Best-Arm Identification with Knapsacks: Minimax Policies

Abhijith Vemulapati
Department of Economics
University of Pennsylvania
Philadelphia, PA 19104
abhiv@sas.upenn.edu

Abstract

A resource-constrained decision maker (DM) designs a continuous-time sequential experiment to determine the best choice out of a set of treatments. The DM divides her attention between observing the treatments until one of the resources runs out. Under the minimax regret criterion, we characterize the optimal policy for two treatments when there is a fixed array of resources. Of additional interest is a prerequisite result in which we characterize the minimax regret optimal policy under a single infinite resource (money) and adaptive stopping. Our analysis relies on a reformulation of the typical optimal stopping problem in which we model diffusions with respect to the cumulative resource expenditure rather than the elapsed time.

1 Introduction

When learning is resource-constrained, how does one optimally acquire information? In this paper, we study a variant of a Bayesian pure exploration bandit/best-arm identification (BAI) problem in which treatments heterogeneously deplete an array of resources. We refer to this problem as best-arm identification with knapsacks (BAIwK). Under minimax regret, we characterize optimal policy in a continuous-time, two-treatment setting. Of additional interest is a prerequisite result in which we characterize the minimax regret optimal policy under a single resource (money) and adaptive stopping.

Our work addresses a wide range of practical settings. In clinical trials, for example, limitations on facility space, manufacturing capacity, and skilled personnel pose logistical constraints that are not reflected in standard models of costly sampling. On the monetary side, trial costs vary significantly with the nature of the treatments used (Moore et al. 2018). Although placebos are relatively inexpensive, experimenters spend a significant portion of their budget on sourcing active controls (DiMasi et al. 2016); this suggests that there is heterogeneity in the cost per unit between treatments.

Despite the relevancy of this problem, the literature on this topic is quite new (Li et al. 2023, Li and Chi Cheung 2024). Furthermore, there is yet no work in the decision theory literature that studies the BAI problem under heterogeneous treatment costs or knapsack constraints. Existing papers in this strand (Adusumilli 2024, Liang et al. 2022) impose the same sampling costs across treatments and no resource constraints. Our work thus provides a useful framework for studying traditional problems of sequential experimentation under treatment-dependent costs.

Studying the continuous time version of the problem poses several advantages. We can leverage the properties of Brownian motions to obtain exact characterizations of the optimal policies and their regrets. Furthermore, recent work on the limit-of-experiments approach in adaptive experiments (Adusumilli (2024), Hirano and Porter (2025)) shows that our proposed policy is asymptotically

Preprint. Under review.

optimal under a large class of parametric reward distributions. One primary difference in our setup from similar papers is that the experimenter collects information as a function of the cumulative expenditure rather than the total time spent on the experiment. We demonstrate how to use this rescaling of information arrival to express the optimal policy as a function of time.

The optimal policy possesses many interesting properties — the sampling strategy is fixed and history-independent. In particular, it is chosen to minimize the estimation variance of the unknown difference in reward means. This counters the underlying intuition of heuristic algorithms proposed in this setup, which is to dynamically shift attention away from sampling treatments that are either ineffective or expensive relative to the remaining resources.

1.1 Outline

Section 2 discusses the state of the current literature on this topic. Section 3 sets up the problem. Section 4 characterizes the minimax optimal strategy for both predetermined and adaptive stopping when there is one resource. Section 5 extends the optimal strategy under predetermined stopping to the case of multiple resource constraints.

While the minimax optimal strategy for BAI under three or more treatments is unknown, we can study a similar setting under Gaussian priors as in Liang et al. (2022). The optimal policy in this setting with heterogeneous costs is of independent interest and we characterize it in Appendix C.

2 Related literature

BAI and multi-armed bandits. There is extensive work on heuristic approaches to BAI and costly dynamic sampling in the operations research literature (Qin and Russo 2022, Kaufmann et al. 1996). The typical benchmarks of such algorithms are fixed confidence (restricting algorithms to a misidentification threshold) and fixed budget (restricting the number of exploration periods). Our work builds on the Bayes framework for BAI (Russo 2016) and extends work on fixed budgets to a knapsack setting. The literature on this subproblem is quite nascent and was first defined in Li et al. (2023) as OAK (Optimal Arms Identification with Knapsacks). The problem we solve closely reflects the BAIwRC problem with deterministic costs proposed by Li and Chi Cheung (2024).

Sequential experimentation with knapsacks has a richer literature in the multi-armed bandit setting; see Badanidiyuru and Kleinberg (2016) and Agrawal and Devanur (2016). Approaching this problem in a decision-theoretic, continuous-time setting would involve solving an PDE that grows in complexity with the number of resources. Even in the case of a two-armed bandit, this would be difficult to solve computationally. By focusing on the pure exploration form of the problem, we have more tractability.

Optimal stopping and rational inattention. A costly optimal stopping problem was first proposed by Wald (1947) and studied in the Bayes and minimax frameworks by Arrow et al. (1949). Generalizing this problem to those with multiple treatments, Liang et al. (2022) and Adusumilli (2024) solved the optimal dynamic treatment strategy for the Gaussian prior and the minimax two-point prior, respectively. Our work extends the results of these papers to heterogeneous costs. More generally, this paper follows a recently burgeoning literature on information acquisition and rational inattention: see Hébert and Woodford (2023), Zhong (2017), Fudenberg et al. (2018), and Morris and Strack (2019).

Continuous-time experimentation. There are two main interpretations of continuous-time information arrival in the rational inattention setting. The first treats treatment assignment as attention allocation (Liang et al. 2022, Fudenberg et al. 2018). In this literature, the DM divides her limited attention between the observation of the signals generated by each treatment. In the other interpretation, treatment allocation in the continuous-time setting is a limit of assignment probabilities in discrete time — this is referred to as *diffusion asymptotics* in Wager and Xu (2023) and in Fan and Glynn (2021). Because of the heterogeneous costs, we will need some additional adjustments to draw equivalence of the former to the latter.

Blackwell ordering. The solution strategy of minimizing the estimation variance of the treatment effect difference is inspired by results from Blackwell (1953) and Greenshtein (1996). Armstrong (2022) also shows that a static treatment assignment rule that minimizes this objective is more efficient than any adaptive sampling strategy. This latter result motivates the treatment assignment strategy chosen in this setting.

3 Setup

We initially restrict our attention to one resource (money) and two treatments. The objective of the decision maker (DM) is to decide which treatment to implement on the population. To do so, she designs a sequential experiment to minimize the regret of her decision.

To this end, we define two related problems: the Wald problem and the BAI problem. The set of actions of the DM common to both games is the treatment assignment strategy during the experiment and the choice of which treatment to implement after the experiment. However, in the Wald problem, the DM also chooses a stopping rule to end the experiment early based on past information, while in the BAI problem, the experiment runs for a fixed length of time. It will turn out that the solution to the Wald problem, shown in Section 4, simplifies easily to that of the BAI problem. This will motivate the main result of Section 5, which describes the minimax optimal policy for BAIwK.

