3.2 Intervention in Markovian Models

3.2.1 Graphs as Models of Interventions

In Chapter 1 (Section 1.3) we saw how causal models, unlike probabilistic models, can serve to predict the effect of interventions. This added feature requires that the joint distribution P be supplemented with a causal diagram—that is, a directed acyclic graph G that identifies the causal connections among the variables of interest. In this section we elaborate on the nature of interventions and give explicit formulas for their effects.

The connection between the causal and associational readings of DAGs is formed through the mechanism-based account of causation, which owes its roots to early works in econometrics (Frisch 1938; Haavelmo 1943; Simon 1953). In this account, assertions about causal influences, such as those specified by the links in Figure 3.1, stand for autonomous physical mechanisms among the corresponding quantities; these mechanisms are represented as functional relationships perturbed by random disturbances. Echoing this tradition, Pearl and Verma (1991) interpreted the causal reading of a DAG in terms of functional, rather than probabilistic, relationships (see (1.40) and Definition 2.2.2); in other words, each child-parent family in a DAG G represents a deterministic function

$$x_i = f_i(pa_i, \epsilon_i), \qquad i = 1, ..., n, \tag{3.2}$$

where pa_i are the parents of variable X_i in G; the ϵ_i $(1 \leq i \leq n)$ are mutually independent, arbitrarily distributed random disturbances. These disturbance terms represent independent background factors that the investigator chooses not to include in the analysis. If any of these factors is judged to be influencing two or more variables (thus violating the independence assumption), then that factor must enter the analysis as an unmeasured (or latent) variable and be represented in the graph by a hollow node, such as Z_0 and B in Figure 3.1. For example, the causal assumptions conveyed by the model in Figure 3.1 correspond to

the following set of equations:

$$\begin{aligned} Z_0 &= f_0(\epsilon_0), & B &= f_B(Z_0, \epsilon_B), \\ Z_1 &= f_1(Z_0, \epsilon_1), & X &= f_X(Z_0, \epsilon_X), \\ Z_2 &= f_2(X, Z_1, \epsilon_2), & Y &= f_Y(X, Z_2, Z_3, \epsilon_Y), \\ Z_3 &= f_3(B, Z_2, \epsilon_3). \end{aligned} \tag{3.3}$$

More generally, we may lump together all unobserved factors (including the ϵ_i) into a set U of background variables and then summarize their characteristics by a distribution function $P(u)$—or by some aspects (e.g. independencies) of $P(u)$. Thus, a full specification of a causal model would entail two components: a set of functional relationships

$$x_i = f_i(pa_i, u_i), \qquad i = 1, \ldots, n, \tag{3.4}$$

and a joint distribution function $P(u)$ on the background factors. If the diagram $G(M)$ associated with a causal model M is acyclic, then M is called *semi-Markovian*. If, in addition, the background variables are independent, M is called *Markovian*, since the resulting distribution of the observed variables would then be Markov relative to $G(M)$ (see Theorem 1.4.1). Thus, the model described in Figure 3.1 is semi-Markovian if the observed variables are $\{X, Y, Z_1, Z_2, Z_3\}$; it would turn Markovian if Z_0 and B were observed as well. In Chapter 7 we will pursue the analysis of general non-Markovian models, but in this chapter all models are assumed to be either Markovian or Markovian with unobserved variables (i.e. semi-Markovian).

Needless to state, we would seldom be in possession of $P(u)$ or even f_i. It is important nevertheless to explicate the mathematical content of a fully specified model in order to draw valid inferences from partially specified models, such as the one described in Figure 3.1.

