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

Wei Chen
Microsoft Research Asia

Yajun Wang
Microsoft

Yang Yuan
Cornell University

ICML, June 18, 2013, revised

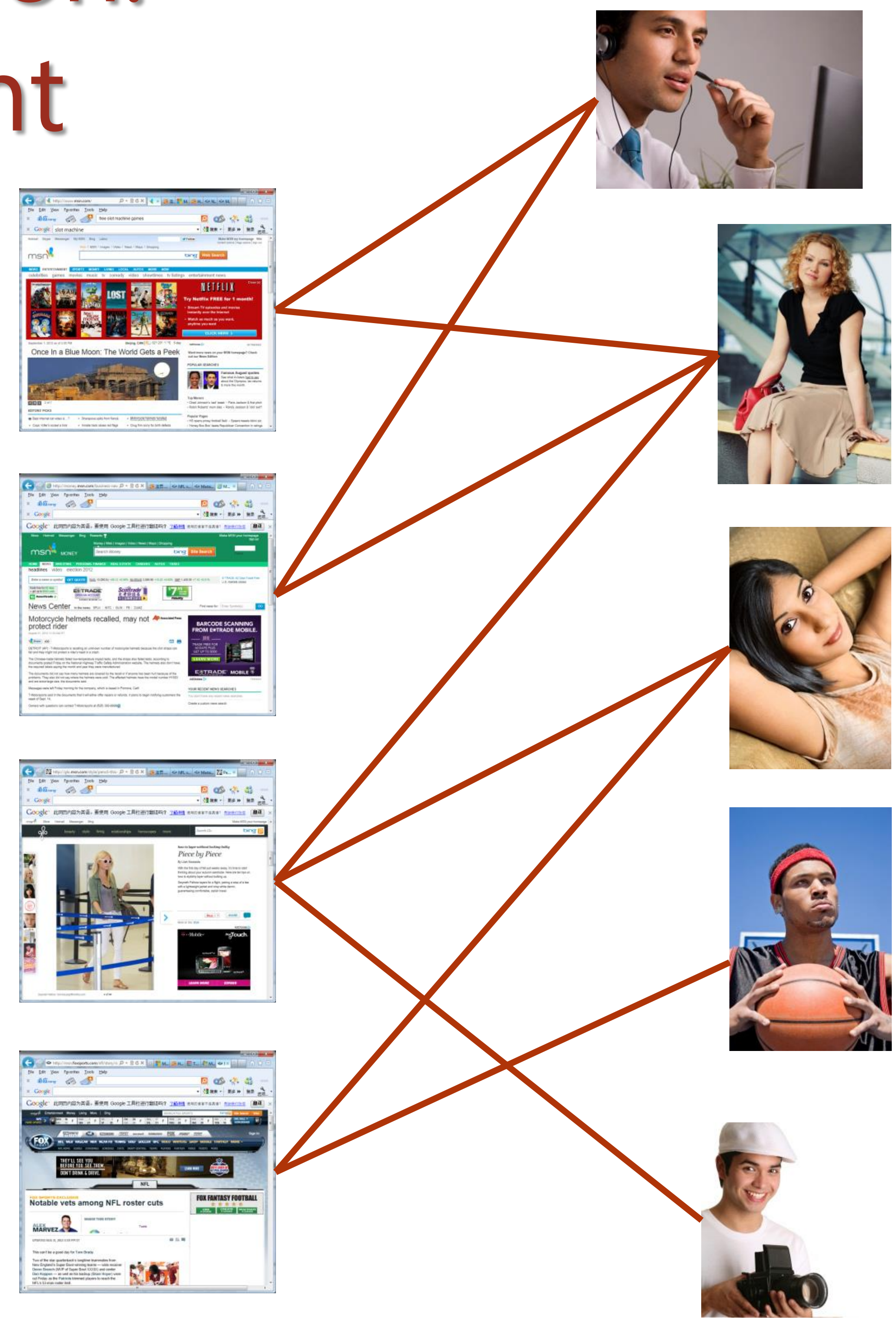
CMAB Outline

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

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

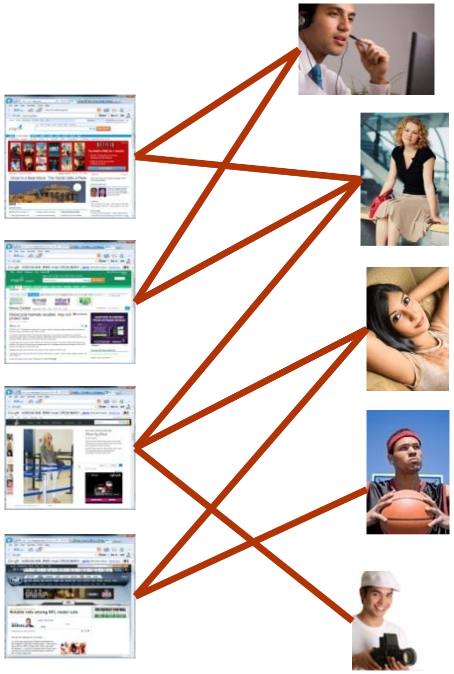
Motivating application: Display ad placement