3.1 The discrete-time Wald problem

The setup of the Wald problem in the Bayesian framework is as follows. There are two treatments indexed by $a \in \{0, 1\}$, which we refer to hereafter as *arms*. In every round $i \in \{1, \ldots, n\}$, the DM pulls an arm $A_i \in \{0, 1\}$. By pulling arm a, the DM incurs a cost c_a and receives a reward $Y_i(a) \sim N(\mu_a, \sigma_a^2)$. While σ_1, σ_0 are assumed to be known, the means $\boldsymbol{\mu} = (\mu_1, \mu_0)$ are unknown and the DM holds a prior belief $p_0(\cdot)$ over them. The history of actions and rewards is denoted by $\mathcal{H}_i = \{A_1, Y_1, \ldots, A_{i-1}, Y_{i-1}\}$, which determines the total number of pulls $q_a(i)$ and the cumulative rewards $x_a(i)$ of each arm up to round i:

$$q_a(i) = \sum_{j=1}^{i} 1[A_j = a], \quad x_a(i) = \sum_{j=1}^{i} Y_i(a) 1[A_j = a].$$

The DM determines the sequence of arm pulls by choosing a *policy*, which consists of three objects: a sampling rule, $\pi_a(i) = \mathbb{P}[A_i = a | \mathcal{H}_i]$, that assigns a probability of pulling each arm in a given round; a stopping rule, τ, that ends the experiment early based on the history $\mathcal{H}_\tau$; and an implementation rule, $\delta \in \{0, 1\}$, which identifies the arm with the highest reward according to $\mathcal{H}_\tau$.

The DM chooses her policy $\boldsymbol{d} = (\pi, \tau, \delta)$ to minimize regret, which is the difference in expected rewards between the full information (oracle) policy and $\boldsymbol{d}$. If $\mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}[\cdot]$ is the expectation under a decision rule $\boldsymbol{d}$ given $\boldsymbol{\mu}$, then the *frequentist regret* faced by the DM is

$$\mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}[\max\{\mu_1, \mu_0\} - (\delta\mu_1 + (1 - \delta)\mu_0 - c_0 q_0 - c_1 q_1)].$$

The expression arises from the fact that the full information policy immediately receives the best arm mean $\max\{\mu_1, \mu_0\}$ at zero cost, while policy $\boldsymbol{d}$ chooses arm δ and incurs sampling cost $c_0 q_0 + c_1 q_1$.

The limit experiment. Following the approach of Wager and Xu (2023), we can express the continuous-time game as the limit of the discrete-time setup outlined above. This requires us to apply the standard "weak signal" scaling, which consists of three changes: (1) setting the mean rewards of the arms to be $\mu_{n,a} := \mu_a / \sqrt{n}$, (2) scaling the rewards $Y_i(a)$ by $n^{-1/2}$, and (3) scaling the length of each round by n^{-1}.

Let time $t = i/n$ be the fraction of the experiment that has progressed up to round i, such that $i = \lfloor nt \rfloor$. Then, the total pulls and cumulative rewards can be expressed as

$$q_a(t) = \frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} 1[A_j = a], \quad x_a(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} Y_j(a) 1[A_j = a].$$

As $n \to \infty$, Wager and Xu (2023) shows that these expressions converge to the following SDEs:

$$\mathrm{d}x_a(t) = \mu_a \pi_a(t) \mathrm{d}t + \sigma_a \sqrt{\pi_a(t)} \mathrm{d}W_a(t), \tag{1}$$

$$\mathrm{d}q_a(t) = \pi_a(t) \mathrm{d}t, \tag{2}$$

where W_0, W_1 are independent Brownian motions. In continuous time, we can interpret $q_a(t)$ as the total time up to t spent observing pulls from arm a, while $x_a(t)$ is the cumulative observed rewards up to time t.

Let $\varepsilon_a \sim N(0, \sigma_a^2)$. To see the intuition for the result above, note that for any time period j,

$$\begin{aligned}\mathbb{E}[Y_j(a)1[A_j = a]] &= \mathbb{E}[\mu_{n,a}1[A_j = a] + \varepsilon_a 1[A_j = a]] \\ &= \pi_a(j)\frac{\mu_a}{\sqrt{n}} \\ \mathbb{V}[Y_j(a)1[A_j = a]] &= \pi_a(j)\sigma_a^2\end{aligned}$$

Then, due to a functional central limit theorem, the partial sum over the rewards converges to a Gaussian process with drift $\mu_a \pi_a(t)$ and variance $\sigma_a^2 \pi_a(t)$. While this explanation requires the rewards to be Gaussian, this is not necessary to obtain the limiting SDEs; this is an important point as it allows us to appeal to results that generalize the policies described in this paper to non-Gaussian settings.

Time change. For the rest of the paper, we rescale the SDEs derived in (1) such that x_a, q_a arrive as functions of the running expenditure r instead of as a function of elapsed time t. This change allows us to leverage the continuous-time techniques used to solve optimal stopping games with the same sampling cost across treatments. The full details of the time rescaling can be found in Appendix A.

Consequently, given that times q_0, q_1 have been allocated to observing the arms, let $r = c_0 q_0 + c_1 q_1$ denote the total cost of the experiment up to time $q_0 + q_1$. We are now interested in the quantities $q_a(r), x_a(r)$, where

$$\mathrm{d}q_a(r) = \pi_a(r)\mathrm{d}r, \quad \mathrm{d}x_a(r) = \pi_a(r)\mu_a \mathrm{d}r + \sqrt{\pi_a(r)}\sigma_a \mathrm{d}W_a(r). \tag{3}$$

Importantly, this rescaling implies the identity $c_0\pi_0(r) + c_1\pi_1(r) = 1$.

Bayes regret. We now recall the definition of regret proposed in the first paragraph and use it to formalize the notion of Bayes regret, now accounting for the extension to continuous time and the change in the time-scale. Let $s(r) = (x_1(r), x_0(r), q_1(r), q_0(r))$ be the state of the experiment up to cost r and define the filtration generated by the state $s(\cdot)$ as $\mathcal{F}_r \equiv \sigma\{s(u); u \leq r\}$. Given $s(r)$, the DM minimizes her regret with respect to her policy $\boldsymbol{d}$. Now, $\boldsymbol{d}$ consists of a sampling rule $\pi_a(r)$, a stopping rule ρ on the total expenditure, and a $\mathcal{F}_r$-measurable implementation rule $\delta \in \{0, 1\}$. If $\mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}[\cdot]$ is the expectation under a decision rule $\boldsymbol{d}$ given $\boldsymbol{\mu}$, then the frequentist regret faced by the DM is now

$$V(\boldsymbol{d}; \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}[\max\{\mu_1, \mu_0\} - \mu_1\delta - \mu_0(1-\delta) + \rho].$$

We can equivalently express the objective function as

$$V(\boldsymbol{d}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}[\max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\delta + \rho].$$

The DM has a prior p_0 over the true value of $\boldsymbol{\mu}$ and updates his beliefs according to Bayes rule. As a result, the Bayes regret is expectation of regret over p_0:

$$V(\boldsymbol{d}; p_0) := \int V(\boldsymbol{d}; \boldsymbol{\mu})\mathrm{d}p_0(\boldsymbol{\mu}).$$

Minimax regret. In the typical bandit setting, policies are often evaluated according to their frequentist minimax regret. Formally, the DM designs a decision rule to minimize regret under the worst possible bandit environment:

$$\max_{\boldsymbol{\mu}} \min_{\boldsymbol{d}} V(\boldsymbol{d}; \boldsymbol{\mu}).$$

In the Bayesian setting, the worst possible bandit environment is equivalent to a *least favorable prior* – the prior belief on $\boldsymbol{\mu}$ that, given the decision rule of the DM, yields the highest regret. In this setting, an adversarial Nature chooses a prior p_0 to maximize Bayes regret. At the same time, the DM chooses $\boldsymbol{d}$ to minimize her Bayes regret given this least favorable prior. This leads to the following definition of a policy that targets minimax regret: A policy $\boldsymbol{d}^* = (\pi^*, \rho^*, \delta^*)$ is **minimax optimal** if there exists a prior p_0^* such that

$$V(\boldsymbol{d}^*, p_0^*) = \sup_{p_0 \in \mathcal{P}} \inf_{\boldsymbol{d}} V(\boldsymbol{d}, p_0) = \inf_{\boldsymbol{d}} \sup_{p_0 \in \mathcal{P}} V(\boldsymbol{d}, p_0),$$

where $\mathcal{P}$ is the set of all probability distributions over $\boldsymbol{\mu}$. Furthermore, p_0^* is the **least favorable prior** corresponding to $\boldsymbol{d}^*$. The task of finding a minimax optimal policy is equivalent to finding the saddle point in a zero-sum game between the DM and Nature.

