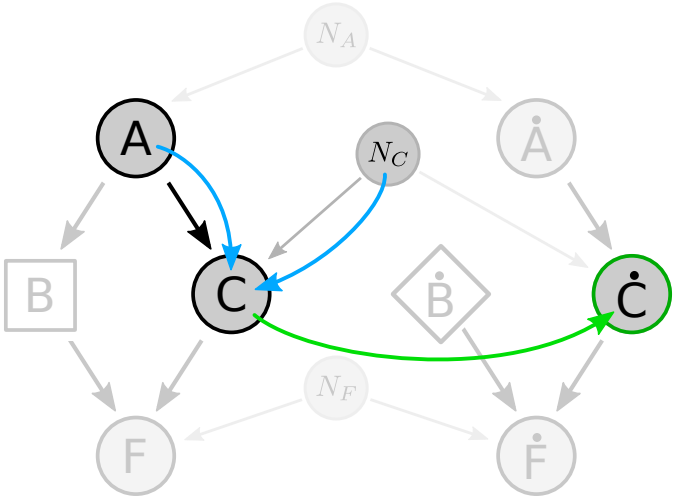
Technical University of Denmark

DTU

Introduction to

Causal Inference



Jeppe Nørregaard

Lars Kai Hansen

Preface

This document was made as part of the course 02463 Active Machine Learning and Agency at the Technical University of Denmark, but works independently the rest of the course. It can be used by anyone who would like an introduction to causality and it has very few prerequisites.

While most of the document only requires a basic understanding of mathematics and probability (percentages etc.), a few parts use more advances probability theory. For ease of use we have marked all sections and exercises if they require an advanced understanding of probability theory. Furthermore we mark exercises with difficulty, as well as whether they require knowledge of the programming language Python.

 Section/subsection/exercise uses probability theory

 Exercise uses Python

 Exercise difficulties from easiest to hardest

The exercises provided are split into two parts. The first set of exercises concerns the topics covered in sections 1-6. The second set of exercises concerns the topics covered in 7-10.

Work in progress

- Improve sections 1, 7, 8.1, 8.2 and 11
- Run through and apply point-environment to main conclusions
- Make page-numbering two sided
- Part of Exercise 2

Version: 0.9 (preprint), March 4, 2020

Acknowledgement

We have been heavily inspired by, borrowed ideas and examples from, and learned extensively from the following resources.

Books

- Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, 2018
- Jonas Martin Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017
- Peter Spirtes, Richard Scheines, and Clark Glymour. *Causation, Prediction, and Search*. Springer New York, 1993

Introductory **presentations**

- Ferenc Huszar. Causal Inference in Everyday Machine Learning, January 2019. Machine Learning Summer School
- Bernhard Schölkopf. Causality, January 2019. Machine Learning Summer School
- Ricardo Silva. Causality, August 2019. Cambridge Advanced Tutorial Lecture Series on Machine Learning

A great **blog** on machine learning

- Ferenc Huszar. inFERENCe, 2012. URL `https://www.inference.vc/`

And a few **articles** with great examples

- Pedro A. Ortega. Bayesian Causal Induction. *preprint*, 2013
- Pedro A. Ortega. Subjectivity, Bayesianism, and causality. *Pattern Recognition Letters*, 64:63–70, 2015
- A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. *UCLA Department of Statistics Papers*, 2011
- R.C. Robinson. Enumeration of acyclic digraph. *Combinatorial mathematics and its applications (R.C. Bose et al.)*, 1970

Contents

Section 1

Introduction

1.1 What is Causality?

This note introduces the subject of *causality* and *causal inference*. Most people have an intuitive idea of what a *cause* is, and what an *effect* is, and what it means when something *causes* something else. However we will need need some more rigid definitions. The definition of causality that we will use in this note is

> **Causality:** Causality is a relationship by which one process or state, a *cause*, contributes to the production of another process or state, an *effect*, where the cause is partly responsible for the effect, and the effect is partly dependent on the cause.

The first key thing to note from the definition is that we have two processes or states; the cause and the effect. Examples of these two could be: the neighbours cat outside and my dog barking, a traffic jam and a road accident, etc. The definition also sets a directional relationship between the cause and the effect. For example it may be that a road accident caused the traffic jam, but there was no traffic jam before the accident (so the accident was not caused by a traffic jam). Finally we note that the cause is *partly* responsible for the cause and the effect is *partly* dependent on the cause. This is important as there may be many causes to an effect and a cause may not guarantee an effect. For example the dog may be barking at something else than the neighbours cat, and sometimes when the neighbours cat is close the dog may be asleep (and probably not barking).

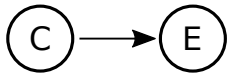
better examples

Furthermore we define

Causal Inference: Causal inference is the process of drawing a conclusion about the causal relationship between two processes or states.

Causal inference is simply when we attempt to figure out what causes what. This will often be the goal of, for example, looking out the window to see what the dog is barking at.
One great tool that's readily available for explaining assumptions and hypotheses is to illustrate them with *causal graphs*

Causal Graph: A causal graph is a graph describing causal relationships.
An example of a causal graph is



where the *cause* (C) has an edge directed towards the *effect* (E).

One example is shown below where the neighbours cat (C) walking by causes the dog to start barking (B).

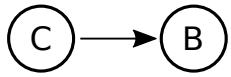


Figure 1: Causal graph for the dog-barking-at-cat situation.

When analysing systems we will often be measuring some variables, while other may not be available even though we know that they are important. We thus make the following definitions

Observable Variable: A variable whose state we can measure/observe. We visualize these by white nodes in graphs.

Hidden Variable: A variables whose state we can not measure/observer. We visualize these by grey nodes in graphs. One example of the graphical representation is

where the can hear the dog bark (B), but we can not see the cat (C) from where we are comfortably sitting in the living room.

1.2 Causality and Statistics

Statistics is an important tool that has been vital for the success of science and engineering. It is the foundation we use to determine whether experiments show the results we expect, or whether we need to change our minds.

The statistics community has though, had quite a bit of trouble with causality. It all boils down to the very famous phrase

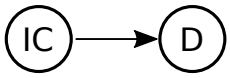
Correlation does not imply causation.

One example is the following fallacious conclusion.

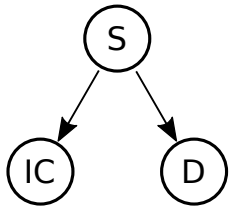
> As ice cream sales increase, the rate of drowning deaths increases sharply.
> Therefore, ice cream consumption causes drowning.

This hypotheses seems implausible. The root of the problem is that more ice cream is sold during the summer and more people go swimming during the summer. If more people go swimming, then more people are in risk of drowning. Thus ice cream and downing are both caused by a third thing: the season. Figure 2 graphs the two different explanations (hypotheses) of the phenomena.

The phrase *"correlation does not imply causation"* is important because people have believed very arbitrary, and sometimes problematic, things due to correlation. For example in Europe during the the Middle Ages people believed that lice where beneficial to your health, because sick people



(a) Hypothesis: Ice-cream (C) causes drowning (D).



(b) Hypothesis: Ice-cream (C) and drowning episodes (D) are both caused by season (S).

Figure 2: Two hypotheses for the correlation between ice-cream and drowning episodes.

rarely had lice. Turns out lice leaves people when their body temperature increases and lice did not at all have health benefits.

For a long time, many statisticians were opposed to the idea that we can even discuss causality. But causality is still quite important for science and engineering. If I put a burning Bunsen burner under a pot of water, the water temperature will increase over time. I can do this experiment millions of times with the same result: the temperature of the water is very correlated with the flame. So did the flame cause the increase in temperature or not? In science we accept experimental results as a way af showing causal relationships, which motivates the revised version of the phrase

> *Correlation does not imply causation. But sometimes it does.*

We are going to learn how to identify those sometimes.

1.3 Confounding

The above problem is an example of a *confounder*.

> **Confounder:** A confounder is a variable that influences both dependent and independent variables in a statistical/causal analysis.

In probability/statistical theory, a dependent variable is a variable that we are attempting to determine based on independent variables. In causality we take that relationship even more serious, where the dependent variable is *caused* by the independent variables.

In the aforementioned problem the dependent variable is drowning episodes while the only independent variable is ice-cream sales, because we are testing a hypothesis in which ice-cream sales cause drownings. The confounding variable is season, which influences both drowning episodes and ice-cream sales, and creates a spurious correlation which in turn motivates the incorrect hypothesis.

Confounders turn out to be absolutely essential for the correct analysis of causal relations and is the most common problem for such an analysis.

1.4 Serious Problems in the Real World

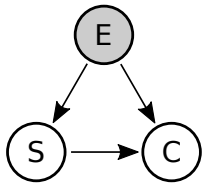
To motivate the study of causality we give two examples with serious consequences to the world.

A Dark Cloud

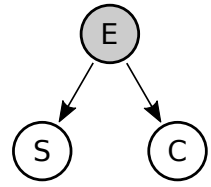
The first example is well known: *Does smoking cause lung cancer?*
Despite the agreement on the topic today, this was a very controversial problem in the late 1950s and early 1960s. Scientists had previously agreed on causal relationships that were simpler. For example it was known that the lack of vitamin C caused scurvy. There was a big problem with smoking though; some people didn't smoke and still got lung cancer, and some

people smoked all their lives and did not get lung cancer. As we will see later, handling these problems can be done using *randomized controlled trials* (experiments). In this case though, creating and experiment where we ask non-smokers to start smoking and take the smokes away from smokers, was not an option. While many scientists argued that smoking did cause lung cancer, others argued that there was a gene that caused lung cancer, while also increasing the craving for smoking. Mathematically it was very difficult to explain which model was correct.

The graph for the current theory is shown in Figure 3a. Here smoking has a causal link to lung cancer. Both smoking and cancer are affected by environment variables, such as genetic heritage influencing risk of cancer and culture influencing probability of being a smoker. The environment factors will usually be very difficult to determine and is thus considered a hidden variable. The alternative hypothesis is illustrated in Figure 3b. Here we also assume a causal relationship from environment to both smoking and cancer, while we assume no causal link between smoking and cancer.



(a) Current hypothesis for smoking and lung cancer.



(b) The alternative hypothesis for smoking and lung cancer from the 1960s.

Figure 3: Two hypotheses for the probability of lung cancer (C) depending on whether the persons smokes (S). We contract other variables, such as genetic heritage, outdoors environment, culture etc., into one variable called environment (E).

The smoking-promoting society abused the problem of confounders to their own advantage! They attempting to discredit the important work done by scientists to determine the health risks of smoking.

Using causal graphs can help us understand the underlying assumptions of hypotheses. It would be very good practice for more researchers to graph all assumptions for their work, as it becomes easier to specify where people disagree. As we will se later, the graphs can also be used to decide how to design experiments and analyse data, in order to test various hypotheses.

A Much Darker Cloud

Today most people seem to accept that smoking does cause lung cancer, but here is a problem with less agreement: *Is climate change caused by humans?*[1]

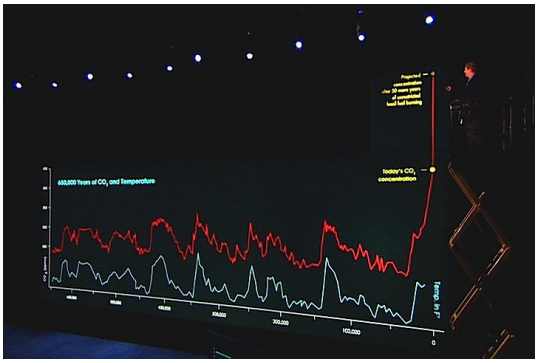


Figure 4: Al Gore comparing estimates of CO_2 and temperature.

Here we have the same problem as with smoking, but worse; we cannot do experiments on the outcome of the planets health, because we only have one! So how can we tell if CO_2 is causing climate change? Back in 2006 Al Gore attempted to warn the world about the problems of climate change in his documentary *An Inconvenient Truth*. One argument he used (with a bit a humour) was to compare the graphs of the estimated CO_2 and temperature over the past several hundred thousand years. The graph can be seen in Figure 4, where he humorously uses a lift to be able to point to the top of the chart.
Unfortunately this approach is not bullet-proof. In Figure 5 we show that the age of Miss America varies quite like the number of murders by steam, hot vapours and hot objects in the US. This brings us back to the problem

[1] it is

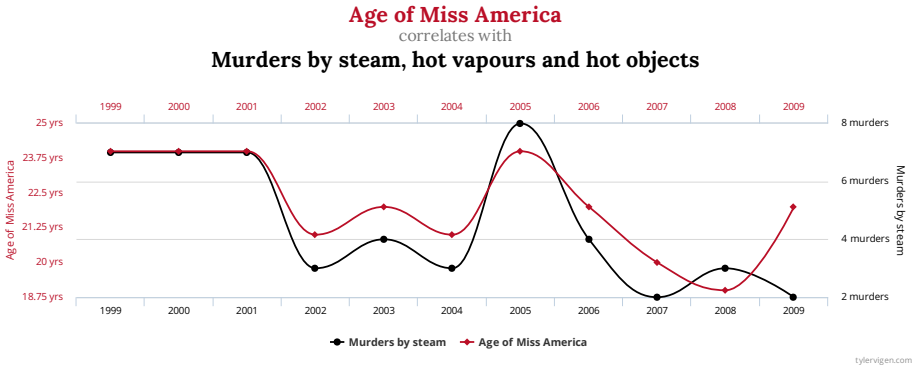


Figure 5: A spurious correlation[2].

with correlation, which will be the first argument used against someone, like Al Gore. We therefore need to carefully consider our causal arguments and experiments.

[2] There is a collection of these laughable, spurious correlations at `http://www.tylervigen.com/spurious-correlations`

Section 2

Interventions

In science we make plenty of causal conclusions. I don't think there exists a universally accepted definition of the *scientific method*, but here is what Wikipedia has to say

> **Scientific Method:** The scientific method is an empirical method of acquiring knowledge. It involves careful observation, applying rigorous scepticism about what is observed, given that cognitive assumptions can distort how one interprets the observation. It involves formulating hypotheses, via induction, based on such observations; experimental and measurement-based testing of deductions drawn from the hypotheses; and refinement (or elimination) of the hypotheses based on the experimental findings.

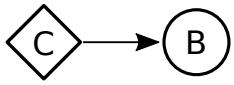
Notice first that the method is *empirical*; it makes observations and bases its inference on those observations. Secondly it concerns *knowledge*. While knowledge is difficult to define precisely one thing is for sure; knowing the causal structures of systems provides a LOT more knowledge than simply knowing the observational distributions. One can only state that more people in the 1960s got lung cancer, while the other can tell us WHY. Thirdly it concerns making hypotheses and making *experimental tests* to eliminate and confirm such hypotheses. Making causal hypotheses and testing them is precisely what this section will introduce you to.

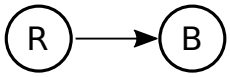
When we do scientific experiments we make controlled situations, in which specific variables are set to specific values/conditions, and then we observe what happens to other variables. For discussing this we will use the term *intervention*.

Intervention: An intervention is when we forcefully set a state or process to be specifically what we want, without affecting any other part of the causal system. We show an intervention graphically by

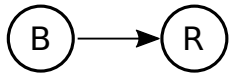
where a variable (V) has been changed by intervention.

For a physical experiment we have therefore intervened on all factors which we control and set when creating the experiment. All states and processes that we are not in control of, including those we measure, are not intervened on. One example could be to get a cat and place it near our house to see if the dog starts barking, which is illustrated below.





(a) Hypothesis 1: Rain changes the state of the barometer.

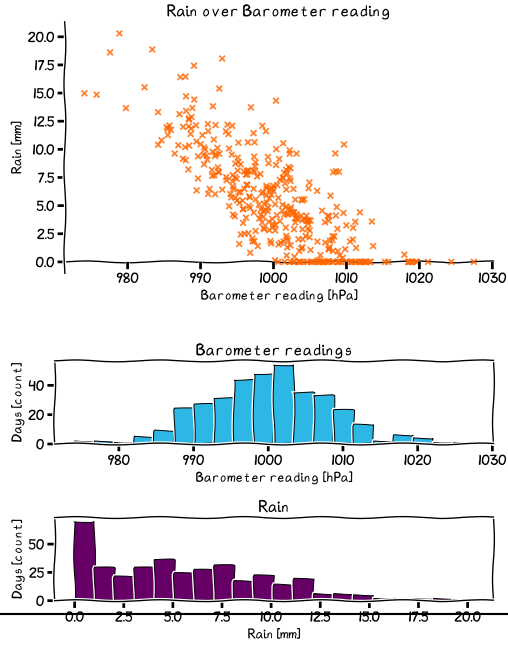


(b) Hypothesis 2: Barometer changes the state of the rain.

Figure 1: The barometer problem.

2.1 The Barometer

Let us consider a simple situation. Over a year (Year 1) we note down rainy days and barometer readings. Figure 2 shows the recorded values and from the scatter plot in the top it seems like the two values are very correlated. Note that we don't inherently know which one should be plotted above which (we don't even know which causes which) and so this was randomly chosen. We also plotted a histogram for each quantity to see how they behave independently of the other.



The figure here may be too small - or at least the text.

Figure 2: Data gathered about rain and barometer readings during Year 1.

We now create two competing hypotheses; the rain causes changes in the barometer (Hypothesis 1) vs. the barometer causes changes in the weather (Hypothesis 2). The hypotheses are graphed in Figure 1. This example is of cause quite trivial, but we can use it to ensure that are methodology in

causality will correctly produce the expected result.

In order to determine which hypothesis is correct we create the following experiment; we again note down rainy days and barometer readings over a year (Year 2), but this time we physically hold the barometer reading to 1000hPa (we do intervention on the barometer). Now there is no longer anything that can affect the barometer (because we are holding it tight). Therefore we know that all causes of the barometer readings must be cancelled (all ingoing arrows). The graphs of the two hypotheses after intervention is seen in Figure 3.

(a) Hypothesis 1: Rain changes the state of the barometer.

(b) Hypothesis 2: Barometer changes the state of the rain.

Figure 3: The barometer problem after intervention on the barometer.

Under hypothesis 1 we expect the rain to do what it always does (nothing has changed for that node), but we expect the barometer to be different from normal behaviour, because we choose its reading. Under hypothesis 2 we again expect the barometer to behave differently, but this time we also expect different behaviour of the rain, because the rain depends on the barometer.

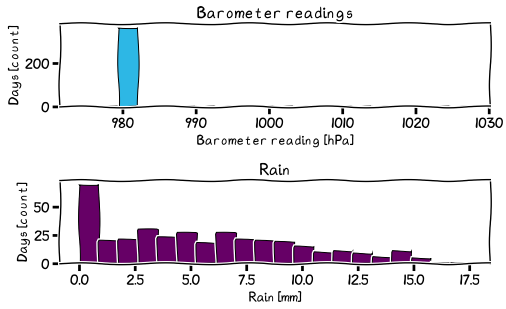


Figure 4: Histogram of barometer readings and rain amount after intervention on barometer during Year 2.

Now we look at our new results from the experiment (Figure 4). The histogram of rainy days seem similar as the one from Figure 2 and it seems plausible that the rain has followed the same distribution this year. The barometer though showed exactly the reading we set it to, so our intervention was certainly successful. If the rain did not change from the intervention while the barometer did, then we can accept hypothesis 1 and reject hypothesis 2 - we inferred a causal relationship!

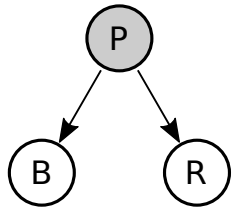
2.2 The Barometer - a Challenging Hypothesis

We could also have attempted to make an experiment by intervening on the rain. This is of cause challenging to do, but let's say someone has invented a machine that can do just that: cause rain no matter what the conditions were before and without changing anything else[1]. How did we avoid having to use this machine for a year? We avoided that because we made one important assumption: that only one of the two hypotheses were possible. Thus falsifying one was equivalent of verifying the other.

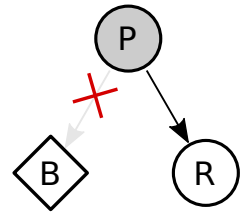
But what if someone introduced a third hypothesis (Hypothesis 3): rain and barometer readings are both caused by atmospheric pressure. Our previous experiment still falsifies hypothesis 2, but it does not verify hypothesis 1. The new hypothesis is graphed in Figure 5a.
Under hypothesis 3 we expect the same results of our experiment as with hypothesis 1. One further complication may be that we can not observe the atmospheric pressure (we measure the atmospheric pressure is measured by some kind of barometer, which is what we are currently investigating).

Let's therefore try out the machine that controls rain! If we set the machine to make constant rain for a year (Year 3 - depressing I know) we expect the following causal graphs for our two competing hypotheses (hypothesis 1 and 3).

[1] I'm guessing the machine pumps a ton of water into the air like a garden hose.

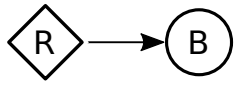


(a) Hypothesis 3: Pressure changes the state of rain and barometer.

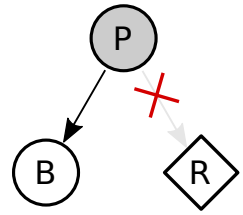


(b) Hypothesis 3: Pressure changes the state of rain and barometer after intervention.

Figure 5: The alternative hypothesis.



(a) Hypothesis 1: Rain changes the state of the barometer.



(b) Hypothesis 3: Pressure changes the state of rain and barometer.

Figure 6: Hypothesis 1 and 3 after intervening on rain.

This time we expect rain to have constant value for both of the hypothesis. For hypothesis 1 we expect the barometer to be constant as caused by the rain, but for hypothesis 3 we expect the barometer to behave normally throughout the year (have varying values depending on pressure).

After running this experiment

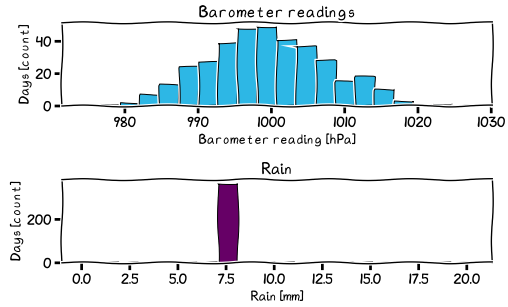


Figure 7: Histogram of barometer readings and rain amount after making it rain during Year 3.

we see the histograms in Figure 7. We note that the barometer is in fact *not* constant throughout the year and that rain thus does *not* cause barometer change. We have falsified hypothesis 1 while verified hypothesis 3: there is a third cause (pressure) that causes both rain and barometer reading.

The important take-home message from the barometer problem is this;

> *When we consider a set of hypotheses or set of causal graphs, we are automatically assuming that* ***anything*** *not mentioned is irrelevant.*

This is quite critical, because experiments can be correctly made, while drawing wrong conclusions because of left-out variables. It also means that even if we make a causal graph with potential links between all nodes (everything influences everything else to some degree), then we are still making the assumption that anything not in the graph is irrelevant - we can never consider all factors in the universe.

Section 3

The Do-operator

In the previous section we have shown how to do interventions in order to test various causal hypotheses. In order to analyse this more thoroughly we need to introduce some mathematical notation and pair it with our knowledge of probability theory.

The table on the next page is a quick recap of probability theory.

Table 1: Probability theory recap.

$P(x)$	Probability (sometimes called *marginal* probability) of observing event x.
$P(\neg x)$	Probability of *not* observing event x.
$P(x, y) = P(x \cap y)$	*Joint* probability of observing both events x and y.
$P(x \mid y)$	*Conditional* probability of observing event x given that we have already observed y.
$P(x \cup y)$	Probability of observing x and/or y.

Normality	For any x	$0 \leq P(x) \leq 1$
Compliment	For any x	$P(\neg x) = 1 - P(x)$
Additivity	x and y are mutually exclusive iff.	$P(x \cup y) = P(x) + P(y)$
Totality	x and y are mutually exclusive and collectively exhaustive iff.	$P(x \cup y) = P(x) + P(y) = 1$
Independence	x and y are independent iff.	$P(x, y) = P(x)\, P(y)$
Overlap	For any x and y	$P(x \cup y) = P(x) + P(y) - P(x, y)$
Multiplication	For any x and y	$P(x, y) = P(x \mid y)\, P(y)$
Conditional	For any x and y	$P(x \mid y) = \frac{P(x, y)}{P(y)}$
Monotonicity	For any x and y	$P(x, y) \leq P(x)$
Bayes rule	The fundamental rule of learning!	$P(y \mid x) = \frac{P(x \mid y)\, P(y)}{P(x)}$

Uppercase letters like A, X and Y denote stochastic variables (parts of systems which we do not know the value of). These symbols are also used for nodes in causal graphs, because nodes in causal graphs represent variables which could take on different values at each experiment. Small case letter like a, x and y denote specific values for the related stochastic variables (for example x is shorthand for $X = x$).

In order to handle interventions mathematically we introduce the *do*-operator

Do-operator: $\mathring{x}$ denotes that we have intervened to force event x to happen (we *do* the event). $P(y \mid \mathring{x})$ is the probability of y given that we have intervened and forced event x. We can also set a variable to a specific value by $P(y \mid x \overset{\circ}{\leftarrow} 2)$. Note that in some literature the do-operator has different notation. For example most use $P\big(y \mid \text{do}(x = 3)\big)$[a].

[a] Pearl's notation.

Here's a couple of things to note about the do-operator. First of all it's *very* different from conditioning - the next section makes that clear. Secondly it is a form of *assignment*, similar to that of computer science. We *causally assign* a value to a variable. We can also do other do-operations like assigning a new probability distribution to a variable. For example here we assign a normal distribution to a variable

$$p(x) \overset{\circ}{\leftarrow} \mathcal{N}(0, 1). \tag{1}$$

This is what we do in randomized trials (more on that later), where we for example randomly select people for a test. We can also make interventions to create restrictions on the system. For example

$$P(z \mid x \overset{\circ}{\leftarrow} y), \tag{2}$$

is the probability of z, given that we force variable x to be equal to variable y (for some system) - y becomes in charge of x.

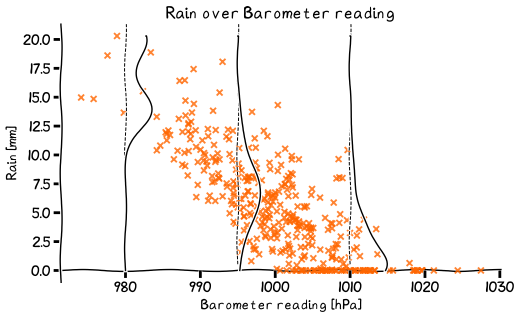
3.1 Interventional Distribution for Barometer

The first thing we need to realize is that generally

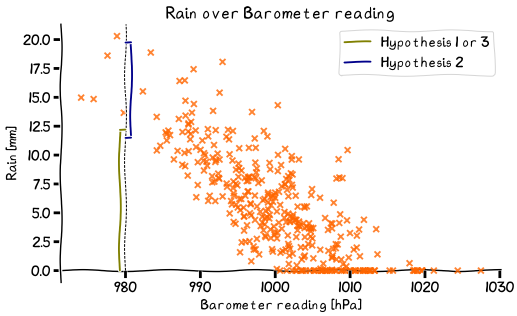
> *Conditional distribution does not equal interventional distribution:*
>
> $$P(y \mid x) \neq P(y \mid \mathring{x})$$

The conditional distribution is the distribution of a variable given that we have observed another variable, while remaining a passive observer. The interventional distribution on the other hand is the distribution of a variables given that we actively change another variable in the system.

