# Learning how to act: making good decisions with machine learning

Finnian Lattimore

December 23, 2015

## 1 Introduction

Learning the outcome of an action is central to many real world problems. Does de-worming children in poor countries improve health and educational outcomes [1, 2]? Would increasing the minimum wage lead to higher unemployment? Will offering this customer a discount improve my revenue? These questions are difficult as they require more than identifying a pattern in data. There are two, very different, approaches to this problem within the machine learning community: reinforcement learning and causal inference.

Reinforcement learning addresses the problem of learning from explicitly taking actions. There is typically some state or environment. An agent chooses an action from those available in the current state. The state then evolves stochastically as a function of the selected action and the agent receives some feedback or reward that is a function of the new state. This setting differs from the standard classification problem in that the agent must learn from feedback on the selected action, rather than being presented with the correct action for a given state. A common modelling assumption is that the state evolves only as a function of the previous state and the action chosen, and given these, is independent of the previous history of states and actions. This is known as a Markov decision process or MDP. A particularly well studied and understood model is the singe state MDP. In this case, there are a set of actions, each associated with a fixed but unknown reward distribution and at each time step our agent selects an action and receives corresponding feedback. This is known as the multi-armed bandit problem.

Causal inference makes use of assumptions to allow the outcome of actions to be predicted from observational data. The key to causal inference is a framework that can model how actions change the state of the world. This framework then allows us to map information collected in one setting to another.

Both approaches can be seen as extensions to the concept of randomised controlled trials. Bandit algorithms deal with the sequential nature of the decision making process, causal inference with the problem that full randomisation is not always feasible, affordable or ethical. The similarities between the problems that these techniques have been developed to address raises the question of if there are problems best addressed by a combination of these approaches and how they can be combined. The goal of my thesis is to explore these questions. In the next sections I review the key literature in causal inference and bandits. I then present a general approach to how causal models might be incorporated into bandit settings and conclude by demonstrating an algorithm that leverages causal assumptions to improve performance in a specific bandit setting.

# 2 Causal Inference

## 2.1 Models of causality and intervention

Predicting the outcome of actions without explicitly taking them requires assumptions about how the actions will change the system. A very powerful and general model that underlies much of the recent work in causality is the causal bayesian network. A causal bayesian network, or directed acyclic graph (DAG), is a bayesian network in which a link $V_i \rightarrow V_j$ is defined to mean $V_i$ directly causes $V_j$. This means that if we intervene and change the value of $V_i$, we expect $V_j$ to change, but if we intervene to change $V_j$, $V_i$ will not change. More generally, if $G$ is a causal network for a distribution $P$ defined over variables $V_1 ... V_N$, then the distribution after an intervention where we set $X \subset V$ to $x$, denoted $do(X = x)$ is obtained by simply dropping the terms for each of the variables in $X$ from the factorization given by the network. This is referred to as the truncated product formula [3].

If the network contains latent (unobservable) variables we will not be able to calculate all the terms in the truncated product formula. However, it may still be possible to determine the post-interventional distribution of specific variables of interest. A general causal query $P(Y|do(X = x))$ is identifiable if it can be shown to be equivalent to an expression containing only pre-interventional quantities. This means that, asymptotically, we can obtain an unbiased estimate for the distribution after an intervention based on purely observational data.

The do calculus is a set of three rules that allow transformations of interventional terms to non-interventional ones, given a causal graph [3]. They are derived directly from d-seperation properties of graphical models and the definition of intervention in causal DAGs. These rules are complete. A query is identifiable if and only if it can be transformed to contain only non-interventional terms via the do-calculus [4, 5]. There is an equivalent algorithm that can take any causal graph and query and determine identifiability [6] (see `http://finnhacks42.github.io/causal_identify` for a javascript demonstration of this algorithm). If a query is not identifiable, it may still be possible to get bounds for causal effects, for example using instrumental variables [7] or by making additional assumptions.

There are two other key frameworks that arise in causal inference. Counterfactuals and structural equation models. Counterfactuals are statements about imagined or alternate realities, are prevalent in everyday language and may play a role in the development of causal reasoning in humans [8]. Causal effects are differences in counterfactual variables; what is the difference between what would happen if we did one thing versus what would happen if we did something else [9, 10, 11, 12, 13].

For example, if we wanted to estimate the causal effect of a medical treatment, then we might let $Y^1$ be a counterfactual random variable representing the (binary) potential outcome if treated. The distribution of $Y^1$ is the distribution we would see in the outcome $Y$ if everyone was treated. Similarly $Y^0$ represents the potential outcome for the placebo. The causal effect of the drug is the difference between the probability of recovery, across the population, if everyone was treated, and the probability of recovery given placebo $P(Y^1) - P(Y^0)$. This quantity can be estimated from observational data if we assume $X \perp\!\!\!\perp Y^0$ and $X \perp\!\!\!\perp Y^1$. These assumptions are referred to as ignoreability assumptions [11]. They state that the treatment each person receives is independent of whether they would recover if treated and if they would recover if not treated. Graphically, this is equivalent to the assumption that there is no variable that is a parent of both the treatment $X$ and the outcome $Y$.

