Learning while Experimenting*

Ettore Damiano
University of Toronto

Li, Hao
University of British Columbia

Wing Suen
University of Hong Kong

December 4, 2018

Abstract. An agent performing risky experimentation can benefit from suspending it to directly learn about the state. "Positive" information acquisition seeks news that would confirm the state that favors experimentation. It is used as a last-ditch effort when the agent is pessimistic about the risky arm before abandoning it. "Negative" information acquisition seeks news that would demonstrate that experimentation is futile. It is used as an insurance strategy to avoid wasteful experimentation when the agent is still optimistic. A higher reward from risky experimentation expands the region of beliefs that the agent optimally chooses information acquisition rather than experimentation.

JEL classification. D83, L15

Keywords. Benefit of information, direct learning, perpetual learning

*We thank the editor and two anonymous referees for providing valuable comments that help improve this paper. Seminar participants at UBC, Shanghai Jiaotong University, Hong Kong University of Science and Technology, Korea University, Osaka University, and at the Canadian Economics Association Annual Conference have given useful feedback to our presentations. Qianjun Lyu has provided excellent research assistance to this project.

1. Introduction

In multi-armed bandit models (Robins 1952), an agent choosing from alternatives with stochastic payoffs (risky arms) faces a trade-off between maximizing the expected payoff based on what he currently knows about the alternatives, and learning about the stochastic processes that generate these payoffs by experimenting with different alternatives. The dynamic trade-off between exploitation and experimentation captured in bandit models has found many applications in economics, ranging from project selection in industrial research and development (Weizman, 1979; Roberts and Weizman, 1981), to job search with firm-specific or industry-specific productivities (Jovanovich, 1979), to monopoly producers learning about market demands (Rothschild, 1984). More recently, bandit problems have been extended to multi-agent settings, including general approaches to strategic experimentation (Bolton and Harris, 1999; Keller, Rady and Cripps, 2005), and applications such as R&D races (Choi, 1991),[1] wage setting (Fellis and Harris, 1996), and price competition (Bergemann and Valimaki, 1996). In all these models, the agent learns about an individual risky alternative by trying it; that is, learning occurs only *through* experimenting.

In this paper, we study a simple bandit model where during experimentation the agent can also engage in a costly dynamic process of information acquisition. Learning about risky arms during experimentation is modeled here as a pure information activity, because rewards can only arrive through experimentation. For example, in the R&D application of bandit problems, while a pharmaceutical firm has a team of laboratory chemists engaged in a process to develop a new drug, it may hire scientists to research about the biochemical foundation behind the potential new drug. A scientific theory will not in itself bring about the new drug, but it can inform the laboratory chemists in their trial and error process. Similarly, while a job-seeker goes through applications and interviews with different firms in an unfamiliar industry, he may be able to acquire information about the industry and his particular fit with it through word-of-mouth in his social network. The same is true for modeling consumer demand for an experience good: besides trying out the good themselves, buyers may be able to obtain useful information about the good by asking their friends who have positive or negative experience with consuming it.

[1] See also Reinganum (1981, 1982), Harris and Vickers (1985, 1987), and Malueg and Tsutsui (1997).

Although it is a pure information activity, learning *while* experimenting changes the agent's belief about the prospects of risky arms and affects the agent's experimentation. The value of information in our model is thus endogenous, and the comparison of different information structures—the main objective of this paper—depends on the current state of the experimentation process (i.e., how optimistic or pessimistic the agent is about the prospects of risky alternatives). To capture learning while experimenting, we use as the benchmark a single-agent, single-risky arm version of the exponential bandits model of Keller, Rady and Cripps (2005). The risky arm is either good and produces a success at an exponential rate, or bad with no possibility of success. Learning through experimentation takes a simple form: starting from a prior belief that the risky arm is good, as the agent continues to experiment without achieving a success, he becomes increasingly pessimistic, eventually abandoning the risky arm when the beliefs drops below a critical threshold.

We consider two opposite information structures in learning while experimenting, both modeled as an exponential processes with uncertain arrival rates. In the case of *positive information*, the agent pays to search for conclusive evidence that a success can indeed be obtained from experimentation. Of course, no such evidence can be found if the risky arm is actually bad. Moreover, even after the evidence arrives, the agent can achieve a success only by continuing with the risky arm, except now the arrival is stochastic but no longer uncertain. In the pharmaceutical firm R&D application, a scientific theory establishing the sound foundation or feasibility of the new drug may be an example of such information structure. In the case of *negative information*, the agent pays to search for conclusive evidence that no success can be obtained from experimentation. Such evidence arrives at a positive rate when the risky arm is bad, but it will never arrive in the opposite state. For the pharmaceutical firm example, negative information acquisition may take the form of doing a toxicological study that demonstrates a fatal flaw in the current approach of developing the drug, or exploring the possibility of a rival or superior drug. Closer to home, a researcher trying to prove a theorem may benefit from negative news coming from a counterexample, and one trying to solve a mathematical problem may benefit from positive news coming from an existence theorem guaranteeing that a solution exists. In both these cases, the arrival of positive or negative news provides conclusive evidence that causes the researcher's belief about the prospect of his research to jump up or down. These types of information acquisition activities are suitably modeled

by exponential bandit processes.

The introduction of information acquisition in experimentation enriches the analysis of optimal experimentation. Without information acquisition, the question is simply when the agent should quit the risky arm; with it, we ask what factors determine whether and when information acquisition is useful to the agent. Is it best for the agent to acquire information about the risky arm when he is still optimistic about achieving a success or when he is already pessimistic? The answer turns out to depend on whether information is positive or negative. This is because the type of information affects the nature of the interaction between learning through experimenting and learning while experimenting.

Positive information reinforces the direct learning through experimenting, because failure to uncover positive evidence speeds up the downgrading of the agent's belief that a success from the risky arm can be achieved. When the agent is optimistic about the risky arm, there is relatively little use in having the positive news, and since information is costly, it is not optimal for the agent to acquire it. In contrast, positive information is more valuable to the decision of whether to quit the risky arm when the agent is pessimistic, as it can potentially avoid quitting before achieving success. We show that the benefit of positive information over optimal experimentation is the highest when the agent is just about to quit, and is lower when the agent is more optimistic. In a sense, positive information may be interpreted as a last-ditch effort in trying the risky arm before irrevocably quitting it. Moreover, the agent optimally quits experimenting at a lower belief when positive information acquisition is used than when it is not used. Finally, it is never optimal to engage in both learning and experimentation; that is, optimal learning requires suspension of experimentation. If learning must be combined with experimentation because the latter cannot be suspended, then the agent will be forced to forgo such combined learning even though the belief justifies using "pure" learning as the last resort.

Unlike positive information, negative information counters learning through experimenting. Failure to uncover negative evidence drives up the agent's belief. When the agent is already so pessimistic that he is about to abandon the risky arm there is little use in having the negative news, but when the agent is relatively optimistic negative information acquisition can be deployed as an insurance strategy to avoid potentially waiting too long to quit. Indeed, we show that the benefit of negative information over optimal experimentation is concave in the agent's belief, so the op-

timal use of negative information acquisition is an interval of intermediate beliefs. Since the agent's belief goes up when he chooses negative information acquisition and there is no negative news, while it goes down when he chooses experimentation but fails to achieve success, the switching point between the two strategies is characterized by a "perpetual learning policy" of alternating between the two in such a way that the belief becomes stationary, until either he achieves a success from the risky arm or quits upon the arrival of negative news. As is the case with positive information acquisition, optimal negative information acquisition requires suspension of experimentation. If negative information acquisition must be combined with experimentation, perpetual learning ceases to be a feature of optimality when negative news arrives at a lower rate than success through experimentation.

In recent independent work, Che and Mierendorff (2017) study the problem of an agent choosing between two possible actions with uncertain payoffs, who can delay the decision to acquire more information about the state, by allocating a fixed budget of "attention" to available information structures. Information acquisition is the only source of learning in their model. In our model, the agent must choose whether to continue with the risky alternative or abandon it. Experimenting with the risky arm is in itself informative, and this learning is augmented by the agent acquiring additional information at a cost. While the models are distinct, some of the insights that emerge from their analysis are similar to ours. Che and Mierendorff (2017) show that as an agent's belief gets closer to the point where he would stop to take one of the alternatives, this agent pursues "contradictory" news that would convince him to choose the other alternative. The logic of this result is the same as that of an agent in our model who chooses positive information acquisition as a last-ditch effort to resurrect experimentation before he abandons it.

Another related work is Moscarini and Smith (2001), who study a problem of an agent choosing between two alternatives. In their paper, learning is modeled as a continuous diffusion process, and they show that the agent increases the level of experimentation (adopts a more costly but more accurate diffusion process) as his belief gets closer to the two stopping thresholds. Although their conclusion appears to be similar to ours in the case of positive learning (the agent chooses positive information acquisition before he is about to quit), the logic behind their result is different. With a diffusion learning process, the agent's belief never jumps; and Moscarini and Smith (2001) show that the marginal benefit of this type of experimentation is

increasing in the value of the problem, which is highest prior to stopping and acting. In our paper, learning is the search for conclusive evidence about the state. The benefit of information in our paper derives from the convexity of the value function. Even though the value is lowest when the agent is about to quit, he has the greatest incentive to acquire positive information at this point for the chance that positive news may cause the value to jump up as the belief jumps to 1.

2. Direct Learning in Experimentation

Consider the following continuous-time bandit model. There is a single arm that yields uncertain returns to an agent. There are two states of the world. In the good state $\mathcal{G}$, the risky arm yields a "success" with a total prize π at a random time according to the exponential distribution with parameter $\lambda > 0$ so long as the agent experiments with it. In the bad state $\mathcal{B}$, the arrival rate of success is 0. The initial belief that the state is $\mathcal{G}$ is denoted as γ_0, and is assumed to be strictly between 0 and 1. Experimenting with the risky arm, or choosing X, has a flow cost $c > 0$. The agent stops experimentation and the control problem ends once he gets the prize π. For simplicity we assume that the agent does not discount.

There is a safe arm, with an arrival rate of "success" equal to 1 in both states. The prize upon "success" from the safe arm is normalized to 0, and the flow cost of choosing the safe arm is also 0. Choosing the safe arm, or choosing Q, is interpreted as quitting experimentation, although formally the control problem does not end until "success" arrives from the safe arm. We make the assumption that experimenting with the risky arm is worthwhile if the state is known to be $\mathcal{G}$:

$$\pi > c/\lambda. \tag{1}$$

Since $1/\lambda$ is the expected duration of achieving success through the risky arm conditional on state $\mathcal{G}$, c/λ is the conditional expected cost of success. Of course, quitting is optimal when the state is known to be $\mathcal{B}$.

In addition to experimenting with the risky arm, we allow the agent to engage in two types of "pure" information acquisition activities. At any point in time, instead of choosing X or Q, the agent may be allowed to choose to conduct "positive" information acquisition (P). If the agent chooses P for a small interval of time $\mathrm{d}t$, conclusive news about the true state arrives with probability $\alpha_P \mathrm{d}t$ for some $\alpha_P > 0$ in state $\mathcal{G}$, and no news arrives if the state is $\mathcal{B}$. The flow cost of positive information

acquisition is $k_P > 0$. Alternatively, the agent may be allowed to conduct "negative" information acquisition (N). If he chooses N for a small interval of time $\mathrm{d}t$, news that confirms the bad state arrives with probability $\alpha_N \mathrm{d}t$ for some $\alpha_N > 0$ if the state is truly $\mathcal{B}$, and no news arrives if the state is $\mathcal{G}$. The flow cost of negative information acquisition is $k_N > 0$.

There is no direct payoff from information acquisition, although the optimal policy is trivial after the agent receives conclusive news about the state. In positive information acquisition, after learning that the state is $\mathcal{G}$, the agent will experiment with the risky arm until a success, which by assumption (1) is optimal for the agent. In negative information acquisition, after learning that the state is $\mathcal{B}$, the agent will optimally quit. In extensions after the main analysis, we consider the case where the agent can choose either positive or negative information acquisition together with experimenting (Sections 3.4 and 4.4), and the case where the agent can acquire positive and negative information independently or jointly (Section 5.2).

If the agent can only choose between experimenting and quitting, our model is a simplified version of the exponential bandit problem introduced by Keller, Rady and Cripps (2005). Exponential bandit problems have become a major workhorse in dynamic game-theoretic models of learning since their contribution, because the potentially intractable problem of computing the Gittins index boils down to determining the optimal stopping time. Our model of information acquisition in experimentation may be thought of as a multi-arm bandit problem with correlated arms, where the agent is choosing between two risky arms—corresponding to experimentation only and pure information acquisition—and a safe option of quitting. The new risky arm is correlated with the original one, unlike in the standard multi-arm bandit problem solved by Gittins (1979), because the potential payoffs from them are determined by the same underlying state. Our paper contributes to the small economics literature on multi-armed bandit with correlated arms (Camargo, 2007; Klein and Rady, 2011).

3. Positive Information Acquisition

In this section, at any moment $t > 0$, the agent makes a choice among Q, X and P for an infinitesimal time interval $[t, t + \mathrm{d}t)$. The formal analysis allows the agent to mix between these three choices. We let $\sigma(t) = (\sigma_Q(t), \sigma_X(t), \sigma_P(t))$ to be the (non-negative) intensities for which these three choices are taken at time t, with $\sigma_Q(t) + \sigma_X(t) + \sigma_P(t) = 1$. However we show in the proof of Proposition 1 below

that the optimal control involves no mixing. Therefore, we sometimes abuse notation by also writing $\sigma(t) \in \{Q, X, P\}$ for the three pure-strategy choices. The entire closed-loop strategy is denoted as $\{\sigma(t)\}$.