In order to realize this consider the barometer problem. In Figure 1a we see three conditional distributions plotted on top of our data. They show the probability density of rain conditioned on our barometer readings at 980hPa, 995hPa and 1010hPal. We can see that the expected rainfall is different depending on our reading of the barometer because the two variables corre-



(a) $P(\text{rain} \mid \text{barometer})$ at Year 1.



(b) The ranges of values expected for the interventional distribution under Hypothesis 1 and 2.

Figure 1: Analysis of Year 1.

late.

Now we assume we are going to intervene on the value of the barometer (like Year 2) and set it to 980. If Hypothesis 2 is true (barometer reading causes rainfall), then we expect that our intervention will change the behaviour of rainfall for that year. In this case we expect the rainfall to be mostly in the range shown in blue in Figure 1b, because that is where most rainfall is conditioned on the barometer reading being 980 - the conditional from Figure 1a.

If on the other hand Hypothesis 1 or 3 is true (rain or pressure causes the barometers reading), then we expect our intervention to have no effect on the rainfall. We thus expect the rainfall to be mostly in the green range of 1b, because that is generally where most rainfall is.

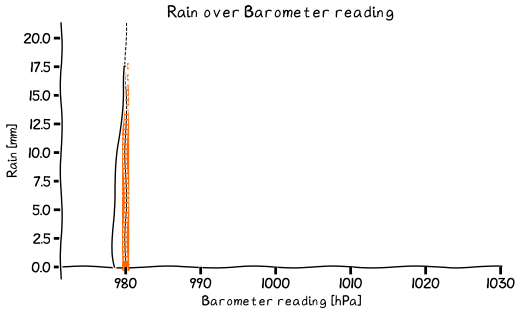


Figure 2: Interventional distribution from Year 2. $P(\text{rain} \mid \text{barometer} = 980)$

Since we did collect interventional data during Year 2, we can plot the interventional distribution as in Figure 2. Here all the datapoints are on-top of each other, because they all have the same barometer reading. The result corresponds well with the expectations of hypothesis 1 or 3 (the barometer reading does not cause anything).

3.2 Interventional Distributions in General

Let's consider the interventional distribution in a more generic setting. We have two variables X and Y and we believe X causes Y. We have contracted all shared causes of X and Y into one node called C - confounders. We have contracted all causes of X which does not affect Y directly into one node called A - ancestors. We have contracted all variables affected by y into one node called D - descendants. Finally we have contracted all nodes between x and y into one node called M - mediators. The situation is shown the Figure 3.

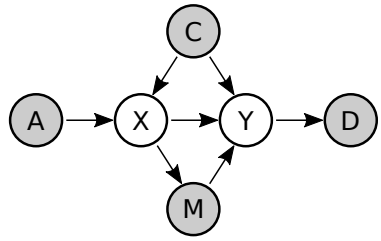


Figure 3: A quite generic causal graph investigating causal relationship between x and y.

Note that there could be additional links between C and D, A and D etc., but it would not change the outcome of the following analysis. One assumption that is important though, is that the graph does not have cycles (at least not any involving x and y)[1].

We notice the following conditional independencies

$$P(y \mid x, a) = P(y \mid x) \tag{3}$$
$$P(x \mid y, d) = P(x \mid y). \tag{4}$$

[1]Causal systems with cycles is in general very difficult to analyse mathematically and is part of contemporary research.

The conditional probability of y becomes

$$P(y \mid x) = \sum_{acm} P(y \mid x, a, c, m) P(a, c, m \mid x) \tag{5}$$

$$= \sum_{cm} P(y \mid x, c, m) P(m \mid c, x) P(c \mid x) \tag{6}$$

$$= \sum_{cm} P(y \mid x, c, m) P(m \mid x) P(c \mid x), \tag{7}$$

where we can remove a because it is independent of y conditioned on x.

We now intervene on x resulting in the following interventional graph

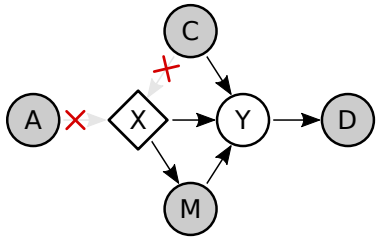


Figure 4: The interventional (on x) version of Figure 3.

The interventional probability of y is

$$P(y \mid \mathring{x}) = \sum_{acm} P(y \mid x, a, c, m) P(m \mid x) P(a, c) \tag{8}$$

$$= \sum_{cm} P(y \mid x, c, m) P(m \mid x) P(c). \tag{9}$$

We now see the difference between the interventional distribution and the conditional distribution by

$$P(y \mid x) = \sum_{cm} P(y \mid x, c, m) P(m \mid x) P(c \mid \boldsymbol{x}), \tag{10}$$

$$P(y \mid \mathring{x}) = \sum_{cm} P(y \mid x, c, m) P(m \mid x) P(c). \tag{11}$$

The conditional distribution conditions all variables in the system on X, while the interventional distribution only conditions the *downstream* variables on X. This difference in conditioning is exactly what makes the

interventional distribution special.

Perhaps more importantly,

> *If there are no confounders for the causal problem at hand, then the conditional distribution* ***equals*** *the interventional distribution, when intervening on the* ***cause****. (assuming no cycles)*

If we know what is the cause, we can therefore reduce our problem of inferring causal relationships to that of correctly handling the confounders of the problem.

We also conclude is that any ancestor (A) of X, which is only an ancestor of Y *through* X, is irrelevant for both distributions, because all information about A passed through X which is known in both cases. Also the descendants (D) of Y, that are only descendants of X *through* Y, are irrelevant because all information they have are already captured in Y.

In this example we knew that the causal direction was from X towards Y, and definitely not the other way. This is not always the case though. The *disambiguation problem* concerns situation in which we don't know the direction. If two events are always observed together, then which of them caused the other? An even more generic causal graph is thus

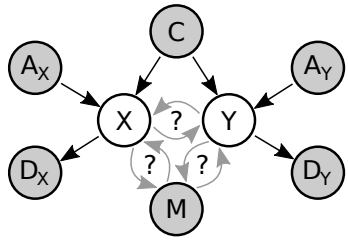


Figure 5: An even more generic causal graph investigating causal relationship between x and y.

Here we have ancestors and descendants for both X and Y, while still only having a single confounders C. The directions of the links between X and Y and the mediators are not know. For this problem we now need to both handle confounders *and* disambiguation of direction (while we are still assuming *no* cycles).

When inferring a causal relationship between two variables we need to handle all confounders as well as the disambiguation problem. (assuming no cycles)

Another complication can be if we want to determine the strength of the *direct* causal relationship between X and Y, in which case the mediators M can also be trouble.

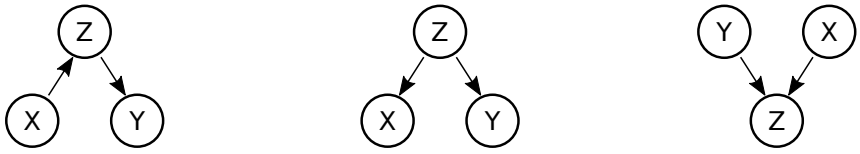
3.3 Causal Patterns

When analysing causal graphs it's very useful to consider three types of arrangements of three causal variables.

Should these three actually be in definition blocks?

The first arrangement we are considering is the *chain junction* as illustrated in Figure 6a. In this arrangement one variable, X, is *directly* causing another, Z, which is *directly* causing a third, Y. We also say that X causes Y, just not directly, and we call Z a *mediator*.

The second arrangement is the *confounding junction* also called a fork, and is seen in Figure 6b. This arrangement is the one we considered in the barometer problem, where pressure the confounding variable. Since confounding is a very central concept in causal analysis, this pattern also becomes quite crucial.



(a) Chain junction (b) Confounding junction / fork (c) Collider

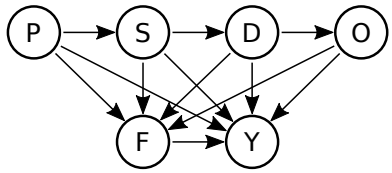
Figure 6: Three important arrangements of causal variables.

The third arrangement can be a bit more tricky in causal analysis - for reason that will later become apparent. We call the arrangement a *collider* and it is illustrated in Figure 6c. It is composed of three causal variables, where two of them (X and Y) directly cause the third one (Z). Z thus holds information about X and Y, while X and Y each holds some information about Z.

3.4 Randomized Trials

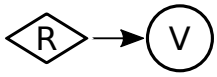
One essential part of scientific studies has been the concept of randomized trials. Let us consider a situation in which a scientist wants to help a set of farmers. The farmers have different fertilizers F and wants to know which one produces highest yield Y based on the subjected plant P, soil fertility S, water drainage D of the area and other factors O. This is actually a situation which the famous statistician R. A. Fisher was concerned with Reference to Book of Why. The farmers have worked in their field for many years and have some guesses for what to use where, but they don't agree and have made no shared documentation of strategies.

We segment their fields into various segments, all of which we can monitor. If we start monitoring the segments we get data on all parameters based on the following causal model.



From this model we can see a problem. We only want a single connection between F and Y, but there are 4 confounders!. Previously we saw that we can solve this with intervention. For example we could simply choose a single fertilizer to use everywhere and then we could determine the effect of that one fertilizer to the yield. This does not solve our task though, as we want to compare the fertilizers and not just evaluate a single one. We therefore need to test all the fertilizer, but also make sure that the confounders are not a problem. This is where the randomized trial becomes relevant.

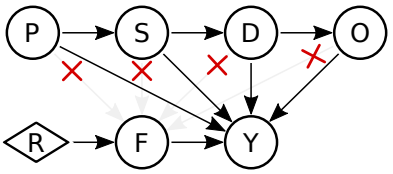
Randomized Trial: An experiment with randomization on causal variables in order to determine their effect.
The graph a randomized variable by



to illustrate that the variable V is still a stochastic variable, but that it now follows the distribution from the randomization R. All incoming links to V can now be ignored.

If we randomly choose a fertilizer for each segment, then we eliminate all relationships that affect the farmers choice of fertilizer. The randomization is a special case of intervention, where we do not intervene with a single value, but rather we replace the original distribution of a variable with a custom made one. By randomization of the fertilizer we get data from a different causal model

Show the do-notation of replacing the probability distribution. Exemplify by rolling a die.



In this causal model there is only one causal link between F and Y, which is the one we want to measure.

Randomized trials/experiments are extremely important for scientific research and is known by many names. A/B testing for example is a special case with two possible values for a causal variable. In medicine A/B testing will typically be done by randomly assigning a new type of medicine (A) or a placebo (B) to patients and measure their improvement.

Section 4

Causal Inference - The Switch

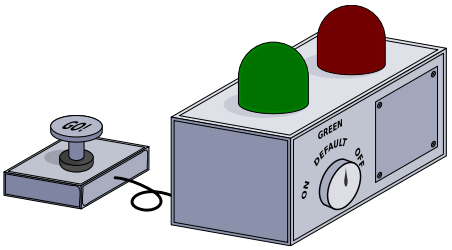


Figure 1: The switch problem.

In this section we will get to the goal of all this causality-talk; the ability to infer causal relationships. We will use our knowledge about intervention and the do-operator to solve a problem similar to the barometer problem, but where the answer is not intuitively known: the switch problem.

We have a box with a green and a red light on it. If we press GO on the related contact, one of four things happen. Either both light turn on, both stay off, the green light turns on or the red light turns on. The process of the outcome is stochastic, but we can get a lot of sample by keep pressing GO and write down what happens.

	$\neg g$	g
$\neg r$	$\frac{3}{8}$	$\frac{1}{8}$
r	$\frac{1}{8}$	$\frac{3}{8}$

Table 1: Recorded outcomes of switch experiment in ratios.

We use the box a ton of times and write down all outcomes. We compute how often each outcome occurs and write the results in a confusion matrix seen in Table 1. It seems that the light are

usually on together and off together, while sometimes they differ.

We use this to hypothesise that one light is the cause of the other light - with some noise. At this point though, we do not know which light causes which.

(a) Hypothesis h_G (b) Hypothesis h_R

Figure 2

In Figure 3 we have illustrated the hypotheses and possible outcomes.

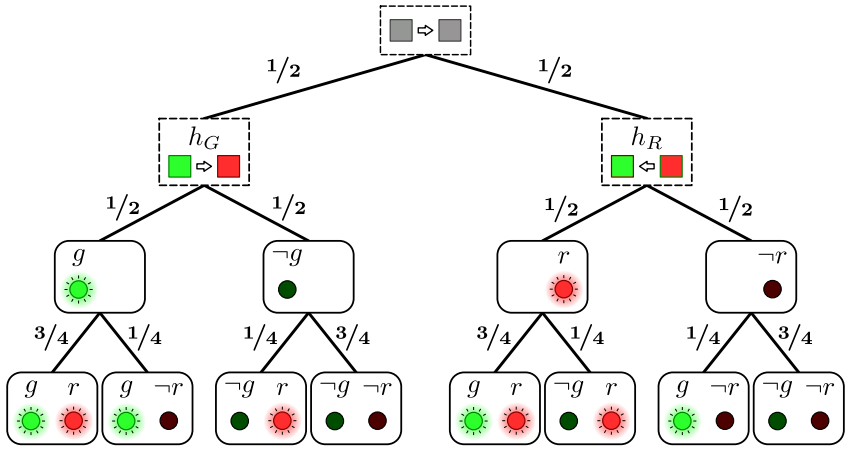


Figure 3: Hypotheses and outcomes of the switch problem.

The dashed-lined boxes are hypotheses. The top one hypothesises that one light causes the other. On the left we hypothesise that green causes red (h_G), and on the right we hypothesise that red causes green (h_R).
Below the hypotheses we have illustrated all outcomes in a hierarchical manner where cause precedes effect (if green causes red, then the state of green is above the state of red).

At each edge we have noted a probability. At the top we indicate that we believe the two hypotheses are equally probable. For both hypotheses there is a 50/50 chance of the cause-light turning on, and then some probabilities for the behaviour of the effect-light. The probabilities are made so that they corresponds to the ratios in Table 1.

4.1 Agent 1: An Observational Sample

We now press GO, see that *both lights turn on* and give this datapoint to *Agent 1*. Agent 1 will attempt to determine which hypothesis is most probable, by applying the probabilities from Figure 3 with the sample and Bayes theorem.

$$P(h_G \mid g, r) = \frac{P(g, r \mid h_G)\, P(h_G)}{P(g, r \mid h_G)\, P(h_G) + P(g, r \mid h_R)\, P(h_R)} \tag{1}$$

$$= \frac{P(r \mid g, h_G)\, P(g \mid h_G)\, P(h_G)}{P(r \mid g, h_G)\, P(g \mid h_G)\, P(h_G) + P(g \mid r, h_R)\, P(r \mid h_R)\, P(h_R)} \tag{2}$$

$$= \frac{\frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2} = P_{A1}(h_G), \tag{3}$$

$$P(h_R \mid g, r) = \frac{P(g, r \mid h_R)\, P(h_R)}{P(g, r \mid h_G)\, P(h_G) + P(g, r \mid h_R)\, P(h_R)} \tag{4}$$

$$= \frac{P(g \mid r, h_R)\, P(r \mid h_R)\, P(h_R)}{P(r \mid g, h_G)\, P(g \mid h_G)\, P(h_G) + P(g \mid r, h_R)\, P(r \mid h_R)\, P(h_R)} \tag{5}$$

$$= \frac{\frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2} = P_{A1}(h_R). \tag{6}$$

Based on this datapoint, Agent 1 concludes that the two hypotheses are equally probable - just like what we started out with. Actually no matter what the outcome is from the lights, and no matter how many datapoints we get, we will not come any closer to determining the causal structure. Since you have made it this far you should of cause have expected this; we can not make causal inference simply from observing the data.

4.2 Agent 2: An Interventional Sample

There is a switch on the side of the box, which allows us to control the state of the green light. It can be set to either DEFAULT, for the currently observed behaviour, or to ON or OFF, which forces the green light to be on or off respectively. If we use the switch to force the green light on, we intervene on the system. The space of hypotheses is the same, but the space of states is now reduces, which is illustrated in Figure 4. Some probabilities are now zero and one, because of the forced green light.

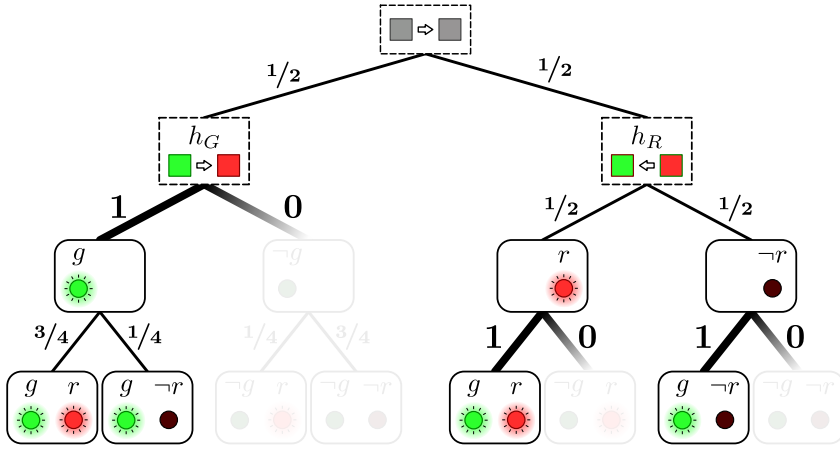


Figure 4: Hypotheses and outcomes of the switch problem during intervention on the green light.

We force green to be on, press GO and see both lights on. The outcome is identical to the observational sample, but we know it is interventional and hand it to *Agent 2*. Agent 2 also computes the probability of each

hypothesis (differences are highlighted in blue)

$$P(h_G \mid \mathring{g}, r) = \frac{P(g, r \mid \mathring{g}, h_G)\, P(h_G)}{P(g, r \mid \mathring{g}, h_G)\, P(h_G) + P(g, r \mid \mathring{g}, h_R)\, P(h_R)} \tag{7}$$

$$= \frac{P(r \mid \mathring{g}, h_G)\, P(g \mid \mathring{g}, h_G)\, P(h_G)}{P(r \mid \mathring{g}, h_G)\, P(g \mid \mathring{g}, h_G)\, P(h_G) + P(g \mid \mathring{g}, r, h_R)\, P(r \mid h_R)\, P(h_R)} \tag{8}$$

$$= \frac{\frac{3}{4} \cdot 1 \cdot \frac{1}{2}}{\frac{3}{4} \cdot 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{3}{5} = 0.6 = P_{A2}(h_G), \tag{9}$$

$$P(h_R \mid \mathring{g}, r) = \frac{P(g, r \mid \mathring{g}, h_R)\, P(h_R)}{P(g, r \mid \mathring{g}, h_G)\, P(h_G) + P(g, r \mid \mathring{g}, h_R)\, P(h_R)} \tag{10}$$

$$= \frac{P(g \mid \mathring{g}, r, h_R)\, P(r \mid h_R)\, P(h_R)}{P(r \mid g, h_G)\, P(g \mid \mathring{g}, h_G)\, P(h_G) + P(g \mid \mathring{g}, r, h_R)\, P(r \mid h_R)\, P(h_R)} \tag{11}$$

$$= \frac{1 \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4} \cdot 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{2}{5} = 0.4 = P_{A2}(h_R). \tag{12}$$

Now we see a difference in the belief in the two hypotheses! When we intervene to force the green light on and red also turns on, Agent 2 concludes that h_G more probable than h_R.

4.3 Prediction On Observational Sample

Agent 1 believes that the two hypotheses concerning the light-box are equiprobable. Agent 2 on the other hand, believes that the probability of h_G being correct is 60%, while the probability of h_R is only 40%.

We now get an observational sample (no intervention) and observe the green light only - it turns on. We then ask the two agents to compute the probability of the red light being on/off.

The probability of the red light being on for h_G can be read off of Figure 3.

$$P(r \mid g, h_G) = \frac{3}{4}. \tag{13}$$

The probability of the red light being on for h_R can be computed using Bayes theorem

$$P(r \mid g, h_R) = \frac{P(g \mid r, h_R)\, P(r \mid h_R)}{P(g \mid r, h_R)\, P(r \mid h_R) + P(g \mid \neg r, h_R)\, P(\neg r \mid h_R)} \tag{14}$$

$$= \frac{\frac{3}{4} \cdot \frac{1}{2}}{\frac{3}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2}} = \frac{\frac{3}{8}}{\frac{4}{8}} = \frac{3}{4}. \tag{15}$$

The agents' predictions are simply a weighted mean of these two values, but since they are identical we trivially have

$$P_{A1}(r \mid g) = P(r \mid g, h_G)\, P_{A1}(h_G) + P(r \mid g, h_R)\, P_{A1}(h_R) \tag{16}$$

$$= \frac{3}{4} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{4} \tag{17}$$

$$P_{A2}(r \mid g) = P(r \mid g, h_G)\, P_{A2}(h_G) + P(r \mid g, h_R)\, P_{A2}(h_R) \tag{18}$$

$$= \frac{3}{4} \cdot \frac{3}{5} + \frac{3}{4} \cdot \frac{2}{5} = \frac{3}{4}. \tag{19}$$

The two agents thus makes identical *observational predictions*, because both hypotheses are capable of perfectly explaining the observational distribution.

4.4 Prediction On Interventional Sample

Now we get an interventional sample, by forcing the green light on and again ask the two agents to predict the probability of the red light.

The probability of the red light being on, under hypothesis h_G, during intervention stays the same

$$P(r \mid \mathring{g}, h_G) = \frac{3}{4}. \tag{20}$$

The probability under hypothesis h_R changes to

$$\begin{aligned} P(r \mid \mathring{g}, h_R) &= \frac{P\big(g \mid \mathring{g}, r, h_R\big)\, P(r \mid h_R)}{P\big(g \mid \mathring{g}, r, h_R\big)\, P(r \mid h_R) + P\big(g \mid \mathring{g}, \neg r, h_R\big)\, P(\neg r \mid h_R)} \\ &= \frac{1 \cdot \frac{1}{2}}{1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}} = \frac{1}{2}. \end{aligned} \tag{21}$$

The interventional predictions of the two agents are

$$P_{A1}\big(r \mid \mathring{g}\big) = P\big(r \mid \mathring{g}, h_G\big)\, P_{A1}(h_G) + P\big(r \mid \mathring{g}, h_R\big)\, P_{A1}(h_R) \tag{22}$$

$$= \frac{3}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{8} + \frac{2}{8} = \frac{5}{8} = 0.625 \tag{23}$$

$$P_{A2}\big(r \mid \mathring{g}\big) = P\big(r \mid \mathring{g}, h_G\big)\, P_{A2}(h_G) + P\big(r \mid \mathring{g}, h_R\big)\, P_{A2}(h_R) \tag{24}$$

$$= \frac{3}{4} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{2}{5} = \frac{9}{20} + \frac{4}{20} = \frac{13}{20} = 0.65. \tag{25}$$

The two agents has *different interventional predictions*, because they have learned different causal structures! The difference in these two number is quite small, but remember this comes from a *single* training point :)

4.5 More Samples for Agent 2

To finish off the switch problem we here show how the causal inference would go about with more data. If we pressed GO a total of 10 times, while forcing green light on, and observed that the red light was on 7 times and off 3 times. We can further improve Agent 1's knowledge about the hypothesis space. This time we remove some trivial terms[1] to simplify the expressions

$$P(h_G \mid \mathring{g}, \{(\neg r)^3, r^7\}) \tag{26}$$

$$= \frac{P\Big(\{(\neg r)^3, r^7\} \;\Big|\; \mathring{g}, h_G\Big)\, P(h_G)}{P\Big(\{(\neg r)^3, r^7\} \;\Big|\; \mathring{g}, h_G\Big)\, P(h_G) + P\Big(\{(\neg r)^3, r^7\} \;\Big|\; \mathring{g}, h_R\Big)\, P(h_R)} \tag{27}$$

$$= \frac{\left(\frac{3}{4}\right)^7 \cdot \left(\frac{1}{4}\right)^3 \cdot \frac{1}{2}}{\left(\frac{3}{4}\right)^7 \cdot \left(\frac{1}{4}\right)^3 \cdot \frac{1}{2} + \left(\frac{1}{2}\right)^7 \cdot \left(\frac{1}{2}\right)^3 \cdot \frac{1}{2}} = \frac{\frac{2187}{2097152}}{\frac{2187}{2097152} + \frac{1}{2048}} \approx 0.68. \tag{28}$$

Using this data, Agent 1 have further increased its belief that h_G is the correct hypothesis and not h_R. Of cause it has still not completely ruled out hypothesis h_R, because that system can also create the same data - it's just less likely.

[1] and abuse mathematical notation a bit

Section 5

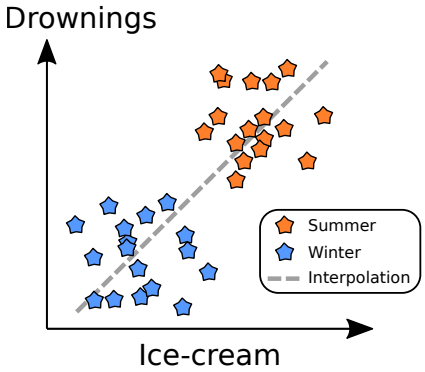
Conditioning

5.1 Conditioning on Confounders

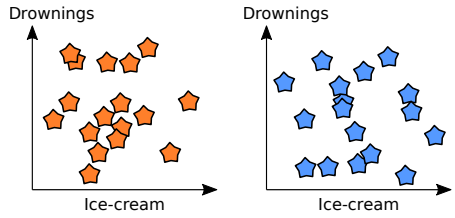
We have looked at how to manipulate a causal graph using intervention and randomization (which is also an intervention). We will now look at a statistical tool which can also help, *conditioning*. Conditioning can help us with confounding variables, which as mentioned is critical for causal analysis (the barometers problem, ice-cream-drowning problem and fertilizer problem are all concerned with confounder issues).