Structural equation models (SEMs) describe a deterministic world, where underlying mechanisms determine the output of any process for a given input. The mechanism (but not the

output) is assumed to be independent of what is fed into it. Linear structural equation models have a long history for causal estimation [14, 15]. Mathematically, each variable is a deterministic function of its direct causes and a noise term that captures unmeasured variables. The noise terms are required to be mutually independent. If there is the possibility that an unmeasured variable influences more than one variable of interest in a study, it must be modelled explicitly as a latent (unobserved) variable. Structural equation models can be represented visually as a network. Each variable is a node and arrows are drawn from causes to their effects. If the network for a structural equation model is acyclic then it implies a recursive factorization of the joint distribution over its variables. In other words, it is a causal bayesian network.

Remarkably for models developed relatively independently in fields with very different approaches and problems, the models we have discussed are functionally very similar. To determine if and how an interventional query can be non-parametrically identified, it is equivalent to specify assumptions graphically in terms of bayesian networks or as structural equation models or as conditional independence statements involving counterfactual variables (ignorability assumptions).

It is possible to pose causal queries in terms of counterfactuals that are not interventional and cannot be phrased in terms of the do-notation. The scientific and philosophical validity of such counterfactual queries remains under question [16, 17], however they are nonetheless widely posed in the form of attribution of causal effects to different pathways and mediation [18, 19, 20].

There are differences between the models we have considered when it comes to these non-interventional queries. Counterfactuals are not defined in causal bayesian networks, as they only encode information on the interventional distribution over variables. Counterfactuals can be defined in terms of structural equation models [3] but there are subtle differences depending on the form of assumptions made. Structural equation models with independent errors allow the identification of quantities in mediation studies, which are not identifiable with the weak ignorability assumptions and cannot be tested experimentally [21].

In practice, differences in focus and approach across different fields eclipse the actual differences in the models. The work on causal graphical models [3, 22] focuses on non-parametric estimation in the population limit and rigorous theoretical foundations. The Neyman-Rubin framework builds on our understanding of randomized experiment and generalizes to quasi-exeperimental and observational settings, with a particular focus on non-random assignment to treatment. This research emphasises estimating average causal effects and provides practical methods for estimation, in particular, propensity scores; a method to control for multiple variables in high dimensional settings with finite data [11]. In economics, inferring causal effects from non-experimental data so as to support policy decisions is central to the field. Economists are often interested in broader measures of the distribution of causal effects than the mean and make extensive use of structural equation models, generally with strong parametric assumptions [23]. In addition, the parametric structural equation models favoured in economics can be extended to analyse cyclic (otherwise referred to as non-recursive) models.

## 2.2 Discovering causal structure

In the previous section we discussed when assumptions about the structure of the variables in a specific problem is sufficient to identify a causal effect. This approach relies on having enough prior knowledge or theory about the problem to allow you to, at least partially, specify the causal network. In this section, we consider the much harder problem of causal inference where you need to learn the network. Causal inference might seem impossible without specific

assumptions about the structure of the variables involved but, amazingly, some aspects of causal structure can be determined from much more general assumptions.

### 2.2.1 Discovery based on conditional independence

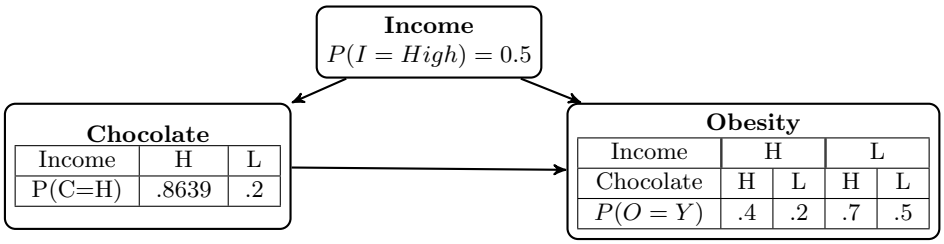
One general approach is to look for clues about the structure of the network in the conditional independence relations in the distribution. Assume there is some acyclic causal network $G$ that generated the distribution $P(\boldsymbol{V})$ from which our data has been sampled. Our goal is to recover the network from this data.

Since $G$ is a bayesian network, if $Z$ d-separates $X$ and $Y$ in $G$ then $(X \perp\!\!\!\perp Y|Z)$ in $P$ . However, we want to work in the other direction, from conditional independence in the distribution to the structure of the network. This requires that we assume the reverse condition: $(X \perp\!\!\!\perp Y|Z)$ in $P$ must imply $Z$ d-separates $X$ and $Y$ in $G$. This assumption, commonly referred to as ***faithfulness***, says there are no additional independence relations that are satisfied in $P$ but not in all distributions $\boldsymbol{P'}$ that are compatible with $G$. Stating that $P$ is faithful to $G$ is equivalent to $G$ is a ***perfect map*** for $P$.