3.2 Best arm identification

In the BAI setting, the terminal expenditure ρ is exogenous. As a result, the decision rule of the DM contains π_0, π_1, and an $\mathcal{F}_\rho$-measurable choice rule $\delta \in \{0, 1\}$. The objective function remains the same. It will turn out that the DM chooses the same sampling strategy and implementation rule, but the characterization of the least-favorable prior will be slightly different.

4 Minimax optimal policy – Wald problem

As a lead-up to the main contribution of the paper, we characterize the minimax optimal solution to the Wald problem under one resource. The solution of the BAI problem emerges as a corollary to Theorem 1 of this section and directly motivates Theorem 2, the main result of the paper.

General approach. We fix a sampling strategy that minimizes the estimation variance of the unknown difference in arm means $\mu_1 - \mu_0$ subject to a constraint on the cost of the observations. This is traditionally known as the Neyman allocation Neyman 1934. Analogous to the Neyman allocation in our setting is the pair of observation times $(q_0^*(r), q_1^*(r))$ that minimize

$$\frac{\sigma_0^2}{q_0} + \frac{\sigma_1^2}{q_1}, \tag{4}$$

subject to the budget constraint $c_0 q_0 + c_1 q_1 \leq r$. We can substitute $q_1 = (r - c_0 q_0)/c_1$ out of the constraint and rewrite the objective in its single-variable, unconstrained form:

$$\min_{q_0 \geq 0} \frac{\sigma_0^2}{q_0} + \frac{\sigma_1^2}{(r - c_0 q_0)/c_1}.$$

As a result, the optimal observation times $q_0^*(r), q_1^*(r)$ up to each r are

$$q_a^*(r) = \frac{\sigma_a/\sqrt{c_a}}{\sigma_0\sqrt{c_0} + \sigma_1\sqrt{c_1}} r,$$

so the marginal allocations $\pi_a(r) = \frac{\mathrm{d}q_a(r)}{\mathrm{d}r}$ will be constant.

We are also interested in characterizing a stopping time and a rule that decides which arm is best at the end of the experiment. The structure of these objects is highly dependent on the prior, but it will turn out that the sampling strategy described above can be associated with a least-favorable prior supported on only two points. Consequently, the forms of the stopping rule ρ^* and the choice rule δ^* will follow from established results, which show that the ρ^* is the first exit of a weighted difference of the signals $x_a(r)$ from a symmetric boundary. Furthermore, δ^* will be determined by the sign of this weighted difference at the stopping point. In the case of a two-point prior, the exit boundary is also constant over time.

Minimax-optimal policies. To describe the minimax optimal policy, let $\Delta_0^* \approx 2.19613, \gamma_0^* \approx 0.536357$ be normalizing constants that pin down the proposed minimax optimal policy of the game. Also, define $\Delta^* = \eta\Delta_0^*, \gamma^* = \gamma_0^*/\eta$, where $\eta = (2/(\sigma_0\sqrt{c_0} + \sigma_1\sqrt{c_1}))^{1/3}$. The following theorem characterizes and verifies the minimax optimal policy of the Wald problem.

Theorem 1. *The policy $\boldsymbol{d}_{\gamma^*} = (\pi^*, \tau^*, \delta^*)$ is minimax optimal, where*

$$\pi_a^* = \frac{\sigma_a/\sqrt{c_a}}{\sigma_0\sqrt{c_0} + \sigma_1\sqrt{c_1}}, \; a \in \{0, 1\},$$
$$\rho^* = \inf_r \left\{ \left| \frac{\sqrt{c_1}x_1(r)}{\sigma_1} - \frac{\sqrt{c_0}x_0(r)}{\sigma_0} \right| \geq \gamma^* \right\},$$
$$\delta^* = 1\left\{ \frac{\sqrt{c_1}x_1(r)}{\sigma_1} - \frac{\sqrt{c_0}x_0(r)}{\sigma_0} \geq 0 \right\}.$$

The corresponding least-favorable prior chosen by Nature is

$$\boldsymbol{\mu} \in \{(\sigma_1 \Delta^* \sqrt{c_1}/2, -\sigma_0 \Delta^* \sqrt{c_0}/2), (-\sigma_1 \Delta^* \sqrt{c_1}/2, \sigma_0 \Delta^* \sqrt{c_0}/2)\},$$

In addition, we can exactly characterize the minimax-optimal policy for the BAI problem.

Corollary 1. *Exogenously fix a total budget ρ. Then, the minimax-optimal policy of the DM is $\boldsymbol{d}^*_{BAI} = (\pi^*, \delta^*)$, where π^*, δ^* are defined in Theorem 1. The corresponding least-favorable prior is*

$$\boldsymbol{\mu} \in \{(\sigma_1 \hat{\Delta}^* \sqrt{c_1}/2, -\sigma_0 \hat{\Delta}^* \sqrt{c_0}/2), (-\sigma_1 \hat{\Delta}^* \sqrt{c_1}/2, \sigma_0 \hat{\Delta}^* \sqrt{c_0}/2)\},$$

where $\hat{\Delta}^* = \operatorname{argmax}_\Delta \Delta \Phi\left(-\frac{\Delta}{2}\sqrt{\rho}\right)$ and $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Proof sketch of theorem 1. The choice of prior from Nature makes the DM indifferent between any sampling strategies. Let $\lambda = 1$ be the state in which $\boldsymbol{\mu} = (\sigma_1 \Delta^* \sqrt{c_1}/2, -\sigma_0 \Delta^* \sqrt{c_0}/2)$ and $\lambda = 0$ otherwise. Given Nature's action and $\lambda = 1$, any sampling strategy π of the DM satisfies

$$\begin{aligned}\frac{\sqrt{c_1} x_1(r)}{\sigma_1} - \frac{\sqrt{c_0} x_0(r)}{\sigma_0} &= \frac{\Delta^*}{2}(c_1 \pi_1 + c_0 \pi_0) r + \sqrt{c_1 \pi_1} W_1(r) - \sqrt{c_0 \pi_0} W_0(r) \\ &= \frac{\Delta^*}{2} r + \tilde{W}(r)\end{aligned}$$

where $\tilde{W}(r)$ is a sum of independent Brownian motions and has variance $c_0 \pi_0 + c_1 \pi_1 = 1$. Similarly, when $\lambda = 0$, any sampling strategy admits the process $-\frac{\Delta^*}{2} r + \tilde{W}(r)$. Thus, the belief process is independent of the sampling strategy π in either state. The optimality of the stopping rule and the implementation rule follow from arguments by Shiryaev (2007) or Morris and Strack (2019). In particular, these results tell us that the stopping rule is the first exit time of $\sqrt{c_1} x_1(r)/\sigma_1 - \sqrt{c_0} x_0(r)/\sigma_0$ from a symmetric interval.

To show that Nature's choice of prior is a best response to the DM, note that the sampling strategy of the DM implies that

$$\frac{\sqrt{c_1} x_1(r)}{\sigma_1} - \frac{\sqrt{c_0} x_0(r)}{\sigma_0} = \pm \frac{\mu_1 - \mu_0}{\sigma_1 \sqrt{c_1} + \sigma_0 \sqrt{c_0}} r + \tilde{W}(r)$$

where the sign of the drift depends on whether $\mu_1 > \mu_0$ or not. Given the DM's decision rule, the regret thus depends only on $\mu_1 - \mu_0$. Since the regret is independent of the labels on μ_1, μ_0, Nature only needs to choose $|\mu_1 - \mu_0|$ to maximize regret. As a result, a mixture over two support points for $\boldsymbol{\mu}$ is a best response.

