Smoothed Adversarial Linear Contextual Bandits with Knapsacks

Vidyashankar Sivakumar (Amazon)
Shiliang Zuo (UIUC)
Arindam Banerjee (UIUC)

International Conference on Machine Learning (ICML)
July 2022

Contextual Bandits with Knapsacks

		Round 1	Round 2	…….
		Reward	Reward	
		0.2	0.01	
K arms		0.5	0.3	
		.	.	
		.	.	
		0.05	0.7	

Goal: Sequentially choose arms over T rounds to maximize rewards.

Contextual Bandits with Knapsacks

		Round 1		Round 2	…….
		Reward	Consumptions	Reward	Consumptions
			R_1,……..,R_d		R_1,……..,R_d
		0.2	0.1,……..,0.35	0.01	0.25,……..,0.7
K arms		0.5	0.6,……..,0.4	0.3	0.6,……..,0.1
		.	.	.	.
		.	.	.	.
		0.05	0.01,……..,0.3	0.7	0.34,……..,0.49

Goal: Sequentially choose arms over T rounds to maximize rewards subject to resource consumptions less than budget B

Applications: Advertising, clinical trials, general resource allocation problems

1. Badanidiyuru, A. et. al. Bandits with Knapsacks. In FOCS, 2013.
2. Devanur, N.R. and Hayes, T.P. The Adwords Problem: Online Keyword Matching with Budgeted Bidders Under Random Permutations. In ACM-EC, 2009.

Smoothed Linear Contextual Bandits with Knapsacks (LinCBwK)

- Smoothed context vector corresponding to each of K arms[1]

$$x_t(a) = \nu_t(a) + g_t(a) \qquad \nu_t(a) \in \mathbb{B}_2^m, \quad g_t(a) \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{m \times m})$$

- Linear rewards and linear consumptions for all d resources

$$\mathbb{E}[r_t(a) | x_t(a), H_{t-1}] = \mu_*^\intercal x_t(a), \quad \mu_* \in \mathbb{S}^{m-1}$$

$$\mathbb{E}[\boldsymbol{v}_t(a) | x_t(a), H_{t-1}] = W_*^\intercal x_t(a), \quad W_* \in \mathbb{R}^{m \times d}$$

- Learner can choose the no-op arm with zero rewards and zero consumptions
- Sequentially choose arms to maximize rewards under resource constraints

$$\max \sum_{t=1}^{T} r_t(a_t) \quad \text{s.t.} \quad \sum_{t=1}^{T} \boldsymbol{v}_t(a_t) \leq \mathbb{I}B$$

1. Kannan et. al. A Smoothed Analysis of Greedy Algorithm for the linear contextual bandit problem. arXiv:1801.04323, 2018.

Benchmark policy

- Probability distribution over arms performs better than fixed arm

$$\mu_* = [1,1], \quad W_* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad x_t(1) = [1.0],\ x_t(2) = [0,1], \quad B = \frac{T}{2}$$

Choosing context 1 and 2 each with probability 0.5 gives optimal rewards

Without constraints picking fixed arm 1 or 2 gives optimal rewards

- We will benchmark algorithm performance against an optimal adaptive policy with knowledge of true parameters and adversarially chosen contexts over all T rounds

A Primal-Dual Approach

Algorithm

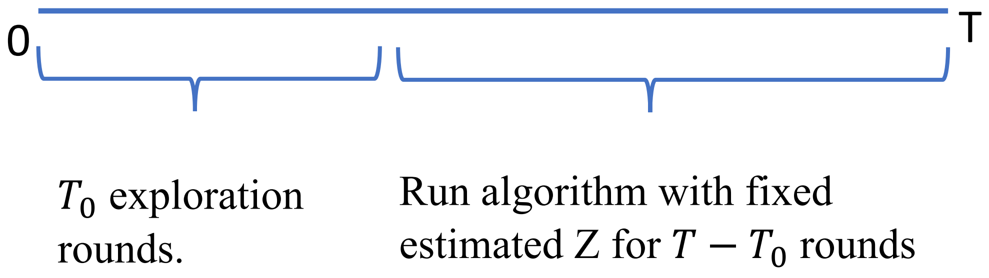
- Estimate $\hat{\mu}_t, \ \hat{W}_t$
- Select arm maximizing $\underbrace{\langle\langle \hat{\mu}_t, x_t(a) \rangle\rangle}_{\text{Reward}} - Z \underbrace{\langle \hat{W}_t x_t(a), \theta_t \rangle}_{\text{Constraints}}$

- Greedy estimates vs UCB estimates
- θ_t is a distribution over resources computed by dual online algorithm based on past resource consumptions
- Optimal regret when $Z = \frac{OPT}{B}$ where OPT is reward of optimal adaptive policy

1. Agrawal, S. and Devanur, N. R. Linear Contextual Bandits with Knapsacks. In NeurIPS, 2016.

Stochastic Smoothed LinCBwK

Decision timeline



Reward/Regret

$$\text{REW} \geq \text{OPT} - O\left(\left(\frac{\text{OPT}}{B} + 1\right) m\sqrt{T}\right)$$

$$B \geq \Omega(m^{2/3} T^{3/4})$$

- Remember smoothed context vector definition

$$x_t(a) = \nu_t(a) + g_t(a) \qquad \nu_t(a) \in \mathbb{B}_2^m, \quad g_t(a) \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{m \times m})$$

- Assume $\{x_t(a)\}_{a=1}^K \sim \mathcal{D}$ iid in each round

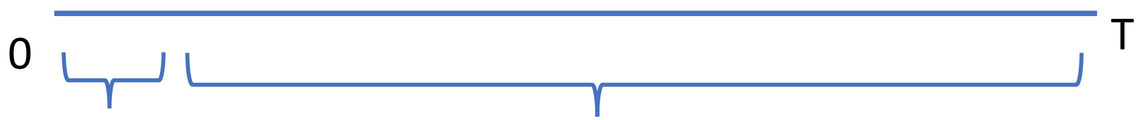
- Remember optimal $Z = \frac{OPT}{B}$. Estimate and extrapolate value of Z after T_0 rounds

- Additive regret

1. Agrawal, S. and Devanur, N. R. Linear Contextual Bandits with Knapsacks. In NeurIPS, 2016.

Adversarial Smoothed LinCBwK

Decision timeline



Explore

Run multiple epochs. New estimate of Z at beginning of epoch. Maximum $\log T$ epochs. $B/2\log T$ budget allocated for each epoch.

Reward/Regret

$$\text{REW} \geq \frac{\text{OPT}}{O(d\log T)} - O\left(\left(\frac{\text{OPT}}{B}+1\right)m\sqrt{T}\right)$$

- Remember smoothed context vector definition

$$x_t(a) = \nu_t(a) + g_t(a) \qquad \nu_t(a) \in \mathbb{B}_2^m, \quad g_t(a) \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{m\times m})$$

- Assume $\{v_t(a)\}_{a=1}^K$ are chosen by an adaptive adversary
- Z cannot be estimated without observing contexts in all rounds
- Doubling trick: Guesstimate OPT after each round, start new epoch if new guesstimate double of previous guesstimate[1]
- Competitive ratio bounds

1. Immorlica et. al. Adversarial Bandits with Knapsacks. In FOCS, 2019

Thank you!

We would like to thank the reviewers for valuable comments and questions. The research was supported by NSF grants IIS 21-31335, OAC 21-30835, DBI 20-21898, and a C3.ai research award.