The optimal control problem ends when success arrives from the risky arm (the time of this event is denoted T_X) or from the safe arm (the time of this event is denoted T_Q). If positive news arrives (the time of this event is denoted T_P), the belief that the state is $\mathcal{G}$ jumps to 1. In this event, assumption (1) ensures that it is optimal to keep choosing X until success arrives. The payoff from this continuation policy is $\pi - c/\lambda$. Even though the control problem does not literally stop when positive news arrives, we use $T = \min\{T_X, T_P, T_Q\}$ to denote the "stopping time" at which either success arrives from the safe or risky arm, or positive news arrives from the information acquisition process, whichever is earlier. We can write the agent's problem in positive information acquisition as choosing $\{\sigma(t)\}$ to maximize:

$$\mathbb{E}\left[\pi \mathbb{1}(T = T_X) + (\pi - c/\lambda)\mathbb{1}(T = T_P) - \int_0^T (\sigma_X(t)c + \sigma_P(t)k_P)\ \mathrm{d}t\right], \tag{2}$$

where the expectation is taken with respect to probability distribution of the stopping time τ given the agent's initial belief γ_0 and the strategy $\{\sigma(t)\}$:

$$\Pr[T > t] = \gamma_0 e^{-\int_0^t (\sigma_X(\tau)\lambda + \sigma_P(\tau)\alpha_P + \sigma_Q(\tau))\,\mathrm{d}\tau} + (1 - \gamma_0)e^{-\int_0^t \sigma_Q(\tau)\,\mathrm{d}\tau}.$$

Starting with any initial belief γ_0, a given strategy $\{\sigma(t)\}$ determines the evolution of the agent's belief. Fix any $t > 0$. Denote as $\gamma(t)$ the agent's belief that the state is $\mathcal{G}$ at time t, given that the decision problem is still on-going, i.e., prior to t success has not yet occurred from choosing X, and conclusive news has not yet arrived from choosing P. If $\sigma(t) = Q$, there is no belief updating. If $\sigma(t) = X$, the belief γ goes down at the rate of λ as the agent chooses X and no success occurs. Bayes' rule requires that the updated belief $\gamma(t)$ conditional on no success satisfies:

$$\frac{\mathrm{d}\gamma(t)}{\mathrm{d}t} = -\gamma(t)(1 - \gamma(t))\lambda. \tag{3}$$

If $\sigma(t) = P$, the belief evolution conditional on no news satisfies:

$$\frac{\mathrm{d}\gamma(t)}{\mathrm{d}t} = -\gamma(t)(1 - \gamma(t))\alpha_P. \tag{4}$$

Of course, the belief jumps to 1 if the risky arm yields success or if positive news arrives.

3.1. No direct learning

Before we characterize the solution to the problem of positive information acquisition (2), we illustrate the main intuition with a heuristic argument. To do so, we first consider the benchmark case where P is not available to the agent. The solution is a special case of our main result in this section, when the cost of positive information acquisition is prohibitively high.

Let $V(\gamma)$ be the value function for the control problem (2), and assume that $V(\cdot)$ is differentiable (a property which will be established in the formal proof of our main result). In the region of beliefs where X is optimal, the principle of dynamic programming gives, for small $\mathrm{d}t$:

$$V(\gamma) = -c\mathrm{d}t + \gamma\lambda\pi\mathrm{d}t + (1 - \gamma\lambda\mathrm{d}t)\left(V(\gamma) + V'(\gamma)\frac{\mathrm{d}\gamma}{\mathrm{d}t}\mathrm{d}t\right).$$

Using equation (3) for $\mathrm{d}\gamma/\mathrm{d}t$, we obtain the differential equation for the value function $V(\gamma)$:

$$\gamma(1-\gamma)\lambda V'(\gamma) = -c + \gamma\lambda\left(\pi - V(\gamma)\right). \tag{5}$$

The right-hand-side of (5) is the expected capital gain from success minus the flow cost of choosing X. Since quitting gives a payoff of 0, the optimal policy is given by a cutoff γ_{QX} such that the agent chooses Q for $\gamma \leq \gamma_{QX}$ and X otherwise. The value of γ_{QX} can be found by value-matching ($V'(\gamma_{QX}) = 0$) and smooth-pasting ($V(\gamma_{QX}) = 0$). This yields

$$\gamma_{QX} = \frac{c}{\lambda\pi},$$

which is strictly between 0 and 1 by assumption (1). The value function $V(\gamma)$ is 0 for $\gamma \leq \gamma_{QX}$, and is the solution to the differential equation (5) with boundary condition $V(\gamma_{QX}) = 0$ for $\gamma \geq \gamma_{QX}$. For future reference, we denote this value function as $U_X(\gamma)$.

Consider the change in expected payoff to a agent who gains access to free positive information for a small time interval of length $\mathrm{d}t$, before going back to his optimal experimentation policy in the benchmark case. Conditional on $\mathcal{G}$, the state is revealed to the agent with probability $\alpha_P\mathrm{d}t$, after which the agent optimally chooses X until success, with expected payoff $\pi - c/\lambda$. Absent the arrival of positive news, the new updated belief $\gamma(t + \mathrm{d}t)$ is given by (4). The change in the agent's payoff is thus

$$\gamma\alpha_P(\pi - c/\lambda)\mathrm{d}t + (1 - \gamma\alpha_P\mathrm{d}t)U_X(\gamma(t + \mathrm{d}t)) - U_X(\gamma).$$

For a first-order approximation of $U_X(\gamma(t+\mathrm{d}t))$ in the above expression, we use (5) for $\gamma > \gamma_{QX}$, and use $U_X(\gamma) = 0$ and $U_X'(\gamma) = 0$ for $\gamma \leq \gamma_{QX}$. The resulting expression is proportional to $\mathrm{d}t$ and to α_P, so we define the *benefit of positive information*, $B_P(\gamma)$, as the time rate of payoff change per unit of news arrival rate:

$$B_P(\gamma) = \begin{cases} (1-\gamma)c/\lambda & \text{if } \gamma > \gamma_{QX}, \\ \gamma(\pi - c/\lambda) & \text{if } \gamma \leq \gamma_{QX}. \end{cases}$$

If the agent is engaged in experimenting (i.e., $\gamma > \gamma_{QX}$), the benefit of positive information is decreasing in γ. The benefit is lower when the agent is more optimistic, because of a smaller change in the agent's belief when news from positive information acquisition arrives (from γ to 1). If the agent is not engaged in experimenting (i.e., $\gamma \leq \gamma_{QX}$), the benefit of positive information is increasing in γ, because it is simply the probability of receiving good news about the state, times the expected payoff from optimally experimenting forever in state $\mathcal{G}$. Thus, $B_P(\gamma)$ is maximized when the belief is γ_{QX} and the agent is just ready to quit. Define

$$B_P^* \equiv \max_{\gamma} B_P(\gamma).$$

As k_P/α_P is the expected cost of verifying the state conditional on $\mathcal{G}$, we use it as the (inverse) measure of the *efficiency of positive information acquisition*. We show in the next subsection that if $k_P/\alpha_P < B_P^*$, then P is optimal for some belief. Otherwise, the value function associated with the maximization problem (2) is identical to $U_X(\gamma)$.

The effects of the c/λ on the benefit of positive information depends on the current belief of the agent. When c/λ is higher, i.e., experimentation is less efficient, there are three forces acting on $B_P(\gamma)$: (i) experimentation becomes less informative than direct positive information acquisition; (ii) the agent quits earlier; and (iii) the payoff from experimentation is lower even if the state is known to be $\mathcal{G}$. When $\gamma \leq \gamma_{QX}$, the first two forces are irrelevant as the agent chooses Q for these beliefs. In this case, a less efficient experimentation technology lowers the benefit from positive information through the third force. When $\gamma > \gamma_{QX}$, the agent adjusts the threshold for quitting in such a way to balance the second and third force, leaving the first effect to dominate. In this case, a less efficient experimentation technology raises the benefit from positive information.

3.2. Optimal learning

Consider now the possibility of learning about the state. In the region of beliefs where P is optimal, the principle of dynamic programming gives, for small $\mathrm{d}t$:

$$V(\gamma) = -k_P \mathrm{d}t + \gamma\alpha_P(\pi - c/\lambda)\mathrm{d}t + (1 - \gamma\lambda\mathrm{d}t)\left(V(\gamma) + V'(\gamma)\frac{\mathrm{d}\gamma}{\mathrm{d}t}\mathrm{d}t\right).$$

Using equation (4) for $\mathrm{d}\gamma/\mathrm{d}t$, we obtain the differential equation for the value function $V(\gamma)$:

$$\gamma(1-\gamma)\alpha_P V'(\gamma) = -k_P + \gamma\alpha_P\left(\pi - c/\lambda - V(\gamma)\right), \tag{6}$$

where the right-hand-side is the expected capital gain from news arrival minus the flow cost of choosing P. Combining equations (5) for the option of choosing X and (6) for the options of P, together with the option of choosing Q (with expected capital gain minus flow cost equal to $-V(\gamma)$), the Hamilton-Jacobi-Bellman (HJB) equation for the value function $V(\cdot)$ corresponding to problem (2) is:

$$\begin{aligned} 0 = \max\big\{&-V(\gamma), -c + \gamma\lambda(\pi - V(\gamma)) - \gamma(1-\gamma)\lambda V'(\gamma), \\ &-k_P + \gamma\alpha_P(\pi - c/\lambda - V(\gamma)) - \gamma(1-\gamma)\alpha_P V'(\gamma)\big\}. \end{aligned} \tag{7}$$

The first term in the maximum operator on the right-hand-side represents the option of Q, the second term represents the option of X, and the third term P. For simplicity of exposition, we assumes that the agent chooses among the pure strategies Q, X and P to obtain the HJB equation (7). Our main result (Proposition 1) allows the agent to choose from general strategies of the form $(\sigma_Q(t), \sigma_X(t), \sigma_P(t))$. Equation (7) is without loss of generality, as we show that the optimal choice at each point in time is always a corner solution.

For $\gamma \in (0,1)$, define $D_X(\gamma)$ and $D_P(\gamma)$ such that

$$\begin{aligned} \gamma(1-\gamma)\lambda D_X(\gamma) &\equiv -c + \gamma\lambda\left(\pi - V(\gamma)\right), \\ \gamma(1-\gamma)\alpha_P D_P(\gamma) &\equiv -k_P + \gamma\alpha_P\left(\pi - c/\lambda - V(\gamma)\right). \end{aligned}$$

The above are essentially the same as (5) and (6). The function $D_X(\gamma)$ is simply $V'(\gamma)$ when X is optimal. Similarly, $D_P(\gamma)$ is equal to $V'(\gamma)$ when P is optimal. We can then rewrite (7) as

$$0 = \max\{-V(\gamma), \gamma(1-\gamma)\lambda(D_X(\gamma) - V'(\gamma)), \gamma(1-\gamma)\alpha_P(D_P(\gamma) - V'(\gamma))\}. \tag{8}$$

By a standard verification theorem (e.g., Oksendal and Sulem, 2005, Theorem 9.8), V is the value function for problem (2) if it is continuously differentiable and satisfies the HJB equation (8).

To find a candidate solution V for the HJB equation, observe from (8) that at a given belief γ, (i) if X is optimal, then $D_X(\gamma) = V'(\gamma) \geq D_P(\gamma)$; and (ii) if P is optimal, then $D_P(\gamma) = V'(\gamma) \geq D_X(\gamma)$. For a conjecture of a continuously differentiable V, we have that $D_X(\hat{\gamma}) = D_P(\hat{\gamma})$ at any belief $\hat{\gamma}$ where the optimal policy changes between X to P. Using the definitions of $D_X(\gamma)$ and $D_P(\gamma)$, this crossing-point, denoted γ_{PX}, is

$$\gamma_{PX} = 1 - \frac{k_P/\alpha_P}{c/\lambda}. \tag{9}$$

Furthermore, $D_P(\gamma) \geq D_X(\gamma)$ if and only if $\gamma \leq \gamma_{PX}$. We therefore conjecture an optimal policy described in case (ii) of Figure 1. At very high belief, the agent is confident about state $\mathcal{G}$ and chooses X. As success does not arrive and the agent becomes less optimistic, he switches to choose P when the belief reaches γ_{PX}. As no positive news arrives, he becomes more pessimistic and eventually chooses Q when the belief reaches γ_{QP}, where

$$\gamma_{QP} = \frac{k_P/\alpha_P}{\pi - c/\lambda} \tag{10}$$

is determined by $D_P(\gamma_{QP}) = 0$ and $V(\gamma_{QP}) = 0$. In Proposition 1 below, we show that such policy is optimal when positive information acquisition is relatively efficient. When P is relatively inefficient, it is never used and the optimal policy is the same as that described in Section 3.1, corresponding to case (i) of Figure 1.

Proposition 1. *Consider the model of positive information acquisition.*

(i) If $k_P/\alpha_P \geq B_P^*$, then the optimal policy is Q when $\gamma \leq \gamma_{QX}$, and X when $\gamma > \gamma_{QX}$;
(ii) if $k_P/\alpha_P < B_P^*$, then the optimal policy is Q when $\gamma \leq \gamma_{QP}$, P when $\gamma \in (\gamma_{QP}, \gamma_{PX}]$, and X when $\gamma > \gamma_{PX}$.

