Bandits with Knapsacks via Online Linear Programming: A (Problem-Dependent) Logarithmic Regret Bound

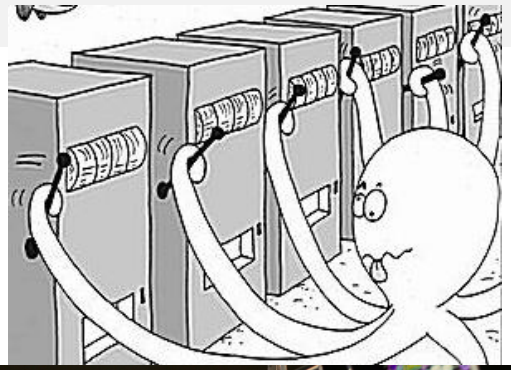
Yinyu Ye

Stanford University

Jan. 2021

(Joint work with Xiaocheng Li and Chunlin Sun)

Multi-Armed Bandits



A fundamental reinforcement learning problem

First studied by Robbins in 1952

Imagine a gambler at a row of slot machines

Each "arm" represents a candidate system

Objective is to identify the best system through experimentation in the most effective way so as to minimize the number of arms pulled.

In this talk, we examine the problem under a setting where there is a more complicated cost structure associated with playing each arm/simulating candidate system

Bandits with Knapsacks

Horizon: T time periods (T known a priori)

Bandits: m arms. Each arm i with an unknown mean reward μ_i, $i \in [m]$

Knapsacks: d types of resources. The total resource capacity $\boldsymbol{B} \in \mathbb{R}^d$.
Each arm i with an unknown mean resource consumption $\boldsymbol{c}_i \in \mathbb{R}^d$, $i \in [m]$

At each time $t \in [T]$, an arm i is selected to play. The realized reward r_t and cost $\boldsymbol{C}_t$ satisfying

$$\mathbb{E}[r_t|i] = \mu_i, \mathbb{E}[\boldsymbol{C}_t|i] = \boldsymbol{c}_i.$$

Goal: To maximize the total reward subject to the resource capacity

Deterministic Linear Program (LP)

With mean reward $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)$ and mean cost $\boldsymbol{C} = (\boldsymbol{c}_1, ..., \boldsymbol{c}_m)$ of all arms, consider the following LP,

$$\max_{\boldsymbol{x}} \sum_{i=1}^{m} \mu_i x_i$$

$$\text{s.t.} \sum_{i=1}^{m} \boldsymbol{c}_i x_i \leq \boldsymbol{B}$$

$$x_i \geq 0, i \in [m]$$

Denote its optimal solution as OPT

x_i represents the optimal fractional number of playing i-th arm if everything is deterministic and known

WLOG, we assume all the entries are in $[0, 1]$ and that $\boldsymbol{B} = (B, B, ..., B)$.

Bandits with Knapsacks (BwK)

Knowledge-wise:

- Known: T
- Unknown: mean reward $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)$ and mean cost/resource consumption $\boldsymbol{C} = (\boldsymbol{c}_1, ..., \boldsymbol{c}_m)$

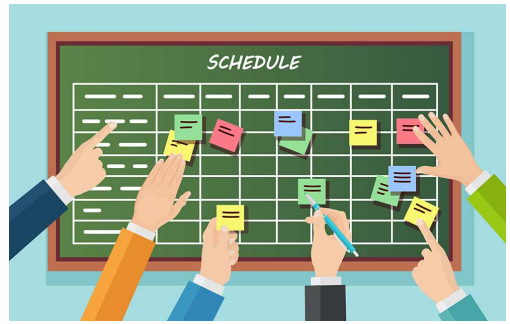
Notation: $i = 1, ..., m$ to index arm, $j = 1, ..., d$ to index constraint, $t = 1, ..., T$ to index time period

Applications

As a general framework, BwK covers a wide range of problems

- Dynamic pricing with limited supply
  - Network revenue management
  - Bundling and volume pricing

- Dynamic procurement and crowd-sourcing markets



- Ad allocation

- Repeated auction
- Network routing and scheduling

- ...

Performance Metric: Regret

τ : Stopping time that one of the resources is depleted

The problem-dependent regret

$$Regret(\mathcal{P}) = OPT - \mathbb{E}\left[\sum_{t=1}^{\tau} r_t\right],$$

where $\mathcal{P}$ encapsulates all the parameters related to the underlying distribution.

