A survey on Multi-Armed, Contextual and Causal bandit algorithms for online learning

Nihaar Shah (ns3413@columbia.edu)

April 2020

1 Introduction

The multi armed bandit problem has been studied since the 1930s (by William Thompson) and has evolved over time with several additional components such as bandits in an adversarial setting, bandits in a stochastic setting and contextual bandits with side information. The need for better sequential decision making under uncertainty and experimental design is of ever increasing importance in fields as diverse as drug trials, web advert placement and personalized medicine to economic policy making and game playing.
The goal of this survey is to first elucidate two classic results within the Multi-Armed bandit and Contextual bandit settings respectively namely Exp3 and Exp4 [Aue+02] algorithms. Secondly, this survey discusses a recent advancement in applying causal inference to the bandits problem which answers the question " Where to intervene?". The approach in [LS18] uses ideas from importance sampling and applies this to a Markovian causal graphs. There will be a focus on proof techniques that were used to derive bounds in each algorithm and any similarities between them.

2 Definitions

Multi-Armed Bandit problem the agent makes a sequence of decisions $1, 2, .., T$ and at each time t, the agent is given a set of K arms from which it must decide which to pull. The agent receives a reward associated with the arm it pulled while the rewards of the remaining arms are unknown.
Stochastic bandits: The agent receives a reward that is sampled from a distribution that is unknown to the agent.
Adversarial bandit environment: The rewards are not sampled from a distribution but chosen by an adversary and can depend on all previous actions of the agent.
Contextual bandits problem: There is a distribution P over $(x, r_1, ..., r_k)$ where x is context (or side information) $a \in \{1, ..., k\}$ is one of the k arms to be pulled and $r_a \in [0, 1]$ is the reward for arm a. The problem is a repeated game: on each round, a sample $(x, r_1, ..., r_k)$ is drawn from P, the context x is announced, and then for precisely one arm 'a' chosen by the player, its reward r_a is revealed.
Simple regret: difference between the return of the optimal action and that of the action chosen by the algorithm after T rounds.

3 Preliminaries

3.1 Importance weighted estimators

A crucial part of the adversarial bandits setting is a mechanism to estimate the reward of unplayed arms i X_i. We can borrow ideas from importance sampling where the goal is to evaluate $E[f(x)] = \int f(x)p(x)dx$ and $x^{(j)} \sim p(x)$ however we can't sample from $p(x)$ but can just evaluate it for a given x. So we use a proposal distribution $q(x)$ from which we can *sample* to draw samples and then re-weight them. Samples for which $q(x) > p(x)$ will be over-represented while those for which $q(x) < p(x)$ will be under-represented. So the weight $w_j = \frac{p(x^{(j)})}{q(x^j)}$ would compensate for this.[Bis06]

Thus,

$$\begin{aligned} E[f(x)] &= \int f(x)\frac{p(x)}{q(x)}q(x) \\ &\approx \frac{1}{J}\Sigma_{j=1}^{J}\frac{p(x^{(j)}}{q(x^{(j)})}f(x^{(j)}) \end{aligned} \tag{1}$$

Inverse Propensity Weighting (IPW) In a similar vein, IPW is often used in causal inference when it is not possible to conduct a controlled experiment but there is observational data that can be used to estimate a potential outcome (counterfactual). This is valid if there are known to be no common causes or confounders between treatment and effect variables. The IPW formula used is: $\hat{\mu}_{a,n} = \frac{1}{n}\Sigma_{i=1}^{n}\frac{Y_i\mathbb{1}[A_i=a]}{\hat{p_n}(A_i=a|X_i)}$ [RRZ94] which has parallels to the Importance weighted formula above: $\hat{p}$ & q, $\Sigma_{i=1}^{n}\frac{\mathbb{1}[A_i=a]}{n}$ & p.

