Versatile Dueling Bandits: Best-of-both World Analyses for Online Learning from Relative Preferences

Aadirupa Saha[1], Pierre Gaillard[2]

[1] Toyota Technological Institute at Chicago (TTIC) (work done at Microsoft Research, NYC)

[2] Inria, Grenoble, France

39th International Conference on Machine Learning, 2022

Motivation: Dueling Bandit (Learning from Preferences)

← **Ratings (Absolute)**

--- How much you score it out of X

Rankings (Relative) →

--- Do you like movie A over B?

Often easier (& more accurate) to elicit ***relative preferences*** than ***absolute scores***

Best-of-Both
- Stochastic
- Adversarial

Corrupted Dueling

Formally: Dueling Bandits
(Learning from pairwise preferences)

Objective: Find the BEST arm(s)
[[Regret minimization (or)
PAC best-arm identification]]

At round t,

Select *two* arms (x_t, y_t)

Unknwown but Fixed

1 2 3 4 … K

Condorcet winner (i^*)

repeat

Observe (noisy) *comparison* $o_t \in \{0,1\}$ ~ P:=

Yue and Joachims. Beat the mean bandit. ICML 2011.

Szorenyi et. al. Online rank elicitation for Plackett-Luce: A dueling bandits approach. NuerIPS 2015.

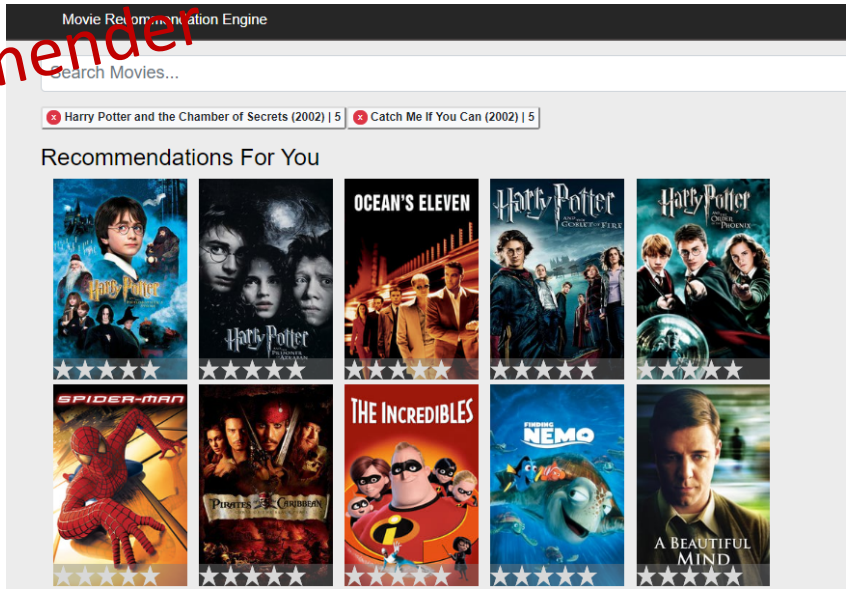
Preference Matrix

	1	2	3	4	5
1	0.5	0.53	0.54	0.56	0.6
2	0.47	0.5	0.53	0.58	0.61
3	0.46	[illegible]	[illegible]	0.54	0.57
4	0.44	[illegible]	0.46	0.5	0.51
5	0.4	0.39	0.43	0.49	0.5

