Think Before You Duel: Understanding Complexities of Preference Learning under Constrained Resources.

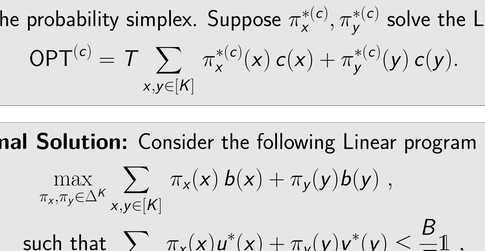
Rohan Deb [1] Aadirupa Saha [2] Arindam Banerjee [1]

[1]University of Illinois, Urbana-Champaign, USA [2]Apple

Dueling Bandits

Dueling Bandits:

- At each round $t \in [T]$ the learner picks two items i, j from K arms, and receives the output of a duel $o_t \sim \text{Ber}(P(x_t, y_t))$:



- **Preference Matrix:** $P(x_t, y_t)$ is the probability of x_t being preferred over y_t.

Condorcet Winner: Condorcet winner $x^{(c)}$ as the arm that is preferred over all the other arms, i.e., $x^{(c)} = i$ iff $P(i,j) > 1/2,\ \forall j \in [K] \setminus \{i\}$. Further the Condorcet score of arm x is defined as $c(x) = P(x, x^{(c)})$.

Borda Winner: *Borda score* of an arm $x \in [K]$ as $b(x) = \frac{1}{K-1}\sum_{x \neq y} P(x,y)$. The *Borda winner* $x^{(b)}$ is defined as the arm that maximizes the Borda score, i.e., $x^{(b)} = \arg\max_x b(x)$.

Constrained Dueling Bandits

- At round t the learner selects two arms $x_t, y_t \in [K]$, it also observes two consumption vectors $u(x_t), v(y_t) \in [0,1]^d$ with $u^*(x) = \mathbb{E}[u(x)]$, and $v^*(x) = \mathbb{E}[v(x)]$.

Condorcet Optimal Solution: Consider the following Linear program (LP) :

$$\max_{\pi_x, \pi_y \in \Delta^K} \sum_{x,y \in [K]} \pi_x(x)\, c(x) + \pi_y(y) c(y)\ ,$$

$$\text{such that} \sum_{x,y \in [K]} \pi_x(x) u^*(x) + \pi_y(y) v^*(y) \leq \frac{B}{T}\mathbb{1}\ ,$$

where Δ^K is the probability simplex. Suppose $\pi_x^{*(c)}, \pi_y^{*(c)}$ solve the LP and define:

$$\text{OPT}^{(c)} = T \sum_{x,y \in [K]} \pi_x^{*(c)}(x)\, c(x) + \pi_y^{*(c)}(y)\, c(y).$$

Borda Optimal Solution: Consider the following Linear program (LP) :

$$\max_{\pi_x, \pi_y \in \Delta^K} \sum_{x,y \in [K]} \pi_x(x)\, b(x) + \pi_y(y) b(y)\ ,$$

$$\text{such that} \sum_{x,y \in [K]} \pi_x(x) u^*(x) + \pi_y(y) v^*(y) \leq \frac{B}{T}\mathbb{1}\ ,$$

where Δ^K is the probability simplex. Suppose $\pi_x^{*(b)}, \pi_y^{*(b)}$ solve the LP and define:

$$\text{OPT}^{(b)} = T \sum_{x,y \in [K]} \pi_x^{*(b)}(x)\, b(x) + \pi_y^{*(b)}(y)\, b(y).$$

Lower Bounds

1. **Condorcet Regret.** We define the Condorcet regret as

$$\text{REG}^{(c)}(T) = \text{OPT}^{(c)} - \mathbb{E}\sum_{t=1}^{\tau} c(x_t) + c(y_t)$$

2. **Borda Regret.** We define the Borda regret as

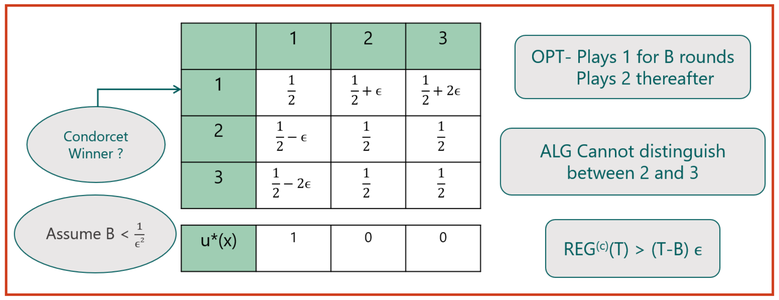
$$\text{REG}^{(b)}(T) = \text{OPT}^{(b)} - \mathbb{E}\sum_{t=1}^{\tau} b(x_t) + b(y_t)$$

Lemma 1: Define the minimum gap in Condorcet scores $\epsilon_{\min}^{(c)} := \min_{i,j \in [K]} (|c(i) - c(j)|)$. Suppose the budget $B = o\left(\frac{K}{\epsilon_{\min}^{(c)\,2}}\right)$. Then there exists a preference matrix P such that

$$\text{REG}^{(c)}(T) = \Omega(T).$$

Further our $\Omega(T)$ regret bound exists even when P satisfies total ordering.