In case (i) of Proposition 1, information acquisition is so inefficient that the agent's optimal policy is the same as when positive information is unavailable. The quitting belief is γ_{QX}. When information acquisition is sufficiently efficient relative to the maximum benefit, we have case (ii) and positive information acquisition becomes optimal for some beliefs. It is easy to verify that $\gamma_{QP} < \gamma_{QX}$ in this case, and so the use of P raises the value function and enables the agent to quit at a more

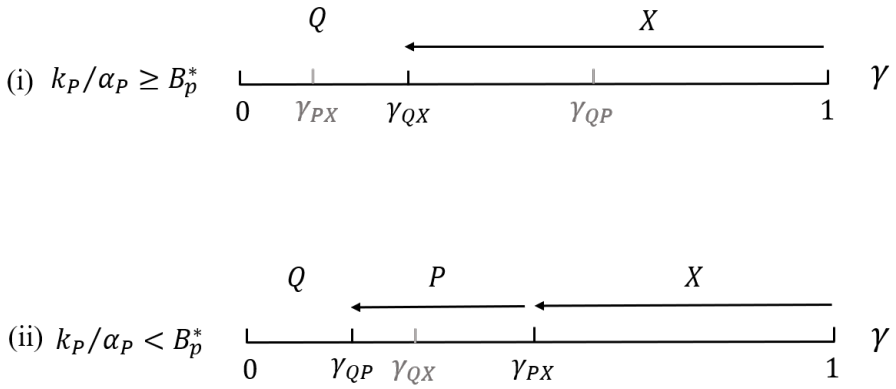

Figure 1. *Optimal policy in the positive information acquisition model. The direction of the arrows indicate how the belief evolves when the risky arm brings no success and information acquisition brings no news.*

pessimistic belief about the state. Furthermore, $\gamma_{QX} < \gamma_{PX}$ in this case, implying that the positive information acquisition region contains the belief that maximizes $B_P(\gamma)$. As we have seen from the discussion of the benefit of positive information, positive information acquisition is used as a last-ditch effort before abandoning the risky arm permanently. Consistent with the fact that $B_P(\gamma)$ decreases in γ whenever the agent experiments with the risky arm, the agent optimally refrains from positive information acquisition if he is sufficiently optimistic about the state.

The result that positive information acquisition is optimally chosen when the agent is relatively pessimistic about state $\mathcal{G}$ holds despite the fact that the agent expects that it is unlike to actually find good news when his belief is low.[2] The main reason is that information is useful only to the extent that it can potentially alter an agent's decision. When the agent is about to quit, positive information acquisition has the greatest potential to correct what would otherwise be a wrong decision, which is why the benefit of positive information is greatest at the quitting point. Calvert (1985) and Suen (2004) make a similar point in the context of static decision making, but the same logic extends to our dynamic learning model.

[2]Observe that γ_{QP} is strictly greater than 0 as long as k_P/α_N is positive. The agent never chooses positive information acquisition when his belief is very low, because he does not expect to find any news in that case.

3.3. Optimal use of direct learning

It is straightforward to see from the formulas for γ_{PX} and γ_{QP} (equations (9) and (10)) that a more efficient technology of positive learning (i.e., a decrease in k_P/α_P) expands the *optimal use of direct learning*, i.e., the range of beliefs $(\gamma_{QP}, \gamma_{PX}]$ for which P is chosen in case (ii) of Proposition 1. Nevertheless, for any $k_P/\alpha_P > 0$, the agent optimally chooses experimentation instead of direct learning if the belief about state $\mathcal{G}$ is sufficiently close to 1. Although learning is cheap and the chance of finding good news is high, good news does not alter the agent's decision to engage the risky arm when he is very optimistic. The agent therefore optimally chooses X, and delays exercising the option to use P until his belief gets lower if success does not arrive.

An increase in the prize π from success raises B_P^* and hence makes it more likely that P is chosen under the optimal policy. Moreover, γ_{QP} decreases in π while γ_{PX} does not depend on π. Recall that $\gamma_{QP} < \gamma_{QX}$ in case (ii) of Proposition 1. This means that an increase in the reward increases the optimal use of positive information acquisition by expanding the region of beliefs for which P is used as the last resort to resurrect risky experimentation. On the other hand, since the agent still has to engage in experimentation if his belief jumps to 1 upon the arrival of good news, the increase in π dose not affect the choice between direct learning about the state and risky experimentation when he is sufficiently optimistic.

Finally, it is straightforward to incorporate discounting into the optimal control problem (2). If the discount rate is $\rho > 0$, the HJB equation becomes:

$$0 = \max\{-V(\gamma) - \rho V(\gamma), \gamma(1-\gamma)\lambda(D_X(\gamma) - V'(\gamma)), \gamma(1-\gamma)\alpha_P(D_P(\gamma) - V'(\gamma))\},$$

where

$$\begin{aligned}\gamma(1-\gamma)\lambda D_X(\gamma) &\equiv -c + \gamma\lambda\left(\pi - V(\gamma)\right) - \rho V(\gamma),\\ \gamma(1-\gamma)\alpha_P D_P(\gamma) &\equiv -k_P + \gamma\alpha_P\left(V(1) - V(\gamma)\right) - \rho V(\gamma).\end{aligned}$$

With discounting, we have $V(1) = (\lambda\pi - c)/(\lambda + \rho)$, which decreases with ρ. Since the agent still has to pursue experimentation and wait for success to arrive to obtain the reward even if he knows that the state is good, discounting reduced the capital gain resulting from obtaining good news. This reduces the value of choosing P relative to the other two options. When positive information acquisition is sufficiently efficient so that P is optimally used, the quitting belief γ_{QP} is determined by

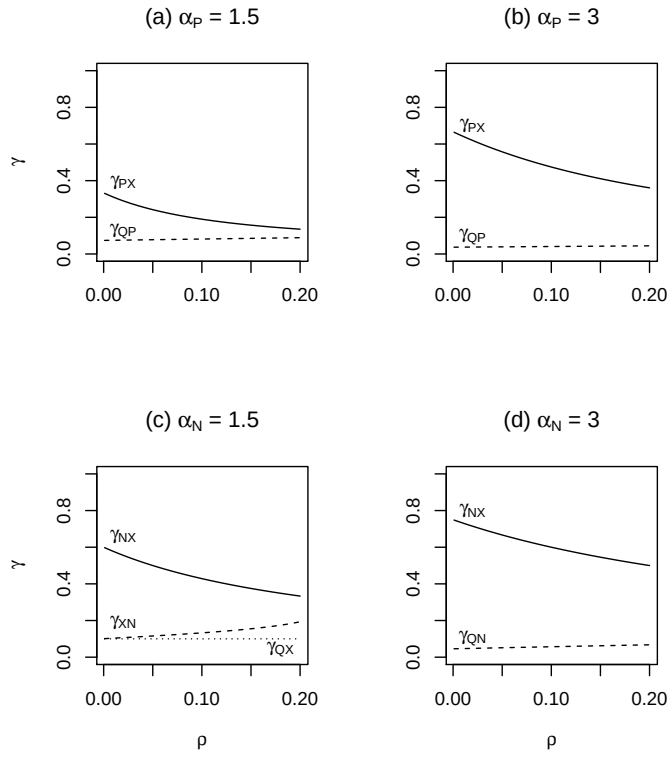

Figure 2. *In panels (a) and (b), P is optimally used in the positive information acquisition model when $\gamma \in (\gamma_{QP}, \gamma_{PX}]$. Greater discounting reduces the region of beliefs for which P is optimally used. Panels (c) and (d) refer to the negative information acquisition model. In panel (c) N is chosen when $\gamma \in (\gamma_{XN}, \gamma_{NX}]$; in panel (d) N is chosen when $\gamma \in (\gamma_{QN}, \gamma_{NX}]$. In both panels, greater discounting reduces the region of beliefs for which N is optimally used. The values of other parameters used in this illustration are: $\pi = 10$, and $\lambda = c = k_P = k_N = 1$.*

$D_P(\gamma) = 0$ and $V(\gamma) = 0$. This gives $\gamma_{QP} = k_P/(\alpha_P V(1))$, so discounting causes the agent to quit earlier. Letting $U_P(\gamma)$ be the solution to the differential equation $D_P(\gamma) = V'(\gamma)$, with boundary condition $V(\gamma_{QP}) = 0$, the upper boundary of the optimal use of positive information is determined by the belief γ_{PX} that satisfies $D_P(\gamma) = D_X(\gamma)$ at $V(\gamma) = U_P(\gamma)$. When $\lambda = \alpha_P$, for any U_P an increase in ρ has the same effect on D_X and D_P apart from reducing $V(1)$ in D_P, so γ_{PX} decreases. In general, however, the equation for γ_{PX} depends on U_P and does not admit an explicit solution, but a numerical analysis confirms that γ_{PX} decreases in ρ. In panels (a) and (b) of Figure 2, we illustrate that the region of positive information acquisition, $(\gamma_{QP}, \gamma_{PX}]$, shrinks when the agent becomes more impatient.

3.4. Combined learning

We have assumed that the agent cannot choose positive information acquisition and experimentation at the same time. Recall that positive information acquisition is a pure information activity, and even after it yields conclusive news about the state, the agent still has to incur the cost of experimentation in order to deliver success. A natural question is then whether it would be optimal for the agent to combine positive information acquisition with experimentation.

To address the above question, we assume that in addition to Q, X and P, the agent can also choose X and P together as a separate option. Denote this choice as $P\dagger$. When the agent makes this choice, the belief in the absence of either success or positive news goes down at the rate of $\lambda + \alpha_P$. Defining $D_{P\dagger}(\gamma)$ according to

$$\gamma(1-\gamma)(\lambda+\alpha_P)D_{P\dagger}(\gamma) \equiv -c - k_P + \gamma(\lambda\pi + \alpha_P(\pi - c/\lambda)) - \gamma(\lambda+\alpha_P)V(\gamma),$$

we can write the HJB equation in this case as

$$\begin{aligned} 0 = \max\{&-V(\gamma), \gamma(1-\gamma)\lambda(D_X(\gamma) - V'(\gamma)), \gamma(1-\gamma)\alpha_P(D_P(\gamma) - V'(\gamma)), \\ &\gamma(1-\gamma)(\lambda+\alpha_P)(D_{P\dagger}(\gamma) - V'(\gamma))\}. \end{aligned} \tag{11}$$

Proposition 2. *Suppose $\sigma(t) \in \{Q, X, P, P\dagger\}$. Under the optimal policy, there is no interval of beliefs such that $\sigma(t) = P\dagger$ when $\gamma(t)$ belongs to that interval.*

The key to Proposition 2 is that $D_{P\dagger}$ is just a linear combination of D_P and D_X. Whenever P is strictly preferred to X, P is also strictly preferred to $P\dagger$. Whenever X is strictly preferred to P, X is also strictly preferred to $P\dagger$. Intuitively, given the linearity inherent in the optimization problem, combining P with X cannot improve on both of them. An immediate implication of Proposition 2 is that optimal positive information acquisition requires the agent to suspend experimentation.

The above result shows that adding the option of $P\dagger$ does not change optimal information acquisition when P is feasible. However, in some applications of our model P may not be feasible unless it is combined with X.[3] The question is then whether the optimal policy with regards to $\{Q, X, P\dagger\}$ will be different from the one

[3]In the motivating example of R&D mentioned in the introduction, for legal reasons or industrial relations, the firm may not be able to suspend product development by its engineers in order to bring in industry or academic scientists to engage in background research about the product.

with regards to $\{Q, X, P\}$. It is obvious from the discussion of Proposition 2 that $P\dagger$ will never be chosen when $k_P/\alpha_P \geq B_P^*$, and so the optimal policy remains the same as in case (i) of Proposition 1. When $k_P/\alpha_B < B_P^*$, the optimal policy will be different but has a similar structure as in case (ii) of Proposition 1—choose Q when $\gamma \leq \gamma_{QP\dagger}$; choose $P\dagger$ when $\gamma \in (\gamma_{QP\dagger}, \gamma_{P\dagger X}]$; and choose X when $\gamma > \gamma_{P\dagger X}$.

Once again, since $D_{P\dagger}$ is a linear combination of D_P and D_X, the marginal benefit of $P\dagger$ over X is the same as P over X, implying that $\gamma_{P\dagger X} = \gamma_{PX}$. However, the marginal benefit of $P\dagger$ over Q is smaller than that of P over X. Formally, since $\gamma_{QP} < \gamma_{PX}$ and $D_X(\gamma_{QP}) < D_P(\gamma_{QP}) = 0$ when $k_P/\alpha_P < B_P^*$, we have

$$(\lambda + \alpha_P) D_{P\dagger}(\gamma_{QP}) = \lambda D_X(\gamma_{QP}) + \alpha_P D_P(\gamma_{QP}) < 0.$$

This implies that $\gamma_{QP\dagger} > \gamma_{QP}$, meaning that the agent would quit strictly before the belief reaches γ_{QP} if only $\{Q, X, P\dagger\}$ is available. Put differently, at belief γ_{QP}, the probability of finding news that confirms state $\mathcal{G}$ is too low to justify the cost if positive information acquisition has to be accompanied by experimenting with the risky arm. However, pure positive information acquisition can still be justified despite the long odds because it is relatively cheap without experimentation. This result is consistent with the earlier claim that optimal positive information acquisition requires the agent to suspend experimentation.

4. Negative Information Acquisition

In this section, at any moment $t > 0$, the agent chooses $\sigma(t) = (\sigma_Q, \sigma_X, \sigma_N)$ in the unit simplex for an infinitesimal time interval $[t, t + \mathrm{d}t)$. We focus on pure strategy choice in the exposition, and sometimes write $\sigma(t) \in \{Q, X, N\}$, although the formal result (Proposition 3 below) allows mixing between these strategies.

When negative information arrives (the time of this event is denoted T_N), the belief jumps to 0. The optimal continuation policy is to choose Q until success arrives from the safe arm (with a payoff of 0). Let $T = \min\{T_X, T_N, T_Q\}$ be the "stopping time," i.e., the time of arrival of success from the risky arm of safe arm, or the time of arrival of negative news, whichever is earlier. We can write the agent's optimal control problem in negative information acquisition as choosing $\{\sigma(t)\}$ to maximize

$$\mathbb{E}\left[\pi \mathbb{1}(T = T_X) - \int_0^T \left(\sigma_X(t)c + \sigma_N(t)k_N\right) \mathrm{d}t\right], \tag{12}$$

where the expectation is taken with respect to the probability distribution of the stopping time τ given the agent's initial belief γ_0 and the strategy $\{\sigma(t)\}$:

$$\Pr[T > t] = \gamma_0 e^{-\int_0^t (\sigma_X(\tau)\lambda + \sigma_Q(\tau))\,\mathrm{d}t} + (1-\gamma_0) e^{-\int_0^t (\sigma_N(\tau)\alpha_N + \sigma_Q(\tau))\,\mathrm{d}\tau}.$$

The belief updating after the agent chooses Q or X is the same as in Section 3. If $\sigma(t) = N$, the updated belief $\gamma(t)$ conditional on the absence of negative news satisfies:

$$\frac{\mathrm{d}\gamma(t)}{\mathrm{d}t} = \gamma(t)(1-\gamma(t))\alpha_N. \tag{13}$$

Thus, in contrast to experimentation and positive information acquisition, the belief that the state is $\mathcal{G}$ goes up at the rate of α_N as the agent chooses N and no negative news arrives. This sign change is critical to the contrasting results we obtain below for negative information acquisition.

