Learning Linear Utility Functions From Pairwise Comparison Queries

Luise Ge, Brendan Juba, Yevgeniy Vorobeychik

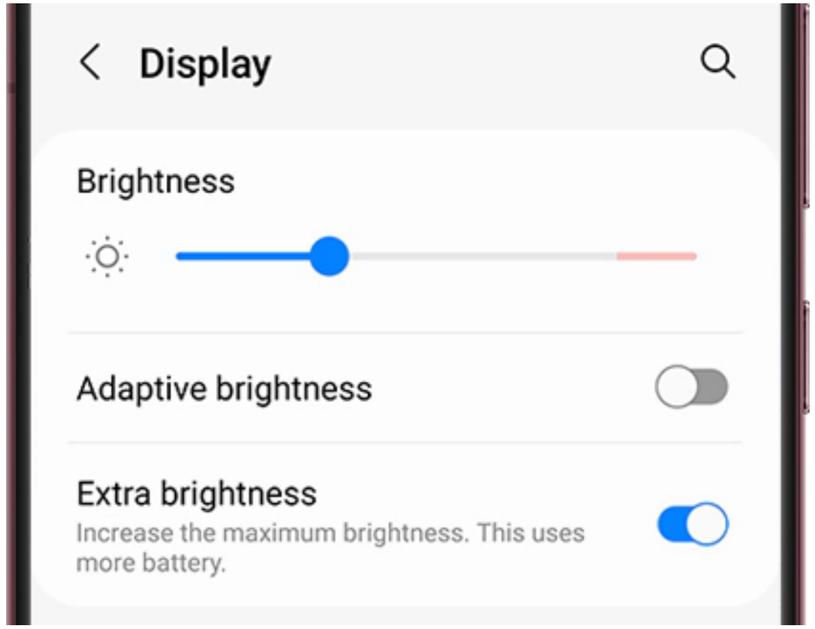
Washington University in St. Louis

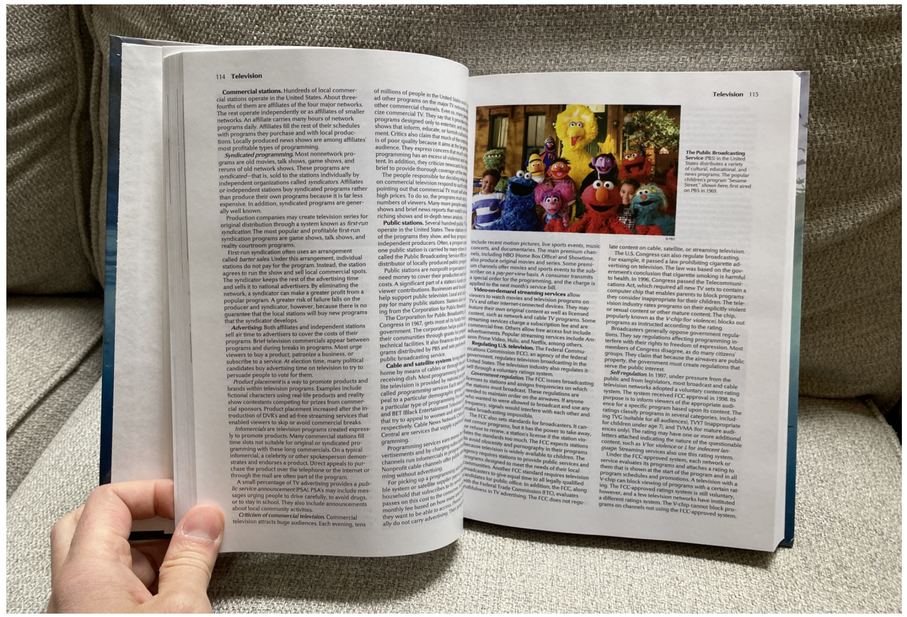
Motivation

Psychology
Economics
Algorithmic Game Theory
Computational Social Choice
AI Alignment

Theory: A fixed set of candidates

Reality: Infinite space for choices





Under what conditions is sample-efficient learning from pairwise comparisons possible?

