

---

# Enhancing Preference-based Linear Bandits via Human Response Time

---

Shen Li<sup>1\*</sup> Yuyang Zhang<sup>2\*</sup> Zhaolin Ren<sup>2</sup> Claire Liang<sup>1</sup> Na Li<sup>2</sup> Julie A. Shah<sup>1</sup>

<sup>1</sup>Massachusetts Institute of Technology <sup>2</sup>Harvard University

{shenli,cyl48}@mit.edu, julie\_a\_shah@csail.mit.edu

{yuyangzhang,zhaolinren}@g.harvard.edu, nali@seas.harvard.edu

## Abstract

Interactive preference learning systems infer human preferences by presenting queries as pairs of options and collecting binary choices. Although binary choices are simple and widely used, they provide limited information about preference strength. To address this, we leverage human response times, which are inversely related to preference strength, as an additional signal. We propose a computationally efficient method that combines choices and response times to estimate human utility functions, grounded in the EZ diffusion model from psychology. Theoretical and empirical analyses show that for queries with strong preferences, response times complement choices by providing extra information about preference strength, leading to significantly improved utility estimation. We incorporate this estimator into preference-based linear bandits for fixed-budget best-arm identification. Simulations on three real-world datasets demonstrate that using response times significantly accelerates preference learning compared to choice-only approaches. Additional materials, such as code, slides, and talk video, are available at <https://shenlirobot.github.io/pages/NeurIPS24.html>

## 1 Introduction

Interactive preference learning from human binary choices is widely used in recommender systems [9, 21, 32, 56], assistive robots [54, 65], and fine-tuning large language models [5, 43, 46, 47, 59]. This process is often framed as a preference-based bandit problem [7, 31], where the system repeatedly presents queries as pairs of options, the human selects a preferred option, and the system infers preferences from these choices. Binary choices are popular because they are easy to implement and impose low cognitive load on users [37, 72, 74]. However, while binary choices reveal preferences, they provide little information about preference strength [77]. To address this, researchers have incorporated additional *explicit human feedback*, such as ratings [50, 58], labels [74], and slider bars [5, 72], but these approaches often complicate interfaces and increase cognitive demands [36, 37].

In this paper, we propose leveraging *implicit human feedback*, specifically response times, to provide additional insights into preference strength. Unlike explicit feedback, response time is unobtrusive and effortless to measure [17], offering valuable information that complements binary choices [2, 16]. For instance, consider an online retailer that repeatedly presents users with a binary query, whether to purchase or skip a recommended product [35]. Since most users skip products most of the time [33], the probability of skipping becomes nearly 1 for most items. This lack of variation in choices makes it difficult to assess how much a user likes or dislikes any specific product, limiting the system’s ability to accurately infer their preferences. Response time can help overcome this limitation. Psychological research shows an inverse relationship between response time and preference strength [17]: users who strongly prefer to skip a product tend to do so quickly, while longer response times can indicate

---

\*First two authors have equal contribution.

weaker preferences. Thus, even when choices appear similar, response time can uncover subtle differences in preference strength, helping to accelerate preference learning.

Leveraging response times for preference learning presents notable challenges. Psychological research has extensively studied the relationship between human choices and response times [17, 19] using complex models like Drift-Diffusion Models [51] and Race Models [12, 66]. While these models align with both behavioral and neurobiological evidence [70], they rely on computationally intensive methods, such as hierarchical Bayesian inference [71] and maximum likelihood estimation (MLE) [52], to estimate the underlying human utility functions from both human choices and response times, making them impractical for real-time interactive systems. Although faster estimators exist [8, 28, 30, 67, 68], they typically estimate the utility functions for a single pair of options without aggregating data across multiple pairs. This limits their ability to leverage structures like linear utility functions, which are widely adopted both in preference learning with large option spaces [21, 24, 41, 54, 56] and in cognitive models for human multi-attribute decision-making [26, 64, 76].

To address these challenges, we propose a computationally efficient method for estimating linear human utility functions from both choices and response times, grounded in the difference-based EZ diffusion model [8, 67]. Our method leverages response times to transform binary choices into richer continuous signals, framing utility estimation as a *linear regression* problem that aggregates data across multiple pairs of options. We compare our estimator to traditional *logistic regression* methods that rely solely on choices [3, 31]. For queries with strong preferences, our theoretical and empirical analyses show that response times complement choices by providing additional information about preference strength. This significantly improves utility estimation compared to using choices alone. For queries with weak preferences, response times add little value but do not degrade performance. **In summary, response times complement choices, particularly for queries with strong preferences.**

Our linear-regression-based estimator integrates seamlessly into algorithms for preference-based bandits with linear human utility functions [3, 31], enabling interactive learning systems to leverage response times for faster learning. We specifically integrated our estimator into the Generalized Successive Elimination algorithm [3] for fixed-budget best-arm identification [29, 34]. Simulations using three real-world datasets [16, 39, 57] consistently show that incorporating response times significantly reduces identification errors, compared to traditional methods that rely solely on choices. *To the best of our knowledge, this is the first work to integrate response times into bandits (and RL).*

Section 2 introduces the preference-based linear bandit problem and the difference-based EZ diffusion model. Section 3 presents our utility estimator, incorporating both choices and response times, and offers a theoretical comparison to the choice-only estimator. Section 4 integrates both estimators into the Generalized Successive Elimination algorithm. Section 5 presents empirical results for estimation and bandit learning. Section 6 discusses the limitations of our approach. Appendix B reviews response time models, parameter estimation techniques, and their connection to preference-based RL.

*Nomenclature:* We use  $[n]$  to denote the set  $\{1, \dots, n\}$ . For a scalar random variable  $x$ , the expectation and variance are denoted by  $\mathbb{E}[x]$  and  $\mathbb{V}[x]$ , respectively. The function  $\text{sgn}(x)$  denotes the sign of  $x$ .

## 2 Problem setting and preliminaries

**Preference-based bandits with a linear utility function.** The learner is given a finite set of options (or “arms”), each represented by a feature vector in  $\mathcal{Z} \subset \mathbb{R}^d$ , and a finite set of binary queries, where each query is the difference between two arms, denoted by  $\mathcal{X} \subset \mathbb{R}^d$ . For instance, if the learner can query any pair of arms, the query space is  $\mathcal{X} = \{z - z': z, z' \in \mathcal{Z}\}$ . In the online retailer example from section 1, the query space is  $\mathcal{X} = \{z - z_{\text{skip}}: z \in \mathcal{Z}\}$ , where  $z$  represents purchasing a product and  $z_{\text{skip}}$  represents skipping (often set as  $\mathbf{0}$ ). For each arm  $z \in \mathcal{Z}$ , the human utility is assumed to be linear in the feature space, defined as  $u_z := z^\top \theta^*$ , where  $\theta^* \in \mathbb{R}^d$  represents the human’s preference parameters. For any query  $x \in \mathcal{X}$ , the utility difference is then defined as  $u_x := x^\top \theta^*$ .

Given a query  $x := z_1 - z_2 \in \mathcal{X}$ , we model human choices and response times using the difference-based EZ-Diffusion Model (dEZDM) [8, 67], integrated with our linear utility structure. (See appendix B.1 for a comparison with other models.) This model interprets human decision-making as a stochastic process in which evidence accumulates over time to compare two options. As shown in fig. 1a, after receiving a query  $x$ , the human first spends a fixed amount of non-decision time, denoted by  $t_{\text{nondec}} > 0$ , to perceive and encode the query. Then, evidence  $E_x$  accumulates over

time following a Brownian motion with drift  $x^\top \theta^*$  and two symmetric absorbing barriers,  $a > 0$  and  $-a$ . Specifically, at time  $t_{\text{nondec}} + \tau$  where  $\tau \geq 0$ , the evidence is  $E_{x,\tau} = x^\top \theta^* \cdot \tau + B(\tau)$ , where  $B(\tau) \sim \mathcal{N}(0, \tau)$  is standard Brownian motion. This process continues until the evidence reaches either the upper barrier  $a$  or lower barrier  $-a$ , at which point a decision is made. The random stopping time,  $t_x := \min \{\tau > 0 : E_{x,\tau} \in \{a, -a\}\}$ , represents the decision time. If  $E_{x,t_x} = a$ , the human chooses  $z_1$ ; if  $E_{x,t_x} = -a$ , they choose  $z_2$ . The choice is represented by the random variable  $c_x$ , where  $c_x = 1$  if  $z_1$  is chosen, and  $-1$  if  $z_2$  is chosen. The total response time,  $t_{\text{RT},x}$ , is the sum of the non-decision time and the decision time:  $t_{\text{RT},x} = t_{\text{nondec}} + t_x$ . The choice probability, expected choice, choice variance, and expected decision time are given as follows [48, eq. (A.16) and (A.17)]:

$$\begin{aligned} \forall x \in \mathcal{X} : \mathbb{P}[c_x = 1] &= \frac{1}{1 + \exp(-2ax^\top \theta^*)}, \quad \mathbb{E}[c_x] = \tanh(ax^\top \theta^*) \\ \mathbb{V}[c_x] &= 1 - \tanh^2(ax^\top \theta^*), \quad \mathbb{E}[t_x] = \begin{cases} \frac{a}{x^\top \theta^*} \tanh(ax^\top \theta^*) & \text{if } x^\top \theta^* \neq 0 \\ a^2 & \text{if } x^\top \theta^* = 0 \end{cases}. \end{aligned} \quad (1)$$

This choice probability matches that of the Bradley and Terry [10] model. If the learner relies solely on choices, then our bandit problem reduces to the transductive linear logistic bandit problem [31].

Figures 1b and 1c illustrate the roles of the parameters  $x^\top \theta^*$  and  $a$ . First, the absolute drift (or the absolute utility difference),  $|x^\top \theta^*|$ , reflects the human's preference strength for the query  $x$ . Larger values indicate stronger preferences, leading to faster decisions and more consistent choices. Smaller values suggest weaker preferences, resulting in slower decisions and less consistent choices. Second, the barrier  $a$  represents the human's conservativeness in decision-making [40]. A more conservative human (higher  $a$ ) requires more evidence to decide, resulting in slower but more consistent choices. In contrast, a less conservative human (lower  $a$ ) decides faster but makes less consistent choices.

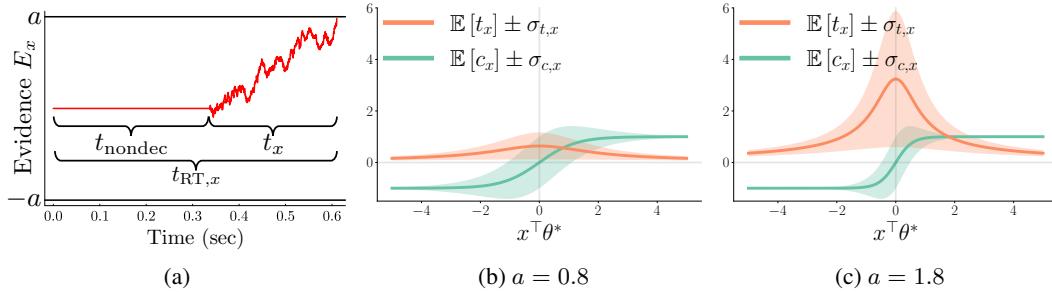


Figure 1: (a) depicts the human decision-making process for a binary query  $x \in \mathcal{X}$ , where the human selects between two arms. The human first spends a fixed non-decision time  $t_{\text{nondec}}$  encoding the query. Then, the human's evidence accumulates according to a Brownian motion with drift  $x^\top \theta^*$ . When the evidence reaches the upper barrier  $a$  or lower barrier  $-a$ , the human makes a choice, denoted by  $c_x = 1$  or  $c_x = -1$ , respectively. The random stopping time of the accumulation process is the decision time  $t_x$ , and the total response time is  $t_{\text{RT},x} = t_{\text{nondec}} + t_x$ . (b) and (c) plot the expected choice  $\mathbb{E}[c_x]$  and the expected decision time  $\mathbb{E}[t_x]$ , with shaded regions representing one standard deviation, plotted as functions of the utility difference  $x^\top \theta^*$  for two barrier values  $a$ .

We adopt the common assumption that  $t_{\text{nondec}}$  is constant across all queries for a given human [16, 76] and further assume that  $t_{\text{nondec}}$  is known to the learner. This assumption enables the learner to perfectly recover  $t_x$  from the observed  $t_{\text{RT},x}$ . In section 5.2, we empirically show that even when  $t_{\text{nondec}}$  is unknown, its impact on the performance of our method that relies on decision times is negligible.

**Learning objective: Best-arm identification with a fixed budget.** We focus on the fixed-budget best-arm identification problem [29, 34]. The learner is provided with a total interaction time budget  $B > 0$ , an arm space  $\mathcal{Z}$ , a query space  $\mathcal{X}$ , and a non-decision time  $t_{\text{nondec}}$ . Both the human's preference vector  $\theta^*$  and the decision barrier  $a$  are unknown. In each episode  $s \in \mathbb{N}$ , the learner selects a query  $x_s \in \mathcal{X}$ , receives human feedback  $(c_{x_s,s}, t_{x_s,s})$  generated by the dEZDM, and consumes  $t_{\text{RT},x_s,s}$  time. When the cumulative interaction time exceeds the budget  $B$  at some episode  $S$ , i.e.,  $\sum_{s=1}^S t_{\text{RT},x_s,s} > B$ , the learner must stop and recommend an arm  $\hat{z} \in \mathcal{Z}$ . The goal is to recommend the unique best arm  $z^* := \arg \max_{z \in \mathcal{Z}} z^\top \theta^*$ , minimizing the error probability  $\mathbb{P}[\hat{z} \neq z^*]$ .

To address this problem, we adopt the Generalized Successive Elimination (GSE) algorithm [1, 3, 75]. GSE divides the total budget  $B$  into multiple phases. In each phase, it strategically samples queries until the phase's budget is exhausted, collecting both human choices and decision times. It then estimates the preference vector  $\theta^*$  and eliminates arms with low estimated utilities. Decision times play a key role in the estimation step by providing complementary information about preference strength, which can enable more accurate estimation of  $\theta^*$  than choices alone. Next, in section 3, we introduce a novel estimator that combines decision times and choices to estimate  $\theta^*$ . Then, in section 4, we discuss how this estimator is integrated into GSE to improve preference learning.

### 3 Utility estimation

This section addresses the problem of estimating human preference  $\theta^*$  from a fixed dataset, denoted by  $\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\}_{x \in \mathcal{X}_{\text{sample}}, i \in [n_x]}$ . Here,  $\mathcal{X}_{\text{sample}}$  denotes the set of queries in the dataset,  $n_x$  denotes the number of samples for each query  $x \in \mathcal{X}_{\text{sample}}$ , and  $s_{x,i}$  denotes the episode when  $x$  is sampled for the  $i$ -th time. Samples from the same query  $x$  are i.i.d., while samples from different queries are independent. Section 3.1 introduces a new estimator, the “choice-decision-time estimator,” which uses both choices and decision times, in contrast to the commonly used “choice-only estimator” that only uses choices [3, 31]. Sections 3.2 and 3.3 theoretically compares these estimators, analyzing both asymptotic and non-asymptotic performance and highlighting the advantages of incorporating decision times. Section 5.1 presents empirical results that validate our theoretical insights.

#### 3.1 Choice-decision-time estimator and choice-only estimator

The choice-decision-time estimator is based on the following relationship between human utilities, choices, and decision times, derived from eq. (1):

$$\forall x \in \mathcal{X}: x^\top \theta^* = \frac{\mathbb{E}[c_x]}{\mathbb{E}[t_x]}. \quad (2)$$

Intuitively, when a human provides consistent choices (i.e., large  $|\mathbb{E}[c_x]|$ ) and makes decisions quickly (i.e., small  $\mathbb{E}[t_x]$ ), it implies a strong preference (i.e., large  $|x^\top \theta^*|$ ). This relationship formulates the estimation of  $\theta^*$  as a *linear regression* problem. Accordingly, the choice-decision-time estimator calculates the empirical means of both choices and decision times, aggregates the ratios across all sampled queries, and applies ordinary least squares (OLS) to estimate  $\theta^*/a$ . Since the ranking of arm utilities based on  $\theta^*/a$  is identical to that based on  $\theta^*$ , estimating  $\theta^*/a$  is sufficient for identifying the best arm. Formally, this estimate of  $\theta^*/a$ , denoted by  $\hat{\theta}_{\text{CH,DT}}$ , is given by:

$$\hat{\theta}_{\text{CH,DT}} := \left( \sum_{x \in \mathcal{X}_{\text{sample}}} n_x x x^\top \right)^{-1} \sum_{x \in \mathcal{X}_{\text{sample}}} n_x x \frac{\sum_{i=1}^{n_x} c_{x,s_{x,i}}}{\sum_{i=1}^{n_x} t_{x,s_{x,i}}}. \quad (3)$$

In contrast, the choice-only estimator is based on eq. (1), which shows that for each query  $x \in \mathcal{X}$ , the random variable  $(c_x + 1)/2$  follows a Bernoulli distribution with mean  $1/[1 + \exp(-x^\top \cdot 2a\theta^*)]$ . Similar to the choice-decision-time estimator, the parameter  $2a$  does not impact the ranking of arms, so estimating  $2a\theta^*$  is sufficient for best-arm identification. This estimation is formulated as a *logistic regression* problem [3, 31], with MLE providing the following estimate of  $2a\theta^*$ , denoted by  $\hat{\theta}_{\text{CH}}$ :

$$\hat{\theta}_{\text{CH}} := \arg \max_{\theta \in \mathbb{R}^d} \sum_{x \in \mathcal{X}_{\text{sample}}} \sum_{i=1}^{n_x} \log \mu(c_{x,s_{x,i}} x^\top \theta), \quad (4)$$

where  $\mu(y) := 1/[1 + \exp(-y)]$  is the standard logistic function. While this MLE lacks a closed-form solution, it can be efficiently solved using optimization methods like Newton's algorithm [25, 44].