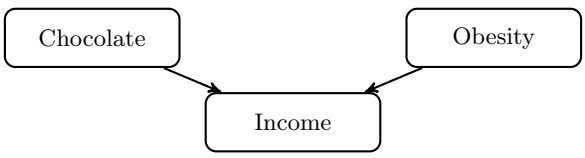
Faithfulness is an assumption. It does not always hold and we cannot verify it from the observational data we wish to use for causal inference. However, most distributions generated by a causal bayesian network will be faithful to that network. For faithfulness to be violated, different causal effects must exactly balance one-another out. For example, consider a simple binary variable model of chocolate consumption, income and obesity, figure 1. If the coefficients in the conditional probability tables are just right then the direct effect of chocolate on obesity will exactly balance the indirect effect through income and obesity will appear independent of chocolate consumption. However, this independence is not stable. It would disappear under a small perturbation to any of the parameters.

Figure 1: Example of a data generated from a causal graph violating the faithfulness assumption

(a) True causal graph generating $P$




(b) Perfect map for $P$, $(C \perp\!\!\!\perp O)$




Given the faithfulness assumption, our causal discovery problem reduces to finding the set of bayesian networks that have exactly the dependency structure as we observe in $P$. A wide range of algorithms have been developed based on this key observation, see table 1. Constraint based methods such as the PC algorithm [22], FCI algorithm [22] and RFCI algorithm [24], perform sequential conditional independence tests and eliminate inconsistent graphs. Search and score based methods, such as GES [25], search over the space of graphs and score them according to how well they fit the independences given a complexity penalising prior. Constraint based

methods are faster, particularly for sparse graphs, but can lack robustness as errors in early conditional independence tests can propagate. Search and score based methods are more robust for small samples sizes but difficult to scale to larger graphs. This has led to the development of hybrid approaches, such as the MMHC algorithm [26]. A key component of causal discovery is the ability to do high dimensional non-parametric conditional independence tests. Developments in kernalized conditional independence tests,[27, 28] have made this possible.

Table 1: A comparison of key causal discovery algorithms

| **Alg.** | **Method** | **Scales (num.vars)** | **$\sim$ Vars** | **Latent** |
|---|---|---|---|---|
| PC | Constraint based | Worst case exponential, polynomial for sparse graphs | 5000 | No |
| FCI | Constraint based | Worst case exponential, polynomial variant FCI+ for sparse graphs | 30 | Yes |
| RFCI | Constraint based | ? | 500 | Yes |
| GES | Search & Score | Worst case exponential | 50 | No |
| MMHC | Hybrid | ? | 5000 | No |

If the end goal of causal discovery is to estimate causal effects, then it may not be necessary to learn the entire graph, only the subset of the graph surrounding target variables of interest. Such local causal discovery techniques can be scaled to problems with many more variables [29]. Once a set of causal graphs has been identified, causal effects of interest can be bounded by combining the results for the all the networks. This procedure is the IDA algorithm [30] and has been found to outperform standard regularization techniques at finding causal effects in a high-dimensional yeast gene expression data set [31]. An implementation is available in the R package pcalg [32]

### 2.2.2 Discovery with functional models

All of the algorithms we have considered so far return a Markov equivalence class. They cannot distinguish between two models that result in the same set of conditional independence relations. Consider the very simple case where we have only two variables and the only possible causal structures are $X \to Y$ or $Y \to X$. These models have the same dependency structure but in one case $P(Y|do(X)) = P(Y|X)$ and in the other $P(Y|do(X)) = P(Y)$. No algorithm relying purely on conditional independence relations can separate these two cases.

One solution is to utilize structural equation models to specify additional assumptions. For example, if we assume that noise is additive, such that $X \to Y \implies Y = f(X) + \epsilon$, then this will only be invertible such that $X = g(Y) + \epsilon'$ for specific pairs of functions $f$ and noise distributions $\epsilon$. Thus in general we will be able to identify the causal direction [33]. This can be extended to post-non-linear additive noise, $Y = h(f(X) + \epsilon)$, [34]. These techniques can also be applied over more than two variables [35].

A more general approach is to leverage the assumption that the functions are independent of inputs [36]. This leads to the idea that $P(X)$ and $P(Y|X)$ are independent if $X \to Y$ but not if $Y \to X$. [37] propose testing for this by applying both semi-supervised and standard techniques

to estimate $P(Y|X)$. Semi-supervised methods, which utilize additional points from $P(X)$ to learn $P(Y|X)$ should only be able to outperform standard methods if $Y \to X$.

Finally, rather than explicitly developing an algorithm based on a specific asymmetry between cause and effect, [38] propose learning what causality looks like from data. They take as input a dataset where each row of data is itself a dataset in which either $X \to Y$ or $Y \to X$ and a corresponding label. Estimates of the distributions $P(X)$, $P(Y)$ and $P(X, Y)$ for each row are then mapped to features in some kernel space via mean kernel embeddings and finally a standard classification algorithm can be trained to learn the labels. New datasets, where the direction of causality is unknown, are then simply mapped to the kernel space and the causal direction is classified according to the trained classifier. In practice, the classifier is trained mostly on simulated data as it is difficult to find a sufficient set of causal problems with only two variables, where the direction of causality is known.