4.1 Discussion

The costs correct the tradeoff in informativeness of the two signals. For example, one will allocate more observation time to the noisier treatment in the setup with $c_0 = c_1$, but there may be great welfare losses in doing so when such treatments are also more expensive. In practice, the rate at which an experimenter purchases a comparative treatment far exceeds the cost of her own drug. Our result suggests that one would need to assign fewer units to these expensive alternatives, especially when there tends to be more precise information about their efficacy. Without accounting for this difference in costs, the DM would thus oversample the comparative treatment.

Corollary 1 motivates the minimax optimal sampling strategy for best arm identification with multiple resource constraints. As long as all resources grow at appropriate rates, the optimal sampling strategy, which minimizes the estimation variance of the difference in treatment effects subject to the resource constraints, is again history-independent.

5 Minimax optimal policy – BAIwK

In this section, we work with an exogenously fixed budget of resources. With multiple constraints, the problem reduces to the one described in the previous section. The reason for this is that the sampling streategy will depend only on the resource constraints that bind at the optimal choice. Our result relies on each resource arriving at a rate proportional its total capacity. Without this assumption, the history-independent strategy may not be optimal.

We do not consider adaptive stopping in this setting because it does not directly relate to a cost that enters the regret. If we want to introduce optimal stopping, we can describe the cost of each arm as the sum of the per-units costs of the resources that it uses. Then, we can apply the main result of Section 4.

5.1 Setup

Let D denote the number of resources and let $\mathrm{A} \in \mathbb{R}_{\geq 0}^{D\times 2}$ be a menu of resource consumptions corresponding to each treatment (element $\mathrm{A}_{i,j}$ is the amount of resource i used up by sampling treatment j). Then, the constraint on the observation times q_0, q_1 is now

$$\mathrm{A}\begin{bmatrix} q_0 \\ q_1 \end{bmatrix} \leq \boldsymbol{\rho}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_D) \in \mathbb{R}^{D\times 1}$ is the total budget of all resources. Let $\mathbf{r} = (r_1, \ldots, r_D)$ represent some arbitrary amount of resources consumed. Assume that information arrives with incremental units of resource 1 and that the other resources arrive at a rate proportional to their total capacity. Formally, we set

$$\mathrm{d}r_d = \frac{\rho_d}{\rho_1}\mathrm{d}r_1$$

and we can then define the SDEs $(x_0(r_1), x_1(r_1))$ using Equation (3). The decision rule $\boldsymbol{d}$ consists of an attention strategy $(\pi_0(r_1), \pi_1(r_1))$ and an $\mathcal{F}_{\rho_1}$measurable implementation rule $\delta \in \{0, 1\}$. Let the Bayes regret $V(\boldsymbol{d}; p_0)$ be defined as in Section 3. Then, to find the minimax optimal policy, the DM and Nature simultaneously solve the game

$$\inf_{\boldsymbol{d}} \sup_{p_0} V(\boldsymbol{d}; p_0).$$

5.2 Minimax optimal decision rule

Assume $\mathrm{rank}(\mathrm{A}) = 2$. The DM minimizes the same objective function as in (4) but with D constraints. Let $\lambda \in \mathbb{R}_{\geq 0}^{D}$ be the Lagrange multiplier on the resource constraint. In Lagrangian form, the DM solves

$$\min_{q_0, q_1, \lambda} \frac{\sigma_0^2}{q_0} + \frac{\sigma_1^2}{q_1} + \lambda^T(\mathrm{A}q - \mathbf{r}) \tag{5}$$

$$\text{s.t} \quad \mathrm{A}q - \mathbf{r} \leq 0 \tag{6}$$

$$q, \lambda \geq 0 \tag{7}$$

The optimal choices q_0^*, q_1^*, λ^* satisfy

$$q_a^*(\mathbf{r}) = \frac{\sigma_a}{\sqrt{\sum_{d=1}^{D} \lambda_d^* \mathrm{A}_{da}}}$$

Note that the problem satisfies the necessary and sufficient KKT conditions (Kuhn and Tucker 1951). As a result, we cannot have an optimal solution for which $\lambda^* = 0$. Im

In general, we can write the observation times q_0, q_1 with respect to r_1 because the level of all other resources is determined by the level of resource 1. Therefore, the optimal allocation $(q_0^*(r_1), q_1^*(r_1))$ satisfies $\mathrm{A}_{d0}q_0^*(r_1) + \mathrm{A}_{d1}q_1^*(r_1) = \frac{\rho_d}{\rho_1}r_1$, where d is any resource for which the constraint in problem (5) is binding (equivalently, any d for which $\lambda_d^* > 0$). Because at least one of λ_d^* is positive, we can take $d = 1$ without loss of generality. In practice, there may be many constraints binding at the same time.

The Lagrangian in (5) scales linearly with r_1 and so we can set $\pi_a^*(r_1) = \frac{q_a^*(\rho_1)}{\rho_1}$. This is the marginal allocation given total observation times q_0^*, q_1^* at the end of the experiment. Then the minimax optimal policy follows from previous arguments.

Theorem 2. *Assume that* $\mathrm{rank}(\mathrm{A}) = 2$ *and that* $\mathrm{d}r_d = \frac{\rho_d}{\rho_1}\mathrm{d}r_1$ *for all* $d \in [D]$*. Consider the solution* $(q_0^*, q_1^*, \lambda^*)$ *to the Lagrange problem described in (5). WLOG assume* $\lambda_1^* > 0$*. A minimax-optimal decision rule* $\boldsymbol{d} = (\pi^*, \delta^*)$ *for best-arm identification with knapsacks is given by*

$$(\pi_0^*, \pi_1^*) = \left(\frac{q_0^*(\rho_1)}{\rho_1}, \frac{q_1^*(\rho_1)}{\rho_1}\right), \quad \delta^* = 1\left\{\frac{\sqrt{A_{11}}x_1(\rho_1)}{\sigma_1} - \frac{\sqrt{A_{10}}x_0(\rho_1)}{\sigma_0} \geq 0\right\}$$

The corresponding least-favorable prior is symmetric and supported on

$$\boldsymbol{\mu} \in \{(\sigma_1 \hat{\Delta}^* \sqrt{\mathrm{A}_{11}}/2, -\sigma_0 \hat{\Delta}^* \sqrt{\mathrm{A}_{10}}/2), (-\sigma_1 \hat{\Delta}^* \sqrt{\mathrm{A}_{11}}/2, \sigma_0 \hat{\Delta}^* \sqrt{\mathrm{A}_{10}}/2)\}$$

where Δ^* is defined in Corollary 1.

Proof. Applying Corollary 1 with attention costs $c_a = \mathrm{A}_{1a}$ as given in the theorem statement, the choice of prior makes the agent indifferent between sampling strategies as long as $c_0\pi_0^* + c_1\pi_1^* = 1$. The result follows due to this fact.

5.2.1 Discussion

Interestingly, a fixed strategy retains minimax optimality in the two-point setting, because the optimal attention allocation that minimizes the estimation variance scales by the same factor as the knapsack size. Thus, one can find where the allocation binds the resource constraints and define the progression of the experiment with respect to one of those limiting resources.

The simplicity of the result relies on the fact that all resources arrive at rates that ensure dynamic consistency of the optimal policy. When this assumption does not hold, the solution to (5) for $\mathbf{r} = \boldsymbol{\rho}$ is no longer consistent with the solution at each value of $\mathbf{r}$. However, it is still optimal to sample such that the total time q_0, q_1 spent observing the arms solves (5).

