The Symmetry between Arms and Knapsacks: A Primal-Dual Approach for Bandits with Knapsacks

Xiaocheng Li[1], Chunlin Sun[2], Yinyu Ye[3]

[1]Imperial College Business School

[2]Institute for Computational and Mathematical Engineering, Stanford University

[3] Department of Management Science and Engineering, Stanford University

1. Multi-Armed Bandit with Knapsack

- **Bandit:** Consider the bandit machine has m arms with fixed but unknown mean reward μ_i and mean cost $\boldsymbol{c}_i \in \mathbf{R}_+^d$ for i in $1, \ldots, m$
- **Knapsack:** Known total budget $\boldsymbol{B} \in \mathbf{R}_+^d$ and the length of time horizon T
- **Rule:** The decision maker selects one arm to pull at each step and then observes the realized reward and cost of the chosen arm
- **Goal:** Maximize the total reward until any type of budgets is depleted or time up.

2. Underlying Linear Programming Problem

- The optimal objective value $\mathrm{OPT_{LP}}$ of the following LP is an upper bound of the expected optimal reward,

$$\begin{aligned} \mathrm{OPT_{LP}} := \max_{\boldsymbol{x}} \ & \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t.} \ & \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{B} \\ & \boldsymbol{x} \ge \boldsymbol{0} \end{aligned}$$

  where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^{\mathrm{T}}, \boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m)$, and $\boldsymbol{x} = (x_1, \ldots, x_m)^{\mathrm{T}} \in \mathbf{R}^m$. Denote $\boldsymbol{x}^*$ as its optimal solution and $\mathrm{OPT_{LP}}$ as its optimal objective value
- If the model is deterministic, the solution of the above LP gives the optimal policy. Here, x_i represents the number of drawing i-th arm.
- Denote $\boldsymbol{\mathcal{I}}^* = \{i \colon x_i^* > 0\}$ as the index set of optimal arms, $\boldsymbol{\mathcal{J}}^* = \{j \colon (\boldsymbol{B} > \boldsymbol{C}\boldsymbol{x}^*)_j\}$ as the index set of binding resources, and $\boldsymbol{\mathcal{I}}'$ and $\boldsymbol{\mathcal{J}}'$ as index sets of sub-optimal arms and non-binding resources, respectively.

3. Motivation- A Regret Upper Bound

- The regret can be bounded by $\sum_{i \in \mathcal{I}'} \Delta_i \mathbf{E}[n_i] + \mathbf{E}\left[\boldsymbol{B}^{(\boldsymbol{\tau})}\right]^{\mathrm{T}} \boldsymbol{y}^*$, where Δ_i is the reduced cost for the i-th arm, n_i is the number of times that i-th arm is pulled, $\boldsymbol{B}^{(\boldsymbol{\tau})}$ is the remaining resources at the termination time, and $\boldsymbol{y}^*$ is the optimal dual price.
- The first term is interpreted as the cost of playing sub-optimal arms; the second term is interpreted as the cost of wasted binding resources.

Thus, to maximize the reward, the decision maker should

- play less sub-optimal arms (optimal arms identification)
- fully consume binding resources (binding resources identification and adaptively procedure to play optimal arms)

4. Symmetry between Arms and Knapsack

- With mild conditions, $|\boldsymbol{\mathcal{I}}^*| = |\boldsymbol{\mathcal{J}}^*|$.
- Denote OPT_i as the optimal objective value of following LP,

$$\begin{aligned} \mathrm{OPT}_i := \max_{\boldsymbol{x}} \ & \boldsymbol{\mu}^\top \boldsymbol{x}, \\ \text{s.t.} \ & \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{B}, \\ & x_i = 0, \boldsymbol{x} \ge \boldsymbol{0}. \end{aligned}$$

  then, $i \in \boldsymbol{\mathcal{I}}^* \Leftrightarrow \mathrm{OPT}_i = \mathrm{OPT_{LP}}$ and $i \in \boldsymbol{\mathcal{I}}' \Leftrightarrow \mathrm{OPT}_i < \mathrm{OPT_{LP}}$.
- Denote OPT_j as the optimal objective value of following LP,

$$\begin{aligned} \mathrm{OPT}_j = \max_{\boldsymbol{x}} \ & \boldsymbol{\mu}^\top \boldsymbol{x} - (\boldsymbol{B} - \boldsymbol{C}\boldsymbol{x})_j, \\ \text{s.t.} \ & \sum_{i=1}^{m} \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{B}, \\ & \boldsymbol{x} \ge 0. \end{aligned}$$