- 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

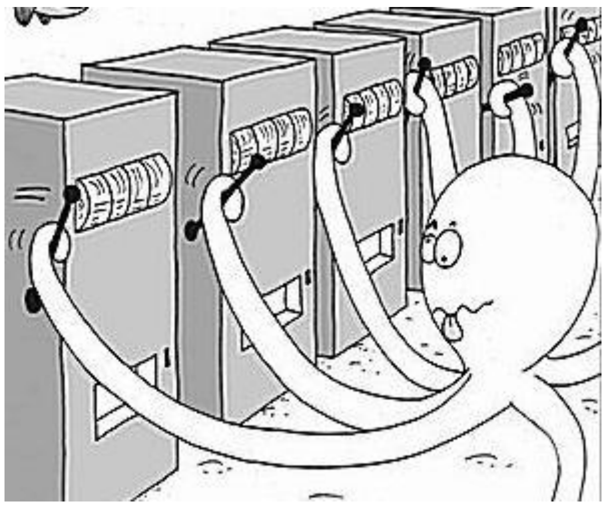
- Combinatorial in nature
- Non-linear optimization objective, based on underlying random events
- Offline optimization may already be hard, need approximation
- Online learning: learn while doing repeated optimization



Multi-armed bandit problem

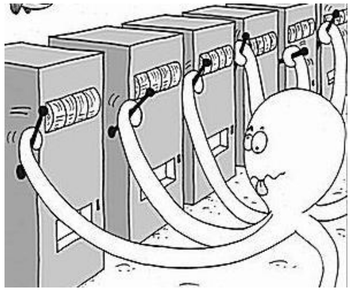
- There are m arms (machines)
- Arm i has an unknown reward distribution with unknown mean μ_i
  - best arm $\mu^* = \max \mu_i$
- In each round, the player selects one arm to play and observes the reward

