Warsaw University of Technology

FACULTY OF
MATHEMATICS AND INFORMATION SCIENCE

Master's diploma thesis

in the field of study Mathematics
and specialisation Mathematical Statistics and Data Science

Optimization in Contextual Multi-Armed Bandits

Hubert Marek Drążkowski

student record book number 316062

thesis supervisor
Prof. Szymon Jaroszewicz

WARSAW 2022

Abstract

ENGLISH TITLE

To be written at the end. 200 words at most. Problem. Method. Findings. Conclusions. Key takeaway.

Keywords: multi-armed bandits, contextual bandits, optimization, online learning, contribution

Streszczenie

POLISH TITLE

To be written at the end

Sowa kluczowe: wielorcy bandyci, kontekstowi bandyci, optymalizacja, uczenie online

Contents

Introduction **11**

1. Definitions and theoretical foundations **14**

Definitions and theoretical foundations **14**

1.1. Problem formulation 14

1.2. What is and what is not MAB 15

1.3. Probability space 15

1.4. Definitions 16

1.5. Canonical Bandit model 20

1.6. Context 21

2. Classical multi-armed bandits theory **22**

2.1. Bayesian interpretation of bandits - Thompson Sampling 22

2.2. Some other section 22

3. Contextual multi-armed bandits theory **23**

4. Experiments **24**

5. Theoretical extensions **25**

Conclusions **26**

Introduction

- What is the thesis about? What is the content of it? What is the Author's contribution to it?
- A brief description of what is the problem of multi armed bandits and online learning with a special emphasis on adding the context.
- Arguments for why this topic is interesting from mathematical, informatics and real life application side, again stressing the contextual variant.
- Stating the distinction between reinforecement learning, game theory, decision theory and multi armed bandits.
- Stating what is the main problem of the thesis and main challange.
- Showing the novelty of the work in the contrast to what was written already and showing the scientific impact it might have.
- A brief description on what essantialy is covered by subsequent paragraphs, how they are linked to each other.

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.

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

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." (Whittle, 1979).

The topic is exponentially raising in popularity. For the search through years 2001 - 2005, there were 1000 google scholar results answering a query for a phrase "bandit algorithm". For a condition for a work to beproduced between 2019 - 2021, the same query resulted in 15000 papers.

In the most general formulation of the problem, the framework could model quite impresive number of applications. A few honorable mentions could be: improved, adaptive A/B testing, advert placement, recommendation services, network routing, dynamic pricing, tree searches, dynamic resource allocation, randomized controlled trials, etc.

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

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"

Bandit problems give a rise to a framework for solving how to make decisions over time under uncertainity. A given fixed limited set of resources must be allocated between alternative choices. The allocation should maximize cumulative gain from those choices over some fixed time. The gains are not known a priori, but expected gain from the alternatives might be learnt with statistics during the process.

In the news article setting, personalized (contextual) algorithm have beaten the regular version by a 12.5% click uplift. (Li et al., 2010)

1. Definitions and theoretical foundations

In this section a common ground for future meditations on the topic will be set up. The chapter will start with

More formal statement of the problem.

A revision of basic assumptions to construct a probability space with special part devoted to the canonical bandit space. A few fundamental deifinitions of building blocks for bandits will be formulated. Reward, An action, An Arm, Regret, History, Enviroment, Policy, Context, Regret. A special case of enviroment class of stochastic bandits will be brought to light.

1.1. Problem formulation

The idea behind the name of multi armed bandits comes from the quest to intuitively explain the problem within the name of the domain. By the use of an intelectual experiment let us immagine going to a casino. In this casino there are multiple machines to possibly gain or loose from. The only way to know which to play is by experimenting. So in other terms there are multiple arms that each give reward upon pulling one. An agent needs to make a sequence of decisions in moments $1, 2...T$. At each time t the agent is given a set of K arms and has to decide which one arm to pull. An agent wants to maximize a cummulative reward over time. Pulling one arm gets the reward sampled from an unknown a priori distribution. In the contextual setting the agent could observe a side information in a given moment in time. A reward then could be dependent on this context.

The clue of the problem is the exploitation vs exploration dillema. Exploration considers acquiring new knowledge, update confidence about reward distributions. Exploitation is related to trying to leverege gained knowledge to maximize the reward.

The answers to such a problem helps us understand two things. First how to efficiently compare distributions. Second, how to dynamically update the confidence about the above.

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 1.1: Reasoning under uncertainity

1.2. What is and what is not MAB

What is online learning? In an online learning an agent (a learner) has to make a sequence of decisions, with a goal to accurately predict the optimal outcome. Predictions are made given some knowledge of quality of previous predictions. Each learner is embeded in some enviroment, cast in an enviroment, so a space of possibilities of information and quality of predictions. Sometimes the bandit problem is understood as a game between a learner and an enviroment.