Because there are possibly multiple binding constraints in the optimal solution to (5), there will be multiple possible Nash equilibria of this form. In practice, one would take r_1 to be the resource for which $(\mathrm{A}_{d0}, \mathrm{A}_{d1}, \rho_d)$ minimizes the closed-form expression for $V(\boldsymbol{d}; p_0)$ in the proof of Corollary 1. The value of $\hat{\Delta}^*$ in this expression also depends on the choice of ρ_d.

6 Conclusion

In this paper, we derive the minimax decision rules of the generalized Wald problem to a sequential experiment with heterogeneous costs. We further verify that a similar result holds in a best-arm identification problem with knapsacks. The optimal rules as a function of total expenditure retain many of the same properties as in Adusumilli (2024), making our results useful for practical settings in which resource costs vary across treatments. An open question in the original optimal stopping problem is what the minimax optimal strategy for three or more treatments would be. We posit that the strategy under heterogeneous costs would translate quite similarly as in this paper if one were to characterize the decision rule with more treatments.

References

Adusumilli, Karun. 2024. "How to sample and when to stop sampling: The generalized Wald problem and minimax policies." *arXiv preprint arXiv:2210.15841.*

Agrawal, Shipra, and Nikhil R. Devanur. 2016. "Linear contextual bandits with knapsacks." In *Advances in Neural Information Processing Systems (NeurIPS),* 3450–3458.

Armstrong, Tim. 2022. "Optimal design of experiments for estimation with selective inference." *Econometrica* 90 (1): 1–30.

Arrow, Kenneth J., David Blackwell, and M.A. Girshick. 1949. "Bayes and minimax solutions of sequential decision problems." *Econometrica* 17 (3/4): 213–244.

Badanidiyuru, Ashwinkumar, and Robert Kleinberg. 2016. "Bandits with Knapsacks." In *Proceedings of the 57th IEEE Symposium on Foundations of Computer Science (FOCS),* 207–216.

Blackwell, David. 1953. "Equivalent Comparison of Experiments." *The Annals of Mathematical Statistics* 24 (2): 265–72.

DiMasi, Joseph, Henry Grabowski, and Ronald Hansen. 2016. "Innovation in the pharmaceutical industry: new estimates of RD costs." *Journal of Health Economics* 47:20–33.

Fan, Lin, and Peter Glynn. 2021. "Diffusion Asymptotics for Thompson Sampling." *arXiv preprint arXiv:2105.09232v2.*

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki. 2018. “Speed, Accuracy, and the Optimal Timing of Choices.” *American Economic Review* 108 (12): 3651–84.

Greenshtein, Eitan. 1996. “Comparison of Sequential Experiments.” *The Annals of Statistics* 24 (1): 436–48.

Hirano, Keisuke, and Jack R. Porter. 2025. “Asymptotic Representations for Sequential Decisions, Adaptive Experiments, and Batched Bandits.” Revised February 2025, https://doi.org/10.48550/arXiv.2302.03117. arXiv: 2302.03117 [econ.EM].

Hébert, Benjamin, and Michael Woodford. 2023. “Rational Inattention when Decisions Take Time.” *Journal of Economic Theory* 208:105612.

Kaufmann, Emilie, Olivier Cappé, and Aurélien Garivier. 1996. “On the complexity of best-arm identification in multi-armed bandit models.” *Journal of Machine Learning Research* 17 (1): 1–42.

Kuhn, Harold W., and Albert W. Tucker. 1951. “Nonlinear programming.” In *Berkeley Symposium on Mathematics, Statistics, and Probability,* 481–492.

Li, Shengjia, Yujia Jin, and Tengyu Ma. 2023. “Best Arm Identification with Knapsacks.” *Proceedings of the 40th International Conference on Machine Learning (ICML).*

Li, Zitian, and Wang Chi Cheung. 2024. “Best Arm Identification with Resource Constraints.” In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics,* edited by Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, 238:253–261. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v238/li24c.html.

Liang, Annie, Xiaosheng Mu, and Vasilis Syrgkanis. 2022. “Dynamically Aggregating Diverse Information.” *Econometrica* 90 (1): 47–80.

Moore, Thomas J., Hanzhe Zhang, Gerard Anderson, and G. Calebi Alexander. 2018. “Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016.” *JAMA Internal Medicine* 178 (11): 1451–1457.

Morris, Stephen, and Philipp Strack. 2019. “The Wald Problem and the Relation of Sequential Sampling and Ex-Ante Information Costs.” *Available at SSRN: https://ssrn.com/abstract=2991567 or http://dx.doi.org/10.2139/ssrn.2991567.*

Neyman, Jerzy. 1934. “On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection.” *Journal of the Royal Statistical Society* 97 (4): 558–625.

Qin, Lihua, and Daniel Russo. 2022. “Best arm identification in stochastic multi-armed bandits with fixed confidence.” *Operations Research* 70 (2): 757–777.

Russo, Daniel. 2016. “Simple Bayesian algorithms for best arm identification.” *Conference on Learning Theory (COLT),* 1417–1418.

Shiryaev, Albert N. 2007. *Probability.* 2nd. Vol. 95. Graduate Texts in Mathematics. Springer. ISBN: 978-0-387-33605-5.

Wager, Stefan, and Kuang Xu. 2023. “Weak Signal Asymptotics for Sequential Experiments.” *Operations Research* 0 (0).

Wald, Abraham. 1947. *Sequential analysis.* John Wiley Sons.

Zhong, Wenxin. 2017. “Rational Inattention and Information Acquisition: An Overview.” *Working Paper.*

A Time Change of the Brownian Motion

Let $r = c_0 q_0(t) + c_1 q_1(t)$, so that $\mathrm{d}r = c_0\pi_0(t)\mathrm{d}t + c_1\pi_1(t)\mathrm{d}t$. Let $\beta(t) := c_0\pi_0(t) + c_1\pi_1(t)$. Define

$$R(t) = \int_0^t \beta(t)\mathrm{d}t, \quad = \inf_{t\geq 0}\{R(t) > r\}$$

By Revuz and Yor (1999), the process $\hat{W}_a(r)$ with respect to r is a Brownian motion, where

$$\hat{W}_a(r) = \int_0^{\tau(r)} \sqrt{\beta(s)}\mathrm{d}W_a(s).$$

The processes x_a, q_a can be rewritten under the time change as

$$\mathrm{d}x_a(r) = \frac{\mu_a \pi_a(\tau(r))}{\beta(\tau(r))}\mathrm{d}r + \sigma_a\sqrt{\frac{\pi_a(\tau(r))}{\beta(\tau(r))}}\mathrm{d}\hat{W}_a(r)$$
$$\mathrm{d}q_a(r) = \frac{\pi_a(\tau(r))}{\beta(\tau(r))}\mathrm{d}r$$

In terms of the optimal policy, we are directly interested in choosing the quantity $\hat{\pi}_a(r) := \frac{\pi_a(\tau(r))}{\beta(\tau(r))}$.

A.1 Time change under the minimax optimal policy

Recall the form of the optimal policy given by Theorem 1:

$$\hat{\pi}_a^* = \frac{\sigma_a/\sqrt{c_a}}{\sigma_0\sqrt{c_0} + \sigma_1\sqrt{c_1}},$$
$$\rho^* = \inf_{r\geq 0}\left\{\frac{\sqrt{c_1}x_1(r)}{\sigma_1} - \frac{\sqrt{c_0}x_0(r)}{\sigma_0} \geq \gamma^*\right\},$$
$$\delta^* = 1\left\{\frac{\sqrt{c_1}x_1(r)}{\sigma_1} - \frac{\sqrt{c_0}x_0(r)}{\sigma_0} \geq 0\right\}.$$

Due to Theorem 1, the sampling rule we are interested in is constant over time and so its unscaled counterpart $\pi_a(t)$ is given by

$$\pi_a(t) = \frac{\sigma_a\sqrt{c_a}}{\sigma_0\sqrt{c_0} + \sigma_1\sqrt{c_1}}$$

Thus, $\tau(r) = \frac{r}{c_0\pi_0 + c_1\pi_1}$ and we can then express the policy in terms of time t rather than running expenditure r.

