3.4 A Calculus of Intervention

This section establishes a set of inference rules by which probabilistic sentences involving interventions and observations can be transformed into other such sentences, thus providing a syntactic method of deriving (or verifying) claims about interventions. Each inference rule will respect the interpretation of the $do(\cdot)$ operator as an intervention that modifies a select set of functions in the underlying model. The set of inference rules that emerge from this interpretation will be called *do-calculus*.

We will assume that we are given the structure of a causal diagram G in which some of the nodes are observable while others remain unobserved. Our objective will be to facilitate the syntactic derivation of causal effect expressions of the form $P(y|\hat{x})$, where X and Y stand for any subsets of observed variables. By "derivation" we mean stepwise reduction of the expression $P(y|\hat{x})$ to an equivalent expression involving standard probabilities of observed quantities. Whenever such reduction is feasible, the causal effect of X on Y is identifiable (see Definition 3.2.4).

3.4.1 Preliminary notation

Let X, Y, and Z be arbitrary disjoint sets of nodes in a causal DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$ (see Figure 3.6 for an illustration). Finally, the expression $P(y|\hat{x}, z) \triangleq P(y, z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y = y$ given that X is held constant at x and that (under this condition) $Z = z$ is observed.

3.4.2 Inference Rules

The following theorem states the three basic inference rules of the proposed calculus. Proofs are provided in Pearl (1995a).

Theorem 3.4.1 (Rules of d Calculus)
Let G be the directed acyclic graph associated with a causal model as defined in (3.3), and let $P(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W we have the following rules.

Rule 1 (Insertion/deletion of observations):

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{3.33}$$

Rule 2 (Action/observation exchange):

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{3.34}$$

Rule 3 (Insertion/deletion of actions):

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}}, \tag{3.35}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Each of these inference rules follows from the basic interpretation of the "hat" $\hat{x}$ operator as a replacement of the causal mechanism that connects X to its preaction parents by a new mechanism $X = x$ introduced by the intervening force. The result is a submodel characterized by the subgraph $G_{\overline{X}}$ (named "manipulated graph" in Spirtes et al. 1993).

Rule 1 reaffirms d-separation as a valid test for conditional independence in the distribution resulting from the intervention $do(X = x)$, hence the graph $G_{\overline{X}}$. This rule follows from the fact that deleting equations from the system does not introduce any dependencies among the remaining disturbance terms (see (3.3)).

Rule 2 provides a condition for an external intervention $do(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $\{X \cup W\}$ blocking all back-door paths from Z to Y (in $G_{\overline{X}}$), since $G_{\overline{X}\underline{Z}}$ retains all (and only) such paths.

Rule 3 provides conditions for introducing (or deleting) an external intervention $do(Z = z)$ without affecting the probability of $Y = y$.

The validity of this rule stems, again, from simulating the intervention $do(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence the graph $G_{\overline{XZ}}$). The reason for limiting the deletion to nonancestors of W-nodes is provided with the proofs of Rules 1–3 in Pearl (1995a).

Corollary 3.4.2 *A causal effect $q = P(y_1, ..., y_k|\hat{x}_1, ..., \hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 3.4.1, that reduces q into a standard (i.e. "hat"-free) probability expression involving observed quantities.*

Whether Rules 1–3 are sufficient for deriving all identifiable causal effects remains an open question. However, the task of finding a sequence of transformations (if such exists) for reducing an arbitrary causal effect expression can be systematized and executed by efficient algorithms (Galles and Pearl 1995; Pearl and Robins 1995), to be discussed in Chapter 4. As we illustrate in Section 3.4.3, symbolic derivations using the hat notation are much more convenient than algebraic derivations that aim at eliminating latent variables from standard probability expressions (as in Section 3.3.2, equation(3.26)).

3.4.3 Causal Effects: Symbolic Derivation of an Example

We will now demonstrate how Rules 1–3 can be used to derive all causal effect estimands in the structure of Figure 3.5. Figure 3.6 displays the subgraphs that will be needed for the derivations that follow.

