Optimal exploration*

David Austen-Smith† César Martinelli‡

December 5, 2018

Abstract

Consider a decision maker who has to choose one of several alternatives, and who is imperfectly informed about the payoff of each of them. In each period, the decision maker has to decide whether to stop and take one of the alternatives, or to continue researching the alternatives. New information is costly and is never conclusive. We provide a dynamic programming formulation of the decision maker's problem with either a finite deadline or no deadline, and give necessary and sufficient conditions for research to take place for some prior beliefs about the alternatives. We show that, at least for short deadlines and situations in which there is a strictly positive probability the decision maker stops searching in the next period under the optimal plan, the decision maker either explores the best alternative and stops after good news, or explores the second best alternative and stops after bad news, with the former path being optimal if the decision maker is relatively optimistic about the payoff of the alternatives. On the other hand, an example shows that searching the third best alternative can be optimal when there are more than three remaining search periods and there is no likelihood of stopping in the next period.

1 Introduction

Consider the following situation. A decision maker has to choose one of several alternatives. There is only imperfect information about the payoff of each of the alternatives. Before choosing one of them, the decision maker can research them

*Austen-Smith is grateful for support under the IDEX Chair, "Information, Deliberation and Collective Choice" at the Institute for Advanced Study in Toulouse (IAST). We thank Arthur Dolgopolov for expert research assistance.

†Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University. Email: dasm@northwestern.edu.

‡Department of Economics, George Mason University. Email: cmarti33@gmu.edu.

sequentially. New information is costly and is never conclusive. In each period, then, the decision maker has to decide whether to stop and take one of the alternatives (presumably, the one with the largest expected payoff), or to continue researching the alternatives. The following examples fit this description:

- (Policy search) A collective decision making group has to choose one of a set of available policies to promote a given group objective. The consequences of choosing any policy are not known surely but the group can devote time and effort to researching the alternatives prior to any final decision. Similarly, legislative committees often have the responsibility of determining the alternative to a given status quo policy, prior to the general assembly voting over whether to maintain the status quo or to reject it in favour of the alternative. The committee first engages in costly research to decide upon which of the available alternatives to propose.

- (House hunting) A family is considering several alternatives for a new home. At each moment in time, the family can visit only one of the new houses, and obtains an imperfect signal of the house quality, until making a final decision on which house to buy.

- (Hiring decision) A firm is considering several candidates for a job. At each moment the firm can acquire new information about one of the candidates, for instance by conducting an interview or organizing a visit, until making a final decision on which candidate to hire.

- (Exploratory drilling) An oil company is hesitating between a number of potential drilling locations. Before exploitation starts, then, it can do some exploratory drilling using drilling methods that are cheaper than a production well.

- (Nest search) A beehive, composed of perfectly altruistic individuals, must decide on the location of a new nest. Locations can be investigated one at a time, by sending a scout who reports truthfully to the hive. The bees have to decide at each moment which location to investigate and when to stop and move jointly to a new nest location.

Each of these applications has been analyzed using some application of the following canonical frameworks, none of which fully captures our set-up:

1. *Bandits.* A defining characteristic of bandit problems is that the alternatives are statistically independent. Although this is true of the alternatives in our model, exploring any alternative here yields only a noisy signal of that alternative's payoff: payoffs from an alternative can be realized only if the

decision maker stops exploring and chooses that alternative. An important consequence of these features is that the exploration-values of alternatives are no longer statistically independent.

2. *Pandora's Box.* In a Pandora box problem, a decision maker obtains a perfectly informative signal about each alternative after investigating it. In our problem, instead, information obtained about an alternative is never conclusive, so that it may be optimal to investigate the same alternative more than once.

3. *Sequential sampling.* In the classical sequential sampling problem, the sequence of experiments is fixed in advance. In the related sequential search problem, experiments correspond to different alternatives arriving randomly over time. In the costly research problem, the variable of interest is the intensity rather than the direction of search. In our problem, the decision maker can decide which experiment to perform at each moment in time before settling for one of the alternatives.

The decision maker of interest here is required to find the optimal sequencing of experiments, taking into account the information available at each moment. As is well known, in both the bandit and Pandora problems, it is possible to construct an index for each alternative with information about that alternative alone, such that the decision maker's optimal choice in any period is the alternative with the highest index. We show that in our problem, the value of exploring each alternative is linked to beliefs about the other alternatives in a way that does not admit construction of an independent index for each of the alternatives. In particular, we show (theorem 5.1) that whether investigating the top or the second alternative is optimal depends on beliefs about the third, even if the third is not immediately investigated.

We offer an analytical characterization of the optimal policy when there are two remaining periods before a final decision must be made. If beliefs about the top two alternatives are close to symmetric, then it is optimal to investigate either one of the top two and to stop afterwards, selecting the alternative with larger posterior expected value. If beliefs about the top two alternatives are asymmetric enough, and initial beliefs are relatively optimistic (in a sense that we make precise), then it is optimal to investigate the top alternative, to stop after good news, and to investigate again after bad news. Finally, if beliefs about the top two alternatives are asymmetric enough, and initial beliefs are relatively pessimistic, then it is optimal to investigate the second alternative, to stop after after bad news, and to investigate again after good news.

The problem we investigate is related to the multi-armed bandit problem. Payoffs from playing any arm (exploring an alternative) in bandit problems are generated in every period. Specifically, the payoff from playing any arm depends only on outcomes in periods during which the arm is played and is therefore statistically independent of any other arm. The literature on bandit problems is considerable: Berry and Fristedt (1985) and Gittins et al. (2011) offer overviews of the theory and Bergemann and Välimäki (2008) provide a succinct summary of some major applications in economics. The key result from the theory is that so long as the arms are independent, the optimal exploration rule is characterized by an index, the Gittins Index (Gittins, 1979), which allows to rank the alternatives in terms of their desirability.[1]In our framework, payoffs only occur once all exploration has stopped permanently and, although the payoff from the chosen alternative at any date is independent of the payoffs associated with other alternatives, the value of exploration is not, as remarked above, so independent, undermining the possibility of an index characterization of an optimal exploration rule.[2]

In the Pandora's problem, introduced by Weitzman (1979), there is a given set of mutually exclusive options and a DM (Pandora) who chooses the order in which to investigate alternatives and when to stop investigating. The reward associated with any given option is unknown ex ante and can be enjoyed only after Pandora stops investigation and chooses the alternative with the highest observed reward. The distribution of rewards for any alternative and its cost of investigation are known. The true payoff from any alternative, however, is fully revealed conditional on exploring that alternative. Assuming that Pandora's objective is to maximize her discounted expected reward net of aggregate search costs, Weitzman (1979) proves that the optimal search rule for his setting is described by a "reservation price rule" akin to the Gittins index. Recent extensions of the Pandora box problem include Olszewski and Weber (2016) and Doval (2018). The key difference between our set-up and theirs is that exploration in Weitzman's environment yields the true value of any alternative but this is never the case for the model below. As a result, the DM in our model may revisit previously investigated alternatives multiple times. Furthermore, we show that no reservation price or index rule exists in our model.[3]

[1]The index property fails if arms are costly to switch (Banks and Sundaram, 1994) or to maintain (Forand, 2015).

[2]Glazebrook (1979) considers what he calls a bandit with an option of settling permanently in one alternative, i.e. a "stoppable bandit." He shows that an index characterization is obtained if the value of the stopping option is almost surely increasing over time. This is precluded in our setting: since we consider an information acquisition problem, exploring an alternative yields a martingale.

[3]Adam (2001) considers a situation in which opening a box both reveals the prize contained in that box and allows the DM to learn about other boxes with the same distribution of rewards. As in Pandora's problem, there is no residual uncertainty about the prize contained in an open box.

Independently from our work, Ke and Villas-Boas (2018) study a continuous time problem of learning about two alternatives when there is an outside option of known value. In their model, signals follow a Brownian motion with a drift given by the true value of the alternative. They show existence of a solution to the infinite horizon problem and characterize it as a system of differential equations, while we study a discrete time problem with many alternatives with possibly finite horizon. Ke and Vilas-Boas also find conditions under which searching the second best of two alternatives can be optimal (Ke and Villas-Boas (2018), Theorem 3 and Figure 3), although in our model there can be many alternatives.

The classical sequential sampling problem was introduced by Wald (1945, 1947) and by Blackwell and Girshick (1954). McCall (1970), Mortensen (1986) and a large literature afterwards investigate the sequential search or job search problem. Moscarini and Smith (2001) introduce a general dynamic model of costly research in which an impatient decision maker (DM) has to choose one from a set of alternatives, the respective payoff of which depends on the true state of the world. Before choosing, the DM can acquire noisy information over time about the state at a (state-independent) cost. The main difference between their setup and ours is that they focus on the intensity of research, while we focus on the direction of research. In particular, they assume decision making in continuous time; the costs of information regarding the true state are strictly convex increasing in the level of information acquisition; and state-specific payoffs across the set of alternatives are not limited to be one of two values as in our model. In addition to characterizing when information acquisition stops and an alternative chosen, their main result, among several, is that the optimal level of information acquisition increases in the Bellman value prior to choosing an alternative.[4] In our model, however, learning occurs piecemeal through sequentially exploring specific alternatives and we focus on the optimal decision at any time regarding which alternative, if any, to explore.

Less closely related contributions in the large literature on strategic experimentation include Bolton and Harris (1999) and Keller et al. (2005). Both papers, discussed along with others by Moscarini and Smith (2001), study dynamic information acquisition models with two alternatives, in which there are many individuals simultaneously making information acquisition decisions through time. All information discovered is public. The additional tensions addressed in these papers, therefore, involve individuals' opportunities to free-ride on others' search.

Section 2 of the paper introduces the model and section 3 formulates the DM's

[4]This result is driven by strict convexity of costs in the level of information acquisition, an economically meaningful assumption that yields a variety of testable predictions from the model regarding R&D.

problem as a dynamic optimization problem, establishing a variety of useful results. The following two sections, respectively, completely characterize the optimal research policy when the number of research periods remaining before a decision is either one (section 4) or two (section 5). Section 6 provides a partial characterization of the optimal policy for a longer horizon, while section 7 provides a computed example. Section 8 concludes. The appendix contains some auxiliary proofs.

2 The exploration problem

Time runs discretely, from $t = 0$ to infinity. There is a finite set of alternatives X, indexed by $x \in X = \{1, \ldots, n\}$. Each alternative has a quality $\omega_x \in \{0,1\}$ indicating whether it is low quality or high quality. The qualities of the alternatives are independent random variables and their realization are not observed by the decision maker (DM). At time 0, beliefs about the quality of each alternative are given by $p^0 = (p_x^0)$, where $p_x^0 \in (0,1)$ indicates the prior probability that alternative x is high quality. There is a deadline $\tau \in \mathbb{N} \cup \{\infty\} = \{0,1,2,\ldots\} \cup \{\infty\}$. At each time $0 \leq t < \tau$, if no alternative has been chosen yet, the DM can either research one of the alternatives in X, or choose one of the alternatives in X. If $\tau \neq \infty$, then the DM must choose an alternative at time τ if no alternative has been chosen yet.