B Proofs of results

B.1 Proof of theorem 1

Proving this theorem will require us to adapt three lemmas from Adusumilli (2024). In particular, we have to show that for general Δ, γ, the strategies of both players are best responses to each other. Then, the final lemma will show that finding a minimax optimal strategy in this game is equivalent to Nature and the DM choosing Δ, γ simultaneously. These will obtain the values Δ^*, γ^* as a result.

Let $\boldsymbol{d}_\gamma$ be the DM's decision rule under these general values of Δ, γ. They take the same form as in the statement of Theorem 1.

Lemma 1. *Suppose Nature's prior is symmetric and supported on*

$$\boldsymbol{\mu} \in \{(\sigma_1\Delta\sqrt{c_1}/2, -\sigma_0\Delta\sqrt{c_0}/2), (-\sigma_1\Delta\sqrt{c_1}/2, \sigma_0\Delta\sqrt{c_0}/2)\}$$

Then, the proposed decision rule $\boldsymbol{d}_{\gamma(\Delta)}$ of the DM is a best response to Nature, where $\gamma(\Delta)$ is defined in (8).

Proof. Due to Shiryaev (2007), Section 4.2.1, the likelihood ratio of the two priors is

$$\ln \varphi^{\pi}(r) = \Delta \left(\frac{\sqrt{c_1} x_1(r)}{\sigma_1} - \frac{\sqrt{c_0} x_0(r)}{\sigma_0} \right)$$

and the Bayes optimal implementation rule will be

$$\delta = 1 \left\{ \frac{\sqrt{c_1} x_1(r)}{\sigma_1} - \frac{\sqrt{c_0} x_0(r)}{\sigma_0} \geq 0 \right\}$$

Let state $\theta = 1$ be when the prior means are $(\sigma_1 \Delta \sqrt{c_1}/2, -\sigma_0 \Delta \sqrt{c_0}/2)$, and let $\theta = 0$ otherwise. From (3), the dynamics of the observed process in this state can be written as

$$\begin{aligned} \frac{\sqrt{c_1} \mathrm{d} x_1(r)}{\sigma_1} - \frac{\sqrt{c_0} \mathrm{d} x_0(r)}{\sigma_0} &= \frac{\Delta}{2} \mathrm{d} r + \sqrt{c_1 \pi_1} \mathrm{d} W_1(r) - \sqrt{c_0 \pi_0} \mathrm{d} W_0(r) \\ &= \frac{\Delta}{2} \mathrm{d} r + \mathrm{d} \tilde{W}(r) \end{aligned}$$

where $\tilde{W}(r)$ is a Brownian motion, since it is a linear combination of of two Brownian motions with $c_0 \pi_0 + c_1 \pi_1 = 1$. When $\theta = 0$, the observed process is $-\frac{\Delta}{2} \mathrm{d} r + \mathrm{d} \tilde{W}(r)$. So, as in Adusumilli (2024), the belief process is independent of the sampling strategy, so π^* is a trivially a best-response to the prior. It remains to verify the stopping rule optimality. Due to Shiryaev (2007), the posterior probability of state $\theta = 1$, $m(r) = \mathbb{P}(\theta = 1 | \mathcal{F}_r)$, evolves according to

$$\mathrm{d} m(r) = \Delta m(r)(1 - m(r)) \mathrm{d} \tilde{W}(r)$$

The stopping rule must be chosen to minimize regret, which can be expressed as

$$\inf_{\rho} \mathbb{E} \left[\frac{\sigma_1 \sqrt{c_1} + \sigma_0 \sqrt{c_0}}{2} \Delta \min\{m_\rho, 1 - m_\rho\} - \rho \right]$$

We can refer to Morris and Strack (2019), which tells us that the induced distribution of the stopping rule has a uniform support over $(\alpha(\Delta), 1 - \alpha(\Delta))$, where

$$\alpha(\Delta) = \underset{\alpha \in [0, \frac{1}{2}]}{\operatorname{argmin}} \left\{ \frac{\sigma_1 \sqrt{c_1} + \sigma_0 \sqrt{c_0}}{2} \Delta \alpha + \frac{2}{\Delta^2} (1 - 2\alpha) \ln \frac{1 - \alpha}{\alpha} \right\}$$

And the resulting stopping rule, by Shiryaev (2007), Section 4.2.1, is

$$\rho^* = \inf_r \left\{ \frac{\sqrt{c_1} x_1}{\sigma_1} - \frac{\sqrt{c_0} x_0}{\sigma_0} \geq \gamma(\Delta) \right\}$$

where

$$\gamma(\Delta) = \frac{1}{\Delta} \ln \frac{1 - \alpha(\Delta)}{\alpha(\Delta)} \tag{8}$$

Thus $\boldsymbol{d}_\gamma$ is a best response to Nature's prior. □

Lemma 2. *Suppose $|\mu_1 - \mu_0| = \frac{\sigma_1 \sqrt{c_1} + \sigma_0 \sqrt{c_0}}{2} \Delta$. Under the sampling strategy $\boldsymbol{d}_\gamma$, the frequentist regret depends on $\boldsymbol{\mu}$ only through $|\mu_1 - \mu_0|$ and has the form given in (9).*

Proof. Under the sampling strategy of the DM and the form of $|\mu_1 - \mu_0|$, we have by equation (3) that when $\mu_1 > \mu_0$,

$$\frac{\sqrt{c_1} x_1(r)}{\sigma_1} - \frac{\sqrt{c_0} x_0(r)}{\sigma_0} = \frac{\Delta}{2} r + \tilde{W}(r)$$

Let $\lambda(r) = \Delta \left\{ \frac{\sqrt{c_1} x_1(r)}{\sigma_1} - \frac{\sqrt{c_0} x_0(r)}{\sigma_0} \right\} = \frac{\Delta^2}{2} r + \Delta \tilde{W}(r)$. Then, an optimal stopping rule of the form $\inf_r \{ \left| \frac{\sqrt{c_1} x_1(r)}{\sigma_1} - \frac{\sqrt{c_0} x_0(r)}{\sigma_0} \right| \geq \gamma \}$ can be written as $\inf_r \{ |\lambda(r)| \geq \Delta \gamma \}$. Using this characterization, the probability of misidentification and the expected cost $\mathbb{E}[\rho_\tau]$ have the same form as in Adusumilli (2024). Then, we can write expected regret as

$$V(\boldsymbol{d}_\gamma, \boldsymbol{\mu}) = \frac{\sqrt{c_0} \sigma_0 + \sqrt{c_1} \sigma_1}{2} \Delta \frac{1 - e^{-\Delta \gamma}}{e^{\Delta \gamma} - e^{-\Delta \gamma}} + \frac{2\gamma}{\Delta} \frac{e^{\Delta \gamma} + e^{-\Delta \gamma} - 2}{e^{\Delta \gamma} - e^{-\Delta \gamma}} \tag{9}$$

When $\mu_1 < \mu_0$, we arrive at the same formula for V. □

Lemma 3. *There exists a unique Nash equilibrium for which the DM chooses* $\boldsymbol{d}_\gamma^* = (\pi^*, \rho_\gamma^*, \delta^*)$ *and nature chooses a symmetric prior supported on* $(\sigma_1\sqrt{c_1}\Delta^*/2, -\sigma_0\sqrt{c_0}\Delta^*/2)$ *and* $(-\sigma_1\sqrt{c_1}\Delta^*/2, \sigma_0\sqrt{c_0}\Delta^*/2)$*, where* $(\Delta^*, \gamma^*) = (\eta\Delta_0^*, \eta^{-1}\gamma_0^*)$*.* Δ_0^*, γ_0^* *are defined in Adusumilli (2024) and* $\eta = \frac{2}{\sigma_0\sqrt{c_0}+\sigma_1\sqrt{c_1}}$*.*