Task 1: Compute $P(z|\hat{x})$

This task can be accomplished in one step, since G satisfies the applicability condition for Rule 2. That is, $X \perp\!\!\!\perp Z$ in $G_{\underline{X}}$ (because the path $X \leftarrow U \rightarrow Y \leftarrow Z$ is blocked by the converging arrows at Y) and we can write

$$P(z|\hat{x}) = P(z|x). \tag{3.36}$$

Task 2: Compute $P(y|\hat{z})$

Here we cannot apply Rule 2 to exchange $\hat{z}$ with z because $G_{\underline{Z}}$ contains

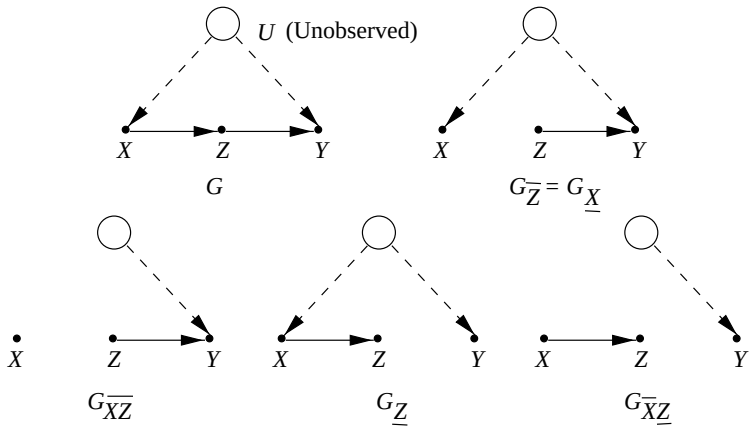

Figure 3.6: Subgraphs of G used in the derivation of causal effects.

a back-door path from Z to Y: $Z \leftarrow X \leftarrow U \rightarrow Y$. Naturally, we would like to block this path by measuring variables (such as X) that reside on that path. This involves conditioning and summing over all values of X:

$$P(y|\hat{z}) = \sum_x P(y|x,\hat{z})P(x|\hat{z}). \tag{3.37}$$

We now have to deal with two terms involving $\hat{z}$, $P(y|x,\hat{z})$ and $P(x|\hat{z})$. The latter can be readily computed by applying Rule 3 for action deletion:

$$P(x|\hat{z}) = P(x) \text{ if } (Z \perp\!\!\!\perp X)_{G_{\overline{Z}}}, \tag{3.38}$$

since X and Z are d-separated in $G_{\overline{Z}}$. (Intuitively, manipulating Z should have no effect on X, because Z is a descendant of X in G.) To reduce the former term, $P(y|x,\hat{z})$, we consult Rule 2:

$$P(y|x,\hat{z}) = P(y|x,z) \text{ if } (Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}, \tag{3.39}$$

noting that X d-separates Z from Y in $G_{\underline{Z}}$. This allows us to write (3.37) as

$$P(y|\hat{z}) = \sum_x P(y|x,z)P(x) = E_x P(y|x,z), \tag{3.40}$$

which is a special case of the back-door formula (equation (3.21)). The legitimizing condition, $(Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}$, offers yet another graphical test for a set X to be sufficient for control of confounding (between Y and

Z) that is equivalent to the ignorability condition of Rosenbaum and Rubin (1983).

Task 3: Compute $P(y|\hat{x})$

Writing

$$P(y|\hat{x}) = \sum_z P(y|z,\hat{x})P(z|\hat{x}), \tag{3.41}$$

we see that the term $P(z|\hat{x})$ was reduced in (3.36) but that no rule can be applied to eliminate the hat symbol $\hat{}$ from the term $P(y|z,\hat{x})$. However, we can legitimately add this symbol via Rule 2:

$$P(y|z,\hat{x}) = P(y|\hat{z},\hat{x}), \tag{3.42}$$

since the applicability condition $(Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\underline{Z}}}$ holds (see Figure 3.6). We can now delete the action $\hat{x}$ from $P(y|\hat{z},\hat{x})$ using Rule 3, since $Y \perp\!\!\!\perp X|Z$ holds in $G_{\overline{XZ}}$. Thus, we have

$$P(y|z,\hat{x}) = P(y|\hat{z}), \tag{3.43}$$

which was calculated in (3.40). Substituting (3.40), (3.43), and (3.36) back into (3.41) finally yields

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x',z)P(x'), \tag{3.44}$$

which is identical to the front-door formula of (3.30).