3.2 Exp3 algorithm

Exponential-weight algorithm for exploration and exploitation (Exp3) is an algorithm for the adversarial bandits setting. The main idea is to maintain a list of weights $w_i(t)$ for each action i in round t and using these weights to calculate a distribution $P_{t,i} = \frac{\exp \eta\hat{X}_{t-1,i}}{\Sigma_{j=1}^{k}\exp(\eta\hat{X}_{t-1,j})}$ and sample an action A_{i_t} to take next. Then it **estimates the rewards** $\hat{X}_t$ (using IPW) for all the actions based on the observed reward X_t. This estimated reward is used to increase and decrease the weights when the payoff is good or bad according to the update rule $w_{t+1,i} = w_{t,i}e^{-\eta\frac{\hat{X}_{t,i}}{p_{t,i}}}$. $\eta \in [0,1]$ in $P_{t,i}$ tunes the trade-off between exploration of new actions ($\eta = 0$ gives uniform chance to all actions) and exploiting actions with known high reward $\eta = 1$.
Estimating Rewards The intuitive reason to estimate rewards using IPW is to compensate for a potentially small probability of getting the observed reward. More formally, it ensures that the conditional expectation of the "estimated reward" is the actual reward. Let $P_{ti} = P(A_t = i|A_1, X_1, .., A_{t-1}, X_{t-1})$ i.e the probability of choosing the ith action at the t-th round conditioned on all previous actions and rewards $(\mathbf{A}, \mathbf{X})$. Then the estimated reward by the end of round t, where $\hat{X}_{ti}$ is given by:

$$\hat{X}_{ti} = \frac{\mathbb{1}[A_t = i]}{P_{ti}}X_t \tag{2}$$

To check if $\hat{X}_{ti}$ (a random variable because it is dependent on A_t, P_t, X_t is a good estimate of the actual x_{ti} we can calculate the mean and variance of this reward estimate. The conditional expectation of the estimate is:

$$E[\hat{X}_{ti}|A_1, X_1, .., A_{t-1}, X_{t-1}] = E_t[\hat{X}_{ti}] = E_t[\frac{A_{ti}}{P_{ti}}x_{ti}] = \frac{x_{ti}}{P_{ti}}E_t[A_{ti}] = x_{ti}$$

This is by noting that $E_t[A_{ti}] = P_{ti}$ and so we find that $\hat{X}_{ti}$ is an unbiased estimator of x_{ti} the true reward. The conditional variance on the other hand is : $\mathbb{V}_t[\hat{X}_{t,i}] = x_{ti}^2(\frac{1-P_{ti}}{P_{t,i}})$ which can be quite large especially if $P_{t,i}$ is small. This is indeed problematic sometimes and should be noted. The regret for n rounds and k arms is $R_n(x) \leq 2\sqrt{nk\log(k)}$ and while the proof isn't in the scope of this survey, the important idea used in it is the inequalities $e^x \leq 1 + x + x^2$ for $x \leq 1$ and $1 + x \leq e^x$ from Taylor's expansion.[LS18]

4 Exp4 algorithm

So far in the MAB problem we have only considered being informed about which actions to take by the observed rewards. In most real-world settings it is common to have side information that can be supplemented to inform the choice of actions for example a news recommendation engine would have access to the location, time, perhaps age and gender of the user. Naively, one could use a K-armed bandit algorithm on each context independently but that would be ignoring any relationships between contexts where learning one could help in inferring others. So we should consider grouping contexts in some way and assign a bandit to each group. The *contextual* bandit problem in the adversarial setting differs from the Exp3 setting in that we get a context vector. Instead of scoring actions, the learner must score experts. "Expert"

refers to a learner for each policy in the hypothesis space of policies. Advice vector $\zeta_{\mathbf{t}}^{\mathbf{N}}$ is the Nth expert's advice vector containing distribution over K arms indicating the recommended probability of playing each arm at time t. The agent's goal is to combine the advice of the experts in such a way that its total reward is close to that of the best expert (in contrast to just the best single action in Exp3). The algorithm is reproduced in Algorithm 1 from [LS18]. Notably, weight update step in line 11 considers the rewards obtained by following expert advice i to proportionally update that expert's weights. The memory cost $O(N)$ for N experts and time complexity is $O(N+K)$ per round. Regret is bounded by $O(\sqrt{TK\log N})$.

Algorithm 1 Exp 4 algorithm

1: Require $\gamma \in (0,1]$. Set $w_{t,i} = 1$ for $i = 1,..,N$
2: **for** t=1,..,T **do**
3: Receive advice $\{\zeta_{\mathbf{t}}^{\mathbf{1}}, \zeta_{\mathbf{t}}^{\mathbf{2}}, ., ., \zeta_{\mathbf{t}}^{\mathbf{N}}\}$
4: **for** j = 1,..,K **do**

$$p_{t,j} = (1-\gamma)\Sigma_{i=1}^{N} \frac{w_{t,i}\zeta_{t,j}^{i}}{\Sigma_{i=1}^{N} w_{t,i}} + \frac{\gamma}{K}$$