The equational model (3.2) is the nonparametric analog of the so-called structural equations model (Wright 1921; Goldberger 1973), except that: the functional form of the equations (as well as the distribution of the disturbance terms) will remain unspecified. The equality signs in structural equations convey the asymmetrical counterfactual relation of "is determined by," and each equation represents a stable autonomous mechanism. For example, the equation for Y states that, regardless of what we currently observe about Y and regardless of any

changes that might occur in other equations, if variables $(X, Z_2, Z_3, \epsilon_Y)$ were to assume the values $(x, z_2, z_3, \epsilon_Y)$, respectively, then Y would take on the value dictated by the function f_Y.

Recalling our discussion in Section 1.4, the functional characterization of each child-parent relationship leads to the same recursive decomposition of the joint distribution that characterizes Bayesian networks:

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid pa_i), \tag{3.5}$$

which, in our example of Figure 3.1, yields

$$\begin{aligned} P(z_0, x, z_1, b, z_2, z_3, y) &= P(z_0)P(x|z_0)P(z_1|z_0)P(b|z_0) \\ &\quad \times P(z_2|x, z_1)P(z_3|z_2, b)P(y|x, z_2, z_3) \end{aligned} \tag{3.6}$$

Moreover, the functional characterization provides a convenient language for specifying how the resulting distribution would change in response to external interventions. This is accomplished by encoding each intervention as an alteration on a select subset of functions while keeping the other functions intact. Once we know the identity of the mechanisms altered by the intervention and the nature of the alteration, the overall effect of the intervention can be predicted by modifying the corresponding equations in the model and using the modified model to compute a new probability function.

The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x_i. Such an intervention, which we call "atomic," amounts to lifting X_i from the influence of the old functional mechanism $x_i = f_i(pa_i, u_i)$ and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by $do(X_i = x_i)$, or $do(x_i)$ for short,[2] amounts

[2] An equivalent notation, using $set(x)$ instead of $do(x)$, was used in Pearl (1995a). The $do(x)$ notation was first used in Goldszmidt and Pearl (1992) and is gaining in popular support. The expression $P(y|do(x))$ is equivalent in intent to $P(Y_x = y)$ in the potential-outcome model introduced by Neyman (1923) and Rubin (1974) and to the expression $P[(X = x) \; \square\!\!\rightarrow \; (Y = y)]$ in the counterfactual theory of Lewis (1973b). The semantical differences among these notions are discussed in Section 3.6.3 and in Chapter 7.

to removing the equation $x_i = f_i(pa_i, u_i)$ from the model and substituting $X_i = x_i$ in the remaining equations. The new model thus created represents the system's behavior under the intervention $do(X_i = x_i)$ and, when solved for the distribution of X_j, yields the causal effect of X_i on X_j, which is denoted $P(x_j|\hat{x}_i)$. More generally, when an intervention forces a subset X of variables to attain fixed values x, then a subset of equations is to be pruned from the model given in (3.4), one for each member of X, thus defining a new distribution over the remaining variables, that completely characterizes the effect of the intervention.[3]

Definition 3.2.1 (Causal Effect)
Given two disjoint sets of variables, X and Y, the causal effect *of X on Y, denoted either as $P(y|\hat{x})$ or as $P(y|do(x))$, is a function from X to the space of probability distributions on Y. For each realization x of X, $P(y|\hat{x})$ gives the probability of $Y = y$ induced by deleting from the model of* (3.4) *all equations corresponding to variables in X and substituting $X = x$ in the remaining equations.*

Clearly, the graph corresponding to the reduced set of equations is a subgraph of G from which all arrows entering X have been pruned (Spirtes et al. 1993). The difference $E(Y|do(x')) - E(Y|do(x''))$ is sometimes taken as the definition of "causal effect" (Rosenbaum and Rubin 1983), where x' and x'' are two distinct realizations of X. This difference can always be computed from the general function $P(y|do(x))$, which is defined for every level x of X and provides a more refined characterization of the effect of interventions.