Learning Set-up

Output : $u(x) = \mathbf{w}^T \phi(x), \|w\|_1 = 1, w > 0,\ \phi(x) \in [0,1]^m$

Input: $\mathscr{D} = \{(x_i, x_i', y_i)\}_{i=1}^n$

Random Utility Model

$$\tilde{u}(x) = u(x) + \tilde{\xi}(x) \qquad x' \succ x \text{ if } \tilde{u}(x') \geq \tilde{u}(x)$$

$$\Pr(x' \succ x) = F(w^T \Delta_\phi(x)), \; \Delta_\phi(x) := \phi(x') - \phi(x), \text{ and}$$

F be the cdf of $\zeta = \tilde{\xi}(\phi(x)) - \tilde{\xi}(\phi(x'))$

Example:

Bradley-Terry (BT)

$$F_{BT}(x) = \frac{1}{1 + \exp(-x)}$$

Thurstone-Mosteller(TM)

$$F_{TM} = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$$

Passive—PAC Learning from Pairwise Comparisons (PAC-PC)

Given a noise distribution $\zeta \sim \mathcal{Q}$, for any input distribution $\mathcal{P}$,

$\forall \epsilon, \delta \in (0,1)$, whenever $n \geq n_A(\epsilon, \delta)$,

the learning algorithm $\mathcal{A}(\{(x_i, x_i', y_i)\}_{i=1}^n)$ with $(x_i, x_i') \sim \mathcal{P}$ i.i.d.,

And y_i follows the RUM with $\mathcal{Q}$,

returns $\hat{u}$ with probability at least $1 - \delta$,

such that $e(\hat{u}, u) \leq \epsilon$ under the same distribution $\mathcal{P}$.

Efficient if $n_A(\epsilon, \delta)$ is polynomial in $\frac{1}{\epsilon}, \frac{1}{\delta}$, VCdim$(\mathcal{U}) = m$

Active learning

We interact with the oracle

Preference prediction:

$$e_1(\hat{u}, u) = \Pr_{(x,x')\sim\mathscr{P}} \left(Z_{\hat{u}}(x, x') \neq Z_u(x, x') \right)$$

Parameter Estimation:

$$e_2(\hat{u}, u) = \|\hat{w} - w^*\|_p^p$$

		Noise-free	RUM noisy
Passive Learning	Preference prediction		
	Parameter Estimation		
Active Learning	Preference prediction		
	Parameter Estimation		

Preference prediction w.o. noise:

$$\Delta_\phi(x) := \phi(x') - \phi(x)$$

$$y = sign(w^T \Delta_\phi(x))$$

		Noise-free	RUM noisy
Passive Learning	Preference prediction	Easy,LP/ Perceptron/ SVM	
	Parameter Estimation		
Active Learning	Preference prediction		
	Parameter Estimation		

Active learning is ***not harder than*** passive learning.

		Noise-free	**RUM noisy**
Passive Learning	Preference prediction	Easy	
	Parameter Estimation		
Active Learning	Preference prediction	Easy	
	Parameter Estimation		

Alabdulmohsin, Ibrahim, Xin Gao, and Xiangliang Zhang. "Efficient active learning of halfspaces via query synthesis." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. No. 1. 2015.

Preference prediction with noise:

(Negative Result). Without any restrictions on the data distribution, we are unable to ***bound the noise, i.e.*** the probability of receiving a wrong label $\eta(x)$ from $\frac{1}{2}$. Efficient learning is impossible.

Tsybakov Noise $\Pr_{x \sim \mathscr{D}_x}[\eta(x) \geq \frac{1}{2} - t] \leq A t^{\frac{\alpha}{1-\alpha}}$

(Positive Result). Suppose that $\mathscr{P}_\phi$ is well behaved and the inverse of noise c.d.f. is bounded by a polynomial. Then the linear utility class $\mathscr{U}$ is efficiently PAC-PC learnable.

