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

Audrey Hu† Liang Zou‡

City University of Hong Kong *University of Amsterdam*

September 17, 2024

*The coauthors would like to express their gratitude to Yeon-Koo Che, Claudio Mezzetti, Carlos Oyarzun, and Zaifu Yang for their invaluable comments, as well as to the seminar participants at the National University of Singapore, Nanyang Technological University, the University of Queensland, the University of Technology Sydney, and the University of York. We are also thankful to the participants of ESEM 2024, EARIE 2024, and the HK Theorists Workshop 2023 for their valuable feedback. Special thanks to Yating Yuan for outstanding research assistance. All remaining errors are our own.

†*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.

‡Email: 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. Departing from the "either-or" binary choices in previous studies, we explore smooth allocation strategies using the first-order approach. Modelling both the action and the posterior as state variables, we obtain clear characterizations of the optimal allocation strategies and comparative statics. The optimal plan significantly enhances the binary strategies, yielding a higher probability of breakthrough and a higher expected payoff. The Goldilocks principle emerges in that the incentives for exploring the risky arm peak at a level that is neither too difficult nor too easy.

Keywords: two-armed bandit; first-order approach; discrete time; exponential distribution, Goldilocks principle.

1 Introduction

This paper examines a discrete-time, two-armed bandit problem, focusing on the fundamental trade-off between exploiting a "safe" arm with a known return and exploring a "risky" arm with uncertain outcomes. The risky arm can either be "good," offering a higher return than the safe arm, or "bad," providing no value. In each period, an agent must allocate a single unit of time between these two options. Success occurs when a "breakthrough" confirms that the risky arm is good. However, if the arm is bad, no breakthrough is possible, leading to wasted exploration time. This bandit model serves as an insightful metaphor for real-world challenges, such as research and development (R&D), pharmaceutical trials, mineral exploration, and seeking proof for major conjectures.

Despite extensive research, most economic applications of this model have been limited to discrete choice sets or convex/linear payoff functions. As a result, optimal solutions often rely on a binary strategy: allocating all available time to either the safe or risky arm and choosing an optimal stopping point if no breakthrough occurs. This approach, consistent with the Gittins and Jones (1974) index theorem, is relatively straightforward to analyze as it avoids the complexity of interior solutions (see Bergemann and Välimäki (2010) for a survey). However, many real-world scenarios involve continuous decision variables. For example, how much capital should a company allocate to R&D annually? How much time should a scholar devote to uncertain but potentially groundbreaking research? How should a monopolist set trial prices to gauge consumer demand for a new product? In such cases, where the objective functions are concave, the optimal strategies are likely to involve interior solutions. The limited research on these solutions is due more to their analytical complexity than to their lack of relevance (see Rothschild (1974)).

The conventional approach to bandit problems typically treats posterior beliefs as state variables and assumes that allocations are time-invariant functions of those beliefs. However, when allocation strategies depend on first-order conditions that

influence subsequent beliefs, the problem becomes analytically complex. Our study introduces a novel approach by first reconfiguring the state space to include both posterior beliefs and allocations as state variables, and then using a "backward recursion" method to obtain a full characterization of the optimal allocation plan. This approach, without loss of generality, allows us to solve a class of two-armed bandit problems—those involving a "breakthrough or nothing" (BorN)—with clear and explicit analytical solutions.

We find that when allocations are allowed to take any value in the interval [0,1], no indexing policy with a stopping time is optimal. The optimal strategy for the exponential BorN bandit involves continuous, non-binary allocations in the range (0,1), with the exploration process never stopping. While it may seem counterintuitive—since repeated failures should indicate that the risky arm is likely "bad"—a conflict arises between Bayes' rule for updating beliefs and the first-order condition for optimality, leading to this continuous exploration strategy.

Notably, this non-stop exploration does not guarantee discovering the "good" arm. Unless the agent begins with a prior belief of 1, there is always a positive probability that a good arm may never yield a breakthrough, no matter how long exploration continues. In the absence of a breakthrough, the optimal allocation gradually converges to zero.

Our analysis uncovers three inefficiencies associated with restricting decisions to binary strategies. First, the cutoff belief under binary strategies is higher than the optimal belief, meaning that the agent may forgo exploration when it would be warranted under the optimal strategy. Second, even when the prior belief is high enough to justify full-time exploration of the risky arm under a binary strategy, the total time devoted to exploration is less than it would be under the optimal continuous strategy, reducing the optimal probability of a breakthrough. Third, efficiency requires balancing the marginal cost and marginal benefit (accounting for the learning effect) in each period. The binary constraint prevents this, resulting in lower expected payoffs during periods without a breakthrough.

Additionally, our study reveals that the optimal allocation strategy follows the Goldilocks principle: the incentive to explore is strongest when the task is neither too difficult nor too easy.

Related Literature

To the best of our knowledge, this is the first study to derive interior optimal allocation strategies for discrete-time BorN (breakthrough or nothing) bandits. In similar BorN settings, Bergemann and Hege (1998, 2005) examined the financing of innovation, assuming that the probability of a breakthrough is linear with investment. They derived an optimal binary strategy and focused on how financing decisions affect the stopping time. Rosenberg et al. (2007) extended Keller et al. (2005) model of strategic experimentation by allowing players to observe each other's actions but not their payoffs, finding that all equilibria involve time-varying cutoff strategies. Heidhues et al. (2015) explored a discrete-time version of the Keller-Rady-Cripps model, showing that payoff observability significantly influences equilibrium properties. Binary actions are also a frequent feature in the learning literature (e.g., Murto and Välimäki (2011)).

Recent economic literature on bandit problems has largely focused on continuous-time models, as noted by Bolton and Harris (1999), who highlighted the tractability of such models. In a BorN framework, Choi (1997) was the first to study R&D races under hazard rate uncertainty. Malueg and Tsutsui (1997) were the first to characterize optimal allocation strategies with interior solutions, although they introduced a fixed cost to discourage continuous experimentation. The seminal work by Keller et al. (2005) on exponential bandits derived a "bang-bang" strategy, where the agent fully allocates time to the risky arm when the posterior belief exceeds a certain threshold, and switches to the safe arm when it falls below. This strategy implies an optimal stopping time, after which no further time is allocated to the risky arm. The bang-bang strategy remains prevalent in both theoretical and empirical studies, either assumed or derived (e.g., Awaya and Krishna (2021); D. Thomas (2021); Besanko and Wu (2013)). Sadler (2021) modeled an agent working on suc-

cessive BorN-exponential bandits, interpreted as ideas, and investigated socially optimal tax or subsidy policies that maximize the value of research spillovers. The references in Sadler (2021) contain additional earlier works related to exponential bandits.

2 The Model

Time is discrete, with countable periods $t \in \mathbb{N} = \{1, 2, ...\}$. A decision-maker (henceforth referred to as the agent), is endowed with one unit of a perfectly divisible resource (referred to as time) per period and faces a two-armed bandit problem. One arm is "safe," and the other is "risky." The safe arm provides a known return of $\ell > 0$ per period. The discount factor is $\delta \in (0, 1)$, so if the agent allocates full time to the safe arm indefinitely, its present value is $L = \ell/(1 - \delta)$. The agent's task is to decide how much time to allocate to the risky arm in each period. Let $a_t \in [0, 1]$ represent the fraction of time allocated to the risky arm in period t, with the remaining time devoted to the safe arm, which yields a certain return of $(1-a_t)\ell$. The risky arm has an unknown type—it can either be "good" or "bad." If it is bad, it yields nothing; if it is good, it is worth full-time exploitation, with an expected present value $G = \gamma L \in (L, \infty)$, where $\gamma > 1$ measures the relative attractiveness of the good risky arm compared to the safe arm.