  Then, $j \in \boldsymbol{\mathcal{J}}^* \Leftrightarrow \mathrm{OPT}_j = \mathrm{OPT_{LP}}$ and $j \in \boldsymbol{\mathcal{J}}' \Leftrightarrow \mathrm{OPT}_j < \mathrm{OPT_{LP}}$.
- Set $\delta = \frac{1}{T}(\mathrm{OPT_{LP}} - \max\{\max_{i \in \boldsymbol{\mathcal{I}}^*}\{\mathrm{OPT}_i\}, \max_{j \in \boldsymbol{\mathcal{J}}'}\{\mathrm{OPT}_j\}\})$.

  It characterizes the hardness of distinguishing optimal arms from non-optimal arms, and binding resources from non-binding resources.
- δ is also a generalization of the sub-optimality measure for multi-armed bandit problem.

5. UCB and LCB

- When the sample size is not large enough, the bias between real mean and sample mean might be large. This difference might mislead the algorithm to find true optimal arms and binding resources. To avoid this case, we apply the upper confidence bound and lower confidence bound techniques by considering both UCB and LCB of rewards and costs of each arm instead of sample mean as literatures.
- Denote $\boldsymbol{\mu}^U, \boldsymbol{\mu}^L$ as the UCB and LCB of $\boldsymbol{\mu}$, and denote $\boldsymbol{C}^U, \boldsymbol{C}^L$ as the UCB and LCB of $\boldsymbol{C}$. The optimal objective value of following two LPs are UCB and LCB of $\mathrm{OPT_{LP}}$.

$$\begin{aligned} \mathrm{OPT}_{\mathrm{LP}}^U := \max_{\boldsymbol{x}} \ & \left(\boldsymbol{\mu}^U\right)^\top \boldsymbol{x}, \\ \text{s.t.} \ & \boldsymbol{C}^L \boldsymbol{x} \le \boldsymbol{B}, \\ & \boldsymbol{x} \ge \boldsymbol{0}. \end{aligned} \qquad \begin{aligned} \mathrm{OPT}_{\mathrm{LP}}^L := \max_{\boldsymbol{x}} \ & \left(\boldsymbol{\mu}^L\right)^\top \boldsymbol{x}, \\ \text{s.t.} \ & \boldsymbol{C}^U \boldsymbol{x} \le \boldsymbol{B}, \\ & \boldsymbol{x} \ge \boldsymbol{0}. \end{aligned}$$

- $\mathrm{OPT}_{\mathrm{LP}}^U$ and $\mathrm{OPT}_{\mathrm{LP}}^L$ converge to $\mathrm{OPT_{LP}}$
- Similarly, define OPT_i^U, OPT_j^U, OPT_i^L and OPT_j^L for all arms and resources. They also satisfy the convergence property.
- As the size of samples are large enough, following inequalities hold: $\mathrm{OPT}_{\mathrm{LP}}^L > \mathrm{OPT}_i^U$ for all optimal-arms and $\mathrm{OPT}_{\mathrm{LP}}^L > \mathrm{OPT}_j^U$ for all non-binding resources. However, $\mathrm{OPT}_{\mathrm{LP}}^L \le \mathrm{OPT}_i^U$ and $\mathrm{OPT}_{\mathrm{LP}}^L \le \mathrm{OPT}_j^U$ always hold for all sub-optimal arms and non-binding resources

