Enhancing Preference-based Linear Bandits via Human Response Time

Shen Li[1 *] **Yuyang Zhang**[2 *] **Zhaolin Ren**[2] **Claire Liang**[1] **Na Li**[2] **Julie A. Shah**[1]
[1]Massachusetts Institute of Technology [2]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.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

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 $\{1, \ldots, n\}$. For a scalar random variable x, the expectation and variance are denoted by $\mathbb{E}[x]$ and $\mathbb{V}[x]$, respectively. The function $\mathrm{sgn}(x)$ denotes the sign of x.

2 Problem setting and preliminaries

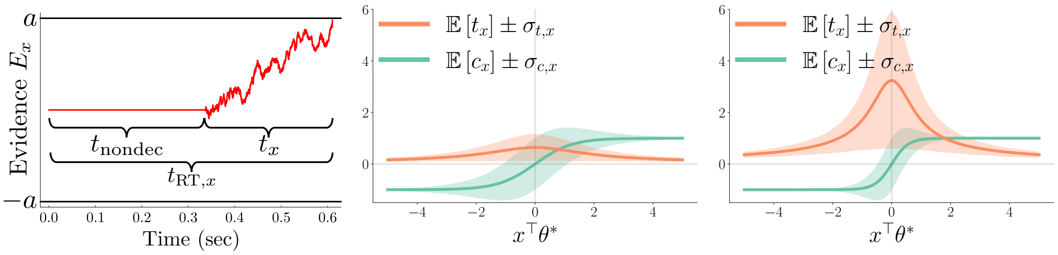
Preference-based bandits with a linear utility function. The learner is given a finite set of options (or "arms"), each represented by a feature vector in $\mathcal{Z} \subset \mathbb{R}^d$, and a finite set of binary queries, where each query is the difference between two arms, denoted by $\mathcal{X} \subset \mathbb{R}^d$. For instance, if the learner can query any pair of arms, the query space is $\mathcal{X} = \{z - z' \colon z, z' \in \mathcal{Z}\}$. In the online retailer example from section 1, the query space is $\mathcal{X} = \{z - z_{\text{skip}} \colon z \in \mathcal{Z}\}$, where z represents purchasing a product and z_{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, $t_x := \min\{\tau > 0\colon E_{x,\tau} \in \{a, -a\}\}$, 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)]:

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

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.



(a)

(b) $a = 0.8$

(c) $a = 1.8$

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_{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_{nondec} is constant across all queries for a given human [16, 76] and further assume that t_{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_{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_{nondec}. Both the human's preference vector θ^* 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 $\widehat{z} \in \mathcal{Z}$. The goal is to recommend the unique best arm $z^* := \arg\max_{z\in\mathcal{Z}} z^\top\theta^*$, minimizing the error probability $\mathbb{P}[\widehat{z} \neq z^*]$.

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 θ^* 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 θ^* than choices alone. Next, in section 3, we introduce a novel estimator that combines decision times and choices to estimate θ^*. 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 θ^* from a fixed dataset, denoted by $\left\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\right\}_{x \in \mathcal{X}_{\text{sample}}, i \in [n_x]}$. 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):

$$\forall x \in \mathcal{X} \colon x^\top \frac{\theta^*}{a} = \frac{\mathbb{E}\left[c_x\right]}{\mathbb{E}\left[t_x\right]}. \tag{2}$$

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 θ^* 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 θ^*/a. Since the ranking of arm utilities based on θ^*/a is identical to that based on θ^*, estimating θ^*/a is sufficient for identifying the best arm. Formally, this estimate of θ^*/a, denoted by $\widehat{\theta}_{\text{CH,DT}}$, is given by:

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

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 $\widehat{\theta}_{\text{CH}}$:

$$\widehat{\theta}_{\text{CH}} \coloneqq \underset{\theta \in \mathbb{R}^d}{\arg\max} \sum_{x \in \mathcal{X}_{\text{sample}}} \sum_{i=1}^{n_x} \log \mu(c_{x,s_{x,i}} \, x^\top \theta), \tag{4}$$

where $\mu(y) \coloneqq 1/[1+\exp(-y)]$ 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 $\widehat{\theta}_{\text{CH,DT}}$). *Given a fixed i.i.d. dataset $\left\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\right\}_{i\in[n]}$ 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 $n \to \infty$, the following holds:*

