
Fixed-Confidence Best-Arm Identification for Causal Mediation Analysis

Harsh Shrivastava¹

Yuta Kawakami¹

Junpei Komiyama^{1,2}

Jin Tian¹

¹Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

²RIKEN AIP, Tokyo, Japan

Abstract

This paper studies the problem of identifying the treatment that maximizes the expected natural direct potential outcome (NDPO), which captures the potential outcome of an intervention while excluding the pathway transmitted through a mediator that researchers may wish to remove from evaluation. We first establish population-level identification of the expected NDPO in a causal bandit setting using observable interventional distributions. We then develop a fixed-confidence best-arm identification (BAI) algorithm based on the Track-and-Stop (TaS) framework, employing a cutting-set method to solve the resulting semi-infinite optimization problem. The proposed algorithm achieves sample-efficient identification with a high-probability correctness guarantee. We prove that it satisfies δ -correctness and asymptotic optimality. Finally, we validate the approach through empirical evaluations on a large-scale real-world advertising dataset (IPinYou).

1 INTRODUCTION

Causal mediation analysis studies how a treatment (X) influences an outcome (Y) through different pathways [Sobel, 1982, Baron and Kenny, 1986, Robins and Greenland, 1992, Avin et al., 2005, Pearl, 2009, Imai et al., 2010a]. To elucidate causal mechanisms in nonlinear models, Pearl [2001] introduced *path-specific effects*, including the expected *natural direct effect* (NDE) via a mediator (Z), defined as $\mathbb{E}[Y_{x,Z_{x_0}}] - \mathbb{E}[Y_{x_0}]$, where $Y_{x,Z_{x_0}}$ denotes the potential outcome under treatment level x with the mediator set to the value it would attain under a reference level x_0 , and Y_{x_0} denotes the potential outcome under treatment level x_0 . The expected NDE measures the causal effect of a treatment on an outcome that is not mediated through the mediator.

In causal bandit problems, the conventional objective is to maximize the interventional mean under arm (treatment) $X = x$, i.e., $\mathbb{E}[Y_x] = \mathbb{E}[Y|do(X = x)]$ [Lattimore et al., 2016, Lee and Bareinboim, 2018, 2019]. However, the interventional mean aggregates effects along all causal pathways, including mechanisms that the researcher may wish to exclude from evaluation (e.g., unethical factor or positioning bias in advertisements). To isolate the influence via the direct pathway, this paper studies the problem of identifying the arm that maximizes the expected *natural direct potential outcome* (NDPO), i.e., $\mathbb{E}[Y_{x,Z_{x_0}}]$, in a causal bandit setting. The NDPO corresponds to the first term of the NDE and is defined via a nested counterfactual.

In a causal bandit setting, actions correspond to interventions on X [Lattimore et al., 2016, Lee and Bareinboim, 2018], and one observes i.i.d. samples from the interventional distribution $\mathbb{P}(Z, Y|do(X = x))$. The identification results of Pearl [2001] are not directly formulated for this setting. Instead, Pearl [2001] established the identification using $\mathbb{P}(Y_{x,z})$ and $\mathbb{P}(Z_{x_0})$. Therefore, we first establish the corresponding identification assumptions of the expected NDPO at the population level using the interventional distribution $\mathbb{P}(Z, Y|do(X = x))$.

Moreover, in the bandit setting, a further objective is to identify the target arm using as few samples as possible while ensuring correctness with high probability. Such a problem has been extensively studied in the fixed-confidence best-arm identification (BAI) framework for multi-armed bandits [Mannor and Tsitsiklis, 2004, Glynn and Juneja, 2004, Audibert and Bubeck, 2010, Jamieson and Nowak, 2014, Garivier and Kaufmann, 2016].

We subsequently investigate the fixed-confidence BAI problem for identifying the expected NDPO-optimal arm, which maximizes the expected NDPO, leveraging the established population-level identification results. To develop a BAI algorithm for identifying the expected NDPO-optimal arm, we first derive a lower bound on the sample complexity required to identify the arm that maximizes the expected

NDPO. Compared with the standard BAI framework [Garivier and Kaufmann, 2016], the BAI for the expected NDPO-optimal arm is more challenging because evaluating the expected NDPO for a given arm x requires combining information from the outcome distribution under x conditional on $Z = z$, $\mathbb{P}(Z, Y|do(X = x))$, with the mediator distribution under the different baseline arm x_0 , $\mathbb{P}(Z|do(X = x_0))$.

We then develop a new BAI algorithm (TaS-NDPO) for identifying the expected NDPO-optimal arm based on the Track-and-Stop (TaS) framework [Garivier and Kaufmann, 2016], using a cutting-set method [Mutapcic and Boyd, 2009] to solve the resulting semi-infinite optimization problem. We prove that it satisfies the desired properties of δ -correctness and asymptotic optimality. δ -correctness guarantees that the algorithm identifies the optimal-arm with probability at least $1 - \delta$ and asymptotic optimality refers to achieving the lower bound of sample complexity in the limit as $\delta \rightarrow 0$.

Finally, we conduct empirical evaluations on a large-scale real-world advertising dataset (IPinYou) to demonstrate the effectiveness of the proposed method.

2 NOTATION AND BACKGROUND

In this section, we introduce notation and background material. Uppercase letters (e.g., X, Z, Y) denote random variables, and lowercase letters (e.g., x, z, y) denote their realizations. Data-dependent quantities such as the stopping time τ_δ and the recommendation $\hat{x}(\tau_\delta)$ are random variables. Let $\mathbf{1}\{\cdot\}$ denote the indicator function. For two probability distributions P and Q over $\mathcal{A}$, the KL divergence is $\text{KL}(P\|Q) := \sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)}$. An event happens almost surely (a.s.) if it occurs with probability 1.

Structural Causal Models. We use the language of Structural Causal Models (SCM) as our basic semantic and inferential framework [Pearl, 2009]. An SCM $\mathcal{M}$ is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, \mathbb{P}_U \rangle$, where $\mathbf{U}$ is a set of exogenous (unobserved) variables following a joint distribution $\mathbb{P}_U$, and $\mathbf{V}$ is a set of endogenous (observable) variables whose values are determined by structural functions $\mathcal{F} = \{f_{V_i}\}_{V_i \in \mathbf{V}}$ such that $v_i := f_{V_i}(\mathbf{pa}_{V_i}, \mathbf{u}_{V_i})$ where $\mathbf{pa}_{V_i} \subseteq \mathbf{V}$ and $U_{V_i} \subseteq \mathbf{U}$. Each SCM $\mathcal{M}$ induces an observational distribution $\mathbb{P}_V$ over $\mathbf{V}$, and a causal graph $G(\mathcal{M})$ over $\mathbf{V}$ in which there exists a directed edge from every variable in $\mathbf{pa}_{V_i}$ to V_i . An intervention of setting a set of endogenous variables $\mathbf{X}$ to constants $\mathbf{x}$, denoted by $do(\mathbf{x})$, replaces the original equations of $\mathbf{X}$ by the constants $\mathbf{x}$ and induces a sub-model $\mathcal{M}_x$. We denote the potential outcome Y under intervention $do(\mathbf{x})$ by $Y_x(\mathbf{u})$, which is the solution of Y in the sub-model $\mathcal{M}_x$ given $\mathbf{U} = \mathbf{u}$.

Causal Mediation Analysis. Let X be a treatment variable (arm), Y be an outcome, and Z be a mediator variable. The following nonparametric SCM ($\mathcal{M}$) is used in causal

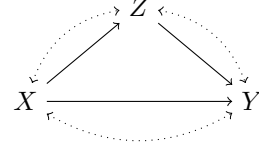


Figure 1: A causal graph representing $\mathcal{M}$.

mediation analysis.

$$\begin{aligned} \mathcal{M} : X &:= f_X(U_X), Z := f_Z(X, U_Z), \\ Y &:= f_Y(X, Z, U_Y), \end{aligned} \quad (1)$$

where U_X , U_Z , and U_Y are latent exogenous variables, with bidirected edges indicating unmeasured confounders affecting the variables. Figure 1 shows a causal graph representing $\mathcal{M}$. Under SCM $\mathcal{M}$, Pearl [2001] defined the unit-level NDE given a reference value x_0 as $Y_{x, Z_{x_0}}(\mathbf{u}) - Y_{x_0}(\mathbf{u})$ for each subject $\mathbf{u}$, and the expected NDE given a reference value x_0 as $\mathbb{E}[Y_{x, Z_{x_0}}] - \mathbb{E}[Y_{x_0}]$. [Pearl, 2001] provided identification conditions of the expected NDE, that is $Y_{x,z} \perp\!\!\!\perp Z_{x_0}$ for any x and z .

Throughout this paper, we consider SCM $\mathcal{M}$ with a discrete arm (treatment) variable $X \in \{0, 1, \dots, K-1\} =: \mathcal{X}$, and a discrete mediator $Z \in \{0, 1, \dots, M-1\} =: \mathcal{Z}$, and a binary outcome $Y \in \{0, 1\} =: \mathcal{Y}$. We fix a baseline arm $x_0 \in \mathcal{X}$. The choice of x_0 is predetermined based on domain knowledge in practice.

Best-arm identification. In the fixed-confidence best-arm identification (BAI) setting, a learner sequentially samples arms and aims to identify the optimal-arm with probability at least $1 - \delta$ for a prescribed confidence level $\delta \in (0, 1)$. A sequential algorithm consists of a sampling rule, a stopping time τ_δ , and a recommendation rule $\hat{x}(\tau_\delta)$. An algorithm is said to be δ -correct if $\mathbb{P}(\tau_\delta < \infty, \hat{x}(\tau_\delta) \neq x^*) \leq \delta$, where x^* denotes the optimal-arm under the true model. The objective is to design a δ -correct algorithm that minimizes the expected stopping time $\mathbb{E}[\tau_\delta]$.

3 RESEARCH OBJECTIVE

In this paper, we focus exclusively on the first term of the NDE, which we call the natural direct potential outcome (NDPO).

Definition 1 (NDPO). For each subject $\mathbf{u}$, we define the unit-level natural direct potential outcome (NDPO) at a reference x_0 as $Y_{x, Z_{x_0}}(\mathbf{u})$. Its population-level expectation is defined as $\mathbb{E}[Y_{x, Z_{x_0}}] := \mathbb{E}_U[Y_{x, Z_{x_0}}(\mathbf{U})]$.

We do not study the contrast with the baseline Y_{x_0} . The expectation of NDPO means: “the expected outcome if we set the arm X to x , while keeping the mediator Z at the level it would have taken under the reference arm x_0 .” NDPO is

a nested counterfactual quantity that measures the influence of the intervention $X = x$, removing the effect transmitted through Z .

In causal bandit problems, the conventional objective is to maximize the interventional mean $\mathbb{E}[Y_x] = \mathbb{E}[Y|do(X = x)]$ [Lattimore et al., 2016, Lee and Bareinboim, 2018, 2019]. This paper studies the maximization of the expectation of the NDPO. We denote the optimal-arm w.r.t. the expected NDPO under SCM $\mathcal{M}$ as follows:

$$x^* := \arg \max_{x \in \mathcal{X}} \mathbb{E}[Y_{x,Z_{x_0}}]. \quad (2)$$

The optimal-arm w.r.t. the expected NDPO is the arm that maximizes the expectation of NDPO. We present motivating examples illustrating why we study the maximization of the expectation of NDPO in Appendix A.

4 BRIDGING CAUSAL INFERENCE AND BEST ARM IDENTIFICATION

Population-level identification of causal quantities has been extensively studied within the causal inference framework [Pearl, 2009]. In this section, we establish the population-level identification of the expected NDPO using interventional distribution $\mathbb{P}(Z, Y|do(X = x))$ for a discrete Y and a connection between the causal inference framework and best arm identification (BAI) theory [Garivier and Kaufmann, 2016]. All proofs are provided in Appendix F.

Population-level causal identification of the expected NDPO under the causal bandit setting. We consider a causal bandit setting in which researchers sequentially collect samples by pulling arms in order to identify the optimal-arm w.r.t. NDPO. Pulling arm x corresponds to performing the intervention $do(X = x)$ [Lattimore et al., 2016]. In our mediation setting, pulling arm x yields i.i.d. samples from the interventional distribution $\mathbb{P}(Z, Y|do(X = x))$.

To identify $\mathbb{E}[Y_{x,Z_{x_0}}]$ from $\mathbb{P}(Z, Y|do(X = x))$, we impose the following assumption on potential outcomes.

Assumption 1 (Assumption for Identification for NDPO). $Y_{x,z} \perp\!\!\!\perp Z_{x'}$ for all $x, x' \in \mathcal{X}$ and $z \in \mathcal{Z}$.

Here, x' denotes an arbitrary reference level. Pearl [2001] considers the reference-specific condition (which we call Assumption 1' here): " $Y_{x,z} \perp\!\!\!\perp Z_{x_0}$ for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ for a fixed x_0 ." Formally, Assumption 1 appears stronger, as it requires independence w.r.t. $(Z_{x'})$ for every (x') , rather than a single reference level (x_0) . However, in practice, such reference-specific assumptions are typically understood to hold irrespective of the particular choice of baseline level. That is, Assumption 1' is implicitly required to hold for any admissible x_0 . More importantly, Pearl [2001] provides a graphical characterization of Assumption 1':

$$Y \perp Z \quad \text{in } G_{\underline{X}, \underline{Z}}, \quad (3)$$

that is, Y is d-separated from Z in the graph obtained by removing all outgoing edges from X and Z . This condition is purely graphical and therefore does not depend on the specific value of x' . Consequently, under the graphical interpretation, Assumptions 1 and 1' are equivalent, as they correspond to the same d-separation condition. Neither Assumption 1 nor Assumption 1' is testable from observed data alone. In applications, what can be assessed is the plausibility of the underlying causal graph. Both assumptions basically correspond to the absence of unmeasured confounding between Z and Y .