3.2.2 Interventions as Variables

An alternative (but sometimes more appealing) account of intervention treats the force responsible for the intervention as a variable within the

[3]The basic view of interventions as equation modifiers originates with Marschak (1950) and Simon (1953). An explicit translation of interventions to "wiping out" equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970) and Sobel (1990). Graphical ramifications of this translation were explicated first in Spirtes et al. (1993) and later in Pearl (1993b).

system (Pearl 1993b). This is facilitated by representing the function f_i itself as a value of a variable, F_i and then writing (3.2) as

$$x_i = I(pa_i, f_i, u_i), \tag{3.7}$$

where I is a three-argument function satisfying

$$I(a, b, c) = f_i(a, c) \text{ whenever } b = f_i.$$

This amounts to conceptualizing the intervention as an external force F_i that alters the function f_i between X_i and its parents. Graphically, we can represent F_i as an added parent node of X_i, and the effect of such an intervention can be analyzed by standard conditionalization—that is, by conditioning our probability on the event that variable F_i attains the value f_i.

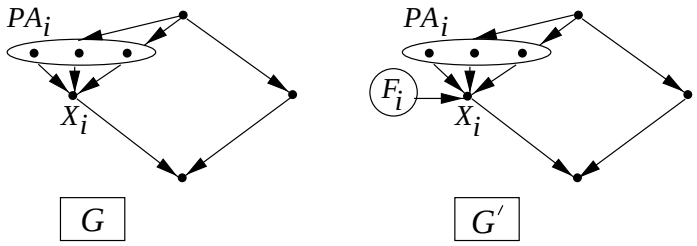

Figure 3.2: Representing external intervention F_i by an augmented network $G' = G \cup \{F_i \to X_i\}$.

The effect of an atomic intervention $do(X_i = x_i')$ is encoded by adding to G a link $F_i \longrightarrow X_i$ (see Figure 3.2), where F_i is a new variable taking values in $\{do(x_i'),\ \text{idle}\}$, x_i' ranges over the domain of X_i, and "idle" represents no intervention. Thus, the new parent set of X_i in the augmented network is $PA_i' = PA_i \cup \{F_i\}$, and it is related to X_i by the conditional probability

$$P(x_i \mid pa_i') = \begin{cases} P(x_i \mid pa_i) & \text{if } F_i = \text{idle}, \\ 0 & \text{if } F_i = do(x_i') \text{ and } x_i \neq x_i', \\ 1 & \text{if } F_i = do(x_i') \text{ and } x_i = x_i'. \end{cases} \tag{3.8}$$

The effect of the intervention $do(x_i')$ is to transform the original probability function $P(x_1, ..., x_n)$ into a new probability function

$P(x_1, ..., x_n|\hat{x}_i')$, given by

$$P(x_1, ..., x_n|\hat{x}_i') = P'(x_1, ..., x_n \mid F_i = do(x_i')), \tag{3.9}$$

where P' is the distribution specified by the augmented network $G' = G \cup \{F_i \rightarrow X_i\}$ and (3.8), with an arbitrary prior distribution on F_i. In general, by adding a hypothetical intervention link $F_i \rightarrow X_i$ to each node in G, we can construct an augmented probability function $P'(x_1, ..., x_n; F_1, ..., F_n)$ that contains information about richer types of interventions. Multiple interventions would be represented by conditioning P' on a subset of the F_i (taking values in their respective $do(x_i')$ domains), and the preintervention probability function P would be viewed as the posterior distribution induced by conditioning each F_i in P' on the value "idle."