Distinction between MAB and other fields Multi armed bandit problem is a part of reinforcement learning (Sutton and Barto, 2018). Whereas, it is a degenerate case of big field of science. In Reinforcement learning current actions can change the future enviroment. In the bandit problem the current actions have no influence over the enviroment, so future distibutions of the rewards or action set. Some researches stress the distinction there rewriting that a reinforcement learning problem has to consider current actions having an impact on the enviroment.

Let us imagine that the reward is not observed at times. This problem belongs to partial monitoring. In the bandit framework the learner observes the reward in each round. In the case of where the enviroment is enxtended to specify more than one agent interacting and influencng the reward distributions conditionall on other agents actions the problem is studied in game theory, and is a part of reinforcement learning also.

1.3. Probability space

A probability space is a special kind of a measure such that the measure of the whole regarded space on which it is defined adds to one (integrates or sums). It is a triple $(\Omega, \mathcal{F}, P)$. Meaning the sample space Ω is an arbitrary non-empty set, the $\sigma - algebra$ $\mathcal{F} \subseteq 2^{\Omega}$ (also called σ-field) a set of subsets of Ω, called event probability, such that:

1. $\mathcal{F}$ contains the sample space: $\Omega \in \mathcal{F}$,

2. $\mathcal{F}$ is closed under complement : if $A \in \mathcal{F}$, then also $(\Omega \setminus A) \in \mathcal{F}$

3. $\mathcal{F}$ is closed under union if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots,$ then also $(\bigcup_{i=1}^{\infty} A_i) \in \mathcal{F}$

The probability measure $P : \mathcal{F} \to [0, 1]$ a function on $\mathcal{F}$ such that:

1. P is countably additive (also called σ-additive): if $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ is a countable collection of pairwise disjoint sets, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.
2. The measure of entire sample space is equal to one: $P(\Omega) = 1$.
3. The measure of the empty space is equal to zero: $P(\emptyset) = 0$.

(Billingsley, 2008)

1.4. Definitions

Give maybe examples to each definition? Should I write what we will not discuss also? Maybe not? Should I write everything in the best general way at once or should I divide and stakc the information? For example context.

Definition 1.1 (An Action). An *Action* is specified by an *enviroment* defined later. In a given game a learner has k options each round of T rounds. Each action is identical to the distribution connected to the action. An enviroment samples a reward from an action specific distribution given the action was played by a learner.

Definition 1.2 (Reward). A *Reward* is a random variable that is assigned to an action. For a given action we have a distribution D_a from which a given reward $X_{a,t}$ is sampled in a time t if the learner has chosen such action. The reward is drawn independently from a fixed distribution. A cumulative reward at time t is $S_t = \sum_{s=1}^{t} X_s$.

Definition 1.3 (History). A *History* is a tuple of 2 element tuples $H_{t+1} = ((A_1, X_1), ...(A_t, X_t)). \in (\mathcal{A} \times R)^t$. It is a product of interaction of a policy and a bandit. The elements within are random variables. This random variable depends on the algorithm chosen to perform the game and a distribution of the rewards for each arm.

Now let us form a fixed sequence:

$$H = ((a_1, x_1), ..., (a_t, x_t)) \in (\mathcal{A} \times R)^t$$

A *feasible t-history* is a sequence that satisfies $P[H_t = H] \geq 0$ for some bandit algorithm. Such an algorithm that fits the described law is then called *H-consistent*. To be precise this algorithm is called *H-induced algorithm*, so such that deterministically creates H. To this end, $\mathcal{H}_t$ be a set of feasible t-histories.

Definition 1.4 (Enviroment). An enviroment class $\mathcal{E}$ specifies the action set $\mathcal{A}$ and the class of distributions of the rewards for each of the specified actions. An *enviroment instance* is a mapping $\nu : H \to X$. Both are often called an *enviroment* and the distinction is based on the context. A given bandit instance, enviroment instance is in some enviroment class $\nu \in E$. In a sense in the game interpretation of the bandit framework. A learner takes an action and the enviroment reveals a reward to him, samples the reward. It is worth noting that in the bandit setting an environment instance generates the reward in response to each action from a distribution that is specific to that action and independent of the previous action choices and rewards. During our meditations we will consider unstructured bandit enviroment class which is defined as follows. This asumptions means that the learner can't infer anything about the distribution of a reward from a different arm than that already currently played. An enviroment class $\mathcal{E}$ in an unstructured enviroment class if $\mathcal{A}$ is finite and there exists set of distributions $\mathcal{D}_a$ for each $a \in \mathcal{A}$ such that

$$\mathcal{E} = \{\nu = P_a : (a \in \mathcal{A}) : P_a \in \mathcal{D}_a \forall a \in \mathcal{A}\}$$

Definition 1.5 (Policy). A *Policy* π is a sequence $(\pi_t)_t^n$ where π_t is a probability kernel from $(\Omega_{t-1}, \mathcal{F}_{t-1})$ to $([k], 2^{[k]})$, where k is the number of possible actions. Every π_t is a mapping from the *History set* to *Actions set* $\pi_t : H_t \to A$.. Sometimes a polcy is also called a strategy. A specific policy is closely related to the algorithm that defines the learner.

