Online Learning and Equilibrium Computation with Ranking Feedback

Mingyang Liu, Yongshan Chen, Zhiyuan Fan, Gabriele Farina,
Asuman Ozdaglar, Kaiqing Zhang

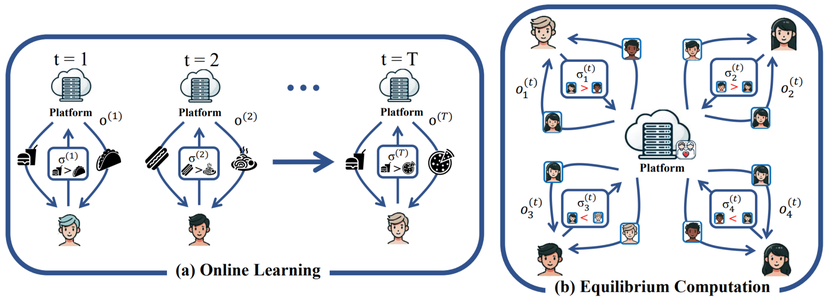
Examples

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.



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

$$\Pr\left(a^{(t)} > b^{(t)} \mid o^{(t)} = \{a, b\}\right) = U^{(t)}(a^{(t)}, b^{(t)})$$
$$u^{(t)}(a) = \sum_{a'\neq a} U^{(t)}(a, a')$$

Ours

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

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

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

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]

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

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

Ranking with Instantaneous Utility

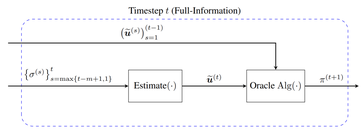
$$\boldsymbol{r}^{(t)} = \boldsymbol{u}^{(t)}.$$

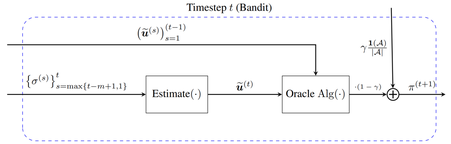
Full-Information Feedback

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

Bandit Feedback

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





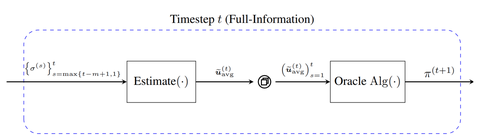
Ranking with Time-average Utility

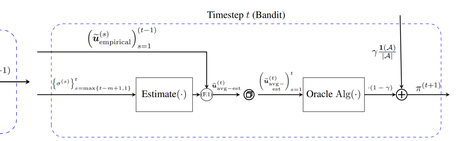
Full-Information Feedback

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

Bandit Feedback

$$\boldsymbol{r}^{(t)} = \frac{\sum_{s=1}^{t} u^{(s)}(a) \sum_{a'\in o^{(s)}} \mathbb{1}\left(a=a'\right)}{\sum_{s=1}^{t} \sum_{a'\in o^{(s)}} \mathbb{1}\left(a=a'\right)}$$
$$o^{(t)} \sim \pi^{(t)}, \quad \left|o^{(t)}\right| = K$$
$$R^{(T)} := \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left(\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} \right\rangle - \frac{1}{K} \sum_{a\in o^{(t)}} u^{(t)}(a)\right)$$





Utility Estimation

Algorithm 1 Utility Estimation with Action Permutations: Estimate$\left(\left\{\sigma^{(s)}\right\}_{s=1}^{m'}\right)$

Input: A set consisting of m' permutations of actions : $\left\{\sigma^{(s)}\right\}_{s=1}^{m'}$ with $|\sigma^{(s)}| = K$ for all $s \in [m']$, and temperature $\tau > 0$.
for $j = 1, 2, \dots, |\mathcal{A}| - 1$ **do**
  for $s = 1, \dots, m'$ **do**
    Calculate $n_{j,1}^{(s)}, n_{j,2}^{(s)}$ defined as

$$n_{j,1}^{(s)} := \sum_{i,k\in[K]} \mathbb{1}\left(\sigma^{(s)}(i) = a^j, \sigma^{(s)}(k) = a^{|\mathcal{A}|} \text{and } i < k\right),$$
$$n_{j,2}^{(s)} := \sum_{i,k\in[K]} \mathbb{1}\left(\sigma^{(s)}(i) = a^j, \sigma^{(s)}(k) = a^{|\mathcal{A}|} \text{and } i > k\right).$$

  end for
  Let $\mathcal{T}_j := \left\{s \in [m'] \colon n_{j,1}^{(s)} + n_{j,2}^{(s)} > 0\right\}$
  Let $\text{sig}^{-1}(x) \colon (0,1) \to \mathbb{R} := \log\frac{x}{1-x}$ be the inverse function of $\text{sig}(\cdot)$. The utility of action a^j is then estimated as

$$\widetilde{u}(a^j) = \begin{cases} \text{Proj}_{[-1,1]}\left(\tau \text{sig}^{-1}\left(\frac{1}{|\mathcal{T}_j|} \cdot \sum_{s\in\mathcal{T}_j}\left(\frac{n_{j,1}^{(s)}}{n_{j,1}^{(s)}+n_{j,2}^{(s)}}\right)\right)\right) & |\mathcal{T}_j| > 0 \\ 0 & |\mathcal{T}_j| = 0. \end{cases}$$

end for
Return $\widetilde{\boldsymbol{u}} = \left(\widetilde{u}(a^1), \widetilde{u}(a^2), \dots, \widetilde{u}\left(a^{|\mathcal{A}|-1}\right), 0\right)$

Overview

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

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

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.

