Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2020

11 Structural Causal Models

So far, we've explored undirected graphical models and Bayesian networks. A given graphical model is used to describe a family of distributions that satisfy some set of conditional independence statements. These models have permitted us to develop graphically-motivated algorithms for answering **statistical** questions, like:

- "what is the most likely position of my vehicle?"
- "what is the probability that message X was sent?"

In addition, we would like to utilize them to answer questions of causal inference:

- "is surgery or chemotherapy a better treatment for breast cancer?"
- "what is the effect of hours slept per night on GPA for university students?"

These questions require that graphical models capture how the probability distributions of some variables *change* in response to other variables being manipulated. In the first example, we need to be able to compute the distribution of some variable measuring how good a treatment is, say "complete remission", in response to the treatment being held constant at either surgery or chemotherapy. In the second example, we need to be able to compute the distribution of the variable "GPA" in response to sleep hours being manipulated. As we shall discuss in this lecture, the directed graphical model with *appropriate* graph structure that is consistent with underlying causal relationships, also called *causal DAG*, will help answer such questions.

Before we introduction causal DAG, let us start with simple example that help us understand notion of causal mechanism or causal structure equations or models. As an important by product, it will help us understand that the distribution after conditioning (i.e. our updated belief after *seeing* a variable take some value) is not the same as the distribution after manipulating (i.e. our updated belief after *making* a variable take that same value). Let $S \in \{0, 1\}$ represent whether or not I'm sick today, and $D \in \{0, 1\}$ represent whether or not I go to the doctor tomorrow. Structurally, we have

$$S = N_S \tag{1}$$
$$D = SD^1 + (1 - S)D^0 \tag{2}$$

where N_S is independent random variable with Bernoulli distribution with pararmeter $\frac{1}{2}$, D^0, D^1 are independent Bernoulli variables with parameters $\frac{1}{5}$ and $\frac{4}{5}$ respectively. In effect (1) represents the *ground truth* mechanism that connects S and D and it leads to natural observations of it. This causal mechanism induces the following joint distribution:

$$P_{SD}(s,d) = .4\mathbb{1}_{s=d} + .1\mathbb{1}_{s\neq d} \tag{3}$$

which induces marginals and conditionals as

- $P_S(s) = .5$, $P_D(d) = .5$ and
- $P_{S|D}(s|d) = .8\mathbb{1}_{s=d} + .2\mathbb{1}_{s\neq d}$, $P_{D|S}(d|s) = .8\mathbb{1}_{d=s} + .2\mathbb{1}_{d\neq s}$.

The conclusions from the joint distribution (and marginals / conditionals) are

1. On average, 50% of days I am sick and 50% of days I visit doctor.
2. If I fell sick today, i.e. $s = 1$, then there is 80% chance that I'll see doctor tomorrow.
3. And if you saw me visit doctor today, i.e. $d = 1$, there is a 80% chance that I fell sick yesterday.
4. That is, S and D are "correlated".

In addition to these conclusions, using the underlying causal mechanism implied by (1), one derives the following conclusions:

5. If I was made sick today (naturally or artificially induced), I will go and see doctor tomorrow with 80% chance (assuming no other active intervention beyond making me sick is in place).
6. If I was made to go see doctor today (e.g. by offering \$100 to me if I made a visit), the likelihood of me being sick yesterday is 50%.

In above conclusions, we have looked at three *different* distributions. The first being *observational* distribution (3) that is used for conclusions (1)-(4). The second being *interventional* distribution with intervention $S = 1$ leading to conclusion (5). Finally, *interventional* distribution with intervention $D = 1$ leading to conclusion (6).

The conclusion (5) states that the distribution of D condition on $S = 1$ under (3) is the same as the distribution of D when S is intervened to 1. And this is implied by causal mechanism (1). It utilize the assumption known as *independence of cause and mechanism*, i.e. the causal mechanism of (1) (specifically, with S impacting D) continues to hold independent of how S is set to 1. The conclusion (6) effectively states that S is generated as per (1) and hence by forcing D to take value 1 does not change the marginal distribution of S.

