_{t+1} = E_t \left\{ \sum_{s=1}^{\infty} \delta^{s-1} \left( \prod_{r=1}^{s-1} (1 - p(\alpha_{t+r})) \right) [g - (1 - \alpha_{t+s})\ell] \right\}. \quad (13)$$

We obtain by re-arranging the terms that

$$V_t(a_t, \pi_{t-1}) = G - [g - (1 - a_t)\ell] - (1 - \pi_{t-1}) E_t \left\{ \sum_{s=1}^{\infty} \delta^s [g - (1 - \alpha_{t+s})\ell] \right\} - \pi_{t-1} (1 - p(a_t)) \delta H_{t+1} \quad (14)$$

In particular, choosing  $a_t = 0$  gives

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

Subtracting this from (14) yields (10).

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  $T = \infty$ ,  $\lim_{t \rightarrow \infty} H_{t+1} = G - L$  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  $\alpha_t$ , allocating full time to the safe arm in period  $t$ , and then switch back to 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)-(4) is given by the solution to the program

$$\max_{a_t \in [0,1]} \hat{V}_t(a_t, \pi_{t-1}) \quad (= -a_t \ell + \delta \pi_{t-1} p(a_t) H_{t+1}) \quad (15)$$

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.

**Lemma 2** *For all  $t = 1, 2, \dots$ ,  $V_t^{*'}(\pi_{t-1}) \geq 0$ . The inequality holds strictly unless  $\alpha_{t+r} \equiv 0$  for all  $r = 0, 1, 2, \dots$*

**Proof.** From (13), we have

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

Therefore, replacing  $a_t$  with  $\alpha_t$  in (14), and differentiating gives

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

Both the above inequalities hold as an equality only if  $\alpha_{t+r} \equiv 0$  for all  $r = 0, 1, 2, \dots$ . Thus, by the envelope theorem,  $V_t^{*'}(\pi_{t-1}) > 0$  as long as the experimentation of the risky arm continues. ■

This lemma confirms that the learning effect on the continuation payoff in (6), given no breakthrough in the current period, is strictly negative.

## 4.1 A three-period example

Suppose  $t = 1, 2, 3$  and in period 3 the agent must stop the experimentation. Suppose after stopping, the time will be fully allocated to the safe arm, i.e.,  $\alpha_s = 0$  for all  $s \geq 3$ . This implies  $H_3 = G - L$ . Denote  $X = [0, 1] \times [\pi_{\min}, \pi_{\max}]$ .

For  $t = 2$ , let  $\pi_1 \in [\pi_{\min}, \pi_{\max}]$  be given. Differentiating (10) w.r.t.  $a_2$  yields

$$\begin{aligned}\frac{\partial}{\partial a_2} V_2(a_2, \pi_1)|_{a_2=\alpha_2} &= -\ell + \delta\pi_1\lambda e^{-\lambda\alpha_2} (G - L) \\ &= \ell \left( -1 + \delta\pi_1\lambda e^{-\lambda\alpha_2} \frac{\gamma - 1}{1 - \delta} \right) = 0\end{aligned}\quad (16)$$

Moreover,  $V_2$  is twice continuously differentiable satisfying, for all  $(a_2, \pi_1) \in X$ ,

$$\begin{aligned}\frac{\partial^2}{\partial a_2^2} V_2(a_2, \pi_1) &= -\ell\delta\pi_1\lambda^2 e^{-\lambda\alpha_2} \frac{\gamma - 1}{1 - \delta} < 0 \\ \frac{\partial^2}{\partial a_2 \partial \pi_1} V_2(a_2, \pi_1) &= \ell\delta\lambda e^{-\lambda\alpha_2} \frac{\gamma - 1}{1 - \delta} > 0\end{aligned}$$

These imply that condition (16) is both necessary and sufficient for the solution  $\alpha_2$  to be optimal, and that the solution  $\alpha_2(\pi_1)$  is continuously differentiable with  $\alpha_2' > 0$  on  $[\pi_{\min}, \pi_{\max}]$ . Solving (16) gives the period 2 optimal strategy

$$\begin{aligned}\alpha_2(\pi_1) &= \frac{1}{\lambda} \ln \frac{\pi_1 \delta \lambda (\gamma - 1)}{1 - \delta} = \frac{1}{\lambda} \ln \frac{\pi_1}{\pi_{\min}} \\ &= \frac{1}{\lambda} \left( \ln \frac{\pi_0 e^{-\lambda a_1}}{1 - \pi_0 (1 - e^{-\lambda a_1})} - \ln \pi_{\min} \right) \\ &=: h(\alpha_1, \pi_0)\end{aligned}\quad (17)$$

