CSE 541 Project, Spring 2024

Instructor: Kevin Jamieson

You have learned a lot about algorithms and their analyses for a variety of multi-armed bandit settings. The aim of this project is to either 1) explore additional multi-armed bandit topics not covered in this course, or 2) apply the methods you have learned to real datasets. If you are currently doing research in any domain, you are highly encouraged to try to find a use for multi-armed bandits in that research.

Bandit algorithms are inherently about collecting data in an online fashion, which makes it difficult to benchmark them on static datasets. We recommend you get creative with making realistic environments (using synthetic data is discouraged). Sometimes enormous bandit-collected datasets are available that you can "replay" by fitting distributions to the data and assume those distributions are accurate (for example, https://nextml.github.io/caption-contest-data/). If you want to test contextual bandits algorithms, one popular thing to do is convert a multi-class classification dataset into a fully-observed contextual bandit dataset (see https://arxiv.org/abs/1802.04064 for an example). Alternatively, you can take a partially collected dataset like the Movielens dataset (https://grouplens.org/datasets/movielens/100k/) of users rating movies, use matrix completion to decompose the partially completed matrix into M = U V^T where each row of U represents a d-dimensional embedding of a user, and each row of V represents a d-dimensional embedding of a movie. Then use V as the feature vectors that you can recommend, and U as the unknown user vectors of users. If you're really ambitious, you could even try finetuning a weak LLM using evaluations from a strong open-source LLM (this is a contextual bandit problem). Some additional project ideas are below.

Deliverables:

Friday May 10: At most one page proposal for your project. Include an initial set of papers that you will reference and datasets you'll use (verify that you can actually access and use them).
Monday June 3: At most three page project report in any format. For empirical work, describe your modeling assumptions, how reasonable the assumptions are (justified by the fit to data), and how you might adapt your approach if some assumptions were stretched or broken. For more theoretical work, include a thorough literature review, describe what is known, what are the highlights of the analysis--why does it work?

Example project ideas:

Streaming Bandits:

Usually all the arms are available to the algorithm and you can run algorithms like UCB to perform regret minimization. However, it might be the case that the arms arrive one by one and in that case you cannot apply algorithms like UCB. This area is broadly classified as Streaming Bandits. Some of the relevant papers in this areas are listed below:
https://proceedings.neurips.cc/paper/2021/file/a2f04745390fd6897d09772b2cd1f581-Paper.pdf
http://proceedings.mlr.press/v139/jin21a/jin21a.pdf
https://arxiv.org/pdf/2004.04666.pdf

Also check the papers that cite on Google Scholar. The aim of the project in this subtopic would be to do a literature survey, find a nice application of this model of streaming bandits, code up some of the important algorithms and compare their results for the mathematical model of the application that you found.

Recommendation Systems & Dueling Bandits

Usually in multi-armed bandits, you pull an arm and observe its reward. However, there might be scenarios where you just want to compare which arm is better. In such settings, dueling bandits is applicable. One famous application is the recommendation where the websites take your feedback by asking questions like :"Which of the following 2 movies do you like more". Dueling bandits is also used a lot to train search engines by interleaving results of two engines and rewarding the one that provides the link that is clicked on. An important paper and survey in this area are listed below:
https://proceedings.mlr.press/v162/saha22a/saha22a.pdf
https://arxiv.org/pdf/1807.11398.pdf

Also check the references of the first paper. The aim of the project in this subtopic would be to do a literature survey, find a nice application of this model of dueling bandits (for eg: recommendation systems), code up some of the important algorithms and compare their results for the mathematical model of the application that you found.

Game theory + Bandits

Multi-armed bandit problems are in general a single agent problem. That is, there is only one agent that chooses an arm and observes the reward. What if there are multiple agents? Such a setting is captured by Game theory. For example, agent X can choose an index i and agent Y can choose an index j and the agent X gets a rewards of $A_{i,j}$ and agent Y gets a reward of $B_{i,j}$. In such settings there is no concept of best arm. Instead there is a notion of equilibrium. One can use the techniques in multi-armed bandit to find such equilibrium. For example in zero sum games, if both players play EXP3 they converge to Nash equilibrium. There are multiple

models that one can think of in this interesting setting. Some of the papers are listed below:
https://arxiv.org/pdf/2006.09517.pdf
https://arxiv.org/pdf/2310.16252.pdf
https://arxiv.org/pdf/2006.05145.pdf
https://arxiv.org/pdf/2204.11417.pdf
https://arxiv.org/pdf/1706.02986.pdf
https://proceedings.neurips.cc/paper/2007/file/08d98638c6fcd194a4b1e6992063e944-Paper.pdf

Also check the references of these papers and check which papers cite these on google scholar. The aim of the project in this subtopic would be to do a literature survey, fix a mathematical model (zero-sum game, general-sum game, multi-player game, extensive form games), code up some of the important algorithms and compare their results for the mathematical model you chose. For instance, you can code up algorithms and see which ones converge faster to an equilibrium.

