Combinatorial Semi-Bandits with Knapsacks

Karthik Abinav Sankararaman
(University of Maryland, College Park)
Aleksandrs Slivkins
(Microsoft Research, New York City)

Problem Formulation

- n atoms, d resources with budget B for each
- T rounds
- For t=1, 2, …T
  - ALG chooses subset of atoms (a.k.a. arm) F_t
  - Nature chooses "outcome matrix": reward and consumption for each resource from fixed distribution over "outcome matrices"
  - ALG observes reward and consumption for all atoms in F_t
    - Reward and consumption for each resource = sum over F_t
  - Stop when one resource exhausted

Constrained to be independent set in matroid

Special Cases: Semi-Bandits (no resources), Bandits with Knapsacks ($|F_t|$=1)

Minimize regret : OPT - E[Total reward of ALG]

OPT

Best dynamic policy if outcome distribution is known

SemiBwK-RRS

- Let LCB=0, UCB=1 when #samples=0
- For t=1, 2, .. , T
  - For each atom: re-compute UCB for rewards and LCB for consumption
  - Solve LP with these estimates
  - Round the LP solution using "Randomized Rounding Schemes" from prior work
    - Same expectation, negative correlation -> Chernoff-like bounds
  - Pull the atoms in this set and observe feedback

Running Time

$$O(\text{Time to solve LP})+\underbrace{O(\text{Time to obtain a random feasible solution for matroids})}_{O(n^2)}$$

$$\begin{array}{ll} \text{maximize} & \mu_{\mathbf{t}}^{+}\cdot\mathbf{x} \\ \text{subject to} & \mathbf{C}_t^{-}(j)\cdot\mathbf{x}\le\frac{B(1-\epsilon)}{T}, \quad j\in[d] \\ & \mathbf{x}\in\mathcal{P} \end{array}$$

$$\mathbb{E}\left[\prod_{i\in S} X_i\right]\le\prod_{i\in S}\mathbb{E}[X_i] \quad \forall S\subseteq[m] \qquad (2.1)$$

$$\mathbb{E}\left[\prod_{i\in S}(1-X_i)\right]\le\prod_{i\in S}\mathbb{E}[1-X_i] \; \forall S\subseteq[m] \qquad (2.2)$$

Applications

1. Dynamic Pricing with Limited Supply:

   B copies each of k products, T rounds. Time t, buyer draws a valuation vector from a fixed distribution. Algorithm assigns prices from finite set S for each of the k products. Buyer buys products with valuation > offered price. Find the "right" price.

2. Dynamic Assortment:

   B copies of d products with fixed price, T rounds. Time t, buyer draws a valuation vector from a fixed distribution. Algorithm shows a subset of k products. Buyer buys products with valuation > fixed price. Find the "right" subset to sell.

- Exponential improvement over naive BwK
- Compared to using [Agrawal, Devanur '16]
  - Factor of $(k|S|)^{1.5}$ improvement for (1)
  - Factor of d/k improvement for (2)

Prior Work

Bandits with Knapsacks (BwK)

- Special Cases
  [Guha, Munagala '07], [Gupta et al., '11], [Tran-Tranh et al., '12], [Combes et al., '15]
- First fully general model
  [Badanidiyuru et al., '13]
- Subsequent follow-up
  [Agrawal, Devanur '14], [Badanidiyuru, et al., '14], [Agrawal, et al., '16], [Agrawal, Devanur '16]

Combinatorial Semi-Bandits

- Adversarial
  [Gyorgy et al., '07]
- i.i.d. setting
  [Gai et al., '10], [Chen, et al., '13], [Kveton, et al., '14], [Kveton, et al., '15], [Combes, et al., '15]
- Subsequent follow-up
  [Kveton et al., '14], [Wen et al., '15], [Krishnamurthy et al., '16]

Open Directions

- Adversarial BwK – rewards and/or resources chosen by adversary
  - Extends adversarial bandits
  - Seems really hard, even for (very) special cases
  - Sub linear regret impossible -> comp. ratio?
- BwK + other classical bandit models
  - Contextual bandits + semi-bandits + BwK: natural direction
  - Prior work combined any 2 out of these 3

Challenges

- Bandits
  - Exploration-Exploitation trade-off
  - Adaptive exploration - remove definite "bad" arms, explore only uncertain arms
- Semi-Bandits
  - Handling exponentially many actions in terms of regret and running time
  - Handling additional feedback
- BwK
  - Adaptive Exploration to resource setting?
  - Exploration consumes resources; Good reward more consumption vs. less reward less consumption
  - OPT no longer best expected per-round reward; Best dynamic policy
- Newer challenges in Semi-BwK
  - Deal with distribution over subsets of atoms
  - In BwK "just" distribution over atoms.

