Bypassing the Monster: A Faster and Simpler Optimal Algorithm for Contextual Bandits under Realizability

Yunzong Xu
MIT

Joint work with David Simchi-Levi (MIT)

July 18
RealML @ ICML 2020

Stochastic Contextual Bandits

- For round $t = 1, \cdots, T$
  - Nature generates a random context x_t according to a fixed unknown distribution $D_{context}$
  - Learner observes x_t and makes a decision $a_t \in \{1, \dots, K\}$
  - Nature generates a random reward $r_t(x_t, a_t) \in [0,1]$ according to an unknown distribution D_{x_t,a_t} with (conditional) mean

$$\mathbb{E}[r_t(x_t, a_t)|x_t = x, a_t = a] = f^*(x, a)$$

- We call f^* the ground-truth reward function
- In statistical learning, people use a function class F to approximate f^*. Some examples of F:
  - Linear class / high-dimension linear class / generalized linear models
  - Reproducing kernel Hilbert spaces
  - Lipschitz and Hölder spaces
  - Neural networks

Challenges

- We are interested in contextual bandits with a general function class F
- Realizability assumption:

$$f^* \in F$$

- **Statistical challenges**: how to achieve the minimax optimal regret for a general function class F?
- **Computational challenges**: how to make the algorithm computational efficient?

- Existing contextual bandits approaches cannot simultaneously address the above two challenges in practice, as they typically
  - Rely on strong parametric/structural assumptions on F (e.g., UCB variants and Thompson Sampling)
  - Become computationally intractable for large F (e.g., EXP4)
  - Assume computationally expensive or statistically restrictive oracles that are only implementable for specific F (a series of work on oracle-based contextual bandits)

Research Question

- Observation: the statistical and computational aspects of "offline regression with a general F" are very well-studied in ML

- Can we reduce general contextual bandits to general offline regression?

- Specifically, for any F, given an offline regression oracle, i.e., a least-squares regression oracle (ERM with square loss):

$$\min_{f \in F} \sum_{t=1}^{s} (f(x_t, a_t) - r_t(x_t, a_t))^2 ,$$

can we design an algorithm that achieves the optimal regret via a few calls to this oracle?

- An open problem mentioned in Agarwal et al. (2012), Foster et al. (2018), Foster and Rakhlin (2020)

Our Contributions

- We provide the first optimal and efficient **offline-regression-oracle-based** algorithm for general contextual bandits (under realizability)
  - The algorithm is much simpler and faster than existing approaches to general contextual bandits
- We provide the first universal and optimal **black-box reduction** from contextual bandits to offline regression
  - Any advances in offline (square loss) regression immediately translate to contextual bandits, statistically and computationally