Muti-armed Bandits, Online Learning and Sequential Prediction

Jian Li

Institute for Interdisciplinary Information Sciences
Tsinghua University
Outline

• Online Learning
• Stochastic Multi-armed Bandits
  • UCB
  • Combinatorial Bandits
• Top-k Arm Identification
• Combinatorial Pure Exploration
• Best Arm Identification
Online Learning

- $t = 1, 2, \ldots, T$

Choose an action $x_t$ (without knowing $f_t$)

Observe the reward $f_t(x_t)$ and the feedback (full information/semi-bandit/bandit feedback)

the environment plays $f_t$
Online Learning

- Adversarial / Stochastic environment

- Feedback
  - full information (Expert Problem): know $f_t$
  - semi-bandit (only makes sense in combinatorial setting)
  - bandit feedback: only knows the value $f_t(x_t)$
    - Exploration-Exploitation Tradeoff
The Expert Problem

A special case – coin guessing game

Imagine the adversary chooses a sequence beforehand (oblivious adversary): TTHHTTHTH……

If the prediction is wrong, cost = 1 for the time slot. Otherwise, cost = -1.

Suppose there is an expert who is really good (who can predict 90% correctly). Can you do (almost) at least this good?
No Regret Algorithms

- Define regret:
  \[ R_T = \sum_{t=1}^{T} c_t(x_t) - \sum_{t=1}^{T} c_t(x^*) \]
  where \( x^* = \arg\min_{x \in X} \sum_{t=1}^{T} c_t(x) \)

- We say an algorithm is “no regret” if \( R_T = o(T) \) (e.g., \( \sqrt{n} \))

- Hedge Algorithm (aka multiplicative weighting) [Freund & Schapire ‘97] can achieve a regret of \( O(\sqrt{n}) \)
  - Deep connection to Adaboost
Universal Portfolio

[Cover 91]

- $n$ stocks
- In each day, the price of each stock will go up or down
- In each day, we need to allocate our wealth between those stocks (without knowing their actually prices on that day)

- We can achieve almost the same asymptotic exponential growth rate of wealth as the best constant rebalanced portfolio chosen in hindsight (i.e., no regret!), using a continuous version of the multiplicative weight algorithm
  - (CRP is no worse than investing the single best stock)
Online Learning

A very active research area in machine learning

- Solving certain classes of convex programs
- Connections to stochastic approximation (SGD: stochastic gradient descent) [Leon Bottou]
- Connections to Boosting: Combining weak learners into strong ones [Freund & Schapire]
- Connections to Differential Privacy: idea of adding noise/regularization/multiplicative weight
- Playing repeated games
- Reinforcement learning (connection to Q-learning, Monte-Carlo tree search)
Outline

• Online Learning

• Stochastic Multi-armed Bandits
  • UCB
  • Combinatorial Bandits
  • Top-k Arm Identification
  • Combinatorial Pure Exploration
  • Best Arm Identification
Exploration-Exploitation Trade-off

• Decision making with limited information

An “algorithm” that we use everyday
• Initially, nothing/little is known
• Explore (to gain a better understanding)
• Exploit (make your decision)

• Balance between exploration and exploitation
• We would like to explore widely so that we do not miss really good choices
• We do not want to waste too much resource exploring bad choices (or try to identify good choices as quickly as possible)
The Stochastic Multi-armed Bandit

- Stochastic Multi-armed Bandit
  - Set of $n$ arms
  - Each arm is associated with an **unknown** reward distribution supported on $[0,1]$ with mean $\theta_i$
  - Each time, sample an arm and receive the reward independently drawn from the reward distribution

Classic problems in stochastic control, stochastic optimization and online learning
Stochastic Multi-armed Bandit

- Statistics, medical trials (Bechhofer, 54), Optimal control, Industrial engineering (Koenig & Law, 85), evolutionary computing (Schmidt, 06), Simulation optimization (Chen, Fu, Shi 08), Online learning (Bubeck Cesa-Bianchi, 12)