For $\gamma \in (0,1)$, define $D_N(\gamma)$ according to

$$-\gamma(1-\gamma)\alpha_N D_N(\gamma) \equiv -k_N - (1-\gamma)\alpha_N V(\gamma).$$

The right-hand-side of the above is the expected capital loss from arrival of negative news minus the flow cost of negative information acquisition. The function $D_N(\gamma)$ is simply $V'(\gamma)$ when N is optimal. Whenever the derivative of the value function exists, the HJB equation for problem (12) is given by

$$0 = \max\left\{-V(\gamma), \gamma(1-\gamma)\lambda(D_X(\gamma) - V'(\gamma)), \gamma(1-\gamma)\alpha_N(V'(\gamma) - D_N(\gamma))\right\}. \tag{14}$$

The first term in the maximum operator on the right-hand-side represents the option of Q, the second term represents the option of X, and the third term N.

4.1. Benefit of learning

As in positive information acquisition, we first derive an expression for the *benefit of negative information*, $B_N(\gamma)$. Consider the change in the expected payoff to the agent who gains access to free negative information for a small time interval of length $\mathrm{d}t$ before going back to his optimal experimentation policy in the benchmark of no direct learning given in Section 3.1. Conditional on $\mathcal{B}$, with probability $\alpha_N \mathrm{d}t$ the agent learns the state and quits. In the absence of the news, the updated belief follows (13). The change in the payoff is given by:

$$(1 - (1-\gamma)\alpha_N \mathrm{d}t) U_X(\gamma(t + \mathrm{d}t)) - U_X(\gamma).$$

We use equation (5) for a first-order approximation of $U_X(\gamma(t+\mathrm{d}t))$. The resulting expression for the payoff change is again proportional to $\mathrm{d}t$ and to α_N, so we define $B_N(\gamma)$ to be the time rate of payoff change per unit of news arrival rate, given by

$$B_N(\gamma) = \begin{cases} \gamma\pi - c/\lambda - U_X(\gamma) & \text{if } \gamma > \gamma_{QX}, \\ 0 & \text{if } \gamma \leq \gamma_{QX}. \end{cases}$$

Since U_X is convex, the benefit of negative information is concave for $\gamma > \gamma_{QX}$. This means that negative information is particularly valuable when the agent has intermediate beliefs about the state. Define

$$B_N^* \equiv \max_\gamma B_N(\gamma).$$

Correspondingly, define $\gamma_N^* \equiv \arg\max_\gamma B_N(\gamma)$, which is uniquely determined by the first order condition, $U_X'(\gamma_N^*) = \pi$. Unlike positive information, the benefit of negative information is 0 when the agent is about to quit, because learning that the state is $\mathcal{B}$ does not change the agent's decision. As the agent becomes more optimistic, the benefit initially increases, but it eventually decreases because negative information acquisition is unlikely to generate any news when the state is likely to be $\mathcal{G}$. Thus, the benefit of negative information stems from its use as an insurance strategy to avoid wasteful experimentation when the agent is still optimistic about the risky arm but the state is actually $\mathcal{B}$.

The benefit of negative information $B_N(\gamma)$ can be related to the benefit of positive information $B_P(\gamma)$ defined earlier. Observe that

$$B_P(\gamma) + B_N(\gamma) = \gamma U_X(1) + (1-\gamma)U_X(0) - U_X(\gamma).$$

Thus, the total benefit of positive and negative information is equal to the value of an experiment that reveals the true state in the benchmark model of Section 3.1 with experimentation only. See Figure 3 for an illustration. Because $U_X(1) = \pi - c/\lambda$ and $U_X(\gamma_{QX}) = 0$, the term $\gamma\pi - c/\lambda$ in the definition of $B_N(\gamma)$ is shown by the chord between $U_X(1)$ and $U_X(\gamma_{QX})$ in Figure 3. The distance between this chord and $U_X(\gamma)$ is the benefit of negative information $B_N(\gamma)$, and the distance between this chord and the value of fully-revealing experiment $\gamma U_X(1) + (1-\gamma)U_X(0)$ is the benefit of positive information $B_P(\gamma)$.

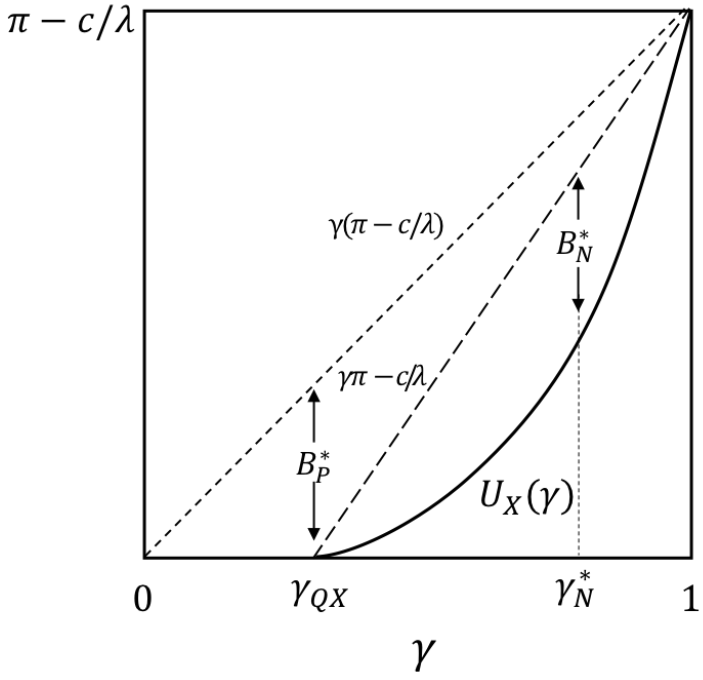

Figure 3. *The benefit of negative information is the distance between the dashed line and $U_X(\gamma)$. This benefit is maximized at γ_N^*. The benefit of positive information is the distance between the dashed line and the dotted line. This benefit is maximized at γ_{QX}.*

4.2. Perpetual learning

In this subsection, we provide a characterization of the optimal policy under negative information acquisition (Proposition 3 below). This is achieved by first using the HJB equation (14) as our guide to conjecture a candidate policy. We use k_N/α_N as the (inverse) measure of the *efficiency of negative information acquisition*, as it represents the expected cost of verifying the state conditional on $\mathcal{B}$. Similar to the case of positive information acquisition, the definition of $B_N(\gamma)$ implies that N is optimal for some beliefs when $k_N/\alpha_N < B_N^*$.

Naturally, we conjecture that X is optimal when the agent's belief γ is sufficiently high. To identify the highest switching point between N and X, we make use of the fact that belief updating operates in opposite directions under N and under X. In particular, because the belief goes up if the agent chooses N and no bad news arrives, while the belief goes down if the agent chooses X and no success arrives,

by alternating between choosing N for an interval of length λdt and choosing X for a small interval $\alpha_N dt$, the agent keeps the belief stationary. This corresponds to the policy: $\sigma_N = \lambda/(\alpha_N + \lambda)$, $\sigma_X = \alpha_N/(\alpha_N + \lambda)$, and $\sigma_Q = 0$. We call this a *perpetual learning policy*, and denote the payoff from such policy $U_S(\gamma)$. It satisfies:

$$U_S(\gamma) = -c\alpha_N dt - k_N \lambda dt + \gamma\lambda\pi\alpha_N dt + (1 - \gamma\lambda\alpha_N dt - (1-\gamma)\lambda\alpha_N dt)\, U_S(\gamma),$$

and hence

$$U_S(\gamma) = \gamma\pi - c/\lambda - k_N/\alpha_N.$$

The payoff $U_S(\gamma)$ is linear in γ, and is shown in Figure 4.[4] The switching point between N and X, denoted as γ_{NX}, is an absorbing point by construction. We have

$$V(\gamma_{NX}) = U_S(\gamma_{NX}),$$

which is a value-matching condition. Furthermore, since the perpetual policy is feasible, optimality of N just below γ_{NX} implies that $V(\gamma) \geq U_S(\gamma)$ for γ slightly below γ_{NX}. Similarly, $V(\gamma) \geq U_S(\gamma)$ for γ slightly above γ_{NX}. Together they imply

$$\lim_{\gamma \uparrow \gamma_{NX}} D_N(\gamma) \leq U_S{}'(\gamma) = \pi \leq \lim_{\gamma \downarrow \gamma_{NX}} D_X(\gamma).$$

The HJB equation (14), however, requires that $D_N(\gamma) \geq V'(\gamma) \geq D_X(\gamma)$ for γ in the neighborhood of γ_{NX}. Thus, we must have

$$D_N(\gamma_{NX}) = \pi = D_X(\gamma_{NX}),$$

implying smooth pasting at the absorbing point. Together with the value-matching condition at γ_{NX}, we can solve for γ_{NX}:

$$\gamma_{NX} = \frac{c/\lambda}{k_N/\alpha_N + c/\lambda}. \tag{15}$$

This is highest value of γ for which N is optimal in the conjectured policy. Value matching and smooth pasting at the absorbing point γ_{NX} are illustrated in Figure 4.

[4]The expression can be easily understood after noting that, under the perpetual learning policy, the expected duration for news arrival in state $\mathcal{B}$ is the same as the expected duration for the arrival of success in state $\mathcal{G}$, both given by $1/(\alpha_N\lambda)$. Thus the expected cost to the agent is just the flow cost of experimentation and information acquisition ($k_N\lambda + c\alpha_N$) times the state-independent expected duration $1/(\alpha_N\lambda)$.

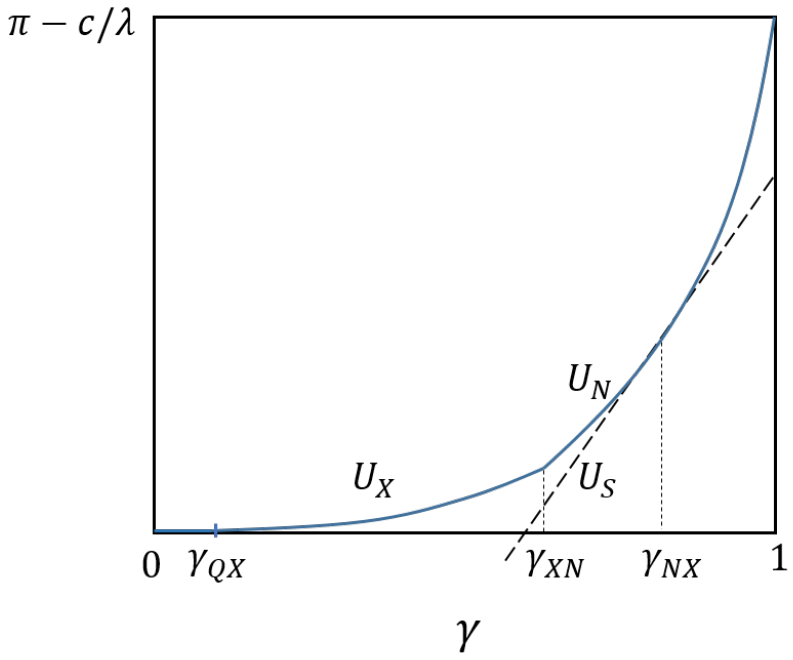

Figure 4. *The dashed line is the payoff $U_S(\gamma)$ from the perpetual learning policy. The value function is equal to U_X for $\gamma \in [0, \gamma_{XN}]$ and to U_N for $\gamma \in (\gamma_{XN}, \gamma_{NX}]$. It is tangent to U_S at the absorbing point γ_{NX}, and has a convex kink at the tipping point γ_{XN}.*

For the lowest value of γ for which N is optimal, we conjecture that it is determined by value matching alone. Let $U_N(\gamma)$ solve the differential equation:

$$\gamma(1-\gamma)U_N'(\gamma) = k_N/\alpha_N + (1-\gamma)U_N(\gamma), \tag{16}$$

with boundary condition $U_N(\gamma_{NX}) = U_S(\gamma_{NX})$. There can be only one intersection between U_N and U_X below γ_{NX}: at γ_{XN} such that $U_X(\gamma_{XN}) = U_N(\gamma_{XN})$, we have[5]

$$\gamma_{XN}(1-\gamma_{XN})(U_N'(\gamma_{XN}) - U_X'(\gamma_{XN})) = U_N(\gamma_{XN}) - U_S(\gamma_{XN}) > 0,$$

where the inequality follows because U_N is strictly convex while U_S is linear, and the two functions are tangent to each another at γ_{NX}. As the switching point between X and N, the intersection γ_{XN} is a tipping point, because the belief goes up for $\gamma > \gamma_{XN}$ while N is used and it goes down for $\gamma < \gamma_{XN}$ while X is used. The value function, given by $U_X(\gamma)$ for $\gamma \leq \gamma_{XN}$ and $U_N(\gamma)$ for $\gamma \in (\gamma_{XN}, \gamma_{NX}]$, does not satisfy smooth pasting at the tipping point, as it kinks "up." See Figure 4.

[5]The expression below assumes that $\gamma_{XN} > \gamma_{QX}$; the intersection is the switching point between Q and N, and we have a convex kink at the switching point trivially.

