Contextual Combinatorial Multi-armed Bandits with Volatile Arms and Submodular Reward

Lixing Chen†; Jie Xu†; Zhuo Lu‡

†University of Miami, Coral Gable, FL 33146 ‡University of South Florida, Tampa, FL 33620

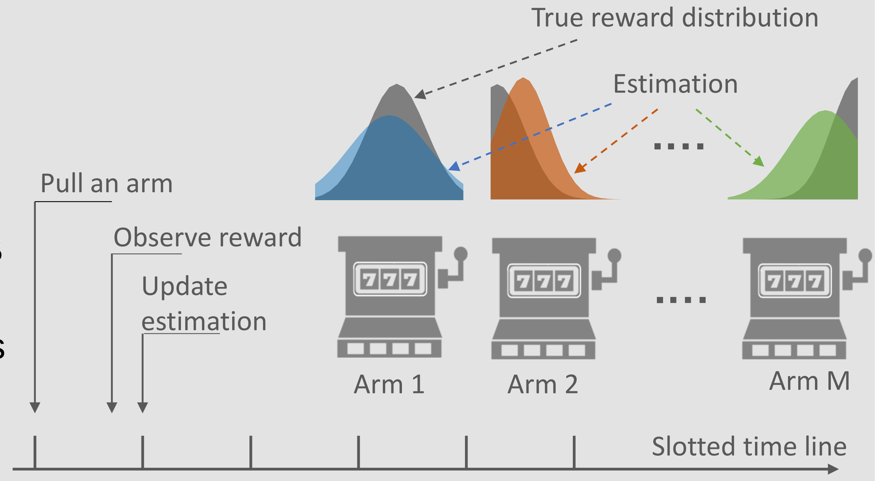
NIPS 2018
Montréal CANADA

Abstract

We study the stochastic contextual combinatorial multi-armed bandit (CC-MAB) framework that is tailored for volatile arms and submodular reward functions. CC-MAB inherits properties from both contextual and combinatorial bandit: it aims to select a set of arms in each round based on contexts associated with the arms. By "volatile arms", we mean that the available arms to select from in each round may change; and by "submodular rewards", we mean that the total reward achieved by selected arms is not a simple sum of individual rewards but demonstrates a feature of diminishing returns determined by the relations between selected arms (e.g. relevance and redundancy). Volatile arms and submodular rewards are often seen in many real-world applications, e.g. recommender systems and crowdsourcing, in which multi-armed bandit (MAB) based strategies are extensively applied. Although there exist works that investigate these issues separately based on standard MAB, jointly considering all these issues in a single MAB problem requires very different algorithm design and regret analysis. Our algorithm CC-MAB provides an online decision-making policy in a contextual and combinatorial bandit setting and effectively addresses the issues raised by volatile arms and submodular reward functions. CC-MAB achieves a sublinear regret $O(cT^{2\alpha+D/3\alpha+D}\log T)$. The performance of CC-MAB is evaluated by experiments conducted on a real-world crowdsourcing dataset, and the result shows that our algorithm outperforms the prior art.

