Combinatorial Multi-Armed Bandit: General Framework, Results and Applications

Wei Chen
Microsoft Research Asia

Yajun Wang
Microsoft

Yang Yuan
Cornell University
CMAB Outline

Motivation and Background

- Motivation from online advertising and viral marketing
- Background on multi-armed bandit (MAB) problem

Combinatorial MAB and Its General Solution

CMAB Applications

Summary and Future Work
Motivating application: Display ad placement

- Bipartite graph of pages and users who are interested in certain pages
  - Each edge has a click-through probability
- Find $k$ pages to put ads to maximize total number of users clicking through the ad
- When click-through probabilities are known, can be solved by approximation
- Question: how to learn click-through prob. while doing optimization?
Main difficulties

- Combinatorial in nature
- Non-linear optimization objective, based on underlying random events
- Offline optimization may already be hard, need approximation
- Online learning: learn while doing repeated optimization
Multi-armed bandit problem

- There are $m$ arms (machines)
- Arm $i$ has an unknown reward distribution with unknown mean $\mu_i$
  - best arm $\mu^* = \max \mu_i$
- In each round, the player selects one arm to play and observes the reward
Multi-armed bandit problem

- Regret after playing \( n \) rounds:
  - \( \text{Regret} = n\mu^* - \mathbb{E}[\sum_{t=1}^{n} R_t(i_t^A)] \)

- Objective: minimize regret in \( n \) rounds

- Balancing exploitation-exploration tradeoff

- Known results:
  - Regret lower bound \( \Omega(\log n) \)
  - Upper Confidence Bound (UCB) algorithm:
    - achieves \( O(\log n) \) regret
Naïve application of MAB to the combinatorial setting

- E.g. online advertising
  - every set of k webpages is treated as an arm
  - reward of an arm is the total click-through counted by the number of people

- Issues
  - combinatorial explosion
  - ad-user click-through information is wasted
Contribution of this paper

- Stochastic combinatorial multi-armed bandit framework
  - handling non-linear reward functions
  - UCB based algorithm and tight regret analysis
  - new applications using CMAB framework
- Comparing with related work
  - linear stochastic bandits [Gai et al. 2012]
    - CMAB is more general, and has much tighter regret analysis
  - online submodular optimizations (e.g. [Streeter & Golovin’08, Hazan & Kale’12])
    - for adversarial case, different approach,
    - CMAB has no submodularity requirement
CMAB Outline

Motivation and Background

Combinatorial MAB and Its General Solution

CMAB Applications

Summary and Future Work

Summary
- Need combinatorial online learning in practice
- Naïve MAB is not feasible
CMAB Outline

Motivation and Background

Combinatorial MAB and Its General Solution

CMAB Applications

Summary and Future Work

- Combinatorial multi-armed bandit (CMAB) framework
- General solution CUCB
Combinatorial multi-armed bandit (CMAB) framework

- A super arm $S$ is a set of (base) arms, $S \subseteq [m]$
- In round $t$, a super arm $S_t^A$ is played according algo $A$
- When a super arm $S$ is played, all based arms in $S$ are played
- Outcomes of all played base arms are observed
- Outcome of arm $i \in [m]$ has an unknown distribution with unknown mean $\mu_i$
Rewards in CMAB

- Reward of super arm $S_t^A$ played in round $t$, $R_t(S_t^A)$, is a function of the outcomes of all played arms.
- Expected reward of playing arm $S$, $\mathbb{E}[R_t(S)]$, only depends on $S$ and the vector of mean outcomes of arms, $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$, denoted $r_\mu(S)$.
  - e.g. independent Bernoulli random variables
- Optimal reward: $\text{opt}_\mu = \max_S r_\mu(S)$
Handling non-linear reward functions

--- two mild assumption on $r_\mu(S)$

- **Monotonicity**
  - if $\mu \leq \mu'$ (pairwise), $r_\mu(S) \leq r_{\mu'}(S)$, for all super arm $S$

- **Bounded smoothness**
  - there exists a strictly increasing function $f(\cdot)$, such that for any two expectation vectors $\mu$ and $\mu'$,
    \[ |r_\mu(S) - r_{\mu'}(S)| \leq f(\Delta), \] where $\Delta = \max_{i \in S} |\mu_i - \mu'_i|$