- [Bechhofer, 58] [Farrell, 64] [Paulson, 64] [Bechhofer, Kiefer, and Sobel, 68], …, [Even-Dar, Mannor, Mansour, 02] [Mannor, Tsitsiklis, 04] [Even-Dar, Mannor, Mansour, 06] [Kalyanakrishnan, Stone 10] [Gabillon, Ghavamzadeh, Lazaric, Bubeck, 11] [Kalyanakrishnan, Tewari, Auer, Stone, 12] [Bubeck, Wang, Viswanatha, 12] … [Karnin, Koren, and Somekh, 13] [Chen, Lin, King, Lyu, Chen, 14]

- Books:
  - Regret analysis of stochastic and nonstochastic multi-armed bandit problems S. Bubeck and N. Cesa-Bianchi, 2012
  - …..
The Stochastic Multi-armed Bandit

- Stochastic Multi-armed Bandit (MAB)

  MAB has MANY variations!

- Goal 1: Minimizing Cumulative Regret (Maximizing Cumulative Reward)

- Goal 2: (Pure Exploration) Identify the (approx) best K arms (arms with largest means) using as few samples as possible (Top-K Arm identification problem)
  - K=1 (best-arm identification)
A Quick Recap

- The Expert problem
  - Feedback: full information
  - Costs: Adversarial

- Stochastic Multi-armed bandits
  - Feedback: bandit information (you only observe what you play)
  - Costs: Stochastic
Upper Confidence Bound

- \( n \) stochastic arms (with unknown distributions)
- In each time slot, we can pull an arm (and get an i.i.d. reward from the reward distribution)
- Goal: maximize the cumulative reward/minimize the regret

\[ T_i(t) \] how many times we have played arm \( i \) up to time \( t \)
Upper Confidence Bound

• UCB Regret bound (Auer, Cesa-Bianchi, Fischer 02)

\[ R_T = \sum_{i=2}^{n} \frac{\log n}{\Delta_i} + (1 + \frac{\pi^2}{3})(\sum_{i=2}^{n} \Delta_i) \]

Gap: \[ \Delta_i = \mu_1 - \mu_i \]

• UCB has numerous extensions: KL-UCB, LUCB, CUCB, CLUCB, Lil-UCB, …..
Outline

- Online Learning
- Stochastic Multi-armed Bandits
  - UCB
  - Combinatorial Bandits
- Top-k Arm Identification
- Combinatorial Pure Exploration
- Best Arm Identification
Combinatorial Bandit - SDUCB

- Stochastic Multi-armed Bandit
  - Set of $n$ arms
  - Each arm is associated with an unknown reward distribution supported on $[0, s]$
  - Each time, we can play a combinatorial set $S$ of arms and receive the reward of the set (e.g., $\text{reward} = \max_{i \in S} X_i$)
- Goal: minimize the regret

- Application: Online Auction
  - Each arm: a user type – the distribution of the valuation
  - Each time we choose $k$ of them
  - The reward is the max valuation

[Chen, Hu, L, Li, Liu, Lu, NIPS16]
Combinatorial Bandit - SDCB

- Stochastic Dominate Confidence Bound

- High level idea: For each arm, maintain an estimate CDF which stochastically dominates the true CDF

- In each iteration, solve the offline optimization problem using the estimate CDF as the input (e.g., find $S$ which maximizes $E[\max_{i \in S} X_i]$)
Combinatorial Bandit - SDCB

- Results: Gap-dependent $O(\ln T)$ regret

\[ M^2 K \sum_{i \in E_B} \frac{2136}{\Delta_{i, \min}} \ln(\lambda T) + \left( \frac{\pi^2}{3} \lambda^{-3}(s - 1) + 1 \right) \alpha M m \]