Then, we have the following identification theorem for a discrete Y :

Theorem 1. Under SCM $\mathcal{M}$ and Assumption 1, the expected NDPO is identified by $\theta(x)$, where

$$\theta(x) = \sum_{y,z} y P_x(y|z) P_{x_0}(z), \quad (4)$$

where $P_x(y|z) = \mathbb{P}(Y = y|Z = z, do(X = x))$ and $P_{x_0}(z) = \mathbb{P}(Z = z|do(X = x))$.

This differs in form from Theorem 1 of Pearl [2001]: we express $\theta(x)$ using the joint interventional distribution $\mathbb{P}(Z, Y|do(X = x))$, rather than using $\mathbb{P}(Y_{x,z})$ and $\mathbb{P}(Z_{x_0})$ [Pearl, 2001]. The equality $\mathbb{P}(Y_{x,z}) = \mathbb{P}(Y|do(X = x), do(Z = z))$ corresponds to intervening on both X and Z , which does not match our causal bandit setting.

Connection to BAI theory. Estimation of mediated effects from a fixed dataset has been widely studied in the causal inference literature. For example, $\theta(x)$ can be directly estimated using plug-in empirical estimators of $P_x(y|z)$ and $P_{x_0}(z)$ [Imai et al., 2010b]. Moreover, VanderWeele [2010], Tchetgen and Shpitser [2012] used parametric or semiparametric regression models to estimate mediated effects.

In contrast, under a bandit setting, data are collected sequentially. In this setting, the researcher is motivated to identify the target using as few samples as possible while guaranteeing correctness with high probability [Garivier and Kaufmann, 2016]. Accordingly, our goal is to develop a BAI algorithm that identifies the arm maximizing the NDPO through $\theta(x)$ with minimal sample complexity, subject to a high-probability correctness guarantee.

We next investigate the BAI problem of identifying the arm that maximizes θ , apart from the causal inference perspective, rather than directly optimizing the expected NDPO.

5 SAMPLE COMPLEXITY LOWER BOUND

We derive a lower bound on the expected stopping time of any δ -correct algorithm that identifies the arm maximizing θ , analogous to the classical fixed-confidence BAI lower

bounds for a binary outcome $Y \in \{0, 1\}$ (for simplicity) [Lai and Robbins, 1985, Garivier and Kaufmann, 2016]. This reveals the fundamental limitations of the BAI framework.

To study BAI for θ , we introduce the following notation. Let $\mathcal{P}$ denote the set of all P , where P is a collection of interventional distributions $(P_x)_{x \in \mathcal{X}}$, where P_x denotes $\mathbb{P}(Z, Y | do(X = x))$. For each $P \in \mathcal{P}$, we define

$$\theta_P(x) := \sum_{z \in \mathcal{Z}} P_x(1|z) P_{x_0}(z), \quad (5)$$

$$x^*(P) := \arg \max_{x \in \mathcal{X}} \theta_P(x). \quad (6)$$

We impose the following assumption.

Assumption 2 (Unique optimum). *For the true model $P \in \mathcal{P}$, the maximizer $x^*(P)$ is unique.*

Let τ_δ denote the stopping time at confidence level δ , and let $\hat{x}_{\tau_\delta}$ be the recommended arm. We define a δ -correct algorithm for the optimal-arm w.r.t. θ as follows:

Definition 2 (δ -correct algorithm). *Let $\delta \in (0, 1)$. An algorithm for identifying the optimal-arm w.r.t. θ is δ -correct if, for every $P \in \mathcal{P}$,*

$$\mathbb{P}_P(\tau_\delta < \infty, \hat{x}_{\tau_\delta} \neq x^*(P)) \leq \delta. \quad (7)$$

To derive the lower bound, we consider an alternative model in which the optimal-arm differs from that in the underlying model. For a fixed model $P \in \mathcal{P}$, define the set of alternatives: $\text{Alt}(P) := \{Q \in \mathcal{P} : x^*(Q) \neq x^*(P)\}$. Equivalently, $Q \in \text{Alt}(P)$ if there exists $x \neq x^*(P)$ such that

$$\theta_Q(x) \geq \theta_Q(x^*(P)). \quad (8)$$

For two models $P, Q \in \mathcal{P}$ and an arm $x \in \mathcal{X}$, define the interventional joint distribution

$$P_x(z, y) := \mathbb{P}_P(Z = z, Y = y | do(X = x)), \quad (9)$$

and similarly $Q_x(z, y)$ for model Q . We denote the corresponding marginals and conditionals by

$$P_x(z) := \sum_{y \in \mathcal{Y}} P_x(z, y), P_x(y|z) := \frac{P_x(z, y)}{P_x(z)}, \quad (10)$$

and similarly for $Q_x(z)$ and $Q_x(y|z)$.

Since $P_x(z, y) = P_x(z)P_x(y|z)$, the KL divergence decomposes as

$$\begin{aligned} \text{KL}(P_x \| Q_x) &= \text{KL}(P_x(Z) \| Q_x(Z)) \\ &+ \sum_{z \in \mathcal{Z}} P_x(z) \text{KL}(P_x(Y|z) \| Q_x(Y|z)). \end{aligned} \quad (11)$$

The KL decomposition in (11) splits $\text{KL}(P_x \| Q_x)$ into a mediator term and a conditional outcome term weighted by $P_x(z)$. Hence discrepancies in $P_x(1|z)$ at rare mediator values contribute little to the total divergence, so distinguishing such differences requires more samples from arm x .

Then, we have the following theorem:

Theorem 2. [Instance-dependent lower bound] *Let $\delta \in (0, 1)$. Under Assumption 2, for any δ -correct algorithm and any $P \in \mathcal{P}$,*

$$\mathbb{E}_P[\tau_\delta] \geq T^*(P) \text{kl}(\delta, 1 - \delta), \quad (12)$$

where $\text{kl}(\delta, 1 - \delta) = \delta \log \frac{\delta}{1 - \delta} + (1 - \delta) \log \frac{1 - \delta}{\delta}$ and

$$(T^*(P))^{-1} = \sup_{w \in \Sigma_K} \inf_{Q \in \text{Alt}(P)} \sum_{x \in \mathcal{X}} w_x \text{KL}(P_x \| Q_x) \quad (13)$$

with $\Sigma_K := \{w \in \mathbb{R}_+^K : \sum_{x \in \mathcal{X}} w_x = 1\}$.

Here $\mathbb{R}_+^K := \{v \in \mathbb{R}^K : v_x \geq 0, \forall x\}$. The vector $w \in \Sigma_K$ represents asymptotic sampling proportions: w_x is the fraction of samples allocated to arm x in the long run. The $\text{kl}(\cdot, \cdot)$ denotes the binary relative entropy. The sup-inf form in (13) chooses proportions w that maximize the worst-case information rate for distinguishing the true model P from alternatives under which the identity of the optimal-arm differs.

Technical difficulties in BAI for θ . The lower bound in Theorem 2 highlights two essential differences from classical best-arm identification. In classical BAI, the objective is $\mu_x = \mathbb{E}_P[Y | do(X = x)]$, and the characteristic time reduces to $(T_{\text{BAI}}^*(P))^{-1} = \sup_{w \in \Sigma_K} \inf_{x \neq x^*(P)} \sum_{a \in \mathcal{X}} w_a \text{KL}(P_a \| Q_a^{(x)})$, where $Q^{(x)}$ denotes an alternative model that coincides with P on all arms except $x^*(P)$ and x , and is chosen so that x becomes optimal under $Q^{(x)}$ [Garivier and Kaufmann, 2016]. This yields a low-dimensional convex program in which, for each suboptimal-arm $x \neq x^*(P)$ (called a *competitor*), the alternative model modifies only the true best arm and the single competing arm x .

In contrast, θ couples outcome behavior under arm x with the mediator distribution under the baseline arm x_0 . The constraint (8) depends jointly on mediator and outcome mechanisms, so the inner minimization cannot be reduced to modifying only the best arm and a single competitor while keeping all other arms fixed, as in classical BAI. Instead, the baseline mediator distribution $Q_{x_0}(Z)$ enters the objective for every arm through $\theta_Q(\cdot)$, creating an explicit coupling across arms. Moreover, the KL decomposition (11) shows that deviations in $P_x(y|z)$ are weighted by the mediator probability $P_x(z)$. If decisive differences occur at mediator values with small probability mass, the effective information rate is reduced, increasing the characteristic time $T^*(P)$.

These two features—the non-separable alternative constraint and the information that accumulates at the level of mediator–outcome cells (x, z) —do not arise in classical BAI. We call each pair $(x, z) \in \mathcal{X} \times \mathcal{Z}$ a *cell*, corresponding to the conditional outcome parameter $P_x(y|z)$ and the frequency with which mediator value z occurs under arm x .

6 A TRACK-AND-STOP ALGORITHM FOR CAUSAL MEDIATION ANALYSIS

We adapt the Track-and-Stop (TaS) framework for fixed-confidence BAI [Garivier and Kaufmann, 2016] to the objective θ . TaS combines forced exploration, a plug-in solution of the characteristic max–min allocation problem, and a generalized likelihood ratio stopping rule. However, the structure of the lower bound in Section 5 prevents a direct application of the classical arm-local TaS analysis. Two modifications are required:

(i) Cell-level information control. By the KL decomposition (11), information accumulates at the mediator–outcome cell level (x, z) . Arm-level forced exploration alone does not ensure that all cell counts $N_{x,z}(t)$ diverge, which is necessary for consistent estimation of $P_x(y|z)$ and evaluation of the likelihood-ratio statistic used in the stopping rule.

(ii) Non-separable alternative optimization. The alternative constraint defining $\text{Alt}(P)$ couples arms through the baseline mediator distribution $P_{x_0}(Z)$. As a result, the inner minimization in the characteristic time cannot be reduced to a single best-arm–competitor pair, and instead yields a bi-convex projection problem.

By modifying these components, we develop a new algorithm, TaS-NDPO, which is presented in Algorithm 1. The following subsections detail the resulting sampling rule and the associated optimization procedure.

6.1 PLUG-IN ESTIMATES

We first define the empirical model and the corresponding plug-in estimators of $\theta(x)$. At round $s = 1, 2, \dots$, the learner selects $X_s \in \mathcal{X}$ and observes $(Z_s, Y_s) \sim P_{X_s}$. Define the arm and cell counts at the round t

$$N_x(t) := \sum_{s=1}^t \mathbf{1}\{X_s = x\}, \quad (14)$$

$$N_{x,z}(t) := \sum_{s=1}^t \mathbf{1}\{X_s = x, Z_s = z\}, \quad (15)$$

$$S_{x,z}(t) := \sum_{s=1}^t \mathbf{1}\{X_s = x, Z_s = z, Y_s = 1\}. \quad (16)$$

Then, the empirical distributions are

$$\hat{P}_x(z; t) := \frac{N_{x,z}(t)}{N_x(t)}, \hat{P}_x(1|z; t) := \frac{S_{x,z}(t)}{N_{x,z}(t)}. \quad (17)$$

The empirical joint model at time t is the collection $\hat{P}(t) := (\hat{P}_x(\cdot; t))_{x \in \mathcal{X}}$, where for each arm $x \in \mathcal{X}$,

$\hat{P}_x(\cdot; t)$ denotes the empirical joint distribution of (Z, Y) under arm x , with probability mass function $\hat{P}_x(z, y; t) := \hat{P}_x(z; t) \hat{P}_x(y|z; t)$ for $(z, y) \in \mathcal{Z} \times \{0, 1\}$. The plug-in estimate of θ is

$$\hat{\theta}_t(x) := \sum_{z \in \mathcal{Z}} \hat{P}_x(1|z; t) \hat{P}_{x_0}(z; t), \quad (18)$$

and the empirical best arm is the smallest index in the argmax set

$$\hat{x}(t) \in \arg \max_{x \in \mathcal{X}} \hat{\theta}_t(x). \quad (19)$$

6.2 COMPUTATION OF THE PLUG-IN ALLOCATION

In contrast to classical BAI, the alternative constraint couples mediator and outcome mechanisms, yielding a non-separable KL projection with bilinear structure of Eq. (13). In this section, after reducing the dimension of the optimization, we introduce a cutting-set method that solves this optimization.

Plug-in allocation. Replacing the true model P by $\hat{P}(t)$ in the characteristic-time optimization, we define the alternative set $\text{Alt}(\hat{P}(t)) := \{Q \in \mathcal{P} : x^*(Q) \neq \hat{x}(t)\}$ and the target allocation

$$\hat{w}(t) \in \arg \max_{w \in \Sigma_K} \inf_{Q \in \text{Alt}(\hat{P}(t))} \sum_{x \in \mathcal{X}} w_x \text{KL}(\hat{P}_x(t) \| Q_x). \quad (20)$$

Decomposition over competitors. Let $P \in \mathcal{P}$ admit a unique maximizer $x^*(P)$. The alternative set is given as

$$\text{Alt}(P) := \bigcup_{x' \neq x^*(P)} \{Q \in \mathcal{P} : \theta_Q(x') \geq \theta_Q(x^*(P))\}. \quad (21)$$

Accordingly, the inner infimum in (20) decomposes into $K - 1$ subproblems indexed by competitors $x' \neq x^*(P)$.

Reduction to three arms. Fix a competitor $x' \neq x^*(P)$ and consider

$$\inf_{Q \in \mathcal{P}} \sum_{a \in \mathcal{X}} w_a \text{KL}(P_a \| Q_a) \text{ s.t. } \theta_Q(x') \geq \theta_Q(x^*(P)). \quad (22)$$

Proposition 1. *[Reduction to three arms] For the inner minimization oracle in (22), it is without loss of optimality to restrict attention to solutions satisfying $Q_a = P_a$ for all $a \notin \{x_0, x^*(P), x'\}$.*

Hence, the inner minimization effectively involves only three arms: $x_0, x^*(P)$, and a single competitor x' .