Proof outline:



Algorithm

Algorithm 1 `Vigilant D-EXP3`

Input: $\eta_w > 0$, $\gamma_w \in (0,1)$, and $\mathcal{O}(\frac{\text{OPT}_w^{(b)}}{B}) \leq Z_w \leq \mathcal{O}(\frac{\text{OPT}_w^{(b)}}{B} + 1)$, $w \in \{x, y\}$
Initialize: $q_1^x(i) = q_1^y(i) = 1/K,\ \forall i \in [K]$
for $t = 1, \ldots, T$ **do**
  Sample $x_t \sim q_t^x, y_t \sim q_t^y$.
  Observe $o_t(x_t, y_t) \sim \text{Ber}(P_t(x_t, y_t))$ and $u_t(x_t),\ v_t(x_t)$.
  Estimate the Lagrangians $\forall\ i \in [K]$

$$\hat{\ell}_t^x(i) = \hat{b}_t(i) + Z_x \lambda_t^{x\top}\left[\frac{B}{2T}\mathbb{1} - \hat{u}_t^x(i)\right],$$

$$\hat{\ell}_t^y(i) = \hat{b}_t(i) + Z_y \lambda_t^{y\top}\left[\frac{B}{2T}\mathbb{1} - \hat{v}_t(i)\right]$$

  Update each arm using EXP-3 for all $i \in [K]$:
  Update λ_t^x and λ_t^y using any online convex optimization on:

$$g_t^x(\lambda) = \left\langle \frac{B}{2T}\mathbb{1} - \hat{u}_t^x(x_t) \right\rangle,\ g_t^y(\lambda) = \left\langle \frac{B}{2T}\mathbb{1} - \hat{v}_t^x(y_t) \right\rangle.$$

Regret Bound

Theorem 1 Assume $B = \mathcal{O}\left(\max\{\frac{K}{\epsilon_{\min}^2}, T^{3/4}\}\right)$. For $\eta_x = \left(\frac{\log K}{T\sqrt{K}}\right)^{2/3}\frac{1}{2Z_x+1}$, $\eta_y = \left(\frac{\log K}{T\sqrt{K}}\right)^{2/3}\frac{1}{2Z_y+1}$ and $\gamma_x = \sqrt{\eta_x K Z_x}, \gamma_y = \sqrt{\eta_y K Z_y}$, the regret of `Vigilant D-EXP3` is bounded by

$$\text{REG}^{(b)}(T) \leq \tilde{\mathcal{O}}\left(\left(\frac{\text{OPT}^{(b)}}{B} + 1\right)(K \log K)^{1/3} T^{2/3}\right)$$

Experiments

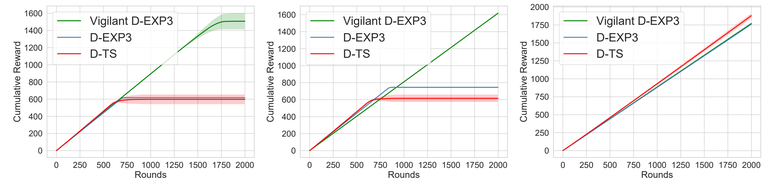


Figure 1. Cumulative Reward across Rounds on synthetic data for three choices of consumptions.

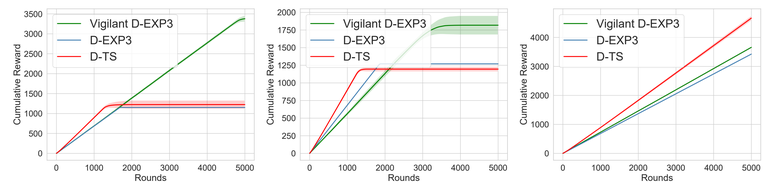


Figure 2. Cumulative Reward across Rounds on car preference dataset for three choices of consumptions.

Conclusion

- We setup the *Constrained Dueling Bandits* problem.
- We show that the 'relative' nature of feedback makes Constrained-DB a difficult problem to solve (by providing lower bounds).
- Under assumptions on the available budget, we provide an EXP3 based algorithm `Vigilant D-EXP3` that attains sub-linear regret.
- Future work: Condorcet setting, extension to contextual bandits, beyond the linear setting, Thompson sampling version

Acknowledgement

The work was supported in part by grants from the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award.