NEYMAN ALLOCATION IS MINIMAX OPTIMAL FOR BEST ARM IDENTIFICATION WITH TWO ARMS

KARUN ADUSUMILLI†

Abstract. This note describes the optimal policy rule, according to the local asymptotic minimax regret criterion, for best arm identification when there are only two treatments. It is shown that the optimal sampling rule is the Neyman allocation, which allocates a constant fraction of units to each treatment in a manner that is proportional to the standard deviation of the treatment outcomes. When the variances are equal, the optimal ratio is one-half. This policy is independent of the data, so there is no adaptation to previous outcomes. At the end of the experiment, the policy maker adopts the treatment with higher average outcomes.

This version: August 29, 2022

I would like to thank Tim Armstrong, Pepe Montiel-Olea and Chao Qin for helpful comments.

†Department of Economics, University of Pennsylvania.
1. Introduction

The goal of running an experiment is often to determine the best possible treatment out of a set of candidate treatments. Suppose an agent is allowed \( n \) periods of experimentation. The agent can adaptively choose which treatment to sample in each period depending on all the information from the previous periods. This is the Best Arm Identification (BAI) problem with a fixed budget. The aim is to describe the optimal sampling and allocation rules for maximizing welfare in the implementation phase following experimentation. However, in-sample outcomes are not included in the welfare calculation, thus differentiating this setup from the multi-armed bandit problem.

Early on this setting was done by Mannor and Tsitsiklis (2004), Chen et al. (2000) and Glynn and Juneja (2004). More recently, there has been a surge of interest in this problem as evidenced by the important work of Garivier and Kaufmann (2016), Russo (2016), Carpentier and Locatelli (2016), Qin et al. (2017), Kasy and Sautmann (2019), among others. The seminal analysis of Russo (2016) characterized the optimal rate of posterior convergence to the best arm in the fixed reward gap regime (i.e., when the mean difference in outcomes between the arms is held fixed, i.e., unchanged over \( n \)). Carpentier and Locatelli (2016) suggest a lower bound on the probability of mis-classification (also under a fixed gap), and a tight bound for this was later derived by Kato et al. (2022) in the small gap regime. However these results do not describe which algorithms to pick under statistical measures of risk such as minimax regret.\(^1\)

This paper focuses on the special case of the BAI problem with just two arms. In this setting, we characterize the optimal policy according to the local asymptotic minimax regret criterion. The local asymptotic regime, also known as the diffusion regime, reduces the problem to the question of choosing the best treatment when the outcomes from each treatment correspond to a Gaussian process. It is shown that the optimal sampling rule is the Neyman allocation. It allocates a constant fraction of units to one of the arms in a manner that is proportional to the treatment standard deviations, with more variable treatments being sampled

\(^1\)Recall that regret is defined as the difference in welfare from employing the treatment chosen by the algorithm as opposed to employing the best treatment.
more often. When the treatment variances are equal, the optimal sampling ratio is one-half. Somewhat surprisingly, the sampling rule is independent of the data, so there is no adaptation to previous outcomes. At the end of the experiment, the agent chooses the treatment with the higher average outcomes.

The Neyman allocation is well known in the design of experiments literature as the allocation rule that minimizes estimation variance of the treatment effect. Despite the difference in goals between estimation and best arm identification, our results show that the optimal sampling rule remains unchanged. Thus, as a practical matter, there is no benefit to running an adaptive experiment (as opposed to a standard RCT) when there are only two treatments, and the goal is to minimize maximum regret.

We emphasize, though, that the above results are intimately connected to the minimax regret criterion. It is far from obvious that minimax optimal rule should be non-adaptive. In the Bayesian formulation of the problem, with normal priors, the optimal sampling rule changes with time, albeit in a deterministic fashion (??). With other priors, the sampling rule could be very different. Indeed, our minimax rule is itself supported by a specific two-point least-favorable prior. The results thus highlight the important role played by the prior, and the sensitivity of the optimal decision rules to it.

Our results are asymptotic in nature, i.e., our policies minimize maximum regret in the limit as \( n \to \infty \). If one instead considers the global minimax regret criterion (i.e., minimax regret for fixed \( n \)), it is known that the sampling ratio \( \gamma = 1/2 \) is optimal under bounded outcomes; see, Stoye (2012). In the global minimax regime, nature chooses the distribution of outcomes for each treatment, not just the means (so it can choose the variances as well). Here, it is easy to see that the minimax regret under unbounded rewards is infinity. The results here are instead derived under local asymptotics. In the large \( n \) limit, previous work by the author, Adusumilli (2021), showed that we can treat the outcome distributions as effectively Gaussian. Furthermore, in this regime, it is without loss of generality to assume the treatment variances to be known, as replacing the unknown variances with consistent estimates does not affect asymptotic regret. The optimal policies under the local asymptotic and global minimax criteria generally do not
coincide, see, e.g., the discussion in Stoye (2012) and Hirano and Porter (2009). In the present setting, they coincide if the treatment variances are equal, but not otherwise.

