HOW TO SAMPLE AND WHEN TO STOP SAMPLING: THE GENERALIZED WALD PROBLEM AND MINIMAX POLICIES

KARUN ADUSUMILLI[†]

ABSTRACT. We study sequential experiments where sampling is costly and a decision-maker aims to determine the best treatment for full scale implementation by (1) adaptively allocating units between two possible treatments, and (2) stopping the experiment when the expected welfare (inclusive of sampling costs) from implementing the chosen treatment is maximized. Working under a continuous time limit, we characterize the optimal policies under the minimax regret criterion. We show that the same policies also remain optimal under both parametric and non-parametric outcome distributions in an asymptotic regime where sampling costs approach zero. The minimax optimal sampling rule is just the Neyman allocation: it is independent of sampling costs and does not adapt to observed outcomes. The decision-maker halts sampling when the product of the average treatment difference and the number of observations surpasses a specific threshold. The results derived also apply to the so-called best-arm identification problem, where the number of observations is exogenously specified.

This version: March 25, 2025

The paper subsumes an unpublished note previously circulated as "Neyman allocation is minimax optimal for best arm identification with two arms" on ArXiv at the following link: https://arxiv.org/abs/2204.05527.

I would like to thank the editor and three anonymous referees for valuable comments that substantially improved the paper. Thanks also to Tim Armstrong, Federico Bugni, David Childers, Pepe Montiel-Olea, Chao Qin, Azeem Shaikh, Tim Vogelsang and seminar participants at various universities and conferences for helpful comments and suggestions.

[†]Department of Economics, University of Pennsylvania.

1. Introduction

Acquiring information is expensive. Experimenters need to carefully choose how many units of each treatment to sample and when to stop sampling. This paper seeks to develop techniques for incorporating the cost of information into experimental design. Specifically, we focus our analysis of costly experimentation within the context of comparative trials where the aim is to determine the best of two treatments.

In the computer science literature, such experiments are referred to as A/B tests. Technology companies like Amazon, Google and Microsoft routinely run hundreds of A/B tests a week to evaluate product changes, such as a tweak to a website layout or an update to a search algorithm. However, experimentation is expensive, especially if the changes being tested are very small and require evaluation on large amounts of data; e.g., Deng et al. (2013) state that even hundreds of millions of users were considered insufficient at Google to detect the treatment effects they were interested in. Clinical or randomized trials are another example of A/B tests. Even here, reducing experimentation costs is a key goal. In fact, this has been a major objective for the FDA since 2004 when it introduced the 'Critical Path Initiative' for streamlining drug development; this in turn led the FDA to promote sequential designs in clinical trials (see, e.g., US Food and Drug Admin., 2018, for the current guidance, which was influenced by the need to reduce experimentation costs). For this reason, many recent clinical trials, such as the ones used to test the effectiveness of Covid vaccines (e.g., Baden et al., 2021), now use multi-stage designs where the experiment can be terminated early if a particularly positive or negative effect is seen in early stages.

In practice, the cost of experimentation directly or indirectly enters the researchers' experimental design when they choose an implicit or explicit stopping time (note that we use stopping time interchangeably with the number of observations in the experiment). For instance, in testing the efficacy of vaccines, experimenters stop after a pre-determined number of infections. In other cases, a power analysis may be used to determine sample size before the start of the experiment. But if the aim is to maximize social welfare (or profits), neither of these procedures is optimal.[1]

In this paper, we develop optimal experimentation designs that maximize welfare while also taking into account the cost of information. In particular, we study optimal sampling and stopping rules in sequential experiments where sampling is costly and the decision

[1]See, e.g., Manski and Tetenov (2016) for a critique on the common use of power analysis for determining the sample size in randomized control trials.

maker (DM) aims to determine the best of two treatments by: (1) adaptively allocating units to one of these treatments, and (2) stopping the experiment when the expected welfare, inclusive of sampling costs, is maximized. We term this the generalized Wald problem, and use minimax regret (Manski, 2021), a natural choice criterion under ambiguity aversion, to determine the optimal decision rule.[2]

We first derive the optimal decision rule in continuous time, under the so-called diffusion regime (Fan and Glynn, 2021; Kuang and Wager, 2024) where information arrives gradually in the form of (continuous) Gaussian increments. Then, we show that analogues of this decision rule are also asymptotically optimal under parametric and non-parametric distributions of outcomes. The asymptotics, which appear to be novel, involve taking the marginal cost of experimentation to 0 at a specific rate. Section 4 delves into the rationale behind these 'small cost asymptotics', and argues that they are practically quite relevant. It is important to clarify here that 'small costs' need not literally imply the monetary costs of experimentation are close to 0. Rather, it denotes that these costs are small compared to the benefit of choosing the best treatment for full-scale implementation.

The optimal decision rule has a number of interesting, and perhaps, surprising properties. First, the optimal sampling rule is history independent and also independent of sampling costs. In fact, it is just the Neyman allocation, which is well known in the Randomized Control Trial (RCT) literature as the (fixed) sampling strategy that minimizes estimation variance; our results state that one cannot do better than this even when allowing for adaptive strategies. Second, it is optimal to stop when the difference in average outcomes between the treatments, multiplied by the number of observations collected up to that point, exceeds a specific threshold. The threshold depends on sampling costs and the standard deviation of the treatment outcomes. Finally, at the conclusion of the experiment, the DM chooses the treatment with the highest average outcomes. The decision rule therefore has a simple form that makes it attractive for applications.

Our results also apply to the best arm identification problem with two arms.[3] Best arm identification shares the same aim of determining the best treatment but the number of observations is now exogenously specified, even as the sampling strategy is allowed to be adaptive. Despite this difference, we find Neyman allocation to be the minimax-regret optimal sampling rule in this context as well. However, by not stopping adaptively, we

[2]We do not consider the minimax risk criterion as it leads to a trivial decision: the DM should never experiment and always apply the status quo treatment.

[3]The results for best arm identification were previously circulated in an unpublished note by the author, accessible from ArXiV at https://arxiv.org/abs/2204.05527. The current paper subsumes these results.

lose on experimentation costs. Compared to best arm identification, we show that the use of an optimal stopping time allows us to attain the same regret, exclusive of sampling costs, with 40% fewer observations on average (under the least-favorable prior); this is independent of model parameters such as sampling costs and outcome variances.

For the most part, this paper focuses on constant sampling costs (i.e., constant per observation). This has been a standard assumption since the classic work of Wald (1947), see also Arrow et al. (1949) and Fudenberg et al. (2018), among others. In fact, many online marketplaces for running experiments, e.g., Amazon Mechanical Turk, charge a fixed cost per query/observation. Note also that the costs may be indirect: for online platforms like Google or Microsoft that routinely run thousands of A/B tests, these could correspond to how much experimentation hurts user experience. Still, one may wonder whether and how our results change under other cost functions and modeling choices, e.g., when data is collected in batches, or, when we measure regret in terms of nonlinear or quantile welfare. We asses this in Section 6. Almost all our results still go through under these variations. We also identify a broader class of cost functions, nesting the constant case, in which the form of the optimal decision stays the same.

1.1. **Related literature.** The question of when to stop sampling has a rich history in economics and statistics. It was first studied by Wald (1947) and Arrow et al. (1949) with the goal being hypothesis testing, specifically, optimizing the trade-off between type I and type II errors, instead of welfare maximization. Still, one can place these results into the present framework by imagining that the distributions of outcomes under both treatments are known, but it is unknown which distribution corresponds to which treatment. This paper generalizes these results by allowing the distributions to be unknown. For this reason, we term the question studied here the generalized Wald problem.

Chernoff (1959) studied the sequential hypothesis testing problem under multiple hypotheses, using large deviation methods. The asymptotics there involve taking the sampling costs to 0, even as there is a fixed reward gap between the treatments. More recently, the stopping rules of Chernoff (1959) were incorporated into the δ-PAC (Probably Approximately Correct) algorithms devised by Garivier and Kaufmann (2016) and Qin et al. (2017) for best arm identification with a fixed confidence. The aim in these studies is to minimize the amount of time needed to attain a pre-specified probability, $1-\delta$, of selecting the optimal arm. However, these algorithms do not directly minimize a welfare criterion, and the constraint of pre-specifying a δ could be misplaced, if, e.g.,

there is very little difference between the first and second best treatments. In fact, under the least-favorable prior, our minimax decision rule mis-identifies the best treatment about 23% of the time. Qin and Russo (2022) study the costly sampling problem under fixed reward gap asymptotics using large deviation methods. The present paper differs in using local asymptotics and in appealing to a minimax regret criterion. However, unlike the papers cited above, we only study binary treatments.

A number of papers (Colton, 1963; Lai et al., 1980; Chernoff and Petkau, 1981) have studied sequential trials in which there is a population of N units, and at each period, the DM randomly selects two individuals from this population, and assigns them to the two treatments. The DM is allowed to stop experimenting at any point and apply a single treatment on the remainder of the population. The setup in these papers is intermediate between our own and two-armed bandits: while the aim, as in here, is to minimize regret, acquiring samples is not by itself expensive and the outcomes in the experimentation phase matter for welfare. This literature also does not consider optimal sampling rules.

The paper is also closely related to the growing literature on information acquisition and design, see, Hébert and Woodford (2017); Fudenberg et al. (2018); Morris and Strack (2019); Liang et al. (2022), among others. Fudenberg et al. (2018) study the question of optimal stopping when there are two treatments and the goal is to maximize Bayes welfare (which is equivalent to minimizing Bayes regret) under normal priors and costly sampling. While their approach relies on an exogenously specified sampling rule, Liang et al. (2022) extend this line of inquiry by allowing for endogenous selection of the sampling rule. In fact, for constant sampling costs, the setup in Liang et al. (2022) is similar to ours but the welfare criterion is different: their framework adopts a Bayesian perspective with normal priors. Although the Neyman allocation plays a key role in the optimal sampling rules under both frameworks, the optimal stopping times have very different qualitative and quantitative properties. A detailed comparison is provided in Section 3.2. These differences in stopping times arise because the minimax regret criterion corresponds to a least-favorable prior with a specific two-point support. Thus, our results highlight the important role played by the prior in determining even the qualitative properties of optimal decisions. This motivates the need for robust decision rules, and the minimax regret criterion is one way to obtain them.

Our results also speak to the literature on drift-diffusion models (DDMs), which are widely used in neuroscience and psychology to study choice processes (Luce et al., 1986;

Ratcliff and McKoon, 2008; Fehr and Rangel, 2011). DDMs are based on the classic binary state hypothesis testing problem of Wald (1947). Fudenberg et al. (2018) extend this model to allow for continuous states, using Gaussian priors, and show that the resulting optimal decision rules are very different, even qualitatively, from the predictions of DDM. In this paper, we show that if the DM is ambiguity averse and uses the minimax regret criterion, then the predictions of the DDM model are recovered even under continuous states. In other words, decision making under ignorance can bring us back to DDM.

Finally, the results in this paper are unique in regards to all the above strands of literature in showing that any discrete time parametric and non-parametric version of the problem can be reduced to the diffusion limit under small cost asymptotics. Diffusion asymptotics were introduced by Fan and Glynn (2021) and Kuang and Wager (2024) to study the properties of Thompson sampling in bandit experiments. The techniques for showing asymptotic equivalence to the limit experiment build on, and extend, previous work on sequential experiments by Adusumilli (2021). Relative to that paper, the novelty here is two-fold: first, we derive a sharp characterization of the minimax optimal decision rule for the Wald problem. Second, we introduce 'small cost asymptotics' that may be of independent interest in other, related problems where there is a 'local-to-zero' cost of continuing an experiment.

2. Setup under incremental learning

Following Fudenberg et al. (2018) and Liang et al. (2022), we start by describing the problem under a stylized setting where time is continuous and information arrives gradually in the form of Gaussian increments. In statistics and econometrics, this framework is also known as diffusion asymptotics (Adusumilli, 2021; Fan and Glynn, 2021; Kuang and Wager, 2024). The benefit of the continuous time analysis is that it enables us to provide a sharp characterization of the minimax optimal decision rule; this is otherwise obscured by the discrete nature of the observations in a standard analysis. Section 4 describes how these asymptotics naturally arise under a limit of experiments perspective when we employ small-cost asymptotics and a local-to-zero scaling for the treatment effect.

The setup is as follows. There are two treatments $0, 1$ corresponding to unknown mean rewards $\boldsymbol{\mu} := (\mu_1, \mu_0)$ and known variances σ_1^2, σ_0^2. It is without loss of generality to take σ_1^2, σ_0^2 to be known in the current setting, as they could otherwise be completely determined in an instant from the quadratic variations of the signal processes $x_1(\cdot)$ and $x_0(\cdot)$, defined below in (2.1). The aim of the decision maker (DM) is to determine which

treatment to implement on the population. To guide her choice, the DM conducts a sequential experiment, while paying a flow cost c as long as the experiment is in progress. At each time-point t, the DM samples a treatment according to the sampling rule $\pi_a(t) \equiv \pi(A = a|\mathcal{F}_t), a \in \{0, 1\}$, which specifies the probability of selecting treatment a given some filtration $\mathcal{F}_t$. The DM then keeps track of the signals, $x_1(t), x_0(t)$ from the two treatments, as well as the fraction of times, $q_1(t), q_0(t)$ each treatment was sampled so far:

$$dx_a(t) = \mu_a \pi_a(t)dt + \sigma_a \sqrt{\pi_a(t)} dW_a(t), \tag{2.1}$$

$$dq_a(t) = \pi_a(t)dt. \tag{2.2}$$

Here, $W_1(t), W_0(t)$ are independent one-dimensional Wiener processes. The experiment ends in accordance with an $\mathcal{F}_t$-adapted stopping time, τ. At the conclusion of the experiment, the DM chooses an $\mathcal{F}_\tau$ measurable implementation rule, $\delta \in \{0, 1\}$, specifying which treatment to implement on the population. The DM's decision thus consists of the triple $\boldsymbol{d} := (\pi, \tau, \delta)$.

Denote $s(t) = (x_1(t), x_0(t), q_1(t), q_0(t))$ and take $\mathcal{F}_t \equiv \sigma\{s(u); u \leq t\}$ to be the filtration generated by the state variables $s(\cdot)$ until time t.[4] Let $\mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}[\cdot]$ denote the expectation under a decision rule $\boldsymbol{d}$, given some value of $\boldsymbol{\mu}$. We evaluate various decision rules by the maximum regret criterion, defined as

$$\begin{aligned} V_{\max}(\boldsymbol{d}) &= \max_{\boldsymbol{\mu} \in \mathbb{R} \times \mathbb{R}} V(\boldsymbol{d}, \boldsymbol{\mu}), \text{ with} \\ V(\boldsymbol{d}, \boldsymbol{\mu}) &:= \mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}} \left[\max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\delta + c\tau \right]. \end{aligned} \tag{2.3}$$

To understand this expression, consider an oracle decision rule $\{\tau = 0, \delta = \mathbb{I}\{\mu_1 > \mu_0\}\}$, which has full knowledge of $\boldsymbol{\mu}$. The oracle would achieve a realized welfare of $\max\{\mu_1, \mu_0\}$. In contrast, a given decision rule $\boldsymbol{d}$ generates a realized welfare of $\mu_0 + (\mu_1 - \mu_0)\delta - c\tau$. The difference between these two welfares, $\max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\delta + c\tau$, is referred to as regret. The quantity $V(\boldsymbol{d}, \boldsymbol{\mu})$ therefore represents the 'frequentist regret', i.e., the expected regret of $\boldsymbol{d}$ given $\boldsymbol{\mu}$. The decision rule, $\boldsymbol{d}^*$, that minimizes $V_{\max}(\boldsymbol{d})$ is known as the minimax-regret optimal decision rule.

Minimax regret is a commonly used decision theoretic criterion when the DM faces ambiguity over the values of $\boldsymbol{\mu}$. In contrast, a Bayesian DM would place some prior p_0

[4] As in Liang et al. (2022), we restrict attention to sampling rules π_a for which a weak solution to the functional SDEs (2.1), (2.2) exists. This is true if either $\pi_a : \{s(z) : z \leq t\} \to [0, 1]$ is continuous, see Karatzas and Shreve (2012, Section 5.4), or, if it is any deterministic function of t.

over $\boldsymbol{\mu}$ and aim to minimize Bayes regret, defined as

$$V(\boldsymbol{d}, p_0) := \int V(\boldsymbol{d}, \boldsymbol{\mu})\, dp_0(\boldsymbol{\mu}). \tag{2.4}$$

We can relate max-regret to Bayes regret as $V_{\max}(\boldsymbol{d}) = \sup_{p_0 \in \mathcal{P}} V(\boldsymbol{d}, p_0)$, where $\mathcal{P}$ denotes the set of all possible probability distributions over $\boldsymbol{\mu}$. This suggests a multiple prior interpretation for the minimax regret criterion. As we show in Section 3, minimax regret can be viewed as the value of a zero-sum game played between nature and the DM, where nature chooses the prior p_0 and the DM chooses the decision rule $\boldsymbol{d}$. The minimax-regret optimal rule, $\boldsymbol{d}^*$, is then Bayes optimal under nature's regret-maximizing choice of the prior, also known as the least-favorable prior.

The decision rules $\boldsymbol{d}$ are dynamic since they are history dependent. But as stated, the max-regret criterion $V_{\max}(\boldsymbol{d})$ is 'static' since it ranks decision rules only at $t = 0$; it implicitly assumes the DM can fully commit to the course of action prescribed by $\boldsymbol{d}$. Nonetheless, the criterion admits a dynamically consistent extension, $V_{\max}(\boldsymbol{d}; t)$, which allows for a consistent conditional ranking of $\boldsymbol{d}$ given any history $\mathcal{F}_t$. This extension is possible because the space of priors $\mathcal{P}$ is unrestricted and therefore 'rectangular' in the sense of Epstein and Schneider (2003). As in Epstein and Schneider (2003), rectangularity implies existence of a $V_{\max}(\boldsymbol{d}; t)$ with a recursive structure, such that $V_{\max}(\boldsymbol{d}; 0) = V_{\max}(\boldsymbol{d})$,

$$\begin{aligned} V_{\max}(\boldsymbol{d}; \tau) &= \sup_{p_0 \in \mathcal{P}} \mathbb{E}_{p_0}\left[\max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\delta \middle| \mathcal{F}_\tau\right], \text{ and} \\ V_{\max}(\boldsymbol{d}; t) &= \sup_{p_0 \in \mathcal{P}} \mathbb{E}_{p_0}\left[c \cdot (\tau \wedge t' - t) + V_{\max}(\boldsymbol{d}; t') \middle| \mathcal{F}_t\right] \ \forall\, t' > t, \end{aligned}$$

where $\mathbb{E}_{p_0}[\cdot | \mathcal{F}_t]$ is the expectation with respect to the posterior of p_0 given $\mathcal{F}_t$. Thus, $\boldsymbol{d}^*$ is dynamically optimal under $V_{\max}(\boldsymbol{d}; t)$, the dynamically consistent extension of $V_{\max}(\boldsymbol{d})$.