# 3 Multi-armed Bandits

In its classic formulation [39] the (stochastic) K-armed bandit describes a sequential decision making problem, with $K$ possible actions or arms. Each arm $i$ is associated with a fixed but unknown reward distribution $\nu_i$ [1]. For each timestep $t$ upto a horizon $T$ the learner selects an action $I_t \in \{1...K\}$ and receives a reward, $g_{I_t,t}$, sampled i.i.d from $\nu_i$. The goal of the learner is to maximize the total reward they receive. This problem introduces the fundamental exploration-exploitation trade-off. The learner must balance playing arms that have yielded good results previously with exploring arms about which they are uncertain.

The performance of bandit algorithms is generally described by the (pseudo) regret, $R(T)$. This is the difference between the expected reward obtained by the algorithm and the expected reward of selecting the best action in every timestep.

$$R(T) = max_{\{i=1...K\}}\mathbb{E}\left[\sum_{t=1}^{T} g_{i,t}\right] - \mathbb{E}\left[\sum_{t=1}^{T} g_{I_t,t}\right] \tag{1}$$

If we let $\mu_i = \mathbb{E}[\nu_i]$ denote the expected reward for each arm $i$ and $\mu^* = max_{\{i=1...K\}}(\mu_i)$ denote the reward for the best arm:

$$R(T) = T\mu^* - \sum_{t=1}^{T} E[\mu_{I_t}] \tag{2}$$

An algorithm is learning if it obtains regret that is sublinear in $T$. The lower bound on the worst case regret for any algorithm (stochastic or adversarial) for the K-armed bandit problem is $\Omega\left(\sqrt{TK}\right)$ [40].

A key algorithm for stochastic bandits, with tractable analysis and strong performance guarantees, is the UCB algorithm [41]. They key to this algorithm is that it keeps track of an upper confidence bound (hence UCB) on the expected reward for each arm and selects the arm with the highest one. This balances exploration and exploitation as an arm with a high upper

[1] In order to obtain regret bounds, some assumptions are required on the distributions $\nu_i$. It sufficient to assume they are sub-gaussian

confidence bound must have either a high expected reward or large uncertainty on the expected reward. Assume for notational simplicity that $\mu_1 > \mu_2 > ... > \mu_K$, such that $\mu^* = \mu_1$, and let $\Delta_i = \mu_i - \mu^*$ be the sub-optimality for each arm. The (problem dependent) regret for UCB is bounded by:

$$R^{ucb}(T) \in \mathcal{O}\left(\sum_{i=2}^{K} \frac{1}{\Delta_i} \log(T)\right) \tag{3}$$

This bound blows up as differences $\Delta_i \to 0$, however the regret itself does not - since although we may not be able to distinguish arms with very small $\Delta_i$ from the optimal arm, we also do not lose much by selecting them. In the worst case, $R^{ucb}(T) = \mathcal{O}\left(\sqrt{TK\log(T)}\right)$ [42]. Subtle modifications to the UCB algorithm can eliminate the logarithmic term in this worst case regret bound. This yields $R^{ucb}(T) = \mathcal{O}\left(\sqrt{TK}\right)$ and closes the gap with the worst case lower bound [43, 44], whilst retaining a good problem dependent bound of the form achieved by UCB [44].

Adversarial bandits are an alternate, widely studied, setting that relaxes the assumption that rewards are generated stochastically. Instead, simultaneously with the learner selecting an action $I_t$, a potentially malicious adversary selects the reward vector $\boldsymbol{g}_t$. As in the stochastic setting, the learner then receives reward $g_{I_t,t}$. The seminal algorithm for adversarial bandits is Exp-3, which, like UCB, obtains regret $\mathcal{O}\left(\sqrt{TK\log(T)}\right)$ regret [40]. Optimal algorithms, with $R(T) = \mathcal{O}\left(\sqrt{TK}\right)$, have also been demonstrated for the adversarial setting [43].

Another problem that has attracted a lot of recent attention [45, 46, 47, 48] within the multi-armed bandit framework is *pure exploration* or *best arm identification*. In this setting, the horizon $T$ represents a fixed budget for exploration after which the algorithm outputs a single best arm $i$. The performance of the algorithm is measured by the simple regret; the expected difference between the mean reward of the (truly) optimal arm and the mean reward of the arm selected by the algorithm, $R_s(T) = \mu^* - \mathbb{E}\left[\mu_i\right]$ . This problem arises naturally in applications where there is a testing or evaluation phase, during which regret is not incurred, followed by a commercialization or exploitation phase. For example, many strategies might be assessed via simulation prior to one being selected and deployed. The simple regret for a K-armed bandit is lower bounded by $\mathcal{O}\left(\sqrt{K/T}\right)$ [45].