If alternative x is researched, then the DM obtains a payoff $u_t = -c \leq 0$ in that period, and gets to observe a signal $s_x^t \in \{0,1\}$. If alternative x is chosen at time t, the DM obtains the payoff $u_t = \omega_x$ in that period, and the payoff $u_{t'} = 0$ in every subsequent period. That is, choosing an alternative is irreversible. DM's utility is the discounted sum of period payoffs, $\sum_{t \geq 0} \delta^t u_t$, where $\delta \in (0,1)$.

Conditional on the quality of the alternative, signals are drawn independently of previous signals of the same or other alternatives, with probability distribution given by

$$\Pr(s_x^t = 1 | \omega_x = 1) = \Pr(s_x^t = 0 | \omega_x = 0) = q,$$
$$\Pr(s_x^t = 1 | \omega_x = 0) = \Pr(s_x^t = 0 | \omega_x = 1) = 1 - q$$

for some $q \in (1/2, 1)$, where $s_x^t = 0$ represents "bad news" and $s_x^t = 1$ represents "good news" about alternative x. Researching alternative x in period t allows the DM to update beliefs about alternative x while keeping beliefs about other alternatives constant.

A *plan* π specifies for each time $0 \leq t < \tau$ such that no alternative is chosen yet, an action $a \in X \cup \{s_x\}_{x \in X}$, with $a = x$ meaning 'research alternative x,' and $a = s_x$ meaning 'stop and choose alternative x,' as a function of the initial beliefs and the history of actions and signals in every previous period. If $\tau \neq \infty$, a plan specifies as

well an action $a \in \{s_x\}_{x \in X}$ in period τ, if the DM has not chosen an alternative yet, as a function of the initial beliefs and the history of actions and signals in every previous period. A plan π is *optimal* if the expected utility of adopting π is larger than or equal to the expected utility of adopting any other plan for any possible initial beliefs.

Since we are interested in optimal plans, we restrict henceforth the set of available actions before the deadline to be $A = X \cup \{s\}$, with $a = s$ meaning 'stop and choose one of the alternatives with maximum (updated) probability of being high quality,' and simply assume that one of the alternatives with maximum (updated) probability of being high quality is chosen at the deadline if society has not chosen yet.

Given the history of previous actions and signals before any period $t \geq 1$, let $p^t = \{p_x^t\}$ denote the (updated) beliefs that each alternative x is high quality. Let $P = [0,1]^n$ be the set of possible beliefs. We say that a plan is *Markovian* if for every $t \geq 0$ there is a function $f_t : P \to A$ such that π specifies action $f_t(p)$ at time t if $p^t = p$. We refer to f_t as the *policy* associated to Markovian plan π at time t. We say that a plan π is *stationary* if there is a function $f : P \to A$ such that π specifies action $f(p)$ after every history such that $p^t = p$ for every $t \geq 0$. That is, a stationary plan specifies the same policy every period. Obviously, stationary plans are of interest if $\tau = \infty$.

3 Dynamic programming formulation

We can describe the DM's problem a dynamic programing problem as follows. Our state space is $P = [0,1]^n$. The set of actions is the same for all states and is given by A for all $t < \tau$ and by $\{s\}$ for $t \geq \tau$. The reward function is given by

$$r(p,a) = \begin{cases} -c & \text{if } a \in X \\ (1-\delta)\max_{x \in X} p_x & \text{if } a = s \end{cases}$$

Note that, in order to adapt the problem to a recursive formulation, we specify a per period utility after stopping such that the expected discounted utility is equal to the highest belief at the time of stopping, which is obtained if choosing optimally at that time.

Given any state $p \in P$, we denote p^{x+} the state given by $p_y^{x+} = p_y$ for all $y \in X \setminus \{x\}$ and

$$p_x^{x+} = \frac{p_x q}{p_x q + (1-p_x)(1-q)} \equiv p_x^+.$$

Similarly, we denote p^{x-} the state given by $p^{x-}_y = p_y$ for all $y \in X \setminus \{x\}$ and

$$p^{x-}_x = \frac{p_x(1-q)}{p_x(1-q)+(1-p_x)q} \equiv p^-_x.$$

Note that p^+_x and p^-_x are the updated beliefs about alternative x after 'good news' and 'bad news,' respectively.[5] Let also

$$\begin{array}{rclcl} q^+_x(p) & = & p_x q + (1-p_x)(1-q) & = & 1-q+p_x(2q-1) \\ q^-_x(p) & = & p_x(1-q)+(1-p_x)q & = & q-p_x(2q-1), \end{array}$$

denote the probability of 'good news' and 'bad news' about alternative x, respectively.

The law of motion, giving us the probability distribution over future states given the current state and action is given by

$$Q(p'|p,a) = \begin{cases} 1 & \text{if } p' = p \text{ and } a = s \\ q^+_x(p) & \text{if } p' = p^{x+} \text{ and } a = x \\ q^-_x(p) & \text{if } p' = p^{x-} \text{ and } a = x \\ 0 & \text{otherwise} \end{cases}.$$

The set of future states with positive probability after state p is given by

$$S(p) = \{p' \in P : Q(p'|p,a) > 0 \text{ for some } a \in A\};$$

note that $S(p)$ is finite.

Note that the parameters of our dynamic programming problem are the number of alternatives n, the cost of research c, the quality of information q, the discount factor δ, and the deadline τ.

Let $\mathcal{W}$ be the space of bounded continuous functions $W : P \to \Re$, equipped with the sup norm, and let the transformation $T : \mathcal{W} \to \mathcal{W}$ be defined by

$$(TW)(p) = \max_{a \in A} \left[r(p,a) + \delta \textstyle\sum_{p' \in S(p)} Q(p'|p,a) W(p') \right]. \tag{1}$$

It is easy to see that T satisfy the Blackwell conditions, that is (a) $W \geq W'$ for all p implies $TW \geq TW'$ for all p, and (b) for any constant b, $T(W+b) = TW + \delta b$. Hence, from theorem 5 in Blackwell (1965), T is a contraction with modulus δ, that is $||TW - TW'|| \leq \delta ||W - W'||$. From the Banach fixed-point theorem, T has a unique fixed point $V \in \mathcal{W}$ satisfying

$$TV = V, \tag{2}$$

[5] For given q, we treat $()^+$ and $()^-$ as functions from $[0,1]$ to $[0,1]$. In particular, $(\rho^+)^- = \rho$ for all $\rho \in [0,1]$. Note that $\rho^+ \gtreqless \nu \iff \rho \gtreqless \nu^-$. Similarly, we treat $()^{x+}$ and $()^{x-}$ as functions from P to P. In particular, $(p^{x+})^{x-} = p$ for all $p \in P$.

and moreover $||T^n W - V|| \leq \delta^n ||W - V||$ for all n.

Let $\mathcal{V}$ be the space of continuous functions $W : P \to [0,1]$, and note that $\mathcal{V}$ is a closed subset of $\mathcal{W}$ and moreover, $W \in \mathcal{V}$ implies $T^n W \in \mathcal{V}$ for all n. Since $\mathcal{V}$ is closed, it follows that $V \in \mathcal{V}$.

Now let

$$W_0(p) = \max_{x \in X} p_x;$$

this is the expected utility of the Markovian plan with policy $f_t(p) = s$ for all p for all $t \geq 0$; by necessity, it is the expected utility of the optimal plan if the deadline is $\tau = 0$. Then iteratively define $W_k \in \mathcal{V}$ by

$$W_k = TW_{k-1}.$$

If $\tau \geq k$, $W_k(p)$ is the expected utility of a Markovian plan that specifies a selection

$$f_t(p) \in \arg\max_{a \in A} \left[r(p,a) + \delta \textstyle\sum_{p' \in S(p)} Q(p'|p,a) W_{k-t-1}(p') \right]$$

for every $0 \leq t < k$, and $f_t(p) = s$ for $t \geq k$.

We have:

Theorem 3.1. *Fix n, c, q, and δ. (i) If $\tau \neq \infty$, a Markovian plan is optimal if and only if its expected utility in period t, as a function of current beliefs, is given by $W_{\tau - t}$ for each $t \leq \tau$. Moreover, there is a Markovian optimal plan. (ii) If $\tau = \infty$, a stationary plan is optimal if and only if its expected utility, as function of current beliefs, is given every period by V satisfying equation 2. Moreover, there is a stationary optimal plan.*

Proof. Item (i) follows from backward induction. For item (ii), necessity and sufficiency of the optimality equation 2 follow from theorem 6(f) in Blackwell (1965), which deals with infinite horizon discounted dynamic programming problems. Existence follows from theorem 7 in Blackwell (1965), which establishes that there is in fact a stationary plan whose expected payoff satisfies equation 2 when the action set is finite. □

Intuitively, $W_k(p)$ is the value associated to the DM problem when there remain k periods ahead before the deadline, and $V(p)$ is the value when there is no deadline.

The following result establishes that in fact the sequence W_t converges to V, and provides a computationally convenient bound.

Corollary 3.1. *The sequence of functions $\{W_k\}$ is monotonically increasing and converges uniformly to V, with $V - W_k \leq (W_k - W_{k-1}) \times \delta/(1-\delta)$ for all p.*

Proof. Since increasing k relaxes the constraint requiring $\hat{\pi}_k$ to stop at time k, it follows that the expected utility of $\hat{\pi}_k$ is larger than or equal to the expected utility of $\hat{\pi}_{k-1}$ for any given initial beliefs, and smaller or equal to the expected utility of the optimal plan, or equivalently $W_{k-1} \leq W_k \leq V$ for every $k \geq 1$ and every $p \in P$. From monotonicity, the definition of W_k, and the fact that T is a contraction, we get $V - W_k = V - T^k W_0 \leq \delta^k (V - W_0)$. Uniform convergence follows.

Again from the fact that T is a contraction, we get $V - W_k = V - TW_{k-1} \leq \delta(V - W_{k-1})$. The bound in the statement of the theorem follows from rewriting the previous inequality. □

The next result provides some convenient facts about the functions W_k and V.

Corollary 3.2. *W_k for $k \geq 0$ and V are (i) permutation invariant and (ii) weakly increasing functions.*

Proof. Part (i) is immediate from the definitions of W_k and V. For part (ii), we claim first that if W_{k-1} is weakly increasing, then W_k is weakly increasing. To see this, note that for $\tilde{p} \in P$ and $\hat{p} \in P$ such that $\tilde{p} \geq \hat{p}$ and for any given alternative x, the distribution over P induced by $Q(p'|\tilde{p},x)$ exhibits first order stochastic dominance over the distribution over P induced by $Q(p'|\hat{p},x)$. Hence, in the problem defining W_k, the expected payoff of searching alternative x

$$-c + \delta \textstyle\sum_{p' \in S(p)} Q(p'|p,x) W_{k-1}(p')$$

is weakly increasing in p if W_{k-1} is weakly increasing. The expected payoff of stopping, $\max_x p_x$, is also weakly increasing in p. Since $W_k(p)$ is the maximum of the expected payoffs of researching the different alternatives and the expected payoff of stopping, the claim follows.