#### 3.2 Asymptotic normality of the two estimators

The choice-decision-time estimator from eq. (3) satisfies the following asymptotic normality result:

**Theorem 3.1** (Asymptotic normality of  $\hat{\theta}_{CH,DT}$ ). *Given a fixed i.i.d. dataset  $\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\}_{i \in [n]}$  for each  $x \in \mathcal{X}_{sample}$ , where  $\sum_{x \in \mathcal{X}_{sample}} xx^\top \succ 0$ , and assuming that the datasets for different  $x \in \mathcal{X}_{sample}$  are independent, then, for any vector  $y \in \mathbb{R}^d$ , as  $n \rightarrow \infty$ , the following holds:*

$$\sqrt{n} y^\top (\hat{\theta}_{CH,DT,n} - \theta^*/a) \xrightarrow{D} \mathcal{N}(0, \zeta^2/a^2).$$

Here, the asymptotic variance depends on a problem-specific constant,  $\zeta^2$ , with an upper bound:

$$\zeta^2 \leq \|y\|_{\left(\sum_{x \in \mathcal{X}_{sample}} \left[\min_{x' \in \mathcal{X}_{sample}} \mathbb{E}[t_{x'}]\right] \cdot xx^\top\right)^{-1}}^2.$$

The proof is provided in appendix C.2. The asymptotic variance upper bound shows that all sampled queries are weighted by a common factor  $\min_{x' \in \mathcal{X}_{sample}} \mathbb{E}[t_{x'}]$ , which is the smallest expected decision time among all the sampled queries in  $\mathcal{X}_{sample}$ . This weight represents the amount of information provided by each query's choices and decision times for utility estimation. A larger weight indicates that all queries in  $\mathcal{X}_{sample}$  provides more information, leading to lower variance and better estimates.

In contrast, the choice-only estimator from eq. 4 has the following asymptotic normality result, as derived from Fahrmeir and Kaufmann [23, corollary 1]:

**Theorem 3.2** (Asymptotic normality of  $\hat{\theta}_{CH}$ ). *Given a fixed i.i.d. dataset  $\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\}_{i \in [n]}$  for each  $x \in \mathcal{X}_{sample}$ , where  $\sum_{x \in \mathcal{X}_{sample}} xx^\top \succ 0$ , and assuming that the datasets for different  $x \in \mathcal{X}_{sample}$  are independent, then, for any vector  $y \in \mathbb{R}^d$ , as  $n \rightarrow \infty$ , the following holds:*

$$\sqrt{n} y^\top (\hat{\theta}_{CH,n} - 2a\theta^*) \xrightarrow{D} \mathcal{N}\left(0, 4a^2 \|y\|_{\left(\sum_{x \in \mathcal{X}_{sample}} [a^2 \mathbb{V}[c_x]] \cdot xx^\top\right)^{-1}}^2\right).$$

This asymptotic variance shows that each sampled query  $x \in \mathcal{X}_{sample}$  is weighted by its own factor  $a^2 \mathbb{V}[c_x]$ , representing the amount of information the query's choices contribute to utility estimation. A larger weight indicates that the query contributes more information, leading to better estimates.

The weights in both theorems highlight the different contributions of choices and decision times to utility estimation. In the choice-only estimator (theorem 3.2), each query is weighted by  $a^2 \mathbb{V}[c_x]$ , which depends on the utility difference  $x^\top \theta^*$  for a fixed barrier  $a$ . As shown by the gray curves in fig. 2a, this weight quickly decays to zero as preferences become stronger (i.e., as  $|x^\top \theta^*|$  increases). This indicates that *choices from queries with strong preferences provide little information*. Intuitively, when preferences are strong, humans consistently select the same option, making it hard to distinguish whether their preference is moderately or very strong. As a result, choices from such queries contribute minimally to utility estimation. This intuition aligns with the online retailer example in section I.

For the choice-decision-time estimator (theorem 3.1), queries are weighted by  $\min_{x' \in \mathcal{X}_{sample}} \mathbb{E}[t_{x'}]$ , which depends on both  $\mathcal{X}_{sample}$  and  $\mathbb{E}[t_x]$ . To better understand this weight, we first plot  $\mathbb{E}[t_x]$  without the 'min' operator as the orange curves in fig. 2a. Comparing the orange and gray curves shows that  $\mathbb{E}[t_x]$  is generally larger than the choice-only weight,  $a^2 \mathbb{V}[c_x]$ . The actual weight in the choice-decision-time estimator, which is the minimum expected decision time across sampled queries, is less than or equal to the orange curve but is likely still higher than the choice-only weight, especially for queries with strong preferences. This suggests that *when preferences are strong, decision times complement choices by capturing preference strength, leading to improved estimation*.

When queries have weak preferences, the choice-decision-time weight may be lower than the choice-only weight. However, since the choice-decision-time weight represents only an upper bound on the asymptotic variance (theorem 3.1), no definitive conclusions can be drawn from the theory alone. Empirically, as shown in section 5.1, decision times add little value but do not degrade performance.

As the barrier  $a$  increases, the choice-decision-time weight rises. In contrast, the choice-only weight increases for queries with weak preferences, but this increase is concentrated in a narrower region, with weights decreasing elsewhere. Intuitively, a higher barrier reflects greater conservativeness in human decision-making, leading to longer decision times and more consistent choices (fig. I). As a result, more queries exhibit strong preferences, making choices from these queries less informative.

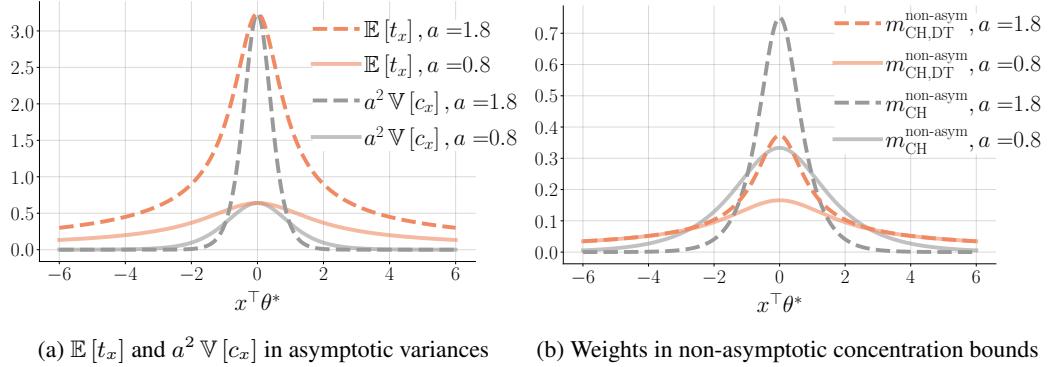


Figure 2: This figure illustrates key terms from our theoretical analyses, highlighting the different contributions of choices and decision times to utility estimation. These terms are functions of the utility difference  $x^\top \theta^*$  and are plotted for two barrier values,  $a$ . (a) compares the weights  $\mathbb{E}[t_x]$  and  $a^2 \mathbb{V}[c_x]$  in the asymptotic variances for the choice-decision-time estimator (orange, theorem 3.1) and the choice-only estimator (gray, theorem 3.2), respectively. This comparison shows that *decision times complement choices, particularly for queries with strong preferences*. (b) compares the weights in the non-asymptotic concentration bounds (theorems 3.3 and 3.4), showing similar trends, though these weights may not be optimal due to proof techniques.

### 3.3 Non-asymptotic concentration of the two estimators for utility difference estimation

In this section, we focus on the simpler problem of estimating the utility difference for a single query, without aggregating data from multiple queries. Comparing the non-asymptotic concentration bounds of both estimators, in this case, provides insights similar to those discussed in section 3.2. Extending this non-asymptotic analysis to the full estimation of the preference vector  $\theta^*$  is left for future work.

Given a query  $x \in \mathcal{X}$ , the task is to estimate the utility difference  $u_x := x^\top \theta^*$  using the fixed i.i.d. dataset  $\{(c_{x,s_{x,i}}, t_{x,s_{x,i}})\}_{i \in [n_x]}$ . Applying the choice-decision-time estimator from eq. (3), we get the following estimate (for details, see appendix C.3.1), which estimates  $u_x/a$  rather than  $u_x$ :

$$\hat{u}_{x,\text{CH,DT}} := \frac{\sum_{i=1}^{n_x} c_{x,s_{x,i}}}{\sum_{i=1}^{n_x} t_{x,s_{x,i}}}. \quad (5)$$

In contrast, applying the choice-only estimator from eq. (4), we get the following estimate (for details, see appendix C.3.2), which estimates  $2au_x$  rather than  $u_x$ :

$$\hat{u}_{x,\text{CH}} := \mu^{-1} \left( \frac{1}{n_x} \sum_{i=1}^{n_x} \frac{c_{x,s_{x,i}} + 1}{2} \right), \quad (6)$$

where  $(c_{x,s_{x,i}} + 1)/2$  is the binary choice coded as 0 or 1, and  $\mu^{-1}(p) := \log(p/(1-p))$  is the logit function (inverse of  $\mu$  introduced in eq. (4)).

Notably, the choice-only estimator in eq. (6) aligns with the EZ-diffusion model's drift estimator [67, eq. (5)]. Moreover, the estimators in Xiang Chiong et al. [73, eq. (6)] and Berlinghieri et al. [8, eq. (7)] combine elements of both estimators from eqs. (5) and (6). In section 5.2 we demonstrate that both estimators from Wagenmakers et al. [67, eq. (5)] and Xiang Chiong et al. [73, eq. (6)] are outperformed by our proposed estimator in eq. (3) for the full bandit problem.

Assuming the utility difference  $u_x \neq 0$ , the choice-decision-time estimator in eq. (5) satisfies the following non-asymptotic concentration bound, proven in appendix C.3.1:

**Theorem 3.3** (Non-asymptotic concentration of  $\hat{u}_{x,\text{CH,DT}}$ ). *For each query  $x \in \mathcal{X}$  with  $u_x \neq 0$ , given a fixed i.i.d. dataset  $\{(c_{x,s_{x,i}}, t_{x,s_{x,i}})\}_{i \in [n_x]}$ , for any  $\epsilon > 0$  satisfying  $\epsilon \leq \min\{|u_x|/(\sqrt{2}a), (1 + \sqrt{2})a|u_x|/\mathbb{E}[t_x]\}$ , the following holds:*

$$\mathbb{P} \left( \left| \hat{u}_{x,\text{CH,DT}} - \frac{u_x}{a} \right| > \epsilon \right) \leq 4 \exp \left( - [m_{\text{CH,DT}}^{\text{non-asym}}(x^\top \theta^*)]^2 n_x [\epsilon \cdot a]^2 \right),$$

where  $m_{\text{CH,DT}}^{\text{non-asym}}(x^\top \theta^*) := \mathbb{E}[t_x] / [(2 + 2\sqrt{2})a]$ .

In contrast, the choice-only estimator in eq. 6 has the following non-asymptotic concentration result, adapted from Jun et al. [31, theorem 5]<sup>2</sup>

**Theorem 3.4** (Non-asymptotic concentration of  $\hat{u}_{x,CH}$ ). *For each query  $x \in \mathcal{X}$ , given a fixed i.i.d. dataset  $\{c_{x,s_{x,i}}\}_{i \in [n_x]}$ , for any positive  $\epsilon < 0.3$ , if  $n_x \geq 1/\hat{\mu}(2au_x) \cdot \max\{3^2 \log(6e)/\epsilon^2, 64 \log(3)/(1 - \epsilon^2/0.3^2)\}$ , the following holds:*

$$\mathbb{P}(|\hat{u}_{x,CH} - 2au_x| > \epsilon) \leq 6 \exp\left(-[m_{CH}^{\text{non-asym}}(x^\top \theta^*)]^2 n_x [\epsilon/(2a)]^2\right),$$

where  $m_{CH}^{\text{non-asym}}(x^\top \theta^*) := a \sqrt{\mathbb{V}[c_x]} / 2.4$ .

The weights  $m_{CH,DT}^{\text{non-asym}}(\cdot)$  and  $m_{CH}^{\text{non-asym}}(\cdot)$  from theorems 3.3 and 3.4, respectively, are functions of the utility difference  $x^\top \theta^*$  for a fixed barrier  $a$ . These weights determine how quickly estimation errors decay as the dataset size  $n_x$  grows, with larger weights indicating faster error reduction. While these weights may not be optimal due to proof techniques, they highlight the distinct contributions of choices and decision times, consistent with our asymptotic analysis in section 3.2. Figure 2b compares the weights for the choice-decision-time estimator (orange,  $m_{CH,DT}^{\text{non-asym}}(\cdot)$ ) and the choice-only estimator (gray,  $m_{CH}^{\text{non-asym}}(\cdot)$ ). For strong preferences, the choice-only weights quickly decay to zero, while the choice-decision-time weights remain relatively large. This supports our key insight that decision times complement choices and improve estimation for queries with strong preferences.

In summary, both asymptotic (section 3.2) and non-asymptotic (section 3.3) analyses demonstrate that the choice-decision-time estimator extracts more information from queries with strong preferences. This finding aligns with prior empirical work [16] and is further supported by our results in section 5.1.

In fixed-budget best-arm identification, our choice-decision-time estimator's ability to extract more information from queries with strong preferences is especially valuable. Bandit learners, such as GSE [3], strategically sample queries, update estimates of  $\theta^*$ , and eliminate lower-utility arms. With the choice-only estimator, learners struggle to extract information from queries with strong preferences. To resolve this, one approach is to selectively sample queries with weak preferences, but this has two drawbacks. First, queries with weak preferences take longer to answer (i.e., require more resources), potentially lowering the 'bang per buck' (information per resource) [4]. Second, since  $\theta^*$  is unknown in advance, learners cannot reliably target queries with weak preferences. In contrast, with our choice-decision-time estimator, learners leverage decision times to gain more information from queries with strong preferences, improving bandit learning performance. We integrate both estimators into bandit learning in section 4 and evaluate their performance in section 5.

## 4 Interactive learning algorithm

We introduce the Generalized Successive Elimination (GSE) algorithm [1, 3, 75] for fixed-budget best-arm identification in preference-based linear bandits, and outline the key options for each GSE component, which we empirically compare in section 5.

The pseudo-code for GSE is shown in algorithm 1. The algorithm uses a hyperparameter  $\eta$  to control the number of phases, the budget per phase, and the number of arms eliminated in each phase. GSE divides the total budget  $B$  evenly across phases and reserves a buffer, sized by another hyperparameter  $B_{\text{buff}}$ , to prevent overspending in any phase (line 4). In each phase, GSE computes an experimental design  $\lambda$ , a probability distribution over the query space, to guide query sampling. We consider two designs: the transductive design [24],  $\lambda_{\text{trans}}$  (line 5), and the weak-preference design [31],  $\lambda_{\text{weak}}$  (line 6). Both designs minimize the worst-case variance of utility differences between surviving arms. The transductive design weights all queries equally, whereas the weak-preference design prioritizes queries with weak preferences to counter the choice-only estimator's difficulty in extracting information from queries with strong preferences (section 3). Since  $\theta^*$  is unknown, the weak-preference design identifies queries with weak preferences based on the previous phase's estimate  $\hat{\theta}_{CH}$ . Then, GSE samples queries based on the design (line 7) and, after exhausting the phase's budget, estimates  $\theta^*$  using either the choice-decision-time estimator  $\hat{\theta}_{CH,DT}$  (line 8) or the choice-only estimator  $\hat{\theta}_{CH}$  (line 9). It then eliminates arms with low estimated utilities (line 10). This process repeats until only one arm remains, which GSE recommends as the best arm (line 12).

<sup>2</sup>In Jun et al. [31, theorem 5], we let  $x_1 = \dots = x_t = 1$  and  $t_{\text{eff}} = d = 1$ .

The key difference between algorithm 1 and previous GSE algorithms [1, 3, 75] is that our setting involves queries with random response times, unknown to the learner. Previous work assumes fixed resource consumption per query and uses deterministic rounding methods [2, 24] to pre-allocate queries. This approach does not handle random resource usage. Instead, we adopt a random sampling procedure [13, 61] in line 7 to allocate queries based on the design. Random resource usage also requires tuning the elimination parameter  $\eta$ , to balance data collection and arm elimination, and the buffer size  $B_{\text{buff}}$ , to prevent overspending. In our empirical study (section 5.2), we manually tune both parameters. Further theoretical analysis is needed to better understand and optimize them.