Consider again the problem of summertime confounding the relationship between ice-cream sales and drowning incidents. Let's say we picked some data for days in the winter and days in the summer. It may seem a bit harsh that there are drownings every day, but let's assume this is a very big country with a population that really likes extreme water-sports.
Figure 1a shows the data for summer and winter. We have also plotted a regression line, which on the face of it explains the relationship between ice-cream sales and drowning episodes. We see a clear relationship between the two variables which creates the mentioned spurious correlation.

In Figure 1b we condition on summer data-points and winter data-points respectively. When we analyse the two groups independently we do not see the spurious correlation any more. Actually for the two groups apart it looks like there is no correlation at all! We have thus solved the confounding problem by conditioning on the confounding variable.



(a) All data.



(b) Data conditioned on summer and winter.

Figure 1: Data for drowning episodes and ice-cream sales during summer and winter.

Conditioning can solve the problem of confounders!

As we have seen in section 3.2, solving confounders is both essential and sufficient for inferring causal relationships, which makes conditioning an extremely important tool.

Consider what conditioning actually is; we set a variable to a constant value, based on observational data. It is like we make one causal model for each value of the conditioned variable. We have illustrated this interpretation in Figure 2, where we condition on variable Z. The square node of Z indicates that this node is a *constant*, while X and Y remain stochastic variables (round). For each possible values of Z we have a square box and a related set of stochastic nodes for X and Y, because these may behave differently for each value of Z. The ancestor A of Z is also duplicated, because the distribution of A is very different depending on what it caused Z to be. When we intervened on nodes we removed their causal links from their ancestors, but when we condition on Z we not **not** remove them, because A is still the cause of Z. We are merely selecting the A for which Z is the conditioned value.

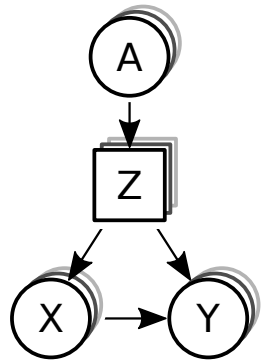


Figure 2: Graphical interpretation of conditioning.

The causal relationships potentially change dramatically depending on the value of Z. It is possible that X causes Y for one value of Z, while their is no causal relationship between X and Y for another value of Z. It could even be that the causal relationship reverses so that Y causes X for a third value of Z. We have further illustrated this situation in Figure 3.
For each value of the conditioned variable Z the causal graph has different distributions and potentially different causal structures.

While conditioning is an extremely important tools there is a couple of catches, which are investigated in the next sections.

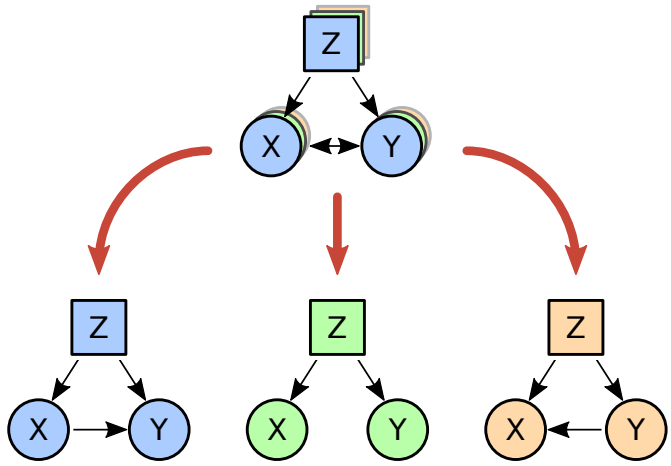


Figure 3: Expanded example of graphical interpretation of conditioning.

5.2 Large Conditioning Space

Let us say you have two confounding variables, which can each take on 10 values. If we condition on both of these we will have a total of $10 \times 10 = 100$ different combinations. In order to make a proper analysis of each conditioned group, we would therefore roughly need 100 times as much data! This problem becomes even more complicated if we are considering continuous variables, because how do you correctly condition on a variable which (in theory) has infinitely many possible values? How do you discretize? Let us say we also collected some data on the ice-cream-drowning problem from spring. Some of these datapoints will be similar to winter, while some of them will be similar to summer. Figure 4 illustrates this data.

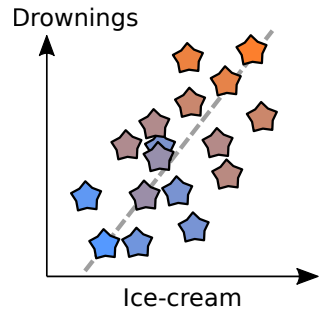


Figure 4: Data conditioned on spring.

We have colourized this data, so that the early spring-datapoints have a

colour similar to winter-datapoints and the late spring-datapoints have a colour similar to summer-datapoints. The spurious correlation between ice-cream sales and drowning episodes have now reappeared! The datapoints close to summer are in the top-right corner, while the winter-like datapoints are in the lower-left corner. This is because we haven't correctly conditioned on our season. In reality spring appears to be a transition period between winter and summer, and can therefore not really be conditioned on by simple measures.

In conclusion

> *Conditioning can become a problem if the space of conditioned variable-states becomes large or continuous.*

5.3 Causes, Mediators and Colliders

Direct Cause and Mediators

Consider the simple causal graph in Figure 5a. If we condition on $X = x$ in this graph we conceptually make X a constant as in Figure 5. If X is constant then Y will always take on the same value. If we sample from this graph we will get various values for Y and always the same value for X. By conditioning on the cause we are trying to analyse we therefore create a problem that wasn't there.

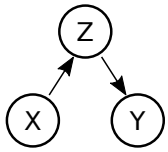
(a) Direct causal link

(b) Conditioning on direct causal link - single value

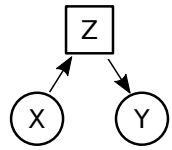
Figure 5: Unfortunate conditioning on direct causal link.

While the above example seems a bit silly, consider instead the *chain junction* from section 3.3. If we condition on the mediator Z, we conceptually make it a constant. No information is ever passed from X to Y, so it

may look like X does not cause Y - even though it does! If we have a theory about Z being a *confounder*, while in reality Z is a *mediator*, then conditioning on it ruins our analysis. We must therefore be sure about Z being a confounder before we condition on it.



(a) Chain junction



(b) Conditioning on mediator - single value

Figure 6: Unfortunate conditioning on mediator.

Conditioning on a mediator disables our ability to see causal relationships through that mediator.

Collider

Now we turn our attention to the *collider* from section 3.3. If we condition on the collider Z we conceptually make Z into a constant z, but we do so by carefully selecting the correct samples where $Z = z$. These samples are the ones where X and Y created the correct circumstances for which Z became z.

In order to understand colliders better we make a simple example with two coins; X and Y. We flip both coins and count the number of heads, Z. Clearly $z \in \{0, 1, 2\}$. If we know X and Y then we know Z exactly. If we know that coin X is heads ($x = 1$), then $z \in \{1, 2\}$, while if X is tails ($x = 0$), then $z \in \{0, 1\}$. Knowing x gives us no way of knowing y unless we also know z.

When the X-coin is known, then the remaining variability in Z comes from Y.

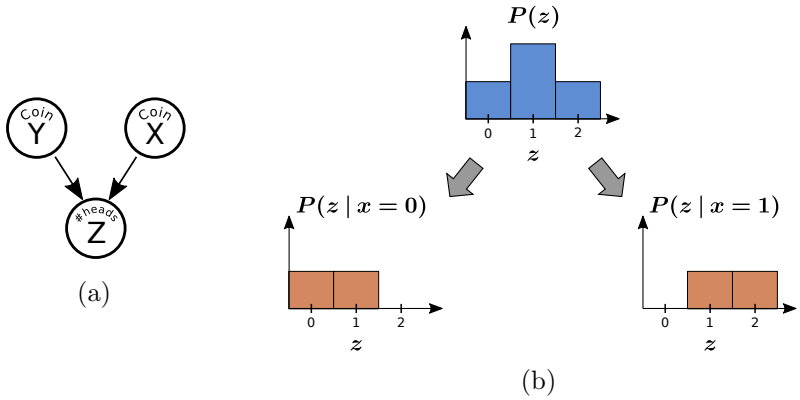


Figure 7: Double coin flip example.

If we intervene on Z then all causal relationship in the collider are eliminated. No information flows anywhere. X and Y can not predict Z and Z can not help us determine X and Y.

But what happens when we condition on Z? If we condition the number of heads to $Z = 1$, then we can not determine X and Y, but we know that $X \neq Y$, so knowing one gives us the other. If we condition on $Z = 1$ it will look like X and Y are *inversely proportional*, because every time one of them is 1 the other will be 0. It thus looks like there is a causal relationship between X and Y, which is clearly wrong!

Our conclusion is therefore

> *Conditioning on a collider can* ***create*** *spurious correlations.*

We will always want to avoid conditioning on a collider.

5.4 Conditioning on Children

There is another problem we have to consider when conditioning on variables. We return to the problem of flipping two coins, but we add a variable which is an exact copy of Y.

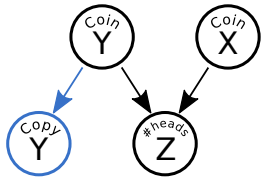
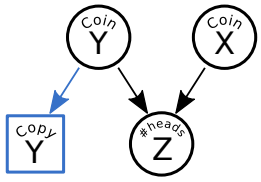
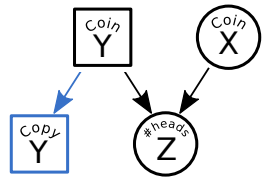


Figure 8: Another case of coin-flipping.
New variable is highlighted in blue.

We still want to determine the causal relationship between X, Y and Z. Let's see what happens it we condition on the copy of Y. We expect to conceptually make Y-copy a constant, but since this variable is identical to Y, then in practise we end up also conditioning on Y!



(a) What we think we get from conditioning on the copy of Y.



(b) What we actually get from conditioning on the copy of Y.

Thus we have ruined our experiment, because we cannot determine Z's dependence of Y if we keep Y constant.

Let's make another experiment where we introduce a die D[1]. We also introduce a "topscore"-variable T for X and D. This variable is 1 if we get heads with X *and* a 6 with D. Otherwise it is 0.
What happens if we condition on $T = 0$? Everytime $X = 0$ then nothing really happens, the relationship between Y, X and Z remains the same.

[1]If you didn't know: "dice" is plural :)

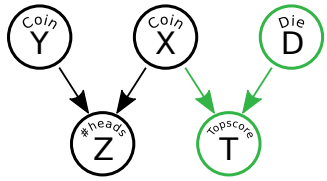


Figure 10: A die has been introduced.

But if $X = 1$ then there is a chance that we will remove that sample. It will therefore look like $X = 0$ more often than $X = 1$, which is not the case. So by conditioning on T we have *partly* conditioned on X, which is not what we wanted.

> *Conditioning on a child of a node X can partly or entirely condition on X as well. We generally want to avoid conditioning on children of the variables we are investigating*

For the coin-and-dice examples these concepts may seem trivial. But consider the situation where we are testing a new drug. Perhaps we want to disregard various types of blood-pressure, sugar-levels etc. and consider how the medicine works conditioned on these. But if any of these are actually *causal children* of the drug (for example if the drug affect blood-pressure), then we end up partly conditioning on the drug itself, which can make us draw incorrect conclusions.

> *Conditioning can be used to solve confounder problems, but should be handled with care. Avoid conditioning on a node you are investigating, or a descendent of a node you are investigating, or on a collider.*

Section 6

Causal Inference - In General

6.1 Inference

Induction inference (here simply called inference) is the process of learning from evidence. It is the opposite process of deduction, in which a learned model is applied to information to produce a conclusion.
If we know the atmospheric pressure and we know how our barometer works, then we can *deduce* the distribution of the barometer state. This seems obvious, because the only thing that influences the barometer is the pressure - knowing the pressure allows us to know the barometer state.
Usually we use barometers the other way around though. We *infer* the distribution of the pressure from looking at the barometer. We know the barometer may not be exact, but we believe the actual atmospheric pressure will be pretty close to what is shown on the barometer.

Inference is governed by Bayes Theorem, which underpins all machine learning[1]

Bayes Theorem: $P(x \mid y) = \dfrac{p(y \mid x)\, p(x)}{p(y)}$

When we apply Bayes theorem we use a known probability distribution $P(y \mid x)$, the *likelihood*, together with some additional information $P(x)$ (the *prior*) and $P(y)$, to compute $P(x \mid y)$, the *posterior* probability. Whe $P(x \mid y)$ is compute from $P(y \mid x)$, we sometimes refer to $P(y \mid x)$ as *forward*

[1] In fact it underpins ALL learning!

probabilities and to $P(x \mid y)$ as *inverse probabilities*.

When we do causal inference we make use of all our tools from probability theory, but we need more than that. The causal tools we need are interventions, randomized interventions and conditioning. We will use these to determine *interventional distributions* - if we know the interventional distributions, then we know the causal structure. From section 3.2 we know that we need to handle confounders as well as disambiguation in order to infer causal relationships. Furthermore we need to make sure we do not deteriorate our causal structure with the tools.

If we are for example interested in the causal relationship between X and Y in the graph on the right, how do we determine what tools to use?

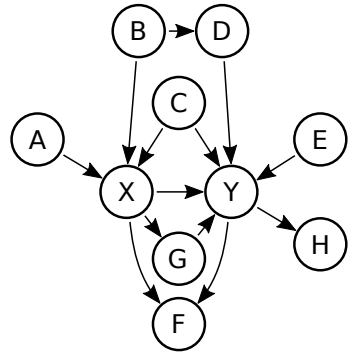
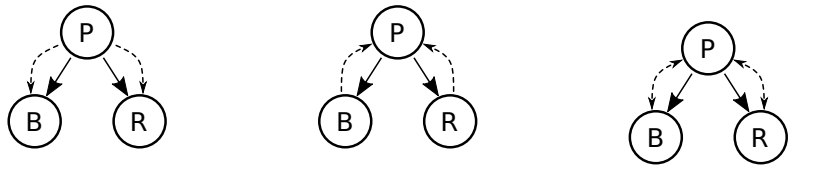


Figure 1: A big causal graph.

6.2 Information and Causal Graphs

One way of analysing causal graphs is from the perspective of information flow. If we know the value of a cause, then that tells us something about the effect from forward probabilities. Similarly if the know the effect then we know something about the cause from inverse probabilities.

If Figure 2 we have revisited the causal graph of the barometer problem.



(a) Forward probability information.

(b) Inverse probability information.

(c) Information flow.

Figure 2: Information flow in the causal model governing atmospheric pressure (P), barometer reading (B) and rainfall (R).

The nodes are still causal variables and the solid links remain causal relationships. Figure 2a shows how we can use the forward probability distributions $P(\text{barometer} \mid \text{pressure})$ and $P(\text{rain fall} \mid \text{pressure})$ to deliver information from pressure to the other variables. Figure 2b shows how we can use the inverse probability distributions $P(\text{pressure} \mid \text{barometer})$ and $P(\text{pressure} \mid \text{rain fall})$ to deliver information from rainfall and the barometer to atmospheric pressure. The total flow of information is therefore as shown in 2c.

We can even distribute information through the pressure node as well - which confirms our first experiments where we noticed that our barometer and the rainfall shared some information.

Thinking about information "flowing" through a causal graph can be a useful way of understanding a problem.

6.3 Causal Patterns - Revisited

We will now revisit the causal patterns defined in section 3.3 and discuss how information flows through them.

We really need some references for these concepts - and the Do-operator!

Chain Junction

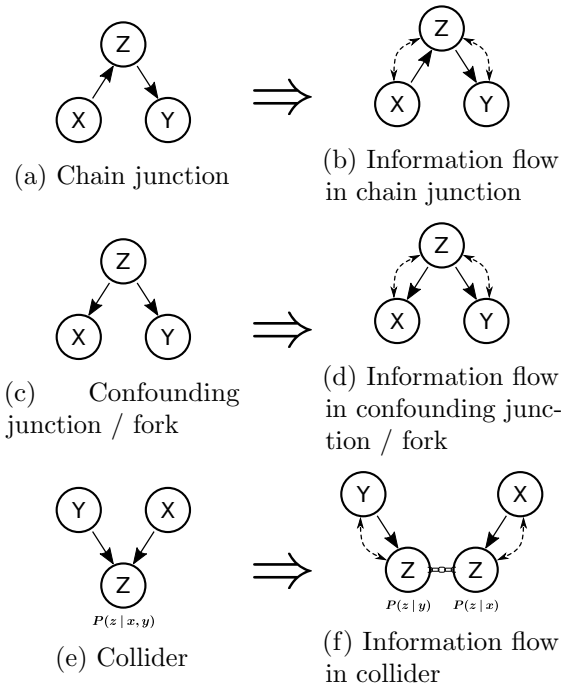
For the *chain junction* seen in Figure 3a we can use forward probabilities to propagate information from X to Z and onwards to Y. We can use inverse probabilities to infer information in the other direction, and so information runs freely through this graph (as shown in Figure 3b).
If we intervene or condition on Z, then conceptually we make it "constant". If it is constant then it carries no information. This will therefore block all information through Z.

Chain junctions lets information through, but can be blocked by intervention or conditioning.

Confounding Junction

For the *confounding junction / fork* seen in Figure 3c we can use inverse probabilities to infer information about Z from X and/or Y. If we can infer information about Z from X, then we can use forward probabilities to get information on Y - and vice versa. Therefore information also runs freely through this type of arrangement (Figure 3d).
Similarly to the chain junction, if we intervene or condition on Z we block of all information through Z. This is what we used when we conditioned on a confounder to help us determine a causal relationship.

Confounding junctions / forks lets information through, but can be blocked by intervention or conditioning.



(a) Chain junction

(b) Information flow in chain junction

(c) Confounding junction / fork

(d) Information flow in confounding junction / fork

(e) Collider

(f) Information flow in collider

Figure 3: Three important arrangements of causal variables.

Collider

We now get to the third arrangement in Figure 3e, the *collider*, which is a bit more tricky. If we know X and/or Y we can say something about Z. If we know Z we can say something about both X and Y. The catch is that we can no longer use X to say anything about Y! The colliding pattern *blocks* the information flow across the node.
One way to think of this is that the collider node holds information from both sides of the graph. If we know that $X = x$, then Z will follow the distribution $P(z \mid x)$, where Y has been marginalized because we do not know it. $P(z \mid x)$ therefore holds all the variance that Z gets from Y.

It is like the Z-node is split for each side of the graph, where one side holds the variability of being caused by Y and the other the variability of being

caused by X. In Figure 3f we have illustrated this by splitting the Z-node into two parts chained together. The left node follows the distribution $P(z \mid y)$, because it is on the side where Y is known but X is unknown and therefore marginalized. The right side similarly follows $P(z \mid x)$. If we know Z then we know both of these chained nodes and can therefore propagate information to both X and Y.

One peculiar thing about colliders is what we saw in section 5.3, conditioning n a collider then *opens up information*!

> *Colliders block information, but are opened up if you condition on them. They still block if intervened upon.*

6.4 Solving a Graph

Using these ideas of information flow, we can solve most causal graphs! Where solve means determining how to find causal relationships. For example consider again the graph on the right where we wish to determine the causal relationship between X and Y.

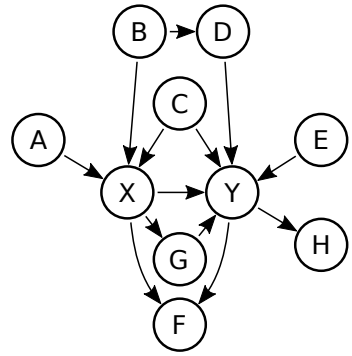


Figure 4: A difficult problem.

There are a lot of nodes, but let us draw all paths between X and Y where information is carried. The nodes A, E and H definitely do not carry any information between X and Y. Furthermore we know that colliders like F blocks information, so F is also not a problem. Below we have sloppily drawn the information paths between X and Y in blue
There are four paths. The BD-path and the C-path are confounders, which we need to handle. The remaining paths are the direct path and the path through G. We want to know how X causes Y and so both of these paths

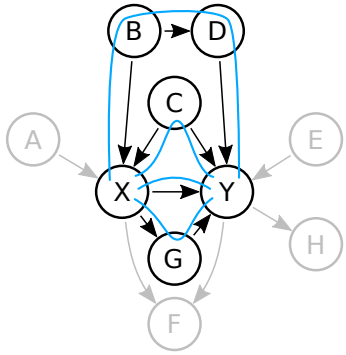
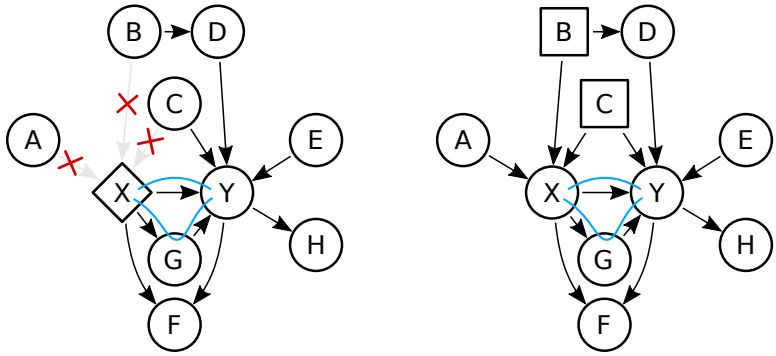


Figure 5: Information flow in the difficult problem.

are needed. If we want to get rid of the confounders we know that we can either condition on all of them or use intervention.



(a) Intervention solution. (b) Conditioning solution.

Figure 6: Solutions to the difficult problem.

If we made an experiment in which we intervened on X we could therefore test for the causal relationship between X and Y. We could also test for the relationships by gathering an observational dataset and conditioning on B and C. We can also conclude that data on A, E and H is not important for the problem. We now know what kind of experiment we should make, what data we should gather and how we should analyse it afterwards.

Using intervention, randomized intervention and conditioning we can block off information in causal graphs, so that the only information flow remaining is the one we wish to measure.

Section 7

Knowledge and Causation

In this section we discuss some aspects of causation which may not be intuitive at first.

7.1 Who Intervened?

If we want to infer causal relationships we will generally be making interventional experiments. But what if someone else makes an intervention? If another person makes an interventional experiment and hands you all data and information, then we might as well use those instead of doing our own experiment - no harm done. But what if you get hands on a dataset without knowing that some of the data is interventional? First of all we saw in section 4 that not having interventional samples put you at a disadvantage. Also you may believe that all the data is observational (because you have no reason to believe otherwise) and your conclusions will probably be wrong. The data does not represent the system you are analysing, because you are analysing a system without interventions.

Let us say you try to increase employees rating of the work environment, by buying a new coffee machine. At first you ask the employees to rate their day for one month. Then you install a new coffee machine to see if the next month's ratings are higher. Hopefully you will see an improvement and your experiment is a success.

What if someone else decided to celebrate the company's anniversary, by buying Starbucks for the whole crew for the first month of your experiment? In this case you have a system where two groups are both intervening

on the same variable; coffee quality. If your new coffee machine does not make better coffee than Starbucks, then you might get worse ratings in the second month, which completely opposes you hypothesis.

If you knew about the anniversary and the Starbucks coffee, then you could have planned your experiment differently. You could have collected ratings during the Starbucks-month and used that as the coffee-intervention - even though "you didn't do it". Afterwards you can decide whether a new coffee machine is worth it.

This illustrates that for interventions we also need to *know* that an intervention is going on. While the example is contrived there are many interesting multiagent systems (such as for example a country's economy) for which experiments are always messy and difficult to perform in isolation. Systems with many actors who all intervene differently at the same time, and each tries to learn from their own experiments.

A famous saying in science is by Isaac Newton *"If I have seen further it is by standing on the shoulders of Giants."*[1] He contributes some of his scientific progress to the people before him, who taught him, wrote books for him to read, and in general gave him the knowledge needed for him to do his work. This has an interesting perspective to causality and intervention. Scientists like Isaac Newton make interventional experiments in order to infer causal relationships about our world. They did that by settings up experiments, intervening on systems, collecting data and analysing this data. When we today base our understanding, science and work on their findings, we use their conclusions as our assumptions. Therefore *their interventional data becomes our causal knowledge.*

[1] Although "on the shoulders of giants" can be traced back to before Newton as well.

7.2 No Free Lunch Theorem

The *No Free Lunch Theorem* can be found in multiple versions, but is here defined as Reference

> **No Free Lunch Theorem:** Any two optimization algorithms are equivalent when their performance is averaged across all possible problems.

This theorem makes a very important foundation for machine learning and AI. It states that for an algorithm that works well on a set of problems, we can always find a similar set of problems with opposite characteristics, for which the algorithm works bad.

The most important consequence of the No Free Lunch Theorem is that we have to make some assumptions in order to learn something. We cannot learn everything only from data - we can not get the perfect learner for free. Luckily for most problems we are interested in there are some shared characteristics which we can use as assumptions. Thus we may (perhaps) be able to come up with algorithms that works well for all problems of human interest.

In causal inference we also have to make some assumptions. As noted in section 2.2, whenever we specify a causal graph we are automatically assuming that everything else in the world does not matter. We restrict our focus to only be about the variables of our model. Similarly in a causal model we make a set of links, which implements our assumptions about the relationships between the variables. Again all links that we do not include are automatically assumed non-existing, which could potentially be wrong.

Therefore in order to do any sort of causal inference we also have to make some assumptions, similarly to the No Free Lunch Theorem of learning. No matter how much and how good our data is, at some point we need to make some assumptions, which currently means we need a human to think about the problem.

7.3 Ladder of Causation

Reference

The Ladder of Causation is a conceptual model by Judea Pearl, whose illustration is shown in figure 1. It distinguished three types of reasoning based on the questions they are able to answer.

The lowest level of causation is called the association level (seeing) - the level of modern AI and of most statistical models. It is able to observe data and make conclusions like, *if this happened, will this other thing also happens?*. It is all about learning probability distributions and conditioning on what you see. Animals also do this; if there is usually food here, there will likely be food again.

The second level of causation is the level of intervention (doing), which is what we have been dealing with in this note. Intervention allows the learner to learn causal models, which in turn can answer questions like *what happens if I do this?* The association level only conditions on observational data, while the association level conditions on interventional data.

The next level of causation concerns imagining and is what section 10 will be dealing with. It considers questions like *what if X had been different?*, in which we *imagine* a case where X is a different value that what it is. It also concerns questions like *why did this happen?*. What *caused* something to happen. It allows models to explain themselves and explain decisions.

