清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

Combinatorial Pure Exploration for Dueling Bandits

Yihan Du[1]

Joint work with Wei Chen[2], Longbo Huang[1] and Haoyu Zhao[1]

[1] IIIS, Tsinghua University

[2]Microsoft Research Asia

Introduction

Motivating example:

- Committee selection
    a) Survey a bystander to learn a sample of the unknown preference probability
    b) Play as few duels as possible to identify the best performing committee

- Preference-based version of the common candidate-position matching
- Scenarios: crowdsourcing, multi-player game, online advertising

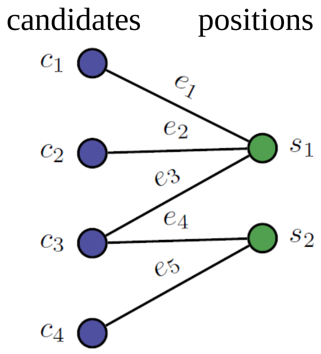
Introduction

- **Multi-Armed Bandit (MAB) [1,2]:** classic online learning problem
  Characterize the exploration-exploitation tradeoff
- **Pure exploration [3,4]:** important variant of MAB
  Identify the best arm with high confidence
- **Combinatorial Pure Exploration for Multi-Armed Bandit (CPE-MAB) [5]:**
  Given a collection of arm subsets with certain combinatorial structures
  Play an arm to identify the best combinatorial subset of arms
- **Dueling Bandit [6]:** with relative feedback
  Applications involving implicit feedback
  E.g., social surveys, market research

Introduction

Combinatorial Pure Exploration for Dueling Bandit (CPE-DB)

- **Bipartite graph $G(C,S,E)$:** candidates, positions
- **Preference matrix P :** define the preference probability of two candidates on one position
- **Preference probability of two matchings $f(M_1, M_2, P)$** is the average preference probability of duels over all positions



	e_1	e_2	e_3	e_4	c_5
e_1	0.5	0.45	1	0	0
e_2	0.55	0.5	0.55	0	0
e_3	0	0.45	0.5	0	0
e_4	0	0	0	0.5	0
e_5	0	0	0	1	0.5

Preference Matrix

Example:

$$M_1 = \{e_1, e_4\}, \qquad M_2 = \{e_2, e_5\}$$

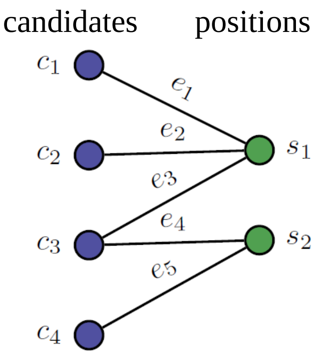
$$f(M_1, M_2, P) = \frac{1}{2}(P_{e_1, e_2} + P_{e_4, e_5})$$

Introduction

Combinatorial Pure Exploration for Dueling Bandit (CPE-DB)

- Goal: find the best matching by exploring the duels at all the positions
- Metric of the "best" matching:
  - a) **Borda winner:** the matching that maximizes the average preference probability over all valid matchings
  - b) **Condorcet winner:** the matching that always wins when compared to others.
- Applications: preference-based version of the common candidate-position matching

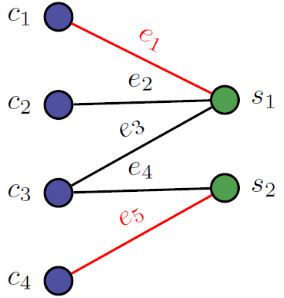
E.g., committee selection, crowdsourcing, online advertising



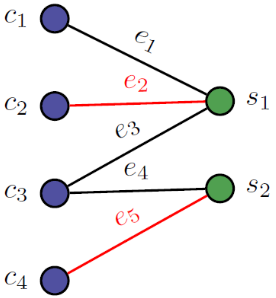
	e_1	e_2	e_3	e_4	c_5
e_1	0.5	0.45	1	0	0
e_2	0.55	0.5	0.55	0	0
e_3	0	0.45	0.5	0	0
e_4	0	0	0	0.5	0
e_5	0	0	0	1	0.5

Preference Matrix

Borda winner



Condorcet winner



Borda Metric - Reduction

Tsinghua University 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University
Microsoft

