The K-armed Dueling Bandits Problem

Yisong Yue, Joseph Broder, Bobby Kleinberg, Thorsten Joachims

Department of Computer Science
Cornell University
Adaptive Information Systems

- **Retrieval Function:** \( f(q) \rightarrow r \)
  - Input: \( q \) (query)
  - Output: \( r \) (ranking by relevance)

- **Conventional Systems**
  - One-size-fits-all
  - Hand-tuned and static retrieval function

- **Room for Improvement**
  - Different users need different retrieval functions
  - Different collections need different retrieval functions

- **Machine Learning**
  - Learn improved retrieval functions
Motivation and Outline

• Setup
  – Corpus of documents [known]
  – Distribution of users and/or queries on corpus [unknown]
  – Set of retrieval functions \{f_1, \ldots, f_K\} [design choice]
  – Each retrieval function \( f_i \) has utility \( U(f_i) \) [unknown]

• Question 1: How can one measure utility?
  – Cardinal vs. ordinal utility measurements
  – Eliciting implicit feedback through interactive experiments

• Question 2: How to efficiently find \( f_i \) with max utility?
  – Efficiently \( \rightarrow \) minimizing regret + computationally efficient
  – Minimize exposure to suboptimal results during learning
  – Dueling Bandits Problem with efficient algorithm
Approaches to Implicit Utility Elicitation

• **Approach 1: Absolute Metrics (cardinal)**
  - Do metrics derived from observed user behavior provide absolute feedback about retrieval quality of f?
  - For example:
    • $U(f) \sim \text{numClicks}(f)$
    • $U(f) \sim 1/\text{abandonment}(f)$

• **Approach 2: Paired Comparison Tests (ordinal)**
  - Do paired comparison tests provide relative preferences between two retrieval functions $f_1$ and $f_2$?
  - For example:
    • $f_1 \succ f_2 \Leftrightarrow \text{pairedCompTest}(f_1, f_2) > 0$
Interleaving(r₁, r₂)

Interpretation: (r₁ > r₂) ↔ clicks(topk(r₁)) > clicks(topk(r₂))

Invariant: For all k, top k of balanced interleaving is union of top k₁ of r₁ and top k₂ of r₂ with k₁=k₂±1.

[Joachims/01]
Balanced Interleaving: Results

Paired Comparison Tests: Summary and Conclusions

- All interleaving experiments reflect the expected order.
- All differences are significant after one month of data.
- Analogous results for Yahoo Search and Bing.
- Low impact (always some good results).
Motivation and Outline

• Setup
  – Corpus of documents [known]
  – Distribution of users and/or queries on corpus [unknown]
  – Set of retrieval functions \{f_1, \ldots, f_K\} [design choice]
  – Each retrieval function $f_i$ has utility $U(f_i)$ [unknown]

• Question 1: How can one measure utility?
  – Cardinal vs. ordinal utility measurements
  – Eliciting implicit feedback through interactive experiments

• Question 2: How to efficiently find $f_i$ with max utility?
  – Efficiently $\Rightarrow$ minimizing regret + computationally efficient
  – Minimize exposure to suboptimal results during learning
  – Dueling Bandits Problem with efficient algorithm
Evaluating Many Retrieval Functions

Task:
Find $f^* \in F$ that gives best retrieval quality over $P(u,q)$?
Tournament

- Can you design a tournament that reliably identifies the correct winner?

  → Noisy Sorting/Max Algorithms:
    - [Feige et al.]: Triangle Tournament Heap $O(n/\varepsilon^2 \log(1/\delta))$ with prob $1-\delta$
    - [Adler et al., Karp & Kleinberg]: optimal under weaker assumptions

Diagram: A series of circles connected by lines, representing the tournament structure.
Problem: Learning on Operational System

• Example:
  – 4 retrieval functions: B > G >> Y > A
  – 10 possible pairs for interactive experiment
    • (B,G)  low cost to user
    • (B,Y)  medium cost to user
    • (Y,A)  high cost to user
    • (B,B)  zero cost to user
    • …

• Minimizing Regret
  – Algorithm gets to decide on the sequence of pairwise tests
  – Don’t present “bad” pairs more often than necessary
  – Trade off (long term) informativeness and (short term) cost