---

**Algorithm 1** Generalized Successive Elimination (GSE) [3]

---

```

1: Input: Arm space  $\mathcal{Z}$ , query space  $\mathcal{X}$ , non-decision time  $t_{\text{nondec}}$ , and total budget  $B$ .
2: Hyperparameters: Elimination parameter  $\eta$  and buffer size  $B_{\text{buff}}$ .
3: Initialization:  $\mathcal{Z}_1 \leftarrow \mathcal{Z}$ .
4: for each phase  $k = 1, \dots, K := \lceil \log_\eta |\mathcal{Z}| \rceil$  with the budget  $B_k := B/K - B_{\text{buff}}$  do
5:   Design 1.  $\lambda_k := \lambda_{\text{trans},k} \leftarrow \arg \min_{\lambda \in \Delta^{|\mathcal{X}|}} \max_{z \neq z' \in \mathcal{Z}_k} \|z - z'\|_{(\sum_{x \in \mathcal{X}} \lambda_x x x^\top)^{-1}}^2$ .
6:   Design 2.  $\lambda_k := \lambda_{\text{weak},k} \leftarrow \arg \min_{\lambda \in \Delta^{|\mathcal{X}|}} \max_{z \neq z' \in \mathcal{Z}_k} \|z - z'\|_{(\sum_{x \in \mathcal{X}} \mu(x^\top \hat{\theta}_{k-1}) \lambda_x x x^\top)^{-1}}^2$ .
7:   Sample queries  $x_j \sim \lambda_k$  and stop at  $J_k$  if  $\sum_{j=1}^{J_k-1} t_{\text{RT},x_j,j} \leq B_k$  and  $\sum_{j=1}^{J_k} t_{\text{RT},x_j,j} > B_k$ .
8:   Estimate 1.  $\hat{\theta}_k := \hat{\theta}_{\text{CH,DT},k} \leftarrow$  apply eq. (3) to all the  $J_k$  samples.
9:   Estimate 2.  $\hat{\theta}_k := \hat{\theta}_{\text{CH},k} \leftarrow$  apply eq. (4) to all the  $J_k$  samples.
10:  Update  $\mathcal{Z}_{k+1} \leftarrow \text{Top-} \left\lceil \frac{|\mathcal{Z}_k|}{\eta} \right\rceil$  arms in  $\mathcal{Z}_k$ , ranked by the estimated utility  $z^\top \hat{\theta}_k$ .
11: end for
12: Output: the single one  $\hat{z} \in \mathcal{Z}_{K+1}$ .

```

---

## 5 Empirical results

This section empirically compares the GSE variations introduced in section 4: (1)  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$ : Transductive design with choice-decision-time estimator. (2)  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH}})$ : Transductive design with choice-only estimator. (3)  $(\lambda_{\text{weak}}, \hat{\theta}_{\text{CH}})$ : Weak-preference design with choice-only estimator.

### 5.1 Estimation performance on synthetic data

We evaluate the estimation performance of the GSE variations on the “sphere” synthetic problem, a standard linear bandit problem in the literature [20, 42, 61]. Details are provided in appendix D.1

Estimation performance, as discussed in section 3, depends on the utility difference  $x^\top \theta^*$  and the barrier  $a$ . We vary  $a$  over a range of values commonly used in psychology [16, 71]. To examine how preference strength impacts estimation, we scale each arm  $z$  to  $c_{\mathcal{Z}} \cdot z$ , effectively scaling each utility difference  $x^\top \theta^*$  to  $c_{\mathcal{Z}} \cdot x^\top \theta^*$ . Small  $c_{\mathcal{Z}}$  values correspond to problems with weak preferences, while large values correspond to strong preferences. For each  $(c_{\mathcal{Z}}, a)$  pair, the system generates 100 random problem instances and runs 100 repeated simulations per instance. In each simulation, the GSE variations sample 50 queries, ignoring the response time budget, and compute  $\hat{\theta}$ . Performance is evaluated by  $\mathbb{P}[\arg \max_{z \in \mathcal{Z}} z^\top \hat{\theta} \neq z^*]$ , which reflects the best-arm identification goal defined in section 2. To isolate the effect of estimation, we allow  $\lambda_{\text{weak}}$  access to the true  $\theta^*$ , enabling it to perfectly compute the terms  $\mu(x^\top \theta^*)$  used in line 6 of algorithm 1.

As shown in fig. 3a, fixing the barrier  $a$  and examining the vertical line, as  $c_{\mathcal{Z}}$  increases and preferences become stronger, the performance of the choice-only estimator with the transductive design first improves and then declines. The initial improvement arises because larger  $c_{\mathcal{Z}}$  increases utility differences between the best arm and others, theoretically simplifying best-arm identification. The subsequent decline, highlighted by the dark curved band, supports our insight from section 3 that choices from queries with strong preferences provide limited information. Fixing  $c_{\mathcal{Z}}$  and examining the horizontal line, performance first improves and then declines. This trend aligns with fig. 2a and section 3.2, where higher barriers  $a$  increase the choice-only weights for queries with weak

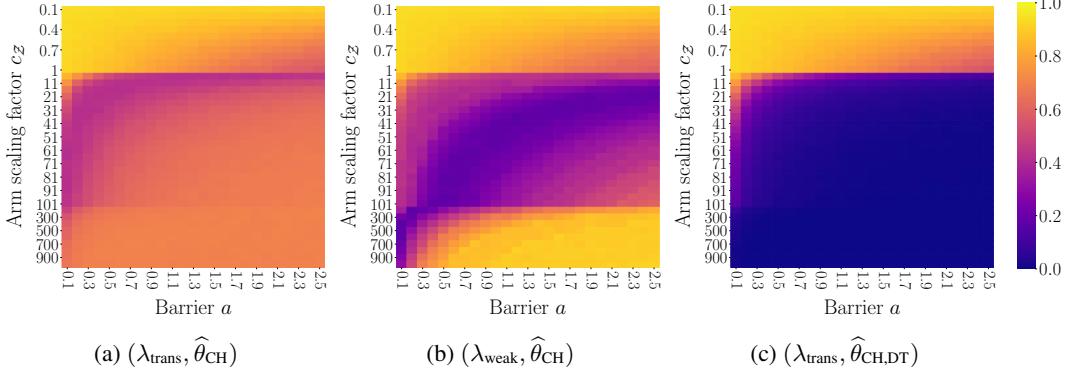


Figure 3: Three heatmaps show estimation error probabilities,  $\mathbb{P}[\arg \max_{z \in \mathcal{Z}} z^\top \hat{\theta} \neq z^*]$ , for three GSE variations, shown as functions of the arm scaling factor  $c_Z$  and barrier  $a$ . Darker colors indicate better estimation. (a) The choice-only estimator  $\hat{\theta}_{\text{CH}}$  with the transductive design  $\lambda_{\text{trans}}$  struggles as  $c_Z$  increases (i.e., preferences become stronger), highlighting that choices from queries with strong preferences provide limited information. (b) The weak-preference design  $\lambda_{\text{weak}}$  improves (a) by sampling queries with weak preferences but assumes perfect knowledge of  $\theta^*$  and equal resource consumption across queries. (c) The choice-decision-time estimator  $\hat{\theta}_{\text{CH,DT}}$  with  $\lambda_{\text{trans}}$  outperforms both choice-only methods in (a) and (b), showing that decision times complement choices and improve estimation, especially for strong preferences.

preferences, initially improving performance. However, as  $a$  grows, fewer queries exhibit increased weights, while most queries' weights decrease, leading to the later performance drop.

In Figure 3b, for moderate  $c_Z$ , the choice-only estimator with the weak-preference design outperforms the transductive design (fig. 3a), demonstrating that focusing on queries with weak preferences improves estimation. However, as  $c_Z$  becomes too large, performance declines because many  $\hat{\mu}(x^\top \theta^*)$  in line 6 of algorithm 1 approach zero, preventing informative queries from being sampled. This advantage of the weak-preference design assumes perfect knowledge of  $\theta^*$  and equal resource consumption across queries. In practice, where  $\theta^*$  is unknown and weak-preference queries require longer response times, the transductive design performs better, as shown in section 5.2.

Figure 3c shows that the choice-decision-time estimator consistently outperforms the choice-only estimators under both the transductive and weak-preference designs, particularly for strong preferences. This suggests that for queries with strong preferences, decision times complement choices and improve estimation, confirming our theoretical insights from section 3, while for queries with weak preferences, decision times add little value but do not degrade performance. The performance also improves with a higher barrier  $a$ , supporting the insights conveyed by fig. 2a and section 3.2.

## 5.2 Fixed-budget best-arm identification performance on real datasets

This section compares the bandit performance of six GSE variations. The first three are as previously defined at the beginning of section 5:  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH}})$ , and  $(\lambda_{\text{weak}}, \hat{\theta}_{\text{CH}})$ .

The 4th GSE variation,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,RT}})$ , evaluates the performance of the choice-decision-time estimator when the non-decision time  $t_{\text{nondec}}$  is unknown. The estimator,  $\hat{\theta}_{\text{CH,RT}}$ , is identical to the original choice-decision-time estimator from Eq. (3), but with response times used in place of decision times.

The 5th GSE variation,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,logit}})$ , is based on Wagenmakers et al. [67 eq. (5)], which states that  $x^\top \cdot (2a\theta^*) = \mu^{-1}(\mathbb{P}[c_x = 1])$ , where  $\mu^{-1}(p) := \log(p/(1-p))$ . By incorporating our linear utility structure, we obtain the following choice-only estimator  $\hat{\theta}_{\text{CH,logit}}$ :

$$\hat{\theta}_{\text{CH,logit}} := \left( \sum_{x \in \mathcal{X}_{\text{sample}}} n_x x x^\top \right)^{-1} \sum_{x \in \mathcal{X}_{\text{sample}}} n_x x \cdot \mu^{-1}(\hat{\mathbf{c}}_x),$$

where  $\hat{\mathbf{c}}_x := \frac{1}{n_x} \sum_{i=1}^{n_x} \frac{1}{2} (c_{x,s_{x,i}} + 1)$  is the empirical mean of the binary choices coded as 0 or 1.

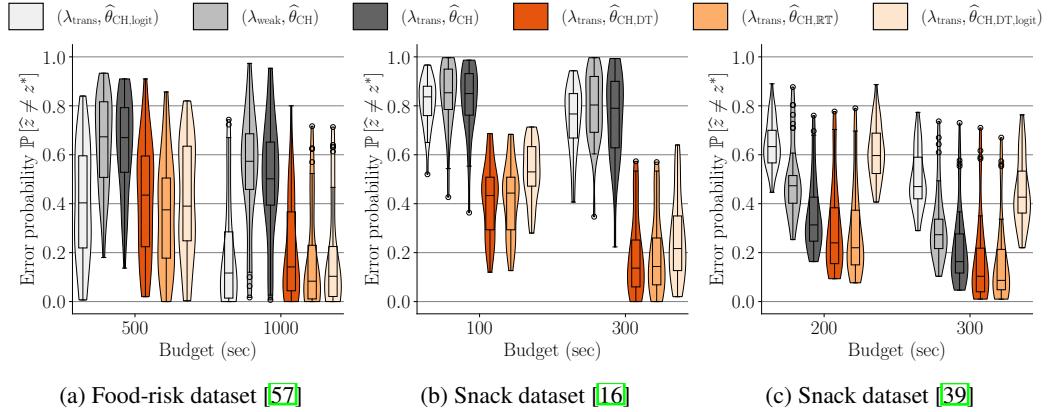


Figure 4: This figure shows violin plots (with overlaid box plots) for datasets (a), (b), and (c), showing the distribution of best-arm identification error probabilities,  $\mathbb{P}[\hat{z} \neq z^*]$ , for all bandit instances across six GSE variations and two budgets. The box plots follow the convention of the `matplotlib` Python package. For each GSE variation and budget, the horizontal line in the middle of the box represents the median of the error probabilities across all bandit instances. Each error probability is averaged over 300 repeated simulations under different random seeds. The box's upper and lower borders represent the third and first quartiles, respectively, with whiskers extending to the farthest points within  $1.5 \times$  the interquartile range. Flier points indicate outliers beyond the whiskers.

The 6th GSE variation,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT,logit}})$ , is based on Xiang Chiong et al. [73, eq. (6)], which states that  $x^\top \theta^* = \text{sgn}(c_x) \sqrt{\mathbb{E}[c_x] / \mathbb{E}[t_x]} \cdot 0.5 \mu^{-1} (\mathbb{P}[c_x = 1])$ . This identity forms the foundation of the estimator in Berlinghieri et al. [8, eq. (7)]. By incorporating our linear utility structure, we obtain the following choice-decision-time estimator  $\hat{\theta}_{\text{CH,DT,logit}}$ :

$$\widehat{\theta}_{\text{CH,DT,logit}} := \left( \sum_{x \in \mathcal{X}_{\text{sample}}} n_x x x^\top \right)^{-1} \sum_{x \in \mathcal{X}_{\text{sample}}} n_x x \cdot \text{sgn}(c_x) \sqrt{\frac{\mathbb{E}[c_x]}{\mathbb{E}[t_x]} \cdot \frac{1}{2} \mu^{-1}(\widehat{\mathfrak{C}}_x)}.$$

We evaluate six GSE variations on bandit instances constructed from three real-world datasets of human choices and response times. Each dataset includes multiple participants. For each participant, we estimated dEZDM parameters, built a bandit instance, and simulated the GSE variations to assess performance. Details on experimental procedures are provided in appendix D. Key results for the three domains are shown in fig. 4, with full results in appendix D. First,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$  consistently outperforms  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH}})$ , demonstrating the benefit of incorporating decision times. Second, both of these variations outperform  $(\lambda_{\text{weak}}, \hat{\theta}_{\text{CH}})$ , as discussed in section 5.1. Third,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$  performs similarly to  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,RT}})$ , suggesting that not knowing the non-decision time has minimal impact. Finally,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,logit}})$  [67] and  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT,logit}})$  [73] do not perform as consistently well as  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$ , highlighting the effectiveness of our proposed choice-decision-time estimator (eq. (3)).

## 6 Conclusion and future work

This work is the first to leverage human response times to improve fixed-budget best-arm identification in preference-based linear bandits. We proposed a utility estimator that combines choices and response times. Both theoretical and empirical analyses show that response times provide complementary information about preference strength, particularly for queries with strong preferences, enhancing estimation performance. When integrated into a bandit algorithm, incorporating response times consistently improved results across three real-world datasets.

One limitation of this approach is its reliance on reliable response time data, which may be challenging in crowdsourcing settings where participants' focus can vary [45]. Future work could integrate eye-tracking data into the DDM framework [26, 38, 39, 57, 76] to monitor attention and filter unreliable responses. Another direction is to relax the assumption of known non-decision times by estimating them directly from data, following methods proposed by Wagenmakers et al. [67].

## References

- [1] A. Alieva, A. Cutkosky, and A. Das. Robust pure exploration in linear bandits with limited budget. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 187–195. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/alieva21a.html>
- [2] C. Alós-Ferrer, E. Fehr, and N. Netzer. Time will tell: Recovering preferences when choices are noisy. *Journal of Political Economy*, 129(6):1828–1877, 2021. doi: 10.1086/713732. URL <https://doi.org/10.1086/713732>
- [3] M. Azizi, B. Kveton, and M. Ghavamzadeh. Fixed-budget best-arm identification in structured bandits. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2798–2804. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/388. URL <https://doi.org/10.24963/ijcai.2022/388>. Main Track.
- [4] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.
- [5] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL <https://arxiv.org/abs/2204.05862>
- [6] C. Baldassi, S. Cerreia-Vioglio, F. Maccheroni, M. Marinacci, and M. Pirazzini. A behavioral characterization of the drift diffusion model and its multialternative extension for choice under time pressure. *Management Science*, 66(11):5075–5093, 2020. doi: 10.1287/mnsc.2019.3475. URL <https://doi.org/10.1287/mnsc.2019.3475>
- [7] V. Bengs, R. Busa-Fekete, A. E. Mesaoudi-Paul, and E. Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021. URL <http://jmlr.org/papers/v22/18-546.html>
- [8] R. Berlinghieri, I. Krajbich, F. Maccheroni, M. Marinacci, and M. Pirazzini. Measuring utility with diffusion models. *Science Advances*, 9(34):eadf1665, 2023. doi: 10.1126/sciadv.adf1665. URL <https://www.science.org/doi/abs/10.1126/sciadv.adf1665>
- [9] V. Bogina, T. Kuflik, D. Jannach, M. Bielikova, M. Kompan, and C. Trattner. Considering temporal aspects in recommender systems: a survey. *User Modeling and User-Adapted Interaction*, 33(1):81–119, 2023. doi: 10.1007/s11257-022-09335-w. URL <https://doi.org/10.1007/s11257-022-09335-w>
- [10] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL <http://www.jstor.org/stable/2334029>
- [11] S. Brown and A. Heathcote. A ballistic model of choice response time. *Psychological review*, 112(1):117, 2005.
- [12] S. D. Brown and A. Heathcote. The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57(3):153–178, 2008. ISSN 0010-0285. doi: <https://doi.org/10.1016/j.cogpsych.2007.12.002>. URL <https://www.sciencedirect.com/science/article/pii/S0010028507000722>
- [13] R. Camilleri, K. Jamieson, and J. Katz-Samuels. High-dimensional experimental design and kernel bandits. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1227–1237. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/camilleri21a.html>
- [14] S. C. Castro, D. L. Strayer, D. Matzke, and A. Heathcote. Cognitive workload measurement and modeling under divided attention. *Journal of experimental psychology: human perception and performance*, 45(6):826, 2019.

[15] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL <https://proceedings.mlr.press/v15/chu11a.html>.

[16] J. A. Clithero. Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, 148:344–375, 2018. ISSN 0167-2681. doi: <https://doi.org/10.1016/j.jebo.2018.02.007>. URL <https://www.sciencedirect.com/science/article/pii/S0167268118300398>.

[17] J. A. Clithero. Response times in economics: Looking through the lens of sequential sampling models. *Journal of Economic Psychology*, 69:61–86, 2018. ISSN 0167-4870. doi: <https://doi.org/10.1016/j.jeop.2018.09.008>. URL <https://www.sciencedirect.com/science/article/pii/S0167487016306444>.

[18] D. R. Cox. *The theory of stochastic processes*. Routledge, 2017.

[19] P. De Boeck and M. Jeon. An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.00102. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00102>.

[20] R. Degenne, P. Menard, X. Shang, and M. Valko. Gamification of pure exploration for linear bandits. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2432–2442. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/degenne20a.html>.

[21] Y. Deldjoo, M. Schedl, and P. Knees. Content-driven music recommendation: Evolution, state of the art, and challenges. *Computer Science Review*, 51:100618, 2024. ISSN 1574-0137. doi: <https://doi.org/10.1016/j.cosrev.2024.100618>. URL <https://www.sciencedirect.com/science/article/pii/S1574013724000029>.

