2016 NDBC

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

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

the environment plays f_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)

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):
TTHHTTTHTH……

time	1	2	3	4	…	T
Expert 1	T	T	H	T	…	T
Expert 2	H	T	T	H	…	H
Expert 3	T	T	T	T	…	T
….						

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:

$$R_T = \sum_{t=1}^{T} c_t(x_t) - \sum_{t=1}^{T} c_t(x^*)$$
$$\text{where } x^* = \text{argmin}_{x \in X} \sum_{t=1}^{T} c_t(x)$$

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

- Hedge Algorithm (aka mulplicative 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 θ_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:
  - Multi-armed Bandit Allocation Indices, John Gittins, Kevin Glazebrook, Richard Weber, 2011
  - 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

Optimism in the Face of Uncertainty

- At time t, construct most optimistic estimate for each arm
  $V_{i,t-1} = \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log t}{T_i(t-1)}}$
- Play arm with max upper bound.
  i.e. play $I_t \in \underset{i \in \{1,\cdots,K\}}{\arg\max} \{V_{i,t-1}\}$
- Proof based on Hoeffding's inequality

$T_i(t)$: 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., $reward = \max_{i \in S} X_i$)
- 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 $\mathrm{E}[\max_{i \in S} X_i]$)

Combinatorial Bandit - SDCB

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

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

- Gap-independent regret

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

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 $\mu_{[1]}, \mu_{[n]}, .., \mu_{[n]}$
- 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

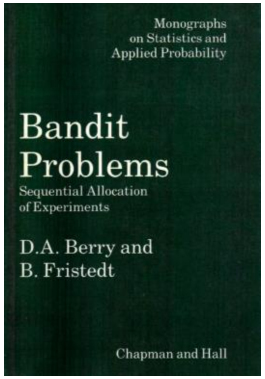
- Regret Minimization
  - Maximizing the cumulative reward

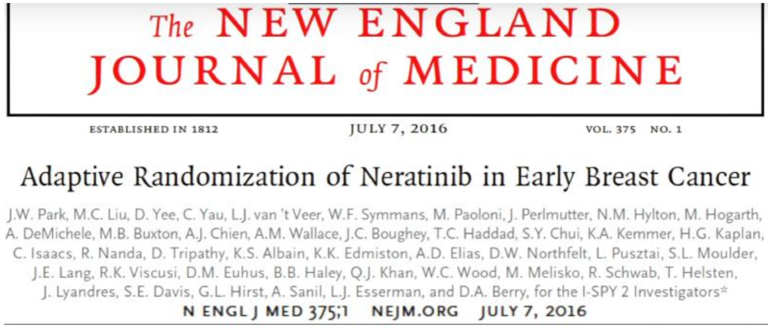
- Best/top-k arm identification
  - Find out the best arm using as few samples as possible

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**



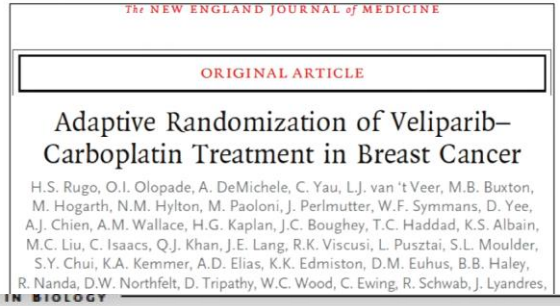

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JULY 7, 2016 VOL. 375 NO. 1

Adaptive Randomization of Neratinib in Early Breast Cancer

J.W. Park, M.C. Liu, D. Yee, C. Yau, L.J. van 't Veer, W.F. Symmans, M. Paoloni, J. Perlmutter, N.M. Hylton, M. Hogarth, A. DeMichele, M.B. Buxton, A.J. Chien, A.M. Wallace, J.C. Boughey, T.C. Haddad, S.Y. Chui, K.A. Kemmer, H.G. Kaplan, C. Isaacs, R. Nanda, D. Tripathy, K.S. Albain, K.K. Edmiston, A.D. Elias, D.W. Northfelt, L. Pusztai, S.L. Moulder, J.E. Lang, R.K. Viscusi, D.M. Euhus, B.B. Haley, Q.J. Khan, W.C. Wood, M. Melisko, R. Schwab, T. Helsten, J. Lyandres, S.E. Davis, G.L. Hirst, A. Sanil, L.J. Esserman, and D.A. Berry, for the I-SPY 2 Investigators*