Since $W_0(p) = \max_x p_x$ is weakly increasing, from the previous claim and induction it follows that W_k is weakly increasing for all $k \geq 0$. Since $\{W_k\}$ converges to V (theorem 3.1), it follows that V is weakly increasing as well. □

For $k \geq 1$, let

$$W_k^x(p) = -c + \delta \textstyle\sum_{p' \in S(p)} Q(p'|p,x) W_{k-1}(p')$$

be the expected payoff of researching alternative x if there are k remaining periods and the DM chooses optimally from next period on. Similarly, if there is no deadline, let

$$V^x(p) = -c + \delta \textstyle\sum_{p' \in S(p)} Q(p'|p,x) V(p')$$

be the expected expected payoff of researching alternative x if the DM adopts from next period on an optimal plan. By definition of W_k and V,

$$W_k(p) = \max\{\max_x p_x, \max_x W_k^x(p)\} \text{ for } k \geq 1$$

and

$$V(p) = \max\{\max_x p_x, \max_x V^x(p)\}.$$

To each $k \geq 1$ we can associate a (k-optimal) policy correspondence $\sigma_k : P \rightrightarrows A$ given by

$$\sigma_k(p) = \arg\max_{a \in A} \left[r(p,a) + \delta \sum_{p' \in S(p)} Q(p'|p,a) W_{k-1}(p') \right].$$

Note that $x \in \sigma_k(p)$ if and only if $W_k(p) = W_k^x(p)$. From theorem 3.1, if there is a finite deadline, a Markovian plan is optimal if and only if the associated policy satisfies $f_t(p) \in \sigma_{\tau - t}(p)$ for every $p \in P$.

Similarly, let the (stationary optimal) policy correspondence $\sigma^* : P \rightrightarrows A$ be defined by

$$\sigma^*(p) = \arg\max_{a \in A} \left[r(p,a) + \delta \sum_{p' \in S(p)} Q(p'|p,a) V(p') \right].$$

Note that $x \in \sigma^*(p)$ if and only if $V(p) = V^x(p)$. From theorem 3.1, if there is no deadline, a stationary plan is optimal if and only if the associated policy satisfies $f(p) \in \sigma^*(p)$ for every $p \in P$.

From corollary 3.1, $\sigma_k(p)$ is a set of approximate best responses to the problem with no deadline, and the approximation gets arbitrarily better as k increases. Moreover, since the action set is finite, for every p there is a finite $\underline{k}$ such that $\sigma_k(p) = \sigma^*(p)$ for all $k \geq \underline{k}$.

Since V and W_k for $k \geq 0$ are continuous, the objective functions in the problems defining $\sigma^*(p)$ and $\sigma_k(p)$ are continuous. It follows from Berge's (1963) maximum theorem that the stationary optimal policy correspondence and the k-optimal policy correspondences are nonempty and upper-semicontinuous.

The optimal policy correspondence partitions the state space P into regions in which different actions are optimal. In particular, we let

$$R_k = \{p \in P : s \notin \sigma_k(p)\} \quad \text{and} \quad R = \{p \in P : s \notin \sigma^*(p)\}$$

represent, respectively the regions in which research is strictly better that stopping in the problem with k periods ahead and in the stationary problem. Naturally, the research area is larger the longer the remaining time before the deadline.

Corollary 3.3. *R_k, R are nested, that is $R_1 \subseteq R_2 \subseteq \cdots \subseteq R$.*

Proof. Note that for $k \geq 1$, $p \in R_k$ if and only if there is some x such that

$$-c + \delta \sum_{p' \in S(p)} Q(p'|p,x) W_{k-1}(p') > \max_{x'} p_{x'},$$

and similarly $p \in R$ if and only if there is some x such that

$$-c + \delta \sum_{p' \in S(p)} Q(p'|p,x) V(p') > \max_{x'} p_{x'}.$$

From monotonicity (theorem 3.1), for every $p \in P$, $W_0(p) \leq W_1(p) \leq \cdots \leq V(p)$. The statement of the corollary follows. □

Regardless of the horizon, the value function is bounded by the probability that at least one of the alternative is good, that is $1 - \prod_x (1 - p_x)$. Hence, a simple upper bound to R is given by

$$R \subseteq \left\{ p : -c + \delta(1 - \textstyle\prod_x (1 - p_x)) > p_{[1]} \right\} \subset (0,1)^n.$$

We may wonder about the relation between the exploration problem and the bandit literature. We can treat the DM as choosing between $n+1$ arms, the first n arms representing the decision to explore each of the alternatives and the reminder arm representing the decision to stop and exploit the top alternative. The value of exploring each alternative and the value of stopping are, of course, not independent. The effect of exploring each alternative on the value of stopping depends not only on the state of that alternative (that is, the beliefs of the decision-maker about that alternative) but also on the states of the other alternatives. Thus, the value of exploring each alternative is not independent of the states of the other alternatives. Per contra, an index formulation (see e.g. equation 2 in Bergemann and Välimäki (2008)) would require writing the value of exploring each alternative as a function of two variables, the state of the alternative and an auxiliary index representing the value of other arms. In our problem, the value of the outside option to keep exploring the alternative changes depending on the states of the other alternatives.

4 Optimal exploration with one period ahead

When there is only one period before the deadline, the DM's problem is whether to stop immediately and take the best alternative, or to investigate one of the alternatives and then choose the best alternative given updated beliefs. In this section, we show that, in the optimal plan one period before the deadline, the DM is indifferent between investigating the top and the second alternative and will not investigate any alternative inferior to the top two. We also provide necessary and sufficient conditions for investigating any alternative to be optimal regardless of the number of periods ahead.

For any vector $p \in [0,1]^n$ we let $(p_{[1]}, \ldots, p_{[n]})$ be a permutation of p such that

$$p_{[1]} \geq p_{[2]} \geq \cdots \geq p_{[n]}.$$

For notational convenience, for given q define

$$w : [0,1]^2 \to [0,1]; \quad w(\rho, \nu) = q(\rho + \nu) - (2q-1)\rho\nu.$$

We have

Theorem 4.1.

$$W_1(p) = \max\{p_{[1]}, -c + \delta w(p_{[1]}, p_{[2]})\}, \quad R_1 = \left\{ p : p_{[2]} > \frac{(1-\delta q)p_{[1]} + c}{\delta q - \delta(2q-1)p_{[1]}} \right\}$$
$$\textit{and} \quad p \in R_1 \Rightarrow \sigma_1(p) = \{x \in X : p_x \geq p_{[2]}\}.$$

Proof. In view of corollary 3.2, we can restrict our attention to vectors of prior beliefs such that $p_1 \geq p_2 \geq \ldots \geq p_n$. The payoffs of researching alternatives 1 and 2 are, respectively,

$$W_1^1(p) = -c + \delta q_1^+(p) p_1^+ + \delta q_1^-(p) \max\{p_1^-, p_2\}$$

and

$$W_1^2(p) = -c + \delta q_2^+(p) \max\{p_2^+, p_1\} + \delta q_2^-(p) p_1.$$

If $p_2 > p_1^-$ (equivalently, $p_2^+ > p_1$), algebraic manipulation yields

$$W_1^1(p) = W_1^2(p) = -c + \delta w(p_1, p_2).$$

If instead $p_2 \leq p_1^-$ (equivalently, $p_2^+ \leq p_1$), we get

$$W_1^1(p) = W_1^2(p) = -c + \delta p_1.$$

The payoff of researching any alternative $\tilde{x}$ such that $p_{\tilde{x}} < p_2$ is

$$W_1^{\tilde{x}}(p) = -c + \delta q_{\tilde{x}}^+(p) \max\{p_{\tilde{x}}^+, p_1\} + \delta q_{\tilde{x}}^-(p) p_1.$$

If $p_{\tilde{x}}^+ > p_1$, we get

$$W_1^{\tilde{x}}(p) = -c + \delta w(p_1, p_{\tilde{x}}),$$

and if $p_{\tilde{x}}^+ \leq p_1$, we get

$$W_1^{\tilde{x}}(p) = -c + \delta p_1.$$

Hence, if $p_2^+ > p_{\tilde{x}}^+ > p_1$,

$$W_1^2(p) - W_1^{\tilde{x}}(p) = \delta(p_2 - p_{\tilde{x}})(q - p_1(2q-1)) > 0;$$

if $p_2^+ > p_1 \geq p_{\tilde{x}}^+$,

$$W_1^2(p) - W_1^{\tilde{x}}(p) = \delta(w(p_1, p_2) - p_1) > 0;$$

and if $p_1 \geq p_2^+ > p_{\tilde{x}}^+$,

$$W_1^2(p) - W_1^{\tilde{x}}(p) = 0.$$

From the preceding argument, researching the top two alternatives has the same expected payoff. Moreover, researching an alternative with a belief smaller than the top two is strictly dominated by researching either of the top two, except in the case in which researching any alternative has payoff equal to $-c + \delta p_1$. Since $-c + \delta p_1 < p_1$, however, in this case the only optimal action is to stop. It follows that it is never optimal to research an alternative with belief strictly below the top two.

Now researching any of the top two alternatives is strictly better than stopping if

$$W_1^1(p) = W_1^2(p) = -c + \delta w(p_1, p_2) > p_1,$$

or equivalently

$$p_2 > \frac{(1 - \delta q)p_1 + c}{\delta q - \delta(2q - 1)p_1}.$$

The statement of the theorem follows. □

We can now derive a necessary and sufficient condition for research to be optimal for some beliefs that holds regardless of whether there is or not a deadline, and regardless of the remaining time before the deadline.

Corollary 4.1. *For each $k \geq 1$, R_k is nonempty if and only if*

$$\frac{\delta q - 1/2}{\sqrt{2\delta(q - 1/2)}} > \sqrt{c},$$

and the same condition holds for R.

Proof. From theorem 4.1, R_1 is nonempty if

$$\frac{(1 - \delta q)p_{[1]} + c}{\delta q - \delta(2q - 1)p_{[1]}} < p_{[1]}$$

or equivalently if

$$\delta p_{[1]}^2(2q - 1) - p_{[1]}(2\delta q - 1) + c < 0.$$

Minimizing the expression in the left-hand side with respect to $p_{[1]}$, we obtain that there is $p_{[1]}$ such that the last inequality holds if and only if the upper bound on c in the statement of the corollary is satisfied.

Note that if R_1 is empty, then $\sigma_1(p) \ni \{s\}$ for all p, which implies $W_1 = W_0$. Recursively, it follows that $W_k = W_0$ for every k, and R_k is empty for all $k \geq 2$. From convergence (theorem 3.1), we get $V = W_0$. But then R is empty as well. Then the bound holds for R_k for $k \geq 1$ and for R as well. □

It is easy to check that increasing δ relaxes the bound in the statement of the corollary, and increasing q relaxes the bound if $\delta q \geq 1/2$. That is, quite intuitively, exploration is optimal if patience and the quality of information are large enough compared to the cost. In particular, $c < 1/4$ is required for exploration to be optimal.

We can rewrite the condition on prior beliefs for research to be optimal from theorem 4.1 as

$$q(p_1 + p_2 - 2p_1p_2) > \frac{p_{[1]} + c}{\delta} - p_1p_2.$$

It is easy to see that if R_1 is nonempty, then it is strictly increasing in q (since $p_1 + p_2 > 2p_1p_2$ for $1 > p_1 \geq p_2 > 0$) and in δ, and strictly decreasing in c.

