Bandits with Knapsacks: Beyond the Worst-Case Analysis

Karthik Abinav Sankararaman (Facebook) Aleksandrs Slivkins (Microsoft Research NYC)

Bandits with Knapsacks (BwK)

K arms, T rounds, d resources.
Resource budgets $B_1, \ldots, B_d$.

In each round $t \in [T]$:
- Choose arm $a_t \in [K]$
- Observe *outcome vector* $\mathbf{o}_t(a_t) \in [0,1]^{d+1}$: reward r_t, consumption $c_{j,t}$ $\forall$ resource $j \in [d]$
- Stop, if some resource runs out of budget

Goal: Maximize the total reward.

- Outcomes $(\mathbf{o}_t(a):\ \text{arms } a \in [K])$ chosen IID
- Benchmark: best fixed distribution over arms
- w.l.o.g. rescale consumption so that $B_j = B$.

Motivating Examples

Dynamic Pricing & Auctions:
d products, limited supply of each.
Seller adjusts prices (resp., auction parameters) over time to maximize total revenue

Crowdsourcing markets:
Many similar tasks, limited budget.
Contractor dynamically adjusts wages to maximize #completed tasks
(extension: d types of tasks, budget for each)

Many more examples in prior work.

Worst-Case Regret : Well-Understood

Optimal $\sqrt{T}$-like regret (upper & lower bounds) (Badanidiyuru, Kleinberg, Slivkins '13).

Achieved by *four* different algorithms.
Our focus: UcbBwK (Agrawal, Devanur '14), based on "optimism under uncertainty".

Optimal(-ish) worst-case regret bounds known for many extensions of BwK
(Agrawal, Devanur '14 '16; Badanidiyuru et al.'14; Agrawal et al., '16; Sankararaman, Slivkins '18).

3 results "beyond the worst case"

- Instance-dependent logarithmic regret: Full characterization: upper & lower bounds. Main open question for stochastic BwK.
- Small per-round regret in all but few rounds
- Large-but-structured action sets: reduction from BwK to bandits

Results: Logarithmic Regret

Without resources: optimal regret $O(\frac{K \log T}{\text{gap}})$,
Reward gap: between the best and 2nd-best arm.
How to generalize it to BwK / resources?

- Lagrange gap: version of "gap" for BwK use Lagrangian functions $\mathcal{L}$ of LP-relaxation.

$$G_{\texttt{lag}}(a) := \mathcal{L}(a^*, \lambda^*) - \mathcal{L}(a, \lambda^*)$$
$$G_{\texttt{lag}} := \min_{\text{arms } a \neq a^*} G_{\texttt{lag}}(a)$$
a^* = best arm, λ^* = optimal dual solution.

- **Theorem:** $O(K G_{\texttt{lag}}^{-1} \log T)$ regret

  Only for **best-arm-optimal instances:** when best fixed distribution over arms is supported on $\{a^*, \texttt{skip}\}$ and is unique.

  Only for $d = 2$ resources: paradigmatic case for most examples of BwK.
- **Theorem:** Both conditions are necessary: essentially, $\Omega(\sqrt{T})$ regret otherwise, for any algorithm & wide family of instances.
- Algorithm: UcbBwK (with a new analysis) Worst-case optimal even if the conditions fail.

Results: Per-round Regret

At each round t, $\texttt{reg}_t := \texttt{opt}/T - \texttt{rew}_t$

Theorem: For all $\varepsilon > 0$, UcbBwK achieves $\texttt{reg}_t < \varepsilon$ in all but $\leq \tilde{O}(K\varepsilon^{-2})$ rounds t.

Assumes $B > \Omega(T)$, paradigmatic case in BwK.

Fairness motivation: each round = single user, reward = user's utility, $\texttt{opt}/T$ = fair share.
Thus, $\texttt{reg}_t$ = deviation from fair share.

In bandits, such result implies $O(\log T)$ regret, but in BwK it does not.

Result: Reduction to Bandits

UCB analysis for X bandits =>
UcbBwK algorithm works for X BwK

Applications: X = {contextual, semi-, MNL}
- Contextual bandits: at each time t, observe context x_t before choosing an action
- Semi-bandits: at each time t, choose $\leq m$ arms, observe the outcome for each of them
- MNL bandits: at each time t, choose $\leq m$ arms, then one "final" arm is chosen via multinomial logistic distribution (MNL).

For each application X, three results:
- worst-case regret: simple corollary. In prior work, each X is a separate paper!
- logarithmic regret (new)
- per-round regret (new)

Caveat: our reduction does not come with a computationally efficient implementation.

Some philosophy: BwK is one of several "problem dimensions" in bandits. Reductions along one "dimension", such as ours, is a good way to handle a "multi-dimensional" problem space

Key Technical Ingredients

Logarithmic Regret Upper Bound
- **LP sensitivity** for each non-optimal arm a, increase expected reward and decrease expected consumption by $\leq \delta(a)$.
  Let X^* be the new optimal LP-solution.
  If $a \in \texttt{support}(X^*)$, then $\delta(a) > G_{\texttt{lag}}$.
- Applied to UcbBwK: each non-optimal arm chosen in $\leq O(K G_{\texttt{lag}}^{-2} \log T)$ rounds
- Careful accounting of reward/consumption $\Rightarrow$ regret $O(K G_{\texttt{lag}}^{-1} \log T)$

Confidence Sum: $\sum_{t \in S \subseteq [T]} \texttt{ConfTerm}(a_t)$ for a given subset S of rounds
- abstracts a key object in a typical analysis of an "optimism under uncertainty" algorithm.
- the main step in such analyses provides a uniform upper-bound on the confidence sum which holds *for any algorithm*
- our reduction inputs such result as a lemma.
- we also use confidence sums to analyze per-round regret of UcbBwK

Gap: two different notions for BwK, both generalize "reward gap" for bandits
- "Lagrange gap" (as defined above)
- "LP gap" for distribution X over arms: optimal LP-value minus LP-value of X. Used to analyze per-round regret.