The classic multi-armed bandit is a powerful tool for sequential decision making. However, the regret grows linearly with the number of (sub-optimal) actions and many real world problems have large or even infinite action spaces. This has led to the development of a wide range of models that assume some structure across the reward distributions for different arms, for example generalized linear bandits [49], dependent bandits [50], X-armed bandits [51] and gaussian process bandits [52], or that consider more complex feedback, for example the recent work on graph feedback [53, 54, 55, 56, 57, 58] and partial monitoring [59, 60].

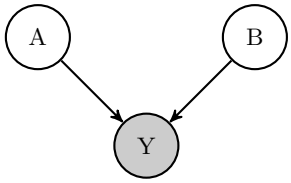
# 4 Unifying the frameworks

A natural way to connect the causal framework with the bandit setting is to model the problem as a causal directed acyclic graph. Each possible assignment of variables to values is an action (bandit arm). See figure 2 for a simple example. The reward could be a general function of the action selected and the final state of the graph. However for simplicity, we will consider the reward to be the value of a single specified node minus the cost of the selected action.

The number of actions or arms grows exponentially with the number of variables in the graph, making it important to use algorithms that take account of the graph structure to reduce the search space.

Modelling a problem as a causal graph only makes sense when rewards are generated stochastically - since causal graphs fundamentally model probability distributions over variables. Thus the connection is to stochastic bandit problems (although adversarial bandits algorithms may be applied to stochastic problems).

Figure 2: A simple causal graphical model and corresponding complete action space. A and B represent binary variables that can be intervened on and Y represents the reward.

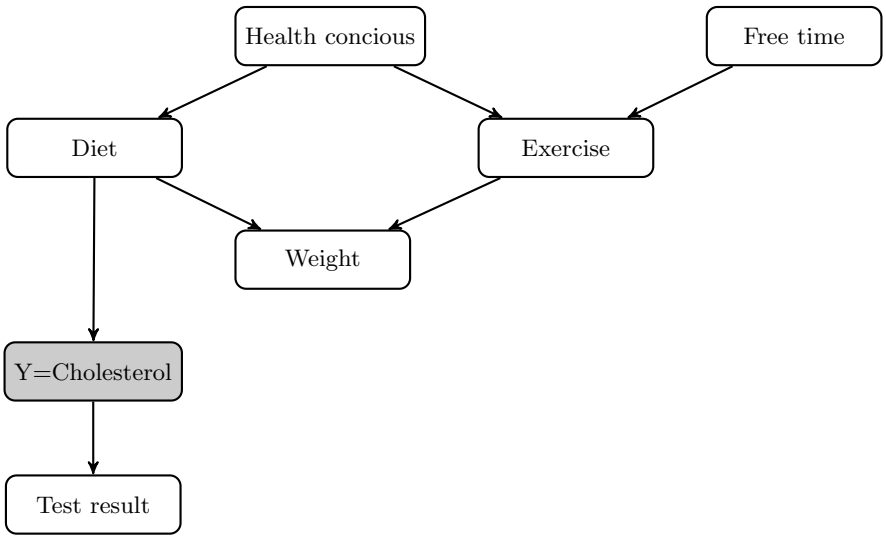



Actions = {do(A=0,B=0), do(A=0,B=1), do(A=1,B=0), do(A=1,B=1), do(A=0), do(A=1), do(B=0), do(B=1), do()}

If we begin by considering the case where the causal graph is known, problems then take on characteristics of different bandit settings depending on the assumptions we make about what subset of actions can be taken, what variables are observable and whether they are observed before or after an action is selected.

If feedback is received only on the reward node then the do-calculus can be applied to eliminate some actions immediately, before any experiments are performed and then a standard bandit algorithm can be run on the remaining actions. See figure 3 as an example.

Figure 3: Example causal graph (based on [61]) where the outcome of interest (reward) is cholesterol level . The do-calculus can be applied to eliminate some actions immediately without the need to do any experiments. For example, no actions involving 'Test Result' need to be considered and interventions on 'Diet' do not need to be considered in conjunction with any other variables.




If we receive feedback on additional nodes, the problem can be more interesting. In addition to

being able to eliminate some actions prior to sampling any data as in the previous case, taking one action may give us some information on actions that were not selected. Consider again the model in figure 2. The causal structure implies:

$$P(Y|do(A=0)) = P(Y|do(), A=0) \tag{4}$$
$$= P(Y|do(B=0), A=0)P(B=0) + P(Y|do(B=1), A=0)P(B=1) \tag{5}$$

Thus we gain information about the reward for the action $do(A=0)$ from selecting the action $do()$ or $do(B=b)$ and then observing $A=0$.

We only get this form of side information for actions that don't specify the value of every variable, ie those in the bottom half of the table in figure 2. Since the reward distribution for actions that set a subset of the variables is the result of marginalizing out other variables, they can only be optimal if they have lower cost. So if the cost of all actions is constant (no matter how many variables must be set), then the problem has the same characteristics as if only the reward node were observable.