To illustrate theorem 4.1, suppose $c = 0.002$, $\delta = 0.98$, and $n = 2$. Figure 4 depicts the research area for increasing values of the quality of information, namely $q = 2/3$ on top and $q = 5/6$ below, with p_1 and p_2 in the horizontal and vertical axis, respectively. In both cases, the area in between the blue lines represents R_1; the (partially overlapping) areas in green and orange represent additions to the research area for longer horizons which will be explained in the next section.

The following corollary establishes some useful facts about R_1 for future reference. Intuitively, it says that beliefs in R_1 cannot be too far off the diagonal, in the sense that bad news about the top alternative or good news about the second top alternative necessarily change the order between the two top.

Corollary 4.2.

(i) $p \in R_1$ *implies* $p^-_{[1]} < p_{[2]}$, *or equivalently,* $p^+_{[2]} > p_{[1]}$.

(ii) For every $p \in P$, $p^{[1]+} \notin R_1$.

(iii) $p^{[2]-} \in R_1$ *implies* $p^-_{[2]} < p_{[3]}$.

(iv) $p^{[2]-} \in R_1$ *implies* $p^{[1]-} \in R_1$.

Figure 4.1: Research areas for $q = 2/3$

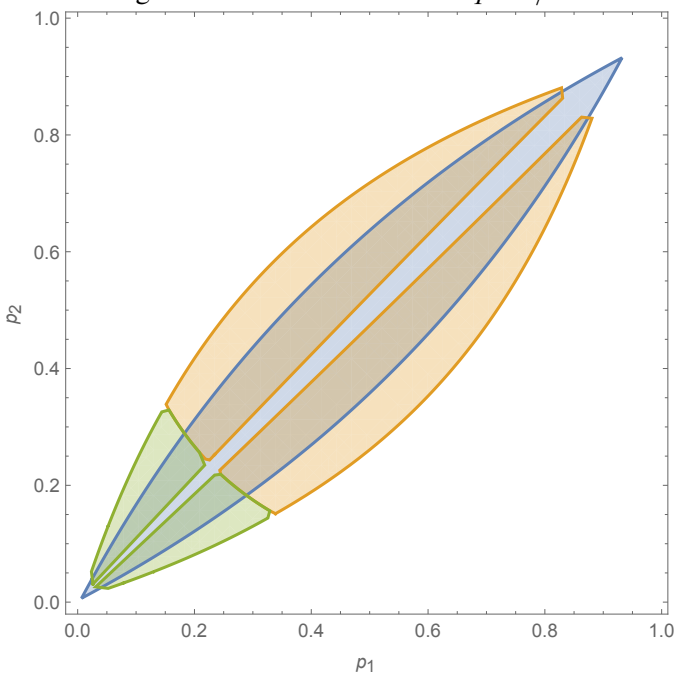
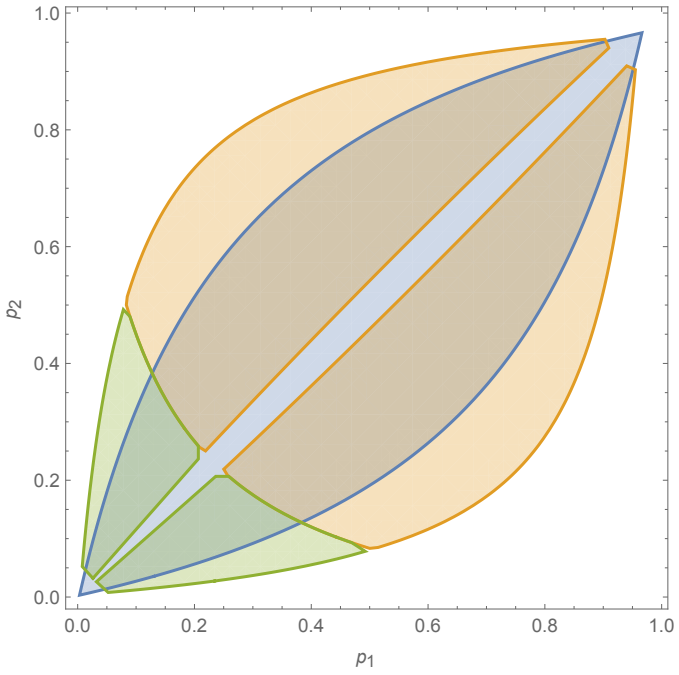


Figure 4.2: Research areas for $q = 5/6$



5 Optimal exploration with two periods ahead

In this section we show that, two periods before the deadline, if it is optimal to investigate any alternative, then either the DM prefers to investigate the top alternative and stop after good news, or prefers to investigate the second alternative and stop after bad news, or is indifferent between the top two and expects to stop the following period with probability one.

Define

$$\hat{W}_2(p) = q^+_{[1]}(p)p^+_{[1]} + q^-_{[1]}(p)(-c + \delta w(p_{[2]}, \max\{p^-_{[1]}, p_{[3]}\}))$$

and

$$\tilde{W}_2(p) = q^+_{[2]}(p)(-c + \delta w(p_{[1]}, p^+_{[2]})) + q^-_{[2]}(p)p_{[1]}.$$

These are the payoffs associated with investigating the first alternative and stopping after good news, and with investigating the second alternative and stopping after bad news. Obviously, $W_2(p) \geq W_2^{[1]}(p) \geq \hat{W}_2(p)$ and $W_2(p) \geq W_2^{[2]}(p) \geq \tilde{W}_2(p)$. Theorem 5.1 below can be summarized by the equation

$$W_2(p) = \max\{W_1(p), \hat{W}_2(p), \tilde{W}_2(p)\}.$$

Theorem 5.1 also provides the boundary between the regions in which $\hat{W}_2(p)$ and $\tilde{W}_2(p)$ are the optimal payoffs. In particular, for given q, δ, c define

$$B : P \to \Re; \quad B(p) = (1-\delta)p_{[1]}p_{[2]} - (1 - p_{[1]} - p_{[2]})c + \delta q^-_{[1]}(p)q^-_{[2]}(p)\max\{p_{[3]} - p^-_{[1]}, 0\}/(2q-1),$$

where we let $p_{[3]} = 0$ if $n = 2$. Note that $B(p) > 0$ if $c = 0$ or $p_{[1]} + p_{[2]} \geq 1$. The theorem establishes that $\hat{W}_2(p) \geq \tilde{W}_2(p)$ if and only if $B(p) \geq 0$. It follows that investigating the second alternative, with the expectation of investigating again only after good news, can be optimal only if there is a fixed cost of investigation and beliefs are sufficiently pessimistic. Intuitively, under these conditions, investigating the second alternative is more likely to end exploration earlier than investigating the top alternative, thus saving expected exploration costs: given $c > 0$ and $p_{[1]} + p_{[2]} < 1$, the difference in the two probabilities of stopping is $q^-_{[2]}(p) - q^+_{[1]}(p) > 0$. Note also that, even if the third alternative (if it exists) is not chosen optimally, beliefs about the third alternative affect which of the top two alternatives is optimal to investigate. This is because the third alternative may become relevant in the following period in case of bad news about the top alternative. In particular, the decision between investigating the top two alternatives cannot be settled optimally without taking into account beliefs about other alternatives.

We have

Theorem 5.1. *(i) If* $p^{[1]-}, p^{[2]+} \notin R_1$, *then* $W_2(p) = W_1(p)$, $p \in R_2 \Leftrightarrow p \in R_1$, *and*

$$p \in R_2 \;\Rightarrow\; \sigma_2(p) = \{x \in X : p_x \geq p_{[2]}\}.$$

(ii) If $p^{[1]-} \in R_1$ *or* $p^{[2]+} \in R_1$, *then*

$$W_2(p) = \begin{cases} \max\{p_{[1]}, \hat{W}_2(p)\} & \text{if} \quad B(p) \geq 0 \\ \max\{p_{[1]}, \tilde{W}_2(p)\} & \text{if} \quad B(p) \leq 0 \end{cases}$$

and

$$p \in R_2 \;\Rightarrow\; \sigma_2(p) = \begin{cases} \{x \in X : p_x = p_{[1]}\} & \text{if} \quad B(p) > 0 \\ \{x \in X : p_x \geq p_{[2]}\} & \text{if} \quad B(p) = 0 \\ \{x \in X : p_x = p_{[2]}\} & \text{if} \quad B(p) < 0 \end{cases}.$$

We can illustrate theorem 5.1 using figure 4. Recall that in the figure we assume $c = 0.002$, $\delta = 0.98$, and $n = 2$, with $q = 2/3$ above and $q = 5/6$ below. In both cases, the green area represents the initial beliefs in which it is optimal to investigate the second alternative, stop after bad news, and investigate either alternative after good news, so that $W_2(p) = \tilde{W}_2(p)$. The orange area represents the initial beliefs in which it is optimal to investigate the top alternative, stop after good news, and investigate either alternative after bad news, so that $W_2(p) = \hat{W}_2(p)$. The blue area that does not overlap the green or orange areas represents initial beliefs such that it is optimal to investigate either alternative and to stop afterwards regardless of whether news are good or bad, so that $W_2(p) = W_1(p) > p_{[1]}$. That is, the non-overlapped blue areas illustrate case (*i*) in the theorem, while the orange and blue areas illustrate case (*ii*). Note that the orange and green areas are the beliefs such that the expected utility of the decision maker increases with a larger horizon.

As figure 4 illustrates, if beliefs about the top two alternatives are nearly equal then it is optimal to investigate only once; if beliefs are asymmetric and relatively optimistic, it is optimal to investigate the top alternative with the expectation of stopping after good news; and if beliefs are asymmetric and relatively pessimistic, it is optimal to investigate the second alternative with the expectation of stopping after bad news. Remarkably, the boundary between pessimistic and optimistic beliefs is independent of the quality of information if there are only two alternatives.

To prove the theorem, we proceed via a series of lemmata. We first claim that any alternative below the top three according to initial beliefs is irrelevant for the optimal plan.

Lemma 5.1. *If* $p_x < p_{[3]}$, *then* $x \notin \sigma_2(p)$, *and moreover,* $p_{[x]} = p'_{[x]}$ *for all* $x \leq 3$ *implies* $W_2(p) = W_2(p')$.

Using corollary 3.2 and lemma 5.1, in the remainder of this section we suppose without loss of generality $n = 3$ and $p_1 \geq p_2 \geq p_3$. Lemma 5.2 establishes that investigating the third alternative is dominated by other plans. (Beliefs about the third alternative are not, however, irrelevant for deciding the optimal plan, as discussed later on.)

Lemma 5.2. *Either $\hat{W}_2(p) \geq W_2^3(p)$ (with strict inequality if $p_3 < p_1$), or $\tilde{W}_2(p) \geq W_2^3(p)$ (with strict inequality if $p_3 < p_2$), or $p_1 > W_2^3(p)$.*

Lemma 5.3 establishes that it cannot be optimal to investigate the second alternative if it is optimal to continue investigating after bad news.

Lemma 5.3. *If $p^{2-} \in R_1$, then $\hat{W}_2(p) \geq W_2^2(p)$, with strict inequality if $p_1 > p_2$.*

Lemma 5.4 tackles case (i) of the theorem, corresponding to the situation where investigating leads to stopping the following period with certainty.

