Contextual multi armed bandits

A trailer

Hubert Drazkowski

Faculty of Mathematics and Information Systems
Warsaw Univeristy of Technology

22.03.2022

Conspectus

1 Intro

2 Notation, definitions, applications

3 A context

4 Problems

5 Bibliography

Intro

Scope

- Working thesis : "Optimization in contextual multi armed bandits."
- Under: Professor Jaroszewicz S.

A few current ideas in the scope.

I Describe why bandits and especially contextual ones are awesome.

II Introduce mathematical objects used in bandits.

III Describe the algorithms and proofs for the multi armed bandits problem that are contextual ones built on.

IV Start with stochastic linear contextual bandits, some theory and a few algorithms.

V Move for stochastic nonlinear contextual bandits, some theory and a few algorithms.

VI Discuss relaxing an assumption (which?).

VII Compare algorithms on a high level.
  - Synthesize ideas, compare bounds etc.
  - Design a simmulation to compute performance of chosen algorithms on a synthetic and real world dataset.

Scope

- Working thesis : "Optimization in contextual multi armed bandits."
- Under: Professor Jaroszewicz S.

A few current ideas in the scope.

! Describe why bandits and especially contextual ones are awesome.

!! Introduce mathematical objects used in bandits.

!! Describe the algorithms and proofs for the multi armed bandits problem that are contextual ones built on.

!V Start with stochastic linear contextual bandits, some theory and a few algorithms.

V Move for stochastic nonlinear contextual bandits, some theory and a few algorithms.

V! Discuss relaxing an assumption (which?).

V!! Compare algorithms on a high level.

  - Synthesize ideas, compare bounds etc.
  - Design a simmulation to compute performance of chosen algorithms on a synthetic and real world dataset.

Map of learning

Actions	don't change state of the world	change state of the world
Learning model of outcomes	**Multi-armed bandits**	**Reinforcement Learning**
Given model of stochastic outcomes	**Decision theory**	**Markov Decision Process**

Table: Reasoning under uncertainity

Other honorable mentions:

- Game theory
- Partiall monitoring

A problem formulation

? What is the problem?

- A given fixed limited set of resources must be allocated between alternative choices. The allocation should maximize a gain from those choices. Expected gain from the alternatives might be learnt with statistics during the process.

A problem formulation

? What is the problem?

- A given fixed limited set of resources must be allocated between alternative choices. The allocation should maximize a gain from those choices. Expected gain from the alternatives might be learnt with statistics during the process.

Description

The idea behind the name of multi armed bandits:

1. There are multiple arms that each give reward upon pulling one.
2. An agent needs to make a sequence of decisions in moments $1, 2...T$.
3. At each time t the agent is given a set of K arms and has to decide which one arm to pull.
4. Agent wants to maximize a cummulative reward over time.
5. Pulling one arm gets the reward sampled from an unknown a priori distribution.

Contextual:

We could observe a side information in a given moment in time. A reward then could be dependent on this context.

Description

The idea behind the name of multi armed bandits:

1. There are multiple arms that each give reward upon pulling one.
2. An agent needs to make a sequence of decisions in moments $1, 2...T$.
3. At each time t the agent is given a set of K arms and has to decide which one arm to pull.
4. Agent wants to maximize a cummulative reward over time.
5. Pulling one arm gets the reward sampled from an unknown a priori distribution.

Contextual:
We could observe a side information in a given moment in time. A reward then could be dependent on this context.

The clue

The clue of the problem is the exploitation vs exploration dillema.

1. Efficiently comparing distributions.
2. Dynamically updating confidence about the above.

The beginnings are due to a certain question. Can we better approach drug testing?

- Thompson (1933) "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples"
- Robbins (1952) "Some aspects of the sequential design of experiments"

Fertile ground

According to Peter Whittle the problem was considered during the second world war.
"Efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage"

years (diff) results	
2001 - 2005 (4)	1000
2006 - 2010 (4)	2700
2011 - 2015 (4)	7000
2016 - 2018 (2)	7000
2019 - 2021 (2)	15000

Table: Google scholar results for a phrase "bandit algorithm"

Notation, definitions, applications

Definitions

Now we will dive into building blocks for the simplest setting for stochastic bandits.