In any event, dynamic consistency is arguably less relevant in the context of the A/B testing examples that are the focus of this paper. In these examples, it is quite reasonable to suppose that the DM is able to commit to the chosen decision rule. For instance, in clinical trials, regulatory agencies explicitly require and enforce adherence to a pre-specified experimental strategy, see, e.g., FDA's guidance for adaptive experiments (US Food and Drug Admin., 2019, Section III.C).

2.1. **Best arm identification.** The best arm identification problem is a special case of the generalized Wald problem where the stopping time is fixed beforehand and set to $\tau = 1$ without loss of generality. This is equivalent to fixing the number of observations

before the start of the experiment; in fact, we show in Section 4 that a unit time interval corresponds to a pre-specified number of observations, n, in a discrete time analysis. Thus, decisions now consist only of $\boldsymbol{d} = (\pi, \delta)$, but π is still allowed to be adaptive. If we further restrict π to be fixed (i.e., non-adaptive), we get back to the typical setting of Randomized Control Trials (RCTs).

Despite these differences, we show in Section 3 that the minimax-regret optimal sampling and implementation rules are the same in all cases; the optimal sampling rule is the Neyman allocation $\pi_a^*(t) = \sigma_a/(\sigma_1 + \sigma_0)$, while the optimal implementation rule is to choose the treatment with the higher average outcomes. Somewhat surprisingly, then, there is no difference in the optimal strategy between best arm identification and standard RCTs (under minimax regret). The presence of τ, however, makes the generalized Wald problem fundamentally different from the other two. We provide a relative comparison of the benefit of optimal stopping in Section 3.3.

2.2. **Bayesian formulation.** It is convenient to first describe minimal regret under the Bayesian approach, given a prior p_0. As noted earlier, we can characterize minimax regret as Bayes regret under a least-favorable prior.

Let $p(\boldsymbol{\mu}|s)$ denote the posterior density of $\boldsymbol{\mu}$ given the current state $s = (x_1, x_0, q_1, q_0) \in \mathbb{R}^4$. By standard results in stochastic filtering, (here, and in what follows, $\propto$ denotes equality up to a normalization constant)

$$\begin{aligned} p(\boldsymbol{\mu}|s) &\propto p(s|\boldsymbol{\mu}) \cdot p_0(\boldsymbol{\mu}) \\ &\propto p_{q_1}(x_1|\mu_1) \cdot p_{q_0}(x_0|\mu_0) \cdot p_0(\boldsymbol{\mu}); \quad p_{q_a}(\cdot|\mu_a) := \mathcal{N}(\cdot|q_a\mu_a, q_a\sigma_a^2) \end{aligned}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2, and the second proportionality follows from the fact $W_1(\cdot), W_0(\cdot)$ are independent Wiener processes.

Define $V^*(s; p_0)$ as the minimal expected Bayes regret given state s, i.e.,

$$V^*(s; p_0) = \inf_{\boldsymbol{d} \in \mathcal{D}} \mathbb{E}_{\boldsymbol{\mu}|s} \left[V\left(\boldsymbol{d}, \boldsymbol{\mu}\right) \right],$$

where $\mathcal{D}$ is the set of all decision rules that satisfy the measurability conditions set out previously. The minimal (ex-ante) Bayes regret, following (2.4), is then related to $V^*(\cdot; p_0)$ as $\inf_{\boldsymbol{d} \in \mathcal{D}} V(\boldsymbol{d}, p_0) = V^*(s_0; p_0)$, where $s_0 := (0, 0, 0, 0)$ represents the initial state. In principle, one could characterize $V^*(\cdot; p_0)$ as a Hamilton-Jacobi-Bellman Variational Inequality (HJB-VI; Øksendal, 2003, Chapter 10), compute it numerically and characterize the optimal Bayes decision rules. However, this can be computationally expensive, and

moreover, does not provide a closed form characterization of the optimal decisions. Analytical expressions can be obtained under two types of priors:

2.2.1. *Gaussian priors.* In this case, the posterior is also Gaussian and its mean and variance can be computed analytically. Liang et al. (2022) derive the optimal decision rule in this setting. See Section 3.2 for a comparison with our proposal. Additional details are provided in Appendix H.

2.2.2. *Two-point priors.* Two point priors are closely related to hypothesis testing and the sequential likelihood ratio procedures of Wald (1947) and Arrow et al. (1949). More importantly for us, the least-favorable prior for minimax regret, described in the next section, has a two point support.

Suppose the prior over $\boldsymbol{\mu} \equiv (\mu_1, \mu_0)$ is supported on the two points $(\bar{a}, \bar{b}), (\underline{a}, \underline{b})$. Let $\lambda = 1$ represent the event $\boldsymbol{\mu} = (\bar{a}, \bar{b})$ and $\lambda = 0$ the event $\boldsymbol{\mu} = (\underline{a}, \underline{b})$. Also, let $(\Omega, \mathcal{F}, \mathbb{P}_\pi)$ denote the relevant probability space given a (possibly) randomized policy π, where $\mathcal{F} := \cup_{t=1}^{\infty} \mathcal{F}_t \cup \sigma(\lambda)$ is the σ-field generated by λ and the filtration $\{\mathcal{F}_t\}_t$ defined previously, and $\mathbb{P}_\pi$ is the joint probability distribution over $\boldsymbol{\mu}$ and the sample paths of $s(t)$ under π. Set P_π^0, P_π^1 to be the probability measures $P_\pi^0(A) := \mathbb{P}_\pi(A|\lambda = 0)$ and $P_\pi^1(A) := \mathbb{P}_\pi(A|\lambda = 1)$ for any $A \in \mathcal{F}$.

Clearly, the likelihood ratio process $\varphi^\pi(t) := \mathbb{E}_{P_\pi^0}\left[\frac{dP_\pi^1}{dP_\pi^0}\middle| \mathcal{F}_t\right]$ is a sufficient statistic for λ.[5] An application of the Girsanov theorem, noting that $W_1(\cdot), W_0(\cdot)$ are independent of each other, gives (see also Shiryaev, 2007, Section 4.2.1)

$$\ln \varphi^\pi(t) = \frac{(\bar{a} - \underline{a})}{\sigma_1^2} x_1(t) + \frac{(\bar{b} - \underline{b})}{\sigma_0^2} x_0(t) - \frac{(\bar{a}^2 - \underline{a}^2)}{2\sigma_1^2} q_1(t) - \frac{(\bar{b}^2 - \underline{b}^2)}{2\sigma_0^2} q_0(t). \tag{2.5}$$

Let m_0 denote the prior probability that $\lambda = 1$. Additionally, given a sampling rule π, let $m^\pi(t) = \mathbb{P}(\lambda = 1|\mathcal{F}_t)$ denote the belief process describing the posterior probability that $\lambda = 1$. Following Shiryaev (2007, Section 4.2.1), $m^\pi(t)$ can be related to $\varphi^\pi(t)$ as

$$m^\pi(t) = \frac{m_0 \varphi^\pi(t)}{(1 - m_0) + m_0 \varphi^\pi(t)}. \tag{2.6}$$

The Bayes optimal implementation rule at the end of the experiment is

$$\begin{aligned} \delta^{\pi,\tau} &= \mathbb{I}\left\{\bar{a} m^\pi(\tau) + \underline{a}(1 - m^\pi(\tau)) \geq \bar{b} m^\pi(\tau) + \underline{b}(1 - m^\pi(\tau))\right\} \\ &= \mathbb{I}\left\{\ln \varphi^\pi(\tau) \geq \ln \frac{(\underline{b} - \underline{a})(1 - m_0)}{(\bar{a} - \bar{b}) m_0}\right\}. \end{aligned} \tag{2.7}$$

[5]Note that $\frac{dP_\pi^1}{dP_\pi^0}$ is a random variable, being the Radon-Nikodym derivative of P_π^1 with respect to P_π^0.

The superscript on δ highlights that the above implementation rule is conditional on a given choice of (π, τ). Relatedly, the Bayes regret at the implementation phase of the experiment (from employing the optimal implementation rule) is

$$\begin{aligned}\varpi^{\pi}(\tau) &:= \mathbb{E}_{\mathbb{P}_\pi}\left[\max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\delta^{\pi,\tau} \middle| \mathcal{F}_\tau\right] \\ &= \mathbb{E}_{\mathbb{P}_\pi}\left[\max\{\mu_1 - \mu_0, 0\} \middle| \mathcal{F}_\tau\right] - \mathbb{E}_{\mathbb{P}_\pi}\left[\mu_1 - \mu_0 \middle| \mathcal{F}_\tau\right]\delta^{\pi,\tau} \\ &= \min\left\{(\bar{a} - \bar{b})m^{\pi}(\tau), (\underline{b} - \underline{a})(1 - m^{\pi}(\tau))\right\}. \end{aligned} \tag{2.8}$$

Hence, for a given sampling rule π, the Bayes optimal stopping time τ^{π}, can be obtained as the solution to the optimal stopping problem

$$\tau^{\pi} = \inf_{\tau \in \mathcal{T}} \mathbb{E}_{\pi}\left[\varpi^{\pi}(\tau) + c\tau\right], \tag{2.9}$$

where $\mathcal{T}$ is the set of all $\mathcal{F}_t$ measurable stopping times, and $\mathbb{E}_{\pi}[\cdot]$ denotes the expectation under the sampling rule π.

3. Minimax regret and optimal decision rules

The minimax regret value can be written as

$$\inf_{\boldsymbol{d} \in \mathcal{D}} V_{\max}(\boldsymbol{d}) = \inf_{\boldsymbol{d} \in \mathcal{D}} \sup_{p_0 \in \mathcal{P}} V(\boldsymbol{d}, p_0). \tag{3.1}$$

Following Wald (1945), we can characterize minimax regret as the value of a zero-sum game played between nature and the DM. Nature's action involves choosing a prior $p_0 \in \mathcal{P}$ over $\boldsymbol{\mu}$, while the DM chooses the decision rule $\boldsymbol{d}$. The equilibrium action of nature is termed the least-favorable prior, and that of the DM, the minimax decision rule. Note that nature's action in the game is static, as it only chooses a prior p_0 at the beginning of the experiment. In contrast, the DM selects a dynamic decision rule $\boldsymbol{d}$ that, to be a best response to nature's choice, must be Bayes optimal with respect to that choice of prior. Consequently, the minimax-regret optimal rule $\boldsymbol{d}^*$ must be consistent with Bayesian updating of the least-favorable prior throughout the experiment.

The following is the main result of this section: Let $\gamma_0^* \approx 0.536357$, $\Delta_0^* \approx 2.19613$ denote universal constants derived from solving a univariate minimax problem (3.9) described later in this section. Also, define $\eta := \left(\frac{2c}{\sigma_1 + \sigma_0}\right)^{1/3}$, $\gamma^* = \gamma_0^*/\eta$ and $\Delta^* = \eta\Delta_0^*$.

Theorem 1. *The zero-sum two player game (3.1) has a Nash equilibrium with a unique minimax-regret value. The minimax-regret optimal decision rule is* $\boldsymbol{d}^* := (\pi^*, \tau^*, \delta^*)$,

where $\pi_a^*(t) = \sigma_a/(\sigma_1 + \sigma_0)$ *for* $a \in \{0, 1\}$,

$$\tau^* = \inf\left\{t : \left|\frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0}\right| \geq \gamma^*\right\},$$

and $\delta^* = \mathbb{I}\left\{\frac{x_1(\tau^*)}{\sigma_1} - \frac{x_0(\tau^*)}{\sigma_0} \geq 0\right\}$. *Furthermore, the least-favorable prior is a symmetric two-point distribution supported on* $(\sigma_1\Delta^*/2, -\sigma_0\Delta^*/2), (-\sigma_1\Delta^*/2, \sigma_0\Delta^*/2)$.

Theorem 1 makes no claim as to the uniqueness of the Nash equilibrium.[6] Even if multiple equilibria were to exist, however, the value of the game $V^* = \inf_{\boldsymbol{d}\in\mathcal{D}} \sup_{p_0\in\mathcal{P}} V(\boldsymbol{d}, p_0)$ would be unique, and $\boldsymbol{d}^*$ would still be minimax-regret optimal.

The optimal strategies under best arm identification can be derived in the same manner as Theorem 1, but the proof is simpler as it does not involve a stopping rule. Let $\Phi(\cdot)$ denote the CDF of the standard normal distribution.

Corollary 1. *The minimax-regret optimal decision rule for best arm identification is* $\boldsymbol{d}^*_{BAI} := (\pi^*, \delta^*)$, *where* π^*, δ^* *are defined in Theorem 1. The corresponding least-favorable prior is a symmetric two-point distribution supported on* $(\sigma_1\bar{\Delta}_0^*/2, -\sigma_0\bar{\Delta}_0^*/2), (-\sigma_1\bar{\Delta}_0^*/2, \sigma_0\bar{\Delta}_0^*/2)$, *where* $\bar{\Delta}_0^* := 2\arg\max_\delta \delta\Phi(-\delta)$.

3.1. **Proof sketch of Theorem 1.** The main challenge with analyzing the game (3.1) is that the action spaces $\mathcal{P}, \mathcal{D}$ of both nature and the DM are infinite dimensional. Therefore, to prove Theorem 1, we first restrict the action spaces of both players and then show that a Nash equilibrium exists within this restricted class.

For nature, we employ the restricted action space, $\mathcal{P}_{\text{rest}} := \{p_\Delta : \Delta \in \mathbb{R}^+\}$, consisting of all 'indifference priors' indexed by $\Delta \in \mathbb{R}$. Specifically, each 'indifference prior' p_Δ is a two-point prior supported on $(\sigma_1\Delta/2, -\sigma_0\Delta/2), (-\sigma_1\Delta/2, \sigma_0\Delta/2)$, with a prior probability of 0.5 at each support point. As for the DM, we employ the restricted action space

$$\mathcal{D}_{\text{rest}} := \left\{\tilde{\boldsymbol{d}}_\gamma = (\pi^*, \tau_\gamma, \delta^{\tau_\gamma}) : \gamma \in \mathbb{R}^+\right\},$$

where

$$\tau_\gamma := \inf\left\{t : \left|\frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0}\right| \geq \gamma\right\}, \text{ and} \tag{3.2}$$

$$\delta^{\tau_\gamma} := \mathbb{I}\left\{\frac{x_1(\tau_\gamma)}{\sigma_1} - \frac{x_0(\tau_\gamma)}{\sigma_0} \geq 0\right\}. \tag{3.3}$$

[6] In fact, this would depend on the topology defined over $\mathcal{D}$ and $\mathcal{P}$.

We demonstrate that for each $p_\Delta \in \mathcal{P}_{\text{rest}}$, there exists a unique $\gamma \in \mathbb{R}^+$ such that $\tilde{\boldsymbol{d}}_\gamma$ is an unconstrained best response of the DM to p_Δ. In other words, $\tilde{\boldsymbol{d}}_\gamma$ is a best response within the unrestricted class $\mathcal{D}$, not just within $\mathcal{D}_{\text{rest}}$. Similarly, for each $\tilde{\boldsymbol{d}}_\gamma \in \mathcal{D}_{\text{rest}}$, there exists a $\Delta \in \mathbb{R}^+$ such that p_Δ is an unconstrained best response of nature to $\tilde{\boldsymbol{d}}_\gamma$. These results imply that any Nash equilibrium within the restricted action space $\mathcal{P}_{\text{rest}} \times \mathcal{D}_{\text{rest}}$ would also be a Nash equilibrium within the unrestricted action space $\mathcal{P} \times \mathcal{D}$. We then formally demonstrate the existence of a Nash equilibrium in the restricted setting and characterize the equilibrium set of actions.

We elaborate on these steps below:

The DM's response to p_Δ. The term 'indifference priors' indicates that these priors make the DM indifferent between any sampling rule π. The intuitive explanation is as follows: Let $\lambda = 1$ represent the event $\boldsymbol{\mu} = (\sigma_1\Delta/2, -\sigma_0\Delta/2)$ and $\lambda = 0$ the event $\boldsymbol{\mu} = (-\sigma_1\Delta/2, \sigma_0\Delta/2)$. These support points are configured in such a way that both treatments provide equal information about λ, making the choice of treatment irrelevant. To illustrate, assume $\sigma_1 = \sigma_0 = 1$. If the DM samples arm 1 for a period of time δt, she would observe the signal process $x_1(t) = (2\lambda - 1)\frac{\Delta}{2}t + W_1(t)$ over that time-period, with a drift of either $\Delta/2$ or $-\Delta/2$ depending on whether $\lambda = 1$ or $\lambda = 0$. Alternatively, sampling arm 0 yields $x_0(t) = -(2\lambda - 1)\frac{\Delta}{2}t + W_0(t)$, with an exactly opposite drift. Since $W_1(\cdot), W_0(\cdot)$ are independent Wiener processes, both sampling strategies are equally informative about λ in the Blackwell sense.

We now describe the formal argument. For both support points of p_Δ, (2.5) implies

$$\ln \varphi^\pi(t) = \left(\frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0}\right) \cdot \Delta. \tag{3.4}$$

Suppose $\lambda = 1$. By (2.1) and (2.2)

$$\begin{aligned}\frac{dx_1(t)}{\sigma_1} - \frac{dx_0(t)}{\sigma_0} &= \frac{\Delta}{2}dt + \sqrt{\pi_1(t)}dW_1(t) - \sqrt{\pi_0(t)}dW_0(t) \\ &= \frac{\Delta}{2}dt + d\tilde{W}(t),\end{aligned} \tag{3.5}$$

where $\tilde{W}(\cdot)$, defined as $d\tilde{W}(t) := \sqrt{\pi_1(t)}dW_1(t) - \sqrt{\pi_0(t)}dW_0(t)$, is a one dimensional Wiener process, being a linear combination of two independent Wiener processes with $\pi_1(t) + \pi_0(t) = 1$. Plugging the above into (3.4) gives

$$d\ln \varphi^\pi(t) = \frac{\Delta^2}{2}dt + \Delta d\tilde{W}(t).$$

In a similar manner, we can show under $\lambda = 0$ that $d \ln \varphi^{\pi}(t) = -\frac{\Delta^2}{2} dt + \Delta d\tilde{W}(t)$. Thus, the evolution of the log-likelihood ratio process, $\ln \varphi^{\pi}(t)$, can be decomposed into two parts: a drift term $(2\lambda - 1)\frac{\Delta^2}{2} dt$ that depends on the state of the world $\lambda \in \{0, 1\}$, and noise $\Delta d\tilde{W}(t)$. Different sampling rules, π, induce the same drift and leave unchanged the distribution of noise, $\tilde{W}(\cdot)$. Therefore, the choice of π does not affect the sample-path distribution of $\varphi^{\pi}(\cdot)$, and consequently, has no bearing on the sample-path distribution of the belief process $m^{\pi}(\cdot)$.

Crucially, this invariance to the choice of π holds at every time point during the experiment, even as p_{Δ} is revised through Bayesian updating. The key to this invariance lies in the fact that Bayesian updating does not alter the support points of the prior, and it is solely these support points, together with π, that govern the evolution of $\varphi^{\pi}(\cdot)$ under Wiener process noise. Now, the precise form of the support points of p_{Δ} ensures the drift of $\varphi^{\pi}(t)$ is independent of π. Then, the linearity property of Wiener processes implies that a linear combination of such processes remains a Wiener process, thereby preserving the independence of the noise process from the choice of π as well.