- Rewards may not be linear, a large class of functions satisfy these assumptions
Offline computation oracle --- allow approximations and failure probabilities

- \((\alpha, \beta)\)-approximation oracle:
  - Input: vector of mean outcomes of all arms \(\mu = (\mu_1, \mu_2, ..., \mu_m)\),
  - Output: a super arm \(S\), such that with probability at least \(\beta\) the expected reward of \(S\) under \(\mu\), \(r_\mu(S)\), is at least \(\alpha\) fraction of the optimal reward:
    \[
    \Pr[r_\mu(S) \geq \alpha \cdot \text{opt}_\mu] \geq \beta
    \]
(α, β)-Approximation regret

- Compare against the αβ fraction of the optimal

\[ \text{Regret} = n \cdot \alpha \beta \cdot \text{opt}_\mu - \mathbb{E}[\sum_{i=1}^{n} r_\mu(S_t^A)] \]

- Difficulty: do not know
  - combinatorial structure
  - reward function
  - arm outcome distribution
  - how oracle computes the solution
Our solution: CUCB algorithm

\[ \bar{\mu} = (\bar{\mu}_1, \bar{\mu}_2, ..., \bar{\mu}_m) \]

**Offline computation oracle**

**Adjustment**

\[ \hat{\mu}_i = \hat{\mu}_i + \sqrt{\frac{3 \ln n}{2T_i}} \]

**Estimation**

\[ \hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_m) \]

**Superarm**

\[ \mu_i: \text{sample mean outcome on arm } i \]

\[ T_i: \# \text{ of times arm } i \text{ is played; key tradeoff between exploration and exploitation} \]
Theorem 1

- The \((\alpha, \beta)\)-approximation regret of the CUCB algorithm in \(n\) rounds using an \((\alpha, \beta)\)-approximation oracle is at most

\[
\sum_{i \in [m], \Delta_{\min}^i > 0} \left( \frac{6 \ln n \cdot \Delta_{\min}^i}{(f^{-1}(\Delta_{\min}^i))^2} + \int_{\Delta_{\min}^i}^{\Delta_{\max}^i} \frac{6 \ln n}{(f^{-1}(x))^2} dx \right) + \left( \frac{\pi^2}{3} + 1 \right) \cdot m \cdot \Delta_{\max}.
\]

- \(\Delta_{\min}^i\) (\(\Delta_{\max}^i\)) are defined as the minimum (maximum) gap between \(\alpha \cdot \text{opt}_\mu\) and reward of a bad super arm containing \(i\). \(\Delta_{\min} = \min_i \Delta_{\min}^i\), \(\Delta_{\max} = \max_i \Delta_{\max}^i\)

- Here, we define the set of bad super arms as

\[
S_B = \{ S \mid r_{\mu}(S) < \alpha \cdot \text{opt}_\mu \}
\]

- Match UCB regret for classic MAB
Proof outline

• If in round $t$, each arm $i$ is sufficiently sampled

$$T_{i,t-1} > \ell_t = \frac{6 \ln t}{(f^{-1}(\Delta_{\text{min}}))^2}$$
times, then with probability $1 - 2mt^{-2}$:

• sample mean $\hat{\mu}_i$ and UCB adjustment is close to true mean $\mu_i$,

$$|\hat{\mu}_{i,T_i,t-1} - \mu_i| \leq \Lambda_{i,t}, \Lambda_{i,t} = \sqrt{\frac{3 \ln t}{2T_{i,t-1}}} \quad \text{(by Hoeffding bound)}$$

$$|ar{\mu}_{i,t} - \mu_i| \leq 2\Lambda_{i,t} \quad \text{(since $\bar{\mu}_{i,t} = \hat{\mu}_{i,T_i,t-1} + \Lambda_{i,t}$)}$$

• UCB adjustment is at least true mean: $\bar{\mu}_t \geq \mu$

• super arm $S_t$ selected in round $t$ is not a bad super arm, why? ...
Proof outline (cont’d)

- define $\Lambda = \sqrt{\frac{3\ln t}{2\ell_t}}$, $\Lambda_t = \max\{\Lambda_{i,t} \mid i \in S_t\}$, thus $\Lambda > \Lambda_t$