- Known parameters:
  - $1, ..., K$ arms that construct an action set $A = (a_1, ...a_K)$,
  - a time horizon T, with rounds $1, ..., T$.
  - an enviroment class $\mathcal{E}$
- Unknown parameters:
  - reward distribution D_a for each arm a,
  - a reward X_t independently sampled from a D_a,
  - an enviroment instance E that lies in some enviroment class $\mathcal{E}$.
- In each round:
  - an algorithm chooses an action a_t from an action set A,
  - observes a reward x_t sampled from D_{at},
  - expands history $H_{t+1} = (A_1, X_1, ...A_t, X_t)$.

Definitions

- A policy is a mapping $\pi : H \to A$.
- An enviroment is a mapping $E : H \to X$.
- A regret R_t relative to a policy π is

$$R(T, \pi) = E_\pi \sum_i^T X_i - \sum_i^T x_i.$$

- A regret relative to a set of policies Π is

$$R(T, \Pi) = \max_{\Pi} X_t * T - E_\pi \sum_i^T X_i.$$

Applications

1. A/B testing
2. Advert placement
3. Recommendagtion services
4. Network routing
5. Dynamic pricing

Tree searches, Resource allocation, Randomized controlled trials, etc. ... an ocean of exploration ...

Applications

1. A/B testing $\sim$ solving non adaptivity, what drug should be tested more often.
2. Advert placement $\sim$ set of adverts, clicks, context, delayed feedback, different metrics.
3. Recommendation services $\sim$ Netflix, large space of actions, short horizon.
4. Network routing $\sim$ every path an action, combinatorally demanding, Monte Carlo Tree Search.
5. Dynamic pricing $\sim$ structured rewards - partial feedback, infinite space of actions.

and a lot to be done.

Applications

1. A/B testing $\sim$ solving non adaptivity, what drug should be tested more often.
2. Advert placement $\sim$ set of adverts, clicks, context, delayed feedback, different metrics.
3. Recommendation services $\sim$ Netflix, large space of actions, short horizon.
4. Network routing $\sim$ every path an action, combinatorally demanding, Monte Carlo Tree Search.
5. Dynamic pricing $\sim$ structured rewards - partial feedback, infinite space of actions.

and a lot to be done.

Applications

1. A/B testing $\sim$ solving non adaptivity, what drug should be tested more often.
2. Advert placement $\sim$ set of adverts, clicks, context, delayed feedback, different metrics.
3. Recommendation services $\sim$ Netflix, large space of actions, short horizon.
4. Network routing $\sim$ every path an action, combinatorally demanding, Monte Carlo Tree Search.
5. Dynamic pricing $\sim$ structured rewards - partial feedback, infinite space of actions.

and a lot to be done.

Applications

1. A/B testing $\sim$ solving non adaptivity, what drug should be tested more often.
2. Advert placement $\sim$ set of adverts, clicks, context, delayed feedback, different metrics.
3. Recommendation services $\sim$ Netflix, large space of actions, short horizon.
4. Network routing $\sim$ every path an action, combinatorally demanding, Monte Carlo Tree Search.
5. Dynamic pricing $\sim$ structured rewards - partial feedback, infinite space of actions.

and a lot to be done.

Applications

1. A/B testing $\sim$ solving non adaptivity, what drug should be tested more often.
2. Advert placement $\sim$ set of adverts, clicks, context, delayed feedback, different metrics.
3. Recommendation services $\sim$ Netflix, large space of actions, short horizon.
4. Network routing $\sim$ every path an action, combinatorally demanding, Monte Carlo Tree Search.
5. Dynamic pricing $\sim$ structured rewards - partial feedback, infinite space of actions.

and a lot to be done.

Tasks

What are the main tasks of the field ?

* Designing algorithms, exploring new ways to dynamically compare multiple distributions, usually in terms of means and setting better measures of confidence in that comparison.
* Proving lower bounds on the regrets of an enviroment classes.
* Proving upper bounds on the regrets of algorithms for an enviroment classes.
* Relaxing assumptions and exploring new enviroment classes.