In summary, the observational distribution between S, D can be captured by either DAG: $S \to D$ or $D \to S$. But only $S \to D$ captures the causal mechanism as suggested by (1)-(3). And it provides a way to reason about distribution induced under intervention. To that end, we start by a definition.

Definition 1. *The distribution of X after setting the variable Y to the value y is the* interventional distribution *given the* intervention $Y = y$*, and is denoted by* $P_X^{do(Y=y)}$.

In the example above, we discussed $P_D^{do(S=1)}$ and $P_S^{do(D=1)}$. It is important to remark that $P_X^{do(Y=y)}$ can be very different from P_X in general. However, as discussed in the above example, if we know the causal mechanism precisely, i.e. (1)-(3), there is a potential to identify the $P_X^{do(Y=y)}$ from the observational distribution of X, Y. We discuss this next.

11.1 Causal DAGs

We'd like to be able to ask any question about the interventional distribution for any possible intervention. We'll denote this enormous set of distributions by P^*:

$$P^* = \{P_{X_\mathcal{V}}^{do(X_I=x_I)} | I \subseteq \mathcal{V}, x_I \in |\mathcal{X}|^{|I|}\}$$

Here denoted $X_\mathcal{V}$ is the set of all of our random variables and we assume that they all come from the same alphabet $\mathcal{X}$ (the only reason to do this is convenience for the notation, otherwise we might want to say "for all assignments of x_I"). Finally, note that when $I = \varnothing$ there are no interventions, so in this case we have our original distribution which we call the *observational distribution*.

Definition 2. P^* *is compatible with a* causal DAG *(aka a* causal Bayesian network*)* G *if,* $\forall$ $P_{x_\mathcal{V}}^{do(X_I=x_I)} \in P^*$*, the following three axioms are satisfied:*

1. ***Factorization*** $P_{X_\mathcal{V}}^{do(X_I=x_I)}$ *factorizes with respect to* G
2. ***Intervention*** $\forall$ $i \in I$, $P_{X_i}^{do(X_I=x'_I)}(x_i) = \mathbb{1}_{x_i=x'_i}$
3. ***Invariance*** $\forall$ $i \notin I$, $P_{X_i|X_{\pi_i}}^{do(X_I=x'_I)}(x_i|x_{\pi_i}) = P_{X_i|X_{\pi_i}}(x_i|x_{\pi_i})$ *as long as* x_{π_i} *is consistent with* x'_I

The Factorization Axiom tells us first that the *observational* distribution factorizes with respect to the DAG G, so that it satisfies all conditional independence statements implied by d-separation in G. Then, it tells us that the distributions we get after interventions still factorize with respect to G, so they don't suddenly have new dependences.

The Intervention Axiom simply tells us that our interventions are precise: when we say that we *set* the X_i to x_i, we mean that it deterministically takes the value x_i. There are other notions of interventions that allow the new value of X_i to be non-deterministic (say, drawn from a Gaussian centered at x_i), but we won't deal with those in this class.

The Invariance Axiom is a formal statement of independence of cause and mechanism: if a variable x_i is taking its values naturally, i.e., it is not intervened on, it still has the same mechanism: the same distribution of values given its parents x_{π_i}. Note the combination of conditioning and intervening on the lefthand side of the equality. All this means is that we're conditioning in the interventional distribution. It's even okay for the conditioning set and the set of intervened variables to contain some (or all) of the same variables.

