Contextual Bandits with Knapsacks for a Conversion Model

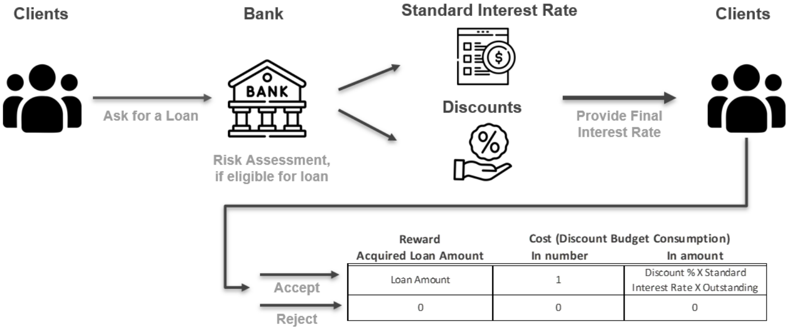
Zhen LI [1] Gilles Stoltz[2,3]

[1] BNP Paribas

[2]Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay

[3]HEC Paris

Motivation Example –
Market Share Expansion for Loans by Incentives and Discounts



- We provide numerical experiments on partially simulated data (based on the UCI Default of Credit Cards dataset)

Contextual Bandits with Knapsacks (CBwK)

Generally speaking, CBwK can be described as following:

$\rightarrow$ Various settings and algorithms based on how r_t and $\mathbf{c}_t$ are generated

For rounds $t = 1, 2, 3, \ldots, T$:

1. Context $\mathbf{x}_t \sim \nu$ is drawn independently of the past
2. Learner observes $\mathbf{x}_t$ and picks action $a_t \in \mathcal{A}$ (finite set)
3. Learner obtains scalar reward r_t and suffers vector costs $\mathbf{c}_t$ (and only gets r_t and $\mathbf{c}_t$ as feedback)

Goals: Maximize $\sum_{t \leqslant T} r_t$ while ensuring $\sum_{t \leqslant T} \mathbf{c}_t \leqslant B\mathbf{1}$

Existing settings and algorithms for CBwK

Setting #1

Badanidiyuru et al. [2014] and Agrawal et al. [2016]

I.i.d. generation of $\Big(\mathbf{x}_t,\ \big(r(a)\big)_{a\in\mathcal{A}},\ \big(\mathbf{c}(a)\big)_{a\in\mathcal{A}}\Big)$

Finite set Π of benchmark policies

Setting #2

Agrawal and Devanur [2016]

I.i.d. contexts $\mathbf{x}_t \sim \nu$ and linear structural assumptions:

$$\mathbb{E}\big[r_t(a)\,\big|\,\mathbf{x}_t \text{ \& past}\big] = \mu_\star^{\mathrm{T}}\mathbf{x}_t(a) \quad \text{and} \quad \mathbb{E}\big[\mathbf{c}_t(a)\,\big|\,\mathbf{x}_t \text{ \& past}\big] = W_\star^{\mathrm{T}}\mathbf{x}_t(a)$$

In both cases

Regret w.r.t. some optimal static policy OPT

(based on Π or the linear assumption)

Setting #3: with conversion model

For rounds $t = 1, 2, 3, \ldots, T$:

1. Context $\mathbf{x}_t \sim \nu$ is drawn independently of the past
2. Learner observes $\mathbf{x}_t$ and picks action $a_t \in \mathcal{A}$
3. Conversion $y_t \in \{0,1\}$ drawn $\sim \mathrm{Ber}\big(\eta(\varphi(a_t, \mathbf{x}_t)^{\mathrm{T}} \boldsymbol{\theta}_\star)\big)$
   Learner observes y_t, gets $r(a_t, \mathbf{x}_t)\, y_t$ and suffers $\mathbf{c}(a_t, \mathbf{x}_t)\, y_t$
   where $\eta(x) = 1/(1+\mathrm{e}^{-x})$, and where r and $\mathbf{c}$ are known functions

