Adversarial Dueling Bandits

Aadirupa Saha[1], Tomer Koren[2], Yishay Mansour[2]
[1] Microsoft Research, New York City
[2] Blavatnik School of Computer Science, Tel Aviv University, & Google Research Tel Aviv.

38th International Conference on Machine Learning, 2021

Problem Overview: Dueling Bandits

Learning from Preferences

Absolute vs. Relative preferences

← **Ratings (Absolute)**

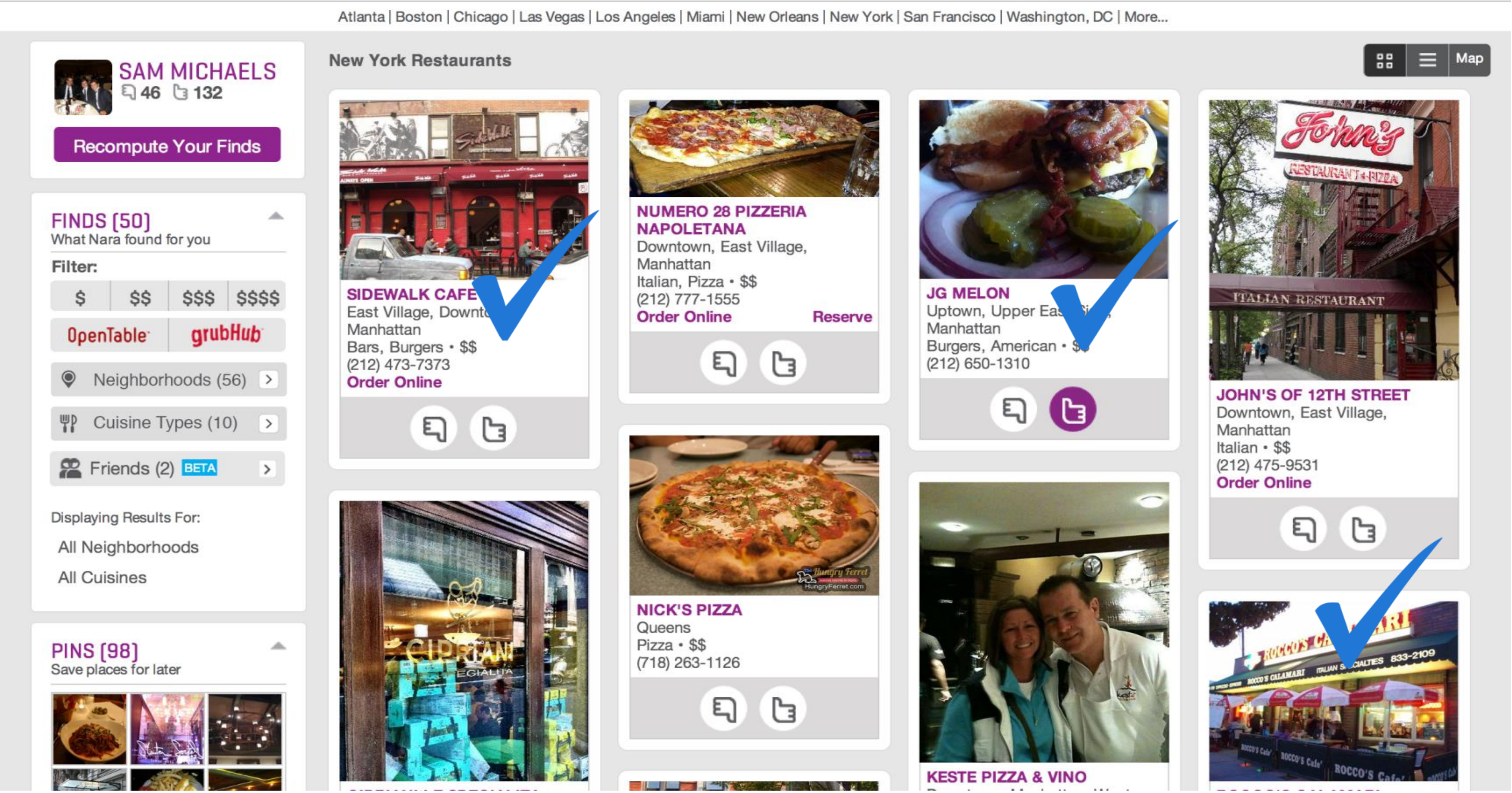
--- How much you score it out of

Rankings (Relative) →

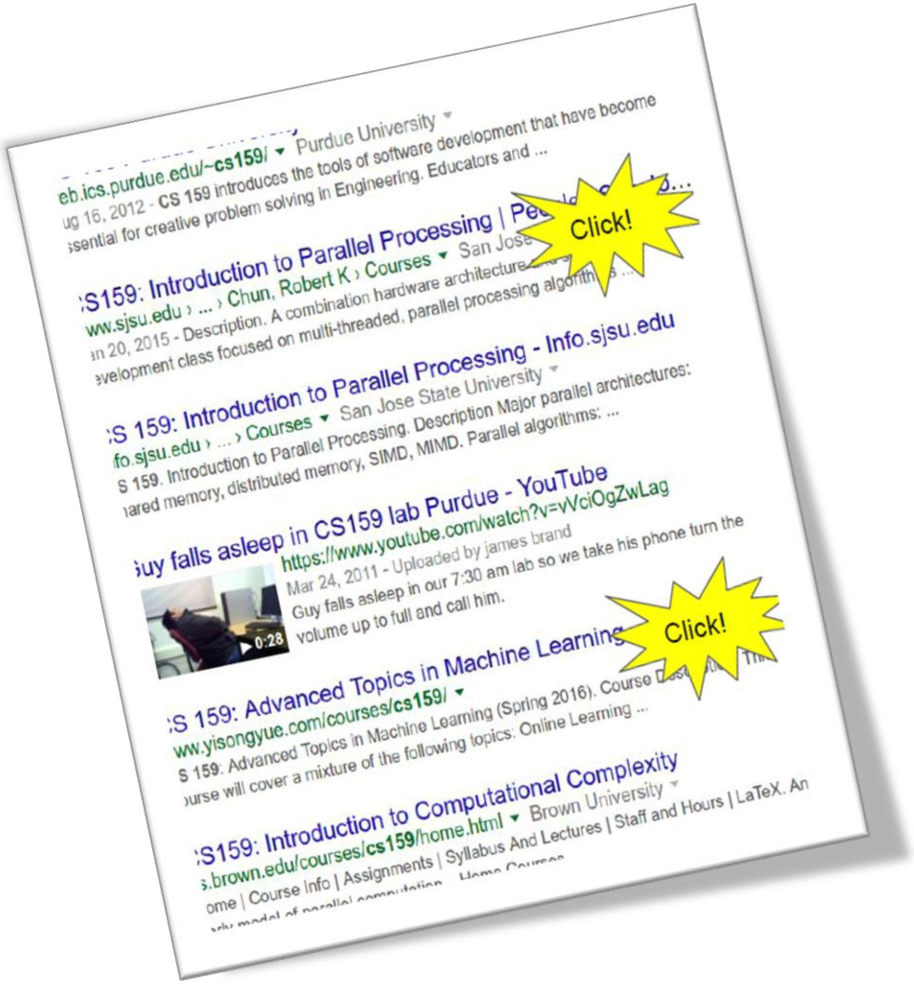
--- Do you like movie A over B?

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

Restaurant recommendation

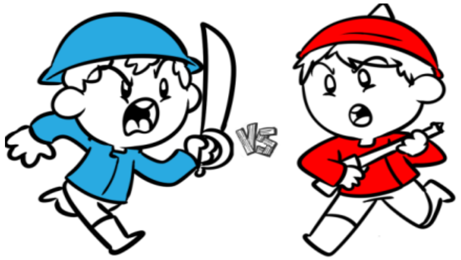


Search engine optimization:




high probability bounds
preference matrix
information retrieval application
remaining arms
optimistic estimates
random variables
nite-horizon setting
Regret bounds
relative upper confidence
Engagement Program
evaluation process
different runs
Dueling Bandit Problem
exploration phase
empirical results
pij lij uij
dealing
IFT BTMT
species
K-Armed Dueling Bandit
regular K-armed bandit
Remi Munos
higher moments
BTM and SAVAGE
Nij wij
Regular UCB
upper confidence bounds
Condorcet winner
exploration horizon
Mean Joachims
preference learning
potential champion
horizonless setting
cumulative regret
pairwise probabilities

More Formally: Dueling Bandits
(Learning from pairwise preferences)



Objective: Regret minimization (or) PAC best-arm identification.

At round t,

Select *two* arms (a_t, b_t)

Stochastic / Adversarial?

1 2 3 4 … K

repeat

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

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	0.47	0.5	0.54	0.57
4	0.44	0.42	0.46	0.5	0.51
5	0.4	0.39	0.43	0.49	0.5

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

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

Adversarial Dueling: Almost no existing works!

1. Gajane et al. *A relative exponential weighing algorithm for adversarial utility-based dueling bandits.* ICML 2015.

    - Very restricted setup of utility-based preferences?