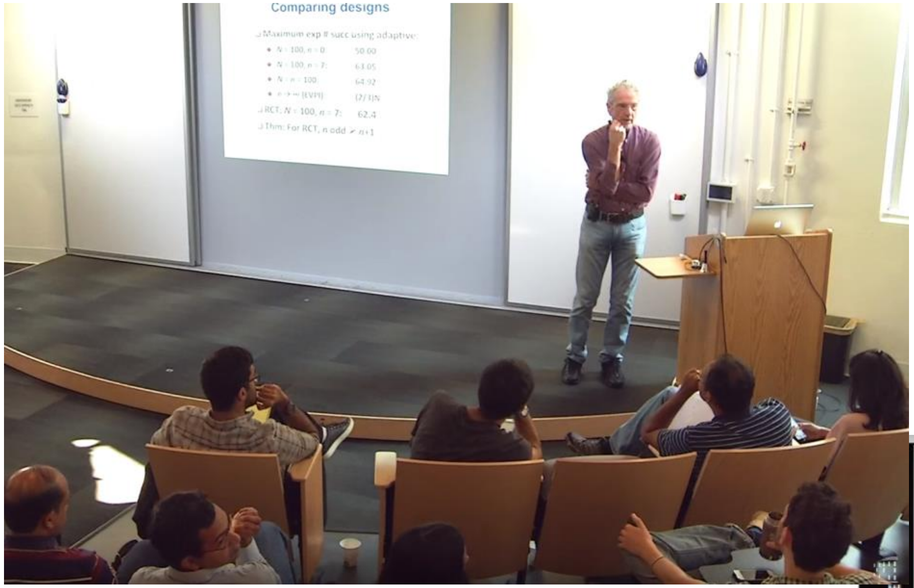
N ENGL J MED 375;1 NEJM.ORG JULY 7, 2016


The NEW ENGLAND JOURNAL of MEDICINE

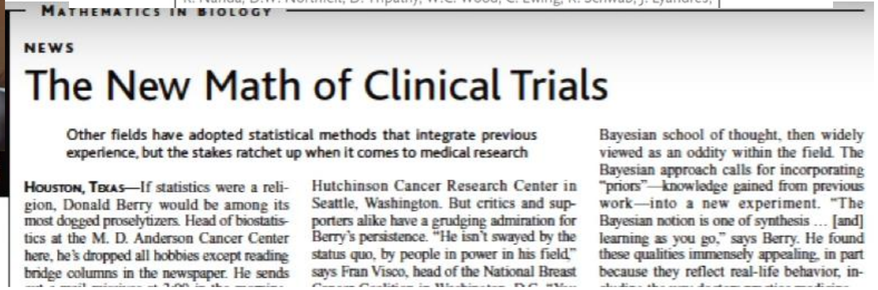
ORIGINAL ARTICLE

Adaptive Randomization of Veliparib–Carboplatin Treatment in Breast Cancer

H.S. Rugo, O.I. Olopade, A. DeMichele, C. Yau, L.J. van 't Veer, M.B. Buxton, M. Hogarth, N.M. Hylton, M. Paoloni, J. Perlmutter, W.F. Symmans, D. Yee, A.J. Chien, A.M. Wallace, H.G. Kaplan, J.C. Boughey, T.C. Haddad, K.S. Albain, M.C. Liu, C. Isaacs, Q.J. Khan, J.E. Lang, R.K. Viscusi, L. Pusztai, S.L. Moulder, S.Y. Chui, K.A. Kemmer, A.D. Elias, K.K. Edmiston, D.M. Euhus, B.B. Haley, R. Nanda, D.W. Northfelt, D. Tripathy, W.C. Wood, C. Ewing, R. Schwab, J. Lyandres,



Don Berry, University of Texas
MD Anderson Cancer Center


MATHEMATICS IN BIOLOGY

NEWS

The New Math of Clinical Trials

Other fields have adopted statistical methods that integrate previous experience, but the stakes ratchet up when it comes to medical research

HOUSTON, TEXAS—If statistics were a religion, Donald Berry would be among its most dogged proselytizers. Head of biostatistics at the M. D. Anderson Cancer Center here, he's dropped all hobbies except reading bridge columns in the newspaper. He sends

Hutchinson Cancer Research Center in Seattle, Washington. But critics and supporters alike have a grudging admiration for Berry's persistence. "He isn't swayed by the status quo, by people in power in his field," says Fran Visco, head of the National Breast

Bayesian school of thought, then widely viewed as an oddity within the field. The Bayesian approach calls for incorporating "priors"—knowledge gained from previous work—into a new experiment. "The Bayesian notion is one of synthesis … [and] learning as you go," says Berry. He found these qualities immensely appealing, in part because they reflect real-life behavior, in-

Applications

- Crowdsourcing:
- Workers are noisy

0.95

0.99

0.5

- 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?