Reduction of CPE-DB for Borda winner to conventional CPE-MAB [5]

CPE-MAB [5] : pull and observe a numerical reward of edge e

Redefine the rewards:

a) Reward of edge e $w(e)$: average preference probability of e over the edges at the same position in $M \in \mathcal{M}$

b) Reward of matching M $w(M)$: sum of the rewards of its containing edges $\Longleftrightarrow$ a factor l times average preference probability of M over all $M \in \mathcal{M}$

Identify Borda winner $\Longleftrightarrow$ (equivalent) identify matching with the maximum reward

But how to learn $w(e)$ efficiently under the dueling bandit setting?

Borda Metric - CLUCB-Borda-PAC

清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

Microsoft

CLUCB-Borda-PAC is built on CLUCB [5]

Naive unbiased sampler for all matchings will cost **exponential time**

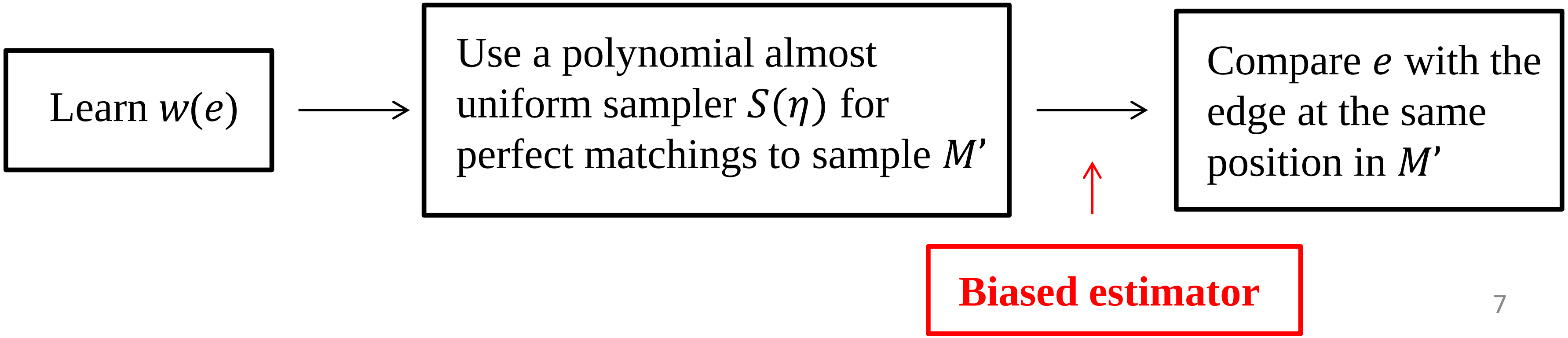
New:

a) Apply a **fully-polynomial almost uniform sampler $\boldsymbol{S(\eta)}$** for perfect matchings [7]

b) The **biased estimator** leads to additional complication in analysis

c) Novel lower bound for CPE-DB under Borda metric

Main idea: efficiently transform numerical observations to equivalent relative observations with $S(\eta)$



Borda Metric - CLUCB-Borda-Exact

清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University
Microsoft

Adapt CLUCB-Borda-PAC to the exact algorithm

Main idea:

a) Use the **"guess gap" (multiple epochs)** technique to obtain the exact solution
b) With **a loss of logarithmic factors** in sample complexity upper bound.

Iterating epoch
$q = 1,2, \ldots$
Cut the confidence and accuracy in **a half**

→

Perform CLUCB-Borda-PAC as a sub-procedure

→

Return **exact** Borda winner

↑

When the accuracy is smaller than the gap

Borda Metric – Theoretical Result

Theorem 1 (CLUCB-Borda-PAC). With probability at least $1-\delta$, CLUCB-Borda-PAC returns an approximate Borda winner with sample complexity

$$\tilde{O}\left(\sum_{e\in E}\min\left\{\frac{\mathrm{width}(G)^2}{(\Delta_e^B)^2},\frac{1}{\varepsilon^2}\right\}\right)$$