Questions:

(1) log T dependence achievable?

(2) What are the key quantities in characterizing the regret upper bound?

Contribution of Our Work

- Problem-independent bounds with $O(\sqrt{T})$-dependency
  Badanidiyuru et. al. (2013), Agrawal and Devanur (2014)
- Problem-dependent bounds with $O(\log T)$-dependency
  - Flajolet and Jaillet (2015): achieves $O(2^{m+d} \log T)$ regret bound but assumes prior knowledge on a number of LP's structural parameters
  - Sankararaman and Slivkins (2020): the setting with one-single constraint and deterministic cost (correctness of the proof remains to be seen)
- Our work:
  - General setting and no prior knowledge
  - $O\left(d^4 + m \log T\right)$ regret

Revisiting LP

The underlying LP

$$\max_{\boldsymbol{x}} \sum_{i=1}^{m} \mu_i x_i$$

$$\text{s.t.} \sum_{i=1}^{m} \boldsymbol{c}_i x_i \leq \boldsymbol{B}$$

$$x_i \geq 0, i \in [m]$$

Notation:

- $\mathcal{I}^* \subset [m]$: optimal basic variables of the LP or optimal arms
- $\mathcal{I}' \subset [m]$: optimal non-basic variables of the LP or sub-optimal arms
- $\mathcal{J}^* \subset [d]$: binding constraints of the LP
- $\mathcal{J}' \subset [d]$: non-binding constraints of the LP

By definition,

$$\mathcal{I}^* \cup \mathcal{I}' = [m], \mathcal{J}^* \cup \mathcal{J}' = [d]$$

Non-degeneracy Assumption

We assume the underlying LP is non-degenerate and it has a unique optimal solution

Denote $\boldsymbol{x}^* = (x_1^*, ..., x_m^*)$. Then, from LP's complementarity, it implies

$$i \in \mathcal{I}^* \Leftrightarrow x_i^* > 0$$

$$j \in \mathcal{J}^* \Leftrightarrow B - \sum_{i=1}^{m} c_{ij} x_i^* = 0$$

and $|\mathcal{I}^*| = |\mathcal{J}^*|$

Any LP satisfies above assumption under a small perturbation.

Dual LP and Reduced Cost

Primal:

$$\begin{aligned} \max \quad & \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{B} \\ & \boldsymbol{x} \geq 0 \end{aligned}$$

Dual:

$$\begin{aligned} \min \quad & \boldsymbol{B}^\top \boldsymbol{y} \\ \text{s.t.} \quad & \boldsymbol{C}^\top \boldsymbol{y} \geq \boldsymbol{\mu} \\ & \boldsymbol{y} \geq 0 \end{aligned}$$

Denote $\boldsymbol{x}^* \in \mathbb{R}^m$ and $\boldsymbol{y}^* \in \mathbb{R}^d$ as optimal solutions
Define reduced cost (profit) for i-th arm

$$\Delta_i := \boldsymbol{c}_i^\top \boldsymbol{y}^* - \mu_i$$

From the strict complementarity, we know

$$\Delta_i = 0 \Leftrightarrow i \in \mathcal{I}^*$$

$$\Delta_i > 0 \Leftrightarrow i \in \mathcal{I}'$$

A suboptimality measure for the arms!

Knapsack Process

Define $\boldsymbol{B}^{(0)} = \boldsymbol{B}$ and

$$\boldsymbol{B}^{(t+1)} = \boldsymbol{B}^{(t)} - \boldsymbol{C}_t$$

Recall that $\boldsymbol{C}_t$ is the resource consumption of the arm played at time period t

$\boldsymbol{B}^{(t)}$ denotes the remaining resource capacity at the end of time period t

Our First Regret Upper Bound

Proposition

The regret of a BwK algorithm has the following upper bound:

$$Regret(\mathcal{P}) \leq \mathbb{E}[\boldsymbol{B}^{(\tau)}]^{\top} \boldsymbol{y}^{*} + \sum_{i \in \mathcal{I}'} \Delta_i \mathbb{E}[n_i(\tau)]$$

- τ: Stopping time of the procedure
- $\boldsymbol{y}^{*}$: dual optimal solution
- $\boldsymbol{B}^{(\tau)}$: remaining resource when the process is terminated
- Δ_i: Reduced cost of the i-th arm
- $n_i(t)$: the number of times that i-th arm is played up to time t

The regret representation inspired by our previous works on online LP: Agrawal et. al. (2014), Li & Ye (2019)