Diakonikolas, Ilias, et al. "Efficiently learning halfspaces with tsybakov noise." 2021. Tsybakov Halfspaces under Well-Behaved Distributions

		Noise-free	RUM noisy
Passive Learning	Preference prediction	Easy	**Efficient under well-behaved distribution+ polynomial c.d.f. inverse bound**
	Parameter Estimation		
Active Learning	Preference prediction	Easy	**Efficient under distribution+noise assumptions**
	Parameter Estimation		

Passive Parameter Estimation w.o. noise:

Counterexamples:

1.Dirac distribution

2.Consider a dataset with each pair satisfying
$\phi(x)^i - \phi(x')^i > 0 \ \forall i \in [1,m]$

Then by assumption $w > 0$, the labels will all be 0, which wouldn't give us any information.

		Noise-free	**RUM noisy**
Passive Learning	Preference prediction	Easy	Polynomial With assumptions
	Parameter Estimation	***Hard***	
Active Learning	Preference prediction	Easy	Polynomial With assumptions
	Parameter Estimation		

Passive Parameter Estimation with noise:

Negative result: Dirac

Positive: $$||\hat{\vec{w}} - \vec{w}^*||_\Sigma \leq C \cdot \sqrt{\frac{m + \log(1/\delta)}{n}}.$$

[1]. Banghua Zhu, Michael Jordan, Jiantao Jiao. 2023 ICML . Principled reinforcement learning with human feedback from pairwise or k-wise comparisons

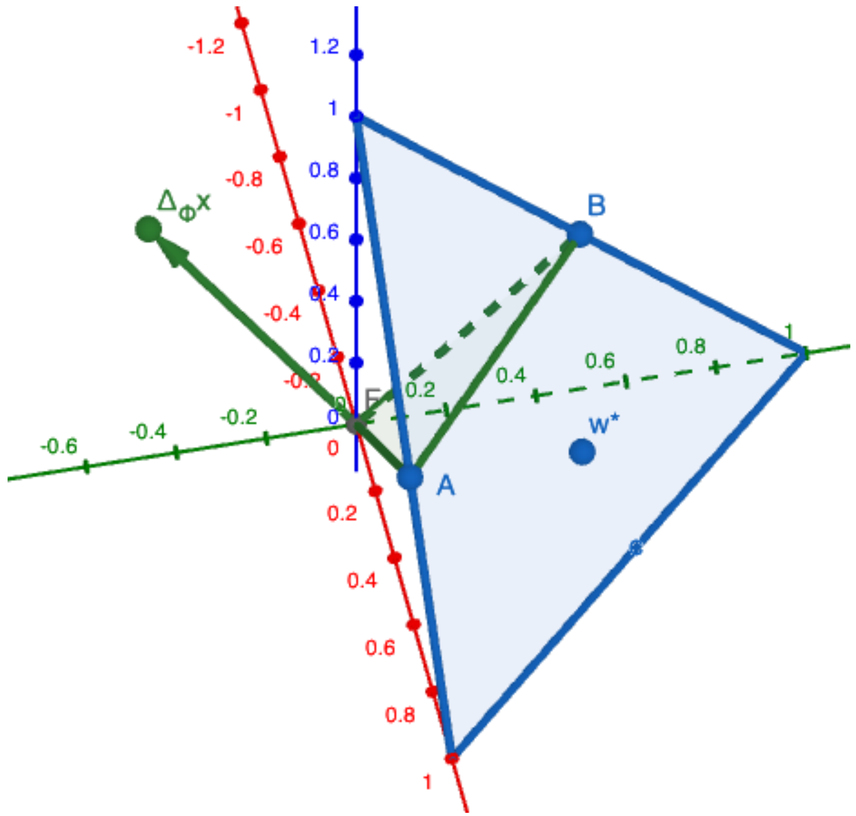
(Extended positive result)

If there exists $\gamma > 0$ such that the noise c.d.f. satisfies $F'(\zeta)^2 - F''(\zeta) \cdot F(\zeta) \geq \gamma$ for all ζ, sample complexity is $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\delta}) + \frac{m}{\epsilon})$

Remark: So having noise could either help or challenge the learning problem.