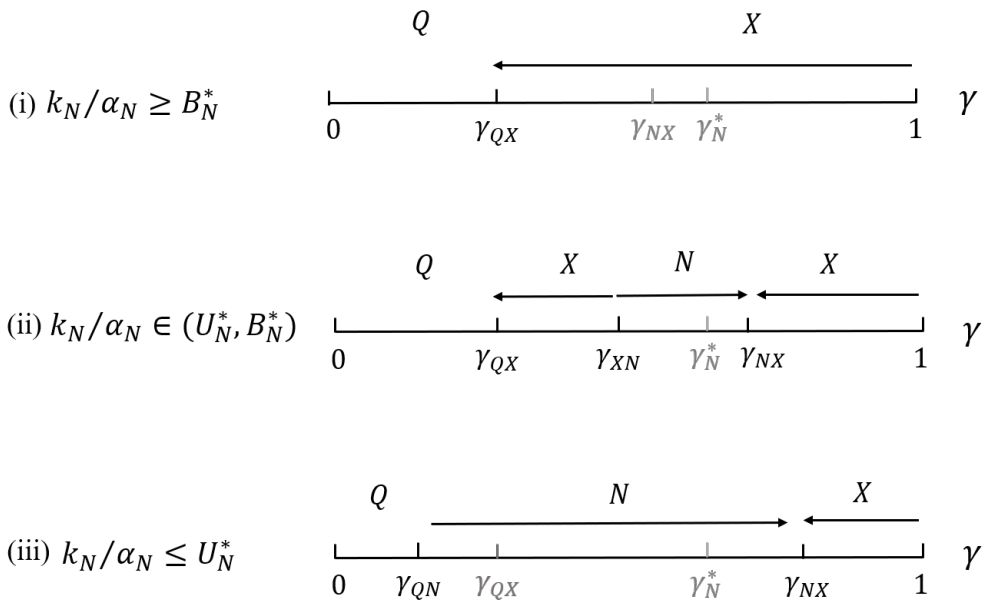

Figure 5. *Optimal policy under negative information. The direction of the arrows indicate how the belief evolves when the risky arm brings no success and information acquisition brings no news. In both case (ii) and case (iii), the belief at γ_{NX} is an absorbing point. The belief at γ_{XN} in case (ii) and the belief at γ_{QN} in case (iii) are tipping point.*

Proposition 3 below provides the characterization of the unique optimal policy in the negative information acquisition model. The proof involves constructing the value function $V(\gamma)$ using the candidate policy and show that it satisfies the HJB equation (14). The resulting V is not differentiable at a tipping point, but since it has a convex kink, we show that it is a viscosity solution to (14), and thus corresponds to the solution to the original problem (12) (see, e.g., Oksendal and Sulem, 2005).

Proposition 3. *Consider the model of negative information acquisition. There is a unique $U_N^* \in (0, B_N^*)$ such that*

(i) if $k_N/\alpha_N \geq B_N^*$, the optimal policy is Q when $\gamma \leq \gamma_{QX}$, and X when $\gamma > \gamma_{QX}$;
(ii) if $k_N/\alpha_N \in (U_N^*, B_N^*)$, then there exists γ_{XN} such that the optimal policy is Q when $\gamma \leq \gamma_{QX}$, X when $\gamma \in (\gamma_{QX}, \gamma_{XN}]$ and $\gamma \geq \gamma_{NX}$, and N when $\gamma \in (\gamma_{XN}, \gamma_{NX}]$;
(iii) if $k_N/\alpha_N \leq U_N^*$, then there exists $\gamma_{QN} < \gamma_{QX}$ such that the optimal policy is Q when $\gamma \leq \gamma_{QN}$, N when $\gamma \in (\gamma_{QN}, \gamma_{NX}]$, and X when $\gamma \geq \gamma_{NX}$.

Figure 5 illustrates the three cases of the optimal policy. Due to the inefficiency of negative information acquisition, in case (i) the optimal policy is the same as when N is unavailable, and is identical to that for the benchmark model with experimentation only. When negative information acquisition is sufficiently efficient relative to the maximum benefit of information B_N^*, it becomes optimal to choose N for some beliefs as the payoff function $U_S(\gamma)$ from the perpetual policy rises above the value function $U_X(\gamma)$ in case (i). In both case (ii) and case (iii) of Proposition 3, the region of beliefs where N is optimally chosen contains γ_N^*, the belief that maximizes $B_N(\gamma)$.[6] As we have seen from the discussion of the benefit of information, negative information acquisition is used as an insurance strategy to avoid unnecessary experimentation. This is achieved by replacing X with N for an interval of beliefs around γ_N^*.

In case (ii) of Proposition 3, the value function is strictly higher than $U_X(\gamma)$ for γ in the negative information acquisition region, $(\gamma_{XN}, \gamma_{NX}]$, while X is chosen for $\gamma < \gamma_{XN}$. In case (iii), when negative information acquisition is very efficient, the value function is strictly higher than $U_X(\gamma)$ whenever the latter is positive. In this case, the agent chooses N at a belief just higher than the quitting belief γ_{QN}. However, negative information should *not* be interpreted as a last-ditch effort before abandoning the risky arm. The agent quits immediately if the starting belief is just below γ_{QN}; above γ_{QN}, negative information acquisition drives the agent's belief up to γ_{NX} (unless bad news is found), at which point the agent optimally adopts the perpetual learning policy, and never quits until either the state is revealed or success arrives from the risky arm.

4.3. Optimal use of negative information

By Proposition 3, the *optimal use* of negative information, i.e., the region of beliefs for which N is optimally chosen, is an interval. It is given by $(\gamma_{XN}, \gamma_{NX}]$ in case (ii), and $(\gamma_{QN}, \gamma_{NX}]$ in case (iii).

It can be readily observed from equation (15) that γ_{NX} increases when k_N/α_N falls. Moreover, because a fall in k_N/α_N raises $U_N(\gamma)$ but leaves $U_X(\gamma)$ unchanged, the intersection of these curves (i.e., γ_{XN} or γ_{QN}, depending on whether case (ii) or

[6] By construction, U_X is tangent to U_S at $\gamma_N^* = \gamma_{NX}$ when $k_N/\alpha_N = B_N^*$. When k_N/α_N decreases, we have $\gamma_{NX} > \gamma_N^*$ in both case (ii) and (iii). At the same time, we have $\gamma_{XN} < \gamma_N^*$ in case (ii) because $U_X'(\gamma_{XN}) < U_N'(\gamma_{XN}) < U_N'(\gamma_{NX}) = \pi = U_X'(\gamma_N^*)$, and $\gamma_{QN} < \gamma_N^*$ in case (iii) because $U_X'(\gamma_{QN}) = 0 < \pi = U_X'(\gamma_N^*)$.

(iii) applies) shifts to the left. Thus, a more efficient technology of negative learning expands the range of beliefs for which N is optimally used. But even though negative information acquisition may be very efficient, as long as k_N/α_N is positive, the agent optimally chooses X instead of N when his belief about state $\mathcal{G}$ is sufficiently close to 1. The reason is not that negative news would not change his decision, but rather the chance of finding any negative news is too low.

An increase in the prize π from success of the risky arm raises both $U_N(\gamma)$ and $U_X(\gamma)$. Solving the relevant differential equations, we find that[7]

$$\frac{\partial U_N(\gamma_{XN})}{\partial \pi} = \gamma_{XN} > \frac{\partial U_X(\gamma_{XN})}{\partial \pi}.$$

Since a higher π raises $U_N(\gamma)$ more than it raises $U_X(\gamma)$, the intersection point γ_{XN} moves to the left. On the other hand, γ_{NX} does not depend on π. Thus, a higher reward π expands the optimal use of negative information $(\gamma_{XN}, \gamma_{NX}]$. The intuition is that choosing N at γ_{XN} can guarantee success under state $\mathcal{G}$ (because the agent's belief would go up to γ_{NX} and he then switches to a perpetual learning policy which mixes between N and X), while choosing X at γ_{XN} does not guarantee success (because the agent may quit if the belief goes below γ_{QX}). If the agent is indifferent between N and X at γ_{XN} for some value of π, he strictly prefers N to X at γ_{XN} when π increases.

It is also straightforward to extend the model of negative information acquisition to allow for a positive discount rate ρ. Discounting does not change the perpetual learning policy, but lowers the payoff function $U_S(\gamma)$. Using value matching and smooth pasting between $U_X(\gamma)$ and $U_S(\gamma)$ to obtain an explicit formula for γ_{NX}, we can verify that it decreases as ρ increases. Thus, the upper boundary of the negative information acquisition region decreases. A higher ρ also lowers both $U_N(\gamma)$ and $U_X(\gamma)$. When negative information acquisition is sufficiently efficient and the optimal policy is given by case (iii) of Proposition 3, only the effect on $U_N(\gamma)$ of

[7]The explicit solution $U_N(\gamma)$ is provided in equation (21) in the appendix. If the state is $\mathcal{B}$, an increase in π has no effect on the payoff from choosing N. If the state is $\mathcal{G}$, the agent chooses N and then switches to the perpetual learning policy and eventually obtains success from the risky arm. A unit increase in π raises his payoff by one unit. Thus, γ_{XN} is the agent's expected increase in payoff when his belief is γ_{XN}. On the other hand, if the agent is choosing X at γ_{XN}, a unit increase in π causes $U_X(\gamma_{XN})$ to increase by less than one unit, because, even in state $\mathcal{G}$, it is possible that the agent may quit before obtaining success.

a marginal increase in ρ is relevant, and so the lower boundary of the negative information acquisition region γ_{QN} increases. If we have case (ii) instead, numerical analysis suggests that it lowers $U_N(\gamma)$ more than it lowers $U_X(\gamma)$, and thus the crossing point γ_{XN} shifts to the right as ρ increases.[8] A less patient agent tends to make less use of negative information. The numerical examples shown in panels (c) and (d) of Figure 2 illustrate this point.

4.4. Fast and slow learning

If the agent combines negative information acquisition with experimentation, the absence of either negative news from N or success from X implies that the updated belief can either go up or down, depending on the comparison between α_N and λ. Denote the combined choice of N and X as $N\dagger$. We refer to the case of upgrading belief ($\alpha_N > \lambda$) as "fast learning," and the case of downgrading belief ($\alpha_N < \lambda$) as "slow learning." For each $\gamma \in (0,1)$, define $D_{N\dagger}(\gamma)$ according to

$$\gamma(1-\gamma)(\alpha_N - \lambda)D_{N\dagger}(\gamma) \equiv c + k_N - \gamma\lambda\pi + (\gamma\lambda + (1-\gamma)\alpha_N)V(\gamma).$$

Since $D_{N\dagger}(\gamma)$ is a linear combination of $D_N(\gamma)$ and $D_X(\gamma)$, as in Section 3.4, combined learning is never strictly optimal when the agent can separate negative information acquisition from experimentation, whether learning is fast or slow.

However, in a model where the agent must not suspend experimentation in order to acquire negative information, fast learning and slow learning have qualitatively different optimal policies. Formally, suppose that the agent can choose one from $\{Q, X, N\dagger\}$ at any moment t. Given the definition of $D_{N\dagger}(\gamma)$ above, we can write the HJB equation as

$$\begin{aligned} 0 = \max\big\{ & -V(\gamma), \gamma(1-\gamma)\lambda(D_X(\gamma) - V'(\gamma)), \\ & \gamma(1-\gamma)(\alpha_N - \lambda)(V'(\gamma) - D_{N\dagger}(\gamma))\big\}. \end{aligned} \tag{17}$$

Since $D_{N\dagger}(\gamma)$ is a linear combination of $D_N(\gamma)$ and $D_X(\gamma)$, negative information acquisition combined with experimentation does not change the benefit of information $B_N(\gamma)$. It follows that regardless of whether learning is fast or slow, $D\dagger$ is not optimal for any belief when $k_N/\alpha_N \geq B_N^*$.

[8] This point is not simple to establish analytically. There are multiple forces at work. Importantly, if the state is $\mathcal{G}$ and the agent chooses N at γ_{XN}, he has a chance of getting the prize only after the belief reaches γ_{NX}. But if he chooses X at the same belief, he has a chance of getting the prize earlier. Higher discounting makes the first option relatively less attractive than the second option.

For fast learning, with $\alpha_N > \lambda$, the term $D_{N\dagger}(\gamma)$ enters the HJB equation (17) above with the same sign as $D_N(\gamma)$ in the corresponding equation (14) in Section 4.2. As a result, the optimal use of $N\dagger$ is also determined by comparing to a perpetual policy that alternates between $N\dagger$ and X. Although the belief goes up at the rate of $\alpha_N - \lambda$ in the absence of news or success, instead of α_N under N in the absence of news alone, keeping the belief unchanged means that the same payoff function $U_S(\gamma)$ defined in Section 4.2 obtains under fast learning. The optimal policy has the same structure as that described in Proposition 3. In particular, there is a unique $U^*_{N\dagger} \in (0, B^*_N)$ such that: (ii′) if $k_N/\alpha_N \in (U^*_{N\dagger}, B^*_N)$, then there exists $\gamma_{XN\dagger}$ such that the optimal policy is Q when $\gamma \leq \gamma_{QX}$, X when $\gamma \in (\gamma_{QX}, \gamma_{XN\dagger}]$ and when $\gamma \geq \gamma_{NX}$; and $N\dagger$ when $\gamma \in (\gamma_{XN\dagger}, \gamma_{NX}]$; (iii′) if $k_N/\alpha_N \leq U^*_{N\dagger}$, then there exists $\gamma_{QN\dagger}$ such that the optimal policy is Q when $\gamma \leq \gamma_{QN\dagger}$, $N\dagger$ when $\gamma \in (\gamma_{QN\dagger}, \gamma_{NX}]$, and X when $\gamma \geq \gamma_{NX}$. However, since the updated belief in the absence of news or success goes up at the reduced rate of $\alpha_N - \lambda$ instead of α_N, the region of beliefs for which $N\dagger$ is optimal shrinks at the lower boundary compared to the corresponding boundary of the region for which N is optimal. In other words, we have $\gamma_{XN\dagger} > \gamma_{XN}$ and $\gamma_{QN\dagger} > \gamma_{QN}$.[9]

For slow learning, with $\alpha_N < \lambda$, the term $D_{N\dagger}(\gamma)$ enters the HJB equation (17) with the opposite sign as $D_N(\gamma)$ in the corresponding equation (14) in Section 4.2. Because the belief always goes down in the case of slow learning, as in positive information acquisition, the optimal use of $N\dagger$ when $k_N/\alpha_N < B^*_N$ is determined by a smooth-pasting condition that requires $D_X(\gamma) = D_{N\dagger}(\gamma)$ at any switching belief between N and X. This requirement gives

$$\lambda\alpha_N V(\gamma) - \gamma\lambda\alpha_N\pi + \alpha_N c + \lambda k_N = 0.$$

Since the right-hand-side of the above is positive at $\gamma = 0$ and $\gamma = 1$, and since $V(\gamma)$ is convex, the smooth-pasting condition is either never satisfied (which happens when $k_N/\alpha_N > B^*_N$), or admits two solutions (when $k_N/\alpha < B^*_N$). This implies that for $k_N/\alpha_N < B^*_N$, $N\dagger$ is chosen for an interval of intermediate beliefs. Let the smaller and larger switching point be represented by $\gamma_{XN\dagger}$ and $\gamma_{N\dagger X}$, respectively.