- Gap-independent regret

\[ 93M \sqrt{mKT \ln(\lambda T)} + \left( \frac{\pi^2}{3} \lambda^{-3}(s - 1) + 1 \right) \alpha M m. \]
Outline

• Online Learning
• Stochastic Multi-armed Bandits
  • UCB
• Combinatorial Bandits
• Top-$k$ Arm Identification
• Combinatorial Pure Exploration
• Best Arm Identification
Best Arm Identification

- Best-arm Identification: Find the best arm out of n arms, with means $\mu[1], \mu[n], \ldots, \mu[n]$
- Goal: use as few samples as possible
- Formulated by Bechhofer in 1954

- Generalization: find out the top-k arms

- Applications: medical trails, A/B test, crowdsourcing, team formation, many extensions….
- Close connections to regret minimization
• Regret Minimization
  • Maximizing the cumulative reward
• Best/top-k arm identification
  • Find out the best arm using as few samples as possible

Your boss:
I want to go to casino tomorrow.
find me the best machine!
Applications

- Clinical Trails
  - One arm – One treatment
  - One pull – One experiment

Don Berry, University of Texas
MD Anderson Cancer Center
Applications

- Crowdsourcing:
- Workers are noisy

How to identify reliable workers and exclude unreliable workers?

Test workers by golden tasks (i.e., tasks with known answers)

- Each test costs money. How to identify the best $K$ workers with minimum amount of money?

<table>
<thead>
<tr>
<th>Top-$K$ Arm Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worker</td>
</tr>
<tr>
<td>Test with golden task</td>
</tr>
</tbody>
</table>
Naïve Solution

- $\epsilon$-approximation: the $i$th arm in our output is at most $\epsilon$ worse than the $i$th largest arm
- Uniform Sampling

Sample each coin $M$ times

Pick the $K$ coins with the largest empirical means

empirical mean: $\#\text{heads}/M$

How large $M$ needs to be (in order to achieve $\epsilon$-approximation)??

$$M = O(\log n)$$

So the total number of samples is $O(n\log n)$
Naïve Solution

Uniform Sampling

- With $M = O(\log n)$, we can get an estimate $\theta'_i$ for $\theta_i$ such that $|\theta_i - \theta'_i| \leq \epsilon$ with very high probability (say $1 - \frac{1}{n^2}$)

  - This can be proved easily using Chernoff Bound (Concentration bound).

  - Then, by union bound, we have accurate estimates for all arms

What if we use $M = O(1)$? (let us say $M = 10$)

- E.g., consider the following example ($K = 1$):
  - $0.9, 0.5, 0.5, \ldots, 0.5$ (a million coins with mean 0.5)

- Consider a coin with mean 0.5,
  
  $\Pr[\text{All samples from this coin are head}] = (1/2)^{10}$

- With const prob, there are more than 500 coins whose samples are all heads
Can we do better??

- Consider the following example:
  - 0.9, 0.5, 0.5, ......................, 0.5  (a million coins with mean 0.5)
  - Uniform sampling spends too many samples on bad coins.

- Should spend more samples on good coins
  - However, we do not know which one is good and which is bad……

- Sample each coin $M=O(1)$ times.
  - If the empirical mean of a coin is large, we DO NOT know whether it is good or bad
  - But if the empirical mean of a coin is very small, we DO know it is bad (with high probability)
Median/Quantile-Elimination

For i=1,2,…

Sample each arm $M_i$ times
Eliminate one quarter arms

Until less 4k arms

When $n \leq 4k$, use uniform sampling

We can find a solution with additive error $\epsilon$

Decrease $\epsilon$, until proper termination condition
Our algorithm

Algorithm 1: ME-AS

1 input: $B$, $\epsilon$, $\delta$, $k$
2 for $\mu = 1/2, 1/4, \ldots$ do
3 \hspace{1em} $S = \text{ME}(B, \epsilon, \delta, \mu, k)$;
4 \hspace{1em} $\{(a_i, \hat{\theta}^{US}(a_i)) \mid 1 \leq i \leq k\} = \text{US}(S, \epsilon, \delta, (1 - \epsilon/2)\mu, k)$;
5 \hspace{1em} if $\hat{\theta}^{US}(a_k) \geq 2\mu$ then
6 \hspace{2em} return $\{a_1, \ldots, a_k\}$;

