FIXED-BUDGET TARGET-TIME BEST-ARM IDENTIFICATION IN TEMPORALLY EVOLVING BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-arm identification (BAI) in multi-armed bandits usually assumes stationary reward distributions while existing works on non-stationary BAI typically chase an arm with largest average reward over a predefined timespan. We introduce the new setting of *fixed-budget target-time BAI* where rewards may evolve away from a target decision time, and the learner must identify the best arm at that target without knowing the relevant temporal scale of non-stationarity in advance. This setting is motivated by laboratory studies that seek to identify the intervention producing the most favorable outcome at a target time –for example, at the beginning or at the end of a prescribed treatment schedule. Observations farther from the target may be less representative (ie *biased*) and deploying or evaluating an arm can be costly. This introduces a unique challenge where a learner has to manage the budget constraint using the right amount of data that can strike an optimal bias-variance tradeoff when performing BAI. In this work, we explore the limits of identifiability for the target-time non-stationary BAI problem and propose an algorithm that is information-theoretically near *optimal* (upto polylog factors) and *do not* require apriori knowledge of drift. Experiments on synthetic and real-world data expose the strengths and weaknesses of the proposed methods over existing BAI algorithms.

1 INTRODUCTION

In this paper, we study a new variant of the problem of Best Arm Identification (BAI) in multi armed bandits. In the considered setup, a learner needs to identify the arm with highest reward at a prespecified target time. However, it differs from the usual stationary formulation of BAI in that arm pulls reveal noisy observations that are sampled from distributions that can be drifted away from the target-time reward distribution. More precisely, associated with each arm, there is an expected reward stream $\mu_1, \ldots, \mu_T$, where 1 denotes the target time and 2 is the reward at the timestamp closest to the timestamp denoted by 1 and so on. The progression of timestamps can be either retrospective (backward running) or prospective (forward running) depending on the application. The reward of the arm, if observed, at time t is given by $Y_t = \mu_t + \epsilon_t$ for some iid subgaussian noise ϵ. The learner is tasked to identify the best arm at time 1 with a budget constraint that it can only perform atmost B arm pulls before declaring the best arm. When the learner decides to explore a particular arm, it receives a reward that is sampled from the next available distribution of that arm based on pulled history of that arm so far. Hence the observed reward stream is emitted from a temporally varying distribution.

As a concrete practical motivation, consider a completed preclinical experiment comparing several treatment protocols. The actionable objective is to select which protocol should be advanced to the next stage of investigation, and this decision is naturally based on treatment efficacy at a prespecified endpoint—the target time—rather than on average performance over the entire treatment course. Although samples may have been collected and stored throughout the duration of the study, obtaining the reward of interest from each sample can require an expensive downstream annotation and expert interpretation. Hence the available annotation budget may permit evaluation of only a subset of all stored samples. When target-time observations do not clearly distinguish the leading treatments, earlier samples offer additional statistical evidence. However, their relevance to the target-time ordering may deteriorate as the biological system evolves.

Pulling an arm corresponds to analyzing the next most recent unevaluated sample from the cohort receiving that treatment protocol. The learner therefore begins with samples closest to the treatment endpoint and consults progressively earlier samples only when the recent evidence is insufficient to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distinguish the leading protocols. This exposes the central tradeoff of the setting: incorporating older observations can reduce statistical uncertainty, but their relevance to the terminal arm-ordering may deteriorate with temporal distance. Though we used the example of preclinical trials as a practical motivation, we emphasize that our setting, formally described in Fig.1, is more broadly applicable. We refer the reader to Appendix A for other motivating applications for the non-stationary BAI framework.

More generally, non-stationary BAI presents two complementary challenges. First, an *exploration* challenge requires allocating the limited pull budget toward promising arms rather than continuing to measure clearly inferior ones. Second, an *estimation* challenge requires reliably inferring the ordering of arms at the target time from non-stationary historical observations. Without structural knowledge of how rewards evolve, the learner must therefore adapt to both which arms to continue sampling and how much historical data to use when estimating their relative ordering. We study the statistical limits of this problem and algorithmic designs that address these two challenges jointly.

Following is the list of our contributions:

- We introduce and formalize fixed-budget target-time BAI with temporally ordered, non-stationary reward streams. Each arm's reward trajectory is fixed before interaction and is not altered by sampling. The reward trajectory can be non-monotone and exhibit repeated reversals in the relative ordering of the arms. The framework therefore separates the decision-relevant target-time ordering from stationary, horizon-average, and learning-curve objectives, and captures applications in which temporally displaced observations are informative but potentially mismatched; see introduction and Appendix A for motivating applications.
- We develop `SignCount-SH` (Algorithm 2), which combines multiresolution pairwise sign estimation with the elimination structure of Sequential Halving (Karnin et al., 2013). We prove a high-confidence simple-regret guarantee (Theorem 5) and a minimax lower bound matching upto poly-logarithmic factors (Theorem 8) over a complexity-bounded problem class.
- We evaluate the method on simulated and real datasets and show performance improvements over existing baselines in non-stationary environments (Section 4). Ablation studies further characterize when the proposed methods are most beneficial.

Notes on technical novelty. Like Sequential Halving (Karnin et al., 2013), our algorithm `SignCount-SH` halves an active arm set at each epoch, but uses a fundamentally different elimination rule. It performs a duel for every pair of active arms by examining their empirical reward differences over multiple window lengths, and ranks and eliminates arms by their aggregate duel win-scores. Because different pairwise winners may be certified at different window lengths, the collective duel outcomes need not be transitive. So winning a direct good-versus-bad arm duel does not by itself ensure that the good arm has the larger aggregate score. Moreover, the active set is data-dependent. These subtleties make the proof substantially different from, and more involved than, the classical Sequential-Halving analysis. We introduce a *pairwise resolution depth* that certifies target-time orderings and a *ranking depth* that guarantees good-over-bad score dominance for any surviving set. Sorting bad arms by ranking depth replaces classical gap sorting and yields the stagewise survival guarantee. A complementary lower bound matches the resulting complexity-class budget up to polylogarithmic factors.

Due to space limitations, a discussion on related work is presented in Appendix B.

2 Setting

In this section, we formalize the problem setting and fix notations. The interaction protocol is summarized in Fig.1. Though we restrict to the scenario where an arm pull reveals one observation, extension of our results to the case where a pull results in a batch of observations is analogous.

Assumptions. We consider the class of instances where the mean gaps $\Delta_{a,b}(t) = \mu_a(t) - \mu_b(t)$ are uniformly bounded. That is $|\Delta_{a,b}(t)| < \infty$ for any time t and for any pair of arms a and b. Further, we assume that the sub-gaussian parameter of the noise of the observations $Y_{i,t}$ is known to be σ. Let $\Delta_a := \mu_1(1) - \mu_a(1)$. Without loss of generality, we suppose that arm 1 is optimal at terminal time 1 and that $0 < \Delta_2 \leq \ldots \leq \Delta_K$.

- The learner is given a maximum arm-pull budget B. Each arm $i \in \{1, \dots, K\} := [K]$ has its own target-indexed reward stream $Y_{i,1}, \dots Y_{i,t}, \dots Y_{i,T}$, where $t = 1$ is closest to the target and larger t is farther from it (older retrospectively or later prospectively). The reward model is
$$Y_{i,t} = \mu_i(t) + \epsilon_{i,t},$$
where $\mu_i(t)$ may vary with depth t and $\epsilon_{i,t}$ is independent mean-zero sub-gaussian noise.
- Let $N_i(b)$ denote the number of times arm i has been pulled after b total pulls across all arms. If the learner pulls arm i at global pull $b+1$, it observes the next unrevealed sample in arm i's target-indexed stream $Y_{i,N_i(b)+1}$.
- The learner chooses pulls adaptively. At each global pull $b+1$, conditional on the history $\mathcal{F}_b = \sigma\left(\left\{I_\ell, Y_{I_\ell, N_{I_\ell}(\ell-1)+1}\right\}_{\ell=1}^{b}\right)$, the learner selects an arm $I_{b+1} \in [K]$ to pull. The selected arm I_{b+1} is $\mathcal{F}_b$-measurable.
- The learner stops after B global pulls. $\sum_{i=1}^{K} N_i(B) = B$.
- The learner outputs a recommendation $\widehat{i} \in [K]$. Let $i^\star \in \arg\max_{i\in[K]} \mu_i(1)$ be the terminal-time best arm. The objective is to control the simple regret by reducing $\mathbb{P}(\mu_{i^*}(1) - \mu_{\widehat{i}}(1) > \epsilon)$ for any $\epsilon \geq 0$, subject to the budget cap B, where the probability is wrt randomness in observations and algorithms.

Figure 1: Problem setting for target-time non-stationary BAI under budget constraints

3 Main Results

In this section, we present our main algorithm, its bounds, followed by a nearly matching information theoretic lower bound on simple regret. All proofs are deferred to Appendix C.

3.1 Algorithm

Consider `Multi-Resolution-Duel` presented in Algorithm 1. The simple observation that underpins our algorithm design is that best-arm identification only requires recovering the *ordering* of arms, rather than accurately estimating their terminal rewards. At resolution k_i, `Multi-Resolution-Duel` compares two arms using the empirical averages of their most recent k_i observations. Increasing k_i reduces statistical noise, but it also incorporates progressively older data. Such smoothing may substantially bias the estimated magnitude of the target-time gap, yet still *preserve its sign*. This makes sign identification potentially a much easier than terminal reward estimation, especially under non-stationarity.

Motivated by this observation, `Multi-Resolution-Duel` scans dyadic resolutions from the most recent, least-smoothed estimate to progressively more averaged estimates, and stops as soon as one scale provides statistically reliable evidence for an ordering. In particular, Line 5 declares a sign only when $|\widehat{d}_{a,b}(k_i)| > 2q_i(\delta)$, where $q_i(\delta)$ quantifies the uncertainty of the empirical gap at that resolution. Thus the algorithm uses additional averaging only when needed to overcome noise, while conservatively avoiding the usage of data that are generated away from terminal time once the sign is already clear. This design reflects a natural inductive bias that observations closer to the target time are more informative about the terminal best arm.

However, it is possible that for strongly adversarial environments, the sign-resolution obtained via Line 6 can be wrong. We note that in such a situation, all the previous resolutions did not provide good statistical signal regarding the sign of the ordering between arms. So recovering the correct terminal sign is impossible without structural assumptions on the drift. This naturally imposes a constraint on the degree of non-stationarity of the environment where meaningful detection of sign is possible. The complexity measure presented in Definition 4 will formalize the problem class where non-trivial simple regret guarantees can be made.

If no scale provides sufficient evidence (i.e, do not pass the condition in Line 5), the comparison is left unresolved and assigned the neutral score $1/2$. All pairwise tests reuse the same arm-level observations, so this comparison structure does not incur a quadratic sampling cost.

The resulting pairwise signs are aggregated by `SignCount-SH`. This algorithm retains the stagewise halving and equal per-epoch-budget allocation structure of classical Sequential Halving (Karnin et al.,

Algorithm 1: MULTI-RESOLUTION-DUEL

Input: Arms a, b. Their reward stream $Y_{r,1}, \ldots, Y_{r,n}$ for $r = a, b$; Confidence parameter δ, noise-std deviation σ.
Output: Estimate of which arm has higher terminal mean

$j \leftarrow \lfloor \log_2 n \rfloor$; $\quad W_{a,b} \leftarrow 1/2$;
for $i = 0, \ldots, j$ **do**
  $k_i \leftarrow 2^i$; $\quad q_i(\delta) = 2\sigma\sqrt{\frac{2}{k_i}\log(\frac{\pi^2 K(i+1)^2}{3\delta})}$;
  $\hat{\mu}_a(k_i) \leftarrow 1/k_i \sum_{r=1}^{k_i} Y_{a,r}$; $\hat{\mu}_b(k_i) \leftarrow 1/k_i \sum_{r=1}^{k_i} Y_{b,r}$; $\hat{d}_{a,b}(k_i) = \hat{\mu}_a(k_i) - \hat{\mu}_b(k_i)$;
  if $|\hat{d}_{a,b}(k_i)| > 2q_i(\delta)$ **then**
    $W_{a,b} \leftarrow \mathbb{I}(\hat{d}_{a,b}(k_i) \geq 0)$, where $\mathbb{I}$ is the binary indicator function;
    break ; `// Exits the loop`
return $W_{a,b}$;

Algorithm 2: SIGNCOUNT-SH

Input: Arms $[K]$, pull budget B, confidence δ, noise-std deviation σ.
Output: Estimated best arm $\widehat{i}$

$L \leftarrow \lceil \log_2 K \rceil$; $\quad A_1 \leftarrow [K]$;
Initialize pull counts $N_i \leftarrow 0$, for all $i \in [K]$;
for $\ell = 1, 2, \ldots, L$ **do**
  $C_i^\ell \leftarrow 0$ for all $i \in [K]$;
  if $|A_\ell| = 1$ **then**
    return the unique arm in A_ℓ;
  $n_\ell \leftarrow |A_\ell|$;
  $q_\ell \leftarrow \lfloor B/(Ln_\ell) \rfloor$;
  for $r = 1, 2, \ldots, q_\ell$ **do**
    for *each arm* $a \in A_\ell$ **do**
      Pull arm a once and observe Y_{a,N_a+1};
      $N_a \leftarrow N_a + 1$;
  $T_\ell \leftarrow N_a$ for any $a \in A_\ell$; `// all active arms have equal depth`
  for *each arm pair* $(a, b) \in A_\ell \times A_\ell$ *with* $a < b$ **do**
    $W_{a,b} \leftarrow$ Result of Algorithm 1 run on data of arms a and b; $Y_{\cdot,1}, \ldots, Y_{\cdot,T_\ell}$;
    $C_a^\ell \leftarrow C_a^\ell + W_{a,b}$; $\quad C_b^\ell \leftarrow C_b^\ell + (1 - W_{a,b})$;
  Let $A_{\ell+1}$ be the $\lceil n_\ell/2 \rceil$ arms in A_ℓ with largest counts $C_{(\cdot)}^{(\ell)}$;
return the unique arm in A_{L+1};

2013), but departs from it in how surviving arms are ranked. Rather than ordering arms by empirical reward estimates, it ranks them by accumulated pairwise sign evidence. This departure is precisely what allows the procedure to exploit temporal scales that may be unsuitable for accurate reward estimation but are still informative about the sign of terminal ordering.

3.2 ANALYSIS OF SIMPLE REGRET

In this section, we analyse the simple regret (i.e, $\mu_{i^*}(1) - \mu_{\widehat{i}}(1)$) of Algorithm 2 and provide nearly matching minimax lowerbounds. Throughout the analysis, we will be referring to several quantities defined in Algorithms 1 and 2. We start by defining a natural population level quantity that will become a central object in our passage towards a complexity measure.

Definition 1. *Let* $s_{a,b} = sign(\mu_a(1) - \mu_b(1)$ *which assumes the value* $+1$ *for positive arguments and* -1 *otherwise. and* $\bar{d}_{a,b}(k_j) = 1/k_j \sum_{i=1}^{j}(\mu_a(i) - \mu_b(i))$. *Let* $\Gamma_{a,b}(j) = s_{a,b}\bar{d}_{a,b}(k_j)$. *The quantity* $\Gamma_{a,b}(j)$ *captures how well the terminal ordering of arms* a *and* b *aligns with the smoothed history at a prefix length* k_j. *Define:*

$$\tau_{a,b}^{(\delta)} = \inf\left\{k_j : \begin{matrix}\Gamma_{a,b}(i) > -q_i(\delta) & \forall i < j, \\ \Gamma_{a,b}(j) > 3q_j(\delta)\end{matrix}\right\}.$$

If no such scale exists, we take $\tau^{(\delta)}_{a,b} = \infty$. The quantity $\tau^{(\delta)}_{a,b}$ will be referred as the resolving depth for arm pair (a, b) (at the confidence δ).

We now explain the intuition behind $\tau^{\delta}_{a,b}$. Consider the averaging estimator $\hat{d}_{a,b}(k_j)$ defined in Algorithm 1. At scale j, the noise level of this estimator is $\sigma/\sqrt{k_j}$. Hence variance of the estimator decays with scale j. However, larger smoothing at high scales can potentially disrupt the alignment of the mean difference with the terminal ordering. The quantity $\tau^{(\delta)}_{a,b}$ is the depth (k_j) corresponding to the first scale j, where there alignment signal is statistically significant so that the `Multi-Resolution-Duel` will trigger the correct sign by at-most the depth k_j. Further, at scales earlier than j, the alignment is not too weak so that `Multi-Resolution-Duel` will not produce a wrong sign output in case it is stopped earlier than depth k_j. Note that $\tau^{(\delta)}_{a,b}$ can be finite even when the alignment is oriented in the opposite direction for intermediate timestamps as long as a later timestamp recovers it correctly.

The following lemma makes these ideas precise.

Lemma 2. *Consider a time $t \geq \tau^{(\delta)}_{a,b}$ and any pair of arms (a, b). With probability at-least $1 - \delta$, `Multi-Resolution-Duel` (Algorithm 1) would output $W_{a,b} = s_{a,b}$.*

Next, we define a ranking certificate. For target accuracy ϵ, define the ϵ-good and ϵ-bad arms as below.

$$\mathcal{G}_\epsilon = \{a : \Delta_a \leq \epsilon\}, \quad M_\epsilon = |\mathcal{G}_\epsilon|, \quad \mathcal{B}_\epsilon = \{a : \Delta_a > \epsilon\}.$$

Definition 3. *Fix a good arm $g \in \mathcal{G}_\epsilon$ and bad arm $a \in \mathcal{B}_\epsilon$. For a third arm b, define*

$$\chi^{(\delta)}_{g,a;b} = \min \left\{ \begin{matrix} \tau^{(\delta)}_{g,b} & \text{if } \mu_g(1) > \mu_b(1), \\ \tau^{(\delta)}_{b,a} & \text{if } \mu_b(1) > \mu_a(1) \end{matrix} \right\}.$$

Let $\rho^{(\delta)}_{g,a} = \max\left\{\tau^{(\delta)}_{g,a}, \max_{b \neq g,a} \chi^{(\delta)}_{g,a;b}\right\}$.

The interpretation is that by depth $t \geq \rho^{(\delta)}_{g,a}$ we have enough statistical signal to conclude the following: a) $g > a$ is certified by Algorithm 1; and b) for every third arm b, either $g > b$ is certified or $b > a$ is certified. The condition (b) will be useful in proving the sign-count ordering $C^\ell_g > C^\ell_a$ for an adequately large epoch ℓ (see Lemma 12). Next we turn the ranking certificate into a complexity measure for identifying best arm under non-stationarity.

Definition 4. *For a fixed good arm $g \in G_\epsilon$, sort the bad arms as:*

$$\rho^{(\delta)}_{g,[1]} \geq \cdots \geq \rho^{(\delta)}_{g,[K-|G_\epsilon|]}.$$

Let θ denote the parameter that describes a bandit instance. Define the complexity measure

$$H^{(\delta)}_{NS}(\theta, \epsilon) = \min_{g \in \mathcal{G}_\epsilon} \max_{1 \leq m \leq K-|G_\epsilon|} (|G_\epsilon| + m)\rho^{(\delta)}_{g,[m]}, \tag{1}$$

and $H^{(\delta)}_{NS}(\theta, \epsilon)$ is taken to be zero if $|G_\epsilon| = K$.

The quantity in Eq.(1), is expressed in terms of an order statistic. While this form is intended to be a generalization to the well-established H_2 style complexity in stationary BAI, one can define an equivalent representation to gain some intuition. Let us first fix a depth t. We remind the reader that t is the temporal distance from the target time 1 in an arm's reward stream (see Fig.1). Let $N^{(\delta)}_g(t) = |\{a \in B_\epsilon : \rho^{(\delta)}_{g,a} \geq t\}|$ be the number of bad arms that requires atleast t samples to correctly certify their sign-count ordering relative to g. These are the relevant hard contenders to arm g, from the bad arm set, at time t.