Lemma 5.4. *If $p^{1-}, p^{2+} \notin R_1$ then $W_2(p) = W_1(p)$, $p \in R_2 \Leftrightarrow p \in R_1$, and $p \in R_2 \Rightarrow \sigma_2(p) = \{x \in X : p_x \geq p_2\}$.*

From the preceding arguments, if investigating does not lead to stopping with certainty the following period, it must be that either the DM investigates the top alternative and investigates again only after bad news, or the DM investigates the second alternative and investigates again only after good news. Lemma 5.5 ranks these two plans using the boundary condition $B(p)$.

Lemma 5.5. $\hat{W}_2(p) \gtreqless \tilde{W}_2(p) \iff B(p) \gtreqless 0$.

Proof of theorem 5.1. From lemmata 5.1 and 5.2, we know that investigating any alternative below the top two cannot be optimal. From corollary 4.2, we know that investigating the top alternative leads to stopping after good news. From lemma 5.3, we know that investigating the second alternative, if optimal, leads to stopping after bad news. Hence, the optimal plan is either stopping, or investigating either alternative and stopping for sure as if there were only one period ahead, or investigating the top alternative and investigating again only after bad news, or investigating the second alternative and investigating again only after good news. The case in which investigating is followed by stopping for sure is handled by lemma 5.4. Lemma 5.5 ranks the last two plans. □

We can check that both cases (i) and (ii) of the theorem are relevant for some beliefs for any parameters δ, q, c such that $R_1 \neq \emptyset$. In particular, if beliefs about the top two alternatives are close, and beliefs about the third alternative lag far behind,

or there is no third alternative, the optimal plan is to search only once either of the top two alternatives and then stop. If instead beliefs about the second and the third alternative are close enough, the optimal plan is to search the top alternative and search again only after bad news.

Corollary 5.1. *Consider $p \in R_1$. (i) If $p_{[1]} = p_{[2]}$ and $p^-_{[2]} \geq p_{[3]}$, then $W_2(p) = W_1(p) > p_{[1]}$. (ii) If $p_{[2]} = p_{[3]}$, then $W_2(p) = \hat{W}_2(p) > W_1(p)$.*

In sum, the previous two sections tell us that, with two or fewer periods before the deadline: (i) optimal exploration is restricted to the top two alternatives; (ii) starting from any initial beliefs such that it is optimal to explore, it is optimal to stop the following period after either good news, bad news, or both; (iii) if it is optimal to stop next period after good news but not after bad news, then it is optimal to explore only the top alternative; (iv) if it is optimal to stop next period after bad news but not after good news, then it is optimal to explore only the second alternative; and (v) if it is optimal to stop next period after both good news and bad news, then it is optimal to explore the top and the second top alternative (i.e. the DM is indifferent). In the next section we show that (iii), (v), and a slightly weaker version of (iv), hold for a longer horizon. The following section provides a numerical example to establish that (i) and (ii) do not generalize.

6 Optimal exploration with three periods ahead

We focus in this section on boundary points, that is, beliefs such that it is optimal to search and there is strictly positive probability of stopping the following period. We show that with three periods ahead of the deadline, it is optimal to investigate either of the top two alternatives at boundary points. We show by counterexample in the following section that this is not true for interior points, that is beliefs such that it is optimal to investigate again after both good and bad news.

Theorem 6.1. *If $x \in \sigma_3(p)$ and $p^{[x]+} \notin R_2$ or $p^{[x]-} \notin R_2$, then $p_x \geq p_{[2]}$.*

We proceed by a series of lemmata.

Lemma 6.1. *If $x \in \sigma_k(p)$ and $p^{[x]+}, p^{[x]-} \notin R_{k-1}$, then $p_x \geq p_{[2]}$.*

The proof follows from similar arguments in the previous two sections and extends to any given horizon.

Lemma 6.2. *If $x \in \sigma_3(p)$ and $p^{[x]-} \in R_2$, $p^{[x]+} \notin R_2$, then $p_x = p_{[1]}$.*

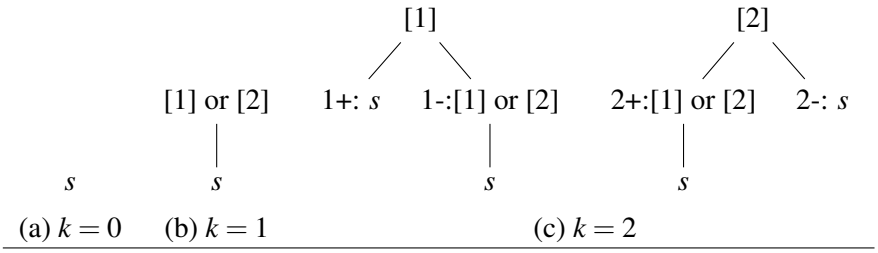
Lemma 6.3. *If $x \in \sigma_3(p)$ and $p^{[x]+} \in R_2$, $p^{[x]-} \notin R_2$, then $p_x \geq p_{[2]}$.*

Theorem 6.1 follows straightforwardly from the three lemmata above.

7 An example with three periods ahead

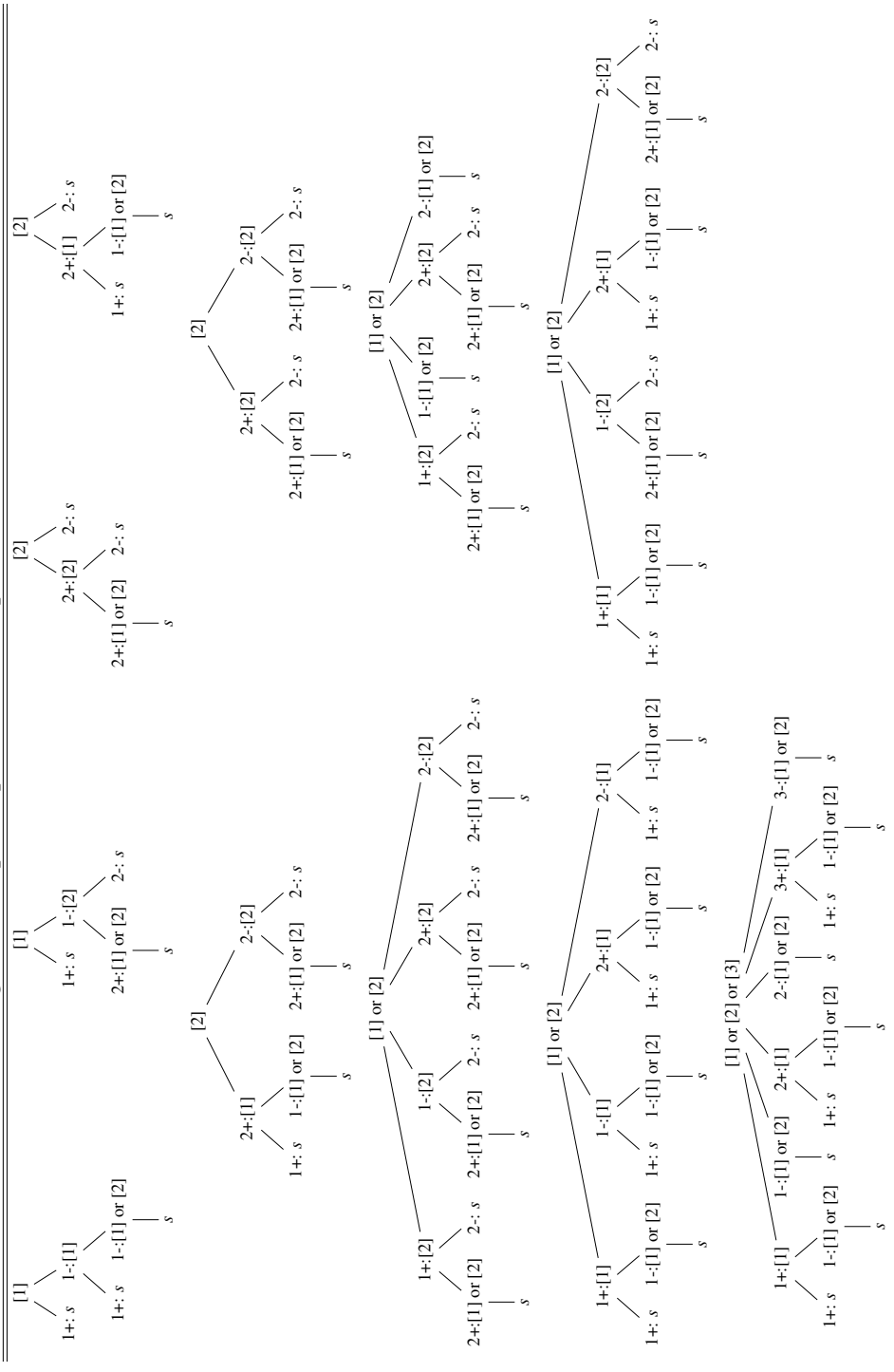
With up to two periods before the deadline, the optimal plan limits exploration to the top two alternatives and, starting from any initial beliefs such that it is optimal to explore, to stop with positive probability the following period. That is, for $k \leq 2$, every search point in the space of beliefs is a boundary point. For longer horizons, neither of these two properties hold. To illustrate this, suppose $n = 3$, $k = 3$, $\delta = 0.98$, $c = 0.002$ and $q = 0.7$. We illustrate in figures 7.1 and 7.2 all plans that are optimal for some beliefs $p \in (0,1)^3$. Plans are described by decision trees, starting with the initial decision node on top. The symbol s indicates that the optimal decision is to stop (and choose the alternative with the largest probability of being high quality). The symbol $[x]$ for $x \in \{1,2,3\}$ indicates that it is optimal to investigate the first, the second, or the third alternative in terms of their probability of being high quality as updated at that node. Note that because of the arrival of news about the alternatives, the order of the alternatives may change over the decision tree. A branch going down to $x-$ indicates the decision after bad news about alternative x, while a branch going down to $x+$ indicates the decision after good news about alternative x. When the decision is to stop after both good and bad news, a unique branch follows a decision node.

Figure 7.1: Optimal plans with three-period deadline ($k \leq 2$)



The shortest possible optimal plan is to stop immediately, as in figure 7.1(a). If it is optimal to explore at most one period ahead, as in figure 7.1(b), the optimal plan is as described by theorem 4.1. That is, it is optimal to explore either of the top two alternatives with indifference between the two. If it is optimal to explore at most two periods ahead, as in figure 7.1(c), the optimal plans are as described by theorem 5.1. That is, it is optimal either to investigate the top alternative, and to stop after good news, or to investigate the second alternative, and to stop after bad news.

Figure 7.2: Optimal plans with three-period deadline ($k = 3$)



If under the optimal plan there may be exploration three periods ahead, there is a variety of optimal plans, as described in figure 7.2. The first row of plans corresponds to boundary points: there is a positive probability of stopping after every investigation decision along each decision tree. In agreement with theorem 6.1, only the top two alternatives can be optimally investigated at boundary points. In particular, in agreement with lemma 6.2, if good news lead to stop, it is optimal to investigate the top alternative. The other rows correspond to plans such that after the first investigation decision there is probability one of investigating again. In particular, the last row corresponds to a plan such that it is optimal to explore indifferently either of the three alternatives.