As the distributions of $\varphi^{\pi}(\cdot), m^{\pi}(\cdot)$ do not depend on π, the Bayes optimal stopping time in (2.9) is also independent of π for indifference priors (standard results in optimal stopping, see e.g., Øksendal, 2003, Chapter 10, imply that the optimal stopping time in (2.9) is a function only of $m^{\pi}(t)$ which is now independent of π). In fact, it has the same form as the optimal stopping time in the Bayesian hypothesis testing problem of Arrow et al. (1949), analyzed in continuous time by Shiryaev (2007, Section 4.2.1) and Morris and Strack (2019). An adaptation of their results (see, Lemma 1 in Appendix A) shows that the Bayes optimal stopping time corresponding to p_{Δ} is

$$\tau_{\gamma(\Delta)} = \inf \left\{ t : \left| \frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0} \right| \geq \gamma(\Delta) \right\}, \tag{3.6}$$

where $\gamma(\Delta)$ is defined in Lemma 1. By (2.7) and (3.4), the corresponding Bayes optimal implementation rule is seen to be $\delta^{\tau_{\gamma(\Delta)}}$, as defined in (3.3). Hence, the decision rule $\tilde{\boldsymbol{d}}_{\gamma(\Delta)}$ is a best response of the DM to nature's choice of p_{Δ}.

Nature's response to $\tilde{\boldsymbol{d}}_{\gamma}$. Lemma 2 in Appendix A shows that the frequentist regret $V\left(\tilde{\boldsymbol{d}}_{\gamma}, \boldsymbol{\mu}\right)$, given some $\boldsymbol{\mu} = (\mu_1, \mu_0)$, depends only on $|\mu_1 - \mu_0|$. To understand this result, observe that

$$V\left(\tilde{\boldsymbol{d}}_{\gamma}, \boldsymbol{\mu}\right) = \max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}\left[\delta^{\tau_{\gamma}}\right] + c\mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}\left[\tau_{\gamma}\right]. \tag{3.7}$$

Clearly, $\tau_\gamma, \delta^{\tau_\gamma}$ depend on the data only through the stochastic process $\sigma_1^{-1}x_1(\cdot)-\sigma_0^{-1}x_0(\cdot)$. Under the Neyman allocation, (2.1) and (2.2) imply

$$\frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0} = \frac{\mu_1-\mu_0}{\sigma_1+\sigma_0}t + \tilde{W}(t) \tag{3.8}$$

for any $\boldsymbol{\mu} \in \mathbb{R}^2$, where $\tilde{W}(\cdot) := \sqrt{\frac{\sigma_1}{\sigma_1+\sigma_0}}W_1(\cdot) - \sqrt{\frac{\sigma_0}{\sigma_1+\sigma_0}}W_0(\cdot)$ is a standard one-dimensional Wiener process. Consequently, the distributions of $\tau_\gamma, \delta^{\tau_\gamma}$ depend on $\boldsymbol{\mu}$ only through $\mu_1 - \mu_0$. This in turn implies, based on (3.7), that $V\left(\tilde{\boldsymbol{d}}_\gamma, \boldsymbol{\mu}\right)$ depends on $\boldsymbol{\mu}$ only through $\mu_1-\mu_0$. This dependence can be further reduced to $|\mu_1 - \mu_0|$ by symmetry since interchanging the treatment labels $0, 1$ would not affect the frequentist regret of $\tilde{\boldsymbol{d}}_\gamma$.

Since $\boldsymbol{\mu}$ affects $V\left(\tilde{\boldsymbol{d}}_\gamma, \boldsymbol{\mu}\right)$ only through $|\mu_1 - \mu_0|$, it is maximized at $|\mu_1 - \mu_2| = (\sigma_1 + \sigma_0)\Delta(\gamma)/2$, where $\Delta(\gamma)$ is some function of γ. Thus, the best response of nature to $\tilde{\boldsymbol{d}}_\gamma$ is to pick any prior that is supported on $\{\boldsymbol{\mu} : |\mu_1 - \mu_0| = (\sigma_1 + \sigma_0)\Delta(\gamma)/2\}$. Therefore, the two-point prior $p_{\Delta(\gamma)}$ is a best response to $\tilde{\boldsymbol{d}}_\gamma$.

The use of Neyman allocation is essential for the above conclusion. With a different sampling rule, the distributions of $\tau_\gamma, \delta^{\tau_\gamma}$ would depend not only on $\mu_1 - \mu_0$, but also on the individual levels of μ_1, μ_0. In such cases, nature could drive the max-regret of the corresponding decision rule to ∞ through an adversarial choice of $\boldsymbol{\mu}$, as we show in Appendix B. This explains why only the Neyman allocation is minimax optimal, even though the DM is indifferent to any sampling rule under p_Δ: it is needed to ensure nature's choice of p_Δ is supported as a best response to $\tilde{\boldsymbol{d}}_\gamma$.

Nash equilibrium. The above observations imply that the overall Nash equilibrium to (3.1) is the same as the Nash equilibrium in the restricted sub-problem where nature chooses an indifference prior, p_Δ, indexed by $\Delta \in \mathbb{R}^+$, and the DM chooses a decision rule $\tilde{\boldsymbol{d}}_\gamma$, indexed by $\gamma \in \mathbb{R}^+$. Thus, the action spaces of nature and the DM in this sub-problem are scalar. Lemma 3 in Appendix A formally demonstrates existence of a Nash equilibrium in the sub-problem using Sion's minimax theorem (Sion, 1958). The equilibrium values of Δ, γ can be computed numerically by writing down the relevant first order conditions for a Nash equilibrium (see, also, Figure A.1 for the best response functions). The universal constants, γ_0^*, Δ_0^* used in Theorem 1 are derived in this manner. Specifically, Lemma 3 demonstrates that these constants solve the following minimax problem, which characterizes the Nash equilibrium in the restricted sub-problem when $\eta = 1$:

$$\min_{\gamma\in\mathbb{R}^+}\max_{\Delta\in\mathbb{R}^+}\left\{\Delta\frac{1-e^{-\Delta\gamma}}{e^{\Delta\gamma}-e^{-\Delta\gamma}} + \frac{2\gamma}{\Delta}\frac{e^{\Delta\gamma}+e^{-\Delta\gamma}-2}{e^{\Delta\gamma}-e^{-\Delta\gamma}}\right\}. \tag{3.9}$$

3.2. **Discussion.**

3.2.1. *Sampling rule.* Perhaps the most striking aspect of the sampling rule is that it is just the Neyman allocation. This rule is non-adaptive (i.e., it is history independent), and is also independent of sampling costs. In fact, Corollary 1 shows that the sampling and implementation rules are identical to those used in the best arm identification problem.

The Neyman allocation is well known in the RCT literature for being the sampling rule that minimizes the estimation variance of the treatment effect $\mu_1 - \mu_0$. Armstrong (2022) shows that it retains its optimality for estimating $\mu_1 - \mu_0$ even when adaptive sampling strategies are allowed. While Armstrong's (2022) result does not apply to best arm identification, Corollary 1 confirms that Neyman allocation is optimal in this context as well. Hence, practitioners should continue using the randomization designs employed in standard (i.e., non-sequential) experiments, even when adaptivity is allowed.

By way of comparison, the optimal sampling rule under Gaussian priors is also non-adaptive, but it varies deterministically with time (Liang et al., 2022). In fact, after a set time point t^* that depends on (σ_1, σ_0) and the prior variance, it too becomes equal to the Neyman allocation; see Appendix H for a detailed description. There are likely priors for which the Bayes optimal sampling rule is adaptive, but analyzing optimal decision rules under general classes of priors (beyond Gaussian or two-point priors) presents a challenging stochastic-filtering problem. As such, it appears difficult to provide any general claims on what priors lead to adaptive sampling.

3.2.2. *Stopping time.* The stopping time τ^* is adaptive but has a simple form: the DM ends the experiment when $\rho(t) := \sigma_1^{-1} x_1(t) - \sigma_0^{-1} x_0(t)$ exceeds $(\frac{\sigma_1+\sigma_0}{2c})^{1/3}\gamma_0^*$ in absolute value. The threshold is decreasing in c and increasing in $\sigma_1 + \sigma_0$. Let $\bar{x}_a(t) := x_a(t)/q_a(t)$ denote the sample average of outcomes from treatment a at time t. Since $q_a(t) = \sigma_a t/(\sigma_1 + \sigma_0)$ under π^*, we can rewrite the optimal stopping rule as

$$\tau^* = \inf\left\{t : t\,|\bar{x}_1(t) - \bar{x}_0(t)| \geq (\sigma_1 + \sigma_0)\gamma^*\right\},$$

meaning the experiment is stopped when the difference in average outcomes multiplied by the duration t exceeds $(\sigma_1 + \sigma_0)\gamma^*$. Furthermore, from the definition of τ^* and (3.8), we can infer that earlier stopping is indicative of larger reward gaps $\mu_1 - \mu_0$, with the average length of the experiment being longest when $\mu_1 - \mu_0 = 0$.

In contrast, Fudenberg et al. (2018) show that when $\sigma_1 = \sigma_0$, the Bayes optimal stopping time under the independent Gaussian prior $\boldsymbol{\mu} \sim \mathcal{N}(0, \varsigma) \times \mathcal{N}(0, \varsigma)$ has the form

$\tau_{\text{Bayes}} = \mathbb{I}\{|\rho(t)| \geq b^*(t; c, \sigma_1, \varsigma)\}$, where the threshold $b^*(t;\cdot)$ is now time-varying. The following intuition, adapted from Fudenberg et al. (2018), helps explain the difference: Suppose that $\rho(t) \approx 0$ for some large t. Under the Gaussian prior, this likely indicates that $\mu_1 - \mu_0$ is close to 0, suggesting no significant difference between the treatments, so the DM should terminate the experiment straightaway. Conversely, under the least-favorable prior p_{Δ^*}, which has a two-point support, $\rho(t) \approx 0$ would be interpreted as noise, so the DM should proceed henceforth as if starting the experiment from scratch. Thus, the properties of the stopping time are very different depending on the prior. The above intuition also suggests that the relation between $\mu_1 - \mu_0$ and stopping times is more complicated under Gaussian priors, and not monotone as under minimax regret.

The stopping time, τ^*, induces a specific probability of mis-identification of the optimal treatment under the least-favorable prior. By Lemmas 2 and 3, this probability is

$$\alpha^* = \frac{1 - e^{-\Delta^*\gamma^*}}{e^{\Delta^*\gamma^*} - e^{-\Delta^*\gamma^*}} = \frac{1 - e^{-\Delta_0^*\gamma_0^*}}{e^{\Delta_0^*\gamma_0^*} - e^{-\Delta_0^*\gamma_0^*}} \approx 0.235. \tag{3.10}$$

Interestingly, α^* is independent of the model parameters c, σ_1, σ_0. This is because the least-favorable prior adjusts the reward gap in response to these quantities.

Another remarkable property, following from Fudenberg et al. (2018, Theorem 1), is that the probability of mis-identification is independent of the stopping time for any given value of $\boldsymbol{\mu}$, i.e., $\mathbb{P}(\delta^* = 1 | \tau^*, \boldsymbol{\mu} = \boldsymbol{b}) = \mathbb{P}(\delta^* = 1 | \boldsymbol{\mu} = \boldsymbol{b})$ for any $\boldsymbol{b} \in \mathbb{R}^2$. This is again different from the setting with Gaussian priors, where earlier stopping is indicative of a higher probability of selecting the best treatment.

3.3. **Benefit of adaptive experimentation.** In both best arm identification and standard RCTs, the number of units of experimentation is specified beforehand. As we have seen previously, the Neyman allocation is minimax optimal under both adaptive and non-adaptive experiments. The benefit of the decision rule, $\boldsymbol{d}^*$, however, is that it enables one to stop the experiment early, thus saving on experimental costs. To quantify this benefit, fix some values of σ_1, σ_0, c, and suppose that nature chooses the least-favorable prior, p_{Δ^*}, for the generalized Wald problem. Note that p_{Δ^*} is in general different from the least-favorable prior for the best arm identification problem.[7]

Let

$$R^* := \int \mathbb{E}_{\boldsymbol{d}^*|\boldsymbol{\mu}} \left[\max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\delta\right] dp_{\Delta^*}$$

[7]However, the two coincide if the parameter values are such that $\eta := \left(\frac{2c}{\sigma_1 + \sigma_0}\right)^{1/3} = \bar{\Delta}_0^*/\Delta_0^* \approx 0.484$, where $\Delta_0^*, \bar{\Delta}_0^*$ are universal constants defined in the contexts of Theorem 1 and Corollary 1.

denote the Bayes regret, under p_{Δ^*}, of the minimax decision rule $\boldsymbol{d}^*$ net of sampling costs. In fact, by symmetry, the above is also the frequentist regret of $\boldsymbol{d}^*$ under both the support points of p_{Δ^*}. Now, let T_{R^*} denote the duration of time required in a non-adaptive experiment to achieve the same Bayes regret R^* (also under the least-favorable prior and net of sampling costs). Then, making use of some results from Shiryaev (2007, Section 4.2.5), we show in Appendix B.2 that

$$\frac{\mathbb{E}[\tau^*]}{T_{R^*}} = \frac{1-2\alpha^*}{2\left(\Phi^{-1}(1-\alpha^*)\right)^2} \ln \frac{1-\alpha^*}{\alpha^*} \approx 0.6. \tag{3.11}$$

In other words, the use of an adaptive stopping time enables us to attain the same regret with 40% fewer observations on average. Interestingly, the above result is independent of σ_1, σ_0, c, though the values of $\mathbb{E}[\tau^*]$ and T_{R^*} do depend on these quantities (it is only the ratio that is constant). Admittedly, (3.11) does not quantify the welfare gain from using an adaptive experiment - this will depend on the sampling costs - but it is nevertheless useful as an informal measure of how much the amount of experimentation can be reduced.

4. Parametric regimes and small cost asymptotics

We now turn to the analysis of parametric models in discrete time. As before, the DM is tasked with selecting a treatment for implementation across a population. To this end, the DM experiments sequentially in periods $j = 1, 2, \ldots$ after paying an 'effective sampling cost' C per period. Let $1/n$ denote the time interval between successive time periods. To analyze asymptotic behavior in this context, we introduce small cost asymptotics, wherein $C = c/n^{3/2}$ for some $c \in (0, \infty)$, and $n \to \infty$.

Are small cost asymptotics realistic? We contend they are, as C is not the actual cost of experimentation, but rather characterizes the tradeoff between these costs and the benefits from full-scale implementation following the experiment. Indeed, one way to motivate this asymptotic regime is to imagine there are $n^{3/2}$ population units in the implementation phase, so the benefit of applying treatment a is $n^{3/2}\mu_a$, but we divide by $n^{3/2}$ throughout. The actual cost of sampling an additional unit is c, which becomes $c/n^{3/2}$ after the division. Moreover, time t is measured in units of n. This framework aligns with the practical observation that sampling costs are relatively small compared to population size, as seen in both online platforms (Deng et al., 2013) and clinical trials.

For example, in Phase 3 clinical trials, the per-unit cost of treatment is relatively high - Moore et al. (2018) estimate the median cost per patient to be around 41,000$.

However, the potential welfare implications for the population are even more substantial, as the decisions from these trials impact millions of people and firms can expect to earn billions of dollars from successful blockbuster drugs. The effective marginal cost of each observation, obtained by dividing the monetary cost by the population size, is therefore quite small and falls well within our asymptotic framework. More generally, our scaling suggests that if the population size is $n^{3/2}$, one should aim to experiment on a sample size of the order n to achieve optimal welfare. This naturally leads to small cost asymptotics.

As with any asymptotic regime, small-cost asymptotics only provide an approximation to the finite sample properties of decision rules (unless the outcomes are truly Gaussian, in which case they would be exact). The actual finite sample performance needs to be assessed using simulations. Nonetheless, asymptotic analysis offers a valuable benchmark: while there may exist decision rules that outperform our proposal in finite samples, it would be difficult to justify using one that is asymptotically inefficient.

4.1. **Setup in parametric regimes.** In each period, the DM assigns a treatment to a single unit of observation according to some sampling rule $\pi_j(\cdot)$. The treatment assignment is a random draw $A_j \sim \text{Bernoulli}(\pi_j)$. This results in an outcome $Y^{(a)} \sim P_\theta^{(a)}$, with $P_\theta^{(a)}$ denoting the population distribution of outcomes under treatment a. In this section, we assume that this distribution is known up to some unknown $\theta^{(a)} \in \mathbb{R}^d$. It is without loss of generality to assume $Y^{(1)}, Y^{(0)}$ are mutually independent (conditional on $\theta^{(1)}, \theta^{(0)}$) as we only ever observe the outcomes from one treatment anyway. After observing the outcome, the DM can decide either to stop sampling, or call up the next unit. At the end of the experiment, the DM prescribes a treatment to apply on the population.

We use the 'stack-of-rewards-representation' for the outcomes from each arm (Lattimore and Szepesvári, 2020, Section 4.6). Specifically, $Y_i^{(a)}$ denotes the outcome for i-th data point corresponding to treatment a. Also, $\mathbf{y}_{nq} := \{Y_i^{(a)}\}_{i=1}^{\lfloor nq \rfloor}$ denotes the sequence of outcomes after $\lfloor nq \rfloor$ observations from treatment a. We can imagine that prior to the experiment, nature draws an infinite stack of outcomes, $\mathbf{y}^{(a)} := \{Y_i^{(a)}\}_{i=1}^{\infty}$, corresponding to each treatment a, and at each period j, if $A_j = a$, the DM observes the outcome at the top of the stack (this outcome is then removed from the stack corresponding to that treatment).

Recall that t is the number of periods elapsed divided by n. Let

$$q_a(t) := \frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} \mathbb{I}(A_j = a),$$

and take $\mathcal{F}_t$ to be the σ-algebra generated by

$$\xi_t := \left\{ \{A_j\}_{j=1}^{\lfloor nt \rfloor}, \{Y_i^{(1)}\}_{i=1}^{\lfloor nq_1(t) \rfloor}, \{Y_i^{(0)}\}_{i=1}^{\lfloor nq_0(t) \rfloor} \right\}, \tag{4.1}$$

the set of all actions and rewards until period nt. The sequence of σ-algebras, $\{\mathcal{F}_t\}_{t \in \mathcal{T}_n}$, where $\mathcal{T}_n := \{1/n, 2/n, \dots\}$, constitutes a filtration. We require $\pi_{nt}(\cdot)$ to be $\mathcal{F}_{t-1/n}$ measurable, the stopping time, τ, to be $\mathcal{F}_{t-1/n}$ measurable, and the implementation rule, δ, to be $\mathcal{F}_\tau$ measurable. The set of all decision rules $\boldsymbol{d} \equiv (\{\pi_{nt}\}_{t \in \mathcal{T}_n}, \tau, \delta)$ satisfying these requirements is denoted by $\mathcal{D}_n$. As unbounded stopping times pose technical challenges, we generally work with $\mathcal{D}_{n,T} \equiv \{\boldsymbol{d} \in \mathcal{D}_n : \tau \leq T \text{ a.s}\}$, the set of all decision rules with stopping times bounded by some arbitrarily large, but finite, T.