5: **end for**
6: Draw action a_t according to $\mathbf{p_t}$(which was just populated) and receive reward r_{a_t}
7: **for** j =1,...,K **do**
8: Calculate the unbiased estimator of r_t

$$\hat{r_{t,j}} = \frac{r_{t,j}}{p_{t,j}} \mathbb{I}[j = a_t]$$

9: **end for**
10: **for** i = 1,..,N **do**
11: Calculate the estimated expected reward and update weight

$$\hat{y}_{t,i} = \zeta_t^i \hat{r}_t$$

$$w_{t+1,i} = w_t e^{\gamma \hat{y}_{t,i}/K}$$

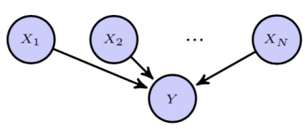
12: **end for**
13: **end for**
[LS18]

5 Causal Bandits: Where to intervene?

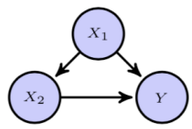
In the contextual bandits setting we encountered the case where in side information may be related which is why it is beneficial to compete different policies that can view all of this side information. Now we consider the case where side information comes from a Causal graph representing relations between the variables in the context. For instance going back to the news recommendation engine, if we knew that age (X_1) and gender (X_2) directly affects articles of a certain type (say health related (Y)) then it can be represented as in fig 1a.
We first consider the parallel bandits algorithm that deals with graphs that look like fig 1a and then more general graphs. Finally there will be a discussion on similarities to previous algorithms to the causal case.

5.0.1 Problem Setup

In each round, the learner can either purely observe by selecting $do()$ or set the value of a single variable. The remaining variables are simultaneously set by independent biased biased coin flips. Formally, when not intervened upon $X_i \sim$Bernoulli(q_i) where $\mathbf{q} = (q_1, ..., q_N) \in [0,1]^N$ so that $q_i = P\{X_i = 1\}$. This approach uses importance sampling ideas and re-weights unobserved actions which is similar to the idea used in EXP3 and EXP4 that we have seen above. This approach also considers rather simple graphs where the treatment and observed variables are direct causes (i.e. share a parent-child relationship). To start with consider the causal graph in fig 1a with



(a) Parallel graph



(b) common-cause

the following definitions:

$$[X_1, ., ., X_i, ., ., X_N]_\tau \in \{0,1\}^N \text{for round t} = \tau$$

At any given round $t = \tau$ the learner can only intervene on at most a single random variable X_γ i.e. set $X_\gamma = j, j \in \{0,1\}$. The remaining $X_i, i \in \{1,2,..,N\} \setminus \gamma$ are drawn from $\{0,1\}$ according to a Bernoulli parametrized by a random vector $\mathbf{q}$:

$$\mathbf{q}_\tau = [q_1, ., ., q_i, ., .q_N]_\tau \in [0,1]$$

$\mathbf{q}$ is not known to the learner and so the algorithm estimates this from the initial period of observation where no interventions are done. So all we have access to is an estimate of $\mathbf{q}$ namely $\hat{\mathbf{q}}$.
Lastly the rewards Y_t are determined by a fixed and unknown mapping $\mathbf{r} : \{0,1\}^N \to [0,1]$.

5.0.2 Algorithm for Parallel graphs

Alg 2 is reproduced from [LLR16]. After observing for the first $T/2$ rounds, the

Algorithm 2 Parallel Bandits

1: Input: Total rounds T and N
2: **for** t $\in T/2$ **do**
3: Perform empty intervention do()
4: Observe $\mathbf{X_t}$ and Y_t
5: **end for**
6: **for** $a = do(X_i = x) \in A$ **do**
7: Count times $X_i = x$ seen: $T_a = \Sigma_{t=1}^{T/2} \mathbb{1}[X_{t,i} = x]$
8: Estimate reward: $\hat{\mu}_a = \frac{1}{T_a} \Sigma_{t=1}^{T/2} \mathbb{1}[X_{t,i} = x] Y_t$
9: Estimate probabilities $\hat{p}_a = \frac{2T_a}{T}$, $\hat{q}_i = \hat{p}_{do(X_i=1)}$
10: **end for**
11: Compute $\hat{m} = m(\hat{\mathbf{q}})$ and $A = \{a \in A : \hat{p} \leq \frac{1}{\hat{m}}\}$
12: Let $T_A := \frac{T}{2|A|}$ be the times to sample each $a \in A$
13: **for** a = do$(X_i = x) \in A$ **do**
14: **for** $t \in 1, .., T_A$ **do**
15: Intervene with a and observe Y_t
16: **end for**
17: Re-estimate $\hat{\mu} = \frac{1}{T_A} \Sigma_{t=1}^{T_A} Y_t$
18: **end for**
19: Return: estimated optimal $\hat{a}* \in \arg\max_{a \in A} \hat{\mu}$