2. Dudik et al. *Contextual Dueling Bandits.* COLT 2015.

    - 1) Contextual scenario, von-Neumann winner. 2) No efficient optimal regret algorithms

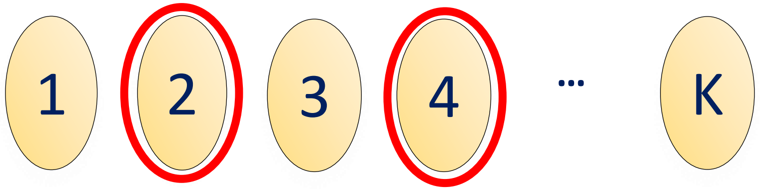
Challenges

- Assumptions on $\boldsymbol{P_t}$?
- Notion of bench-mark (best arm) to measure regret ?
- Only 1 bit feedback per $\boldsymbol{P_t}$!

Our Problem Setup

Problem Setup:

Adversarial Dueling Bandits: Sequential games of T rounds



Finite Action/arm space:
$[K] = \{1,2,\ldots K\}$

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

- Environment chooses P_t
- Play duel $(x_t,y_t) \in [K]x[K]$
- Receive feedback $o_t \sim Ber[P_t(x_t,y_t)]$

End

Regret objective:

Regret w.r.t. cumulative Borda-winner

Borda Regret $R_T := \sum_{t=1}^{T} b_t(i^*) - \frac{1}{2}(b_t(x_t) + b_t(y_t))$

where, Borda-score of Item-i: $b_t(i) := \frac{1}{K-1}\sum_{j \neq i} P_t(i,j)$

and, cumulative Borda-winner: $i^* := \arg\max_{i \in [K]} \sum_{t=1}^{T} b_t(i)$

Summary of Results

Our results:

Lower Bound

(A trivial lower bound from MAB)

$$\Omega(K/\Delta \log T)$$

- Oure results: $\Omega(\min(\Delta T, K/\Delta^2))$
  - Gap-dependent: $\Omega(K/\Delta^2)$
  - Worst-case: $\Omega(K^{1/3}T^{2/3})$

Our upper bounds

- Expected regret:

$$\mathbf{E}[R_T] \leq 6(K \log K)^{1/3} T^{2/3}$$

- (1-δ)-High probability regret:

$$R_T = \tilde{O}(K^{1/3}T^{2/3})$$

- Fixed-gap setting:

$$O((K/\Delta^2) \log(2KT/\delta))$$

$$\Delta = \min_{i \in [K] \setminus \{i^*\}} \Delta_i \quad \text{where} \quad \Delta_i = b(i^*) - b(i)$$

Lower Bound

Hard instance constructions:

Indistinguishable!

Borda-winner

$$P_1 = \begin{array}{c} \begin{bmatrix} 0.5 & \dots & 0.5 & 0.9+\epsilon & \dots & 0.9+\epsilon \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.5 & \dots & 0.5 & 0.9 & \dots & 0.9 \\ 0.1-\epsilon & \dots & 0.1 & 0.5 & \dots & 0.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.1-\epsilon & \dots & 0.1 & 0.5 & \dots & 0.5 \end{bmatrix} \end{array} \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \text{(good arms)} \\ \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \text{(bad arms)} \end{array}$$

Needs to be explored
for (1/ϵ)^2 times!

- Oure results:
  - Gap-dependent: $\Omega(K/\Delta^2)$
  - Worst-case: $\Omega(K^{1/3}T^{2/3})$

Proposed Algorithm
Regret Upper Bound

Algorithmic ideas:

Algorithm 1 **Dueling-EXP3 (D-EXP3)**

1: **Input:** Item set indexed by $[K]$, learning rate $\eta > 0$, parameters $\gamma \in (0, 1)$
2: **Initialize:** Initial probability distribution $q_1(i) = 1/K, \ \forall i \in [K]$
3: **for** $t = 1, \ldots, T$ **do**
4: Sample $x_t, y_t \sim q_t$ i.i.d. (with replacement)
5: Receive preference $o_t(x_t, y_t) \sim \mathrm{Ber}(P_t(x_t, y_t))$
6: Estimate scores, for all $i \in [K]$:

$$\boxed{\tilde{s}_t(i) = \frac{\mathbf{1}(x_t = i)}{K q_t(i)} \sum_{j \in [K]} \frac{\mathbf{1}(y_t = j) o_t(x_t, y_t)}{q_t(j)}}$$

7: Update, for all $i \in [K]$:

$$\tilde{q}_{t+1}(i) = \frac{\exp(\eta \sum_{\tau=1}^{t} \tilde{s}_\tau(i))}{\sum_{j=1}^{K} \exp(\eta \sum_{\tau=1}^{t} \tilde{s}_\tau(j))} \quad ; \quad q_{t+1}(i) = (1 - \gamma)\tilde{q}_{t+1}(i) + \frac{\gamma}{K}$$