⇒ Dueling Bandits Problem
Regret for the Dueling Bandits Problem

• Given:
  – A finite set $H$ of candidate retrieval functions $f_1 \ldots f_K$
  – A pairwise comparison test $f \succ f'$ on $H$ with $P(f \succ f')$

• Regret:
  – $R(A) = \sum_{t=1..T} [P(f^* \succ f_t) + P(f^* \succ f'_t) - 1]$
  – $f^*$: best retrieval function in hindsight (assume single $f^*$ exists)
  – $(f,f')$: retrieval functions tested at time $t$

Example:

<table>
<thead>
<tr>
<th>Time Step</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>...</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>$(f_9,f_{12}) \rightarrow f_9$</td>
<td>$(f_5,f_9) \rightarrow f_5$</td>
<td>$(f_1,f_3) \rightarrow f_3$</td>
<td></td>
</tr>
<tr>
<td>Regret</td>
<td>$P(f^* \succ f_9) + P(f^* \succ f_{12}) - 1 = 0.9$</td>
<td>$P(f^* \succ f_5) + P(f^* \succ f_9) - 1 = 0.78$</td>
<td>$= 0.01$</td>
<td></td>
</tr>
</tbody>
</table>

[with Yisong Yue, Josef Broder, Bobby Kleinberg]
Tournament

• Can you design a tournament that has low regret?
  ➔ Don’t know!
Algorithm: Interleaved Filter 1

- **Algorithm**

InterleavedFilter1(T, W={f₁…fₖ})

- Pick random f' from W
- δ = 1/(TK²)
- WHILE |W| > 1
  - FOR f ∈ W DO
    » duel(f', f)
    » update Pₙ
  - t = t + 1
  - cₜ = (log(1/δ)/t)⁰.⁵
  - Remove all f from W with Pₙ < 0.5 - cₜ [WORSE WITH PROB 1 - δ]
  - IF there exists f'' with Pₙ > 0.5 + cₜ [BETTER WITH PROB 1 - δ]
    » Remove f' from W
    » f' = f''; t = 0
- UNTIL T: duel(f', f')

[with Yisong Yue, Josef Broder, Bobby Kleinberg]
IF1: Main Result

• **Theorem:** The expected regret of IF1 is

\[ E[R_T] = O \left( \frac{K \log K}{\epsilon_{1,2} \log T} \right) \]

where \( \epsilon_{1,2} = P(f_1 > f_2) - 0.5 \) and K is the number of bandits.

• **Assumptions:**
  - Strong Stochastic Transitivity: \( \epsilon_{i,k} \geq \max\{\epsilon_{i,j}, \epsilon_{j,k}\} \)
  - Stochastic Triangle Inequality: \( \epsilon_{i,k} \leq \epsilon_{i,j} + \epsilon_{j,k} \)
  - \( \epsilon \)-winner exists
Assumptions

• Preference Relation: \( f_i \succ f_j \iff P(f_i \text{ beats } f_j) = 0.5 + \varepsilon_{i,j} > 0.5 \)

• Weak Stochastic Transitivity: \( f_i \succ f_j \) and \( f_j \succ f_k \) \( \Rightarrow f_i \succ f_k \)

\[
\begin{align*}
    f_1 & \succ f_2 \succ f_3 \succ f_4 \succ f_5 \succ f_6 \succ \ldots \succ f_K
\end{align*}
\]

• Strong Stochastic Transitivity: \( \varepsilon_{i,k} \geq \max\{\varepsilon_{i,j}, \varepsilon_{j,k}\} \)

\[
\begin{align*}
    \varepsilon_{1,4} & \geq \varepsilon_{2,4} \geq \varepsilon_{3,4} \geq 0.5 \geq \varepsilon_{5,4} \geq \varepsilon_{6,4} \geq \ldots \geq \varepsilon_{K,4}
\end{align*}
\]

• Stochastic Triangle Inequality: \( f_i \succ f_j \succ f_k \) \( \Rightarrow \varepsilon_{i,k} \leq \varepsilon_{i,j} + \varepsilon_{j,k} \)

\[
\begin{align*}
    \varepsilon_{1,2} & = 0.01 \text{ and } \varepsilon_{2,3} = 0.01 \Rightarrow \varepsilon_{1,3} \leq 0.02
\end{align*}
\]