Thus,  $\alpha_2(\pi_1)$  can be also seen as a function  $h(a_1, \pi_0)$  satisfying

$$\frac{\partial h}{\partial a_1} = -\frac{1 - \pi_0}{1 - \pi_0 (1 - e^{-\lambda a_1})} < 0 \quad (18)$$

$$\frac{\partial h}{\partial \pi_0} = \frac{1}{\pi_0 \lambda (1 - \pi_0 (1 - e^{-\lambda a_1}))} > 0 \quad (19)$$

For  $t = 1$ , we have

$$\begin{aligned}H_2(\alpha_1, \pi_0) &= g - (1 - \alpha_2)\ell + \delta e^{-\lambda\alpha_2} (G - L) \\ &= \ell \left( \gamma - (1 - h(\alpha_1, \pi_0)) + \frac{1 - \pi_0 (1 - e^{-\lambda a_1})}{\lambda \pi_0 e^{-\lambda a_1}} \right) \text{ by (17)} \\ \delta e^{-\lambda\alpha_2} (G - L) &= (1 - \delta) (G - (1 - \alpha_2)L) + \delta e^{-\lambda\alpha_2} (G - L) \\ &= (1 - \delta) (G - L) + (1 - \delta) \alpha_2 L + \delta e^{-\lambda\alpha_2} (G - L)\end{aligned}$$



which is continuously differentiable. Further, condition (16) implies

$$\begin{aligned}
\frac{\partial H_2}{\partial \alpha_2} &= \ell - \delta \lambda e^{-\lambda \alpha_2} (G - L) \\
&= \ell - \delta \lambda e^{-\lambda \alpha_2} (G - L) (\pi_1 + 1 - \pi_1) \\
&= -\delta \lambda e^{-\lambda \alpha_2} (G - L) (1 - \pi_1) \\
&\rightarrow \left| \frac{\partial H_2}{\partial \alpha_2} \right| < H_2
\end{aligned} \tag{20}$$

The period 1 optimal strategy  $\alpha_1$  is now the solution of

$$\frac{\partial}{\partial a_1} V_1(a_1, \pi_0)|_{a_1=\alpha_1} = -\ell + \delta \pi_0 \lambda e^{-\lambda \alpha_1} H_2(h(\alpha_1, \pi_0)) = 0 \tag{21}$$

We have

$$\begin{aligned}
\frac{\partial^2}{\partial a_1^2} V_1(a_1, \pi_0) &= -\delta \pi_0 \lambda^2 e^{-\lambda a_1} H_2 + \delta \pi_0 \lambda e^{-\lambda a_1} \frac{\partial H_2}{\partial \alpha_2} \frac{\partial h}{\partial a_1} \\
&= \delta \pi_0 \lambda e^{-\lambda a_1} H_2 \left( -\lambda + \frac{1}{H_2} \frac{\partial H_2}{\partial \alpha_2} \frac{\partial h}{\partial a_1} \right) \\
&< \delta \pi_0 \lambda^2 e^{-\lambda a_1} H_2 \left( -1 + \frac{1 - \pi_0}{1 - \pi_0 (1 - e^{-\lambda a_1})} \right) \text{ by (20)} \\
&< 0
\end{aligned} \tag{22}$$

Likewise, we have

$$\begin{aligned}
\frac{\partial^2}{\partial a_1 \partial \pi_0} V_1(a_1, \pi_0) &= \delta \lambda e^{-\lambda a_1} H_2 + \delta \pi_0 \lambda e^{-\lambda a_1} \frac{\partial H_2}{\partial \alpha_2} \frac{\partial h}{\partial \pi_0} \\
&= \delta \lambda e^{-\lambda a_1} H_2 \left( 1 + \frac{\pi_0}{H_2} \frac{\partial H_2}{\partial \alpha_2} \frac{\partial h}{\partial \pi_0} \right) \\
&> \delta \lambda e^{-\lambda a_1} H_2 \left( 1 - \frac{1 - \pi_1}{(1 - \pi_0 (1 - e^{-\lambda a_1}))} \right) \text{ by (20)} \\
&> \delta \lambda e^{-\lambda a_1} H_2 \left( 1 - \frac{1 - \pi_0}{(1 - \pi_0 (1 - e^{-\lambda a_1}))} \right) > 0
\end{aligned} \tag{23}$$

So the condition in (20) is also sufficient. We thus have  $\alpha_1$  implicitly defined by

$$L(\alpha_1, \pi_0) := \delta \pi_0 \lambda e^{-\lambda \alpha_1} H_2(\alpha_1, \pi_0) - \ell = 0$$

The function  $L(\alpha_1, \pi_0)$  is continuously differentiable on  $[0, 1] \times [\pi_{\min}, \pi_{\max}]$ . The inequalities (22)-(23) ensure that, by the implicit function theorem,  $L(\alpha_1, \pi_0) = 0$  gives the solution (implicitly)  $\alpha_1(\pi_0)$  on  $[\pi_{\min}, \pi_{\max}]$  such that  $\alpha_1' > 0$ .