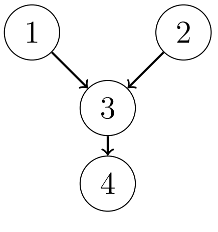
These axioms allow us to derive *any* interventional distribution $P_{X_\mathcal{V}}^{do(X_I=x'_I)}$ from just the observational distribution and the DAG G.

$$\begin{aligned}
P_{X_\mathcal{V}}^{do(X_I=x'_I)}(x_\mathcal{V}) &= \prod_{i\in\mathcal{V}} P_{X_i|X_{\pi_i}}^{do(X_I=x'_I)}(x_i|x_{\pi_i}) && (4)\\
&= \prod_{i\in I} P_{X_i|X_{\pi_i}}^{do(X_I=x'_I)}(x_i|x_{\pi_i}) \prod_{i\notin I} P_{X_i|X_{\pi_i}}^{do(X_I=x'_I)}(x_i|x_{\pi_i}) \\
&= \prod_{i\in I} P_{X_i|X_{\pi_i}}^{do(X_I=x'_I)}(x_i|x_{\pi_i}) \prod_{i\notin I} P_{X_i|X_{\pi_i}}(x_i|x_{\pi_i}) && (5)\\
&= \prod_{i\in I} \mathbb{1}_{x_i=x'_i} \prod_{i\notin I} P_{X_i|X_{\pi_i}}(x_i|x_{\pi_i}) && (6)\\
&= \mathbb{1}_{x_I=x'_I} \prod_{i\notin I} P_{X_i|X_{\pi_i}}(x_i|x_{\pi_i})\,. && (7)
\end{aligned}$$

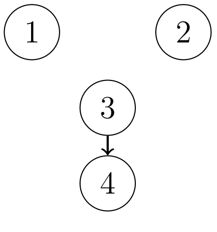
Here (4) follows from the Factorization Axiom, (5) follows from the Invariance Axiom, and (6) follows from the Intervention Axiom and the fact that a deterministic marginal implies a deterministic conditional. The final form (7) will be referred to as the *truncated factorization*, since it is the same factorization as the observational distribution without the conditional probabilities of the intervened nodes.

Note that this interventional distribution factorizes with respect to a subgraph $G' = (\mathcal{V}, \mathcal{E}')$ of $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E}' = \mathcal{E} \setminus \{(i,j) \mid j \in I\}$, i.e., all of the edges coming into the intervened nodes are removed. The process of removing these edges is sometimes called *graph surgery*, and the new graph is referred to as the *mutilated graph*, but we'll refer to it just as the *interventional graph* and denoted the intervened nodes with a superscript.

For example, given an intervention at node 3 in G below, the interventional graph $G^{(\{3\})}$ is as depicted.



G



$G^{(\{3\})}$

Returning to our example of the model for S and D, we first observe that the observational distribution P_{SD} can be factored in two ways: $P_{SD} = P_S P_{D|S}$, corresponding to the DAG G_1 in Fig. 1 and $P_{SD} = P_D P_{D|S}$, corresponding to the DAG G_2 in Fig. 2. Both DAGs are valid Bayesian networks for our joint distribution, but only G_1 gives us the correct interventional distributions.

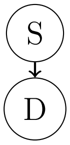


Figure 1: G_1

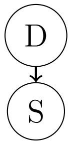


Figure 2: G_2

For G_1:

$$\begin{aligned}
P_{SD}^{do(D=1)}(s,d) &= \mathbb{1}_{d=1} P_S(s) = .5\mathbb{1}_{d=1} \\
P_{SD}^{do(S=1)}(s,d) &= \mathbb{1}_{s=1} P_{D|S}(d|s) = \mathbb{1}_{s=1}(.8\mathbb{1}_{d=1} + .2\mathbb{1}_{d=0})
\end{aligned}$$

While for G_2:

$$\begin{aligned}
P_{SD}^{do(D=1)}(s,d) &= \mathbb{1}_{d=1} P_{S|D}(s|d) = \mathbb{1}_{d=1}(.8\mathbb{1}_{s=1} + .2\mathbb{1}_{s=0}) \\
P_{SD}^{do(S=1)}(s,d) &= \mathbb{1}_{s=1} P_D(d) = .5\mathbb{1}_{s=1}
\end{aligned}$$