8: **end for**

Unbiased estimate of Borda-score from 1 bit feedback! $\mathbf{E}[\tilde{s}_t(i)] = s_t(i)$

Experiments:

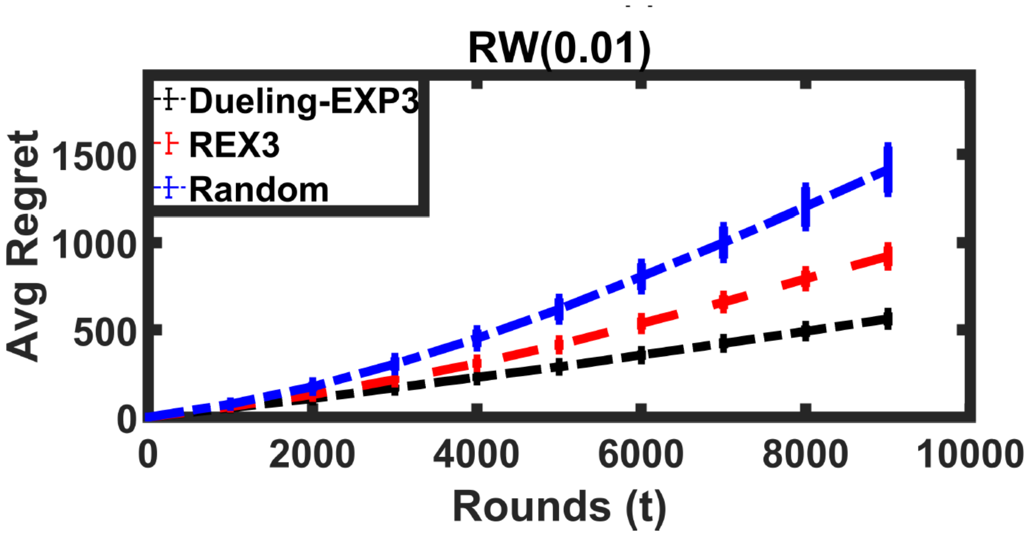
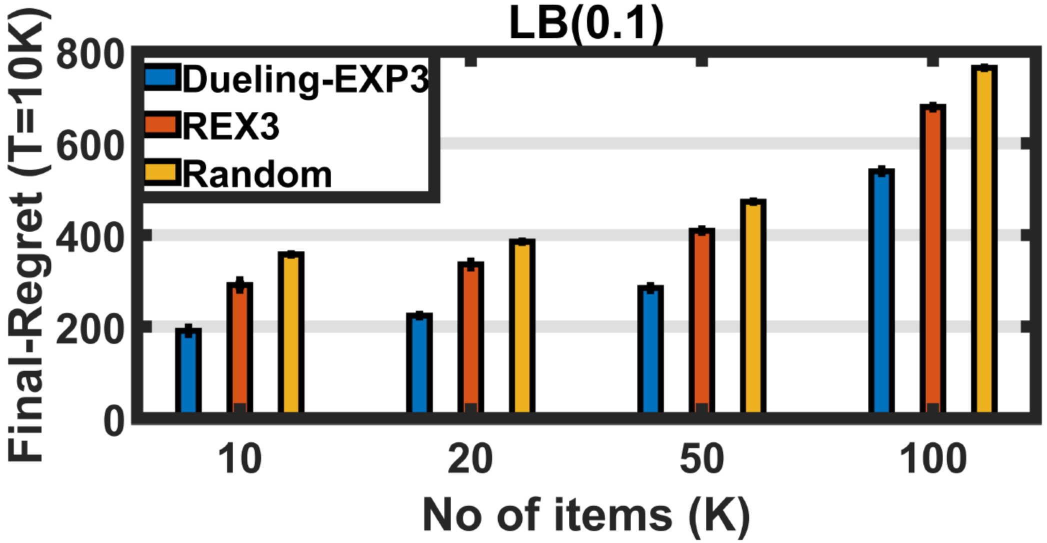


Figure 1. Averaged cumulative regret over time



In a nutshell:

- Problem formulation: Adversarial Dueling Bandits with Borda regret
- Upper bound algorithm: (Expected + High probability + Gap-dependent) regret
- Lower bound justifies tightness and algorithm's optimality

Future Works:

- Other notions of winners: Cordorcet, von-Neumann etc.
- Better rates? Under what assumptions we can attain $\Theta(\sqrt{KT})$?
- Extending dueling-bandits: Feedback graphs? General side information/ partial monitoring games?

Thanks!

Questions @ aasa@microsoft.com