• \( \varepsilon \)-Winner exists: \( \varepsilon = \max_i \{ P(f_1 \text{ beats } f_i) - 0.5 \} = \varepsilon_{1,2} > 0 \)
IF1: Proof Outline

$$E[R_T] \leq \left(1 - \frac{1}{T}\right) E[R_T^{IF1}] + \frac{1}{T} O(T) = O(E[R_T^{IF1}])$$

1. The probability that IF1 returns suboptimal bandit is less than 1/T
   a) Probability that a match has wrong winner is at most $\delta=1/(T K^2)$. 
   b) Upper bound on the number of matches: $K^2$

2. Bound expected regret $E[R_T^{IF1}]$ of IF1
   a) Bound number of duels in a match: $O(1/\epsilon^2)$
   b) Bound regret per match
   c) Bound the number of rounds before IF1 terminates
Lemma 1a: Probability that a Match has Wrong Winner is at most $\delta = 1/(T K^2)$

- **Proof:**
  - Reminder: Confidence interval $c_t = (\log(1/\delta)/t)^{0.5}$
  - If we declare the wrong winner between $f_i$ and $f_j$, then observed $P_t$ must have been outside confidence interval.
  - $P(|P_t - E[P_t]| \geq c_t) \leq 2 \exp(-2tc_t^2) = 2\delta^2 = 2/(T^2 K^4)$
  - Union bound over all time steps: $2T/(T^2 K^4) \leq 1/(T K^2) = \delta$
IF1: Proof Outline

\[ E[R_T] \leq \left(1 - \frac{1}{T}\right) E[R_T^{IF1}] + \frac{1}{T}O(T) = O(E[R_T^{IF1}]) \]

1. The probability that IF1 returns suboptimal bandit is less than 1/T
   a) Probability that a match has wrong winner is at most \( \delta = 1/(T K^2) \).
   b) Upper bound on the number of matches: \( K^2 \)

2. Bound expected regret \( E[R_T^{IF1}] \) of IF1
   a) Bound number of duels in a match
   b) Bound regret per match
   c) Bound the number of rounds before IF1 terminates
Lemma 2a: Bound Number of Duels in a Match

- Consider: match between $f_i$ and $f_j$ with $P(f_i$ beats $f_j) = 0.5 + \varepsilon_{i,j}$
- If match is $t$ duels long, then $P_t - c_t \leq 0.5$, otherwise the match would terminate.
- $P(n > t) \leq P(P_t - c_t \leq 0.5) = P(E[P_t] - P_t \geq \varepsilon_{i,j} - c_t)$
- For any $m \geq 4$ and $t = (m \log(TK^2)/\varepsilon_{i,j}^2)$, we have $c_t \leq 0.5 \varepsilon_{i,j}^2$.
- Hoeffding bound $\Rightarrow O(1/\varepsilon_{i,j}^2 \log(TK)) \ $whp
IF1: Proof Outline

\[
E[R_T] \leq \left(1 - \frac{1}{T}\right) E[R_T^{IF1}] + \frac{1}{T} O(T) = O(E[R_T^{IF1}])
\]

1. The probability that IF1 returns suboptimal bandit is less than 1/T
   a) Probability that a match has wrong winner is at most \( \delta = 1/(T K^2) \).
   b) Upper bound on the number of matches: \( K^2 \)

2. Bound expected regret \( E[R_T^{IF1}] \) of IF1
   a) Bound number of duels in a match: \( O(1/\varepsilon_{i,j}^2 \log(TK)) \) whp
   b) Bound regret per match
   c) Bound the number of rounds before IF1 terminates
Lemma 2b: Bound Regret per Match

• Proof:
  – Let current incumbent \( f' = f_j \):
  
  \[
  \text{whp (Lemma 2a)}
  \]
  
  – Each duel(\( f_i, f_j \)) incurs (\( \epsilon_{1,j} + \epsilon_{1,i} \)) regret:

  ________
  – Case \( f_i \gtrless f_j \): Then \( \epsilon_{1,j} + \epsilon_{1,i} \leq 2 \epsilon_{1,j} \) (SST) and regret is bounded
  
  \[
  2\epsilon_{1,j} \cdot O(1/\epsilon^2_{1,j} \log(TK)) = O(1/\epsilon_{1,j} \log(TK)) \leq O(1/\epsilon_{1,2} \log(TK))
  \]

  ________
  – Case \( f_i \prec f_j \) and \( \epsilon_{j,i} \leq \epsilon_{1,j} \):
  
  \[
  3\epsilon_{1,j} \cdot O(1/\epsilon^2_{1,j} \log(TK)) = O(1/\epsilon_{1,j} \log(TK)) \leq O(1/\epsilon_{1,2} \log(TK))
  \]

  ________
  – Case \( f_i \prec f_j \) and \( \epsilon_{j,i} > \epsilon_{1,j} \):
  
  \[
  3\epsilon_{j,i} \cdot O(1/\epsilon^2_{j,i} \log(TK)) = O(1/\epsilon_{1,j} \log(TK)) \leq O(1/\epsilon_{1,2} \log(TK))
  \]
IF1: Proof Outline

\[ E[R_T] \leq \left(1 - \frac{1}{T}\right) E[R_{T}^{IF1}] + \frac{1}{T}O(T) = O(E[R_{T}^{IF1}]) \]

1. **The probability that IF1 returns suboptimal bandit is less than 1/T**
   a) Probability that a match has wrong winner is at most \( \delta = 1/(T K^2) \).
   b) Upper bound on the number of matches: \( K^2 \)

2. **Bound expected regret \( E[R_{T}^{IF1}] \) of IF1**
   a) Bound number of duels in a match: \( O(1/\epsilon_{i,j}^2 \log(TK)) \) whp
   b) Bound regret per match: \( O(1/\epsilon_{1,2} \log(TK)) \) whp
   c) Bound the number of rounds before IF1 terminates
Lemma 2c: Bound the Number of Rounds before IF1 Terminates

- Random walk: \( X_i = 1 \) if \( f_i \) becomes incumbent, \( X_i = 0 \) else

\[ \sum X_i = \text{number of steps in random walk} = \text{number of rounds} \]

- Note: If IF1 does not make a mistake, then only forward steps.

- Strong Stochastic Transitivity: \( \forall i: p_{i+1,1} \geq p_{i+1,2} \geq \ldots \geq p_{i+1,i} \)

\[ \Rightarrow \text{Worst case: } p_{i+1,1} = p_{i+1,2} = \ldots = p_{i+1,i} = 1/i \]

\[ \sum X_i = O(\log K) \text{ rounds whp} \]
IF1: Proof Outline

\[ E[R_T] \leq \left(1 - \frac{1}{T}\right) E[R_T^{IF1}] + \frac{1}{T} O(T) = O(E[R_T^{IF1}]) \]

1. Theorem: IF1 incurs expected regret bounded by

\[ E(R_T) \leq O \left(\frac{K \log K}{\varepsilon_{1,2}} \log T\right) \]

2. Bound expected regret \( E[R_T^{IF1}] \) of IF1
   a) Bound number of duels in a match: \( O(1/\varepsilon_{i,j}^2 \log(TK)) \) whp ✔
   b) Bound regret per match: \( O(1/\varepsilon_{1,2} \log(TK)) \) whp ✔
   c) Bound rounds before IF1 terminates: \( O(\log K) \) whp ✔
Lower Bound

- **Theorem:** Any algorithm for the dueling bandits problem has regret

\[ R_T \leq \Omega \left( \frac{K}{\epsilon_{1,2} \log T} \right) \]

- **Proof:** [Karp/Kleinberg/07][Kleinberg/etal/07]

- **Intuition:**
  - Magically guess the best bandit, just verify guess
  - Worst case: \( \forall f_i > f_j: P(f_i > f_j) = 0.5 + \epsilon \)
  - Lemma 2a: Need \( O(1/\epsilon^2 \log T) \) duels to get 1-1/T confidence.
**Algorithm: Interleaved Filter 2**

- **Algorithm**
  
  InterleavedFilter1(T,W={f₁…fₖ})
  