[9]This follows because $N\dagger$ is equivalent to an interior solution with $\sigma_N = \sigma_X = 1/2$, and such a policy is sub-optimal by Proposition 3. Thus, if we define $U_{N\dagger}(\gamma)$ as the solution to the counterpart of (16), with N replaced by $N\dagger$ and with the same boundary condition $U_{N\dagger}(\gamma_{NX}) = U_S(\gamma_{NX})$, then $U_{N\dagger}(\gamma) < U_N(\gamma)$ for all $\gamma < \gamma_{NX}$. Moreover, the crossing point between $U_{N\dagger}(\gamma)$ and $U_X(\gamma)$ is to the right of the crossing point between $U_N(\gamma)$ and $U_X(\gamma)$.

The optimal policy is: (ii″) for $k_N/\alpha_N < B_N^*$, choose Q when $\gamma \leq \gamma_{QX}$, choose X when $\gamma \in (\gamma_{QX}, \gamma_{XN\dagger}]$ and when $\gamma \geq \gamma_{N\dagger X}$, and choose $N\dagger$ when $\gamma \in (\gamma_{XN\dagger}, \gamma_{N\dagger X}]$. However, notice that in the case of slow learning, $\gamma_{XN\dagger}$ is *not* a tipping point, because the updated beliefs goes down both to the left and to the right of this point when there is no news or no success. Nonetheless, slow learning cannot be optimally used a last-ditch effort like positive information acquisition, as the switching point $\gamma_{XN\dagger}$ is always above γ_{QX}. In other words, even when k_N/α_N is very low, there is no counterpart to case (iii) of Proposition 3 in slow learning.

5. Positive Versus Negative Information Acquisition

In the previous two sections we have presented two separate models, positive information acquisition and negative information acquisition. In this section we connect the two models in two ways. First, we make a comparison of them in terms of the region of their optimal use. Such a comparison furthers our understanding of the comparative statics of positive versus negative information acquisition. Second, we consider a fully dynamic model in which at any moment the agent can choose between positive and negative information acquisition in addition to experimentation and quitting. This optimal dynamic choice helps address the question of when and what kind of direct learning the agent should optimally engage in.

5.1. Ex ante comparison

From the earlier discussion, we find that both the benefit of positive information $B_P(\gamma)$ and the benefit of negative information $B_N(\gamma)$ are single-peaked in the belief about the state. The benefit $B_P(\gamma)$ is maximized at γ_{QX}, while the benefit $B_N(\gamma)$ is maximized at γ_N^*. The fact that $\gamma_{QX} < \gamma_N^*$ (see Figure 3) suggests positive information acquisition tends to be used when the agent is more pessimistic than when negative information acquisition is used. This is intuitive because positive information acquisition is used as a last-ditch effort and negative information acquisition is used as an insurance strategy. In the following proposition, we say that an interval of beliefs is *more pessimistic* than another interval if both the upper bound and the lower bound of the first are lower than those of the second.

Proposition 4. *If $k_P/\alpha_P = k_N/\alpha_N \leq \min\{B_P^*, B_N^*\}$, then positive information acquisition is optimally used at beliefs that are more pessimistic than the beliefs when negative information acquisition is optimally used.*

Without restricting positive and negative information acquisition to have the same efficiency, we can still ensure that when both are optimally used for some beliefs, it can never happen that the lower bound of the positive information acquisition region is higher than the upper bound of the negative information acquisition region. This claim follows because

$$\gamma_{QP} \leq \gamma_{QX} < \gamma_N^* \leq \gamma_{NX}.$$

However, the information acquisition regions for P and N are no longer ordered, because γ_{PX} may exceed γ_{NX}. This is the case if, example, k_P/α_P is close to 0 while k_N/α_N is close to B_N^*. With the restriction of $k_P/\alpha_P = k_N/\alpha_N = k/\alpha$, the optimal use of P may or may not overlap with the optimal use of N. For example, if k/α is just below B_N^* and $B_N^* < B_P^*$, then γ_{XN} is close to γ_{NX}, which by Proposition 4 strictly exceeds γ_{PX}; hence the two optimal uses do not overlap. If k/α is close to 0, γ_{QN} goes to 0 while γ_{PX} goes to 1; hence they do overlap.

5.2. Dynamic choice

Imagine that the agent can choose $\{Q, X, P, N\}$ at any moment. The HJB equation becomes

$$\begin{aligned} 0 = \max \big\{ &-V(\gamma), \gamma(1-\gamma)\lambda(D_X(\gamma) - V'(\gamma)), \gamma(1-\gamma)\alpha_P(D_P(\gamma) - V'(\gamma)), \\ &\gamma(1-\gamma)\alpha_N(V'(\gamma) - D_N(\gamma)) \big\}. \end{aligned} \tag{18}$$

The new comparison in (18) is between N and P. Since the agent's belief goes up when he chooses N and there is no negative news, and it goes down when he chooses P and there is no positive news, the switching point between N and P is determined by the comparison with a new perpetual policy.[10] It is straightforward to show that the payoff from permanently alternating between N and P to keep the belief stationary until the arrival of either the positive or negative news yields the payoff function

$$U_Y(\gamma) = \gamma(\pi - c/\lambda) - k_N/\alpha_N - k_P/\alpha_P.$$

As in the case of the absorbing point γ_{NX} between N and X, there is a unique candidate for an absorbing point between N and P, denoted as γ_{NP}, defined by smooth pasting and value matching with U_Y. This gives

$$\gamma_{NP} = \frac{k_P/\alpha_P}{k_N/\alpha_N + k_P/\alpha_P},$$

[10] Specifically, this policy is $\sigma_P = \alpha_N/(\alpha_P + \alpha_N)$, $\sigma_N = \alpha_P/(\alpha_P + \alpha_N)$, and $\sigma_X = \sigma_Q = 0$.

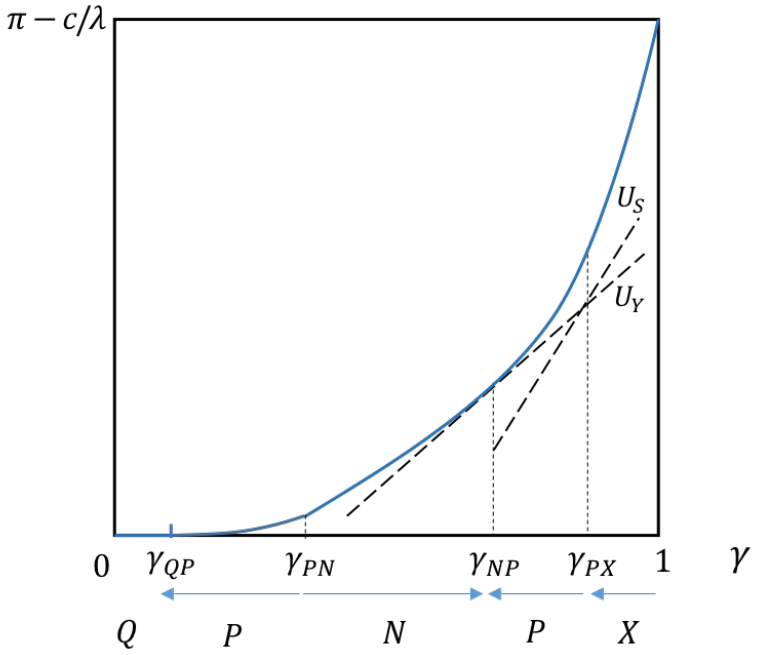

Figure 6. *Optimal policy in the dynamic choice model when k_P/α_N and k_N/α_N are equal, and are both sufficiently small. The value function is tangent to U_Y at the absorbing point γ_{NP}, where the agent mixes between N and P until the true state is revealed. The kink of the value function at γ_{PN} is a tipping point.*

with N chosen below the absorbing point and P chosen above it.

A full analysis of this model involves a large number of cases depending on the parameters, and is not particularly insightful. Instead we focus on the special case where $k_P/\alpha_P = k_N/\alpha_N = k/\alpha$, and k/α is small relative to both c/λ and $\pi - c/\lambda$ (the payoff when the state is known to be $\mathcal{G}$).

Proposition 5. *Let $\sigma(t) \in \{Q, X, P, N\}$ and assume $k_P/\alpha_P = k_N/\alpha_N = k/\alpha$. For any k/α sufficiently small, there exists $\gamma_{PN} \in (\gamma_{QP}, \gamma_{NP})$ such that the optimal policy is: Q when $\gamma \leq \gamma_{QP}$; P when $\gamma \in (\gamma_{QP}, \gamma_{PN}]$ and when $\gamma \in [\gamma_{NP}, \gamma_{PX}]$; N when $\gamma \in (\gamma_{PN}, \gamma_{NP}]$; and X when $\gamma > \gamma_{PX}$.*

Figure 6 illustrates the value function corresponding to the optimal policy described in Proposition 5. The value function has a convex kink at the tipping point γ_{PN}. Below it, the agent optimally chooses P as a last-ditch effort to find good news before he quits, which mimics the optimal policy in case (ii) of Proposition 1; above

it, the agent chooses N, and switches to a mix of N and P if no negative news is found and the belief goes up to γ_{NP}. Because information acquisition is efficient relative to experimentation, the agent chooses X only when the belief is close to 1. If success does not arrive, he switches to P, and eventually switches to a mix of N and P when no positive news is found.

Proposition 5 shows positive information acquisition is optimally used in two ways when both P and N are sufficiently efficient. For $\gamma \in (\gamma_{QP}, \gamma_{PN}]$, the chance of getting good news is relatively low. Positive information acquisition is used as a last-ditch effort to resurrect a potentially valuable project before the agent quits. This is the same as what we have in Section 3. However, with N now available at the same efficiency level, the region where P is optimally used as a last-ditch effort shrinks from $(\gamma_{QP}, \gamma_{PX}]$ to $(\gamma_{QP}, \gamma_{PN}]$. Positive information acquisition is also optimal for $\gamma \in [\gamma_{NP}, \gamma_{PX}]$, but plays a different role. With γ_{NP} as an absorbing point, the agent optimally chooses P when the belief is above γ_{NP} and the chance of getting positive news is relatively high, and the agent optimally chooses N when the belief is below γ_{NP} and the chance of getting negative news is relatively high. Therefore, similar to a finding in Che and Mierendorff (2017), positive information acquisition is part of the optimal strategy of pursuing "confirmatory" news when information acquisition is efficient.

In contrast, negative information acquisition is only used as part of confirmatory learning around the absorbing point γ_{NP} with positive information acquisition. The use of N as an insurance strategy to avoid wasteful experimentation in Section 4 has disappeared, because N is no longer used in an interval of beliefs with an upper-bound γ_{NX} as in case (ii) and case (iii) of Proposition 3. This is in spite of the result in Proposition 4 that $\gamma_{NX} > \gamma_{PX}$, so that the agent optimally uses negative information when he is more optimistic compared to when he uses positive information. However, that result presumes that P and N are only available on their own. When both are available at the same sufficiently efficient level, Proposition 5 shows that the absorbing point γ_{NX} between N and X is no longer relevant. The high efficiency of positive information acquisition has raised the value function entirely above U_S, so that X can only be used on its own, and not alternated with N in perpetual learning at γ_{NX}.

6. Discussion

There are potentially two broad directions where the model can be further extended. One is to generalize the idea of information acquisition in an experimentation model beyond positive and negative structures, maintaining the key feature that information acquisition has no direct payoffs. In the present paper, only two states are possible, one in which the risky arm is good and is expected to deliver a success, and the other in which the arm is bad and no success is ever possible. As a result, there is limited scope to address the issue of which type of information, and when, the agent should try to acquire. In a general environment with more than two possible states, we may model the dynamic trade-off between a narrow but in-depth search for information about the state versus a broad but cursory one. For example, in the R&D application of the bandit problem, a firm might be engaged in a product development process, where the potential new product can have several features that may be represented by a multi-dimensional state space. How to model the state space and build different types of information structures remains a challenge.

The second direction is to extend the single-agent model to a game between rival agents competing to be the first to achieve success from a risky arm. For example, imagine firms in the same market competing to develop a new product. There may be some uncertainty in the true state of the world, which could either be such that firms face identical and independent prospects of successfully developing a new product, or no success is possible by any firm because there is no demand for it or because the technology is not feasible. Information acquisition can be introduced in such a competitive framework, by assuming that agents have an independent, reversible and costly option of uncovering conclusive news about the state. The challenge is that if agents' information acquisition activities are not observable and news from information acquisition is not shared, even if they start with the same information about the state, private information will emerge among them. Recent papers that have introduced private information to bandit problems, including Moscarini and Squintani (2010), Farrell and Simcoe (2012), and Guo and Roesler (2016), may suggest a way to analyze the strategic interactions that come with learning while experimenting.