Ignoring ties for a moment, at $t = \rho^{(\delta)}_{g,[m]}$ there are exactly m bad arms with resolution depth (to resolve means to correctly certify the sign-count ordering wrt g) at-least t. Further due to Lemma 12, the m^{th} bad arm gets resolved at depth t. Thus $N^{(\delta)}_g(t) = m$ when $t \in (\rho^{(\delta)}_{g,[m+1]}, \rho^{(\delta)}_{g,[m]}])$. Consequently, we have the equivalent representation

$$H^{(\delta)}_{\text{NS}}(\theta, \epsilon) = \min_{g \in \mathcal{G}_\epsilon} \max_{t \in \{\rho^{(\delta)}_{g,a} : a \in \mathcal{B}_\epsilon\}} (|G_\epsilon| + N^{(\delta)}_g(t))t = \min_{g \in \mathcal{G}_\epsilon} \sup_{t \geq 0, N^{(\delta)}_g(t) \geq 1} (|G_\epsilon| + N^{(\delta)}_g(t))t.$$

The restriction $N_g^{(\delta)}(t) \geq 1$ is there because beyond a depth with $N_g^{(\delta)}(t) = 0$, there are no bad arms to differentiate from ϵ-good arms and the simple regret is guaranteed to be within ϵ. We note that $N_g^{(\delta)}(t)$ is a non-increasing function of t. If $H_{\text{NS}}^{(\delta)}(\theta, \epsilon) \leq C$, then one obtains the envelope $N_g^{(\delta)}(t) \leq C/t - |G_\epsilon|$ for some anchor $g \in G_\epsilon$. Faster the decay of $N_g^{(\delta)}(t)$ is, faster the bad arms becomes resolvable against that good arm and hence leading to lower statistical difficulty for BAI.

We are now ready to upper bound the simple regret of Algorithm 2 in terms of the quantity $H_{\text{NS}}^{(\delta)}$.

Theorem 5. *Suppose the budget $B \geq 4 \log_2 K \cdot H_{NS}^{(\delta)}(\theta, \epsilon)$. Then with probability at-least $1 - \delta$, Algorithm 2 terminates with an arm $\hat{i}$ that is ϵ-optimal. i.e, $\Delta_{\hat{i}} = \mu_1(1) - \mu_{\hat{i}}(1) \leq \epsilon$.*

Proof sketch. The proof tracks an ϵ-good anchor arm $g^\star$ attaining the minimum in Eq.(1). We split the analysis into two cases: a) an early stage where epochs satisfy $|A_\ell| > 2|\mathcal{G}_\epsilon|$ and b) a complementary late stage. Consider an epoch ℓ in the early stage with $g^* \in A_\ell$. For the sake of contradiction assume that $g^* \notin A_{\ell+1}$. Then we argue that there are at-least m bad arms that empirically outrank the arm g^* in terms of their signcount scores for an m that is sufficiently large. The large enough m and the minimum budget constraint ensures that the magnitude of the quantity $\rho_{g^*,[m]}$ is at-most the depth T_ℓ defined in Algorithm 2. This implies that there can be at-most $m - 1$ bad arms whose ordering can remain unresolved against the arm g^* by the end of epoch ℓ, and hence can outrank g^* empirically with high probability. This yields a contradiction to the fact that there must be at-least m bad arms that must outrank g^* to eliminate it from $A_{\ell+1}$. This proves the survival of g^* into the next epoch. Once the active set becomes small enough in the late stage, T_ℓ becomes large enough, which alongside the budget constraint ensures that either $g^\star$ survives or every surviving arm is ϵ-optimal. The full proof is given in the appendix. □

Remark 6. *We emphasize that the ϵ in the theorem is merely an analysis parameter and the algorithm do not use it as a hyperparamater. Setting $\epsilon < \Delta_2$ provides guarantee for the probability of identifying the best arm.*

We recover the familiar H_2 type guarantee for BAI in stationary case upto logarithmic terms as an immediate corollary.

Corollary 7 (Recovery of stationary guarantee)**.** *Let $H_2(\theta) = \max_{2 \leq i \leq K} \frac{i}{\Delta_i^2}$. It can be shown that*

$$H_{NS}^{(\delta)}(\theta, 0) \lesssim K + \sigma^2 H_2(\theta) \left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\Delta_2^2}\right) \right].$$

Thus due to Theorem 5, $B \gtrsim \log K \left(K + \sigma^2 H_2(\theta)\ \text{polylog}(K, 1/\delta)\right)$ implies $\Pr_\theta(\widehat{a} \neq 1) \leq \delta$.

The additive K is present even in the noiseless limit $\sigma \to 0$ as every arm must be observed at least once. Classical fixed-budget Sequential Halving guarantees are often presented as exponential error bounds of the form $\Pr(\widehat{a} \neq 1) \leq \exp\left(-c \frac{B}{H_2(\theta)\ \text{polylog}(K)}\right)$. This can be inverted to provide the stationary guarantee similar to Corollary 7 for any δ. However, we remark that Corollary 7 is applicable for only one prespecified value of δ since it is an input parameter to Algorithm 2. Thus the classical exponential tail bound from Karnin et al. (2013) is stronger. It remains as an open question if such uniform tail bounds are possible in the non-stationary setting. We now proceed to construct a lowerbound showing that the budget condition in Theorem 5 is near minimax optimal.

Theorem 8. *[Finite-confidence minimax lower bound] Assume Gaussian noise with variance σ^2, $K \geq 8$, $\delta \in (0, 1/4]$, and $0 < \epsilon \leq \sigma/8$. For $h > 0$, define $\Theta_h^{(\delta,\epsilon)} := \left\{\theta : H_{\text{NS}}^{(\delta)}(\theta, \epsilon) \leq h\right\}$, and let Π_B denote the class of adaptive fixed-budget policies using at most B pulls. For $\pi \in \Pi_B$, let $\widehat{i}$ denote its recommended arm. Define the simple regret as $r_B := \mu_1(1) - \mu_{\widehat{i}}(1)$. There exist universal constants $c, C > 0$ such that for every*

$$h \geq C \frac{K\sigma^2}{4\epsilon^2} \left[\log(eK/\delta) + \log\log(e^e + \sigma^2/4\epsilon^2)\right], \tag{2}$$

$$\text{if } B \leq c\, h \frac{\log(1/\delta)}{\log(eK/\delta) + \log\log(e^e + h/K)}, \tag{3}$$

then, for every $\pi \in \Pi_B$, $\sup_{\theta \in \Theta_h^{(\delta,\epsilon)}} \mathbb{P}_\theta^\pi(r_B > \epsilon) > \delta$. Moreover, the hard instance can be chosen to satisfy $H_{\text{NS}}^{(\delta)}(\theta, \epsilon) \geq c\, h$.

Remark 9 (Finite-confidence near-minimaxity). *Fix ϵ, δ and an admissible h as in Theorem 8. Over $\Theta_h^{(\delta,\epsilon)}$, no policy can guarantee $\sup_{\theta \in \Theta_h^{(\delta,\epsilon)}} \mathbb{P}_\theta^\pi(r_B > \epsilon) \le \delta$ when $B \lesssim h \log(1/\delta)/[\log(eK/\delta) + \log\log(e^e + h/K)]$, whereas Theorem 5 shows that* `SignCount-SH` *provides this guarantee when $B \gtrsim h \log K$. Thus,* `SignCount-SH` *is finite-confidence near-minimax over $\Theta_h^{(\delta,\epsilon)}$, up to poly-logarithmic factors in K and $1/\delta$ and a mild doubly logarithmic factor in h.*

4 Experimental Results

We compare against three fixed-budget BAI baselines and one ablation of our method. *Vanilla-SH* is the original Sequential Halving algorithm of Karnin et al. (2013). *Successive Rejects* follows Audibert et al. (2010). We also evaluate *P1-RAGE* and *Adjacent-BAI*, obtained by specializing the non-stationary linear bandit method of Xiong et al. (2024) and Maynard-Zhang et al. (2026) respectively to the standard-basis arm set. Finally, *Full-prefix SH* is an ablation: it uses exactly the same Sequential Halving allocation and cumulative observations as `SignCount-SH`, but replaces multiresolution sign based duel with full-prefix empirical mean based duel. It therefore isolates the effect of Algorithm 1.

In the implementation of our method, we remark that at Line 17 of Algorithm 2, there may be multiple arms with same sign-count scores C_a^ℓ. To break ties we adopt a simple optimism in the face of uncertainity principle: we optimistically treat the reward process of such tied arms as stationary and use the rule of the stationary baseline Vanilla-SH to estimate their rewards. This entails using only the observations accumulated in the current epoch to compute empirical means. Ties are then broken with priority induced by those empirical means.

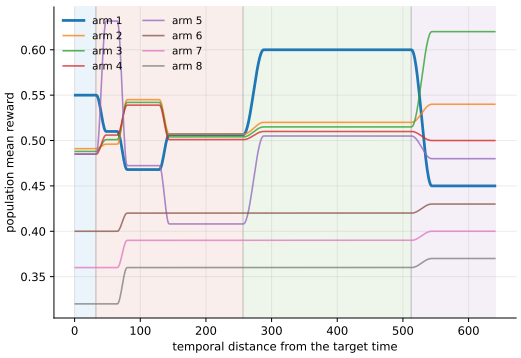


Figure 2: *Reward trajectories for instance used in simulation.*

4.1 Simulation Experiments

Setup and Metrics. We consider the reward trajectory shown in Fig.2, which comprises of four distinct regimes, shaded in different colors (see Appendix D for the exact equations). Observations are obtained by adding independent Gaussian noise to the reward trajectory. Unless otherwise stated, we use $\sigma = 0.02$ and the confidence parameter $\delta = 0.05$ throughout. We consider 4 target times, $\{1, 49, 289, 545\}$, planted near the start of each regime. For a BAI run for the target time t_0, observations from times $t \ge t_0$ are supplied to the algorithms. We report the mean target-time best arm identification probability as well as the mean simple regret in Fig.3 averaged across the target times and 1000 independent runs as a function of normalized budget B/K.

We can see from Fig.3 that at smaller values of budget where most of the data is from distribution similar to that at target-time, all methods perform favorably owing to the stationarity of the data

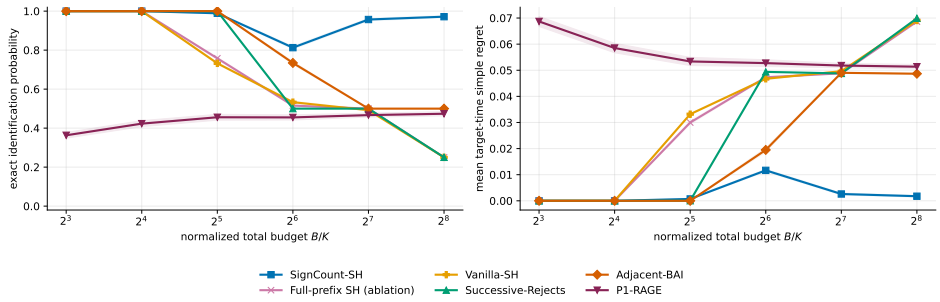


Figure 3: *Average target-time best-arm identification probability (left) and mean target-time simple regret (right) against normalized budget at $\sigma = 0.02$, $\delta = 0.05$. Shaded bands are 95% confidence intervals over 1,000 runs and 4 target times. Numerical values of the metrics are shown in Tables 2 and 3.*

distribution. As budget increases, more biased data enters the algorithms causing the performance of stationary baselines to fall. On the other hand, `SignCount-SH` is relatively more robust towards this bias as it maintains better performance even at large budgets while also having competitive performance at small budgets. This is made possible thanks to the adaptive selection of data by Algorithm 1 which resolves pairwise duels at the dyadic depth that is closest to the target time where a statistically relevant sign certificate is obtainable. We provide evidence for this phenomenon of adaptive data selection in Appendix G. We refer the reader to Appendix D for ablations wrt the noise level and other numerical results.

4.2 Real Data Experiments: NYC Yellow-Taxi Pickup Zone Recommendation

Dataset and Setup. For experiments on a real dataset, we consider the public 2024 New York City TLC yellow-taxi data, which consists of 12 monthly trip files containing pickup timestamps and taxi-zone identifiers (New York City Taxi and Limousine Commission, 2024). Pickup zones are arms, and we divide time into non-overlapping 90-minute slots. The population-level reward of an arm in a slot is the fraction of all trips in that slot whose pickup occurred in the corresponding zone. An observation of an arm is generated by drawing a uniform random subsample of trips from the relevant slot and returning the fraction whose pickup occurred in that arm's zone. This is an unbiased noisy estimate of the arm's population-level reward. Each such batch estimate counts as one observation toward the budget, irrespective of the number of trips in the subsample. Due to the non-stationary nature of taxi demand, the (population-level) reward of each arm changes throughout the day (see Fig. 8).

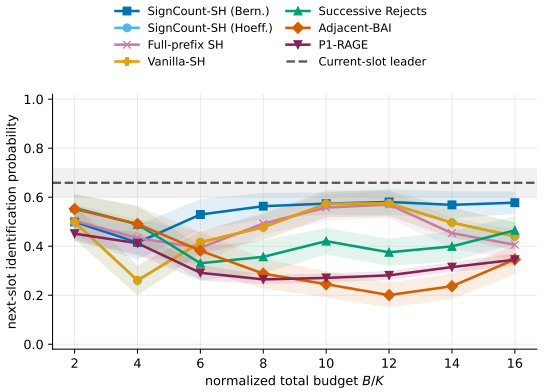


Figure 4: *Mean probability of recommending best arm for the next time slot. Bands are pointwise 95% confidence intervals over 17 weekly averages, and 50 random trip-subsampling runs. The dashed "Current-slot leader" is an oracle follow-the-leader strategy that recommends the population maximizer of reward at t for $t+1$. Numerical values are reported in Table 1.*

Suppose we are at the end of a 90-minute slot t. Once observations through slot t are available, that slot has already ended, so identifying its best arm is retrospective and cannot guide an action within the same slot. The earliest actionable decision is therefore a recommendation for slot $t+1$. Accordingly, we place every BAI method in the same follow-the-leader (FTL) wrapper: it identifies the best pickup zone at t from the noisy observations defined above and recommends that zone for $t+1$. For each target slot, a method may request at most B observations in total.

The best zone at $t+1$ may differ from the current best zone in highly non-stationary environments. If such switches occur frequently, no algorithm can reliably predict the next best zone in a minimax sense (Besbes et al., 2015). Indeed, in the NYC trip data, the current best zone persists to the next slot approximately 65% of the time. We emphasize that the purpose of this experiment is not to construct the optimal next-slot forecaster, but to compare the downstream performance of different BAI methods under the same actionable wrapper on a real-world dataset. Since the resulting performance metrics combine identification error at t with changes in the arm ordering between t and $t+1$, we separately report a leakage-safe, current-slot BAI evaluation in Appendix K.

Our setting differs from the usual forecasting setting in two respects: (a) each target-slot decision is subject to a fixed observation budget, and (b) we seek to control the pointwise misidentification probability and pointwise simple regret, rather than cumulative forecasting regret.

We fix four pickup zones—Midtown Center, Upper East Side North, Upper East Side South, and Midtown East—as arms. The morning 90-minute slots starting at 09:00, 10:30, and 12:00 are used as target slots for which we seek to recommend the best pickup zone. We consider weekdays from September through December 2024, leading to a total of 255 target slots. The fixed pickup zones and time slots are decided via a chronological design-validation split as follows. January–April data are used to select the four zones with the largest mean pickup shares. May–August data are used to select the slot duration such that sufficiently many trips are available for stable reward estimates while

retaining meaningful temporal variation. Thus, neither choice uses outcomes from the temporally subsequent September–December test period where the results are reported.

Variants of Algorithm 2 In the real-world data, we find that the threshold $q_i(\delta)$ used in Line 3 of Algorithm 1 is often too conservative leading to unresolved duel outcomes (ie $W = 1/2$). To mitigate this issue, we design a Bernstein based threshold that takes into account the empirical variance of the observed rewards (see Appendix E). Variant of Algorithm 2 that uses this threshold in the sub-routine `Multi-Resolution-Duel` is labeled as `SignCount-SH (Bernstein)`, while the original unmodified algorithm is termed as `SignCount-SH (Hoeffding)`. We remark that establishing the near-minimax optimality of `SignCount-SH (Hoeffding)` is the main scope of this paper. Analogous simple-regret guarantees for `SignCount-SH (Bernstein)` can be obtained by adapting our analysis using Bernstein concentration (Appendix F).

Table 1: *Mean probability of recommending best arm at next timeslot with 95% confidence intervals.*

Method	$B/K = 2$	$B/K = 4$	$B/K = 6$	$B/K = 8$	$B/K = 16$
SignCount-SH (Bernstein)	0.500 [0.434, 0.566]	0.399 [0.333, 0.466]	**0.522** [**0.461**, **0.582**]	**0.558** [**0.508**, **0.608**]	**0.573** [**0.532**, **0.615**]
SignCount-SH (Hoeffding)	0.499 [0.433, 0.565]	0.261 [0.203, 0.318]	0.416 [0.373, 0.460]	0.478 [0.428, 0.528]	0.441 [0.387, 0.494]
Full-prefix SH	0.504 [0.442, 0.567]	0.433 [0.360, 0.506]	0.393 [0.330, 0.457]	0.493 [0.449, 0.536]	0.406 [0.347, 0.465]
Vanilla-SH	0.499 [0.433, 0.565]	0.261 [0.203, 0.318]	0.416 [0.373, 0.460]	0.478 [0.428, 0.528]	0.441 [0.387, 0.494]
Successive Rejects	**0.556** [**0.502**, **0.610**]	0.489 [0.417, 0.562]	0.331 [0.264, 0.397]	0.357 [0.292, 0.421]	0.465 [0.411, 0.519]
Adjacent-BAI	0.552 [0.495, 0.609]	**0.491** [**0.421**, **0.561**]	0.382 [0.307, 0.456]	0.288 [0.230, 0.346]	0.346 [0.288, 0.404]
P1-RAGE	0.451 [0.423, 0.480]	0.412 [0.369, 0.456]	0.292 [0.262, 0.321]	0.265 [0.248, 0.281]	0.345 [0.322, 0.368]

Results and Interpretation. We plot the mean identification probability in Fig.4 (results on regret is deferred to Appendix H due to space constraints), averaged over the target slots described above. The dashed line in Fig.4 denotes the average performance of an oracle-FTL that recommends the current slot's population-level best arm (approximated by using all the trip data available at the slot) to the next slot. We see that `SignCount-SH (Bernstein)` attains superior performance especially with larger budgets where more data can be misleading due to drift. Further its performance is comparable to the oracle-FTL baseline. In contrast, `SignCount-SH (Hoeffding)` didn't resolve any signtests in the multi-resolution duel due to its conservative threshold, falling back to the performance of the vanilla-SH baseline via the tie breaking rule. We see that stationary baselines attain suboptimal performance especially under larger budgets, due to the bias caused by data from older history. We also report the mean fraction of history used by Algorithm 1 in Appendix G, demonstrating its adaptive usage of historical data when resolving the signs. We use a subsample size of $\min\{4000, N_s/2\}$ trips, where N_s is the total number of trips available at a timeslot s, to compute the observed reward at that slot for a specific arm. Reducing the sample size can increase the noise in the observed rewards making the identification task more challenging. Ablations on the sample-size are reported in Appendix I. We see that the performance of all algorithms deteriorates with reduced sample size while `SignCount-SH (Bernstein)` exhibits a more graceful deterioration.

We note that performance of any algorithm is strongly dependent on how challenging the underlying non-stationarity of rewards and their mutual gaps are. To test this, we report ablations on an evening timespan in Appendix J where the persistence of best arm across time slots is weaker and the average gap between arms are smaller in comparison to the morning timespan. The relative gains of the proposed methods over the stationary baselines are diminished in this more challenging environment.

5 Conclusion

We presented a novel framework for fixed-budget best-arm identification in non-stationary environments, where the objective is to identify the best arm at a prescribed target time. We proposed `SignCount-SH`, which requires no apriori knowledge of the drift, and established a confidence-dependent simple-regret guarantee. We also proved that, over instances with confidence-dependent complexity at most h, any algorithm returns an ϵ-optimal arm with probability at least $1 - \delta$ requires a budget scaling linearly with h up to logarithmic factors, matching the sufficient budget scaling of `SignCount-SH`. Synthetic and real-data experiments demonstrate both the benefits of adapting the amount of historical data used and the regimes in which the target-time ordering cannot be reliably resolved.

Developing sharper adaptive instance-dependent algorithms, proving guarantees that hold simultaneously for all confidence levels without fixing δ in advance, and extensions to structured arms and broader real-world settings are promising directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools for polishing some parts of the writing, assisting with the lower bound proof, developing the code for experiments and finding related work. LLM-generated materials were verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-Yadkori, Peter Bartlett, Victor Gabillon, Alan Malek, and Michal Valko. Best of both worlds: Stochastic and adversarial best-arm identification. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 918–949. PMLR, 2018.

Ayya Alieva, Ashok Cutkosky, and Abhimanyu Das. Robust pure exploration in linear bandits with limited budget. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 187–195. PMLR, 2021.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 126–135. PMLR, 2017.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pp. 41–53, 2010.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019.

MohammadJavad Azizi, Branislav Kveton, and Mohammad Ghavamzadeh. Fixed-budget best-arm identification in structured bandits. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2798–2804. International Joint Conferences on Artificial Intelligence Organization, 2022.