  - Pick random f' from W
  - \( \delta = 1/(TK^2) \)
  - WHILE \(|W| > 1\)
    - FOR b ∈ W DO
      » duel(f',f)
      » update \( P_f \)
    - t=t+1
    - \( c_t = (\log(1/\delta)/t)^{0.5} \)
    - Remove all f from W with \( P_f < 0.5 - c_t \) [WORSE WITH PROB 1-\( \delta \)]
    - IF there exists f'' with \( P_{f''} > 0.5 + c_t \) [BETTER WITH PROB 1-\( \delta \)]
      » Remove f' from W
    » Remove all f from W that are empirically inferior to f'
    » f'=f''; t=0
  - UNTIL T: duel(f',f')

---

[with Yisong Yue, Josef Broder, Bobby Kleinberg]
Why is it Safe to Remove Empirically Inferior Bandits?

- **Lemma:** Mistakenly pruning a bandit has probability at most \( \delta = 1/(T K^2) \).

- **Proof:**
  - Mistake: \( f_p \succ f_w \succ f_i \) (pruned: \( f_p \), winner: \( f_w \), incumbent: \( f_i \))
  - \( B_{n,w,p} \): Given \( w \) is winner after \( n \) duels, \( f_p \) mistakenly pruned.
  - To show: \( P(B_{n,w,p}) \leq 1-\delta \) for all \( n \) and \( w \).
  - Suppose \( P(b_w \succ b_i) = \alpha \) and given \( B_{n,w,p} : P(b_p \succ b_i) \geq \alpha \).
    \[ \Rightarrow E(S_{w,i}+S_{i,p}) \leq n. \]
  - Duels won \( S_{w,i} - 0.5n < \sqrt{n \log(1/\delta)} \) and \( S_{i,p} > 0.5n \)
    \[ \Rightarrow S_{w,i} + S_{i,p} - n > \sqrt{n \log(1/\delta)} \]
  - Hoeffding \( P(S_{w,i} + S_{i,p} - n > \sqrt{n \log(1/\delta)}) \leq \delta \)
Bound the Number of Matches of IF2

- **Lemma:** Assuming IF2 is mistake free, then it plays $O(K)$ matches in expectation.
- **Intuition:**
Regret Bound for IF2

\[ E[R_T] \leq \left(1 - \frac{1}{T}\right) E[R_T^{IF1}] + \frac{1}{T} O(T) = O(E[R_T^{IF1}]) \]

- **Lemma**: Mistakenly pruning a bandit has probability at most \( \delta = 1/(T K^2) \).
- **Lemma**: Assuming IF2 is mistake free, then it plays \( O(K) \) matches in expectation.

**Theorem**: IF2 incurs expected regret bounded by

\[ E(R_T) \leq O \left( \frac{K}{\epsilon_{1,2}} \log T \right) \]
Experiments: Synthetic Data

- **Lower-Bound data:** $\forall f_i \succ f_j: P(f_i \succ f_j) = 0.5 + \varepsilon$

- **Bradley-Terry data**
Experiment: Simulated Web Search

- **Microsoft Web Search Data (Chris Burges) with manual relevance assessment**
- **Feedback** $f_i \succ f_j$:
  - Draw query at random
  - Preference $f_i \succ f_j$ (probabilistically) based on NDCG difference of rankings produced by $f_i$ and $f_j$
Why not a log-Gap?

• **To achieve log-gap:**
  - Log number of rounds need to be played
  - Most inferior bandits must not get eliminated anyway without pruning.

  ![Diagram of bandits and log-gaps](image)

• **Experiment results**
  - Typically 2-4 rounds largely independent of number of bandits
  - Many bandits much worse, so eliminated before round ends

  ![Diagram with bandits and values](image)
Summary

• **Dueling Bandits Problem**
  – Only ordinal information about payoffs
  – Algorithms proposes two alternatives, user provides noisy preference.
  – Preference can be interleaving, direct comparison, etc.

• **Interleaved Filter Algorithm**
  – Regret based on win/loss against optimal bandit
  – Strategy: keep incumbent, compare against others, prune inferior
  – $O(K/\epsilon \log T)$ regret like for bandits with absolute feedback

• **Further Question**
    • Lower variability
    • Relax strong stochastic transitivity
  – Algorithm for finite and convex sets of bandit [Yue & Joachims, 2009]