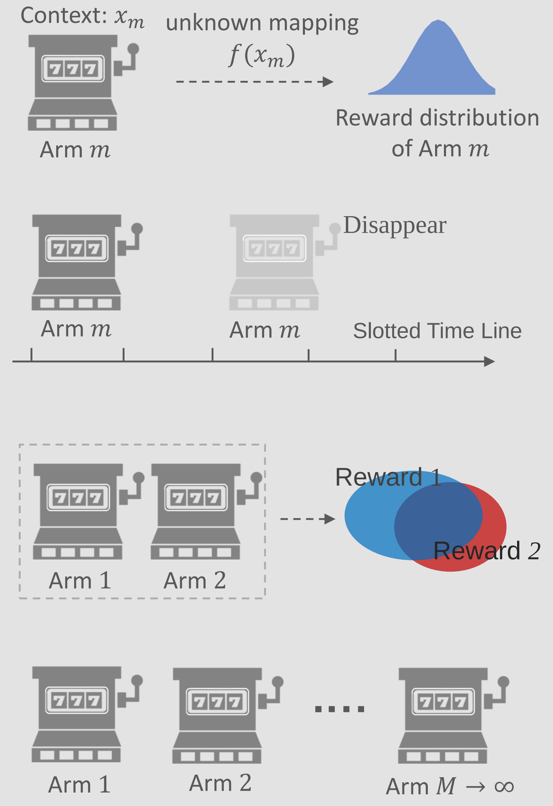
Introduction

- **Multi-armed Bandit (MAB)**
  - Exploration-Exploitation Tradeoff
    - exploration: learn the expected reward of different arms
    - exploitation: pull the arm that yielded highest reward in the past
  - Objective
    - maximize cumulative reward over time horizon by balancing exploration and exploitation.
  - Performance Metric: Regret
    - gap between the cumulative reward achieved by the designed algorithm and that achieved by an Oracle that always selecting the best arm.
  - Sublinear Regret
    - a sublinear regret in the time horizon T guarantees an asymptotically optimal performance.
  - Applications
    - clinical trials
    - crowdsourcing
    - recommender systems
  - Limitation of standard MAB
    - pull one arm each slot
    - independent arm rewards
    - constant arm set
    - a finite number of arms



Key Features and Contributions

- **Contextual Combinatorial Multi-armed Bandit (CC-MAB)**
  - Arm Context
    - context associated with each arm determines the reward distribution
  - Volatile Arms
    - arms may "appear" or "disappear" across time slots
  - Combinatorial Bandit and submodularity
    - pull multiple arms in each round
    - submodular: diminishing returns due to the relations between arms (e.g. redundancy)
  - Infinite Arm Set
    - allow infinitely many arms in MAB framework



Preliminaries

- **Sequential Decision-making using CC-MAB**
  - Time varying arm set $\mathcal{M}^t$ that captures volatile arms.
  - Context-aware arm quality $\boldsymbol{r}^t = \{r(x_m^t)\}_{m\in\mathcal{M}^t}$
    - each arm m is associated with a context (side information) x_m.
    - arm quality is a random variable parameterized by context $r(x_m)$.
  - Limited budget B: select in each slot a set of arm $\mathcal{S}^t \subseteq \mathcal{M}^t, |\mathcal{S}^t| \le B$
  - Submodular reward
    - $u(\boldsymbol{r}, \{m\}\cup\mathcal{S}) - u(\boldsymbol{r},\mathcal{S}) \ge u(\boldsymbol{r},\{m\}\cup\mathcal{B}) - u(\boldsymbol{r},\mathcal{B})$, for $\mathcal{S}\subseteq\mathcal{B}\subseteq\mathcal{M}$, and $m \notin \mathcal{B}$
    - marginal utility $\Delta(\boldsymbol{r}, m|\mathcal{S}) \triangleq u(\boldsymbol{r},\{m\}\cup\mathcal{S}) - u(\boldsymbol{r},\mathcal{S})$
  - Utility maximization in finite time horizon T
    - $\max_{\mathcal{S}^1,\dots,\mathcal{S}^T} \sum_{t=1}^{T} \mathbb{E}\left[u(\boldsymbol{r}^t,\mathcal{S}^t)\right]$ s.t. $|\mathcal{S}^t| \le B, \ \mathcal{S}^t \subseteq \mathcal{M}^t, \ \forall t$
- **Oracle Solution**
  - Oracle knows expected quality $\mu(x) = \mathbb{E}[r(x)]$ a priori.
  - Solving per-slot problem: $\mathcal{S}^{*,t}(\boldsymbol{x}^t) = \arg\max_{\mathcal{S}\subseteq\mathcal{M}^t,|\mathcal{S}|\le B} u(\boldsymbol{\mu}^t,\mathcal{S})$
    - GA approximates the optimal solution with polynomial runtime.

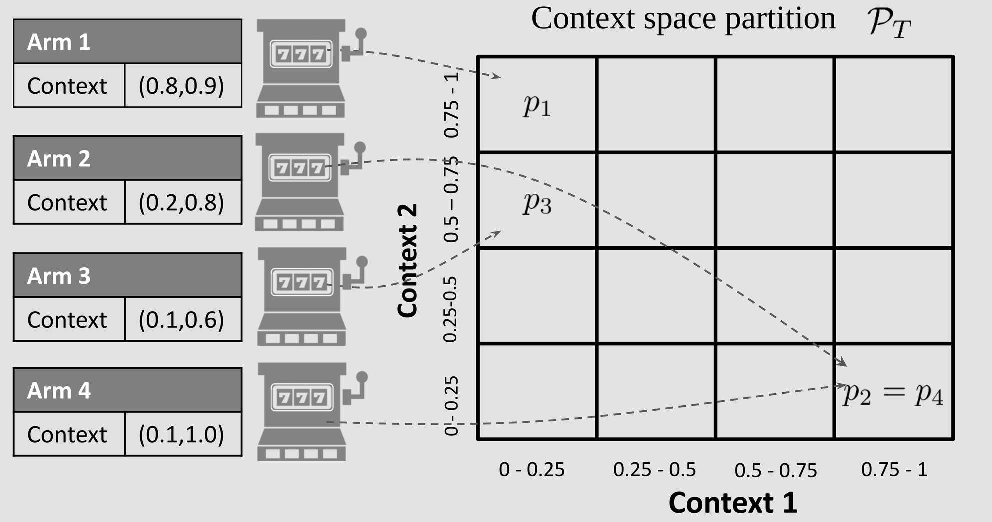
Algorithm 1 Greedy Algorithm (GA)