Dheeraj Baby, Yifei Tang, Hieu Duy Nguyen, Yu-Xiang Wang, and Rohit Pyati. Adaptive estimation and learning under temporal distribution shift. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2176–2202. PMLR, 2025.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, volume 5809 of *Lecture Notes in Computer Science*, pp. 23–37. Springer, 2009.

Romain Camilleri, Kevin Jamieson, and Julian Katz-Samuels. High-dimensional experimental design and kernel bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1227–1237. PMLR, 2021a.

Romain Camilleri, Zhihan Xiong, Maryam Fazel, Lalit Jain, and Kevin G. Jamieson. Selective sampling for online best-arm identification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11071–11082, 2021b.

Leonardo Cella, Massimiliano Pontil, and Claudio Gentile. Best model identification: A rested bandit formulation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1362–1372. PMLR, 2021.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1843–1854. PMLR, 2020.

Rémy Degenne, Pierre Menard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2432–2442. PMLR, 2020.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Computational Learning Theory*, volume 2375 of *Lecture Notes in Computer Science*, pp. 255–270. Springer, 2002.

Tanner Fiez, Lalit Jain, Kevin G. Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*, pp. 174–188. Springer, 2011.

Elise Han, Chengpiao Huang, and Kaizheng Wang. Model assessment and selection under temporal distribution shift. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Matthew Hoffman, Bobak Shahriari, and Nando de Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 365–374. PMLR, 2014.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 240–248. PMLR, 2016.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10007–10017, 2020.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Julian Katz-Samuels, Lalit Jain, Zohar Karnin, and Kevin G. Jamieson. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10371–10382, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Bingde Liu and John P. Cartlidge. Nonstationary continuum-armed bandit strategies for automated trading in a simulated financial market. In *Proceedings of the 13th International Defence and Homeland Security Simulation Workshop*. Caltek, 2023.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic MDPs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7447–7458. PMLR, 2021.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Leo Maynard-Zhang, Zhihan Xiong, Kevin Jamieson, and Maryam Fazel. On the complexity of best-arm identification in non-stationary linear bandits. In *Proceedings of the 39th Annual Conference on Learning Theory (COLT)*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5021–5052. PMLR, 2026.

Alessio Mazzetto and Eli Upfal. An adaptive algorithm for learning with unknown distribution drift. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

Marco Mussi, Alessandro Montenegro, Francesco Trovò, Marcello Restelli, and Alberto Maria Metelli. Best arm identification for stochastic rising bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36953–36989. PMLR, 2024.

New York City Taxi and Limousine Commission. TLC Trip Record Data. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page, 2024. Yellow taxi trip records, accessed August 31, 2026.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Chao Tao, Saúl Blanco, and Yuan Zhou. Best arm identification in linear bandits with linear dimension dependency. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4877–4886. PMLR, 2018.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4300–4354. PMLR, 2021.

Zhihan Xiong, Romain Camilleri, Maryam Fazel, Lalit Jain, and Kevin Jamieson. A/B testing and best-arm identification for linear bandits with robustness to non-stationarity. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2024.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 843–851. PMLR, 2018.

Junwen Yang and Vincent Y. F. Tan. Minimax optimal fixed-budget best arm identification in linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yao Zhao, Connor Stephens, Csaba Szepesvari, and Kwang-Sung Jun. Revisiting simple regret: Fast rates for returning a good arm. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 42110–42158, 23–29 Jul 2023.

A FURTHER MOTIVATING EXAMPLES

In addition to the example of biomedical study mentioned in the introduction, we highlight few more scenarios where our proposed setting is meaningful. Our framework arises when the expected performance of an arm changes over time, but the scientific or operational objective concerns its performance at a particular decision time. The framework admits two complementary interpretations. In a retrospective setting, the candidate arms have already generated historical artifacts, such as biological samples or manufactured units, but their rewards remain unknown until a costly measurement is performed. The learner begins with observations closest to the decision time and moves backward into history only when additional information is needed. In a prospective setting, the target is performance at a fixed initial time and the budget is instead consumed by deploying candidate arms. Clearly inferior arms can then be discontinued, while only promising arms continue to receive resources. In either interpretation, observations farther from the target time may reduce statistical uncertainty but can introduce bias because they are generated under different conditions. The learner must therefore balance temporal relevance against variance reduction while allocating a fixed budget adaptively.

Prospective selection of a treatment protocol for a recurring seasonal contamination event. Consider a reservoir that experiences a similar contamination pulse near the start of each spring runoff season. During winter, animal waste, decaying organic matter, and other contaminants accumulate in the surrounding watershed. The first sustained snowmelt or heavy spring rainfall then washes a mixture of microorganisms, suspended sediment, and organic matter into the reservoir. Because the watershed, land use, and seasonal trigger are broadly similar from year to year, the initial contamination state is approximately recurrent, although not identical. The facility uses the current year's event to identify which of K fixed-duration treatment protocols should be deployed rapidly when a comparable event begins the following year.

The facility cannot test all protocols with arbitrarily many simultaneous replications. Each experiment consumes contaminated water, treatment chemicals, reactor capacity, technician time, and measurement resources. It therefore evaluates all protocols in an initial batch, eliminates clearly inferior ones, and uses the remaining budget to test only the surviving protocols in later batches.

The setting is non-stationary because the untreated water collected this year evolves while the batches are conducted. Microbial concentrations and species proportions may change, cells may enter different physiological states, suspended particles may settle, and nutrients or dissolved oxygen may be depleted. Later tests are therefore not exact replications of tests started at the beginning of the seasonal event. They are still useful when this drift is gradual, but they provide increasingly biased information about initial-time effectiveness. Observed outcomes are also noisy because water portions are not perfectly homogeneous and because treatment response, finite-volume sampling, detection limits, and laboratory measurements all introduce variability.

A stationary best-arm identification method would treat all observations as samples from a fixed reward distribution. With equal sampling across times, it would estimate each protocol's average effectiveness over the sampled portion of the evolving event, rather than its effectiveness at the initial state. Moreover, because successive elimination algorithms for BAI allocates many of the remaining tests to surviving protocols in later batches, their estimates may be weighted even more heavily toward later conditions. The resulting ranking can therefore differ from the ranking at event onset. The non-stationary BAI formulation instead uses later tests to reduce variance while controlling for the bias caused by temporal drift.

Automated trading-policy selection. Another retrospective application arises in automated trading. Consider a system with K approved short-horizon policies, each specifying a different rule for when and how aggressively to trade. Their relative performance may change as the level of market activity and the difficulty of completing trades evolve. At a recalibration time, the system must select one policy for the next short operating period. Recent historical intervals—short segments of recorded prices and completed trades—are available, but evaluating a candidate policy on one such interval may require a detailed simulation of the actions it would have taken and the resulting profit or loss. This is fundamentally an off-policy evaluation problem as the interaction of a policy with the market environment can change some states of the market while the logged data in the interval is produced by some roll-in policy. Since these evaluations can be computationally expensive, only a limited number can be performed.

The system may therefore evaluate all policies on the most recent intervals, eliminate clearly inferior candidates, and use progressively older intervals to refine the estimates of the surviving policies. The selected policy is deployed for a short period during which the current trading conditions are expected to remain sufficiently similar, after which the procedure is repeated. Prior work has shown that adapting automated trading strategies under changing market conditions can be formulated as a non-stationary bandit problem and has proposed evaluating alternative strategy parameters in a simulator driven by recent market data (Liu & Cartlidge, 2023). Our formulation differs by considering a finite candidate set, a fixed evaluation budget, temporally ordered historical observations, and a terminal best-policy identification objective.

B Related Work

The distinction between our setting and prior work is easiest to state along three axes: the quantity to be identified, the way observations are revealed, and the assumed temporal structure. We seek the arm with the largest mean at a fixed target time. Pulling an arm reveals its next target-indexed observation, and all arms compete for one fixed pull budget. The mean trajectories away from the target may be non-monotone and may reverse their relative ordering. These features separate our problem from several closely related literatures.

Stationary best-arm identification. Classical stochastic BAI assumes that every pull of arm a is an independent sample from one fixed distribution with mean μ_a. Early work established fixed-confidence procedures and sample-complexity lower bounds (Even-Dar et al., 2002; Mannor & Tsitsiklis, 2004). The complementary fixed-budget formulation, initiated by Bubeck et al. (2009), allocates a prescribed number of pulls and minimizes the probability of recommending a suboptimal arm. Successive Rejects (Audibert et al., 2010) and Sequential Halving (Karnin et al., 2013; Zhao et al., 2023) are canonical elimination procedures for this problem; instance-dependent lower bounds and complexity characterizations are developed, for example, by Kaufmann et al. (2016). Our allocation rule inherits the halving structure of Sequential Halving, but the statistical comparison is different. Averaging all observations of an arm is valid in the stationary model because every observation has the same mean. Here observations farther from the target can carry a different, or even opposite, pairwise ordering. Consequently, classical inverse-gap complexities are replaced by resolving and ranking depths that depend on the complete history of pairwise mean gaps.

Stationary linear and structured BAI. When arms share a known feature representation, linear BAI can exploit that geometry through optimal experimental design. The fixed-confidence literature develops static and adaptive allocations, transductive designs, and instance-optimal procedures (Soare et al., 2014; Tao et al., 2018; Xu et al., 2018; Fiez et al., 2019; Degenne et al., 2020; Jedra & Proutiere, 2020). Closely related work studies Gaussian width, high-dimensional, and kernel-aware design criteria (Katz-Samuels et al., 2020; Camilleri et al., 2021a; Allen-Zhu et al., 2017). Fixed-budget variants and related correlated-arm selection methods use, among other tools, Bayesian models, XY-designs, and G-optimal allocations (Hoffman et al., 2014; Alieva et al., 2021; Azizi et al., 2022; Yang & Tan, 2022). These works are relevant to how a sampling budget should be distributed across statistically coupled arms. They nevertheless assume that one stationary parameter determines all sampled rewards. Our difficulty is orthogonal: even in the unstructured multi-armed specialization, successive pulls of an arm come from different positions in a prespecified temporal stream, and the desired arm is defined by its mean at the target position rather than by a stationary parameter.

Non-stationary linear BAI with a horizon-average target. Robust BAI with arbitrary, potentially adversarial reward sequences was studied for multi-armed bandits by Abbasi-Yadkori et al. (2018). The closest fixed-budget non-stationary linear BAI formulations are those of Xiong et al. (2024) and Maynard-Zhang et al. (2026). These methods allow an unpredictable sequence of parameters $\theta_1, \ldots, \theta_B$. Let $\mathcal{X} \subset \mathbb{R}^d$ denote the arm set, where $x \in \mathcal{X}$ is an arm's feature vector. At round t, arm x yields $Y_t(x) = x^\top \theta_t + \eta_t$, where $\mathbb{E}[\eta_t \mid \theta_{1:B}] = 0$. Consequently, these methods seek

$$
\begin{aligned}
x^{\mathrm{avg}} &\in \operatorname*{argmax}_{x \in \mathcal{X}} \frac{1}{B} \sum_{t=1}^{B} \mathbb{E}[Y_t(x) \mid \theta_{1:B}] \\
&= \operatorname*{argmax}_{x \in \mathcal{X}} x^\top \left(\frac{1}{B} \sum_{t=1}^{B} \theta_t \right),
\end{aligned}
$$

the arm with the largest expected reward averaged over the interaction horizon. P1-RAGE (Xiong et al., 2024) combines robust exploration with adaptive linear designs, while Adjacent-BAI (Maynard-Zhang et al., 2026) characterizes how the geometry of a linear arm set can improve the error exponent. Our objective is different: the desired arm maximizes $\mu_a(1)$ at one prespecified target time, not its mean over the interaction horizon. Moreover, our temporal index is arm-specific and ordered by distance from the target: allocating another pull to arm a reveals the next observation in that arm's stream. Thus an allocation changes how far into each arm's history the learner looks. A method that consistently estimates a horizon-average reward can therefore be inconsistent for the target-time ordering. We include both methods as experimental baselines precisely because they are robust non-stationary BAI procedures with a different estimand.

Target-time selection under temporal distribution shift. Related statistical-learning work adapts how much recent history is used to estimate the current distribution's mean when the drift magnitude is unknown (Mazzetto & Upfal, 2023), while Baby et al. (2025) obtain pointwise final-time estimation guarantees using a wavelet denoising estimator. Han et al. (2024) study model assessment and selection for the current data distribution using samples from the current and historical epochs. Together, these methods address the same fundamental bias–variance tension: incorporating more history reduces sampling noise but can increase temporal mismatch. Their feedback models, however, differ from ours. The estimation works observe one temporal data stream, and in model selection the loss of every fixed candidate can be evaluated on a labeled example. Thus historical data provide full-information estimates or comparisons, with no shared bandit budget to allocate among candidates. In our problem, an arm's reward is revealed only when that arm is pulled. The learner must therefore solve both an exploration problem—which arms should receive the remaining pulls—and an estimation problem—which temporal resolution should be used to compare the surviving arms. Our sign-count construction addresses these two decisions jointly and its complexity accounts for arm elimination as well as temporal resolution.

Resource-dependent and rested-arm identification. Non-stochastic BAI for hyperparameter optimization also exposes an ordered sequence of losses as additional resources are assigned to a candidate (Jamieson & Talwalkar, 2016); Hyperband combines this idea with successive halving across resource levels (Li et al., 2018). Related stochastic rested-bandit models treat arms as learning processes whose expected performance changes with their own number of pulls (Cella et al., 2021; Mussi et al., 2024). These formulations are similar in that observations from one arm need not be identically distributed and elimination can save resources. Their targets and structural assumptions are nevertheless different. They typically identify the best limiting model or the best arm after training, and exploit deterministic convergence envelopes or monotone, concave improvement with pull count. In our setting the target mean $\mu_a(1)$ is fixed before sampling, a pull only reveals another pre-existing target-indexed observation, and the trajectory away from the target need not be monotone. In particular, multiple historical ranking reversals are allowed as long as a suitable target-aligned resolving scale exists.

Online non-stationary bandits. Another large body of literature studies online reward maximization when the best arm changes during interaction. In the multi-armed setting, this literature includes discounted and sliding-window methods for switching rewards (Garivier & Moulines, 2011), variation-budget methods (Besbes et al., 2014), and ADSWITCH, which achieves near-optimal dynamic regret without knowing the number of abrupt distribution changes (Auer et al., 2019). Related parameter-free adaptation extends to contextual bandits (Chen et al., 2019). The MASTER reduction of Wei & Luo (2021) also specializes directly to multi-armed bandits by using UCB1 as its base learner; it adapts without prior knowledge to either the number or the total magnitude of changes. Beyond bandits, analogous ideas appear in non-stationary reinforcement learning through sliding-window optimism and restarted model-free algorithms (Cheung et al., 2020; Mao et al., 2021). The performance criterion in these lines of work is cumulative or dynamic regret: rewards collected while learning are part of the objective, and the learner must repeatedly track good actions or policies. Our problem is pure exploration. Intermediate rewards have value only as information, and performance is assessed by one recommendation for a fixed target time. This change in objective is consequential: older data are often discarded in online tracking, whereas in target-time BAI they can be essential for distinguishing close arms when their historical pairwise ordering remains aligned with the target. The main question is therefore not how rapidly to forget the past, but how much target-aligned history is statistically useful for each comparison under a fixed global budget. Also related, but tangential, is

selective sampling for online BAI, which decides when streaming observations should be queried (Camilleri et al., 2021b); it does not address temporal resolution for a prespecified target-time mean.

C Omitted Proofs

C.1 Upperbound Analysis of Simple Regret

For the sake of simplicity in presentation, we assume that K is a power of 2 and B is a multiple of $K \log_2 K$. Extension to the general case is straight-forward. We start by conditioning on a (fictitious) good event.

Lemma 10. *[Global Good Event $\mathcal{E}_\delta$] Let $\bar{\mu}_a(k_j) = 1/k_j \sum_{r=1}^{k_j} \mu_a(r)$. Consider the event $\mathcal{E}_\delta$ where $(|\hat{\mu}(k_j) - \bar{\mu}(k_j))| \leq q_j(\delta)/2$ for all arms a and scales j. This event obeys $P(\mathcal{E}_\delta) \geq 1 - \delta$. Further, under $\mathcal{E}_\delta$, it holds that*

$$|\hat{\mu}_a(k_j) - \hat{\mu}_b(k_j) - (\bar{\mu}_a(k_j) - \bar{\mu}_b(k_j))| \leq q_j(\delta), \forall a, b, j$$

Remark 11. *We note that the event $\mathcal{E}_\delta$ is a fictitious event as in the observations needed to compute quantities such as $\hat{\mu}_a(k_j)$ may not even be requested by the best arm identification algorithm. The only purpose of this good event is to condition the randomness in observations so that the analysis of algorithm can be done through deterministic arguments thereof.*

Proof. Consider a fixed arm a and fixed dyadic scale j. Let $\bar{\mu}_a(k_j) = 1/k_j \sum_{r=1}^{k_j} \mu_a(r)$. We have from sub-gaussian concentration that

$$P\left(|\hat{\mu}_a(k_j) - \bar{\mu}_a(k_j))| \geq q_j(\delta)/2\right) \leq \frac{6\delta}{\pi^2 K(j+1)^2}$$

Now taking union bound over all arms a and scales j yields that $P(\mathcal{E}_\delta) \geq 1 - \delta$. The second statment of the lemma is immediate from applying triangle inequality.

□

Lemma 2. *Consider a time $t \geq \tau_{a,b}^{(\delta)}$ and any pair of arms (a, b). With probability at-least $1 - \delta$, `Multi-Resolution-Duel` (Algorithm 1) would output $W_{a,b} = s_{a,b}$.*

Proof. We take the case $\mu_a(1) > \mu_b(1)$. The alternate is similar.

Let $k_j = \lfloor \log_2 t \rfloor$.Consider an earlier scale $i < j$. We have

$$\begin{aligned} \hat{d}_{a,b}(k_i) &= \Gamma_{a,b}(i) + \hat{d}_{a,b}(k_i) - \bar{d}_{a,b}(k_i) \\ &> -2q_i(\delta), \end{aligned}$$

with probability at-least $1 - \delta$ due to Lemma 10. Thus if `Multi-Resolution-Duel` stops at scale i, then it must have $\hat{d}_{a,b}(k_i) > 2q_i(\delta)$ and would output $W_{a,b} = 1$ correctly.

On the other hand, suppose `Multi-Resolution-Duel` stops at scale j. Then we must have

$$\begin{aligned} \hat{d}_{a,b}(k_j) &= \Gamma_{a,b}(j) + \hat{d}_{a,b}(k_j) - \bar{d}_{a,b}(k_j) \\ &> 2q_j(\delta), \end{aligned}$$

with probability at-least $1 - \delta$. This would again result in $W_{a,b} = 1$ which is the correct sign direction. □

Lemma 12 (Ranking Certificate)**.** *Suppose we are the end of an epoch ℓ in Algorithm 2. Consider the case where a good arm $g \in \mathcal{G}_\epsilon$ and a bad arm $a \in \mathcal{B}_\epsilon$ are both in the active set A_ℓ. If $T_\ell \geq \rho_{g,a}^{(\delta)}$, then $C_g^\ell > C_a^\ell$ with probability at-least $1 - \delta$.*

Proof. We have

$$\begin{aligned} C_g^\ell - C_a^\ell &= (W_{g,a}^\ell - W_{a,g}^\ell) + \sum_{b \in A^\ell \setminus \{g,a\}} (W_{g,b}^\ell - W_{a,b}^\ell) \\ &= 1 + \sum_{b \in A^\ell \setminus \{g,a\}} (W_{g,b}^\ell - W_{a,b}^\ell), \end{aligned}$$

with probability at-least $1-\delta$, due to Lemma 2 and by the fact that $T_\ell \geq \rho^{(\delta)}_{g,a}$. Since $T_\ell \geq \chi^{(\delta)}_{g,a;b}$, for any third arm b, then either one of the following hold empirically with probability at-least $1-\delta$ due to Lemma 2. (i) $W^\ell_{g,b} = 1$ or (ii) $W^\ell_{b,a} = 1$. In either case we can see that $W^\ell_{g,b} - W^\ell_{a,b} \geq 0$.

Thus we conclude that $C^\ell_g > C^\ell_a$ with high probability.