2. Setup in the diffusion regime

We start by describing the continuous time version of the problem. There are two treatments 0, 1 corresponding to mean rewards \( \mu_1, \mu_0 \) and reward variances \( \sigma_1, \sigma_0 \). To begin with, we assume \( \sigma_1, \sigma_0 \) are known. The experiment runs until time \( t = 1 \). At each instant, the agent can choose to direct attention to one of the treatments by choosing the sampling rule \( \pi_1(\cdot) \). Here \( \pi_1(\cdot) \) denotes the probability that treatment 1 is sampled, and we also define \( \pi_0 = 1 - \pi_1 \). At the end of the experiment, the agent selects a treatment for full-scale implementation, according to \( \pi_{fs}(\cdot) \in \{0, 1\} \).

Let \( x_1(t), x_0(t) \) denote the cumulative outcomes from both treatments, and \( q_1(t), q_0(t) \) the number of times each treatment has been sampled so far. Under the policy \( \pi_1(\cdot) \), these state variables evolve as diffusion processes (for \( a \in \{0, 1\} \)):

\[
\begin{align*}
dx_a(t) &= \mu_a \pi_a dt + \sigma_a \sqrt{\pi_a} dW_a(t) \\
dq_a(t) &= \pi_a dt.
\end{align*}
\]

Here \( W_1(t), W_0(t) \) are Weiner processes tracking the variability of rewards for each treatment. Since we only sample one treatment at any given instant, it is without loss of generality to assume \( W_1(\cdot), W_0(\cdot) \) are independent.

Denote \( s(t) = (x_1(t), x_0(t), q_1(t), q_0(t)) \). We require \( \pi_1(\cdot) \) to be adapted to the filtration \( \{\mathcal{F}_t\}_{t \in \mathbb{R}^+} \), where \( \mathcal{F}_t = \sigma\{s(u); u \leq t\} \) is the \( \sigma \)-algebra generated by the state variables \( s(\cdot) \) until time \( t \). Furthermore, \( \pi_{fs} \) needs to be \( \mathcal{F}_{t-1} \) measurable. The aim is choose \( \pi_1, \pi_{fs} \) optimally in order to minimize maximum regret

\[
R_{max}(\pi_1, \pi_{fs}) = \max_{(\mu_1, \mu_0)} R \left( \pi_1, \pi_{fs}, (\mu_1, \mu_0) \right),
\]

where

\[
R \left( \pi_1, \pi_{fs}, (\mu_1, \mu_0) \right) = \mathbb{E}_{\pi_1|\mu_1, \mu_0} \left[ \max\{\mu_1 - \mu_0, 0\} - (\mu_1 - \mu_0)\pi_{fs} \right],
\]

where \( \mathbb{E}_{\pi_1|\mu_1, \mu_0}[\cdot] \) denotes the expectation under the sampling rule \( \pi_1 \) given \( \mu_1, \mu_0 \).
3. Minimax Optimal Policy in the Diffusion Regime

Following Wald (1945), we model the minimax problem as a game played between nature and the agent. Nature chooses a prior $m_0$ over the vector of mean rewards $(\mu_1, \mu_0)$ and the agent chooses the policy functions $(\pi_1, \pi_{fs})$. The equilibrium value of this game gives minimax regret. The optimal choice of the agent is termed the minimax policy, and that of nature, the least favorable prior.

In general, solving the two-player game can be daunting. However, a useful starting point is the conjecture, made in Adusumilli (2021), that the least favorable has the same number of support points as the number of treatments, which is 2 in this setting (we will formally verify this conjecture here).

Following the above ansatz, consider a prior supported on the two points $(a_1, b_1), (a_0, b_0)$, where $a_1 > b_1$ and $b_0 > a_0$. Let $\theta = 1$ denote the state when nature chooses $(a_1, b_1)$, and $\theta = 0$ the state when nature chooses $(a_0, b_0)$. Also let $(\Omega, \mathbb{P}, \mathcal{F}_t)$ denote the relevant probability space, where $\mathcal{F}_t$ is defined above. Define the probability measures $P_0, P_1$ as $P_0 := \mathbb{P}(A | \theta = 0)$ and $P_1 := \mathbb{P}(A | \theta = 1)$ for any $A \in \mathcal{F}_t$.