Implications of the Regret Upper Bound

Two tasks to accomplish to minimize the regret:

Task I: Control the number of plays $n_i(\tau)$ for sub-optimal arms $i \in \mathcal{I}'$ which corresponds to the first component in the regret bound

$$\sum_{i \in \mathcal{I}'} \Delta_i \mathbb{E}[n_i(\tau)]$$

Playing each arm $i \in \mathcal{I}'$ will induce a cost of Δ_i

Task II: Control the remaining binding resources $\boldsymbol{B}_j^{(\tau)}$ for $j \in \mathcal{J}^*$ which corresponds to the second component in the regret bound

$$\mathbb{E}[\boldsymbol{B}^{(\tau)}]^\top \boldsymbol{y}^*$$

Recall τ is the time that one of the resource is exhausted

Task II is overlooked in the existing BwK literature.

Our Solution

A two-phase algorithm

- Phase I: Distinguish the basic variable set $\mathcal{I}^*$ from the non-basic variable set with as fewer number of plays as possible
- Phase II: Use the arms in $\mathcal{I}^*$ to exhaust the resource through an adaptive procedure

Phase I

Identify $\mathcal{I}^*$ from sampling (playing arms)

Objective: Identify $\mathcal{I}^*$ but avoid playing arms in $\mathcal{I}'$ too many times

From an LP's perspective: Identify the optimal basis with noisy observations of the LP's input

LP Stability under Perturbation (I)

Primal LP:

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{B} \\ & \boldsymbol{x} \geq 0, \end{aligned}$$

Perturbed LP:

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \hat{\boldsymbol{\mu}}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \hat{\boldsymbol{C}}\boldsymbol{x} \leq \boldsymbol{B} \\ & \boldsymbol{x} \geq 0, \end{aligned}$$

Denote

- Primal LP: optimal basis $\mathcal{I}^*$, binding constraints $\mathcal{J}^*$
- Perturbed LP: optimal basis $\hat{\mathcal{I}}^*$, binding constraints $\hat{\mathcal{J}}^*$

Question: What is the condition for $\mathcal{I}^* = \hat{\mathcal{I}}^*$ and $\mathcal{J}^* = \hat{\mathcal{J}}^*$?

Motivation for the question: The perturbed LP can be viewed as an estimation for the primal LP based on playing arms and making observations

LP Stability under Perturbation (II)

Lemma (LP Stability)

If the following conditions hold for all $i = 1, ..., m$:

$$\|\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i\|_\infty \leq \frac{\sigma\Delta}{\min\{md, d^2\}},$$

$$|\hat{\mu}_i - \mu_i| \leq \frac{\sigma\Delta}{\sqrt{\min\{md, d^2\}}},$$

then the two LPs share the same optimal basis, i.e., $\mathcal{I}^* = \hat{\mathcal{I}}^*$ and $\mathcal{J}^* = \hat{\mathcal{J}}^*$.

- σ is a constant related to the smallest singular value of matrix $\boldsymbol{C}$
- $\Delta = \min_{i \in \mathcal{I}'} \Delta_i$ smallest sub-optimality gap (smallest non-zero reduced cost)
- Perturbed coefficients $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{C}}$ can be interpreted as the estimators in bandits problem

Interpretations of the Parameters

σ : Indicate the linearity between LP's columns

$\Delta = \min_{i \in \mathcal{I}'} \Delta_i$: Represent the sub-optimality gap for the "best" non-basic variable

A smaller value for the above two quantities increases the difficulty of identifying $\mathcal{I}^*$

High Probability Event

Denote $\hat{\mu}_i(t)$ and $\hat{C}_{ji}(t)$ as the sample mean of history observations up to time t for μ_i and C_{ji}, respectively.

Lemma (Concentration)

With probability no less than $1 - \frac{4md}{T^2}$, *the following*

$$\mu_i \in \left(\hat{\mu}_i(t) - \sqrt{\frac{2 \log T}{n_i(t)}}, \hat{\mu}_i(t) + \sqrt{\frac{2 \log T}{n_i(t)}} \right),$$

$$C_{ji} \in \left(\hat{C}_{ji}(t) - \sqrt{\frac{2 \log T}{n_i(t)}}, \hat{C}_{ji}(t) + \sqrt{\frac{2 \log T}{n_i(t)}} \right),$$

hold for all $i \in [m], j \in [d], t \in T$.