Algorithm 2: Median Elimination (ME)

1 input: $B$, $\epsilon$, $\delta$, $\mu$, $k$
2 $S_1 = B$, $\epsilon_1 = \epsilon/16$, $\delta_1 = \delta/8$, $\mu_1 = \mu$, and $\ell = 1$;
3 while $|S_{\ell}| > 4k$ do
4 \hspace{1em} sample every arm $a \in S_{\ell}$ for $Q_{\ell} = (12/\epsilon_{\ell}^2)(1/\mu_{\ell}) \log(6k/\delta_{\ell})$ times;
5 \hspace{1em} for each arm $a \in S_{\ell}$ do
6 \hspace{2em} its empirical value $\hat{\theta}(a)$ = the average of the $Q_{\ell}$ samples from $a$;
7 \hspace{2em} $a_1, \ldots, a_{|S_{\ell}|}$ = the arms sorted in non-increasing order of their empirical values;
8 \hspace{2em} $S_{\ell+1} = \{a_1, \ldots, a_{|S_{\ell}|/2}\}$;
9 \hspace{2em} $\epsilon_{\ell+1} = 3\epsilon_{\ell}/4$, $\delta_{\ell+1} = \delta_{\ell}/2$, $\mu_{\ell+1} = (1 - \epsilon_{\ell})\mu_{\ell}$, and $\ell = \ell + 1$;
10 return $S_{\ell}$;

Algorithm 3: Uniform Sampling (US)

1 input: $S$, $\epsilon$, $\delta$, $\mu_s$, $k$
2 sample every arm $a \in S$ for $Q = (96/\epsilon^2)(1/\mu_s) \log(4|S|/\delta)$ times;
3 for each arm $a \in S$ do
4 \hspace{1em} its US-empirical value $\hat{\theta}^{US}(a)$ = the average of the $Q$ samples from $a$;
5 \hspace{1em} $a_1, \ldots, a_{|S|}$ = the arms sorted in non-increasing order of their US-empirical values;
6 return $\{(a_1, \hat{\theta}^{US}(a_1)), \ldots, (a_k, \hat{\theta}^{US}(a_k))\}$
(worst case) Optimal bounds

<table>
<thead>
<tr>
<th>problem</th>
<th>sample complexity</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$-AS</td>
<td>$O\left(\frac{n}{\varepsilon^2} \frac{1}{\theta_k(B)} \log \frac{n}{\delta}\right)$</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>$O\left(\frac{n}{\varepsilon^2} \frac{1}{\theta_k(B)} \log \frac{k}{\delta}\right)$</td>
<td>new</td>
</tr>
<tr>
<td>lower bound</td>
<td>$\Omega\left(\frac{n}{\varepsilon^2} \log \frac{k}{\delta}\right)$</td>
<td>[11]</td>
</tr>
<tr>
<td></td>
<td>$\Omega\left(\frac{n}{\varepsilon^2} \frac{1}{\theta_k(B)} \log \frac{k}{\delta}\right)$</td>
<td>new</td>
</tr>
<tr>
<td>$k_{avg}$ -AS</td>
<td>$O\left(\frac{n}{\varepsilon^2} \frac{1}{\theta_{avg}(B)} \left(1 + \frac{\log(1/\delta)}{k}\right)\right)$</td>
<td>[14]</td>
</tr>
<tr>
<td>upper bound</td>
<td>$O\left(\frac{n}{\varepsilon^2} \frac{1}{\theta_{avg}(B)} \left(1 + \frac{\log(1/\delta)}{k}\right)\right)$</td>
<td>[16]</td>
</tr>
<tr>
<td></td>
<td>$\Omega\left(\frac{n}{\varepsilon^2} \frac{1}{\theta_{avg}(B)} \left(1 + \frac{\log(1/\delta)}{k}\right)\right)$</td>
<td>new</td>
</tr>
<tr>
<td>lower bound</td>
<td>$\Omega\left(\frac{n}{\varepsilon^2} \frac{1}{\theta_{avg}(B)} \left(1 + \frac{\log(1/\delta)}{k}\right)\right)$</td>
<td>[16]</td>
</tr>
<tr>
<td></td>
<td>$\Omega\left(\frac{n}{\varepsilon^2} \frac{1}{\theta_{avg}(B)} \left(1 + \frac{\log(1/\delta)}{k}\right)\right)$</td>
<td>new</td>
</tr>
</tbody>
</table>