One advantage of the augmented network representation is that it is applicable to *any* change in the functional relationship f_i and not merely to the replacement of f_i by a constant. It also displays clearly the ramifications of spontaneous changes in f_i, unmediated by external control. Figure 3.2 predicts, for example, that only descendants of X_i would be effected by changes in f_i and hence the marginal probability $P(z)$ will remain unaltered for every set Z of nondescendants of X_i. Likewise, Figure 3.2 dictates that the conditional probability $P(y|x_i)$ remains invariant to changes in f_i for any set Y of descendants of X_i, provided X_i d-separates F_i from Y. Kevin Hoover (1990, 1999) used such invariant features to determine the direction of causal influences among economic variables (e.g., employment and money supply) by observing the changes induced by sudden modifications in the processes that govern these variables (e.g., tax reform, labor dispute). Indeed, whenever we obtain reliable information (e.g., from historical or institutional knowledge) that an abrupt local change has taken place in a specific mechanism f_i that constrains a given family $(X_i,\ PA_i)$ of variables, we can use the observed changes in the marginal and conditional probabilities surrounding those variables to determine whether X_i is indeed the child (or dependent variable) of that family, thus determining the direction of causal influences in the domain. The statistical features that remain invariant under such changes, as well as the causal assumptions underlying this invariance, are displayed in the augmented network G'.

3.2.3 Computing the Effect of Interventions

Regardless of whether we represent interventions as a modification of an existing model (Definition 3.2.1) or as a conditionalization in an augmented model (equation (3.9)), the result is a well-defined transformation between the preintervention and postintervention distributions. In the case of an atomic intervention $do(X_i = x_i')$, this transformation can be expressed in a simple *truncated-factorization* formula that follows immediately from (3.2) and Definition 3.2.1:[4]

$$P(x_1, ..., x_n|\hat{x}_i') = \begin{cases} \prod_{j \neq i} P(x_j|pa_j) & \text{if } x_i = x_i', \\ 0 & \text{if } x_i \neq x_i'. \end{cases} \tag{3.10}$$

Equation (3.10) reflects the removal of the term $P(x_i \mid pa_i)$ from the product of (3.5), since pa_i no longer influence X_i. For example, the intervention $do(X = x')$ will transform the pre-intervention distribution given in (3.6) into the product

$$\begin{aligned} P(z_0, z_1, b, z_2, z_3, y|\hat{x}') &= P(z_0)P(z_1|z_0)P(b|z_0) \\ &\quad \times P(z_2|x', z_1)P(z_3|z_2, b)P(y|x', z_2, z_3). \end{aligned}$$

Graphically, the removal of the term $P(x_i|pa_i)$ is equivalent to removing the links between PA_i and X_i while keeping the rest of the network intact. Clearly, the transformation defined in (3.10) satisfies the condition of Definition 1.3.1 as well as the properties of (1.38)–(1.39).

Multiplying and dividing (3.10) by $P(x_i'|pa_i)$, the relationship to the preintervention distribution becomes more transparent:

$$P(x_1, \ldots, x_n|\hat{x}_i') = \begin{cases} \dfrac{P(x_1, \ldots, x_n)}{P(x_i' \mid pa_i)} & \text{if } x_i = x_i', \\ 0 & \text{if } x_i \neq x_i'. \end{cases} \tag{3.11}$$

If we regard a joint distribution as an assignment of mass to a collection of abstract points $(x_1, \ldots, x_n)$, each representing a possible state

[4] Equation (3.10) can also be obtained from the G-computation formula of Robins (1986, p. 1423; see Section 3.6.4) and the manipulation theorem of Spirtes et al. (1993) (according to this source, said formula was "independently conjectured by Fienberg in a seminar in 1991"). Additional properties of the transformation defined in (3.10) and (3.11) are given in Goldszmidt and Pearl (1992) and Pearl (1993b).

of the world, then the transformation described in (3.11) reveals some interesting properties of the change in mass distribution that take place as a result of an intervention $do(X_i = x_i')$ (Goldszmidt and Pearl 1992). Each point $(x_1, \ldots, x_n)$ is seen to increase its mass by a factor equal to the inverse of the conditional probability $P(x_i'|pa_i)$ corresponding to that point. Points for which this conditional probability is low would boost their mass value substantially, while those possessing a pa_i value that anticipates a natural (noninterventional) realization of x_i' (i.e., $P(x_i'|pa_i) \approx 1$) will keep their mass unaltered. In standard Bayes conditionalization, each excluded point $(x_i \neq x_i')$ transfers its mass to the entire set of preserved points through a renormalization constant. However, (3.11) describes a different transformation: each excluded point $(x_i \neq x_i')$ transfers its mass to a select set of points that share the same value of pa_i. This can be seen from the constancy of both the total mass assigned to each stratum pa_i and the relative masses of points within each such stratum:

$$\begin{aligned} P(pa_i|do(x_i')) &= P(pa_i); \\ \frac{P(s_i, pa_i|do(x_i'))}{P(s_i', pa_i|do(x_i'))} &= \frac{P(s_i, pa_i)}{P(s_i', pa_i)}. \end{aligned}$$

Here S_i denotes the set of all variables excluding $\{PA_i \cup X_i\}$. This select set of mass-receiving points can be regarded as "closest" to the point excluded by virtue of sharing the same history, as summarized by pa_i (see Sections 4.1.3 and 7.4.3).

Another interesting form of (3.11) obtains when we interpret the division by $P(x_i'|pa_i)$ as conditionalization on x_i' and pa_i:

$$P(x_1, \ldots, x_n|\hat{x}_i') = \begin{cases} P(x_1, \ldots, x_n|x_i', pa_i)P(pa_i) & \text{if } x_i = x_i', \\ 0 & \text{if } x_i \neq x_i'. \end{cases} \quad (3.12)$$

This formula becomes familiar when used to compute the effect of an intervention $do(X_i = x_i')$ on a set of variables Y disjoint of $(X_i \cup PA_i)$. Summing (3.12) over all variables except $Y \cup X_i$ yields the following theorem.

Theorem 3.2.2 (Adjustment for Direct Causes)
Let PA_i denote the set of direct causes of variable X_i, and let Y be any

set of variables disjoint of $\{X_i \cup PA_i\}$. *The effect of the intervention* $do(X_i = x_i')$ *on* Y *is given by*

$$P(y|\hat{x}_i') = \sum_{pa_i} P(y|x_i', pa_i)P(pa_i), \tag{3.13}$$

where $P(y|x_i', pa_i)$ *and* $P(pa_i)$ *represent preintervention probabilities.*

Equation (3.13) calls for conditioning $P(y|x_i')$ on the parents of X_i and then averaging the result, weighted by the prior probability of $PA_i = pa_i$. The operation defined by this conditioning and averaging is known as "adjusting for PA_i."

Variations of this adjustment have been advanced by many philosophers as probabilistic definitions of causality and causal effect (see Section 7.5). Good (1961), for example, calls for conditioning on "the state of the universe just before" the occurrence of the cause. Suppes (1970) calls for conditioning on the entire past, up to the occurrence of the cause. Skyrms (1980, p. 133) calls for conditioning on "maximally specific specifications of the factors outside of our influence at the time of the decision which are causally relevant to the outcome of our actions ...". The aim of conditioning in these proposals is, of course, to eliminate spurious correlations between the cause (in our case, $X_i = x_i'$) and the effect ($Y = y$); clearly, the set of parents PA_i can accomplish this aim with great economy. In the structural account that we pursue in this book, causal effects are defined in a radically different way. The conditioning operator is not introduced into (3.13) as a remedial "adjustment" aimed at eradicating spurious correlations. Rather, it emerges formally from the deeper principle represented in (3.10)—that of preserving all the invariant information that the preintervention distribution can provide.