Noting that $W_1(\cdot), W_0(\cdot)$ are independent of each other, it follows by the Girsanov theorem (see also Shiryaev, 2007, Section 4.2.1) that

$$\ln \frac{dP_1}{dP_0}(\mathcal{F}_t) = \frac{(a_1 - a_0)}{\sigma_1^2} x_1(t) + \frac{(b_1 - b_0)}{\sigma_0^2} x_0(t) - \frac{(a_1^2 - a_0^2)}{2\sigma_1^2} q_1(t) - \frac{(b_1^2 - b_0^2)}{2\sigma_0^2} q_0(t).$$

(3.1)

Let $m_1$ denote the prior probability that $\theta = 1$. Furthermore, let $m^\pi_1 = \mathbb{P}(\theta = 1 | \mathcal{F}_t)$ denote the posterior probability that $\theta = 1$. Following Shiryaev (2007, Section 4.2.1), the belief process $m^\pi_1$ can be related to the likelihood ratio process $\varphi(t) := \frac{dP_1}{dP_0}(\mathcal{F}_t)$ as

$$m^\pi_1 = \frac{m_1 \varphi(t)}{(1 - m_1) + m_1 \varphi(t)}.$$  

The optimal allocation rule at the end of the experiment is then

$$\pi_{FS} = I \{ a_1 m^\pi_1 + a_0 (1 - m^\pi_1) \geq b_1 m^\pi_1 + b_0 (1 - m^\pi_1) \}$$

$$= I \left\{ \ln \varphi(t) \geq \ln \frac{(b_0 - a_0)(1 - m_1)}{(a_1 - b_1)m_1} \right\}$$

(3.2)

We can obtain the Nash equilibrium of this game under the following steps:
1. **Indifference priors.** Consider two point priors supported on 
\((\sigma_1 \Delta/2, -\sigma_0 \Delta/2), (-\sigma_1 \Delta/2, \sigma_0 \Delta/2)\) for some \(\Delta > 0\). For these priors, the agent is indifferent between the choice of any measurable \(\pi_1\). Intuitively, either treatment is equally informative about the true value of \(\theta \in \{0, 1\}\) here, so it does not matter which treatment the agent directs attention to.

To see this formally, observe that under these priors, \((3.1)\) implies
\[
\ln \varphi(t) := \ln \frac{dP^1}{dP^0}(\mathcal{F}_t) = \Delta \left\{ \frac{x_1(t)}{\sigma_1} - \frac{x_0(t)}{\sigma_0} \right\}.
\]
Suppose \(\theta = 1\). Then the evolution equations \((2.1), (2.2)\) imply
\[
\frac{dx_1(t)}{\sigma_1} - \frac{dx_0(t)}{\sigma_0} = \frac{\Delta}{2} dt + \sqrt{\pi_1} dW_1(t) - \sqrt{\pi_0} dW_0(t)
= \frac{\Delta}{2} dt + d\tilde{W}(t),
\]
where \(\tilde{W}(t) := \sqrt{\pi_1} dW_1(t) - \sqrt{\pi_0} dW_0(t)\) is a one dimensional Weiner process, being a linear combination of two independent Weiner processes with \(\pi_1 + \pi_0 = 1\).

Plugging the above into \((3.3)\) gives
\[
d\ln \varphi(t) = \frac{\Delta^2}{2} dt + \Delta d\tilde{W}(t).
\]
In a similar manner, we can show under \(\theta = 0\) that
\[
d\ln \varphi(t) = -\frac{\Delta^2}{2} dt + \Delta d\tilde{W}(t).
\]
In either case, the choice of \(\pi_1\) does not affect the evolution of the likelihood-ratio process \(\varphi(t)\), and consequently has no bearing on the evolution of the beliefs \(m_t^\pi\).