Top-1 arm (PAC) [Even-Dar et al. 02]
We solve the average (additive) version in [Zhou, Chen, L ICML’14]
We extend the result to both (multiplicative) elementwise and average in [Cao, L, Tao, Li, NIPS’15]
Outline

- Online Learning
- Stochastic Multi-armed Bandits
  - UCB
  - Combinatorial Bandits
- Top-k Arm Identification
- Combinatorial Pure Exploration
- Best Arm Identification
A More General Problem

Combinatorial Pure Exploration

- A general combinatorial constraint on the feasible set of arms
  - Best-\(k\)-arm: the uniform matroid constraint
  - First studied by [Chen et al. NIPS14]

- E.g., we want to build a MST. But each time we obtain improved bounds for general matroid constraints
  - We obtain improved bounds for general matroid constraints
    - Our bounds even improve previous results on Best-\(k\)-arm

[Chen, Gupta, L. COLT’16]
Application

- A set of jobs
- A set of workers
- Each worker can only do one job
- Each job has a reward distribution

- Goal: choose the set of jobs with the largest total expected reward

Feasible sets of jobs that can be completed form a transversal matroid
Our Results

- A generalized definition of gap

$$\Delta_{e}^{\mathcal{M},\mu} := \begin{cases} 
\text{OPT}(\mathcal{M}) - \text{OPT}(\mathcal{M}_{S \setminus \{e\}}) & e \in \text{OPT}(\mathcal{M}) \\
\text{OPT}(\mathcal{M}) - (\text{OPT}(\mathcal{M}_{/\{e\}}) + \mu(e)) & e \notin \text{OPT}(\mathcal{M}) 
\end{cases}$$

- Exact identification
  - [Chen et al.]
    $$\left(\sum_{e \in S} \Delta_{e}^{-2} (\ln \delta^{-1} + \ln n + \ln \sum_{e \in S} \Delta_{e}^{-2})\right)$$

- Previous best-k-arm [Kalyanakrishnan]:
  $$O\left(\sum_{i=1}^{n} \Delta_{[i]}^{-2} (\ln \delta^{-1} + \ln \sum_{i=1}^{n} \Delta_{[i]}^{-2})\right)$$

- Ours:
  $$O\left(\sum_{e \in S} \Delta_{e}^{-2} (\ln \delta^{-1} + \ln k + \ln \ln \Delta_{e}^{-1})\right)$$

- Our result is even better than previous best-k-arm result
- Our result matches Karnin’ et al. result for best-1-arm
Our Results

- **PAC: Strong eps-optimality (stronger than elementwise opt)**
  - Ours: $O(n\varepsilon^{-2} \cdot (\ln k + \ln \delta^{-1}))$
  - Generalizes [Cao et al.][Kalyanakrishnan et al.]
  - Optimal: Matching the LB in [Kalyanakrishnan et al.]