Figure 2 shows a graphical example of the behavior of  $\alpha_1$  as  $\delta$  and  $\lambda$  change.

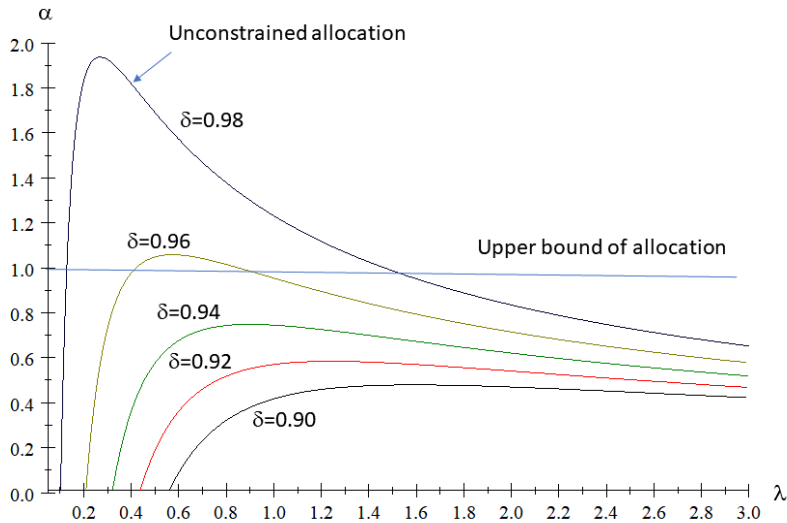


Figure 2: Behavior of  $\alpha_1$  as  $\delta$  and  $\lambda$  change, assuming  $\pi_0 = 0.1$  and  $\gamma = 3$ .

## 4.2 Optimal risk premium

The maximization program in (15) has also a very transparent economic interpretation: If  $\boldsymbol{\alpha} = (\alpha_t)_{t=1}^{\infty}$  is an optimal allocation plan, then each  $\alpha_t$  maximizes the risk premium in period  $t$  given the subsequent planned actions. This objective, although equivalent to the one in (2)-(4), 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*

$$\hat{V}_t^* = \frac{e^{\alpha_t \lambda} - 1}{\lambda} (\ell + \eta_t) - \alpha_t \ell \quad (24)$$

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

**Proof.** Since  $V_t$  in (10) is concave in  $a_t$ , so is  $\hat{V}_t$ . Thus,  $\alpha_t$  is characterized by the first-order condition

$$\frac{\partial}{\partial a_t} \hat{V}_t(a_t, \pi_{t-1})|_{a_t=\alpha_t} = -\ell + \delta\pi_{t-1}\lambda e^{-\alpha_t\lambda} H_{t+1} = \eta_t \quad (25)$$

where  $\eta_t$  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 (10), 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 (24) has a natural interpretation as follows. The term  $\alpha_t\ell$  is the opportunity cost of investment of time  $\alpha_t$  in the risky arm. The term  $\ell + \eta_t$  derives from (25), which equals the *marginal* benefit of investing  $\alpha_t$  in the risky arm:  $\delta\pi_{t-1}\lambda e^{-\alpha_t\lambda} H_{t+1}$ . The overall benefit of investing  $\alpha_t$  in the risky arm is then given by the first term in (24), 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  $\alpha_t$ , depending on the posterior  $\pi_{t-1}$  only indirectly through  $\alpha_t$ .

Compared with (6), we can see from (25) 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  $\{\pi_t, \pi_{t+1}, \dots\}$ , as does  $V_{t+1}^*$ .

Of course, we still have the burden to solve for the optimal allocation plan  $\alpha$  for the more general case.

