Preference-Based Multi-Armed Bandits

Kelsey Ball

The University of Texas at Austin

September 24, 2022

Agenda

- Problem Setting
- Algorithms
- Simulations
- Related/Open Problems

Problem Setting

Typical Multi-Armed Bandit (MAB) setting:

1. Play arm i
2. Observe numerical reward $X_t(i)$

Preference-based MAB setting:

1. Play two arms i, j
2. Observe (with noise) which arm is better

Motivation

- Preferential feedback (e.g., a pairwise comparison) is sometimes more readily available than scalar estimates of reward
- Examples:
    - Eye doctor examination
    - Ranker evaluation for information retrieval systems
    - TrueSkill: Xbox gamer ranking system

Problem Formulation

- Set of k arms $\mathcal{A} = \{a_1, \dots, a_k\}$
- Characterized by preference relation $Q = [q_{i,j}] \in [0,1]^{k \times k}$ where $q_{i,j}$ is the probability of observing a preference for a_i over a_j
- We say $a_i \succ a_j$ if $q_{i,j} > 1/2$
- A "tie" or indifference is modeled as $q_{i,j} = 1/2$; thus $q_{i,i} = 1/2$ for all $i \in [k]$.

How is Regret Defined?

First, define $\Delta_{i,j}$ as a notion of distinguishability between arms:

$$\Delta_{i,j} = q_{i,j} - 1/2$$

$q_{i,j}$	$\Delta_{i,j}$	Interpretation
0	-1/2	i never beats j
1/2	0	i, j indistinguishable
1	1/2	i always beats j

Note: $\Delta_{i,j} > 0$ implies $a_i \succ a_j$

How is Regret Defined?

- Main Idea: Player incurs small regret by choosing two nearly optimal arms

$$R_n = \frac{1}{2} \sum_{t=1}^{n} \Delta_{i^*,i(t)} + \Delta_{i^*,j(t)}$$

- Note: For an optimal arm $i*$,

$$\Delta_{i^*,j(t)} \geq 0$$

  so regret will be non-negative.
- Note: Regret is zero only if player compares the optimal arm to itself; i.e. commits to choice of best arm and refrains from gathering more information

Interleaved Filter (IF) [YJ09]

Overview

- Explore-then-exploit algorithm
- Explore phase: successive elimination of suboptimal arms (with high probability), until one remains
- Exploit phase: repeatedly compare best (hypothesized) arm to itself
- Expected regret bound:

$$E[R_n] = O\left(\frac{k}{\min_{j \neq i^*} \Delta_{i^*,j}} \log n\right)$$

Interleaved Filter (IF) [YJ09]

IF makes strong assumptions on underlying preference matrix Q:

- There exists a total ordering $a_1 \succ a_2 \succ \cdots \succ a_k$ such that $a_i \succ a_j \implies \Delta_{i,j} > 0$
- Strong Stochastic Transitivity (SST): for $a_i \succ a_j \succ a_k$,

$$\Delta_{i,k} \geq \max\{\Delta_{i,j}, \Delta_{j,k}\}$$

- Stochastic Triangle Inequality (diminishing returns):

$$\Delta_{i,k} \leq \Delta_{i,j} + \Delta_{j,k}$$

Trick For Bounding Explore-Exploit Algorithms

Explore-exploit algorithms can be constructed in such a way that the regret is determined solely by the explore phase:

- Show that the explore phase returns the best arm w.p. $\geq 1 - \frac{1}{n}$.
- If, instead, it returns a suboptimal arm (w.p. $\leq \frac{1}{n}$), we can upper bound the total regret by n.
- Thus,

$$\begin{aligned} E[R_n] &= \left(1 - \frac{1}{n}\right) E[R_n^{explore}] + \frac{1}{n} \cdot n \\ &= O\left(E[R_n^{explore}] + 1\right) \end{aligned}$$

Therefore the expected regret is upper bounded by the expected regret of the explore phase.

