Enhancing Preference-based Linear Bandits via Human Response Time

Shen Li*, Yuyang Zhang*, Zhaolin Ren, Claire Liang, Na Li, Julie A. Shah

Question for you ···

Which one would you like to have now?

[1]

Long response time ➡ Weak preference

Short response time ➡ Strong preference

Research Questions

1. ***How*** to combine response times with choices to improve preference learning?
2. ***Under what conditions*** do response times provide additional value beyond choices?

Problem Formulation: Linear Bandit

- Human preference vector $\theta \in \mathbb{R}^d$
- Each arm $z \in Z$ with a utility $z^\top \theta$
- Each query x with a utility difference $x^\top \theta$

$x = z_1 - z_2$

Query x → Choice $c_x \in \{1, -1\}$; Response time $t_x > 0$

Estimating θ using feedback (see contributions 1 and 2).

Bandit algorithm

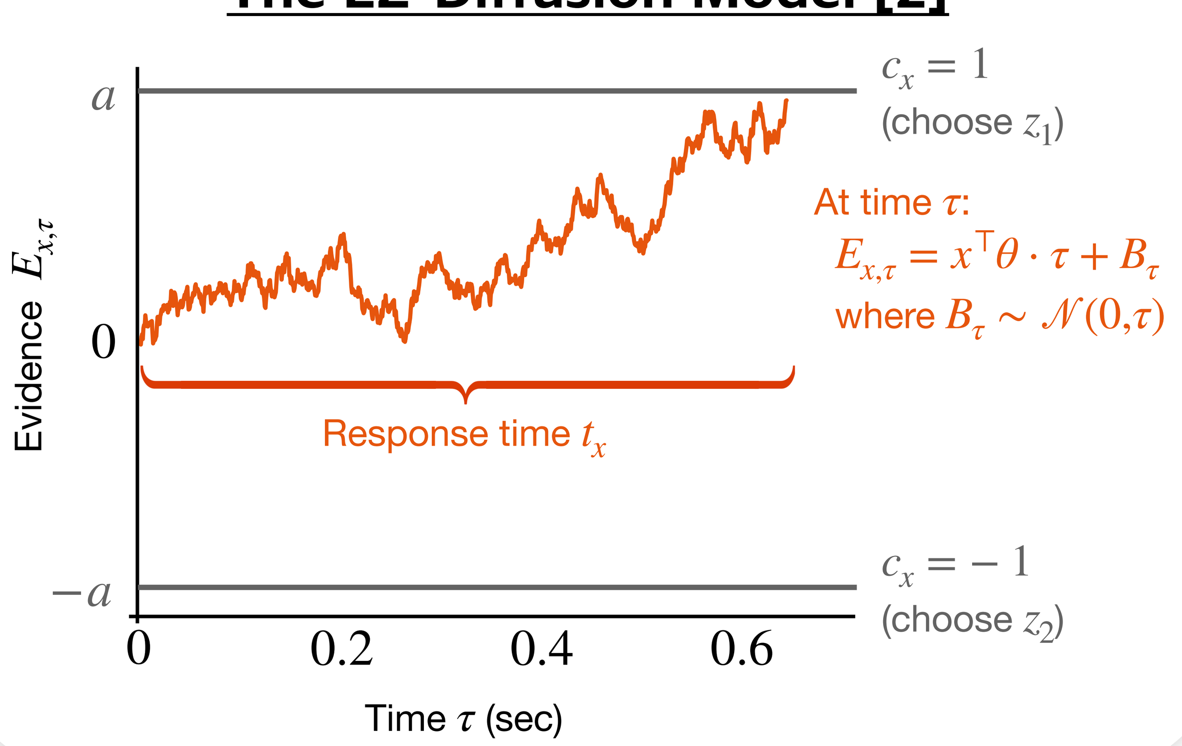
If a time budget is exhausted, then terminate.

Recommended arm $\hat{z}$

Goal: $\hat{z}$ is the best arm $z^* = \arg\max_{z \in Z} z^\top \theta$.

Integrating utility estimators into bandit learning (see panel "Bandit Learning").

The EZ-Diffusion Model [2]



Contribution 1: Utility Estimation Using Both Choices and Response Times

Estimate θ, given a fixed dataset that contains i.i.d. data $\{c_{x,i}, t_{x,i}\}_{i \in [n_x]}$ for each query $x \in \mathcal{X}$.