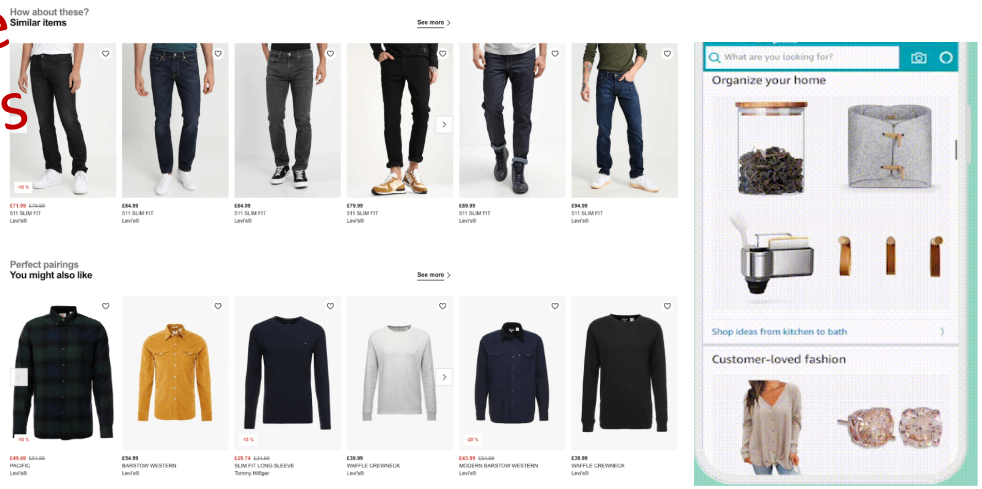
(In reality) User Preferences are Variable

Search Engines

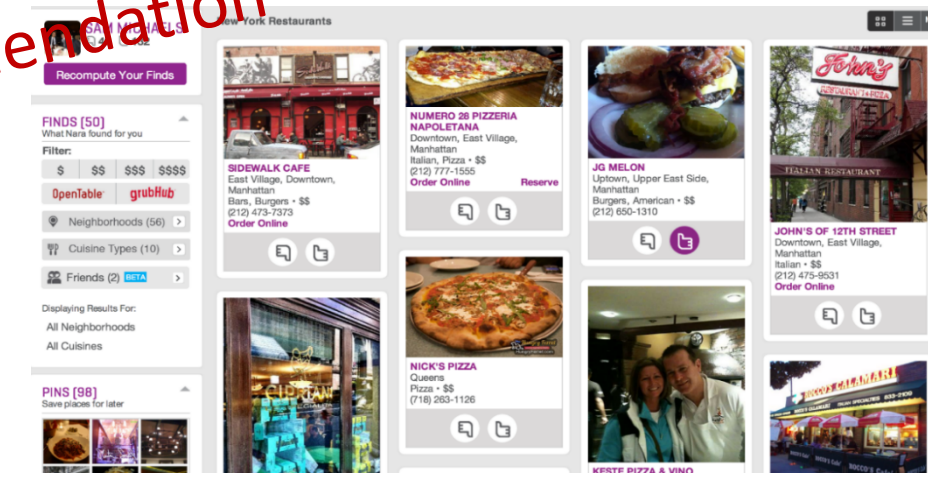
Movie Recommender



Online Retails



Restaurant Recommendation



Problem Overview: Dueling Bandit Regret (for any Adversarial Preference Sequence)

Adversarial DB (against Fixed benchmark)

	1	2	3	4	5
1	0.5	0.79	0.98	0.16	0.06
2	0.21	0.5	0.43	0.08	0.27
3	0.02	0.57	0.5	0.14	0.07
4	0.84	0.92	0.86	0.5	0.51
5	0.94	0.73	0.93	0.49	0.5

P_1 (t = 1)

	1	2	3	4	5
1	0.5	0.13	0.94	0.16	0.06
2	0.87	0.5	0.43	0.08	0.71
3	0.06	0.57	0.5	0.14	0.07
4	0.84	0.92	0.86	0.5	0.51
5	0.94	0.29	0.93	0.49	0.5

P_2 (t = 2)

	1	2	3	4	5
1	0.5	0.83	0.94	0.16	0.06
2	0.17	0.5	0.43	0.18	0.71
3	0.06	0.57	0.5	0.14	0.07
4	0.84	0.82	0.86	0.5	0.51
5	0.94	0.29	0.93	0.49	0.5

P_3 (t = 3)

	1	2	3	4	5
1	0.5	0.83	0.94	0.16	0.06
2	0.17	0.5	0.43	0.18	0.71
3	0.06	0.57	0.5	0.14	0.07
4	0.84	0.82	0.86	0.5	0.51
5	0.94	0.29	0.93	0.49	0.5

P_4 (t = 4) ....