Proof. Rewrite the value function as $V(d_\gamma, \boldsymbol{\mu}) = \frac{\sqrt{c_0}\sigma_0+\sqrt{c_1}\sigma_1}{2} R(\gamma, \Delta)$ where

$$R(\gamma, \Delta) = \Delta\frac{1-e^{-\Delta\gamma}}{e^{\Delta\gamma}-e^{-\Delta\gamma}} + \frac{2\eta^3\gamma}{\Delta}\frac{e^{\Delta\gamma}+e^{-\Delta\gamma}-2}{e^{\Delta\gamma}-e^{-\Delta\gamma}}$$

where $\eta^3 = \frac{2}{\sqrt{c_0}\sigma_0+\sqrt{c_1}\sigma_1}$. For $\eta = 1$, we can compute (γ_0, Δ_0) and scale $\gamma^* = \eta^{-1}\gamma_0^*,\ \Delta^* = \gamma^*\Delta_0^*$ according to η. The strategies according to γ^* and Δ^* will be minimax optimal. The rest of the proof follows from Adusumilli (2024). □

Finally, we prove that the proposed sampling strategy is minimax optimal for best arm identification.

B.2 Proof of corollary 1

Proof. From Lemma 2, the DM is indifferent between sampling strategies given the form of Nature's prior. Thus her decision rule (π^*, δ^*) is a best response to Nature. To determine a best response of Nature to the DM, suppose $|\mu_1 - \mu_0| = \frac{\sqrt{c_0}\sigma_0+\sqrt{c_1}\sigma_1}{2}$ and that $\mu_1 > \mu_0$. Under π^*,

$$\frac{\sqrt{c_1}\mathrm{d}x_1(r)}{\sigma_1} - \frac{\sqrt{c_0}\mathrm{d}x_0(r)}{\sigma_0} = \frac{\Delta}{2}\mathrm{d}r + \mathrm{d}\tilde{W}(r)$$

As a result, expected regret at the end of the experiment can be written as

$$V(\boldsymbol{d}, \boldsymbol{\mu}) = (\mu_1-\mu_0)\mathbb{P}\left(\frac{\sqrt{c_1}x_1(\rho)}{\sigma_1} - \frac{\sqrt{c_0}x_0(\rho)}{\sigma_0} < 0\right) = \frac{\sigma_1\sqrt{c_1}+\sigma_0\sqrt{c_0}}{2}\Delta\Phi\left(-\frac{\Delta}{2}\sqrt{\rho}\right)$$

Under (π^*, δ^*), the regret depends only on $\mu_1 - \mu_0$, as before. Thus, Nature's best response is to choose $\Delta^* = \mathrm{argmax}_\Delta\, \Delta\Phi\left(-\frac{\Delta}{2}\sqrt{\rho}\right)$. We arrive at the same expression for expected regret when $\mu_1 \leq \mu_0$, so the result holds. □

C Gaussian prior — one resource

Setup. We take the same setup as in Liang et al. (2022), but diffusions evolve according to the total expenditure r. There are K treatments, each with processes $(x_k(r), q_k(r), \pi_k(r))$ as defined in Section 3. The set of information $s(r) = \{(x_k(r), q_k(r), \pi_k(r))\}_{k=1}^K$ is used to define a filtration $\mathcal{F}_r = \sigma\{s(u); u \leq r\}$. Let $\mathbf{q} = (q_1, \ldots, q_K)$ and take as given a menu of costs $\mathbf{c} = (c_1, \ldots, c_K)$. The new budget constraint at each point r is $\mathbf{c}^\intercal\mathbf{q} \leq r$.

Assume that the prior belief over the unknown means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ is distributed according to $\boldsymbol{\mu} \sim N(\boldsymbol{\mu}^0, \Sigma)$, where Σ has full rank. Define a vector of weights $\alpha = (\alpha_1, \ldots, \alpha_K)$ and a payoff-relevant state $\omega = \alpha_1\mu_1 + \ldots + \alpha_K\mu_K$ that linearly depends on the unknown means.

When the total budget ρ is exhausted, the DM chooses an action a and receives a payoff $u(\rho, a, \omega)$. The DM aims to maximize this payoff with respect to her information $\mathcal{F}_\rho$ at the termination of the experiment. Formally, the DM chooses π_0, π_1 to maximize

$$\mathbb{E}[\max_a u(\rho, a, \omega)|\mathcal{F}_\rho]$$

In this appendix, we focus on characterizing the sampling strategy independent of the stopping time, but the form of the stopping time follows straightforwardly from Liang et al. (2022).

Connection to BAI. When $K = 2$, the BAI problem under welfare can be written as $\mathbb{E}[\max\{\mu_2, \mu_1\}|\mathcal{F}_\rho] = \mathbb{E}[\max\{\mu_2 - \mu_1, 0\} + \mu_1|\mathcal{F}_\rho]$. Unlike the minimax-optimal decision rule, the Bayes-optimal decision rule under Bayes risk is the same for both regret and welfare.

By Doob's Optional Stopping Theorem, $\mathbb{E}[\mu_1|\mathcal{F}_\rho]$ is equal to its prior mean μ_1^0. Thus the allocation strategy π_1, π_2 affects only on the difference $\mu_2 - \mu_1$. Adapting this to Liang et al. (2022) framework, we can take $\alpha = (1, 1)$ and $\boldsymbol{\mu} = (-\mu_1, \mu_2)$. Then, we can write the payoff under the optimal action a^* as $u(\rho, a^*, \omega) = \max\{\omega, 0\}$ — this is the reward differential of choosing the treatment with higher mean.

Sampling strategy. Since normal signals can be ranked by their posterior precisions, which are deterministic under normal priors, the dynamic problem of allocating attention reduces to a static one that is a slight variant of Liang et al. (2022). Let $V(\mathbf{q})$ be the posterior variance given that attention q_i has been given to each source. Then, the DM chooses total attention $\mathbf{q}$ to solve

$$\min_{\mathbf{q}:\mathbf{c}^\intercal \mathbf{q} \leq r} V(\mathbf{q}) \tag{10}$$

C.1 BAI setting with two arms

From the discussion in the previous section, we can find the optimal sampling strategy for BAI by solving the game in Liang et al. (2022) with $K = 2, \boldsymbol{\alpha} = \mathbf{1}$, and $\boldsymbol{\mu} = (-\mu_1, \mu_2)$. For $i \neq j$, let $\text{cov}_i = \Sigma_{ii} + \Sigma_{ij}$ and let $x_i = |\Sigma|$. This setup corresponds to the one proposed by Fudenberg et al. (2018). We begin by characterizing the uniformly optimal strategy. The dynamic optimality of this strategy follows from Liang et al. (2022), with the results adapted to the progression of information with respect to r.

Theorem 3. *Assume that* $\text{cov}_1 + \sqrt{\frac{c_1}{c_2}}\text{cov}_2 \geq 0$. *In addition, assume WLOG that* $\text{cov}_1 \geq \sqrt{\frac{c_1}{c_2}}\text{cov}_2$. *A uniformly optimal strategy exists and is given by the following cumulative attention allocation as a function of r:*

$$q_1^*(r) = \begin{cases} \frac{r}{c_1} \text{ if } r \leq r^* = \frac{1}{x_2}(\sqrt{c_1 c_2}\text{cov}_1 - c_1\text{cov}_2), \\ \frac{x_1 r - \text{cov}_2\sqrt{c_1 c_2} + c_2 \text{cov}_1}{x_1 c_1 + x_2\sqrt{c_1 c_2}} \text{ otherwise} \end{cases} \tag{11}$$

and $q_2^* = \frac{1}{c_2}(r - c_1 q_1^*(r))$.