Computing Equilibrium in Games

Assumption 6.1. *The (full-information) online learning algorithm* Alg *satisfies the following condition: for any $T > 0$, $t \in [T]$, and sequences of utilities $\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}, \left(\boldsymbol{u}'^{(s)}\right)_{s=1}^{t} \in \left(\mathbb{R}^{\mathcal{A}}\right)^t$, we have*

$$\left\|\text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right) - \text{Alg}\left(\left(\boldsymbol{u}'^{(s)}\right)_{s=1}^{t}\right)\right\| \le L\left\|\sum_{s=1}^{t}\boldsymbol{u}^{(s)} - \sum_{s=1}^{t}\boldsymbol{u}'^{(s)}\right\|,$$

where $L = \Theta\left(T^{-c}\right)$ for some constant $c \in (0,1)$.

Assumption 7.1 (Sublinear variation of strategies)**.** *The (full-information) online learning algorithm* Alg *needs to satisfy the following condition: for any $T > 0$, $t \in [T-1]$, and sequence of utility vectors $\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t+1} \in \left([-1,1]^{\mathcal{A}}\right)^{t+1}$, we have $\left\|\text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right) - \text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t+1}\right)\right\| \le \eta$, where $\eta = \Theta\left(T^{-w}\right)$ for some constant $w \in (0,1)$.*

Theorem 7.2. *Consider* **InstUtil Rank** *with constant $\tau > 0$ and Algorithm 2. For any $\delta \in (0,1)$, $T > 0$, and any full-information no-regret learning algorithm with numeric utility feedback,* Alg*, that satisfies Assumption 7.1, by choosing M, m, γ according to Theorem 5.2, we have that with probability at least $1-\delta$, the algorithm finds an ϵ-CCE, with*

$$\epsilon \le \max_{i\in[N]}\left\{\frac{1}{T}R_i^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}_i^{(t)}\right)_{t=1}^{T}\right)\right\} + \mathcal{O}\left(\eta^{\frac{1}{3}}\left(\log\left(\frac{T}{\delta}\right)\right)^{\frac{1}{3}}\right) \quad \text{(Full-Information)}$$

$$\epsilon \le \max_{i\in[N]}\left\{\frac{1}{T}R_i^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}_i^{(t)}\right)_{t=1}^{T}\right)\right\} + \mathcal{O}\left(\eta^{\frac{1}{5}}\log\left(\frac{T}{\delta}\right)\right). \quad \text{(Bandit)}$$

Theorem 7.3. *Consider* **AvgUtil Rank** *with constant $\tau > 0$ and Algorithm 3. For any $\delta \in (0,1)$, $T > 0$, and any full-information no-regret learning algorithm with numeric utility feedback,* Alg*, that satisfies Assumption 6.1, by choosing M, m, γ according to Theorem 6.2, we have that with probability at least $1-\delta$, the algorithm finds an ϵ-CCE under full-information feedback, with*

$$\epsilon \le \max_{i\in[N]}\left\{\frac{1}{T}R_i^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}_i^{(t)}\right)_{t=1}^{T}\right)\right\} + \mathcal{O}\left(LT^{\frac{8}{9}}\log\left(\frac{T}{\delta}\right)\right). \quad \text{(Full-Information)}$$

When M, m, and γ are chosen as in Theorem 6.3, and both Assumption 6.1 and Assumption 7.1 hold, we have that with probability at least $1-\delta$, the algorithm finds an ϵ-CCE under bandit feedback, with

$$\epsilon \le \max_{i\in[N]}\left\{\frac{1}{T}R_i^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}_i^{(t)}\right)_{t=1}^{T}\right)\right\} + \widetilde{\mathcal{O}}\left(\left(\log\left(\frac{1}{\delta}\right)\right)^2\left(L^{\frac{1}{3}}\eta^{\frac{1}{6}} + L^{\frac{1}{2}}\right)T^{\frac{4}{9}}\right). \quad \text{(Bandit)}$$

LLM Routing with Ranking Feedback

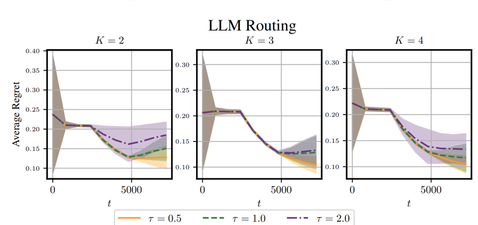


Figure: $R^{(T)}$ with AvgUtil under bandit feedback for different temperatures τ and numbers of proposed actions K in the online learning setting.