Reduced inner problem. Let $x^* = x^*(P)$ and fix $x' \neq x^*$. After restriction, the free variables are $q_0(z) := Q_{x_0}(z), r^*(z) := Q_{x^*}(1|z), r'(z) := Q_{x'}(1|z), z \in \mathcal{Z}$. The feasible set is $\Delta_{|\mathcal{Z}|} \times [0, 1]^{|\mathcal{Z}|} \times [0, 1]^{|\mathcal{Z}|}$, where $\Delta_{|\mathcal{Z}|} :=$

$\{q \in \mathbb{R}_+^{|\mathcal{Z}|} : \sum_{z \in \mathcal{Z}} q(z) = 1\}$ is the probability simplex over $\mathcal{Z}$. This set is compact and convex. The objective becomes

$$\begin{aligned} & w_{x_0} \sum_z \text{KL}(\hat{P}_{x_0}(z) \| q_0(z)) \\ & + w_{x^*} \sum_z \hat{P}_{x^*}(z) \text{KL}(\hat{P}_{x^*}(1|z) \| r^*(z)) \\ & + w_{x'} \sum_z \hat{P}_{x'}(z) \text{KL}(\hat{P}_{x'}(1|z) \| r'(z)). \end{aligned} \quad (23)$$

The original constraint is

$$\sum_z r'(z) q_0(z) \geq \sum_z r^*(z) q_0(z). \quad (24)$$

This inequality must hold with equality at an optimal solution. Indeed, if there were an optimal solution for which the inequality were strict, then there would exist a sufficiently small $\alpha > 0$ such that replacing

$$\begin{aligned} r^*(z) &\leftarrow (1 - \alpha) r^*(z) + \alpha \hat{P}_{x^*}(1|z), \\ r'(z) &\leftarrow (1 - \alpha) r'(z) + \alpha \hat{P}_{x'}(1|z). \end{aligned} \quad (25)$$

reduces the objective value in (23), while not changing the sign of the difference between the left-hand side and right-hand side of the constraint. Therefore, no optimal solution can satisfy the constraint with strict inequality, and the constraint reduces to the scalar bilinear equality

$$\sum_z r'(z) q_0(z) = \sum_z r^*(z) q_0(z). \quad (26)$$

Proposition 2. *[Bi-convex structure] The optimization problem defined by (23)–(26) is convex in (r^*, r') for fixed q_0 , and convex in q_0 for fixed (r^*, r') , but not jointly convex.*

Cutting-set method. Even after reducing the dimension to three, the resulting problem is still a linear semi-infinite program in $w = (w_{x_0}, w_{x^*}, w_{x'})$. This is because the constraint involving $(r'(z), q_0(z), r^*(z))$ must hold over a continuous domain.

The cutting-set method [Mutapcic and Boyd, 2009] starts with some initial set $\mathcal{Q}^{(1)} \subset \text{Alt}(\hat{P})$.¹ At each iteration $s = 1, 2, \dots$, the method maintains a finite active set $\mathcal{Q}^{(s)} \subset \text{Alt}(\hat{P})$ and proceeds as follows: (i) solving $\max_{w \in \Sigma_K} \min_{Q \in \mathcal{Q}^{(s-1)}} \sum_x w_x \text{KL}(\hat{P}_x \| Q_x)$ on w , which is a finite-dimensional linear programming, and (ii) computing a most-violated alternative model via the reduced bi-convex problem, and add it to $\mathcal{Q}^{(s)}$.

¹In our implementation, the active set is initialized with one constraint for each competitor arm. This avoids degenerate initial allocations and ensures that each competitor is represented in the master problem. If a candidate allocation assigns zero mass to an arm a , then the competitor subproblem with $x' = a$ is selected as a hardest constraint and added to the active set; the subsequent master update then accounts for this arm and assigns it positive mass when needed.

Optimization properties. Under mild conditions, including boundedness of the objective and empirical means bounded away from 0 and 1, the cutting-set method converges to a global optimum [Mutapcic and Boyd, 2009, Section 5.2] assuming that (ii) is exact. These conditions are satisfied in our setting when the empirical estimates remain in $[\varepsilon, 1 - \varepsilon]$ for some $\varepsilon > 0$.

Regarding the optimality of (ii), we solve it via alternating minimization. The implementation details are provided in Appendix C. Under standard regularity conditions, block coordinate descent for bi-convex problems converges to a stationary point [Tseng, 2001]. However, in general, such a stationary point may not be globally optimal.

For ease of discussion, our theoretical analysis in Section 7 assumes access to the exact optimizer.

Computation. In our experiments, the optimization routine is not a major computational bottleneck: TaS-NDPO takes 22.5 seconds on average for runs with up to 100,000 samples; detailed runtime comparisons are reported in Appendix E.2.

6.3 SAMPLING RULE

We use the convention that after t pulls we have counts $N_x(t), N_{x,z}(t)$ (Section 6.1), and we choose the next arm X_{t+1} based on plug-in quantities computed from data up to time t . Given the plug-in allocation $\hat{w}(t)$, the algorithm follows a D-tracking rule augmented with (i) competitor coverage and (ii) cell-level forcing.

D-tracking. When neither coverage nor forcing is active, we apply the D-tracking rule [Garivier and Kaufmann, 2016], i.e., X_{t+1} is updated by the smallest index of $\arg \max_{x \in \mathcal{X}} (t \hat{w}_x(t) - N_x(t))$.

Competitor coverage. Let $C_x(t)$ denote the number of times arm x has been selected as the competitor up to time t , with initialization $C_x(0) = 0$ and update

$$C_x(t) = C_x(t-1) + \mathbf{1}\{x'(t) = x\}. \quad (27)$$

Fix $a \in (0, 1)$ and define $h(t) := \lceil t^a \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. After observing t samples, we select the next competitor $x'(t+1)$ as follows. If $\min_{x \in \mathcal{X}} C_x(t) < h(t)$, choose

$$x'(t+1) \in \arg \min \{C_x(t) : C_x(t) < h(t)\}, \quad (28)$$

breaking ties by the smallest index. Otherwise, choose $x'(t+1)$ as any competitor $x' \neq \hat{x}(t)$ attaining the minimum among the competitor-wise inner problems described in Section 6.2, breaking ties by the smallest index. This rule guarantees that each arm is selected as a competitor at least $h(t) - 1 = \Omega(t^a)$ times (for all sufficiently large t).

Cell-level forcing. Define the relevant arm set at time $t+1$ by $\mathcal{A}_{\text{rel}}(t+1) = \{x_0, \hat{x}(t), x'(t+1)\}$. Let $g(t)$ be nonde-

Algorithm 1 Algorithm of TaS-NDPO

Require: Confidence $\delta \in (0, 1)$; baseline arm x_0 ; $h(t) = \lceil t^a \rceil$ for some $a \in (0, 1)$; nondecreasing $g(t)$ such as $g(t) \rightarrow \infty$ and $g(t) = o(h(t))$.

- 1: Pull each arm once and update counts; set $C_x \leftarrow 0$ for all $x \in \mathcal{X}$.
- 2: **for** $t = |\mathcal{X}|, |\mathcal{X}| + 1, \dots$ **do**
- 3: Compute $\hat{P}(t), \hat{\theta}_t(\cdot), \hat{x}(t) \in \arg \max_{x \in \mathcal{X}} \hat{\theta}_t(x)$, and $\hat{w}(t)$.
- 4: Compute $Z_t(x, \hat{x}(t))$ for all $x \neq \hat{x}(t)$.
- 5: **if** $\min_{x \neq \hat{x}(t)} Z_t(x, \hat{x}(t)) \geq \beta(t, \delta)$ **then**
- 6: **return** $\hat{x}(t)$
- 7: **end if**
- 8: **Competitor coverage:**
- 9: **if** $\min_{x \in \mathcal{X}} C_x < h(t)$ **then**
- 10: $x'(t+1)$ is updated by the smallest index of $\arg \min\{C_x : C_x < h(t)\}$.
- 11: **else**
- 12: $x'(t+1) \leftarrow$ a minimizer of the competitor-wise inner problems described in Section 6.2.
- 13: **end if**
- 14: $C_{x'(t+1)} \leftarrow C_{x'(t+1)} + 1$,
 $\mathcal{A}_{\text{rel}} \leftarrow \{x_0, \hat{x}(t), x'(t+1)\}$,
 $\mathcal{D} \leftarrow \{u \in \mathcal{A}_{\text{rel}} : \min_{z \in \mathcal{Z}} N_{u,z}(t) < g(t)\}$.
- 15: **Arm selection:**
- 16: **if** $\mathcal{D} \neq \emptyset$ **then**
- 17: X_{t+1} is updated by the smallest index of $\arg \min_{u \in \mathcal{D}} (\min_{z \in \mathcal{Z}} N_{u,z}(t), N_u(t))$.
- 18: **else**
- 19: X_{t+1} is updated by the smallest index of $\arg \max_{x \in \mathcal{X}} (t\hat{w}_x(t) - N_x(t))$.
- 20: **end if**
- 21: Pull X_{t+1} ; observe (Z_{t+1}, Y_{t+1}) ; update counts.
- 22: **end for**
- 23: **return** $\hat{x}(t)$

creasing with $g(t) \rightarrow \infty$ and $g(t) = o(h(t))$. For any arm x , define its minimum cell count $m_x(t) := \min_{z \in \mathcal{Z}} N_{x,z}(t)$. Define the set of deficient relevant arms (computed from time- t counts) $\mathcal{D}(t) := \{x \in \mathcal{A}_{\text{rel}}(t+1) : m_x(t) < g(t)\}$. If $\mathcal{D}(t) \neq \emptyset$, we force exploration by choosing $X_{t+1} \in \arg \min_{u \in \mathcal{D}(t)} (m_u(t), N_u(t))$, i.e., we minimize $m_u(t)$ and break ties by the smallest total count $N_u(t)$ (and then by the smallest index). Otherwise, we perform D-tracking breaking ties by the smallest index.

Role of coverage and forcing. Competitor coverage ensures each arm is compared often enough, while cell-level forcing ensures all relevant cell counts $N_{x,z}(t)$ grow, which is needed for consistency and for the likelihood-ratio statistic in the stopping rule. Both mechanisms act on a vanishing fraction of rounds since $h(t) = t^a$ with $a \in (0, 1)$ and $g(t) = o(h(t))$, so the dominant sampling dynamics are governed by D-tracking.

6.4 STOPPING RULE

The stopping rule should guarantee δ -correctness while stopping as soon as the data exclude every model under which a suboptimal-arm could be optimal. As in classical TaS, we use a likelihood-based test against the closest alternative model.

Likelihood function. For any $Q \in \mathcal{P}$, define the negative log-likelihood ratio $\mathcal{L}_t(Q) := \sum_{a \in \mathcal{X}} N_a(t) \text{KL}(\hat{P}_a(t) \| Q_a)$, where $\hat{P}_a(t)$ denotes the empirical joint distribution of (Z, Y) under arm a .

GLRT statistic. At time t , for each competitor $x \neq \hat{x}(t)$, define the generalized likelihood ratio statistic (GLRT)

$$Z_t(x, \hat{x}(t)) := \inf_{Q \in \mathcal{P} : \theta_Q(x) \geq \theta_Q(\hat{x}(t))} \mathcal{L}_t(Q), \quad (29)$$

where $\theta_Q(x) = \sum_{z \in \mathcal{Z}} Q_x(1|z) Q_{x_0}(z)$.

The statistic $Z_t(x, \hat{x}(t))$ measures the minimal empirical KL divergence to a model under which arm x is at least as good as the current empirical best arm $\hat{x}(t)$.

Stopping condition. Let $\beta(t, \delta)$ be a nondecreasing threshold. Define the stopping time

$$\tau_\delta := \inf\{t \geq 1 : \min_{x \neq \hat{x}(t)} Z_t(x, \hat{x}(t)) \geq \beta(t, \delta)\}, \quad (30)$$

and output $\hat{x}(\tau_\delta)$. A concrete choice of $\beta(t, \delta)$ ensuring δ -correctness is provided in Appendix B.

7 THEORETICAL GUARANTEES

In this section, we establish δ -correctness and asymptotic optimality of Algorithm 1. We first assume

Assumption 3 (Uniform interior (KL-regularity)). *There exists $\varepsilon \in (0, 1/2)$ such that for every model $R \in \mathcal{P}$, every arm $x \in \mathcal{X}$, and every mediator value $z \in \mathcal{Z}$,*

$$R_x(z) \geq \varepsilon, R_x(1|z) \in [\varepsilon, 1 - \varepsilon]. \quad (31)$$

This assumption guarantees that all KL divergences in the characteristic-time optimization are finite and continuous. It ensures that the relevant probability vectors remain in the interior of the simplex, which is needed for the continuity and compactness arguments in the asymptotic analysis.

Then, our algorithm has δ -correctness.

Theorem 3. [δ -correctness] *Assume $\beta(t, \delta)$ is chosen as in Appendix B. For any $\delta \in (0, 1)$ and any $P \in \mathcal{P}$, Algorithm 1 satisfies*

$$\mathbb{P}_P(\tau_\delta < \infty, \hat{x}(\tau_\delta) \neq x^*(P)) \leq \delta. \quad (32)$$

For the asymptotic optimality result, we also assume uniqueness of the characteristic allocation.

Assumption 4 (Unique characteristic allocation). *The maximizer $w^*(P)$ of the characteristic optimization problem is unique.*

We show the following lemmas:

Lemma 1. *Under Assumption 3. Let $h(t) = \lceil t^a \rceil$ for some $a \in (0, 1)$ and let $g(t)$ be nondecreasing with $g(t) \rightarrow \infty$ and $g(t) = o(h(t))$. Under the sampling rule of Section 6.3,*

$$N_x(t) \rightarrow \infty, N_{x,z}(t) \rightarrow \infty, \forall x \in \mathcal{X}, z \in \mathcal{Z}, \text{ a.s.} \quad (33)$$

Since $N_{x,z}(t) \rightarrow \infty$ for all x, z , the strong law yields

$$\hat{P}_x(z; t) \rightarrow P_x(z), \hat{P}_x(1|z; t) \rightarrow P_x(1|z) \text{ a.s.} \quad (34)$$

Hence $\hat{P}(t) \rightarrow P$ almost surely.