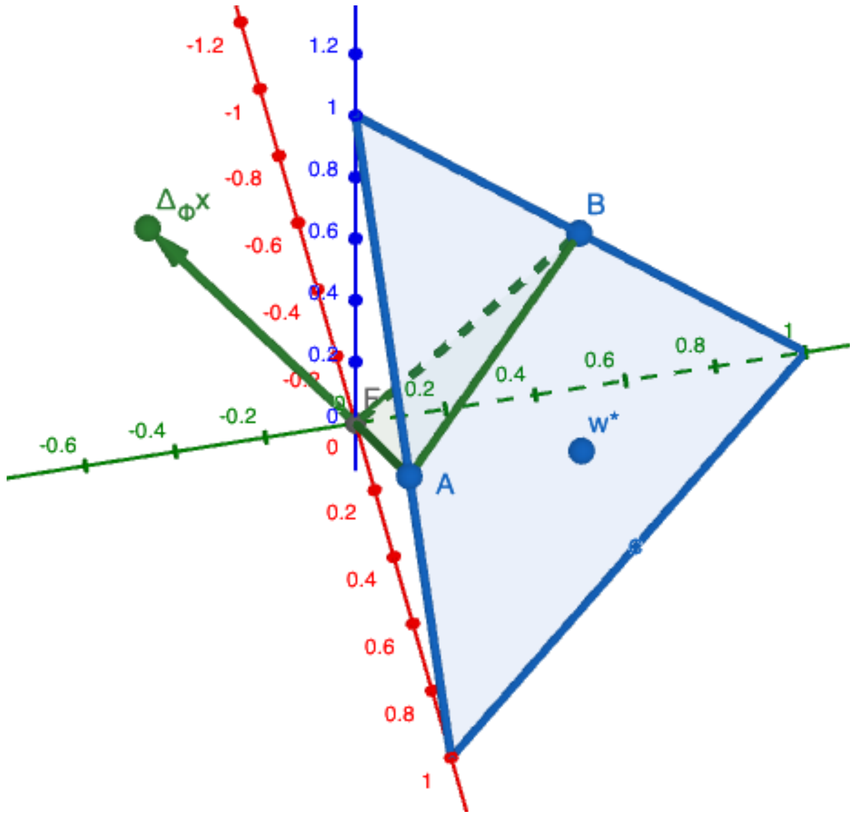
		Noise-free	**RUM noisy**
Passive Learning	Preference prediction	Easy	Polynomial With assumptions
	Parameter Estimation	Hard	***Polynomial in specifc cases***
Active Learning	Preference prediction	Easy	Polynomial With assumptions
	Parameter Estimation		

Active Parameter Estimation w.o. noise:



Sample complexity: $\mathcal{O}(m \log(\frac{\sqrt{m}}{\epsilon}))$

Active Parameter Estimation with noise:



Idea: use repetitive queries

$$\mathcal{O}(\frac{1}{(p_0 - 1/2)^2}\log(\frac{1}{\delta})), p_0 = F(\frac{\epsilon}{\sqrt{m-1}}) > 1/2$$

Eg. For BT, passive v.s. active

$$\mathcal{O}(\frac{1}{\epsilon}(\log(\frac{1}{\delta}) + m)) \qquad \mathcal{O}(\frac{1}{\exp(\frac{\epsilon}{\sqrt{m}})}\log(\frac{1}{\delta}))$$

		Noise-free	RUM noisy
Passive Learning	Preference prediction	Easy	Polynomial With assumptions
	Parameter Estimation	Hard	Hard, Polynomial in specifc cases
Active Learning	Preference prediction	Easy	Polynomial With assumptions
	Parameter Estimation	Easy	Polynomial when 1/F is polynomial

Takeaways:

1. Not impossible!

2. Prediction is easier than estimation.

3. Some noise model can help with learning. But what model?

4. Active learning is promising.
But how to make sure your made-up queries are meaningful?

Axioms for AI Alignment from Human Feedback

Luise Ge
Wash U

Daniel Halpern
Harvard

Evi Micha
Harvard

Ariel D. Procaccia
Harvard

Itai Shapira
Harvard

Yevgeniy Vorobeychik
Wash U

Junlin Wu
Wash U

Abstract

In the context of reinforcement learning from human feedback (RLHF), the reward function is generally derived from maximum likelihood estimation of a random utility model based on pairwise comparisons made by humans. The problem of learning a reward function is one of preference aggregation that, we argue, largely falls within the scope of social choice theory. From this perspective, we can evaluate different aggregation methods via established axioms, examining whether