learner has access to $(\mathbf{X}, \mathbf{Y})$ where $\mathbf{X} \in \{0,1\}^{T/2 \times N}$ and $\mathbf{Y} \in R^{T/2}$. Consequently the learner has the following distribution $P(\mathbf{Y}|\mathbf{X})$. This is shown in lines 1-4 of algorithm 2. In our particular causal graph in fig 1a due to the direct causal relation between X_i and Y we can conclude that intervening on $do(X_i = j)$ has the same effect on Y as observing that $X_i = j$ i.e. $P(Y|do(X_i = j) = P(Y|X_i = j)$. Effectively we have the interventional distribution.
Now the set of all possible actions consists of setting each of the N variables X_i to $0 or 1$ therefore $|A| = 2^{N-1}$. The algorithm iterates through each of these possible interventions and counts how many times it was observed so far. Say $X_7 = 0$ was observed 42 times out of $T/2 = 100$ rounds then it estimates the probability of this intervention $P(X_7 = 0) = 42/100$ and its complement $P(X_7 = 1) = 1 - 0.42 = 0.58$ so $q_7 = 0.58$. This is done in lines 7 and 9. Line 8 estimates the reward distribution based on the action-reward pairs observed.
The observations will be skewed towards those X_is for which $P(X_i = j)$ is large because these would be visible more frequently. To compensate for those actions for

which $P(X_i = j)$ is small, the remaining $T/2$ rounds are split to estimate the rewards for these infrequent actions.
We can treat the $m(\mathbf{q})$ as a function that outputs a parameter that favours low probability actions due to this expression $I_\tau = \{i : \min\{q_i, 1 - q_i\} < \frac{1}{\tau}\}$ yet doesn't favor these actions too much if given a choice to select fewer such actions as seen by - $m(\mathbf{q}) = \min\{\tau : |I_\tau \leq \tau\}$. In a sense this is a tradeoff between exploration and exploitation as done in line 11. This is reminsicent of the γ parameter in Exp3 and η in Exp4 because they serve a similar role in the trade-off.
Finally the algorithm creates a *subset* of such actions that fit the criteria of low enough probability (line 11) and evenly splits the remaining $T/2$ rounds to intervene on them (line 12). The reward distribution is updated for these selected actions only (line 17). Finally the action that yielded the best approximate reward is considered optimal (line 19).

Theorem: Algorithm 1 satisfies $R_T \in O(\sqrt{\frac{m(q)}{T}\log(\frac{NT}{m(q)})})$

Proof: This proof is from the supplementary material of [LLR16] and here shall be reproduced in parts that are important. The proof requires some lemmas. **Lemma 1:** Let $i \in \{1, ..., N\}$ and $\delta > 0$ then:

$$P\{|\hat{q}_i - q_i| \geq \sqrt{\frac{6q_i}{T}\log\frac{2}{\delta}}\} \leq \delta \tag{3}$$

Proof: $\hat{q}_i = \frac{2}{T}\Sigma_{t=1}^{T/2} X_{t,i}$ where $X_{t,i} \sim Bernoulli(q_i)$ so applying Chernoff's bound:

$$P\{|\hat{q}_i - q_i| \geq \epsilon\} \leq 2e^{-\frac{T\epsilon^2}{6q_i}}$$

Where we can solve for ϵ after setting $\delta = 2e^{\frac{-T\epsilon^2}{6q_i}}$
Lemma 2: Let $X_1, X_2, ...$ be a sequence of random variables with $X_i \in [0, 1]$ and $E[X_i] = p$ and $\delta \in [0, 1]$ then:

$$P\{\exists t \geq n_0 : |\frac{1}{t}\Sigma_{s=1}^{t} X_s - p| \geq \sqrt{\frac{2}{n_0}\log\frac{2}{\delta}}\} \leq 4\delta \tag{4}$$

