Online Learning and Equilibrium Computation with Ranking Feedback

Mingyang Liu[1], Yongshan Chen[2,3], Zhiyuan Fan[1], Gabriele Farina[1], Asuman Ozdaglar[1], Kaiqing Zhang[2]

[1] Massachusetts Institute of Technology

[2] University of Maryland, College Park

[3] Northeastern University

Table of contents

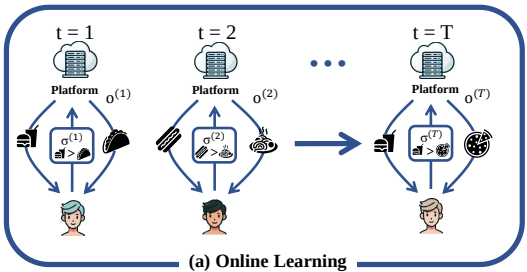
Why ranking feedback?

Numeric v.s. Ranking Feedback

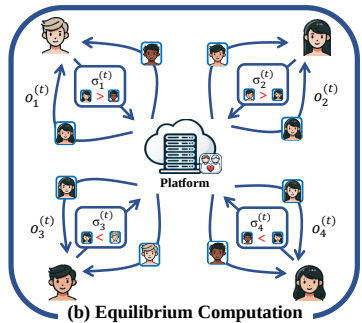
- Easier for humans to rank than score,
- Numeric utilities may be private or unavailable.

Applications

- LLM routing with user preference rankings,
- Recommendation / matching system.



(a) Online Learning



(b) Equilibrium Computation

Comparison to Dueling Bandit

$$R^{(T)} = \max_{\widehat{a} \in \mathcal{A}} \sum_{t=1}^{T} u^{(t)}(\widehat{a}) - \frac{1}{2}\left(u^{(t)}(a^{(t)}) + u^{(t)}(b^{(t)})\right). \tag{1.1}$$

Adversarial Dueling Bandit [Saha et al., 2021]

Receive $U^{(1)}, \ldots, U^{(T)} \in [0,1]^{\mathcal{A} \times \mathcal{A}}$

$$\Pr\left(a^{(t)} > b^{(t)} \mid o^{(t)} = \left\{a^{(t)}, b^{(t)}\right\}\right) = U^{(t)}(a^{(t)}, b^{(t)})$$

$$u^{(t)}(a) = \sum_{a' \neq a} U^{(t)}(a, a')$$

Ours

Receive $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(T)} \in [-1,1]^{\mathcal{A}}$

$$\Pr\left(a^{(t)} > b^{(t)} \mid o^{(t)} = \{a, b\}\right) = \frac{\exp(r^{(t)}(a))}{\exp(r^{(t)}(a)) + \exp(r^{(t)}(b))}$$

$$\boldsymbol{r}^{(t)} = \begin{cases} \boldsymbol{u}^{(t)} & \text{Instantaneous Utility} \\ \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{u}^{(s)} & \text{Time-average Utility (Full-Info)} \\ \frac{\sum_{s=1}^{t} u^{(s)}(a) \sum_{a' \in o^{(s)}} \mathbb{1}\left(a=a'\right)}{\sum_{s=1}^{t} \sum_{a' \in o^{(s)}} \mathbb{1}(a=a')} & \text{Time-average Utility (Bandit)} \end{cases}$$

Online Learning with Ranking Feedback

Procedure

- Update the strategy $\pi^{(t)} \in \Delta^{\mathcal{A}}$
- Propose a multiset (possibly include repeated elements) of actions $o^{(t)} \subseteq \mathcal{A}$
- Receive ranking $\sigma^{(t)}$ over options in $o^{(t)}$

Plackett-Luce (PL) model [Luce, 1959, Plackett, 1975]

$$\mathbb{P}\left(\sigma^{(t)} \mid o^{(t)}\right) = \prod_{k_1=1}^{K} \frac{\exp\left(\frac{1}{\tau} r^{(t)}\left(\sigma^{(t)}(k_1)\right)\right)}{\sum_{k_2=k_1}^{K} \exp\left(\frac{1}{\tau} r^{(t)}\left(\sigma^{(t)}(k_2)\right)\right)}. \tag{1.2}$$

$r^{(t)}$ is a function of $\left\{\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(t)}\right\}$ to be specified later.