Relative strength of Arm-i against left arm

Relative strength of Arm-i aginst right arm

Static Regret (against fixed benchmark):

$$R_T(i) := \max_{i \in [K]} \sum_{t=1}^{T} \frac{\left\{\left[P_t(i, x_t) - \frac{1}{2}\right] + \left[P_t(i, y_t) - \frac{1}{2}\right]\right\}}{2}$$

Average strength of Arm-i

Solving DB by
Reducing to “Sparring”-MAB

Adversarial Dueling (against Fixed benchmark)

History of Sparring

- **Ailon et al. Reducing Dueling Bandits to Cardinal Bandits. ICML'14**
- **Gajane et al. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. ICML'15**

But: only Utility based preferences $P_t(a,b) = \frac{(u_a - u_b)+1}{2}$

- ***** Saha & Gupta. Optimal and Efficient Dynamic Regret Algorithms for Non-Stationary Dueling Bandits. ICML'22 *****

---results apply for any **general sequence $P_1, P_2, \ldots P_T$**

(Adv) MAB – vs – (Adv) MAB

[SPARRING!]

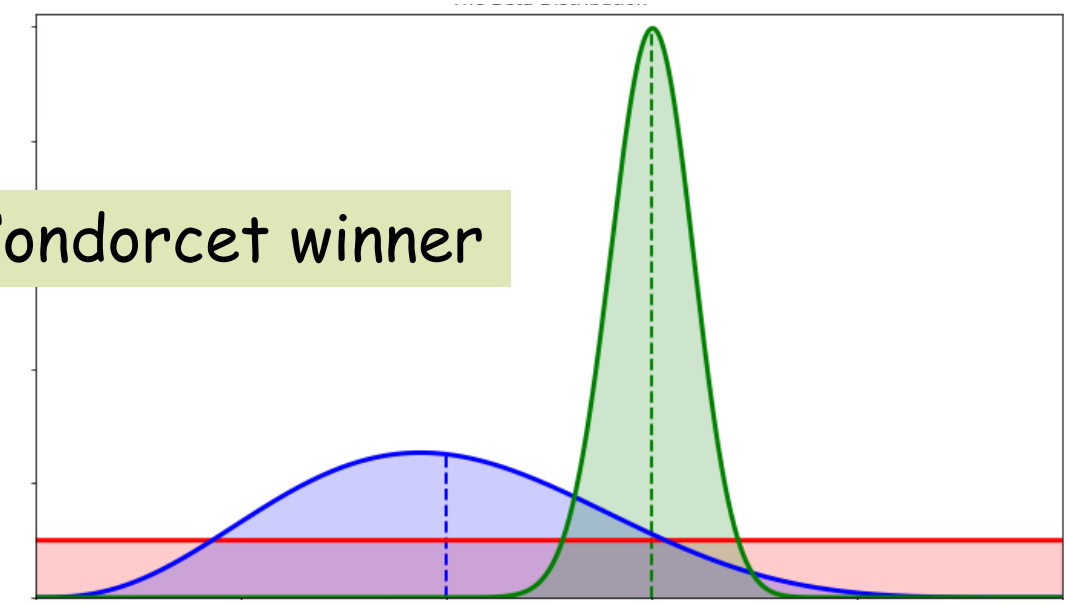
Stochastic K-armed DB (against Fixed Benchmark)

At round $t = 1,2, \ldots T$

	1	2	3	4	5
1	0.5	0.53	0.54	0.56	0.6
2	0.47	0.5	0.93	0.98	0.91
3	0.46	0.07	0.5	0.54	0.57
4	0.44	0.02	0.46	0.5	0.51
5	0.4	0.09	0.43	0.49	0.5

P

Condorcet winner



some "good" arm

Regret: $\sum_{t=1}^{T} \frac{\left\{\left[P(i^*, x_t) - \frac{1}{2}\right] + \left[P(i^*, y_t) - \frac{1}{2}\right]\right\}}{2}$

- Existing Results:
  - Gap-dependent: $\Theta\left(\sum_{t=1}^{T} \frac{\log T}{\Delta_i}\right)$
  - Worst-case: $\Theta(\sqrt{KT})$