Proof: Hoeffding's bound and union bound are useful:

$$\begin{aligned} P\{\exists t \geq n_0 : |\frac{1}{t}\Sigma_{s=1}^{t} X_s - p| \geq \sqrt{\frac{2}{n_0}\log\frac{2}{\delta}}\} &\leq \Sigma_{t=n_0}^{\infty} P\{|\frac{1}{t}\Sigma_{s=1}^{t} X_s - p| \geq \sqrt{\frac{2}{n_0}\log\frac{2}{\delta}}\} \\ &\leq 2\Sigma_{t=n_0}^{\infty} e^{\frac{-t}{n_0}\log\frac{2}{\delta}} \leq 4\delta \end{aligned} \tag{5}$$

Lemma 3 Let $\delta \in (0, 1)$ and assume $T \geq 48m\log\frac{2N}{\delta}$ Then

$$P\{2m(\mathbf{q})/3 \leq m(\hat{\mathbf{q}})\} \geq 1 - \delta \tag{6}$$

Proof: Let F be the event that there exists i such that $1 \leq i \leq N$ for which:

$$|\hat{q}_i - q_i| \leq \sqrt{\frac{6q_i}{T}\log\frac{2N}{\delta}} \tag{7}$$

By the Union Bound and Lemma 3 $P\{F\} \leq \delta$. When F does not hold we have $2m(\mathbf{q})/3 \leq m(\hat{\mathbf{q}}) \leq 2m(\mathbf{q})$. From the definition of $m(\mathbf{q})$ and our assumption on $\mathbf{q}$ for $i > m$ we have $q_i \geq q_m \geq 1/m$ and so by Lemma 3:

$$\begin{aligned} \frac{3}{4} &\geq \frac{1}{2} + \sqrt{\frac{3}{T}\log\frac{2N}{\delta}} \geq q_i + \sqrt{\frac{6q_i}{T}\log\frac{2N}{\delta}} \geq \hat{q}_i \\ &\geq q_i - \sqrt{\frac{6q_i}{T}\log\frac{2N}{\delta}} \geq q_i - \sqrt{\frac{q_i}{8m}} \geq \frac{1}{2m} \end{aligned} \tag{8}$$

By pigeonhole principle then $m(\hat{\mathbf{q}}) \leq 2m$. For the other direction, since failure event F doesn't hold, we have for $i \leq m$:

$$\hat{q}_i \leq q_i + \sqrt{\frac{6q_i}{T}\log\frac{2N}{\delta}} \leq \frac{1}{m}(1 + \sqrt{\frac{1}{8}}) \leq \frac{3}{2m} \tag{9}$$

Therefore, $m(\hat{\mathbf{q}}) \geq 2m(\mathbf{q})/3$
Proof of Theorem 1 Let $\delta = m = m(\mathbf{q})/N$. Then by lemma 6:

$$P\{2m/3 \leq m(\hat{\mathbf{q}}) \leq 2m\} \geq 1 - \delta \tag{10}$$

Recalling that $A = \{a \in \mathbb{A} : \hat{p}_a \leq 1/m(\hat{\mathbf{q}})\}$ Then for $a \in A$ the algorithm estimates μ_a from $T/(2m(\hat{\mathbf{q}})) \geq T/(4m)$ samples. Therefore by Hoeffding's inequality and union bound:

$$P\{\exists a \in A : |\mu_a - \hat{\mu}_a| \geq \sqrt{\frac{8m}{T} \log \frac{2N}{\delta}}\} \leq \delta \tag{11}$$

For arms not in a we have $\hat{p}_a \geq 1/m(\hat{\mathbf{q}}) \geq 1/(2m)$. Therefore if $a = do(X_i = j)$ then

$$\hat{p}_a = \frac{2}{T} \Sigma_{t=1}^{T/2} \mathbb{1}\{X_i = j\} \geq \frac{1}{2m} \tag{12}$$

$\Sigma_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = j\} \geq T/(4m)$ Therefore from lemma 4 we have:

$$P(\Sigma_{t=1}^{T/2} \mathbb{1}\{X_i = j\} \geq \frac{T}{4m} \wedge |\hat{\mu}_a - \mu_a| \geq \sqrt{\frac{8m}{T} \log \frac{2N}{\delta}})$$

So with probability at least $1 - 6\delta$ $|\hat{\mu}_a - \mu_a| \leq \sqrt{\frac{8m}{T} \log \frac{2N}{\delta}} = \epsilon$
If this occurs then $\mu_{a*_T} \geq \hat{\mu}_{a*_T} - \epsilon \geq \mu_{a*} - 2\epsilon$
Therefore $\mu* - E[\hat{\mu}_{\hat{a*}_T}] \leq 6\delta + \epsilon \leq \frac{6m}{T} + \sqrt{\frac{32m}{T} \log \frac{NT}{m}}$. QED