Appendix

Proof of Proposition 1. By the principle of dynamic programming,

$$V(\gamma) = \max_{\sigma_Q,\sigma_X,\sigma_P} (\gamma\lambda\pi - c)\sigma_X \mathrm{d}t + (\gamma\alpha_P(\pi - c/\lambda) - k_P)\sigma_P \mathrm{d}t \\ + (1 - \sigma_Q \mathrm{d}t - \gamma\lambda\sigma_X \mathrm{d}t - \gamma\alpha_P\sigma_P \mathrm{d}t)(V(\gamma) + V'(\gamma)\mathrm{d}\gamma).$$

where $\mathrm{d}\gamma = -\gamma(1-\gamma)(\lambda\sigma_X + \alpha_P\sigma_P)\mathrm{d}t$. Using the definitions of D_X and D_P, we have the HJB equation which allows for mixing:

$$0 = \max_{\sigma_Q,\sigma_X,\sigma_P} - V(\gamma)\sigma_Q + \gamma(1-\gamma)(D_X(\gamma) - V'(\gamma))\lambda\sigma_X \\ + \gamma(1-\gamma)(D_P(\gamma) - V'(\gamma))\alpha_P\sigma_P, \tag{19}$$

subject to $\sigma_Q, \sigma_X, \sigma_P \geq 0$ and $\sigma_Q + \sigma_X + \sigma_P = 1$.

In the following, for $\sigma = X, P$, we use $\tilde{U}_\sigma(\gamma; \hat{\gamma}, \hat{v})$ to represent the solution to the differential equation, $D_\sigma(\gamma) = U'(\gamma)$ (where U replaces V in the equation that defines D_σ), with boundary condition $U(\hat{\gamma}) = \hat{v}$.

Case (i). The candidate value function $V(\gamma)$ corresponding to the policy is:

$$V(\gamma) = U_X(\gamma) = \begin{cases} 0 & \text{if } \gamma \leq \gamma_{QX}, \\ \tilde{U}_X(\gamma; \gamma_{QX}, 0) & \text{if } \gamma > \gamma_{QX}. \end{cases}$$

It is straightforward to show that $k_P/\alpha_P \geq B_P^*$ implies $\gamma_{PX} \leq \gamma_{QX} \leq \gamma_{QP}$. For $\gamma < \gamma_{QX}$, we have $V'(\gamma) = V(\gamma) = 0$. By definition of γ_{QX}, we have $D_X(\gamma) < V'(\gamma)$. Also, $D_P(\gamma) < V'(\gamma)$ because $\gamma < \gamma_{QX} \leq \gamma_{QP}$. This implies that the HJB equation (19) holds, with $\sigma_Q = 1$ solving the maximization problem. For $\gamma > \gamma_{QX}$, we have $V(\gamma) > 0$ and $V'(\gamma) = D_X(\gamma)$. Moreover, $D_P(\gamma) < V'(\gamma)$ because $\gamma > \gamma_{QX} \geq \gamma_{PX}$. It follows that (19) holds, with $\sigma_X = 1$ solving the maximization problem on the right-hand-side. Since V is continuously differentiable and is a solution to the HJB equation (19), by Theorem 9.8 of Oksendal and Sulem (2005), it is the value function corresponding to problem (2).

Case (ii). The candidate value function corresponding to the policy is:

$$V(\gamma) = \begin{cases} 0 & \text{if } \gamma \leq \gamma_{QP}, \\ \tilde{U}_P(\gamma; \gamma_{QP}, 0) & \text{if } \gamma \in (\gamma_{QP}, \gamma_{PX}], \\ \tilde{U}_X(\gamma; \gamma_{PX}, \tilde{U}(\gamma_{PX})) & \text{if } \gamma > \gamma_{PX}. \end{cases}$$

Observe that $\gamma_{QP} < \gamma_{QX} < \gamma_{PX}$ because $k_P/\alpha_P < B_P^*$ in this case. For $\gamma < \gamma_{QP}$, we have $V'(\gamma) = V(\gamma) = 0$. Since $D_P(\gamma) < V'(\gamma)$ by definition of γ_{QP}, and since $D_X(\gamma) < V'(\gamma)$ by definition of γ_{QX} and by $\gamma_{QX} > \gamma_{QP}$, the HJB equation (19) holds, with $\sigma_Q = 1$ solving the maximization problem on the right-hand-side. For $\gamma \in (\gamma_{QP}, \gamma_{PX})$, we have $V(\gamma) > 0$ and $V'(\gamma) = D_P(\gamma)$. Also, $D_X(\gamma) < V'(\gamma)$ because $\gamma < \gamma_{PX}$. It follows that the HJB equation (19) holds, with $\sigma_P = 1$ solving the maximization problem. For $\gamma > \gamma_{PX}$, we have $V'(\gamma) = D_X(\gamma)$. Also, $D_P(\gamma) < V'(\gamma)$ because $\gamma > \gamma_{PX}$. Therefore the HJB equation (19) holds, with $\sigma_X = 1$ solving the maximization problem on the right-hand-side. Since V is a continuously differentiable function that satisfies (19), it is the value function for problem (2). ∎

Proof of Proposition 2. Suppose to the contrary that $\sigma(t) = P\dagger$ for t such that $\gamma(t)$ belongs to some interval (γ', γ''). Then the HJB equation (11) implies that, for all $\gamma \in (\gamma', \gamma'')$,

$$D_{P\dagger}(\gamma) = V'(\gamma) \geq \max\{D_P(\gamma), D_X(\gamma)\}.$$

But since

$$(\lambda + \alpha_P) D_{P\dagger}(\gamma) = \lambda D_X(\gamma) + \alpha_P D_P(\gamma),$$

the HJB equation would imply that $D_X(\gamma) = D_P(\gamma)$ for all $\gamma \in (\gamma', \gamma'')$, which is a contradiction because $D_X(\gamma) = D_P(\gamma)$ implies $\gamma = \gamma_{PX}$. ∎

Proof of Proposition 3. Using a similar derivation as in the proof of Proposition 1, we can show that the HJB equation that allows for mixing is

$$\begin{aligned} 0 = \max_{\sigma_Q, \sigma_X, \sigma_N} & -V(\gamma)\sigma_Q + \gamma(1-\gamma)(D_X(\gamma) - V'(\gamma))\lambda\sigma_X \\ & + \gamma(1-\gamma)(V'(\gamma) - D_N(\gamma))\alpha_N\sigma_N, \end{aligned} \tag{20}$$

subject to $\sigma_Q, \sigma_X, \sigma_N \geq 0$ and $\sigma_Q + \sigma_X + \sigma_N = 1$. Denote as $\tilde{U}_X(\gamma; \hat{\gamma}, \hat{v})$ the solution to the differential equation, $D_X(\gamma) = U'(\gamma)$ (where U replaces V in the equation that defines D_X), with boundary condition $U(\hat{\gamma}) = \hat{v}$.

Case (i). The candidate value function is $V(\gamma) = U_X(\gamma)$. For $\gamma < \gamma_{QX}$, we have $V'(\gamma) = V(\gamma) = 0$. Since $D_X(\gamma) < V'(\gamma)$ by definition of γ_{QX}, and since $D_N(\gamma) > V'(\gamma)$ by definition of D_N, the HJB equation (20) holds, with $\sigma_Q = 1$ solving the maximization problem of the right-hand-side. For $\gamma > \gamma_{QX}$, we have $V(\gamma) > 0$ and $V'(\gamma) = D_X(\gamma)$. Moreover,

$$\gamma(1-\gamma)(D_N(\gamma) - V'(\gamma)) = k_N/\alpha_N - (\gamma\pi - c/\lambda - U_X(\gamma)) = k_N/\alpha_N - B_N(\gamma).$$

Since $k_N/\alpha_N \geq B_N^*$, we have $D_N(\gamma) \geq V'(\gamma)$. Hence (20) holds, with $\sigma_X = 1$ solving the maximization problem. The value function V is by construction continuously differentiable, and thus corresponds to the solution to the original problem (12). This completes the proof for case (i).

Now, suppose that $k_N/\alpha_N < B_N^*$. Recall that $U_N(\gamma)$ crosses $U_X(\gamma)$ below γ_{NX} at most once, and at the crossing point U_N is steeper than U_X. Before proceeding to cases (ii) and (iii), we show that there is a unique value of $k_N/\alpha_N \in (0, B_N^*)$ such that $U_N(\gamma)$ crosses $U_X(\gamma)$ at γ_{QX}, with $U_N(\gamma_{QX}) = 0$. As U_N solves $D_N(\gamma) = V'(\gamma)$ with boundary condition $V(\gamma_{NX}) = U_S(\gamma_{NX})$, an increase in k_N/α_N strictly decreases $U_N(\gamma_{QX})$, because γ_{NX} is smaller, and $U_S(\gamma)$ is smaller and $D_N(\gamma)$ is greater for all γ. At $k_N/\alpha_N = 0$, we obtain the explicit solution $U_N(\gamma) = \gamma(\pi - c/\lambda)$, and thus $U_N(\gamma_{QX}) > 0$. At $k_N/\alpha_N = B_N^*$, by definition of B_N^* we have that U_X is tangent to U_S at γ_{NX}. The strict single-crossing property of $U_N(\gamma) - U_X(\gamma)$ implies that $U_N(\gamma) < U_X(\gamma)$ for all $\gamma < \gamma_{NX}$, and hence $U_N(\gamma_{QX}) < U_X(\gamma_{QX}) = 0$. Thus, there is $U_N^* \in (0, B_N^*)$ such that $U_N(\gamma_{QX}) > 0$ if and only if $k_N/\alpha_N < U_N^*$. If $k_N/\alpha_N \in [U_N^*, B_N^*)$, then $U_N(\gamma)$ crosses $U_X(\gamma)$ to the right of γ_{QX} at a point denoted γ_{XN}; this corresponds to case (ii) of the proposition. If $k_N/\alpha_N < U_N^*$, the crossing point is to the left of γ_{QX}, which we denote as γ_{QN}; this corresponds to case (iii) of the proposition.

Case (ii). Consider the candidate value function:

$$V(\gamma) = \begin{cases} U_X(\gamma) & \text{if } \gamma \leq \gamma_{XN}, \\ U_N(\gamma) & \text{if } \gamma \in (\gamma_{XN}, \gamma_{NX}], \\ \tilde{U}_X(\gamma; \gamma_{NX}, U_S(\gamma_{NX})) & \text{if } \gamma > \gamma_{NX}. \end{cases}$$

By the strict single-crossing property of $U_N(\gamma) - U_X(\gamma)$, the constructed V has a convex kink at γ_{XN}, and is therefore strictly convex for all $\gamma > \gamma_{QX}$. It follows that

$$\gamma(1-\gamma)(D_N(\gamma) - D_X(\gamma)) = V(\gamma) - U_S(\gamma) \geq 0$$

for all $\gamma > \gamma_{QX}$, with strict inequality except at $\gamma = \gamma_{NX}$.

For $\gamma < \gamma_{QX}$, the argument for verifying that $V(\gamma)$ satisfies (20) is the same as in case (i). For $\gamma \in (\gamma_{QX}, \gamma_{XN})$ and $\gamma > \gamma_{NX}$, we have $D_X(\gamma) = V'(\gamma)$ by construction. Moreover, $D_N(\gamma) > V'(\gamma)$. Hence the HJB equation (20) holds with $\sigma_X = 1$ solving the maximization problem on the right-hand-side. For $\gamma \in (\gamma_{XN}, \gamma_{NX})$, we have

$D_N(\gamma) = V'(\gamma)$ by construction. Moreover, $D_X(\gamma) < V'(\gamma)$. Hence the HJB equation (20) holds with $\sigma_N = 1$ solving the maximization problem.

Finally, $V(\gamma)$ is continuously differentiable by construction, except at the point γ_{XN}. At γ_{XN}, the left-derivative is less than the right-derivative:

$$V'_-(\gamma_{XN}) = D_X(\gamma_{XN}) < D_N(\gamma_{XN}) = V'_+(\gamma_{XN}).$$

For every $v \in [V'_-(\gamma_{XN}), V'_+(\gamma_{XN})]$, we have

$$\begin{aligned} 0 \geq \max_{\sigma_Q, \sigma_X, \sigma_N} & -V(\gamma_{XN})\sigma_Q + \gamma_{XN}(1-\gamma_{XN})(D_X(\gamma_{XN}) - v)\lambda\sigma_X \\ & + \gamma_{XN}(1-\gamma_{XN})(v - D_N(\gamma_{XN}))\alpha_N\sigma_N. \end{aligned}$$

By a standard verification theorem (Oksendal and Sulem, 2005, Theorem 9.8), the value function $V(\gamma)$ is a viscosity solution of the HJB equation, and therefore corresponds to the solution to the original maximization problem (12).

Case (iii). The candidate value function is constructed as follows:

$$V(\gamma) = \begin{cases} 0 & \text{if } \gamma \leq \gamma_{QN}, \\ U_N(\gamma) & \text{if } \gamma \in (\gamma_{QN}, \gamma_{NX}], \\ \tilde{U}_X(\gamma; \gamma_{NX}, U_S(\gamma_{NX})) & \text{if } \gamma > \gamma_{NX}. \end{cases}$$

Following a similar argument as in case (ii), we can show that $D_N(\gamma) \geq D_X(\gamma)$ for all $\gamma > \gamma_{QN}$, with strictly inequality except for $\gamma = \gamma_{NX}$.