A final consequence of these axioms is another version of independence of cause and mechanism: if we do some combination of conditioning and intervening on the parents of a node, the conditional distribution of that node in the interventional distribution is the same as if we had seen the parents take those values in the observational distribution. Formally:

Corollary 1. *If $\pi_i = A \cup B$ and $A \cap B = \varnothing$ then*

$$P_{X_i|X_A}^{do(X_B=x'_B)}(x_i|x_A) = P_{X_i|X_{\pi_i}}(x_i|x_A, x'_B)$$

Proof. From the law of total probability, i.e. $P_X(x) = \sum_b P_Y(y)P_{X|Y}(x|y)$, we have

$$\begin{aligned}
P_{X_i|X_A}^{do(X_B=x'_B)}(x_i|x_A) &= \sum_{x_B} P_{X_B}^{do(X_B=x'_B)}(x_B) P_{X_i|X_{\pi_i}}^{do(X_B=x'_B)}(x_i|x_A, x'_B) \\
&= \sum_{x_B} \mathbb{1}_{x_B=x'_B} P_{X_i|X_{\pi_i}}^{do(X_B=x'_B)}(x_i|x_A, x'_B) \qquad (8) \\
&= P_{X_i|X_{\pi_i}}^{do(X_B=x'_B)}(x_i|x_A, x'_B) \\
&= P_{X_i|X_{\pi_i}}(x_i|x_A, x'_B) \qquad (9)
\end{aligned}$$

(8) follows from the Intervention Axiom and (9) follows from the Invariance Axiom. □

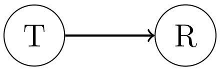


Figure 3: Model 1 for kidney stone example

Stronger versions of this statement holds (e.g. A and B can overlap as long as the values conditioned on and the values of the intervention are the same, $A \cup B$ could be a superset of π_i as long as A doesn't contain descendants of i, etc.), but we will not need these here.

11.2 Estimating Causal Effect

Given a causal DAG and an observational distribution, we can begin to answer our causal questions like "is Treatment A or Treatment B better for this disease?" or "how does gender affect university admission rate?". A famous puzzle known as Simpson's paradox highlights the difficulty of this question.

A 1986 medical study compared the efficacy two treatments (denoted by T): open surgical procedures (Treatment A) and a minimally-invasive procedure called percutaneous nephrolithotomy (Treament B) for alleviating kidney stones. The efficacy of each treatment was based on the recovery (R) of the patients, with $R = 1$ denoting that the patients had no kidney stones 3 months after treatment and $R = 0$ denoting that they still had kidney stones. As a doctor, we might be interested in the quantities $P_{R|do(T=A)}(1)$ and $P_{R|do(T=B)}(1)$ when deciding which of the two treatments to give a new patient. In fact, their difference $P_{R|do(T=A)}(1) - P_{R|do(T=B)}(1) = \mathbb{E}_{do(T=A)}[R] - \mathbb{E}_{do(T=B)}[R]$ is a standard measure known as the *average treatment effect* (ATE).

The study offered hopeful results: 289 out of 350 patients receiving Treatment B were successfully treated, while only 273 out of 350 patients receiving Treatment A had successful results. That is,

$$P_{R|T}(1|A) = \frac{273}{350} \cong .78$$

$$P_{R|T}(1|B) = \frac{289}{350} \cong .83$$

Given the causal DAG in Fig. 3, we have $P_{R|do(T=A)}(1) = P_{R|T}(1|A)$, and we may conclude that we should prefer treatment A.