A breakthrough occurs when the agent obtains conclusive evidence that the risky arm is good. Before this happens, the probability of a breakthrough in any period t, given that the risky arm is good, follows an exponential distribution: $F(a_t) = 1 - e^{-\lambda a_t}$, where $\lambda > 0$ represents how easy it is to achieve a breakthrough. Following standard approaches in bandit problems, we assume memoryless experiments, where the probability of success in each period depends solely on the current allocation.

The agent starts with a prior belief $\pi_0 \in [0, 1]$ that the risky arm is good. According to Bayes' rule, if the agent allocates $a_1..., a_{t-1}$ in periods 1 to $t - 1$

without a breakthrough, the posterior belief after period t is updated as follows:

$$\pi_t = \begin{cases} 1^+ & \text{if a breakthrough occurs in period } t \\ \frac{\pi_{t-1}e^{-\lambda a_t}}{1-\pi_{t-1}\left(1-e^{-\lambda a_t}\right)} & \text{if no breakthrough in } t \end{cases} \tag{1}$$

If a breakthrough occurs, $\pi_t = 1$, meaning that the agent knows with certainty that the risky arm is good. The sequence $(\pi_t)_{t=0}^{\infty}$ forms a martingale, meaning that the conditional expectation $E(\pi_t|\pi_{t-1})$ equals π_{t-1} for all $a_t \in [0,1]$ and $t \in \mathbb{N}$.

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, and the plan stops once a breakthrough occurs.

Using dynamic programming, the agent's optimal allocation plan $\boldsymbol{\alpha}$ satisfies the Bellman equation for all $t \in \mathbb{N}$:

$$V^*(\pi_{t-1}) = \max_{a\in[0,1]} V(a, \pi_{t-1}) \text{ s.t. } (1) \tag{2}$$

where $V : [0,1]^2 \to [L, G]$ represents the agent's expected conditional payoff:

$$V(a, \pi_{t-1}) = (1-a)\ell + \delta\pi_{t-1}\left(1-e^{-\lambda a}\right)G + \delta(1-\pi_{t-1}\left(1-e^{-\lambda a}\right))V^*(\pi_t) \tag{3}$$

The interpretation is straightforward: suppose the risky arm has yielded nothing by period $t-1$, and the agent's past allocations imply a posterior belief π_{t-1}. If the agent allocates α_t to the risky arm in period t, he receives an immediate return $(1-\alpha_t)\ell$ from the safe arm. There is a probability $\pi_{t-1}\left(1-e^{-\lambda\alpha_t}\right)$ that a breakthrough will occur, rewarding the agent with G. If no breakthrough occurs, the agent updates his belief and continues to the next period.

We first take a look at the necessary conditions for an optimal allocation plan. By standard arguments, $V(a,\pi)$ is continuously differentiable in both arguments. Therefore, for all $a = \alpha_t$ at which (2) has an interior solution, the first-order condition must be satisfied:

$$\begin{aligned} \frac{\partial}{\partial a}V(\alpha_t, \pi_{t-1}) &= -\ell + \delta\pi_{t-1}\lambda e^{-\lambda\alpha_t}\left[G - V(\alpha_{t+1}, \pi_t)\right] && (4) \\ &\quad +\delta(1-\pi_{t-1}\left(1-e^{-\lambda\alpha_t}\right))\frac{\partial V(\alpha_{t+1},\pi_t)}{\partial \pi_t}\frac{\partial \pi_t}{\partial a} && (5) \\ &= 0 \end{aligned}$$

where we apply the envelope theorem for $\partial V/\partial \pi_t$, and

$$\frac{\partial \pi_t}{\partial a} = -\frac{(1-\pi_{t-1})\,\pi_{t-1}}{\left(1-\pi_{t-1}\left(1-e^{-\lambda\alpha_t}\right)\right)^2}\lambda e^{-\lambda\alpha_t} < 0 \tag{6}$$

The term in (5) captures the learning effect inherent in the bandit problem. Since $\partial V/\partial \pi_t \geq 0$ (holding any plan fixed, increasing π_t increases the probability of a breakthrough), this effect is negative. Myopic solutions, on the other hand, ignore this learning effect and equate the marginal cost ℓ with the current-period marginal benefit $\delta \pi_{t-1} \lambda e^{-\lambda\alpha_t}\left[G - V(\alpha_{t+1}, \pi_t)\right]$. As a result, optimal allocations that account for learning effects are generally lower than the myopic allocations.

3 The Two-States Approach

Program (2) suggests that in the search for an optimal solution, one may restrict attention to pure stationary Markovian allocation plans, meaning that $\alpha_t = \alpha(\pi_{t-1})$ for all t, where α is a time-invariant deterministic function of the posterior belief. While this restriction is harmless in binary-action models (see Section 3.3), finding the optimal function $\alpha(\cdot)$ in the general case remains a challenging open question.

Rather than focusing solely on posterior beliefs, we treat both the allocation and the posterior belief as state variables. This approach allows us to consider the pair (α_t, π_{t-1}) as functions of the adjacent states, thereby offering a dynamic view of the balance between exploration and exploitation in the optimal allocation strategies. Additionally, this approach significantly simplifies the analysis.

By replacing $V^*(\pi_{t-1})$ with $V(\alpha_t, \pi_{t-1})$, we transform the problem into a mathematically equivalent form to (2)-(3):

$$V(\alpha_t, \pi_{t-1}) - G = \max_{a\in[0,1]} \left(V(a, \pi_{t-1}) - G\right) \text{ s.t. } (1)$$

where

$$V(a, \pi_{t-1}) - G = (1-a)\ell - (1-\delta)G + \left(1-\pi_{t-1}(1-e^{-\lambda a})\right)\delta\left[V(\alpha_{t+1}, \pi_t) - G\right] \tag{7}$$

To simplify notation, denote

$$\begin{aligned} c_t &= (1-\delta)G - (1-\alpha_t)\ell \\ q_t &= \pi_{t-1}(1-e^{-\lambda\alpha_t}) \end{aligned}$$

Then, for $a = \alpha_t$, expanding (7) yields

$$\begin{aligned} V(\alpha_t, \pi_{t-1}) - G &= -c_t + (1-q_t)\,\delta\,[V(\alpha_{t+1}, \pi_t) - G] \qquad (8) \\ &= -c_t - \sum_{s=1}^{\infty} \delta^s \left(\prod_{r=0}^{s-1} (1-q_{t+r}) \right) c_{t+s} \qquad (9) \end{aligned}$$

Lemma 1 *Suppose* $\boldsymbol{\alpha}$ *is optimal. Then,* $\forall t \in \mathbb{N}$,

$$\frac{\partial}{\partial a} V(a, \pi_{t-1}) = -\ell + \delta \pi_{t-1} \lambda e^{-\lambda a} H_{t+1} \qquad (10)$$

where

$$H_{t+1} = c_{t+1} + \delta e^{-\lambda \alpha_{t+1}} H_{t+2} > 0 \qquad (11)$$

Proof. The term in large brackets in (9) represents the conditional probability of no breakthrough over the next s periods. This can be re-written as:

$$\prod_{r=0}^{s-1} (1-q_{t+r}) = 1 - \pi_{t-1} + \pi_{t-1} e^{-\lambda \sum_{r=0}^{s-1} \alpha_{t+r}} \qquad (12)$$