In table 7.1 we have tabulated the frequency with which of the possible plans is optimal, under the assumption that the initial vector of beliefs p is distributed uniformly over the cube $(0,1)^3$. We have grouped the plans according to initial action and maximum length of exploration. An empty cell indicates impossibility.

Table 7.1: Frequency of optimal plans (percentage)

	s	[1]	[2]	[1] or [2]	[1], [2] or [3]
$k=0$	50.7620				
$k=1$				0.3712	
$k=2$		4.7606	0.1552	0.0000	
$k=3$		32.7538	1.3764	9.7060	0.1148

The most populated among the exploration cells in table 7.1 corresponds to plans with the maximum possible length of three periods and starting with investigating the top alternative. Among these plans, and among all plans involving exploration, the most frequent is the first plan to the left in the top row of figure 7.2, with a frequency of 31.8532%. It corresponds to investigating the top alternative, stopping after good news, and investigating again the top alternative after bad news, in which case good news lead to stop and bad news lead to investigate again. The second most populated cell corresponds to plans with the maximum length in which it is indifferent to start with the top or the second top alternative. Among these plans, the most frequent is the first plan to the left in the fourth row of figure 7.2, with a frequency of 8.7496%. It corresponds to investigating the top alternative after both good news and bad news, and then stopping after good news and investigating again after bad news. The third most populated cell corresponds to the third plan to the left in figure 7.1. It requires to investigate the top alternative, stop after good news, and investigate either of the top two after bad news. For the parameters chosen, then, under a uniform distribution of initial beliefs, stop-

ping most likely occurs after good news about the top alternative, except near the deadline.

8 Conclusions

Sequential experimentation problems are ubiquitous in economics, politics, and other fields (e.g. biology), and have received a great deal of attention over the years. The current paper studies a simple experimentation problem, focusing on the sequence of exploration, or alternatives to research, prior to making a final choice from among the available alternatives. Conditional on being chosen, any alternative from a given available set yields a fixed per period payoff of either one with some (interior) probability, or zero with the complementary probability. In any one period, a decision maker (DM) either chooses either to research an alternative at a fixed cost or make a final choice of an available alternative. Research yields only an informative but noisy signal of the value of the alternative under consideration (either "good news" signaling a positive payoff, or "bad news" signaling a zero payoff); payoffs from any alternative accrue only if the DM stops exploration and chooses the alternative. Exploring any given alternative repeatedly is feasible but resuming exploration after making a choice is prohibitively costly.

Save for characterizing when it is optimal to research (corollary 4.1), identifying the optimal policy in general seems analytically intractable. The complication here is that the (endogenous) value of exploring any alternative in a given period is not independent of the values of other alternatives, despite payoff values of the feasible alternatives being statistically independent. There is thus no available Gittins-type of index theorem available. A characterization of the optimal policy function at boundary points,[6] however, is available for when there are at most three remaining search periods ($k \leq 3$). In particular, all research points are boundary points when there are at most two remaining search periods (theorems 4.1 and 5.1), although this is not true for ($k \geq 3$) (figure 7.2).

In any period, order alternatives by decreasing likelihood that they yield a payoff of one. Then the main results for when there are at most three available search periods are easily summarized. Assume the optimal plan is followed in the current period. If it is optimal to investigate an alternative x now but to stop next period after

1. both good news and bad news, then x must be one of the top two alternatives for any number of search periods ($k \geq 1$), with the DM being indifferent between the top and second ranked alternatives when $k = 1$;

[6] That is, beliefs such that it is optimal to research and there is strictly positive probability of stopping the following period.

2. good news but not after bad news, then x must be the top ranked alternative if $k \in \{2,3\}$;

3. bad news but not good news, then x must be one of the top two alternatives when $k = 3$; when $k = 2$, x must be the second best alternative if beliefs are 'sufficiently pessimistic' and the top alternative if beliefs are 'sufficiently optimistic.'

(1) is a general property of optimal exploration regardless of the search horizon. (2) and (3) essentially say that, when there are at most three remaining search periods, receiving good news about an alternative implies it is optimal either to stop and choose that alternative, or to explore it further in the next period; receiving bad news about the alternative implies it is optimal either to stop and take the best alternative, or to explore that best alternative in the next period. These properties seem intuitive and we conjecture they generalize to the arbitrary search horizon case.

Two main motivations for the model are the nest-searches of eusocial insects (see e.g. Seeley, 2010) and, closer to home, searching for the best policy to implement some given collective goal. In this latter case, the model can be interpreted as a common value committee deliberation, whereby the committee engages in researching alternatives prior to reaching a decision. Natural extensions in this context include relaxing the common value assumption to some extent, for example by having individuals independently choose whether to pay the fixed cost to explore an alternative in any period (presuming only one such search per period) and reporting the findings back to the committee. In this case, reporting signals truthfully is individually rational but there is clearly a free-rider problem in regard to searching. We expect search will stop too soon relative to the optimal plan. Similarly, if discount rates are individually specific and the committee votes over when to stop search, the optimality of search depends on both the distribution of discount rates among individuals and the distribution of likelihoods that alternatives have positive value in any period (see Chan et al., 2017). We leave such extensions for another time.

Appendix: Additional proofs

Proof of corollary 4.2:

Suppose without loss of generality $p_1 \geq p_2 \geq \ldots \geq p_n$. For part (i), from theorem 4.1, if $p \in R_1$ then $W_1^1(p) = W_1^2(p) > p_1$. Using $\delta < 1$ and $c \geq 0$, $W_1^1(p) > p_1$ implies $p_1^- < p_2$ (that is, researching alternative 1 is only worth if bad news lead to a change in the best alternative).

For part (ii), suppose $p^{1+} \in R_1$. Then using part (i), $p_1 = (p_1^+)^- < p_2$, a contradiction.

For part (iii), suppose $p_2^- \in R_1$ and $p_2^- \geq p_3$. Then, using part (i), $p_1^- = (p_{[1]}^{2-})^- < p_{[2]}^{2-} = p_2^-$, a contradiction.

For part (iv), note that the condition for $p^{2-} \in R_1$ is

$$\max\{p_2^-, p_3\} > \frac{(1-\delta q)p_1 + c}{\delta q - \delta(2q-1)p_1},$$

while the condition for $p^{1-} \in R_1$, depending on whether $p_1^- \leq p_2$ or $p_1^- > p_2$ is either

$$\max\{p_1^-, p_3\} > \frac{(1-\delta q)p_2 + c}{\delta q - \delta(2q-1)p_2} \quad \text{or} \quad p_1^- > \frac{(1-\delta q)p_2 + c}{\delta q - \delta(2q-1)p_2},$$

which is weaker than the condition above in either case. □

Proof of lemma 5.1:

From theorem 4.1, we know that $p_{[1]} = p'_{[1]}, p_{[2]} = p'_{[2]}$ implies $W_1(p) = W_1(p')$. For any alternative x, we have

$$W_2^x(p) = -c + \delta(q_x^+ W_1(p_x^+, p_{-x}) + q_x^- W_1(p_x^-, p_{-x})).$$

Note that $p_{\hat{x}} = p_{[3]} > p_x$ implies $q_{\hat{x}}^+(p) > q_x^+(p)$, $q_{\hat{x}}^-(p) < q_x^-(p)$, $W_1(p_{\hat{x}}^+, p_{-\hat{x}}) \geq W_1(p_x^+, p_{-x})$ and $W_1(p_{\hat{x}}^-, p_{-\hat{x}}) = W_1(p_x^-, p_{-x})$. But then $W_2^{\hat{x}}(p) > W_2^x(p)$, which implies $x \notin \sigma_2(p)$. Hence, x will not be in the top two alternatives the following period in the optimal plan. It follows that it does not affect the expected utility of the DM. □

Proof of lemma 5.2:

It is easy to see that if $p_3 \leq p_2^-$ (equivalently, $p_3^+ \leq p_2$), then investigating the third alternative is dominated. We suppose henceforth that $p_3 > p_2^-$. Suppose first

that $p^{3-}, p^{3+} \in R_1$. Then

$$\begin{aligned}
&(\hat{W}_2(p) - W_2^3(p))/\delta \\
&= q_1^+(p)p_1^+ + q_1^-(p)(-c + \delta w(p_2, \max\{p_1^-, p_3\}) \\
&\qquad -q_3^+(p)(-c + \delta w(p_1, p_3^+)) - q_3^-(p)(-c + \delta w(p_1, p_2)) \\
&> q_1^+(p)(-c + \delta p_1^+) + q_1^-(p)(-c + \delta w(p_2, \max\{p_1^-, p_3\})) \\
&\qquad -q_3^+(p)(-c + \delta w(p_1, p_3^+)) - q_3^-(p)(-c + \delta w(p_1, p_2)) \\
&\propto q_1^+(p)p_1^+ + q_1^-(p)w(p_2, \max\{p_1^-, p_3\}) - q_3^+(p)w(p_1, p_3^+) - q_3^-(p)w(p_1, p_2) \\
&= [qp_1 + q_1^-(p)(qp_2 + q_2^-(p)\max\{p_1^-, p_3\})] - [qp_1 + q_1^-(p)(qp_3 + q_3^-(p)p_2)] \\
&\geq 0.
\end{aligned}$$

Suppose next $p^{3-} \in R_1$ but $p^{3+} \notin R_1$. Then

$$\begin{aligned}
&(\hat{W}_2(p) - W_2^3(p))/\delta \\
&= q_1^+(p)p_1^+ + q_1^-(p)(-c + \delta w(p_2, \max\{p_1^-, p_3\}) \\
&\qquad -q_3^+(p)\max\{p_1, p_3^+\} - q_3^-(p)(-c + \delta w(p_1, p_2)) \\
&\geq qp_1 - q_3^+(p)\max\{p_1, p_3^+\} - \delta(q_3^-(p)w(p_1, p_2) - q_1^-(p)w(p_2, \max\{p_1^-, p_3\})).
\end{aligned}$$

with strict inequality if $p_1 > p_3$. We have two cases; if $p_1 \leq p_3^+$,

$$\begin{aligned}
&(\hat{W}_2(p) - W_2^3(p))/\delta \\
&\geq q(p_1 - p_3) - \delta(q_3^-(p)w(p_1, p_2) - q_1^-(p)w(p_2, p_3)) \\
&> (q - \delta q^2)(p_1 - p_3) > 0.
\end{aligned}$$

If instead $p_1 > p_3^+$,

$$\begin{aligned}
&(\hat{W}_2(p) - W_2^3(p))/\delta \\
&> (2q-1)p_1(1 - p_3) - \delta(q_3^-(p)w(p_1, p_2) - q_1^-(p)w(p_2, p_1^-)) > 0.
\end{aligned}$$

Suppose $p^{3-} \notin R_1$ but $p^{3+} \in R_1$. Then

$$\begin{aligned}
&(\tilde{W}_2(p) - W_2^3(p))/\delta \\
&\quad = q_2^+(p)(-c + \delta w(p_1, p_2^+)) + q_2^-(p)p_1 \\
&\qquad -q_3^+(p)(-c + \delta w(p_1, p_3^+)) - q_3^-(p)p_1 \geq 0,
\end{aligned}$$

with strict inequality if $p_2 > p_3$.

