Combinatorial Bandits for Maximum Value Reward Function under Value-Index Feedback

Yiliu Wang[1]; Wei Chen[2]; Milan Vojnovic[3]
[1]Allen Institute, [2]Microsoft Research, [3]London School of Economics

Abstract

We investigate the combinatorial multi-armed bandit problem where an action is to select k arms from a set of base arms, and its reward is the maximum of the sample values of these k arms, under a weak feedback structure that only returns the value and index of the arm with the maximum value.

Summary of our contributions:

- Novel feedback structure: much weaker than the semi-bandit, slightly stronger than full bandit.
- Novel concept: biased arm replacement.
- Distribution dependent and independent regret bound: comparable to the bounds obtained under semi-bandit.

Motivation

- In many real-world applications we make sequential decisions…
  - Online shopping
  - Digital advertising
  - Portfolio selection
- … in a non sequential order
  - Users may go back and forth on the list
  - List may not be presented in a one-dimensional order (e.g. two-dimensional grids)
  - User pays attention to a subset of items, and select the most valuable one among them to click

Figure 1. Application Scenario: Online Recommendation

Problem Formulation

Consider an agent and a system of n base arms $X_1, \ldots, X_n$.

- At each time step t, the agent chooses a subset of items of size k.
- The agent receives the maximum value of the set and the index achieving this max value as feedback.

Key challenge: limited feedback. When item i wins with value v_i, for other competing items j, we do not know if j has intrinsically lower value, or j was just unlucky (not being paid attention).

Model Analysis: Binary case

- Each arm is a binary random variable that takes v_i with probability p_i.
- For learning purpose, decompose into two sets of base arms Z_i and V_i.

  Z = $\{Z_1, \ldots, Z_n\}$: Bernoulli random variables with means $p_1, \ldots, p_n$.
  V = $\{V_1, \ldots, V_n\}$: Deterministic with means $v_1, \ldots, v_n$.

In each round an action is played, we obtain information on some base arms and we call them triggered base arms.

- Warm Up: if the agent knows the ordering of v_i's, the problem reduces to (weighted) cascading bandit.
- Real Challenge: Generally, we have no information on Z_i if V_i is not triggered yet.

Reward function: $r_S(\boldsymbol{p}, \boldsymbol{v}) = \sum_{i \in S} v_i p_i \prod_{j \in S, j<i} (1 - p_i)$

- Monotonicity: reward nondecreasing for any p_i and v_i
- Relative triggering probability modulated (RTPM) condition: $\left| r_S(\boldsymbol{p}, \boldsymbol{v}) - r_S(\boldsymbol{p}', \boldsymbol{v}') \right| \leq 2 \sum_{i \in S} q_i^{\boldsymbol{p},S} v_i' \left| p_i - p_i' \right| + \sum_{i \in S} \tilde{q}_i^{\boldsymbol{p},S} \left| v_i - v_i' \right|$
  - $q_i^{\boldsymbol{p},S} = \prod_{j<i} (1 - p_i)$: triggering probability for arm Z_i
  - $\tilde{q}_i^{\boldsymbol{p},S} = q_i^{\boldsymbol{p},S} \cdot p_i$: triggering probability for arm V_i

New Technique: Biased arm replacement

- Key idea: pretend Z_i is triggered before v_i is observed.
- For any arm with unknown value v_i, replace its parameters s.t.

$$(p_i, v_i) \to (p_i', v_i') : \quad v_i' = 1, \text{and } p_i' = p_i \cdot v_i$$

The estimates are biased, yet achieved the purpose.

- Distribution dependent regret $O((k/\Delta)\log T)$
- Distribution independent regret $O(\sqrt{mkT \log T})$

$$\mathrm{Reg}(t) = \Delta_{S_t} \leq \underbrace{\left(r_{S_t}\left(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{v}}_t\right) - r_{S_t}\left(\boldsymbol{p}_t', \boldsymbol{v}_t'\right) \right)}_{\text{estimation error}} + \underbrace{\left(r_{S_t}\left(\boldsymbol{p}_t', \boldsymbol{v}_t'\right) - r_{S_t}\left(\boldsymbol{p}, \boldsymbol{v}\right) \right)}_{\text{replacement error}}$$

Both can be bounded similarly by applying the RTPM

Arbitrary distributions with finite support

- Each arm X_i takes values $v_{i,0} = 0 < v_{i,1} < \cdots < v_{i,s_i} \leq 1$.
- Key idea: Multi-valued arms can be turned into a set of binary arms.

$$X_i \to \{X_{i,j} | j = 0,1,\ldots s_i\} \text{ s.t. } X_i \overset{d}{=} \max_j X_{i,j}$$

To handle unknown support size:

- Keep a dynamic counter for known support size.
- Use a fictitious arm with value 1 for unknown values.

- Achieves $(1 - \epsilon)$ regret using a PTAS offline oracle:

$$O\left(k \sum_i \frac{s_i}{\Delta_{\min}^i} \log T \right)$$

Simulation results

Take three different sets of distributions with the same support

$$v_i = 0.1 * i \text{ for } i = 1,2,\ldots,9.$$

- For $i = 1,2,\ldots,6$, $p_i = 0.2$, and for $i = 7,8,9$, $p_i = 0.5$.
- Low-risk low-reward item: Change first arm to arm with small v_i but large p_i, $p_1 = 0.9$
- High-risk high-reward item: Change last arm to arm with large v_i but small p_i, $p_9 = 0.1$

Our regret curve closely aligns with that of the CUCB method under the semi-bandit setting, indicating that we do not incur substantial losses despite receiving much less feedback.

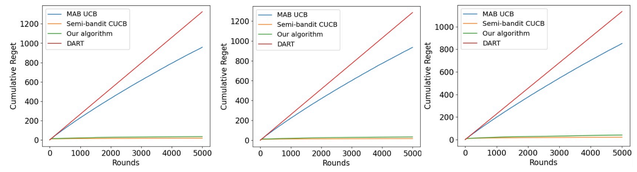

Figure 2. Regret for our algorithm and benchmarks used for comparison, for different distributions of arm outcomes.

Contact

<Yiliu Wang>
<Allen Institute>
Email: yiliu.wang@alleninstitute.org

References

1. Mridul Agarwal, Vaneet Aggarwal, Abhishek Kumar Umrawal, and Chris Quinn. DART: Adaptive accept reject algorithm for non-linear combinatorial bandits. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pp. 6557–6565, 2021b.
2. Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In International conference on machine learning, pp. 151–159. PMLR, 2013.
3. Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. Journal of Machine Learning Research, 17(50):1–33, 2016b.
4. Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. Advances in Neural Information Processing Systems, 28, 2015b.
5. Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In Artificial Intelligence and Statistics, pp. 535–543. PMLR, 2015a.
6. Xutong Liu, Jinhang Zuo, Siwei Wang, Carlee Joe-Wong, John Lui, and Wei Chen. Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms. Advances in Neural Information Processing Systems, 35:14904–14916, 2022.
7. Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilis- tically triggered arms and its applications. Advances in Neural Information Processing Systems, 30, 2017.
8. Idan Rejwan and Yishay Mansour. Top-k combinatorial bandits with full-bandit feedback. In Algorithmic Learning Theory, pp. 752–776. PMLR, 2020.