Bandits with Knapsacks

Ashwinkumar Badanidiyuru, Robert Kleinberg, Aleksandrs Slivkins

FOCS'13

Problem description: Bandits with knapsacks (BwK):

- A learner has a fixed set of potential actions, a.k.a. arms, in T time rounds.
- In each time, the learner chooses an arm and observe: a reward, a resource consumption vector.
- A pre-specified budget vector
- The process stops when any type of resource consumption exceeds its budget or time is run up

Applications

- Dynamic pricing with limited supply

  The algorithm is a seller which has a number of identical items for sale and the agents arrive sequentially. Each agent has a private value v_t for an item and buys an item if v_t exceeds p_t, the price offered by the seller. (Arms: prices)

Applications

- Network routing.

  - Connection requests arrive one by one, each of which consists of a pair of terminals.

  - The system chooses a routing protocol (arm) for each connection (a mapping from terminal pairs to a path).

  - Toal bandwidth consumption on each edge/node <= capacity.

  - Goal: maximizing the # of successful connections.

Related work

- MAB problem: single/deterministic resource consumption

  - Lai and Robbins [1985], Gyorgy et al. [2007]

- Stochastic packing problem: full information of past and present

  - Devanur et al. [2001]

Formal problem model

- A fixed and finite set of **m** arms (action set **X**)

- In each t, picks arm $x_t \in X$, receives reward $r_t \in [0, 1]$ and consumes $c_{t,i} \in [0, 1]$ amount of resource i.
  A fixed constraint $B_i \in \mathbb{R}_+$ is on the consumption of resource $i \in \{1, \cdots, d\}$.

- Algorithm stops at the earliest time τ (<= **T**) when any constraint is violated.

- Goal: maximizing total reward.

Benchmark and regret

- Benchmark: the optimal policy with a time-invariant mixture of arms

- Regret: OPT - REW (total reward of algorithm)

Preliminaries

- I.i.d. assumption: reward and resource consumption of each arm

- Reduction to uniform budgets: multiplying B_{min}/B_i on both sides of constraint

- Confidence radius: $rad(\hat{v}, N) = \sqrt{\frac{C_{rad}\hat{v}}{N} + \frac{C_{rad}}{N}}$

$$C_{rad} = O(\log(dT|X|))$$

$$E[v] = [\hat{v} - rad(\hat{v}, N), \hat{v} + rad(\hat{v}, N)$$

LP-relaxation

E[# of times arm x is chosen]

E[reward of arm x]

$$
\begin{array}{llcll}
\max & \sum_{x \in X} \xi_x \, r(x,\mu) & & & \text{in } \xi_x \in \mathbb{R}, \text{for each } x \in X \\
\text{s.t.} & \sum_{x \in X} \xi_x \, c_i(x,\mu) & \leq & B & \text{for each resource } i \\
 & \xi_x & \geq & 0 & \text{for each arm } x.
\end{array}
$$

E[resource consumption of type i, arm x]

$$
\begin{array}{llcll}
\min & B \sum_i \eta_i & & & \text{in } \eta_i \in \mathbb{R}, \text{for each resource } i \\
\text{s.t.} & \sum_i \eta_i \, c_i(x,\mu) & \geq & r(x,\mu) & \text{for each arm } x \in X \\
 & \eta_i & \geq & 0 & \text{for each resource } i.
\end{array}
$$

The primal-dual algorithm

Algorithm `PrimalDualBwK`

Initialization
In the first m rounds, pull each arm once.
$v_1 = \mathbf{1} \in [0,1]^d$.
$\{v_t \in [0,1]^d$ is the round-t estimate of the optimal solution η^* to (`LP-dual`) in Section 3.$\}$
$\{$We interpret $v_t(i)$ as an estimate of the (fictional) unit cost of resource i, for each i.$\}$
Set $\epsilon = \sqrt{\ln(d)/B}$.
for rounds $t = m+1, \ldots, \tau$ *(i.e., until resource budget exhausted)* **do**
For each arm $x \in X$,
Compute UCB estimate for the expected reward, $u_{t,x} \in [0,1]$.
Compute LCB estimate for the resource consumption vector, $L_{t,x} \in [0,1]^d$.
Expected cost for one pull of arm x is estimated by $\texttt{EstCost}_x = L_{t,x} \cdot v_t$.
Pull arm $x = x_t \in X$ that maximizes $u_{t,x}/\texttt{EstCost}_x$, the optimistic *bang-per-buck* ratio.
Update estimated unit cost for each resource i:

$$v_{t+1}(i) = v_t(i)\,(1+\epsilon)^{\ell},\ \ell = L_{t,x}(i).$$

Main result

- Theorem 4.2. Consider an instance of BwK with d resources, m = |X|, and B = min_i B_i. The regret of algorithm PrimalDualBwK satisfies:

$$\mathtt{OPT}_{\mathtt{LP}} - \mathtt{REW} \leq O\left(\sqrt{\log(dT)}\right)\left(\sqrt{m\,\mathtt{OPT}_{\mathtt{LP}}} + \mathtt{OPT}_{\mathtt{LP}}\sqrt{\frac{m}{B}}\right) + O(m)\,\log(dT)\log(T).$$

- Lower-bound: $\Omega\left(\min\left(\mathtt{OPT},\ \mathtt{OPT}\sqrt{\frac{m}{B}} + \sqrt{m\,\mathtt{OPT}}\right)\right)$

Proof sketch

- Bound the ratio in the deterministic case: reward and resource consumptions always equal to the corresponding expectation.
- Bound the regret by the analysis of estimation error or reward and resource consumptions.

Adapted algorithm for deterministic case

Algorithm 1 Algorithm `PrimalDualBwK`, adapted for deterministic outcomes

1: **Initialization**
2: In the first m rounds, pull each arm once.
3: For each arm $x \in X$, let $r_x \in [0,1]$ and $C_x \in [0,1]^d$
4: denote the reward and the resource consumption vector revealed in Step 2.
5: $v_1 = \mathbf{1} \in [0,1]^d$.
6: $\{v_t \in [0,1]^d$ is the round-t estimate of the optimal solution η^* to (`LP-dual`) in Section 3.$\}$
7: {We interpret $v_t(i)$ as an estimate of the (fictional) unit cost of resource i, for each i.}
8: Set $\epsilon = \sqrt{\ln(d)/B}$.
9: **for** rounds $t = m+1, \ldots, \tau$ *(i.e., until resource budget exhausted)* **do**
10: For each arm $x \in X$,
11: *Expected cost* for one pull of arm x is estimated by $\texttt{EstCost}_x = C_x \cdot v_t$.
12: Pull arm $x = x_t \in X$ that maximizes $r_x / \texttt{EstCost}_x$, the *bang-per-buck* ratio.
13: Update estimated unit cost for each resource i:

$$v_{t+1}(i) = v_t(i)\,(1+\epsilon)^{\ell},\ \ \ell = C_x(i).$$

Ratio in the deterministic case

$$
\begin{aligned}
B \geq \bar{y}^\intercal C \xi^* &= \frac{1}{\mathtt{REW}} \sum_{m<t<\tau} (r^\intercal z_t)(y_t^\intercal C \xi^*) \\
&\geq \frac{1}{\mathtt{REW}} \sum_{m<t<\tau} (r^\intercal \xi^*)(y_t^\intercal C z_t) \\
&\geq \frac{\mathtt{OPT_{LP}}}{\mathtt{REW}} \left[(1-\epsilon) \sum_{m<t<\tau} y^\intercal C z_t - \frac{\ln d}{\epsilon} \right] \\
&\geq \frac{\mathtt{OPT_{LP}}}{\mathtt{REW}} \left[B - \epsilon B - m - 1 - \frac{\ln d}{\epsilon} \right].
\end{aligned}
$$

- Choose $\epsilon = \sqrt{\frac{lnd}{B}}$
- Regret is bounded by $OPT_{LP} \cdot O(\sqrt{\frac{lnd}{B}} + \frac{m}{B})$