[22] J. Drugowitsch. Fast and accurate monte carlo sampling of first-passage times from wiener diffusion models. *Scientific reports*, 6(1):20490, 2016.

[23] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985. ISSN 00905364, 21688966. URL <http://www.jstor.org/stable/2241164>.

[24] T. Fiez, L. Jain, K. G. Jamieson, and L. Ratliff. Sequential experimental design for transductive linear bandits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2019/file/8ba6c657b03fc7c8dd4dff8e45defcd2-Paper.pdf>.

[25] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf>.

[26] G. Fisher. An attentional drift diffusion model over binary-attribute choice. *Cognition*, 168:34–45, 2017. ISSN 0010-0277. doi: <https://doi.org/10.1016/j.cognition.2017.06.007>. URL <https://www.sciencedirect.com/science/article/pii/S0010027717301695>.

[27] D. Fudenberg, P. Strack, and T. Strzalecki. Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12):3651–84, December 2018. doi: 10.1257/aer.20150742. URL <https://www.aeaweb.org/articles?id=10.1257/aer.20150742>.

[28] D. Fudenberg, W. Newey, P. Strack, and T. Strzalecki. Testing the drift-diffusion model. *Proceedings of the National Academy of Sciences*, 117(52):33141–33148, 2020. doi: 10.1073/pnas.2011446117. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2011446117>.

[29] V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2012/file/2a0a0a0a0a0a0a0a0a0a0a0a0a0a0a0a-Paper.pdf>.

sociates, Inc., 2012. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2012/file/8b0d268963dd0cfb808aac48a549829f-Paper.pdf>.

[30] R. P. Grasman, E.-J. Wagenmakers, and H. L. van der Maas. On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology*, 53(2):55–68, 2009. ISSN 0022-2496. doi: <https://doi.org/10.1016/j.jmp.2009.01.006>. URL <https://www.sciencedirect.com/science/article/pii/S0022249609000066>

[31] K.-S. Jun, L. Jain, B. Mason, and H. Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5148–5157. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/jun21a.html>

[32] M. Karimi, D. Jannach, and M. Jugovac. News recommender systems – survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227, 2018. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2018.04.008>. URL <https://www.sciencedirect.com/science/article/pii/S030645731730153X>

[33] N. Karpov and Q. Zhang. Instance-sensitive algorithms for pure exploration in multinomial logit bandit. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7096–7103, Jun. 2022. doi: 10.1609/aaai.v36i7.20669. URL <https://ojs.aaai.org/index.php/AAAI/article/view/20669>

[34] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL <http://jmlr.org/papers/v17/kaufman16a.html>

[35] A. Konovalov and I. Krajbich. Revealed strength of preference: Inference from response times. *Judgment and Decision Making*, 14(4):381–394, 2019. doi: 10.1017/S1930297500006082.

[36] P. Koppol, H. Admoni, and R. Simmons. Iterative interactive reward learning. In *Participatory Approaches to Machine Learning, International Conference on Machine Learning Workshop*, 2020.

[37] P. Koppol, H. Admoni, and R. Simmons. Interaction considerations in learning from humans. In Z.-H. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 283–291. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/40. URL <https://doi.org/10.24963/ijcai.2021/40>. Main Track.

[38] I. Krajbich. Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, 29:6–11, 2019. ISSN 2352-250X. doi: <https://doi.org/10.1016/j.copsyc.2018.10.008>. URL <https://www.sciencedirect.com/science/article/pii/S2352250X18301866>. Attention & Perception.

[39] I. Krajbich, C. Armel, and A. Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298, 2010. doi: 10.1038/nn.2635. URL <https://doi.org/10.1038/nn.2635>

[40] V. Lerche, A. Voss, and M. Nagler. How many trials are required for parameter estimation in diffusion modeling? a comparison of different optimization criteria. *Behavior Research Methods*, 49(2):513–537, 2017. doi: 10.3758/s13428-016-0740-2. URL <https://doi.org/10.3758/s13428-016-0740-2>

[41] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW ’10*, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL <https://doi.org/10.1145/1772690.1772758>

[42] Z. Li, K. Jamieson, and L. Jain. Optimal exploration is no harder than Thompson sampling. In S. Dasgupta, S. Mandt, and Y. Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1684–1692. PMLR, 02–04 May 2024. URL <https://proceedings.mlr.press/v238/li24h.html>

[43] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

[44] T. P. Minka. A comparison of numerical optimizers for logistic regression. *Unpublished draft*, 2003. URL <https://tminka.github.io/papers/logreg/minka-logreg.pdf>.

[45] C. E. Myers, A. Interian, and A. A. Moustafa. A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology*, 13, 2022. ISSN 1664-1078. doi: 10.3389/fpsyg.2022.1039172. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.1039172>.

[46] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[47] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf>.

[48] J. Palmer, A. C. Huk, and M. N. Shadlen. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, 5(5):1–1, 05 2005. ISSN 1534-7362. doi: 10.1167/5.5.1. URL <https://doi.org/10.1167/5.5.1>.

[49] M. L. Pedersen, M. J. Frank, and G. Biele. The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, 24(4):1234–1251, 2017. doi: 10.3758/s13423-016-1199-y. URL <https://doi.org/10.3758/s13423-016-1199-y>.

[50] M. Pérez-Ortiz, A. Mikhailuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2020. doi: 10.1109/TIP.2019.2936103.

[51] R. Ratcliff and G. McKoon. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, 20(4):873–922, 04 2008. ISSN 0899-7667. doi: 10.1162/neco.2008.12-06-420. URL <https://doi.org/10.1162/neco.2008.12-06-420>.

[52] R. Ratcliff and F. Tuerlinckx. Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9(3):438–481, 2002. doi: 10.3758/BF03196302. URL <https://doi.org/10.3758/BF03196302>.

[53] R. Ratcliff, P. L. Smith, S. D. Brown, and G. McKoon. Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4):260–281, 2016. ISSN 1364-6613. doi: <https://doi.org/10.1016/j.tics.2016.01.007>. URL <https://www.sciencedirect.com/science/article/pii/S1364661316000255>.

[54] D. Sadigh, A. Dragan, S. Sastry, and S. Seshia. Active preference-based learning of reward functions. In *Proceedings of Robotics: Science and Systems*, Cambridge, Massachusetts, July 2017. doi: 10.15607/RSS.2017.XIII.053.

[55] M. Shvartsman, B. Letham, E. Bakshy, and S. L. Keeley. Response time improves gaussian process models for perception and preferences. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[56] N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*, 197:116669, 2022. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2022.116669>. URL <https://www.sciencedirect.com/science/article/pii/S0957417422001543>.

[57] S. M. Smith and I. Krajbich. Attention and choice across domains. *Journal of Experimental Psychology: General*, 147(12):1810, 2018.

[58] T. Somers, N. R. Lawrence, and G. A. Hollinger. Efficient learning of trajectory preferences using combined ratings and rankings. In *Robotics: Science and Systems Conference Workshop on Mathematical Models, Algorithms, and Human-Robot Interaction*, 2017.

[59] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf>.

[60] T. Strzalecki. *Stochastic Choice Theory*. Econometric Society Monographs. Cambridge University Press, 2025. URL [https://scholar.harvard.edu/sites/scholar.harvard.edu/files/tomasz/files/manuscript\\_01.pdf](https://scholar.harvard.edu/sites/scholar.harvard.edu/files/tomasz/files/manuscript_01.pdf).

[61] C. Tao, S. Blanco, and Y. Zhou. Best arm identification in linear bandits with linear dimension dependency. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4877–4886. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/tao18a.html>.

[62] A. W. Thomas, F. Molter, I. Krajbich, H. R. Heekeren, and P. N. C. Mohr. Gaze bias differences capture individual choice behaviour. *Nature Human Behaviour*, 3(6):625–635, 2019. doi: 10.1038/s41562-019-0584-8. URL <https://doi.org/10.1038/s41562-019-0584-8>.

[63] A. Tirinzoni and R. Degenne. On elimination strategies for bandit fixed-confidence identification. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18586–18598. Curran Associates, Inc., 2022. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2022/file/760564ebba4797d0dcf1678e96e8cfc-Paper-Conference.pdf>.

[64] J. S. Trueblood, S. D. Brown, and A. Heathcote. The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, 121(2):179, 2014.

[65] M. Tucker, E. Novoseller, C. Kann, Y. Sui, Y. Yue, J. W. Burdick, and A. D. Ames. Preference-based learning for exoskeleton gait optimization. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2351–2357, 2020. doi: 10.1109/ICRA40945.2020.9196661.

[66] M. Usher and J. L. McClelland. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 108(3):550, 2001.

[67] E.-J. Wagenmakers, H. L. J. Van Der Maas, and R. P. P. P. Grasman. An ez-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 14(1):3–22, 2007. doi: 10.3758/BF03194023. URL <https://doi.org/10.3758/BF03194023>.

[68] E.-J. Wagenmakers, H. L. J. van der Maas, C. V. Dolan, and R. P. P. P. Grasman. Ez does it! extensions of the ez-diffusion model. *Psychonomic Bulletin & Review*, 15(6):1229–1235, 2008. doi: 10.3758/PBR.15.6.1229. URL <https://doi.org/10.3758/PBR.15.6.1229>.

[69] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[70] R. Webb. The (neural) dynamics of stochastic choice. *Management Science*, 65(1):230–255, 2019. doi: 10.1287/mnsc.2017.2931. URL <https://doi.org/10.1287/mnsc.2017.2931>.

[71] T. V. Wiecki, I. Sofer, and M. J. Frank. Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, 7, 2013. ISSN 1662-5196. doi: 10.3389/fninf.2013.00014. URL <https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2013.00014>.

[72] N. Wilde, E. Biyik, D. Sadigh, and S. L. Smith. Learning reward functions from scale feedback. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 353–362. PMLR, 08–11 Nov 2022. URL <https://proceedings.mlr.press/v164/wilde22a.html>.

[73] K. Xiang Chiong, M. Shum, R. Webb, and R. Chen. Combining choice and response time data: A drift-diffusion model of mobile advertisements. *Management Science*, 70(2):1238–1257, 2024. doi: 10.1287/mnsc.2023.4738. URL <https://doi.org/10.1287/mnsc.2023.4738>.

[74] Y. Xu, H. Zhang, K. Miller, A. Singh, and A. Dubrawski. Noise-tolerant interactive learning using pairwise comparisons. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/e11943a6031a0e6114ae69c257617980-Paper.pdf>.

[75] J. Yang and V. Tan. Minimax optimal fixed-budget best arm identification in linear bandits. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 12253–12266. Curran Associates, Inc., 2022. URL <https://proceedings.neurips.cc/paper%5Ffiles/paper/2022/file/4f9342b74c3bb63f6e030d8263082ab6-Paper-Conference.pdf>.

[76] X. Yang and I. Krajbich. A dynamic computational model of gaze and choice in multi-attribute decisions. *Psychological Review*, 130(1):52, 2023.

[77] H. Yu, R. M. Aronson, K. H. Allen, and E. S. Short. From “thumbs up” to “10 out of 10”: Reconsidering scalar feedback in interactive reinforcement learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4121–4128, 2023. doi: 10.1109/IROS55552.2023.10342458.

[78] C. Zhang, C. Kemp, and N. Lipovetzky. Goal recognition with timing information. *Proceedings of the International Conference on Automated Planning and Scheduling*, 33(1):443–451, Jul. 2023. doi: 10.1609/icaps.v33i1.27224. URL <https://ojs.aaai.org/index.php/ICAPS/article/view/27224>.

[79] C. Zhang, C. Kemp, and N. Lipovetzky. Human goal recognition as bayesian inference: Investigating the impact of actions, timing, and goal solvability. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’24, page 2066–2074, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.

## A Broader impacts

Incorporating human response times in human-interactive AI systems provides significant benefits, such as efficiently eliciting user preferences, reducing cognitive loads on users, and improving accessibility for users with disabilities and various cognitive abilities. These benefits can greatly improve recommendation systems, assistive robots, online shopping platforms, and fine-tuning for large language models. However, using human response times also raises concerns about privacy, manipulation, and bias against individuals with slower response times. Governments and law enforcement should work together to mitigate these negative consequences by establishing ethical standards and regulations. Businesses should always obtain user consent before recording response times.

## B Literature review

### B.1 Bounded accumulation models for choices and response times

Bounded Accumulation Models (BAMs) describe human decision-making using an accumulator (or sampling rule) and a stopping rule [70]. In binary choice tasks, such as two-alternative forced choice tasks, a widely used BAM is the drift-diffusion model (DDM) [51], which models decisions as Brownian motion with fixed boundaries. To capture differences in human response times for correct and incorrect answers, Ratcliff and McKoon [51] allows drift, starting point, and non-decision time to vary across trials. Wagenmakers et al. [67] later introduced the EZ-diffusion model (EZDM), a simplified version of DDM with closed-form solutions for choice and response time moments, making parameter estimation easier and more robust. EZDM assumes deterministic drift, starting point, and non-decision time, fixed across trials, with the starting point equidistant from the boundaries. Berlinghieri et al. [8] specialized EZDM to the difference-based EZDM (dEZDM), where the drift represents the utility difference between two options. For binary queries with arms  $z_1$  and  $z_2$ , the drift is modeled as  $u_{z_1} - u_{z_2}$ , where  $u_{z_1}$  and  $u_{z_2}$  are the utilities of  $z_1$  and  $z_2$ .

As discussed in section 2, we impose a linear utility structure on the dEZDM, where each arm's utility is given by  $u_z = z^\top \theta^*$ , with  $\theta^*$  denotes the human preference vector. This approach is supported by both bandit and psychology literature. In bandits, linear utility models scale efficiently with a large number of arms [15, 41]. In psychology, linear combinations of attributes are commonly used in multi-attribute decision-making models [26, 64, 76]. The standard dEZDM in [8, Definition 1] is a special case of our dEZDM with a linear utility structure, where arms correspond to the standard basis vectors in Euclidean space  $\mathbb{R}^d$ . This mirrors the relationship between multi-armed bandits and linear bandits.

Similarly to our approach, Shvartsman et al. [55] parameterize the human utility function as a Gaussian process and propose a moment-matching Bayesian inference method that uses both choices and response times to estimate latent utilities. Unlike our work, their focus is solely on estimation and does not address bandit optimization. Integrating their estimation techniques into bandit optimization presents an interesting avenue for future research.

Another widely used BAM is the race model [11, 66], which naturally extends to queries with more than two options. In race models, each option has its own accumulator, and the decision ends when any accumulator reaches its barrier. BAMs can also model human attention during decision-making. For example, the attentional-DDM [38, 39, 76] jointly models choices, response times, and eye movements across different options or attributes. Similarly, Thomas et al. [62] introduce the gaze-weighted linear accumulator model to study gaze bias at the trial level. To incorporate learning effects, Pedersen et al. [49] combine reinforcement learning (RL) with DDM, where the human adjusts the drift through RL. In contrast, our work uses RL for AI decision-making when interacting with humans. BAMs also connect to Bayesian RL models of human cognition. For example, Fudenberg et al. [27] propose a model where humans balance decision accuracy and time cost, showing it is equivalent to a DDM with time-decaying boundaries. Neurophysiological evidence supports BAMs. For instance, EEG recordings demonstrate that neurons exhibit accumulation processes and decision thresholds [70]. Additionally, diffusion processes have been used to model neural firing rates [53].

### B.2 Parameter estimation for bounded accumulation models

BAMs often lack closed-form density functions, so hierarchical Bayesian inference is commonly used for parameter estimation [71]. While flexible, these methods are computationally intensive, making them impractical for real-time applications in online learning systems. Faster estimators [8, 67, 73] usually estimate parameters for individual option pairs without leveraging data across pairs. To address this, we propose a computationally efficient method for estimating linear human utility functions, which we integrate into bandit learning. In section 5.2 we empirically show that our estimator outperforms those from prior work [67, 73].

In practice, using response time data requires pre-processing and model fitting, as outlined by Myers et al. [45]. Additionally, Alós-Ferrer et al. [2], Baldassi et al. [6], Fudenberg et al. [28] propose statistical tests to assess the suitability of various DDM extensions for a given dataset.

### B.3 Uses of response times

Response times serve multiple purposes, as highlighted by Clithero [17]. A primary use is improving choice prediction. For instance, Clithero [16] showed that DDM predicts choice probabilities more accurately than the logit model, with parameters estimated through Bayesian Markov chain Monte Carlo. Similarly, Alós-Ferrer et al. [2] demonstrated that response times enhance the identifiability of human preferences compared to using choices alone.

Response times also shed light on human decision-making processes. Castro et al. [14] applied DDM analysis to explore how cognitive workload, induced by secondary tasks, influences decision-making. Analyzing response times has been a long-standing method in cognitive testing to assess mental capabilities [19]. Additionally, Zhang et al. [78, 79] introduced a framework that uses human planning time to infer their intended goals.

Response times can also enhance AI decision-making. In dueling bandits and preference-based RL [7], human choice models are commonly used for preference elicitation. One such model, the random utility model, can be derived from certain BAMs [2]. For example, as discussed after eq. (1), both the Bradley-Terry model [10] and dEZDM [8, 67] yield logistic choice probabilities in the form  $\mathbb{P}[z_1 \succ z_2] = \sigma_{logistic}(u_{z_1} - u_{z_2}) = 1 / (1 + \exp(-c \cdot (u_{z_1} - u_{z_2})))$ , where  $u_{z_1}$  and  $u_{z_2}$  denote the utilities of  $z_1$  and  $z_2$  and  $c$  is some constant [7, section 3.2]. Our work leverages this connection between random utility models and choice-response-time models to estimate human utilities using both choices and response times.