Lemma 2. *[Consistency of plug-in allocation] Under Assumption 2, if $\hat{P}(t) \rightarrow P$ almost surely, then $\hat{w}(t) \rightarrow w^*(P)$ almost surely.*

Lemma 3. *[Sublinearity of forced rounds] Let $F(t)$ denote the number of rounds up to time t in which either competitor coverage or cell-level forcing overrides the D -tracking update. Then $F(t) = O(h(t)) + O(g(t))$. In particular, if $h(t) = t^a$ with $a \in (0, 1)$ and $g(t) = o(h(t))$, then $F(t) = o(t)$ almost surely.*

Lemma 4. *[Tracking of sampling proportions] Under Assumptions 3 and 4, the sampling proportions satisfy $N_x(t)/t \rightarrow w_x^*(P)$ for all $x \in \mathcal{X}$ almost surely.*

This follows from Lemmas 2 and 3.

Let $T^*(P)$ denote the characteristic time in Theorem 2.

Lemma 5. *[Asymptotic growth of the GLRT] For every competitor $x \neq x^*(P)$,*

$$\mathbb{P}_P \left(\liminf_{t \rightarrow \infty} t^{-1} Z_t(x, \hat{x}(t)) \geq \inf_{Q \in \mathcal{P}: \theta_Q(x) \geq \theta_Q(x^*(P))} \sum_{a \in \mathcal{X}} w_a^*(P) \text{KL}(P_a \| Q_a) \right) = 1. \quad (35)$$

Consequently,

$$\mathbb{P}_P \left(\liminf_{t \rightarrow \infty} t^{-1} \min_{x \neq x^*(P)} Z_t(x, \hat{x}(t)) \geq T^*(P)^{-1} \right) = 1. \quad (36)$$

Then, our algorithm has asymptotic optimality.

Theorem 4. *[Almost sure asymptotic optimality] Fix a true model $P \in \mathcal{P}$ satisfying Assumptions 2, 3 and 4. Then Algorithm 1 satisfies*

$$\mathbb{P}_P \left(\limsup_{\delta \rightarrow 0} \frac{\tau_\delta}{\text{kl}(\delta, 1 - \delta)} \leq T^*(P) \right) = 1. \quad (37)$$

Table 1: Stopping-time statistics on the IPinYou. The error rate is the proportion of runs in which the algorithm chooses a suboptimal-arm. Lower values show better performance.

Method	Median	Interquartile range	Error
TaS-NDPO (ours)	12,928	[4,864, 33,984]	0.00
TaS-NDPO (arm-level)	25,856	[11,520, 38,848]	0.00
Baseline-First	18,304	[8,960, 38,016]	0.01
Uniform	21,248	[9,472, 38,016]	0.00

Finally, under Assumption 1, the expected NDPO is identified by θ . Therefore, the δ -correctness and asymptotic optimality of our algorithm for maximizing θ translate directly into δ -correctness and asymptotic optimality for the expected NDPO under Assumption 1.

8 EXPERIMENTS

We evaluate our method on simulations based on synthetic and real-world data. The real-world experiments use two datasets: the IPinYou advertising dataset and the framing dataset from the `mediation` package. Additional synthetic experiments are reported in Appendix E. These experiments empirically support the validity of the lower bound in Theorem 2, illustrate cases where the arm maximizing $\mathbb{E}[Y|do(X=x)]$ differs from the expected NDPO-optimal arm, and study sensitivity to gap structure, mediator sparsity, baseline-arm choice, and mediator cardinality. Runtime comparisons are reported in Appendix E.2.

Dataset. We use the IPinYou dataset [Zhang et al., 2015], a publicly available benchmark for computational advertising, available at (<https://contest.ipinyou.com>). The dataset contains approximately $N = 3.16 \times 10^6$ ad impressions. From this dataset, we select $K = 10$ creatives with sufficient support (i.e., each with at least 30,000 impressions). The mediator Z corresponds to discretized slot visibility with three levels, and the outcome Y is a binary click indicator. We impose Assumption 1, corresponding to the absence of unmeasured mediator–outcome confounding under treatment interventions, which guarantees identification of the expected NDPO from interventional data.

Algorithms. We compare the following algorithms:

- **TaS-NDPO (ours).** Algorithm 1.
- **TaS-NDPO (arm-level).** Variant of TaS-NDPO by replacing cell-level forcing with classical arm-level forced exploration.
- **Baseline-First.** The algorithm first estimates the baseline distribution $P_{x_0}(z)$ using uniform sampling, and then samples arms uniformly to estimate $P_x(1|z)$.

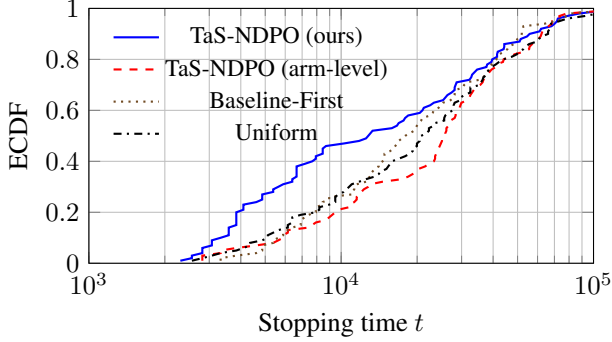


Figure 2: ECDF of stopping times on the IPinYou. Higher curves correspond to faster identification.

- **Uniform.** Uniform sampling over arms combined with the same plug-in estimator and stopping rule as ours.

To avoid zero empirical probabilities when some counts are small, we use Laplace smoothing [Manning, 2008], i.e., $\hat{P}_x(z; t) = (N_{x,z}(t) + \alpha) / (N_x(t) + \alpha|\mathcal{Z}|)$ with $\alpha = 0.05$.

Results. We first compare the arms that maximize the interventional mean $\mathbb{E}[Y|do(X = x)]$ with those that maximize the expected NDPO. On the IPinYou dataset, creative 10,722 (ID) achieves the highest interventional mean, while creative 10,720 maximizes the expected NDPO. Examining mediator distributions reveals that creative 10,722 appears in the top slot 23.7% of the time, compared to only 9.9% for creative 10,720. Thus, part of its click-rate advantage is mediated through favorable slot positioning, which is removed under the NDPO perspective.

We next evaluate all BAI algorithms with a confidence level $\delta = 0.05$ over 100 independent runs. Table 1 summarizes stopping time statistics, and Figure 2 reports the empirical cumulative distribution function (ECDF) of stopping times. As shown in Table 1, all methods exhibit empirical error rates below 0.05. Thus, all methods empirically satisfy the δ -correctness requirement. The TaS-NDPO (ours) achieves a median stopping time of 12,928 samples, reducing sample complexity by roughly 50% compared to the arm-level variant (25,856). It also improves upon uniform sampling and the two-stage baseline while maintaining zero empirical error. Figure 2 further illustrates the distributional behavior of the stopping times: the ECDF curve of TaS-NDPO lies uniformly above the alternatives, indicating consistently faster identification across runs. The separation is most pronounced in the lower quantiles, demonstrating that cell-level exploration in our algorithm accelerates early evidence accumulation. These results support that our algorithm substantially improves sample efficiency for identifying the optimal-arm w.r.t. the expected NDPO.

Additional real-world validation. We also evaluate the algorithms on the framing dataset from the `mediation` package at (<https://cran.r-project.org/web/>

Table 2: Stopping-time statistics on the framing dataset. The error rate is the proportion of runs in which the algorithm chooses a suboptimal arm. Lower values show better performance.

Method	Median	Interquartile range	Error
TaS-NDPO (ours)	5,888	[2,624, 9,408]	0.00
TaS-NDPO (arm-level)	6,400	[3,840, 10,304]	0.00
Baseline-First	8,704	[7,168, 9,984]	0.00
Uniform	7,680	[3,392, 11,518]	0.02

`packages/mediation/index.html`), using the same evaluation protocol as above. Table 2 reports the median stopping time, interquartile range, and empirical error rate. TaS-NDPO achieves the lowest median stopping time and zero empirical error, providing additional real-world validation beyond the IPinYou dataset.

The additional sensitivity experiments in Appendix E show that the benefit of cell-level forcing is regime-dependent. The gains are most substantial in settings with rare mediator-outcome cells, while the arm-level Track-and-Stop variant remains competitive in easier regimes. The gap-scaling experiment varies the NDPO gap while holding the mediator structure fixed.

9 CONCLUSION

We study the fixed-confidence BAI problem of selecting the arm that maximizes the expected NDPO and develop a TaS-based algorithm that is δ -correct and asymptotically optimal. We then evaluate its efficiency through simulation experiments.

We focus on the expected NDPO in this paper. Other related quantities include the controlled direct effect and the natural indirect effect, defined as $\mathbb{E}[Y_{x,z}] - \mathbb{E}[Y_{x_0,z}]$ and $\mathbb{E}[Y_{x_0,z_x}] - \mathbb{E}[Y_{x_0}]$, respectively [Pearl, 2001]. In Appendix D, we consider the corresponding BAI problems for optimizing $\mathbb{E}[Y_{x,z}]$ over x (for fixed z) and for optimizing $\mathbb{E}[Y_{x_0,z_x}]$ over x . The former reduces to a standard BAI problem, whereas the latter closely parallels our algorithm for the expected NDPO. Furthermore, we focused on the case of a binary outcome since it covers a wide range of applications related to website optimization. Extending our statistical framework to multi-category outcomes (e.g., five-star reviews) is relatively easy.

An interesting direction for future work is to extend these ideas to path-specific objectives in more general causal structures [Avin et al., 2005, Shpitser and Sherman, 2018, Malinsky et al., 2019].

Acknowledgements

The authors thank the anonymous reviewers for their time and thoughtful comments. J. Komiyama was supported by the MBZUAI Start-up Fund [BF0121].

References

- Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, January 2010.
- Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 357–363, 2005.
- Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.
- Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2009.
- Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.
- Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 36th Conference on Winter Simulation, WSC '04*, page 577–585. Winter Simulation Conference, 2004. ISBN 0780387864.
- Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010a.
- Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71, 2010b.
- Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th annual conference on information sciences and systems (CISS)*, pages 1–6. IEEE, 2014.
- Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17:1:1–1:42, 2016.
- T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8.
- Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.
- Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.
- Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4164–4172, 2019.
- Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088. PMLR, 2019.
- Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing, 2008.
- Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.
- Almir Mutapcic and Stephen Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 24(3):381–406, 2009.
- Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, UAI'01*, page 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.
- James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155, 1992.
- Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020. doi: 10.1287/opre.2019.1911.
- Ilya Shpitser and Eli Sherman. Identification of personalized effects associated with causal pathways. *Uncertain Artif Intell*, 2018, Aug 2018.
- Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 1982.
- Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816, 2012.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, June 2001. ISSN 1573-2878. doi: 10.1023/A:1017501703105.

Tyler J VanderWeele. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, 21(4):540–551, 2010.

Eric Zeng, Miranda Wei, Theo Gregersen, Tadayoshi Kohno, and Franziska Roesner. Polls, clickbait, and commemorative \$2 bills: problematic political advertising on news and media websites around the 2020 u.s. elections. In *Proceedings of the 21st ACM Internet Measurement Conference, IMC '21*, page 507–525, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391290.

Weinan Zhang, Shuai Yuan, Jun Wang, and Xuehua Shen. Real-time bidding benchmarking with ipinyou dataset, 2015. URL <https://arxiv.org/abs/1407.7073>.

Fixed-Confidence Best-Arm Identification for Causal Mediation Analysis (Supplementary Material)

Harsh Shrivastava¹

Yuta Kawakami¹

Junpei Komiyama^{1,2}

Jin Tian¹

¹Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

²RIKEN AIP, Tokyo, Japan

A MOTIVATING EXAMPLES

We present motivating examples illustrating why we study the maximization of the expectation of NDPO.

Example 1 (Advertisement). We present a motivating example illustrating why we study the maximization of the expectation of NDPO. We consider the problem of selecting an advertisement creative (X) from multiple candidates, where the outcome (Y) represents a click indicator, purchase amount, or conversion. Researchers often select the optimal advertisement by maximizing the expected potential outcome under arm x . In practice, however, an advertisement may improve the outcome Y partly through undesirable mechanisms—for example by exploiting user confusion, eliciting accidental clicks, or leveraging platform-specific presentation artifacts. We treat such undesired factors as a mediator (Z). Improving outcomes primarily by changing Z should not be considered a legitimate success (e.g., due to ethical or regulatory concerns). In reality, for example, Zeng et al. [2021] pointed out the widespread use of misleading and manipulative tactics in online political advertising. The NDPO ignores effects achieved through such undesirable factors. Consequently, selecting the optimal advertisement by maximizing the NDPO leads to fairer advertising decision-making, based purely on the direct effect rather than on effects mediated by undesirable factors.

Example 2 (Medicine). In medical decision-making, physicians aim to select treatments that optimally improve patient outcomes. However, some treatments may achieve their effects through adverse intermediate responses (e.g., fever or other adverse reactions). In such cases, it is desirable to evaluate treatment effects that exclude pathways operating through these adverse mediators. Researchers therefore seek to maximize the direct effect, thereby maximizing the treatment benefit while avoiding pathways through these adverse mediators. The indirect effect is also of interest, as it quantifies the effect transmitted through these adverse mediators, and minimizing this effect is another important research objective. In this paper, however, we focus on maximizing the direct effect.

Example 3 (Politics). In political decision-making, policymakers often aim to select policies that effectively reduce inflation. However, some policies may achieve this through undesirable mechanisms, such as increasing unemployment or reducing economic growth. In such cases, it may be preferable to evaluate policies based on their effects that exclude these undesirable pathways. Although the direct effect cannot generally be reproduced by real-world interventions, it provides a useful measure of the desirability of treatment policies under a preference for avoiding specific mediating pathways. In some applications, such as political science, researchers may focus on the indirect effect to understand the underlying mechanisms through which an intervention exerts its influence. In contrast, our objective is to identify the treatment policy that is most preferred under this criterion.

Example 4 (Education). In educational decision-making, teachers often choose among candidate classes or programs to improve students' performance. However, some classes may increase test scores through undesirable mechanisms, such as excessively stressful learning environments. In such cases, it may be preferable to evaluate options based on their effects that exclude these undesirable pathways.

Example 5 (Legal). In legal decision-making, policymakers often design laws or regulations to reduce crime. However, some policies may achieve this through undesirable mechanisms, such as excessive surveillance. In such cases, it may be preferable to evaluate policies based on their effects that exclude these undesirable pathways.