6. Primal-dual Adaptive Algorithm

Algorithm 1 Primal-dual Adaptive Algorithm for BwK

Input: Resource capacity $\boldsymbol{B}$, T

%% Phase I: Identification of $\mathcal{I}^*$ and $\mathcal{J}'$

Initialize $\hat{\mathcal{I}}^* = \hat{\mathcal{J}}' = \emptyset$, $t = 0$

Initialize the knapsack process $\boldsymbol{B}^{(0)} = \boldsymbol{B}$

while $|\hat{\mathcal{I}}^*| + |\hat{\mathcal{J}}'| < d$ **do**

  Play each arm $i \in [m]$ once

  Update $t = t + m$ and the knapsack process $\boldsymbol{B}^{(t)}$

  Update the estimates $\hat{\boldsymbol{\mu}}(t)$ and $\hat{\boldsymbol{C}}(t)$

  Solve the LCB problem and obtain $\mathrm{OPT}_{\mathrm{LP}}^L(t)$

  for $i \notin \hat{\mathcal{I}}^*$ **do**

    Solve the following UCB problem for OPT_i

$$\begin{aligned} \mathrm{OPT}_i^U(t) := \max_{\boldsymbol{x}} \ & \left(\boldsymbol{\mu}^U(t)\right)^\top \boldsymbol{x}, \\ \text{s.t.} \ & \boldsymbol{C}^L(t)\boldsymbol{x} \le \boldsymbol{B}, \\ & x_i = 0, \boldsymbol{x} \ge \boldsymbol{0}. \end{aligned}$$

    if $\mathrm{OPT}_{\mathrm{LP}}^L(t) > \mathrm{OPT}_i^{\mathrm{U}}(t)$ **then**

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

    end if

  end for

  for $j \notin \hat{\mathcal{J}}'$ **do**

    Solve the following UCB problem for OPT_j

$$\begin{aligned} \mathrm{OPT}_j^U(t) := \min_{\boldsymbol{y}} \ & \boldsymbol{B}^\top \boldsymbol{y} - B, \\ \text{s.t.} \ & (\boldsymbol{C}^L(t))^\top \boldsymbol{y} \ge \boldsymbol{\mu}^U(t) + \boldsymbol{C}_{j,\cdot}^U(t), \\ & \boldsymbol{y} \ge 0. \end{aligned}$$

    if $\mathrm{OPT}_{\mathrm{LP}}^L(t) > \mathrm{OPT}_j^{\mathrm{U}}(t)$ **then**

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

    end if

  end for

end while

Update $t = t + 1$

%% Phase II: Exhausting the binding resources

while $t \le \tau$ **do**

  Solve the following LP

$$\begin{aligned} \max_{\boldsymbol{x}} \ & \left(\boldsymbol{\mu}^U(t-1)\right)^\top \boldsymbol{x}, \\ \text{s.t.} \ & \boldsymbol{C}^L(t-1)\boldsymbol{x} \le \boldsymbol{B}^{(t-1)}, \\ & x_i = 0, \ i \notin \mathcal{I}^*, \\ & \boldsymbol{x} \ge \boldsymbol{0}. \end{aligned}$$

  Denote its optimal solution as $\tilde{\boldsymbol{x}}$

  Normalize $\tilde{\boldsymbol{x}}$ into a probability and randomly play an arm according to the probability

  Update estimates $\hat{\boldsymbol{\mu}}(t)$, $\hat{\boldsymbol{C}}(t)$, and $\boldsymbol{B}^{(t)}$

  Update $t = t + 1$

end while

7. Theoretical Analysis

- During the Phase one, each arm will be played for no more than $O(\frac{\log T}{\delta^2})$ times. Moreover, with probability no less than $1 - O\left(\frac{1}{T^2}\right)$, the algorithm can identify true optimal arms and binding resources.
- If the budget is large enough, $\mathbf{E}\left[\boldsymbol{B}^{(\boldsymbol{\tau})}\right] = \mathrm{O}(\frac{1}{\delta^2})$
- The regret of this algorithm can be bounded by $O(\frac{\log T}{\delta^2})$