- **PAC: Average eps-optimality**
  - Ours: $O(n\varepsilon^{-2}(1 + \ln \delta^{-1}/k))$. (under mild condition)
  - Generalizes [Zhou et al.]
  - Optimal  (under mild condition): matching the lower bound in [Zhou et al.]
Our technique

- What is more interesting is our technique
  - Sampling-and-Pruning technique
    - Originally developed by Karger, and used by Karger, Klein, Tarjan for the expected linear time MST

- High level idea (for MST)
  - Sample a subset of edges (uniformly and random, w.p. 1/100)
  - Find the MST $T$ over the sampled edges
  - Use $T$ to prune a lot of edges (w.h.p. we can prune a constant fraction of edges)
  - Iterate over the remaining edges
Outline

• Online Learning
• Stochastic Multi-armed Bandits
  • UCB
  • Combinatorial Bandits
• Top-k Arm Identification
• Combinatorial Pure Exploration
• Best Arm Identification
Best Arm Identification

• Some classical results:
  • Mannor-Tsitsiklis lower bound:
    \[ \Omega \left( \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \delta^{-1} \right) \]
    \[ \Delta_{[i]} = \mu[1] - \mu[i] \]
    It is an instance-wise lower bound
  • A PAC algorithm – Median Elimination [Even-Dar et al.]
    • Find an \( \epsilon \)-optimal arm using \( \epsilon^{-2} n \log \delta^{-1} \) samples
    • The bound is worst-case optimal
Are we done? – a misclaim

<table>
<thead>
<tr>
<th>Source</th>
<th>Sample Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even-Dar et al. [12]</td>
<td>$\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln n + \ln \Delta_{[i]}^{-1} \right)$</td>
</tr>
<tr>
<td>Gabillon et al. [16]</td>
<td>$\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \sum_{i=2}^{n} \Delta_{[i]}^{-2} \right)$</td>
</tr>
<tr>
<td>Jamieson et al. [19]</td>
<td>$\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \left( \sum_{j=2}^{n} \Delta_{[j]}^{-2} \right) \right)$</td>
</tr>
<tr>
<td>kalyanakrishnan et al. [23]</td>
<td>$\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \sum_{i=2}^{n} \Delta_{[i]}^{-2} \right)$</td>
</tr>
<tr>
<td>Jamieson et al. [19]</td>
<td>$\ln \delta^{-1} \cdot \left( \ln \ln \delta^{-1} \cdot \sum_{i=2}^{n} \Delta_{[i]}^{-2} + \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \Delta_{[i]}^{-1} \right)$</td>
</tr>
<tr>
<td>Karnin et al. [24], Jamieson et al. [20]</td>
<td>$\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \ln \Delta_{[i]}^{-1} \right)$</td>
</tr>
</tbody>
</table>

Mannor-Tsitsiklis lower bound: $\Omega \left( \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \delta^{-1} \right)$

Farrell’s lower bound (2 arms): $\Delta_{[2]}^{-2} \ln \ln \Delta_{[2]}^{-1}$

**Attempting to believe**: Karnin’s upper bound is tight

Jamieson et al.: “The procedure cannot be improved in the sense that the number of samples required to identify the best arm is within a constant factor of a lower bound based on the law of the iterated logarithm (LIL)”.
Are we done? – a misclaim