2. **Form of \(\pi_{fs}\) under the indifference prior.** Suppose nature chooses the indifference prior for some \(\Delta > 0\). Then \((3.2)\) and \((3.3)\) imply
\[
\pi_{fs} = \mathbb{I} \left\{ \Delta \left\{ \frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \right\} \geq \ln \left( \frac{1 - m_1}{m_1} \right) \right\}
= \mathbb{I} \left\{ \frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \geq c \right\},
\]
where \(c := \Delta^{-1} \ln \left( \frac{1-m_1}{m_1} \right)\). Note that for any given \(\Delta\) we can induce \(c\) to be any value we like by varying \(m_1\). This implies that the choice of \(c\) is essentially equivalent to the choice of \(m_1\).
3. Determining the equilibrium values of $\pi_1, c$. Based on the above observations, consider the ansatz that the minimax policy of the agent is of the form $\pi_1 = \gamma$ for some $\gamma \in [0, \infty)$ and

$$\tilde{\pi}_{fs} = \mathbb{I} \left\{ \frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \geq c \right\}.$$ 

Suppose nature chooses $(\tilde{a}_1, \tilde{b}_1)$, where $\tilde{a}_1 > \tilde{b}_1$. Under the ansatz for the minimax policy,

$$\frac{dx_1(t)}{\sigma_1} - \frac{dx_0(t)}{\sigma_0} = \left( \tilde{a}_1 \frac{\gamma}{\sigma_1} - \tilde{b}_1 \frac{1 - \gamma}{\sigma_0} \right) dt + d\tilde{W}(t),$$

so the expected regret under a given sampling probability $\gamma$, and threshold constant $c$ is

$$R\left(\gamma, c, (\tilde{a}_1, \tilde{b}_1)\right) = (\tilde{a}_1 - \tilde{b}_1) \mathbb{P} \left( \frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \leq c \right)$$

$$= (\tilde{a}_1 - \tilde{b}_1) \mathbb{P} \left( \left\{ \tilde{a}_1 \frac{\gamma}{\sigma_1} - \tilde{b}_1 \frac{1 - \gamma}{\sigma_0} \right\} + \tilde{W}(1) \leq c \right)$$

$$= (\tilde{a}_1 - \tilde{b}_1) \Phi \left( c - \left\{ \tilde{a}_1 \frac{\gamma}{\sigma_1} - \tilde{b}_1 \frac{1 - \gamma}{\sigma_0} \right\} \right). \quad (3.4)$$

If $\gamma/\sigma_1 > (1 - \gamma)/\sigma_0$, nature can choose $\tilde{a}_1, \tilde{b}_1$ in such a way that the expected regret above is arbitrarily large (nature can set $\tilde{a}_1 - \tilde{b}_1$ to be arbitrarily large and $\tilde{a}_1$ to be negative with $|\tilde{a}_1| \gg \tilde{a}_1 - \tilde{b}_1$; the latter ensures the term inside the $\Phi(\cdot)$ function in (3.4) is close to $\infty$). This suggests that the optimal sampling rule should satisfy $\gamma^* \leq \sigma_1/(\sigma_1 + \sigma_0)$.

Similarly, if nature chooses $(\tilde{a}_0, \tilde{b}_0)$ where $\tilde{b}_0 > \tilde{a}_0$, the expected regret is

$$R\left(\gamma, c, (\tilde{a}_0, \tilde{b}_0)\right) = (\tilde{b}_0 - \tilde{a}_0) \Phi \left( -c - \left\{ \tilde{b}_0 \frac{1 - \gamma}{\sigma_0} - \tilde{a}_0 \frac{\gamma}{\sigma_1} \right\} \right). \quad (3.5)$$

Now if $\gamma/\sigma_1 < (1 - \gamma)/\sigma_0$, nature can again achieve infinite regret in (3.5). Combined with the previous observation that $\gamma^* \leq \sigma_1/(\sigma_1 + \sigma_0)$, we are led to conclude only $\gamma^* = \sigma_1/(\sigma_1 + \sigma_0)$ can prevent infinite regret in both scenarios. So setting $\gamma$ to this value and combining (3.4) - (3.5), we find the max regret under $(\gamma^*, c)$ to be

$$R^{\max}(\gamma^*, c) = \max \left\{ \max_\delta \delta \Phi \left( c - \frac{\delta}{\sigma_1 + \sigma_0} \right), \max_\delta \delta \Phi \left( -c - \frac{\delta}{\sigma_1 + \sigma_0} \right) \right\}. \quad (3.6)$$

Clearly, $R^{\max}(\gamma^*, c)$ is minimized when $c^* = 0$. 