We hypothesize that our key insight, that response times provide complementary information, especially for queries with strong preferences, extends beyond the dEZDM and the specific logistic link function  $\sigma_{logistic}$ . Many psychological models capture both choices and response times but lack closed-form choice distributions. In such cases, the choice probability is often expressed as  $\mathbb{P}[z_1 \succ z_2] = \sigma^\dagger(u_{z_1}, u_{z_2})$ , where  $\sigma^\dagger$  is a function of  $u_{z_1}$  and  $u_{z_2}$  without a closed form. Fixing  $u_{z_2}$  and varying  $u_{z_1}$  defines the psychometric function  $\sigma^\dagger(\cdot, u_{z_2})$ , which typically exhibits an “S” shape [60, fig. 1.1]. As preferences become stronger,  $\sigma^\dagger$  flattens, similar to figs. 1b and 1c, suggesting that choices carry less information. We conjecture that response times remain a valuable complementary signal in such cases.

If we further assume the choice probability depends only on the utility difference,  $u_{z_1} - u_{z_2}$ , then  $\mathbb{P}[z_1 \succ z_2] = \sigma^\ddagger(u_{z_1} - u_{z_2})$ , where the link function  $\sigma^\ddagger$  is typically assumed to be strictly monotonic and bounded within  $[0, 1]$  [7, section 3.2]. These properties naturally produce an “S”-shaped curve that flattens as preferences become stronger, again suggesting that choices provide less information. In such cases, we conjecture that response times can complement choices to enhance learning.

In summary, BAMs, like DDMs and race models, offer a strong theoretical framework for understanding human decision-making, supported by both behavioral and neurophysiological evidence. These models have been widely applied to choice prediction and the study of human cognitive processes. Our work connects BAMs with bandit algorithms by introducing a computationally efficient estimator for online preference learning. Future research could explore other BAM variants to further examine the benefits of incorporating response times.

## C Proofs

### C.1 Parameters of the difference-based EZ-Diffusion Model (dEZDM) [8, 67]

Given a human preference vector  $\theta^*$ , for each query  $x \in \mathcal{X}$ , the utility difference is defined as  $u_x := x^\top \theta^*$ . In the dEZDM model (introduced in section 2), with barrier  $a$ , according to Wagenmakers et al. [67], eq. (4), (6), and (9)], the human choice  $c_x$  has the following properties:

$$\mathbb{P}(c_x = 1) = \frac{1}{1 + \exp(-2au_x)}, \quad \mathbb{P}(c_x = -1) = \frac{\exp(-2au_x)}{1 + \exp(-2au_x)}.$$

Thus, the expected choice is  $\mathbb{E}[c_x] = \tanh(au_x)$ , and the choice variance is  $\mathbb{V}[c_x] = 1 - \tanh(au_x)^2$  (restating eq. (1)).

The human decision time  $t_x$  has the following properties:

$$\begin{aligned} \mathbb{E}[t_x] &= \begin{cases} \frac{a}{u_x^2} \tanh(au_x) & \text{if } u_x \neq 0 \\ a^2 & \text{if } u_x = 0 \end{cases} \quad (\text{restating eq. (2)}), \\ \mathbb{V}[t_x] &= \begin{cases} \frac{a}{u_x^3} \frac{\exp(4au_x) - 1 - 4au_x \exp(2au_x)}{(\exp(2au_x) + 1)^2} & \text{if } u_x \neq 0 \\ 2a^4/3 & \text{if } u_x = 0 \end{cases}. \end{aligned}$$

From this, we obtain the following key relationship:

$$\frac{\mathbb{E}[c_x]}{\mathbb{E}[t_x]} = \frac{u_x}{a} = x^\top \left( \frac{1}{a} \theta^* \right) \quad (\text{restating eq. (2)}).$$

All these parameters depend solely on the utility difference  $u_x := x^\top \theta^*$  and the barrier  $a$ .

### C.2 Asymptotic normality of the choice-decision-time estimator for estimating the human preference vector $\theta^*$

We now present the proof of the asymptotic normality result for the choice-decision-time estimator,  $\widehat{\theta}_{\text{CH,DT}}$ , as stated in theorem 3.1, which is restated as follows:

**Theorem 3.1** (Asymptotic normality of  $\widehat{\theta}_{\text{CH,DT}}$ ). *Given a fixed i.i.d. dataset  $\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\}_{i \in [n]}$  for each  $x \in \mathcal{X}_{\text{sample}}$ , where  $\sum_{x \in \mathcal{X}_{\text{sample}}} x x^\top \succ 0$ , and assuming that the datasets for different  $x \in \mathcal{X}_{\text{sample}}$  are independent, then, for any vector  $y \in \mathbb{R}^d$ , as  $n \rightarrow \infty$ , the following holds:*

$$\sqrt{n} y^\top \left( \widehat{\theta}_{\text{CH,DT},n} - \theta^*/a \right) \xrightarrow{D} \mathcal{N}(0, \zeta^2/a^2).$$

Here, the asymptotic variance depends on a problem-specific constant,  $\zeta^2$ , with an upper bounded:

$$\zeta^2 \leq \|y\|_{\left( \sum_{x \in \mathcal{X}_{\text{sample}}} \left[ \min_{x' \in \mathcal{X}_{\text{sample}}} \mathbb{E}[t_{x'}] \right] \cdot x x^\top \right)^{-1}}^2.$$

*Proof.* To simplify notation, we define:

$$\widehat{c}_x = \frac{1}{n} \sum_{i=1}^n c_{x,s_{x,i}}, \quad \mathcal{C}_x = \mathbb{E}[c_x], \quad \widehat{\mathcal{T}}_x = \frac{1}{n} \sum_{i=1}^n t_{x,s_{x,i}}, \quad \mathcal{T}_x = \mathbb{E}[t_x]. \quad (7)$$

For brevity, we abbreviate  $\mathcal{X}_{\text{sample}}$  as  $\mathcal{X}$  and  $\widehat{\theta}_{\text{CH,DT},n}$  as  $\widehat{\theta}$ . The estimator  $\widehat{\theta}$  can be expressed as:

$$\widehat{\theta} = \left( \sum_{x' \in \mathcal{X}} n x' x'^\top \right)^{-1} \sum_{x \in \mathcal{X}} n x \frac{\widehat{c}_x}{\widehat{\mathcal{T}}_x} \quad (\text{restating eq. (3)}).$$

We rewrite  $\theta^*/a$  as:

$$\begin{aligned} \theta^*/a &= \left( \sum_{x' \in \mathcal{X}} n x' x'^\top \right)^{-1} \sum_{x \in \mathcal{X}} n x x^\top \frac{\theta^*}{a} \\ &= \left( \sum_{x' \in \mathcal{X}} n x' x'^\top \right)^{-1} \sum_{x \in \mathcal{X}} n x \frac{\mathcal{C}_x}{\mathcal{T}_x}. \end{aligned} \quad (8)$$

Therefore, for any vector  $y \in \mathbb{R}^d$ , we have:

$$y^\top \left( \hat{\theta} - \frac{\theta^*}{a} \right) = y^\top \left( \sum_{x' \in \mathcal{X}} nx' x'^\top \right)^{-1} \sum_{x \in \mathcal{X}} nx \left( \frac{\hat{\mathcal{C}}_x}{\hat{\mathcal{T}}_x} - \frac{\mathcal{C}_x}{\mathcal{T}_x} \right) =: \sum_{x \in \mathcal{X}} \xi_x \left( \frac{\hat{\mathcal{C}}_x}{\hat{\mathcal{T}}_x} - \frac{\mathcal{C}_x}{\mathcal{T}_x} \right), \quad (9)$$

where  $\xi_x$  is defined as  $\xi_x := y^\top (\sum_{x' \in \mathcal{X}} nx' x'^\top)^{-1} nx$ . In eq. (9), the only random variables are  $\hat{\mathcal{C}}_x$  and  $\hat{\mathcal{T}}_x$ . For simplicity, for any  $x_i \in \mathcal{X} := \{x_1, \dots, x_{|\mathcal{X}|}\}$ , we slightly abuse the notation and use  $\xi_i, c_i, t_i, \mathcal{C}_i, \mathcal{T}_i, \hat{\mathcal{C}}_i$  and  $\hat{\mathcal{T}}_i$  denote  $\xi_{x_i}, c_{x_i}, t_{x_i}, \mathcal{C}_{x_i}, \mathcal{T}_{x_i}, \hat{\mathcal{C}}_{x_i}$ , and  $\hat{\mathcal{T}}_{x_i}$ , respectively. By applying the multidimensional central limit theorem, we have:

$$\begin{aligned} \sqrt{n} \begin{bmatrix} \hat{\mathcal{C}}_1 - \mathcal{C}_1 \\ \hat{\mathcal{T}}_1 - \mathcal{T}_1 \\ \vdots \\ \hat{\mathcal{C}}_{|\mathcal{X}|} - \mathcal{C}_{|\mathcal{X}|} \\ \hat{\mathcal{T}}_{|\mathcal{X}|} - \mathcal{T}_{|\mathcal{X}|} \end{bmatrix} &\xrightarrow{D} \mathcal{N} \left( 0, \begin{bmatrix} \mathbb{V}[c_1] & \text{cov}[c_1, t_1] & & & \\ \text{cov}[t_1, c_1] & \mathbb{V}[t_1] & & & \\ & & \ddots & & \\ & & & \mathbb{V}[c_{|\mathcal{X}|}] & \text{cov}[c_{|\mathcal{X}|}, t_{|\mathcal{X}|}] \\ & & & \text{cov}[t_{|\mathcal{X}|}, c_{|\mathcal{X}|}] & \mathbb{V}[t_{|\mathcal{X}|}] \end{bmatrix} \right) \\ &= \mathcal{N} \left( 0, \text{diag} \left[ \mathbb{V}[c_1], \mathbb{V}[t_1], \dots, \mathbb{V}[c_{|\mathcal{X}|}], \mathbb{V}[t_{|\mathcal{X}|}] \right] \right). \end{aligned} \quad (10)$$

In the first line of eq. (10), the block-diagonal structure of the covariance matrix emerges because  $(\hat{\mathcal{C}}_i, \hat{\mathcal{T}}_i)_{i \in [|\mathcal{X}|]}$  are independent of each other. For any fixed  $x_i$ , to derive the second line of eq. (10), we use the fact that:

$$\begin{aligned} \mathbb{E}[t_i c_i] &= \mathbb{P}(c_i = 1) \mathbb{E}[1 \cdot t_i | c_i = 1] + \mathbb{P}(c_i = -1) \mathbb{E}[-1 \cdot t_i | c_i = -1] \\ &\stackrel{(i)}{=} (\mathbb{P}(c_i = 1) - \mathbb{P}(c_i = -1)) \mathbb{E}[t_i | c_i = 1] \\ &= \mathbb{E}[c_i] \mathbb{E}[t_i], \end{aligned} \quad (11)$$

where (i) is because  $\mathbb{E}[t_i | c_i = 1] = \mathbb{E}[t_i | c_i = -1]$  [48, eq. (A.7) and (A.9)]. Therefore, eq. (11) implies that  $\text{cov}(c_i, t_i) = 0$ <sup>3</sup>, which justifies the second line of eq. (10).

Now, let us define the function  $g(c_1, t_1, \dots, c_{|\mathcal{X}|}, t_{|\mathcal{X}|}) := \sum_{i \in [|\mathcal{X}|]} \xi_i c_i / t_i$ . The gradient of  $g$  is:

$$\nabla g|_{(c_1, t_1, \dots, c_{|\mathcal{X}|}, t_{|\mathcal{X}|})} = [\xi_1/t_1 \quad -\xi_1 c_1/t_1^2 \quad \dots \quad \xi_{|\mathcal{X}|}/t_{|\mathcal{X}|} \quad -\xi_{|\mathcal{X}|} c_{|\mathcal{X}|}/t_{|\mathcal{X}|}^2]^\top. \quad (12)$$

Using the multivariate delta method, we obtain:

$$\begin{aligned} \sqrt{n} \sum_{i \in [|\mathcal{X}|]} \xi_i \left( \frac{\hat{\mathcal{C}}_i}{\hat{\mathcal{T}}_i} - \frac{\mathcal{C}_i}{\mathcal{T}_i} \right) \\ = \sqrt{n} \left( g \left( \hat{\mathcal{C}}_1, \hat{\mathcal{T}}_1, \dots, \hat{\mathcal{C}}_{|\mathcal{X}|}, \hat{\mathcal{T}}_{|\mathcal{X}|} \right) - g \left( \mathcal{C}_1, \mathcal{T}_1, \dots, \mathcal{C}_{|\mathcal{X}|}, \mathcal{T}_{|\mathcal{X}|} \right) \right) \\ \xrightarrow{D} \mathcal{N} \left( 0, \nabla g^\top|_{(\mathcal{C}_1, \mathcal{T}_1, \dots, \mathcal{C}_{|\mathcal{X}|}, \mathcal{T}_{|\mathcal{X}|})} \begin{bmatrix} \mathbb{V}[c_1] & & & & \\ & \mathbb{V}[t_1] & & & \\ & & \ddots & & \\ & & & \mathbb{V}[c_{|\mathcal{X}|}] & \\ & & & & \mathbb{V}[t_{|\mathcal{X}|}] \end{bmatrix} \nabla g|_{(\mathcal{C}_1, \mathcal{T}_1, \dots, \mathcal{C}_{|\mathcal{X}|}, \mathcal{T}_{|\mathcal{X}|})} \right) \\ = \mathcal{N} \left( 0, \sum_{i \in [|\mathcal{X}|]} \xi_i^2 \left( \frac{1}{\mathcal{T}_i^2} \mathbb{V}(c_i) + \frac{\mathcal{C}_i^2}{\mathcal{T}_i^4} \mathbb{V}(t_i) \right) \right) \\ = \mathcal{N} \left( 0, \frac{1}{a^2} \sum_{i \in [|\mathcal{X}|]} \xi_i^2 \left( \frac{a^2}{\mathcal{T}_i^2} \mathbb{V}(c_i) + \frac{a^2 \mathcal{C}_i^2}{\mathcal{T}_i^4} \mathbb{V}(t_i) \right) \right) \end{aligned} \quad (13)$$

<sup>3</sup>Equation (11) implies that for any query  $x_i$ , the human choice  $c_i$  and decision time  $t_i$  are uncorrelated. Moreover, they are independent, as discussed by Drugowitsch [22, the discussion above eq. (7)] and Baldassi et al. [16, proposition 3].

By applying the identities outlined in appendix C.1, we can establish the following identity:

$$\forall i \in [|\mathcal{X}|]: \frac{a^2}{\mathcal{T}_i^2} \mathbb{V}(c_i) + \frac{a^2 \mathcal{C}_i^2}{\mathcal{T}_i^4} \mathbb{V}(t_i) = \frac{1}{\mathcal{T}_i}. \quad (14)$$

Substituting this identity into eq. (13), we obtain:

$$\sqrt{n} \sum_{i \in [|\mathcal{X}|]} \xi_i \left( \frac{\widehat{\mathcal{C}}_i}{\widehat{\mathcal{T}}_i} - \frac{\mathcal{C}_i}{\mathcal{T}_i} \right) \xrightarrow{D} \mathcal{N} \left( 0, \frac{1}{a^2} \sum_{i \in [|\mathcal{X}|]} \xi_i^2 \frac{1}{\mathcal{T}_i} \right). \quad (15)$$

Finally, the asymptotic variance can be upper bounded as follows:

$$\begin{aligned} & \frac{1}{a^2} \sum_{i \in [|\mathcal{X}|]} \xi_i^2 \frac{1}{\mathcal{T}_i} \\ & \leq \frac{1}{a^2} \frac{1}{\min_{i \in [|\mathcal{X}|]} \mathcal{T}_i} \sum_{i \in [|\mathcal{X}|]} \xi_i^2 \\ & = \frac{1}{a^2} \frac{1}{\min_{i \in [|\mathcal{X}|]} \mathcal{T}_i} \cdot \left( \sum_{x \in \mathcal{X}} y^\top \left( \sum_{x' \in \mathcal{X}} n x' x'^\top \right)^{-1} n^2 x x^\top \left( \sum_{x' \in \mathcal{X}} n x' x'^\top \right)^{-1} y \right) \quad (16) \\ & = \frac{1}{a^2} \frac{1}{\min_{i \in [|\mathcal{X}|]} \mathcal{T}_i} \cdot y^\top \left( \sum_{x' \in \mathcal{X}} x' x'^\top \right)^{-1} y \\ & = \frac{1}{a^2} y^\top \left( \sum_{x' \in \mathcal{X}} \left[ \min_{i \in [|\mathcal{X}|]} \mathcal{T}_i \right] x' x'^\top \right)^{-1} y \\ & \equiv \frac{1}{a^2} \|y\|_{\left( \sum_{x' \in \mathcal{X}} \left[ \min_{i \in [|\mathcal{X}|]} \mathcal{T}_i \right] x' x'^\top \right)^{-1}}^2. \end{aligned}$$

□

### C.3 Non-asymptotic concentration of the two estimators for estimating the utility difference $u_x$ given a query $x$

#### C.3.1 The choice-decision-time estimator

Section 3.3 focuses on the problem of estimating the utility difference for a single query. Given a query  $x \in \mathcal{X}$ , the objective is to estimate the utility difference  $u_x := x^\top \theta^*$  using an i.i.d. dataset, denoted by  $\{(c_{x,s_{x,i}}, t_{x,s_{x,i}})\}_{i \in [n_x]}$ .