<table>
<thead>
<tr>
<th>Source</th>
<th>Sample Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Even-Dar et al. [12]</td>
<td>(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln n + \ln \Delta_{[i]}^{-1} \right))</td>
</tr>
<tr>
<td>Gabillon et al. [16]</td>
<td>(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \sum_{i=2}^{n} \Delta_{[i]}^{-2} \right))</td>
</tr>
<tr>
<td>Jamieson et al. [19]</td>
<td>(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \left( \sum_{j=2}^{n} \Delta_{[j]}^{-2} \right) \right))</td>
</tr>
<tr>
<td>kalyanakrishnan et al. [23]</td>
<td>(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \sum_{i=2}^{n} \Delta_{[i]}^{-2} \right))</td>
</tr>
<tr>
<td>Jamieson et al. [19]</td>
<td>(\ln \delta^{-1} \cdot \left( \ln \ln \delta^{-1} \cdot \sum_{i=2}^{n} \Delta_{[i]}^{-2} + \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \Delta_{[i]}^{-1} \right))</td>
</tr>
<tr>
<td>Karnin et al.[24], Jamieson et al.[20]</td>
<td>(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \ln \Delta_{[i]}^{-1} \right))</td>
</tr>
<tr>
<td>This paper (Thm 2.5)</td>
<td>(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \ln \min(n, \Delta_{[i]}^{-1}) + \Delta_{[2]}^{-2} \ln \ln \Delta_{[2]}^{-1} \right))</td>
</tr>
<tr>
<td>This paper (clustered instances) Thm B.22</td>
<td>(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \left( \ln \delta^{-1} + \Delta_{[2]}^{-2} \ln \ln \Delta_{[2]}^{-1} \right))</td>
</tr>
</tbody>
</table>

Mannor-Tsitsiklis lower bound: \(\Omega \left( \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \delta^{-1} \right)\)

Farrell’s lower bound (2 arms): \(\Delta_{[2]}^{-2} \ln \ln \Delta_{[2]}^{-1}\)

**Attempting to believe**: Karnin’s upper bound is tight
- Of course, to completely close the problem, we need to show the remaining generalization from Farrell’s LB to \(n\) arms: \(\sum \Delta_{[i]}^{-2} \log \log \Delta_{[i]}^{-1}\)
New Upper and Lower Bounds

- Our new upper bound (strictly better than Karnin’s)

\[
O\left(\Delta_{[2]}^{-2} \ln \ln \Delta_{[2]}^{-1} + \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \delta^{-1} + \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \ln \min(n, \Delta_{[i]}^{-1})\right)
\]

Farrell’s LB  M-T LB  LnLn n term seems strange……..
New Upper and Lower Bounds

• Our new upper bound (strictly better than Karnin’s)

\[ O\left(\Delta_{[2]}^{-2} \ln \ln \Delta_{[2]}^{-1} + \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \delta^{-1} + \sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \ln \min(n, \Delta_{[i]}^{-1})\right) \]

Farrell’s LB \hspace{2cm} M-T LB \hspace{2cm} \ln \ln n \text{ term seems strange………}

• It turns out the \( \ln \ln n \) term is fundamental.

• Our new lower bound (not instance-wise)

\[ \Omega\left(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \ln n\right) \]
Open Question

• (almost) Instance optimal algorithm for best arm

\[ H_i = \sum_{u \in G_i} \Delta_{[u]}^{-2} \]

• Gap Entropy: \[ \text{Ent}(I) = \sum_{G_i \neq \emptyset} p_i \log p_i^{-1} \]
\[ p_i = H_i / \sum_j H_j. \]

• Gap Entropy Conjecture:
  • An instance-wise lower bound \[ \mathcal{L}(I, \delta) = \Theta \left( H(I)(\ln \delta^{-1} + \text{Ent}(I)) \right). \]
\[ H(I) = \sum_{i=2}^{n} \Delta_{[i]}^{-2}. \]
  • An algorithm with sample complexity:
\[ O \left( \mathcal{L}(I, \delta) + \Delta_{[2]}^{-2} \ln \ln \Delta_{[2]}^{-1} \right). \]
Thanks.
lapordge@gmail.com
Reference

- Leon Bottou, Online Learning and Stochastic Approximations
- T Cover, Universal Portfolios. Mathematical finance, 1991
- S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. JMLR, 2004
- Y. Zhou, X. Chen, and J. Li. Optimal pac multiple arm identification with applications to crowdsourcing. In ICML 2014
- W. Cao, J. Li, Y. Tao, and Z. Li. On top-k selection in multi-armed bandits and hidden bipartite graphs. In NIPS 2015

Some materials about MW from Daniel Golovin’s slides; Some material about UCB from Sumeet Katariya’s slides