B STOPPING CONDITION

This appendix justifies the stopping rule introduced in Section 6, and in particular the choice of the confidence radius $\beta(t, \delta)$ ensuring δ -correctness.

B.1 UNIFORM CONCENTRATION FOR DISCRETE DISTRIBUTIONS

We begin with a standard concentration inequality for empirical distributions on a finite alphabet.

Lemma 6 (Concentration on a finite alphabet). *Let $\hat{P}_n$ be the empirical distribution obtained from n i.i.d. samples drawn from a categorical distribution P supported on an alphabet of size M . Then, for any $u > 0$,*

$$\Pr\left(\text{KL}(\hat{P}_n \| P) \geq u\right) \leq (n+1)^M e^{-nu}. \quad (38)$$

Lemma 6 is a classical concentration inequality for discrete symbols (see, e.g., Cover and Thomas, 2009).

Lemma 7 (Concentration on a combination of finite alphabets). *Let $i = 1, 2, \dots, I$ for some finite positive integer I . For each i , let $\hat{P}_{n_i, i}$ be the empirical distribution obtained from n_i i.i.d. samples drawn from a categorical distribution P_i supported on an alphabet of size M_i . Let $M = \max_i M_i$. Then, for any $u > 0$ and for any $(n_1, n_2, \dots, n_I)$ such that $n = \sum_{i=1}^I n_i$,*

$$\Pr\left(\sum_{i=1}^I n_i \text{KL}(\hat{P}_{n_i, i} \| P_i) \geq nu\right) \leq (n+1)^{3MI} e^{-nu} \quad (39)$$

holds.

Proof. Let $\mathcal{V}$ be the number of possible combinations of empirical means $(\hat{P}_{n_1, 1}, \hat{P}_{n_2, 2}, \dots, \hat{P}_{n_I, I})$ such that $\sum_i n_i = n$. Since there are I empirical distribution of categories at most M and each category's empirical mean is in $\{0, 1/n_i, 2/n_i, \dots, 1\}$ such that $n_i \leq n$, we have $|\mathcal{V}| \leq n^{2MI}$. Let

$$\mathcal{P}_u = \left\{ (\hat{P}_{n_1, 1}, \hat{P}_{n_2, 2}, \dots, \hat{P}_{n_I, I}) \in \mathcal{V} : \sum_{i=1}^I n_i \text{KL}(\hat{P}_{n_i, i} \| P_i) \geq nu \right\}. \quad (40)$$

Then, Lemma 6 states that the probability that each such set of empirical distributions in $\mathcal{P}_u$ realizes is at most

$$\prod_i (n_i + 1)^{M_i} e^{-n_i u} \leq (n+1)^{MI} e^{-nu}, \quad (41)$$

and thus

$$\Pr\left(\sum_{i=1}^I n_i \text{KL}(\hat{P}_{n_i, i} \| P_i) \geq nu\right) \leq |\mathcal{P}_u| \times (\text{The right-hand side of (41)}) \quad (42)$$

$$\leq n^{2MI} \times (n+1)^{MI} e^{-nu} \quad (43)$$

$$\leq (n+1)^{3MI} e^{-nu}. \quad (44)$$

□

B.2 APPLICATION TO MEDIATOR AND OUTCOME DISTRIBUTIONS

We apply Lemma 7 to both components of the causal model. For each arm x , let $\hat{P}_{Z|x, n}$ denote the empirical distribution of the mediator Z after n pulls of arm x . Let $M_Z := |\mathcal{Z}|$ and $M_Y := |\mathcal{Y}|$ denote the alphabet sizes of the mediator and outcome, respectively. In our setting $M_Y = 2 < M_Z$.

Time-uniform confidence allocation. To obtain bounds that hold uniformly over all $t \geq 1$, define

$$\delta_t := \frac{\delta}{\pi^2 t^{2+2KM_Z}}. \quad (45)$$

Since $\sum_{t=1}^{\infty} 1/(\pi^2 t^2) \leq 1$, a union bound over all t and all possible combinations of $(N_x(t), N_{x,z}(t))$ (number of such possible combinations is at most t^{2KM_Z}) yields an event of probability at least $1 - \delta$. Namely, by setting

$$\beta(t, \delta) = \log\left(\frac{\pi^2 t^{2+2KM_Z}}{\delta}\right) + 6M_Z K \log(t+1), \quad (46)$$

we have

$$\forall t \{\mathcal{L}_t(Q) \leq \beta(t, \delta)\} \quad (47)$$

holds with probability at least $1 - \delta$.

C IMPLEMENTATION DETAILS FOR THE REDUCED INNER OPTIMIZATION

This appendix describes the numerical solution of the reduced inner optimization problem appearing in the plug-in allocation (20) and in the GLRT statistic (29), after reduction to a single competitor and restriction to the relevant arms.

C.1 REDUCED WEIGHTED INNER PROBLEM

Fix a competitor $x' \neq x^*$ and let

$$\gamma_0 := w_{x_0}, \quad \alpha_{s,z} := w_{x^*} \hat{P}_{x^*}(z), \quad \alpha_{p,z} := w_{x'} \hat{P}_{x'}(z), \quad (48)$$

where w denotes the current allocation vector.

Let

$$P_0(z) := \hat{P}_{x_0}(z), \quad p_s(z) := \hat{P}_{x^*}(1|z), \quad p_p(z) := \hat{P}_{x'}(1|z). \quad (49)$$

After the three-arm reduction (Section 6.2), the inner problem reduces to

$$\begin{aligned} \min_{\substack{q_0 \in \Delta(\mathcal{Z}) \\ q_s, q_p \in (0,1)^{|\mathcal{Z}|}}} & \gamma_0 \text{KL}(P_0 \| q_0) + \sum_{z \in \mathcal{Z}} \alpha_{s,z} \text{KL}(p_s(z) \| q_s(z)) \\ & + \sum_{z \in \mathcal{Z}} \alpha_{p,z} \text{KL}(p_p(z) \| q_p(z)) \end{aligned} \quad (50)$$

$$\text{s.t.} \quad \sum_{z \in \mathcal{Z}} q_0(z) (q_p(z) - q_s(z)) \geq 0. \quad (51)$$

The feasible set is compact and the objective is continuous. Moreover, the problem is convex in (q_s, q_p) for fixed q_0 and convex in q_0 for fixed (q_s, q_p) , but not jointly convex.

We solve (50)–(51) via block coordinate descent (alternating optimization).

C.2 UPDATE OF (q_s, q_p) FOR FIXED q_0

For fixed q_0 , the subproblem in (q_s, q_p) is convex:

$$\begin{aligned} \min_{q_s, q_p} & \sum_z \alpha_{s,z} \text{KL}(p_s(z) \| q_s(z)) + \sum_z \alpha_{p,z} \text{KL}(p_p(z) \| q_p(z)) \\ \text{s.t.} & \sum_z q_0(z) (q_p(z) - q_s(z)) \geq 0. \end{aligned} \quad (52)$$

If the constraint is inactive at the unconstrained minimizer $(q_s, q_p) = (p_s, p_p)$, then this is the solution.

Otherwise, the constraint is active and we introduce a Lagrange multiplier $\lambda \geq 0$. For each z , the subproblem separates and we must solve

$$\min_{q \in (0,1)} \alpha \text{KL}(p||q) + b q, \quad (53)$$

where $b = \lambda q_0(z)$ (with opposite signs for q_s and q_p).

Using

$$\frac{\partial}{\partial q} \text{KL}(p||q) = \frac{q - p}{q(1 - q)}, \quad (54)$$

the stationarity condition yields the quadratic equation

$$bq^2 - (\alpha + b)q + \alpha p = 0. \quad (55)$$

Among its real roots in $(0, 1)$, the one minimizing the objective is selected. The multiplier λ is chosen by one-dimensional bisection so that the constraint (51) holds with equality when active. Monotonicity of the constraint function in λ ensures existence and uniqueness of the solution.

C.3 UPDATE OF q_0 FOR FIXED (q_s, q_p)

For fixed (q_s, q_p) , define

$$d(z) := q_p(z) - q_s(z). \quad (56)$$

The subproblem in q_0 is

$$\min_{q_0 \in \Delta(\mathcal{Z})} \text{KL}(P_0||q_0) \text{ s.t. } \sum_z q_0(z) d(z) \geq 0. \quad (57)$$

If the constraint is satisfied at $q_0 = P_0$, then this is optimal.

Otherwise, introducing a multiplier $\mu \geq 0$ and using KKT conditions yields the closed-form solution

$$q_0(z) = \frac{P_0(z)}{\nu - \mu d(z)}, \quad (58)$$

where ν is determined by the normalization condition $\sum_z q_0(z) = 1$. The multiplier μ is obtained by one-dimensional bisection so that the constraint holds with equality when active.

C.4 ALTERNATING OPTIMIZATION AND CONVERGENCE

The algorithm alternates between:

- updating (q_s, q_p) given q_0 ,
- updating q_0 given (q_s, q_p) .

Each block subproblem is solved exactly and is convex. The objective in (50) is non-increasing at every iteration, and the feasible set is compact. Therefore, by standard results on block coordinate descent for bi-convex problems [Tseng, 2001], every limit point of the sequence is a stationary point of (50)–(51).

In practice, the dimensionality is $O(|\mathcal{Z}|)$ and convergence is rapid. This procedure is used both in the plug-in allocation computation and in evaluation of the GLRT statistic.

D BAI ALGORITHMS FOR OTHER MEDIATION MEASURES

Finding the best arm in terms of $\mathbb{E}[Y_{x,z}]$, which is the first term of the controlled direct effect (CDE) at $Z = z$, is relatively straightforward using a best arm identification algorithm. For ease of discussion, let us consider $\mathbb{E}[Y_{x,z}]$ at $z = 1$

$$\mathbb{E}_P[Y_{x,z=1}]. \quad (59)$$

This boils down to standard bandits/BAI but we only count the sample when it is shown at a specific position $z = 1$. Namely,

$$N_x(t) = \sum_{s=1}^t \mathbf{1}\{X_t = x, Z = 1\}, \hat{P}_x(t) = \frac{\sum_{s=1}^t Y_t \mathbf{1}\{X_t = x, Z = 1\}}{N_x(t)} \quad (60)$$

and one can run Track-and-Stop or Top-two Thompson sampling [Russo, 2020] based on these statistics.

Regarding the Natural Indirect effect (NIE), the first term of NDE is

$$\mathbb{E}_P[Y_{x_0, Z_x}] = \sum_z P_{x_0}(1|z) P_x(z). \quad (61)$$

We can consider the alternative distributions as:

$$\text{Alt}(P) := \left\{ Q \in \mathcal{P} : \exists x \neq x^*(P) \text{ such that } \sum_{z \in \mathcal{Z}} Q_{x_0}(1|z) Q_x(z) > \sum_{z \in \mathcal{Z}} Q_{x_0}(1|z) Q_{x^*(P)}(z) \right\}. \quad (62)$$

We hypothesize that our Track-and-Stop algorithm with a cutting-plane procedure (Section 6.2) can also achieve asymptotically optimal sample complexity for identifying the best arm w.r.t. the $\mathbb{E}[Y_{x_0, Z_x}]$ -criterion. In summary, finding the best arm for $\mathbb{E}[Y_{x,z}]$ is straightforward. Finding the best arm for $\mathbb{E}[Y_{x_0, Z_x}]$ can be done using our framework.

E SYNTHETIC EXPERIMENTS

In this appendix, we conduct synthetic experiments.

Setting. We consider instances with $K = 3$ arms and $M = 3$ mediator values. Full parameter specifications are as follows: Let $x_0 = 0$ and $\mathcal{Z} = \{0, 1, 2\}$. Fix $P_0(Z) = (0.8, 0.15, 0.05)$. Let $P(Z|X = 1) = (0.1, 0.2, 0.7)$ and $P(Z|X = 2) = P(Z|X = 0) = P_0(Z)$. Set $P(Y = 1|X = 0, Z = z) = 0.2$, $P(Y = 1|X = 1, Z \in \{0, 1\}) = 0.2$, $P(Y = 1|X = 1, Z = 2) = 0.9$, and $P(Y = 1|X = 2, Z = z) = 0.235 + \Delta$ for all z . Then $\theta(1) = 0.235$ and $\theta(2) = 0.235 + \Delta$, so the expected NDPO gap equals Δ , while $\mathbb{E}[Y_1] = 0.69$ is the largest interventional mean. To avoid zero empirical probabilities when some counts are small, we use Laplace smoothing [Manning, 2008], i.e., $\hat{P}_x(z; t) = (N_{x,z}(t) + \alpha) / (N_x(t) + \alpha|\mathcal{Z}|)$ with $\alpha = 0.05$. All methods use the same confidence level $\delta = 0.05$ and the same stopping threshold $\beta(t, \delta)$ from Appendix B. Optimization tolerances and forcing schedules are matched whenever applicable. Performance was robust to moderate changes in these settings.

Algorithms. We compare the following algorithms:

- **TaS-NDPO (ours).** Algorithm 1.
- **TaS-NDPO (arm-level).** Variant of TaS-NDPO by replacing cell-level forcing with classical arm-level forced exploration.
- **Baseline-First.** The algorithm first estimates the baseline distribution $P_{x_0}(z)$ using uniform sampling, and then samples arms uniformly to estimate $P_x(1|z)$.
- **Uniform.** Uniform sampling over arms combined with the same plug-in estimator and stopping rule as ours.
- **TaS-IM (diagnostic).** TaS algorithm for optimizing the interventional mean (IM) $\mathbb{E}[Y_x]$.

Gap-scaling family. Mediator distributions are fixed while the NDPO gap

$$\Delta := \theta(x^*) - \max_{x \neq x^*} \theta(x) \quad (63)$$

is controlled through the outcome model of arm 2. This isolates the effect of gap magnitude.

To validate Theorem 2, we vary Δ and measure the median stopping time. Figure 3 shows log-log scaling. The empirical slope aligns with the reference $1/\Delta^2$, confirming the predicted instance-dependent complexity.