We begin by applying the choice-decision-time estimator from eq. (3), which is derived by solving the following least squares problem:

$$\hat{\theta}_{\text{CH,DT}} = \arg \min_{\theta \in \mathbb{R}^d} \sum_{x \in \mathcal{X}_{\text{sample}}} n_x \left( x^\top \theta - \frac{\sum_{i \in [n_x]} c_{x,s_{x,i}}}{\sum_{i \in [n_x]} t_{x,s_{x,i}}} \right)^2.$$

Similarly, the utility difference for a single query is estimated as the solution to the following least squares problem, yielding the estimate:

$$\hat{u}_{x,\text{CH,DT}} = \arg \min_{u \in \mathbb{R}} \left( u - \frac{\sum_{i \in [n_x]} c_{x,s_{x,i}}}{\sum_{i \in [n_x]} t_{x,s_{x,i}}} \right)^2 = \frac{\sum_{i \in [n_x]} c_{x,s_{x,i}}}{\sum_{i \in [n_x]} t_{x,s_{x,i}}} \quad (\text{restating eq. (5)}).$$

The resulting estimate,  $\hat{u}_{x,\text{CH,DT}}$ , approximates  $u_x/a$  rather than  $u_x$ . However, since the ranking of arm utilities is preserved between  $u_x/a$  and  $u_x$ , estimating  $u_x/a$  is sufficient for the purpose of best-arm identification.

For the case where the utility difference  $u_x \neq 0$ , the non-asymptotic concentration inequality for this estimator is presented in theorem 3.3. To prove this, we first introduce lemma C.1, which demonstrates that for any given query  $x$ , the decision time is a sub-exponential random variable.

To simplify notation, we define:

$$\hat{c}_x = \frac{1}{n_x} \sum_{i=1}^{n_x} c_{x,s_{x,i}}, \quad c_x = \mathbb{E}[c_x], \quad \hat{\mathcal{T}}_x = \frac{1}{n_x} \sum_{i=1}^{n_x} t_{x,s_{x,i}}, \quad \mathcal{T}_x = \mathbb{E}[t_x], \quad \hat{u}_{x,\text{CH,DT}} = \frac{\hat{c}_x}{\hat{\mathcal{T}}_x}. \quad (17)$$

**Lemma C.1.** *If  $u_x \neq 0$ , then  $(t_x - \mathcal{T}_x)$  is sub-exponential  $\text{SE}(\nu_x^2, \alpha_x)$ , where  $\nu_x = \sqrt{2}a/|u_x|$  and  $\alpha_x = 2/u_x^2$ .*

*Proof.* For simplicity, we will omit the subscript  $x$  throughout the proof and assume, without loss of generality, that  $u > 0$ .

Our objective is to establish the following inequality, which holds for all  $s \in (-u^2/2, u^2/2)$ :

$$\mathbb{E}(\exp(s(t - \mathcal{T}))) \leq \exp\left(\frac{2a^2/u^2}{2}s^2\right). \quad (18)$$

This implies that  $(t - \mathcal{T})$  is sub-exponential  $\text{SE}(\nu^2, \alpha)$ , as defined by Wainwright [69, Definition 2.7].

**Step 1: Transform eq. (18) into a more manageable inequality (eq. (24)).**

Using Cox [18, eq. (128)], with  $\Delta := u^2 - 2s$ ,  $\theta_1 := -u - \sqrt{\Delta}$  and  $\theta_2 := -u + \sqrt{\Delta}$ , we have<sup>4</sup>:

$$\begin{aligned}
\mathbb{E}(\exp(st)) &= \frac{\exp(a\theta_1) - \exp(2a\theta_2 + a\theta_1)}{\exp(2a\theta_1) - \exp(2a\theta_2)} - \frac{\exp(a\theta_2) - \exp(2a\theta_1 + a\theta_2)}{\exp(2a\theta_1) - \exp(2a\theta_2)} \\
&= \frac{\exp(a\theta_1)[1 + \exp(a\theta_1 + a\theta_2)]}{\exp(2a\theta_1) - \exp(2a\theta_2)} - \frac{\exp(a\theta_2)[1 + \exp(a\theta_2 + a\theta_1)]}{\exp(2a\theta_1) - \exp(2a\theta_2)} \\
&= \frac{[\exp(a\theta_1) - \exp(a\theta_2)][1 + \exp(a\theta_2 + a\theta_1)]}{\exp(2a\theta_1) - \exp(2a\theta_2)} \\
&= \frac{1 + \exp(a\theta_2 + a\theta_1)}{\exp(a\theta_1) + \exp(a\theta_2)} \\
&= \frac{\exp(-au) + \exp(au)}{\exp(-a\sqrt{\Delta}) + \exp(a\sqrt{\Delta})} \\
&=: \frac{N}{D(s)}.
\end{aligned} \tag{19}$$

In the last line, we define  $N = 2 \cosh(au)$  and  $D(s) = 2 \cosh(a\sqrt{\Delta})$ . Thus, we arrive at:

$$\mathbb{E}(\exp(s \cdot (t - \mathcal{T}))) = \frac{N}{D(s)} \cdot \frac{1}{\exp(s \cdot \mathcal{T})} = \frac{N}{\exp(sa \tanh(au)/u) D(s)}. \tag{20}$$

To prove the original inequality in eq. (18), it is now sufficient to show:

$$D(s) \cdot \exp\left(\frac{a}{u} \tanh(au)s + \frac{a^2}{u^2}s^2\right) \geq N. \tag{21}$$

For  $s = 0$ , the inequality holds trivially, as:

$$D(0) \cdot 1 = 2 \cosh(au) = N. \tag{22}$$

For  $s \neq 0$ , taking the derivative of the left-hand side of eq. (21) yields:

$$\begin{aligned}
&\frac{d}{ds} \left( D(s) \cdot \exp\left(\frac{a}{u} \tanh(au)s + \frac{a^2}{u^2}s^2\right) \right) \\
&= \exp\left(\frac{a}{u} \tanh(au)s + \frac{a^2}{u^2}s^2\right) \cdot \left( -\frac{2a}{\sqrt{\Delta}} \sinh(a\sqrt{\Delta}) + 2 \cosh(a\sqrt{\Delta}) \cdot \left( \frac{a}{u} \tanh(au) + 2\frac{a^2}{u^2}s \right) \right) \\
&= 2 \exp\left(\frac{a}{u} \tanh(au)s + \frac{a^2}{u^2}s^2\right) \cosh(a\sqrt{\Delta}) \cdot \left( -\frac{a}{\sqrt{\Delta}} \tanh(a\sqrt{\Delta}) + \frac{a}{u} \tanh(au) + 2\frac{a^2}{u^2}s \right).
\end{aligned} \tag{23}$$

In step 2, we will prove the following inequality:

$$-\frac{a}{\sqrt{\Delta}} \tanh(a\sqrt{\Delta}) + \frac{a}{u} \tanh(au) + 2\frac{a^2}{u^2}s \begin{cases} \geq 0, & \forall s \geq 0, \\ < 0, & \forall s < 0, \end{cases} \tag{24}$$

Equation (24) implies that  $D(s) \cdot \exp\left(\frac{a}{u} \tanh(au)s + \frac{a^2}{u^2}s^2\right) \geq N$ , which finishes the proof.

**Step 2. Prove eq. (24).**

---

<sup>4</sup>In Cox [18, eq. (128)], setting  $a = 2a$  and  $x_0 = a$  leads to the desired result.

For  $s \geq 0$ , the following holds:

$$\begin{aligned}
& -\frac{a}{\sqrt{\Delta}} \tanh(a\sqrt{\Delta}) + \frac{a}{u} \tanh(au) + 2\frac{a^2}{u^2}s \\
& \stackrel{(i)}{\geq} a \tanh(a\sqrt{\Delta}) \left( \frac{1}{u} - \frac{1}{\sqrt{\Delta}} \right) + 2\frac{a^2}{u^2}s \\
& = a \tanh(a\sqrt{\Delta}) \frac{-2s}{u\sqrt{\Delta}(\sqrt{\Delta}+u)} + 2\frac{a^2}{u^2}s \\
& = -2s \cdot \frac{a^2}{u(\sqrt{\Delta}+u)} \cdot \frac{\tanh(a\sqrt{\Delta})}{a\sqrt{\Delta}} + 2\frac{a^2}{u^2}s \\
& \stackrel{(ii)}{\geq} -2s\frac{a^2}{u^2} \cdot 1 + 2\frac{a^2}{u^2}s \\
& = 0.
\end{aligned} \tag{25}$$

Here, (i) follows from  $\tanh(au) \geq \tanh(a\sqrt{\Delta}) = \tanh(a\sqrt{u^2 - 2s})$  and (ii) follows from  $\tanh(x)/x \leq 1$ .

For  $s < 0$ , the following holds:

$$\begin{aligned}
& -\frac{a}{\sqrt{\Delta}} \tanh(a\sqrt{\Delta}) + \frac{a}{u} \tanh(au) + 2\frac{a^2}{u^2}s \\
& \stackrel{(i)}{\leq} a \tanh(a\sqrt{\Delta}) \left( \frac{1}{u} - \frac{1}{\sqrt{\Delta}} \right) + 2\frac{a^2}{u^2}s \\
& = -2s \cdot \frac{a^2}{u(\sqrt{\Delta}+u)} \cdot \frac{\tanh(a\sqrt{\Delta})}{a\sqrt{\Delta}} + 2\frac{a^2}{u^2}s \\
& \stackrel{(ii)}{\leq} -2s\frac{a^2}{u^2} \cdot 1 + 2\frac{a^2}{u^2}s \\
& = 0.
\end{aligned} \tag{26}$$

Here, (i) follows from  $\tanh(au) \leq \tanh(a\sqrt{\Delta}) = \tanh(a\sqrt{u^2 - 2s})$  and (ii) follows from  $\tanh(x)/x \leq 1$ .

By combining both cases, we conclude that the inequality in eq. (24) holds, which completes Step 2 and proves the desired result.  $\square$

Next, we prove theorem 3.3, which provides the non-asymptotic concentration inequality for the estimator from eq. (5), restated as follows:

**Theorem 3.3** (Non-asymptotic concentration of  $\hat{u}_{x,\text{CH,DT}}$ ). *For each query  $x \in \mathcal{X}$  with  $u_x \neq 0$ , given a fixed i.i.d. dataset  $\{(c_{x,s_{x,i}}, t_{x,s_{x,i}})\}_{i \in [n_x]}$ , for any  $\epsilon > 0$  satisfying  $\epsilon \leq \min\{|u_x|/(\sqrt{2}a), (1 + \sqrt{2})a|u_x|/\mathbb{E}[t_x]\}$ , the following holds:*

$$\mathbb{P}\left(\left|\hat{u}_{x,\text{CH,DT}} - \frac{u_x}{a}\right| > \epsilon\right) \leq 4 \exp\left(-\left[m_{\text{CH,DT}}^{\text{non-asym}}(x^\top \theta^*)\right]^2 n_x [\epsilon \cdot a]^2\right),$$

where  $m_{\text{CH,DT}}^{\text{non-asym}}(x^\top \theta^*) := \mathbb{E}[t_x] / [(2 + 2\sqrt{2})a]$ .

*Proof.* For clarity, we will omit the subscripts  $x$  throughout this proof. Based on lemma C.1, we define the constants  $\nu := \sqrt{2}a/|u|$  and  $\alpha := 2/u^2$ .

We begin by introducing  $\epsilon_C := \mathcal{T}/(\sqrt{2} + \sqrt{2}\nu|\mathcal{C}|/\mathcal{T}) \cdot \epsilon$  and  $\epsilon_{\mathcal{T}} := \nu\epsilon_C$ . From the identities provided in appendix C.1, we know that  $\nu|\mathcal{C}|/\mathcal{T} = \sqrt{2}a/|u| \cdot |u|/a = \sqrt{2}$ . This allows us to simplify the constants  $\epsilon_C$  and  $\epsilon_{\mathcal{T}}$  as:

$$\epsilon_C = \frac{\mathcal{T}}{\sqrt{2}(\sqrt{2}+1)}\epsilon \quad \text{and} \quad \epsilon_{\mathcal{T}} = \frac{\nu\mathcal{T}}{\sqrt{2}(\sqrt{2}+1)}\epsilon. \tag{27}$$

For any  $\epsilon$  satisfying the following condition:

$$\epsilon \leq \min \left\{ \frac{1}{\nu}, \frac{\sqrt{2}(1+\sqrt{2})\nu}{\alpha\mathcal{T}} \right\}, \quad (28)$$

we observe that  $\epsilon_{\mathcal{T}} < \min \{ \mathcal{T}(1 - 1/\sqrt{2}), \nu^2/\alpha \}$ . We can now apply lemma C.2 to derive the following:

$$\mathbb{P} \left( \left| \widehat{\mathcal{T}} - \mathcal{T} \right| > \epsilon_{\mathcal{T}} \right) \leq 2 \exp \left( -\frac{n\epsilon_{\mathcal{T}}^2}{2\nu^2} \right). \quad (29)$$

Thus, by combining the results, we conclude:

$$\begin{aligned} \mathbb{P} \left( \left| \frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}} \right| > \epsilon \right) &= \mathbb{P} \left( \left| \frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}} \right| > \sqrt{2} \frac{\epsilon_{\mathcal{C}} + \epsilon_{\mathcal{T}} \cdot |\mathcal{C}|/\mathcal{T}}{\mathcal{T}} \right) \\ &\stackrel{(i)}{\leq} \mathbb{P} \left( \left| \widehat{\mathcal{C}} - \mathcal{C} \right| > \epsilon_{\mathcal{C}} \right) + \mathbb{P} \left( \left| \widehat{\mathcal{T}} - \mathcal{T} \right| > \epsilon_{\mathcal{T}} \right) \\ &\stackrel{(ii)}{\leq} 2 \exp \left( -\frac{n\epsilon_{\mathcal{C}}^2}{2} \right) + 2 \exp \left( -\frac{n\epsilon_{\mathcal{T}}^2}{2\nu^2} \right) \\ &\stackrel{(iii)}{=} 4 \exp \left( -\frac{n\epsilon_{\mathcal{C}}^2}{2} \right) \\ &= 4 \exp \left( -\frac{\mathcal{T}^2}{4(1+\sqrt{2})^2} \cdot n\epsilon^2 \right). \end{aligned} \quad (30)$$

Here, (i) follows from lemma C.3, (ii) uses lemma C.2 and eq. (29), and (iii) follows from eq. (27).  $\square$

### Supporting Details

**Lemma C.2.** *For each query  $x$  with  $u_x \neq 0$ , and constants  $\epsilon_{\mathcal{C}} > 0$  and  $\epsilon_{\mathcal{T}} \in (0, \nu_x^2/\alpha_x]$ , the following inequalities hold:*

$$\mathbb{P} \left( \left| \widehat{\mathcal{C}}_x - \mathcal{C}_x \right| \geq \epsilon_{\mathcal{C}} \right) \leq 2 \exp \left( -\frac{n\epsilon_{\mathcal{C}}^2}{2} \right), \quad \mathbb{P} \left( \left| \widehat{\mathcal{T}}_x - \mathcal{T}_x \right| \geq \epsilon_{\mathcal{T}} \right) \leq 2 \exp \left( -\frac{n\epsilon_{\mathcal{T}}^2}{2\nu_x^2} \right). \quad (31)$$

Here, the constants are  $\nu_x := \sqrt{2}a/|u_x|$  and  $\alpha_x := 2/u_x^2$ .