Fit the estimation error into the final ratio

$$
\begin{aligned}
B &\geq \bar{y}^\intercal C\xi^* && (\xi^* \textit{ is primal feasible}) \\
&= \frac{1}{\mathtt{REW_{UCB}}} \sum_{m<t<\tau} (u_t^\intercal z_t)(y_t^\intercal C\xi^*) \\
&\geq \frac{1}{\mathtt{REW_{UCB}}} \sum_{m<t<\tau} (u_t^\intercal z_t)(y_t^\intercal L_t\xi^*) && (\textit{clean execution}) \\
&\geq \frac{1}{\mathtt{REW_{UCB}}} \sum_{m<t<\tau} (u_t^\intercal \xi^*)(y_t^\intercal L_t z_t) \\
&\geq \frac{1}{\mathtt{REW_{UCB}}} \sum_{m<t<\tau} (r^\intercal \xi^*)(y_t^\intercal L_t z_t) && (\textit{clean execution}) \\
&\geq \frac{\mathtt{OPT_{LP}}}{\mathtt{REW_{UCB}}} \left[(1-\epsilon) y^\intercal \left(\sum_{m<t<\tau} L_t z_t \right) - \frac{\ln d}{\epsilon} \right] \\
&= \frac{\mathtt{OPT_{LP}}}{\mathtt{REW_{UCB}}} \left[(1-\epsilon) y^\intercal \left(\sum_{m<t<\tau} C_t z_t \right) - (1-\epsilon) y^\intercal \left(\sum_{m<t<\tau} E_t z_t \right) - \frac{\ln d}{\epsilon} \right]
\end{aligned}
$$

Fit the estimation error into the final ratio

- LCB of resource consumption is close to actual resource consumption
- UCB of reward is close to actual resource consumption

$$\mathtt{REW}_{\mathtt{UCB}} \geq \mathtt{OPT}_{\mathtt{LP}} \left[1 - \epsilon - \frac{m+1}{B} - \frac{1}{B} \left\| \sum_{m<t<\tau} E_t z_t \right\|_\infty - \frac{\ln d}{\epsilon B} \right].$$

$$\mathtt{REW} \geq \mathtt{REW}_{\mathtt{UCB}} - \sum_{m<t<\tau} (u_t - r_t)^\intercal z_t = \mathtt{REW}_{\mathtt{UCB}} - \sum_{m<t<\tau} \delta_\tau^\intercal z_t.$$

$$\mathtt{REW} \geq \mathtt{OPT}_{\mathtt{LP}} - \left[2\mathtt{OPT}_{\mathtt{LP}} \left(\sqrt{\frac{\ln d}{B}} + \frac{m+1}{B} \right) + m + 1 \right] - \frac{\mathtt{OPT}_{\mathtt{LP}}}{B} \left\| \sum_{m<t<\tau} E_t z_t \right\|_\infty - \left| \sum_{m<t<\tau} \delta_t^\intercal z_t \right|$$

- With high probability, we have the following:

$$\left| \sum_{m<t<\tau} \delta_t z_t \right| \leq O\left(\sqrt{C_{\mathtt{rad}}\, m \mathtt{REW}} + C_{\mathtt{rad}}\, m \, \log T\right)$$

$$\left\| \sum_{m<t<\tau} E_t z_t \right\|_\infty \leq O\left(\sqrt{C_{\mathtt{rad}}\, m B} + C_{\mathtt{rad}}\, m \, \log T\right).$$

- Plug these two terms into the regret

$$\mathtt{REW} \geq \mathtt{OPT}_{\mathtt{LP}} - \left[2\mathtt{OPT}_{\mathtt{LP}}\left(\sqrt{\frac{\ln d}{B}} + \frac{m+1}{B}\right) + m + 1\right] - \frac{\mathtt{OPT}_{\mathtt{LP}}}{B}\left\|\sum_{m<t<\tau} E_t z_t\right\|_\infty - \left|\sum_{m<t<\tau} \delta_t^\intercal z_t\right|$$

$$\mathtt{OPT}_{\mathtt{LP}} - \mathtt{REW} \leq O\left(\sqrt{\log(dT)}\right)\left(\sqrt{m\,\mathtt{OPT}_{\mathtt{LP}}} + \mathtt{OPT}_{\mathtt{LP}}\sqrt{\frac{m}{B}}\right) + O(m)\,\log(dT)\log(T).$$

Lessons learned

- Techniques of learning problems may be embedded in the online algorithm of classical problems
- Necessary assumptions may be hidden in the problem setting (e.g. small size resource consumption)

Q & A

Thank you