The transformation of (3.10) can easily be extended to more elaborate interventions in which several variables are manipulated simultaneously. For example, if we consider the compound intervention $do(S = s)$ where S is a subset of variables, then (echoing (1.37)) we should delete from the product of (3.5) all factors $P(x_i|pa_i)$ corresponding to variables in S and obtain the more general truncated factorization

$$P(x_1, \ldots, x_n|\hat{s}) = \begin{cases} \prod_{i|X_i \notin S} P(x_i|pa_i) & \text{for } x_1, \ldots, x_n \text{ consistent with } s, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

Likewise, we need not limit ourselves to simple interventions that set variables to constants. Instead, we may consider a more general modification of the causal model whereby some mechanisms are *replaced*. For example, if we replace the mechanism that determines the value of X_i by another equation, one that involves perhaps a new set PA_i^* of variables, then the resultant distribution would obtain by replacing the factor $P(x_i|pa_i)$ with the conditional probability $P^*(x_i|pa_i^*)$ induced by the new equation. The modified joint distribution would then be given by $P^*(x_1, \ldots, x_n) = P(x_1, \ldots, x_n)P^*(x_i|pa_i^*)/P(x_i|pa_i)$.

An Example: Process Control

To illustrate these operations, let us consider an example involving process control; analogous applications in the areas of health management, economic policy making, product marketing, or robot motion planning should follow in a straightforward way. Let the variable Z_k stand for the state of a production process at time t_k, and let X_k stand for a set of variables (at time t_k) that is used to control that process (see Figure 3.3). For example, Z_k could stand for such measurements as

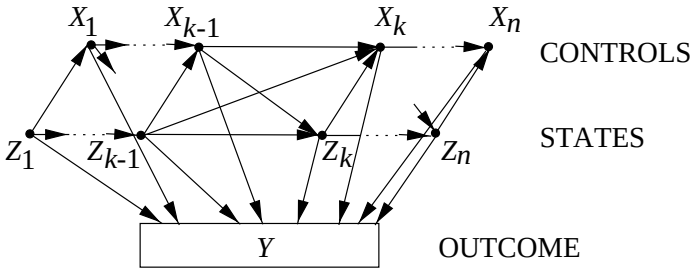

Figure 3.3: Dynamic causal diagram illustrating typical dependencies among the control variables $X_1, \ldots, X_n$, the state variables $Z_1, \ldots, Z_n$, and the outcome variable Y of a sequential process.

temperature and pressure at various location in the plant, and X_k could stand for the rate at which various chemicals are permitted to flow in strategic conduits. Assume that data are gathered while the process is controlled by a strategy S in which each X_k is determined by (i) monitoring three previous variables (X_{k-1}, Z_k, and Z_{k-1}), and (ii) choosing

$X_k = x_k$ with probability $P(x_k|x_{k-1}, z_k, z_{k-1})$. The performance of S is monitored and summarized in the form of a joint probability function $P(y, z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n)$, where Y is an outcome variable (e.g., the quality of the final product). Finally, let us assume (for simplicity) that the state Z_k of the process depends only on the previous state Z_{k-1} and on the previous control X_{k-1}. We wish to evaluate the merit of replacing S with a new strategy, S^*, in which X_k is chosen according to a new conditional probability $P^*(x_k|x_{k-1}, z_k, z_{k-1})$.

Based on our previous analysis (equation (3.14)), the performance $P^*(y)$ of the new strategy S^* will be governed by the distribution

$$\begin{aligned} & P^*(y, z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \qquad (3.15) \\ &= P^*(y|z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \\ &\quad \times \textstyle\prod_k P^*(z_k|z_{k-1}, x_{k-1}) \prod_k P^*(x_k|x_{k-1}, z_k, z_{k-1}). \end{aligned}$$

Because the first two terms remain invariant and the third one is known, we have

$$\begin{aligned} P^*(y) &= \sum_{z_1,\ldots,z_n,x_1,\ldots,x_n} P^*(y, z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \\ &= \sum_{z_1,\ldots,z_n,x_1,\ldots,x_n} P(y|z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \\ &\quad \times \prod_k P(z_k|z_{k-1}, x_{k-1}) \prod_k P^*(x_k|x_{k-1}, z_k, z_{k-1}). \qquad (3.16) \end{aligned}$$