The mean outcomes under a parameter θ are denoted by $\mu_a(\theta) := \mathbb{E}_{P_\theta^{(a)}}[Y_i^{(a)}]$. Following Hirano and Porter (2009), for each $a \in \{0, 1\}$, we consider local perturbations of the form $\{\theta_0^{(a)} + h_a/\sqrt{n}; h_a \in \mathbb{R}^d\}$, with h_a unknown, around a reference parameter $\theta_0^{(a)}$. As in that paper, $\theta_0^{(a)}$ is chosen such that $\mu_1(\theta_0^{(1)}) = \mu_0(\theta_0^{(0)}) = 0$; the last equality, which sets the quantities to 0, is not necessary and is simply a convenient re-centering. This choice of $\theta_0^{(a)}$ defines the hardest instance of the generalized Wald problem. When $\mu_1(\theta_0^{(1)}) \neq \mu_0(\theta_0^{(0)})$, determining the best treatment is trivial under large n, and many decision rules, including the one we propose here (in Section 4.4), would achieve zero asymptotic regret.

Let $P_h^{(a)} := P^{(a)}_{\theta_0^{(a)} + h/\sqrt{n}}$ and take $\mathbb{E}_h^{(a)}[\cdot]$ to be its corresponding expectation. We assume $P_\theta^{(a)}$ is differentiable in quadratic mean around $\theta_0^{(a)}$ with score functions $\psi_a(Y_i)$ and information matrices $I_a := \mathbb{E}_0^{(a)}[\psi_a \psi_a^\intercal]$. For each $h \in \mathbb{R}^d$, denote

$$\mu_{n,a}(h) := \mu_a(\theta_0^{(a)} + h/\sqrt{n}) \approx \dot{\mu}_a^\intercal h/\sqrt{n},$$

where $\dot{\mu}_a := \nabla_\theta \mu_a(\theta_0^{(a)})$. To reduce some notational overhead, we set $\theta_0^{(1)} = \theta_0^{(0)} = \theta_0$, and also suppose that $\mu_{n,a}(h) = -\mu_{n,a}(-h)$ for all h. The latter is always true asymptotically. Both simplifications can be easily dispensed with, at the expense of some additional notation: we emphasize that our results do not fundamentally require $\theta_0^{(1)}, \theta_0^{(0)}$ to be the same or even have the same dimension.

4.2. **Bayes and minimax regret under fixed n.** Let $P_{n,h}^{(a)}$ denote the joint probability over $\mathbf{y}_{nT}^{(a)} := \left\{Y_1^{(a)}, \dots, Y_{nT}^{(a)}\right\}$ - the largest possible (under $\tau \leq T$) iid sequence of outcomes that can be observed from treatment a - when $Y^{(a)} \sim P_h^{(a)}$. Define $\boldsymbol{h} := (h_1, h_0)$, take $P_{n,\boldsymbol{h}}$ to be the joint probability $P_{n,h_1}^{(1)} \times P_{n,h_0}^{(0)}$, and $\mathbb{E}_{n,\boldsymbol{h}}[\cdot]$ its corresponding expectation.

The frequentist regret of decision rule $\boldsymbol{d}$ is defined as

$$\begin{aligned} V_n(\boldsymbol{d}, \boldsymbol{h}) &= V_n\left(\boldsymbol{d}, (\mu_{n,1}(h_1), \mu_{n,0}(h_0))\right) \\ &:= \sqrt{n}\mathbb{E}_{n,\boldsymbol{h}}\left[\max\left\{\mu_{n,1}(h_1) - \mu_{n,0}(h_0), 0\right\} - \left(\mu_{n,1}(h_1) - \mu_{n,0}(h_0)\right)\delta + \frac{c}{n^{3/2}}n\tau\right] \\ &= \sqrt{n}\mathbb{E}_{n,\boldsymbol{h}}\left[\max\left\{\mu_{n,1}(h_1) - \mu_{n,0}(h_0), 0\right\} - \left(\mu_{n,1}(h_1) - \mu_{n,0}(h_0)\right)\delta\right] + c\mathbb{E}_{n,\boldsymbol{h}}[\tau], \end{aligned}$$

where the multiplication by $\sqrt{n}$ in the second line of the above equation is a normalization ensuring $V_n(\boldsymbol{d}, \boldsymbol{h})$ converges to a non-trivial quantity.

Let ν denote a dominating measure over $\{P_\theta : \theta \in \Theta\}$, and define $p_\theta := dP_\theta/d\nu$. Also, take M_0 to be some prior over $\boldsymbol{h}$, and m_0 its density with respect to some other dominating measure ν_1. By Adusumilli (2021), the posterior density (wrt ν_1), $p_n(\cdot|\mathcal{F}_t)$, of $\boldsymbol{h}$ depends only on $\mathbf{y}^{(a)}_{nq_a(t)} = \{Y_i^{(a)}\}_{i=1}^{\lfloor nq_a(t)\rfloor}$ for $a \in \{0, 1\}$. Hence,

$$\begin{aligned} p_n(\boldsymbol{h}|\mathcal{F}_t) &= p_n\left(\boldsymbol{h}|\mathbf{y}^{(1)}_{nq_1(t)}, \mathbf{y}^{(0)}_{nq_0(t)}\right) \\ &\propto \left\{\prod_{i=1}^{\lfloor nq_1(t)\rfloor} p^{(1)}_{\theta_0+h_1/\sqrt{n}}(Y_i^{(1)})\right\}\left\{\prod_{i=1}^{\lfloor nq_0(t)\rfloor} p^{(0)}_{\theta_0+h_0/\sqrt{n}}(Y_i^{(0)})\right\} m_0(\boldsymbol{h}). \end{aligned} \tag{4.2}$$

The fixed n Bayes regret of a decision $\boldsymbol{d}$ is given by $V_n(\boldsymbol{d}, m_0) := \int V_n(\boldsymbol{d}, \boldsymbol{h}) dm_0(\boldsymbol{h})$.

Following definition (4.1), let ξ_τ denote the set of all actions and rewards generated over the course of the experiment. From the form of $V_n(\boldsymbol{d}, \boldsymbol{h})$, it is clear that the Bayes optimal implementation rule is $\delta^*(\xi_\tau) = \mathbb{I}\{\mu_{n,1}(\xi_\tau) \geq \mu_{n,0}(\xi_\tau)\}$, and the resulting Bayes regret at the terminal state is

$$\varpi_n(\xi_\tau) := \mu_n^{\max}(\xi_\tau) - \max\left\{\mu_{n,1}(\xi_\tau), \mu_{n,0}(\xi_\tau)\right\}, \tag{4.3}$$

where $\mu_{n,a}(\xi_\tau) := \mathbb{E}_{\boldsymbol{h}|\xi_\tau}[\mu_{n,a}(h_a)]$ and $\mu_n^{\max}(\xi_\tau) := \mathbb{E}_{\boldsymbol{h}|\xi_\tau}[\max\{\mu_{n,1}(h_1), \mu_{n,0}(h_0)\}]$. We can thus associate each combination, (π, τ), of sampling rules and stopping times with the distribution $\mathbb{P}_{\pi,\tau}$ that they induce over $(\varpi_n(\xi_\tau), \tau)$. Thus,

$$V_n\left(\boldsymbol{d}, m_0\right) = \mathbb{E}_{\pi,\tau}\left[\sqrt{n}\varpi_n(\xi_\tau) + c\tau\right].$$

For any given $T < \infty$, the minimal Bayes regret in the fixed n setting is therefore

$$V^*_{n,T}(m_0) = \inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} \mathbb{E}_{\pi,\tau}\left[\sqrt{n}\varpi_n(\xi_\tau) + c\tau\right].$$

While our interest is in minimax regret, $V^*_{n,T} := \inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} \sup_{\boldsymbol{h}} V_n(\boldsymbol{d}, \boldsymbol{h})$, the minimal Bayes regret is a useful theoretical device as it provides a lower bound, $V^*_{n,T} \geq V^*_{n,T}(m_0)$ for any prior m_0.

4.3. **Lower bound on minimax regret.** We impose the following assumptions (here, and in what follows, $|\cdot|$ denotes the Euclidean norm):

Assumption 1. *(i) The class* $\{P_\theta^{(a)}; \theta \in \mathbb{R}^d\}$ *is differentiable in quadratic mean around* θ_0 *for each* $a \in \{0,1\}$. *(ii)* $\mathbb{E}_0^{(a)}[\exp|\psi_a(Y_i^{(a)})|] < \infty$ *for* $a \in \{0,1\}$. *(iii) There exist* $\dot{\mu}_1, \dot{\mu}_0$ *and* $\epsilon_n^{(1)}, \epsilon_n^{(0)} \to 0$ *s.t* $\sqrt{n}\mu\left(P_h^{(a)}\right) := \sqrt{n}\mu_{n,a}(h) = \dot{\mu}_a^\intercal h + \epsilon_n^{(a)}|h|^2$ *for each* $a \in \{0,1\}$ *and* $h \in \mathbb{R}^d$.

The assumptions are standard, with the only onerous requirement being Assumption 1(ii), which requires the score function to have bounded exponential moments. This is needed due to the proof techniques, which are adapted from Adusumilli (2021).

Let V^* denote the asymptotic minimax regret, defined as the value of the minimax problem in (3.1).

Theorem 2. *Suppose Assumptions 1(i)-(iii) hold. Then,*

$$\sup_{\mathcal{J}} \lim_{T\to\infty} \liminf_{n\to\infty} \inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} \sup_{\boldsymbol{h}\in\mathcal{J}} V_n(\boldsymbol{d},\boldsymbol{h}) \geq V^*,$$

where the outer supremum is taken over all finite subsets $\mathcal{J}$ *of* $\mathbb{R}^d \times \mathbb{R}^d$.

It is straightforward to extend Theorem 2 to best arm identification. We omit the formal statement for brevity. The proof proceeds as follows: Let $\sigma_a^2 := \dot{\mu}_a^\intercal I_a^{-1}\dot{\mu}_a$,

$$h_a^* := \frac{\sigma_a \Delta^*}{2\dot{\mu}_a^\intercal I_a^{-1}\dot{\mu}_a} I_a^{-1}\dot{\mu}_a,$$

and take m_0^* to be the symmetric two-prior supported on $(h_1^*, -h_0^*)$ and $(-h_1^*, h_0^*)$. This is the parametric counterpart to the least-favorable prior described in Theorem 1. Clearly, there exist subsets $\mathcal{J}$ such that

$$\inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} \sup_{\boldsymbol{h}\in\mathcal{J}} V_n(\boldsymbol{d},\boldsymbol{h}) \geq \inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} V_n(\boldsymbol{d}, m_0^*).$$

In Appendix A, we show

$$\lim_{T\to\infty} \lim_{n\to\infty} \inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} V_n(\boldsymbol{d}, m_0^*) = V^*. \tag{4.4}$$

To prove (4.4), we build on previous work in Adusumilli (2021). Standard techniques, such as asymptotic representation theorems (van der Vaart, 2000), are not easily applicable here due to the continuous time nature of the problem. We instead employ a three step approach: First, we replace $P_{n,\boldsymbol{h}}$ with a simpler family of measures whose likelihood ratios (under different values of $\boldsymbol{h}$) are the same as those under Gaussian distributions. Then,

for this family, we write down a HJB-Variational Inequality (HJB-VI) to characterize the optimal value function under fixed n. PDE approximation arguments then let us approximate the fixed-n value function with that under continuous time. The latter is shown to be V^*.

The definition of asymptotic minimax risk used in Theorem 2 is standard, see, e.g., van der Vaart (2000, Theorem 8.11), apart from the $\lim_{T\to\infty}$ operation. The theorem asserts that V^* is a lower bound on minimax regret under any bounded stopping time. The bound T can be arbitrarily large. Our proof techniques require bounded stopping times as various approximation results, e.g., the SLAN property (see, equation (A.10) in Appendix E), are only valid when the experiment is of bounded duration.[8] Nevertheless, we conjecture that there is no loss in setting $T = \infty$ in practice.

4.4. **Attaining the bound.** We now describe a decision rule $\boldsymbol{d}_n = (\pi_n, \tau_n, \delta_n)$ that is asymptotically minimax optimal. Let $\sigma_a^2 = \dot{\mu}_a^\intercal I_a^{-1} \dot{\mu}_a$ for each a and

$$\rho_n(t) := \frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0}, \text{ where } \quad x_a(t) := \frac{\dot{\mu}_a^\intercal I_a^{-1}}{\sqrt{n}} \sum_{i=1}^{\lfloor nq_a(t)\rfloor} \psi_a(Y_i^{(a)}).$$

Note that $x_a(t)$ is the efficient influence function process for estimation of $\mu_a(\theta)$. We assume $\dot{\mu}_a, I_a, \sigma_a$ are known; but in practice, they should be replaced with consistent estimates (from a vanishingly small initial sample) so that they do not require knowledge of the reference parameter θ_0. As described in the next section, this can be done without affecting the asymptotic results.

Take π_n to be any sampling rule such that

$$\left| \frac{q_a(t)}{t} - \frac{\sigma_a}{\sigma_1 + \sigma_0} \right| \leq B \lfloor nt \rfloor^{-b_0} \text{ uniformly over bounded } t, \tag{4.5}$$

for some $B < \infty$ and $b_0 > 1/2$. To simplify matters, we suppose that π_n is deterministic, e.g., $\pi_{n,1}(t) = \mathbb{I}\{q_1(t) \leq t\sigma_1/(\sigma_1 + \sigma_0)\}$. Fully randomized rules, $\pi_{n,1}(t) = \sigma_1/(\sigma_0 + \sigma_1)$, do not satisfy the 'fine-balance' condition (4.5) and we indeed found them to perform poorly in simulations. We further employ

$$\tau_{n,T} = \inf\{t : |\rho_n(t)| \geq \gamma^*\} \wedge T$$

as the stopping time, and as the implementation rule, set $\delta_{n,T} = \mathbb{I}\{\rho_n(\tau_{n,T}) \geq 0\}$.

[8] For any given $\boldsymbol{h}$, the dominated convergence theorem implies $\lim_{T\to\infty} \inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} V_n(\boldsymbol{d}, \boldsymbol{h}) = \inf_{\boldsymbol{d}\in\mathcal{D}_n} V_n(\boldsymbol{d}, \boldsymbol{h})$. However, to allow $T = \infty$ in Theorem 1, we need to show that this equality holds uniformly over n. In specific instances, e.g., when the parametric family is Gaussian, this is indeed the case, but we are not aware of any general results in this direction.

Intuitively, $\boldsymbol{d}_{n,T} = (\pi_n, \tau_{n,T}, \delta_{n,T})$ is the finite sample counterpart of the minimax optimal decision rule $\boldsymbol{d}^*$ from Section 3. The following theorem shows that it is asymptotically minimax optimal in that it attains the lower bound of Theorem 2.

Theorem 3. *Suppose Assumptions 1(i)-(iii) hold. Then,*

$$\sup_{\mathcal{J}} \lim_{T\to\infty} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in\mathcal{J}} V_n(\boldsymbol{d}_{n,T}, \boldsymbol{h}) = V^*,$$

where the outer supremum is taken over all finite subsets $\mathcal{J}$ of $\mathbb{R}^d \times \mathbb{R}^d$.

An important implication of Theorem 3 is that the minimax optimal decision rule only involves one state variable, $\rho_n(t)$. This is even though the state space in principle includes all the past observations until period i, for a total of at least $2i$ variables. The theorem thus provides a major reduction in dimension.

4.5. **Unknown variances.** Replacing σ_1, σ_0 (and other population quantities) with consistent estimates has no effect on asymptotic regret. We suggest two approaches to attain the minimax lower bounds when these parameters are unknown.

The first approach uses 'forced exploration' (see, e.g., Lattimore and Szepesvári, 2020, Chapter 33, Note 7): we set $\pi_n^*(t) = 1/2$ for the first $\bar{n} = n^a$ observations, where $a \in (0,1)$. This corresponds to a time duration of $\bar{t} = n^{a-1}$. We use the data from these periods to obtain consistent estimates, $\hat{\sigma}_1^2, \hat{\sigma}_0^2$ of σ_1^2, σ_0^2. From $\bar{t}$ onwards, we apply the minimax optimal rule $\boldsymbol{d}_{n,T}$ after plugging-in $\hat{\sigma}_1, \hat{\sigma}_0$ in place of σ_1, σ_0. Note that when applying $\boldsymbol{d}_{n,T}$, we should start $x_1(\cdot), x_0(\cdot)$ from their values at $\bar{t}$ to ensure the information accrued before $\bar{t}$ is also taken into account. This strategy is asymptotically minimax optimal for any $a \in (0,1)$.

Our second suggestion is to place a prior on σ_1, σ_0, and continuously revise their values using the posterior means. We recommend using an inverse-gamma prior and computing the posterior by treating the scores $\psi_a(Y_i^{(a)})$ as Gaussian (which is justified asymptotically). This approach has the advantage of not requiring any tuning parameters.

Admittedly, both proposals treat estimation of σ_1, σ_0 as separate and somewhat less critical than the estimation of the population mean parameters. However, this merely reflects the significant asymmetry in the complexity of estimating these parameters in the continuous time setting. As noted in Section 2, σ_1, σ_0 can be learnt instantly in continuous time from the quadratic variations of $x_1(t), x_0(t)$. Conversely, running the sequential experiment for a brief period would only marginally update the prior over $\boldsymbol{\mu}$.

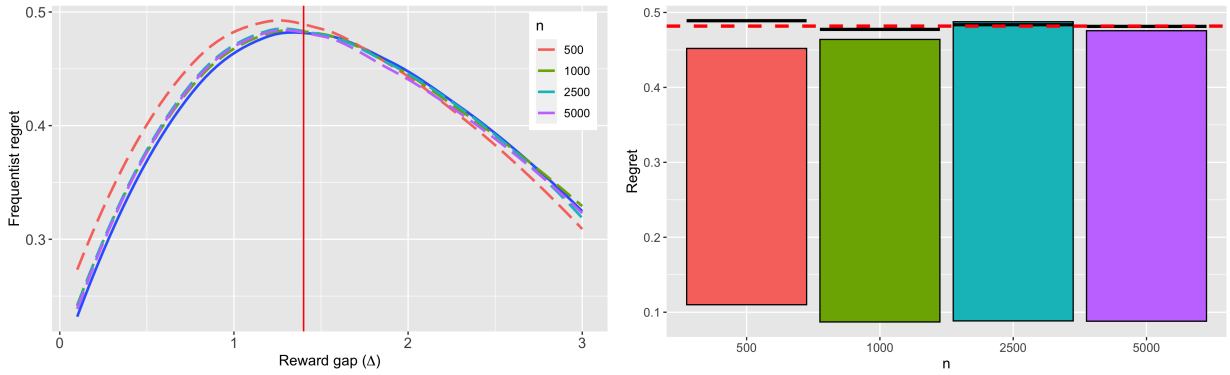
Furthermore, in the finite n setting, it is important to recognize that $\theta^{(a)}$ characterizes the entire distribution of $Y^{(a)}$, including its mean and variance. The quantities σ_1, σ_0 do not represent the variances of the underlying probability distributions - which are subject to change anyway under the local sequence $\theta_0^{(a)} + h_a/\sqrt{n}$ - but are rather the information matrices evaluated at the reference parameters $\theta_0^{(1)}, \theta_0^{(0)}$. Conceptually, under local asymptotics, these reference parameters are assumed to be known in advance. While one would aim, in practice, to construct procedures that adapt to or are invariant to these quantities, the impact of estimating them cannot be accounted for in the local asymptotic framework itself. This is not to diminish the importance of efficiently estimating σ_1, σ_0; rather, the issue lies outside the scope of first-order asymptotic theory, which is just too coarse an approximation for this purpose. Addressing this would require employing higher-order asymptotics.