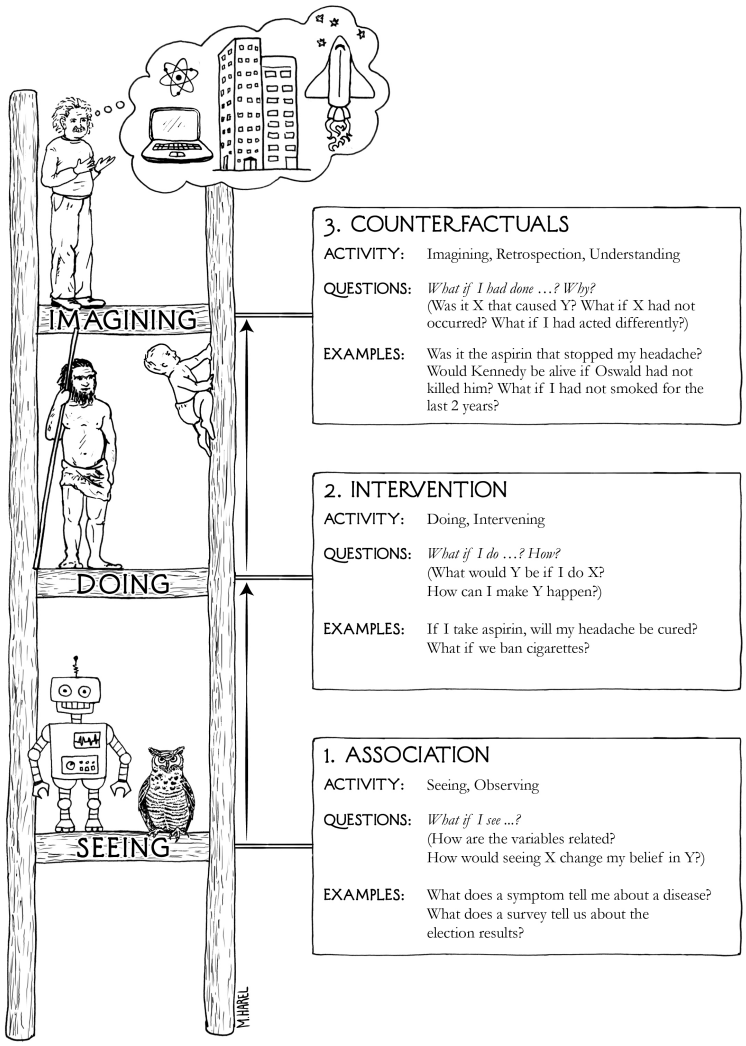


Figure 1: The Ladder of Causation

Section 8

Causality and AI

This section is to ensure a general understanding of mathematical models (specifically machine learning) and how they relate to causal models. It is useful to solidify these concepts before advancing our causal models.

8.1 Mathematics, Science and Machine Learning

The exact definitions of mathematics and science is a controversial topic, but here we discuss some general properties of the fields and how machine learning relates.

In science we attempt to learn about the physical universe. We do so by making models (theories/hypotheses) of how we believe the universe works. We then *test* these models in the universe through experiments, which is the cornerstone of science. With experiments we control systems we want to analyse, we intervene on variables we want to examine, we collect data on outcomes, and we use the data to verify/falsify the models. Models that are verified we keep and improve, while models that are falsified we stop using. This makes science *empirical*, as we use experiments and data to decide what to believe. Furthermore scientific models are *causal*. They help us predict future events, while also explain their causes.

Mathematics is fundamentally quite different. In mathematics we *prove* our ideas. We do not rely on empirical experiments and data in order to verify/falsify ideas. Newly proposed ideas remain *conjectures* until they

are proven or disproven. If disproven we disregard them, while if we prove them we consider them theorems or lemmas. Mathematics is therefore not empirical and also not inherently causal.

While mathematics is not build on experiments, it is common to make experiments to get a sense of what might be provable. If you have a conjecture that you believe to be true, then you could experiment by testing it a ton of times and see if it always works. If it does always work, then it seem plausible that you may be able to prove it. Still only when proven do we consider it a mathematical theorem/lemma.
Furthermore, mathematics is perhaps the most useful "tool" know to humanity. As we prove more and more theorems in mathematics they become tools that we can use in science and all kinds of other fields. Therefore basically all experiments will be *using* mathematics, but not *extending* mathematics.

One field of mathematics that is crucial to science is statistics. In statistics we prove properties of statistical distributions/models. We use these models as hypotheses for how data behaves. If we assume a distribution/model for a data process, we can make provable conclusions about that process. Also we can utilize proven tests for determining whether a distribution/model is plausibly *correct* for the given problem.

In machine learning we have given up on finding the *true* model of the world. We have come to realize that it may be too time-consuming to correctly model every detail of the world. Instead we make very flexible models, which can hypothetically learn anything, and then *train* them for their purpose using data. We therefore make the models do the inference for us, which saves us a lot of time. This is why some defined machine learning as "improving from experience without being explicitly programmed". We do not prove the correctness of the model as we do in mathematics. Machine learning could thus be considered *empirical mathematics*, where we make mathematical models, but verify them empirically.

In science *we* (people) make *explicit causal models* about the universe. In machine learning *computers* makes *automatic models* for *practical purposes*.

The formalisation of causality in this document introduces ideas which could enable machine learning to also make causal models. In this case machine learning will be able to make automated science - it can help us discover causal models in huge datasets and problems too complicated for us to grasp.

8.2 Models and Physics - an example

Figure and rewrite

Consider the following scenario. We have a set of rubber balls that we roll from a starting point out onto the floor of a gym. The balls have the same weight, but vary in size, and we roll them with varying angle and force. We now collect a dataset, where we provide the initial speed, the size, the angle of the roll, as well as the final position of each ball we roll. We train a fancy machine learning model, say a neural network, to predict the final position in (x, y)-coordinates, based on the other information. Sometimes the angle of the ball will deviate a bit and the floor friction will not be completely constant, so the balls are not completely deterministic. Still, after 200 throws the neural network predicts the final position with very high precision.

We also write out a formula for the final position, based on our understanding of physics. Using the teachings of physics for rigid bodies, we can derive such a formula. We only need to determine a couple of constants for the friction of the floor, air resistance etc. which we can fit using our dataset. This formula is also capable of predicting the final position of the balls quite precisely.

Now we get another type of ball which is of the same varying sizes, but weighs more. Weighing more will give the ball more initial, kinetic energy for a given initial speed. If we directly predict with both models, they both predict incorrectly. But in the physical model we can easily incorporate the new weight. With almost no effort we can make the physical model handle the new situation and predict correctly again! What do we do with the neural network? We could try to explain to it that the balls have changed

weight, but how exactly do we do that? We would probable have to roll quite a few of the new balls, in order to get data to retrain the neural network.

How about if we go outside and roll balls across a parking lot with asphalt. Again both models are off, because the friction of the asphalt is completely different from that in the gym. We could get more data for refitting the two models. Alternatively we could look up information about friction of asphalt and apply to the physical model. Using this approach we are again able to manipulate the physical model to handle new circumstances, without getting new data. And perhaps even more interestingly, we have used *others experience* (the friction info on asphalt) for using in our own problem.

So what is the big difference between two two models? Well, the first thing to note is that we broke the most important rule of machine learning: the test data *must* come from the same distribution as the training data! This assumption underpins everything machine learning builds on, and we broke it when we tested the models on different systems than what we started with. But physical models *can* help us in new situations. With physical models we can even compute how systems behave on Mars, even though the "training data" we used to learn about physics is not from Mars.
The second thing to note is that technically the neural network is wrong. No matter how great neural networks are, we have to face the fact that a ball rolling on a surface *are not neural networks*. They are mechanical systems following the laws of physics. In fact, the best we can hope for when using the neural network, is that it *learns the physical model*. The situation in which the neural network will be most precise, is the situation where it *contains* the physical model of the balls. We did make physical model to predict position of the balls, but on the other hand the balls *must follow the laws of physics*.

The physical model is a *causal* model. It does not simply try to model the distribution of final positions, but actually models the physics of the system, and if our understanding of physics is correct, then a ball rolling across a surface *is that modelled system*. The causal model is capable adapting to interventions in the system, as well as answer other questions

like, why did the ball not roll very far? - because of a higher friction of asphalt.

8.3 Machine Learning Recap

Here we make a quick recap of models used in machine learning (and other branches too). It's useful for expanding to more detailed causal models. While not all previous sections are needed to understand this section, it is a good ideas to read sections 8.1 and 8.2. For the rest of this section we use the following notation: x is inputs, y is a real-valued output and c is a discrete-valued output. $P(c)$ denotes a probability of discrete event c, while $p(y)$ denotes the probability-density of real-value y.

In AI and ML we commonly distinguish between two types of problems:

Classification: Models the relationship from inputs to a **discrete**-valued output.
Regression: Models the relationship from inputs to a **real**-valued output.

Basic Models

The most basic version of these two methods chooses the most probable/-expected value of the problem. The *pure discriminative model* determine the most probable class for an inputs

$$\arg\max_{c} \; P(c \mid x). \tag{1}$$

The model makes discrete decisions, but does not tell us anything about how confident it is. The basic regression model computes the expectation of the output variable based on the inputs

$$\mathbb{E}[y \mid x] = \int y \cdot p(y \mid x) \, \mathrm{d}x. \tag{2}$$

This model provides one good estimate of the output, but again does not model the confidence of the prediction.

Figure 1 illustrates the basic models. The classifier divides the space into parts according to class and the regression model fits a line through the data.

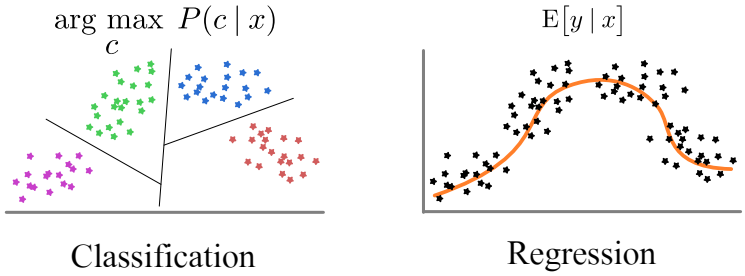


Figure 1: Basic models

Probabilistic Models

If we take more probabilistic approach to classification get a *conditional discriminative model*, which models the probability mass of all classes

$$P(c \mid x). \tag{3}$$

Using this type of model we can easily provide the same predictions as with the pure discriminative model, but we also know how confident the model is and how likely all other class are.
Similarly by upgrading the regression model we get the conditional density of the output

$$p(y \mid x). \tag{4}$$

Again we can use this model to select the most probable value, but we can also say something about the much the output may vary from this estimate.

The probabilistic models are illustrated in Figure 2. The classifier makes boundaries between classes with soft transitions to illustrates areas of uncertainty. Notice that it is quite arbitrary what the classifier decides to do in regions with no data (the centre part). The regression model again

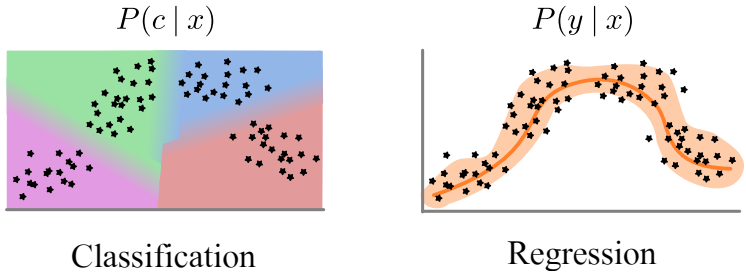


Figure 2: Probabilistic models

fits a line, but this time with some notion of uncertainty.

Knowing the confidence and variance of outputs is crucial for many tasks, and to be honest there is very few reasons to settle for a basic model today, as we can usually make probabilistic versions.

Generative Models

We now get to the final level of statistical models

> **Generative Model:** A generative model models both inputs and outputs - it models the whole system. Crucially a generative model is capable of generating new samples.

The generative model provides the joint probability of inputs and output

$$P(c, x) \text{ or } p(y, x). \tag{5}$$

We can use this model to provide conditional probabilities (for both classes and regression targets) and it becomes a very natural starting point for designing conditional discriminative models. Since we model the joint distribution, the difference between inputs and outputs become less relevant. The difference between regression and classification also becomes blurry as a regression model with a discrete input could be identical to a classification model.

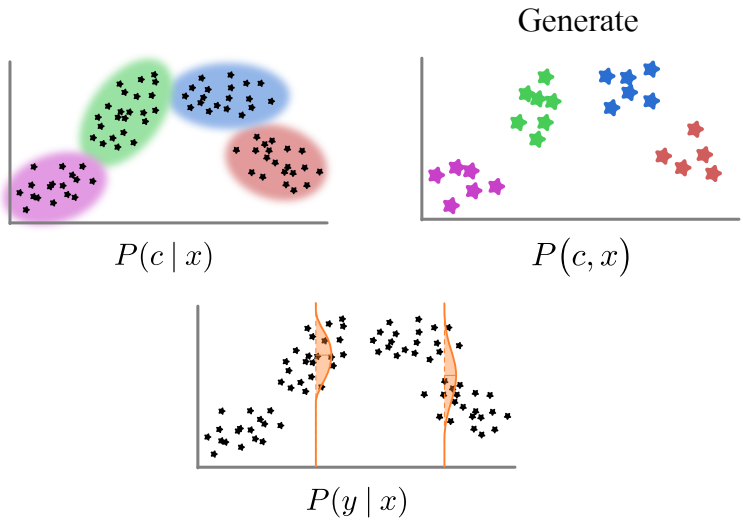


Figure 3: Generative models

Figure 3 illustrates the generative models. On the left we classify by computing the class-conditional probabilities. We are also able to determine areas with no information (centre region), as we know where the input distributions are limited. In a sense we can do outlier detection, and be warned if data is from a region for which the model has limited information. In the middle we generate data using the joint distribution.
On the right we condition on the input data, which provides mean and variance estimates for regression purposes.

Section 9

Full Causal Models

Up until now we have considered probability distributions for various problems and augmented them with graphical causal models. In this section we are going to expand these tools and make the first direct connection to artificial intelligence (AI) and machine learning (ML).

9.1 Generative Process Models

We now upgrade out generative models.

> **Generative Process Models:** A generative model which models the *process* of the data and not only the distribution.

The difference between a generative model and a generative process model is a bit vague and weird, but in nutshell there can be many different generative models that all correctly models a problem, but there is only one generative process model. The generative process model is the *correct* generative model, which is correct even when we intervene and which can extrapolate to new systems. We can therefore use such a model for causality.

The table below recaps the taxonomy of models

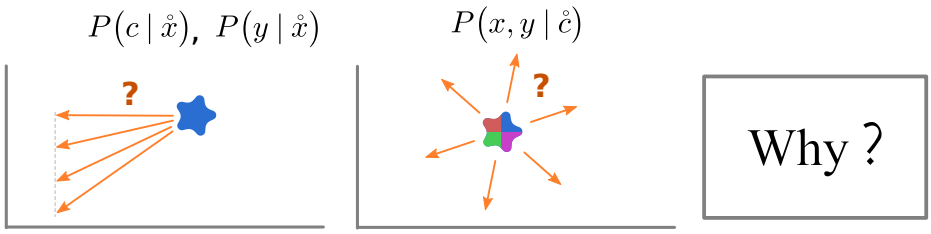


Figure 1: Generative process models. We can use these models to ask what happens when the input is intervened on (left), when the class is intervened on (center), as well as *why* a particular outcome was observed.

Level	Model type	Modelled probability/density	Questions and usages
1	Basic	$\arg\max_c P(c \mid x)$, $\mathbb{E}[y \mid x]$	Single prediction.
2	Probabilistic	$P(c \mid x)$, $p(y \mid x)$	Prediction with uncertainty.
3	Generative	$P(c, x)$, $p(y, x)$	Generation and outlier warning.
4	Generative Process	$P(c, x)^\star$, $p(y, x)^\star$	What happens if we force x/c? What if x/c had been different? Why did this happen?

9.2 Structural Equation Models

In causal literature we use a specific type of generative process model called a Structured Equation Model - SEM. An SEM is a set of ordered assignments of the following form

$$\begin{aligned} a &\leftarrow f_a() \\ b &\leftarrow f_b(a) \\ c &\leftarrow f_c(a, b) \\ &\vdots \end{aligned}$$

The model first determines a from some stochastic function $f_a()$. It then computes b using another stochastic function $f_b(a)$, so that uses a as input. The formula for c uses a and b and so on. Note that while each variable has access to the previous ones, they do not *have* to use them. The ordered set of assignments creates dependencies where each variable depends on ealier ones.

Let's take an example

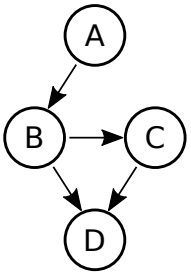
$$A \leftarrow N_A, \qquad N_A \sim \mathcal{N}(0,1) \tag{1}$$

$$B \leftarrow \frac{1}{2} \cdot A + N_B, \qquad N_B \sim \mathcal{N}(0,1) \tag{2}$$

$$C \leftarrow B + N_C, \qquad N_C \sim \mathcal{N}(0,1) \tag{3}$$

$$D \leftarrow B \cdot C + \frac{1}{3} \cdot N_D, \qquad N_D \sim \mathcal{N}(0,1). \tag{4}$$

We see that A depends on no other variable. B depends on A, C depends on B, and D depends on both B and C. Furthermore the stochasticity of the variables is represented by a stochastic variable for each (which in this example are all normally distributed).



For any SEM, we can graph the dependencies of the assignments. If the SEM models a causal system, then

this graph is the causal graph of the system - back to
things we know about! If we graph the above example we get the graph on the right, where we see the expected dependencies.

A causal graph is a simplified SEM, where we know which parameters depend on which, but we do not know *how* they depend. With the SEM we know all the details.

Using the above SEM with a programming language with sampling libraries (such as Python) we can sample A - we just need to draw samples from a normal distribution. Similarly we can sample from B by first sampling from A and computing B with samples from N_B's normal distribution. We can continue to do this and sample the whole process - we have a generative process model.

We can use the causal graph to do all the analyses we have previously discussed, but can also go a bit deeper. When asked how B depends on A we can actually specify *how much* B depends on A, because we can compute how much of B's variance that comes from A (one third). Similarly we can say that two thirds of B's variance is *independent* of A. SEM's gives a much more detailed view of the causal process.

Section 10

Counterfactuals

Turn all the important points into point-boxes.

10.1 A Counterfactual Question

We will now discuss a different type of causal question; "what if"-questions. For that purpose we start with an example.

> **Counterfactual:** A counterfactual question is one where we seek information about how the world would have been different given some specific change. For example; in a situation where X *did* happened, what if X had *not* happened?

At 23:40 on the 14^{th} of April 1912, the RMS Titanic hit an iceberg which ultimately sank the ship after 2 hours and 40 minutes. The iceberg was spotted by Frederick Fleet, but only too late for the ship to navigate around it. Let us consider the following question; *if the RMS Titanic had started its voyage one day later, would it have made the crossing?*

At first we may get some information on the event. We know the date of the iceberg collision, and could probably get information about the weather and the preceding winter, perhaps even estimate the amount of ice in the waters based on the tales from the survivors. Let's gather all this information and put it into a variable U called utility information. Furthermore we know that the Titanic left from Southampton on the 10^{th} of April and that it did sink; $D = 10$ and S. Let's first try to frame the

problem mathematically. We want to determine the probability that the ship did not sink, given that it left the next day, while also conditioning on the known information U as well as the fact that it left on the 10^{th} and it did sink

$$P(\neg S \mid D = 11, U = u, D = 10, S). \tag{1}$$

Here we see the first obvious problem, which is what makes counterfactuals quite different. We are searching the probability of *not sinking* $\neg S$, given that we know that the *ship sank* S. Furthermore we are conditioning on the date being the 11^{th}, $D = 11$, while also conditioning on the known fact that the ship did actually leave on the 10^{th}, $D = 10$.

We first need to get rid of this confusion. Some variables hold information that we know. These are just like any other stochastic variable and we keep them as is. The other variables hold *counterfactual* information. They either hold the query that we wish to know about (in our case $\neg S$) or causal changes that we wish to investigate (in our case $D = 11$). We will here mark all counterfactual variables by a black dot $\bullet$, so that

$$P(\neg \dot{S} \mid D \dot{\leftarrow} 11, U = u, D = 10, S). \tag{2}$$

A common interpretation of counterfactuals is that we consider a parallel universe, where everything is *exactly* the same, except for the counterfactual variables. We therefore ask; in a parallel universe where U (all auxiliary information) is the same, what is the probability of $\neg \dot{S}$ given $D \dot{\leftarrow} 11$, when we know that in *our* universe we observed $D = 10$ and S.

10.2 Party Time!

The Titanic is an example of how important and relevant counterfactual questions can be. In order to learn about this topic though, we start with a simpler example; a party.

Bob is not too fond of parties but really likes Alice. If Alice is at the party, then Bob usually goes unless they are having a disagreement. Carl really likes partying, but Alice is his ex-girlfriend and sometimes he will avoid the party if she is there. Finally Bob and Carl cannot stand each other and if they are both at a party, they usually fight. Sometimes they even fight without being at a party.

We wrap this knowledge into the following causal model, where we have also specified the probabilities. A denotes Alice being at the party, B means Bob is at the party, C Carl and F means that there is a fight.

	value	if	$P(N_x)$
$A \leftarrow$	0	$N_A = 0$	0.40
	1	$N_A = 1$	0.60
$B \leftarrow$	0	$N_B = 0$	0.07
	A	$N_B = 1$	0.90
	$1 - A$	$N_B = 2$	0.03
$C \leftarrow$	0	$N_C = 0$	0.05
	$1 - A$	$N_C = 1$	0.85
	1	$N_C = 2$	0.10
$F \leftarrow$	0	$N_F = 0$	0.05
	$C \cdot B$	$N_F = 1$	0.90
	$C \cdot B + (1 - C) \cdot (1 - B)$	$N_F = 2$	0.05

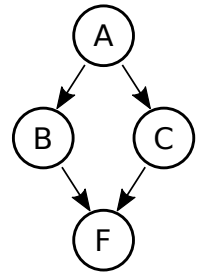


Figure 1: Causal graph for the party.

We have good intuitive explanations for each of the variables. N_A contains the variance of whether Alice attends the party. If $N_B = 0$ then Bob can't attend the party, if $N_B = 1$ then Bob will go if Alice goes, and if $N_B = 2$ then Bob will do the opposite of Alice, because they are having a

disagreement that night. With Carls attendance we see that Carl can't go to the party when $N_C = 0$, will go no matter what if $N_C = 2$, and might go but will avoid Alice if $N_C = 1$. N_F basically sets the mood of the party - either there is not fight, or Carl and Bob will fight if they are at the party, or they will fight if they are either both at the party or both somewhere else.

Before we start analysing this problem we note that we can expand on the causal graph in a useful way. In the graph on the right we have also made nodes for the noise/variance variables. Each noise variable is hidden because we cannot observe it. For example we can check whether Alice is at the party, but we do not know what specific factors made her come to the party. Furthermore we have modelled the problem in a way where all variance of each variable is collected in these variables, when conditioned on ancestors: if we know that Alice is at the party then all remaining variance of Bob's attendance is in N_B.

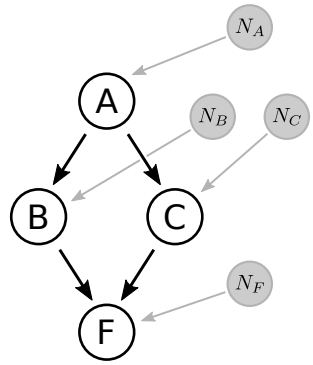


Figure 2: Causal graph for the party with noise/variance variables.

A Counterfactual Party

Last weekend there was a big party which Bob did not attend. We wish to determine the probability of a fight if Bob *had* gone

$$P(\dot{F} \mid \dot{B}, \neg B). \tag{3}$$

Here we have noted the only known information; Bob was not at the party $\neg B$. We are thinking of a parallel world where we make Bob go $\dot{B}$ and see if there is a fight $\dot{F}$.

First we make a new causal graph where we duplicate all the nodes: one for the real universe and one for the counterfactual universe.

All observable nodes of the two universes have their own nodes, because we can make analyses where we inspect each side of the variables. Conversely we cannot observe any of the hidden variables, but we assume that everything we do not touch is identical for the two universes, so they only have one node each.

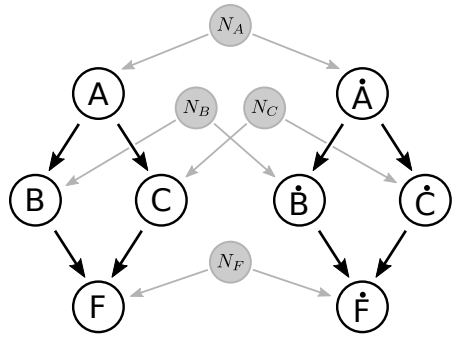


Figure 3: Causal graph for the party with counterfactual variables.

Now we can return to the query for $P(\dot{F} \mid \dot{B}, \neg B)$. We know that $B = 0$ in our universe. In the counterfactual we *intervene* to make $B \dot{\leftarrow} 1$. Finally in our situation A, F and C are hidden. The causal graph that we are now analysing looks like this
We have one constant node, one interventional node that we forced ourselves, and the remaining are hidden nodes. N_B is no longer relevant as both its children are constant/intervened on.

We wish to determine the probability $P(\dot{F} \mid \dot{B}, \neg B)$, for which the variables A and C are marginalized. We therefore need sum the probability of F, given we force $\dot{B}$, and conditioned on all possible values of A and C. The

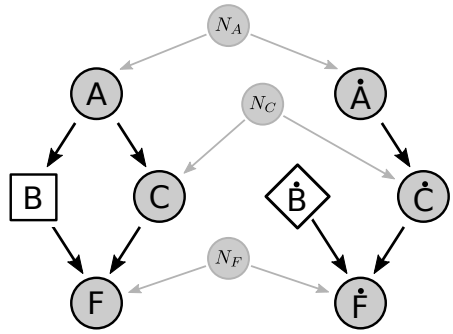


Figure 4: Counterfactual query graph for the party.

computation becomes

$$P(\dot{F} \mid \dot{B}, \neg B) = \sum_{a=0}^{1} \sum_{c=0}^{1} P(\dot{F} \mid \dot{B}, C = c, A = a)\, P(C = c, A = a \mid \neg B). \tag{4}$$

The fight depends on Carl, and Carl depends on Alice, so let's start with Alice. We compute the probability of Alice being at the party with inverse probabilities (Bayes Theorem)

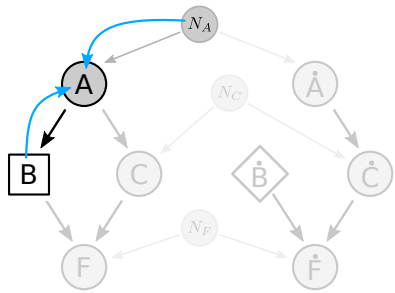
$$P(A \mid \neg B) = \frac{P(\neg B \mid A)\, P(A)}{P(\neg B \mid A)\, P(A) + P(\neg B \mid \neg A)\, P(\neg A)} \tag{5}$$

$$= \frac{(0.07 + 0.03) \cdot 0.6}{(0.07 + 0.03) \cdot 0.6 + (0.07 + 0.90) \cdot 0.4} \approx 0.13, \tag{6}$$

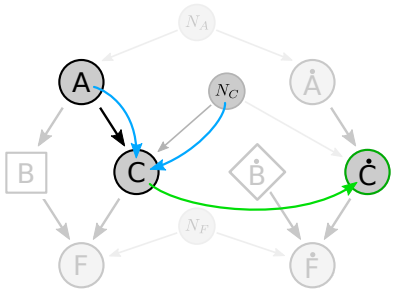
$$P(\neg A \mid \neg B) \approx 1 - 0.093 \approx 0.87 \tag{7}$$

So Alice is most likely not at the party. We have illustrated the flow of information in Figure 5a. We know the constant B and we know the distribution of N_A, which we can use to infer the distribution of A.