Borda hardness:

$$H^B := \sum_{e\in E}\frac{1}{(\Delta_e^B)^2}$$

Theorem 2 (CLUCB-Borda-Exact). With probability at least $1-\delta$, CLUCB-Borda-Exact returns the Borda winner with sample complexity

$$\tilde{O}\left(\mathrm{width}(G)^2 H^B\right)$$

Theorem 3 (Borda lower bound). There exists a problem instance of CPE-DB with Borda winner where any correct algorithm has sample complexity

$$\tilde{\Omega}\left(H^B\right)$$

Remark: our algorithms are tight on the hardness metric H_B

Condorcet Metric – CAR-Cond

清華大学 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University
Microsoft

Identify Condorcet winner ⟺ (equivalent) $\max_{x=\chi_{M_1}} \min_{y=\chi_{M_2}} \frac{1}{\ell} x^T P y$ (the optimal value = $\frac{1}{2}$)

$\chi_M \in \{0,1\}^m$: vector representation of matching M

This discrete optimization problem has **exponential search space**

How to **efficiently** solve it ?

Use **continuous relaxation** and just consider $\max_{x \in \mathcal{P}(\mathcal{M})} \min_{y \in \mathcal{P}(\mathcal{M})} \frac{1}{\ell} x^T P y$

$\mathcal{P}(\mathcal{M})$: convex hull of all vectors χ_M in decision class $\mathcal{M}$

Design an offline **oracle (FPTAS)** to solve this optimization problem

Projected subgradient descent, Frank-Wolfe algorithm

Condorcet Metric – CAR-Cond

清華大学 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University
Microsoft

Online part:

a) For each edge e, we force e in / out of the convex hull (polytope) $\mathcal{P}(\mathcal{M})$:

b) Check the optimal value of $\max_{x \in \mathcal{P}(\mathcal{M}, A_1, R_1)} \min_{y \in \mathcal{P}(\mathcal{M}, A_2, R_2)} \frac{1}{\ell} x^T Q y$ (the optimal value = $\frac{1}{2}$)

c) Determine whether or not e is in Condorcet winner

Theorem 4 (CAR-Cond). With probability at least $1 - \delta$, CAR-Cond returns the Condorcet winner with sample complexity

$$\tilde{O}\left(\sum_{j=1}^{\ell} \sum_{e \neq e',\ e, e' \in E_j} \frac{1}{(\Delta^C_{e,e'})^2}\right)$$

Further design **CAR-Parallel** using the **verification** technique to improve the result for small δ

Remark: When $l = 1$, the problem **reduces** to the original Condorcet dueling bandit problem

The result of our CAR-Parallel **matches the state-of-the-art [8]**

Conclusion

I. Formulate **CPE-DB**, with metrics Borda winner and Condorcet winner.

II. For Borda winner, establish **reduction** to CPE-MAB [5], propose efficient **PAC and exact** algorithms, **nearly optimal** for a subclass of problems.

III. For Condorcet winner, design offline **FPTAS** and online CAR-Cond, which is the **first polynomial** algorithm for CPE-DB with Condorcet winner.

Future Work

I. Find a **lower bound for polynomial algorithms** in CPE-DB with Condorcet winner

II. Study a more **general** CPE-DB model and other preference functions $f(M_1, M_2, P)$

References

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multi-armed bandit problem. Machine learning 47, 2-3 (2002), 235–256.

[2] Shipra Agrawal and Navin Goyal. 2012. Analysis of Thompson sampling for the multi-armed bandit problem. In COLT. 39–1.

[3] Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. Journal of machine learning research, 7(Jun):1079–1105, 2006.

[4] Chen, L. and Li, J. Open problem: Best arm identification: Almost instance-wise optimality and the gap entropy conjecture. In Conference on Learning Theory, pp. 1643–1646, 2016.

[5] Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. Combinatorial pure exploration of multi-armed bandits. In Advances in Neural Information Processing Systems, pp. 379–387, 2014.

[6] Yisong Yue and Thorsten Joachims. 2011. Beat the mean bandit. In ICML. 241–248.

[7] Jerrum, M., Sinclair, A., and Vigoda, E. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. Journal of the ACM (JACM), 51(4):671–697, 2004.

[8] Karnin, Z. S. Verification based solution for structured mab problems. In Advances in Neural Information Processing Systems, pp. 145–153, 2016.

THANK YOU！

Email: duyh18@mails.tsinghua.edu.cn