5. Numerical illustration

A/B testing is commonly used in online platforms for optimizing websites. Consequently, to assess the finite sample performance of our proposed policies, we run a Monte-Carlo simulation calibrated to a realistic example of such an A/B test. Suppose there are two candidate website layouts, with exit rates γ_0, γ_1, and we want to run an A/B test to determine the one with the lowest exit rate.[9] The outcomes are binary, $Y^{(a)} \sim \text{Bernoulli}(\gamma_a)$. This is a parametric setting with score functions $\psi_a(Y_i^{(a)}) = Y_i^{(a)}$. We calibrate $\gamma_0 = 0.4$, which is a typical value for an exit rate. The cost of experimentation is normalized to $c = 1$ and we consider various values of n, corresponding to different 'population sizes' (recall that the benefit during implementation is scaled as $n^{3/2}\gamma_a$). We then set $\gamma_1 = \gamma_0 + \Delta/\sqrt{n}$, and describe the results under varying Δ. Local asymptotics provide a good approximation in practice because raw performance gains are generally small - typically, $|\gamma_1 - \gamma_0|$ is of the order 0.05 or less (see, e.g., Deng et al., 2013) - but these gains can translate into large profits when applied at scale, i.e., when n is large.

Since $\sigma_a = \sqrt{\gamma_a(1-\gamma_a)}$ is unknown, we employ 'forced sampling' with $\bar{n} = \max(50, 0.05n)$, i.e., using about 5% of the sample, to estimate σ_1, σ_0. Note that the asymptotically optimal sampling rule is always 1/2 in the Bernoulli setting, so forced sampling is in fact asymptotically costless. We also experimented with a beta prior to continuously update σ_a, but found the results to be somewhat inferior, (see Appendix C for details). Figure 5.1,

[9]The exit rate is defined as the fraction of viewers of a webpage who exit from the website it is part of (i.e., without viewing other pages in that website).



A: Frequentist regret profiles

B: Performance under least-favorable prior

Note: The solid curve in Panel A is the regret profile of $\boldsymbol{d}^*$; the vertical line denotes Δ^*. We only plot the results for $\Delta > 0$ as the values are close to symmetric. The dashed line in Panel B is V^*, the asymptotic minimax regret. The solid horizontal lines denote the Bayes regret in finite samples, under the least-favorable prior. The bars describe the interquartile range of regret.

FIGURE 5.1. Finite sample performance of $\boldsymbol{d}_n$

Panel A plots the finite sample frequentist regret profiles of our policy rules $\boldsymbol{d}_n := \boldsymbol{d}_{n,\infty}$ (with $T = \infty$) for various values of n, along with that of the minimax optimal policy $\boldsymbol{d}^*$ under the diffusion regime; the regret profile of the latter is derived analytically in Lemma 3. Diffusion asymptotics provide a very good approximation to the finite sample properties of $\boldsymbol{d}_n$, even for such relatively small values of n as $n = 1000$. In practice, A/B tests often involve tens, even hundreds, of thousands of observations. The max-regret of $\boldsymbol{d}_n$ is also very close to the asymptotic lower bound V^* (the max-regret of $\boldsymbol{d}^*$).

Figure 5.1, Panel B displays some summary statistics for the Bayes regret of $\boldsymbol{d}_n$ under the least-favorable prior, p_{Δ^*}. The regret distribution is positively skewed and heavy tailed. The finite sample Bayes regret is again very close to V^*.

Appendix C reports additional simulation results using Gaussian outcomes.

6. Variations and extensions

We now consider various modifications of the basic setup and analyze if, and how, the optimal decisions change. Appendix G discusses extensions to multiple treatments.

6.1. **Batching.** In practice, it may be that data is collected in batches instead of one at a time, and the DM can only make decisions after processing each batch. Let B_n denote the number of observations considered in each batch. In the context of Section 4, this corresponds to a time duration of B_n/n. An analysis of its proof shows that Theorem 2 continues to hold as long as $B_n/n \to 0$. Thus, $\boldsymbol{d}_{n,T}$ remains asymptotically minimax optimal in this scenario.

Even for $B_n/n \to m \in (0,1)$, the optimal decision rules remain broadly unchanged. Asymptotically, we have equivalence to Gaussian experiments, so we can analyze batched experiments under the diffusion framework by imagining that the stopping time is only allowed to take on discrete values $\{0, 1/m, 2/m, \dots\}$. It is then clear from the discussion in Section 3.1 that the optimal sampling and implementation rules remain unchanged. The discrete nature of the setting makes determining the optimal stopping rule difficult, but it is easy to show that the decision rule $(\pi^*, \tau_m^*, \delta^{\tau_m^*})$, where

$$\tau_m^* := \inf\left\{t \in \{0, 1/m, 2/m, \dots\} : \left|\frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0}\right| \geq \gamma^*\right\}$$

and $\delta^{\tau_m^*} := \mathbb{I}\left\{\sigma_1^{-1}x_1(\tau_m^*) - \sigma_0^{-1}x_0(\tau_m^*) \geq 0\right\}$, while not being exactly optimal, has a minimax regret that is arbitrarily close to V^* for large enough m (note that no batched experiment can attain a minimax regret that is lower than V^*).

6.2. **Alternative cost functions.** All our results so far were derived under constant sampling costs. The same techniques apply to other types of flow costs as long as these depend only on $\rho(t) := \sigma_1^{-1}x_1(t) - \sigma_0^{-1}x_0(t)$. In particular, suppose that the frequentist regret is given by

$$V(\boldsymbol{d}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{d}|\boldsymbol{\mu}}\left[\max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\delta + \int_0^\tau c(\rho(t))dt\right],$$

where $c(z)$ is the flow cost of experimentation when $\rho(t) = z$. We require $c(\cdot)$ to be (i) positive, (ii) bounded away from 0, i.e., $\inf_z c(z) \geq \underline{c} > 0$, and (iii) symmetric, i.e., $c(z) = c(-z)$. By (3.8), $(\sigma_1 + \sigma_0)\rho(t)/t$ is an estimate of the treatment effect $\mu_1 - \mu_0$, so the above allows for situations in which sampling costs depend on the magnitude of the estimated treatment effects. While we are not aware of any real world examples of such costs, they could arise if there is feedback between the observations and sampling costs, e.g., if it is harder to find subjects for experimentation when the treatment effect estimates are higher. When there are only two states, the 'ex-ante' entropy cost of Sims (2003) is also equivalent to a specific flow cost of the form $c(\cdot)$ above, see Morris and Strack (2019).[10]

For the above class of cost functions, we show in Appendix D that the minimax optimal decision rule $\boldsymbol{d}^*$ and the least-favorable prior p_Δ^* have the same form as in Theorem 1, but the values of γ^*, Δ^* are different and need to be calculated by solving the minimax

[10]However, we are not aware of any extension of this result to continuous states.

problem

$$\min_{\gamma} \max_{\Delta} \left\{ \left(\frac{\sigma_1 + \sigma_0}{2} \right) \frac{\left(1 - e^{-\Delta\gamma}\right) \Delta}{e^{\Delta\gamma} - e^{-\Delta\gamma}} + \frac{\left(e^{\Delta\gamma} - 1\right) \zeta_\Delta(\gamma) + \left(1 - e^{-\Delta\gamma}\right) \zeta_\Delta(-\gamma)}{e^{\Delta\gamma} - e^{-\Delta\gamma}} \right\},$$

where

$$\zeta_\Delta(x) := 2 \int_0^x \int_0^y e^{\Delta(z-y)} c(z) dz dy.$$

Beyond this class of sampling costs, however, it is easy to conceive of scenarios in which the optimal decision rule differs markedly from the one we obtain here. For instance, Neyman allocation would no longer be the optimal sampling rule if the costs for sampling each treatment were different. Alternatively, if $c(\cdot)$ were to depend on t, the optimal stopping time could have a very different form. The analysis of these cost functions is not covered by the present techniques.

6.3. **Non-parametric outcomes.** In Appendix E, we extend Theorems 2 and 3 to the non-parametric setting, where there is no a-priori information about the distributions $P^{(1)}, P^{(0)}$ of $Y_i^{(1)}$ and $Y_i^{(0)}$. The minimax optimal rule retains the same form as in Section 4.4, but with $x_a(t)$ now defined as $n^{-1/2} \sum_{i=1}^{\lfloor nq_a(t) \rfloor} Y_i^{(a)}$, and σ_a^2 as the variance of $Y_i^{(a)}$ at some reference distribution $P_0^{(a)}$ (as in Section 4, $P_0^{(1)}, P_0^{(0)}$ are to be chosen such that $\mathbb{E}_{P_0^{(1)}}[Y_i^{(1)}] = \mathbb{E}_{P_0^{(0)}}[Y_i^{(0)}]$). One can obtain these same results by simply assuming the outcomes to be Gaussian.

The above results can also be extended to different regret measures. Specifically, instead of $\mu(\cdot)$ denoting the mean functional in the definition of regret $\max\{\mu(P^{(1)}) - \mu(P^{(0)}), 0\} - (\mu(P^{(1)}) - \mu(P^{(0)}))\delta + c\tau$, it can denote other functionals of the outcome distribution in the implementation phase (we still need costs to be linear and additively separable). For instance, $\mu(\cdot)$ could be the quantile function. In Appendix E.4, we show that the decision rule $\boldsymbol{d}_{n.T}$ from Section 4.4 is still minimax optimal if we just redefine $x_a(t)$ to now be the efficient influence function process $n^{-1/2} \sum_{i=1}^{\lfloor nq_a(t) \rfloor} \psi_a(Y_i^{(a)})$, where $\psi_a(\cdot)$ is the efficient influence function corresponding to $\mu(P^{(a)})$.

7. Conclusion

This paper proposes a minimax-regret optimal procedure for determining the best treatment when sampling is costly. The optimal sampling rule is simply the Neyman allocation, while the optimal stopping rule advises that the experiment be terminated when the average difference in outcomes multiplied by the number of observations exceeds a specific threshold. While these rules were derived under diffusion asymptotics, it is shown

that finite sample counterparts of these rules remain optimal under both parametric and non-parametric regimes. The form of these rules is robust to a number of different variations of the original problem, e.g., under batching, different cost functions etc. Given the simple nature of these rules, and the potential for large efficiency gains (requiring, on average, 40% fewer observations than standard approaches), we believe they hold a lot of promise for practical use.

Data availability statement. The data and code underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.14792035.

References

Adusumilli, K. (2021): "Risk and optimal policies in bandit experiments," *arXiv preprint arXiv:2112.06363*.

Armstrong, T. B. (2022): "Asymptotic efficiency bounds for a class of experimental designs," *arXiv preprint arXiv:2205.02726*.

Arrow, K. J., D. Blackwell, and M. A. Girshick (1949): "Bayes and minimax solutions of sequential decision problems," *Econometrica*, 213–244.

Baden, L. R. et al. (2021): "Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine," *New England Journal of Medicine*, 384, 403–416.

Berger, J. O. (2013): *Statistical decision theory and Bayesian analysis*: Springer Science & Business Media.

Chernoff, H. (1959): "Sequential design of experiments," *The Annals of Mathematical Statistics*, 30, 755–770.

Chernoff, H., and A. J. Petkau (1981): "Sequential medical trials involving paired data," *Biometrika*, 68, 119–132.

Colton, T. (1963): "A model for selecting one of two medical treatments," *Journal of the American Statistical Association*, 58, 388–400.

Deng, A., Y. Xu, R. Kohavi, and T. Walker (2013): "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 123–132.

Epstein, L. G., and M. Schneider (2003): "Recursive multiple-priors," *Journal of Economic Theory*, 113, 1–31.

Fan, L., and P. W. Glynn (2021): "Diffusion Approximations for Thompson Sampling," *arXiv preprint arXiv:2105.09232*.

Fehr, E., and A. Rangel (2011): "Neuroeconomic Foundations of Economic Choice–Recent Advances," *Journal of Economic Perspectives*, 25, 3–30.

Fudenberg, D., P. Strack, and T. Strzalecki (2018): "Speed, accuracy, and the optimal timing of choices," *American Economic Review*, 108, 3651–84.

Garivier, A., and E. Kaufmann (2016): "Optimal best arm identification with fixed confidence," in *Conference on Learning Theory*, 998–1027, PMLR.

Hébert, B., and M. Woodford (2017): "Rational inattention and sequential information sampling,"Technical report, National Bureau of Economic Research.

Hirano, K., and J. R. Porter (2009): "Asymptotics for statistical treatment rules," *Econometrica*, 77, 1683–1701.

Karatzas, I., and S. Shreve (2012): *Brownian motion and stochastic calculus* Volume 113: Springer Science & Business Media.

Kuang, X., and S. Wager (2024): "Weak signal asymptotics for sequentially randomized experiments," *Management Science*, 70, 7024–7041.

Lai, T., B. Levin, H. Robbins, and D. Siegmund (1980): "Sequential medical trials," *Proc. Natl. Acad. Sci. U.S.A.*, 77, 3135–3138.

Lattimore, T., and C. Szepesvári (2020): *Bandit algorithms*: Cambridge University Press.

Liang, A., X. Mu, and V. Syrgkanis (2022): "Dynamically aggregating diverse information," *Econometrica*, 90, 47–80.

Luce, R. D. et al. (1986): *Response times: Their role in inferring elementary mental organization* (8): Oxford University Press on Demand.

Manski, C. F. (2021): "Econometrics for decision making: Building foundations sketched by haavelmo and wald," *Econometrica*, 89, 2827–2853.

Manski, C. F., and A. Tetenov (2016): "Sufficient trial size to inform clinical practice," *Proc. Natl. Acad. Sci. U.S.A.*, 113, 10518–10523.

Moore, T. J., H. Zhang, G. Anderson, and G. C. 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, 1451–1457.

Morris, S., and P. Strack (2019): "The Wald problem and the relation of sequential sampling and ex-ante information costs," *Available at SSRN 2991567*.

Øksendal, B. (2003): *Stochastic differential equations*: Springer.

Qin, C., D. Klabjan, and D. Russo (2017): "Improving the expected improvement algorithm," *Advances in Neural Information Processing Systems*, 30.

Qin, C., and D. Russo (2022): "Adaptivity and confounding in multi-armed bandit experiments," *arXiv preprint arXiv:2202.09036*.

Ratcliff, R., and G. McKoon (2008): "The diffusion decision model: theory and data for two-choice decision tasks," *Neural Computation*, 20, 873–922.

Reikvam, K. (1998): "Viscosity solutions of optimal stopping problems," *Stochastics and Stochastic Reports*, 62, 285–301.

Shiryaev, A. N. (2007): *Optimal stopping rules*: Springer Science & Business Media.

Sims, C. A. (2003): "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

Sion, M. (1958): "On general minimax theorems.," *Pacific Journal of Mathematics*, 8, 171–176.

US Food and Drug Admin. (2018): "FDA In Brief: FDA launches new pilot to advance innovative clinical trial designs as part agency's broader program to modernize drug development and promote innovation in drugs targeted to unmet needs," "https://www.fda.gov/news-events/fda-brief/fda-brief-fda-modernizes-clinical-trial-designs-and-approaches-drug-development-pro

——— (2019): "Guidance document: Adaptive Design Clinical Trials for Drugs and Biologics," "https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry".

van der Vaart, A. W. (2000): *Asymptotic statistics*: Cambridge university press.

van der Vaart, A. W., and J. Wellner (1996): *Weak convergence and empirical processes: with applications to statistics*: Springer Science & Business Media.

Wald, A. (1945): "Statistical decision functions which minimize the maximum risk," *Annals of Mathematics*, 265–280.

——— (1947): "Sequential analysis,"Technical report.

Appendix A. Proofs

A.1. **Proof of Theorem 1.** The proof makes use of the following lemmas:

Lemma 1. *Suppose nature sets p_0 to be a symmetric two-point prior supported on $(\sigma_1\Delta/2, -\sigma_0\Delta/2), (-\sigma_1\Delta/2, \sigma_0\Delta/2)$. Then the decision $d(\Delta) = (\pi^*, \tau_{\gamma(\Delta)}, \delta^*)$, where $\gamma(\Delta)$ is defined in (A.5), is a best response by the DM.*

Proof. The prior is an indifference-inducing one, so by the argument given in Section 3.1, the DM is indifferent between any sampling rule π. Thus, $\pi^*_a(t) = \sigma_a/(\sigma_1 + \sigma_0)$ is a best-response to this prior. Also, the prior is symmetric with $m_0 = 1/2$, so by (2.7) and (3.4), the Bayes optimal implementation rule, for any given given stopping time τ, is

$$\delta^\tau = \mathbb{I}\left\{\ln \varphi^{\pi^*}(\tau) \geq 0\right\} = \mathbb{I}\left\{\frac{x_1(\tau)}{\sigma_1} - \frac{x_0(\tau)}{\sigma_0} \geq 0\right\}.$$

It remains to compute the Bayes optimal stopping time. Let $\lambda = 1$ denote the event $\boldsymbol{\mu} = (\sigma_1\Delta/2, -\sigma_0\Delta/2)$, with $\lambda = 0$ otherwise. The discussion in Section 3.1 implies that, conditional on λ, the distribution of the likelihood ratio process $\varphi^\pi(t)$ does not depend on π and evolves as

$$d\ln\varphi(t) = (2\lambda - 1)\frac{\Delta^2}{2}dt + \Delta d\tilde{W}(t), \tag{A.1}$$

where $\tilde{W}(\cdot)$ is a one-dimensional Wiener process. By a similar argument as in Shiryaev (2007, Section 4.2.1), this in turn implies that the posterior probability $m^\pi(t) := \mathbb{P}^\pi(\lambda = 1|\mathcal{F}_t)$ is also independent of π and evolves as[11]

$$dm(t) = \Delta m(t)(1 - m(t))d\tilde{W}(t). \tag{A.2}$$

Therefore, by (2.9) the optimal stopping time also does not depend on π and is given by

$$\tau(\Delta) = \inf_{\tau\in\mathcal{T}} \mathbb{E}\left[\varpi(m(\tau)) + c\tau\right], \text{ where} \tag{A.3}$$

$$\varpi(m) := \frac{(\sigma_1 + \sigma_0)}{2}\Delta\min\left\{m, 1 - m\right\}. \tag{A.4}$$

Inspection of the objective function in (A.3) shows that this is exactly the same objective as in the Bayesian hypothesis testing problem, analyzed previously by Arrow et al. (1949) and Morris and Strack (2019). We follow the analysis of the latter paper. Morris

[11]Shiryaev (2007, Section 4.2.1) analyzes Bayesian updating under binary states, but the setup and notation are slightly different from here. However, the likelihood ratio (LR) processes are the same in both cases, making the problems, and the derivation of (A.2), equivalent. Specifically, in Shiryaev (2007, Equation 4.53), the LR process evolves as $d\ln\varphi(t) = \frac{r}{\sigma^2}(d\xi(t) - \frac{r}{2}dt)$ where $d\xi(t) := r\theta dt + \sigma d\tilde{W}(t)$. Here, $r, \sigma \in \mathbb{R}^+$ are known, and $\theta \in \{0, 1\}$ denotes the unknown binary state of the world. Equating r, σ, θ with $\Delta, 1, \lambda$ then gives the same LR process as (A.1).

and Strack (2019) show that instead of choosing the stopping time τ, it is equivalent to imagine that the DM chooses a probability distribution G over the posterior beliefs $m(\tau)$ at an 'ex-ante' cost

$$c(G) = \frac{2c}{\Delta^2}\int (1-2m)\ln\frac{1-m}{m}dG(m),$$

subject to the constraint $\int m dG(m) = m_0 = 1/2$. The precise form of $c(G)$ is due to Morris and Strack (2019, Proposition 3). Under the distribution G, the expected regret, exclusive of sampling costs, for the DM is

$$\int \varpi(m)dG(m) = \frac{(\sigma_1+\sigma_0)}{2}\Delta\int \min\{m,1-m\}dG(m).$$

Hence, the stopping time, τ, that solves (A.3) is the one that induces the distribution G^*, defined as

$$\begin{aligned} G^* &= \underset{G:\int m dG(m)=\frac{1}{2}}{\arg\min}\left\{c(G)+\int \varpi(m)dG(m)\right\} \\ &= \underset{G:\int m dG(m)=\frac{1}{2}}{\arg\min}\int f(m)dG(m), \end{aligned}$$

where

$$f(m) := \frac{2c}{\Delta^2}(1-2m)\ln\frac{1-m}{m} + \frac{(\sigma_1+\sigma_0)}{2}\Delta\min\{m,1-m\}.$$

Clearly, $f(\cdot)$ is strictly convex on $[0,1/2]$ and $f(m) = f(1-m)$. Hence, setting

$$\alpha(\Delta) := \underset{\alpha\in\left[0,\frac{1}{2}\right]}{\arg\min}\left\{\frac{(\sigma_1+\sigma_0)}{2}\Delta\alpha + \frac{2c}{\Delta^2}(1-2\alpha)\ln\frac{1-\alpha}{\alpha}\right\},$$

it is easy to see that G^* is a two-point distribution, supported on $\alpha(\Delta), 1-\alpha(\Delta)$ with equal probability $1/2$.