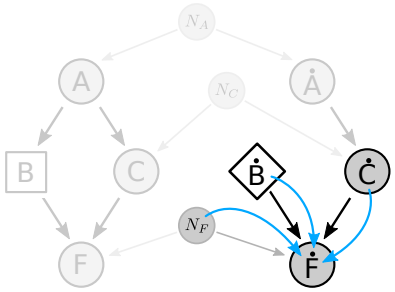
Now that we know the distribution of Alice, we can determine the distribution of Carl. In Figure 5b we use the information about A and distribution of N_C to infer the distribution of C. This is done by first getting the



(a) Determining distribution of Alice A.



(b) Determining distribution of Carl C and therefore also $\dot{C}$.



(c) Determining the distribution of F.

Figure 5: Flow of information for determining the counterfactual probability $P(\dot{F} \mid \dot{B}, \neg B)$.

conditionals

$$P(C \mid A) = 0.10 \qquad P(C \mid \neg A) = 0.95 \tag{8}$$

and then computing the probability of Carls attendance

$$P(C \mid \neg B) = \sum_{a=0}^{1} P(C \mid A = a)\, P(A = a \mid \neg B) \approx 0.10 \cdot 0.13 + 0.95 \cdot 0.87 \approx 0.84, \tag{9}$$

$$P(\neg C \mid \neg B) \approx 1 - 0.84 \approx 0.16. \tag{10}$$

We see that Carl is most likely at the party if Bob is not. Also, since we know that $C = \dot{C}$, we can transfer the knowledge about C to its counterfactual twin.

Now we know the distribution of $\dot{C}$, and we already knew the distribution of N_F and the forced $\dot{B}$. This allows us to finally consider the probability of a fight F. Computing the counterfactual probability of a fight is seen in Figure 5c and is done by

$$\begin{aligned} &P(F \mid \dot{B}, \neg B) && (11)\\ &\quad = P(F \mid \dot{B}, C)\, P(C \mid \neg B) + P(F \mid \dot{B}, \neg C)\, P(\neg C \mid \neg B) && (12)\\ &\quad = 0.95 \cdot 0.84 + 0.00 \cdot 0.16 = 0.79. && (13) \end{aligned}$$

We can therefore conclude, that if Bob was not at the party, then there most likely would have been a fight between him and Carl, if we forced him to go.

10.3 Counterfactuals in General

We will now analyse the concept of counterfactual reasoning from a more generic point of view. Counterfactuals is all about hypothesising about the effect of making a specific change to a sample. In contrast, conditional and interventional distributions are about systems, and how they behave under conditioning and intervention. Counterfactuals are also about systems, but they apply to a *specific sample*.

If we gather all variables that we make counterfactual interventions on into X, and gather all variables that we are interested into Y, then a generic graph of the causal system is seen below.

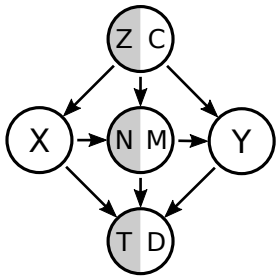


Figure 6: Generic causal graph investigating counterfactuals with X and Y.

We know that there is a causal relationships directed from X to Y. We have some mediators M between the two, some confounders C, and some descendants D. In counterfactuals it is quite crucial how much you know about the system. We have therefore split the confounders into visible confounders C and hidden confounders Z, split the mediators into visible M and hidden N, as well as split the descendants into visible D and hidden T. Furthermore note that it may be a direct link between X and Y and between Z/C and T/D, but we simply let these run through the mediator node for simplicity.

Since we are doing counterfactuals we assume that we know the generative process model - so there is not ambiguity about cause and effect etc. and

we now all relationships. Now we observe a sample $(x_0, y_0, c_0, m_0, d_0)$ and ask, what is the probability that we would have observed y_1 instead, given that we had intervened and forced x_1

$$P\Big(\dot{Y} = y_1 \;\Big|\; X \dot{\leftarrow} x_1, x_0, y_0, c_0, m_0, d_0\Big). \tag{14}$$

First thing to note is that we keep all the known auxiliary information in place (c_0, m_0, d_0) and also condition on the true values of the counterfactual variables (x_0, y_0). Knowing more about the system (for example moving variables from Z to C) with allow us to compute a more precise counterfactual probability, while loosing information (moving variables from C to Z) will make the counterfactual probability less precise.

Hidden Variables

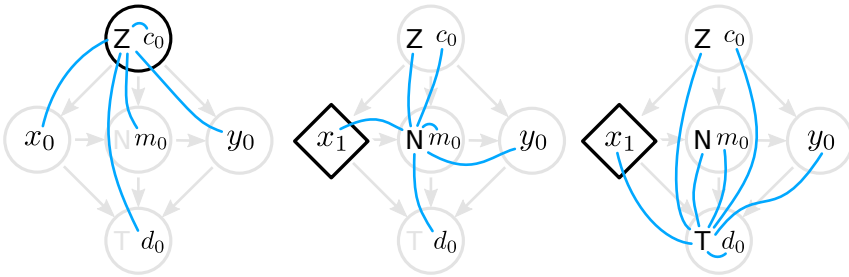
For predicting Y we need to figure out how the hidden variables might have behaved in the given situation. We thus inspect the distribution for Z, N and T. We know that they produced the sample $(x_0, y_0, c_0, m_0, d_0)$, and so their distributions should be conditioned on these

$$P(z, n, t \mid x_0, y_0, c_0, m_0, d_0) \tag{15}$$

which will involve forward probabilities and Bayes theorem for computing backwards probabilities. Also we are now performing a counterfactual intervention on $\dot{X}$ which may change some of the hidden variables.

We first determine the probability of Z - the highest, hidden node. This distribution is conditioned on $(x_0, y_0, c_0, m_0, d_0)$, because it help produce that outcome, which we illustrate in Figure 7a. It is not conditioned on $\dot{x_1}$, because Z is not affected by X.
In Figure 7b we move down to N. Here we replace x_0 by $\dot{x_1}$, because the value of N did not *cause* x_0, but is itself *caused* by the new value $\dot{x_1}$. The distribution of N is also conditioned on the value of its ancestor Z, whose distribution we have just determined, as well as all the other known information.

Finally in Figure 7c we consider the distribution of T. We condition the distribution of T on z, n, $\dot{x}$ and k, because it is the descendant of all these



(a) Prediction of Z (b) Prediction of N (c) Prediction of T

Figure 7: Information used for prediction of hidden variables.

variables.

We let $k = (y_0, c_0, m_0, d_0)$ for reducing clutter and conclude that

$$P(z, n, t \mid \dot{x_1}, x_0, k) = \underbrace{P(t \mid \dot{x_1}, k, z, n)}_{\text{lowest node}} \underbrace{P(n \mid \dot{x_1}, k), z}_{\text{center node}} \underbrace{P(z \mid x_0, k)}_{\text{highest node}}. \quad (16)$$

Prediction

If we know all variables, then the probability of y depends only on its ancestors (not D and T)

$$P(y \mid x, c, z, m, n, d, t) = P(y \mid x, c, z, m, n), \quad (17)$$

which can be computed directly from the generative process model.

We can now analyse the probability in (14) by marginalizing over all

unknown variables Z, N and T

$$P\Big(\overset{\bullet}{Y} = y_1 \;\Big|\; X \overset{\bullet}{\leftarrow} x_1, x_0, y_0, c_0, m_0, d_0\Big) \tag{18}$$

$$= \sum_{znt} \underbrace{P(y_1 \mid \overset{\bullet}{x_1}, c_0, m_0, d_0, z, n, t)}_{\substack{\text{probability of } y_1 \text{ conditioned} \\ \text{on all variables}}} \underbrace{P(z, n, t \mid \overset{\bullet}{x_1}, x_0, y_0, c_0, m_0, d_0)}_{\text{probability of hidden variables}}$$

$$= \sum_{znt} \underbrace{P(y_1 \mid \overset{\bullet}{x_1}, c_0, m_0, d_0, z, n, t)}_{Y} \underbrace{P(t \mid \overset{\bullet}{x_1}, k, z, n)}_{T} \underbrace{P(n \mid \overset{\bullet}{x_1}, k, z)}_{N} \underbrace{P(z \mid x_0, k)}_{Z} \qquad [(16)]$$

$$= \sum_{zn} \underbrace{P(y_1 \mid \overset{\bullet}{x_1}, c_0, m_0, z, n)}_{Y} \underbrace{P(n \mid \overset{\bullet}{x_1}, k, z)}_{N} \underbrace{P(z \mid x_0, k)}_{Z} \qquad [(17)]$$

$$= \sum_{zn} \underbrace{P(y_1 \mid \overset{\bullet}{x_1}, c_0, m_0, z, n)}_{Y} \underbrace{P(n \mid \overset{\bullet}{x_1}, y_0, c_0, m_0, d_0, z)}_{N} \underbrace{P(z \mid x_0, y_0, c_0, m_0, d_0)}_{Z}. \qquad [\text{k}]$$

There are some important things to note here. First of all, the final computation of the probability of Y is not special. $P(y_1 \mid x_1, c_0, m_0, z, n)$ is simply evaluating our model for y_1 conditioned on (x_1, c_0, m_0, z, n). Secondly the *hidden descendants* T are not used in any way and are irrelevant for the analysis. The only relevant descendants are the visible ones, because they provide information about the states of the other variables (so we included d_0 in k). Thirdly the ancestors of X uses x_0, while the descendants of X uses $\overset{\bullet}{x_1}$. This is because Z helped create x_0, while the descendants are affected by the counterfactual intervention on X.

Variants of the Counterfactual Question

Say there is a fire in a building, which you have just left as part of the fire drill. You could ask "would the fire have been smaller if I had closed the basement window?" (do to less oxygen for the fire). It might be that someone else already closed it, in which case it would not have made a difference. If no one had the time to close that window, then it probably would have made a difference. This is a situation where X is hidden and

we simply wish to know what the probability of Y is given that we *force* X.

Consider another situation where you just left the voting booth of an election. The results won't be released in several hours, but still you wonder whether it would have made a difference if you hadn't voted. This is a counterfactual question where you know X, but you don't know Y. Also you should always vote.

Thus it is possible to make counterfactuals where you don't know X and/or Y. In this case these variables won't appear in the conditionals of (18).

10.4 Counterfactuals and Interventions

I think this can improved with respect to marginalizing X

In the previous example we knew $(x_0, y_0, c_0, m_0, d_0)$. Let's consider what happens if we know less information. First we eliminate x_0 and y_0. These no longer appear in the conditionals in (18) and our probability become a bit more imprecise for the specific situation.

Now say we don't know any of the confounders (all of C is moved in to Z), which makes the probability more imprecise again. We can continue to move all of M in to N, and all of D into T. Now we do not know about any variables of the situation and the probability becomes

$$P\Big(\dot{Y} = y_1 \;\Big|\; X \overset{\bullet}{\leftarrow} x_1\Big) = \sum_{zn} P(y_1 \mid \dot{x_1}, z, n)\, P(n \mid \dot{x_1}, z)\, P(z) \tag{19}$$

$$= p(y_1 \mid X \overset{\circ}{\leftarrow} x_1), \tag{20}$$

which can be verified by inspecting (11). Thus the interventional distribution is the counterfactual distribution where we do not know anything! Or phrased differently, the interventional distribution is the expected counterfactual distribution over the observable population.

This makes intuitive sense: the interventional distribution evaluates the probability of an event given that we force on a variable, without knowing

anything else about the system. A counterfactual can therefore be considered a conditional, interventional distribution, where we intervene while conditioning on a *specific sample*.

Section 11

Perspectives on Causality

11.1 Why?

Yeah, Why?

The word *why* is a bit overloaded. At the same time, it is the cornerstone of some of the most important scientific questions, and so important in causal analysis that Judea Pearl named his book *The Book of Why*.

In order to specify what kinds of questions we will be concerned with, we first note the following three types of why-questions.

1. Why does event X happen?
2. Why does event X cause Y?
3. Why are we doing this?

The first question is also the simplest one. If a house has burned down, it is reasonable to ask why and investigate. This is a cause-and-effect question, where we use causal models to search for a likely cause for the effect. For example we know that faulty electrical systems is a common cause of fires, so if we find a device at the center of the fire, it may be a plausible explanation. The common methodology for solving this question is well understood using Bayesian approaches. Under multiple hypotheses for why there was a fire, we use probability theory to find the one that most precisely describes the outcome and evidence.

The last question is a confusing one which commonly deteriorates some of the most heated debates - usually because people do not agree what

such a question means. It's questions like *why are we doing this?* and *why are we here?*. It concerns the *meaning* or *purpose* with an action, thing or person, which implies that someone else is giving it meaning or purpose. We are also going to discuss these types of questions here.

The remaining question is the middle one; why did X cause Y? This is a causal question with great use and it is all about mediators and direct causal links.

Scurvy

Finish

11.2 Bias in Decisions

Finish

11.3 Explaining Decisions

Finish

11.4 Causality and AI

Why is it important?
Can it be done and what is the current status?
Fairness and AI, Schoelkopfs presentation page 252.
inFERENCe also mentions fairness, but it requires counterfactuals.
Perhaps you do need to move this to the end after all.

Exercises

1^{st}

Exercise

The exercises here concern the topics covered in sections 1-6.

1.1 Problems and Graphs

For each of the descriptions below, draw the related causal graph.

A Small Graph

J is a mediator between ancestor H and descendant G. P is a confounder of H and G, while Q is a child of J.

Ideal Gas Law

We have a nitrogen tank used for cooling with. We wish to know what its pressure is. We know various specs on the tank and can also do some measurements on it, but we don't know what we need. Do a Wikipedia search on the *Ideal Gas Law* and determine what factors we need to know about the tank, in order to determine the pressure inside the tank. Make a causal diagram illustrating this.

Relaxing is Dangerous?

A study from 2005 found that

> *... embarking on the Golden Years at age 55 doubled the risk for death before reaching age 65, compared with those who toiled beyond age 60.*

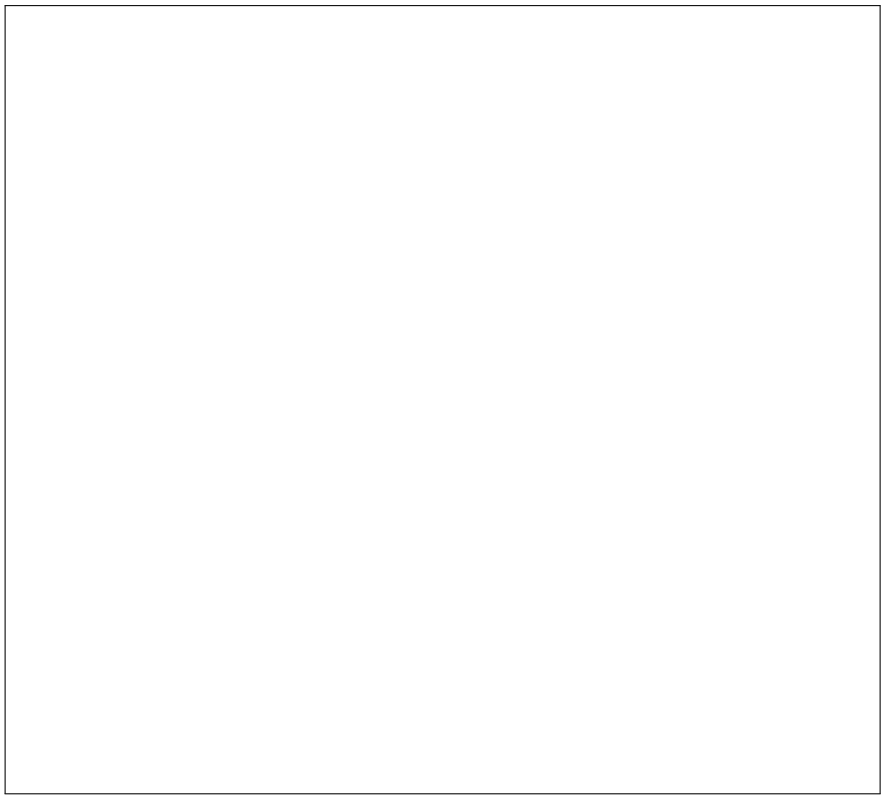
That is; people who retired early seemed to die early[1].
Make a causal graph of how you think the following variables interacts: death (D), current age (A), health-condition (H), job type (J), retirement age (R), personal economy (E). Assume everyone in the data already is retired.

[1] Luckily, a couple of later studies found the opposite to be true:
`https://www.nytimes.com/2018/01/29/upshot/early-retirement-longevity-health-wellness.html`

A Big Graph

In the following we enumerate the English alphabet in the normal order, starting with $a = 1$.

F is the parent of J, who is also a child of H. The 5th prime is parent of the 6th prime, who is a parent of the 7th prime, who is a parent of the first prime. If a node's number is odd and greater than 2^4, then that node is a parent of the last letter in the alphabet. Letter 5^2 is the child of letter 2^3. The vowels between letters "j" and "x" are children of J and parents of 2^4. The only letter whose number is a divisor of 289 is a child of node $(2^4) - 1$.



1.2 A Paradox

Initial Analysis

We are going to analyse some numbers on the fall 1973 admissions to the University of California, Berkeley. We have a hypothesis that the university has a bias against the admission of women! Consider the table below and compute the percentages of admission.

	Applicants	Admitted	Admitted %
Men	2590	1269	
Women	1835	556	

Is there a bias?

Extended Analysis

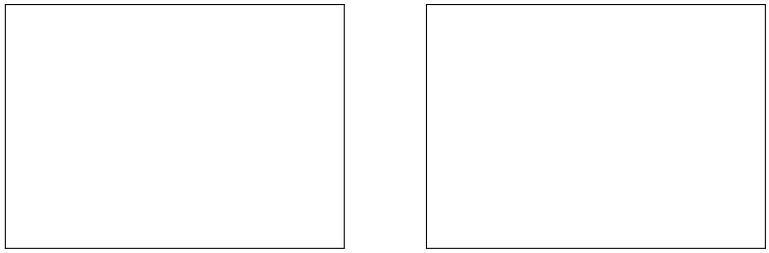
We have some more data on the admission rates for each of 6 departments. We wish to find out which departments are most problematic. Fill out the percentages below to find any bias within the departments. Let's find all departments where the men/women admissions differs by more that 5%. That is, ignore small differences list 50% and 52%, but consider a difference like 50% and 55% a bias. Note the biases in the tick-boxes on the right of the table (bias against women/no-one/men).

	Men			Women			Bias against W/-/M
Depart.	Appl.	Admit.	Admit. %	Appl.	Admit.	Admit. %	
A	825	511		108	88		☐ ☐ ☐
B	560	353		25	17		☐ ☐ ☐
C	325	120		593	202		☐ ☐ ☐
D	417	138		375	131		☐ ☐ ☐
E	191	131		393	94		☐ ☐ ☐
F	272	16		341	24		☐ ☐ ☐
Total	2590	1269		1835	556		☐ ☐ ☐

Is there a bias? Who is the admissions biased against?

Causal Analysis

In order to figure out this problem we will analyse it using the causal tools we have learned. Hypothesis 1 was that there is a causal link from gender G to admission rate A - and no other variables are considered. In Hypothesis 2 we also consider department D. What could the causal links be between G, A and D? Consider the graph for both Hypothesis 1 and 2 and draw them if Figure 1.



(a) Hypothesis 1 (b) Hypothesis 2

Figure 1

Can you use these graphs to determine whether there is a bias against women?

Simpson's Paradox

This is a case of *Simpson's Paradox*, which is perhaps the most famous statistical paradox. The paradox happens when a confounder (the department) inverts the analysed relationship (gender causes admission) in the aggregated data (total admissions), as opposed to within the subgroups. The true relationship is the one found in the subgroup, because the confounder no longer has influence.

Another famous example is from a medical study comparing the success rates of two treatments for kidney stones, whose data is shown on the right. For the total, Treatment B seems best, but this is because Treatment B got more of the easy cases (small stones) and fewer of the difficult cases (large stones). In reality Treatment A is superior.

	Treatment A	Treatment B
Small stones	93% (81 / 87)	87% (234 / 270)
Large stones	73% (192 / 263)	69% (55 / 80)
Total	78% (273 / 350)	83% (289 / 350)

When we have information and data on the confounder we can eliminate the problem with conditioning. The problem really arises when we can't access the data or perhaps don't even know about the confounder. If we have no information about the size of the kidney stones, then the only possible conclusion is - incorrectly - that Treatment B is best for treating kidney stones. This is where random trials have their strengths. By randomizing who gets which treatment, A or B, we can ensure that no confounder has influenced the decision. This is the point of randomization in A/B testing.

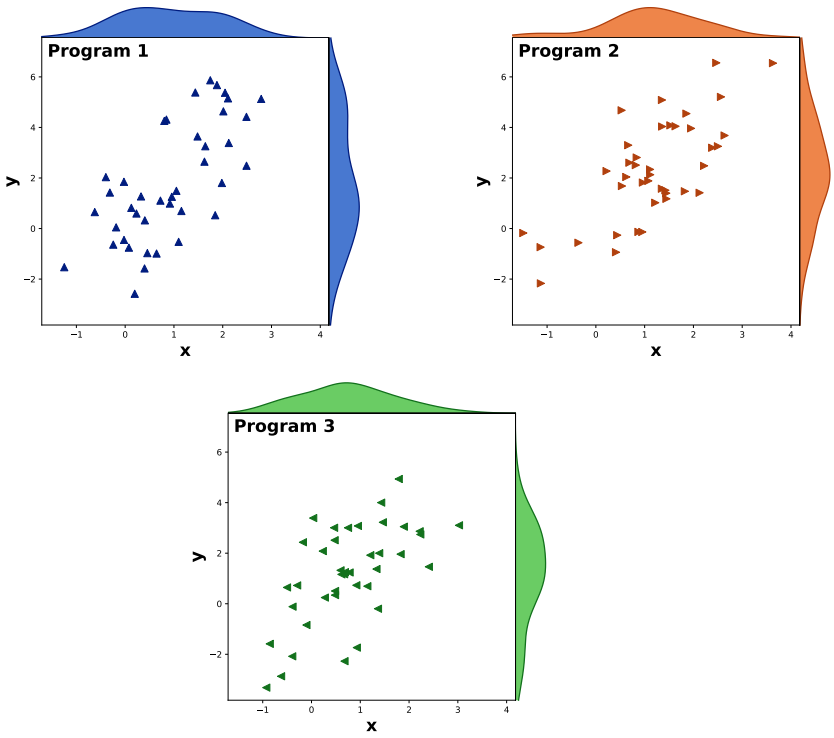
The Berkeley and Kidney Stones examples are from

- P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187, 1975
- S A Julious and M A Mullee. Confounding and Simpson's paradox. *BMJ*, 309, 1994

1.3 Intervention on Programs

Initial Analysis

In this exercise we will use the script `intervention_on_programs_1.py`. This script runs an experiment on three programs. Each of the programs produces 2D points like the one shown in the figure below, with density histograms along the axes



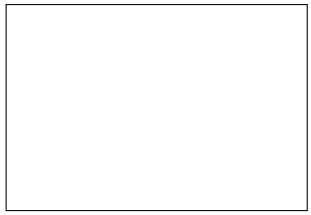
We want to know what the causal graph is for each program and how much they differ. The first thing we should consider is whether the distributions above are the same. Do you think they are?

It might be tough to say anything quantitatively about how equal these

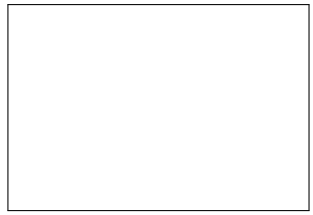
distributions are just from inspection. In the script there are two settings `n_samples` and `fit_normal_distribution`. You can use these to get more information from the distribution. Try to figure out how many samples you need before you can see the parameters fitted. Are some of the distributions identical?

Inference of Causal Structures

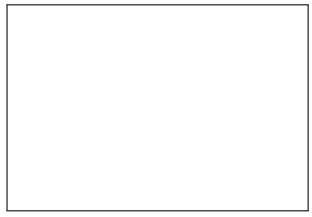
We want to determine the causal structures of the programs, but first we will make some hypotheses. There are three basic causal structures between X and Y that can explain the data above. See if you can come up with these and sketch them here



(a) Hypothesis 1



(b) Hypothesis 2



(c) Hypothesis 3

Figure 3

If we intervene on the programs they will act differently depending on their causal structure. Use the table below to check off the variable that you expect to change under intervention of X, or of Y, or under no intervention at all.

	No intervention	Intervention on X	Intervention on Y
	Change in $\boxed{X} : \boxed{Y}$	Change in $\boxed{X} : \boxed{Y}$	Change in $\boxed{X} : \boxed{Y}$
Hypothesis 1	□ : □	□ : □	□ : □
Hypothesis 2	□ : □	□ : □	□ : □
Hypothesis 3	□ : □	□ : □	□ : □

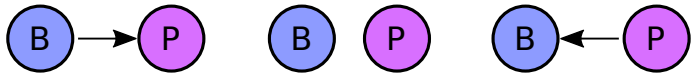
In there script there are two settings parameters: `x` and `y`. You can use these to set intervene on the programs. Try intervening on each of the variables and note below how the programs behave

	No intervention	Intervention on X	Intervention on Y
	Change in $\boxed{X} : \boxed{Y}$	Change in $\boxed{X} : \boxed{Y}$	Change in $\boxed{X} : \boxed{Y}$
Program 1	□ : □	□ : □	□ : □
Program 2	□ : □	□ : □	□ : □
Program 3	□ : □	□ : □	□ : □

So which program corresponds to each of the hypotheses?

1.4 A New Switch

In this exercise we will again consider the switch problem. Let's assume that we have another switch-box. This time the lights are blue and purple and we don't know whether blue causes purple, whether purple causes blue or whether there is no causal relationship.