7
We thus find that the optimal choices of \( \gamma, c \) for the agent are \( \sigma_1/(\sigma_1 + \sigma_0), 0 \). It remains to show that this constitutes an equilibrium. To this end, suppose that the agent chooses \( \gamma^* = \sigma_1/(\sigma_1 + \sigma_0), c^* = 0 \). Then it follows from (3.4) - (3.6) that the optimal response of nature is to choose \((\tilde{a}, \tilde{b})\) such that
\[
|\tilde{a} - \tilde{b}| = \eta^* := \arg \max_\delta \delta \Phi \left( -\frac{\delta}{\sigma_1 + \sigma_0} \right) = (\sigma_1 + \sigma_0) \arg \max_\delta \delta \Phi (-\delta).
\]
Nature is otherwise indifferent between any \((\tilde{a}, \tilde{b})\) satisfying the above condition. In particular, a prior supported on the two points \((\sigma_1 \Delta^*/2, -\sigma_0 \Delta^*/2), (-\sigma_1 \Delta^*/2, \sigma_0 \Delta^*/2)\), where \(\Delta^* := 2\eta^*/(\sigma_1 + \sigma_0) = 2\arg \max_\delta \delta \Phi (-\delta)\), would be a best response to the agent’s actions. This is an indifference prior, so the agent’s actions are best responses to it as well: the agent is indifferent between any \(\pi_a\) as noted earlier, and we can induce \(c^* = 0\) by setting \(m_1 = 1/2\). We have thus obtained a Nash equilibrium to the game. The result is summarized below:

**Theorem 1.** The minimax optimal decision rule is \(d^* := (\pi_1^*, \pi_{fs}^*)\), where \(\pi_a^* = \sigma_a/(\sigma_1 + \sigma_0)\) for \(a \in \{0, 1\}\), and \(\pi_{fs}^* = I\left\{ \frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \geq 0 \right\}\). Furthermore, the least favorable prior is a symmetric two-point distribution supported on \((\sigma_1 \Delta^*/2, -\sigma_0 \Delta^*/2)\) and \((-\sigma_1 \Delta^*/2, \sigma_0 \Delta^*/2)\) where \(\Delta^* = 2\arg \max_\delta \delta \Phi (-\delta)\).

**Discussion of the minimax policy.** Perhaps the most striking feature of the minimax optimal policy is that it is independent of the data. The policy assigns a fixed proportion \(\gamma = \sigma_1/(\sigma_1 + \sigma_0)\) of units to treatment 1. This is just the Neyman allocation. It is same sampling rule one would also employ if the aim were to minimize the estimation variance of the treatment effect \(\mu_1 - \mu_0\).

The implementation rule at the end of the experiment is given by
\[
\pi_{fs}^* = I\left\{ \frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \geq 0 \right\}.
\]
Under the optimal sampling rule, treatment 1 is sampled \(\sigma_1/\sigma_0\) times more often than treatment 0. This implies \(\frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0}\) is proportional to the difference in average outcomes between both treatments. In other words, the treatment with higher average outcomes is chosen for full-scale implementation.
4. LOCAL ASYMPTOTICS

So far the optimal policies have been characterized under the diffusion regime. Let \( V^* \) denote the value of maximum regret from the previous section

\[
V^* := \max_\delta \delta \Phi \left( -\frac{\delta}{\sigma_1 + \sigma_0} \right) = (\sigma_1 + \sigma_0) \max_\delta \delta \Phi (-\delta).
\]

Then \( V^* \) is also the lower bound on asymptotic minimax regret in both parametric and non-parametric regimes. In fact the bound is tight in the sense that it can be achieved through analogues of the minimax optimal policy described above. Formal statements follow:

4.1. Parametric regime. Let \( \{ (P^{(1)}_{\theta^{(1)}}, P^{(0)}_{\theta^{(0)})} : \theta^{(1)}, \theta^{(0)} \in \mathbb{R} \} \) denote the set of candidate probability measures for the joint distribution of outcomes under both treatments. It is without loss of generality to assume \( P^{(1)}_{\theta^{(1)}}, P^{(0)}_{\theta^{(0)}} \) are mutually independent (conditional on \( \theta^{(1)}, \theta^{(0)} \)) as we only ever observe the outcomes from one treatment anyway.