We now claim that this distribution is induced by the stopping time $\tau_{\gamma(\Delta)}$, where

$$\gamma(\Delta) := \frac{1}{\Delta}\ln\frac{1-\alpha(\Delta)}{\alpha(\Delta)}. \tag{A.5}$$

To this end, observe that by (2.6) and (3.5),

$$m(t) = \frac{\exp\left\{\Delta\left(\frac{x_1(t)}{\sigma_1}-\frac{x_0(t)}{\sigma_0}\right)\right\}}{1+\exp\left\{\Delta\left(\frac{x_1(t)}{\sigma_1}-\frac{x_0(t)}{\sigma_0}\right)\right\}}.$$

We can then write $\tau_{\gamma(\Delta)}$ in terms of $m(t)$ as

$$\tau_{\gamma(\Delta)} = \inf\left\{t : m(t)\notin[\alpha(\Delta),1-\alpha(\Delta)]\right\}.$$

This immediately implies that the support points of $m(\tau_{\gamma(\Delta)})$ are $\alpha(\Delta)$, $1-\alpha(\Delta)$; also,

$$\mathbb{E}[m(\tau_{\gamma(\Delta)})] = (1-\alpha(\Delta))\,\mathbb{P}\left(m(\tau_{\gamma(\Delta)}) = 1-\alpha(\Delta)\right) + \alpha(\Delta)\mathbb{P}\left(m(\tau_{\gamma(\Delta)}) = \alpha(\Delta)\right).$$

But $m(\cdot)$ is a martingale, so Doob's optional stopping theorem implies $\mathbb{E}[m(\tau_{\gamma(\Delta)})] = \mathbb{E}[m(0)] = 1/2$. Equating the two expressions for $\mathbb{E}[m(\tau_{\gamma(\Delta)})]$ gives

$$\mathbb{P}\left(m(\tau_{\gamma(\Delta)}) = 1-\alpha(\Delta)\right) = \mathbb{P}\left(m(\tau_{\gamma(\Delta)}) = \alpha(\Delta)\right) = 1/2.$$

Thus, $m(\tau_{\gamma(\Delta)})$ is distributed as G^*, and the stopping time $\tau_{\gamma(\Delta)}$ is therefore the best response to nature's prior. □

Lemma 2. *Suppose $\boldsymbol{\mu}$ is such that $|\mu_1 - \mu_0| = \frac{\sigma_1+\sigma_0}{2}\Delta$. Then, for any $\gamma, \Delta > 0$,*

$$V\left(\tilde{\boldsymbol{d}}_\gamma, \boldsymbol{\mu}\right) = \frac{(\sigma_1+\sigma_0)}{2}\Delta\frac{1-e^{-\Delta\gamma}}{e^{\Delta\gamma}-e^{-\Delta\gamma}} + \frac{2c\gamma}{\Delta}\frac{e^{\Delta\gamma}+e^{-\Delta\gamma}-2}{e^{\Delta\gamma}-e^{-\Delta\gamma}}.$$

Thus, the frequentist regret of $\tilde{\boldsymbol{d}}_\gamma$ depends on $\boldsymbol{\mu}$ only through $|\mu_1-\mu_0|$.

Proof. Suppose that $\mu_1 > \mu_0$. Define

$$\xi(t) := \left(\frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0}\right)\cdot\Delta.$$

Note that under $\tilde{\boldsymbol{d}}_\gamma$ and $\boldsymbol{\mu}$,

$$\frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0} = \frac{\Delta}{2}t + \tilde{W}(t),$$

where $\tilde{W}(\cdot)$ is one-dimensional Brownian motion. Hence $\xi(t) = \frac{\Delta^2}{2}t + \Delta\tilde{W}(t)$. We can write the stopping time τ_γ in terms of $\xi(t)$ as

$$\tau_\gamma = \inf\left\{t : \left|\frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0}\right| \geq \gamma\right\} = \inf\left\{t : |\xi(t)| \geq \Delta\gamma\right\},$$

and the implementation rule as $\delta^{\tau_\gamma} = \mathbb{I}\left\{\xi(\tau_\gamma) \geq 0\right\} = \mathbb{I}\left\{\xi(\tau_\gamma) = \Delta\gamma\right\}$.

Now, noting the form of $\xi(t)$, we can apply similar arguments as in Shiryaev (2007, Section 4.2, Lemma 5), to show that

$$\mathbb{E}\left[\tau_\gamma|\boldsymbol{\mu}\right] = \frac{2}{\Delta^2}\frac{\Delta\gamma\left(e^{\Delta\gamma}+e^{-\Delta\gamma}-2\right)}{e^{\Delta\gamma}-e^{-\Delta\gamma}}.$$

Furthermore, following Shiryaev (2007, Section 4.2, Lemma 4), we also have

$$\mathbb{P}(\delta^{\tau_\gamma} = 0|\boldsymbol{\mu}) = \mathbb{P}(\xi(\tau_\gamma) = -\Delta\gamma|\boldsymbol{\mu}) = \frac{1-e^{-\Delta\gamma}}{e^{\Delta\gamma}-e^{-\Delta\gamma}}.$$

Hence, the frequentist regret is given by

$$V\left(\tilde{\boldsymbol{d}}_\gamma, \boldsymbol{\mu}\right) = \frac{\sigma_1+\sigma_0}{2}\Delta\mathbb{P}(\delta^{\tau_\gamma}=0|\boldsymbol{\mu}) + c\mathbb{E}\left[\tau_\gamma|\boldsymbol{\mu}\right]$$
$$= \frac{(\sigma_1+\sigma_0)}{2}\Delta\frac{1-e^{-\Delta\gamma}}{e^{\Delta\gamma}-e^{-\Delta\gamma}} + \frac{2c\gamma}{\Delta}\frac{e^{\Delta\gamma}+e^{-\Delta\gamma}-2}{e^{\Delta\gamma}-e^{-\Delta\gamma}}.$$

While the above was shown under $\mu_1 > \mu_0$, an analogous argument under $\mu_1 < \mu_0$ gives the same expression for $V\left(\tilde{\boldsymbol{d}}_\gamma, \boldsymbol{\mu}\right)$. □

Lemma 3. *Consider a two-player zero-sum game in which nature chooses $\Delta \in \mathbb{R}^+$ indexing the indifference prior p_Δ and the DM chooses $\gamma \in \mathbb{R}^+$ indexing the decision rule $\boldsymbol{d}_\gamma = (\pi^*, \tau_\gamma, \delta^{\tau_\gamma})$. There exists a unique Nash equilibrium to this game at $\Delta^* = \eta\Delta_0^*$ and $\gamma^* = \eta^{-1}\gamma_0^*$, where $\eta, \Delta_0^*, \gamma_0^*$ are defined in Section 3.*

Proof. By Lemma 2, the frequentist regret under a given choice of $\Delta := 2|\mu_1-\mu_0|/(\sigma_1+\sigma_0)$ and γ is given by $\frac{(\sigma_1+\sigma_0)}{2}R(\gamma,\Delta)$, where

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

Lemma 2 further implies that the frequentist regret $V(\boldsymbol{d}^*, \boldsymbol{\mu})$ depends on $\boldsymbol{\mu}$ only through Δ. Therefore, the frequentist regret under both support points of p_Δ must be the same. Hence, the Bayes regret, $V(\boldsymbol{d}_\gamma, p_\Delta)$, is the same as the frequentist regret at each support point, i.e.,

$$V(\boldsymbol{d}_\gamma, p_\Delta) = \frac{(\sigma_1+\sigma_0)}{2}R(\gamma,\Delta). \tag{A.6}$$

We aim to find a Nash equilibrium in a two-player game in which natures chooses p_Δ, equivalently Δ, to maximize $R(\gamma,\Delta)$, while the DM chooses $\boldsymbol{d}_\gamma$, equivalently γ, to minimize $R(\gamma,\Delta)$.

For $\eta = 1$, the unique Nash equilibrium to this game is given by $\Delta = \Delta_0^*$ and $\gamma = \gamma_0^*$. We start by first demonstrating the existence of a Nash equilibrium. This is guaranteed by Sion's minimax theorem (Sion, 1958) as long as $R(\gamma,\Delta)$ is continuous in both arguments (which is easily verified), and 'convex quasi-concave' on $\mathbb{R}^+\times\mathbb{R}^+\backslash\{0\}$.[12] To show convexity in the first argument, write $R(\cdot,\Delta) = R_1(\alpha(\cdot,\Delta),\Delta)$ where

$$R_1(\alpha,\Delta) := \Delta\alpha + \frac{2}{\Delta^2}(1-2\alpha)\ln\frac{1-\alpha}{\alpha};\ \text{and}$$
$$\alpha(\gamma,\Delta) := \frac{1-e^{-\Delta\gamma}}{e^{\Delta\gamma}-e^{-\Delta\gamma}}.$$

[12] In fact, convexity can be replaced with quasi-convexity for the theorem.

Now, for any fixed $\Delta > 0$, it is easy to verify that $R_1(\cdot, \Delta)$ and $\alpha(\cdot, \Delta)$ are convex over the domain $\mathbb{R}^+$. Since the composition of convex functions is also convex, this proves convexity of $R(\cdot, \Delta)$. To prove $R(\gamma, \cdot)$ is quasi-concave, write $R(\gamma, \cdot) = R_2(\gamma, \alpha(\gamma, \cdot))$, where

$$R_2(\gamma, \alpha) := \frac{1}{\gamma}\alpha \ln \frac{1-\alpha}{\alpha} + 2\gamma^2 \frac{(1-2\alpha)}{\ln \frac{1-\alpha}{\alpha}}.$$

Now, $\alpha \ln \frac{1-\alpha}{\alpha}$ and $(1-2\alpha)/\ln \frac{1-\alpha}{\alpha}$ are concave over $\mathbb{R}^+\backslash\{0\}$, so $R_2(\gamma, \cdot)$ is also concave over $\mathbb{R}^+\backslash\{0\}$ for any fixed $\gamma > 0$. Concavity implies the level set $\{\alpha : R_2(\gamma, \alpha) \geq \nu\}$ is a closed interval in $\mathbb{R}^+\backslash\{0\}$ for any $\nu \in \mathbb{R}$. But $\alpha(\gamma, \cdot)$ is positive and strictly decreasing, so for a fixed $\gamma > 0$,

$$\{\Delta : R(\gamma, \Delta) \geq \nu\} \equiv \{\Delta : R_2(\gamma, \alpha(\gamma, \Delta)) \geq \nu\}$$

is also a closed interval in $\mathbb{R}^+\backslash\{0\}$, and therefore, convex, for any $\nu \in \mathbb{R}$. This proves quasi-concavity of $R(\gamma, \cdot)$ whenever $\gamma > 0$. At the same time, $R(\gamma, \Delta) = \Delta/2$ when $\gamma = 0$; hence, $R(\gamma, \cdot)$ is in fact quasi-concave for any $\gamma \geq 0$. We thus conclude by Sion's theorem that a Nash equilibrium exists. It is then routine to numerically compute Δ_0^*, γ_0^* through first-order conditions and show that these values are unique; we skip these calculations, which are straightforward. Figure A.1 provides a graphical illustration of the Nash equilibrium.

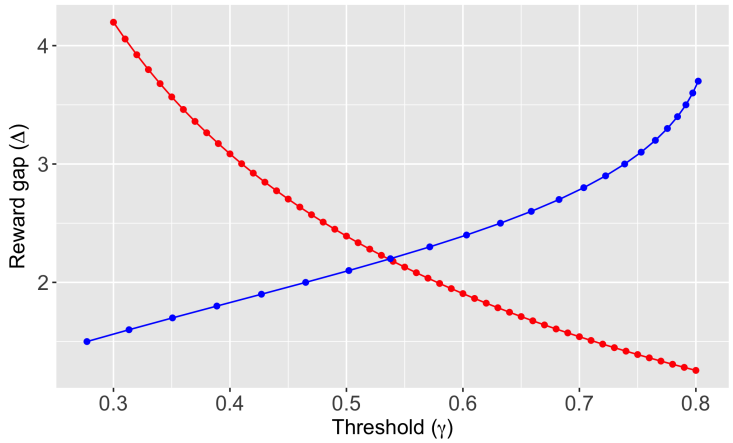
It remains to determine the Nash equilibrium under general η. By the form of $R(\gamma, \Delta)$, if γ_0^* is a best response to Δ_0^* for $\eta = 1$, then $\eta^{-1}\gamma_0^*$ is a best response to $\eta\Delta_0^*$ for general η. Similarly, if Δ_0^* is a best response to γ_0^* for $\eta = 1$, then $\eta\Delta_0^*$ is a best response to $\eta^{-1}\gamma_0^*$ for general η. This proves $\Delta^* := \eta\Delta_0^*$ and $\gamma^* := \eta^{-1}\gamma_0^*$ is a Nash equilibrium in the general case. □

We now complete the proof of Theorem 1: By Lemmas 1 and 3, $\boldsymbol{d}^*$ is the optimal Bayes decision corresponding to p_0^*. We now show

$$\sup_{\boldsymbol{\mu}} V(\boldsymbol{d}^*, \boldsymbol{\mu}) = V(\boldsymbol{d}^*, p_0^*), \tag{A.7}$$

which implies $\boldsymbol{d}^*$ is minimax optimal according to the verification theorem in Berger (2013, Theorem 17). To this end, recall from Lemma 2 that the frequentist regret $V(\boldsymbol{d}^*, \boldsymbol{\mu})$ depends on $\boldsymbol{\mu}$ only through $\Delta := 2|\mu_1 - \mu_0|/(\sigma_1 + \sigma_0)$. Furthermore, by Lemma 3, Δ^* is the best response of nature to $\boldsymbol{d}^*$. These results imply

$$\sup_{\boldsymbol{\mu}} V(\boldsymbol{d}^*, \boldsymbol{\mu}) = \frac{(\sigma_1 + \sigma_0)}{2} \sup_{\Delta} R(\gamma^*, \Delta) = \frac{(\sigma_1 + \sigma_0)}{2} R(\gamma^*, \Delta^*).$$



Note: The upward-sloping curve describes the best response of Δ to a given γ, while the downward sloping curve describes the best response of γ to a given Δ. The point of intersection is the Nash equilibrium. This is for $\eta = 1$.

FIGURE A.1. Best responses and Nash equilibrium

But by (A.6), we also have $V(\boldsymbol{d}^*, p_0^*) = \frac{(\sigma_1+\sigma_0)}{2} R(\gamma^*, \Delta^*)$. This proves (A.7).

A.2. **Proof of Corollary 1.** We employ the same strategy as in the proof of Theorem 1. Suppose nature employs the indifference prior p_Δ, for any $\Delta > 0$. Then by similar arguments as earlier, the DM is indifferent between any sampling rule π, and the optimal implementation rule is $\delta^* = \mathbb{I}\left\{\frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \geq 0\right\}$.

We now determine Nature's best response to the DM choosing $\boldsymbol{d}^* = (\pi^*, \delta^*)$, where π^* is the Neyman allocation. Consider an arbitrary $\boldsymbol{\mu} = (\mu_1, \mu_0)$ such that $|\mu_1 - \mu_0| = \frac{\sigma_1+\sigma_0}{2}\Delta$. Suppose $\mu_1 > \mu_0$. Under π^*,

$$\frac{dx_1(t)}{\sigma_1} - \frac{dx_0(t)}{\sigma_0} = \frac{\Delta}{2}dt + d\tilde{W}(t),$$

where $\tilde{W}(\cdot)$ is the standard Wiener process, so the expected regret under $\boldsymbol{d}^*, \boldsymbol{\mu}$ is

$$V\left(\boldsymbol{d}^*, \boldsymbol{\mu}\right) = (\mu_1 - \mu_0)\mathbb{P}\left(\frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \leq 0\right) = \frac{\sigma_1 + \sigma_0}{2}\Delta\Phi\left(-\frac{\Delta}{2}\right). \tag{A.8}$$

An analogous argument shows that the same expression holds when $\mu_0 > \mu_1$ as well. Consequently, nature's optimal choice of $\boldsymbol{\mu}$ is to set Δ to $\bar{\Delta}^* = 2\arg\max_\delta \delta\Phi\left(-\delta\right)$, but is otherwise indifferent between any $\boldsymbol{\mu}$ such that $|\mu_1 - \mu_0| = \frac{\sigma_1+\sigma_0}{2}\bar{\Delta}^*$. Thus, $p_{\bar{\Delta}^*}$ is a best response by nature to the DM's choice of $\boldsymbol{d}^* = (\pi^*, \delta^*)$.