If the information on the value of additional nodes is available prior to selecting an action the problem resembles a contextual bandit. For example if we observe $A=0$ then, in deciding between the actions $do(B=0)$ and $do(B=1)$, we would want information on $P(Y|A=0, B=0)$ and $P(Y|A=0, B=1)$. Note, side information can still arise if we learn the value of some variables prior to selecting an action and some afterwards.

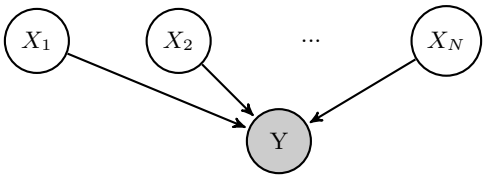
### 4.1 Incorporating side information induced by casual graph

As a first step towards more general problems, we develop and evaluate a bandit algorithm that incorporates the side information induced by a specific causal graph.

Assume we have a known causal model with binary variables $\boldsymbol{X} = \{X_1..X_N\}$ that independently cause a target variable of interest $Y$, figure 4. This is an expansion of the model shown in figure 2. We can run sequential experiments on the system, where at each timestep $t$ we can either do nothing, $do()$, or select a single variable, $X_i$, on which to intervene, $do(X_{t,i} = J_t)$, and subsequently observe the complete result, $(\boldsymbol{X}_t, Y_t)$.

As an example, consider a farmer wishing to optimize the yield of her crop. She can invest in a green house to control temperature, a watering system to control soil moisture, fertilizers to set soil nutrients, etc. We assume only a single intervention is feasible due to cost and that each of these variables are independent of one-another (this may not always be the case - temperature could be related to rainfall for example). After having selected which variable to control, she plants her crops and observes the values of the remaining input variables and the yield. This repeats across many growing seasons, and the goal is to maximize the total cumulative yield.

Figure 4: Assumed Causal Structure




Let $\boldsymbol{q} \in [0,1]^N$ be a fixed vector where $q_i = P(X_i = 1)$. In each time-step $t$ upto a known horizon $T$:

1. The learner chooses an $I_t \in \{0, \ldots, N\}$ and $J_t \in \{0, 1\}$, where $I_t = 0$ represents $do()$.
2. Variables not set by the learner take their values probabilistically according to $\boldsymbol{q}$

$$X_{t,i} = \begin{cases} \sim \text{Bernoulli}(q_i) & \text{if } i \neq I_t \\ J_t & \text{otherwise}. \end{cases}$$

3. The learner observes $(\boldsymbol{X}_t, Y_t)$ and receives reward $Y_t \sim \text{Bernoulli}(r(\boldsymbol{X}_t))$ where $r : \{0, 1\}^N \to [0, 1]$ is unknown and arbitrary.

The problem could be treated as a standard bandit problem with $K = 2N + 1$ arms,

$$\{do(), do(X_1 = 0), do(X_1 = 1) \ \ldots \ do(X_N = 1)\}$$

yielding a regret $\mathcal{O}\left(\sqrt{TN}\right)$. However, this doesn't leverage the side information induced by the causal structure:

$$\begin{aligned} P(Y_t|do(X_{i,t} = j)) &= P(Y_t|do(), X_{i,t} = j) && (6) \\ &= P(Y_t|do(X_{l,t} = 0), X_{i,t} = j)(1 - q_l) + P(Y_t|do(X_{l,t} = 1), X_{i,t} = j)q_l && (7) \end{aligned}$$

The hardness of this problem depends strongly on the vector $\boldsymbol{q}$. This is intuitive from a consideration of the extreme cases. If $\boldsymbol{q} = \mathbf{0}$ (or $\mathbf{1}$), then half of the actions we wish to explore will only occur if we explicitly select them. For these actions we get no side information by playing other actions. So at best we could expect a factor of two improvement over simply running a standard bandit algorithm. If $\boldsymbol{q} = \frac{\mathbf{1}}{\mathbf{2}}$, the observe action, $do()$, effectively becomes a revealing action and we would anticipate obtaining a regret with minimal dependence on the number of variables $N$.

We consider a simple explore-exploit based algorithm, that will explore for $h$ timesteps, sampling actions in a way that depends on $\boldsymbol{q}$. We then select the arm with the highest expected reward for the remaining $T - h$ time steps. During the exploration phase, we will balance purely observing, $I_t = 0$, which takes advantage of equation 6 to reveal the reward for frequently occurring arms, with explicit sampling of infrequently occurring arms.

Without loss of generality, we can assume $q_i \in [0, \frac{1}{2}]$ and $q_1 \leq q_2 ... \leq q_N$. Let:

$$m \in [2, N] = \left\{ m : q_m > \frac{1}{m} \right\} \tag{8}$$

$m$ is a measure of the number of infrequently occurring arms. If the problem is completely balanced $q_1 ... q_N = \frac{1}{2}$ then $m = 2$. If the problem is completely unbalanced, $q_1 ... q_N = 0$ then $m = N$. Let $q_{ij} = \mathrm{P}\{X_i = j\}$ and $\mu_{ij} = P(Y|do(X_i = j))$.