Top-K Arm Identification	
Worker	Bernoulli arm with mean θ_i (θ_i: i-th worker's reliability)
Test with golden task	Obtain a binary-valued sample (correct/wrong)

Naïve Solution

- ϵ-approximation: the ith arm in our output is at most ϵ worse than the the ith largest arm
- Uniform Sampling

Sample each coin M times

Pick the K coins with the largest empirical means

empirical mean: #heads/M

How large M needs to be (in order to achieve ϵ-approximation)??

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

So the total number of samples is O(nlogn)

Naïve Solution

Uniform Sampling

- With M=O(logn), we can get an estimate θ_i' for θ_i such that $|\theta_i - \theta_i'| \le \epsilon$ 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, ……………………, 0.5 (a million coins with mean 0.5)
  - Consider a coin with mean 0.5,

    Pr[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 M_i : *increasing expoentially*

Eliminate one quarter arms

Until less 4k arms

When $\mathrm{n} \leq 4k$， use uniform sampling

We can find a solution with additive error ϵ

Decrease ϵ， until proper termination condition

Our algorithm

Algorithm 1: ME-AS

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

Algorithm 2: Median Elimination (ME)

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

Algorithm 3: Uniform Sampling (US)

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

(worst case) Optimal bounds

Table 1: Comparison of our and previous results (all bounds are in expectation)

problem		sample complexity	source
k-AS	upper bound	$O\left(\frac{n}{\epsilon^2}\frac{1}{\theta_k(B)}\log\frac{n}{\delta}\right)$	[14]
		$O\left(\frac{n}{\epsilon^2}\frac{1}{\theta_k(B)}\log\frac{k}{\delta}\right)$	**new**
	lower bound	$\Omega\left(\frac{n}{\epsilon^2}\log\frac{k}{\delta}\right)$	[11]
		$\Omega\left(\frac{n}{\epsilon^2}\frac{1}{\theta_k(B)}\log\frac{k}{\delta}\right)$	**new**
k_{avg}-AS	upper bound	$O(\frac{n}{\epsilon^2}\frac{1}{\theta_k(B)}\log\frac{n}{\delta})$	[14]
		$O\left(\frac{n}{\epsilon^2}\frac{1}{(\theta_{\text{avg}}(B))^2}\left(1+\frac{\log(1/\delta)}{k}\right)\right)$	[16]
		$O\left(\frac{n}{\epsilon^2}\frac{1}{\theta_{\text{avg}}(B)}\left(1+\frac{\log(1/\delta)}{k}\right)\right)$	**new**
	lower bound	$\Omega\left(\frac{n}{\epsilon^2}\left(1+\frac{\log(1/\delta)}{k}\right)\right)$	[16]
		$\Omega\left(\frac{n}{\epsilon^2}\frac{1}{\theta_{\text{avg}}(B)}\left(1+\frac{\log(1/\delta)}{k}\right)\right)$	**new**

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]

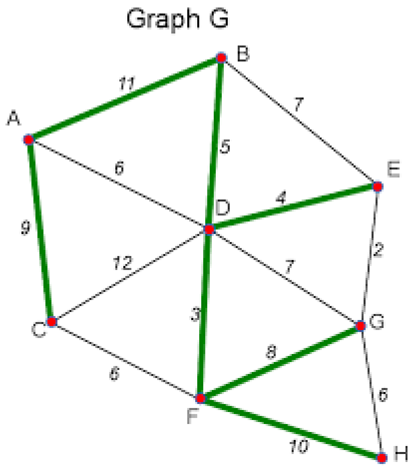
Outline

- 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 get a noisy estimate of the true cost of each edge



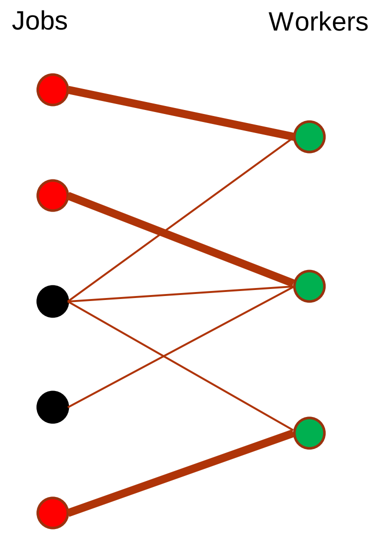
- We obtain improved bounds for general matroid constaints
  - Our bounds even improve previous results on Best-k-arm