Input: $\mathcal{M}^t$, $\boldsymbol{r}^t$, $u(\cdot,\cdot)$, B.
Initialization: $\mathcal{S}_0 \leftarrow \emptyset, k \leftarrow 0$;
while $k \le B$ **do**:
  $k = k+1$;
  select $m_k = \arg\max_{m_k\in\mathcal{M}^t\setminus\mathcal{S}_{k-1}} \Delta(\boldsymbol{\mu}^t, \{m_k\}|\mathcal{S}_{k-1})$;
  $\mathcal{S}_k = \mathcal{S}_{k-1}\cup\{m_k\}$
end while

  - performance guarantee: $u(\boldsymbol{\mu}^t,\mathcal{S}^t) \ge (1-\frac{1}{e})u(\boldsymbol{\mu}^t,\mathcal{S}^{*,t})$
- **Regret**
  - Utility loss against Oracle solution
    - $R(T) = (1-\frac{1}{e})\cdot\sum_{t=1}^{T}\mathbb{E}[u(\boldsymbol{r}^t,\mathcal{S}^{*,t})] - \sum_{t=1}^{T}\mathbb{E}[u(\boldsymbol{r}^t,\mathcal{S}^t)]$

Technique Overview

- **Context Space Partition**
  - Partition: $\mathcal{X}\to\mathcal{P}_T, (h_T)^D$ hypercubes of identical size $\frac{1}{h_T}\times\cdots\times\frac{1}{h_T}$
  - Counters: each hypercube $p\in\mathcal{P}_T$ keeps a counter $C^t(p)$ to record the number of times that an arm with $x\in p$ is chosen.
  - Experiences: each hypercube $p\in\mathcal{P}_T$ keeps an experience $\mathcal{E}^t(p)$ to store the observed quality of chosen arms with $x \in p$.
  - Quality estimation: $\hat{r}^t(p) = \frac{1}{C^t(p)}\sum_{r\in\mathcal{E}^t(p)} r$



Counter and Experience Update

Selected arms: $\mathcal{S}^t = \{1,2,4\}$

Update counters:

$C^{t+1}(p_1) = C^t(p_1) + 1$
$C^{t+1}(p_2) = C^t(p_2) + 2$

Update experiences:

$\mathcal{E}^{t+1}(p_1) = \mathcal{E}^{t+1}(p_1)\cup\{r_1\}$
$\mathcal{E}^{t+1}(p_2) = \mathcal{E}^{t+1}(p_2)\cup\{r_2, r_4\}$

- **Exploration-Exploitation Tradeoff**
  - Under-explored hypercubes and arms
    - under-explored hypercubes: $\mathcal{P}_T^{\mathrm{ue},t} \triangleq \{p\in\mathcal{P}_T \mid \exists\, m\in\mathcal{M}^t, x_m^t\in p, C^t(p)\le K(t)\}$
    - under-explored arms: $\mathcal{M}^{\mathrm{ue},t} \triangleq \{m\in\mathcal{M}^t \mid p_m^t \in \mathcal{P}_T^{\mathrm{ue},t}\}$
  - Explore when $\mathcal{M}^{\mathrm{ue},t}\ne\emptyset$
    - If $|\mathcal{M}^{\mathrm{ue},t}| \ge B$: $\mathcal{S}^t \leftarrow$ randomly select B arms in $\mathcal{M}^{\mathrm{ue},t}$
    - If $|\mathcal{M}^{\mathrm{ue},t}| < B$: $\mathcal{S}^t \leftarrow$ pick all arms in $\mathcal{M}^{\mathrm{ue},t}$ and other $B - |\mathcal{M}^{\mathrm{ue},t}|$ arms using GA
      $m_k = \arg\max_{m_k\in\mathcal{M}^t\setminus\{\mathcal{M}^{\mathrm{ue},t}\cup\mathcal{S}_{k-1}\}} \Delta(\hat{\boldsymbol{r}}^t, m_k|\{\mathcal{S}_{k-1}\cup\mathcal{M}^{\mathrm{ue},t}\}), \ k = 1,..,B-|\mathcal{M}^{\mathrm{ue},t}|)$
  - Exploit when $\mathcal{M}^{\mathrm{ue},t} = \emptyset$
    - choose the best arms based on quality estimation
      $\mathcal{S}^t \leftarrow m_k = \arg\max_{m_k\in\mathcal{M}^t\setminus\cup_i^{k-1}m_i} \Delta(\hat{\boldsymbol{r}}^t, \{m_k\}|\cup_i^{k-1} m_i), \ k = 1,\dots,B$