Main Theoretical Result

- UCB-based algorithm achieving the following bound

There exists a randomized algorithm for this problem which achieves the following regret

$$\tilde{O}(\sqrt{n}(OPT/\sqrt{B}+\sqrt{T+OPT}))$$

with the following assumptions $B>3(\alpha n+\sqrt{\alpha n T})$ where $\alpha=\Theta(\log(ndT))$

- "Shape" of regret consistent with BwK literature
- Optimal for special cases: BwK, Combinatorial Semi-Bandits
- Orders of magnitude improvement over using prior work to this model
- Matroid assumption general enough for many applications
  - Exponential improvement over using naive BwK

Proof Overview

- Step 1: Optimal value of LP at each time-step at least 1/T * OPT w.h.p.
- Step 2: High probability bound on difference between total reward of algorithm and optimal value
  - Negative correlation implies strong concentration – exploit this!
  - Combine with concentration bounds from [Babiaoff et al (EC '12)]

Lemma 6.9. Consider SemiBwK without stopping. Then with probability at least $1-nTe^{-\Omega(\alpha)}$:

$$\left|\sum_{t\in[T]} r_t-\sum_{t\in[T]}\mu_{\mathbf{t}}^{+}\cdot\mathbf{x_t}\right|\le O\left(\sqrt{\alpha n\sum_{t\in[T]} r_t}+\sqrt{\alpha n T}+\alpha n\right)$$

- Step 3: W.H.P. algorithm doesn't run out of resources in T time-steps
  - Regret analyzed conditioned on this clean event

Lemma 6.10. Consider SemiBwK without stopping. Then with probability at least $1-nTe^{-\Omega(\alpha)}$:

$$\forall j\in[d] \quad \left|\sum_{t\in[T]}\chi_t(j)-\sum_{t\in[T]}\mathbf{C_t^-(j)}\cdot\mathbf{x_t}\right|\le\sqrt{\alpha n B_\epsilon}+\alpha n+\sqrt{\alpha n T}.$$

From LP constraints we have, $\sum_{t\le T}\mathbf{C_t^-(j)}\cdot\mathbf{x_t}\le(1-\epsilon)B$.

Hence combining this with Lemma 6.10, we have

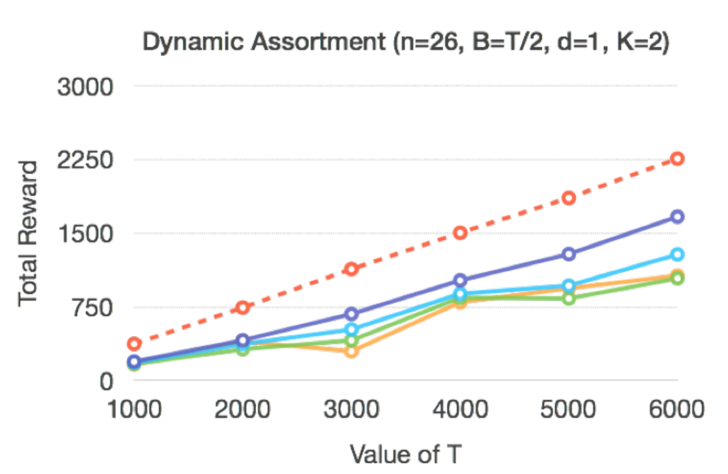
$$\sum_{t\le T}\mathbf{C_t(j)}\cdot\mathbf{Y_t}\le(1-\epsilon)B+\epsilon B\le B.$$

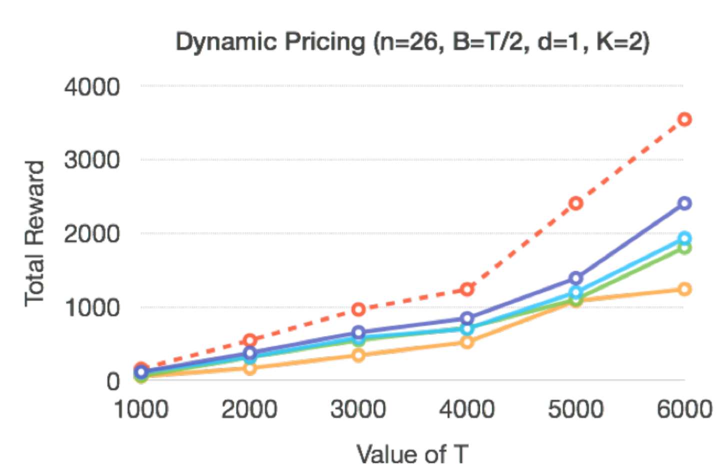
- Concentration Theorem

Let $\mathcal{Z}_T=\{\zeta_{t,a}: a\in\mathcal{A}, t\in[T]\}$ be a family of random variables taking values in $[0,1]$. Assume random variables $\{\zeta_{t,a}: a\in\mathcal{A}\}$ satisfy property (2.1) given $\mathcal{Z}_{t-1}$ and have expectation $\frac{1}{2}$ given $\mathcal{Z}_{t-1}$, for each round t. Let $Z=\frac{1}{nT}\sum_{a\in\mathcal{A},t\in[T]}\zeta_{t,a}$ be the average. Then for some absolute constant c,

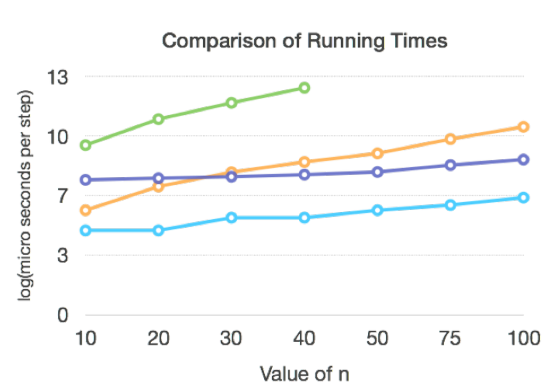
$$\Pr[Z\ge\tfrac{1}{2}+\eta]\le c\cdot e^{-2m\eta^2} \qquad (\forall\eta>0).$$

Numerical Results





SemiBwK-RRS OMM linBwK-sample PD-BwK LP-OPT



Karthik A. Sankararaman
kabinav@cs.umd.edu
Aleksandrs Slivkins
slivkins@microsoft.com

Full Paper
https://arxiv.org/abs/1705.08110