Interleaved Filter (IF) [YJ09]

Explore

- Maintain candidate best arm $\hat{b}$
- Compare $\hat{b}$ with all other arms via round robin
- Prune any arms that are inferior w.p. $1 - \delta$
- If any arm b' is superior to $\hat{b}$ w.p. $1 - \delta$, prune $\hat{b}$ from candidate set and update $\hat{b} \leftarrow b'$

Exploit

- Repeatedly play $\hat{b}$, $\hat{b}$

Interleaved Filter (IF) [YJ09]

How to prune inferior arms with high probability?

- Maintain empirical estimate $\hat{P}_{i,j}^{(t)}$ of $Pr(a_i \succ a_j)$ as fraction of wins in t comparisons
- Maintain confidence interval for $\hat{P}_{i,j}^{(t)}$:

$$\hat{C}_{i,j}^{(t)} = (\hat{P}_{i,j}^{(t)} - c, \hat{P}_{i,j}^{(t)} + c), \quad c = \sqrt{\frac{\log \frac{1}{\delta}}{t}}$$

  such that $\hat{P}_{i,j}^{(t)} \in \hat{C}_{i,j}^{(t)}$ for all t w.h.p.
- If $\hat{P}_{\hat{b},b'} > 1/2$ and $1/2 \notin \hat{C}_{\hat{b},b'}$, prune b'

Interleaved Filter (IF) [YJ09]

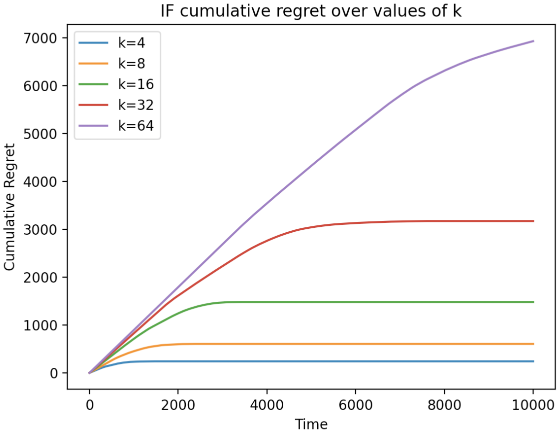

Figure: Cumulative regret over different values of k for a fixed time horizon, averaged over 10 runs.

Beat The Mean (BTM) [YJ11]

Overview

- Elimination algorithm that favors arms with the fewest comparisons and pairs them with another arm uniformly at random (the "mean" arm)
- Relaxes strong transitivity assumption: there exists some $\gamma \geq 1$ such that
$$\gamma\Delta_{i,k} \geq \max\{\Delta_{i,j}, \Delta_{j,k}\}$$
- Gives high probability bound on regret in addition to bound on expected regret. Both are of order
$$O\left(\frac{\gamma^7 k}{\min_{j\neq i^*} \Delta_{i^*,j}} \log n\right)$$
- Matches IF bound when $\gamma = 1$, i.e. strong transitivity holds

Beat The Mean (BTM) [YJ11]

Explore

In each round t, while candidate pool $|W_t| > 1$:

- Select arm b with fewest comparisons
- Select arm b' uniformly at random from W_t
- Compare b, b'
- Update $\hat{P}_b$ or $\hat{P}_{b'} := \frac{\#wins}{\#comparisons}$
- If (empirically) best and worst arm separated by large enough margin, eliminate worst and start new round

Exploit

- Let $\hat{b}$ be the unique arm in W_t. Repeatedly play $\hat{b}$, $\hat{b}$.

Beat The Mean (BTM) [YJ11]

- What margin is needed to separate empirically best and worst arms?
- If

$$\min_{b' \in W_t} \hat{P}_{b'} + c \leq \max_{b \in W_t} \hat{P}_b - c, \quad c_{\delta,\gamma}(n) = 3\gamma^2 \sqrt{\frac{1}{n} \log \frac{1}{\delta}}$$

  Then remove $\arg\min_{b' \in W_t} \hat{P}_{b'}$ from W_t.