□

Theorem 5. *Suppose the budget $B \geq 4\log_2 K \cdot H^{(\delta)}_{NS}(\theta,\epsilon)$. Then with probability at-least $1-\delta$, Algorithm 2 terminates with an arm $\hat{i}$ that is ϵ-optimal. i.e, $\Delta_{\hat{i}} = \mu_1(1) - \mu_{\hat{i}}(1) \leq \epsilon$.*

Proof. Let $g^* \in \mathcal{G}_\epsilon$ be the good arm that attains the minimum in Eq.(1). Consider an epoch ℓ where $|A_\ell| > 2|\mathcal{G}_\epsilon|$. Throughout the proof we condition on the good event $\mathcal{E}_\delta$ which happens with probability at-least $1-\delta$ by Lemma 10.

Let $g^* \in A_\ell$. For the sake of contradiction, suppose that $g^* \notin A_{\ell+1}$. Then the $|A_\ell|/2$ arms selected into $A_{\ell+1}$ must each have sign-count score at least $C^\ell_{g^*}$, where equality may occur due to tie-breaking. Among them, at most $|\mathcal{G}_\epsilon|-1$ can be good arms. Hence $A_{\ell+1}$ contains at least

$$m = \frac{|A_\ell|}{2} - |\mathcal{G}_\epsilon| + 1$$

arms from $\mathcal{B}_\epsilon$ whose sign-count scores are at least $C^\ell_{g^*}$.

From Eq.(1), we have

$$H^{(\delta)}_{\mathrm{NS}}(\theta,\epsilon) \geq (|\mathcal{G}_\epsilon| + m)\rho^{(\delta)}_{g^*,[m]}.$$

Consequently

$$\begin{aligned}
\rho^{(\delta)}_{g^*,[m]} &\leq \frac{H^{(\delta)}_{\mathrm{NS}}(\theta,\epsilon)}{|A_\ell|/2+1} \\
&\leq \frac{2H^{(\delta)}_{\mathrm{NS}}(\theta,\epsilon)}{|A_\ell|} \\
&\leq \frac{2\log_2 K \cdot H^{(\delta)}_{\mathrm{NS}}(\theta,\epsilon)\cdot T_\ell}{B} \\
&\leq T_\ell.
\end{aligned}$$

Hence by Lemma 12, at-most $m-1$ bad arms can outrank g^* in terms of the sign counts. This is a contradiction. Applying the argument inductively yields that if $|A_\ell| > 2|\mathcal{G}_\epsilon|$, then $g^* \in A_{\ell+1}$.

Now consider an epoch where $|A_\ell| \leq 2|\mathcal{G}_\epsilon|$. We have

$$\begin{aligned}
\rho^{(\delta)}_{g^*,[1]} &\leq \frac{H^{(\delta)}_{\mathrm{NS}}(\theta,\epsilon)}{|\mathcal{G}_\epsilon|+1} \\
&\leq \frac{2H^{(\delta)}_{\mathrm{NS}}(\theta,\epsilon)}{|A_\ell|} \\
&\leq \frac{4\log_2 K \cdot H^{(\delta)}_{\mathrm{NS}}(\theta,\epsilon)\cdot T_\ell}{B} \\
&\leq T_\ell.
\end{aligned}$$

This relation when combined with Lemma 12 implies that $C^\ell_{g^*} > C^\ell_a$ for every $a \in \mathcal{B}_\epsilon \cap A^\ell$, regardless of how ties are broken. Thus no bad arm can remain unresolved against g^* at depth T_ℓ. Then there can be two scenarios. Either $g^* \in A_{\ell+1}$ in which case the inductive argument moves forward. In the scenario where $g^* \notin A_{\ell+1}$, we can say that no arms in $\mathcal{B}_\epsilon$ must be present in $A_\ell + 1$. In this situation, $A_{\ell+1} \subseteq \mathcal{G}_\epsilon$ at which point the algorithm will eventually terminate with an arm that is ϵ-good. □

Next, we prove Corollary 7. We restate it in a more convenient form.

Lemma 13 (Fixed-confidence stationary recovery)**.** *Suppose that the mean trajectories are stationary, i.e* $\mu_a(s) = \mu_a, \forall a \in [K],\ s \geq 1$*, and assume that the best arm is unique. Without loss of generality, let arm* 1 *be optimal and* $0 = \Delta_1 < \Delta_2 \leq \ldots \leq \Delta_K$*. Let* $H_2(\theta) = \max_{2 \leq i \leq K} \frac{i}{\Delta_i^2}$*.*

Consider exact best-arm identification, i.e., $\epsilon = 0$*. There exists a universal constant* $C > 0$ *such that, for every* $\delta \in (0, 1)$*,*

$$B \geq C \log K \left(K + \sigma^2 H_2(\theta) \left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\Delta_2^2}\right) \right] \right) \tag{4}$$

implies

$$\Pr_\theta(\widehat{a} \neq 1) \leq \delta. \tag{5}$$

Proof. Under stationarity, for every pair a, b and every dyadic window $k_j = 2^j$,

$$\overline{d}_{a,b}(k_j) = \mu_a - \mu_b.$$

Hence the signed population prefix signal is independent of the window size:

$$\begin{aligned} \Gamma_{a,b}(j) &= s_{a,b}\overline{d}_{a,b}(k_j) \\ &= |\mu_a - \mu_b|. \end{aligned}$$

In particular, the condition preventing a statistically significant wrong-sign decision at earlier scales is automatically satisfied.

Let

$$\Delta_{a,b} := |\mu_a - \mu_b|.$$

The pair is therefore resolved once

$$\Delta_{a,b} > 3q_j(\delta), \tag{6}$$

where

$$q_j(\delta) = 2\sigma\sqrt{\frac{2}{k_j}\log\left(\frac{\pi^2 K(j+1)^2}{3\delta}\right)}. \tag{7}$$

Combining (6) and (7), a pair (a, b) is resolved at scale $k_j = 2^j$ whenever

$$k_j > 72\frac{\sigma^2}{\Delta_{a,b}^2}\left[\log\left(\frac{\pi^2 K}{3\delta}\right) + 2\log(j+1)\right]. \tag{8}$$

We now exhibit a dyadic scale satisfying (8).

For this argument only, write

$$x := \frac{\sigma^2}{\Delta_{a,b}^2}, \qquad L := \log\left(\frac{eK}{\delta}\right) + \log\log(e^e + x).$$

Let $j^\star$ be the first dyadic index such that

$$k_{j^\star} \geq \max\{1, 512xL\}.$$

We first consider the nontrivial case $512xL \geq 1$. By the choice of $j^\star$ and dyadic rounding,

$$512xL \leq k_{j^\star} < 1024xL. \tag{9}$$

Since

$$\begin{aligned} \log\left(\frac{\pi^2 K}{3\delta}\right) &\leq 2\log\left(\frac{eK}{\delta}\right) \\ &\leq 2L, \end{aligned} \tag{10}$$

it remains to control $\log(j^\star + 1)$.

Since $k_{j^\star} = 2^{j^\star}$,

$$j^\star + 1 \le \frac{\log(ek_{j^\star})}{\log 2}.$$

Therefore,

$$\log(j^\star + 1) \le \log\left(\frac{1}{\log 2}\right) + \log\log(ek_{j^\star}). \tag{11}$$

Using (9),

$$\begin{aligned}\log(ek_{j^\star}) &< \log(1024exL) \\ &= \log(1024e) + \log x + \log L. \end{aligned} \tag{12}$$

By the definition of L,

$$\begin{aligned}\log x &\le \log(e^e + x) \\ &= \exp(\log\log(e^e + x)) \\ &\le e^L.\end{aligned}$$

Moreover, $L \ge 2$, and hence

$$\log L \le L \le e^L.$$

Since $\log(1024e) < 1.1e^L$ for $L \ge 2$, (12) gives

$$\log(ek_{j^\star}) < 3.1e^L.$$

Taking logarithms once more,

$$\begin{aligned}\log\log(ek_{j^\star}) &< L + \log(3.1) \\ &< 1.6L,\end{aligned}$$

where the last inequality again uses $L \ge 2$. Substituting into (11) gives

$$\log(j^\star + 1) < 1.8L. \tag{13}$$

Combining (10) and (13),

$$\begin{aligned}72x\left[\log\left(\frac{\pi^2 K}{3\delta}\right) + 2\log(j^\star + 1)\right] &< 72x(2L + 3.6L) \\ &< 404xL \\ &< 512xL \\ &\le k_{j^\star}.\end{aligned}$$

Thus $k_{j^\star}$ satisfies (8), and the pair is resolved by this scale.

If $512xL < 1$, then $j^\star = 0$ and $k_{j^\star} = 1$. In this case,

$$\begin{aligned}72x\log\left(\frac{\pi^2 K}{3\delta}\right) &\le 144x\log\left(\frac{eK}{\delta}\right) \\ &\le 144xL \\ &< \frac{144}{512} < 1,\end{aligned}$$

so the pair is already resolved at the first scale.

We have therefore shown that

$$\tau_{a,b}^{(\delta)} \le 1024\left[1 \vee \left\{\frac{\sigma^2}{\Delta_{a,b}^2}\left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\Delta_{a,b}^2}\right)\right]\right\}\right]. \tag{14}$$

Using $1 \vee y \le 1 + y$, this gives, for a universal constant C,

$$\tau_{a,b}^{(\delta)} \le C \left(1 + \frac{\sigma^2}{\Delta_{a,b}^2} \left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\Delta_{a,b}^2}\right) \right] \right).$$

The leading constant term must be retained throughout: since dyadic scales satisfy $k_j \ge 1$, we have $\tau_{a,b}^{(\delta)} \ge 1$ however large $\Delta_{a,b}$ is, and after taking order statistics in Definition 4 this contributes the additive K appearing in (16). See Remark **??**.

We next bound the ranking depth of a suboptimal arm $a \ge 2$. Recall that

$$\rho_{1,a}^{(\delta)} = \max\left\{ \tau_{1,a}^{(\delta)}, \max_{b \ne 1,a} \chi_{1,a;b}^{(\delta)} \right\}.$$

For every third arm b, either $\mu_b \le \mu_a$, in which case

$$\mu_1 - \mu_b \ge \Delta_a,$$

or $\mu_b > \mu_a$, in which case

$$\Delta_a = (\mu_1 - \mu_b) + (\mu_b - \mu_a).$$

Therefore, in the latter case,

$$\max\{\mu_1 - \mu_b,\, \mu_b - \mu_a\} \ge \frac{\Delta_a}{2}.$$

Thus, for every third arm b, at least one of the pairwise comparisons entering $\chi_{1,a;b}^{(\delta)}$ has gap at least $\Delta_a/2$. Applying (14) to that comparison gives

$$\rho_{1,a}^{(\delta)} \le C \left(1 + \frac{\sigma^2}{\Delta_a^2} \left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\Delta_a^2}\right) \right] \right). \tag{15}$$

For exact best-arm identification,

$$\mathcal{G}_0 = \{1\}, \qquad M_0 := |\mathcal{G}_0| = 1.$$

Therefore,

$$H_{\mathrm{NS}}^{(\delta)}(\theta, 0) = \max_{1 \le m \le K-1} (m+1) \rho_{1,[m]}^{(\delta)}.$$

Since the gaps are ordered as

$$\Delta_2 \le \Delta_3 \le \cdots \le \Delta_K,$$

the m-th largest ranking depth $\rho_{1,[m]}^{(\delta)}$ is attained at the m-th smallest gap, namely Δ_{m+1}. Hence (15) yields

$$\begin{aligned} H_{\mathrm{NS}}^{(\delta)}(\theta, 0) &\le C \max_{1 \le m \le K-1} (m+1) \left(1 + \frac{\sigma^2}{\Delta_{m+1}^2} \left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\Delta_{m+1}^2}\right) \right] \right) \\ &\le C \max_{1 \le m \le K-1} (m+1) + C\sigma^2 \max_{2 \le i \le K} \frac{i}{\Delta_i^2} \left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\Delta_i^2}\right) \right] \\ &\le CK + C\sigma^2 H_2(\theta) \left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\Delta_2^2}\right) \right], \end{aligned} \tag{16}$$

where the second inequality splits the maximum of a sum into the sum of maxima, and the third uses $\max_{1 \le m \le K-1}(m+1) = K$ together with the monotonicity of $\log\log(e^e + \sigma^2/\Delta_i^2)$ in Δ_i. The additive CK is present even as $\sigma \to 0$, and it is dominated by the second term whenever $\Delta_K^2 \le \sigma^2 \log(eK/\delta)$.

Finally, the general upper bound for Sign-Count Sequential Halving from Theorem 5 states that

$$B \ge C H_{\mathrm{NS}}^{(\delta)}(\theta, 0) \log K \tag{17}$$

implies

$$\Pr_\theta(\widehat{a} \ne 1) \le \delta.$$

Combining (16) and (17) proves (4)–(5).

□

C.2 Lower bound Analysis

Intuition for the lower bound. The hard family is constructed so that identifying the terminally best arm requires accumulating evidence over many observations. Under instance θ_i, arm i has terminal mean 2ϵ, whereas all other arms have terminal mean zero. However, we restrict $\epsilon \lesssim \sigma$ to place ourselves in the statistically non-trivial regime. Consequently, the single terminal observation is not enough to reveal the best arm reliably: relative to the all-zero null instance, it contributes only

$$\frac{(2\epsilon)^2}{2\sigma^2} = \frac{2\epsilon^2}{\sigma^2}$$

units of KL information, which is bounded by a constant. After this terminal observation, the special arm differs from the null only through the smaller persistent mean η. Each additional observation of that arm contributes only

$$\frac{\eta^2}{2\sigma^2}$$

units of KL information. Thus, distinguishing the special arm with high confidence requires repeatedly sampling it and accumulating these small information gains.

To formalize this difficulty, consider first the null instance in which all arms have mean zero. An arbitrary algorithm need not treat the arms symmetrically: it may sample or recommend some arms much more often than others. Nevertheless, fewer than half of the arms can receive more than $2B/K$ expected samples, and fewer than half can be recommended with probability greater than $2/K$. Hence by pigeonhole principle, there exists an arm $i^\star$ that is neither heavily sampled nor heavily favored under the null:

$$\mathbb{E}_0^\pi[N_{i^\star}] \le \frac{2B}{K}, \qquad \mathbb{P}_0^\pi(\widehat{a} = i^\star) \le \frac{2}{K}.$$

We then consider the alternative instance $\theta_{i^\star}$ in which precisely this arm is the terminally best arm.

If the algorithm is to achieve failure probability at most δ on $\theta_{i^\star}$, its probability of recommending $i^\star$ must increase from at most $2/K$ under the null to at least $1-\delta$ under the alternative. Such a large change in behavior requires the two transcript distributions to have KL divergence of order at least $\log(1/\delta)$. On the other hand, because $i^\star$ receives at most $2B/K$ expected observations under the null, the total KL divergence is at most of order

$$\frac{\epsilon^2}{\sigma^2} + \frac{B\eta^2}{K\sigma^2}.$$

The first term is only a constant, while the second term is the information accumulated from repeatedly observing the weak persistent signal η. Therefore, if

$$B \ll K\frac{\sigma^2}{\eta^2}\log\left(\frac{1}{\delta}\right),$$

the observations do not contain enough information to change the algorithm's behavior from rarely selecting $i^\star$ under the null to selecting it with probability at least $1-\delta$ under the alternative. Hence some member of the hard family must have ϵ-simple-regret failure probability larger than δ.

Finally, the parameter η controls both this information-theoretic difficulty and the confidence-dependent ranking complexity $H_{\mathrm{NS}}^{(\delta)}$. By choosing η appropriately, the hard family can be placed at any admissible complexity level, allowing the information lower bound to be expressed directly in terms of the same $H_{\mathrm{NS}}^{(\delta)}$ that appears in the upper bound for Sign-Count Sequential Halving.

Finite-confidence minimax lower bound. We next establish a lower bound directly in terms of the same confidence-dependent complexity $H_{\mathrm{NS}}^{(\delta)}(\theta, \epsilon)$ that appears in the upper bound. Thus ϵ and δ are fixed in advance, and we ask how large the budget must be for any algorithm to guarantee

$$\mathbb{P}_\theta^\pi\left(r_B > \epsilon\right) \le \delta$$

uniformly over instances of a given $H_{\mathrm{NS}}^{(\delta)}$ complexity.

Since Gaussian noise with variance σ^2 is a subclass of the σ-sub-Gaussian model, it is sufficient for the lower bound to consider Gaussian observations.

Hard family. Assume $K \geq 8$, $\delta \in (0, 1/4]$, and

$$0 < \epsilon \leq \frac{\sigma}{8}.$$

Fix a parameter

$$0 < \eta \leq 2\epsilon.$$

For every $i \in [K]$, define an instance θ_i by

$$\mu_i(1) := 2\epsilon, \qquad \mu_i(s) := \eta, \qquad s \geq 2,$$
$$\mu_a(s) := 0, \qquad a \neq i,\ s \geq 1.$$

Thus arm i is the unique ϵ-good arm under θ_i. In particular,

$$\widehat{a} \neq i \quad \Longleftrightarrow \quad r_B = 2\epsilon > \epsilon. \tag{18}$$

Lemma 14 (Pairwise resolution depth of the hard family)**.** *Under the hard instance θ_i, for every $a \neq i$,*

$$\tau_{i,a}^{(\delta)} = \Theta\left(\frac{\sigma^2}{\eta^2}\left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\eta^2}\right)\right]\right). \tag{19}$$

The constants in the $\Theta(\cdot)$ notation are universal.

Proof. For every $a \neq i$ and every prefix length k,

$$\overline{d}_{i,a}(k) = \eta + \frac{2\epsilon - \eta}{k}.$$

This quantity is positive at every scale. Hence the condition in the definition of $\tau_{i,a}^{(\delta)}$ that prevents an earlier wrong-sign resolution is automatically satisfied. Therefore $\tau_{i,a}^{(\delta)}$ is the first dyadic scale $k_j = 2^j$ such that

$$\overline{d}_{i,a}(k_j) > 3q_j(\delta), \tag{20}$$

where

$$q_j(\delta) = 2\sigma\sqrt{\frac{2}{k_j}\log\left(\frac{\pi^2 K(j+1)^2}{3\delta}\right)}. \tag{21}$$

For this proof only, write

$$x := \frac{\sigma^2}{\eta^2}, \qquad L := \log\left(\frac{eK}{\delta}\right) + \log\log(e^e + x).$$

Since $\eta \leq 2\epsilon \leq \sigma/4$, we have

$$x \geq 16.$$

We first derive a necessary condition for resolution. From $K \geq 8$ and $\delta \leq 1/4$, the logarithm in (21) is larger than one, and hence

$$q_j(\delta) \geq \frac{\sigma}{\sqrt{k_j}} \geq \frac{\sigma}{k_j}.$$

Using $2\epsilon \leq \sigma/4$,

$$\frac{2\epsilon - \eta}{k_j} \leq \frac{\sigma}{4k_j} \leq \frac{q_j(\delta)}{4}.$$

Therefore, if (20) holds, then

$$3q_j(\delta) < \eta + \frac{2\epsilon - \eta}{k_j}$$
$$\leq \eta + \frac{q_j(\delta)}{4},$$

which implies

$$q_j(\delta) < \frac{4}{11}\eta. \tag{22}$$

By (21), condition (22) implies

$$k_j > \frac{121}{2} x \left[\log\left(\frac{\pi^2 K}{3\delta}\right) + 2\log(j+1) \right]. \tag{23}$$

Since

$$\log\left(\frac{\pi^2 K}{3\delta}\right) \geq \log\left(\frac{eK}{\delta}\right),$$

(23) implies $k_j > x$, and therefore

$$j + 1 \geq \log_2 x + 1.$$

For $x \geq 16$,

$$\begin{aligned} \log\log(e^e + x) &\leq \log\log(2x) \\ &\leq 2\log(\log_2 x + 1) \\ &\leq 2\log(j+1). \end{aligned}$$

Here the second inequality follows from the elementary relation

$$\log(2x) \leq (\log_2 x + 1)^2, \qquad x \geq 16.$$

Consequently, the bracket in (23) is at least L, and hence every resolving scale must satisfy

$$k_j > 60xL. \tag{24}$$

We next exhibit a scale at which resolution must occur. Let $j^\star$ be the first dyadic index such that

$$k_{j^\star} \geq 512xL.$$

By dyadic rounding,

$$k_{j^\star} < 1024xL.$$

We show that $q_{j^\star}(\delta) < \eta/3$. First,

$$\begin{aligned} \log\left(\frac{\pi^2 K}{3\delta}\right) &\leq \log\left(\frac{eK}{\delta}\right) + 1 \\ &\leq \frac{5}{4}L, \end{aligned} \tag{25}$$

where the final inequality uses $L \geq \log(32e) > 4$.