Define H_{t+1} by

$$H_{t+1} = c_{t+1} + \delta e^{-\lambda \alpha_{t+1}} H_{t+2} \quad \left(= c_{t+1} + \sum_{s=1}^{\infty} \delta^s e^{-\lambda \sum_{r=1}^{s} \alpha_{t+r}} c_{t+s} \right). \qquad (13)$$

Substituting (12) and (13) into (9), and replacing α_t with a, we get

$$V(a, \pi_{t-1}) = \delta G + (1-a)\ell - (1-\pi_{t-1})\, C_{t+1} - \pi_{t-1} e^{-\lambda a} \delta H_{t+1} \qquad (14)$$

where $C_{t+1} = \sum_{s=1}^{\infty} \delta^s c_{t+s}$. Since both C_{t+1} and H_{t+1} are functions of planned actions from period $t+1$ onward, the envelope theorem implies

$$\frac{\partial}{\partial a} V(a, \pi_{t-1}) = -\ell + \delta \pi_{t-1} \lambda e^{-\lambda a} H_{t+1}$$

as given in (10). ■

This lemma demonstrates that tracking future posteriors is unnecessary, and that the first-order conditions are simplified as shown in (10). In comparison to (4)-(5), we observe from (10) that the function H_{t+1} captures the overall marginal benefit, including learning effects, of investing time in the risky arm. The simplified form of $\partial V/\partial a_t$ arises because H_{t+1} does not depend directly on the sequence of posterior beliefs $\{\pi_t, \pi_{t+1}, ...\}$, unlike V.

3.1 Basic properties of the optimal allocation plan

If exploring the risky arm is deemed unprofitable, the agent's optimal payoff derives entirely from the returns of the safe arm, such that $V \equiv L$. To avoid this trivial case, we assume $\delta\lambda\left(\gamma-1\right) > 1-\delta$ and set the prior belief $\pi_0 > \pi_{\min}$, where:

$$\pi_{\min} = \frac{1-\delta}{\delta\lambda\left(\gamma-1\right)} \tag{15}$$

Note that:

$$\frac{\partial}{\partial a}V(0,\pi_{\min}) = -\ell + \delta\pi_{\min}\lambda(G-L) = 0$$

which implies that $\alpha_t > 0$ if and only if $\pi_{t-1} > \pi_{\min}$.

We now introduce a "no stopping" result that sharply contrasts with existing literature on exponential bandits under binary strategies.

Proposition 1 *Suppose $\pi_0 \in (\pi_{\min}, 1)$, and $\boldsymbol{\alpha}$ is optimal. Then $\alpha_t > 0$ for all $t \in \mathbb{N}$. That is, experimentation with the risky arm never stops without a breakthrough.*

Proof. It suffices to show that $\pi_{t-1} > \pi_{\min}$ implies $\pi_t > \pi_{\min}$ for all $t \in \mathbb{N}$. We prove this by contradiction. Pick any t such that $\pi_{t-1} > \pi_{\min}$, meaning $\alpha_t > 0$. Suppose $\pi_t \leq \pi_{\min}$. Then, $\alpha_{t+1} = 0$, implying $H_{t+1} = G - L$. From (10) in Lemma 1, α_t, π_{t-1} and π_t must satisfy the following conditions:

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

where the second inequality comes from the assumption $\pi_t \leq \pi_{\min}$. Cancelling terms, these two conditions imply

$$\frac{\pi_{t-1}}{\pi_t} \geq e^{\lambda \alpha_t}. \tag{16}$$

However, by (1), for $\alpha_t > 0$, we have

$$\frac{\pi_{t-1}}{\pi_t} = e^{\lambda \alpha_t} (1 - \pi_{t-1}) + \pi_{t-1} < e^{\lambda \alpha_t} \tag{17}$$

The contradiction between (16) and (17) proves $\pi_t > \pi_{\min}$, and therefore $\alpha_t > 0$. This confirms the proposition. ■

The general "no-stop" result in this proposition might seem surprising. The proof highlights a key reason: for stopping to be optimal at any time $t+1$, two conditions must be met simultaneously. One condition, in (16), derives from the first-order condition for optimality, while the other, in (17), derives from the Bayes rule. Stopping in any period t leads to a conflict between these conditions, making it impossible.

A natural question arising from Proposition 1 is: since the agent never stops experimenting with the risky arm, does this mean the good arm will be discovered with probability 1? In other words, if, in period 0, Nature assigns a probability π_0 for the risky arm to be good and $1 - \pi_0$ for it to be bad, and the risky arm is indeed good (though the agent is unaware of the outcome, knowing only the ex-ante probability π_0), will the agent eventually discover the truth through persistent trials?

Our next proposition addresses this question.

If the risky arm is good, the probability of no breakthrough by period T is $e^{-\lambda \sum_{t=1}^{T} \alpha_t}$. By Bayes' rule, the odds ratios of the posterior beliefs are updated as follows:

$$\begin{aligned} \frac{\pi_T}{1 - \pi_T} &= e^{-\lambda \alpha_T} \frac{\pi_{T-1}}{1 - \pi_{T-1}} \\ &= e^{-\lambda \sum_{t=1}^{T} \alpha_t} \frac{\pi_0}{1 - \pi_0} \end{aligned} \tag{18}$$

Proposition 2 *Suppose $\pi_0 \in (\pi_{\min}, 1)$ and $\boldsymbol{\alpha}$ is optimal.*

(i) $\pi_t \to \pi_{\min}$ and $\alpha_t \to 0$ as $t \to \infty$.

(ii) The agent's breakthrough probability is

$$\Pr(\textit{Breakthrough}) = \frac{\pi_0 - \pi_{\min}}{1 - \pi_{\min}} \tag{19}$$

As a result, given that the risky arm is good, the conditional probability is:

$$\Pr(\textit{Breakthrough} \,|\textit{Risky arm good}) = \frac{\pi_0 - \pi_{\min}}{\pi_0 (1 - \pi_{\min})} \tag{20}$$

(iii) Both probabilities above are increasing functions of π_0, δ, λ, and γ.

Proof. (i) The sequence of no-breakthrough posteriors $(\pi_t)_{t=0}^{\infty}$ forms a decreasing sequence, bounded below by $\pi_{\min}$. Thus, by the Monotone Convergence Theorem, π_t tends to a limit $\pi_\infty \geq \pi_{\min}$. Taking the limit as $T \to \infty$ in (18) yields

$$\frac{\pi_\infty}{1 - \pi_\infty} = e^{-\lambda \sum_{t=1}^{\infty} \alpha_t} \frac{\pi_0}{1 - \pi_0} \tag{21}$$

This implies $\lim_{t\to\infty} \alpha_t = 0$ because the sum $\sum_{t=1}^{\infty} \alpha_t$ is finite. Consequently, taking the limit in (11), we obtain

$$\begin{aligned} \lim_{t\to\infty} H_{t+1} &= g - (1 - \lim_{t\to\infty} \alpha_{t+1})\ell + \delta \lim_{t\to\infty} e^{-\lambda\alpha_{t+1}} \lim_{t\to\infty} H_{t+2} \\ &= g - \ell + \delta \lim_{t\to\infty} H_{t+2} \end{aligned}$$

Solving this gives

$$\lim_{t\to\infty} H_{t+1} = \frac{g - \ell}{1 - \delta} = \frac{\gamma - 1}{1 - \delta}\ell$$

By Lemma 1, for all t such that $\alpha_t \in (0, 1)$, we have

$$\frac{\partial}{\partial a} V(\alpha_t, \pi_{t-1}) = -\ell + \delta \pi_{t-1} \lambda e^{-\lambda\alpha_t} H_{t+1} = 0$$

Thus, as $t \to \infty$, we obtain

$$-\ell + \delta \pi_\infty \lambda \frac{\gamma - 1}{1 - \delta}\ell = 0 \Longrightarrow \pi_\infty = \pi_{\min}$$

Thus, the proof of (i) is complete.