## 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  $\alpha = (\alpha_t)_{t=1}^{\infty}$  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  $t \rightarrow \infty$ , in which each  $\alpha_t$  is determined by*

$$\alpha_t = \frac{1}{\lambda} \ln \frac{\delta \pi_{t-1} (1 + \lambda(\gamma - 1 + \alpha_{t+1}))}{\delta \pi_{t-1} + 1 - \delta} \in (0, 1) \quad (26)$$

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

$$\pi_t = \frac{\pi_{t-1} e^{-\lambda \alpha_t}}{1 - \pi_{t-1} (1 - e^{-\lambda \alpha_t})}, \quad (27)$$

converges to  $\pi_{\min}$  as  $t \rightarrow \infty$ . (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  $s = 1, 2, \dots$ . It follows then from (10) of Lemma 1 that  $\alpha_t$ ,  $\pi_{t-1}$  and  $\pi_t$  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

$$\frac{\pi_{t-1}}{\pi_t} \geq e^{\lambda \alpha_t}. \quad (28)$$

By the Bayes rule, (27) holds and thus  $\alpha_t > 0$  implies

$$\frac{\pi_{t-1}}{\pi_t} = e^{\lambda\alpha_t} (1 - \pi_{t-1}) + \pi_{t-1} < e^{\lambda\alpha_t} \quad (29)$$

The contradiction between (28) and (29) proves  $\alpha_t > 0$  for all  $t \geq 1$ .

(ii) By Lemma 1, differentiating  $V_t$  in (10) and applying (11) give us

$$\begin{aligned} \frac{\partial}{\partial a_t} V_t(a_t, \pi_{t-1}) &= -\ell + \delta\pi_{t-1}\lambda e^{-\lambda a_t} H_{t+1} \\ &= -\ell + \delta\pi_{t-1}\lambda e^{-\lambda a_t} (g - (1 - \alpha_{t+1})\ell + \delta e^{-\lambda\alpha_{t+1}} H_{t+2}) \end{aligned} \quad (30)$$

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 (30), replacing  $\pi_t$  with the right-hand side of (27), and rearranging we derive

$$\begin{aligned} &\frac{\partial}{\partial a_t} V_t(a_t, \pi_{t-1}) \\ &= -\ell + \delta\pi_{t-1}\lambda e^{-\lambda a_t} \left( g - (1 - \alpha_{t+1})\ell + \frac{\ell}{\pi_t\lambda} \right) \\ &= \ell \left[ e^{-\lambda a_t} \delta\pi_{t-1} (1 + \lambda(\gamma - 1 + \alpha_{t+1})) - \delta\pi_{t-1} - (1 - \delta) \right] \end{aligned} \quad (31)$$

If  $\alpha_t = 1$ , we define  $\tau = t$ . In this case (31)  $\geq 0$  at  $a_t = 1$ , implying

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

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

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

If  $\alpha_t < 1$ , then (30) = 0 yields

$$\alpha_t = \frac{1}{\lambda} \ln \frac{\delta\pi_{t-1} (1 + \lambda(\gamma - 1 + \alpha_{t+1}))}{\delta\pi_{t-1} + 1 - \delta}, \quad (32)$$

confirming (26). What remains is to show that  $(\alpha_t)_{t=1}^\infty$  is a strictly decreasing sequence that converges to 0 as  $t \rightarrow \infty$ .

Given  $\alpha_t > 0$  for all  $t \geq 1$ , (27) implies that  $(\pi_t)_{t=0}^\infty$  is a strictly decreasing sequence bounded from below by  $\pi_{\min}$ . Thus, by the monotone convergence theorem,  $\pi_t$  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  $\alpha_t$  increases in  $\pi_{t-1}$  and  $\alpha_{t+1}$ , as can be easily checked from (32), for  $t = T - 1$  we have

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

This shows that  $(\alpha_t)_{t=1}^T$  is a strictly decreasing sequence. Since  $T$  is arbitrary, taking limit as  $T \rightarrow \infty$  verifies that  $\alpha_t > \alpha_{t+1}$  for all  $t \geq 1$ . Finally, since  $\alpha_t > 0$  implies  $\pi_{t-1} > \pi_t$ ,  $\lim_{t \rightarrow \infty} \pi_t = \pi_\infty$  implies  $\lim_{t \rightarrow \infty} \alpha_t = 0$ , 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  $\alpha_t$  is sufficiently close to 0, it does not 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,  $\alpha_t$  gradually declines toward 0 but there exists no “last period” in which  $\alpha_t$  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 (28)) derives from the first-order conditions for the optimality of the allocation plan, and the other (see (28)) 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  $\alpha_t$  is determined by the current posterior belief  $\pi_{t-1}$  and, recursively, the next-period allocation plan  $\alpha_{t+1}$ . In the next proposition, we solve  $(\alpha_t)_{t=1}^\infty$  in which  $\alpha_t$  is determined solely by the posterior  $\pi_{t-1}$  and other exogenous variables. We then conduct a comparative statics analysis of the properties of  $\alpha_t$ .