(a) Hypothesis h_B (b) Hypothesis h_N (c) Hypothesis h_P

Figure 4

Furthermore we do not know any probabilities before hand. The space of hypotheses and outcome states is

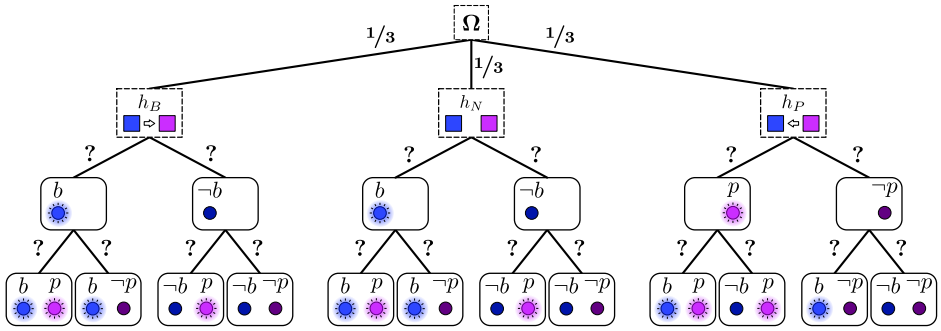


Figure 5: Advanced switch problem.

Using script `a_new_switch.py` we can get samples from the switch. We want to determine all marginal and conditional probabilities for analysing the problem. Use the script to determine these and fill them out below. Can we rule out any of the hypotheses?

$P(b)$	:	$P(\neg b)$	:	$P(p)$	:	$P(\neg p)$
	:		:		:	

$P(p \mid b)$	:	$P(\neg p \mid b)$	:	$P(p \mid \neg b)$	:	$P(\neg p \mid \neg b)$
	:		:		:	

$P(b \mid p)$	:	$P(\neg b \mid p)$	:	$P(b \mid \neg p)$	:	$P(\neg b \mid \neg p)$
	:		:		:	

We will be analysing the causal structure of the switch by intervention. In order to determine which hypothesis is correct we need to know what to expect from the intervention under each hypothesis. Fill out the table below according to what we expect to happen.

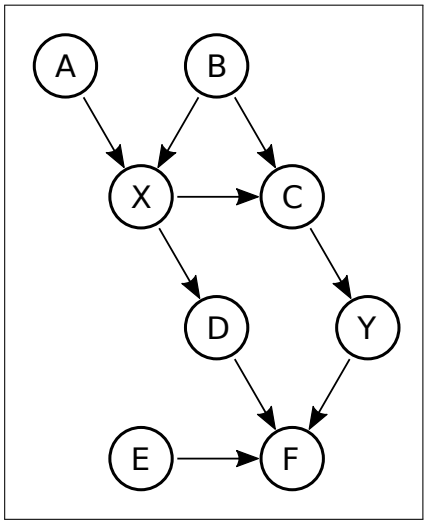
	Intervention on Blue				Intervention on Purple			
	$P(p \mid \mathring{b})$	$P(\neg p \mid \mathring{b})$	$P(p \mid \mathring{\neg b})$	$P(\neg p \mid \mathring{\neg b})$	$P(b \mid \mathring{p})$	$P(\neg b \mid \mathring{p})$	$P(b \mid \mathring{\neg p})$	$P(\neg b \mid \mathring{\neg p})$
Hypothesis h_B								
Hypothesis h_N								
Hypothesis h_P								

The method `sample_switch()` has two parameters for intervention: `intervene_blue` and `intervene_purple`. Use intervention to determine what hypothesis is correct.

1.5 Information Flow

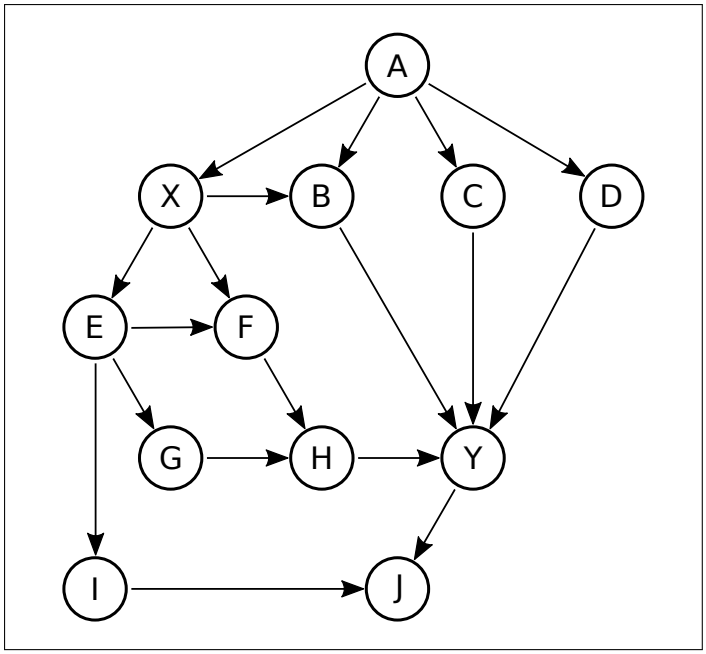
Flow Graph 1

On the right there is a causal graph. We wish to determine the causal relationship between X and Y. Draw all flow-paths between X and Y, and determine two experiments for determining the causal relationships.



Flow Graph 2

In the graph below we again want to determine the causal relationship between X and Y. Draw all information flows between the two nodes of interest.



How would you design experiments to determine the causal relationship?

2^{nd} Exercise

The exercises here concern the topics covered in 7-10.
Some of the exercises spoil solutions to parts of 1^{st} Exercise.

Gaussian Distribution

For some of these exercises we will use Gaussian/normal distributions. The following is a recap of a few properties of Gaussians.

A univariate Gaussian distribution is parameterized by

$$X \sim \mathcal{N}(\mu, \sigma^2), \tag{1}$$

$$\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+, p(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

A bivariate Gaussian distribution is parameterized by

I don't think we need the bivariate one.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

$$p(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$, and ρ is the correlation between the two dimensions.

The sum of two normally distributed, stochastic variables is again a normally distributed

$$\begin{aligned} X &\sim \mathcal{N}(\mu_x, \sigma_x^2), \\ Y &\sim \mathcal{N}(\mu_y, \sigma_y^2), \end{aligned} \tag{3}$$

$$\begin{aligned} Z &= aX + bY, & Z &= aX - bY, \\ Z &\sim \mathcal{N}(a\mu_x + b\mu_y, \sigma_z^2), & Z &\sim \mathcal{N}(a\mu_x - b\mu_y, \sigma_z^2), \\ \sigma_z^2 &= a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y, & \sigma_z^2 &= a^2\sigma_x^2 + b^2\sigma_y^2 - 2ab\rho\sigma_x\sigma_y, \end{aligned}$$

where ρ is correlation.

Categorical Distribution

We will also use a categorical distribution parameterized by

$$\begin{aligned} X &\sim \mathcal{C}(\mathbf{a}), \\ p(X = i) &= \mathbf{a}_i. \end{aligned} \tag{4}$$

For example

$$\begin{aligned} X &\sim \mathcal{C}([0.3, 0.4, 0.2, 0.1]) \\ p(X = 0) = 0.3, \quad p(X =&1) = 0.4, \quad p(X = 2) = 0.2, \quad p(X = 3) = 0.1. \end{aligned} \tag{5}$$

Bernoulli Distribution

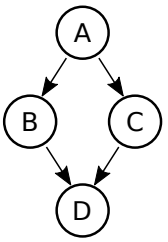
The Bernoulli distribution is a special case of the Categorical, with only two categories

$$\begin{aligned} X &\sim \mathcal{B}(0.8) \\ p(X = 0) = 0.2, \quad &p(X = 1) = 0.8. \end{aligned} \tag{6}$$

2.1 Signal Through Variables

We will here analyse the following model

$$
\begin{aligned}
A &\leftarrow N_A, & N_A &\sim \mathcal{N}(2,2) \\
B &\leftarrow \frac{1}{2}\cdot A + N_B, & N_B &\sim \mathcal{N}\left(-1/2, 2\right) \\
C &\leftarrow A + N_C, & N_C &\sim \mathcal{N}(1,1) \\
D &\leftarrow \frac{2}{3}\cdot B - \frac{1}{3}\cdot C + N_D, & N_D &\sim \mathcal{N}(1,1).
\end{aligned}
$$



Since all ancestors are normal distributions we know that D itself must also be normally distributed. Let's first consider the means of the four variables. Using (3) determine the means for the four variables

$\mathbb{E}\left[A\right]$	$\mathbb{E}\left[B\right]$	$\mathbb{E}\left[C\right]$	$\mathbb{E}\left[D\right]$

Now that we have those in place, we want to determine how much A controls the other variables. That is; how much of B's, C's and D's variance is determined by A. Using (3) you can determine/compute the variance of A, B and C

$\text{var}\left[A\right]$	$\text{var}\left[B\right]$	$\text{var}\left[C\right]$

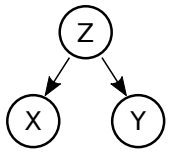
Determining the variance of D requires computing the correlation between B and C. Alternatively we can derive the final expression of D. Insert the definitions of A, B and C to determine the form of D as a function of N_A, N_B, N_C and N_D

What happened here and why is it interesting?

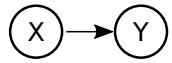
What is the variance of D?

2.2 Modelling Programs

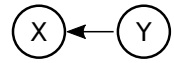
For this exercise we return to the programs of exercise 1.3. The causal graphs which we inferred in that exercise are seen below. Furthermore the data looks Gaussian.



(a) Program 1



(b) Program 2

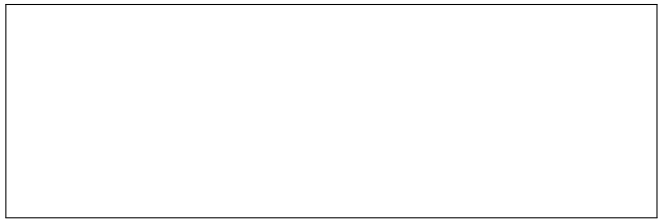


(c) Program 3

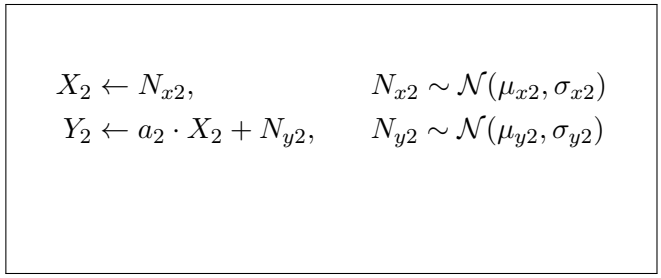
Figure 1

The Programs

We are going to assume that all the programs are normal distributions and sums of normal distributions - which are also normal distributions. First sketch out the program-structures - we've started by writing program 2. We call a_2 the *mixing coefficient* of program 2, as it determines how much X is "mixed into" Y. It does not matter what you name the mixing coefficients etc.



Program 1



Program 2

Program 3

Let's start with the two smaller programs; 2 and 3. Can you come up with a simple strategy for determining the parameters of their normal distributions?

Plan some experiments (2-9) and run the three programs. Measure the means and variances of the data and store in the table below. Write what interventions you do (for example $X \overset{\circ}{\leftarrow} 3.14$) - if you do any.

Experiment		Program 1		Program 2		Program 3		Interventi
		X	Y	X	Y	X	Y	
1	Mean							
	Variance							
2	Mean							
	Variance							
3	Mean							
	Variance							
4	Mean							
	Variance							
5	Mean							
	Variance							
6	Mean							
	Variance							
7	Mean							
	Variance							
8	Mean							
	Variance							
9	Mean							
	Variance							

Table 1: Experiments

Can you determine the noise-distributions for program 2 and 3?
For example N_{x2} and N_{y2}?

In order to completely determine programs 2 and 3, we need to determine the mixing coefficients like a_2. Using (3), with the measured means and/or variances, derive the coefficients for programs 2 and 3.

In order to solve problem 1, we need to intervene on the confounder Z. In the programs, it is possible to do intervention on this variable by setting `_z`. Do an intervention experiment on the programs, where you intervene on Z and note the results in the data-table from earlier. Assume the distribution of Z is

$$Z \sim \mathcal{N}(0, 1). \tag{7}$$

Can you determine the noise-distributions of X and Y, as well as the mixing coefficients?

2.3  An Advanced Switch  

Make a switch (or two?) with more lights - say 4. We don't draw out all possible hypotheses and we don't make the tree.
They start with zero and we expect them to find causal model with equations.
Note that they have already tried inferring such probabilities in Exercise 1, we just didn't tell them.

2.4 Counterfactuals 1

Consider the following causal model

$$
\begin{aligned}
X &\leftarrow N_x & N_x &\sim \mathcal{B}\left(\frac{7}{10}\right) \\
Y &\leftarrow N_y & N_y &\sim \mathcal{C}\left(\left[\frac{1}{5}, \frac{3}{5}, \frac{1}{5}\right]\right) \\
Z &\leftarrow X + Y - N_z & N_z &\sim \mathcal{C}\left(\left[\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right]\right)
\end{aligned}
$$

All Known

Say we observe $x = 1, y = 1, z = 0$.
Determine the value of z, given that we counterfactually do $x \overset{\bullet}{\leftarrow} 0$.

Unknown z

Now we observe $x = 1, y = 1$.
Determine the probability of $z = 1$, given that we counterfactually do $x \overset{\bullet}{\leftarrow} 0$.

$$p(\overset{\bullet}{z} = 1 \mid x \overset{\bullet}{\leftarrow} 0, x = 1, y = 1).$$

Unknown y

Now say we observe $x = 1, z = 0$.
We wish to determine the counterfactual probability

$$p(\dot{z} = 0 \mid x \overset{\bullet}{\leftarrow} 0, z = 0).$$

In order to determine this probability, we need to determine the conditional distributions of N_z and N_y. We know what x is, so what is the constraint on the remaining variables influencing Z?

Note down all values for N_y and N_z which fulfils this constraint, and their probabilities

N_y	N_z	$p(N_y, N_z)$

Now marginalize over N_y and N_z to determine the counterfactual probability

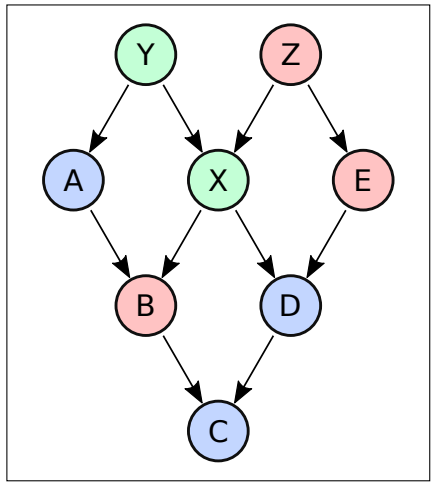
2.5 Counterfactuals 2

The example is too boring. If they are all additions, then nothing interesting happens.

Consider the following causal model

$$
\begin{aligned}
N_x, N_y &\sim \mathcal{B}\left(\tfrac{1}{3}\right) \\
N_a, N_c, N_d &\sim \mathcal{B}\left(\tfrac{2}{3}\right) \\
N_z, N_b, N_e &\sim \mathcal{C}\left(\left[\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{2}{4}\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
Y &\leftarrow N_y & Z &\leftarrow N_z \\
A &\leftarrow Y + N_a & X &\leftarrow Y + Z + N_x \\
E &\leftarrow Z + N_e & B &\leftarrow A + X + N_b \\
D &\leftarrow X + E + N_d & C &\leftarrow B + D + N_c
\end{aligned}
$$



Round 1

We want to determine

$$p\big(\dot{d} \,|\, x \overset{\bullet}{\leftarrow} 0, x = 1, z = 1\big).$$

First write out D as an expression of noise-variables and known variables and insert the counterfactual variables

Now we can make a table of values of N_d and N_e, together with the values for D and their probabilities

N_d	N_e	D	$p(N_d, N_e)$

What is the distribution of D?

Round 2

We want to determine

$$p\big(\overset{\bullet}{d} \mid x \overset{\bullet}{\leftarrow} 0, x = 1, z = 1, d = 2\big).$$

You can utilize a lot of information from Round 1. First of all we already know D as an expression of noise-variables. Furthermore we have the table of N_d, N_e and D values. Find all rows of the table where $d = 2$. Note them below and compute the conditional probabilities

N_d	N_e	$p(N_d, N_e)$	$p(N_d, N_e \mid d = 2)$	D

What is the distribution of D?

Round 3

We want to determine

$$p\big(\overset{\bullet}{d} \mid x \overset{\bullet}{\leftarrow} 0, x = 1, d = 3\big).$$

We have information about D, but not E and Z. Determine the values of the noise-variables N_d, N_e and N_z, for which $D = 3$. Note their probabilities and conditional probabilities

N_d	N_e	N_z	$p(N_d, N_e, N_z)$	$p(N_d, N_e, N_z \mid d = 3)$
				$2/9$
				$2/9$
				$1/9$
				$2/9$
				$2/9$

2.6  Counterfactuals 3 

Round 4

Determine

$$p\big(\dot{c} \mid x \twoheadleftarrow 3, x = 1, z = 1, y = 2, c = 16\big)$$

Round 5

Determine

$$p\big(\dot{c} \mid x \twoheadleftarrow 3, x = 1, c = 16\big)$$

Project

Project in Causal Inference

Project TODO:

- Change the structure of the code base. Make three directories
  - document - code for making the main document, which other should not edit
  - exercises - code for the exercises which the students will access
  - project - project code

  Keep all runnable scripts in the root of each directory, and make the script runnable from there. Thus make three PyCharm projects.

1.1 Project Description

In this project we will infer a causal model and compete to see who can do it most efficiently!

Your teacher will have defined a causal model, which is running on a server. You can get samples from the causal model by simply emailing the server. You can ask for as many experiments and samples as you want, and make any combination of interventions, but there is a catch to greed: the winner will be the team that can correctly determine the causal structure using the least experiments and samples!

Details of experiments and competition are below. You need to ask your teacher for the following information:

Server email	
Number of nodes in causal system	$N =$
Cost of sample	$C_s =$
Cost of experiment	$C_e =$
Cost of incorrect guess	$C_i =$

1.2 Experiments

For each experiment you want to do, you have to send one email to the server. You write what experiment you want to do in the subject line (the body of the email is irrelevant). The format for communicating with the server is

`n_samples, <interventions>`,

where `n_samples` is the number of samples you want to get for this experiment and `<interventions>` is a comma-separated list of interventions you want to do (without the `<` and `>`).

Here's a couple of examples on the query-format

	X	Y	Z	F	G	I
0	1.028379	1.0	0.0	0.485250	1.276530	1.193008
1	2.670319	1.0	0.0	0.810593	-0.301816	1.027937
2	5.294258	0.0	1.0	0.319836	-3.224374	0.254132
3	-3.494897	0.0	1.0	0.481506	0.640079	1.165772
4	2.440380	1.0	1.0	0.816298	-4.420412	2.269149

(a) Readable data-format

Table 1: Example of file from experiment server.

`10`	Get 10 observational samples.
`10,`	Also get 10 observational samples.
`5, X=1`	Get 5 interventional samples where we force $X \overset{\circ}{\leftarrow} 1$
`25, Z=3.14, X=1, Y=0`	Get 25 interventional samples where we force $Z \overset{\circ}{\leftarrow} 3.14$, $X \overset{\circ}{\leftarrow} 1$ and $Y \overset{\circ}{\leftarrow} 0$

After sending an email to the server it will perform the experiment and send the results back. The received email will have two files with the same data. One file will have the data in a human-readable format as in Table 1a. The other file will be in CSV-format and will have higher precision, but will not be as easy to inspect by people. This format is exemplified in Table 1b.

1.3 Inferring the Graph

We want you to determine the causal graph of the model your teacher has set up. Your teacher should have informed you about how many nodes there are, but so far that is all we know. To get you started we should consider how many possible graphs you might be working with. The number of causal graphs (assuming no cycles) is the number of *labelled, acyclic, directed graphs* of a certain size. Robinson 1970 has shown that the number of these graphs possible for N nodes is

$$a_N = \sum_{k=1}^{N} (-1)^{k-1} \binom{N}{k} 2^{k(N-k)} a_{N-k}, \tag{1}$$

where $a_1 = 1$. How many possible graphs can you make with the N nodes provided in this project?[1]

Do you think it might be possible for you to try all combinations?

If not, then you may have to figure out a more structured approach. A good start is to get some samples from the observational distribution and see what the data looks like. Through correlation you may be able to eliminate some possible causal structures. Then you should start thinking about what interventions would be best for determining the causal graph.

[1] If all else fails you can look up the A003024 sequence in OEIS.

1.4 Making a Guess

You can make a guess of the causal graph by sending

```
guess: <edges>
```

to the server, there `<edges>` is a list of edges. The format of the list of edges is like you would initialize a list of tuples in python, where each tuple has two strings; first the name of the ancestor then the name of the descendant.

The following are examples of attempts at guessing the graph in Figure 7 on page 45

`guess: [('Y', 'Z'), ('X', 'Z')]`	correct guess
`guess: [('X', 'Z'), ('y', 'z')]`	correct guess
`guess: [('Y', 'Z')]`	incorrect guess
`guess: [('Y', 'Z'), ('Z', 'X')]`	incorrect guess
`guess: [('X', 'Z'), ('Y', 'Z'), ('Z', 'H')]`	incorrect guess

Notice that the order of the edges and the casing of the names do not matter.

Also, every time you make an incorrect guess, the server will remember and decrease your performance.

1.5 Competition

The competition is as follows: we hope for all students to determine the causal graph of the problem. After they have guessed the correct graph, we compute a cost of

$$\text{cost} = C_e \cdot (\#\text{ experiments}) + C_s \cdot (\#\text{ samples}) + C_i \cdot (\#\text{ incorrect guesses}). \tag{2}$$

The winner is the team with the lowest cost.

If you make many small experiments, then $C_e \cdot (\#\text{ experiments})$ may get big and give you a bad score. Similarly if you make all your examples huge, then $C_s \cdot (\#\text{ samples})$ may give you a bad score. You therefore need to balance the two to get a good score. Also if you guess correct the first time, then you don't have to be bothered with C_i, but every time you guess wrong it increases the cost!

Finally note that you should prioritise your studies and your understanding of the course higher than your competition score :P.

Project in Causal Inference - Teacher

Read the project description made for the student before this document. Also it does not matter if your students read this document - there will not be any useful spoilers.

2.1 Project Scope

The idea of the project is that the students will attempt to guess a causal graph which you define. This is a very open problem, which can utilize many interesting tools of mathematics, computer science and machine learning, but can also be solved with fairly simple approaches in a less rigorous manner.

The simplest approach is probably to simply make histograms of the data and compare these under interventions. Relatively quickly you can determine the first links in the causal graph and take it from there.

More advanced approaches will include modelling the data (normal distributions etc.). This gives the students precise measures of when a variable has changed, as well as an idea about how many samples are necessary for reliably determining these changes. The methodology can potentially include tools like information theory, active learning and Bayesian optimization to determine the best experiments to perform to most efficiently determine the causal structure.

We recommend making a competition with prizes for the winners, but also recommend basing any grades on a project report (or something similar), which is independent of the competition scores.

2.2 Server

We have attempted to make the server as easy as possible to use, but note that (depending on version of this document) it has had limited testing and may be improved in the future. In the supplied code base there is a `project`-directory. This directory has the following files at its root for you to use

1. `define_server.py`
2. `server_run.py`
3. `server_reset.py`
4. `server_inspect.py`
5. `plot_causal_graph.py`

The first file `define_server.py` is where you will define the problem which the students will be facing, as well set some settings for the server. The original version of the document is seen in Figure **??** on page **??**.

Gmail

Currently the system is set up for using a Gmail-account. You can (probably) use other email-providers, but you might have to rewrite some of the interactions with the server. It is likely that we will experiment with other email-providers as well, so it might be easier in future versions. If you already have a Gmail-account, then you should most definitely NOT use that account. We will be saving the password for the account as free-text on your computer and also remove some security settings.

For setting up Gmail

1. Make a Gmail-account dedicated for the course
2. Allow less-secure apps by
   (a) Go to `https://myaccount.google.com`
   (b) *Security*
   (c) Find *Less secure app access*
   (d) Turn on access

3. Write email-account name and password in `define_server.py` (more on that later)
4. Run competition
5. Delete account afterwards

Setup

For setting up the server you have to edit a couple of fields of the `ServerSettings`-class in `define_server.py`. Refer to Figure **??** to see lines and an example of setup.

- Line 9: write the email of the account used for the server.
- Line 10: write the password of the account used for the server.
- Lines 13, 14 and 15 are information about the email-provider. If you use Gmail, then keep as is.
- Line 18 notes how often the server will check the email (currently every 2 seconds).

Access

Not everyone should have access to the server. This is first of all to avoid the server crashing from spam etc., but also so that you know exactly which student got which data. In lines 20-26 we write down all emails that are allowed to participate in the competition, separated by line-breaks. In the example we have included the server's own email which we use for testing, but that is not necessary - just write the students emails.

When the students ask for data, the server will remember how much data they have received. This will be stored per-email. Thus if the students work in teams it is easiest to have a single student in each team with access to the server. This way you know exactly how much data each team has, because each team is uniquely identified by a single email address.

Also the responses from the server may sometimes end up in spam filters - make sure the students check that.

2.3 Causal Model

It's time to define the causal model for the students to analyse. We have also attempted to make this process as easy as we could. You define the model by creating the `_sample()`-method of the `ExperimentSystem`-class in line 39.

You define a causal variable by assigning some samples to its name as shown in lines 2, 3 and 4. In those lines we use three methods defined in the class for ease of use. All they do is create samples from a normal-distribution, a Bernoulli distribution (binary variable) and a categorical distribution. These samples are assigned to their respective nodes in the caused model - here we use the names `X`, `Y` and `Z`.

You can also use other samplers as long as they have exactly `n_samples` values. We illustrate this in line 8, where we use `numpy`'s beta-distribution sampler for setting variable `F`.

You can combine causal variables to make causal structures, by referring to the names. This is illustrated in line 11, where we make variable `G` a causal descendant of nodes `X` and `F`.

For ensuring that the system works as well, please do *not* make any temporary variables at all. Any numbers that you save temporarily should be saved in a named causal variable. If you want to make hidden variables that the students can not see, name them something that starts with `_`, as shown in line 15. Variable `_H` is hidden and is in this example used as a hidden confounder for variable `I`.

When the students make experiments with the example system they receive data on `X`, `Y`, `Z`, `F`, `G` and `I`, while variable `_H` is hidden from them.

If you want to see the full data including all hidden variables, then you pass a query with a password. For example

`5, X=1, password=Open_Sesame`	Get 5 interventional samples where we force $X \overset{\circ}{\leftarrow} 1$, and use password `Open_Sesame`.

The password is set in line 31 and should of course be changed in case the students read this document :)

The server will return data-tables with samples, where the causal variables are per default arranged in the order of their definition. This gives the students a hint, because the first variables will be the ancestors and the last variables will be the descendants. In order to remove this hint you can define some other ordering as done in line 21.