If using both **choices** and **response times**:

$$x^\top \frac{\theta}{a} = \frac{\mathbb{E}[c_x]}{\mathbb{E}[t_x]}$$

$$\hat{\theta}_{choices,\ times} = \arg\min_{\theta} \sum_{x \in \mathcal{X}} n_x \left(x^\top \theta - \frac{\frac{1}{n_x} \sum_{i \in [n_x]} c_{x,i}}{\frac{1}{n_x} \sum_{i \in [n_x]} t_{x,i}} \right)^2$$

If using **choices** (same as Bradley-Terry [4]):

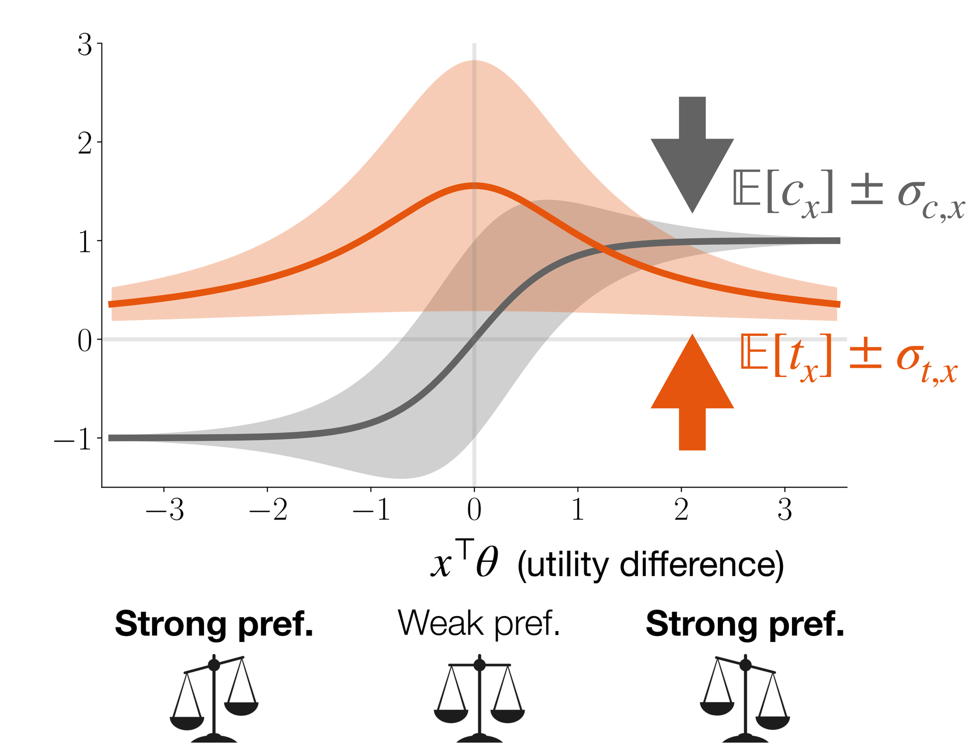
$$\mathbb{P}\left[c_x = 1\right] = \frac{1}{1 + \exp\left(-c_x \cdot x^\top \cdot 2a\theta\right)}$$

$$\hat{\theta}_{choices} = \arg\max_{\theta} \sum_{x \in \mathcal{X},\ i \in [n_x]} \log \frac{1}{1 + \exp\left(-c_{x,i} \cdot x^\top \cdot \theta\right)}$$

Contribution 2: A Key Insight:

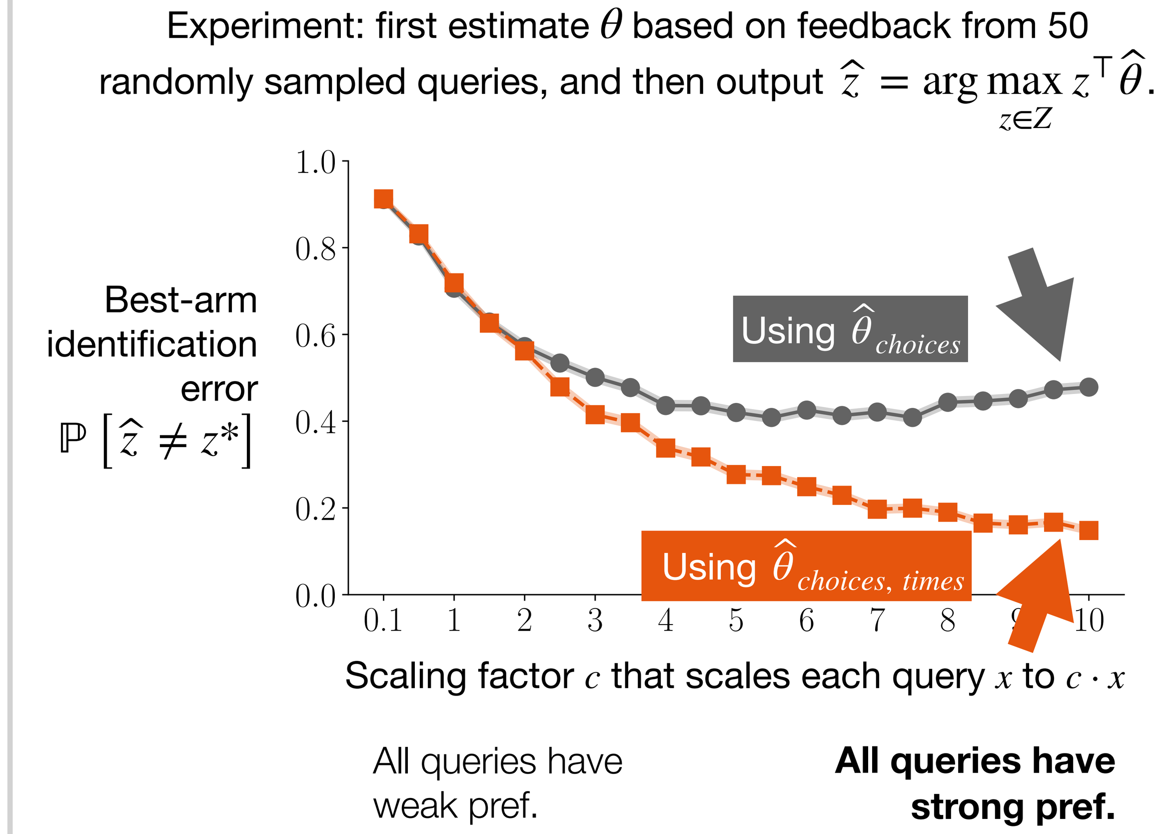
Response times from queries with strong preferences provide extra information beyond choices, which accelerates preference learning.

EZ-diffusion model confirms the key insight:



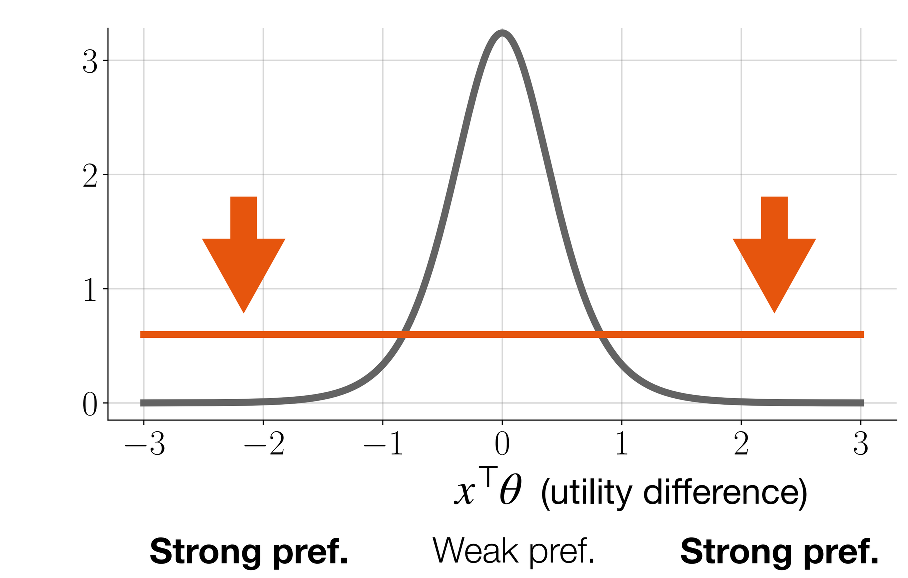
Simulation result confirms the key insight:

Experiment: first estimate θ based on feedback from 50 randomly sampled queries, and then output $\hat{z} = \arg\max_{z \in Z} z^\top \hat{\theta}$.



Asymptotic variances confirms the key insight:

(The plot shows an example where all $x^\top \theta \in [-3, 3]$.)