$\mu_a = E[X_a]$, where $X_a \sim D_a$

Definition 1.6 (Regret). A *Regret* of the learner is always taken relative to a policy π or to a set of policies Π. The regret measures the quality of a strategy taken in a given enviroment instance. It is the difference between the total expected reward using policy π for n rounds and the total expected reward collected by the learner over n rounds. The regret relative to a set of policies Π is the maximum regret relative to any policy $\pi \in \Pi$ in the set. A *competitor class* Π is the set used to measure the performance of a learner to the theoretically best possible strategy.

1. A regret R_T relative to a policy π is

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

2. A regret relative to a set of policies Π is

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

The regret clearly depends on an enviroment. The *worst case regret* is the maximum regret over all enviroments from an enviroment class.

Definition 1.7. An *immediate regret* measures the difference between the reward from an anction taken at given moment and that optimal action that could have been taken. It is also called *suboptimal gap, action gap, instant/immediate regret.* Mathematically it is expressed with

$$\Delta_a(\nu) = \mu * (\nu) - \mu_a(\nu)$$

.

There is a useful way to express regret with the suboptimal gap operator, which is shown in lemma ??.

Lemma 1.8 (Decomposition of a regret lemma). For any policy π and a stochastic bandit enviroment ν with $\mathcal{A}$ finite or countable and horizon $TinN$, the regret Rn of the policy π in ν satisfies

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a E[T_a(n)].$$

where, $T_a(t) = \sum_{s=1}^{t} 1[A_s = a]$, where $1[\cdot]$ denotes the indicator function.

Proof. We can begin the proof rewriting the sum of rewards expressed with an indicator

$$S_n = \sum_t X_t = \sum_t \sum_a X_t 1[A_t = a].$$

Furtheromore, the regret

$$R_T = T\mu * -E[S_T] = \sum_{a \in \mathcal{A}} \sum_{t}^{T} = 1E[\mu * -X_t[A_t = a]],$$

Finally

$$E[\mu* - X_t[A_t = a]|A_t] = 1[A_t = a]E[\mu* - X_t|A_t] = 1[A_t = a](\mu* - \mu_{A_t}) = 1[A_t = a]\mu* - \mu_{a_t} = 1[A_t = a]\Delta_a,$$

which ends the proof. □

Throught this master thesis we will consider a stochastic bandit framework.

Definition 1.9 (Stochastic bandit). A *Stochastic bandit* is an enviroment class understood as a collection of distributions $\nu = \{P_a : a \in \mathcal{A}\}$, where $\mathcal{A}$ is the set of avaliable actions.

Lemma 1.10 (Properties of a regret for a stochastic bandit lemma). Let ν be a stochastic bandit enviroment. Then the following properties are true:

1. For all policies π, the regret is non negative $R_n \geq 0$.

2. The policy π that plays $A_t \in_a \mu_a$ achieves $R_n(\pi, \nu) = 0$ in the time horizon.

3. If $R_n(\pi, \nu) = 0$ for soem policy π then $P(\mu_{A_t} = \mu*) = 1$ for all times t in $[t]$

Those mean that one can always find a policy for which the regret is zero and for all other it is non negative.

Proof. □

How does the process look like? Now we will dive into building blocks for the simplest setting for stochastic bandits. Let as assume that at the beginnig of the game the learner faces such a collection of objects.

- 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 $\mathcal{E}$.

Then 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 $H_{t+1} = (A_1, X_1, ...A_t, X_t)$.

1.5. Canonical Bandit model

A special case of a probability space that is usually considered in the multi armed bandit model is the canonical bandit model. In this formal characterisation I will closely follow (Lattimore and Szepesvári, 2020). For the cases considered in the master thesis the space of possible actions, the action set will be countable. That excludes the uncountable sets like in the application of dynamic pricing. We will consider a finite horizon $T \in N$.
For each $t \in [T]$, let $\Omega_t = ([k] \times R)^t \in R^{2t}$ and $\mathcal{F}_t = \mathcal{B}(\Omega_t)$. Random variables A_i, X_i are coordinate projections