The mean outcomes under a parameter \( \theta \) are given by \( \mu_a(\theta) := \mathbb{E}_{P^{(a)}_\theta}[Y_i] \). Following Hirano and Porter (2009) and Adusumilli (2021), for \( a \in \{0, 1\} \), we consider local perturbations of the form \( \{ \theta^{(a)}_0 + h^{(a)}/\sqrt{n}; h^{(a)} \in \mathbb{R} \} \) around a reference parameter \( \theta^{(a)}_0 \). As in those papers, \( \theta^{(a)}_0 \) is chosen such that \( \mu_a(\theta^{(a)}_0) = 0 \) for each \( a \in \{0, 1\} \). This defines the hardest instance of the best arm identification problem, with \( \mu_a(\theta^{(a)}_0 + h/\sqrt{n}) \approx \nabla_{\theta} \mu_a(\theta_0) \). Denote \( P^{(a)}_h := P^{(a)}_{\theta^{(a)}_0 + h/\sqrt{n}} \) and let \( \mathbb{E}^{(a)}_h[\cdot] \) denote its corresponding expectation. We assume \( P^{(a)}_h \) is differentiable in quadratic mean around \( \theta^{(a)}_0 \) with score functions \( \psi_a(Y_i) \) and information matrices \( I_a := \left( \mathbb{E}^{(a)}_0[\psi^2] \right)^{-1} \). Let \( \hat{\mu}_{n,a} \) denote the best regular estimator of the mean outcomes for treatment \( a \) in the sense

\[
\sqrt{n} \left( \hat{\mu}_{n,a} - \mu_a(h) \right) \xrightarrow{d}_{P^{(a)}_h} \mathcal{N}(0, \hat{\mu}_a I_a^{-1} \hat{\mu}_a). \tag{4.1}
\]

In what follows, we define \( \sigma_a^2 := \hat{\mu}_a I_a^{-1} \hat{\mu}_a. \)

Suppose \( V_n(\pi_1, \pi^{fa}; h_1, h_0) \) denotes the frequentist regret due to the sampling rule \( \pi_1 \) and implementation rule \( \pi^{fa} \) when the local parameters are \( \theta^{(1)}_0 + h_1/\sqrt{n} \) and \( \theta^{(0)}_0 + h_0/\sqrt{n} \). Define \( n_a := n \sigma_a/(\sigma_1 + \sigma_0); a \in \{0, 1\} \). Our key result in this
section is that the policy rules \( \pi_1^*, \pi^{fs} \), given below are minimax optimal:

\[
\pi_1^* = \frac{\sigma_1}{\sigma_1 + \sigma_0}; \quad \pi^{fs} = \mathbb{I} \{ \hat{\mu}_{n_{1,1}} - \hat{\mu}_{n_{0,0}} \geq 0 \}.
\]

This is shown under the following assumptions.

**Assumption 1.**

(i) The class \( \{ P_{\theta}^{(a)} ; \theta \in \mathbb{R} \} \) is differentiable in quadratic mean around \( \theta_0^{(a)} \) for \( a \in \{0, 1\} \).

(ii) \( \mathbb{E}_0^{(a)}[\exp |\psi_a(Y)|] < \infty \) for \( a \in \{0, 1\} \).

(iii) For each \( a \in \{0, 1\} \) there exists \( |\hat{\mu}_a| < \infty \) s.t \( \sqrt{n} \mu(P_h^{(a)}) = \hat{\mu}_a^h + o(|h|^2) \).

(iv) The estimators \( \hat{\mu}_a \) are best regular as defined in (4.1).

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

(i) \( \lim_{n \to \infty} \inf_{\pi_1, \pi^{fs}} \sup_{|h_1|, |h_0| \leq C} V_n(\pi_1, \pi^{fs}; h_1, h_0) \geq V^* \) for any \( C < \infty \).

(ii) If, further, Assumption 1(iv) holds,

\[
\sup_{\mathcal{J}} \lim_{n \to \infty} \sup_{|h_1|, |h_0| \in \mathcal{J}} V_n(\pi_1^*, \pi^{fs}; h_1, h_0) = V^*,
\]

where the outer supremum is taken over all finite subsets \( \mathcal{J} \) of \( \mathbb{R} \).

The first part of Theorem 1 says that \( V^* \) provides a lower bound on minimax regret. This is shown in Adusumilli (2021). Strictly speaking, the results in Adusumilli (2021) characterize the minimax regret using PDE methods. However, from the results in Section 3, it is straightforward to see that the regret can be alternatively characterized using \( V^* \) defined above.

The second part of the theorem says that \( \pi_1^*, \pi^{fs} \) attain this bound, thereby proving that they are asymptotically minimax optimal. The proof of this uses similar arguments in Hirano and Porter (2009, Lemma 3 & Theorem 3.2), and is therefore omitted.