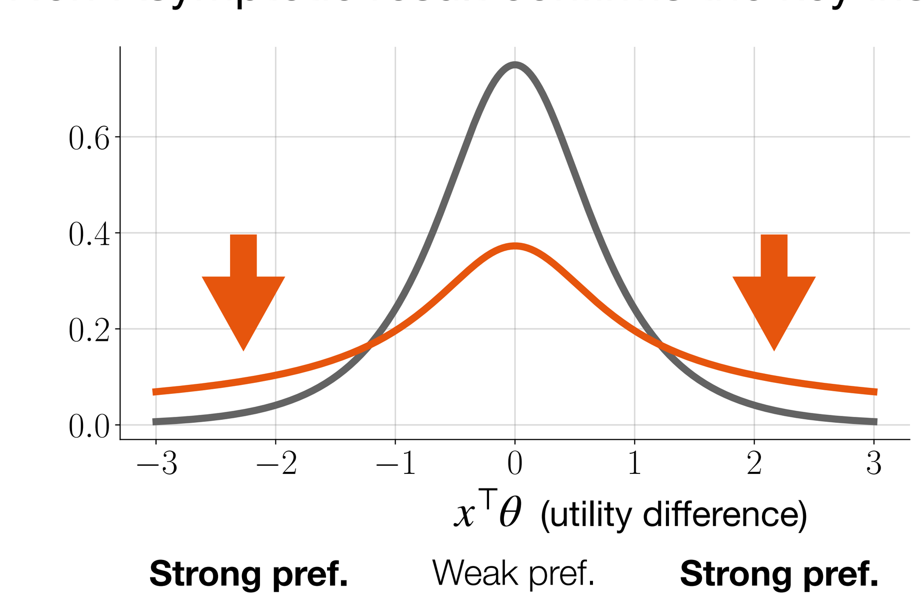
Given a fixed dataset with n choices and response times for each query in $\mathcal{X}$, then, for each arm z, the utility estimation error satisfies:

$$\sqrt{n}\left(z^\top \hat{\theta} - z^\top \theta\right) \xrightarrow{D} \mathcal{N}\left(0, \mathrm{AVar}_z\right).$$

If using $\hat{\theta}_{choices,\ times}$, then $\mathrm{AVar}_z \leq z^\top \left(\sum_{x \in \mathcal{X}} \boxed{\min_{\tilde{x} \in \mathcal{X}} \mathbb{E}[t_{\tilde{x}}]}\, x x^\top \right)^{-1} z$

If using $\hat{\theta}_{choices}$, then $\mathrm{AVar}_z = z^\top \left(\sum_{x \in \mathcal{X}} \boxed{a^2 \mathrm{Var}[c_x]}\, x x^\top \right)^{-1} z$

Non-Asymptotic result confirms the key insight:



For each query x with $x^\top \theta \neq 0$:

given a fixed i.i.d. dataset with n_x choices and response times, for any $\epsilon > 0$, if ϵ is sufficiently small and n_x is sufficiently large, then, the utility estimation error satisfies:

$$\mathbb{P}\left[\left|x^\top \hat{\theta} - x^\top \theta\right| > \epsilon\right] \leq 6 \exp\left(-M_x^2 \cdot n_x \cdot \epsilon^2\right).$$

If using $\hat{\theta}_{choices,\ times}$, then $M_x = \dfrac{1}{2 + 2\sqrt{2}} \dfrac{\mathbb{E}[t_x]}{a}$

If using $\hat{\theta}_{choices}$, then $M_x = \dfrac{1}{2.4} a \sqrt{\mathbb{V}[c_x]}$