*Proof.* Since  $c_x \in \{-1, 1\}$ , by applying Hoeffding's inequality [69, proposition 2.5], we obtain:

$$\mathbb{P} \left( \left| \widehat{\mathcal{C}}_x - \mathcal{C}_x \right| \geq \epsilon_{\mathcal{C}} \right) \leq 2 \exp \left( -\frac{n\epsilon_{\mathcal{C}}^2}{2} \right). \quad (32)$$

From lemma C.1, we know that  $t_x$  is sub-exponential  $SE(\nu_x^2, \alpha_x)$ . By applying Wainwright [69, proposition 2.9 and eq. (2.18)], we obtain:

$$\mathbb{P} \left( \left| \widehat{\mathcal{T}}_x - \mathcal{T}_x \right| \geq \epsilon_{\mathcal{T}} \right) \leq 2 \exp \left( -\frac{n\epsilon_{\mathcal{T}}^2}{2\nu_x^2} \right), \quad \forall \epsilon_{\mathcal{T}} \in (0, \nu_x^2/\alpha_x]. \quad (33)$$

$\square$

**Lemma C.3.** *Consider constants  $\mathcal{C} \in \mathbb{R}$ ,  $\mathcal{T} > 0$ ,  $\epsilon_{\mathcal{C}} > 0$ , and  $\epsilon_{\mathcal{T}} \in (0, (1 - 1/\sqrt{2})\mathcal{T})$ . For any  $\widehat{\mathcal{C}} \in [\mathcal{C} - \epsilon_{\mathcal{C}}, \mathcal{C} + \epsilon_{\mathcal{C}}]$  and  $\widehat{\mathcal{T}} \in [\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}]$ , the following inequality holds*

$$\left| \frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}} \right| \leq \sqrt{2} \frac{\epsilon_{\mathcal{C}} + \epsilon_{\mathcal{T}} \cdot |\mathcal{C}|/\mathcal{T}}{\mathcal{T}}. \quad (34)$$

*Proof.* The maximum value of  $|\widehat{\mathcal{C}}/\widehat{\mathcal{T}} - \mathcal{C}/\mathcal{T}|$  is attained at the extremum of  $\widehat{\mathcal{C}}/\widehat{\mathcal{T}}$ . Since  $\widehat{\mathcal{C}}/\widehat{\mathcal{T}}$  is linear in  $\widehat{\mathcal{C}}$ , the extremum of  $\widehat{\mathcal{C}}/\widehat{\mathcal{T}}$  is attained at  $C^* \in \{\mathcal{C} - \epsilon_C, \mathcal{C} + \epsilon_C\}$  for any  $\widehat{\mathcal{T}} \in [\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}] > 0$ . Given that  $\widehat{\mathcal{T}} > 0$ , the extremum of  $C^*/\widehat{\mathcal{T}}$  is attained at  $T^* \in \{\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}\}$ . Therefore, the extremum of  $\widehat{\mathcal{C}}/\widehat{\mathcal{T}}$  lies in the set:

$$\max_{\substack{\widehat{\mathcal{C}} \in [\mathcal{C} - \epsilon_C, \mathcal{C} + \epsilon_C] \\ \widehat{\mathcal{T}} \in [\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}]}} \frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} \in \left\{ \frac{\mathcal{C} - \epsilon_C}{\mathcal{T} - \epsilon_{\mathcal{T}}}, \frac{\mathcal{C} - \epsilon_C}{\mathcal{T} + \epsilon_{\mathcal{T}}}, \frac{\mathcal{C} + \epsilon_C}{\mathcal{T} - \epsilon_{\mathcal{T}}}, \frac{\mathcal{C} + \epsilon_C}{\mathcal{T} + \epsilon_{\mathcal{T}}} \right\}. \quad (35)$$

For any combination  $(s_C, s_{\mathcal{T}}) \in \{\pm 1\} \times \{\pm 1\}$ , and using the function  $\epsilon_{\mathcal{T}} \leq (1 - 1/\sqrt{2})\mathcal{T}$ , we have:

$$\left| \frac{\mathcal{C} + s_C \epsilon_C}{\mathcal{T} + s_{\mathcal{T}} \epsilon_{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}} \right| = \left| \frac{s_C \epsilon_C \mathcal{T} - s_{\mathcal{T}} \epsilon_{\mathcal{T}} \mathcal{C}}{\mathcal{T}(\mathcal{T} + s_{\mathcal{T}} \epsilon_{\mathcal{T}})} \right| \leq \frac{\epsilon_C \mathcal{T} + \epsilon_{\mathcal{T}} |\mathcal{C}|}{\mathcal{T}(\mathcal{T} - \epsilon_{\mathcal{T}})} \leq \sqrt{2} \frac{\epsilon_C \mathcal{T} + \epsilon_{\mathcal{T}} |\mathcal{C}|}{\mathcal{T}^2}. \quad (36)$$

By combining these results, we conclude that:

$$\max_{\substack{\widehat{\mathcal{C}} \in [\mathcal{C} - \epsilon_C, \mathcal{C} + \epsilon_C] \\ \widehat{\mathcal{T}} \in [\mathcal{T} - \epsilon_{\mathcal{T}}, \mathcal{T} + \epsilon_{\mathcal{T}}]}} \left| \frac{\widehat{\mathcal{C}}}{\widehat{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}} \right| = \max_{(s_C, s_{\mathcal{T}}) \in \{\pm 1\} \times \{\pm 1\}} \left| \frac{\mathcal{C} + s_C \epsilon_C}{\mathcal{T} + s_{\mathcal{T}} \epsilon_{\mathcal{T}}} - \frac{\mathcal{C}}{\mathcal{T}} \right| \leq \sqrt{2} \frac{\epsilon_C + \epsilon_{\mathcal{T}} |\mathcal{C}| / \mathcal{T}}{\mathcal{T}}.$$

□

### C.3.2 The choice-only estimator

We now apply the logistic-regression-based choice-only estimator from eq. (4) to estimate the utility difference for a single query. Recall that for each query  $x \in \mathcal{X}$ , the human choice  $c_x \in \{-1, 1\}$ . We define the binary-encoded choice as  $e_x := (c_x + 1) / 2 \in \{0, 1\}$ . We reformulate the MLE in eq. (4) into a utility difference estimation problem for a single query, leading to the following optimization problem:

$$\begin{aligned} \widehat{u}_{x, \text{CH}} &= \arg \max_{u \in \mathbb{R}} \sum_{i \in [n_x]} \log \mu(c_{x, s_{x,i}} u) \\ &= \arg \max_{u \in \mathbb{R}} \sum_{i \in [n_x]} \log \left[ (\mu(u))^{e_{x, s_{x,i}}} \cdot (\mu(-u))^{1 - e_{x, s_{x,i}}} \right]. \end{aligned}$$

The first-order optimality condition provides the optimal solution:

$$\widehat{u}_{x, \text{CH}} = \mu^{-1} \left( \frac{1}{n_x} \sum_{i \in [n_x]} e_{x, s_{x,i}} \right) \quad (\text{restating eq. (6)}),$$

where  $\mu^{-1}(p) := \log(p/(1-p))$  is the logit function (also known as the log-odds), defined as the inverse of the function  $\mu(\cdot)$  introduced in eq. (4).

The resulting estimate,  $\widehat{u}_{x, \text{CH}}$ , from eq. (6) gives an estimate of  $2au_x$ , not  $u_x$ . However, since the ranking of arm utilities based on  $2au_x$  is the same as that based on the true  $u_x$ , estimating  $2au_x$  suffices for identifying the best arm.

The non-asymptotic concentration inequality for this estimator is stated in theorem 3.4. This result is directly adapted from Jun et al. [31, theorem 5], by letting  $x_1 = \dots = x_t = 1$  and  $t_{\text{eff}} = d = 1$ .

## D Experiment details

Our empirical experiments (Sec. 5) were conducted on a MacBook Pro (M3 Pro, Nov 2023) with 36 GB of memory.

Our implementation is available via <https://shenlirobot.github.io/pages/NeurIPS24.html>. The code is written in Julia and builds on the implementation by Tirinzoni and Degenne [63], where the transductive and weak-preference designs are solved using the Frank–Wolfe algorithm [24]. Their code is accessible at <https://github.com/AndreaTirinzoni/bandit-elimination>. Simulations and Bayesian inference for the DDM are implemented using the Julia package `SequentialSamplingModels.jl`, available at <https://itsdfish.github.io/SequentialSamplingModels.jl/dev/#SequentialSamplingModels.jl>.

For a query  $x \in \mathcal{X}$ , the estimators from Wagenmakers et al. [67] and Xiang Chiong et al. [73], analyzed in section 3.3 and benchmarked in section 5.2, require calculating  $\mu^{-1}(p) := \log(p/(1-p))$ , where  $\mu^{-1}(\cdot)$  is the logit function and  $p := 1/n_x \cdot \sum_{i=1}^{n_x} (c_{x,s_{x,i}} + 1)/2$  represents the empirical mean of the human binary choices coded as 0 or 1. Since  $p = 0$  or  $p = 1$  makes this calculation undefined, we follow Wagenmakers et al. [67] the discussion below fig. 6] and approximate  $p$  as  $1 - 1/(2n_x)$  when  $p = 1$  and  $1/(2n_x)$  when  $p = 0$ .

### D.1 The “Sphere” Synthetic Problem for Evaluating Estimation Performance in section 5.1

We evaluate estimation performance using the “sphere” synthetic problem, a standard benchmark in linear bandit literature [20, 42, 61]. In this problem, the arm space  $\mathcal{Z} \subset \{z \in \mathbb{R}^5: \|z\|_2 = 1\}$  contains 10 randomly generated arms. To define the true preference vector  $\theta^*$ , we select the two arms  $z$  and  $z'$  that are closest in direction, i.e.,  $(z, z') \in \arg \max_{z, z' \in \mathcal{Z}} z^\top z'$ , and set  $\theta^* = z + 0.01(z' - z)$ . In this way,  $z$  is the best arm. The query space is  $\mathcal{X} := \{z - z': z \in \mathcal{Z}\}$ .

## D.2 Processing the food-risk dataset with choices (-1 or 1) [57]

We accessed the food-risk dataset with choices (-1 or 1) [57] through Yang and Krajbich [76]’s repository (<https://osf.io/d7s6c/>). This dataset includes the choices and response times of 42 participants, each responding to between 60 and 200 queries. Each query compares two arms, with each arm containing two food items. By selecting an arm, participants had an equal chance of receiving either food item, hence the name “food risk” (or “food-gamble”) task. Additionally, participants’ eye movements were tracked during the experiment. Yang and Krajbich [76] modeled each participant’s choices, response times, and eye movements using the attentional DDM [39], where the drift for each query is a linear combination of the participant’s ratings of the four food items in the query, with the weights adjusting based on their eye movements. The ratings,  $\in \{-10, -9, \dots, 0, \dots, 9, 10\}$ , were collected before the participants interacted with the binary queries.

In our work, for each participant, we define each arm’s feature vector as the participant’s ratings of the two corresponding food items, augmented with second-order polynomials. We fit each participant’s data to a difference-based EZ-diffusion model [8, 67] with a linear utility structure, as introduced in section 2. For each participant, using Bayesian inference with non-informative priors [16], we estimated the preference vector  $\theta^* \in \mathbb{R}^5$ , non-decision time  $t_{\text{nondec}}$ , and barrier  $a$ . Across participants, the barrier  $a$  ranged from 0.715 to 2.467, with a mean of 1.437, and  $t_{\text{nondec}}$  ranged from 0.206 to 1.917 seconds, with a mean of 0.746 seconds. This procedure generated one bandit instance per participant, with a preference vector  $\theta^* \in \mathbb{R}^5$ , an arm space  $\mathcal{Z} \subset \mathbb{R}^5$  where  $|\mathcal{Z}| \in [31, 95]$ , and a query space  $\mathcal{X} := \{z - z': z \in \mathcal{Z}\}$ . Then, we used the fitted models to simulate human feedback for bandit experiments.

For each bandit instance, we benchmarked six GSE variations (introduced in section 5.2):  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,RT}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH}})$ ,  $(\lambda_{\text{weak}}, \hat{\theta}_{\text{CH}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,logit}})$ , and  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT,logit}})$ . For each GSE variation, we ran 300 repeated simulations under different random seeds, with human choices and response times sampled from the dEZDM with the identified parameters. Since each bandit instance contains a different number of arms, rather than tuning the elimination parameter  $\eta$  in algorithm 1 for each instance, we set  $\eta = 2$ , following the convention in previous bandit research, e.g., Azizi et al. [3] section 3]. We manually tuned the buffer size  $B_{\text{buff}}$  in algorithm 1 to 20, 30, or 50 seconds based on empirical performance, ensuring the budget was not exceeded in each phase. The full results are shown in fig. 5, with selected results highlighted in fig. 4a.

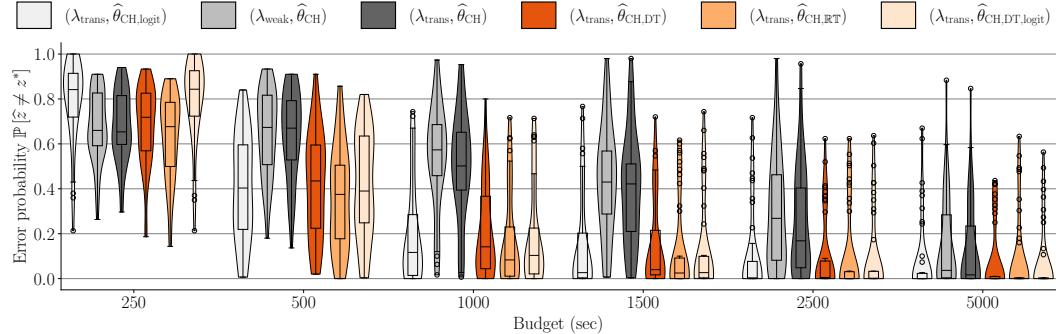


Figure 5: A violin plot overlaid with a box plot showing the best-arm identification error probability,  $\mathbb{P}[\hat{z} \neq z^*]$ , as a function of budget for each GSE variation, simulated using the food-risk dataset with choices (-1 or 1) [57], as described in appendix D.2. The box plots follow the convention of the `matplotlib` Python package. For each GSE variation and budget, the horizontal line in the middle of the box represents the median of the error probabilities across all bandit instances. Each error probability is averaged over 300 repeated simulations under different random seeds. The box’s upper and lower borders represent the third and first quartiles, respectively, with whiskers extending to the farthest points within  $1.5 \times$  the interquartile range. Flier points indicate outliers beyond the whiskers.

### D.3 Processing the snack dataset with choices (yes or no) [16]

We accessed the snack dataset with choices (yes or no) [16] through the supplementary material provided by Alós-Ferrer et al. [2] at <https://www.journals.uchicago.edu/doi/abs/10.1086/713732>. This dataset consists of training and testing data. The training data was collected from a “YN” task, where 31 participants provided binary feedback (“Yes” or “No”) and response times for queries comparing each of the 17 snack items to a fixed reference snack, with each query repeated 10 times. The reference snack, assigned a utility of 0, remained fixed throughout the experiment. The testing data was collected using a two-alternative forced-choice task, where participants provided binary choices and response times for queries comparing two snack items, with each query repeated once. Clithero [16] fit a difference-based EZ-diffusion model [8, 67] to the training data using Bayesian inference with non-informative priors, without imposing a linear utility structure, and tested the model using the testing data.

In our work, we fit each participant’s training data to a difference-based EZ-diffusion model with a linear utility structure, as described in section 2 and used the fitted model to simulate human feedback for bandit experiments. We preprocessed the data by removing outliers, following Clithero [16] footnote 22], excluding trials with response times below 200 ms or greater than five standard deviations above the mean. After cleaning, the number of trials per participant ranged from 167 to 170. Since the dataset does not provide feature vectors for the 17 non-reference snack items, we used one-hot encoding to represent each snack item as a feature vector in  $\mathbb{R}^{17}$ . This allowed us to construct a bandit instance for each participant with a preference vector  $\theta^* \in \mathbb{R}^{17}$ , an arm space  $\mathcal{Z} \subset \mathbb{R}^{17}$  with  $|\mathcal{Z}| = 17$ , and a query space  $\mathcal{X} := \{z - \mathbf{0} : z \in \mathcal{Z}\}$  to represent comparisons with the reference snack. We applied Bayesian inference with non-informative priors [16] to estimate each participant’s preference vector  $\theta^*$ , non-decision time  $t_{\text{nondec}}$ , and barrier  $a$ . Across participants, the barrier  $a$  ranged from 0.759 to 1.399, with a mean of 1.1, and  $t_{\text{nondec}}$  ranged from 0.139 to 0.485 seconds, with a mean of 0.367 seconds.

For each of the six GSE variations (introduced in section 5.2):  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,RT}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH}})$ ,  $(\lambda_{\text{weak}}, \hat{\theta}_{\text{CH}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,logit}})$ , and  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT,logit}})$ , we tuned the elimination parameter  $\eta$  in algorithm 1 using the following procedure: We considered  $\eta \in \{2, 3, 4, 5, 6, 7, 8, 9\}$ , resulting in the number of phases  $:= \lceil \log_\eta |\mathcal{Z}| \rceil = \lceil \log_\eta (17) \rceil$  (line 4 of algorithm 1) being  $\{5, 3, 3, 2, 2, 2, 2, 2\}$ , respectively. We excluded  $\eta > \lceil 17/2 \rceil = 9$ , as those cases also result in 2 phases, the same as  $\eta \in \{5, 6, 7, 8, 9\}$ . Then, for each  $\eta$ , for each of the 31 bandit instances, and for each budget  $\in \{50, 75, 100, 125, 150, 200, 250, 300\}$  seconds, we ran 50 repeated simulations per GSE variation under different random seeds, sampling human feedback from the fitted dEZDM. We then aggregated the results into a single best-arm identification error probability for each GSE variation,  $\eta$ , bandit instance, and budget. These error probabilities were compiled into violin and box plots, as shown in fig. 6.

For each GSE variation, we selected the  $\eta$  that minimized the median error probability, as shown in the box plots in fig. 6. If multiple  $\eta$  values yielded the same median, we used the third quartile, and if necessary, the first quartile, to break ties. Based on this approach, we selected:  $\eta = 6$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$ ,  $\eta = 6$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,RT}})$ ,  $\eta = 9$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH}})$ ,  $\eta = 9$  for  $(\lambda_{\text{weak}}, \hat{\theta}_{\text{CH}})$ ,  $\eta = 9$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,logit}})$ , and  $\eta = 5$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT,logit}})$ .

After tuning  $\eta$ , we manually set the buffer size  $B_{\text{buff}}$  in algorithm 1 to 10 seconds based on empirical results, ensuring the budget was not exceeded in any phase. We then benchmarked each GSE variation on all 31 bandit instances using its own manually tuned  $\eta$  and  $B_{\text{buff}}$ . Each variation was evaluated over 300 repeated simulations with different random seeds, where human choices and response times were sampled from the dEZDM with the identified parameters. The full results are shown in fig. 7, with selected results presented in fig. 4b.