(ii) Given that the risky arm is good, the agent's breakthrough probability can now be computed from (21), substituting $\pi_{\min}$ for π_∞ :

$$1 - e^{-\lambda \sum_{t=1}^{\infty} \alpha_t} = 1 - \frac{\pi_{\min}}{1-\pi_{\min}} \frac{1-\pi_0}{\pi_0} = \frac{\pi_0 - \pi_{\min}}{\pi_0 (1-\pi_{\min})}$$

which gives (20). Multiplying both sides by π_0 (the probability that the risky arm is good) gives (19).

(iii) The result is straightforward to verify, noting that $\pi_{\min}$ is a decreasing function of δ, λ, and γ. ■

In this proposition, the conditional breakthrough probability given by (20) shows that, as long as the agent has imprecise information ($\pi_0 < 1$) regarding the true, good state of the risky arm there is always a positive probability of making mistakes. Although the agent never stops experimenting, the time allocated to the risky arm becomes infinitesimally small over time. As a result, the total probability of achieving a breakthrough is strictly less than 1 even when the risky arm is good. Naturally, this probability approaches 1 as π_0 approaches 1, illustrating the value of precise information in this bandit problem. Proposition 2 also formally establishes a necessary condition for $\boldsymbol{\alpha}$ to be optimal: (α_t, π_{t-1}) must tend to $(0, \pi_{\min})$ as $t \to \infty$.

Part (iii) of Proposition 2 focuses on the comparative statics of the breakthrough probabilities with respect to π_0, λ, δ, and γ. Before interpreting these results, we will first examine the comparative statics of the total maximum time allocated to the risky arm.

Proposition 3 *Define the maximum time to be allocated to the risky arm by* $A = \sum_{t=1}^{\infty} \alpha_t$.

(i) A *is an increasing function of* (π_0, δ, γ).

(ii) Fix any (π_0, δ, γ) *and consider* A *as a function of* λ. *There exists a unique* λ^* *that maximizes* A *such that* $A'(\lambda) > 0$ *for* $\lambda < \lambda^*$ *and* $A'(\lambda) < 0$ *for* $\lambda > \lambda^*$.

Proof. By (20), we have

$$A = \frac{1}{\lambda} \ln \left(\frac{\pi_0}{(1-\pi_0)} \frac{1-\pi_{\min}}{\pi_{\min}} \right)$$

(i) The right-hand side increases in π_0 and decreases in $\pi_{\min}$. Given that $\pi_{\min}$ is a decreasing function of δ and γ, A is an increasing function of (π_0, δ, γ).

(ii) Substituting (15), we get

$$A'(\lambda) = -\frac{1}{\lambda}\left(A - \frac{\delta(\gamma-1)}{\delta - \lambda\delta + \lambda\gamma\delta - 1}\right)$$

Therefore, $A' = 0$ implies

$$\begin{aligned} A &= \frac{\delta(\gamma-1)}{\delta - \lambda\delta + \lambda\gamma\delta - 1} \\ \text{and } A'' &= -\frac{1}{\lambda}\left(A' - \frac{\partial}{\partial\lambda}\frac{\delta(\gamma-1)}{\delta - \lambda\delta + \lambda\gamma\delta - 1}\right) \\ &= -\frac{1}{\lambda}A^2 < 0 \end{aligned}$$

The statement (ii) is thus confirmed. ■

Since A is related to the breakthrough probabilities given in (19)-(20), the comparative statics results in Proposition 3(i) Proposition 2(iii) can be interpreted as follows. In (19), increasing π_0 has two effects: a direct effect that the risky arm has a higher probability of being good, and an indirect effect that encourages the agent to commit a higher level of total time A to the exploration of the arm. Both effects are positive, contributing to a higher probability of breakthrough. In (20), only the indirect effect is present, but it still contributes to a higher breakthrough probability. Likewise, increasing δ or γ makes the exploration more attractive, resulting in a positive, indirect effect on the breakthrough probability through a higher level of A.

Regarding the hazard rate, increasing λ has a direct positive effect on the breakthrough probability but may or may not encourage more intensive exploration of the risky arm (see Figure 1). Interestingly, Proposition 3(ii) introduces a new observation: $A(\lambda)$ follows the Goldilocks principle, meaning that optimal allocation increases with λ when the task is difficult (low λ) but decreases when the task is easy (high λ). The intuition is that when there is high confidence that a breakthrough will occur as long as the risky arm is good, a higher λ makes breakthroughs easier,

reducing the agent's need to allocate additional time due to opportunity cost. However, when the task is sufficiently difficult, increasing λ makes it more achievable, prompting the agent to invest more time in reaching a breakthrough.

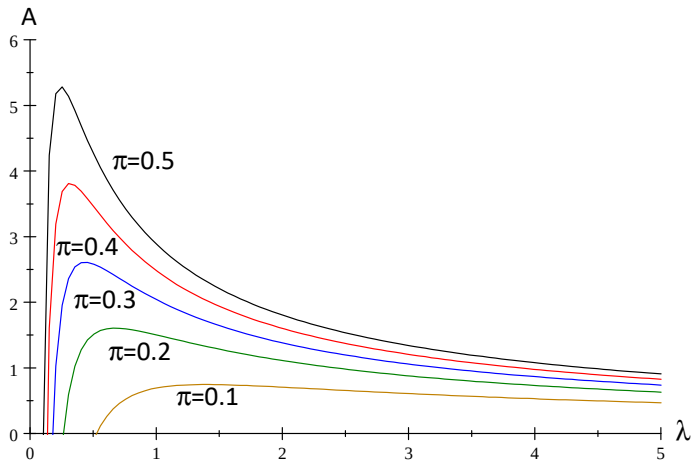

Figure 1: **Goldilocks principle**: the agent allocates the maximum total time to the experiment when the job is neither too difficult (low λ) nor too easy (high λ). The figure depicts the behavior of the planned total exploration time from the current period onward, A, as a function of λ and the (updated) belief π, for $\delta = 0.95$ and $\gamma = 2$.

This finding contrasts with Malueg and Tsutsui (1997), who predicted that optimal allocation should always increase with λ. In the binary-choice literature, Choi (1997) was the first (and only) to find indeterminate comparative statics predictions regarding the effect of λ. Choi attributed this to hazard rate uncertainty, showing that increasing λ could extend the experiment's stopping time when the prior π_0 is low, but shorten it when the prior is high. In contrast, our study offers clearer insights, showing that A peaks at a unique $\lambda^* \in (0, \infty)$, which is a function of all the exogenous variables.