Confidence intervals for μ_i and C_{ji}: A standard method in bandits literature

Phase I of Our Algorithm

Identify the optimal basis $\mathcal{I}^*$ with as fewer samples as possible

From the LP stability lemma, we consider two cases:

- Known parameters Δ, σ: Direct application of the lemma
- No prior knowledge: More challenging! A primal-dual perspective

Case I – Known Parameters Δ, σ

Algorithm:

1. Play each arm for $\frac{m^3 d^2 \log T}{\sigma^2 \Delta^2}$ times (omitting the constant here)
2. Compute $\hat{\boldsymbol{C}}(t)$ and $\hat{\boldsymbol{\mu}}(t)$ based on sample mean
3. Solve the following LP

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \hat{\boldsymbol{\mu}}(t)^\top \boldsymbol{x} \\ \text{s.t.} \quad & \hat{\boldsymbol{C}}(t)\boldsymbol{x} \leq \boldsymbol{B} \\ & \boldsymbol{x} \geq 0 \end{aligned}$$

Denote its optimal basis and binding constraints as $\hat{\mathcal{I}}^*$ and $\hat{\mathcal{J}}^*$

Analysis: From LP stability lemma, we know that $\hat{\mathcal{I}}^* = \mathcal{I}^*$ and $\hat{\mathcal{J}}^* = \mathcal{J}^*$ hold with probability $1 - \frac{4md}{T^2}$

Case II – No Prior Knowledge – Intuition

There are in total $O(2^{m+d})$ possible candidates for the optimal basis and binding constraints, $(\mathcal{I}^*, \mathcal{J}^*)$

Our approach: Through experimentation, eliminate all the "false" candidate. The remaining ones are the true candidates

The elimination procedure can be very efficient by taking advantage of the LP's structure.

Case II – No Prior Knowledge – Motivation

Lesson learned so far: the optimal arm set $\mathcal{I}^*$ and the binding constraint set $\mathcal{J}^*$ are equally important

$$\downarrow$$

A new notion of sub-optimality from a primal-dual perspective

$$\begin{aligned} OPT_i := \max \quad & \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{b} \\ & x_i = 0, \boldsymbol{x} \geq 0 \end{aligned} \qquad \begin{aligned} OPT_j := \min \quad & \boldsymbol{b}^\top \boldsymbol{y} \\ \text{s.t.} \quad & \boldsymbol{C}^\top \boldsymbol{y} \geq \boldsymbol{\mu} \\ & y_j \text{ free }, \boldsymbol{y}_{-j} \geq 0 \end{aligned}$$

OPT_i: importance of arm i (if i-th arm not played)
OPT_j: importance of constraint j (if j-th constraint wrongly exhausted)

Define (as the bottleneck of BwK problem)

$$\delta = OPT - \max \left\{ \max_{i \in \mathcal{I}'} \ OPT_i, \ \max_{j \in \mathcal{J}^*} \ OPT_j \right\},$$

Lower/Upper Confidence Bound for LP

At each time t,

$$\max_{\boldsymbol{x}} \sum_{i=1}^{m} \left(\hat{\mu}_i(t) - \sqrt{\frac{2\log T}{n_i(t)}} \right) x_i \qquad \max_{\boldsymbol{x}} \sum_{i=1}^{m} \left(\hat{\mu}_i(t) + \sqrt{\frac{2\log T}{n_i(t)}} \right) x_i$$

$$\text{s.t.} \sum_{i=1}^{m} \left(\hat{\boldsymbol{C}}_i(t) + \sqrt{\frac{2\log T}{n_i(t)}} \right) x_i \leq \boldsymbol{B} \qquad \text{s.t.} \sum_{i=1}^{m} \left(\hat{\boldsymbol{C}}_i(t) - \sqrt{\frac{2\log T}{n_i(t)}} \right) x_i \leq \boldsymbol{B}$$

$$\boldsymbol{x} \geq 0 \qquad \boldsymbol{x} \geq 0$$

Provide a lower bound and upper bound for LP

No Prior Knowledge – Algorithm

1. Initialize $\hat{\mathcal{I}}^* = \hat{\mathcal{J}}' = \emptyset$, n=0
2. While $|\hat{\mathcal{I}}^*| + |\hat{\mathcal{J}}'| < d$

   (1) Play each arm once and update estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{C}}$

   (2) Compute an UCB estimate of OPT_i for each arm i and OPT_j for each constraint j; Denote the estimate as OPT_i^U and OPT_j^U