Decision differences. In this model, the interventional-mean-optimal arm differs from the expected NDPO-optimal arm. TaS-IM selects arm 1, while TaS-NDPO consistently identifies arm 2, demonstrating that optimizing expected NDPO can lead to different decisions.

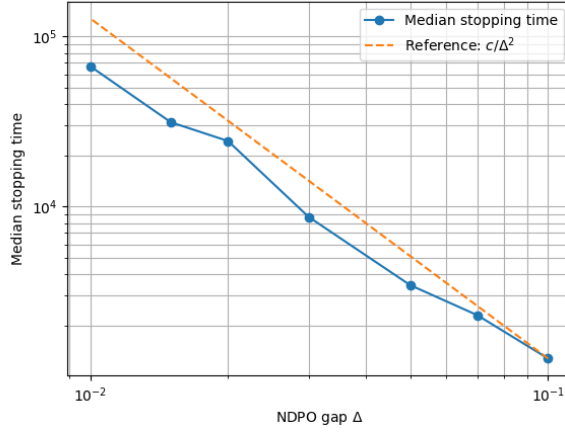


Figure 3: Median stopping time vs. NDPO gap Δ (log–log scale). Dashed line: $1/\Delta^2$.

Table 3: Median stopping time as a function of mediator sparsity η . Lower is better. Bold indicates the best method for each η .

η	TaS-NDPO (ours)	TaS-NDPO (arm-level)	Uniform	Baseline-First
0.5	2048	2048	2560	6912
0.1	20368	20608	28928	21664
0.05	29184	36864	33664	31960
0.01	36480	47104	46080	41216
0.005	41216	57088	46208	47520
0.001	45056	54528	47616	49024

E.1 ADDITIONAL SENSITIVITY EXPERIMENTS

We further evaluate the effect of mediator sparsity, mediator cardinality, and baseline-arm choice. The gap-scaling experiment above varies the NDPO gap while holding the mediator structure fixed. In contrast, the experiments below vary the mediator structure or the reference arm. Overall, the results show that the advantage of cell-level forcing is regime-dependent, with substantial gains in settings with rare mediator–outcome cells. The arm-level Track-and-Stop baseline already represents a strong adaptive method and performs well in easier regimes.

Mediator sparsity. Table 3 reports median stopping times as the mediator sparsity parameter η varies. Smaller values of η correspond to rarer mediator–outcome cells. Our method is best or tied for best across all values of η . The improvement over arm-level forcing becomes more pronounced as rare cells become important, showing the benefit of explicit cell-level forcing in sparse mediator regimes.

Mediator cardinality. Table 4 reports median stopping times as the number of mediator values M varies. The results are not monotone in M , confirming that the benefit of cell-level forcing depends on the structure of the instance. Our method performs best for $M = 2$ and $M = 16$, while the arm-level variant is best for $M = 8$ and Baseline-First is slightly better for $M = 4$. This shows that arm-level exploration can be competitive in easier regimes, whereas cell-level forcing is most useful when the relevant mediator–outcome cells drive the difficulty of identification.

Baseline-arm choice. Table 5 reports median stopping times for different choices of the baseline arm x_0 . Changing x_0 changes the expected NDPO target and may also change the NDPO-best arm. For $x_0 = 0$, the instance is harder and our method achieves the smallest stopping time. For $x_0 = 1$ and $x_0 = 2$, the instance is easier, and TaS-NDPO, the arm-level variant, and Uniform all stop at the same median time. Baseline-First is slower in these easy cases because it separately estimates the baseline distribution before sampling the remaining arms.

Table 4: Median stopping time as a function of mediator cardinality M . Lower is better. Bold indicates the best method for each M .

M	TaS-NDPO (ours)	TaS-NDPO (arm-level)	Uniform	Baseline-First
2	13312	15360	15872	13440
4	21632	23592	25432	21504
8	21504	14976	18560	19584
16	10752	20864	22016	22016

Table 5: Median stopping time as a function of baseline-arm choice x_0 . Lower is better. Bold indicates the best method for each baseline choice in a 4-arm instance.

x_0	NDPO best arm	TaS-NDPO (ours)	TaS-NDPO (arm-level)	Uniform	Baseline-First
0	3	59904	69248	70144	61952
1	1	512	512	512	5376
2	2	512	512	512	5376

E.2 RUNTIME COMPARISON

We also measure the wall-clock runtime of each method. Table 6 compares statistical efficiency, measured by median stopping time, with computational cost, measured by average runtime per run. TaS-NDPO incurs a moderate computational overhead relative to non-adaptive baselines, but its runtime is comparable to the arm-level adaptive baseline and it achieves the lowest median stopping time.

F PROOFS

In this appendix, we provide the proofs of the theorems, propositions, and lemmas stated in the main text.

F.1 PROOF OF THEOREM 1

Theorem 1. Under SCM $\mathcal{M}$ and Assumption 1, the expected NDPO is identified by $\theta(x)$, where

$$\theta(x) = \sum_{y,z} y P_x(y|z) P_{x_0}(z), \quad (4)$$

where $P_x(y|z) = \mathbb{P}(Y = y|Z = z, do(X = x))$ and $P_x(z) = \mathbb{P}(Z = z|do(X = x))$.

Proof. Under Assumption 1, we have

$$\begin{aligned}
& \mathbb{E}[Y_{x,Z_{x_0}}] \\
&= \sum_{y,z} y \mathbb{P}(Y_{x,z} = y|Z_{x_0} = z) \mathbb{P}(Z_{x_0} = z) \\
&= \sum_{y,z} y \mathbb{P}(Y_{x,z} = y) \mathbb{P}(Z_{x_0} = z) (\because \text{Assumption 1}) \\
&= \sum_{y,z} y \mathbb{P}(Y_{x,z} = y|Z_x = z) \mathbb{P}(Z_{x_0} = z) (\because \text{Assumption 1}) \\
&= \sum_{y,z} y \mathbb{P}(Y_x = y|Z_x = z) \mathbb{P}(Z_{x_0} = z) (\because \text{Counterfactual Consistency}) \\
&= \sum_{y,z} y \mathbb{P}(Y = y|Z = z, do(X = x)) \mathbb{P}(Z = z|do(X = x_0)).
\end{aligned} \quad (64)$$

Then, we have the theorem. □

Table 6: Comparison of statistical efficiency and computational cost. The median stopping time measures sample complexity, and the average runtime is measured in seconds per run. Lower is better.

Method	Median stopping time	Average runtime (s)
TaS-NDPO (ours)	36,480	22.5
TaS-NDPO (arm-level)	47,104	25.3
Uniform	46,080	6.75
Baseline-First	46,336	6.43

F.2 PROOF OF THEOREM 2

Theorem 2. [Instance-dependent lower bound] Let $\delta \in (0, 1)$. Under Assumption 2, for any δ -correct algorithm and any $P \in \mathcal{P}$,

$$\mathbb{E}_P[\tau_\delta] \geq T^*(P) \text{kl}(\delta, 1 - \delta), \quad (12)$$

where $\text{kl}(\delta, 1 - \delta) = \delta \log \frac{\delta}{1-\delta} + (1 - \delta) \log \frac{1-\delta}{\delta}$ and

$$(T^*(P))^{-1} = \sup_{w \in \Sigma_K} \inf_{Q \in \text{Alt}(P)} \sum_{x \in \mathcal{X}} w_x \text{KL}(P_x \| Q_x) \quad (13)$$

with $\Sigma_K := \{w \in \mathbb{R}_+^K : \sum_{x \in \mathcal{X}} w_x = 1\}$.

Proof. The proof follows the standard change-of-measure argument [Lai and Robbins, 1985].

Let $N_x(t)$ denote the (random) number of times arm x is sampled up to time t , and let τ_δ be the stopping time of a δ -correct algorithm. For any alternative model $Q \in \text{Alt}(P)$, the transportation inequality (Lemma 1 in Garivier and Kaufmann, 2016) yields

$$\sum_{x=0}^{K-1} \mathbb{E}_P[N_x(\tau_\delta)] \text{KL}(P_x \| Q_x) \geq \text{kl}(\delta, 1 - \delta), \quad (65)$$

where P_x and Q_x denote the joint distributions of (Z, Y) under arm x for models P and Q , respectively.

Define the sampling proportions

$$w_x := \frac{\mathbb{E}_P[N_x(\tau_\delta)]}{\mathbb{E}_P[\tau_\delta]}, \quad \sum_x w_x = 1. \quad (66)$$

Dividing both sides by $\mathbb{E}_P[\tau_\delta]$ gives

$$\mathbb{E}_P[\tau_\delta] \sum_x w_x \text{KL}(P_x \| Q_x) \geq \text{kl}(\delta, 1 - \delta). \quad (67)$$

Under the joint mediator–outcome observation model, the KL divergence for a fixed arm x decomposes as

$$\text{KL}(P_x \| Q_x) = \sum_{z \in \mathcal{Z}} P_{z|x} \text{KL}(P_{Y|x,z} \| Q_{Y|x,z}) + \text{KL}(P_{Z|x} \| Q_{Z|x}). \quad (68)$$

Substituting this decomposition yields

$$\mathbb{E}_P[\tau_\delta] \geq \frac{\text{kl}(\delta, 1 - \delta)}{\sum_x w_x [\sum_z P_{z|x} \text{KL}(P_{Y|x,z} \| Q_{Y|x,z}) + \text{KL}(P_{Z|x} \| Q_{Z|x})]}. \quad (69)$$

Since the above bound holds for all $Q \in \text{Alt}(P)$, we may take the infimum over Q , and since the strategy is arbitrary, we may take the supremum over $w \in \Sigma_K$. Thus,

$$\mathbb{E}_P[\tau_\delta] \geq \frac{\text{kl}(\delta, 1 - \delta)}{\sup_{w \in \Sigma_K} \inf_{Q \in \text{Alt}(P)} \sum_x w_x \text{KL}(P_x \| Q_x)}. \quad (70)$$

Finally, using the inequality Kaufmann et al. [2016] of $\text{kl}(\delta, 1 - \delta) \geq \log(1/(2.4\delta)) = \log(1/\delta) + \Theta(1)$ completes the proof. $\square$

F.3 PROOF OF PROPOSITION 1

Proposition 1. *[Reduction to three arms] For the inner minimization oracle in (22), it is without loss of optimality to restrict attention to solutions satisfying $Q_a = P_a$ for all $a \notin \{x_0, x^*(P), x'\}$.*

Proof. The constraint depends only on $Q_{x_0}, Q_{x^*(P)}, Q_{x'}$. For any $a \notin \{x_0, x^*(P), x'\}$, the term $\text{KL}(P_a \| Q_a)$ is uniquely minimized at $Q_a = P_a$ by strict convexity of KL divergence. Any deviation from P_a strictly increases the objective while leaving the constraint unchanged, and hence cannot be optimal. $\square$

F.4 PROOF OF PROPOSITION 2

Proposition 2. *[Bi-convex structure] The optimization problem defined by (23)–(26) is convex in (r^*, r') for fixed q_0 , and convex in q_0 for fixed (r^*, r') , but not jointly convex.*

Proof. For fixed q_0 , the objective is a sum of KL divergences in r^* and r' , which are convex functions. The constraint (26) is affine in (r^*, r') when q_0 is fixed. Similarly, for fixed (r^*, r') , the objective is convex in q_0 , and the constraint is affine in q_0 . Joint convexity fails due to bilinearity of the constraint. $\square$

F.5 PROOFS FOR SECTION 7

Throughout, fix a true model $P \in \mathcal{P}$. Recall that for each arm $x \in \mathcal{X}$, pulling x yields i.i.d. samples $(Z_t, Y_t) \sim P_x$, and that the algorithm is adaptive but non-anticipating. We use $\mathbb{P}_P$ and $\mathbb{E}_P$ for probability and expectation under P .

We also recall the empirical joint model $\hat{P}(t) = (\hat{P}_x(t))_{x \in \mathcal{X}}$, where $\hat{P}_x(t)$ is the empirical distribution of (Z, Y) under arm x up to time t . For $Q \in \mathcal{P}$, define the negative log-likelihood ratio

$$\mathcal{L}_t(Q) = \sum_{x \in \mathcal{X}} N_x(t) \text{KL}(\hat{P}_x(t) \| Q_x), \quad (71)$$

which is equivalent to the decomposed form used in (29). The GLRT statistic is

$$Z_t(x, y) = \inf_{Q \in \mathcal{P}: \theta_Q(x) \geq \theta_Q(y)} \mathcal{L}_t(Q). \quad (72)$$

F.5.1 Proof of Theorem 3 (δ -correctness)

Theorem 3. *[δ -correctness] Assume $\beta(t, \delta)$ is chosen as in Appendix B. For any $\delta \in (0, 1)$ and any $P \in \mathcal{P}$, Algorithm 1 satisfies*

$$\mathbb{P}_P(\tau_\delta < \infty, \hat{x}(\tau_\delta) \neq x^*(P)) \leq \delta. \quad (32)$$

Proof. For each $t \geq 1$, define the likelihood-based confidence set

$$\mathcal{U}_t := \{Q \in \mathcal{P} : \mathcal{L}_t(Q) \leq \beta(t, \delta)\}. \quad (73)$$

Step 1: time-uniform containment of the true model. By the uniform concentration inequality stated and proven in Appendix B, the threshold $\beta(t, \delta)$ can be chosen so that

$$\mathbb{P}_P(P \in \mathcal{U}_t \text{ for all } t \geq 1) \geq 1 - \delta. \quad (74)$$

Let $\mathcal{E} := \{P \in \mathcal{U}_t \forall t \geq 1\}$ denote this event.