Finally, suppose $p^{3-}, p^{3+} \notin R_1$ but $W_2^3(p) \geq p_1$. Note that $W_2^3(p) \geq p_1$ implies $p_3^+ > p_1$ and hence $p_2^+ > p_1$. Then

$$\begin{aligned}
&(\tilde{W}_2(p) - W_2^3(p))/\delta \\
&\quad \geq q_2^+(p)p_2^+ + q_2^-(p)p_1 - q_3^+(p)p_3^+ - q_3^-(p)p_1 \\
&\quad = q_1^-(p)(p_2 - p_3) \geq 0,
\end{aligned}$$

with strict inequality if $p_2 > p_3$. □

Proof of lemma 5.3:

From corollary 4.2(iii), we know that $p^{2-} \in R_1$ implies $p_3 > p_2^-$. Note that then $W_2^2(p^{2-}) = -c + \delta w(p_1, p_3)$. Suppose first that $p^{2+} \in R_1$. Then:

$$\begin{aligned}
&(\hat{W}_2(p) - W_2^2(p))/\delta \\
&\quad \geq q_1^+(p)p_1^+ + q_1^-(p)(-c + \delta w(p_2, p_3)) - q_2^+(p)W_1(p^{2+}) - q_2^-(p)W_1(p^{2-}) \\
&\quad > q_1^+(p)(-c + \delta p_1^+) + q_1^-(p)(-c + \delta w(p_2, p_3)) \\
&\qquad\qquad -q_2^+(p)(-c + \delta w(p_1, p^{2+})) - q_2^-(p)(-c + \delta w(p_1, p_3)) \\
&\quad \propto\ q_1^+(p)p_1^+ + q_1^-(p)w(p_2, p_3) - q_2^+(p)w(p_2^+, p_1) - q_2^-(p)w(p_1, p_3) \\
&\quad = q_1^-(p)[w(p_2, p_3) - qp_2 - q_2^-(p)p_3] \\
&\quad = 0.
\end{aligned}$$

Suppose instead that $p^{2+} \notin R_1$. Then:

$$\begin{aligned}
&(\hat{W}_2(p) - W_2^2(p))/\delta \\
&\quad \geq q_1^+(p)p_1^+ + q_1^-(p)(-c + \delta w(p_2, p_3)) - q_2^+(p)p^{2+} - q_2^-(p)W_1(p^{2-}) \\
&\quad = q(p_1 - p_2) + (q_2^-(p) - q_1^-(p))c + \delta(q_1^-(p)w(p_2, p_3) - q_2^-(p)w(p_1, p_3)) \\
&= (q + (2q-1)c)(p_1 - p_2) - \delta(q - (2q-1)p_3)(p_1 - p_2) \\
&= (p_1 - p_2)[(2q-1)(c + \delta p_3) + q(1 - \delta q)] \geq 0,
\end{aligned}$$

with strict inequality if $p_1 > p_2$. □

Proof of lemma 5.4:

From corollary 4.2(ii), we have $p^{1+} \notin R^1$. Thus, $p^{1-} \notin R_1$ implies $W_2^1(p) = W_1^1(p) = W_1(p)$, where the last equality follows from theorem 4.1. Also from corollary 4.2(iv), $p^{1-}, p^{2+} \notin R_1$ implies $p^{2-}, p^{2+} \notin R_1$, which implies $W_2^2(p) = W_1^2(p) = W_1(p)$. In either case, $p \in R_2$ if and only if $p \in R_1$. As in theorem 4.1, investigating either of the top two alternatives is optimal if it is worth investigating.□

Proof of lemma 5.5:

Suppose $p_1^- \geq p_3$. (The third alternative is irrelevant.) Then

$$\begin{aligned}
&(\hat{W}_2(p) - \tilde{W}_2(p))/\delta \\
&\quad = q_1^+(p)p_1^+ - q_2^-(p)p_1 + (q_1^-(p) - q_2^+(p))c \\
&\qquad\qquad +\delta(q_1^-(p)w(p_2, p_1^-) - q_2^+(p)w(p_1, p_2^+)) \\
&\quad = (2q-1)(p_1p_2 - (1 - p_1 - p_2)c) \\
&\qquad +\delta(q_1^-(p)qp_2 + q(1-q)p_1 - q_2^+(p)qp_1 - q^2p_2 + (2q-1)^2) \\
&\quad = (2q-1)((1-\delta)p_1p_2 - (1 - p_1 - p_2)c).
\end{aligned}$$

Suppose instead $p_1^- < p_3$. (The third alternative is relevant.) Using the preceding equation

$$\begin{aligned}\hat{W}_2(p) - \tilde{W}_2(p))/\delta &= ((2q-1)((1-\delta)p_1p_2 - (1-p_1-p_2)c) \\ &\quad +\delta q_1^-(p)(w(p_2,p_3) - w(p_2,p_1^-)) \\ &= (2q-1)((1-\delta)p_1p_2 - (1-p_1-p_2)c) \\ &\quad +\delta q_1^-(p)q_2^-(p)(p_3 - p_1^-).\end{aligned}$$

From both cases we get that

$$\begin{aligned}(\hat{W}_2(p) - \tilde{W}_2(p))/\delta = (2q-1)((1-\delta)p_1p_2 - (1-p_1-p_2)c) \\ +\delta q_1^-(p)q_2^-(p)\max\{p_3 - p_1^-, 0\} = (2q-1)B(p),\end{aligned}$$

as desired. □

Proof of corollary 5.1: For part (i), note that, given the premise, $p^{[2]+} = p^{[1]+} \notin R_1$ (by corollary 4.2(ii)), and $p^{[2]-} \notin R_1$ (by corollary 4.2(iii)). The desired result follows from theorem 5.1(i). For part (ii), note that, given the premise, $p_{[1]}^{[2]-} = p_{[1]}$ and $p_{[2]}^{[2]-} = p_{[2]}$, so that $p \in R_1$ implies $p^{[2]-} \in R_1$. The desired result follows from theorem 5.1(ii). □

Proof of lemma 6.2:

We proceed by contradiction. Assume without loss of generality that $p_1 = \max_{y \in X} p_y$, and suppose there is some p such that it is optimal to explore some alternative x with $p_x < p_1$, and moreover next period it is optimal to stop after good news but not after bad news. If $p_1 \geq p_x^+$, it cannot be optimal to stop after good news about x but not after bad news, since the payoff of exploring is larger after good news, and after bad news alternative 1 is still available. Hence $p_x^+ > p_1$ and the plan's payoff is

$$q_x^+ p_x^+ + q_x^- W_2(p^{x-}).$$

Since there is exploration at p^{x-}, we have $W_2(p^{x-}) \geq p_1$. Let the payoff of the optimal plan at $p \in P$ conditional on the true value of alternatives 1 and x be given by $u(p|\omega_1, \omega_x)$, and let $u(p^{x-}|1,1) = v$, $u(p^{x-}|0,0) = w$, $u(p^{x-}|0,1) = r$ and $u(p^{x-}|1,0) = t$. Thus, $W_2(p^{x-}) = p_1p_xv + (1-p_1)(1-p_x)w + (1-p_1)p_xr + p_1(1-p_x)t \geq p_1$. We claim that $r \geq 0$. It is simple to show that $W_2(p^{x-}) \geq p_1$ and $r \geq w$ imply $r \geq 0$; we prove below that $r \geq w$.

Now consider switching alternatives x and 1 at beliefs p and any posterior beliefs; this plan payoff is

$$q_1^+ p_1^+ + q_1^- \tilde{W}_2(p^{1-}),$$

where $\tilde{W}_2(p^{1-}) = p_1 p_x v + (1-p_1)(1-p_x)w + (1-p_1)p_x t + p_1(1-p_x)r$. The optimal plan and this deviation have the same expected payoffs, qw and $q+(1-q)v$ in states of the world $(\omega_1,\omega_x) = (0,0),(1,1)$, respectively. Using the states of the world $(\omega_1,\omega_x) = (1,0),(0,1)$, we get

$$\begin{aligned}
&(q_1^+ p_1^+ + q_1^- \tilde{W}_2(p^{1-})) - (q_x^+ p_x^+ + q_x^- W_2(p^{x-})) \\
&\qquad = (p_1(1-p_x)(q+(1-q)r) + (1-p_1)p_x qt) \\
&\qquad\qquad - (p_1(1-p_x)qt) + (1-p_1)p_x(q+(1-q)r)) \\
&\qquad = p_1(1-p_x)(q+(1-q)r-qt) + (1-p_1)p_x(qt-q-(1-q)r) \\
&\qquad = (p_1-p_x)(q(1-t)+(1-q)r) > 0,
\end{aligned}$$

where we use $t<1$ and $r \geq 0$. But this contradicts x being optimal.

We still have to prove that $r \geq w$. More generally, for any pair of alternatives y,x, any $p \in P$ and any horizon k, let the payoff of the optimal plan at p conditional on the true value of alternatives x and y be given by $u_k(p|\omega_y,\omega_x)$. We claim that, for $k \leq 2$,

$$u_k(p|1,0) \geq u_k(p|0,0).$$

The result is immediate if the optimal decision at p with horizon k is to stop. The result is also easy to establish for $k=1$. Consider the case $k=2$. If the decision is to investigate any alternative $z \neq y,x$, we have

$$\begin{aligned}
&u_2(p|1,0) - u_2(p|0,0) \\
&\qquad = q^+(p_z)(u_1(p^{z+}|1,0) - u_1(p^{z+}|0,0)) \\
&\qquad\qquad + q^-(p_z)(u_1(p^{z-}|1,0) - u_1(p^{z-}|0,0)) \geq 0.
\end{aligned}$$

The desired result from $u_1(p|1,0) \geq u_1(p|0,0)$. The argument is similar if the decision is to investigate x. If the decision instead is to investigate y, we have

$$\begin{aligned}
&u_2(p|1,0) - u_2(p|0,0) \\
&\qquad = (qu_1(p^{y+}|1,0) + (1-q)u_1(p^{y-}|1,0)) \\
&\qquad\qquad - (qu_1(p^{y-}|0,0) + (1-q)u_1(p^{y+}|0,0)). \quad (3)
\end{aligned}$$

Since we are investigating y and we are in $k=2$, we must have either $p_y = p_{[1]}$ or $p_y = p_{[2]} < p_{[1]}$. In either situation, we do not need to consider the case in which the optimal policy is to stop after both good and bad news, since then $u_2(p|1,0) - u_2(p|0,0) = u_1(p|1,0) - u_1(p|0,0) \geq 0$.

Suppose first that $p_y = p_{[1]}$. Then equation 3 becomes

$$\begin{aligned}
&u_2(p|1,0) - u_2(p|0,0) \\
&= q + (1-q)u_1(p^{y-}|1,0) - qu_1(p^{y-}|0,0) \\
&= q + (1-q)(u_1(p^{y-}|1,0) - u_1(p^{y-}|0,0)) - (2q-1)u_1(p^{y-}|0,0) \\
&> q - (2q-1) > 0,
\end{aligned}$$

where we use $1 > u_1(p|1,0) \geq u_1(p|0,0)$ for all p.