3.2 Optimal allocation plan

Let us partition the bandit environment into two complementary scenarios:

Scenario I. $\delta\pi_0\left(e^{-\lambda}\left(1+\lambda\gamma\right)-1\right)\leq\left(1-\delta\right)$.

Scenario II. $\delta\pi_0\left(e^{-\lambda}\left(1+\lambda\gamma\right)-1\right)>\left(1-\delta\right)$.

Proposition 4 *Given any $\pi_0\in(\pi_{\min},1)$, there exists a unique optimal allocation plan $\boldsymbol{\alpha}$. (i) Under **Scenario I**, the sequence of the joint states $\{(\alpha_t,\pi_{t-1})\}_{t=1}^{\infty}$ satisfies the backward recursive relation*

$$\alpha_t = \frac{1}{\lambda}\ln\left(1+\delta\lambda\left(\gamma-1+\alpha_{t+1}\right)-\frac{1-\delta}{\pi_t}\right)\in(0,1) \tag{22}$$

$$\pi_{t-1} = \frac{\pi_t}{\pi_t+e^{-\lambda\alpha_t}(1-\pi_t)}\in(\pi_{\min},\pi_0) \tag{23}$$

*for all $t\in\mathbb{N}$. (ii) Under **Scenario II**, there is a unique switching period $\tau\in\{0\}\cup\mathbb{N}$ such that $\alpha_t=1$ for $t\leq\tau$[1] and $\{(\alpha_t,\pi_{t-1})\}_{t=\tau+1}^{\infty}$ satisfies (22)-(23).*

Proof. See Appendix. ■

Proposition 4 demonstrates that, in any period t, the pair of optimal allocation and belief (α_t,π_{t-1}) is uniquely determined by the subsequent states. Therefore, α_t and π_{t-1} are functions of the subsequent period allocation and belief (α_{t+1},π_t). The mapping provided in (22)-(23) is time-invariant, offering a remarkably simple algorithm for determining the optimal allocation plan (see Figure 2).

As seen in the proof of this proposition, we introduce a novel approach featuring a form of backward recursion to derive the mapping (22)-(23). This approach consists of three steps:

1. *Initial Assumption*: We begin by considering an arbitrarily large but finite T, assuming the experiment stops after period T. In this context, the posterior belief in the last period is $\pi_{T-1}=(1+\varepsilon_T)\pi_{\min}$. Crucially, we treat π_{T-1} as a free variable rather than a function of the prior belief and the history of the past allocations.

[1] If $\tau=0$, without any consequence we define $\alpha_0=1$.

2. *Backward Optimization:*. Using backward induction while ensuring consistency with Bayes' rule for posteriors, we derive an optimal sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T)_{t=1}^T$ parameterized by ε_T.

3. *Existence and Uniqueness*: We show the existence of a unique ε_T^* such that the sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T^*)_{t=1}^T$ has the initial prior equal to π_0. Finally, by taking the limit as $T \to \infty$ and applying the transversality condition $(\alpha_T, \pi_{T-1}) \to (0, \pi_{\min})$, we establish the existence and uniqueness of the optimal allocation plan.

Since the mapping (22)-(23) is bijective, the functional relationship between any two adjacent pairs of states can also be expressed by a forward-moving law of motion, given the optimal (α_1, π_0). Our next proposition explores this forward-moving approach.

Proposition 5 *There is a unique optimal no-breakthrough allocation plan* $\boldsymbol{\alpha}$ *associated with a unique switching time* $\tau \in \{0\} \cup \mathbb{N}$ *such that* $\alpha_t = 1$ *for all* $t \leq \tau$*, and the sequence of the joint states* $\{(\alpha_t, \pi_{t-1})\}_{t=\tau+1}^{\infty}$ *obeys the law of motion*

$$\alpha_{t+1} = \frac{1}{\delta\lambda}\left(e^{\lambda\alpha_t} + \frac{1-\delta}{\pi_t} - 1\right) - (\gamma - 1) \in (0, 1) \tag{24}$$

$$\pi_t = \frac{\pi_{t-1}e^{-\lambda\alpha_t}}{1 - \pi_{t-1}(1 - e^{-\lambda\alpha_t})} \in (\pi_{\min}, \pi_0) \tag{25}$$

Proof. This is a straightforward corollary of Proposition 4, given the equivalence between (22)-(23) and (24)-(25). ■

Compared with Proposition 4, the forward motion described in Proposition 5 has the advantage of being more familiar and easier to implement. is more familiar and easier to implement. The agent may begin by allocating full time to exploring the risky arm. As time passes without a breakthrough, the agent will, at a certain time τ, switch to a strategy with an allocation less than 1. Unlike the bang-bang strategy, where the allocation drops from 1 to 0 at a specific cutoff posterior

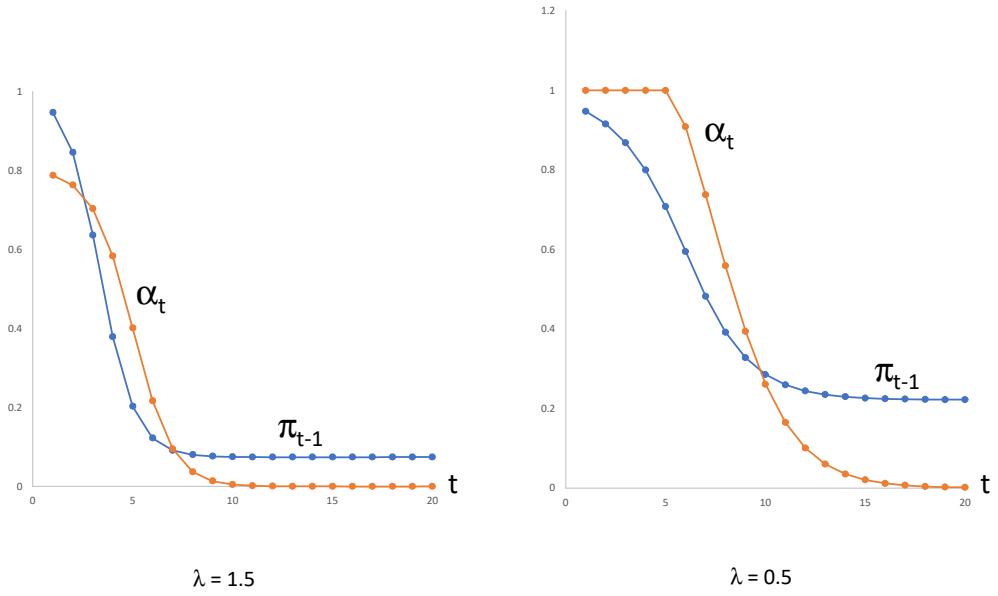

Figure 2: Both figures assume $\delta = 0.9$, $\gamma = 2$, starting with $\pi_0 = 0.95$. They differ only in λ.

level, Proposition 5 predicts more moderate downward adjustments in response to no breakthroughs. At the switching time, the initial condition $(\alpha_{\tau+1}, \pi_\tau)$ must be optimal; otherwise, the motion law described in (24)-(25) lacks an optimality foundation. Fortunately, $(\alpha_{\tau+1}, \pi_\tau)$ can be derived or estimated with arbitrary precision using the backward recursion algorithm from Proposition 4.