A context

Problem protocol

W should add an unknown parameter $\theta \in \theta = (\theta_a \in R^d : a \in A)$ specific for an arm.
The natural regret in this setting is built on the same notion as for standard bandits

$$R_n = E[\sum_{z \in Z} \max_{a \in A} \sum_{t \in [T]: z_t = z} (x_{ta} - X_t)].$$

Then for each round

1. An algorithm observes a context.
2. An algorithm picks an arm.
3. A reward dependent on the context is realized.
4. An algorithm updates history.

In the news article setting, personalized (contextual) algorithm have beaten the regular version by a 12.5% click uplift. [Li at al. 2012]

Versions

Let as pick a context $z, z' \in Z$ from a context space.

Lipschitz bandits

$$E(X_a|z) - E(X_a|z') \leq L|z - z'|$$

Linear bandits

$$E(X_a|z) = z\theta_a$$

Policy class bandits

Let us take a policy $\pi : Z \to A$ and a distribution P_z over contexts.

$$E(\pi) = E_{z \in P_z}[E(\pi(X)|z)]$$

Versions

Let as pick a context $z, z' \in Z$ from a context space.

Lipschitz bandits

$$E(X_a|z) - E(X_a|z') \leq L|z - z'|$$

Linear bandits

$$E(X_a|z) = z\theta_a$$

Policy class bandits

Let us take a policy $\pi : Z \to A$ and a distribution P_z over contexts.

$$E(\pi) = E_{z \in P_z}[E(\pi(X)|z)]$$

Versions

Let as pick a context $z, z' \in Z$ from a context space.

Lipschitz bandits

$$E(X_a|z) - E(X_a|z') \leq L|z - z'|$$

Linear bandits

$$E(X_a|z) = z\theta_a$$

Policy class bandits

Let us take a policy $\pi : Z \to A$ and a distribution P_z over contexts.

$$E(\pi) = E_{z \in P_z}[E(\pi(X)|z)]$$

Ideas for algorithms

1. 1. UCB (upper confidence bound): find an estimator $\hat{\mu}_n(X_a)$ of a mean reward and another one which measure the uncertainity $\hat{\sigma}_n(X_a)$ Then solve for

$$a_{t+1} = arg \max_{a \in A}(\hat{\mu}_n(a) + \hat{\sigma}_n(a))$$

2. 2. Thompson sampling: specify prior on θ that gowern rewards, calculate posteriors. The uncertainiy comes from the prior but reduces with the amount of data etc.
3. 3. ϵ - greedy Current best mean reward should be chosen, but with a changing probability over all arms. Experiment randomly across arms with lower probability that decreases to zero as more observations come and the current best is chosen more frequently.

Problems

Some possible topics

There are some possible angles to attack the field

- Moving beyond reliazability assumption (about knowing the true function class of reward)
- Tackling nonstationary distributions of X_t
- Inventing more efficient algorithms
- Model selection
- Causal interpretation of bandits
- Adapting bandits to certain applications
- Infinately many arms
- ... and probably many more fine grained

Bibliography

Books

- Tor Lattimore, Csaba Szepesvári (2020). Bandit Algorithms, Cambridge University Press
- Sebastien Bubeck, Nicolo Cesa – Bianchi (2012). Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems, Now Publihsers
- Aleksandrs Slivinks (2019). Introduction to Multi-Armed Bandits, Foundations and Trends in Machine Learning, Vol 12, No 1-2, 1-286.

Articles

- Chih-Chun Wang, Sanjeev Kulkarani, Vincent Poor (2005). "Bandit problems with side observations", IEEE Transactions on Automatic Control, 50, 338-355.
- Li Zhou (2015). "A survey on Contextual Multi-armed Bandits", arXive.
- Lihong Li, Wei Chu, John Langford, and Robert E Schapire. "A contextual-bandit approach to personalized news article recommendation." In Proceedings of the 19th International Conference on World Wide Web, pages 661–670. ACM, 2010.
- Bouneffouf, D., Rish, I., Aggarwal, C. (2020). "Survey on Applications of Multi-Armed and Contextual Bandits". 2020 IEEE Congress on Evolutionary Computation (CEC).
- Dimakopoulou M., Zhou Z., Athey S., Imbens G. (2018) "Estimation Considerations in Contextual Bandits" arXive.

The End