Goals: Maximize $\displaystyle\sum_{t \leqslant T} r(a_t, \mathbf{x}_t)\, y_t$ while ensuring $\displaystyle\sum_{t \leqslant T} \mathbf{c}(a_t, \mathbf{x}_t)\, y_t \leqslant B\,\mathbf{1}$

Contrib. #1: Protocol coupling rewards and costs through conversions

Regret definition

Short-hand notation $P(a, \mathbf{x}) = \eta(\boldsymbol{\varphi}(a, \mathbf{x})^{\mathrm{T}}\, \boldsymbol{\theta}_\star)$

Regret is (as well) w.r.t. some optimal static policy based:

$$OPT(\nu, P, B) = \max_{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})} T\, \mathbb{E}_{\mathbf{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} r(a, \mathbf{X})\, P(a, \mathbf{X})\, \pi_a(\mathbf{X}) \right]$$

$$\text{under} \qquad T\, \mathbb{E}_{\mathbf{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \mathbf{c}(a, \mathbf{X})\, P(a, \mathbf{X})\, \pi_a(\mathbf{X}) \right] \leqslant B\,\mathbf{1}$$

Reward goal: Minimize $\qquad OPT(\nu, P, B) - \sum_{t \leqslant T} r(a_t, \mathbf{x}_t)\, y_t$

New policy

1. If budget constraints violated, play no-op a_{null}
2. Otherwise,
   - Compute high-proba. upper bound $U_{t-1}(a, \mathbf{x})$ on $P(a, \mathbf{x})$
     MLE + projection, adapted from Faury et al. [2020]
   - Compute policy, i.e., mapping $\mathcal{X} \to \mathcal{P}(\mathcal{A})$:

$$\mathbf{p}_t = \underset{\pi:\mathcal{X}\to\mathcal{P}(\mathcal{A})}{\operatorname{argmax}} \; T\,\mathbb{E}_{\mathbf{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}} r(a,\mathbf{X})\, U_{t-1}(a,\mathbf{X})\, \pi_a(\mathbf{X})\right]$$

$$\text{under} \qquad T\,\mathbb{E}_{\mathbf{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}} \mathbf{c}(a,\mathbf{X})\, U_{t-1}(a,\mathbf{X})\, \pi_a(\mathbf{X})\right] \leqslant B_T \mathbf{1}$$

   - Based on context $\mathbf{x}_t$, draw action $a_t \sim \mathbf{p}_t(\mathbf{x}_t)$

$\rightarrow$ **Contrib. #2:** Algorithm based on primal formulation
(Compare to the dual formulation of, e.g., Agrawal and Devanur [2016])

Performance

Regret bound:

$$OPT(\nu, P, B) - \sum_{t \leqslant T} r(a_t, \mathbf{x}_t)\, y_t = \widetilde{\mathcal{O}}\Big(\big(1 + OPT(\nu, P, B)/B\big)\sqrt{T} \Big)$$

Orders in magnitude in T comparable to other CBwK regret bounds (Badanidiyuru et al. [2014] and Agrawal and Devanur [2016])

Summary of key restrictions and assumptions:

– Setting #1: Finite set Π of benchmark policies
– Setting #2: Heavy assumption of linear structure
– Setting #3: Finite set $\mathcal{X}$ of contexts

Reference

S. Agrawal and N. Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems (NeurIPS'16)*, volume 29, 2016.

S. Agrawal, N.R. Devanur, and L. Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT'16)*, volume PMLR:49, pages 4–18, 2016.

A. Badanidiyuru, J. Langford, and A. Slivkins. Resourceful contextual bandits. In *Proceedings of the 27th Conference on Learning Theory (COLT'14)*, volume PMLR:35, pages 1109–1134, 2014.

L. Faury, M. Abeille, C. Calauzenes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, volume PMLR:119, pages 3052–3060, 2020.