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

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

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

The proof is provided in appendix C.2. The asymptotic variance upper bound shows that all sampled queries are weighted by a common factor $\min_{x'\in\mathcal{X}_{\text{sample}}} \mathbb{E}\left[t_{x'}\right]$, which is the smallest expected decision time among all the sampled queries in $\mathcal{X}_{\text{sample}}$. 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 $\mathcal{X}_{\text{sample}}$ 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 $\widehat{\theta}_{\text{CH}}$). *Given a fixed i.i.d. dataset $\left\{x, c_{x,s_{x,i}}, t_{x,s_{x,i}}\right\}_{i\in[n]}$ 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 $n \to \infty$, the following holds:*

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

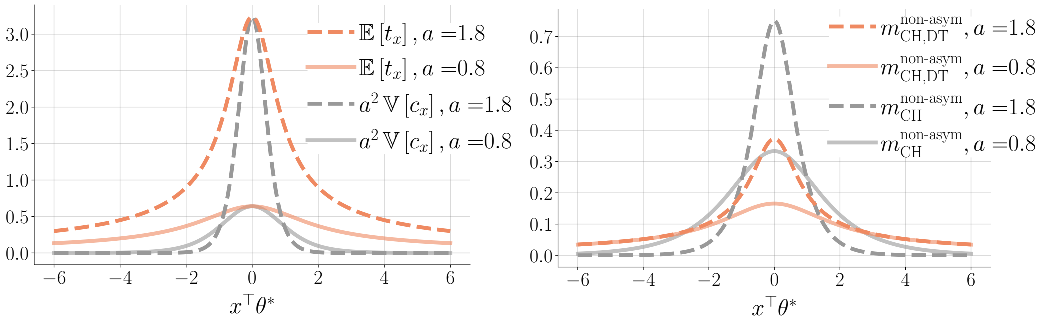
This asymptotic variance shows that each sampled query $x \in \mathcal{X}_{\text{sample}}$ is weighted by its own factor $a^2\,\mathbb{V}\left[c_x\right]$, representing the amount of information the query's choices contribute to utility estimation. A larger weight indicates that the query contributes 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 $a^2\,\mathbb{V}\left[c_x\right]$, 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 1.

For the choice-decision-time estimator (theorem 3.1), queries are weighted by $\min_{x'\in\mathcal{X}_{\text{sample}}} \mathbb{E}\left[t_{x'}\right]$, which depends on both $\mathcal{X}_{\text{sample}}$ and $\mathbb{E}\left[t_x\right]$. To better understand this weight, we first plot $\mathbb{E}\left[t_x\right]$ without the 'min' operator as the orange curves in fig. 2a. Comparing the orange and gray curves shows that $\mathbb{E}\left[t_x\right]$ is generally larger than the choice-only weight, $a^2\,\mathbb{V}\left[c_x\right]$. 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. 1). As a result, more queries exhibit strong preferences, making choices from these queries less informative.



(a) $\mathbb{E}\left[t_x\right]$ and $a^2\,\mathbb{V}\left[c_x\right]$ in asymptotic variances

(b) Weights in non-asymptotic concentration bounds

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 $\mathbb{E}\left[t_x\right]$ and $a^2\,\mathbb{V}\left[c_x\right]$ 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 θ^* 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:

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

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:

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

where $(c_{x,s_{x,i}}+1)/2$ is the binary choice coded as 0 or 1, and $\mu^{-1}(p) := \log\left(p/(1-p)\right)$ is the logit function (inverse of μ 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 $\widehat{u}_{x,\text{CH,DT}}$)**.** *For each query $x \in \mathcal{X}$ with $u_x \neq 0$, given a fixed i.i.d. dataset $\left\{\left(c_{x,s_{x,i}}, t_{x,s_{x,i}}\right)\right\}_{i\in[n_x]}$, for any $\epsilon > 0$ satisfying $\epsilon \leq \min\left\{|u_x|/(\sqrt{2}a), \left(1+\sqrt{2}\right)a|u_x|/\mathbb{E}\left[t_x\right]\right\}$, the following holds:*