Beat The Mean (BTM) [YJ11]

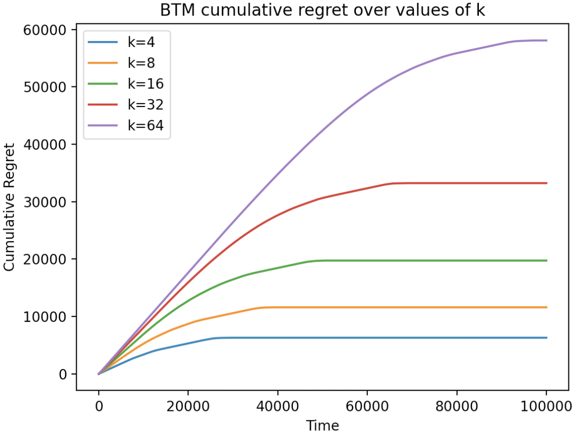

Figure: Cumulative regret over different values of k for a fixed time horizon, averaged over 10 runs.

Relative UCB (RUCB) [ZWMR14]

Main Idea:

- For the first arm, choose a hypothetical best arm
- For the second arm, choose arm with the best chance of beating the first arm

Improvements over IF/BTM:

- Horizonless (does not need knowledge of n)
- Relaxed assumptions on preference matrix (does not require total ordering, strong stochastic transitivity, or stochastic triangle inequality; only requires a best arm)

Relative UCB (RUCB) [ZWMR14]

RUCB Algorithm

In each round t:

- Get candidate set of plausible best arms i.e.:

$$W_t = \{i : \hat{q}_{i,j}(t) + c_{i,j}(t) > 1/2 \ \ \forall j \neq i\}$$

- Select one candidate arm b uniformly at random from W_t
- Use UCB to choose the other candidate arm b':

$$b' = \arg\max_{j \neq b} U_{j,b}$$

  where $U_{j,b} = \hat{q}_{j,b}(t) + c_{j,b}$
- Compare b, b' and update $\hat{q}_{b,b'}(t), \hat{q}_{b',b}(t)$

Relative UCB (RUCB) [ZWMR14]

- Expected and high probability bounds:

$$R_n \leq O\left(k^2 + \sum_{i \neq i^*} \frac{\log n}{\Delta_{i,i^*}^2}\right)$$

- Not directly comparable to IF/BTM bounds, which only depend on $\min_{j \neq i^*} \Delta_{i^*,j}$

Relative UCB (RUCB) [ZWMR14]

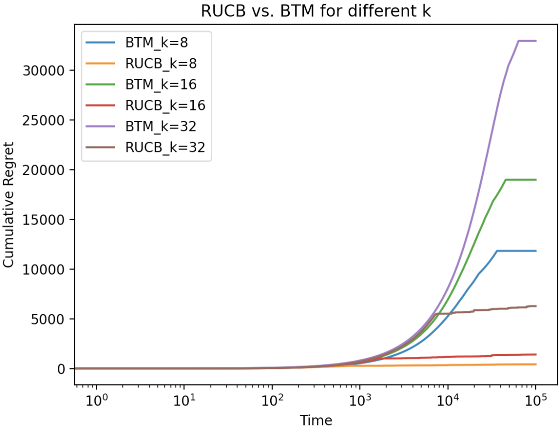

Figure: Cumulative regret for RUCB vs. BTM over different values of k, averaged over 10 runs.

Related Tasks/Problem Settings

- (ε, δ)-PAC learning: the best arm, a ranking of arms, the top-k arms, or Q.
- Non-coherent preference matrices (e.g. allow preferential cycles) – require alternative notions of regret/target concepts
- Multi-Armed Dueling Bandits: player may select an arbitrary subset of arms and observe preferential feedback

Open Problems

- Statistical tests to determine whether the assumptions of the preference matrix (transitivity, triangle inequality) hold, given sample access to Q
- Combining preference-based and real-valued MAB settings (player may choose whether to pull a single arm and observe numerical reward, or multiple and observe preferential reward)

Reference I

[YJ09] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1201–1208, 2009.

[YJ11] Yisong Yue and Thorsten Joachims. Beat the mean bandit. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 241–248. Citeseer, 2011.

[ZWMR14] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In International conference on machine learning, pages 10–18. PMLR, 2014.