We have thereby shown $p_{\bar{\Delta}^*}, \boldsymbol{d}^*$ form a Nash equilibrium. That $\boldsymbol{d}^*$ is minimax optimal then follows by similar arguments as in the proof of Theorem 1.

A.3. **Proof of Theorem 2.** Our aim is to show (4.4). The outline of the proof is as follows: First, as in Adusumilli (2021), we replace the true marginal and posterior distributions with suitable approximations. Next, we apply dynamic programming arguments and viscosity solution techniques to obtain a HJB-variational inequality (HJB-VI) for the value function in the experiment. Finally, the HJB-VI is connected to the problem of determining the optimal stopping time under diffusion asymptotics.

Step 0 (Definitions and preliminary observations). Under m_0^*, let $\gamma = 1$ denote the event $\boldsymbol{h} = (h_1^*, -h_0^*)$ and $\gamma = 0$ the event $\boldsymbol{h} = (-h_1^*, h_0^*)$. Also, let $\mathbf{y}_{nq}^{(a)} := \{Y_i^{(a)}\}_{i=1}^{\lfloor nq \rfloor}$ denote the stacked representation of outcomes $Y_i^{(a)}$ from the first nq observations corresponding to treatment a, and for any $\boldsymbol{h} := (h_1, h_0)$, take $P_{nq_1,nq_0,\boldsymbol{h}}$ to be the distribution corresponding to the joint density $p_{nq_1,h_1}(\mathbf{y}_{nq_1}^{(1)}) \cdot p_{nq_0,h_1}(\mathbf{y}_{nq_0}^{(0)})$, where

$$p_{nq,h_a}(\mathbf{y}_{nq_1}^{(a)}) := \prod_{i=1}^{nq} p_{n,h_a}(Y_i^{(a)}).$$

Also, define $\bar{P}_n$ as the marginal distribution of $\left(\mathbf{y}_{nT}^{(1)}, \mathbf{y}_{nT}^{(0)}\right)$, i.e., it is the probability measure whose density, with respect to the dominating measure $\nu(\mathbf{y}_{nT}^{(1)}, \mathbf{y}_{nT}^{(0)}) := \prod_{a\in\{0,1\}} \nu(Y_1^{(a)}) \times \cdots \times \nu(Y_{nT}^{(a)})$, is

$$\bar{p}_n\left(\mathbf{y}_{nT}^{(1)}, \mathbf{y}_{nT}^{(0)}\right) = \int p_{nT,h_1}(\mathbf{y}_{nT}^{(1)}) \cdot p_{nT,h_1}(\mathbf{y}_{nT}^{(0)}) dm_0^*(\boldsymbol{h}).$$

Due to the two-point support of m_0^*, the posterior density $p_n(\cdot|\xi_t)$ can be associated with a scalar,

$$m_n(\xi_t) := m_n\left(\mathbf{y}_{nq_1(t)}^{(1)}, \mathbf{y}_{nq_0(t)}^{(0)}\right) := P_n\left(\gamma = 1 | \mathbf{y}_{nq_1(t)}^{(1)}, \mathbf{y}_{nq_0(t)}^{(0)}\right).$$

That the posterior depends on ξ_t only via $\mathbf{y}_{nq_1(t)}^{(1)}, \mathbf{y}_{nq_0(t)}^{(0)}$ is an immediate consequence of Adusumilli (2021, Lemma 1). Recalling the definition of $\varpi_n(\cdot)$ in (4.3), we have $\varpi_n(\xi_t) = \varpi_n(m_n(\xi_t))$, where, for any $m \in [0,1]$,

$$\begin{aligned}\varpi_n(m) &:= \min\left\{\{\mu_{n,0}(h_0^*) - \mu_{n,1}(-h_1^*)\}(1-m), \{\mu_{n,1}(h_1^*) - \mu_{n,0}(-h_0^*)\} m\right\} \\ &= (\mu_{n,1}(h_1^*) - \mu_{n,0}(-h_0^*)) \min\{m, 1-m\}.\end{aligned}$$

The first equation above always holds, while the second holds under the simplification $\mu_{n,a}(h) = -\mu_{n,a}(-h)$ described in Section 4.

Let

$$z_{a,nq_a} := \frac{I_a^{-1/2}}{\sqrt{n}} \sum_{i=1}^{\lfloor nq_a \rfloor} \psi_a(Y_i^{(a)}), \tag{A.9}$$

denote the (standardized) score process. Under quadratic mean differentiability - Assumption 1(i) - the following SLAN property holds for both treatments:

$$\sum_{i=1}^{\lfloor nq_a \rfloor} \ln \frac{dp^{(a)}_{\theta_0+h/\sqrt{n}}}{dp^{(a)}_{\theta_0}}(Y_i^{(a)}) = h^\intercal I_a^{1/2} z_{a,nq_a} - \frac{q_a}{2} h^\intercal I_a h + o_{P^{(a)}_{nT,\theta_0}}(1), \text{ uniformly over bounded } q_a. \tag{A.10}$$

See Adusumilli (2021, Lemma 2) for the proof.[13]

As in Adusumilli (2021), we now define approximate versions of the true marginal and posterior by replacing the actual likelihood $\prod_a p^{(a)}_{nq_a,h_a}(\mathbf{y}^{(a)}_{nT})$ with

$$\prod_a \lambda^{(a)}_{nq,h_a}(\mathbf{y}^{(a)}_{nq}) := \prod_a \frac{d\Lambda^{(a)}_{nq,h_a}(\mathbf{y}^{(a)}_{nq})}{d\nu}, \text{ where}$$
$$\lambda^{(a)}_{nq,h}(\mathbf{y}^{(a)}_{nq}) := \exp\left\{h^\intercal I_a^{1/2} z_{a,nq} - \frac{q}{2} h^\intercal I_a h\right\} p^{(a)}_{nq,\theta_0}(\mathbf{y}^{(a)}_{nq}) \ \forall\ q \in [0,T]. \tag{A.11}$$

In other words, we approximate the true likelihood with the first two terms in the SLAN expansion (A.10). The construction of the approximate marginal and posterior is described below.

(Approximate marginal:) Denote by $\tilde{P}_{nq_1,nq_0,\boldsymbol{h}}$ the measure whose density is $\lambda^{(1)}_{nq_1,h_1}(\mathbf{y}^{(1)}_{nq_1}) \cdot \lambda^{(0)}_{nq_0,h_0}(\mathbf{y}^{(0)}_{nq_0})$, and take $\tilde{\tilde{P}}_{nq_1,nq_0}$ to be its marginal over $\mathbf{y}^{(1)}_{nq_1}, \mathbf{y}^{(0)}_{nq_0}$ given the prior $m^*_0(\boldsymbol{h})$. Note that the density (wrt ν) of $\tilde{\tilde{P}}_{nq_1,nq_0}$ is

$$\tilde{\tilde{p}}_{nq_1,nq_0}\left(\mathbf{y}^{(1)}_{nq_1}, \mathbf{y}^{(0)}_{nq_0}\right) = \int \lambda^{(1)}_{nq_1,h^{(1)}}\left(\mathbf{y}^{(1)}_{nq_1}\right) \cdot \lambda^{(0)}_{nq_0,h^{(0)}}\left(\mathbf{y}^{(0)}_{nq_0}\right) dm^*_0(\boldsymbol{h}). \tag{A.12}$$

Also, define $\tilde{\tilde{p}}_n\left(\mathbf{y}^{(1)}_{nT}, \mathbf{y}^{(0)}_{nT}\right) := \tilde{\tilde{p}}_{nT,nT}\left(\mathbf{y}^{(1)}_{nT}, \mathbf{y}^{(0)}_{nT}\right)$. Then, $\tilde{\tilde{p}}_n\left(\mathbf{y}^{(1)}_{nT}, \mathbf{y}^{(0)}_{nT}\right)$ approximates the true marginal $\bar{p}_n\left(\mathbf{y}^{(1)}_{nT}, \mathbf{y}^{(0)}_{nT}\right)$.

(Approximate posterior:) Next, let $\tilde{\varphi}(t)$ be the approximate likelihood ratio

$$\tilde{\varphi}(t) = \frac{\lambda^{(1)}_{nq_1,h_1^*}\left(\mathbf{y}^{(1)}_{nq_1(t)}\right) \cdot \lambda^{(0)}_{nq_0,-h_0^*}\left(\mathbf{y}^{(0)}_{nq_0(t)}\right)}{\lambda^{(1)}_{nq_1,-h_1^*}\left(\mathbf{y}^{(1)}_{nq_1(t)}\right) \cdot \lambda^{(0)}_{nq_0,h_0^*}\left(\mathbf{y}^{(0)}_{nq_0(t)}\right)} = \exp\left\{\Delta^* \rho_n(t)\right\},$$

where

$$\rho_n(t) := \frac{\dot{\mu}_1^\intercal z_{1,nq_1(t)}}{\sigma_1} - \frac{\dot{\mu}_0^\intercal z_{0,nq_0(t)}}{\sigma_0}. \tag{A.13}$$

Based on the above, we can approximate the true posterior, $m_n(\xi_t)$, by

$$\frac{\tilde{\varphi}(t)}{1+\tilde{\varphi}(t)} = \frac{\exp\left\{\Delta^* \rho_n(t)\right\}}{1+\exp\left\{\Delta^* \rho_n(t)\right\}} := \tilde{m}(\rho_n(t)), \tag{A.14}$$

[13] It should be noted that the score process in that paper is defined slightly differently, as $I_a^{-1/2} z_{a,nq_a}$ under the present notation.

where $\tilde{m}(\rho) := \exp(\Delta^*\rho)/(1+\exp(\Delta^*\rho))$ for $\rho \in \mathbb{R}$. When $\rho_n(t) = \rho$, the approximate posterior $\tilde{m}(\rho)$ in turn implies an approximate posterior, $\tilde{p}_n(\boldsymbol{h}|\rho)$, over $\boldsymbol{h}$ that takes the value $(h_1^*, -h_0^*)$ with probability $\tilde{m}(\rho)$ and $(-h_1^*, h_0^*)$ with probability $1-\tilde{m}(\rho)$.

Step 1 (Posterior and probability approximations). Set $V_{n,T}^* = \inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} V_n^*(\boldsymbol{d}, m_0^*)$. Using dynamic programming arguments, it is straightforward to show that there exists a non-randomized sampling rule and stopping time that minimizes $V_n^*(\boldsymbol{d}, m_0)$ for any prior m_0. We therefore restrict $\mathcal{D}_{n,T}$ to the set of all deterministic rules, $\bar{\mathcal{D}}_{n,T}$. Under deterministic policies, the sampling rules π_{nt}, states ξ_t and stopping times τ are all deterministic functions of $\mathbf{y}_{nT}^{(1)}, \mathbf{y}_{nT}^{(0)}$. Recall that $\mathbf{y}_{nT}^{(1)}, \mathbf{y}_{nT}^{(0)}$ are the stacked vector of outcomes under nT observations of each treatment. It is useful to think of $\{\pi_{nt}\}_{t=1/n}^{T}, \tau$ as quantities mapping $(\mathbf{y}_{nT}^{(1)}, \mathbf{y}_{nT}^{(0)})$ to realizations of regret.[14] Taking $\bar{\mathbb{E}}_n[\cdot]$ to be the expectation under $\bar{P}_n$, we then have

$$V_n^*(\boldsymbol{d}, m_0^*) = \bar{\mathbb{E}}_n\left[\sqrt{n}\varpi_n\left(m_n\left(\xi_\tau\right)\right) + c\tau\right],$$

for any deterministic $\boldsymbol{d} \in \bar{\mathcal{D}}_{n,T}$.

Now, take $\bar{\tilde{\mathbb{E}}}_n[\cdot]$ to be the expectation under $\bar{\tilde{P}}_n$, and define

$$\tilde{V}_n(\boldsymbol{d}, m_0^*) = \bar{\tilde{\mathbb{E}}}_n\left[\sqrt{n}\varpi_n\left(\tilde{m}\left(\rho_n(\tau)\right)\right) + c\tau\right]. \tag{A.15}$$

By Lemma 7 in Appendix F,

$$\lim_{n\to\infty} \sup_{\boldsymbol{d}\in\bar{\mathcal{D}}_{n,T}} \left|V_n^*(\boldsymbol{d}, m_0^*) - \tilde{V}_n(\boldsymbol{d}, m_0^*)\right| = 0.$$

This in turn implies $\lim_{n\to\infty}\left|V_{n,T}^* - \tilde{V}_{n,T}^*\right| = 0$, where $\tilde{V}_{n,T}^* := \inf_{\boldsymbol{d}\in\bar{\mathcal{D}}_{n,T}} \tilde{V}_n^*(\boldsymbol{d}, m_0^*)$.

Step 2 (Recursive formula for $\tilde{V}_{n,T}^*$). We now employ dynamic programming arguments to obtain a recursion for $\tilde{V}_{n,T}^*$. This requires a bit of care since $\bar{\tilde{P}}_n$ is not a probability, even though it does integrate to 1 asymptotically.

[14] Note that π, τ still need to satisfy the measurability restrictions, and some components of $\mathbf{y}_{nT}^{(a)}$ may not be observed as both treatments cannot be sampled nT times.

Recall that $\tilde{p}_n(\boldsymbol{h}|\rho)$ is the probability measure on $\boldsymbol{h}$ that assigns probability $\tilde{m}(\rho)$ to $(h_1^*, -h_0^*)$ and probability $1-\tilde{m}(\rho)$ to $(-h_1^*, h_0^*)$. Define

$$\tilde{p}_n(Y^{(a)}|\rho) = p_{\theta_0}^{(a)}(Y^{(a)}) \cdot \int \exp\left\{\frac{1}{\sqrt{n}} h_a^\intercal \psi_a(Y^{(a)}) - \frac{1}{2n} h_a^\intercal I_a h_a\right\} d\tilde{p}_n(\boldsymbol{h}|\rho),$$

$$\tilde{\tilde{p}}_n(\mathbf{y}_{-nq_1}^{(1)}, \mathbf{y}_{-nq_0}^{(0)}|\rho, q_1, q_0) = \int \frac{\lambda_{nT,h_1}^{(1)}\left(\mathbf{y}_{nT}^{(1)}\right) \cdot \lambda_{nT,h_0}^{(0)}\left(\mathbf{y}_{nT}^{(0)}\right)}{\lambda_{nq_1,h_1}^{(1)}\left(\mathbf{y}_{nq_1}^{(1)}\right) \cdot \lambda_{nq_0,h_0}^{(0)}\left(\mathbf{y}_{nq_0}^{(0)}\right)} d\tilde{p}_n(\boldsymbol{h}|\rho), \quad \text{and}$$

$$\eta(\rho, q_1, q_0) = \int d\tilde{\tilde{p}}_n\left(\mathbf{y}_{-nq_1}^{(1)}, \mathbf{y}_{-nq_0}^{(0)}|\rho, q_1, q_0\right), \tag{A.16}$$

where $\mathbf{y}_{-nq}^{(a)} := \{Y_{nq+1}^{(a)}, \ldots, Y_{nT}^{(a)}\}$. In words, $\tilde{\tilde{p}}_n(\mathbf{y}_{-nq_1}^{(1)}, \mathbf{y}_{-nq_0}^{(0)}|\rho, q_1, q_0)$ is the approximate probability density over the future values of the stacked rewards $\{Y_i^{(a)}\}_{i=nq_a+1}^{nT}$ given the current state ρ, q_1, q_0. Note that $\eta(\rho, q_1, q_0)$ is the normalization constant of $\tilde{\tilde{p}}_n(\mathbf{y}_{-nq_1}^{(1)}, \mathbf{y}_{-nq_0}^{(0)}|\rho, q_1, q_0)$.

By Lemma 8 in Appendix F, $\tilde{V}_{n,T}^* = \tilde{V}_{n,T}^*(0,0,0,0)$, where $\tilde{V}_{n,T}^*(\cdot)$ solves the recursion

$$\tilde{V}_{n,T}^*(\rho, q_1, q_0, t) = \min\Bigg\{\sqrt{n}\eta(\rho, q_1, q_0)\varpi_n(\tilde{m}(\rho)),$$
$$\frac{\eta(\rho, q_1, q_0)c}{n} + \min_{a\in\{0,1\}} \int \tilde{V}_{n,T}^*\left(\rho + \frac{(2a-1)\dot{\mu}_a^\intercal I_a^{-1}\psi_a(Y^{(a)})}{\sqrt{n}\sigma_a}, q_1 + \frac{a}{n}, q_0 + \frac{1-a}{n}, t + \frac{1}{n}\right) d\tilde{p}_n(Y^{(a)}|\rho)\Bigg\}, \tag{A.17}$$

for $t \leq T$, and

$$\tilde{V}_{n,T}^*(\rho, q_1, q_0, T) = \sqrt{n}\eta(\rho, q_1, q_0)\varpi_n(\tilde{m}(\rho)).$$

The function $\eta(\cdot)$ accounts for the fact $\tilde{\tilde{P}}_n$ is not a probability.

Now, Lemma 9 in Appendix F shows that

$$\sup_{\rho, q_1, q_0} |\eta(\rho, q_1, q_0) - 1| \leq M n^{-\vartheta} \tag{A.18}$$

for some $M < \infty$ and any $\vartheta \in (0, 1/2)$. Furthermore, by Assumption 1(iii),

$$\lim_{n\to\infty} \sup_{m\in[0,1]} \left|\sqrt{n}\varpi_n(m) - \varpi(m)\right| = 0, \tag{A.19}$$

where $\varpi(m) := \frac{\sigma_1+\sigma_0}{2}\Delta^* \min\{m, 1-m\}$. Since $\varpi(\cdot)$ is uniformly bounded, it follows from (A.19) that $\sqrt{n}\varpi_n(\cdot)$ is also uniformly bounded. Then, (A.18) and (A.19) imply

$$\lim_{n\to\infty} \left|\tilde{V}_{n,T}^*(0) - \breve{V}_{n,T}^*(0)\right| = 0,$$

where $\breve{V}_{n,T}(\rho,t)$ is defined as the solution to the recursion

$$\breve{V}^*_{n,T}(\rho,t) = \min\left\{\varpi(\tilde{m}(\rho)), \frac{c}{n} + \min_{a\in\{0,1\}} \int \breve{V}^*_{n,T}\left(\rho + \frac{(2a-1)\dot{\mu}_a^\intercal I_a^{-1}\psi_a(Y^{(a)})}{\sqrt{n}\sigma_a}, t+\frac{1}{n}\right) d\tilde{p}_n(Y^{(a)}|\rho)\right\}$$
$$\text{for } t \le T, \tag{A.20}$$
$$\breve{V}^*_{n,T}(\rho,T) = \varpi(\tilde{m}(\rho)).$$

We can drop the state variables q_1, q_0 in $\breve{V}^*_{n,T}(\cdot)$ as they enter the definition of $\tilde{V}^*_{n,T}(\rho,q_1,q_0,t)$ only via $\eta(\rho,q_1,q_0)$, which was shown in (A.18) to be uniformly close to 1.