5.1 General graphs

The more general problem is where the graph is known, but arbitrary. In general, unlike in the parallel graph case it can't be claimed $P(Y|X_i = j) = P(Y|do(X_i = j))$ since causation is not correlation. However, if all the variables are observable, any causal distribution $P(X_1, .., X_N|do(X_i = j))$ can be expressed in terms of the observational distributions via the truncated factorization formula [Pea10]:

$$P(X_1, .., X_N|do(X_i = j)) = \Pi_{k \neq i} P(X_k|Pa_{X_k})\delta(X_i - j)$$

The naive way of applying parallel bandits is to apply this truncated factorization to write expression $P(Y|a)$ for each action a in terms of the observational quantities and playing those actions for which the observational estimates were poor. However this is not optimal because we ignored the information we could have learned about the reward for intervening on one variable from rounds in which we act on other variables. This loss of related information is reminiscent of the reasons for using expert advice in the contextual setting. Here, a simple example to illustrate this lost information is a causal chain where each child is deterministically determined by its parent's value. Then performing a single action $do(X_1 = 1)$ can inform us of the reward from all interventions $do(X_i = 1)$ for $i = 2, 3..K$. Also consider the graph in fig 1b where we plan to intervene $do(X_2 = 1)$ so the incoming arrows into X_2 get deleted and we can identify this causal effect from the expression $P(Y|do(X_2 = j)) = \Sigma_{X_1} P(X_1, X_2 = j, Y) = \Sigma_{X_1} P(X_1) P(Y|X_1, X_2 = j)) = P(X_1 = 0)P(Y|X_1 = 0, X_2 = j) + P(X_1 = 1)P(Y|X_1 = 1, X_2 = j)$. If we deterministically set $X_2 = X_1 = j = 1$ then clearly we won't observe $P(Y|X_1 = 0, X_2 = 1)$ and consequently the estimate for this intervention would also be poor. So, we need an estimator for each action that incorporates information obtained from every other action.
If we assume that conditional interventional distributions $P(Pa_Y|a)$ (but not $P(Y|a)$. Let η be the distribution on available interventions $a \in A$ so $\eta_a \geq 0$ and $\Sigma_a \eta_a = 1$. Define $Q = \Sigma_a \eta_a P(Pa_Y|a)$ to be the mixture distribution over the interventions with respect to η. Then the algorithm samples T actions from η and uses them to estimate the returns μ_a simultaneously via a truncated importance weighted estimator.

6 Discussion

The importance weight estimator in the general graphs causal bandit algorithm aims to use a subset of actions distributed as η that can be intervened on to estimate all of the actions $a \in A$. This is done by using $R_a(X) = \frac{P\{Pa_Y(X)|a\}}{Q\{Pa_Y(X)\}}$. This idea is similar to the IPW reward estimation method in Exp4 $\hat{r}_{t,j} = \frac{r_{t,j}}{p_{t,j}} \mathbb{I}(j = a_t)$. There is also a parallel to be drawn between the η in Exp3, γ parameter in Exp4 and the $m(q)$ parameter in Causal bandits setting - both are used to trade-off exploration and exploitation albeit in different settings. In conclusion, this survey has discussed three settings and 4 algorithms of the bandit problem: Exp3 in the Multi-Armed case, Exp4 in the contextual case, Parallel-bandits and General bandits in the Causal

case. At appropriate times in the discussion certain similarities were pointed out. Notably there is a gap of 14 years between the first two and latter two algorithms, and still similar ideas are continued. There are some other papers that look at the do-calculus approach to selecting which variables of the causal graph (including a semi-Markovian graph) to intervene on after finding certain sets of equivalent arms in the graph. Those approaches and other future work in the intersection of causality and bandits is promising and should be explored.

References

[RRZ94] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. "Estimation of regression coefficients when some regressors are not always observed". In: *Journal of the American statistical Association* 89.427 (1994), pp. 846–866.

[Aue+02] Peter Auer et al. "The nonstochastic multiarmed bandit problem". In: *SIAM journal on computing* 32.1 (2002), pp. 48–77.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Pea10] Judea Pearl. "An introduction to causal inference". In: *The international journal of biostatistics* 6.2 (2010).

[LLR16] Finnian Lattimore, Tor Lattimore, and Mark D Reid. "Causal bandits: Learning good interventions via causal inference". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1181–1189.

[LS18] Tor Lattimore and Csaba Szepesvári. "Bandit algorithms". In: *preprint* (2018), p. 28.