In the special case where S^* is deterministic and time-invariant, X_k becomes a function of X_{k-1}, Z_k, and Z_{k-1}:

$$x_k = g(x_{k-1}, z_k, z_{k-1}).$$

Then the summation over $x_1, \ldots, x_n$ can be performed, yielding

$$\begin{aligned} P^*(y) &= \sum_{z_1,\ldots,z_n} P(y|z_1, z_2, \ldots, z_n, g_1, g_2, \ldots, g_n) \\ &\quad \times \prod_k P(z_k|z_{k-1}, g_{k-1}), \qquad (3.17) \end{aligned}$$

where g_k is defined recursively as

$$g_1 = g(z_1) \text{ and } g_k = g(g_{k-1}, z_k, z_{k-1}).$$

In the special case of a strategy X^* composed of elementary actions $do(X_k = x_k)$, the function g degenerates into a constant, x_k, and we obtain

$$\begin{aligned} P^*(y) &= P(y|\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) \\ &= \sum_{z_1,.,z_n} P(y|z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \prod_k P(z_k|z_{k-1}, x_{k-1}) \end{aligned} \tag{3.18}$$

which can also be obtained from (3.14).

The planning problem illustrated by this example is typical of Markov decision processes (MDPs) (Howard 1960; Dean and Wellman 1991; Bertsekas and Tsitsiklis 1996), where the target of analysis is finding the best next action $do(X_k = x_k)$, given the current state Z_k and past actions. In MDPs, we are normally given the transition functions $P(z_{k+1}|z_k, \hat{x}_k)$ and the cost function to be minimized. In the problem we have just analyzed, neither function is given; instead, they must be learned from data gathered under past (presumably suboptimal) strategies. Fortunately, because all variables in the model were measured, both functions were identifiable and could be estimated directly from the corresponding conditional probabilities as follows:

$$P(z_{k+1}|z_k, \hat{x}_k) = P(z_{k+1}|z_k, x_k);$$
$$P(y|z_1, z_2, \ldots, z_n, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) = P(y|z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n).$$

In Chapter 4 (Section 4.4) we will deal with partially observable Markov decision processes (POMDPs), where some states Z_k are unobserved; learning the transition and cost functions in those problems will require a more intricate method of identification.

It is worth noting that, in this example, to predict the effect of a new strategy it is necessary first to measure variables (Z_k) that are affected by some control variables (X_{k-1}). Such measurements are generally shunned in the classical literature on experimental design (Cox 1958, p. 48) because they lie on the causal pathways between treatment and outcome and thus tend to confound the desired effect estimate. However, our analysis shows that, when properly processed, such measurements may be indispensable in predicting the effect of certain control programs. This will be especially true in semi-Markovian models (i.e.,

DAGs involving unmeasured variables), which are analyzed in Section 3.3.2.

Summary

The immediate implication of the analysis provided in this section is that—given a causal diagram in which all direct causes (i.e. parents) of intervened variables are observable—one can infer postintervention distributions from preintervention distributions; hence, under such assumptions we can estimate the effects of interventions from passive (i.e. nonexperimental) observations, using the truncated factorization of (3.14). Yet the more challenging problem is to derive causal effects in situations like Figure 3.1, where some members of PA_i are unobservable and so prevent estimation of $P(x_i'|pa_i)$. In Sections 3.3 and 3.4 we provide simple graphical tests for deciding when $P(x_j|\hat{x}_i)$ is estimable in such models. But first we need to define more formally what it means for a causal quantity Q to be estimable from passive observations, a question that falls under the technical term *identification*.

3.2.4 Identification of Causal Quantities

Causal quantities, unlike statistical parameters, are defined relative to a causal model M and not relative to a joint distribution $P_M(v)$ over the set V of observed variables. Since nonexperimental data provides information about $P_M(v)$ alone, and since several models can generate the same distribution, the danger exists that the desired quantity will not be discernible unambiguously from the data—even when infinitely many samples are taken. Identifiability ensures that the added assumptions we make about M (e.g., the causal graph or the zero coefficients in structural equations) will supply the missing information without explicating M in full detail.