Task 4: Compute $P(y,z|\hat{x})$

We have

$$P(y,z|\hat{x}) = P(y|z,\hat{x})P(z|\hat{x}).$$

The two terms on the r.h.s. were derived before in (3.36) and (3.43), from which we obtain

$$\begin{aligned} P(y,z|\hat{x}) &= P(y|\hat{z})P(z|x) \\ &= P(z|x)\textstyle\sum_{x'} P(y|x',z)P(x'). \end{aligned} \tag{3.45}$$

Task 5: Compute $P(x,y|\hat{z})$

$$\begin{aligned} P(x,y|\hat{z}) &= P(y|x,\hat{z})P(x|\hat{z}) \\ &= P(y|x,z)P(x). \end{aligned} \tag{3.46}$$

The first term on the r.h.s. is obtained by Rule 2 (licensed by $G_{\underline{Z}}$) and the second term by Rule 3 (as in (3.38)).
Note that, in all the derivations, the graph G has provided both the license for applying the inference rules and the guidance for choosing the right rule to apply.

3.4.4 Causal Inference by Surrogate Experiments

Suppose we wish to learn the causal effect of X on Y when $P(y|\hat{x})$ is not identifiable and, for practical reasons of cost or ethics, we cannot control X by randomized experiment. The question arises of whether $P(y|\hat{x})$ can be identified by randomizing a surrogate variable Z that is easier to control than X. For example, if we are interested in assessing the effect of cholesterol levels (X) on heart disease (Y), a reasonable experiment to conduct would be to control subjects' diet (Z), rather than exercising direct control over cholesterol levels in subjects' blood.

Formally, this problem amounts to transforming $P(y|\hat{x})$ into expressions in which only members of Z obtain the hat symbol. Using Theorem 3.4.1, it can be shown that the following conditions are sufficient for admitting a surrogate variable Z:

(i) X intercepts all directed paths from Z to Y; and

(ii) $P(y|\hat{x})$ is identifiable in $G_{\overline{Z}}$.

Indeed, if condition (i) holds then we can write $P(y|\hat{x}) = P(y|\hat{x},\ \hat{z})$, because $(Y \perp\!\!\!\perp Z|X)_{G_{\overline{XZ}}}$. But $P(y|\hat{x},\ \hat{z})$ stands for the causal effect of X on Y in a model governed by $G_{\overline{Z}}$, which—by condition (ii), is identifiable. Translated to our cholesterol example, these condition require that there be no direct effect of diet on heart conditions and no confounding of cholesterol levels and heart disease, unless we can neutralize such confounding by additional measurements.

Figures 3.9(e) and 3.9(h) (in Section 3.5.2) illustrate models in which both conditions hold. With Figure 3.9(e), for example, we obtain this estimand

$$P(y|\hat{x}) = P(y|x,\ \hat{z}) = \frac{P(y,\ x|\hat{z})}{P(x|\hat{z})}. \tag{3.47}$$

This can be established directly by first applying Rule 3 to add $\hat{z}$,

$$P(y|\hat{x}) = P(y|\hat{x},\ \hat{z}) \text{ because } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\overline{Z}}},$$

and then applying Rule 2 to exchange $\hat{x}$ with x:

$$P(y|\hat{x},\ \hat{z}) = P(y|x,\ \hat{z}) \text{ because } (Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}\overline{Z}}}.$$

According to (3.47), only one level of Z suffices for the identification of $P(y|\hat{x})$ for any values of y and x. In other words, Z need not be varied at all; it can simply be held constant by external means and, if the assumptions embodied in G are valid, the r.h.s. of (3.47) should attain the same value regardless of the (constant) level at which Z is being held. In practice, however, several levels of Z will be needed to ensure that enough samples are obtained for each desired value of X. For example, if we are interested in the difference $E(Y|\hat{x}) - E(Y|\hat{x}')$, where x and x' are two treatment levels, then we should choose two values z and z' of Z that maximize the number of samples in x and x' (respectively) and then estimate

$$E(Y|\hat{x}) - E(Y|\hat{x}') = E(Y|x, \hat{z}) - E(Y|x', \hat{z}').$$