Figure 2 depicts the dynamics of the joint states (π_{t-1}, α_t) for Scenario I (left) and Scenario II (right). Consistent with the Goldilocks principle, when λ is relatively high ($\lambda = 1.5$), expecting an easy breakthrough the agent does not allocate full time to the risky arm. But after a few periods of failure, the posterior quickly drops toward $\pi_{\min}$ indicating that the risky arm is most likely bad. As a result, α_t quickly drops toward 0. When λ is reasonably encouraging but not sufficiently low ($\lambda = 0.5$), the agent allocates full time to the risky arm, switching to a more moderate allocation strategy after some periods of failure.

3.3 Optimal vs. binary strategies

Replacing $a \in [0,1]$ in (2) with $a \in \{0,1\}$, we arrive at the exponential bandit problem under binary strategies (see, e.g., Heidhues et al. (2015)). For the single-agent case, the binary optimal plan is relatively straightforward to derive.

In each period, the agent's decision is to compare which choice—$a = 0$ or $a = 1$—yields the higher expected payoff for that period. Let π^* represent the cutoff belief that the risky arm is good, where the agent is indifferent between choosing 0 and 1. Then, from (2)-(3),

$$V(1,\pi^*) - L = \delta\left\{\pi^*(1-e^{-\lambda})(G-L) + L\right\} - L = 0$$

implies, recalling that $G/L = \gamma$, the cutoff belief π^* is given by

$$\pi^* = \frac{(1-\delta)}{\delta(1-e^{-\lambda})(\gamma-1)} \tag{26}$$

The optimal action plan is a pure stationary Markovian plan, defined as follows:[2]

$$\beta(\pi) := \begin{cases} 1 & \text{if } \pi \geq \pi^* \\ 0 & \text{if } \pi < \pi^* \end{cases} \tag{27}$$

If $\pi_0 < \pi^*$, the agent will not explore the risky arm at all. If $\pi_0 \geq \pi^*$, the agent starts with $\alpha_1 = 1$ and continues this until either a breakthrough occurs or period T is reached without a breakthrough. Therefore, the maximum scheduled time for exploring the risky arm is given by:

$$T = \min\{t \in \mathbb{N} : \pi_{t+1} = \frac{\pi_0 e^{-\lambda t}}{1-\pi_0(1-e^{-\lambda t})} < \pi^*\}$$

If the risky arm is good, the probability of no breakthrough until period T is $e^{-\lambda T}$. Since $\pi_T \geq \pi^*$, from (18) we derive

$$e^{-\lambda T} = \frac{1-\pi_0}{\pi_0}\frac{\pi_T}{1-\pi_T} \geq \frac{1-\pi_0}{\pi_0}\frac{\pi^*}{1-\pi^*} \tag{28}$$

[2]When the agent is indifferent between 0 and 1, we assume he chooses 1 so that any choice of 0 implies the same choice indefinately afterward.

which implies

$$T = \frac{1}{\lambda}\ln\left(\frac{\pi_0}{1-\pi_0}\frac{1-\pi_T}{\pi_T}\right) \leq \frac{1}{\lambda}\ln\left(\frac{\pi_0}{1-\pi_0}\frac{1-\pi^*}{\pi^*}\right) \tag{29}$$

Now we let $\boldsymbol{\alpha} = (\alpha_t)_{t=1}^{\infty}$, the optimal allocation plan presented in Propositions 4-5, be referred to as Plan A, and let the binary plan β, as described in (27), be referred to as Plan B. Let V_A and V_B denote the expected payoffs under Plan A and Plan B, respectively. The values of V_A and V_B can be expressed recursively according to (8), and by Proposition 4 and (27), these values can be computed using inputs of the corresponding allocation plans:

$$\begin{aligned} V_A(\alpha_t, \pi^A_{t-1}) &= \delta G + (1-\alpha_t)\ell + \left(1 - \pi^A_{t-1}(1-e^{-\lambda\alpha_t})\right)\delta\left[V_A(\alpha_{t+1}, \pi^A_t) - G\right] \\ V_B(\beta_t, \pi^B_{t-1}) &= \delta G + (1-\beta_t)\ell + \left(1 - \pi^B_{t-1}(1-e^{-\lambda\beta_t})\right)\delta\left[V_B(\beta_{t+1}, \pi^B_t) - G\right] \end{aligned}$$

Here, $\{\pi^A_{t-1}\}$ and $\{\pi^B_{t-1}\}$ are the sequences of the (no breakthrough) posteriors under Plan A and Plan B, and V_A satisfies the transversality condition $\lim_{t\to\infty} V_A(\alpha_t, \pi^A_{t-1}) = L$. For $t > T$, $\beta_t \equiv 0$ so that $V_B(\beta_t, \pi^B_{t-1}) \equiv L$.

Proposition 6 *Suppose* $\pi_0 \in (\pi_{\min}, 1)$.

(i) Plan A provides a stronger incentive to begin exploring the risky arm compared to Plan B, as $\pi_{\min} < \pi^*$.

(ii) Plan A involves a greater total time commitment to exploring the risky arm than Plan B, i.e., $A > T$.

(iii) As a result, the probability of discovering a good risky arm is higher under Plan A than under Plan B, as $1 - e^{-\lambda A} > 1 - e^{-\lambda T}$.

(iv) In any period t *preceded by no breakthrough, the expected payoff under Plan A is strictly higher than under Plan B, i.e.,* $V_A(\alpha_t, \pi^A_{t-1}) > V_B(\beta_t, \pi^B_{t-1})$ *for all* $t \in \mathbb{N}$.

Proof. (i) Since $e^{-\lambda} > 1 - \lambda$, then

$$\pi^* = \frac{(1-\delta)}{\delta(1-e^{-\lambda})(\gamma-1)} > \frac{1-\delta}{\delta\lambda(\gamma-1)} = \pi_{\min}$$

The conclusions in (ii) and (iii) hold because

$$\begin{aligned} A &= \frac{1}{\lambda}\ln\left(\frac{\pi_0}{(1-\pi_0)}\frac{1-\pi_{\min}}{\pi_{\min}}\right) \\ &> \frac{1}{\lambda}\ln\left(\frac{\pi_0}{(1-\pi_0)}\frac{1-\pi^*}{\pi^*}\right) \quad \text{by (i)} \\ &\geq T \quad \text{by (29)} \end{aligned}$$

(iv) Plan B is the solution of (2)-(3) under the constraint $a \in \{0,1\}$, whereas Plan A is not bound by this constraint. Moreover, for t sufficiently large (i.e., $> \tau$), Plan A involves always the interior solutions, i.e., $\alpha_t \in (0,1)$ for $t > \tau$. This fact implies that Plan B is strictly suboptimal in that even given the same posterior at the start of any period t, $V_A(\alpha_t, \pi) > V_B(\beta_t, \pi)$. Moreover, $1 = \beta_t \geq \alpha_t$ for all $t \leq T$ implies $\pi^A_{t-1} \geq \pi^B_{t-1}$, and $V(a,\pi)$ strictly decreases in π. Therefore $V_A(\alpha_t, \pi^A_{t-1}) > V_B(\beta_t, \pi^B_{t-1})$ for all $t \in \mathbb{N}$. ■

Part (i) of this proposition implies that when $\pi_0 \in (\pi_{\min}, \pi^*)$, the risky arm will be considered unprofitable under Plan B. However, this is not the case under Plan A. In this regard, Plan A offers stronger incentives to engage in risky arm exploration. Parts (ii) and (iii) of the proposition demonstrate that even when the prior belief is sufficiently high for the agent to invest full time exploring the risky arm under the binary constraint, the total time devoted to exploration is still lower than under the optimal allocation policy. As a result, the probability of achieving a breakthrough is strictly lower with the binary constraint. Table 1 illustrates a numerical example of conclusion (iv) of the proposition.