However, when stratifying patients according to the diameter D of their kidney stones (ℓ for large vs. s for small), Treatment A was found to be better in both strata: of patients with small stones, 81 out of 87 (93%) had successful results under Treatment A but only 234 out of 270 (87%) had successful results for Treatment B. Similarly, out of patients with large stones, 192 out of 263 (73%) of patients had successful results with Treatment A but only 55 out of 80 (69%) has successful results with Treatment B. That is,

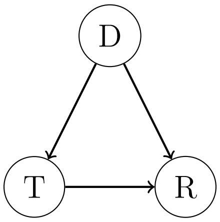


Figure 4: Model 2 for kidney stone example

$$P_{R|T,D}(1|A,s) = \frac{81}{87} \cong .93$$
$$P_{R|T,D}(1|B,s) = \frac{234}{270} \cong .87$$
$$P_{R|T,D}(1|A,\ell) = \frac{192}{263} \cong .73$$
$$P_{R|T,D}(1|B,\ell) = \frac{55}{80} \cong .69$$

Apparently, the size of the kidney stones has an effect on the probability of recovery: patients with larger kidney stones are less likely to recover. Astutely, we might also notice that the size of the kidney stone seems to have an effect on which treatment a patient is given. If the doctors already expected Treatment A to perform better, perhaps they preferentially gave it to patients with large kidney stones, who required a better treatment. Thus, we may reasonably develop a new model where D is a cause of both R and T, as in Fig. 4.

Computing $P_R^{do(T=A)}(1)$ from observational quantities is more complicated in this model:

$$\begin{aligned}
P_R^{do(T=A)}(R=1) &= \sum_d P_{R|D}^{do(T=A)}(1|d) P_D^{do(T=A)}(d) && (10)\\
&= \sum_d P_{R|D}^{do(T=A)}(1|s) P_D(d) && (11)\\
&= \sum_d P_{R|D,T}(1|d,A) P_D(d) && (12)\\
&= \frac{192}{263}\frac{343}{700} + \frac{81}{87}\frac{357}{700} && (13)\\
&\cong .83\,, && (14)
\end{aligned}$$

where (10) is simply marginalization over the joint distribution $P_{R,D}^{do(T=A)}$, (11) follows from the Invariance Axiom and the fact that D has no parents in G, and (12) follows from independence of cause and mechanism. Observe the form of (12): the overall interventional distribution $P_R^{do(T=A)}(1)$ is computed as a mixture of the conditional distribution $P_{R|D,T}$ for each strata of patients, with mixture weights P_D given by the relative sizes of each strata.

A similar calculation shows that for $P_R^{do(T=B)}(1) = \frac{55}{80}\frac{343}{700} + \frac{234}{270}\frac{357}{700} \cong .78$. According to these calculations, we conclude that treatment A is the better treatment overall. This matches the common statistical wisdom that one should condition on confounding variables; however, not all available variables should be conditioned on (e.g. we don't want to conclude smoking doesn't cause premature death because we condition on the variable that mediates this effect, the presence of lung cancer), and a causal DAG lets us read off exacly which sets of variables are okay to condition on:

Definition 3. *Given an causal DAG over nodes $\mathcal{V}$ and two nodes X_i and X_j s.t. $X_j \notin \pi_{X_i}$, $X_A \subset V \setminus \{X_i, X_j\}$ is a* valid adjustment set *for (X_i, X_j) if*

$$P_{X_j}^{do(X_i=x_i)}(x_j) = \sum_{x_A} P_{X_j|X_i,X_A}(x_j|x_i,x_A)P_{X_A}(x_A)$$

Proposition 1 (Parent Adjustment). *π_{X_i} is a valid adjustment set of (X_i, X_j)*

Proof.

$$\begin{aligned} P_{X_\mathcal{V}}^{do(X_i=x_i')}(x_\mathcal{V}) &= \mathbb{1}_{x_i=x_i'} \prod_{j\neq i} P_{X_j|X_{\pi_j}}(x_j|x_{\pi_j}) \\ &= \mathbb{1}_{x_i=x_i'} \frac{P_{X_\mathcal{V}}(x_\mathcal{V})}{P_{X_i|X_{\pi_i}}(x_i|x_{\pi_i})} \\ &= \mathbb{1}_{x_i=x_i'} \frac{P_{X_\mathcal{V}}(x_\mathcal{V})P_{X_{\pi_i}}(x_{\pi_i})}{P_{X_i,X_{\pi_i}}(x_i,x_{\pi_i})} \\ &= \mathbb{1}_{x_i=x_i'} P_{X_\mathcal{V}|X_{\pi_i}}(x_\mathcal{V}|x_{\pi_i},x_i)P_{X_{\pi_i}}(x_{\pi_i}) \end{aligned} \tag{15}$$

Where (15) comes from multiplying and dividing by $P_{X_i|X_{\pi_i}}(x_i|x_{\pi_i})$.