Suppose instead that $p_y = p_{[2]} < p_{[1]}$. If $p_x < p_{[1]}$, equation 3 becomes

$$\begin{aligned}
&u_2(p|1,0) - u_2(p|0,0) \\
&= qu_1(p^{y+}|1,0) + (1-q)p_{[1]} - qp_{[1]} - (1-q)u_1(p^{y+}|0,0) \\
&= q(u_1(p^{y+}|1,0) - p_{[1]}) + (1-q)(p_{[1]} - u_1(p^{y+}|0,0)).
\end{aligned}$$

We have $u_1(p^{y+}|1,0) - p_{[1]} = -c + \delta(q + (1-q)p_{[1]}) - p_{[1]}$ which, using the characterization of R_1 in theorem 4.1, is positive given $p^{y+} \in R_1$. Similarly $p_{[1]} - u_1(p^{y+}|0,0) = p_{[1]} + c - \delta(qp_{[1]}) > 0$.

Finally, if $p_x = p_{[1]}$, equation 3 becomes

$$u_2(p|1,0) - u_2(p|0,0) = qu_1(p^{y+}|1,0) - (1-q)u_1(p^{y+}|0,0)).$$

If the DM searches y, we get

$$u_2(p|1,0) - u_2(p|0,0) = q(-c + \delta q) + (1-q)c.$$

From corollary 4.1, we have $c < 1/4$ and $\delta q > 1/2$, so the above expression is positive. Similarly, if the DM searches x, we get

$$u_2(p|1,0) - u_2(p|0,0) = q > 0,$$

which finishes the proof. □

Proof of lemma 6.3:

As in the previous proof, we proceed by contradiction. Assume without loss of generality that $p_1 = p_{[1]}$ and $p_{[2]} = p_2$, and suppose there is some p such that it is optimal to explore some alternative x with $p_x < p_2$, and moreover next period it is optimal to stop after bad news but not after good news. The plan's payoff is

$$q_x^+ W_2(p^{x+}) + q_x^- p_1.$$

Let the payoff of the optimal plan at $p \in P$ conditional on the true value of alternatives 2 and x be given by $u(p|\omega_2, \omega_x)$, and let $u(p^{x+}|1,1) = v$, $u(p^{x+}|0,0) = w$,

$u(p^{x+}|0,1) = r$ and $u(p^{x+}|1,0) = t$. Thus, $V_2(p^{x+}) = p_2 p_x v + (1-p_2)(1-p_x)w + (1-p_2)p_x r + p_2(1-p_x)t \geq p_1$.

Now consider switching alternatives x and 2 at beliefs p and any posterior beliefs; this plan payoff is

$$q_2^+ \tilde{W}_2(p^{2+}) + q_2^- p_1,$$

where $\tilde{W}_2(p^{2+}) = p_2 p_x v + (1-p_2)(1-p_x)w + (1-p_2)p_x t + p_2(1-p_x)r$. The optimal plan and this deviation have the same expected payoffs, $(1-q)w + qp_1$ and $qv + (1-q)p_1$ in states of the world $(\omega_2, \omega_x) = (0,0), (1,1)$, respectively. Using the states of the world $(\omega_2, \omega_x) = (1,0), (0,1)$, we get

$$\begin{aligned}
&(q_2^+ \tilde{W}_2(p^{2+}) + q_2^- p_1) - (q_x^+ W_2(p^{x+}) + q_x^- p_1) \\
&\quad = (p_2(1-p_x)(qr + (1-q)p_1) + (1-p_2)p_x((1-q)t + qp_1)) \\
&\qquad - (p_2(1-p_x)(qt + (1-q)p_1) + (1-p_2)p_x((1-q)r + qp_1)) \\
&\quad = p_2(1-p_x)q(r-t) + (1-p_2)p_x(1-q)(t-r) \\
&\quad = (p_2(1-p_x)q - (1-p_2)p_x(1-q))(r-t),
\end{aligned}$$

which is positive if $r > t$. We prove below that either $r > t$, which contradicts x being optimal, or $r = t$ so that switching alternatives x and 2 leads to the same payoff, but the plan after the switch is itself suboptimal, which contradicts x being optimal.

Suppose first that 1) $p_x^+ \geq p_1$. 1.1) Suppose the decision at p_x^+ with two periods ahead is to stop after both good and bad news. In that case it is optimal to explore either x or 1. If we explore x we get $r = q + (1-q)p_1 > qp_1 = t$. If instead we explore 1 we get $r = qp_1 + q_1^- > qp_1 = t$. 1.2) Suppose the decision at p_x^+ with two periods ahead is to stop after good news and search again after bad news. In this case it is optimal to explore x. We get $r = q + (1-q)W_1(p) > qW_1(p) = t$, where p is the original belief vector. 1.3) Suppose the decision at p_x^+ with two periods ahead is to stop after bad news and search again after good news. In that case it is optimal to explore 1. We get $r = q_1^- + q_1^+ u_1(p^{1+,x+}|0,1)$ and $t = q_1^+ u_1(p^{1+,x+}|1,0)$, where $r > t$ since $u_1(p^{1+,x+}|0,1) > u_1(p^{1+,x+}|1,0)$ regardless of whether the DM explore 1 or x at $p^{1+,x+}$ with one period ahead.

Suppose 2) $p_2 < p_x^+ < p_1$. 2.1) If the decision at p_x^+ with two periods ahead is to stop after both good and bad news, we get the same result as in the previous paragraph.

2.2) Suppose the decision at p_x^+ with two periods ahead is to stop after good news and search again after bad news. In this case it is optimal to explore 1. We get $r = qp_1 + (1-q)u_1(p^{1-,x+}|0,1)$ and $t = qp_1 + (1-q)u_1(p^{1-,x+}|1,0)$. If the DM explores x or 1 at $p^{1-,x+}$ with one period ahead we get the desired result; if the DM explores 2 we get $r = t$ so that switching alternatives x and 2 leads to the same

payoff. Note that after switching the DM explores x at $p^{1-,2+}$, which is suboptimal since by the premise $p_x < p_1^-$ and by theorem 4.1 it is not optimal to explore third alternatives with one period ahead.

2.3) Suppose the decision at p_x^+ with two periods ahead is to stop after bad news and search again after good news. In that case it is optimal to explore x. We get $r = qu_1(p^{x++}|0,1) + (1-q)p_1$ and $t = (1-q)u_1(p^{x++}|1,0) + qp_1$. If we explore 1 at p^{x++} with one period ahead, we get $u_1(p^{x++}|0,1) = qp_1 + (1-q)$ and $u_1(p^{x++}|1,0) = qp_1$ which yields $r > t$. If we explore x at p^{x++} with one period ahead, we get $u_1(p^{x++}|0,1) = q + (1-q)p_1$ and $u_1(p^{x++}|1,0) = qp_1$ which yields $r > t$.

3) Suppose $p_x^+ \leq p_2$. 3.1) It cannot be the case that the decision at p_x^+ with two periods ahead is to stop after both good and bad news since then positive news about x is irrelevant and we can explore either 1 or 2 three periods ahead and stop after both good and bad news.

3.2) Suppose the decision at p_x^+ with two periods ahead is to stop after good news and search again after bad news. In this case it is optimal to explore 1. We get $r = qp_1 + (1-q)u_1(p^{1-,x+}|0,1)$ and $t = qp_1 + (1-q)u_1(p^{1-,x+}|1,0)$. The DM cannot explore optimally x at $p^{1-,x+}$ with one period ahead since $p^+ < p_2$. If the DM explores 1 we get $r = qp_1 + (1-q)(qp_1 + q_1^-) > qp_1 + (1-q)qp_1$. If the DM explores 2 at $p^{1-,x+}$ with one period ahead, we get $r = t$ so that switching alternatives x and 2 leads to the same payoff. Note that after switching the DM explores x at $p^{1-,2+}$, which is suboptimal since by the premise $p_x < p_1^-$ and by theorem 4.1 it is not optimal to explore third alternatives with one period ahead.

3.3) Suppose the decision at p_x^+ with two periods ahead is to stop after bad news and search again after good news. In that case it is optimal to explore 2. But then after bad news we choose 1 and after good news we explore either 1 or 2, so searching x in the first place cannot be optimal. □

References

Adam, K., 2001. Learning while searching for the best alternative. Journal of Economic Theory 101, 252–280.

Banks, J., Sundaram, R., 1994. Switching costs and the Gittins index. Econometrica 62 (3), 687–694.

Bergemann, D., Välimäki, J., 2008. Bandit problems. In: Durlauf, S., Blume, L. (Eds.), The New Palgrave Dictionary of Economics, 2nd Edition. Vol. 1. Palgrave Macmillan, New York, pp. 336–340.

Berry, D. A., Fristedt, B., 1985. Bandit Problems: Sequential Allocation of Experiments. Chapman and Hall, London.

Blackwell, D., 1965. Discounted dynamic programming. The Annals of Mathematical Statistics 36 (1), 226–235.

Blackwell, D., Girshick, M. A., 1954. The theory of games and statistical decisions. Wiley, New York.

Bolton, P., Harris, C., 1999. Strategic experimentation. Econometrica 67 (2), 349–374.

Chan, J., Lizzeri, A., Suen, W., Yariv, L., 2017. Deliberating collective decisions. Review of Economic Studies 85 (2), 929–963.

Doval, L., 2018. Whether or not to open Pandora's box. Journal of Economic Theory 175, 127–158.

Forand, J.-G., 2015. Keeping your options open. Journal of Economic Dynamics and Control 53, 47–68.

Gittins, J., Glazebrook, K., Weber, R., 2011. Multi-Armed Bandit Allocation Indices, 2nd Edition. Wiley.

Gittins, J. C., 1979. Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B (Methodological) 41 (2), 148–177.

Glazebrook, K. D., 1979. Stoppable families of alternative bandit processes. Journal of Applied Probability 16 (4), 843–854.

Ke, T. T., Villas-Boas, J. M., 2018. Optimal learning before choice, MIT and UC Berkeley.

Keller, G., Rady, S., Cripps, M., 2005. Strategic experimentation with exponential bandits. Econometrica 73 (1), 39–68.

McCall, J., 1970. Economics of information and job search. Quarterly Journal of Economics 84, 113–126.

Mortensen, D., 1986. Job search and labor market analysis. In: Ashenfelter, O. C., Layard, R. (Eds.), Handbook of Labor Economics. Vol. 2. North Holland, pp. 849–919.

Moscarini, G., Smith, L., 2001. The optimal level of experimentation. Econometrica 69 (6), 1629–1644.

Olszewski, W., Weber, R., 2016. A more general Pandora rule? Journal of Economic Theory 160, 429–437.

Seeley, T. D., 2010. Honeybee Democracy. Princeton University Press.

Wald, A., 1945. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics 16 (2), 117–186.

Wald, A., 1947. Foundations of a general theory of sequential decision functions. Econometrica 15 (4), 279–313.

Weitzman, M., 1979. Optimal search for the best alternative. Econometrica 47 (3), 641–654.