Period t	α_t	π^A_{t-1}	V_A	β_t	π^B_{t-1}	V_B
20	0.000666334	0.074148148	**10.00066648**	0	0.064089558	**10**
19	0.001497119	0.074302462	**10.00209935**	0	0.064089558	**10**
18	0.004101149	0.074726697	**10.00601196**	0	0.064089558	**10**
17	0.011452222	0.075923163	**10.0170383**	0	0.064089558	**10**
16	0.031471594	0.079302198	**10.04817894**	0	0.064089558	**10**
15	0.08235546	0.088803474	**10.13576379**	0	0.064089558	**10**
14	0.192669699	0.115135149	**10.37866229**	0	0.064089558	**10**
13	0.369854021	0.184741567	**11.02174453**	0	0.064089558	**10**
12	0.557589366	0.343403576	**12.49154383**	0	0.064089558	**10**
11	0.689155649	0.595215881	**14.83230322**	0	0.064089558	**10**
10	0.756919538	0.820678269	**16.93479695**	1	0.234829613	**10.64188839**
9	0.78576814	0.937003143	**18.01739151**	1	0.57902243	**13.36626138**
8	0.796768769	0.980057438	**18.39830904**	1	0.86041747	**16.02042038**
7	0.800699624	0.993914876	**18.46497373**	1	0.965066868	**17.10363236**
6	0.80204805	0.998165004	**18.36640486**	1	0.991987929	**17.40213455**
5	0.802498059	0.999448669	**18.15711593**	1	0.998201068	**17.47503653**
4	0.802645313	0.999834536	**17.91903891**	1	0.999598042	**17.49223443**
3	0.802692793	0.999950358	**17.73445812**	1	0.999910283	**17.49624037**
2	0.802707928	0.999985108	**17.62491116**	1	0.99997998	**17.4971671**
1	0.802712709	0.999995533	**17.57045789**	1	0.999995533	**17.49738043**

Table 1: The expected payoff under Plan A is strictly higher than under Plan B. Both plans start from $\pi_0 = 0.99999$, assuming $\delta = 0.9$, $\lambda = 1.5$, $\ell = 1$, $\gamma = 2$ and $G = 20$.

4 Concluding Remarks

This study offers a first-of-its-kind analysis of Breakthrough-or-Nothing (BorN) bandit problems, where allocation strategies can be perfectly divided between the safe and risky arms. The study makes a significant theoretical contribution to the understanding of discrete-time bandits by showing that the problem's complexity can be reduced by expanding the state space and transforming the problem. This approach facilitates clear analytical solutions and comparative statics results. Specifically, we demonstrate that when decision variables are continuous, the optimal allocation strategy diverges from binary approaches, increasing the incentive to explore the risky arm and the expected payoff, and rendering cessation after a finite number of unsuccessful trials suboptimal. These findings go beyond merely complementing the existing literature—they substantially enhance it.

Although our analysis focuses on the exponential setting, we argue that, under certain conditions, the results can be extended to more general probability functions. The analytical framework we develop is flexible and adaptable to various scenarios involving multiple agents and strategic interactions, which we expect will lead to new insights. A compelling direction for future research is exploring how this transformation or re-formulation of BorN bandits can simplify the analysis of other problem types.

5 Appendix

Proof of Proposition 4. (i) Consider **Scenario I**. Let $1 < T < \infty$ be arbitrarily large, and assume that exploration stops in period $T+1$, so that $H_{T+1} = G - L$. Denote the expected payoff under this early stopping constraint by $V(a, \pi|T)$. For convenience, we simplify the notation by omitting a subscript t from V.

Let $\pi_{T-1} = (1+\varepsilon_T)\pi_{\min}$ represent the belief in period T, where $0 < \varepsilon_T < \bar{\varepsilon} := \frac{\pi_0}{\pi_{\min}} - 1$. In light of Proposition 2(i), ε_T tends to 0 as T approaches infinity.

The method we adopt follows a form of *backward recursion*. We start by treating π_{T-1} (or ε_T) as a free variable, rather than a function of the prior belief and the history of the past allocations. Using backward induction, while ensuring the posteriors remain consistent with Bayes' rule, we derive an optimal sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T)_{t=1}^{T}$ parameterized by ε_T.

Next, we determine a unique ε_T such that $\pi_0|\varepsilon_T$ equals the initial belief π_0. Finally, we employ a limit argument to establish the existence and uniqueness of the optimal plan $(\alpha_t)_{t=1}^{\infty}$. For notational convenience, we will suppress the parameter ε_T in the following derivations until it is needed.

First, consider the problem $\max_{a\in[0,1]} V(a, \pi_{T-1}|T)$. Given Proposition 2(i), for a sufficiently large T, we assume $\alpha_T \in (0,1)$. By Lemma 1,

$$\frac{\partial}{\partial a} V(\alpha_T, \pi_1|T) = -\ell + \delta\pi_{T-1}\lambda e^{-\lambda\alpha_T}(G-L) = 0$$

yields

$$\alpha_T = \frac{1}{\lambda} \ln \frac{\pi_{T-1}}{\pi_{\min}} = \frac{1}{\lambda} \ln(1+\varepsilon_T) = \arg\max_{a\in[0,1]} V(a, \pi_{T-1}|T) \tag{30}$$

where we used (15) and the fact that $G - L = (\frac{\gamma-1}{1-\delta})\ell$, as well as the concavity of $V(a, \pi|T)$ in a.

Now, for any $t < T$, assume $(\alpha_{t+1}, \pi_t) \in (0,1) \times (\pi_{\min}, \pi_0)$, such that

$$\alpha_{t+1} = \arg\max_{a\in[0,1]} V(a, \pi_t|T)$$

By (10), $\alpha_{t+1} \in (0,1)$ implies

$$\frac{\partial}{\partial a} V(\alpha_{t+1}, \pi_t|T) = -\ell + \delta\pi_t \lambda e^{-\lambda\alpha_{t+1}} H_{t+2} = 0$$

or equivalently,

$$H_{t+2} = \frac{\ell}{\delta\pi_t \lambda e^{-\lambda\alpha_{t+1}}} \tag{31}$$

Using (11) and, we derive

$$\frac{\partial}{\partial a} V(\alpha_t, \pi_{t-1}|T) = -\ell + \delta\pi_{t-1}\lambda e^{-\lambda\alpha_t} \left(c_{t+1} + \delta e^{-\lambda\alpha_{t+1}} H_{t+2}\right)$$

Substituting H_{t+2} from (31), we get

$$\frac{\partial}{\partial a} V(\alpha_t, \pi_{t-1}|T) = -\ell + \delta\pi_{t-1}\lambda e^{-\lambda\alpha_t} \left(c_{t+1} + \frac{\ell}{\pi_t\lambda}\right) \tag{32}$$

By substituting (1) and using $c_t = (1-\delta)G - (1-\alpha_t)\ell = (\gamma - (1-\alpha_t))\,\ell$, and after rearranging terms, we obtain

$$\frac{\partial}{\partial a} V(\alpha_t, \pi_{t-1}|T) = \left[\delta\pi_{t-1} M(\alpha_t, \alpha_{t+1}) - (1-\delta)\right] \ell \tag{33}$$

where $M(\alpha_t, \alpha_{t+1}) = e^{-\lambda\alpha_t}\left(1 + \lambda\left(\gamma - 1 + \alpha_{t+1}\right)\right) - 1$.