Step 3 (PDE approximation and relationship to optimal stopping). For any $\rho\in\mathbb{R}$, let

$$\varpi(\rho) := \varpi(\tilde{m}(\rho)) = \frac{(\sigma_1+\sigma_0)\Delta^*}{2}\min\left\{\frac{\exp(\Delta^*\rho)}{1+\exp(\Delta^*\rho)}, \frac{1}{1+\exp(\Delta^*\rho)}\right\}.$$

Lemma 10 in Appendix F shows that $\breve{V}^*_{n,T}(\cdot)$ converges locally uniformly to $V^*_T(\cdot)$, the unique viscosity solution of the HJB-VI

$$\min\left\{\varpi(\rho) - V^*_T(\rho,t), c + \partial_t V^*_T + \frac{\Delta^*}{2}(2\tilde{m}(\rho)-1)\partial_\rho V^*_T + \frac{1}{2}\partial^2_\rho V^*_T\right\} = 0 \text{ for } t\le T,$$
$$V^*_T(\rho,T) = \varpi(\rho). \tag{A.21}$$

Note that the sampling rule does not enter the HJB-VI. This is a consequence of the choice of the prior, m^*_0.

There is a well known connection between HJB-VIs and the problem of optimal stopping that goes by the name of smooth-pasting or the high contact principle, see Øksendal (2003, Chapter 10) for an overview. In the present context, letting $W(t)$ denote one-dimensional Brownian motion, it follows by Reikvam (1998) that

$$V^*_T(0,0) = \inf_{\tau\le T}\mathbb{E}\left[\varpi(\rho_\tau) + c\tau\right], \text{ where}$$
$$d\rho_t = \frac{\Delta^*}{2}(2\tilde{m}(\rho_t)-1)dt + dW(t);\ \rho_0 = 0,$$

and τ is the set of all stopping times adapted to the filtration $\mathcal{F}_t$ generated by ρ_t.

Step 4 (Taking $T\to\infty$). Through steps 1-3, we have shown

$$\lim_{n\to\infty}\inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}}\sup_{\boldsymbol{h}} V_n(\boldsymbol{d},\boldsymbol{h}) \ge \lim_{n\to\infty}\inf_{\boldsymbol{d}\in\mathcal{D}_{n,T}} V_n(\boldsymbol{d},m^*_0) = V^*_T(0,0).$$

We now argue that

$$\lim_{T\to\infty} V^*_T(0,0) = V^*_\infty := \inf_\tau \mathbb{E}\left[\varpi(\rho_\tau) + c\tau\right].$$

Suppose not: Then, there exists $\epsilon > 0$, a sequence $\{T_j\}_j$ with $T_j \uparrow \infty$, and some stopping time $\bar{\tau}$ such that $V(\bar{\tau}) := \mathbb{E}[\varpi(\rho_{\bar{\tau}}) + c\bar{\tau}] < V^*_{T_j}(0,0) - \epsilon$ for all j (note that we always have $V^*_T(0,0) \geq V^*_\infty$ by definition). Now, $\varpi(\cdot)$ is uniformly bounded, so by the dominated convergence theorem, $\lim_{j\to\infty} \mathbb{E}\left[\varpi(\rho_{\bar{\tau}\wedge T_j})\right] = \mathbb{E}[\varpi(\rho_{\bar{\tau}})]$. Hence,

$$\begin{aligned}\lim_{j\to\infty} V^*_{T_j}(0,0) &\leq \lim_{j\to\infty} \mathbb{E}\left[\varpi(\rho_{\bar{\tau}\wedge T_j}) + c\left(\bar{\tau}\wedge T_j\right)\right] \\ &= \mathbb{E}[\varpi(\rho_{\bar{\tau}})] + \lim_{j\to\infty} c\mathbb{E}[\bar{\tau}\wedge T_j] \leq V(\bar{\tau}).\end{aligned}$$

This is a contradiction.

It remains to show V^*_∞ is the same as V^*, the value of the two-player game in Theorem 1. Define

$$m_t = \frac{\exp(\Delta^* \rho_t)}{1+\exp(\Delta^*\rho_t)}.$$

By a change of variables from ρ_t to m_t, we can write $V^*_\infty := \inf_\tau \mathbb{E}[\varpi(m_t) + c\tau]$, where $dm_t = \Delta^* m_t(1-m_t)dW_t$ by Ito's lemma. But by way of the proof of Lemma 1, see (A.3), this is just V^*. The theorem can therefore be considered proved.

A.4. **Proof of Theorem 3.** For any $\boldsymbol{h} = (h_1, h_0)$, let $P_{n,\boldsymbol{h}}$ denote the joint distribution with density $p^{(1)}_{nT,\theta_0+h_1/\sqrt{n}}(\mathbf{y}^{(1)}_{nT}) \cdot p^{(0)}_{nT,\theta_0+h_0/\sqrt{n}}(\mathbf{y}^{(0)}_{nT})$. Take $\mathbb{E}_{n,\boldsymbol{h}}[\cdot]$ to be the corresponding expectation. We can write $V_n(\boldsymbol{d}_{n,T}, \boldsymbol{h})$ as

$$V_n(\boldsymbol{d}_{n,T}, \boldsymbol{h}) = \mathbb{E}_{n,\boldsymbol{h}}\left[\sqrt{n}\left(\mu_{n,1}(h_1) - \mu_{n,0}(h_0)\right)\mathbb{I}\{\delta_{n,T} \geq 0\} + c\tau_{n,T}\right].$$

Define $\mu(\boldsymbol{h}) = (\dot{\mu}_1^\intercal h_1, \dot{\mu}_0^\intercal h_0)$, $\Delta\mu(\boldsymbol{h}) = \dot{\mu}_1^\intercal h_1 - \dot{\mu}_0^\intercal h_0$ and $\Delta_n\mu(\boldsymbol{h}) = \mu_{n,1}(h_1) - \mu_{n,0}(h_0)$. In addition, we also define $\tilde{q}_a(t) := \sigma_a t/(\sigma_1+\sigma_0)$.

Step 1 (Weak convergence of $\rho_n(t)$). Denote $P_{n,0} = P_{n,(0,0)}$. By the SLAN property (A.10), independence of $\mathbf{y}^{(1)}_{nT}, \mathbf{y}^{(0)}_{n,T}$ given $\boldsymbol{h}$, and the central limit theorem,

$$\ln \frac{dP_{n,\boldsymbol{h}}}{dP_{n,0}}\left(\mathbf{y}^{(1)}_{nT}, \mathbf{y}^{(0)}_{n,T}\right) = \sum_{a\in\{0,1\}} \left\{h_a^\intercal I_a^{1/2} z_{a,nT} - \frac{T}{2} h_a^\intercal I_a h_a\right\} + o_{Pn,0}(1) \tag{A.22}$$

$$\xrightarrow[P_{n,0}]{d} \mathcal{N}\left(\frac{-T}{2}\sum_{a\in\{0,1\}} h_a^\intercal I_a h_a, T\sum_{a\in\{0,1\}} h_a^\intercal I_a h_a\right). \tag{A.23}$$

Therefore, by Le Cam's first lemma, $P_{n,\boldsymbol{h}}$ and $P_{n,0}$ are mutually contiguous.

We now determine the distribution of $\rho_n(t)$. We start by showing

$$\left|\frac{\dot{\mu}_a^\intercal I_a^{-1}}{\sigma_a\sqrt{n}} \sum_{i=1}^{\lfloor nq_a(t)\rfloor} \psi_a(Y_i^{(a)}) - \frac{\dot{\mu}_a^\intercal I_a^{-1}}{\sigma_a\sqrt{n}} \sum_{i=1}^{\lfloor n\tilde{q}_a(t)\rfloor} \psi_a(Y_i^{(a)})\right| = o_{P_{n,0}}(1), \tag{A.24}$$

uniformly over $t \le T$. Choose any $b \in (1/2, 1)$. For $t \le n^{-b}$, we must have $q_a(t), \tilde{q}_a(t) \le n^{-b}$, so (A.24) follows from Assumption 1(ii), which implies

$$\sup_{1 \le i \le nT} |\psi_a(Y_i^{(a)})| = O_{P_{n,0}}(n^{1/r}), \text{ for any } r > 0. \tag{A.25}$$

As for the other values of t, by (4.5) and (A.25),

$$\frac{\dot{\mu}_a^\intercal I_a^{-1}}{\sigma_a \sqrt{n}} \left\{ \sum_{i=1}^{\lfloor nq_a(t) \rfloor} \psi_a(Y_i^{(a)}) - \sum_{i=1}^{\lfloor n\tilde{q}_a(t) \rfloor} \psi_a(Y_i^{(a)}) \right\} \lesssim \sqrt{n}\, |q_a(t) - \tilde{q}_a(t)| \sup_{1 \le i \le nT} |\psi_a(Y_i^{(a)})| = o_{P_{n,0}}(1),$$

uniformly over $t \in (n^{-b}, T]$.

Now, (A.24) implies

$$\rho_n(t) = \frac{\dot{\mu}_1^\intercal I_1^{-1}}{\sigma_1 \sqrt{n}} \sum_{i=1}^{\lfloor n\tilde{q}_1(t) \rfloor} \psi_1(Y_i^{(1)}) - \frac{\dot{\mu}_0^\intercal I_0^{-1}}{\sigma_0 \sqrt{n}} \sum_{i=1}^{\lfloor n\tilde{q}_0(t) \rfloor} \psi_0(Y_i^{(0)}) + o_{P_{n,0}}(1) \text{ uniformly over } t \le T. \tag{A.26}$$

By Donsker's theorem, and recalling that $\tilde{q}_a(t) = \sigma_a t/(\sigma_1 + \sigma_0)$,

$$\frac{\dot{\mu}_a^\intercal I_a^{-1}}{\sigma_a \sqrt{n}} \sum_{i=1}^{\lfloor n\tilde{q}_a(\cdot) \rfloor} \psi_a(Y_i^{(a)}) \xrightarrow[P_{n,0}]{d} \sqrt{\frac{\sigma_a}{\sigma_1 + \sigma_0}} W_a(\cdot),$$

where $W_1(\cdot), W_0(\cdot)$ can be taken to be independent Wiener processes due to the independence of $\mathbf{y}_{nT}^{(1)}, \mathbf{y}_{n,T}^{(0)}$ under $P_{n,0}$. Combined with (A.26), we conclude

$$\rho_n(\cdot) \xrightarrow[P_{n,0}]{d} \tilde{W}(\cdot), \tag{A.27}$$

where $\tilde{W}(\cdot) = \sqrt{\frac{\sigma_1}{\sigma_1+\sigma_0}} W_1(\cdot) - \sqrt{\frac{\sigma_0}{\sigma_1+\sigma_0}} W_0(\cdot)$ is another Wiener process.

Let Z denote the normal random variable in (A.23). Equations (A.23) and (A.27) imply that $\rho_n(\cdot), \ln(dP_{n,\boldsymbol{h}}/dP_{n,0})$ are asymptotically tight, and therefore, the joint $(\rho_n(\cdot), \ln(dP_{n,\boldsymbol{h}}/dP_{n,0}))$ is also asymptotically tight under $P_{n,0}$. Furthermore, for any $t \in [0, T]$, it can be shown using (A.26) and (A.22) that

$$\begin{pmatrix} \rho_n(t) \\ \ln \frac{dP_{n,\boldsymbol{h}}}{dP_{n,0}} \end{pmatrix} \xrightarrow[P_{n,0}]{d} \begin{pmatrix} \tilde{W}(t) \\ Z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \frac{-T}{2} \sum_a h_a^\intercal I_a h_a \end{pmatrix}, \begin{bmatrix} t & \frac{\Delta\mu(\boldsymbol{h})}{\sigma_1+\sigma_0} t \\ \frac{\Delta\mu(\boldsymbol{h})}{\sigma_1+\sigma_0} t & T \sum_a h_a^\intercal I_a h_a \end{bmatrix} \right).$$

Based on the above, an application of Le Cam's third lemma as in van der Vaart and Wellner (1996, Theorem 3.10.12) then gives

$$\rho_n(\cdot) \xrightarrow[P_{n,\boldsymbol{h}}]{d} \rho(\cdot) \quad \text{where} \quad \rho(t) := \frac{\Delta\mu(\boldsymbol{h})}{\sigma_1 + \sigma_0} t + \tilde{W}(t). \tag{A.28}$$

Step 2 (Weak convergence of $\delta_{n,T}, \tau_{n,T}$). Let $\mathbb{D}[0,T]$ denote the metric space of all functions from $[0,T]$ to $\mathbb{R}$ equipped with the sup norm. For any element $z(\cdot) \in \mathbb{D}[0,T]$, define $\tau_T(z) = T \wedge \inf\{t : |z(t)| \geq \gamma\}$ and $\delta_T(z) = \mathbb{I}\{z(\tau_T(z)) > 0\}$.

Now, under $\boldsymbol{h} = (0,0)$, $\rho(\cdot)$ is the Wiener process, whose sample paths take values (with probability 1) in $\bar{\mathbb{C}}[0,T]$, the set of all continuous functions such that $\gamma, -\gamma$ are regular points (i.e., if $z(t) = \gamma$, $z(\cdot) - \gamma$ changes sign infinitely often in any time interval $[t, t+\epsilon]$, $\epsilon > 0$; a similar property holds under $z(t) = -\gamma$). The latter is a well known property of Brownian motion, see Karatzas and Shreve (2012, Problem 2.7.18), and it implies $z(\cdot) \in \bar{\mathbb{C}}[0,T]$ must 'cross' the boundary within an arbitrarily small time interval after hitting γ or $-\gamma$. It is then easy to verify that if $z_n \to z$ with $z_n \in \mathbb{D}[0,T]$ for all n and $z \in \bar{\mathbb{C}}[0,T]$, then $\tau_T(z_n) \to \tau_T(z)$ and $\delta_T(z_n) \to \delta_T(z)$. By construction, $\tau_{n,T} = \tau_T(\rho_n)$ and $\delta_{n,T} = \delta_T(\rho_n)$, so by (A.27) and the extended continuous mapping theorem (van der Vaart and Wellner, 1996, Theorem 1.11.1)

$$(\tau_{n,T}, \delta_{n,T}) \xrightarrow[P_{n,0}]{d} (\tau_T^*, \delta_T^*),$$

where $\tau_T^* := \tau_T(\rho)$ and $\delta_T^* := \delta_T(\rho)$.

For general $\boldsymbol{h}$, $\rho(\cdot)$ is distributed as in (A.28). By the Girsanov theorem, the probability law induced on $\mathbb{D}[0,T]$ by the process $\frac{\Delta\mu(\boldsymbol{h})}{\sigma_1+\sigma_0}t + \tilde{W}(t)$ is absolutely continuous with respect to the probability law induced by $\tilde{W}(t)$. Hence, with probability 1, the sample paths of $\rho(\cdot)$ again lie in $\bar{\mathbb{C}}[0,T]$. Then, by similar arguments as in the case with $\boldsymbol{h} = (0,0)$, but now using (A.28), we conclude

$$(\tau_{n,T}, \delta_{n,T}) \xrightarrow[P_{n,\boldsymbol{h}}]{d} (\tau_T^*, \delta_T^*). \tag{A.29}$$

Step 3 (Convergence of $V_n(\boldsymbol{d}_{n,T}, \boldsymbol{h})$). From (3.8) and the discussion in Section 3.1, it is clear that the distribution of $\rho(t)$ is the same as that of $\sigma_1^{-1}x_1(t) - \sigma_0^{-1}x_0(t)$ in the diffusion regime. Thus, the joint distribution, $\mathbb{P}$, of (τ_T^*, δ_T^*), defined in Step 2, is the same as the joint distribution of

$$\left(\tau_T^* := \tau^* \wedge T, \delta_T^* := \mathbb{I}\left\{\frac{x_1(\tau^* \wedge T)}{\sigma_1} - \frac{x_0(\tau^* \wedge T)}{\sigma_0} \geq 0\right\}\right)$$

in the diffusion regime, when the optimal sampling rule π^* is used. Therefore, defining $\boldsymbol{d}_T^* \equiv (\pi^*, \tau_T^*, \delta_T^*)$ and $\mathbb{E}[\cdot]$ to be the expectation under $\mathbb{P}$, we obtain

$$V(\boldsymbol{d}_T^*, \mu(\boldsymbol{h})) = \mathbb{E}\left[\Delta\mu(\boldsymbol{h})\delta_T^* + c\tau_T^*\right],$$

where $V(\boldsymbol{d}, \boldsymbol{\mu})$ denotes the frequentist regret of $\boldsymbol{d}$ in the diffusion regime. Now, recall that by the definitions stated early on in this proof,

$$V_n(\boldsymbol{d}_{n,T}, \boldsymbol{h}) = \mathbb{E}_{n,\boldsymbol{h}}\left[\sqrt{n}\Delta_n \mu(\boldsymbol{h})\delta_{n,T} + c\tau_{n,T}\right].$$

Since δ_n, τ_n are bounded and $\sqrt{n}\Delta_n\mu(\boldsymbol{h}) \to \Delta\mu(\boldsymbol{h})$ by Assumption 1(iii), it follows from (A.29) that for each $\boldsymbol{h}$,

$$\lim_{n\to\infty} V_n(\boldsymbol{d}_{n,T}, \boldsymbol{h}) = V(\boldsymbol{d}_T^*, \mu(\boldsymbol{h})). \tag{A.30}$$

For any given $\boldsymbol{h}$ and $\epsilon > 0$, a dominated convergence argument as in Step 4 of the proof of Theorem 2 shows that there exists $\bar{T}_{\boldsymbol{h}}$ large enough such that

$$V(\boldsymbol{d}_T^*, \mu(\boldsymbol{h})) \leq V(\boldsymbol{d}^*, \mu(\boldsymbol{h})) + \epsilon \tag{A.31}$$

for all $T \geq \bar{T}_h$. Fix a finite subset $\mathcal{J}$ of $\mathbb{R}$ and define $\bar{T}_{\mathcal{J}} = \sup_{\boldsymbol{h}\in\mathcal{J}} T_{\boldsymbol{h}}$. Then, (A.30) and (A.31) imply

$$\liminf_{n\to\infty} \sup_{\boldsymbol{h}\in\mathcal{J}} V_n(\boldsymbol{d}_{n,T}, \boldsymbol{h}) \leq \sup_{\boldsymbol{h}\in\mathcal{J}} V(\boldsymbol{d}_T^*, \mu(\boldsymbol{h})) \leq \sup_{\boldsymbol{h}\in\mathcal{J}} V(\boldsymbol{d}^*, \mu(\boldsymbol{h})) + \epsilon,$$

for all $T \geq \bar{T}_{\mathcal{J}}$. Since the above is true for any $\mathcal{J}$ and $\epsilon > 0$,

$$\begin{aligned}\sup_{\mathcal{J}} \lim_{T\to\infty} \liminf_{n\to\infty} \sup_{\boldsymbol{h}\in\mathcal{J}} V_n(\boldsymbol{d}_{n,T}, \boldsymbol{h}) &\leq \sup_{\mathcal{J}} \sup_{\boldsymbol{h}\in\mathcal{J}} V(\boldsymbol{d}^*, \mu(\boldsymbol{h})) \\ &\leq \sup_{\boldsymbol{\mu}} V(\boldsymbol{d}^*, \boldsymbol{\mu}) = V^*.\end{aligned}$$

The inequality can be made an equality due to Theorem 2. We have thereby proved Theorem 3.