By marginalizing over $X_{\mathcal{V}\setminus\{X_j\}}$, we obtain, as desired,

$$\begin{aligned} P_{X_j}^{do(X_i=x_i)}(x_j) &= \sum_{x_{\mathcal{V}\setminus\{j\}}} \mathbb{1}_{x_i=x_i'} P_{X_\mathcal{V}|X_{\pi_i}}(x_\mathcal{V}|x_{\pi_i},x_i)P_{X_{\pi_i}}(x_{\pi_i}) \\ &= \sum_{x_{\mathcal{V}\setminus\{i,j\}}} P_{X_\mathcal{V}|X_{\pi_i}}(x_\mathcal{V}|x_{\pi_i},x_i')P_{X_{\pi_i}}(x_{\pi_i}) \\ &= \sum_{x_{\pi_i}} P_{X_{\pi_i}}(x_{\pi_i}) \sum_{x_{\mathcal{V}\setminus\{i,j\}\setminus\pi_i}} P_{X_\mathcal{V}|X_{\pi_i}}(x_\mathcal{V}|x_{\pi_i},x_i') \\ &= \sum_{x_{\pi_i}} P_{X_{\pi_i}}(x_{\pi_i})P_{X_j|X_i,X_{\pi_i}}(x_j|x_i',x_{\pi_i}) \end{aligned}$$

□

However, this requires that $P_{X_j|X_i,X_{\pi_i}}$ and $P_{X_{\pi_i}}$ be known. In some cases, this might not be true. For example, the size of the kidney stones might not be a direct cause of the assigned treatment, but it might effect the assignment through the variable S, the doctor's sympathy. This situation is depicted in Fig. 5. The following proposition provides an alternative way of forming adjustment sets from the data that is available.

Proposition 2 (Backdoor Adjustment)**.** *$X_A \subset V \setminus \{X_i, X_j\}$ is a valid adjustment set for (X_i, X_j) if X_A is a* backdoor *with respect to (X_i, X_j), that is, if:*

- *X_A contains no descendants of X_i*
- *X_A blocks all paths from X_i to X_j entering with arcs into X_i*

Proof. First, note that $X_i \perp X_A | X_{\pi_i}$ since a node is independent of non-descendants given its parents. Next, we show that $X_j \perp\!\!\!\perp X_{\pi_i} \mid X_i, X_A$. Consider $k \in \pi_i$. For any path from i to k, we have two cases:

1. i is on the path: then conditioning on i blocks the path
2. i is not on the path: then we can add $k \to i$ to the path, in which case by the definition of A, the path is blocked when conditioning on A. We need to show that adding i to the conditioning set does not open the path. This happens if i is a descendant of some collider c on the path, and that is the path is not blocked in any other way. However, in this case there would be a backdoor path $j \rightsquigarrow c \to \ldots \to i$ that is unblocked, but this must be blocked by A. So the path is blocked some other way.

Conditioning on i blocks all paths between k and j passing through i. By the definition of A, all other paths from k to j are blocked by A. So all of π_i are d-separated from j by A and i.