$$A_t = (a_1, x_1, ...a_T, x_T) = a_t$$

$$X_t = (a_1, x_1, ...a_T, x_T) = x_t$$

The probability measure $(\Omega_T, \mathcal{F}_T)$ depends on both the enviroment and the policy. Let $v = (P_i)_{i=1}^k$ be a stochastic bandit where each P_i is a probability measure on $(R, \mathcal{B}(R))$. Two conditions have to be satisified in order to reflect the sequential nature of the interaction of the learner and an enviroment. First the conditional distribution of action A_t given the history H_t is

$$\pi_t(\cdot|H_t)$$

almost surely. Where π_1, π_2 ... is a sequence of probablity kernels that characterise the learner. The learner can't use the future observations in current decisions. Second, the conditional distribution of reward X_t given $H_t \cup A_t$ is P_{A_t} almost surely. Thus a probability measure on $(T, \mathcal{F}_T)$ has ot satisfy those assumptions. The Radon-Nikodym derivative with respect to a σ finite measure on $(R, \mathcal{B}(R))$ α for which P_i is absolutely continuous with respect to that measure.

$$\pi_i : R \to R$$

such that $\int_B \pi d\alpha = P_i(B)$ for all $B \in \mathcal{B}(R)$

$$p_{\nu\pi}(a_1, x_1, ..., a_T, x_T) = \prod_{t=1}^{T} \pi(a_t|H_T) p_{a_t}(x_t)$$

Counting measure with $\rho(B) = |B|$, the density $p_{\nu\pi} : \Omega \to R$ is defined with the respect to the product measure $(\rho \times \alpha)^T$ The $p_{\nu\pi}$ is a distribution on $([k] \times \mathcal{A})$

$\mathbf{P}_{\nu\pi} = \int_B p_{\nu\pi}(\omega)(\rho \times \alpha)^T(d\omega)$ for all $B \in \mathcal{F}_n$

1.6. Context

Definition 1.11 (Context). A *context* is an information present at each round about the conditions on which the distribution of an action will be sampled. For example in the recommendation system this might be an aditional information about a user.

(Canonical model for contextual bandit) The extension of the canonical bandit model should take into account context. Hence, let $\mathcal{A}$ and $mathcalC$ be finite sets. We will consider a stochastic bandit enviroment with an addition that a learner first observes a context $C_t \in C$. We will assume that the sample of contexts $C_1, ..., C_n$ are identically independently distribuited from a distributiin defined on a set $\mathcal{C}$. They then choose an action $A_t \in \mathcal{A}$ and receive a reward $X_t \sim P_{A_t, C_t}$. A need of formal construction that would add context.

In order to adapt the setting fully we 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

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

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.

Rewrite it in an unified way with the previous protocole

2. Classical multi-armed bandits theory

2.1. Bayesian interpretation of bandits - Thompson Sampling

This corresponds to the Bayesian viewpoint where the objective is to minimise the average cumulative regret with respect to a prior on the environment class.

Bayesian regret. Let us define to be a prior probabiliyu measure on $\mathcal{E}$. Then the Bayesian regret is

$$BR_n = (\pi,) = \int_{\mathcal{E}} R_n(\pi, \nu) d(\nu)$$

compare to

$$BR_n = (\pi,) = E_{\mathcal{E}\sim}[E[R_T(\pi, \nu)]|\mathcal{E}]$$

2.2. Some other section

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 in time 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.

3. Contextual multi-armed bandits theory

Let as pick a context $c, c' \in C$ from a context space. Lipschitz bandits

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

Linear bandits

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

Policy class bandits Let us take a policy $\pi : Z \to A$ and a distribution P_c over contexts.

$$E(\pi) = E_{c \in P_c}[E(\pi(X)|c)]$$

Four principles of stochastic linear bandits. 1) Thompson sampling 2) Optimisation based algorithms 3) Information directed sampling 4) Epsilon Greedy

4. Experiments

5. Theoretical extensions

Conclusions

- Complementary to the introduction, a very brief, essential, concluding refreshment of what was done in the paragraphs.
- What is the answer to the posted problem.
- What are the specific results and main conclusions of the work.
- What are possible extensions to the work.

Bibliography

Billingsley, P. (2008). *Probability and measure.* John Wiley & Sons.

Lattimore, T. and C. Szepesvári (2020). *Bandit algorithms.* Cambridge University Press.

Li, L., W. Chu, J. Langford, and R. E. Schapire (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: An introduction.* MIT press.

Whittle, P. (1979). Discussion of dr gittins paper. *Journal of the Royal Statistical Society 41*, 164–177.

List of symbols and abbreviations

nzw.	nadzwyczajny
*	star operator
~	tilde

If you don't need it, delete it.

List of Figures

If you don’t need it, delete it.

List of tables

1.1 Reasoning under uncertainity 15

If you don't need it, delete it.

List of appendices

1. Appendix 1

2. Appendix 2

3. In case of no appendices, delete this part.