We suggest computing a cost for the students as the sum of

$$\text{cost} = C_e \cdot (\#\text{ experiments}) + C_s \cdot (\#\text{ samples}) + C_i \cdot (\#\text{ incorrect guesses}).$$

In order to allow the server to compute the cost you should fill these in, in the lines starting at line 33.

You can plot the causal graph with the script `plot_causal_graph.py` to see if it matches you expectations.

Note that the system does not care about casing of the names, so do not name nodes the same name but with different case (for example `x` and `X`).

2.4 Running Server

The server can be run by running the script called `server_run.py`. While the server is running it will display an output similar to that of Figure 1. In the example the server first check the emails and concludes that there has been no queries. It then receives two queries from two students, which it successfully understood and answered. It then continues to wait for emails. The server writes out who it has received emails from and what their subject lines were.

If the server receives an email it does not understand, it will write a comment to the student about the format of the emails. One example of the server receiving a bad email is seen in Figure 2, where a student wrote an incorrect query.

```
2019-11-22 13:33:29 -> Checking emails
2019-11-22 13:33:31 ->       No new emails.
2019-11-22 13:33:31 -> Sleeping 2s

2019-11-22 13:33:33 -> Checking emails
2019-11-22 13:33:36 ->       Email 15: SUCCESS, [15, Y=0], [
   student_email_1@university.com]
2019-11-22 13:33:39 ->       Email 16: SUCCESS, [20, X=1], [
   student_email_2@university.com]
2019-11-22 13:33:39 -> Sleeping 2s

2019-11-22 13:33:41 -> Checking emails
2019-11-22 13:33:43 ->       No new emails.
2019-11-22 13:33:43 -> Sleeping 2s

2019-11-22 13:33:45 -> Checking emails
2019-11-22 13:33:47 ->       No new emails.
2019-11-22 13:33:47 -> Sleeping 2s
```

Figure 1: Example of output from server when running.

```
2019-11-22 13:36:51 -> Checking emails
2019-11-22 13:36:53 ->       Email 17: Cannot parse subject line,
                             [X=1] [student_email_1@university.com]
   <--
2019-11-22 13:36:53 -> Sleeping 2s
```

Figure 2: Example of output from server which received incorrect query.

2.5 Results

The script `server_inspect.py` is used to see how everything is progressing. An example of the output of this program is seen in Figure 3

```
Server results.

24 emails reveived.
22 where successfully parsed (91.6%).


--------------------------------------------------------------------------------

User information
                          n_emails  n_samples  n_experiments  guesses
  incorrect_guesses  bad_emails
student_1@university.com         9         64              3        5
           4           1
student_2@university.com         7         58              3        3
           2           1
student_3@university.com         8         80              4        4
           3           0
--------------------------------------------------------------------------------



--------------------------------------------------------------------------------

Competition table
                          n_experiments  n_samples  incorrect_guesses  cost
student_2@university.com              3         58                  2   258
student_3@university.com              4         80                  3   370
student_1@university.com              3         64                  4   404

______________________________
      experiment_cost  sample_cose  incorrect_guess_cost
Costs              20            1                    70
--------------------------------------------------------------------------------
```

Figure 3: Example of output from `server_inspect.py`.

First the server prints how many emails it has successfully understood. In the example the students have sent 2 emails with incorrect formatting.

The first table shows some utility information about the students

- how many emails they have sent
- how many samples and experiments they have made
- how many guesses and correct guesses they have made
- how many incorrectly formatted emails they have sent

The second table shows the competition data. Is shows the number of experiments, samples and incorrect guesses for each student as well as their final score.

At the bottom we also see the costs of experiments, samples and incorrect guesses for reference.

Each row in the bottom table is only filled out once the student has guessed the correct graph and will remain static afterwards - even if the student keeps querying the system out of interest.

```python
import numpy as np

from project.src.causal_system import CausalSystem


class ServerSettings:

    # Email information  -> you need to turn on "Less secure app access" on Gmail
    username = "intervention.experiment@gmail.com"  #
    server_password = "hnJk9FYAWN03Er2p"  #

    # Email provider information
    imap_host = "imap.gmail.com"  #
    smtp_host = "smtp.gmail.com"  #
    smtp_port = 587  #

    # Other settings
    check_email_delay = 2  #

    # Emails with access to experiment   #
    allowed_emails = """
    intervention.experiment@gmail.com
    jepno@dtu.dk
    student_email_1@university.com
    student_email_2@university.com
    """  #


# Define the causal system
class ExperimentSystem(CausalSystem):  #
    _project_password = "Open_Sesame"  #

    # Score settings for competition
    experiment_cost = 20
    sample_cost = 1
    incorrect_guess_cost = 70

    # Causal model
    def _sample(self, n_samples):  #
        # Using predefined distributions
```

Figure 4: `define_server.py`

Continued...

```
        # Using predefined distributions
        self["X"] = self.normal(mu=2, std=3)    #
        self["Y"] = self.binary(p_success=0.8)    #
        self["Z"] = self.categorical([7, 2, 1])    #

        # You can also use numpy sampling
        #  - but you have to do it correctly by using _n_samples
        self["F"] = np.random.beta(a=5, b=3, size=n_samples)  #

        # Combining is fine
        self["G"] = self.normal(mu=0, std=1) * self["X"] + self["F"]  #

        # This variable is hidden (because of the "_")
        # It will not be returned in the data unless you have the password
        self["_H"] = self.normal(mu=0, std=1)  #

        # This variable has a hidden confounder
        self["I"] = self["F"] + self["_H"] + self.normal(mu=0, std=0.2)
    #

        # Ordering without hint of causal structure
        self._ordering = ['I', 'Z', 'F', '_H', 'X', 'G', 'Y']  #
```

Figure 5: `define_server.py`

Exercises - With Solutions

1^{st}

Exercise

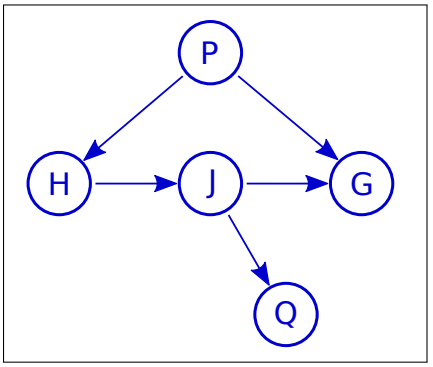
The exercises here concern the topics covered in sections 1-6.

1.1 Problems and Graphs

For each of the descriptions below, draw the related causal graph.

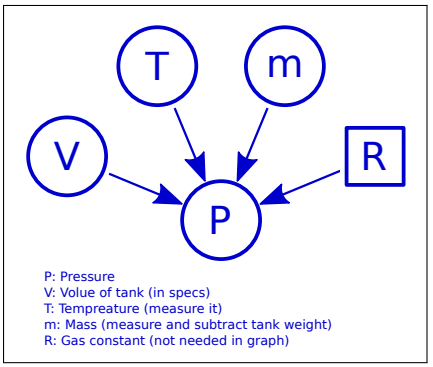
A Small Graph

J is a mediator between ancestor H and descendant G. P is a confounder of H and G, while Q is a child of J.



Ideal Gas Law

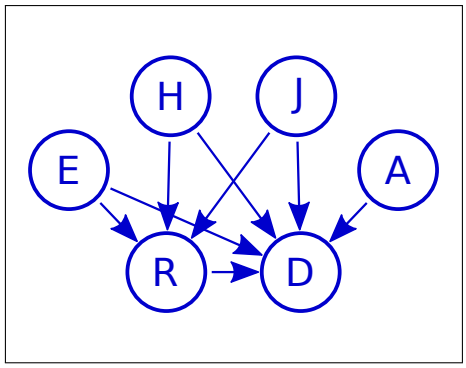
We have a nitrogen tank used for cooling with. We wish to know what its pressure is. We know various specs on the tank and can also do some measurements on it, but we don't know what we need. Do a Wikipedia search on the *Ideal Gas Law* and determine what factors we need to know about the tank, in order to determine the pressure inside the tank. Make a causal diagram illustrating this.



Relaxing is Dangerous?

A study from 2005 found that

> *... embarking on the Golden Years at age 55 doubled the risk for death before reaching age 65, compared with those who toiled beyond age 60.*



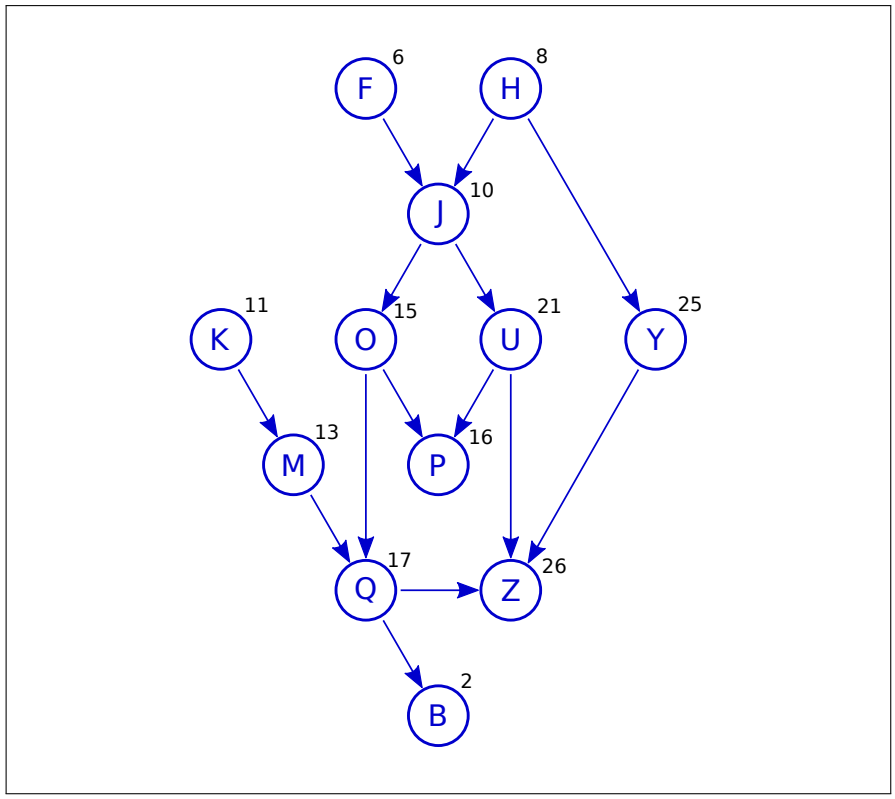
That is; people who retired early seemed to die early[1].
Make a causal graph of how you think the following variables interacts: death (D), current age (A), health-condition (H), job type (J), retirement age (R), personal economy (E). Assume everyone in the data already is retired.
(E), (H) and (J) can all influence retirement age (but your current age post-retirement does not). All factors can influence risk of death. There are a LOT of confounders for this problem, so if the research was not done right, it may not be trustworthy.

[1] Luckily, a couple of later studies found the opposite to be true:
`https://www.nytimes.com/2018/01/29/upshot/early-retirement-longevity-health-wellness.html`

A Big Graph

In the following we enumerate the English alphabet in the normal order, starting with $a = 1$.

F is the parent of J, who is also a child of H. The 5th prime is parent of the 6th prime, who is a parent of the 7th prime, who is a parent of the first prime. If a node's number is odd and greater than 2^4, then that node is a parent of the last letter in the alphabet. Letter 5^2 is the child of letter 2^3. The vowels between letters "j" and "x" are children of J and parents of 2^4. The only letter whose number is a divisor of 289 is a child of node $(2^4) - 1$.



1.2 A Paradox

Initial Analysis

We are going to analyse some numbers on the fall 1973 admissions to the University of California, Berkeley. We have a hypothesis that the university has a bias against the admission of women! Consider the table below and compute the percentages of admission.

	Applicants	Admitted	Admitted %
Men	2590	1269	49%
Women	1835	556	30%

Is there a bias?
It seems like there is a bias against women.

Extended Analysis

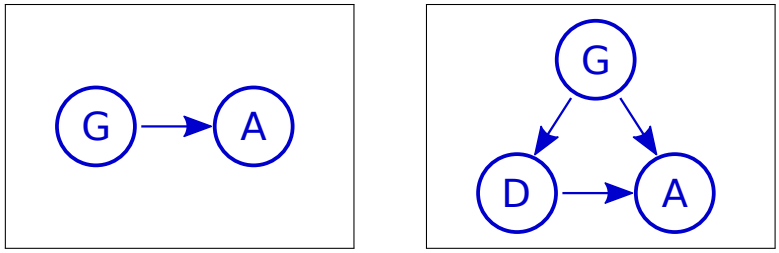
We have some more data on the admission rates for each of 6 departments. We wish to find out which departments are most problematic. Fill out the percentages below to find any bias within the departments. Let's find all departments where the men/women admissions differs by more that 5%. That is, ignore small differences list 50% and 52%, but consider a difference like 50% and 55% a bias. Note the biases in the tick-boxes on the right of the table (bias against women/no-one/men).

	Men			Women			Bias against W/-/M
Depart.	Appl.	Admit.	Admit. %	Appl.	Admit.	Admit. %	
A	825	511	62%	108	88	82%	☐ ☐ ☑
B	560	353	63%	25	17	68%	☐ ☐ ☑
C	325	120	37%	593	202	34%	☐ ☑ ☐
D	417	138	33%	375	131	35%	☐ ☑ ☐
E	191	131	28%	393	94	24%	☐ ☑ ☐
F	272	16	6%	341	24	7%	☐ ☑ ☐
Total	2590	1269	49%	1835	556	30%	☑ ☐ ☐

Is there a bias? Who is the admissions biased against?
If we look at the overall stats it seems like there is a bias against women, but if we look at the departments it seems like there is a bias against men.

Causal Analysis

In order to figure out this problem we will analyse it using the causal tools we have learned. Hypothesis 1 was that there is a causal link from gender G to admission rate A - and no other variables are considered. In Hypothesis 2 we also consider department D. What could the causal links be between G, A and D? Consider the graph for both Hypothesis 1 and 2 and draw them if Figure 1.



(a) Hypothesis 1 (b) Hypothesis 2

Figure 1

The gender could cause a bias in both hypotheses, but the gender could also cause choice of department which could affect admission rate.

Can you use these graphs to determine whether there is a bias against women?
If hypothesis 1 is correct, then conditioning on the department should not make a difference - but it does. Hypothesis 2 therefore seems more likely. If we want to assess the magnitude of the direct causal link between gender and admission rate under Hypothesis 2, then we have to condition on department (just like we did). When conditioning on department there actually seem to be a small bias against men.

Simpson's Paradox

This is a case of *Simpson's Paradox*, which is perhaps the most famous statistical paradox. The paradox happens when a confounder (the department) inverts the analysed relationship (gender causes admission) in the aggregated data (total admissions), as opposed to within the subgroups. The true relationship is the one found in the subgroup, because the confounder no longer has influence.

Another famous example is from a medical study comparing the success rates of two treatments for kidney stones, whose data is shown on the right. For the total, Treatment B seems best, but this is because Treatment B got more of the easy cases (small stones) and fewer of the difficult cases (large stones). In reality Treatment A is superior.

	Treatment A	Treatment B
Small stones	93% (81 / 87)	87% (234 / 270)
Large stones	73% (192 / 263)	69% (55 / 80)
Total	78% (273 / 350)	83% (289 / 350)

When we have information and data on the confounder we can eliminate the problem with conditioning. The problem really arises when we can't access the data or perhaps don't even know about the confounder. If we have no information about the size of the kidney stones, then the only possible conclusion is - incorrectly - that Treatment B is best for treating kidney stones. This is where random trials have their strengths. By randomizing who gets which treatment, A or B, we can ensure that no confounder has influenced the decision. This is the point of randomization in A/B testing.

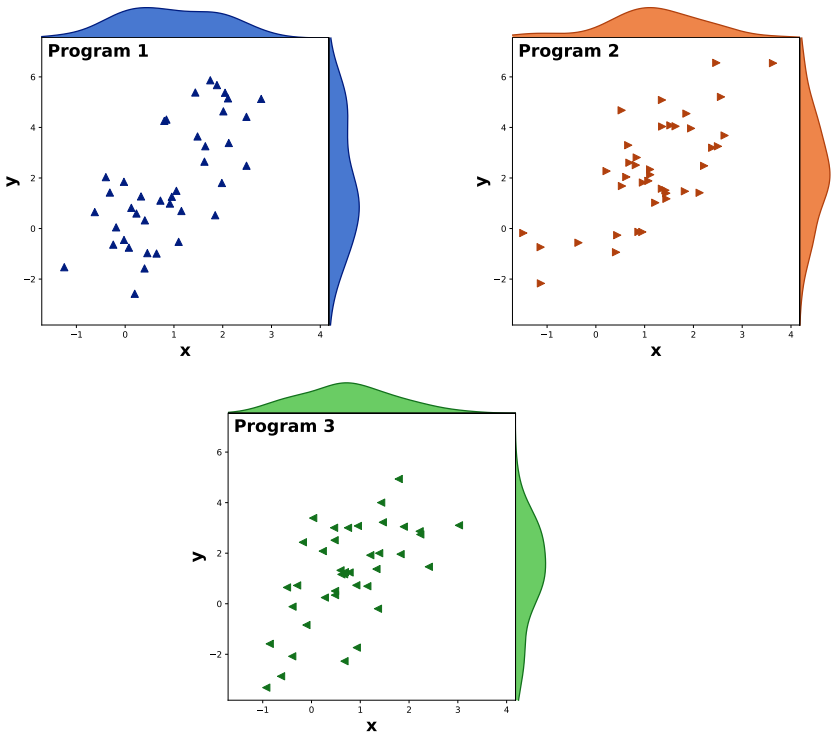
The Berkeley and Kidney Stones examples are from

- P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187, 1975
- S A Julious and M A Mullee. Confounding and Simpson's paradox. *BMJ*, 309, 1994

1.3 Intervention on Programs

Initial Analysis

In this exercise we will use the script `intervention_on_programs_1.py`. This script runs an experiment on three programs. Each of the programs produces 2D points like the one shown in the figure below, with density histograms along the axes



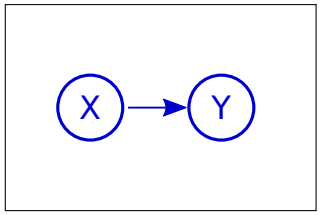
We want to know what the causal graph is for each program and how much they differ. The first thing we should consider is whether the distributions above are the same. Do you think they are?
Maybe - difficult to say with this few points

It might be tough to say anything quantitatively about how equal these

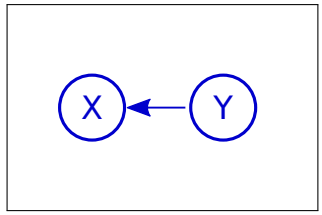
distributions are just from inspection. In the script there are two settings `n_samples` and `fit_normal_distribution`. You can use these to get more information from the distribution. Try to figure out how many samples you need before you can see the parameters fitted. Are some of the distributions identical?
I ran about 500 points. The distributions seem quite similar. If I run 50.000 points then they seem really identical.

Inference of Causal Structures

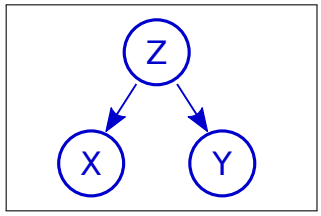
We want to determine the causal structures of the programs, but first we will make some hypotheses. There are three basic causal structures between X and Y that can explain the data above. See if you can come up with these and sketch them here



(a) Hypothesis 1



(b) Hypothesis 2



(c) Hypothesis 3

Figure 3

If we intervene on the programs they will act differently depending on their causal structure. Use the table below to check off the variable that you expect to change under intervention of X, or of Y, or under no intervention

at all.

	No intervention	Intervention on X	Intervention on Y
	Change in $\boxed{X}:\boxed{Y}$	Change in $\boxed{X}:\boxed{Y}$	Change in $\boxed{X}:\boxed{Y}$
Hypothesis 1	☐ : ☐	☑ : ☑	☐ : ☑
Hypothesis 2	☐ : ☐	☑ : ☐	☑ : ☑
Hypothesis 3	☐ : ☐	☑ : ☐	☐ : ☑

The order of the answers of cause change depending on order of the hypotheses.

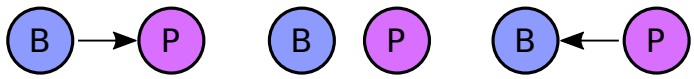
In there script there are two settings parameters: `x` and `y`. You can use these to set intervene on the programs. Try intervening on each of the variables and note below how the programs behave

	No intervention	Intervention on X	Intervention on Y
	Change in $\boxed{X}:\boxed{Y}$	Change in $\boxed{X}:\boxed{Y}$	Change in $\boxed{X}:\boxed{Y}$
Program 1	☐ : ☐	☑ : ☐	☐ : ☑
Program 2	☐ : ☐	☑ : ☑	☐ : ☑
Program 3	☐ : ☐	☑ : ☐	☑ : ☑

So which program corresponds to each of the hypotheses?
Using our ordering we have Program 1 is Hypothesis 3, Program 2 is Hypothesis 1 and Program 3 is Hypothesis 2.

1.4 A New Switch

In this exercise we will again consider the switch problem. Let's assume that we have another switch-box. This time the lights are blue and purple and we don't know whether blue causes purple, whether purple causes blue or whether there is no causal relationship.



(a) Hypothesis h_B (b) Hypothesis h_N (c) Hypothesis h_P

Figure 4

Furthermore we do not know any probabilities before hand. The space of hypotheses and outcome states is

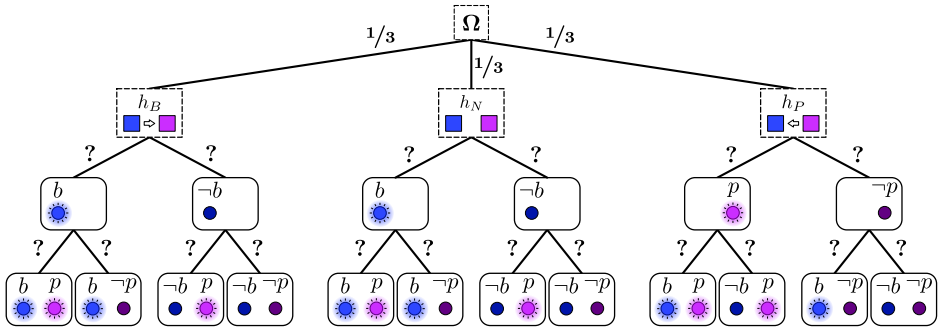


Figure 5: Advanced switch problem.

Using script `a_new_switch.py` we can get samples from the switch. We want to determine all marginal and conditional probabilities for analysing the problem. Use the script to determine these and fill them out below. Can we rule out any of the hypotheses?

We can rule out h_N, because the variables are clearly not independent:

$$P(p \mid b) \neq P(p), \qquad P(b \mid p) \neq P(b), \qquad P(b, p) \neq P(b)\, P(p).$$

$P(b)$	:	$P(\neg b)$	:	$P(p)$	:	$P(\neg p)$
0.50	:	0.50	:	0.50	:	0.50

$P(p \mid b)$	:	$P(\neg p \mid b)$	:	$P(p \mid \neg b)$	:	$P(\neg p \mid \neg b)$
0.65	:	0.35	:	0.35	:	0.65

$P(b \mid p)$	:	$P(\neg b \mid p)$	:	$P(b \mid \neg p)$	:	$P(\neg b \mid \neg p)$
0.65	:	0.35	:	0.35	:	0.65

We will be analysing the causal structure of the switch by intervention. In order to determine which hypothesis is correct we need to know what to expect from the intervention under each hypothesis. Fill out the table below according to what we expect to happen.

	Intervention on Blue				Intervention on Purple			
	$P(p \mid b^{\circ})$	$P(\neg p \mid b^{\circ})$	$P(p \mid \neg b^{\circ})$	$P(\neg p \mid \neg b^{\circ})$	$P(b \mid p^{\circ})$	$P(\neg b \mid p^{\circ})$	$P(b \mid \neg p^{\circ})$	$P(\neg b \mid \neg p^{\circ})$
Hypothesis h_B	0.65	0.35	0.35	0.65	0.50	0.50	0.50	0.50
Hypothesis h_N	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Hypothesis h_P	0.50	0.50	0.50	0.50	0.35	0.65	0.65	0.35

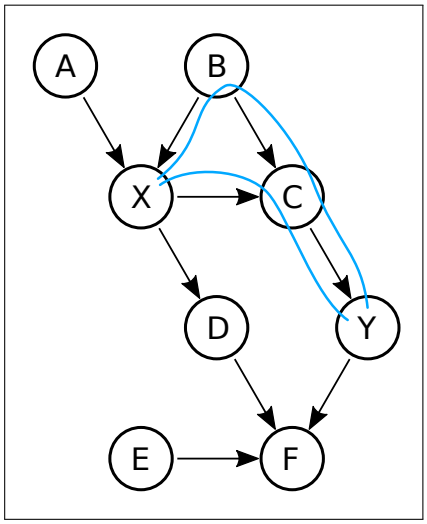
The method `sample_switch()` has two parameters for intervention: `intervene_blue` and `intervene_purple`. Use intervention to determine what hypothesis is correct.

Hypothesis h_P is correct.

1.5 Information Flow

Flow Graph 1

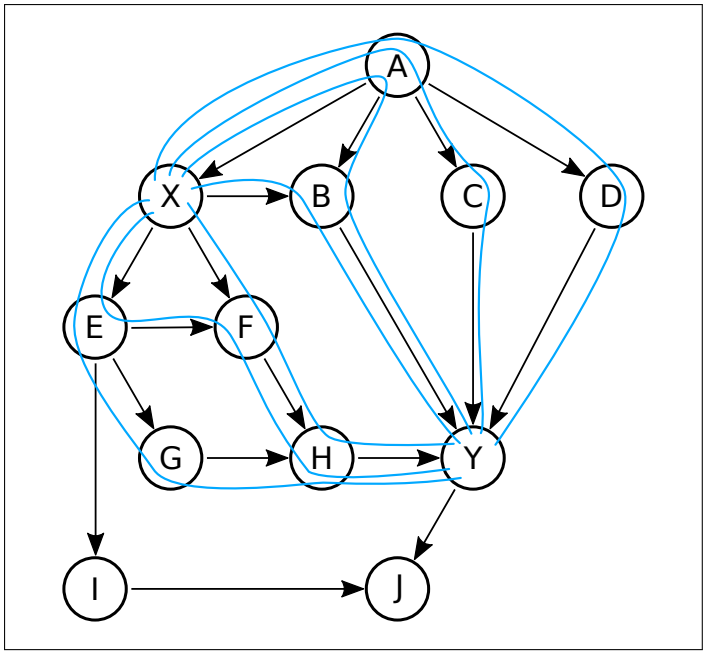
On the right there is a causal graph. We wish to determine the causal relationship between X and Y. Draw all flow-paths between X and Y, and determine two experiments for determining the causal relationships.