Moreover, since $k_{j^\star} < 1024xL$,

$$j^\star + 1 < \log_2(2048xL).$$

By the definition of L,

$$\log x \leq e^L, \qquad \log L \leq e^L.$$

Since $L > 4$, we also have $\log 2048 \leq e^L$. Thus

$$\log(2048xL) \leq 3e^L.$$

It follows that

$$\begin{aligned} \log(j^\star + 1) &< \log\log(2048xL) - \log\log 2 \\ &\leq L + \log 3 - \log\log 2 \\ &\leq \frac{4}{3}L. \end{aligned} \tag{26}$$

Combining (25) and (26),

$$\log\left(\frac{\pi^2 K}{3\delta}\right) + 2\log(j^\star + 1) < 4L.$$

Hence

$$\begin{aligned} 72x\left[\log\left(\frac{\pi^2 K}{3\delta}\right) + 2\log(j^\star + 1)\right] &< 288xL \\ &< k_{j^\star}. \end{aligned}$$

By (21), this is exactly sufficient to guarantee

$$q_{j^\star}(\delta) < \frac{\eta}{3}.$$

Since

$$\overline{d}_{i,a}(k_{j^\star}) \geq \eta > 3q_{j^\star}(\delta),$$

the pair is resolved by scale $k_{j^\star}$. Therefore

$$\tau_{i,a}^{(\delta)} < 1024xL. \tag{27}$$

Combining (24) and (27) proves (19). □

Lemma 15 (Confidence-dependent complexity of the hard family)**.** *For every hard instance θ_i,*

$$H_{\mathrm{NS}}^{(\delta)}(\theta_i, \epsilon) = \Theta\left(K\frac{\sigma^2}{\eta^2}\left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\eta^2}\right)\right]\right).$$

Proof. Under θ_i, arm i is the unique ϵ-good arm, so

$$M_\epsilon = 1.$$

By Lemma 14, every comparison between i and an arm $a \neq i$ has the same resolution depth, up to universal constants.

In fact, for every bad arm $a \neq i$ and every third arm $b \notin \{i, a\}$, we have

$$\mu_i(1) > \mu_b(1) = \mu_a(1).$$

Thus the comparison $i > b$ is the eligible comparison appearing in $\chi_{i,a;b}^{(\delta)}$, and hence

$$\chi_{i,a;b}^{(\delta)} = \tau_{i,b}^{(\delta)}.$$

All comparisons $i > b$ have the same population prefix trajectory. Therefore

$$\rho_{i,a}^{(\delta)} = \Theta\left(\frac{\sigma^2}{\eta^2}\left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\eta^2}\right)\right]\right), \qquad a \neq i.$$

All $K - 1$ bad arms therefore have the same ranking depth up to universal constants. Since $M_\epsilon = 1$,

$$\begin{aligned} H_{\mathrm{NS}}^{(\delta)}(\theta_i, \epsilon) &= \max_{1 \leq m \leq K-1} (m+1)\rho_{i,[m]}^{(\delta)} \\ &= \Theta\left(K\frac{\sigma^2}{\eta^2}\left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\eta^2}\right)\right]\right), \end{aligned}$$

as claimed. □

Lemma 16 (Binary separation required for δ-correctness)**.** *Let $\delta \in (0, 1/4]$. If*

$$p \leq \frac{1}{4}, \qquad\qquad q \geq 1 - \delta,$$

then

$$\mathrm{kl}(p, q) \geq \frac{1}{4}\log\left(\frac{1}{\delta}\right),$$

where kl *denotes the binary relative entropy.*

Proof. Since $q \geq 1 - \delta \geq 3/4 > p$, binary relative entropy is increasing in its second argument, and hence

$$\mathrm{kl}(p, q) \geq \mathrm{kl}(p, 1 - \delta).$$

Expanding,

$$\mathrm{kl}(p, 1 - \delta) = (1 - p) \log\left(\frac{1}{\delta}\right) - h(p) - p \log(1 - \delta),$$

where $h(p)$ is the binary entropy. Since $p \leq 1/4$, $h(p) \leq \log 2$, and $-p \log(1 - \delta) \geq 0$,

$$\mathrm{kl}(p, 1 - \delta) \geq \frac{3}{4} \log\left(\frac{1}{\delta}\right) - \log 2.$$

Finally, $\delta \leq 1/4$ implies $\log(1/\delta) \geq 2 \log 2$, and therefore

$$\mathrm{kl}(p, q) \geq \frac{1}{4} \log\left(\frac{1}{\delta}\right).$$

□

Lemma 17 (Finite-confidence information lower bound)**.** *For every possibly adaptive fixed-budget policy π, if*

$$B \leq \frac{1}{8} K \frac{\sigma^2}{\eta^2} \log\left(\frac{1}{\delta}\right), \tag{28}$$

then there exists an index $i \in [K]$ such that

$$\mathbb{P}_i^\pi \left(r_B > \epsilon\right) > \delta. \tag{29}$$

Here $\mathbb{P}_i^\pi$ denotes the probability law of the complete interaction transcript generated when policy π is run on θ_i.

Proof. Suppose, toward a contradiction, that

$$\mathbb{P}_i^\pi \left(r_B > \epsilon\right) \leq \delta, \qquad \forall i \in [K]. \tag{30}$$

By (18), this implies

$$\mathbb{P}_i^\pi \left(\widehat{a} = i\right) \geq 1 - \delta, \qquad \forall i \in [K]. \tag{31}$$

Introduce the null instance θ_0, under which every arm has mean zero at every time. Let $\mathbb{P}_0^\pi$ denote its transcript law, $\mathbb{E}_0^\pi$ the corresponding expectation, and N_i the number of observations obtained from arm i.

Since the policy uses at most B observations,

$$\sum_{i=1}^{K} \mathbb{E}_0^\pi[N_i] \leq B.$$

Also,

$$\sum_{i=1}^{K} \mathbb{P}_0^\pi(\widehat{a} = i) = 1.$$

Therefore there exists strictly fewer than $K/2$ arms where $E_0^\pi[N_i] \geq 2B/K$. Similarly there exists strictly fewer than $K/2$ arms where $\mathbb{P}_0^\pi(\widehat{a} = i) \geq 2/K$. Consequently, by pigeonhole principle, there exists an index $i^\star$ such that

$$\mathbb{E}_0^\pi[N_{i^\star}] \leq \frac{2B}{K}, \qquad \mathbb{P}_0^\pi(\widehat{a} = i^\star) \leq \frac{2}{K}. \tag{32}$$

Fix this index for the remainder of the proof.

The null instance and $\theta_{i^\star}$ differ only in observations from arm $i^\star$. The first observation from that arm has mean shift 2ϵ, whereas every subsequent observation has mean shift η. By the chain rule for KL divergence and the Gaussian KL formula,

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}_0^\pi \,\|\, \mathbb{P}_{i^\star}^\pi\right) &= \frac{1}{2\sigma^2}\mathbb{E}_0^\pi\left[(2\epsilon)^2\mathbf{1}\{N_{i^\star} \geq 1\} + \eta^2(N_{i^\star}-1)_+\right] \\ &\leq \frac{2\epsilon^2}{\sigma^2} + \frac{\eta^2}{2\sigma^2}\mathbb{E}_0^\pi[N_{i^\star}] \\ &\leq \frac{2\epsilon^2}{\sigma^2} + \frac{B\eta^2}{K\sigma^2}. \end{aligned} \tag{33}$$

Now consider the event

$$E := \{\widehat{a} = i^\star\}.$$

By (32) and (31),

$$\begin{aligned}\mathbb{P}_0^\pi(E) &\leq \frac{2}{K} \leq \frac{1}{4}, \\ \mathbb{P}_{i^\star}^\pi(E) &\geq 1-\delta.\end{aligned}$$

The data-processing inequality for KL divergence and Lemma 16 therefore imply

$$\mathrm{KL}\left(\mathbb{P}_0^\pi \,\|\, \mathbb{P}_{i^\star}^\pi\right) \geq \frac{1}{4}\log\left(\frac{1}{\delta}\right). \tag{34}$$

On the other hand, $\epsilon \leq \sigma/8$ implies

$$\frac{2\epsilon^2}{\sigma^2} \leq \frac{1}{32}.$$

Under (28), (33) gives

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}_0^\pi \,\|\, \mathbb{P}_{i^\star}^\pi\right) &\leq \frac{1}{32} + \frac{1}{8}\log\left(\frac{1}{\delta}\right) \\ &< \frac{1}{4}\log\left(\frac{1}{\delta}\right),\end{aligned}$$

where the final inequality follows from $\delta \leq 1/4$. This contradicts (34).

Hence the policy cannot satisfy (30) on every member of the hard family. Therefore there exists an i satisfying (29). □

Theorem 8. *[Finite-confidence minimax lower bound] Assume Gaussian noise with variance σ^2, $K \geq 8$, $\delta \in (0, 1/4]$, and $0 < \epsilon \leq \sigma/8$. For $h > 0$, define $\Theta_h^{(\delta,\epsilon)} := \left\{\theta : H_{\mathrm{NS}}^{(\delta)}(\theta, \epsilon) \leq h\right\}$, and let Π_B denote the class of adaptive fixed-budget policies using at most B pulls. For $\pi \in \Pi_B$, let $\widehat{i}$ denote its recommended arm. Define the simple regret as $r_B := \mu_1(1) - \mu_{\widehat{i}}(1)$. There exist universal constants $c, C > 0$ such that for every*

$$h \geq C\frac{K\sigma^2}{4\epsilon^2}\left[\log(eK/\delta) + \log\log(e^e + \sigma^2/4\epsilon^2)\right], \tag{2}$$

$$\text{if } B \leq c\,h\frac{\log(1/\delta)}{\log(eK/\delta) + \log\log(e^e + h/K)}, \tag{3}$$

then, for every $\pi \in \Pi_B$, $\sup_{\theta \in \Theta_h^{(\delta,\epsilon)}} \mathbb{P}_\theta^\pi(r_B > \epsilon) > \delta$. Moreover, the hard instance can be chosen to satisfy $H_{\mathrm{NS}}^{(\delta)}(\theta, \epsilon) \geq c\,h$.

Proof. For the hard family above, define

$$F(\eta) := K\frac{\sigma^2}{\eta^2}\left[\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{\sigma^2}{\eta^2}\right)\right].$$

The function $F(\eta)$ is continuous and strictly increases as η decreases, with

$$F(\eta) \longrightarrow \infty \qquad \text{as } \eta \downarrow 0.$$

The lower bound on h in (2) therefore ensures that we may choose some

$$0 < \eta \leq 2\epsilon$$

such that

$$F(\eta) = \frac{h}{C_0}, \tag{35}$$

where C_0 is a sufficiently large fixed universal constant.

By Lemma 15, C_0 can be chosen so that every member θ_i of this hard family satisfies

$$c_0 h \leq H_{\mathrm{NS}}^{(\delta)}(\theta_i, \epsilon) \leq h \tag{36}$$

for a universal $c_0 > 0$.

It remains to show that the budget condition (3) is small enough to invoke Lemma 17. Let

$$x := \frac{\sigma^2}{\eta^2}.$$

Equation (35) gives

$$\frac{h}{C_0 K} = x \left[\log\left(\frac{eK}{\delta}\right) + \log\log(e^e + x) \right]. \tag{37}$$

In particular,

$$\frac{h}{K} \geq x.$$

Hence

$$\log\left(\frac{eK}{\delta}\right) + \log\log\left(e^e + \frac{h}{K}\right) \geq \log\left(\frac{eK}{\delta}\right) + \log\log(e^e + x). \tag{38}$$

Combining (37) and (38),

$$\begin{aligned} \frac{h}{\log(eK/\delta) + \log\log\left(e^e + \frac{h}{K}\right)} &\leq C_0 K x \\ &= C_0 K \frac{\sigma^2}{\eta^2}. \end{aligned}$$

Therefore, by choosing the universal constant c in (3) small enough, that budget condition implies

$$B \leq \frac{1}{8} K \frac{\sigma^2}{\eta^2} \log\left(\frac{1}{\delta}\right).$$

Lemma 17 then guarantees that for at least one member θ_i of this hard family,

$$\mathbb{P}_i^{\pi}\left(r_B > \epsilon\right) > \delta.$$

By (36), this instance also has $H_{\mathrm{NS}}^{(\delta)}(\theta_i, \epsilon)$ of order h, which proves the theorem. □

D Additional Simulation Results

Functional form for rewards in Fig.2. Here, we describe the details for the reward trajectories and noise used in the simulation. We use $K = 8$ arms with streams of length $T = 640$ and independent Gaussian noise. Unless it is being varied, the noise standard deviation is $\sigma = 0.02$, the smallest terminal gap is $\Delta_{\min} = 0.059$, and the confidence input to `SignCount-SH` is $\delta = 0.05$. We vary the normalized budget over

$$B/K \in \{8, 16, 32, 64, 128, 256\},$$

and, at $B/K = 128$, vary

$$\sigma \in \{0.01, 0.015, 0.02, 0.025, 0.03, 0.04\}.$$

The simulated mean trajectories are specified exactly as follows. Define six regime vectors

$$\begin{aligned}
\boldsymbol{v}^{(0)} &= (0.55, 0.491, 0.488, 0.485, 0.485, 0.40, 0.36, 0.32),\\
\boldsymbol{v}^{(1)} &= (0.51, 0.496, 0.501, 0.506, 0.6316666667, 0.40, 0.36, 0.32),\\
\boldsymbol{v}^{(2)} &= (0.468, 0.545, 0.542, 0.539, 0.4723333333, 0.42, 0.39, 0.36),\\
\boldsymbol{v}^{(3)} &= (0.5065, 0.507, 0.504, 0.501, 0.4080777778, 0.42, 0.39, 0.36),\\
\boldsymbol{v}^{(4)} &= (0.60, 0.52, 0.515, 0.51, 0.505, 0.42, 0.39, 0.36),\\
\boldsymbol{v}^{(5)} &= (0.45, 0.54, 0.62, 0.50, 0.48, 0.43, 0.40, 0.37).
\end{aligned}$$

For integers $u < v$, let

$$c_{u,v}(s) := \frac{1}{2} - \frac{1}{2}\cos\left(\frac{\pi(s-u)}{v-u}\right), \qquad u \le s \le v,$$

and write

$$\boldsymbol{\mu}(s) = \begin{cases}
\boldsymbol{v}^{(0)}, & 1 \le s \le 32,\\
(1 - c_{33,48}(s))\boldsymbol{v}^{(0)} + c_{33,48}(s)\boldsymbol{v}^{(1)}, & 33 \le s \le 48,\\
\boldsymbol{v}^{(1)}, & 49 \le s \le 64,\\
(1 - c_{65,80}(s))\boldsymbol{v}^{(1)} + c_{65,80}(s)\boldsymbol{v}^{(2)}, & 65 \le s \le 80,\\
\boldsymbol{v}^{(2)}, & 81 \le s \le 128,\\
(1 - c_{129,144}(s))\boldsymbol{v}^{(2)} + c_{129,144}(s)\boldsymbol{v}^{(3)}, & 129 \le s \le 144,\\
\boldsymbol{v}^{(3)}, & 145 \le s \le 256,\\
(1 - c_{257,288}(s))\boldsymbol{v}^{(3)} + c_{257,288}(s)\boldsymbol{v}^{(4)}, & 257 \le s \le 288,\\
\boldsymbol{v}^{(4)}, & 289 \le s \le 512,\\
(1 - c_{513,544}(s))\boldsymbol{v}^{(4)} + c_{513,544}(s)\boldsymbol{v}^{(5)}, & 513 \le s \le 544,\\
\boldsymbol{v}^{(5)}, & 545 \le s \le 640.
\end{cases}$$

The observed reward of arm i at time s is

$$Y_{i,s} = \mu_i(s) + \xi_{i,s}, \qquad \xi_{i,s} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

The 4 target times considered in the experiment are $\{1, 49, 289, 545\}$.

Regime orderings. Recall the different regimes shaded in Fig.2. Away from the smooth transitions, the four qualitative portions of the trajectory have the following arm orderings, from largest to smallest mean. The regime 1 plateau has $1 \succ 2 \succ 3 \succ 4 \succ 5 \succ 6 \succ 7 \succ 8$. Within the contender-dominated regime 2, the three successive plateaus have $5 \succ 1 \succ 4 \succ 3 \succ 2 \succ 6 \succ 7 \succ 8$, $2 \succ 3 \succ 4 \succ 5 \succ 1 \succ 6 \succ 7 \succ 8$, and the weaker reversal $2 \succ 1 \succ 3 \succ 4 \succ 6 \succ 5 \succ 7 \succ 8$. The sustained return of arm 1 restores $1 \succ 2 \succ 3 \succ 4 \succ 5 \succ 6 \succ 7 \succ 8$ in regime 3, whereas regime 4 has $3 \succ 2 \succ 4 \succ 5 \succ 1 \succ 6 \succ 7 \succ 8$. The cosine transitions interpolate continuously between these plateau orderings.

Numerical values for the plot in Fig.3 are reported in Tables 2 and 3.

Table 2: *Average target-time best-arm identification probability against normalized budget at* $\sigma = 0.02$, $\delta = 0.05$. *Shaded bands are 95% confidence intervals over* 1,000 *runs and* 4 *target times.*

Method	$B/K = 16$	$B/K = 32$	$B/K = 64$	$B/K = 128$	$B/K = 256$
`SignCount-SH`	**1.0000** [1.0000, 1.0000]	0.9890 [0.9858, 0.9922]	**0.8128** [0.8060, 0.8195]	**0.9570** [0.9512, 0.9629]	**0.9715** [0.9666, 0.9764]
Full-prefix SH	**1.0000** [1.0000, 1.0000]	0.7580 [0.7523, 0.7637]	0.5155 [0.5118, 0.5192]	0.5000 [0.5000, 0.5000]	0.2500 [0.2500, 0.2500]
Vanilla-SH	0.9998 [0.9993, 1.0000]	0.7323 [0.7283, 0.7362]	0.5330 [0.5278, 0.5382]	0.4920 [0.4893, 0.4947]	0.2500 [0.2500, 0.2500]
Successive Rejects	**1.0000** [1.0000, 1.0000]	0.9995 [0.9988, 1.0000]	0.5000 [0.5000, 0.5000]	0.5000 [0.5000, 0.5000]	0.2503 [0.2498, 0.2507]
Adjacent-BAI	**1.0000** [1.0000, 1.0000]	**1.0000** [1.0000, 1.0000]	0.7343 [0.7274, 0.7411]	0.5000 [0.5000, 0.5000]	0.5000 [0.5000, 0.5000]
P1-RAGE	0.4230 [0.4076, 0.4384]	0.4558 [0.4409, 0.4706]	0.4550 [0.4418, 0.4682]	0.4668 [0.4567, 0.4768]	0.4738 [0.4627, 0.4848]

Table 3: *Mean target-time simple regret against normalized budget at* $\sigma = 0.02$, $\delta = 0.05$. *Shaded bands are 95% confidence intervals over* 1,000 *runs and* 4 *target times.*

Method	$B/K = 16$	$B/K = 32$	$B/K = 64$	$B/K = 128$	$B/K = 256$
`SignCount-SH`	**0.00000** [0.00000, 0.00000]	0.00072 [0.00051, 0.00092]	**0.01166** [0.01124, 0.01208]	**0.00261** [0.00225, 0.00296]	**0.00176** [0.00146, 0.00206]
Full-prefix SH	**0.00000** [0.00000, 0.00000]	0.02998 [0.02937, 0.03059]	0.04731 [0.04706, 0.04755]	0.04883 [0.04880, 0.04887]	0.06864 [0.06859, 0.06869]
Vanilla-SH	0.00003 [0.00000, 0.00009]	0.03317 [0.03291, 0.03344]	0.04673 [0.04639, 0.04708]	0.04963 [0.04940, 0.04987]	0.06907 [0.06903, 0.06910]
Successive Rejects	**0.00000** [0.00000, 0.00000]	0.00003 [0.00000, 0.00008]	0.04938 [0.04932, 0.04943]	0.04879 [0.04875, 0.04883]	0.06990 [0.06986, 0.06994]
Adjacent-BAI	**0.00000** [0.00000, 0.00000]	**0.00000** [0.00000, 0.00000]	0.01949 [0.01873, 0.02025]	0.04902 [0.04895, 0.04908]	0.04867 [0.04864, 0.04870]
P1-RAGE	0.05851 [0.05670, 0.06032]	0.05337 [0.05181, 0.05494]	0.05273 [0.05145, 0.05402]	0.05180 [0.05095, 0.05265]	0.05137 [0.05051, 0.05222]

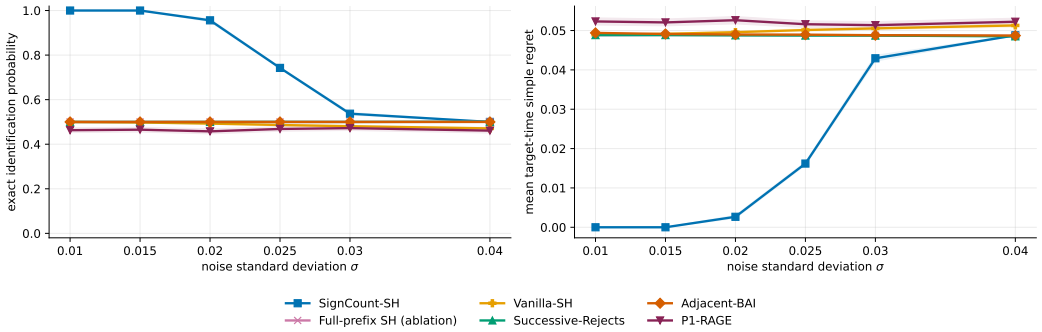


Figure 5: *Exact best-arm identification probability (left) and mean simple regret (right) as the noise standard deviation σ varies at $B/K = 128$. Shaded bands are 95% confidence intervals over 1000 runs and 4 target times.*

Ablation on noise level. We now investigate how the performance of different BAI methods deteriorate as a function of the noise level of the observed rewards. Fig. 5 plots the performance metrics as a function of the noise level. We can see that the performance of `SignCount-SH` degrades with increasing noise level as expected.

