Adversarial Bandits with Knapsacks
Nicole Immorlica\textsuperscript{1}, Karthik Abinav Sankararaman\textsuperscript{2}, Robert Schapire\textsuperscript{1}, Aleksandrs Slivkins\textsuperscript{1}
\textsuperscript{1}Microsoft Research, New York City \textsuperscript{2}The University of Maryland, College Park

**Overview**

- **BwK:** general model for multi-armed bandits with resource consumption
- **First algorithm for Adversarial BwK,** matching lower bound.
- **Subroutine:** new algorithm for Stochastic BwK, with much simpler analysis.
- **Modular algorithm ⇒ several extensions.**

**Motivating Examples**

- **Dynamic Pricing/Auctions:**
  - \(d\) products, limited supply of each.
  - Seller adjusts prices (resp., auction params) 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.**

**Prior Work on Stochastic BwK**

- **Special cases:** Badanidiyuru\textsuperscript{+} ’12; Babaioff\textsuperscript{+} ’12; Tran-Thanh\textsuperscript{+} ’12; Krause & Singla ’13; Ding\textsuperscript{+} ’13; ...
- **BwK:** model & optimal algorithm: Badanidiyuru, Kleinberg, Slivkins ’13
- **Extensions:** Agrawal & Devanur ‘14; Badanidiyuru, Langford, Slivkins ’14; Agrawal, Devanur, Li ’16; Sankararaman & Slivkins ’18

Simultaneous work on Adversarial BwK: special cases with \(\text{ratio} = 1\) (ask us)

**BwK: General Framework**

\(K\) arms, \(d\) resources, budgets \(B_1, \ldots, B_d\)

\(K\) in each round \(t \in [T]\):
- Choose arm \(a_t \in [K]\)
- Observe outcome vector \(o_t(a_t) \in [0, 1]^{d+1}\):
  - 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.

Outcome matrix: \(M_t = (o_t(a) : \ arms\ \ a \in [K])\).

- **Stochastic BwK:** \(M_t\) chosen IID.
- **Adversarial BwK:** \(M_t\) chosen adversarially.

**WLOG** rescale s.t. all budgets are \(B = \min_j B_j\).

**Benchmark**

\[ \text{OPT} = \text{best fixed distribution over arms} \]

\[ \text{can be } \text{d times better than best fixed arm} \]

\[ \mathbb{E}[\text{REW}] \geq \frac{\text{OPT}}{\text{ratio}} - \text{regret.} \]

REW = algorithm’s total reward

**Lower bound for Adversarial BwK**

Simple construction for \(\text{ratio} \geq \frac{1}{2}\):
- 2 arms, 1 resource, \(B = T/2\)
- Arm 1: consumption 1 in each round.
- Arm 2: 0 reward, 0 consumption.

Rew. for Arm 1

\[ \text{Instance 1} \quad t \in [1, T/2) \quad \text{Low} \]
\[ \text{Instance 2} \quad t \in [T/2, T) \quad \text{Medium} \]

More nuanced construction ⇒ \(\text{ratio} \geq \Omega(\log T)\).

**Main Algorithm (MAIN)**

Two adversarial online learning algorithms:
1. \(\text{ALG}_1\) for bandit feedback (e.g., EXP3.P)
2. \(\text{ALG}_2\) for full-feedback (e.g., Hedge).

**Adversarial BwK**

Use MAIN with \(T_0\) = random guess for OPT
\[ \Rightarrow \text{ratio} = O(d^2 \log T) \text{ vs. oblivious adversary} \]

**Challenge:** F1, F3 don’t hold, F2 doesn’t help.

**Proof Sketch** completely new analysis
\begin{enumerate}
  \item \(\text{LP relaxation: pick best stopping time } \tau\),
  \[ \mathbb{E}[o_t] \rightarrow \sum_t r_t o_t. \]
  \item \(\forall T_0\) \(\text{REW} \geq \min(T_0, \text{OPT} - dT_0) - \text{regret.} \)
  \[ T_0 = O(\text{OPT}) \Rightarrow \text{REW} \geq \text{OPT} / (d + 1)^2. \]
  \item \(T_0 = O(\text{OPT})\) with prob. \(1/\log_{d+1} T\).
\end{enumerate}

**High-prob guarantee vs. adaptive adversary**

Algorithm: each phase runs MAIN with fixed \(T_0\)
- Start with small guess \(T_0\), increase it adaptively.
- Observed data → IPS estimates → approx. LP; Increase \(T_0\) based on the approx. LP value.

Analysis: much more complicated, applies (a,b) to the last complete phase.

**Extensions**

\(\text{ALG}_1\) for \(X\) bandits ⇒ MAIN for \(X\) BwK, where \(X = \{\text{contextual, semi-, convex}\}\)
- No new research needed.
- **Stochastic BwK:** each extension was a paper (with slightly stronger regret bounds)
- **Adversarial BwK:** all results new.

**Caveat:** need \(\text{ALG}_1\) to have high-probability regret bound vs. adaptive adversary.

\[ \text{Regret} \ O \left( \sqrt{\frac{d}{T}} \sqrt{TK} \right) \text{ (optimal for } B = \Omega(T)) \]