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

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

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

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

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

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

The weights $m_{\text{CH,DT}}^{\text{non-asym}}(\cdot)$ and $m_{\text{CH}}^{\text{non-asym}}(\cdot)$ 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, $m_{\text{CH,DT}}^{\text{non-asym}}(\cdot)$) and the choice-only estimator (gray, $m_{\text{CH}}^{\text{non-asym}}(\cdot)$). 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 θ^*, 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 θ^* 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 η 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_{buff}, to prevent overspending in any phase (line 4). In each phase, GSE computes an experimental design λ, a probability distribution over the query space, to guide query sampling. We consider two designs: the transductive design [24], λ_{trans} (line 5), and the weak-preference design [31], λ_{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 θ^* is unknown, the weak-preference design identifies queries with weak preferences based on the previous phase's estimate $\widehat{\theta}_{\text{CH}}$. Then, GSE samples queries based on the design (line 7) and, after exhausting the phase's budget, estimates θ^* using either the choice-decision-time estimator $\widehat{\theta}_{\text{CH,DT}}$ (line 8) or the choice-only estimator $\widehat{\theta}_{\text{CH}}$ (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).

[2] In Jun et al. [31, theorem 5], we let $x_1 = \cdots = x_t = 1$ 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 [3, 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 η, to balance data collection and arm elimination, and the buffer size B_{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. $\lambda_k := \lambda_{\text{trans},k} \leftarrow \arg\min_{\lambda \in \blacktriangle^{|\mathcal{X}|}} \max_{z \neq z' \in \mathcal{Z}_k} \|z - z'\|^2_{\left(\sum_{x \in \mathcal{X}} \lambda_x x x^\top\right)^{-1}}$.
6:     Design 2. $\lambda_k := \lambda_{\text{weak},k} \leftarrow \arg\min_{\lambda \in \blacktriangle^{|\mathcal{X}|}} \max_{z \neq z' \in \mathcal{Z}_k} \|z - z'\|^2_{\left(\sum_{x \in \mathcal{X}} \dot{\mu}(x^\top \widehat{\theta}_{k-1})\, \lambda_x x x^\top\right)^{-1}}$.
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. $\widehat{\theta}_k := \widehat{\theta}_{\text{CH,DT},k} \leftarrow$ apply eq. (3) to all the $J_k$ samples.
9:     Estimate 2. $\widehat{\theta}_k := \widehat{\theta}_{\text{CH},k} \leftarrow$ apply eq. (4) to all the $J_k$ samples.
10:    Update $\mathcal{Z}_{k+1} \leftarrow$ Top-$\left\lceil \frac{|\mathcal{Z}_k|}{\eta} \right\rceil$ arms in $\mathcal{Z}_k$, ranked by the estimated utility $z^\top \widehat{\theta}_k$.
11: **end for**
12: **Output:** the single one $\widehat{z} \in \mathcal{Z}_{K+1}$.
```

5 Empirical results

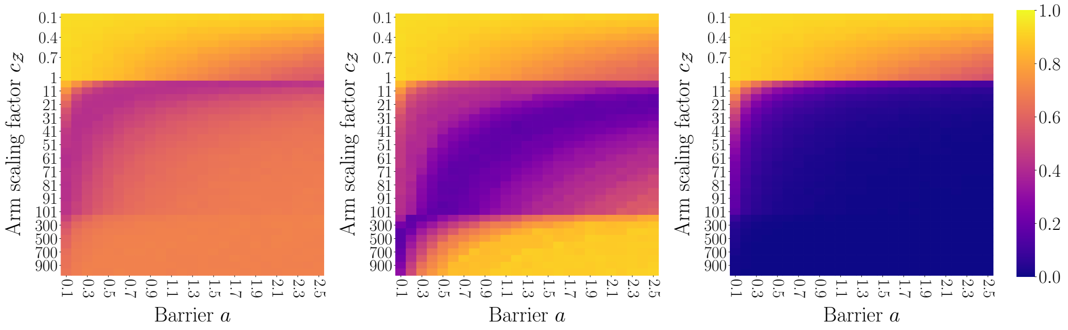
This section empirically compares the GSE variations introduced in section 4: (1) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$: Transductive design with choice-decision-time estimator. (2) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$: Transductive design with choice-only estimator. (3) $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$: 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 $\widehat{\theta}$. Performance is evaluated by $\mathbb{P}[\arg\max_{z \in \mathcal{Z}} z^\top \widehat{\theta} \neq z^*]$, which reflects the best-arm identification goal defined in section 2. To isolate the effect of estimation, we allow λ_{weak} access to the true θ^*, enabling it to perfectly compute the terms $\dot{\mu}(x^\top \theta^*)$ 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



(a) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$ (b) $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$ (c) $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$