4.2. **Non-parametric regime.** Let \( \mathcal{P}_1, \mathcal{P}_0 \) denote a candidate class of probability measures for the two treatments with bounded variances, and dominated by some measure \( \nu \). Also, let \( P_{0}^{(1)} \in \mathcal{P}_1 \) and \( P_{0}^{(0)} \in \mathcal{P}_0 \) denote reference probability distributions. Following Van der Vaart (2000), for each treatment \( a \in \{0, 1\} \) we consider smooth one-dimensional sub-models of the form \( \{ P_{t,h}^{(a)} : t \leq \eta \} \) for some
\( \eta > 0 \), where \( h(\cdot) \) is a measurable function satisfying
\[
\int \left[ \frac{(dP_{t,h}^{(a)})^{1/2}}{t} - \frac{(dP_0^{(a)})^{1/2}}{t} - \frac{1}{2} h \left( dP_0^{(a)} \right)^{1/2} \right]^2 d\nu \to 0 \text{ as } t \to 0. \tag{4.2}
\]

In analogy with the parametric setting, we compute minimax regret under the local (i.e., local to \( P_0^{(1)}, P_0^{(0)} \)) sequence of probability measures \( P_{t,h}^{(a)} \).

It is well known, see e.g. Van der Vaart (2000), that (4.2) implies \( \int h dP_0^{(a)} = 0 \) and \( \int h^2 dP_0^{(a)} < \infty \). The set of all such candidate \( h \) is termed the tangent space \( T(P_0^{(a)}) \). This is a subset of the Hilbert space \( L^2(P_0^{(a)}) \), endowed with the inner product \( (f, g)_a = \mathbb{E}_{P_0^{(a)}}[fg] \) and norm \( \|f\|_a = \langle f, f \rangle_a \). The mean rewards under \( P \in \mathcal{P}_a \) are given by \( \mu_a(P) = \int xdP(x) \). To obtain non-trivial regret bounds, we suppose \( \mu_a(P_0^{(a)}) = 0 \) for \( a \in \{0, 1\} \). The rationale for this is similar to setting \( \mu_a(\theta_0^{(a)}) = 0 \) in the parametric setting. Let \( \psi_a(x) := x - \int xdP_0^{(a)}(x) = x \) and \( \sigma_a^2 := \int x^2dP_0^{(a)}(x) \). Then, \( \psi_a(\cdot) \) is the influence function corresponding to \( \mu_a \), in the sense that under some mild assumptions on \( \{P_{t,h}^{(a)}\} \),
\[
\frac{\mu(P_{t,h}^{(a)}) - \mu(P_0^{(a)})}{t} - \langle \psi_a, h \rangle = \frac{\mu(P_{t,h}^{(a)})}{t} - \langle \psi_a, h \rangle = o(t). \tag{4.3}
\]

It will be shown that the following policy rules are minimax optimal:
\[
\pi_1^* = \frac{\sigma_1}{\sigma_1 + \sigma_0}; \quad \pi_1^{\text{is}} = \mathbb{I} \left\{ \frac{x_1(1)}{\sigma_1} - \frac{x_0(1)}{\sigma_0} \geq 0 \right\},
\]
where \( x_a(t) := n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \mathbb{I} \{A_i = a\} Y_i \) with \( A_i \in \{0, 1\} \) indicating which treatment was sampled in period \( i \). The following assumptions are taken from Adusumilli (2021):

**Assumption 2.** (i) The sub-models \( \{P_{t,h}^{(a)}; h \in T(P_0)\} \) satisfy (4.2).

(ii) \( \mathbb{E}_{P_0^{(a)}}[\exp |Y|] < \infty \) for \( a \in \{0, 1\} \).

(iii) For each \( a \in \{0, 1\} \), \( \sqrt{n} \mu(P_{1/\sqrt{n},h}^{(a)}) = \langle \psi_a, h \rangle + o \left( \|h\|^2 \right) \).

For the statement of the theorem below, let \( \mathcal{H}_I^{(a)} \) denote a finite subset of the tangent space \( T(P_0^{(a)}) \) of dimension \( I \); this is the space spanned by \( I \) sub-elements from an orthonormal basis \( \{\phi_1, \phi_2, \ldots \} \) for the closure of \( T(P_0^{(a)}) \).
Theorem 3. Suppose that Assumptions 2(i)-(iii) hold. Then,

$$\sup_{I_1, I_0} \lim_{n \to \infty} \inf_{\pi \in \Pi} \sup_{h_1 \in H^{(1)}_{I_1}, h_0 \in H^{(0)}_{I_0}} V_n(\pi_1, \pi^{fs}; h_1, h_0) \geq V^*$$

where the outer supremum is taken over all finite subsets $I_1, I_0$ of the tangent spaces $T(P^{(1)}_0), T(P^{(0)}_0)$. Furthermore,

$$\sup_{I_1, I_0} \lim_{n \to \infty} \sup_{h_1 \in H^{(1)}_{I_1}, h_0 \in H^{(0)}_{I_0}} V_n(\pi^*_1, \pi^{*fs}; h_1, h_0) = V^*.$$ 

The first part of the above theorem shows that the minimax regret is lower bounded by $V^*$. This is already shown in Adusumilli (2021). The definition of minimax regret here is the same as in Van der Vaart (2000, Theorem 25.21). The second part of the theorem shows that $\pi^*_1, \pi^{*fs}$ attain this bound and are therefore asymptotically minimax optimal. The proof of this is similar to Hirano and Porter (2009, Lemma 3'), and is omitted.