If $\alpha_t = 1$, then

$$0 \le \frac{\partial}{\partial a} V(1, \pi_{t-1}|T) = \left[\delta\pi_{t-1} M(1, \alpha_{t+1}) - (1-\delta)\right] \ell < \left[\delta\pi_0 M(1,1) - (1-\delta)\right] \ell$$

which implies

$$\delta\pi_0 \left(e^{-\lambda}\left(1 + \lambda\gamma\right) - 1\right) > 1 - \delta$$

However, this inequality is ruled out under **Scenario I**. Thus, we must have $\alpha_t \in (0,1)$, and by induction, all period allocations are in $(0,1)$. This, combined with (32), implies

$$\frac{\partial}{\partial a}V(\alpha_t, \pi_{t-1}|T) = -\ell + \delta\pi_{t-1}\lambda e^{-\lambda\alpha_t}\left(c_{t+1} + \frac{\ell}{\pi_t\lambda}\right) = 0 \tag{34}$$

for all $t \leq T$. The Bayes rule now gives

$$\pi_{t-1}(\alpha_t, \pi_t) = \frac{\pi_t}{\pi_t + e^{-\lambda\alpha_t}(1-\pi_t)} \tag{35}$$

Substituting (35) into (34), and solving for α_t, we obtain

$$\begin{aligned} \alpha_t &= \frac{1}{\lambda}\ln\left(1 + \delta\lambda\left(\gamma - 1 + \alpha_{t+1}\right) - \frac{1-\delta}{\pi_t}\right) \\ &= \arg\max_{a\in[0,1]} V_t(a, \pi_{t-1}|T) \end{aligned} \tag{36}$$

Thus, by induction, we have derived a sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T)_{t=1}^T$ that maximizes V under an initial prior $\pi_0|\varepsilon_T$. We now show that a unique ε_T exists such that $\pi_0|\varepsilon_T = \pi_0$, consistent with the given prior. From (35)-(36), it is clear that π_{t-1} has positive partial derivatives with respect to α_t and π_t, and α_t has positive partial derivatives with respect to α_{t+1} and π_t. Thus, both α_t and π_{t-1} are continuous, increasing functions of (α_{t+1}, π_t). By induction, $(\alpha_t, \pi_{t-1}|\varepsilon_T)$ can be regarded as a pair of continuous and increasing functions of ε_T on $[0, \bar{\varepsilon}]$ for all $t \leq T$.

In particular, $\pi_0|\varepsilon_T$ is a continuous and increasing function of ε_T. When $\varepsilon_T = \bar{\varepsilon}$, Bayes' rule implies $\pi_0|\bar{\varepsilon} > \pi_{T-1} = \pi_0$. When $\varepsilon_T = 0$, we have $\pi_{T-1} = \pi_{\min}$ and by (35), this implies $\pi_{T-2} = \pi_{\min}$, and so on. By backward induction, we conclude that $\pi_0|0 = \pi_{\min} < \pi_0$.

Therefore, by the Intermediate Value Theorem and monotonicity of $\pi_0|\cdot$, there exists a unique $\varepsilon_T^* \in (0, \bar{\varepsilon})$ such that $\pi_0|\varepsilon_T^* = \pi_0$. Consequently, we arrive at a unique sequence $(\alpha_t, \pi_{t-1}|\varepsilon_T^*)_{t=1}^T$ that solves the program in (2)-(3), satisfying (22)-(23), given any prior $\pi_0 \in (\pi_{\min}, 1)$ and stopping time $T+1$.

Next, consider $T \to \infty$. By Proposition 2, we have $\varepsilon_T^* \to 0$ so that, by continuity, for each $t < T$

$$\lim_{T\to\infty}\left(\alpha_t, \pi_{t-1}|\varepsilon_T^*\right) = \lim_{\varepsilon_T\to 0}\left(\alpha_t, \pi_{t-1}|\varepsilon_T^*\right) = \left(\alpha_t, \pi_{t-1}|0\right).$$

This allows us to define $(\alpha_t, \pi_{t-1})_{t=1}^{\infty} = (\alpha_t, \pi_{t-1}|0)_{t=1}^{\infty}$. By continuity, it is straightforward to see that the equations in (22)-(23) hold for each t as $T \to \infty$.

Finally, let $V(T)$ denote the optimal value of the agent's expected payoff under the constraint that the experiment must stop after T. Let V^* denote the optimal unconstrained value. Since $L \leq V^* \leq G$,

$$0 \leq V^* - V(T) \leq \delta^{T+1}(G - L)$$

Thus $\lim_{T\to\infty} V(T) = V^*$.

(ii) Consider now **Scenario II**. Define

$$\pi^s = \frac{(1-\delta)}{\delta\left(e^{-\lambda}(1+\lambda\gamma) - 1\right)}$$

Take any $t \geq 1$, and we will show that $\alpha_{t+1} = 1$ implies $\alpha_t = 1$. From part (i), $\alpha_{t+1} = 1$ implies $\pi_{t-1} > \pi^s$, and

$$\frac{\partial}{\partial a} V(1, \pi_t) = -\ell + \delta\pi_t \lambda e^{-\lambda} H_{t+2} \geq 0$$

which implies

$$H_{t+2} \geq \frac{\ell}{\delta\pi_t \lambda e^{-\lambda}}. \tag{37}$$

Similar to the derivation of (33), now using (37) we get

$$\begin{aligned} \frac{\partial}{\partial a} V(1, \pi_{t-1}) &\geq \left[\delta\pi_{t-1} M(1,1) - (1-\delta)\right]\ell \\ &= \left[\delta\pi_{t-1}\left(e^{-\lambda}(1+\lambda\gamma) - 1\right) - (1-\delta)\right]\ell \\ &> 0 \end{aligned}$$

where the last inequality holds because $\pi_{t-1} > \pi^s$. Thus, $\alpha_t = 1$, and by induction, $\alpha_1 = ... = \alpha_{t+1} = 1$.

If $\alpha_1 \in (0,1)$, then $\tau = 0$. If $\alpha_1 = 1$, then there exists a unique $\tau = \max\{t \in \mathbb{N} : \alpha_t = 1\}$ such that $\alpha_t \in (0,1)$ for all $t > \tau$. Part (i) of the proposition then implies that $\{(\alpha_t, \pi_{t-1})\}_{t=\tau+1}^{\infty}$ satisfies (22)-(23). ■

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. 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.

Choi, J. P. (1997). Herd behavior, the "penguin effect," and the suppression of informational diffusion: an analysis of informational externalities and payoff interdependency. *The RAND Journal of Economics*, pages 407–425.

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.

Heidhues, P., Rady, S., and Strack, P. (2015). Strategic experimentation with private payoffs. *Journal of Economic Theory*, 159:531–551.

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.

Sadler, E. (2021). Dead ends. *Journal of Economic Theory*, 191:105167.