**Proposition 3** *Suppose Assumptions 1-2 hold. Given  $\delta, \gamma, \lambda$ , 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*

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

and (iv) the sign of  $\frac{\partial \alpha_t}{\partial \lambda}$  can be positive or negative, depending on the values  $(\pi_{t-1}, \delta, \gamma, \lambda)$ .

**Proof.** We first show that  $\pi_0 \leq \pi_{\max}$  implies  $\alpha_t < 1$  for all  $t \geq 1$ . Inspecting (31) in the proof of Proposition 2, if  $\alpha_t = 1$ , then (31)  $\geq 0$  at  $a_t = 1$ . Consequently,

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

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  $r = 1, 2, \dots$ . 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  $\boldsymbol{\alpha}$  with  $\pi_0 \in (\pi_{\min}, 1)$ . By Lemma 1,

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

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  $\alpha_T$  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  $\alpha_T$  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 26 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 (27), let  $\pi_t = \pi_t(a_t, \pi_{t-1}, \lambda)$ . It is continuously differentiable, with

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

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

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

Since  $h$  and  $\pi_t$  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:

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

By (26) 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

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

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  $\alpha_t$  satisfying properties (i)-(iv). If there is no exogenous stopping constraint, letting  $T \rightarrow \infty$  and noting from Proposition 2 that  $\alpha_t \rightarrow 0$  implies  $H_{t+1} \rightarrow G - L$  as  $t \rightarrow \infty$ , the optimal plan  $(\alpha_t)_{t=1}^\infty$  is obtained in the limit. ■

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  $\lambda$  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  $\lambda$  encourages the agent to work harder on achieving a breakthrough.

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

**Corollary 1** *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  $\ell$  holding  $\gamma$  fixed.*

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

## 6 Concluding Remarks

This study offers an intricate analysis of the bandit problem, where optimal allocation strategies are responsive to evolving beliefs. Our findings make a theoretical contribution to the understanding of discrete-time bandits, demonstrating that the problem's complexity can be reduced via a mathematically equivalent transformation of objectives. This re-formulation paves the way for transparent analytical solutions and comparative-statics results. In the case of exponential bandits, we

show that when decision variables are allowed to be continuous, the optimal allocation strategy diverges from binary strategies, making the cessation following finite periods of unsuccessful trials suboptimal. The results presented in this study do more than just complement the existing literature; they significantly enrich it.

Although our analysis and results are specific to the Breakthrough-or-Nothing (BorN) setting, we argue that under certain sufficient conditions, they can be extended to more general probability functions. The analytical framework developed in this study is flexible and can be adapted to various settings involving multiple agents and strategic interactions. We anticipate that these applications will lead to novel insights. An intriguing question for future research is the extent to which the transformation or re-formulation of the BorN bandits can simplify the analysis of other problem types.

## References

- Awaya, Y. and Krishna, V. (2021). Startups and upstarts: disadvantageous information in r&d. *Journal of Political Economy*, 129(2):534–569.
- Bergemann, D. and Hege, U. (1998). Venture capital financing, moral hazard, and learning. *Journal of Banking & Finance*, 22(6-8):703–735.
- Bergemann, D. and Hege, U. (2005). The financing of innovation: Learning and stopping. *RAND Journal of Economics*, pages 719–752.
- Bergemann, D. and Välimäki, J. (2010). The dynamic pivot mechanism. *Econometrica*, 78(2):771–789.
- Besanko, D., Tong, J., and 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.

- Besanko, D. and 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.
- Bolton, P. and Harris, C. (1999). Strategic experimentation. *Econometrica*, 67(2):349–374.
- D. Thomas, C. (2021). Strategic experimentation with congestion. *American Economic Journal: Microeconomics*, 13(1):1–82.
- Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. *Progress in Statistics*, pages 241–266.
- Keller, G., Rady, S., and Cripps, M. (2005). Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68.
- Malueg, D. A. and Tsutsui, S. O. (1997). Dynamic r&d competition with learning. *The RAND Journal of Economics*, pages 751–772.
- Murto, P. and Välimäki, J. (2011). Learning and information aggregation in an exit game. *The Review of Economic Studies*, 78(4):1426–1461.
- Rosenberg, D., Solan, E., and Vieille, N. (2007). Social learning in one-arm bandit problems. *Econometrica*, 75(6):1591–1611.
- Rothschild, M. (1974). A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202.
- Sather, A. (2021). R&d spending as a percentage of revenue by industry [s&p500]. <https://investingforbeginners.com/rd-spending-as-a-percentage-of-revenue-by-industry/>.
- Stokey, N. L., Lucas, R. E., and Prescott, E. C. (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press.