E VARIANCE-ADAPTIVE VERSION OF SIGNCOUNT-SH

Here we describe the variance-adaptive version of `SignCount-SH` used in the experiments. This implementation is the variable-batch-size specialization of the method analyzed in Appendix F.

The Hoeffding calibration in Line 3 of Algorithm 1 can be conservative because it does not use the variance of the reward observations. We therefore replace it with an arm-level empirical-Bernstein radius.

For a fixed window s, let

$$X_{a,s,1}, \dots, X_{a,s,m_s} \in \{0,1\}$$

be the pickup-zone indicators used to estimate the reward of arm a. Define

$$\widehat{p}_{a,s} := \frac{1}{m_s} \sum_{u=1}^{m_s} X_{a,s,u}.$$

The unbiased empirical variance of the individual Bernoulli observations is

$$\begin{aligned} \widehat{w}_{a,s} &:= \frac{1}{m_s - 1} \sum_{u=1}^{m_s} (X_{a,s,u} - \widehat{p}_{a,s})^2 \\ &= \frac{m_s}{m_s - 1} \widehat{p}_{a,s}(1 - \widehat{p}_{a,s}). \end{aligned}$$

Consequently,

$$\frac{\widehat{w}_{a,s}}{m_s} = \frac{\widehat{p}_{a,s}(1 - \widehat{p}_{a,s})}{m_s - 1}$$

is an unbiased estimator of $\mathrm{Var}(\widehat{p}_{a,s})$.

For each temporal depth s, define

$$\beta_s := \log\left(\frac{2\pi^2 K s^2}{3\delta}\right), \qquad c_s := \sqrt{\frac{2\beta_s}{m_s - 1}}. \tag{39}$$

Following Equation (42), define the arm-level variance upper bound

$$\widehat{u}_{a,s} := \min\left\{\frac{1}{4}, \left(\sqrt{\widehat{w}_{a,s}} + c_s\right)^2\right\}.$$

At dyadic resolution $k_i = 2^i$, let

$$\ell_i := \log\left(\frac{2\pi^2 K(i+1)^2}{3\delta}\right) \tag{40}$$

and define the arm-level empirical-Bernstein radius

$$r_{a,i}^{\mathrm{EB}} := \sqrt{\frac{2\ell_i}{k_i^2}\sum_{s=1}^{k_i}\frac{\widehat{u}_{a,s}}{m_s}} + \frac{\ell_i}{3k_i \min_{s\le k_i} m_s}. \tag{41}$$

For a pair of arms (a, b), define

$$q_{a,b,i}^{\mathrm{EB}} := r_{a,i}^{\mathrm{EB}} + r_{b,i}^{\mathrm{EB}}.$$

The Bernstein version of MULTI-RESOLUTION-DUEL (Algorithm 1) replaces the threshold in Line 3 by $q_{a,b,i}^{\mathrm{EB}}$ and replaces the test in Line 5 by

$$\left|\widehat{d}_{a,b}(k_i)\right| > 2q_{a,b,i}^{\mathrm{EB}}.$$

The empirical prefix gap remains

$$\widehat{d}_{a,b}(k) := \frac{1}{k}\sum_{s=1}^{k}\left(\widehat{p}_{a,s} - \widehat{p}_{b,s}\right).$$

To see the form of Equation (41), write

$$p_{a,s} := \mathbb{E}[X_{a,s,u}].$$

Under independent Bernoulli sampling,

$$\mathrm{Var}(\widehat{p}_{a,s}) = \frac{p_{a,s}(1-p_{a,s})}{m_s}.$$

Therefore, the variance of the arm-level prefix estimate is

$$\mathrm{Var}\left(\frac{1}{k}\sum_{s=1}^{k}\widehat{p}_{a,s}\right) = \frac{1}{k^2}\sum_{s=1}^{k}\frac{p_{a,s}(1-p_{a,s})}{m_s}.$$

The first term in Equation (41) is the variance-dependent term obtained after replacing the unknown Bernoulli variances by the upper confidence bounds $\widehat{u}_{a,s}$. The second term is the lower-order Bernstein term; the factor $\min_{s\le k_i} m_s$ arises from the largest weight of any individual observation in the prefix.

The logarithmic factors in Equations (39) and (40) depend on K, rather than $\binom{K}{2}$. As in Lemma 19, confidence events are constructed simultaneously for the K arm-level estimates. The pairwise bound then follows by adding the two arm-level radii, without a separate union bound over arm pairs.

F VARIANCE-ADAPTIVE SIGNCOUNT−SH UNDER BATCHED OBSERVATIONS

In this section, we analyze the variance-adaptive extension of `SignCount-SH` (Algorithm 2) under a variable-batch observation model. Each batch query costs one unit of the BAI budget, irrespective of the number of observations in the batch.

Batched observation model. Let $(m_t)_{t\ge 1}$ be a deterministic sequence of integers satisfying $m_t \ge 2$, known to the learner. The batch size m_t may vary with the temporal depth t, but is the same for all arms at that depth. When arm $a \in [K]$ is pulled for the tth time, the learner observes

$$X_{a,t,1}, \ldots, X_{a,t,m_t} \in [0, 1].$$

For every fixed (a, t), the observations within the batch are independent and identically distributed with

$$\mathbb{E}[X_{a,t,s}] = \mu_a(t), \qquad \mathrm{Var}(X_{a,t,s}) = v_{a,t}.$$

We assume that all observations are independent across arms, temporal depths, and positions within a batch. Both $\mu_a(t)$ and $v_{a,t}$ may depend on a and t.

The reward associated with this batch query is its empirical mean

$$Y_{a,t} := \frac{1}{m_t} \sum_{s=1}^{m_t} X_{a,t,s}.$$

Thus,

$$\mathbb{E}[Y_{a,t}] = \mu_a(t), \qquad \mathrm{Var}(Y_{a,t}) = \frac{v_{a,t}}{m_t}.$$

Observing $Y_{a,t}$ together with its underlying batch costs one unit of the BAI budget.

The empirical variance within the batch is

$$\widehat{v}_{a,t} := \frac{1}{m_t - 1} \sum_{s=1}^{m_t} (X_{a,t,s} - Y_{a,t})^2 .$$

For a prefix of length k, define

$$\widehat{\mu}_a(k) := \frac{1}{k} \sum_{t=1}^{k} Y_{a,t}, \qquad \overline{\mu}_a(k) := \frac{1}{k} \sum_{t=1}^{k} \mu_a(t),$$

and, for every pair (a, b),

$$\widehat{d}_{a,b}(k) := \widehat{\mu}_a(k) - \widehat{\mu}_b(k), \qquad \overline{d}_{a,b}(k) := \overline{\mu}_a(k) - \overline{\mu}_b(k).$$

Upper confidence bounds on the batch variances. For $t \geq 1$, let

$$\beta_t := \log\left(\frac{2\pi^2 K t^2}{3\delta}\right), \qquad c_t := \sqrt{\frac{2\beta_t}{m_t - 1}}.$$

Define the empirical variance upper bound

$$\widehat{u}_{a,t} := \min\left\{\frac{1}{4}, \left(\sqrt{\widehat{v}_{a,t}} + c_t\right)^2\right\}, \tag{42}$$

and its deterministic population envelope

$$\overline{u}_{a,t} := \min\left\{\frac{1}{4}, \left(\sqrt{v_{a,t}} + 2c_t\right)^2\right\}.$$

The truncation at $1/4$ follows because every random variable supported on $[0, 1]$ has variance at most $1/4$.

Lemma 18 (Concentration of the empirical standard deviation)**.** *Let $Z_1, \ldots, Z_n$ be independent and identically distributed random variables supported on $[0, 1]$, with variance v, and let*

$$\widehat{v} := \frac{1}{n-1} \sum_{s=1}^{n} \left(Z_s - \frac{1}{n} \sum_{r=1}^{n} Z_r\right)^2 .$$

For every $\eta \in (0, 1)$,

$$\mathbb{P}\left(\sqrt{v} > \sqrt{\widehat{v}} + \sqrt{\frac{2\log(1/\eta)}{n-1}}\right) \leq \eta$$

and

$$\mathbb{P}\left(\sqrt{\widehat{v}} > \sqrt{v} + \sqrt{\frac{2\log(1/\eta)}{n-1}}\right) \leq \eta.$$

Lemma 18 is Theorem 10 of Maurer & Pontil (2009). Apply it to the batch at depth t, with $n = m_t$ and

$$\eta_{a,t} := \frac{3\delta}{2\pi^2 K t^2}.$$

A union bound over $a \in [K]$, $t \geq 1$, and the two tails gives

$$\mathbb{P}\left(\left|\sqrt{\widehat{v}_{a,t}} - \sqrt{v_{a,t}}\right| \leq c_t \text{ for every } a \in [K], t \geq 1\right)$$
$$\geq 1 - 2\sum_{a=1}^{K}\sum_{t=1}^{\infty}\eta_{a,t} = 1 - \frac{\delta}{2}, \tag{43}$$

where we used $\sum_{t=1}^{\infty} t^{-2} = \pi^2/6$.

Let $\mathcal{E}_{\text{var}}$ denote the event in (43). On this event,

$$v_{a,t} \leq \widehat{u}_{a,t} \leq \overline{u}_{a,t} \qquad \text{for every } a \in [K], t \geq 1. \tag{44}$$

Indeed, the first inequality follows from

$$\sqrt{v_{a,t}} \leq \sqrt{\widehat{v}_{a,t}} + c_t,$$

whereas the second follows from

$$\sqrt{\widehat{v}_{a,t}} \leq \sqrt{v_{a,t}} + c_t.$$

Arm-level empirical-Bernstein radii. As in MULTI-RESOLUTION-DUEL (Algorithm 1), let $k_i = 2^i$ and define

$$m_{\min}(k_i) := \min_{1 \leq t \leq k_i} m_t.$$

Let

$$\ell_i := \log\left(\frac{2\pi^2 K(i+1)^2}{3\delta}\right). \tag{45}$$

For arm a and resolution i, define the observable arm-level radius

$$r_{a,i}^{\text{EB}} := \sqrt{\frac{2\ell_i}{k_i^2}\sum_{t=1}^{k_i}\frac{\widehat{u}_{a,t}}{m_t}} + \frac{\ell_i}{3k_i m_{\min}(k_i)}.$$

The corresponding pairwise radius is

$$q_{a,b,i}^{\text{EB}} := r_{a,i}^{\text{EB}} + r_{b,i}^{\text{EB}}.$$

The Bernstein version of MULTI-RESOLUTION-DUEL (Algorithm 1) replaces Line 5 by

$$\text{if } \left|\widehat{d}_{a,b}(k_i)\right| > 2q_{a,b,i}^{\text{EB}}, \text{ then return its empirical sign.} \tag{46}$$

The logarithmic factor in (45) depends on K, rather than K^2. We first construct a simultaneous confidence event for the K arm-level prefix estimates. Pairwise confidence bounds then follow by the triangle inequality, without an additional union bound over arm pairs.

Lemma 19 (Global variable-batch empirical-Bernstein event)**.** *With probability at least $1-\delta$, simultaneously for every arm $a \in [K]$ and every dyadic resolution i,*

$$|\widehat{\mu}_a(k_i) - \overline{\mu}_a(k_i)| \leq r_{a,i}^{\text{EB}}. \tag{47}$$

Consequently, simultaneously for every pair (a,b) and every resolution i,

$$\left|\widehat{d}_{a,b}(k_i) - \overline{d}_{a,b}(k_i)\right| \leq q_{a,b,i}^{\text{EB}}. \tag{48}$$

Proof. Fix an arm a and a resolution i. Write

$$Z_{a,t,s} := \frac{X_{a,t,s} - \mu_a(t)}{k_i m_t}, \qquad 1 \leq t \leq k_i, \quad 1 \leq s \leq m_t.$$

Then

$$\widehat{\mu}_a(k_i) - \overline{\mu}_a(k_i) = \sum_{t=1}^{k_i}\sum_{s=1}^{m_t} Z_{a,t,s}.$$

These centered random variables are independent and satisfy

$$|Z_{a,t,s}| \le \frac{1}{k_i m_{\min}(k_i)}.$$

Their total variance is

$$\sum_{t=1}^{k_i}\sum_{s=1}^{m_t} \mathrm{Var}(Z_{a,t,s}) = \frac{1}{k_i^2}\sum_{t=1}^{k_i}\frac{v_{a,t}}{m_t}.$$

Bernstein's inequality therefore gives

$$\mathbb{P}\left(|\widehat{\mu}_a(k_i) - \overline{\mu}_a(k_i)| > \sqrt{\frac{2\ell_i}{k_i^2}\sum_{t=1}^{k_i}\frac{v_{a,t}}{m_t}} + \frac{\ell_i}{3k_i m_{\min}(k_i)}\right)$$
$$\le 2e^{-\ell_i} = \frac{3\delta}{\pi^2 K(i+1)^2}. \tag{49}$$

Taking a union bound over $a \in [K]$ and $i \ge 0$ shows that (49) holds simultaneously with probability at least

$$1 - \sum_{a=1}^{K}\sum_{i=0}^{\infty}\frac{3\delta}{\pi^2 K(i+1)^2} = 1 - \frac{\delta}{2}.$$

Intersect this event with $\mathcal{E}_{\mathrm{var}}$, which has probability at least $1-\delta/2$. On $\mathcal{E}_{\mathrm{var}}$, Equation (44) gives $v_{a,t} \le \widehat{u}_{a,t}$. Hence the deviation bound in (49) is at most $r_{a,i}^{\mathrm{EB}}$. This proves (47) simultaneously with probability at least $1-\delta$.

Finally,

$$\left|\widehat{d}_{a,b}(k_i) - \overline{d}_{a,b}(k_i)\right|$$
$$\le |\widehat{\mu}_a(k_i) - \overline{\mu}_a(k_i)| + |\widehat{\mu}_b(k_i) - \overline{\mu}_b(k_i)| \le q_{a,b,i}^{\mathrm{EB}},$$

which proves (48). No union bound over the $\binom{K}{2}$ arm pairs is required. □

Deterministic envelopes and resolving depths. To define a deterministic instance complexity, let

$$\overline{r}_{a,i}^{\mathrm{EB}} := \sqrt{\frac{2\ell_i}{k_i^2}\sum_{t=1}^{k_i}\frac{\overline{u}_{a,t}}{m_t}} + \frac{\ell_i}{3k_i m_{\min}(k_i)},$$

and

$$\overline{q}_{a,b,i}^{\mathrm{EB}} := \overline{r}_{a,i}^{\mathrm{EB}} + \overline{r}_{b,i}^{\mathrm{EB}}.$$

Also define the deterministic lower envelope

$$\underline{q}_i^{\mathrm{EB}} := \frac{2\ell_i}{3k_i m_{\min}(k_i)}.$$

On $\mathcal{E}_{\mathrm{var}}$,

$$\underline{q}_i^{\mathrm{EB}} \le q_{a,b,i}^{\mathrm{EB}} \le \overline{q}_{a,b,i}^{\mathrm{EB}} \qquad \text{for every } a, b, i. \tag{50}$$

Recall from Definition 1 that

$$s_{a,b} := \mathrm{sign}\left(\mu_a(1) - \mu_b(1)\right)$$

and

$$\Gamma_{a,b}(i) := s_{a,b}\overline{d}_{a,b}(k_i).$$

Define the variable-batch empirical-Bernstein resolving depth by

$$\tau_{a,b}^{(\delta),\mathrm{EB}} := \inf\left\{k_j : \begin{array}{ll}\Gamma_{a,b}(i) > -\underline{q}_i^{\mathrm{EB}} & \text{for every } i < j,\\ \Gamma_{a,b}(j) > 3\overline{q}_{a,b,j}^{\mathrm{EB}}.\end{array}\right\}. \tag{51}$$

As in Definition 1, we set $\tau_{a,b}^{(\delta),\mathrm{EB}} = \infty$ if no such scale exists.

Lemma 20 (Correctness of the variable-batch Bernstein comparison)**.** *Consider a pair of arms (a, b) and suppose that the Bernstein version of* MULTI-RESOLUTION-DUEL *(Algorithm 1) has observed both arms to a common depth*

$$t \geq \tau_{a,b}^{(\delta),\mathrm{EB}}.$$

On the event of Lemma 19, the Bernstein comparison in (46) returns the terminal ordering $s_{a,b}$.

Proof. Let $k_j = \tau_{a,b}^{(\delta),\mathrm{EB}}$. At an earlier resolution $i < j$, Equations (48), (50), and (51) imply

$$\begin{aligned} s_{a,b}\widehat{d}_{a,b}(k_i) &\geq \Gamma_{a,b}(i) - q_{a,b,i}^{\mathrm{EB}} \\ &> -\underline{q}_i^{\mathrm{EB}} - q_{a,b,i}^{\mathrm{EB}} \\ &\geq -2q_{a,b,i}^{\mathrm{EB}}. \end{aligned}$$

Therefore, if the test stops at resolution $i < j$, it cannot cross the threshold in the wrong direction.

If the test has not stopped before resolution j, then

$$\begin{aligned} s_{a,b}\widehat{d}_{a,b}(k_j) &\geq \Gamma_{a,b}(j) - q_{a,b,j}^{\mathrm{EB}} \\ &> 3\overline{q}_{a,b,j}^{\mathrm{EB}} - q_{a,b,j}^{\mathrm{EB}} \\ &\geq 2q_{a,b,j}^{\mathrm{EB}}. \end{aligned}$$

Hence the test crosses the threshold at resolution j in the correct direction. □

Simple-regret guarantee. Define $\chi_{g,a;b}^{(\delta),\mathrm{EB}}$, $\rho_{g,a}^{(\delta),\mathrm{EB}}$, and $H_{\mathrm{NS,EB}}^{(\delta)}(\theta, \epsilon)$ exactly as in Definitions 3 and 4, with $\tau_{a,b}^{(\delta)}$ replaced by $\tau_{a,b}^{(\delta),\mathrm{EB}}$.

Theorem 21 (Variable-batch Bernstein simple-regret guarantee)**.** *Suppose*

$$B \geq 4 \log_2 K \, H_{\mathrm{NS,EB}}^{(\delta)}(\theta, \epsilon).$$

Then the variable-batch Bernstein version of `SignCount-SH` *(Algorithm 2) returns an arm $\widehat{a}$ satisfying*

$$\mu_1(1) - \mu_{\widehat{a}}(1) \leq \epsilon$$

with probability at least $1 - \delta$.

Proof. Condition on the global event in Lemma 19, which has probability at least $1 - \delta$. This event plays the same role as the global good event in Lemma 10. Conditional on it, Lemma 20 is the variable-batch Bernstein counterpart of the pairwise comparison guarantee in Lemma 2.

The proof of the ranking certificate in Lemma 12 uses only the pairwise comparison guarantee and is otherwise deterministic. It therefore applies after replacing $\tau^{(\delta)}$ and $\rho^{(\delta)}$ by $\tau^{(\delta),\mathrm{EB}}$ and $\rho^{(\delta),\mathrm{EB}}$, respectively. Finally, the proof of the simple-regret guarantee in Theorem 5 uses the ranking certificate together with the allocation rule of `SignCount-SH` (Algorithm 2). The same proof therefore applies with $H_{\mathrm{NS}}^{(\delta)}$ replaced by $H_{\mathrm{NS,EB}}^{(\delta)}$.

The result also holds under arbitrary tie-breaking among arms with equal sign-count scores. □

G FINAL-PAIR SIGN-RESOLUTION DEPTH

We inspect the depths selected by Algorithm 1 to empiricially certify its adaptive selection of historical data. For this, we consider the last epoch where there are only two arms remaining. The largest historical stream will be observed for the two surviving arms in the final epoch.

