Bandits with Constraints

Paul Lagrée

Université Paris-Saclay

February 12, 2016

Overview

1. Motivations
2. Bandits with Knapsacks
   - Model
   - Problem-Independent Analysis
   - Problem-Dependent Analysis
   - Single limited resource
   - Other
3. Bandits with Concave Rewards and Convex Knapsacks

Motivations

- MAB algorithms gradually focus on best arms
- Conversely, exploration is mostly done at the beginning of the process of the algorithm.
- **Question:** How to enforce continuous exploration (keeping a meaningful problem formulation)?

Introduce **contraints** in the problem formulation.

Motivations – Examples

- **Advertising:** daily-budget on each ad
  Consumption can be stochastic (cost per click) or deterministic (cost per impression)

- **Dynamic pricing with limited supply:** At each time t, fix a price p_t to item limited in supply by B. The client buys or leaves.

Model – Bandits with Knapsacks BwK

At each time t:

- Choose an arm $A_t \in \{1, \dots, K\}$ given past observations
- Observe a random **reward** $X_t \in [0, 1]$ and a **consumption vector** $(C_{t,1}, \dots, C_{t,D}) \in [0, 1]^D$
- Each resource j is **constrained** by an initial endowment $B^j \geq 0$
- The decision maker can pull arms as long as he does not run out of **any resource**

Regret

We define τ the random stopping time as

$$\tau = \min \left\{ t \in \mathbb{N} \ : \ \exists j \in \{1, \dots, D\}, \sum_{s=1}^{t} C_{s,j} > B^j \right\}$$

and $\text{OPT} = \mathbb{E}[R_{OPT}(B^1, \dots, B^D)]$ the expected total reward of the **optimal dynamic policy**.

Definition – Regret

$$Reg_{B^1,\dots,B^D} = \text{OPT} - \mathbb{E}[\sum_{t=1}^{\tau-1} X_t]$$

Computational complexity

- Finding the optimal pulling strategy, **given the latent distributions**, is a challenge in itself [Papadimitriou, 1999]
- How to evaluate the regret?

Theorem (Upper Bound [Badanidiyuru, 2013])

The expected total payoff of the oracle policy is upper bounded by the optimal value of the following linear program

$$\sup_{\boldsymbol{\xi} \geq 0} \boldsymbol{\mu} \cdot \boldsymbol{\xi} \text{ subject to } \boldsymbol{c^j} \cdot \boldsymbol{\xi} \leq B^j, \quad j = 1, \dots, D$$

Lower bound – Distribution Independent

Theorem (Lower Bound [Badanidiyuru, 2013])

Any algorithm for BwK *muster incur regret*

$$\Omega\left(\min\left(\mathrm{OPT}, \mathrm{OPT}\sqrt{\frac{K}{B}} + \sqrt{K\,\mathrm{OPT}}\right)\right)$$

where B *denotes the smallest budget.*

Algorithm Idea I

Algorithm [Agrawal and Devanur, 2014]

1. Find an optimal solution $(p_{t,k})$ to

$$\sup_{\boldsymbol{p}} \sum_{k=1}^{K} (\hat{\mu}_{k,t} + \epsilon_{k,t}) p_k$$

$$\text{s.t. for all } j, (\hat{c}^j_{k,t} - \epsilon_{k,t}) p_k \leq (1-\epsilon)\frac{B^j}{T} \text{ and } \sum_k p_k = 1$$

2. Select arm A_t according to $(p_{k,t})$ probabilities.

Theorem (Upper Bound [Agrawal and Devanur, 2014])

For a well chosen ϵ, the regret is bounded by $\tilde{O}(\mathrm{OPT}\sqrt{K/B} + \sqrt{K\,\mathrm{OPT}})$ with high probability.

Algorithm Idea II

Algorithm [Flajolet, 2015]

1. Find an optimal solution $\boldsymbol{\xi_t} = (\xi_{t,1}, \ldots, \xi_{t,K})$ to

$$\sup_{\boldsymbol{\xi}} \sum_{k=1}^{K} (\hat{\mu}_{k,t} + \beta \epsilon_{k,t}) \xi_k$$

$$\text{s.t. for all } j, \sum_{k} \hat{c}^j_{k,t} \xi_k \leq B^j$$

2. Identify arms involved in the optimal basic solution (k such that $\xi_{k,t} > 0$ and play one of these arms (to be specified).

Parameters will be specified according to the problem specification.

Algorithm Idea II – Comments

- Optimistic **but** only on the rewards
- This can be interpreted as an extension of index-based policies. If we denote $\mathcal{F}_t$ the set of basic feasible solutions.
  1. Step 1: $\boldsymbol{\xi}_t = \arg\max_{\xi \in \mathbf{F}_t} \sum_k \xi_{k,t} \hat{\mu}_{k,t} + \beta \sum_k \boldsymbol{\xi}_{k,t} \epsilon_{k,t}$
  2. Step 2: if only resource is time, unambiguous step: only one active constraint.

Single limited resource

- **Single resource** with consumption limited by a global budget B
- If resource is time: **original** problem
- **Unique** best arm: fixed optimal policy
- **Attention:** Resource consumption can be **stochastic**

Theorem ([Flajolet, 2015])

The regret of algorithm II is upper bounded with order of $O(\ln(B))$.

Other specific problems

[Flajolet, 2015] also give upper bounds for the following scenarios:

- Single limited resource and time horizon
- Arbitrarly many resources with **deterministic** costs

Model – BwCR

At each time step t,

- Select arm $A_t \in \{1, \dots, K\}$
- Observe vector $\boldsymbol{v}_t \in [0,1]^D$ such that $\mathbb{E}[\boldsymbol{v}_t | A_t] = V_{\bullet, A_t}$ and where V is a $D \times K$ matrix

Goal:

1. Make the average of the observed vectors $\frac{1}{T}\sum_t \boldsymbol{v}_t$ lie in a given **convex** set S
2. At the same time, maximize $f(\frac{1}{T}\sum_t \boldsymbol{v}_t)$ for a given **concave** function.

Special Case

- With $\boldsymbol{v}_t = (X_t, C_{t,1}, \ldots, C_{t,C})$ we can **recover** the orginal BwK problem.
- Take f the function which associate the **first component** of $\boldsymbol{v}$
- and **define** S by the C constraints: $S = \{\boldsymbol{x} : \boldsymbol{x}_{-1} \leq \frac{B}{T}\}$

Discussion.

References

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins (2013)
Bandits with Knapsacks
Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, 207 – 216.

Shipra Agrawal and Nikhil R. Devanur (2014)
Bandits with Concave Rewards and Convex Knapsacks
Proceedings of the Fifteenth ACM Conference on Economics and Computation, 989 – 1006.

Papadimitriou and Tsitsiklis (1999)
The complexity of optimal queuing network control
Mathematics of Operations Research, 293 – 305.

Arthur Flajolet and Patrick Jaillet (2015)
Low regret bounds for Bandits with Knapsacks
Work in progress.

Richard Combes, Chong Jiang and Rayadurgam Srikant (2015)
Bandits with Budgets: Regret Lower Bounds and Optimal Algorithms
SIGMETRICS, 245 – 257.