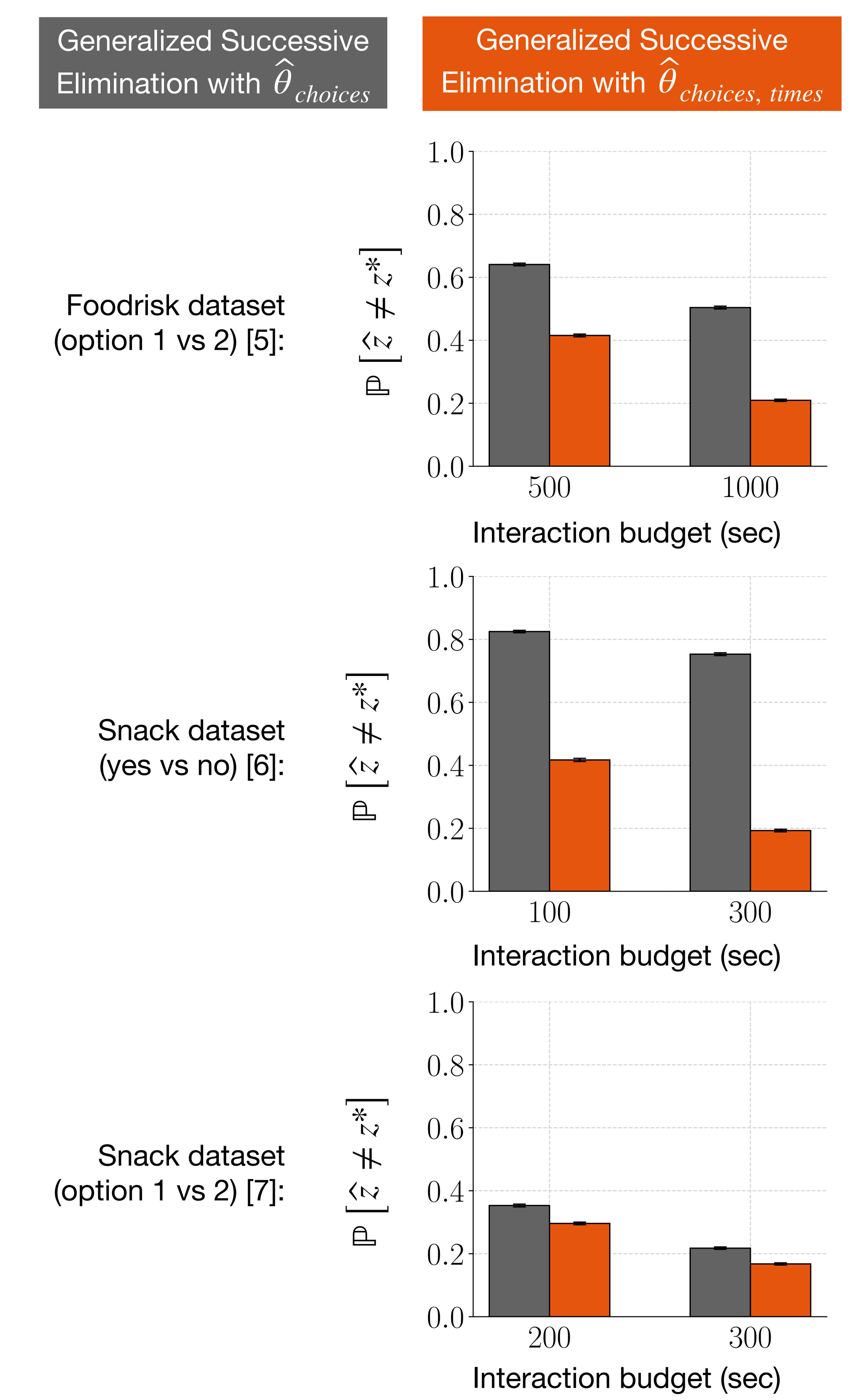
Bandit Learning

Algorithm: Generalized Successive Elimination [3]

- Split the total budget B evenly into S phases.
- For each phase $s = 1, \ldots, S$:
  - Compute the experimental design λ_s (a distribution over queries).
  - Sample queries according to λ_s till the budget is exhausted.
  - Estimate θ and eliminate the arms with low estimated utilities.
- Recommend the one arm remaining.

- Hyperparameters
  - Elimination parameter η determines that
    1. $S = \left\lceil \log_\eta |Z| \right\rceil$
    2. Only keep the top $\left\lceil \frac{|Z_s|}{\eta} \right\rceil$ arms at the end of each phase.
  - Buffer size B_{buff} determines each phase's budget $\frac{B}{S} - B_{buff}$, to prevent over-consuming the budget.

Simulation results based on real-world datasets



Scan for full paper

Scan for Shen Li's website (I am on the faculty job market!)

Harvard John A. Paulson School of Engineering and Applied Sciences

References

[1] Alós-Ferrer, C., Fehr, E., & Netzer, N. (2021). Time will tell: Recovering preferences when choices are noisy. Journal of Political Economy.

[2] Wagenmakers, E. J., Van Der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. Psychonomic bulletin & review.

[3] Azizi, M. J., Kveton, B., & Ghavamzadeh, M. (2022). Fixed-budget best-arm identification in structured bandits. IJCAI.

[4] Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*.

[5] Smith, S. M., & Krajbich, I. (2018). Attention and choice across domains. Journal of Experimental Psychology: General.

[6] Clithero, J. A. (2018). Improving out-of-sample predictions using response times and a model of the decision process. Journal of Economic Behavior & Organization.

[7] Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. Nature neuroscience.