Definition 3.2.3 (Identifiability)
Let $Q(M)$ be any computable quantity of a model M. We say that Q is identifiable *in a class $\boldsymbol{M}$ of models if, for any pairs of models M_1 and M_2 from $\boldsymbol{M}$, $Q(M_1) = Q(M_2)$ whenever $P_{M_1}(v) = P_{M_2}(v)$. If our observations are limited, and permit only a partial set F_M of features*

(*of* $P_M(v)$) *to be estimated, we define* Q *to be* identifiable from F_M *if* $Q(M_1) = Q(M_2)$ *whenever* $F_{M_1} = F_{M_2}$.

Identifiability is essential for integrating statistical data (summarized by $P(v)$) with incomplete causal knowledge of $\{f_i\}$, as it enables us to estimate quantities Q consistently from large samples of P without specifying the details of M; the general characteristics of the class $\boldsymbol{M}$ suffice. For the purpose of our analysis, the quantity Q of interest is the causal effect $P_M(y|\hat{x})$, which is certainly computable from a given model M (using Definition 3.2.1) but which we often need to compute from an incomplete specification of M—in the form of general characteristics portrayed in the graph G associated with M. We will therefore consider a class $\boldsymbol{M}$ of models that have the following characteristics in common:

(i) they share the same parent-child families (i.e., the same causal graph G); and

(ii) they induce positive distributions on the observed variables (i.e., $P(v) > 0$).

Relative to such classes, we now have the following.

Definition 3.2.4 (Causal-Effect Identifiability)
The causal effect *of* X *on* Y *is said to be* identifiable *from a graph* G *if the quantity* $P(y|\hat{x})$ *can be computed uniquely from any positive probability of the observed variables—that is, if* $P_{M_1}(y|\hat{x}) = P_{M_2}(y|\hat{x})$ *for every pair of models* M_1 *and* M_2 *with* $P_{M_1}(v) = P_{M_2}(v) > 0$ *and* $G(M_1) = G(M_2) = G$.

The identifiability of $P(y|\hat{x})$ ensures that it is possible to infer the effect of action $do(X = x)$ on Y from two sources of information:

(i) passive observations, as summarized by the probability function $P(v)$; and

(ii) the causal graph G, which specifies (qualitatively) which variables make up the stable mechanisms in the domain or, alternatively, which variables participate in the determination of each variable in the domain.

Restricting identifiability to positive distributions assures us that the condition $X = x$ is represented in the data in the appropriate context, thus avoiding a zero denominator in (3.10). It would be impossible to infer the effect of action $do(X = x)$ from data in which X never attains the value x in the context wherein the action is applied. Extensions to some nonpositive distributions are feasible but will not be treated here. Note that, to prove nonidentifiability, it is sufficient to present two sets of structural equations that induce identical distributions over observed variables but have different causal effects.

Using the concept of identifiability, we can now summarize the results of Section 3.2.3 in the following theorem.

Theorem 3.2.5 *Given a causal diagram G of any Markovian model in which a subset V of variables are measured, the causal effect $P(y|\hat{x})$ is identifiable whenever $\{X \cup Y \cup PA_X\} \subseteq V$, that is, whenever X, Y, and all parents of variables in X are measured. The expression of $P(y|\hat{x})$ is then obtained by adjusting for PA_x, as in* (3.13).

A special case of Theorem 3.2.5 holds when *all* variables are assumed to be observed.

Corollary 3.2.6 *Given the causal diagram G of any Markovian model in which all variables are measured, the causal effect $P(y|\hat{x})$ is identifiable for every two subsets of variables X and Y and is obtained from the truncated factorization of* (3.14).

We now turn our attention to identification problems in semi-Markovian models.