$\Delta_i = [\, P(CW, i) - 0.5 \,]$ – Gap of arm-i

Saha & Gupta. Optimal and Efficient Dynamic Regret Algorithms for Non-Stationary Dueling Bandits. ICML'22

Our Key Idea:
Sparring + OMD as MAB blackbox

Algorithm: Versatile DB (Best-of-Both DB)

Algorithm 3 `Versatile-DB` **(Best-of-Both DB)**

input: Arm set: $[K]$, Regularizers: $(\Psi_t)_{t=1,2,\ldots}$

init: $\widehat{L}_{i,0} \leftarrow \mathbf{0}_K$ for $i \in \{+1,-1\}$

for $t = 1, 2, \ldots$ **do**

choose $p_{i,t} = \nabla(\Psi_t + \mathcal{I}_\Delta)^*(-\widehat{L}_{i,t-1})$ ← OMD based MAB weight update[*]

For $i \in \{+1,-1\}$, sample $k_{i,t}$ from the distribution $(p_{i,t}(1), \ldots, p_{i,t}(K))$

Observe preference feedback $o_t(k_{+1,t}, k_{-1,t})$ ← Sparring Step (OMD vs OMD)

Compute $\widehat{\ell}_{i,t}(k)$ for $i \in \{+1,-1\}$ and $k \in [K]$

$$\widehat{\ell}_{i,t}(k) = \begin{cases} \ell_{i,t}(k)/p_{i,t}(k) & \text{if } k = k_{i,t} \\ 0 & \text{otherwise} \end{cases}$$

← "Estimated loss" for MAB blackboxes

update $\widehat{L}_{i,t} = \widehat{L}_{i,t-1} + \widehat{\ell}_{i,t}$

end for

[*] Zimmert & Seldin. An Optimal Algorithm for Stochastic and Adversarial Bandits . AISTATS'19, JMLR'21.

Result1: Best of Both Dueling Bandits

+

Result2: First Optimal Instance dependent bound for Condorcet-DB Regret

Best-of-Both Duels! (Bridging)

Stochastic

Adversarial

Optimal gap-dependent bound for **Condorcet Regret**

Long Standing (since 2014, Zoghi et al'14)

if i^* is the Condorcet winner

Regret:

for ALL arms $i^* \in [K]$

[Tight]

$$= O\left(\sum_{t=1}^{T} \frac{\log T}{\Delta_i}\right)$$

$$\frac{K^2}{2} + 4\sum_{i=2}^{K} \frac{(i-1)\log(KT/\delta)}{\Delta_i} \leftarrow \sum_{t=1}^{T} \frac{\left\{\left[P(i^*, x_t) - \frac{1}{2}\right] + \left[P(i^*, y_t) - \frac{1}{2}\right]\right\}}{2} \rightarrow 4\sqrt{KT}.$$

$$\Delta_i = P(i^*, i) - \frac{1}{2}$$

Saha and Gaillard. Non-Stationary Dueling Bandits. arXiv/under Submission, 2022

Result3: Dueling Bandits with Adversarial Corruption

Corrupted DB (Adversarial Corruption)

True outcome: $(o_1, o_2, \dots, o_T)$

Corrupted outcome: $(\widetilde{o_1}, \widetilde{o_2}, \dots, \widetilde{o_T})$

Corruption: $$C := \sum_{t=1}^{T} \sum_{k \neq k^{(\mathrm{CW})}} \left| o_t(k^{(\mathrm{CW})}, k) - \tilde{o}_t(k^{(\mathrm{CW})}, k) \right|$$

Our regret guarantee:

Optimal additive dependency on the "Total Corruption" C.

$$\overline{R}_T \leq \sum_{k \neq k^*} \frac{4 \log T + 12}{\Delta_k} + 4 \log T + \frac{1}{\Delta_{\min}} + \frac{3}{2}\sqrt{K} + 8 + C$$

Experiments:

Comparing against baselines

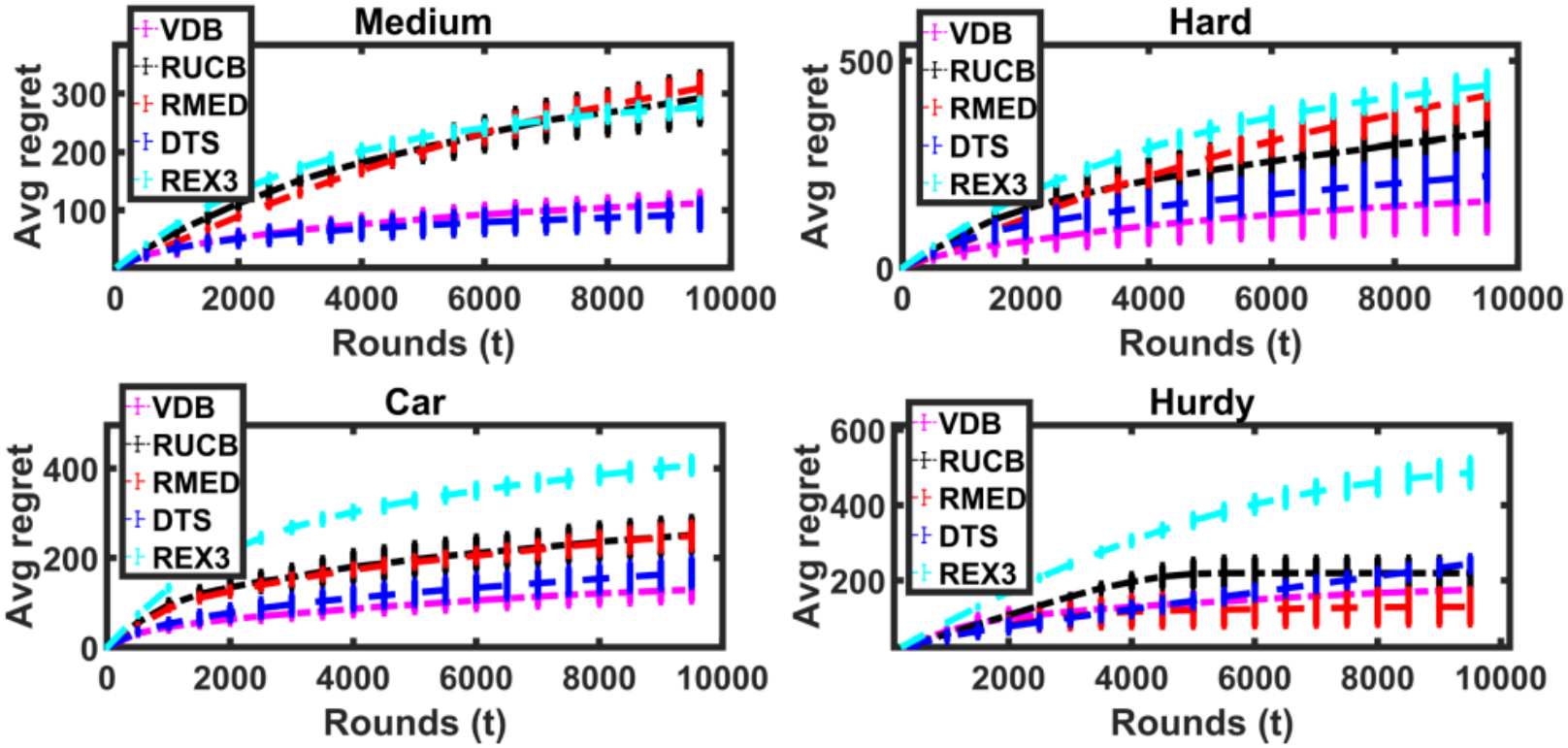


Figure 1. Averaged cumulative regret over time

Experiments:

Comparing baselines in Corrupted Preferences

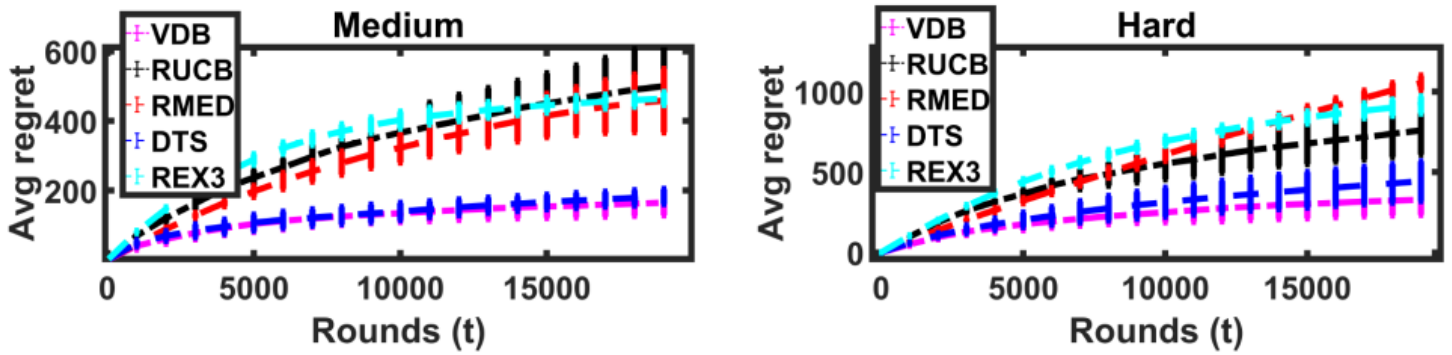


Figure 2. Averaged cumulative regret (20% corrupted feedback)

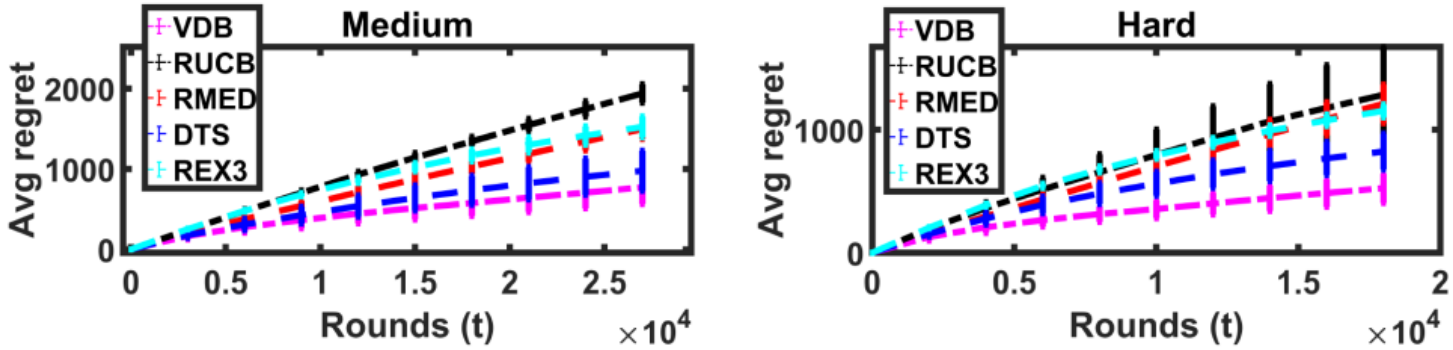


Figure 3. Averaged cumulative regret (40% corrupted feedback)

Summarizing:

- Problem formulation: Best-of-Both Dueling Bandits with optimal & worst case and instance-dependent Regret guarantee
- Upper bounds with "Sparring-OMD": $O(\sqrt{KT})$ regret for adversarial, and $O(\sum \frac{log\ T}{\Delta_i})$ regret for Condorcet Regret
- Lower bounds to justify our algorithms' optimality

Future Works:

- Generalizability? Can we solve any DB problem with its corresponding MAB counterpart?
- Extensions: Contextual Scenario, Infinite/Structured Arms, Subsetwise preferences, Feedback graphs?
- Other generalization of Dueling Bandits: Delay tolerant, Cost-sensitive, etc.

Thanks!

Questions @ aadirupa@ttic.edu