We want to estimate the causal path $X \to C \to Y$ while ignoring the flow of information of the path $X \leftarrow B \to C \to Y$. In order to do this we can make an observational experiment and condition on B, or we could make an interventional experiment where we intervene on X.

Flow Graph 2

In the graph below we again want to determine the causal relationship between X and Y. Draw all information flows between the two nodes of interest.



How would you design experiments to determine the causal relationship?
We can again intervene on X if possible.
We can also condition on B, C and D, but the better solution is to only condition on A.

2nd Exercise

The exercises here concern the topics covered in 7-10.
Some of the exercises spoil solutions to parts of 1st Exercise.

Gaussian Distribution

For some of these exercises we will use Gaussian/normal distributions. The following is a recap of a few properties of Gaussians.

A univariate Gaussian distribution is parameterized by

$$X \sim \mathcal{N}(\mu, \sigma^2), \tag{1}$$

$$\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+, p(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

A bivariate Gaussian distribution is parameterized by

I don't think we need the bivariate one.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

$$p(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$, and ρ is the correlation between the two dimensions.

The sum of two normally distributed, stochastic variables is again a normally distributed

$$\begin{aligned} X &\sim \mathcal{N}(\mu_x, \sigma_x^2), \\ Y &\sim \mathcal{N}(\mu_y, \sigma_y^2), \end{aligned} \tag{3}$$

$$\begin{aligned} Z &= aX + bY, & Z &= aX - bY, \\ Z &\sim \mathcal{N}(a\mu_x + b\mu_y, \sigma_z^2), & Z &\sim \mathcal{N}(a\mu_x - b\mu_y, \sigma_z^2), \\ \sigma_z^2 &= a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y, & \sigma_z^2 &= a^2\sigma_x^2 + b^2\sigma_y^2 - 2ab\rho\sigma_x\sigma_y, \end{aligned}$$

where ρ is correlation.

Categorical Distribution

We will also use a categorical distribution parameterized by

$$\begin{aligned} X &\sim \mathcal{C}(\mathbf{a}), \\ p(X = i) &= \mathbf{a}_i. \end{aligned} \tag{4}$$

For example

$$\begin{aligned} X &\sim \mathcal{C}([0.3, 0.4, 0.2, 0.1]) \\ p(X = 0) = 0.3, \quad p(X =&1) = 0.4, \quad p(X = 2) = 0.2, \quad p(X = 3) = 0.1. \end{aligned} \tag{5}$$

Bernoulli Distribution

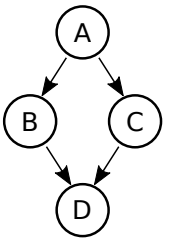
The Bernoulli distribution is a special case of the Categorical, with only two categories

$$\begin{aligned} X &\sim \mathcal{B}(0.8) \\ p(X = 0) = 0.2, \quad &p(X = 1) = 0.8. \end{aligned} \tag{6}$$

2.1 Signal Through Variables

We will here analyse the following model

$$
\begin{aligned}
A &\leftarrow N_A, & N_A &\sim \mathcal{N}(2,2) \\
B &\leftarrow \frac{1}{2}\cdot A + N_B, & N_B &\sim \mathcal{N}\left(-1/2, 2\right) \\
C &\leftarrow A + N_C, & N_C &\sim \mathcal{N}(1,1) \\
D &\leftarrow \frac{2}{3}\cdot B - \frac{1}{3}\cdot C + N_D, & N_D &\sim \mathcal{N}(1,1).
\end{aligned}
$$



Since all ancestors are normal distributions we know that D itself must also be normally distributed. Let's first consider the means of the four variables. Using (3) determine the means for the four variables

$\mathbb{E}[A]$	$\mathbb{E}[B]$	$\mathbb{E}[C]$	$\mathbb{E}[D]$
2	$\frac{1}{2}\cdot 2 - \frac{1}{2} = \frac{1}{2}$	$2+1=3$	$\frac{2}{3}\cdot\frac{1}{2} - \frac{1}{3}\cdot 3 + 1 = \frac{1}{3}$

Now that we have those in place, we want to determine how much A controls the other variables. That is; how much of B's, C's and D's variance is determined by A. Using (3) you can determine/compute the variance of A, B and C

$\text{var}[A]$	$\text{var}[B]$	$\text{var}[C]$
2	$\frac{1}{2}\cdot 2 + 2 = 3$	$2+1=3$

Determining the variance of D requires computing the correlation between B and C. Alternatively we can derive the final expression of D. Insert the definitions of A, B and C to determine the form of D as a function of N_A, N_B, N_C and N_D

$$
\begin{aligned}
D &\leftarrow \frac{2}{3} \cdot B - \frac{1}{3} \cdot C + N_D \\
&= \frac{2}{3} \cdot \underbrace{\left(\frac{1}{2} \cdot A + N_B\right)}_{B} - \frac{1}{3} \cdot \underbrace{(A + N_C)}_{C} + N_D \\
&= \frac{1}{3} \cdot A + \frac{2}{3} \cdot N_B - \frac{1}{3} \cdot A - \frac{1}{3} \cdot N_C + N_D \\
&= \frac{2}{3} \cdot N_B - \frac{1}{3} \cdot N_C + N_D.
\end{aligned}
$$

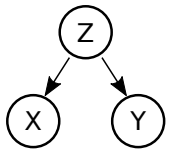
What happened here and why is it interesting?
The signal of A through B and C exactly cancels out, making D completely independent of A!
From the graph we would expect D to be caused by A, but it turns out it is not.
While it may rarely work out this precisely in practice, there are situations where signals can cancel each other out, so that expected causes has almost no influence of effects.

What is the variance of D?

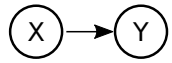
$$\text{var}[D] = \frac{2}{3} \cdot 2 - \frac{1}{3} \cdot 1 + 1 = 1.$$

2.2 Modelling Programs

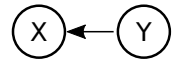
For this exercise we return to the programs of exercise 1.3. The causal graphs which we inferred in that exercise are seen below. Furthermore the data looks Gaussian.



(a) Program 1



(b) Program 2



(c) Program 3

Figure 1

The Programs

We are going to assume that all the programs are normal distributions and sums of normal distributions - which are also normal distributions. First sketch out the program-structures - we've started by writing program 2. We call a_2 the *mixing coefficient* of program 2, as it determines how much X is "mixed into" Y. It does not matter what you name the mixing coefficients etc.

$$
\begin{aligned}
Z_1 &\leftarrow N_{z1}, & N_{z1} &\sim \mathcal{N}(\mu_{z1}, \sigma_{z1}) \\
X_1 &\leftarrow a_{1x} \cdot Z_1 + N_{x1}, & N_{x1} &\sim \mathcal{N}(\mu_{x1}, \sigma_{x1}) \\
Y_1 &\leftarrow a_{1y} \cdot Z_1 + N_{y1}, & N_{y1} &\sim \mathcal{N}(\mu_{y1}, \sigma_{y1})
\end{aligned}
$$

Program 1

$$
\begin{aligned}
X_2 &\leftarrow N_{x2}, & N_{x2} &\sim \mathcal{N}(\mu_{x2}, \sigma_{x2}) \\
Y_2 &\leftarrow a_2 \cdot X_2 + N_{y2}, & N_{y2} &\sim \mathcal{N}(\mu_{y2}, \sigma_{y2})
\end{aligned}
$$

Program 2

$$
\begin{aligned}
Y_3 &\leftarrow N_{y3}, & N_{y3} &\sim \mathcal{N}(\mu_{y3}, \sigma_{y3}) \\
X_3 &\leftarrow a_3 \cdot Y_3 + N_{x3}, & N_{x3} &\sim \mathcal{N}(\mu_{x3}, \sigma_{x3})
\end{aligned}
$$

Program 3

Let's start with the two smaller programs; 2 and 3. Can you come up with a simple strategy for determining the parameters of their normal distributions?

For program 2 we can measure X's distribution directly, and if we intervene and set $X \overset{\circ}{\leftarrow} 0$ we can measure Y's distribution directly. For program 3 it's the other way around.

Plan some experiments (2-9) and run the three programs. Measure the means and variances of the data and store in the table below. Write what interventions you do (for example $X \overset{\circ}{\leftarrow} 3.14$) - if you do any.

Experiment		Program 1		Program 2		Program 3		Interventi
		X	Y	X	Y	X	Y	
1	Mean	1.0	2.0	1.0	2.0	1.0	2.0	-
	Variance	1.0	4.0	1.0	4.0	1.0	4.0	
2	Mean	0.0	2.0	0.0	[0.6]	0.0	2.0	$X \overset{\circ}{\leftarrow} 0$
	Variance	0.0	4.0	0.0	[2.0]	0.0	4.0	
3	Mean	1.0	0.0	1.0	0.0	[0.3]	0.0	$Y \overset{\circ}{\leftarrow} 0$
	Variance	1.0	0.0	1.0	0.0	[0.5]	0.0	
4	Mean							
	Variance							
5	Mean							
	Variance							
6	Mean							
	Variance							
7	Mean							
	Variance							
8	Mean	[0.4]	[0.8]	1.0	2.0	1.0	2.0	$Z \overset{\circ}{\leftarrow} 0$
	Variance	[0.3]	[1.2]	1.0	4.0	1.0	4.0	
9	Mean	[1.6]	[3.2]	1.0	2.0	1.0	2.0	$Z \overset{\circ}{\leftarrow} 2$
	Variance	[0.3]	[1.2]	1.0	4.0	1.0	4.0	

Table 1: Experiments

The numbers of the table are just suggestions. The last rows are for determining program 1, which needs additional interventions. We have marked the places where an intervention on a variable has changed *another* variable - which indicates a causal relationship.

Can you determine the noise-distributions for program 2 and 3?
For example N_{x2} and N_{y2}?

Yes we can! :D
From observational data we can determine the distributions of N_{x2} and N_{y3}.
From the interventional data where $X \overset{\circ}{\leftarrow} 0$ we can determine N_{y2}.
From the data where $Y \overset{\circ}{\leftarrow} 0$ we can determine N_{x3}.

$$N_{x2} \sim \mathcal{N}(1,1) \qquad N_{y2} \sim \mathcal{N}(0.6,2)$$
$$N_{x3} \sim \mathcal{N}(0.3,0.5) \qquad N_{y3} \sim \mathcal{N}(2,4)$$

In order to completely determine programs 2 and 3, we need to determine the mixing coefficients like a_2. Using (3), with the measured means and/or variances, derive the coefficients for programs 2 and 3.

We know what the observational means/variances are, and what the means/variances of the noise-distributions are. From this we can determine the coefficients of the programs by

$$\mathbb{E}[Y_2] = a_2 \cdot \mu_{x2} + \mu_{y2} = a_2 \cdot 1 + 0.6 = 2 \quad \Leftrightarrow \quad a_2 = \frac{2-0.6}{1} = 1.4,$$
$$\mathrm{var}[Y_2] = a_2^2 \cdot \sigma_{x2}^2 + \sigma_{y2}^2 = a_2^2 \cdot 1 + 2 = 4 \quad \Leftrightarrow \quad a_2 = \sqrt{\frac{4-2}{1}} \approx 1.41,$$

$$\mathbb{E}[X_3] = a_3 \cdot \mu_{y3} + \mu_{x3} = a_3 \cdot 2 + 0.3 = 1 \quad \Leftrightarrow \quad a_3 = \frac{1-0.3}{2} = 0.35,$$
$$\mathrm{var}[X_3] = a_3^2 \cdot \sigma_{y3}^2 + \sigma_{x3}^2 = a_3^2 \cdot 4 + 0.5 = 1 \quad \Leftrightarrow \quad a_3 = \sqrt{\frac{1-0.5}{4}} \approx 0.35.$$

Thus we can use either the means or the variances to determine the mixing coefficients - or both to verify our results.

In order to solve problem 1, we need to intervene on the confounder Z. In the programs, it is possible to do intervention on this variable by setting `_z`. Do an intervention experiment on the programs, where you intervene on Z and note the results in the data-table from earlier. Assume the distribution of Z is

$$Z \sim \mathcal{N}(0, 1). \tag{7}$$

Can you determine the noise-distributions of X and Y, as well as the mixing coefficients?
Yes we can.
If we make an experiment where $Z \overset{\circ}{\leftarrow} 0$, then we can read off the noise-distributions directly as

$$N_{x1} \sim \mathcal{N}(0.4, 0.3)$$
$$N_{y1} \sim \mathcal{N}(0.8, 1.2).$$

From the interventional distribution where $Z \overset{\circ}{\leftarrow} 2$ we have

$$\mathbb{E}[X_1] = a_{x1} \cdot \mathring{\mu}_{z1} + \mu_{x1} = a_{x1} \cdot 2.0 + 0.4 = 1.6 \quad \Leftrightarrow \quad a_{x1} = \frac{1.6 - 0.4}{2.0} = 0.6,$$
$$\text{var}[X_1] = a_{x1}^2 \cdot \mathring{\sigma}_{z1}^2 + \sigma_{x1}^2 = a_{x1}^2 \cdot 0.0 + 0.3 = 0.3 \quad \Leftrightarrow \quad 0.3 = 0.3,$$

$$\mathbb{E}[Y_1] = a_{y1} \cdot \mathring{\mu}_{z1} + \mu_{y1} = a_{y1} \cdot 2.0 + 0.8 = 3.2 \quad \Leftrightarrow \quad a_{y1} = \frac{3.2 - 0.8}{2.0} = 1.2,$$
$$\text{var}[Y_1] = a_{y1}^2 \cdot \mathring{\sigma}_{z1}^2 + \sigma_{y1}^2 = a_{y1}^2 \cdot 0.0 + 1.2 = 1.2 \quad \Leftrightarrow \quad 1.2 = 1.2$$

Knowing the mixing coefficients we can now use the observational means and variances to find

$$\mathbb{E}[X_1] = a_{x1} \cdot \mu_{z1} + \mu_{x1} = 0.6 \cdot \mu_{z1} + 0.4 = 1.0 \quad \Leftrightarrow \quad \mu_{z1} = \frac{1.0 - 0.4}{0.6} = 1.0,$$
$$\text{var}[X_1] = a_{x1}^2 \cdot \sigma_{z1}^2 + \sigma_{x1}^2 = 0.6^2 \cdot \sigma_{z1}^2 + 0.3 = 1.0 \quad \Leftrightarrow \quad \sigma_{z1}^2 = \frac{1.0 - 0.3}{0.6^2} = 1.95,$$

$$\mathbb{E}[Y_1] = a_{y1} \cdot \mu_{z1} + \mu_{y1} = 1.2 \cdot \mu_{z1} + 0.8 = 2.0 \quad \Leftrightarrow \quad \mu_{z1} = \frac{2.0 - 0.8}{1.2} = 1.0,$$
$$\text{var}[Y_1] = a_{y1}^2 \cdot \sigma_{z1}^2 + \sigma_{y1}^2 = 1.2^2 \cdot \sigma_{z1}^2 + 1.2 = 4.0 \quad \Leftrightarrow \quad \sigma_{z1}^2 = \frac{4.0 - 1.2}{1.2^2} = 1.95$$

The true values used in the program is $\mu_{z1} = 1.0$ and $\sigma_{z1}^2 = 2.0$, which can be found if we use higher precision.

2.3  An Advanced Switch  

Make a switch (or two?) with more lights - say 4. We don't draw out all possible hypotheses and we don't make the tree.
They start with zero and we expect them to find causal model with equations.
Note that they have already tried inferring such probabilities in Exercise 1, we just didn't tell them.

2.4 Counterfactuals 1

Consider the following causal model

$$\begin{aligned} X &\leftarrow N_x & N_x &\sim \mathcal{B}\left(\frac{7}{10}\right) \\ Y &\leftarrow N_y & N_y &\sim \mathcal{C}\left(\left[\frac{1}{5}, \frac{3}{5}, \frac{1}{5}\right]\right) \\ Z &\leftarrow X + Y - N_z & N_z &\sim \mathcal{C}\left(\left[\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right]\right) \end{aligned}$$

All Known

Say we observe $x = 1, y = 1, z = 0$.
Determine the value of z, given that we counterfactually do $x \overset{\bullet}{\leftarrow} 0$.

First we determine all the noise-variables we can.
From the values of x and y we have

$$X = N_x = 1, \qquad Y = N_y = 1.$$

From this we find that

$$Z = X + Y - N_z = N_x + N_y - N_z = 1 + 1 - N_z = 0 \quad \Leftrightarrow \quad N_z = 2.$$

So the counterfactual z will be

$$z = \dot{N}_x + N_y - N_z = 0 + 1 - 2 = -1.$$

This is hardly a counterfactual question; we know all information.

Unknown z

Now we observe $x = 1, y = 1$.
Determine the probability of $z = 1$, given that we counterfactually do $x \overset{\bullet}{\leftarrow} 0$.

$$p(\dot{z} = 1 \mid x \overset{\bullet}{\leftarrow} 0, x = 1, y = 1).$$

We know all variables determining z, except for one; the noise-variable N_z. We can marginalize over this variable in order to determine the probability of $z = 1$

$$p(\dot{z} = 1 \mid x \overset{\bullet}{\leftarrow} 0, x = 1, y = 1) = \sum_{i=0}^{2} p(z = 1 \mid x = 0, y = 1, N_z = i) \cdot p(N_z = i)$$
$$= p(z = 1 \mid x = 0, y = 1, N_z = 0) \cdot p(N_z = 0) = 1 \cdot \frac{2}{5} = \frac{2}{5}.$$

This also, is not really a counterfactual question; we are simply computing the forward probabilities of z.

Unknown y

Now say we observe $x = 1, z = 0$.
We wish to determine the counterfactual probability

$$p(\dot{z} = 0 \mid x \overset{\bullet}{\leftarrow} 0, z = 0).$$

In order to determine this probability, we need to determine the conditional distributions of N_z and N_y. We know what x is, so what is the constraint on the remaining variables influencing Z?
We have

$$z = 1 + Y - N_z = 1 + N_y - N_z = 0 \quad \Leftrightarrow \quad N_y + 1 = N_z.$$

Note down all values for N_y and N_z which fulfils this constraint, and their probabilities

N_y	N_z	$p(N_y, N_z)$ $= p(N_y) \cdot p(N_z)$
0	1	$\frac{1}{5} \cdot \frac{2}{5} = \frac{2}{25}$
1	2	$\frac{3}{5} \cdot \frac{1}{5} = \frac{3}{25}$

Now marginalize over N_y and N_z to determine the counterfactual probability

$$
\begin{aligned}
p(\overset{\bullet}{z} = 0 \mid x \overset{\bullet}{\leftarrow} 0, z = 0) &= \sum_{i=0}^{2} \sum_{j=0}^{2} p(z = 0 \mid x = 0, N_y = i, N_z = j) \cdot p(N_y = i, N_z = j) \\
&= p(z = 0 \mid x = 0, N_y = 0, N_z = 1) \cdot p(N_y = 0, N_z = 1) \\
&\quad + p(z = 0 \mid x = 0, N_y = 1, N_z = 2) \cdot p(N_y = 1, N_z = 2) \\
&= 1 \cdot \frac{2}{25} + 1 \cdot \frac{3}{25} = \frac{5}{25}.
\end{aligned}
$$

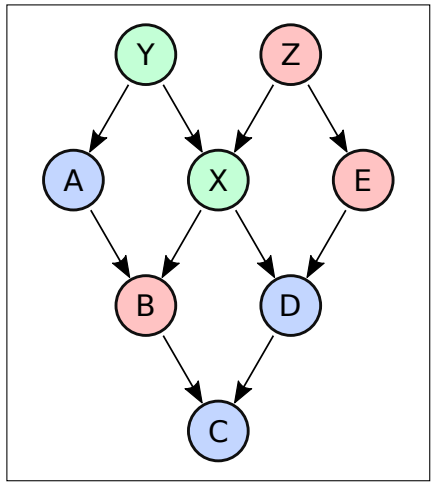
2.5 Counterfactuals 2

The example is too boring. If they are all additions, then nothing interesting happens.

Consider the following causal model

$$
\begin{aligned}
N_x, N_y &\sim \mathcal{B}\left(\tfrac{1}{3}\right) \\
N_a, N_c, N_d &\sim \mathcal{B}\left(\tfrac{2}{3}\right) \\
N_z, N_b, N_e &\sim \mathcal{C}\left(\left[\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{2}{4}\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
Y &\leftarrow N_y & Z &\leftarrow N_z \\
A &\leftarrow Y + N_a & X &\leftarrow Y + Z + N_x \\
E &\leftarrow Z + N_e & B &\leftarrow A + X + N_b \\
D &\leftarrow X + E + N_d & C &\leftarrow B + D + N_c
\end{aligned}
$$



Round 1

We want to determine

$$p\big(\dot{d} \,|\, x \dot{\leftarrow} 0, x = 1, z = 1\big).$$

First write out D as an expression of noise-variables and known variables and insert the counterfactual variables

$$D \leftarrow X + E + N_d = X + Z + N_e + N_d.$$

For the counterfactual case we have

$$\dot{D} \leftarrow \dot{x} + z + N_e + N_d = 0 + 1 + N_e + N_d = 1 + N_e + N_d.$$

Now we can make a table of values of N_d and N_e, together with the values for D and their probabilities

N_d	N_e	D	$p(N_d, N_e)$	$= p(N_d) \cdot p(N_e)$
0	0	1	$1/3 \cdot 1/4 = 1/12$	
0	1	2	$1/3 \cdot 1/4 = 1/12$	
0	2	3	$1/3 \cdot 2/4 = 2/12$	
1	0	2	$2/3 \cdot 1/4 = 2/12$	
1	1	3	$2/3 \cdot 1/4 = 2/12$	
1	2	4	$2/3 \cdot 2/4 = 4/12$	

What is the distribution of D?

$D \sim \mathcal{C}\left(\left[0, \frac{1}{12}, \frac{3}{12}, \frac{4}{12}, \frac{4}{12}\right]\right)$.

Round 2

We want to determine

$$p\big(\overset{\bullet}{d} \mid x \overset{\bullet}{\leftarrow} 0, x = 1, z = 1, d = 2\big).$$

You can utilize a lot of information from Round 1. First of all we already know D as an expression of noise-variables. Furthermore we have the table of N_d, N_e and D values. Find all rows of the table where $d = 2$. Note them below and compute the conditional probabilities

N_d	N_e	$p(N_d, N_e)$	$p(N_d, N_e \mid d = 2)$	D
0	1	$1/12$	$1/3$	2
1	0	$2/12$	$2/3$	2

What is the distribution of D?
D can only take on the value 2.

Round 3

We want to determine

$$p\big(\overset{\bullet}{d} \mid x \overset{\bullet}{\leftarrow} 0, x = 1, d = 3\big).$$

We have information about D, but not E and Z. Determine the values of the noise-variables N_d, N_e and N_z, for which $D = 3$. Note their probabilities and conditional probabilities

N_d	N_e	N_z	$p(N_d, N_e, N_z)$	$p(N_d, N_e, N_z \mid d = 3)$
1	1	0	$2/3 \cdot 1/4 \cdot 1/4 = 1/24$	$2/9$
1	0	1	$2/3 \cdot 1/4 \cdot 1/4 = 1/24$	$2/9$
0	1	1	$1/3 \cdot 1/4 \cdot 1/4 = 1/48$	$1/9$
0	2	0	$1/3 \cdot 2/4 \cdot 1/4 = 1/24$	$2/9$
0	0	2	$1/3 \cdot 1/4 \cdot 2/4 = 1/24$	$2/9$

2.6  Counterfactuals 3 

Round 4

Determine

$$p\big(\dot{c} \mid x \twoheadleftarrow 3, x = 1, z = 1, y = 2, c = 16\big)$$

Round 5

Determine

$$p\big(\dot{c} \mid x \twoheadleftarrow 3, x = 1, c = 16\big)$$

IN PROGRESS

Notes and ideas

1 Motivation/real world examples

1. Spurious collections: `http://www.tylervigen.com/spurious-correlations`
   For example "Divorce rate in Maine" and "Per capita consumption of margarine"
2. The whole story of smoking from the 1950s? Book of Why goes quite a bit into this.
   Silvia also uses this as example of confounder (the alternative theory in the 50s)
   - Slide 66-69 from Silvia's slides have an interesting story, which builds on the smoking story. Could be used to show how one persons interventions becomes another persons prior knowledge.
3. "An Inconvenient Truth"
   Al Gore makes a point when comparing two lines in
   `http://smallpond.ca/jim/ref/inconvenientTruth/full/00_23_53.jpg`
   Is he correct that there is a causative link?
   Is climate change the "new smoking?"
4. Yield of field
5. From Silvia (slide 4): "Eating makes you faithful" and "Relaxing makes you die"
6. Alternative project idea
   - Use the example from Klaus-Robert Mullers paper on Relevance propagation

- Make 10 image databases - some of them have problems
- Train algorithm and use explainability (relevance propagation) to find problems
- Eliminate problematic databases

2 Examples usable with models/-graphs

1. Switch example from Pedro Ortega - with the trees for plausible models
2. Man writing down a number/name and making a calligraphy of that number/name
   - Used in "Elements of Causal Inference"
   - Ask the man for a specific number change both
   - Replace calligraphy of write-down and change only one
3. Barometer: rain influences barometer but not the other way around
4. Light switch example: apparent loss of independence
   Silvia page 33, Spirtes et al. 2000 (page 59).
   The independence is moved because we now also model some of the "noise".
   This is a really good point!
5. Simpsons paradox (of cause)
   - Solve using intervention (do-operator)
   - Silvia does this, but I think we can explain it simpler
   - Use the Berkeley example: `https://en.wikipedia.org/wiki/Simpson%27s_paradox#UC_Berkeley_gender_bias`

3 Exercises

- Intervention on programs

1. Do something like in inFERENCe `https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/`
2. Make some programs that produce the same distributions
3. Allow students to do intervention on the programs (via some api) to get interventional distributions
4. Have the students identify the generating process (they can actually identify the generating distributions)

- Counterfactuals
  1. inFERENCe 3 makes an example with a beard - perhaps change the example but the idea is good.
  2. Present students with a small table of datapoints, a causal graph and a set of structural equations.
  3. Have them compute counterfactual probabilities from this information
  4. **advanced:** make then infer some of the probability distributions in the structural equations first with interventional data (perhaps given beforehand).
- Physics simulator? Climate simulator? Do an experiment in the simulators and show causality.
- Perhaps a simulator of the body or medicine is better, because you do not intuitively know the answer.
- A project idea could be to plan an experiment based on some causal assumptions etc., and submit the experiment to us. Once the experiment is submitted it "runs" on a server and they get the results.