Step 2: stopping implies no alternative remains in $\mathcal{U}_t$. Fix $t \geq 1$ and an arm $x \neq \hat{x}(t)$. By definition,

$$Z_t(x, \hat{x}(t)) = \inf_{Q: \theta_Q(x) \geq \theta_Q(\hat{x}(t))} \mathcal{L}_t(Q). \quad (75)$$

Hence the condition $Z_t(x, \hat{x}(t)) \geq \beta(t, \delta)$ implies that for every Q such that $\theta_Q(x) \geq \theta_Q(\hat{x}(t))$ we have $\mathcal{L}_t(Q) \geq \beta(t, \delta)$, i.e., no such Q lies in $\mathcal{U}_t$. Equivalently,

$$\forall Q \in \mathcal{U}_t : \quad \theta_Q(\hat{x}(t)) \geq \theta_Q(x). \quad (76)$$

At the stopping time τ_δ , the algorithm stops only if

$$\min_{x \neq \hat{x}(\tau_\delta)} Z_{\tau_\delta}(x, \hat{x}(\tau_\delta)) \geq \beta(\tau_\delta, \delta), \quad (77)$$

so (76) holds for all competitors $x \neq \hat{x}(\tau_\delta)$. Thus, for all $Q \in \mathcal{U}_{\tau_\delta}$,

$$\theta_Q(\hat{x}(\tau_\delta)) \geq \theta_Q(x) \quad \forall x \neq \hat{x}(\tau_\delta), \quad (78)$$

which means $\hat{x}(\tau_\delta) \in \arg \max_{x \in \mathcal{X}} \theta_Q(x)$ for every $Q \in \mathcal{U}_{\tau_\delta}$ if $\tau_\delta < \infty$.

Step 3: correctness on the event $\mathcal{E}$. On the event $\mathcal{E}$, we have $P \in \mathcal{U}_{\tau_\delta}$ (since $\tau_\delta \geq 1$), and by Assumption 2, the conclusion above gives $\hat{x}(\tau_\delta) = x^*(P)$ on $\mathcal{E}$. Therefore,

$$\mathbb{P}_P(\hat{x}(\tau_\delta) \neq x^*(P), \tau_\delta < \infty) \leq \mathbb{P}_P(\mathcal{E}^c) \leq \delta, \quad (79)$$

using (74). This proves δ -correctness. $\square$

F.5.2 Proof of Lemma 1

Lemma 1. Under Assumption 3. Let $h(t) = \lceil t^a \rceil$ for some $a \in (0, 1)$ and let $g(t)$ be nondecreasing with $g(t) \rightarrow \infty$ and $g(t) = o(h(t))$. Under the sampling rule of Section 6.3,

$$N_x(t) \rightarrow \infty, N_{x,z}(t) \rightarrow \infty, \forall x \in \mathcal{X}, z \in \mathcal{Z}, a.s. \quad (83)$$

Proof. Fix an arm $x \in \mathcal{X}$. We first show that $N_x(t) \rightarrow \infty$ almost surely.

Step 1: x enters the relevant set infinitely often. Let $C_x(t) = \sum_{s=1}^t \mathbf{1}\{x'(s) = x\}$ be the competitor counter. By construction of the competitor-coverage rule, for every t ,

$$\min_{u \in \mathcal{X}} C_u(t) \geq h(t) - 1. \quad (80)$$

In particular, $C_x(t) \geq h(t) - 1$ for all sufficiently large t , hence $C_x(t) \rightarrow \infty$. Moreover, whenever $x'(s) = x$ we have $x \in \mathcal{A}_{\text{rel}}(s)$. Therefore, x belongs to $\mathcal{A}_{\text{rel}}(t)$ infinitely often.

Step 2: contradiction argument for $N_x(t)$. Suppose by contradiction that $N_x(t)$ is bounded on some event E with $\mathbb{P}_P(E) > 0$. Then there exists $B < \infty$ such that on E ,

$$N_x(t) \leq B \quad \forall t, \quad \text{and hence} \quad m_x(t) := \min_{z \in \mathcal{Z}} N_{x,z}(t) \leq N_x(t) \leq B \quad \forall t. \quad (81)$$

Since $g(t) \rightarrow \infty$ and is nondecreasing, there exists T such that $g(t) > B$ for all $t \geq T$. Thus, on E , for every $t \geq T$ with $x \in \mathcal{A}_{\text{rel}}(t)$ we have $m_x(t) < g(t)$, i.e., $x \in \mathcal{D}(t)$.

Let

$$\mathcal{T} := \{t \geq T : x \in \mathcal{A}_{\text{rel}}(t)\}. \quad (82)$$

By Step 1, $\mathcal{T}$ is infinite. For each $t \in \mathcal{T}$, the set $\mathcal{D}(t)$ is nonempty and the forcing rule selects

$$X_{t+1} \in \arg \min_{u \in \mathcal{D}(t)} (m_u(t), N_u(t)). \quad (83)$$

Assume that on E the arm x is selected only finitely many times. Then there exists $T' \geq T$ such that for all $t \in \mathcal{T}$ with $t \geq T'$, the forced choice satisfies $X_{t+1} \neq x$. Since only finitely many arms exist, there must be an arm $u^* \neq x$ that is selected infinitely often among these forcing rounds.

By Assumption 3, $P_{u^*}(z) > 0$ for all z , and hence, conditional on selecting arm u^* infinitely often, the strong law yields $N_{u^*,z}(t) \rightarrow \infty$ for every z , so in particular $m_{u^*}(t) \rightarrow \infty$ almost surely. Therefore, on E we have $m_{u^*}(t) > B$ for all sufficiently large t .

But for all t we also have $m_x(t) \leq B$. Hence, for all sufficiently large $t \in \mathcal{T}$, the arm x satisfies $m_x(t) < m_{u^*}(t)$ and belongs to $\mathcal{D}(t)$, so x must be chosen by the minimization of $m_u(t)$ in the forcing rule, a contradiction. We conclude that $N_x(t) \rightarrow \infty$ almost surely.

Step 3: divergence of all cell counts. Fix $z \in \mathcal{Z}$. Consider the subsequence of rounds at which $X_s = x$. Conditional on these pull times, the mediator observations are i.i.d. with law $P_x(\cdot)$. By the strong law of large numbers applied to $\mathbf{1}\{Z = z\}$ along this subsequence,

$$\frac{N_{x,z}(t)}{N_x(t)} \rightarrow P_x(z) \quad \text{almost surely.} \quad (84)$$

By Assumption 3, $P_x(z) > 0$, and since $N_x(t) \rightarrow \infty$, it follows that $N_{x,z}(t) \rightarrow \infty$ almost surely.

Since x and z were arbitrary, the claim holds for all (x, z) . $\square$

F.5.3 Proof of Lemma 2

Lemma 2. [Consistency of plug-in allocation] Under Assumption 2, if $\hat{P}(t) \rightarrow P$ almost surely, then $\hat{w}(t) \rightarrow w^*(P)$ almost surely.

Proof. Define for any model $R \in \mathcal{P}$ and $w \in \Sigma_K$,

$$\Phi_R(w) := \inf_{Q \in \text{Alt}(R)} \sum_{x \in \mathcal{X}} w_x \text{KL}(R_x \| Q_x), \quad \text{Alt}(R) := \{Q \in \mathcal{P} : x^*(Q) \neq x^*(R)\}. \quad (85)$$

By definition, $\hat{w}(t) \in \arg \max_{w \in \Sigma_K} \Phi_{\hat{P}(t)}(w)$.

Step 1: Stabilization of the plug-in best arm. Since $x^*(P)$ is unique, there exists a gap

$$\Delta := \theta_P(x^*(P)) - \max_{x \neq x^*(P)} \theta_P(x) > 0, \quad (86)$$

where $\theta_P(x)$ denotes the objective value of arm x under model P . Because $\theta_R(x)$ is a continuous function of R on $\mathcal{P}$ (it is a finite sum/product of probabilities on a finite alphabet), $\hat{P}(t) \rightarrow P$ almost surely implies $\theta_{\hat{P}(t)}(x) \rightarrow \theta_P(x)$ for every x . Hence, on an event of probability one, there exists a (random) time T_0 such that for all $t \geq T_0$,

$$\theta_{\hat{P}(t)}(x^*(P)) > \max_{x \neq x^*(P)} \theta_{\hat{P}(t)}(x), \quad (87)$$

which implies $\hat{x}(t) = x^*(\hat{P}(t)) = x^*(P)$ for all $t \geq T_0$.

Step 2: For large t , the alternative set becomes fixed. On the same event, for all $t \geq T_0$ we have

$$\text{Alt}(\hat{P}(t)) = \{Q \in \mathcal{P} : x^*(Q) \neq \hat{x}(t)\} = \{Q \in \mathcal{P} : x^*(Q) \neq x^*(P)\} =: \text{Alt}^*, \quad (88)$$

which no longer depends on t .

We now work on $t \geq T_0$ and treat Alt^* as fixed.

Step 3: Compactness of Alt^* . Since $\mathcal{P}$ is a product of simplices over a finite alphabet, it is compact. Moreover, for each $x \neq x^*(P)$, the set

$$\mathcal{C}_x := \{Q \in \mathcal{P} : \theta_Q(x) \geq \theta_Q(x^*(P))\} \quad (89)$$

is closed because $\theta_Q(\cdot)$ is continuous in Q . Therefore

$$\text{Alt}^* = \bigcup_{x \neq x^*(P)} \mathcal{C}_x \quad (90)$$

is a finite union of closed sets, hence closed, and thus compact as a closed subset of compact $\mathcal{P}$.

Step 4: Continuity of the value function on a fixed alternative set. Fix $w \in \Sigma_K$ and define

$$G(R, w, Q) := \sum_{x \in \mathcal{X}} w_x \text{KL}(R_x \| Q_x), \quad Q \in \text{Alt}^*. \quad (91)$$

Under Assumption 3, all probabilities involved are bounded away from 0 and 1 uniformly over $R, Q \in \mathcal{P}$. Hence $\text{KL}(R_x \| Q_x)$ is finite and continuous in (R, Q) for each x , and therefore G is continuous in (R, w, Q) on $\mathcal{P} \times \Sigma_K \times \text{Alt}^*$. Since Alt^* is compact, the value function

$$\tilde{\Phi}_R(w) := \inf_{Q \in \text{Alt}^*} G(R, w, Q) \quad (92)$$

is continuous in (R, w) by Berge's maximum theorem.

Step 5: Continuity of the argmax and convergence of $\hat{w}(t)$. Define $f(R, w) := \tilde{\Phi}_R(w)$ on $\mathcal{P} \times \Sigma_K$. We have shown f is continuous and Σ_K is compact. Let $w^*(P)$ denote the unique maximizer of $w \mapsto f(P, w)$.

Consider any sequence $t_n \rightarrow \infty$ and write $R_n := \hat{P}(t_n)$ and $w_n := \hat{w}(t_n)$. By compactness of Σ_K , (w_n) has a convergent subsequence (not relabeled) with limit $\bar{w} \in \Sigma_K$. On the almost sure event where $\hat{P}(t) \rightarrow P$ and $t_n \geq T_0$ eventually, we have $R_n \rightarrow P$ and $\text{Alt}(\hat{P}(t_n)) = \text{Alt}^*$ for all large n . Since w_n maximizes $f(R_n, \cdot)$,

$$f(R_n, w_n) \geq f(R_n, w) \quad \text{for all } w \in \Sigma_K. \quad (93)$$

Taking limits along the convergent subsequence and using continuity of f gives

$$f(P, \bar{w}) \geq f(P, w) \quad \text{for all } w \in \Sigma_K, \quad (94)$$

so $\bar{w} \in \arg \max_{w \in \Sigma_K} f(P, w) = \{w^*(P)\}$ by uniqueness. Hence $\bar{w} = w^*(P)$.

Because every convergent subsequence of $(\hat{w}(t))$ has the same limit $w^*(P)$, it follows that $\hat{w}(t) \rightarrow w^*(P)$ almost surely. $\square$

F.5.4 Proof of Lemma 3

Lemma 3. [Sublinearity of forced rounds] Let $F(t)$ denote the number of rounds up to time t in which either competitor coverage or cell-level forcing overrides the D -tracking update. Then $F(t) = O(h(t)) + O(g(t))$. In particular, if $h(t) = t^a$ with $a \in (0, 1)$ and $g(t) = o(h(t))$, then $F(t) = o(t)$ almost surely.

Proof. Write $F(t) = F_{\text{cov}}(t) + F_{\text{cell}}(t)$, where $F_{\text{cov}}(t)$ (resp. $F_{\text{cell}}(t)$) counts the rounds up to t on which competitor coverage (resp. cell-level forcing) is active.

Step 1: competitor coverage is deterministically $O(h(t))$. Recall $C_x(t) = \sum_{s=1}^t \mathbf{1}\{x'(s) = x\}$ and the coverage rule that, whenever $\min_x C_x(t-1) < h(t)$, selects $x'(t)$ among arms with $C_x(t-1) < h(t)$ and increments only that counter. Thus, for each arm x , the rule can select x under the condition $C_x(t-1) < h(t)$ at most $h(t)$ times up to time t . Summing over $K := |\mathcal{X}|$ arms yields the deterministic bound

$$F_{\text{cov}}(t) \leq \sum_{x \in \mathcal{X}} C_x(t) \mathbf{1}\{C_x(t) < h(t) + 1\} \leq K h(t) + K, \quad (95)$$

hence $F_{\text{cov}}(t) = O(h(t))$ and therefore $F_{\text{cov}}(t) = o(t)$ since $a \in (0, 1)$.