[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

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

- 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]:

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

  - 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:

$$\Omega\left(\sum_{i=2}^{n} \Delta_{[i]}^{-2} \ln \delta^{-1}\right) \qquad \Delta_{[i]} = \mu_{[1]} - \mu_{[i]}$$

    It is an instance-wise lower bound
  - A PAC algorithm – Median Elimination [Even-Dar et al.]
    - Find an ϵ-optimal arm using $\epsilon^{-2} n \log \delta^{-1}$ samples
    - The bound is worst-case optimal

Are we done? – a misclaim

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

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

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

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 of the Farrell's LB to n arms: $\sum \Delta_{[i]}^{-2} \log\log \Delta_{[i]}^{-1}$

Misclaim!

New Upper and Lower Bounds

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

$$O\Big(\underbrace{\Delta_{[2]}^{-2}\ln\ln\Delta_{[2]}^{-1}}_{\text{Farrell's LB}} + \underbrace{\sum\nolimits_{i=2}^{n}\Delta_{[i]}^{-2}\ln\delta^{-1}}_{\text{M-T LB}} + \underbrace{\sum\nolimits_{i=2}^{n}\Delta_{[i]}^{-2}\ln\ln\min(n,\Delta_{[i]}^{-1})}_{\ln\ln n \text{ term seems strange........}}\Big)$$

[Chen, Li. ArXiv 15]

New Upper and Lower Bounds

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

$$O\Big(\underbrace{\Delta_{[2]}^{-2}\ln\ln\Delta_{[2]}^{-1}}_{\text{Farrell's LB}} + \underbrace{\sum\nolimits_{i=2}^{n}\Delta_{[i]}^{-2}\ln\delta^{-1}}_{\text{M-T LB}} + \underbrace{\sum\nolimits_{i=2}^{n}\Delta_{[i]}^{-2}\ln\ln\min(n,\Delta_{[i]}^{-1})}_{\text{lnln}n\text{ term seems strange……..}}\Big)$$

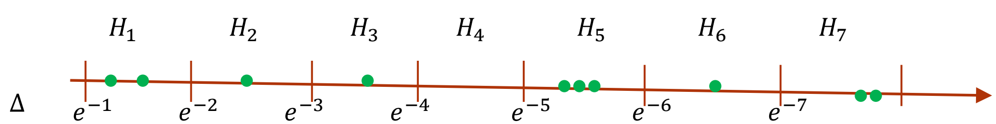
- It turns out the lnln$n$ term is fundamental.
- Our new lower bound (not instance-wise)

$$\Omega\left(\sum\nolimits_{i=2}^{n}\Delta_{[i]}^{-2}\ln\ln n\right)$$

Open Question

- (almost) Instance optimal algorithm for best arm

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



- Gap Entropy: $\mathsf{Ent}(I) = \sum_{G_i \neq \emptyset} p_i \log p_i^{-1}.$ $\quad p_i = H_i / \sum_j H_j.$

- Gap Entropy Conjecture:
  - An instance-wise lower bound $\mathcal{L}(I,\delta) = \Theta\left(H(I)(\ln \delta^{-1} + \mathsf{Ent}(I))\right).$

    $H(I) = \sum_{i=2}^{n} \Delta_{[i]}^{-2}.$
  - 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

- [book] Cesa-Bianchi, Nicolo, and Gábor Lugosi. Prediction, learning, and games. Cambridge university press, 2006.
- Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs, JMLR2002
- Leon Bottou, Online Learning and Stochastic Approximations
- T Cover, Universal Portfolios. Mathematical finance, 1991
- Farrell. Asymptotic behavior of expected sample size in certain one sided tests. The Annals of Mathematical Statistics 1964
- E. Even-Dar, S. Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In COLT 2002
- S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. JMLR, 2004
- Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In ICML, 2013
- K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. COLT, 2014
- S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits.In NIPS, 2014
- 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
- L. Chen, **J. Li**. On the Optimal Sample Complexity for Best Arm Identification, ArXiv, 2016
- L. Chen, A. Gupta, and **J. Li**. Pure exploration of multi-armed bandit under matroid constraints. In COLT2016.
- W. Chen, W. Hu, F. Li, **J. Li**, Y. Liu, P. Lu. Combinatorial Multi-Armed Bandit with General Reward Functions, In NIPS 2016.

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