   (3) Compute an LCB estimate for the primal LP; Denote it as OPT^L

   (4) For all $i \notin \hat{\mathcal{I}}^*$, if $OPT^L > OPT_i^U$, then

$$\hat{\mathcal{I}}^* = \hat{\mathcal{I}}^* \cup \{i\}$$

   For all $j \notin \hat{\mathcal{J}}'$, if $OPT^L > OPT_j^U$, then

$$\hat{\mathcal{J}}' = \hat{\mathcal{J}}' \cup \{j\}$$

Analysis

Intuition for the algorithm:

- $OPT^L > OPT_i^U$: the i-th arm is important
- $OPT^L > OPT_j^U$: the j-th constraint is non-binding

Proposition

With probability no less than $1 - \frac{4md}{T^2}$, *the algorithm terminates within* $O(\frac{m \log T}{\sigma^2 \delta^2})$ *rounds and it can identify the optimal basis of the underlying LP, i.e.,* $\hat{\mathcal{I}}^* = \mathcal{I}^*$ *and* $\hat{\mathcal{J}}' = \mathcal{J}'$.

Take-away: Without any prior knowledge, the algorithm identifies the structure of the underlying LP in $O(m \log T)$ rounds

Surprising: There are $O(2^{m+d})$ possible configurations of $(\mathcal{I}^*, \mathcal{J}')$, but identifiable with $O(m \log T)$ sample complexity

Phase II of Our Algorithm

After identifying $\mathcal{I}^*$, we will not play any arm in $\mathcal{I}'$ afterwards

The remaining task: To exhaust the resource using arms in $\mathcal{I}^*$

We propose an adaptive method. Without the method, the regret of an BwK algorithm is $\Omega(\sqrt{T})$

Filling the Knapsacks/Exhausting the Resources

For simplicity, suppose after the $t' = O(\frac{m \log T}{\sigma^2 \delta^2})$ rounds of exploration, $\mathcal{I}^*$ and $\mathcal{J}^*$ are identified.

Adaptive algorithm for filling the knapsacks:

For $t = t' + 1, ..., T$

1. Solve the UCB-LP and denote its optimal solution as $\tilde{\boldsymbol{x}}$

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \sum_{i=1}^{m} \left(\hat{\mu}_i(t) + \sqrt{\frac{2 \log T}{n_i(t)}} \right) x_i \\ \text{s.t.} \quad & \sum_{i=1}^{m} \left(\hat{\boldsymbol{C}}_i(t) - \sqrt{\frac{2 \log T}{n_i(t)}} \right) x_i \leq \boldsymbol{B}^{(t-1)} \\ & \boldsymbol{x} \geq 0, x_i = 0 \text{ for } i \in \mathcal{I}' \end{aligned}$$

2. Normalize $\tilde{\boldsymbol{x}}$ into a probability and play an arm accordingly
3. Update the knapsack process $\boldsymbol{B}^{(t)}$ (remaining resource)

Interpretation:

$\hat{\mu}_i(t)$ and $\hat{\boldsymbol{C}}_i(t)$ are sample mean estimate

$\sqrt{\frac{2 \log T}{n_i(t)}}$ term encourages exploration to better learn the parameters

Analysis of the Adaptive Algorithm

The algorithm adaptively changes the right-hand-side of the LP (aligned with our works on online LP, Li & Ye (2019), Chen et. al. (2021))

Proposition

Under the adaptive algorithm,

$$\mathbb{E}[B_j^{(\tau)}] = O\left(\frac{d^3}{\sigma^2}\right)$$

for $j \in \mathcal{J}^*$

Surprising: The remaining resource is not dependent on T. Without the adaptive mechanism, $\tau \approx T$ implies $\mathbb{E}[B_j^{(\tau)}] \approx O(\sqrt{T})$.

Implication: The regret's dependency on T only comes from the first task – identifying the LP's structure

Combining the Two Phases

Proposition

The regret of our two-phase algorithm

$$O\left(\frac{d^4}{\sigma^2} + \frac{m \log T}{\sigma^2 \delta^2}\right)$$

Summary:

- LP-based characterization of the BwK problem
- The symmetry between arms and knapsack
- The importance of (i) identifying optimal arms and (ii) stabilizing the knapsack process
- Adaptive algorithm

Open question: Can x_i be updated by a fast algorithm rather than solving an LP at each time t?

Thank you!