Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem

Junpei Komiyama 1, Junya Honda 1, Hisashi Kashima 2, Hiroshi Nakagawa 1
1. Univ. of Tokyo 2. Kyoto Univ.

at COLT2015, Paris, France.
Motivation: preference elicitation

Which kind of Sushi is the most popular?
I prefer left (fatty tuna) to right (eel).

Alice
I prefer left (flatfish) to right (shrimp).

Bob
I prefer right (shrimp) to left (fatty tuna).

Charlie
I prefer right (shrimp) to left (fatty tuna).

Q. to how many users can we recommend the most popular one?
Dueling bandit problem
[Yue+ COLT2009]

- K arms: $[K] = \{1, \ldots, K\}$
- At each round $t=1, \ldots, T$, an algorithm selects a pair of arms $(l(t), m(t))$, and receives the preferred one (binary: $l(t)$ or $m(t)$).
  - $l(t) = m(t)$ is possible, which yields no useful information.
- Probability that $i$ is preferred over $j$ is $\mu_{i,j} \in [0,1]$. 
Condorcet assumption
[Urvoy+ ICML2013]

- Condorcet winner $i^*$ is the arm s.t. $\mu_{i^*,j} > 1/2$ for any $j \neq i^*$.
  - Preferred by more people to every other arm.

- $\text{Regret}(T) = \sum_{t=1}^{T} \frac{\mu_{i^*,l(t)} + \mu_{i^*,m(t)} - 1}{2}$
  - Regret increases unless $(l(t), m(t)) = (i^*, i^*)$.
  - Find $i^*$ and select $(i^*, i^*)$ in most rounds for sublinear regret.
Main result 1: Regret lower bound

- Challenge: compare arm $i$ with which arm?

- Let $b^*(i) = \arg\min_{j: \mu_{i,j} < \frac{1}{2}} \frac{\mu_{i^*,i} + \mu_{i^*,j} - 1}{d(\mu_{i,j}, \frac{1}{2})}$. To achieve optimal $\log T$ factor of regret,

  - Compare $i$ with $b^*(i)$: check that each arm $i \neq i^*$ is NOT the Condorcet winner with minimal regret.

  - Usually $b^*(i) = i^*$, but not always the case.
Main result 2: The RMED algorithms


- Let \( I_i(t) = \sum_{j \in [K]: \hat{\mu}_{i,j} \leq \frac{1}{2}} N_{i,j}(t) d\left(\hat{\mu}_{i,j}, \frac{1}{2}\right) \).

  - # of comparisons
  - KL-divergence from Bernoulli(0.5).

- \( \exp(-I_i(t)) \) : likelihood that the arm \( i \) is the Condorcet winner.
Main routine of RMED2FH

Initial phase: draw all pair of arms $\log \log T$ times.

while $t < T$ do

1. Put each arm $i$ s.t. $\exp(-I_i(t)) > 1/t$ into the candidate set.
2. For each arm $i$ in the candidate set, compare $i$ with the estimated $\hat{b}^*(i)$.

$\hat{b}^*(i) = b^*(i)$ with high-prob

- Regret: not only $O(K \log T)$, but also the constant factor matches the lower bound (optimal!).
Numerical Experiment

Sushi dataset [Kamishima KDD2003]
- Condorcet winner is mild fatty tuna.

![Graph showing performance comparison between existing and proposed algorithms]

Existing algs.

Proposed algs: RMED1/2.

About 1/4 regret, compared with existing algorithms
Summary

- We derive regret lower bound in the dueling bandit problem.
- We propose the RMED algorithms.
  - RMED2FH compares $i$ with $b^*(i)$ with high-prob, and achieves optimal regret.

Thx for listening. More details are in our poster!