To prove that X_A is a valid adjustment, we must show that $P_{X_j}^{do(X_i=x_i')}(x_j) = \sum_{x_A} P_{X_A}(x_A) P_{X_j|X_i,X_A}(x_j|x_i, x_A)$. We'll drop subscripts on the distributions to save

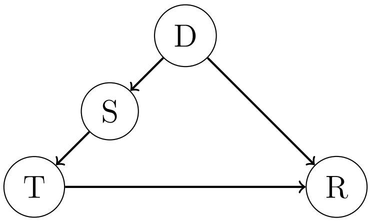


Figure 5: Model 3 for kidney stone example

space, since it's clear what distribution it is from the argument.

$$\begin{aligned}
P^{do(X_i=x_i')}(x_j) &= \sum_{x_{\pi_i}} P(x_{\pi_i}) P(x_j|x_i', x_{\pi_i}) && (16)\\
&= \sum_{x_{\pi_i}} P(x_{\pi_i}) \sum_{x_A} P(x_A|x_i', x_{\pi_i}) P(x_j|x_i', x_{\pi_i}, x_A) && (17)\\
&= \sum_{x_{\pi_i}} P(x_{\pi_i}) \sum_{x_A} P(x_A|x_{\pi_i}) P(x_j|x_i', x_{\pi_i}, x_A) && (18)\\
&= \sum_{x_{\pi_i}} P(x_{\pi_i}) \sum_{x_A} P(x_A|x_{\pi_i}) P(x_j|x_i', x_A) && (19)\\
&= \sum_{x_A} P(x_j|x_i', x_A) \sum_{x_{\pi_i}} P(x_{\pi_i}) P(x_A|x_{\pi_i}) && (20)\\
&= \sum_{x_A} P(x_j|x_i', x_A) P(x_A) && (21)
\end{aligned}$$

Where (16) comes from the parent adjustment formula, (17) follows from the law of total probability, (18) follows from $X_i \perp\!\!\!\perp X_A | X_{\pi_i}$, and (19) follows from $X_j \perp\!\!\!\perp X_{\pi_i} \mid X_i, X_A$.

□

11.2.1 Randomized Control

As we saw in the kidney stone example, understanding the effect of different treatments boils down to understanding the causal mechanisms of the causal DAG. First, we started with Model 1 (Figure 3), and calculated the implications of this DAG. Then, we added a confounding variable D in Model 2 (Figure 4), and because we knew the distribution of D, we could calculate the new implications. What if there are confounders that we don't about, or we don't want to consider in our calculations? How can we account for unknown confounding variables in that case?

To explore this, let's use the same causal DAG as Model 2 in Figure 4. Let T be a binary variable, $T \in \{0, 1\}$, and let $R, D \in \mathbb{R}$. We can write the distribution R as:

$$R = R^0(1 - T) + R^1 T + D + \varepsilon \tag{22}$$

where R^0 is the distribution of R when $T = 0$, R^1 is the distribution of R when $T = 1$, and ε is a independent zero-mean error term. This expression shows that R is a combination of the causal effects from T and D. For simplicity, we can ignore ε in our calculations. We want to understand the average effect of each treatment, which is equivalent to $\mathbb{E}^{do(T=1)}[R] = \mathbb{E}[R|T = 1] = \mathbb{E}[R^1] + \mathbb{E}[D|T = 1]$ and $\mathbb{E}^{do(T=0)}[R] = \mathbb{E}[R^0] + \mathbb{E}[D|T = 0]$. The difference between the average effects of treatments is then

$$\mathbb{E}^{do(T=1)}[R] - \mathbb{E}^{do(T=0)}[R] = \mathbb{E}[R^1] - \mathbb{E}[R^0] + \mathbb{E}[D|T = 1] - \mathbb{E}[D|T = 0].$$

In order to determine the average treatment effect, $\mathbb{E}[R^1] - \mathbb{E}[R^0]$, we need a way to remove the extra effect of $\mathbb{E}[D|T = 1] - \mathbb{E}[D|T = 0]$. Because we intervened on T and assigned its value independently from D, $T \perp\!\!\!\perp D$ is true and $\mathbb{E}[D|T = 1] - \mathbb{E}[D|T = 0] = \mathbb{E}[D] - \mathbb{E}[D] = 0$. In the observational world, $T \perp\!\!\!\perp D$ may not hold, but by intervening on T randomly, we effectively break the edge between D and T and remove the difference in effect from any confounding variables. This is known as randomized control.