Suppose we select $I_t = 0$ for the first $h/2$ timesteps. This is at worst half the optimal. We use these observations to estimate the rewards for all the frequently occurring arms, $\{do(X_i = 0) \ \forall i\}$, and $\{X_i = 1 : i \geq m\}$

$$\hat{\mu}_{ij} = \frac{\sum_{t=1}^{h/2} \mathbb{1}\{Y_t = 1, X_{t,i} = j\}}{\frac{h}{2} q_{ij}} \tag{9}$$

We then explicitly play each of the $m$ infrequently occurring arms $\frac{h}{2m}$ times and estimate their rewards as:

$$\hat{\mu}_{ij} = \frac{2m}{h} \sum_{t=1}^{h/2m} \mathbb{1}\{Y_t = 1 | X_i = j\} \tag{10}$$

We then simply play the arm with the highest estimated reward for the remaining $T - h$ timesteps. We can show that we can select an $h$ (based on $T$, $N$ and $m$) that leads to worst case regret $R_T = \tilde{\mathcal{O}}\left(T^{2/3} m^{1/3}\right)$. This result can still be obtained in the case where $\boldsymbol{q}$ is unknown and must be learned during the exploration time. Comparing this to the standard bandit regret, $R_T = \mathcal{O}\left(\sqrt{TK}\right)$, we expect the causal algorithm to do better if $m < \frac{K^{3/2}}{\sqrt{T}}$. We can also show that the simple regret $R_s \in \tilde{\mathcal{O}}\left(\sqrt{m/T}\right)$. If $m << K$, this represents a significant improvement over standard bandit setting. [2]

In this algorithm, we have only used the side information provided by the $do()$ action about other actions. Since the $do()$ action fully reveals the value of alternate actions we could have incorporated this information via the graph feedback model [53], where at each timestep the feedback graph $G_t$ is selected stochastically, dependent on $\boldsymbol{q}$, and revealed after an action has been chosen. The feedback graph is distinct from the causal graph. A link $A \to B$ in $G_t$ indicates that selecting the action $A$ reveals the reward for action $B$. For this specific problem, $G_t$ will always be a star graph with the action $do()$ connected to half the remaining actions. The Exp3-IX algorithm [57] was developed for the adversarial version of this problem and has regret $\mathcal{O}\left(\sqrt{\bar{\alpha}T}\right)$, where $\bar{\alpha}$ is the average independence number of $G_t$. In our case $\bar{\alpha} = \frac{N}{2}$ so we again obtain the regret of the standard bandit algorithm. The issue here is that a malicious adversary can select the same graph each time, such that the rewards for half the arms are never revealed by the informative action. This is equivalent to a, nominally, stochastic selection of feedback graph where $\boldsymbol{q} = \mathbf{0}$

[54] consider a stochastic version of the graph feedback problem, but with a fixed graph available to the algorithm before it must select an action. In addition, their algorithm is not optimal for all graph structures and fails, in particular, to provide improvements for star like graphs as in our case. [56] improve the dependence of the algorithm on the graph structure but still assume the graph is fixed and available to the algorithm before the action is selected.

More generally, assuming causal structure creates more complex types of side information, such as that shown in equation 7. In this case, selecting one action does not fully reveal an alternate action but provides some information towards an estimate. The quality of the estimate notably depends not only on the number of times that action was selected. For example, to get a good estimate for $X_1 = 1$ by intervening on $X_2$ requires us to sample both $X_2 = 0$ and $X_2 = 1$, in proportions dependent on $q_2$. This more complex side information does not fit within the graph feedback framework.

Partial monitoring is a very general framework for for decoupling the feedback from the action and reward. It can be used to classify problems into one of four categories, trivial with no regret, easy with $R_T = \tilde{\Theta}\left(\sqrt{T}\right)$, hard with $R_T = \Theta\left(T^{2/3}\right)$ and hopeless with $R_T = \Omega\left(T\right)$ [60]. Partial monitoring algorithms yield results that are optimal with respect the the horizon $T$ but not other parameters, such as $K$, which is the key focus of incorporating causal structure.

[2] $\tilde{\mathcal{O}}\left(.\right)$ and $\tilde{\Theta}\left(.\right)$ suppress polylogarithmic factors

### 4.2 Open questions

There are many open questions still to be addressed. For the specific setting described above I need to derive lower bounds for the worst case regret so as to determine how close the current approach is to optimal.

It is clearly desirable to obtain an algorithm that could be generalized to an arbitrary causal graph. This will likely require the development of a bandit algorithm that can leverage more complex forms of side information like that in equation 7, in addition to an efficient algorithm for eliminating actions that need not be considered at all.

Another open class of problem to consider are cases where the causal structure is not (fully) known and also needs to be learnt within the online setting.

## 5 Paid work statement

I am not currently undertaking or planning to undertake any paid work.