For $\gamma < \gamma_{QN}$, we have $V(\gamma) = V'(\gamma) = 0$. Since $\gamma_{QN} < \gamma_{QX}$ in this case, $D_X(\gamma) < V'(\gamma)$. Furthermore, $D_N(\gamma) > V'(\gamma)$. Hence, the HJB equation (20) holds with solution $\sigma_Q = 1$ to the maximization problem. For $\gamma \in (\gamma_{QN}, \gamma_{NX})$, we have $D_N(\gamma) = V'(\gamma)$ by construction. Moreover, $D_X(\gamma) < V'(\gamma)$. Hence V satisfies (20) with $\sigma_N = 1$ solving the maximization problem on the right-hand-side. For $\gamma > \gamma_{NX}$, we have $D_X(\gamma) = V'(\gamma)$ by construction. Moreover, $D_X(\gamma) < V'(\gamma)$. Hence V satisfies (20) with solution $\sigma_X = 1$ to the maximization problem.

The function $V(\gamma)$ is continuously differentiable except at the point γ_{QN}. Further, at γ_{QN} we have

$$V'_-(\gamma_{QN}) = 0 < D_N(\gamma_{QN}) = V'_+(\gamma_{QN}).$$

For every $v \in [V'_-(\gamma_{QN}), V'_+(\gamma_{QN})]$, we have

$$0 \geq \max_{\sigma_Q, \sigma_X, \sigma_N} (D_X(\gamma_{QN}) - v)\lambda\sigma_X + (v - D_N(\gamma_{QN}))\alpha_N\sigma_N,$$

where the inequality follows because $\gamma_{QN} < \gamma_{QX}$ implies that $D_X(\gamma_{QN}) < 0$. Thus the value function $V(\gamma)$ is a viscosity solution of the HJB equation, and is the solution to problem (12). ∎

Proof of Proposition 4. Denote $k_P/\alpha_P = k_N/\alpha_N = k/\alpha$. A direct comparison of the upper bound of the optimal uses establishes that $\gamma_{PX} < \gamma_{NX}$.

For the lower bound, when $k/\alpha > U_N^*$, from case (ii) of Proposition 3, we have $\gamma_{QP} \leq \gamma_{QX} < \gamma_{XN}$. When $k/\alpha \leq U_N^*$, in case (iii) of Proposition 3, we have

$$U_N(\gamma) = \gamma(\pi - c/\lambda) - k/\alpha + \gamma\left(\log\left(\frac{k/\alpha}{c/\lambda}\frac{\gamma}{1-\gamma}\right) - 1\right)k/\alpha. \tag{21}$$

At $\gamma = \gamma_{QP}$, the first two terms vanish. The third term is negative, because from $\gamma_{QP} \leq \gamma_{QX} < \gamma_N^* < \gamma_{NX}$ we have

$$\frac{k/\alpha}{c/\lambda}\frac{\gamma_{QP}}{1-\gamma_{QP}} < \frac{k/\alpha}{c/\lambda}\frac{\gamma_{NX}}{1-\gamma_{NX}} = 1.$$

As a result, $U_N(\gamma_{QP}) < 0 = U_N(\gamma_{QN})$. Since U_N is an increasing function, we have $\gamma_{QP} < \gamma_{QN}$. Thus, for any $k/\alpha \leq \min\{B_P^*, B_N^*\}$, we have $\gamma_{QP} < \min\{\gamma_{XN}, \gamma_{QP}\}$. ∎

Proof of Proposition 5. For each $\sigma = X, P, N$, denote as $\tilde{U}_\sigma(\gamma; \hat{\gamma}, \hat{v})$ to represent the solution to the differential equation, $D_\sigma(\gamma) = U'(\gamma)$ (where U replaces V in the equation that defines D_σ), with boundary condition $U(\hat{\gamma}) = \hat{v}$. The candidate value function is constructed as follows:

$$V(\gamma) = \begin{cases} 0 & \text{if } \gamma \leq \gamma_{QP}, \\ \tilde{U}_P(\gamma; \gamma_{QP}, 0) \equiv \tilde{U}_{P1}(\gamma) & \text{if } \gamma \in (\gamma_{QP}, \gamma_{PN}], \\ \tilde{U}_N(\gamma; \gamma_{NP}, U_Y(\gamma_{NP})) \equiv \tilde{U}_N(\gamma) & \text{if } \gamma \in (\gamma_{PN}, \gamma_{NP}], \\ \tilde{U}_P(\gamma; \gamma_{NP}, U_Y(\gamma_{NP})) \equiv \tilde{U}_{P2}(\gamma) & \text{if } \gamma \in (\gamma_{NP}, \gamma_{PX}], \\ \tilde{U}_X(\gamma; \gamma_{PX}, \tilde{U}_{P2}(\gamma_{PX})) \equiv \tilde{U}_X(\gamma) & \text{if } \gamma > \gamma_{PX}. \end{cases}$$

We begin by showing that the following four conditions hold when k/α is sufficiently small. First, $\gamma_{QP} < \gamma_{QX} < \gamma_{PX}$, so we have case (ii) of Proposition 1 if N is unavailable. This is satisfied if $k/\alpha < (\pi - c/\lambda)c/(\lambda\pi)$.

Second, $\gamma_{QP} < \gamma_{NP} < \gamma_{PX}$, so that the switching point γ_{NP} between N and P is potentially valid. This is satisfied if $k/\alpha < \min\{c/(2\lambda), (\pi - c/\lambda)/2\}$.

Third, $\tilde{U}_N(\gamma)$ and $\tilde{U}_{P1}(\gamma)$ cross each other once at some γ_{PN} which satisfies $\tilde{U}'_N(\gamma_{PN}) > \tilde{U}'_{P1}(\gamma_{PN})$ and $\gamma_{QP} < \gamma_{PN} < \gamma_{NP}$, so that V has a convex kink at the tipping point γ_{PN}. At any crossing point γ_{PN}, the sign of $\tilde{U}'_N(\gamma_{PN}) - \tilde{U}'_{P1}(\gamma_{PN})$ is the same as

$$\gamma_{PN}(1-\gamma_{PN})\left(D_N(\gamma_{PN}) - D_P(\gamma_{PN})\right) = \tilde{U}_N(\gamma_{PN}) - U_Y(\gamma_{PN}),$$

which is positive because $\tilde{U}_N$ is strictly convex while U_Y is linear, and they are tangent to each other at γ_{NP}. Solving the relevant differential equations gives

$$\begin{aligned}
\tilde{U}_N(\gamma) &= \gamma(\pi - c/\lambda) - \left(1 + 2\gamma - \gamma \log \frac{\gamma}{1-\gamma}\right) k/\alpha, \\
\tilde{U}_{P1}(\gamma) &= \gamma(\pi - c/\lambda) - \left(1 + (1-\gamma)\log\left(\frac{\gamma}{1-\gamma}\frac{1-\gamma_{QP}}{\gamma_{QP}}\right)\right) k/\alpha.
\end{aligned}$$

Clearly, we have $\tilde{U}_N(\gamma_{QP}) < \tilde{U}_{P1}(\gamma_{QP})$; we also have $\tilde{U}_N(\gamma_{PN}) > \tilde{U}_{P1}(\gamma_{PN})$ if $\log((1-\gamma_{QP})/\gamma_{QP}) > 2$. The third condition is satisfied if k/α is sufficiently small.

Fourth, $\tilde{U}_X(\gamma_{NX}) > U_S(\gamma_{NX})$, so γ_{NX} is not a potential switching point between N and X. Solving the differential equation yields

$$\tilde{U}_{P2}(\gamma) = \gamma(\pi - c/\lambda) - \left(1 + 2(1-\gamma) + (1-\gamma)\log\frac{\gamma}{1-\gamma}\right) k/\alpha.$$

We have

$$\tilde{U}_X(\gamma_{NX}) - U_S(\gamma_{NX}) > \tilde{U}_{P2}(\gamma_{NX}) - U_S(\gamma_{NX}) = \left(\frac{c/\lambda}{k/\alpha} - \log\frac{c/\lambda}{k/\alpha} - 2\right)(1-\gamma_{NX})k/\alpha.$$

For sufficiently small k/α, the term in the first bracket on the right-hand-side above is positive and hence the fourth condition holds.

Now, we show that V satisfies the HJB equation (18). For $\gamma < \gamma_{QP}$, we have $V'(\gamma) = 0$. By definition of γ_{QP}, we have $D_P(\gamma) < V'(\gamma)$. Further, $D_X(\gamma) < V'(\gamma)$ because $\gamma < \gamma_{QP} < \gamma_{QX}$ by the first condition above. Finally, $D_N(\gamma) > V'(\gamma)$. Thus, the HJB equation (18) holds with $\sigma = Q$ solving the maximization problem on the right-hand-side.

For $\gamma \in (\gamma_{QP}, \gamma_{PN})$, we have $V'(\gamma) = D_P(\gamma)$. Since $\gamma < \gamma_{PN} < \gamma_{PX}$ by the second condition, $D_X(\gamma) < V'(\gamma)$. Moreover, since V has a convex kink at γ_{NP} and is tangent to U_Y at γ_{NP},

$$\gamma(1-\gamma)(D_N(\gamma) - D_P(\gamma)) = V(\gamma) - U_Y(\gamma) \geq 0,$$

with strict inequality except for $\gamma = \gamma_{NP}$. Thus, $D_N(\gamma) > V'(\gamma)$ for $\gamma \in (\gamma_{QP}, \gamma_{PN})$. The HJB equation (18) holds with $\sigma = P$ solving the maximization problem.

For $\gamma \in (\gamma_{PN}, \gamma_{NP})$, we have $V'(\gamma) = D_N(\gamma)$. We have just argued above that $D_P(\gamma) - V'(\gamma) < 0$. By the second condition, this in turn implies $D_X(\gamma) - V'(\gamma) < 0$ because $\gamma < \gamma_{NP} < \gamma_{PX}$. Thus, (18) holds with $\sigma = N$ solving the maximization problem.

For $\gamma \in (\gamma_{NP}, \gamma_{PX})$, we have $V'(\gamma) = D_P(\gamma)$. As before, $D_N(\gamma) > V'(\gamma)$, and $D_X(\gamma) - V'(\gamma) < 0$ because $\gamma < \gamma_{PX}$. Thus, (18) holds with $\sigma = P$ solving the maximization problem.

Finally, for $\gamma > \gamma_{PX}$, we have $V'(\gamma) = D_X(\gamma)$ by construction. Since $\gamma > \gamma_{PX}$, we have $D_P(\gamma) < V'(\gamma)$. Observe that

$$\gamma(1-\gamma)(D_N(\gamma) - V'(\gamma)) = V(\gamma) - U_S(\gamma).$$

By the fourth condition, $V(\gamma_{NX}) > U_S(\gamma_{NX})$. By definitions of D_X and U_S, we have $V'(\gamma_{NX}) < \pi$, and if we define $\hat{\gamma}$ such that $V'(\hat{\gamma}) = \pi$, then $\hat{\gamma} > \gamma_{NX}$ and hence $V(\hat{\gamma}) > U_S(\hat{\gamma})$. Since V is convex and U_S is linear, we have $V(\gamma) > U_S(\gamma)$ for all $\gamma > \gamma_{PX}$, and thus $D_N(\gamma) > V'(\gamma)$. The HJB equation (18) holds with $\sigma = X$ solving the maximization problem.

We have established that $V(\gamma)$ satisfies the HJB equation (18) whenever it is differentiable. Moreover, V is continuously differentiable by construction, except at the point γ_{PN}, where it has a convex kink. Thus, V is a viscosity solution to the HJB equation. By the same verification theorem invoked in the proof of Proposition 3, V is the value function for the control problem involving $\{Q, X, P, N\}$. ∎

References

Bergemann, D., and Valimaki, J. "Learning and strategic pricing," *Econometrica* 64 (1996): 1125–1150.

Bergemann, D., and Valimaki, J. "Experimentation in markets," *Review of Economic Studies* 67 (2000): 213–234.

Bolton, P., and Harris, C. "Strategic experimentation," *Econometrica* 67 (1999): 349–374.

Calvert, R. "The value of biased information: a rational choice model of political advice," *Journal of Politics* 47 (1985): 530–555.

Camargo, B. "Good news and bad news in two-armed bandits," *Journal of Economic Theory* 133 (2007): 558–566.

Che, Y. K., and Mierendorff, K. "Optimal sequential decision with limited attention," Columbia University working paper, 2017.

Choi, J. P. "Dynamic R&D competition under 'hazard rate' uncertainty," *RAND Journal of Economics* 22 (1991): 596–610.

Farrell, J., and Simcoe, T. "Choosing the rules for consensus standardization," *RAND Journal of Economics* 43 (2012): 235–252.

Gittins, J. C. "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society* 41 (1979): 148–177.

Guo, Y., and Roesler, A. K. "Private learning and exit decisions in collaboration," Northwestern University working paper, 2016.

Harris, C., and Vickers, J. "Perfect equilibrium in a model of a race," *Review of Economic Studies* 52 (1985): 193–209.

Harris, C., and Vickers, J. "Racing with uncertainty," *Review of Economic Studies* 54 (1987): 1–21.

Moscarini, G., and Squintani, F. "Competitive experimentation withe private information: the survivor's curse," *Journal of Economic Theory* 145 (2010): 639–660.

Klein, N., and Rady, S. "Negatively correlated bandits," *Review of Economic Studies* 78 (2011): 693–732.

Keller, G., Cripps, M., and Rady, S. "Strategic experimentation with exponential bandits," *Econometrica* 73 (2005): 39–68.

Malueg, D., and Tsutsui, S. "Dynamic competition with learning," *RAND Journal of Economics* 28 (1997): 751–772.

Oksendal, B., and Sulem, A. *Applied Stochastic Control of Jump Diffusions,* Berlin: Springer (2005).

Reinganum, J. "Dynamic games of innovation," *Journal of Economic Theory* 25 (1981): 21–41.

Reinganum, J. "A dynamic game of R&D: Patent protection and competitive behavior," *Econometrica* 50 (1982): 671–688.

Roberts, K., and Weitzman, M. "Funding criteria for research, development and exploration of projects," *Econometrica* 49 (1981): 1261–1288.

Suen, W. "The self-perpetuation of biased beliefs," *Economic Journal* 114 (2004): 377–396.

Weitzman, M. "Optimal search for the best alternative," *Econometrica* 47 (1979): 641–654.