Adversarial Bandits with Knapsacks

Nicole Immorlica[1], Karthik Abinav Sankararaman[2], Robert Schapire[1], Aleksandrs Slivkins[1]

[1]Microsoft Research, New York City [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 $\Rightarrow$ 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+ '12; Babaioff+ '12; Tran-Thanh+ '12; Krause & Singla '13; Ding+ '13; ...
- **BwK: model & optimal algorithm:** Badanidiyuru, Kleinberg, Slivkins '13
- **Extensions:** Agrawal & Devanur '14 '16; Badanidiyuru, Langford, Slivkins '14; Agrawal, Devanur, Li '16; Sankararaman & Slivkins '18

Simultaneous work on Adversarial BwK: special cases with **ratio** $= 1$ (ask us)

BwK: General Framework

K arms, d resources, 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.

Outcome matrix: $\mathbf{M}_t = (\mathbf{o}_t(a):\ \text{arms } a \in [K])$.
- **Stochastic BwK:** $\mathbf{M}_t$ chosen IID.
- **Adversarial BwK:** $\mathbf{M}_t$ chosen adversarially.

WLOG rescale s.t. all budgets are $B = \min_j B_j$.

Benchmark

OPT = best fixed distribution over arms (can be d times better than best fixed arm)

$$\mathbb{E}[\mathrm{REW}] \geq \frac{\mathrm{OPT}}{\textbf{ratio}} - \textbf{regret}.$$

REW = algorithm's total reward

Lower bound for Adversdarial BwK

Simple construction for **ratio** $\geq \frac{5}{4}$.
- 2 arms, 1 resource, $B = T/2$
- Arm 1: consumption 1 in each round.
- Arm 2: 0 reward, 0 consumption.

Rew. for Arm 1	$t \in [1, T/2]$	$t \in (T/2, T]$
Instance 1	Medium	Low
Instance 2	Medium	High

More nuanced construction $\Rightarrow$ **ratio** $\geq \Omega(\log T)$.

Main Algorithm (MAIN)

Two adversarial online learning algorithms:

(1) $\texttt{ALG}_1$ for bandit feedback (*e.g.*, EXP3.P)
(2) $\texttt{ALG}_2$ for full-feedback (*e.g.*, Hedge).

playing a repeated zero-sum game.

In each round $t \in [T]$:
- Simultaneously: $\texttt{ALG}_1$ picks arm $a_t \in [K]$, $\texttt{ALG}_2$ picks resource $j_t \in [d]$.
- Outcome vector $\mathbf{o}_t(a_t)$ is observed.
- Reward for $\texttt{ALG}_1$, cost for $\texttt{ALG}_2$:
$$\mathcal{L}_t(a_t, j_t) := r_t + 1 - \tfrac{T_0}{B}\, c_{t,j_t}$$
- ... revealed to $\texttt{ALG}_2$ for each resource j.

$T_0 = T$ for Stochastic BwK, parameter othw.

Stochastic BwK

$\mathbb{E}[\mathcal{L}_t]$ is the Lagrangian for linear relaxation

maximize	$T \cdot \sum_{a\in[K]} X(a)\ \mathbb{E}[r_t(a)]$
s.t.	$\sum_{a\in[K]} X(a) = 1$
$\forall j \in [d]$	$\sum_{a\in[K]} X(a)\ \mathbb{E}[c_{j,t}(a)] \leq B/T$
$\forall a \in [K]$	$0 \leq X(a) \leq 1.$

Proof Sketch Use facts from prior work:
F1: $\mathrm{OPT} \leq$ LP-value.
F2: Minimax Lagrangian $\Rightarrow$ Nash equilibrium.
F3: Average play $\rightarrow$ Nash equilibrium.

BwK $\xrightarrow{\text{F1}}$ LP $\xrightarrow{\text{F2}}$ 1-round game $\xrightarrow{\text{F3}}$ T-round game $\xrightarrow{\text{F4}}$ BwK

F4 (new): Large reward at stopping time.

Regret $\tilde{\mathcal{O}}\left(\frac{T}{B}\sqrt{TK}\right)$ (optimal for $B = \Omega(T)$)

Adversarial BwK

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

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

Proof Sketch completely new analysis

(a) LP relaxation: pick best stopping time τ,
$\mathbb{E}[\mathbf{o}_t] \ \rightarrow \frac{1}{\tau}\sum_{t\in[\tau]} \mathbf{o}_t$.
(b) $\forall T_0$ $\mathrm{REW} \geq \min(T_0, \mathrm{OPT} - dT_0) -$ regret.
$T_0 = \mathcal{O}(\mathrm{OPT}) \implies \mathrm{REW} \gtrsim \mathrm{OPT}/(d+1)^2$.
(c) $T_0 = \mathcal{O}(\mathrm{OPT})$ with prob. $1/\log_{d+1} T$.

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 $\rightarrow$ IPS estimates $\rightarrow$ approx. LP; Increase T_0 based on the approx. LP value.

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

Extensions

$\texttt{ALG}_1$ for X bandits $\rightarrow$ MAIN for X BwK, where X = {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 $\texttt{ALG}_1$ to have high-probability regret bound vs. adaptive adversary.