Multi-armed bandit problem

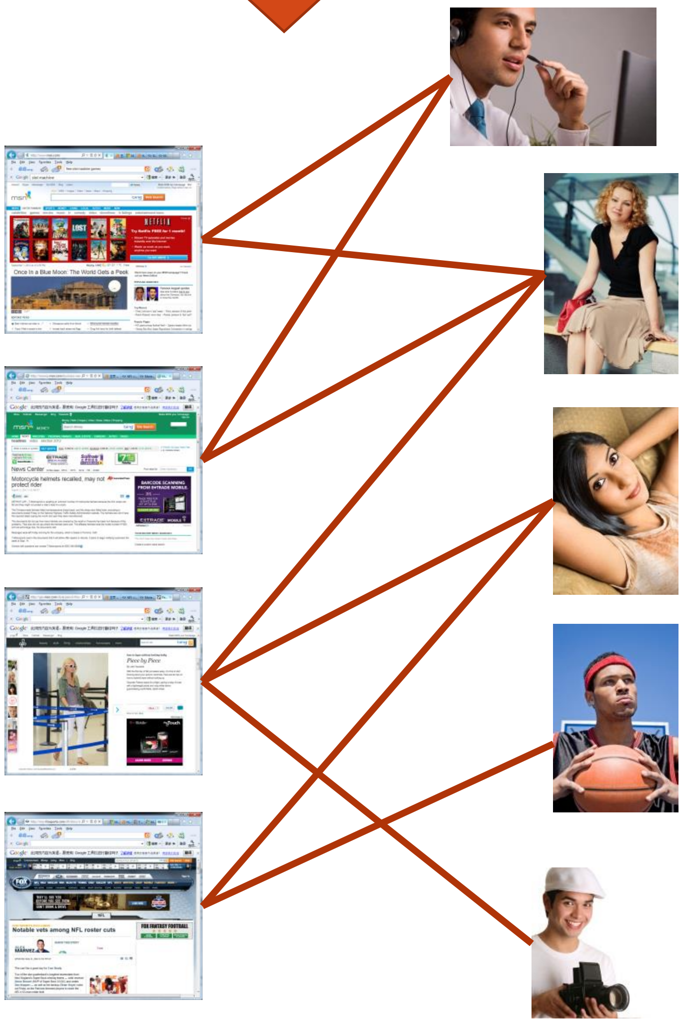


- Regret after playing n rounds:
  - Regret = $n\mu^* - \mathbb{E}[\sum_{t=1}^{n} R_t(i_t^A)\ ]$
- Objective: minimize regret in n rounds
- Balancing exploitation-exploration tradeoff
- Known results:
  - Regret lower bound $\Omega(\log n)$
  - Upper Confidence Bound (UCB) algorithm:
    - achieves $O(\log n)$ 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 μ_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, $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_m)$, denoted $r_{\boldsymbol{\mu}}(S)$
  - e.g. independent Bernoulli random variables
- Optimal reward: $\mathrm{opt}_{\boldsymbol{\mu}} = \max_S r_{\boldsymbol{\mu}}(S)$

Handling non-linear reward functions --- two mild assumption on $r_{\boldsymbol{\mu}}(S)$

- Monotonicity
  - if $\boldsymbol{\mu} \leq \boldsymbol{\mu}'$ (pairwise), $r_{\boldsymbol{\mu}}(S) \leq r_{\boldsymbol{\mu}'}(S)$, for all super arm S
- Bounded smoothness
  - there exists a strictly increasing function $f(\cdot)$, such that for any two expectation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, $|r_{\boldsymbol{\mu}}(S) - r_{\boldsymbol{\mu}'}(S)| \leq f(\Delta)$, 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

- (α, β)-approximation oracle:
  - Input: vector of mean outcomes of all arms $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_m)$,
  - Output: a super arm S, such that with probability at least β the expected reward of S under $\boldsymbol{\mu}$, $r_{\boldsymbol{\mu}}(S)$, is at least α fraction of the optimal reward:

$$\Pr\left[r_{\boldsymbol{\mu}}(S) \geq \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}\right] \geq \beta$$