Fairness + Bandits

Recently fairness has been an emerging topic in ML. Can you think of ways to make your algorithms fair? For example, in medical trials you would like your drugs to be fairly distributed across all the groups that volunteer. Some of the papers in this area are listed below:

https://arxiv.org/pdf/2205.13930.pdf
https://arxiv.org/pdf/2007.06699.pdf
https://arxiv.org/pdf/2208.12584.pdf

Also check the references of these papers and check which papers cite these on google scholar. The aim of the project in this subtopic would be to do a literature survey, fix a mathematical model which captures fairness in some way, code up some of the important algorithms and compare their results for the mathematical model you chose.

Transductive (stochastic) linear bandits and its applications

In the class, we have studied the pure exploration problem in classical linear bandits. That is, we are given a set of arms $\mathcal{X}$. Everytime we pull the arm $x \in \mathcal{X}$, we are given the feedback $r = x^\top \theta^* + \xi$, where θ *is some unknown hidden vector and* ξ *are some random variables. Therefore, our goal is to find* $\arg\max_x x^\top \theta$. Please look at lecture notes for detailed algorithms.

In practice, however, we may not be allowed to pull every arm in $\mathcal{X}$. For example, in drug discovery, thousands of compounds are evaluated in order to determine which ones are effective at combating a disease, but you are only allowed to test on a limited number of

compounds that are approved by the government. On the other hand, you may also choose arms out of $\mathcal{X}$. You can refer to 1906.08399.pdf (arxiv.org) for more detailed examples.

We call this problem transductive bandits where you are given an available item set (i.e. arm set) $\mathcal{Z}$ which is different from $\mathcal{X}$. Here your goal is to find some real world application of transductive linear bandits and implement existing algorithms on that, discussing how the relationship between $\mathcal{X}$ and $\mathcal{Z}$ affect your final performance. You are also welcome to further extend this setting into contextual bandits or other bandits problems.

Adversarial Linear Bandits:

Unlike the previous setting, θ^* vector can change over time. Similar to EXP3, one can think of minimizing regret with respect to the best arm. There exists an EXP3 algorithm for adversarial linear bandit (see Chapter 27 of Lattimore book). However, the algorithm has a limitation. The guarantees hold only in expectation.

Goal: Can you implement an algorithm/heuristic that is easy to implement, has better guarantees than the existing algorithm and justify why it might work with high probability. For reference, see EXP3-IX algorithm in Chapter 12 of Lattimore book to understand how the classical EXP3 can be converted to an algorithm which has high probability guarantees. Also look at the following paper: https://www.mit.edu/~rakhlin/papers/bandit_merged.pdf

You are also welcome to further extend this setting into contextual bandits.

Multi-task representation learning bandits

In the class, we have studied various bandit algorithms for single task, that is, we are aiming to optimize only one object (i.e. one reward functions). In reality, there usually exists multiple different but related tasks. Therefore, instead of implementing bandits algorithm for each tasks separately, we can play those tasks together and therefore leveraging the relationship between each tasks.

There exist various ways to represent the relationship between tasks. You are welcome to propose a relationship structures yourself. Here, we suggests the shared representation structure. Specifically, for each task $t \in [T]$, we have

$r_t = x^\top \theta_t^* + \xi$, where ξ are some random variable θ_t^* *is some unknown hidden vector for task* t *which can be further decomposed as* $\theta_t^* = B^* w_t$ *, where* $B^* \in \mathbb{R}^{d \times k}$ is some unknown shared representation function mapping x from high dimension d to some low dimension k and $w_t^* \in \mathbb{R}^k$ is some unknown task specified parameter. Therefore, you don't need to learn $d \times T$ dimension unknown vectors. Please refer to 2010.06531.pdf (arxiv.org) for more details.

One common example for such structure is that, in computer vision tasks, despite various goals (e.g. classification, object detection, edge detection, style transfer…) They all share the same neural-net encoder (i.e. backbone).

In this project, you are expected to find one multi-task application, which can either be the shared representation structure or any other structure you can think of. (Although if you decide to do the latter, you may not be guaranteed to have an existing algorithm). Then implement a proper algorithm on the application and report your results. Of course you are also welcome to design your own problem or algorithms under this scope.

Other related reference:

cella21a.pdf (mlr.press)

du23b.pdf (mlr.press)

2202.10066.pdf (arxiv.org)

Casual bandits

Medical drug testing, policy setting, and other scientific processes are commonly framed and analyzed in the language of sequential experimental design and, in special cases, as bandit problems. However, not all experiments can be explicitly and actively conducted. For example, when predicting the effect of changes to childcare subsidies on workforce participation, or school choice on grades. In this case, people need to leverage the casual model between different variables.

1606.03203.pdf (arxiv.org) is the first work to combine the multi-arm bandits with the causal inference in which interventions are treated as arms in a bandit problem but their influence on the reward — along with any other observations—is assumed to conform to a known causal graph.