Analytical Results

- **Assumption (Hölder Condition):** There exists $L > 0$, $\alpha > 0$ such that for any two contexts, it holds that $|\mu(x)-\mu(x')| \le L\|x-x'\|^\alpha$.
- **Regret Upper Bound:** Let $K(t) = t^{\frac{2\alpha}{3\alpha+D}}\log(t)$ and $h_T = \lceil T^{\frac{1}{3\alpha+D}}\rceil$. If CC-MAB is run with these parameters and Hölder condition holds true, the regret $R(T)$ is bounded by

$$R(T) \le (1-\frac{1}{e})\cdot Br^{\max}2^D\left(\log(T)T^{\frac{2\alpha+D}{3\alpha+D}} + T^{\frac{D}{3\alpha+D}}\right) + (1-\frac{1}{e})\cdot B^2 r^{\max}\binom{M^{\max}}{B}\frac{\pi^2}{3} + \left(3BLD^{\alpha/2} + \frac{2Br^{\max}+2BLD^{\alpha/2}}{(2\alpha+D)/(3\alpha+D)}\right)T^{\frac{2\alpha+D}{3\alpha+D}}.$$

The leading order of the above regret $R(T)$ is $O(cT^{\frac{2\alpha+D}{3\alpha+D}}\log(T))$, where $c = (1-\frac{1}{e})Br^{\max}2^D$.

- **Discussion on Regret Upper Bound**
  - regret leading order is sublinear
  - If α is large enough, the regret bound of CC-MAB is close to that of continuum bandit, i.e., $O(cT^{\frac{2}{3}}\log^{\frac{1}{3}}(T))$.
  - the regret bound is looser when budget B is large. If $B \ge \mathcal{M}^t$ the regret should be 0.
- **Arm arrival pattern**

  If Hölder condition holds true and the arm arrival pattern satisfies $\mathbb{E}[\Pr(B < M^t)] = \beta$, let R^{ub} be the original regret upper bound, the regret is now bounded by $R(T) \le \beta R^{\mathrm{ub}}$.

Experiment

- **Crowdsourcing on Yelp Dataset**
  - users are employed to review businesses.
  - each user-business pair is an arm
  - Dixit-Stiglitz model as submodular reward: $u_i = \left(\sum_j (r_{ij})^p\right)^{1/p}, p \ge 1$.
  - context-aware arm quality