Proof. For $r \leq r^*$, we have $q_1 \leq \frac{\text{cov}_1 - \sqrt{\frac{c_1}{c_2}}\text{cov}_2}{\sqrt{\frac{c_1}{c_2}}x_2}$. Thus, $\sqrt{\frac{c_1}{c_2}}(x_2 q_1 + \text{cov}_2) \leq x_1 q_2 + \text{cov}_1$. We also have that $-\sqrt{\frac{c_1}{c_2}}(x_2 q_1 + \text{cov}_2) \leq x_1 q_2 + \text{cov}_1$ due to the assumption $\text{cov}_1 + \sqrt{\frac{c_1}{c_2}}\text{cov}_2 \geq 0$. It follows that $\frac{c_1}{c_2} \leq \frac{(x_1 q_2 + \text{cov}_1)^2}{(x_2 q_1 + \text{cov}_2)^2}$. The RHS is the marginal rate of substitution between the two treatments, so all attention is placed on treatment 1. Now, if $r > r^*$, there is an interior solution where $\frac{\partial V/\partial q_1}{\partial V/\partial q_2} = \frac{c_1}{c_2}$ holds for strictly positive attention allocations. Using this first order condition and the budget constraint gives us that

$$(q_1^*(r), q_2^*(r)) = \left(\frac{x_1 r - \text{cov}_2\sqrt{c_1 c_2} + c_2\text{cov}_1}{x_1 c_1 + x_2\sqrt{c_1 c_2}}, r - q_1^*\right)$$

□

The uniformly optimal strategy increases weakly in r and the results of Liang et al. (2022) can be applied directly to verify uniqueness and dynamic optimality. Similarly, the stopping rule boundary will take the same form as in that paper but as a function of r.

C.2 General case

Now, relaxing the assumption $K = 2$ and taking a general $\boldsymbol{\alpha}$, we repeat the statement of Theorem 2 from Liang et al. (2022), with a slight omission. The optimal strategy no longer grows in proportion with $\boldsymbol{\alpha}$ but depends also on $\mathbf{c}$. Theorem 4 describes the optimal strategy in this setting. We reference Assumption 6 of Liang et al. (2022), which requires the inverse of the prior covariance matrix to be diagonally dominant.

Theorem 4. *Assume Σ^{-1} is diagonally dominant. There exist $r_1 < r_2 < \cdots < r_m < +\infty$ and nested sets $\emptyset \subset B_1 \subset B_2 \subset \cdots \subset B_m \subset \{1, \ldots, K\}$ such that there exists a deterministic optimal attention allocation strategy. This strategy has $m \leq K$ stages. In each stage $[r_{k-1}, r_k)$, $\pi(r)$ is constant and supported on B_k. The optimal attention allocation at any expenditure $r \geq r_{m-1}$ is proportional to* $\left(\frac{\alpha_1}{\sqrt{c_1}}, \ldots, \frac{\alpha_K}{\sqrt{c_K}}\right)$.

The following series of lemmas proves the theorem.

Lemma 4. *Suppose Σ^{-1} is diagonally dominant. Given an arbitrary attention vector $\boldsymbol{q}$, define $\gamma = (\Sigma^{-1} + \text{diag}(\boldsymbol{q}))^{-1}\boldsymbol{\alpha}$ and denote by B the set of indices such that $|\gamma_i|/c_i$ is maximized. Then γ_i/c_i is the same positive number for every $i \in B$.*

Proof. Because $c_i > 0$, we can directly apply Lemma 8 of Liang et al. (2022). $\square$

Lemma 5. *Suppose Σ^{-1} is diagonally dominant. If the $\underline{r}$-optimal vector satisfies $\partial_1 V(\boldsymbol{q}(\underline{r}))/c_1 = \cdots = \partial_K V(\boldsymbol{q}(\underline{r}))/c_K$, then, for each $k \in [K]$, the r-optimal attention at time $r \geq \underline{r}$ is given by*

$$q_k(r) = q_k(\underline{r}) + \frac{r - \underline{r}}{\alpha_1/\sqrt{c}_1 + \cdots + \alpha_K/\sqrt{c_K}}(\alpha_k/\sqrt{c_k})$$

Proof. Similar to Liang et al. (2022), we require that $\partial_k V/c_k$ remain equal across all treatments as attention increases in the given direction. Thus we need to show that the directional derivative with respect to each q_k is proportional only to its cost c_k. Let

$$\begin{aligned}\delta_k := \sum_{j=1}^{K} \partial_{kj} V \cdot \alpha_j/\sqrt{c_j} &= 2\sum_{j=1}^{K} \gamma_k \gamma_j (\Sigma^{-1} + Q)^{-1}_{kj} \cdot \alpha_j/\sqrt{c_j} \\ &= 2\sum_{j=1}^{K} \sqrt{\frac{c_k}{c_1}}\sqrt{\frac{c_j}{c_1}}\gamma_1^2 (\Sigma^{-1} + Q)^{-1}_{kj} \cdot \alpha_j/\sqrt{c_j} \\ &= 2\frac{\sqrt{c_k}}{c_1}\gamma_1^2\gamma_k = 2\frac{\sqrt{c_k}}{c_1}\sqrt{\frac{c_k}{c_1}}\gamma_1^3 = \frac{c_k}{c_1\sqrt{c_1}}\gamma_1^3\end{aligned}$$

Thus $\partial_k V/c_k$ is equal across all treatments in the direction $\alpha_k/\sqrt{c_k}$. $\square$

The following lemma follows quite straightforwardly from Liang et al. (2022) and has been stated here for the sake of completeness. By repeatedly applying the result, the theorem is proved.

Lemma 6. *Suppose Σ^{-1} is strictly diagonally-dominant. Choose any expenditure $\underline{r}$ and denote*

$$B = \operatorname*{argmin}_i \partial_i V(\boldsymbol{q}(\underline{r}))/c_i = \operatorname*{argmax}_i |\gamma_i|/c_i$$

Then there exists $\beta \in \Delta^{K-1}$ supported on B and $\bar{r} > \underline{r}$ such that $\boldsymbol{q}(r) = \boldsymbol{q}(\underline{r}) + (r - \underline{r}) \cdot \beta$ at times $r \in [\underline{r}, \bar{r}]$.

The vector β depends only on $\Sigma, \boldsymbol{\alpha}, B,$ and $\boldsymbol{c}$. The expenditure $\bar{r}$ is the lowest expenditure higher than $\underline{r}$ at which $\operatorname{argmin}_i \partial_i V(\boldsymbol{q}(\bar{r}))/c_i$ is a strict superset of B. Moreover, when $|B| < K$, it holds that $\bar{r} < \infty$, whereas when $|B| = K$, it holds that $\bar{r} = \infty$ and β is proportional to $\left(\frac{\alpha_1}{\sqrt{c_1}}, \cdots, \frac{\alpha_K}{\sqrt{c_K}}\right)$.

Discussion. The DM starts by sampling the treatment for which the ratio of its marginal reduction of V to its cost $\partial_k V/c_k$ is largest. One by one, she incorporates the treatment with the next highest reduction-to-cost ratio.

An immediate extension of interest is the case of multiple treatments and resources. This approach requires some additional care as the set of binding budget constraints may change as the experiment progresses. Drawing from section 5, the idea will be to sample the treatment with the highest reduction to shadow cost ratio, where the shadow cost $c_k^*(\mathbf{r}) = \sum_{d=1}^{D} \lambda_d(\mathbf{r})\mathrm{A}_{da}$ may change with time.