4.3. Unknown variances. Replacing $\sigma_1, \sigma_0$ with consistent estimates has no effect on asymptotic risk. Since the optimal allocation rule depends on these quantities, we can proceed as follows: Let $\tilde{n} := n^\rho$ for some $\rho \in (0, 1)$. For the first $\tilde{n}$ observations, we sample the treatments in equal proportions and use the outcomes generated to obtain consistent estimates, $\hat{\sigma}_1, \hat{\sigma}_0$ of $\sigma_1, \sigma_0$ (alternatively, one could obtain these estimates from a pilot experiment). We then apply the Neyman allocation rule $\hat{\sigma}_a/(\hat{\sigma}_1 + \hat{\sigma}_0)$ for all the observations from $\tilde{n}$ onwards. The resulting policies also attain the lower bounds on asymptotic minimax regret, in both parametric and non-parametric settings. In practice, the choice of $\rho$ would matter, and further work is needed to determine this.

5. Conclusion

This paper describes the asymptotic minimax optimal policy for best arm identification with two arms. The optimal sampling rule involves sampling in fixed proportions and there is no adaptation to past outcomes. In this setting, the goals of estimation and minimizing regret (at least according to the minimax criterion) coincide. Crucial to this result is the observation that nature’s least favorable
prior has a two-point support and makes the agent indifferent between any sampling rule. Going beyond two arms, we expect the least favorable prior to generally have as many support points as the number of arms. It is unknown, however, if there exist indifference inducing priors beyond the two arm case.

REFERENCES

K. Adusumilli, “Risk and optimal policies in bandit experiments,” arXiv preprint arXiv:2112.06363, 2021.

A. Carpentier and A. Locatelli, “Tight (lower) bounds for the fixed budget best arm identification bandit problem,” in Conference on Learning Theory. PMLR, 2016, pp. 590–604.

C.-H. Chen, J. Lin, E. Yücesan, and S. E. Chick, “Simulation budget allocation for further enhancing the efficiency of ordinal optimization,” Discrete Event Dynamic Systems, vol. 10, no. 3, pp. 251–270, 2000.

A. Garivier and E. Kaufmann, “Optimal best arm identification with fixed confidence,” in Conference on Learning Theory. PMLR, 2016, pp. 998–1027.

P. Glynn and S. Juneja, “A large deviations perspective on ordinal optimization,” in Proceedings of the 2004 Winter Simulation Conference, 2004, vol. 1. IEEE, 2004.

K. Hirano and J. R. Porter, “Asymptotics for statistical treatment rules,” Econometrica, vol. 77, no. 5, pp. 1683–1701, 2009.

M. Kasy and A. Sautmann, “Adaptive treatment assignment in experiments for policy choice,” 2019.

M. Kato, K. Ariu, M. Imaizumi, M. Uehara, M. Nomura, and C. Qin, “Best arm identification with a fixed budget under a small gap,” stat, vol. 1050, p. 11, 2022.

S. Mannor and J. N. Tsitsiklis, “The sample complexity of exploration in the multi-armed bandit problem,” Journal of Machine Learning Research, vol. 5, no. Jun, pp. 623–648, 2004.

C. Qin, D. Klabjan, and D. Russo, “Improving the expected improvement algorithm,” Advances in Neural Information Processing Systems, vol. 30, 2017.

D. Russo, “Simple bayesian algorithms for best arm identification,” in Conference on Learning Theory. PMLR, 2016, pp. 1417–1418.
A. N. Shiryaev, *Optimal stopping rules*. Springer Science & Business Media, 2007, vol. 8.

J. Stoye, “New perspectives on statistical decisions under ambiguity,” *Annu. Rev. Econ.*, vol. 4, no. 1, pp. 257–282, 2012.

A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000.

A. Wald, “Statistical decision functions which minimize the maximum risk,” *Annals of Mathematics*, pp. 265–280, 1945.