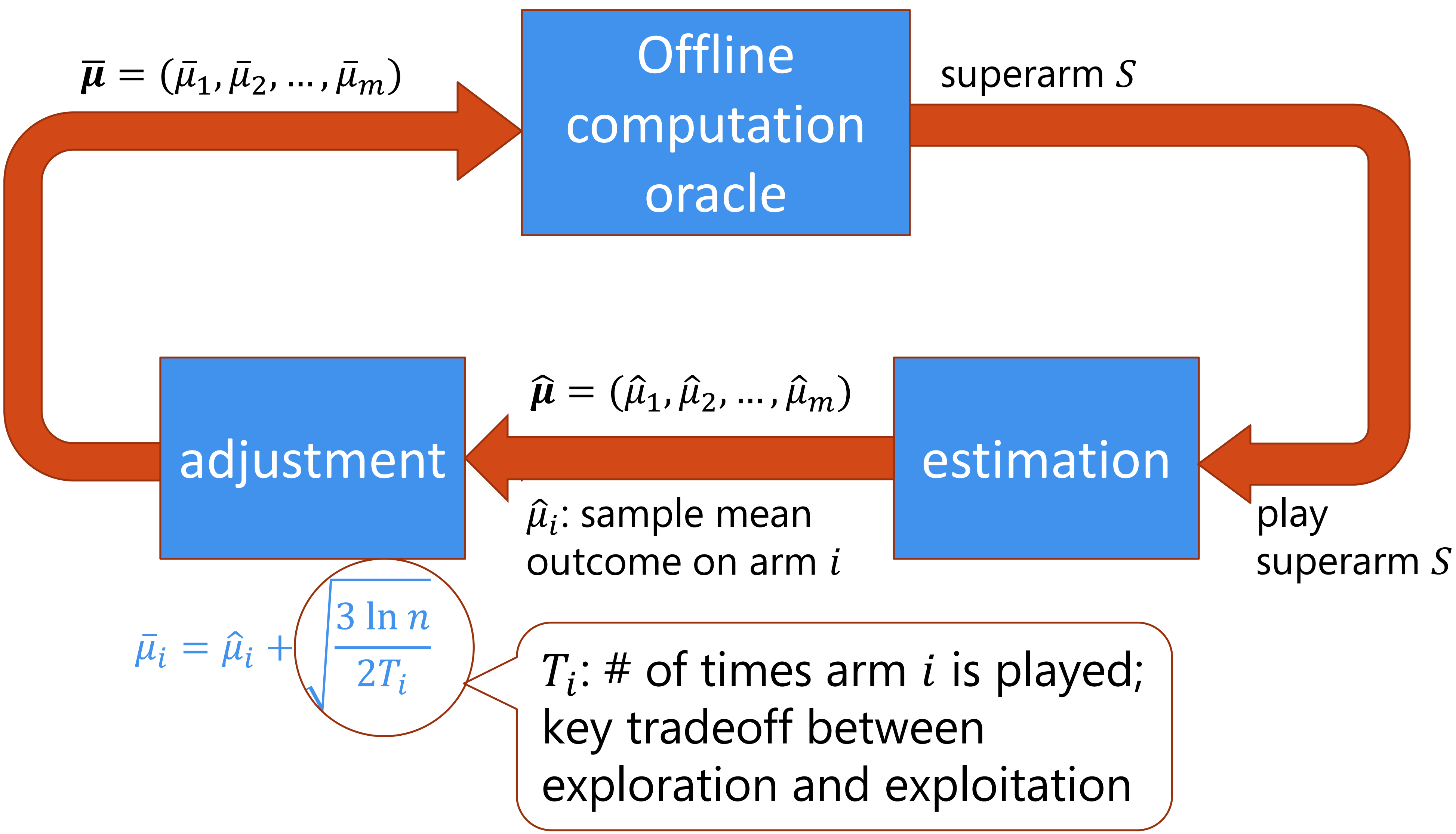
(α, β)-Approximation regret

- Compare against the $\alpha\beta$ fraction of the optimal

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

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

Our solution: CUCB algorithm



Theorem 1

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

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

  - $\Delta_{\min}^i$ ($\Delta_{\max}^i$) are defined as the minimum (maximum) gap between $\alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}$ 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

$$\mathcal{S}_{\mathrm{B}} = \{ S \mid r_{\boldsymbol{\mu}}(S) < \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}} \}$$

- Match UCB regret for classic MAB

Proof outline

- If in round t, each arm i is sufficiently sampled $T_{i,t-1} > \ell_t = \frac{6 \ln t}{\left(f^{-1}(\Delta_{\min})\right)^2}$ times, then with probability $1 - 2mt^{-2}$:
  - sample mean $\hat{\mu}_i$ and UCB adjustment is close to true mean μ_i,

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

    $|\bar{\mu}_{i,t} - \mu_i| \leq 2\Lambda_{i,t}$ (since $\bar{\mu}_{i,t} = \hat{\mu}_{i,T_{i,t-1}} + \Lambda_{i,t}$)
  - UCB adjustment is at least true mean: $\bar{\boldsymbol{\mu}}_t \geq \boldsymbol{\mu}$
  - super arm S_t selected in round t is not a bad super arm, why? …

Proof outline (cont'd)

- define $\Lambda = \sqrt{\frac{3\ln t}{2\ell_t}}$, $\Lambda_t = \max\{\Lambda_{i,t} | i \in S_t\}$, thus $\Lambda > \Lambda_t$
- Then we have: $r_{\boldsymbol{\mu}}(S_t) + f(2\Lambda)$
  - $> r_{\boldsymbol{\mu}}(S_t) + 2f(2\Lambda_t)$ {strict monotonicity of f}
  - $\geq r_{\bar{\boldsymbol{\mu}}_t}(S_t)$ {bounded smoothness of $r_{\boldsymbol{\mu}}(S)$}
  - $\geq \alpha \cdot \mathrm{opt}_{\bar{\boldsymbol{\mu}}_t}$ {α-approximation w.r.t. $\bar{\boldsymbol{\mu}}_t$}
  - $\geq \alpha \cdot r_{\bar{\boldsymbol{\mu}}_t}(S^*_{\boldsymbol{\mu}})$ {definition of $\mathrm{opt}_{\bar{\boldsymbol{\mu}}_t}$}
  - $\geq \alpha \cdot r_{\boldsymbol{\mu}}(S^*_{\boldsymbol{\mu}}) = \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}}$ {monotonicity of $r_{\boldsymbol{\mu}}(S)$}
- Since $f(2\Lambda) = \Delta_{\min}$, contradiction to def'n of $\Delta_{\min}$, so S_t is not a bad super arm with probability $1 - 2mt^{-2}$.

Proof outline (cont'd)

- When some arm is not sufficiently sampled, pay regret $\Delta_{\max}$. Get a loose bound:

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

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

Theorem 2

- Consider a CMAB problem with an (α, β)-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
  $$\sum_{i \in E, \Delta^i_{\min} > 0} \frac{12 \cdot |E|^2 \cdot \ln n}{\Delta^i_{\min}} + \left(\frac{\pi^2}{3} + 1\right) \cdot |E| \cdot \Delta_{\max}$$
- 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
- Separat arning

Future wo
- contextu tions

Questions?