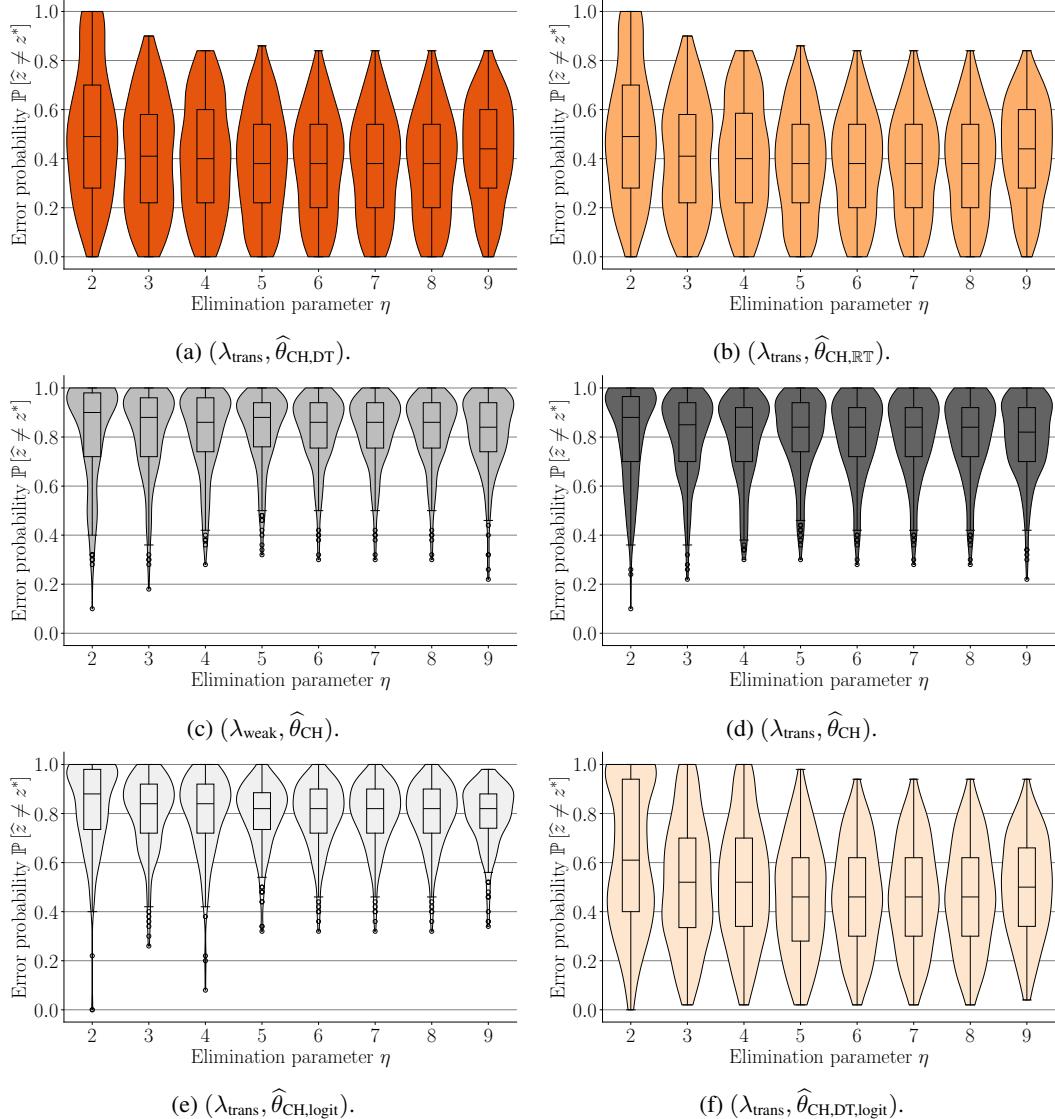


Figure 6: Violin plots overlaid with box plots, used for tuning the elimination parameter  $\eta$  in algorithm I for each GSE variation, simulated based on the snack dataset with choices (yes or no) [16], as discussed in appendix D.3. Each plot shows the best-arm identification error probability,  $\mathbb{P}[\hat{z} \neq z^*]$ , as a function of  $\eta$ . The box plots follow the convention of the matplotlib Python package. The horizontal line in each box represents the median of the error probabilities across all bandit instances and budgets. Each error probability is averaged over 50 repeated simulations under different random seeds. The top and bottom borders of the box represent the third and first quartiles, respectively, while the whiskers extend to the farthest points within  $1.5 \times$  the interquartile range. Flier points are the outliers past the end of the whiskers.

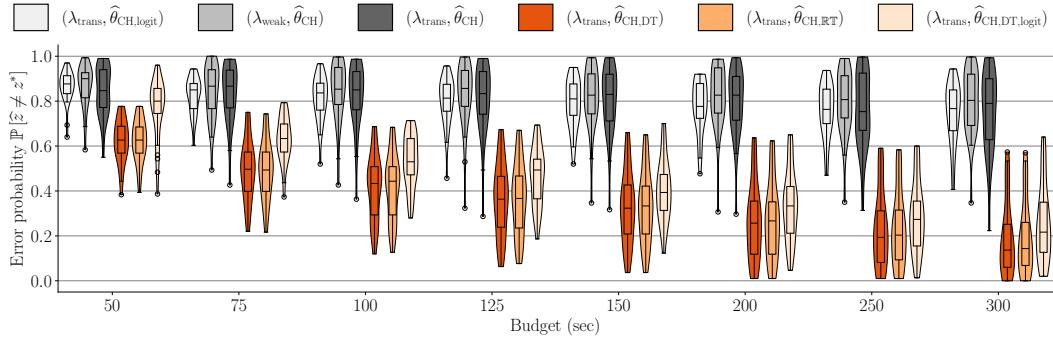


Figure 7: A violin plot overlaid with a box plot showing the best-arm identification error probability,  $\mathbb{P}[\hat{z} \neq z^*]$ , as a function of budget for each GSE variation, simulated using the snack dataset with choices (yes or no) [16], as described in appendix D.3. The box plots follow the convention of the [matplotlib Python package](#). For each GSE variation and budget, the horizontal line in the middle of the box represents the median of the error probabilities across all bandit instances. Each error probability is averaged over 300 repeated simulations under different random seeds. The box's upper and lower borders represent the third and first quartiles, respectively, with whiskers extending to the farthest points within  $1.5 \times$  the interquartile range. Flier points indicate outliers beyond the whiskers.

#### D.4 Processing the snack dataset with choices (-1 or 1) [39]

We accessed the snack dataset with choices (-1 or 1) [39] via Fudenberg et al. [27]’s replication package at <https://www.aeaweb.org/articles?id=10.1257/aer.20150742>. This dataset contains choices and response times from 39 participants, each responding to between 49 and 100 queries comparing two snack items. Participants’ eye movements were tracked during the experiment. Krajbich et al. [39] modeled each participant’s choices, response times, and eye movements using the attentional DDM, where the drift for each query is a linear combination of the participant’s ratings of both snack items in the query, with the weights influenced by their eye movements. The ratings,  $\in \{-10, -9, \dots, 0, \dots, 9, 10\}$ , were collected before participants interacted with the binary queries.

In our work, to avoid creating trivial bandit problems by encoding snack items as 1-dimensional vectors (as done in appendix D.2), we defined the feature vector for each snack item with a participant rating  $r_z \in \{-10, -9, \dots, 0, \dots, 9, 10\}$  as a one-hot vector in  $\mathbb{R}^{21}$ , where the  $(r_z + 11)$ -th element is 1 and the rest are 0. The preference vector  $\theta^*$  is structured as  $\beta^* \cdot [-10, -9, \dots, 0, \dots, 9, 10]^\top \in \mathbb{R}^{21}$ , where  $\beta^*$  is participant-specific and unknown to the learner. This ensures that, for each arm  $z$ , the participant’s utility is  $u_z := z^\top \theta^* = r_z \beta^*$ . In this way, each participant’s data generated a bandit instance with a preference vector  $\theta^* \in \mathbb{R}^{21}$ , a set of arms  $\mathcal{Z} \subset \mathbb{R}^{21}$  with  $|\mathcal{Z}| = 21$ , and a query space  $\mathcal{X} := \{z - z': z \in \mathcal{Z}\}$ .

We fit each participant’s data to a difference-based EZ-diffusion model [8, 67] using the linear utility structure described above. For each participant, using Bayesian inference with non-informative priors [16], we estimated the preference vector  $\theta^*$  (or equivalently, the parameter  $\beta^*$ ), non-decision time  $t_{\text{nondec}}$ , and barrier  $a$ . Across participants, the barrier  $a$  ranged from 0.75 to 2.192 with a mean of 1.335, and  $t_{\text{nondec}}$  ranged from 0.387 to 1.22 seconds with a mean of 0.641 seconds. We then used these fitted models to simulate human feedback for bandit experiments, assuming the learner did not know the underlying structure  $\theta^* = \beta^* \cdot [-10, -9, \dots, 0, \dots, 9, 10]^\top$ .

For each of the following GSE variations (introduced in section 5.2):  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,RT}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH}})$ ,  $(\lambda_{\text{weak}}, \hat{\theta}_{\text{CH}})$ ,  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,logit}})$ , and  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT,logit}})$ , we tuned the elimination parameter  $\eta$  in algorithm 1 using the following procedure: We considered  $\eta \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ , which resulted in the number of phases :=  $\lceil \log_\eta |\mathcal{Z}| \rceil = \lceil \log_\eta (17) \rceil$  (line 4 of algorithm 1) being  $\{5, 3, 3, 2, 2, 2, 2, 2, 2, 2\}$ , respectively. We excluded cases where  $\eta > \lceil 21/2 \rceil = 11$ , as these result in 2 phases, identical to when  $\eta \in \{5, 6, 7, 8, 9, 10, 11\}$ . Then, for each  $\eta$ , for each of the 39 bandit instances, and for each budget  $\in \{150, 200, 250, 300, 350, 400, 450, 500\}$  seconds, we ran 50 repeated simulations per GSE variation under different random seeds, sampling human feedback from the fitted dEZDM. We then aggregated the results into a single best-arm identification error probability for each GSE variation,  $\eta$ , bandit instance, and budget. These error probabilities were compiled into violin and box plots, as shown in fig. 8.

For each GSE variation, we selected the  $\eta$  that minimized the median error probability, as shown in the box plots in fig. 8. If multiple  $\eta$  values yielded the same median, we used the third quartile, and if necessary, the first quartile, to break ties. Based on this approach, we selected:  $\eta = 4$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT}})$ ,  $\eta = 4$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,RT}})$ ,  $\eta = 4$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH}})$ ,  $\eta = 2$  for  $(\lambda_{\text{weak}}, \hat{\theta}_{\text{CH}})$ ,  $\eta = 5$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,logit}})$ , and  $\eta = 5$  for  $(\lambda_{\text{trans}}, \hat{\theta}_{\text{CH,DT,logit}})$ .

After tuning  $\eta$ , we manually set the buffer size  $B_{\text{buff}}$  in algorithm 1 to 20 seconds based on empirical results, ensuring the budget was not exceeded in any phase. We then benchmarked each GSE variation on all 39 bandit instances using its own manually tuned  $\eta$ . Each variation was evaluated over 300 repeated simulations with different random seeds, where human choices and response times were sampled from the dEZDM with the identified parameters. The full results are shown in fig. 9, with selected results presented in fig. 4c.

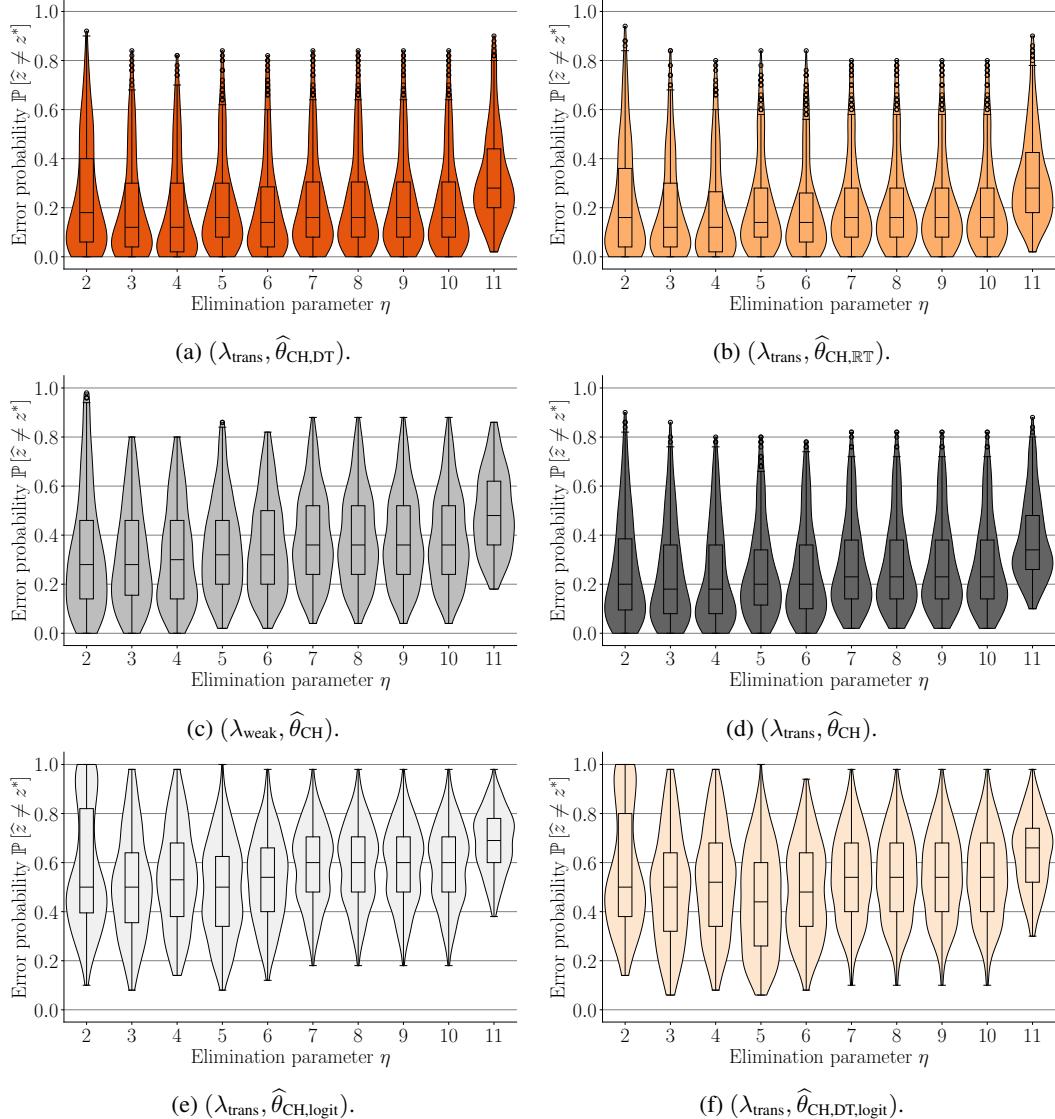


Figure 8: Violin plots overlaid with box plots, used for tuning the elimination parameter  $\eta$  in algorithm I for each GSE variation, simulated based on the snack dataset with choices (-1 or 1) [39], as discussed in appendix D.4. Each plot shows the best-arm identification error probability,  $\mathbb{P}[\hat{z} \neq z^*]$ , as a function of  $\eta$ . The box plots follow the convention of the `matplotlib` Python package. The horizontal line in each box represents the median of the error probabilities across all bandit instances and budgets. Each error probability is averaged over 50 repeated simulations under different random seeds. The top and bottom borders of the box represent the third and first quartiles, respectively, while the whiskers extend to the farthest points within  $1.5 \times$  the interquartile range. Flier points are the outliers past the end of the whiskers.

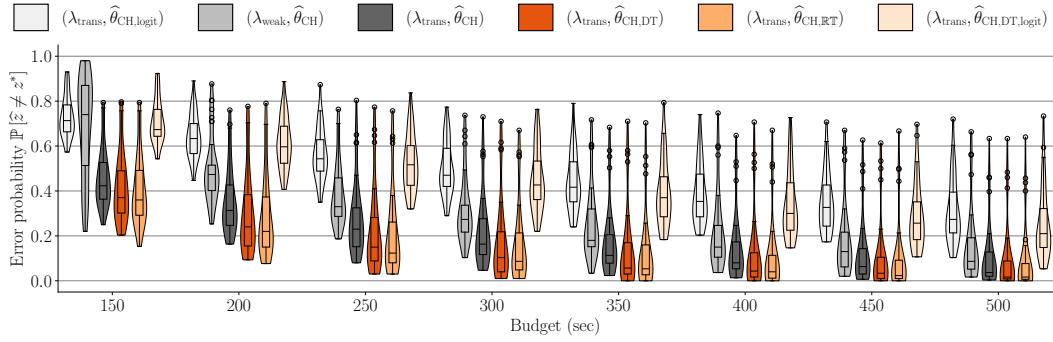


Figure 9: A violin plot overlaid with a box plot showing the best-arm identification error probability,  $\mathbb{P}[\hat{z} \neq z^*]$ , as a function of budget for each GSE variation, simulated using the snack dataset with choices (-1 or 1) [39], as described in appendix D.4. The box plots follow the convention of the [matplotlib Python package](#). For each GSE variation and budget, the horizontal line in the middle of the box represents the median of the error probabilities across all bandit instances. Each error probability is averaged over 300 repeated simulations under different random seeds. The box's upper and lower borders represent the third and first quartiles, respectively, with whiskers extending to the farthest points within  $1.5 \times$  the interquartile range. Flier points indicate outliers beyond the whiskers.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction clearly state the key contribution, that incorporating response times in preference learning, and align with the paper's scope and findings.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitations are discussed in section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions are stated within theorems, and full proofs are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The full algorithm is detailed in algorithm I, and implementation details, including dataset conversions to bandit instances, are provided in section 5 and appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release code with instructions to reproduce the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Implementation details, hyperparameter tuning, and procedures for data processing and conversion to bandit instances are provided in appendix D

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Key empirical results (fig. 4) use violin and box plots to visualize variability, and we clarify sources of randomness in the text.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: Details on compute resources are provided in appendix [D](#).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: We reviewed the NeurIPS ethics guidelines and adhered to them throughout the research process.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: Societal impacts are discussed in appendix [A](#).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release data or models prone to misuse but focuses on estimation methods for preference learning. Potential societal impacts are discussed in appendix A.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite original papers and sources for code and datasets (section 5 and appendix D).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We will release code with detailed documentation and an appropriate license.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: This work does not involve crowdsourcing or human-subject research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing or human-subject research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.