In this project, you are expected to do a survey on different casual graph settings and choose one casual dataset (which can be either simulated or from reality, but we may require certain complexity if you use a simulated dataset). Then try to find a proper algorithm to implement on the chosen dataset, study how the properties of graphs affect the final performance. You are also welcome to develop your own algorithm of course.

Other related reference:

lu20a.pdf (mlr.press)

2206.07883.pdf (arxiv.org)

Causal Bandits: Online Decision-Making in Endogenous Settings

Nonstationary bandits

In the class we have studied the classical stochastic bandits where there exists an stationary hindsight distribution (distribution of reward for multi-arm bandits and the joint-distribution of context and reward for contextual bandits). In reality, however, these distributions can be time-varying and applying classical stochastic bandits algorithm against a fixed benchmark is meaningless and wrong. For example, customer preferences vary over the course of a year, so we should not aim for a single best arm or policy.

There exists various settings and solutions to deal with this non-stationary environment. (1) 2102.05406.pdf (arxiv.org) and 2112.13838.pdf (arxiv.org) try to minimize the regret against changing optimal policies/arms, such regret measurement is called dynamic regret. (2)2302.09739.pdf (arxiv.org) and 2307.15154.pdf (arxiv.org) try to either minimize the reget or identify the best arm that agnostically achieves optimal performance against a fixed policy/arm in both stationary and non-stationary scenarios.

In this project, you are expected to do a survey on different settings of nonstationary bandits and choose one to investigate by finding some practical scenarios. Notice that some algorithms might not be easily implementable, so it's ok to implement some suboptimal algorithm as long as you have a good discussion on your result.

Alternatively, you can try to design other types of variation yourself with simulated data.

Other possible topics:

1. Applications of A/B Testing

A/B testing refers to the task of determining the best option among two alternatives that yield random outcomes. This a broad area with tons of applications and usually closely related with different bandits algorithms. Can you do a comprehensive survey of this area, come up with a few interesting applications, code up a few algorithms and show some meaningful experimental results for a suitable mathematical model for the applications you chose. Here are two papers to get started:

http://proceedings.mlr.press/v35/kaufmann14.pdf
https://www.sciencedirect.com/science/article/pii/S0164121224000542

2. Reinforcement learning

Most of the topics we talked about can be extended to reinforcement learning (especially episodic markov decision process) settings, which is more complicated due to their multi-step, state-involved sequential decision nature. This topic is a little beyond the scope of this class so we will skip the details here. But if anyone is interested in working on a RL project you can do a little bit of keywords search or look at this webpage:

https://courses.cs.washington.edu/courses/cse599m/19sp/projects.html
Note the webpage is 5 yrs old and many problems listed here might have been solved. Nevertheless, this can give you an idea what kind of project you can work on in RL.

3.LLMs + Bandits

This is more like an open topic. Large language models are a hot research topic these days. Can you do a comprehensive survey on works in the intersection of multi-armed bandits and LLMs. Following this choose a mathematical model involving bandits, code up a language model and perform some meaningful experiments. This is quite a broad area. You are free to choose any meaningful model which involves both bandits and language models. For eg: you can do the training process by adaptively collecting data using bandit algorithms and then training your LLM to perform some task or you can use bandits algorithm to choose a proper prompt
2305.03495.pdf (arxiv.org)
2401.06692.pdf (arxiv.org)

4. Bandits with Heavy Tail
Usually we study bandits under the assumption that the distributions are sub-gaussian. What if they are heavy tailed instead. Can you study the existing Bandit literature like Stochastic bandits, linear bandits under such distributions. Look at one such paper:
http://sbubeck.com/BCL13.pdf
Also the definition of heavy tailed distribution:
https://en.wikipedia.org/wiki/Heavy-tailed_distribution

5. Best of Both worlds.
Typically bandit algorithms are designed separately for stochastic setting and for adversarial bandit setting. Can you think of single algorithm which works for both the settings simultaneously? Look at one such paper:
https://jmlr.csail.mit.edu/papers/volume22/19-753/19-753.pdf

Look for similar papers in other bandit areas.

6. Reinforcement Learning from Human Feedback (RLHF)
RLHF is a generalization of Dueling Bandits. Can you study existing literature and think of some nice project idea? Here is a paper you can look at:
https://arxiv.org/pdf/2312.14925.pdf
7. Adversarial Transductive Linear Bandits
Can you implement an algorithm/heuristic that is easy to implement and has good guarantees for adversarial transductive linear bandits. Formally, the problem is defined as follows. Say an adversary chooses a sequence $\theta_1, \theta_2, \ldots \theta_T$ beforehand (this is an oblivious adversary). In each round t, you choose a best arm candidate $z_t \in \mathcal{Z}$ and an observation arm $x_t \in \mathcal{X}$. Then you observe $\langle x_t, \theta_t \rangle$. Now you aim to minimize the following regret:

$$R(T) = \max_{z \in \mathcal{Z}} \sum_{t=1}^{T} \langle z, \theta_t \rangle - \sum_{t=1}^{T} \langle z_t, \theta_t \rangle$$