We look at the depth selected by the sign-resolution test before resolving the comparison among the finalists. Let d_B be the number of pulls accumulated by each finalist under budget B, and let $\widehat{k}$ be the first dyadic prefix at which algorithm stops and output a $W_{a,b} \neq 1/2$. We record

$$U_B := \begin{cases} \widehat{k}/d_B, & W_{a,b} \neq 1/2, \\ 1, & W_{a,b} = 1/2, \end{cases} \qquad Q_B := \mathbf{1}\{W_{a,b} \neq 1/2\}.$$

Thus U_B is the fraction of allocated finalist history indicated by the sign-resolution test. If the test never resolves, we assign $U_B = 1$. The statistic Q_B records how often the sign test resolves the final pair.

NYC taxi experiment. For $B/K \in \{4, 8, 12, 16\}$, the final-epoch allocations are respectively $d_B \in \{6, 12, 18, 24\}$. Fig.6 displays the mean value of these statistics (the numbers in the plots describe the mean value). We find that Bernstein variant of Algorithm 1 (`SignCount-SH (Bernstein)`) indeed adapts to non-stationarity by selecting relevant portion of the historical data when resolving signs. Further in about 40% of the target slots, the sign test successfully fires leading to a value of $W \neq 1/2$. Hence the algorithm do not always resort back to tie breaking rules.

In contrast, the hoeffding version shows no adaptivity to historical data due to its conservatively large thresholds. Hence `SignCount-SH (Hoeffding)` simply falls back to Vanilla-SH via the tie-breaking rule.

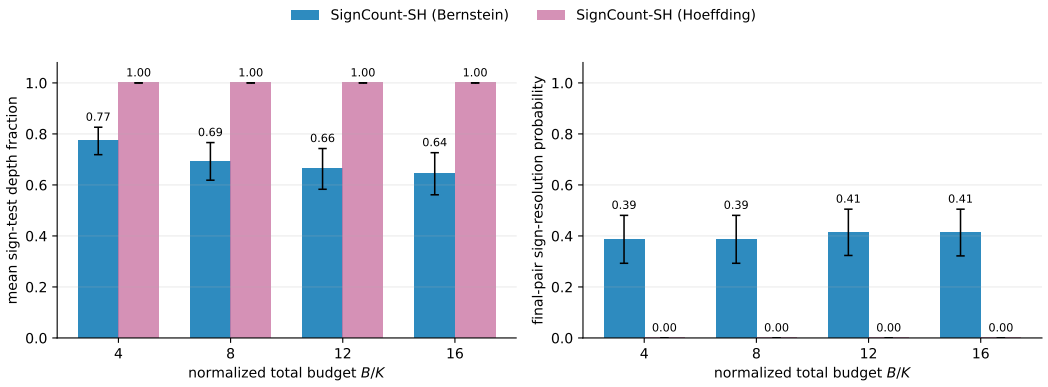


Figure 6: *Final-pair sign-resolution depth in the NYC taxi experiment. The left panel reports U_B, the mean sign-test depth fraction: the first resolving prefix divided by the finalist's allocated depth, with an unresolved comparison assigned fraction one. The stage-local tie-breaking fallback is excluded. The right panel reports Q_B, the probability that the sign test itself resolves the last pair. Error bars are pointwise 95% confidence intervals over 17 weekly averages.*

Simulation experiment. We repeat the audit for the Hoeffding SignCount-SH method on the simulation instance at $\sigma = 0.02$.

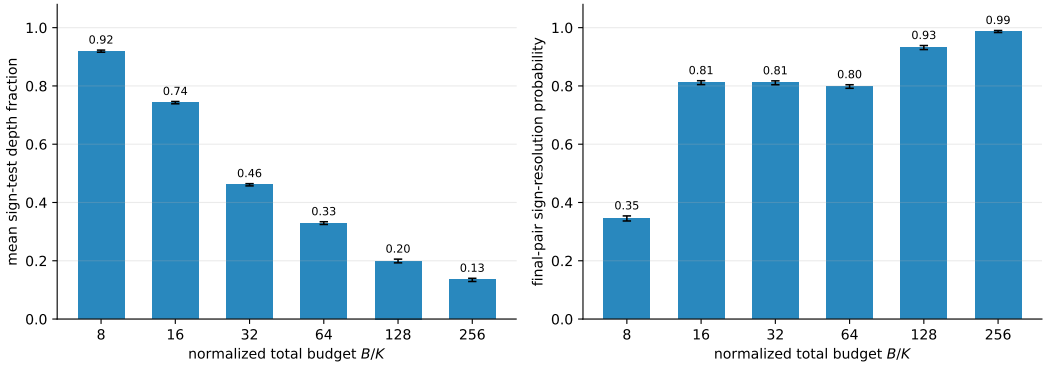


Figure 7: *Final-pair sign-resolution depth for Hoeffding SignCount-SH on the gradual-crossing simulation at $\sigma = 0.02$. The Y axis are U_B (left) and Q_B (right). Error bars are pointwise 95% intervals over 1,000 independent simulation replicates and 4 target times*

Fig.7 demonstrates that the sign test increasingly resolves the final comparison without consuming the full allocated history. As B/K increases from 8 to 256, the mean depth fraction U_B decreases from 0.919 to 0.134, while the resolution probability Q_B increases from 0.345 to 0.987. Thus, the decline in U_B is not caused by more abstentions: the test resolves more frequently while using a smaller fraction of the available history. This behavior reflects adaptation across target times, with easier target-indexed streams resolving at shallow scales and the more difficult target-1 stream becoming resolvable only at larger budgets.

H Additional Details and Results Continued from Section 4.2

In this section, we provide the omitted tables and figures from Section 4.2 due to space constraints. We use a subsample size of $\min\{4000, N_s/2\}$ trips, where N_s is the total number of trips available at a timeslot s, to compute the unbiased estimate of the rewards at that slot for a specific arm.

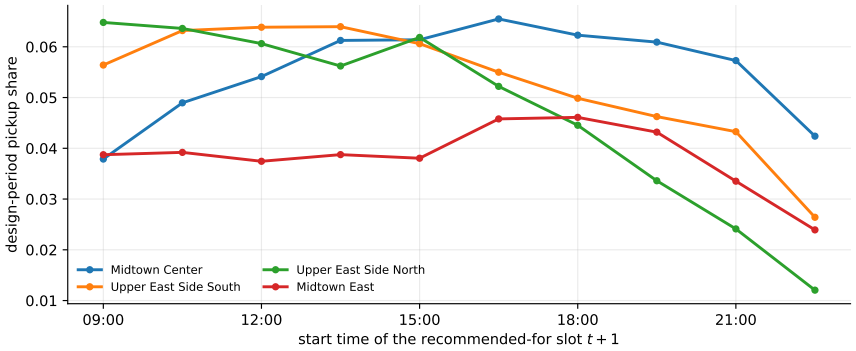


Figure 8: *Probability of pickup across 90-minute timeslots. The probabilities show non-stationary behavior causing the best pickup zone to change across time.*

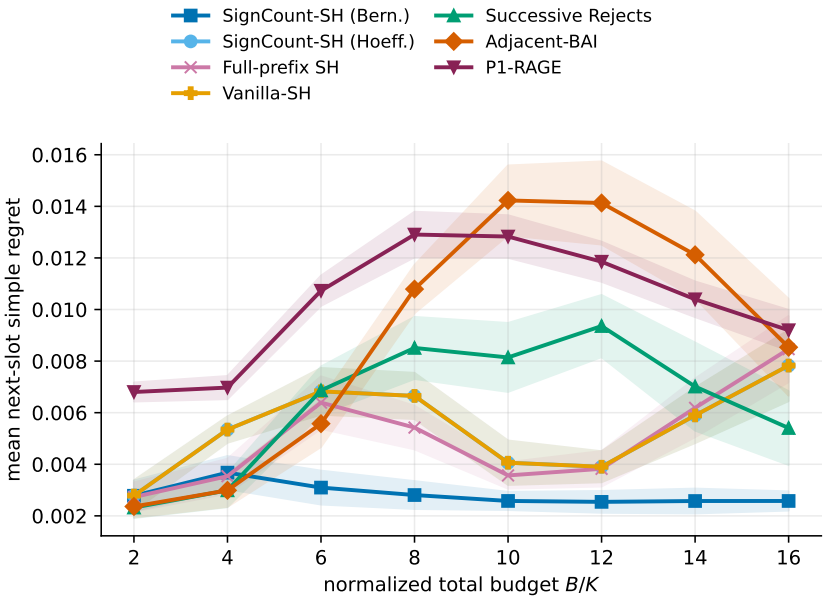


Figure 9: *Average one-step regret for recommending best arm for the next time slot. Bands are 95% confidence intervals.*

Table 4: *Mean one-step simple regret with 95% confidence intervals.*

Method	$B/K=2$	$B/K=4$	$B/K=6$	$B/K=8$	$B/K=16$
SignCount-SH (Bernstein)	0.00278 [0.00219, 0.00337]	0.00383 [0.00320, 0.00445]	**0.00322** [**0.00254**, **0.00390**]	**0.00292** [**0.00235**, **0.00348**]	**0.00266** [**0.00227**, **0.00305**]
SignCount-SH (Hoeffding)	0.00279 [0.00219, 0.00338]	0.00535 [0.00483, 0.00586]	0.00683 [0.00592, 0.00773]	0.00665 [0.00576, 0.00754]	0.00783 [0.00645, 0.00920]
Full-prefix SH	0.00273 [0.00217, 0.00329]	0.00354 [0.00286, 0.00421]	0.00639 [0.00537, 0.00741]	0.00542 [0.00458, 0.00627]	0.00847 [0.00719, 0.00974]
Vanilla-SH	0.00279 [0.00220, 0.00338]	0.00535 [0.00483, 0.00586]	0.00683 [0.00592, 0.00773]	0.00665 [0.00576, 0.00754]	0.00783 [0.00645, 0.00920]
Successive Rejects	**0.00232** [**0.00192**, **0.00272**]	0.00299 [0.00234, 0.00364]	0.00686 [0.00594, 0.00779]	0.00851 [0.00731, 0.00971]	0.00540 [0.00398, 0.00683]
Adjacent-BAI	0.00236 [0.00194, 0.00279]	**0.00299** [**0.00235**, **0.00363**]	0.00557 [0.00467, 0.00648]	0.01079 [0.00986, 0.01172]	0.00854 [0.00666, 0.01041]
P1-RAGE	0.00680 [0.00643, 0.00717]	0.00697 [0.00653, 0.00742]	0.01073 [0.01014, 0.01131]	0.01291 [0.01202, 0.01379]	0.00920 [0.00841, 0.00998]

I Ablation on Sample Noise in NYC Taxi-Trip Dataset

We use a subsample size of $\min\{4000, N_s/2\}$ trips, where N_s is the total number of trips available at a timeslot s, to compute the unbiased estimate of the rewards at that slot for a specific arm. Reducing the sample size can increase the noise in the observed rewards making the identification task more challenging. Ablations by setting the sample size to $\min\{4000, qN_s\}$ for various sampling fractions

q is displayed in Fig.10. Performance of all algorithms deteriorates with reduced sample size while `SignCount-SH (Bernstein)` exhibits a more graceful deterioration.

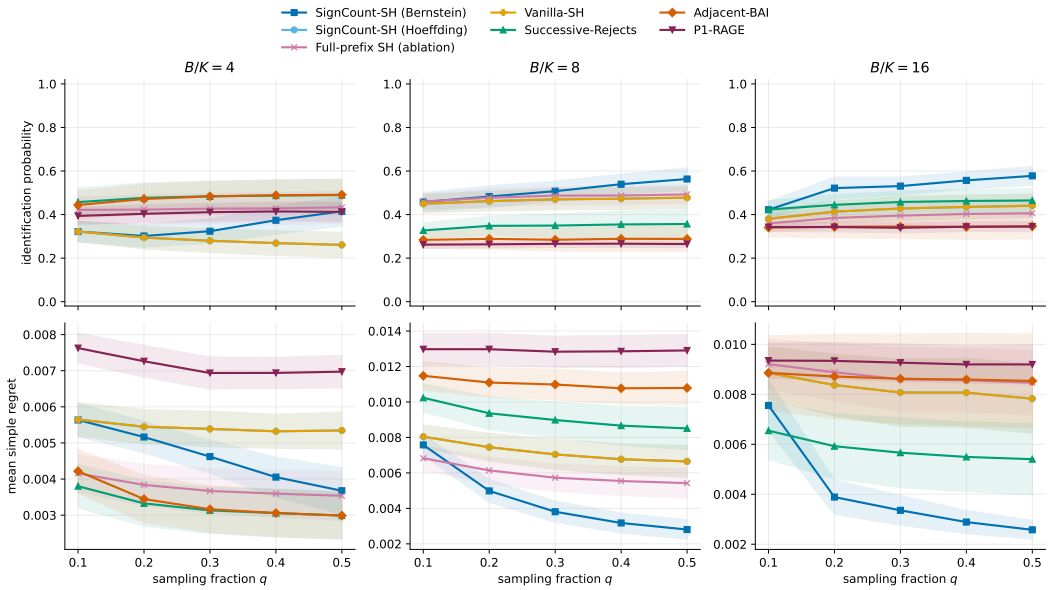


Figure 10: *Mean probability of recommending best arm at next timeslot (top) and simple regret (bottom) as the sampling fraction q varies in $m_s(q) = \min\{4000, \lfloor qN_s \rfloor\}$. m_s trips are used to construct noisy reward observations that is revealed to BAI algorithms. Smaller values of q lead to more noise in observations.*

Table 5: Morning mean next-slot best arm identification probability under the reward-query batch-size sweep with 95% confidence intervals.

B/K	Method	$q = 0.10$	$q = 0.20$	$q = 0.30$	$q = 0.40$	$q = 0.50$
4	SignCount-SH (Bernstein)	0.322 [0.275, 0.369]	0.302 [0.250, 0.354]	0.324 [0.268, 0.379]	0.374 [0.310, 0.438]	0.415 [0.347, 0.483]
	SignCount-SH (Hoeffding)	0.322 [0.276, 0.368]	0.294 [0.243, 0.346]	0.280 [0.226, 0.334]	0.269 [0.213, 0.325]	0.261 [0.203, 0.318]
	Full-prefix SH	0.421 [0.363, 0.480]	0.423 [0.357, 0.489]	0.427 [0.359, 0.496]	0.428 [0.356, 0.501]	0.433 [0.360, 0.506]
	Vanilla-SH	0.322 [0.276, 0.369]	0.294 [0.243, 0.346]	0.280 [0.226, 0.334]	0.269 [0.213, 0.325]	0.261 [0.203, 0.318]
	Successive Rejects	**0.457 [0.393, 0.522]**	**0.477 [0.409, 0.545]**	**0.484 [0.414, 0.553]**	0.487 [0.416, 0.559]	0.489 [0.417, 0.562]
	Adjacent-BAI	0.444 [0.380, 0.508]	0.472 [0.403, 0.540]	0.484 [0.415, 0.552]	**0.489 [0.419, 0.559]**	**0.491 [0.421, 0.561]**
	P1-RAGE	0.394 [0.355, 0.433]	0.403 [0.362, 0.445]	0.411 [0.370, 0.452]	0.414 [0.372, 0.456]	0.412 [0.369, 0.456]
8	SignCount-SH (Bernstein)	0.458 [0.415, 0.501]	**0.483 [0.433, 0.532]**	**0.507 [0.462, 0.553]**	**0.539 [0.494, 0.585]**	**0.563 [0.512, 0.615]**
	SignCount-SH (Hoeffding)	0.449 [0.408, 0.491]	0.462 [0.418, 0.506]	0.470 [0.422, 0.518]	0.473 [0.423, 0.522]	0.478 [0.428, 0.528]
	Full-prefix SH	**0.460 [0.424, 0.495]**	0.478 [0.438, 0.517]	0.487 [0.445, 0.529]	0.487 [0.445, 0.529]	0.493 [0.449, 0.536]
	Vanilla-SH	0.450 [0.408, 0.491]	0.462 [0.418, 0.506]	0.470 [0.422, 0.518]	0.473 [0.423, 0.522]	0.478 [0.428, 0.528]
	Successive Rejects	0.327 [0.293, 0.361]	0.348 [0.304, 0.392]	0.349 [0.296, 0.403]	0.355 [0.294, 0.415]	0.357 [0.292, 0.421]
	Adjacent-BAI	0.284 [0.247, 0.321]	0.289 [0.241, 0.336]	0.284 [0.233, 0.336]	0.289 [0.231, 0.347]	0.288 [0.230, 0.346]
	P1-RAGE	0.262 [0.245, 0.279]	0.263 [0.248, 0.279]	0.266 [0.249, 0.282]	0.266 [0.249, 0.283]	0.265 [0.248, 0.281]
16	SignCount-SH (Bernstein)	0.422 [0.384, 0.460]	**0.522 [0.471, 0.572]**	**0.531 [0.490, 0.572]**	**0.557 [0.518, 0.596]**	**0.578 [0.535, 0.621]**
	SignCount-SH (Hoeffding)	0.380 [0.343, 0.418]	0.413 [0.365, 0.462]	0.428 [0.373, 0.483]	0.435 [0.381, 0.489]	0.441 [0.387, 0.494]
	Full-prefix SH	0.359 [0.320, 0.398]	0.385 [0.334, 0.435]	0.395 [0.339, 0.452]	0.403 [0.345, 0.460]	0.406 [0.347, 0.465]
	Vanilla-SH	0.380 [0.343, 0.418]	0.413 [0.365, 0.462]	0.428 [0.373, 0.483]	0.435 [0.381, 0.489]	0.441 [0.387, 0.494]
	Successive Rejects	**0.423 [0.383, 0.464]**	0.445 [0.398, 0.491]	0.458 [0.408, 0.508]	0.462 [0.410, 0.515]	0.465 [0.411, 0.519]
	Adjacent-BAI	0.340 [0.298, 0.383]	0.344 [0.294, 0.394]	0.345 [0.290, 0.400]	0.343 [0.287, 0.400]	0.346 [0.288, 0.404]
	P1-RAGE	0.343 [0.323, 0.363]	0.343 [0.321, 0.365]	0.339 [0.316, 0.363]	0.344 [0.322, 0.366]	0.345 [0.322, 0.368]

Table 6: Morning mean next-slot simple regret under the reward-query batch-size sweep with 95% confidence intervals.