Figure 3: Three heatmaps show estimation error probabilities, $\mathbb{P}[\arg\max_{z\in\mathcal{Z}} z^\top\widehat{\theta} \neq z^*]$, for three GSE variations, shown as functions of the arm scaling factor $c_\mathcal{Z}$ and barrier a. Darker colors indicate better estimation. (a) The choice-only estimator $\widehat{\theta}_{\text{CH}}$ with the transductive design λ_{trans} struggles as $c_\mathcal{Z}$ increases (i.e., preferences become stronger), highlighting that choices from queries with strong preferences provide limited information. (b) The weak-preference design λ_{weak} improves (a) by sampling queries with weak preferences but assumes perfect knowledge of θ^* and equal resource consumption across queries. (c) The choice-decision-time estimator $\widehat{\theta}_{\text{CH,DT}}$ with λ_{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 $c_\mathcal{Z}$, 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 $c_\mathcal{Z}$ becomes too large, performance declines because many $\dot{\mu}(x^\top\theta^*)$ 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 θ^* and equal resource consumption across queries. In practice, where θ^* 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

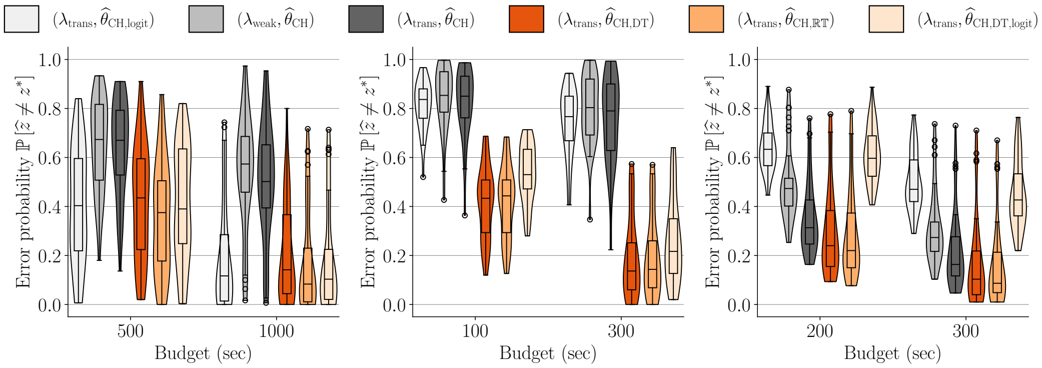
This section compares the bandit performance of six GSE variations. The first three are as previously defined at the beginning of section 5: $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, and $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$.

The 4th GSE variation, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, evaluates the performance of the choice-decision-time estimator when the non-decision time t_{nondec} is unknown. The estimator, $\widehat{\theta}_{\text{CH},\mathbb{RT}}$, 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, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$, 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 $\mu^{-1}(p) := \log\left(p/\left(1-p\right)\right)$. By incorporating our linear utility structure, we obtain the following choice-only estimator $\widehat{\theta}_{\text{CH,logit}}$:

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

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



(a) Food-risk dataset [57]

(b) Snack dataset [16]

(c) Snack dataset [39]

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, $\mathbb{P}[\widehat{z} \neq z^*]$, 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, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$, is based on Xiang Chiong et al. [73, eq. (6)], which states that $x^\top \theta^* = \operatorname{sgn}(c_x) \sqrt{\mathbb{E}[c_x] / \mathbb{E}[t_x] \cdot 0.5\, \mu^{-1} (\mathbb{P}[c_x = 1])}$. 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 $\widehat{\theta}_{\text{CH,DT,logit}}$:

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

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, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$ consistently outperforms $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH}})$, demonstrating the benefit of incorporating decision times. Second, both of these variations outperform $(\lambda_{\text{weak}}, \widehat{\theta}_{\text{CH}})$, as discussed in section 5.1. Third, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$ performs similarly to $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH},\mathbb{RT}})$, suggesting that not knowing the non-decision time has minimal impact. Finally, $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,logit}})$ [67] and $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT,logit}})$ [73] do not perform as consistently well as $(\lambda_{\text{trans}}, \widehat{\theta}_{\text{CH,DT}})$, highlighting the effectiveness of our proposed choice-decision-time estimator (eq. (3)).

6 Conclusion and future work

This work is the first to leverages 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].