## References


[1] Edward Miguel and Michael Kremer. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, pages 159–217, 2004.

[2] Calum Davey, Alexander M Aiken, Richard J Hayes, and James R Hargreaves. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International journal of epidemiology*, page dyv128, 2015.

[3] Judea Pearl. *Causality: models, reasoning and inference*. MIT Press, Cambridge, 2000.

[4] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In Thomas S. Richardson and R Dechter, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2006.

[5] Yimin Huang and Marco Valtorta. Pearl's Calculus of Intervention Is Complete. In Thomas S. Richardson and R Dechter, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2006.

[6] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *The Journal of Machine Learning Research*, 9:1941–1979, 2008.

[7] Joshua Angrist and Jorn-Stephan Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2009.

[8] Deena S Weisberg and Alison Gopnik. Pretense, counterfactuals, and Bayesian causal models: why what is not real really matters. *Cognitive science*, 37(7):1368–81, 2013.

[9] DB Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 1974.

[10] DB Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 1978.

[11] PR Rosenbaum and DB Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[12] DB Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, mar 2005.

[13] DB Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, sep 2008.

[14] S Wright. Correlation and causation. *Journal of agricultural research*, 1921.

[15] T Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 11(1):1–12, 1943.

[16] AP Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 2000.

[17] AP Dawid. Statistical Causality from a Decision-Theoretic Perspective. *arXiv preprint arXiv:1405.2292*, 2014.

[18] Judea Pearl. Interpretation and Identification of Causal Mediation. *Psychological methods*, jun 2014.

[19] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1):51–71, feb 2010.

[20] Tyler J VanderWeele and Sonia Hernández-Diaz. Is there a direct effect of pre-eclampsia on cerebral palsy not through preterm birth? *Paediatric and perinatal epidemiology*, 25(2):111–5, mar 2011.

[21] Thomas S Richardson and James M Robins. Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(128), 2013.

[22] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[23] James Heckman. Econometric causality. *International Statistical Review*, 2008.

[24] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, feb 2012.

[25] David Maxwell Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554, 2002.

[26] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, mar 2006.

[27] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592, 2007.

[28] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv: ...*, 2012.

[29] Constantin F Aliferis and Ioannis Tsamardinos. Algorithms for large-scale local causal discovery and feature selection in the presence of limited sample or large causal neighbourhoods. Technical Report October, Technical report, Technical report DSL-02-08, Department of Biomedical Informatics, Vanderbilt University, 2002.

[30] Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, dec 2009.

[31] Marloes H. Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010.

[32] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, VV(Ii), 2012.

[33] Patrick Hoyer, Dominik Janzing, and Joris Mooij. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, 2009.

[34] Kun Zhang and A Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. *NIPS 2008 Workshop on Causality. URL http://www ...*, 2008.

[35] Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[36] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniusis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, may 2012.

[37] Dominik Janzing and Jonas Peters. On causal and anticausal learning. In *International Conference on Machine Learning*, 2012.

[38] David Lopez-Paz, Krikamol Muandet, and Benjamin Recht. The Randomized Causation Coefficient. *arXiv preprint arXiv:1409.4366*, sep 2014.

[39] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–536, 1952.

[40] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, 1995.

[41] P Auer, N Cesa-bianchi, and P Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[42] Sébastien Bubeck. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[43] JY Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd annual Conference On Learning Theory*, pages 773—-818, 2009.

[44] Tor Lattimore. Optimally Confident UCB : Improved Regret for Finite-Armed Bandits. (1):1–16, 2015.

[45] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pages 23–37. Springer, 2009.

[46] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13—-p, 2010.

[47] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

[48] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1238–1246, 2013.

[49] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[50] Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal. Multi-armed bandit problems with dependent arms. *Proceedings of the 24th international conference on Machine learning - ICML '07*, pages 721–728, 2007.

[51] Sébastien Bubeck, R Munos, Gilles Stoltz, and C Szepesvári. X-armed bandits. *Multi-Armed Bandits*, pages 1–38, 2010.

[52] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[53] Shie Mannor and Ohad Shamir. From Bandits to Experts: On the Value of Side-Observations. pages 1–9, 2011.

[54] Marc Lelarge and Inria Ens. Leveraging Side Observations in Stochastic Bandits. *Uai*, 2012.

[55] Noga Alon and Nicolò Cesa-Bianchi. From Bandits to Experts: A Tale of Domination and Independence. *arXiv preprint arXiv: ...*, pages 1–22, 2013.

[56] Swapna Buccapatnam, Atilla Eryilmaz, and B Shroff Ness. Stochastic Bandits with Side Observations on Networks. *ACM SIGMETRICS'14, June 2014, Austin, Texas.*

[57] Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. *Neural Information Processing Systems*, pages 1–9, 2014.

[58] Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online Learning with Feedback Graphs : Beyond Bandits. *Colt*, pages 1–26, 2015.

[59] Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *COLT/EuroCOLT*, pages 208–223. Springer, 2001.

[60] Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring-classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

[61] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT Press, 2009.