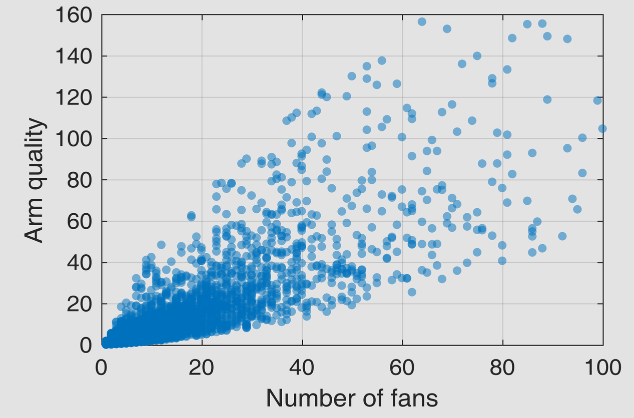


Fig. 1. Arm quality distribution

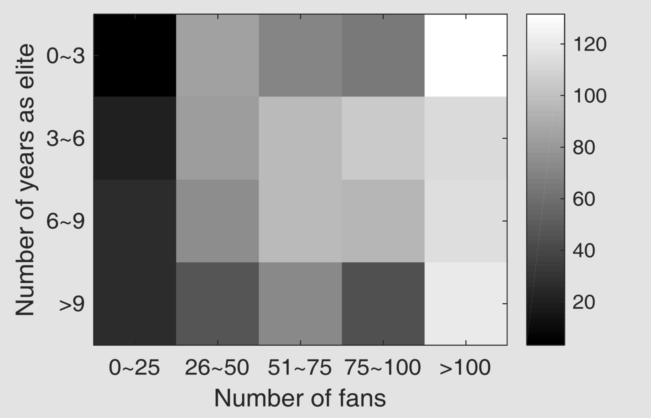


Fig. 2. expected quality of hypercubes

- **Results**

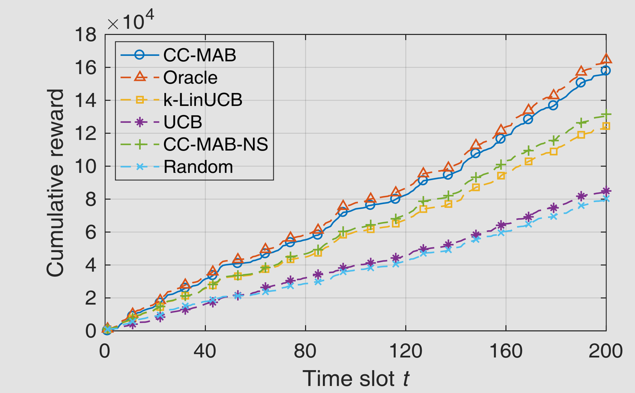


Fig. 3. Comparison of cumulative rewards

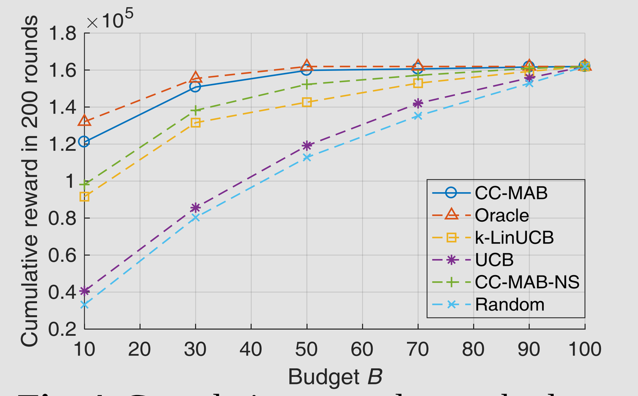


Fig. 4. Cumulative rewards over budgets

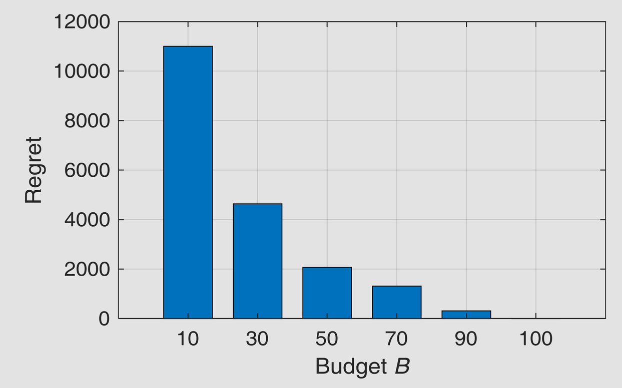


Fig. 5. Regret over budgets

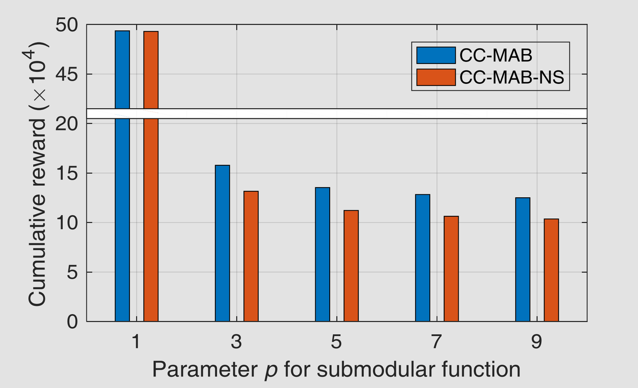


Fig. 6. Impact of submodularity