B/K	Method	$q = 0.10$	$q = 0.20$	$q = 0.30$	$q = 0.40$	$q = 0.50$
4	SignCount-SH (Bernstein)	0.00564 [0.00517, 0.00610]	0.00516 [0.00473, 0.00560]	0.00462 [0.00416, 0.00508]	0.00405 [0.00350, 0.00460]	0.00368 [0.00303, 0.00432]
	SignCount-SH (Hoeffding)	0.00565 [0.00518, 0.00612]	0.00545 [0.00497, 0.00592]	0.00539 [0.00490, 0.00588]	0.00532 [0.00485, 0.00579]	0.00535 [0.00483, 0.00586]
	Full-prefix SH	0.00416 [0.00364, 0.00467]	0.00384 [0.00327, 0.00441]	0.00367 [0.00306, 0.00428]	0.00360 [0.00295, 0.00425]	0.00354 [0.00286, 0.00421]
	Vanilla-SH	0.00565 [0.00518, 0.00611]	0.00545 [0.00498, 0.00592]	0.00539 [0.00490, 0.00588]	0.00532 [0.00485, 0.00579]	0.00535 [0.00483, 0.00586]
	Successive Rejects	**0.00380 [0.00320, 0.00440]**	**0.00333 [0.00270, 0.00396]**	**0.00313 [0.00250, 0.00377]**	**0.00306 [0.00240, 0.00372]**	0.00299 [0.00234, 0.00364]
	Adjacent-BAI	0.00422 [0.00361, 0.00482]	0.00345 [0.00279, 0.00410]	0.00316 [0.00252, 0.00381]	0.00306 [0.00240, 0.00371]	**0.00299 [0.00235, 0.00363]**
	P1-RAGE	0.00763 [0.00722, 0.00804]	0.00726 [0.00682, 0.00770]	0.00694 [0.00649, 0.00738]	0.00694 [0.00651, 0.00737]	0.00697 [0.00653, 0.00742]
8	SignCount-SH (Bernstein)	0.00758 [0.00700, 0.00816]	**0.00499 [0.00434, 0.00563]**	**0.00381 [0.00322, 0.00440]**	**0.00318 [0.00261, 0.00374]**	**0.00281 [0.00227, 0.00335]**
	SignCount-SH (Hoeffding)	0.00804 [0.00738, 0.00870]	0.00745 [0.00667, 0.00823]	0.00704 [0.00620, 0.00789]	0.00677 [0.00591, 0.00763]	0.00665 [0.00576, 0.00754]
	Full-prefix SH	**0.00683 [0.00614, 0.00752]**	0.00614 [0.00538, 0.00690]	0.00573 [0.00495, 0.00650]	0.00554 [0.00473, 0.00636]	0.00542 [0.00458, 0.00627]
	Vanilla-SH	0.00804 [0.00738, 0.00870]	0.00745 [0.00667, 0.00823]	0.00704 [0.00620, 0.00789]	0.00677 [0.00591, 0.00763]	0.00665 [0.00576, 0.00754]
	Successive Rejects	0.01024 [0.00944, 0.01103]	0.00936 [0.00845, 0.01028]	0.00898 [0.00798, 0.00998]	0.00867 [0.00757, 0.00976]	0.00851 [0.00731, 0.00971]
	Adjacent-BAI	0.01148 [0.01068, 0.01228]	0.01110 [0.01027, 0.01192]	0.01099 [0.01012, 0.01185]	0.01077 [0.00991, 0.01164]	0.01079 [0.00986, 0.01172]
	P1-RAGE	0.01298 [0.01210, 0.01386]	0.01298 [0.01208, 0.01388]	0.01284 [0.01197, 0.01370]	0.01286 [0.01196, 0.01376]	0.01291 [0.01202, 0.01379]
16	SignCount-SH (Bernstein)	0.00756 [0.00655, 0.00856]	**0.00389 [0.00320, 0.00458]**	**0.00336 [0.00277, 0.00395]**	**0.00289 [0.00242, 0.00335]**	**0.00258 [0.00220, 0.00295]**
	SignCount-SH (Hoeffding)	0.00886 [0.00782, 0.00990]	0.00837 [0.00714, 0.00960]	0.00807 [0.00673, 0.00942]	0.00807 [0.00669, 0.00945]	0.00783 [0.00645, 0.00920]
	Full-prefix SH	0.00921 [0.00822, 0.01020]	0.00889 [0.00773, 0.01004]	0.00860 [0.00733, 0.00987]	0.00857 [0.00731, 0.00982]	0.00847 [0.00719, 0.00974]
	Vanilla-SH	0.00886 [0.00782, 0.00990]	0.00837 [0.00714, 0.00960]	0.00807 [0.00673, 0.00942]	0.00807 [0.00669, 0.00945]	0.00783 [0.00645, 0.00920]
	Successive Rejects	**0.00655 [0.00538, 0.00771]**	0.00592 [0.00463, 0.00721]	0.00566 [0.00427, 0.00706]	0.00549 [0.00410, 0.00688]	0.00540 [0.00398, 0.00683]
	Adjacent-BAI	0.00886 [0.00737, 0.01035]	0.00871 [0.00703, 0.01040]	0.00862 [0.00684, 0.01041]	0.00859 [0.00674, 0.01045]	0.00854 [0.00666, 0.01041]
	P1-RAGE	0.00935 [0.00865, 0.01005]	0.00934 [0.00861, 0.01007]	0.00927 [0.00852, 0.01002]	0.00920 [0.00846, 0.00993]	0.00920 [0.00841, 0.00998]

J Ablation on Non-Stationarity in NYC Taxi-Trip Dataset

We repeat the complete real-world evaluation on a shifted afternoon block. The recommended-for slots begin at 14:00, 15:30, and 17:00, and the corresponding decision slots begin at 12:30, 14:00, and 15:30. We find that the full-population leader in decision slot t remains a population leader in slot $t+1$ in 63.1% of the afternoon tasks, compared with 65.9% of the morning tasks. The average gap between a next-slot best arm and its strongest runner-up is 0.0107 in the morning and 0.0050 in the afternoon. Thus both of these effects contribute to afternoon setting being more challenging than morning setting.

We report analogous results as in Section 4.2. Though at budgets $B/K = 6, 8, 16$, our proposed methods perform better, their relative advantage over stationary baselines are reduced due to the more challenging environment.

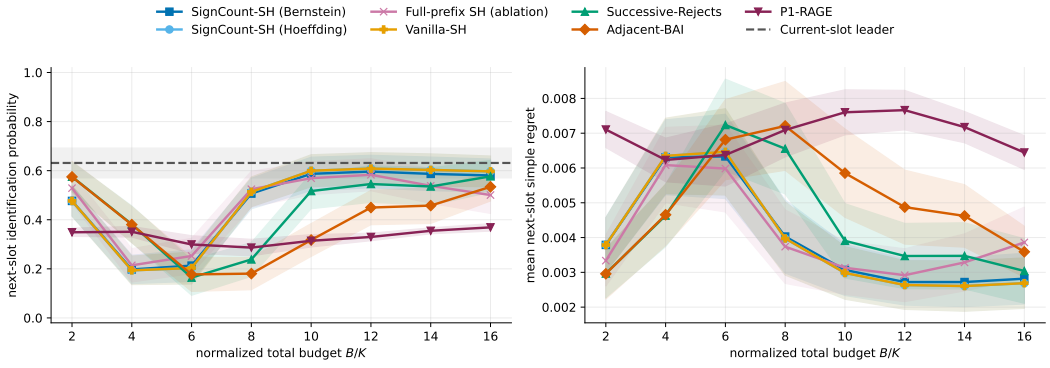


Figure 11: *Afternoon timespan mean next-slot best arm identification probability (left) and mwan simple regret (right). Bands are 95% confidence intervals.*

Table 7: *Afternoon next-slot best arm identification probability with 95% confidence intervals.*

Method	$B/K=2$	$B/K=4$	$B/K=6$	$B/K=8$	$B/K=16$
SignCount-SH (Bernstein)	0.477 [0.415, 0.538]	0.197 [0.139, 0.254]	**0.213 [0.149, 0.277]**	0.508 [0.446, 0.570]	0.582 [0.515, 0.648]
SignCount-SH (Hoeffding)	0.477 [0.415, 0.538]	0.194 [0.136, 0.251]	0.203 [0.139, 0.268]	**0.514 [0.453, 0.574]**	**0.597 [0.534, 0.660]**
Full-prefix SH	0.529 [0.468, 0.590]	0.214 [0.146, 0.282]	**0.252 [0.181, 0.324]**	**0.524 [0.452, 0.597**	0.501 [0.426, 0.575]
Vanilla-SH	0.477 [0.415, 0.538]	0.194 [0.136, 0.252]	0.203 [0.139, 0.268]	**0.514 [0.453, 0.574]**	**0.597 [0.534, 0.660]**
Successive Rejects	**0.575 [0.515, 0.635]**	**0.382 [0.307, 0.456]**	0.164 [0.092, 0.236]	0.238 [0.170, 0.307]	0.577 [0.512, 0.641]
Adjacent-BAI (MAB)	0.575 [0.512, 0.637]	0.380 [0.307, 0.454]	0.177 [0.108, 0.246]	0.180 [0.115, 0.246]	0.534 [0.477, 0.592]
P1-RAGE	0.349 [0.332, 0.366]	0.351 [0.330, 0.372]	0.299 [0.264, 0.335]	0.287 [0.253, 0.320]	0.369 [0.354, 0.384]

Table 8: *Afternoon mean next-slot simple regret with 95% confidence intervals.*

Method	$B/K=2$	$B/K=4$	$B/K=6$	$B/K=8$	$B/K=16$
SignCount-SH (Bernstein)	0.00379 [0.00300, 0.00457]	0.00632 [0.00524, 0.00740]	**0.00633 [0.00513, 0.00754]**	0.00402 [0.00297, 0.00507]	0.00281 [0.00208, 0.00355]
SignCount-SH (Hoeffding)	0.00379 [0.00301, 0.00457]	0.00635 [0.00526, 0.00743]	0.00646 [0.00522, 0.00770]	**0.00397 [0.00292, 0.00502]**	**0.00269 [0.00197, 0.00341]**
Full-prefix SH	0.00334 [0.00261, 0.00406]	0.00608 [0.00500, 0.00717]	**0.00598 [0.00474, 0.00723]**	**0.00374 [0.00268, 0.00479]**	0.00386 [0.00285, 0.00487]
Vanilla-SH	0.00379 [0.00300, 0.00457]	0.00635 [0.00526, 0.00743]	0.00647 [0.00523, 0.00770]	0.00397 [0.00292, 0.00502]	**0.00269 [0.00197, 0.00341]**
Successive Rejects	0.00297 [0.00227, 0.00366]	**0.00463 [0.00374, 0.00553]**	0.00724 [0.00592, 0.00855]	0.00656 [0.00526, 0.00786]	0.00304 [0.00211, 0.00396]
Adjacent-BAI (MAB)	**0.00296 [0.00223, 0.00368]**	0.00466 [0.00377, 0.00554]	0.00681 [0.00566, 0.00796]	0.00721 [0.00593, 0.00849]	0.00359 [0.00286, 0.00432]
P1-RAGE	0.00711 [0.00659, 0.00762]	0.00623 [0.00561, 0.00686]	0.00637 [0.00548, 0.00727]	0.00709 [0.00632, 0.00786]	0.00645 [0.00596, 0.00693]

K Target-Time BAI on the NYC Taxi Dataset

Motivation and leakage-safe protocol. The main real-data experiment evaluates a practically motivated recommendation for slot $t+1$. That metric necessarily combines two sources of error: whether the algorithm identifies the best arm from observations available at t, and whether the arm ordering changes between t and $t+1$. To isolate the first component, we repeat the experiment as target-time BAI: the algorithm observes the same 24-slot reverse-chronological history ending at t, recommends an arm for t, and is evaluated against the arm ordering in t itself.

Because the target slot is now also observable to the algorithm, we reinstate the deterministic trip-level split constructed before any experiment. Each trip is assigned, independently of its pickup zone, to either an observation pool or a held-out truth pool. All reward queries use only the observation pool, whereas correctness and regret use only the disjoint truth pool.

The observation pool contains approximately one half of the trips in each slot. If its size in historical slot s is N_s^{obs}, a reward query uses $m_s^{\mathrm{obs}} \min\{4000, N_s^{\mathrm{obs}}\}$ sampled trips. This approximately matches the effective $\min\{4000, \lfloor N_s/2 \rfloor\}$ query size in the forecast experiment, while making

observation and evaluation samples exactly disjoint. A query still costs one budget unit irrespective of m_s^{obs}.

We retain the same four design-selected zones, budgets, algorithms and timespans as in Section 4. The figures and tables below indicate that `SignCount-SH (Bernstein)` performs favorably across many budgets in the morning timespan. However, in the afternoon timespan where the inter-arm gaps are more challenging, `SignCount-SH (Bernstein)` falls back to the performance of Vanilla-SH owing to the tie-breaks.

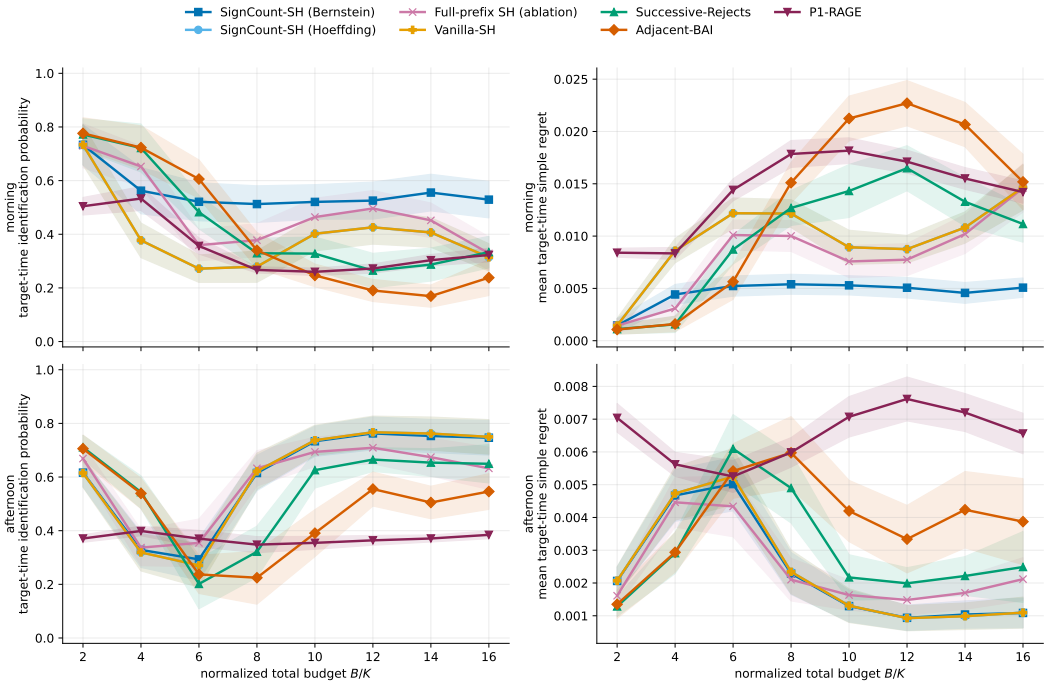


Figure 12: *Mean Target-time best-arm identification probability (left) and mean simple regret (right) for the morning (top) and afternoon (bottom) spans. Algorithms use only the observation-pool histories ending at t and are scored against the disjoint truth pool in t. Bands are 95% confidence intervals.*

Table 9: *Morning mean target-time best-arm identification probability with 95% confidence intervals.*

Method	$B/K = 2$	$B/K = 4$	$B/K = 6$	$B/K = 8$	$B/K = 16$
SignCount-SH (Bernstein)	0.733 [0.659, 0.808]	0.562 [0.477, 0.648]	0.521 [0.450, 0.592]	**0.513** [**0.445**, **0.581**]	**0.529** [**0.461**, **0.597**]
SignCount-SH (Hoeffding)	0.733 [0.659, 0.808]	0.377 [0.313, 0.442]	0.272 [0.221, 0.322]	0.279 [0.221, 0.337]	0.314 [0.264, 0.365]
Full-prefix SH	0.730 [0.648, 0.811]	0.653 [0.577, 0.729]	0.360 [0.302, 0.417]	0.378 [0.316, 0.440]	0.329 [0.267, 0.390]
Vanilla-SH	0.733 [0.659, 0.808]	0.377 [0.313, 0.442]	0.272 [0.221, 0.322]	0.279 [0.221, 0.337]	0.314 [0.264, 0.365]
Successive-Rejects	0.772 [0.714, 0.830]	0.722 [0.632, 0.811]	0.482 [0.421, 0.544]	0.329 [0.273, 0.385]	0.333 [0.275, 0.392]
Adjacent-BAI	**0.776** [**0.718**, **0.834**]	**0.723** [**0.642**, **0.804**]	**0.606** [**0.535**, **0.677**]	0.340 [0.274, 0.406]	0.238 [0.172, 0.305]
P1-RAGE	0.504 [0.472, 0.537]	0.533 [0.489, 0.577]	0.356 [0.327, 0.386]	0.267 [0.256, 0.278]	0.323 [0.298, 0.348]

Table 10: *Morning target-time mean simple regret with 95% confidence intervals.*

Method	$B/K = 2$	$B/K = 4$	$B/K = 6$	$B/K = 8$	$B/K = 16$
SignCount-SH (Bernstein)	0.00145 [0.00077, 0.00213]	0.00442 [0.00347, 0.00537]	**0.00523** [**0.00426**, **0.00619**]	**0.00540** [**0.00445**, **0.00634**]	**0.00507** [**0.00415**, **0.00599**]
SignCount-SH (Hoeffding)	0.00145 [0.00077, 0.00213]	0.00860 [0.00745, 0.00975]	0.01219 [0.01073, 0.01364]	0.01215 [0.01083, 0.01347]	0.01467 [0.01251, 0.01682]
Full-prefix SH	0.00144 [0.00079, 0.00209]	0.00308 [0.00217, 0.00399]	0.01010 [0.00868, 0.01153]	0.01001 [0.00854, 0.01148]	0.01460 [0.01230, 0.01690]
Vanilla-SH	0.00145 [0.00077, 0.00213]	0.00860 [0.00745, 0.00975]	0.01219 [0.01073, 0.01364]	0.01215 [0.01083, 0.01347]	0.01467 [0.01251, 0.01682]
Successive-Rejects	0.00110 [0.00065, 0.00156]	**0.00156** [**0.00079**, **0.00234**]	0.00872 [0.00733, 0.01010]	0.01269 [0.01102, 0.01436]	0.01116 [0.00942, 0.01289]
Adjacent-BAI	**0.00107** [**0.00062**, **0.00152**]	0.00160 [0.00087, 0.00233]	0.00564 [0.00396, 0.00732]	0.01510 [0.01317, 0.01704]	0.01518 [0.01255, 0.01782]
P1-RAGE	0.00841 [0.00799, 0.00884]	0.00834 [0.00778, 0.00891]	0.01442 [0.01337, 0.01547]	0.01785 [0.01657, 0.01913]	0.01421 [0.01315, 0.01527]

Table 11: *Afternoon mean target-time best-arm identification probability with 95% confidence intervals.*

Method	$B/K = 2$	$B/K = 4$	$B/K = 6$	$B/K = 8$	$B/K = 16$
SignCount-SH (Bernstein)	0.616 [0.564, 0.669]	0.328 [0.263, 0.393]	0.293 [0.216, 0.370]	0.615 [0.549, 0.681]	0.747 [0.682, 0.811]
SignCount-SH (Hoeffding)	0.616 [0.563, 0.668]	0.319 [0.250, 0.389]	0.271 [0.190, 0.352]	0.621 [0.554, 0.689]	**0.750** [**0.686**, **0.814**]
Full-prefix SH	0.669 [0.619, 0.719]	0.337 [0.270, 0.404]	0.355 [0.266, 0.443]	**0.633** [**0.567**, **0.698**]	0.632 [0.554, 0.711]
Vanilla-SH	0.615 [0.562, 0.668]	0.319 [0.250, 0.388]	0.271 [0.190, 0.352]	0.621 [0.554, 0.689]	**0.750** [**0.686**, **0.814**]
Successive-Rejects	**0.711** [**0.666**, **0.755**]	**0.544** [**0.483**, **0.604**]	0.201 [0.109, 0.294]	0.321 [0.226, 0.417]	0.649 [0.578, 0.720]
Adjacent-BAI	0.706 [0.655, 0.756]	0.539 [0.483, 0.596]	0.237 [0.166, 0.308]	0.225 [0.126, 0.323]	0.546 [0.480, 0.612]
P1-RAGE	0.371 [0.357, 0.385]	0.399 [0.381, 0.417]	**0.370** [**0.340**, **0.401**]	0.348 [0.318, 0.378]	0.384 [0.370, 0.399]

Table 12: *Afternoon target-time mean simple regret with 95% confidence intervals.*

Method	$B/K = 2$	$B/K = 4$	$B/K = 6$	$B/K = 8$	$B/K = 16$
SignCount-SH (Bernstein)	0.00206 [0.00164, 0.00248]	0.00467 [0.00388, 0.00546]	0.00502 [0.00420, 0.00583]	0.00229 [0.00165, 0.00293]	**0.00109** [**0.00064**, **0.00154**]
SignCount-SH (Hoeffding)	0.00207 [0.00164, 0.00249]	0.00473 [0.00392, 0.00554]	0.00525 [0.00440, 0.00610]	0.00234 [0.00168, 0.00301]	0.00110 [0.00063, 0.00157]
Full-prefix SH	0.00161 [0.00128, 0.00195]	0.00446 [0.00371, 0.00522]	**0.00434** [**0.00341**, **0.00526**]	**0.00210** [**0.00145**, **0.00275**]	0.00212 [0.00147, 0.00277]
Vanilla-SH	0.00207 [0.00165, 0.00249]	0.00473 [0.00392, 0.00554]	0.00525 [0.00440, 0.00610]	0.00234 [0.00168, 0.00301]	0.00110 [0.00063, 0.00157]
Successive-Rejects	**0.00128** [**0.00098**, **0.00159**]	**0.00291** [**0.00229**, **0.00353**]	0.00610 [0.00505, 0.00714]	0.00490 [0.00383, 0.00596]	0.00249 [0.00141, 0.00357]
Adjacent-BAI	0.00135 [0.00093, 0.00178]	0.00294 [0.00233, 0.00354]	0.00542 [0.00459, 0.00624]	0.00597 [0.00487, 0.00708]	0.00387 [0.00256, 0.00518]
P1-RAGE	0.00704 [0.00661, 0.00748]	0.00562 [0.00526, 0.00598]	0.00525 [0.00476, 0.00574]	0.00599 [0.00554, 0.00643]	0.00656 [0.00595, 0.00718]