Step 2: cell-level forcing is $O(g(t))$ almost surely. Fix an arm $x \in \mathcal{X}$ and define the cell counts $N_{x,z}(t) = \sum_{s \leq t} \mathbf{1}\{X_s = x, Z_s = z\}$ and $N_x(t) = \sum_z N_{x,z}(t)$. Let

$$p_{\min} := \min_{x \in \mathcal{X}, z \in \mathcal{Z}} P_x(z) \in (0, 1) \quad (96)$$

which exists and is strictly positive by Assumption 3. Consider the event

$$\mathcal{E}_x := \{\forall z \in \mathcal{Z} : \lim_{n \rightarrow \infty} \frac{N_{x,z}(t_n)}{n} = P_x(z) \text{ along the pull times of } x\}. \quad (97)$$

By the strong law of large numbers applied to the i.i.d. mediator draws under repeated pulls of x , we have $\mathbb{P}(\mathcal{E}_x) = 1$. On $\mathcal{E}_x$, there exists an (a.s. finite) random time T_x such that for all $t \geq T_x$ and all $z \in \mathcal{Z}$,

$$N_{x,z}(t) \geq \frac{p_{\min}}{2} N_x(t). \quad (\star)$$

Consequently, for $t \geq T_x$,

$$m_x(t) := \min_{z \in \mathcal{Z}} N_{x,z}(t) \geq \frac{p_{\min}}{2} N_x(t). \quad (98)$$

Therefore, if x is cell-deficient at time $t \geq T_x$, i.e. $m_x(t) < g(t)$, then necessarily

$$N_x(t) < \frac{2}{p_{\min}} g(t). \quad (\dagger)$$

Now let $F_{\text{cell},x}(t)$ be the number of times up to t that the algorithm selects arm x *because of cell-level forcing*. Trivially, $F_{\text{cell},x}(t) \leq N_x(t)$. Moreover, once $(\dagger)$ fails (i.e. once $N_x(t) \geq \frac{2}{p_{\min}} g(t)$) and $t \geq T_x$, the arm x cannot be cell-deficient anymore, hence it cannot be selected due to cell forcing. Thus, on $\mathcal{E}_x$ and for all large enough t ,

$$F_{\text{cell},x}(t) \leq \frac{2}{p_{\min}} g(t) + T_x. \quad (99)$$

Summing over all $x \in \mathcal{X}$ and using finiteness of $\mathcal{X}$ gives, on $\bigcap_x \mathcal{E}_x$ (an event of probability one),

$$F_{\text{cell}}(t) = \sum_{x \in \mathcal{X}} F_{\text{cell},x}(t) \leq \frac{2K}{p_{\min}} g(t) + \sum_{x \in \mathcal{X}} T_x = O(g(t)) \quad \text{a.s.} \quad (100)$$

Since $g(t) = o(h(t))$ and $h(t) = t^a$ with $a \in (0, 1)$, we have $g(t) = o(t)$ and hence $F_{\text{cell}}(t) = o(t)$ almost surely.

Step 3: conclusion. Combining the bounds yields, almost surely,

$$F(t) = F_{\text{cov}}(t) + F_{\text{cell}}(t) = O(h(t)) + O(g(t)) = o(t). \quad (101)$$

□

F.5.5 Proof of Lemma 4

Lemma 4. *[Tracking of sampling proportions] Under Assumptions 3 and 4, the sampling proportions satisfy $N_x(t)/t \rightarrow w_x^*(P)$ for all $x \in \mathcal{X}$ almost surely.*

Proof. We work on the almost sure event on which both

$$\hat{w}(t) \rightarrow w^*(P) \quad \text{and} \quad F(t) = o(t) \quad (102)$$

hold, where $F(t)$ denotes the number of forced rounds up to time t .

Let $\mathcal{F} \subseteq \mathbb{N}$ be the (random) set of forced rounds and define $n(t) := t - F(t)$, the number of non-forced rounds up to time t . Then $n(t) \rightarrow \infty$ and $n(t)/t \rightarrow 1$.

Let $\tau_1 < \tau_2 < \dots$ denote the (decision) times at which the arm selection rule uses D-tracking rather than forcing. Since at time t the algorithm chooses X_{t+1} , the non-forced pull counts are defined by

$$\tilde{N}_x(k) := \sum_{j=1}^k \mathbf{1}\{X_{\tau_j+1} = x\}. \quad (103)$$

We have the decomposition

$$N_x(t) = \tilde{N}_x(n(t)) + N_x^{\text{F}}(t), \quad 0 \leq N_x^{\text{F}}(t) \leq F(t). \quad (104)$$

Step 1: induced dynamics on non-forced rounds. On each non-forced decision time τ_k , the algorithm selects

$$X_{\tau_k+1} \in \arg \max_{x \in \mathcal{X}} (\tau_k \hat{w}_x(\tau_k) - N_x(\tau_k)). \quad (105)$$

Using (104),

$$\tau_k \hat{w}_x(\tau_k) - N_x(\tau_k) = k \hat{w}_x(\tau_k) - \tilde{N}_x(k) + R_{k,x}, \quad (106)$$

where

$$R_{k,x} = (\tau_k - k) \hat{w}_x(\tau_k) - N_x^F(\tau_k). \quad (107)$$

Since $\tau_k - k = F(\tau_k) = o(\tau_k) = o(k)$ and $0 \leq N_x^F(\tau_k) \leq F(\tau_k) = o(k)$, we have uniformly in x ,

$$\sup_{x \in \mathcal{X}} |R_{k,x}| = o(k). \quad (108)$$

Thus the update rule on non-forced rounds is an $o(k)$ perturbation of the ideal D-tracking rule

$$\arg \max_x (k \hat{w}_x(\tau_k) - \tilde{N}_x(k)). \quad (109)$$

Step 2: convergence along non-forced rounds. Since $\hat{w}(\tau_k) \rightarrow w^*(P)$, the deterministic stability argument for D-tracking (cf. [Garivier and Kaufmann, 2016, Appendix B.2–B.3]) yields

$$\frac{\tilde{N}_x(k)}{k} \rightarrow w_x^*(P) \quad \text{for all } x \in \mathcal{X}. \quad (110)$$

Step 3: lifting back to real time. Using (104),

$$\frac{N_x(t)}{t} = \frac{n(t)}{t} \cdot \frac{\tilde{N}_x(n(t))}{n(t)} + \frac{N_x^F(t)}{t}. \quad (111)$$

Since $n(t)/t \rightarrow 1$, $\tilde{N}_x(n)/n \rightarrow w_x^*(P)$, and $N_x^F(t)/t \leq F(t)/t \rightarrow 0$, we conclude

$$\frac{N_x(t)}{t} \rightarrow w_x^*(P) \quad \text{for all } x \in \mathcal{X}. \quad (112)$$

□

F.5.6 Proof of Lemma 5

Lemma 5. [Asymptotic growth of the GLRT] For every competitor $x \neq x^*(P)$,

$$\mathbb{P}_P \left(\liminf_{t \rightarrow \infty} t^{-1} Z_t(x, \hat{x}(t)) \geq \inf_{Q \in \mathcal{P}: \theta_Q(x) \geq \theta_Q(x^*(P))} \sum_{a \in \mathcal{X}} w_a^*(P) \text{KL}(P_a \| Q_a) \right) = 1. \quad (35)$$

Consequently,

$$\mathbb{P}_P \left(\liminf_{t \rightarrow \infty} t^{-1} \min_{x \neq x^*(P)} Z_t(x, \hat{x}(t)) \geq T^*(P)^{-1} \right) = 1. \quad (36)$$

Proof. Work on an event of probability one on which the following hold simultaneously: (i) $\hat{P}(t) \rightarrow P$, (ii) $N_a(t)/t \rightarrow w_a^*(P)$ for all $a \in \mathcal{X}$, and (iii) $\hat{x}(t) = x^*(P)$ for all sufficiently large t . All three were established in Section 7.

Fix a competitor $x \neq x^*(P)$. For all sufficiently large t , we have $\hat{x}(t) = x^*(P)$, and therefore

$$Z_t(x, \hat{x}(t)) = \inf_{\substack{Q \in \mathcal{P}: \\ \theta_Q(x) \geq \theta_Q(x^*(P))}} \sum_{a \in \mathcal{X}} N_a(t) \text{KL}(\hat{P}_a(t) \| Q_a). \quad (113)$$

Let (t_n) be a subsequence with $t_n \rightarrow \infty$. For each n , choose Q_n in the constraint set such that

$$Z_{t_n}(x, \hat{x}(t_n)) \geq \sum_{a \in \mathcal{X}} N_a(t_n) \text{KL}(\hat{P}_a(t_n) \| (Q_n)_a) - \frac{1}{n}. \quad (114)$$

Since $\mathcal{P}$ is compact, we may extract a subsequence (not relabeled) such that $Q_n \rightarrow Q_\infty \in \mathcal{P}$.

Because $\hat{P}(t_n) \rightarrow P$ and $\theta_Q(\cdot)$ is continuous in Q , the constraint $\theta_{Q_n}(x) \geq \theta_{Q_n}(x^*(P))$ passes to the limit, hence Q_∞ satisfies $\theta_{Q_\infty}(x) \geq \theta_{Q_\infty}(x^*(P))$.

Under Assumption 3, $\text{KL}(\cdot \| \cdot)$ is continuous on $\mathcal{P} \times \mathcal{P}$. Thus for each arm a ,

$$\text{KL}(\hat{P}_a(t_n) \| (Q_n)_a) \rightarrow \text{KL}(P_a \| (Q_\infty)_a). \quad (115)$$

Combining with $N_a(t_n)/t_n \rightarrow w_a^*(P)$ yields

$$\lim_{n \rightarrow \infty} \frac{1}{t_n} \sum_{a \in \mathcal{X}} N_a(t_n) \text{KL}(\hat{P}_a(t_n) \| (Q_n)_a) = \sum_{a \in \mathcal{X}} w_a^*(P) \text{KL}(P_a \| (Q_\infty)_a). \quad (116)$$

Therefore,

$$\liminf_{n \rightarrow \infty} \frac{1}{t_n} Z_{t_n}(x, \hat{x}(t_n)) \geq \sum_{a \in \mathcal{X}} w_a^*(P) \text{KL}(P_a \| (Q_\infty)_a) \geq \inf_{\substack{Q \in \mathcal{P}: \\ \theta_Q(x) \geq \theta_Q(x^*(P))}} \sum_{a \in \mathcal{X}} w_a^*(P) \text{KL}(P_a \| Q_a). \quad (117)$$

Since the sequence (t_n) was arbitrary and limit infimum corresponds to one of such subsequences, we conclude

$$\liminf_{t \rightarrow \infty} \frac{1}{t} Z_t(x, \hat{x}(t)) \geq \inf_{\substack{Q \in \mathcal{P}: \\ \theta_Q(x) \geq \theta_Q(x^*(P))}} \sum_{a \in \mathcal{X}} w_a^*(P) \text{KL}(P_a \| Q_a) \quad \text{w.p. 1.} \quad (118)$$

Taking the minimum over $x \neq x^*(P)$ and recalling the definition of $T^*(P)$ from Theorem 2 yields

$$\liminf_{t \rightarrow \infty} \frac{1}{t} \min_{x \neq x^*(P)} Z_t(x, \hat{x}(t)) \geq \frac{1}{T^*(P)}. \quad (119)$$

□

E.5.7 Proof of Theorem 4 (Almost-sure asymptotic optimality)

Theorem 4. [Almost sure asymptotic optimality] Fix a true model $P \in \mathcal{P}$ satisfying Assumptions 2, 3 and 4. Then Algorithm 1 satisfies

$$\mathbb{P}_P \left(\limsup_{\delta \rightarrow 0} \frac{\tau_\delta}{\text{kl}(\delta, 1 - \delta)} \leq T^*(P) \right) = 1. \quad (37)$$

Proof. Let

$$Z(t) := \min_{x \neq \hat{x}(t)} Z_t(x, \hat{x}(t)) \quad (120)$$

be the GLRT statistic used in the stopping rule

$$\tau_\delta = \inf \{t \geq 1 : Z(t) \geq \beta(t, \delta)\}. \quad (121)$$

Step 1: Linear growth of the GLRT. By Lemma 5,

$$\liminf_{t \rightarrow \infty} \frac{Z(t)}{t} \geq \frac{1}{T^*(P)} \quad \text{w.p. 1} \quad (122)$$

Hence, for every $\varepsilon > 0$, there exists a (random) finite time t_ε (w.p. 1) such that for all $t \geq t_\varepsilon$,

$$Z(t) \geq \frac{t}{(1 + \varepsilon)T^*(P)} \quad \text{w.p. 1} \quad (123)$$

Step 2: Dominating $\beta(t, \delta)$ by a polynomial. From (46)

$$\beta(t, \delta) = \log\left(\frac{\pi^2 t^{2+2KM_Z}}{\delta}\right) + 6M_Z K \log(t+1). \quad (124)$$

For all $t \geq 1$, using $\log(t+1) \leq \log(2t) = \log 2 + \log t$, we obtain

$$\begin{aligned} \beta(t, \delta) &\leq \log(1/\delta) + \log(\pi^2) + (2 + 2KM_Z) \log t + 6KM_Z(\log 2 + \log t) \\ &= \log\left(\frac{C t^\alpha}{\delta}\right), \end{aligned} \quad (125)$$

where

$$\alpha := 2 + 8KM_Z, \quad C := \pi^2 2^{6KM_Z}. \quad (126)$$

Step 3: Solving the implicit inequality. Combining (123) and (125), for all $t \geq t_\varepsilon$, the condition $Z(t) \geq \beta(t, \delta)$ is implied by

$$\frac{t}{(1+\varepsilon)T^*(P)} \geq \log\left(\frac{C t^\alpha}{\delta}\right). \quad (127)$$

Let $a := (1+\varepsilon)T^*(P)$. Then it suffices that

$$t \geq a \log\left(\frac{C t^\alpha}{\delta}\right). \quad (128)$$

Applying Lemma 18 of Garivier and Kaufmann [2016] yields that, for all sufficiently small δ ,

$$\tau_\delta \leq a \left(\log(C/\delta) + \alpha \log \log(C/\delta) + O(1) \right) \quad \text{w.p. 1} \quad (129)$$

Step 4: Limsup bound. Dividing (129) by $\text{kl}(\delta, 1-\delta)$ and letting $\delta \downarrow 0$ gives, w.p. 1,

$$\limsup_{\delta \rightarrow 0} \frac{\tau_\delta}{\text{kl}(\delta, 1-\delta)} \leq (1+\varepsilon)T^*(P), \quad (130)$$

since $\text{kl}(\delta, 1-\delta) = \log(1/\delta) + O(1)$ and $\log \log(1/\delta) = o(\text{kl}(\delta, 1-\delta))$. As $\varepsilon > 0$ is arbitrary, letting $\varepsilon \downarrow 0$ yields

$$\limsup_{\delta \rightarrow 0} \frac{\tau_\delta}{\text{kl}(\delta, 1-\delta)} \leq T^*(P) \quad \text{w.p. 1.} \quad (131)$$

□