Setting 1: Ranking with Instantaneous Utility

$$\boldsymbol{r}^{(t)} = \boldsymbol{u}^{(t)}. \tag{1.3}$$

Full-Information Feedback

$$o^{(t)} = \mathcal{A}$$

$$R^{(T),\text{external}} := \max_{\widehat{\pi} \in \Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \pi^{(t)} \right\rangle$$

Bandit Feedback

$$o^{(t)} \sim \pi^{(t)}, \quad \left| o^{(t)} \right| = K$$

$$R^{(T)} := \max_{\widehat{\pi} \in \Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left(\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} \right\rangle - \frac{1}{K} \sum_{a \in o^{(t)}} u^{(t)}(a) \right)$$

Setting 2: Ranking with Time-average Utility

Full-Information Feedback

$$\boldsymbol{r}^{(t)} = \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{u}^{(s)}$$

$$o^{(t)} = \mathcal{A}$$

$$R^{(T),\text{external}} := \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \pi^{(t)} \right\rangle$$

Bandit Feedback

$$\boldsymbol{r}^{(t)} = \frac{\sum_{s=1}^{t} u^{(s)}(a) \sum_{a'\in o^{(s)}} \mathbb{1}\left(a = a'\right)}{\sum_{s=1}^{t} \sum_{a'\in o^{(s)}} \mathbb{1}\left(a = a'\right)}$$

$$o^{(t)} \sim \pi^{(t)}, \quad \left|o^{(t)}\right| = K$$

$$R^{(T)} := \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left(\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} \right\rangle - \frac{1}{K} \sum_{a\in o^{(t)}} u^{(t)}(a) \right)$$

Overview

$$P^{(T)} := \sum_{t=2}^{T} \left\| \boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)} \right\| \leq \mathcal{O}(T^q). \tag{2.1}$$

Lower Bound	Full-Information	Bandit
InstUtil Rank	$\Omega(T)$ for $\tau \leq \mathcal{O}(1)$	
AvgUtil Rank	$\widetilde{\Omega}(T)$ for $\tau \leq \mathcal{O}\left(\frac{1}{T \log T}\right)$	$\Omega(T)$ for $\tau \leq \mathcal{O}\left(\frac{1}{\log T}\right)$

Upper Bound ($\tau = \mathcal{O}(1)$, Sublinear Regret)	Full-Information	Bandit
InstUtil Rank	$q < 1$	
AvgUtil Rank	✓	$q < \frac{1}{3}$

Table: Summary of our contributions, including the *negative results* (top) and the *positive results* (bottom). ✓indicates that no additional assumptions are needed. Here, $\tau > 0$ denotes the temperature parameter of the ranking model in the PL model.

Impossibility Results on InstUtil Rank (Proof Sketch)

When $\tau = 0.1$ and $\mathcal{A} = \{a, b\}$,

$$\text{Instance 1} = \begin{cases} (-0.5, 0) & w.p. \quad \frac{4}{13} \\ (0.15, 0) & w.p. \quad \frac{9}{13} \end{cases} \qquad \text{Instance 2} = \begin{cases} (-0.02, 0) & w.p. \quad \approx 0.58 \\ (0.1, 0) & w.p. \quad \approx 0.42 \end{cases}$$

Then,

$$\Pr\left(a > b \,|\, o^{(t)} = \{a, b\}\right) \text{ is the same in both instances.}$$

However,

$$\mathbb{E}_{\text{Instance 1}}\left[u(a)\right] = -0.05 \qquad \mathbb{E}_{\text{Instance 2}}\left[u(a)\right] \approx -0.03.$$

Routing Large Language Models

Experimental Setup

- Dataset: HH-RLHF [Bai et al., 2022]
- $\mathcal{A}$ ={Qwen3-32B [Yang et al., 2025], Phi-4 [Abdin et al., 2024], GPT-4o [Hurst et al., 2024], and Llama-3.1-70B [Dubey et al., 2024]}
- A reward model generates the utility to simulate human preference (huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2)

Routing Large Language Models

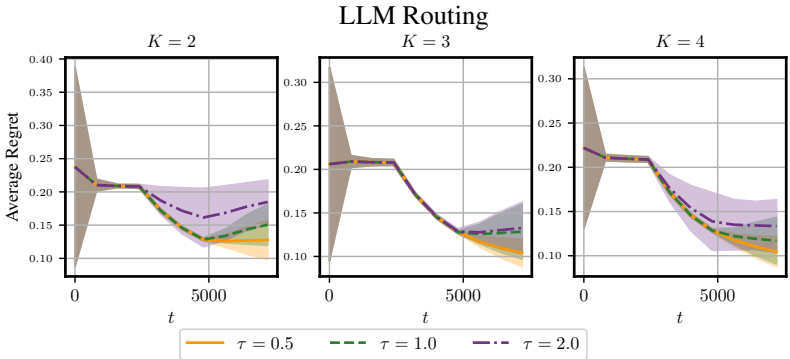


Figure: $R^{(T)}$ with AvgUtil under bandit feedback for different temperatures τ and numbers of proposed actions K in the online learning setting.

Thank You!

Any questions?

References I

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning (ICML)*, 2021.

R Duncan Luce. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

References II

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.