- Then we have: $r_\mu(S_t) + f(2\Lambda)$
  - $> r_\mu(S_t) + 2f(2\Lambda_t)$ \{strict monotonicity of $f$\}
  - $\geq r_{\bar{\mu}_t}(S_t)$ \{bounded smoothness of $r_\mu(S)$\}
  - $\geq \alpha \cdot \text{opt}_{\bar{\mu}_t}$ \{$\alpha$-approximation w.r.t. $\bar{\mu}_t$\}
  - $\geq \alpha \cdot r_{\bar{\mu}_t}(S^*_\mu)$ \{definition of $\text{opt}_{\bar{\mu}_t}$\}
  - $\geq \alpha \cdot r_\mu(S^*_\mu) = \alpha \cdot \text{opt}_\mu$ \{monotonicity of $r_\mu(S)$\}

- Since $f(2\Lambda) = \Delta_{\min}$, contradiction to def’n of $\Delta_{\min}$, so $S_t$ is not a bad super arm with probability $1 - 2mt^{-2}$. 
When some arm is not sufficiently sampled, pay regret $\Delta_{\text{max}}$. Get a loose bound:

$$\left( \frac{6 \ln n}{(f^{-1}(\Delta_{\text{min}}))^2 + \frac{\pi^2}{3} + 1} \right) \cdot m \cdot \Delta_{\text{max}}$$

To tighten the bound, fine-tune bad super arms, sufficient sampling, and regret gaps.
Theorem 2

- Consider a CMAB problem with an $(\alpha, \beta)$-approximation oracle. If the bounded smoothness function $f(x) = \gamma \cdot x^\omega$ for some $\gamma > 0$ and $\omega \in (0,1]$, the regret of CUCB is at most:

$$\frac{2\gamma}{2 - \omega} \cdot (6m \ln n)^{\omega/2} \cdot n^{1-\omega/2} + \left(\frac{\pi^2}{3} + 1\right) \cdot m \cdot \Delta_{\max}.$$ 

- When $\omega = 1$, the distribution-independent bound is $O(\sqrt{mn \ln n})$.
CMAB Outline

Motivation and Background
Combinatorial MAB and Its General Solution
CMAB Applications
Summary and Future Plan

- Combinatorial multi-armed bandit (CMAB) framework
- General solution CUCB
CMAB Outline

Motivation and Background

Combinatorial MAB and Its General Solution

CMAB Applications

Summary and Future Plan

- Online advertising
- Linear reward bandits
Application to ad placement

- Bipartite graph $G = (L, R, E)$
- Each edge is a base arm
- Each set of edges linking $k$ webpages is a superarm
- Bounded smoothness function $f(\Delta) = |E| \cdot \Delta$
- $(1 - 1/e, 1)$-approximation regret

$$\sum_{i \in E, \Delta^i_{\text{min}} > 0} \frac{12 \cdot |E|^2 \cdot \ln n}{\Delta^i_{\text{min}}} + \left(\frac{\pi^2}{3} + 1\right) \cdot |E| \cdot \Delta_{\text{max}}$$
- Improvement based on clustered arms is available
Application to linear bandit problems

- Linear bandits: matching, shortest path, spanning tree (in networking literature)
- Maximize weighted sum of rewards on all arms
- Our result significantly improves the previous regret bound on linear rewards [Gai et al. 2012]
  - indicating that our general framework does not lose fidelity
Application to social influence maximization

- Require a new model extension to allow probabilistically triggered arms
- Use the same CUCB algorithm
- See full report arXiv:1111.4279 for complete details
CMAB Outline

Motivation and Background

Combinatorial MAB and Its General Solution

CMAB Applications

Summary and Future Work

• Online advertising
• linear reward bandits
CMAB Outline

Motivation and Background

Combinatorial MAB and Its General Solution

CMAB Applications

Summary and Future Work

Summary
• Separation of computation and learning
• Future work
• Contextual CMAB, partial observations
Summary and future work

• Summary
  • Avoid combinatorial explosion while utilizing low-level observed information
  • Modular approach: separation between online learning and offline optimization
  • Handles non-linear reward functions
  • New applications of the CMAB framework

• Future work
  • Combinatorial bandits in adversarial and contextual bandit settings
  • Combinatorial bandits where outcomes of underlying arms are only indirectly observed
CMAB Outline

Motivation and Background

Combinatorial MAB and Its General Solution

CMAB Applications

Summary and Future Work

Summary
• Separating computation and learning
• Contextual CMAB, partial observations

Future work

Questions?