CHAPTER FOUR

Actions, Plans, and Direct Effects

He whose actions exceed his wisdom,
his wisdom shall endure.
Rabbi Hanina ben Dosa
(1st century A.D.)

Preface

So far, our analysis of causal effects has focused on primitive interventions of the form $do(x)$, which stood for setting the value of variable X to a fixed constant, x, and asking for the effect of this action on the probabilities of some response variables Y. In this chapter we introduce several extensions of this analysis.

First (Section 4.1), we discuss the status of actions vis-à-vis observations in probability theory, decision analysis, and causal modeling, and we advance the thesis that the main role of causal models is to facilitate the evaluation of the effect of *novel* actions and policies that were unanticipated during the construction of the model.

In Section 4.2 we extend the identification analysis of Chapter 3 to conditional actions of the form "do x if you see z" and stochastic policies of the form "do x with probability p if you see z." We shall see that the evaluation and identification of these more elaborate interventions can be obtained from the analysis of primitive interventions. In Section 4.3, we use the intervention calculus developed in Chapter 3 to give a graphical characterization of the set of semi-Markovian models for which the causal effect of one variable on another can be identified.

We address in Section 4.4 the problem of evaluating the effect of sequential plans – namely, sequences of time-varying actions (some taken concurrently) designed to produce a certain outcome. We provide a graphical method of estimating the effect of such plans from nonexperimental observations in which some of the actions are influenced by their predecessors, some observations are influenced by the actions, and some confounding variables are unmeasured. We show that there is substantial advantage to analyzing a plan into its constituent actions rather than treating the set of actions as a single entity.

Finally, in Section 4.5 we address the question of distinguishing direct from indirect effects. We show that direct effects can be identified by the graphical method developed in Section 4.4. An example using alleged sex discrimination in college admission will serve to demonstrate the assumptions needed for proper analysis of direct effects.

4.1 INTRODUCTION

4.1.1 Actions, Acts, and Probabilities

Actions admit two interpretations: reactive and deliberative. The reactive interpretation sees action as a consequence of an agent's beliefs, disposition, and environmental inputs, as in "Adam ate the apple because Eve handed it to him." The deliberative interpretation sees action as an option of choice in contemplated decision making, usually involving comparison of consequences, as in "Adam was wondering what God would do if he ate the apple." We shall distinguish the two views by calling the first "act" and the second "action." An act is viewed from the outside, an action from the inside. Therefore, an act can be predicted and can serve as evidence for the actor's stimuli and motivations (provided the actor is part of our model). Actions, in contrast, can neither be predicted nor provide evidence since (by definition) they are pending deliberation and turn into *acts* once executed.

The confusion between actions and acts has led to Newcomb's paradox (Nozick 1969) and other oddities in the so-called evidential decision theory, which encourages decision makers to take into consideration the evidence that an action would provide, if enacted. This bizarre theory seems to have loomed from Jeffrey's influential book *The Logic of Decision* (Jeffrey 1965), in which actions are treated as ordinary events (rather than interventions) and, accordingly, the effects of actions are obtained through conditionalization rather than through a mechanism-modifying operation like $do(x)$. (See Stalnaker 1972; Gibbard and Harper 1976; Skyrms 1980; Meek and Glymour 1994; Hitchcock 1996.)

Traditional decision theory[1] instructs rational agents to choose the option x that maximizes expected utility,[2]

$$U(x) = \sum_y P(y \mid do(x))u(y),$$

where $u(y)$ is the utility of outcome y; in contrast, "evidential decision" theory calls for maximizing the conditional expectation

$$U_{\text{ev}}(x) = \sum_y P(y \mid x)u(y),$$

in which x is (improperly) treated as an observed proposition.

The paradoxes that emerge from this fallacy are obvious: patients should avoid going to the doctor "to reduce the probability that one is seriously ill" (Skyrms 1980, p. 130);

---

[1] I purposely avoid the common title "causal decision theory" in order to suppress even the slightest hint that any alternative, noncausal theory can be used to guide decisions.

[2] Following a suggestion of Stalnaker (1972), Gibbard and Harper (1976) used $P(x \mathrel{\square\!\!\rightarrow} y)$ in $U(x)$, rather than $P(y \mid do(x))$, where $x \mathrel{\square\!\!\rightarrow} y$ stands for the subjunctive conditional "y if it were x." The semantics of the two operators are closely related (see Section 7.4), but the equation-removal interpretation of the $do(x)$ operator is less ambiguous and clearly suppresses inference from effect to cause.

workers should never hurry to work, to reduce the probability of having overslept; students should not prepare for exams, lest this would prove them behind in their studies; and so on. In short, all remedial actions should be banished lest they increase the probability that a remedy is indeed needed.

The oddity in this kind of logic stems from treating actions as acts that are governed by past associations instead of as objects of free choice, as dictated by the semantics of the $do(x)$ operator. This "evidential" decision theory preaches that one should never ignore genuine statistical evidence (in our case, the evidence that an act normally provides regarding whether the act is needed), but decision theory proper reminds us that actions – by their very definition – render such evidence irrelevant to the decision at hand, for actions *change* the probabilities that acts normally obey.[3]

The moral of this story can be summarized in the following mnemonic rhymes:

> Whatever evidence an act might provide
> On facts that preceded the act,
> Should never be used to help one decide
> On whether to choose that same act.

Evidential decision theory was a passing episode in the philosophical literature, and no philosopher today takes the original version of this theory seriously. Still, some recent attempts have been made to revive interest in Jeffrey's expected utility by replacing $P(y \mid x)$ with $P(y \mid x, K)$, where K stands for various background contexts, chosen to suppress spurious associations (as in (3.13)) (Price 1991; Hitchcock 1996). Such attempts echo an overly restrictive empiricist tradition, according to which rational agents live and die by one source of information – statistical associations – and hence expected utilities should admit no other operation but Bayes's conditionalization. This tradition is rapidly giving way to a more accommodating conception: rational agents should act according to theories of actions; naturally, such theories demand action-specific conditionalization (e.g. $do(x)$) while reserving Bayes's conditionalization for representing passive observations (see Goldszmidt and Pearl 1992; Meek and Glymour 1994; Woodward 1995).

In principle, actions are not part of probability theory, and understandably so: probabilities capture normal relationships in the world, whereas actions represent interventions that perturb those relationships. It is no wonder, then, that actions are treated as foreign entities throughout the literature on probability and statistics; they serve neither as arguments of probability expressions nor as events for conditioning such expressions.

Even in the statistical decision-theoretic literature (e.g. Savage 1954), where actions are the main target of analysis, the symbols given to actions serve merely as indices for distinguishing one probability function from another, not as entities that stand in logical relationships to the variables on which probabilities are defined. Savage (1954, p. 14) defined "act" as a "function attaching a consequence to each state of the world," and he treated a chain of decisions, one leading to other, as a single decision. However, the

[3] Such evidence is rendered irrelevant within the actor's own probability space; in multiagent decision situations, however, each agent should definitely be cognizant of how other agents might interpret each of his pending "would-be" acts.

logic that leads us to infer the consequences of actions and strategies from more elementary considerations is left out of the formalism. For example, consider the actions: "raise taxes," "lower taxes," and "raise interest rates." The consequences of all three actions must be specified separately, prior to analysis; none can be inferred from the others. As a result, if we are given two probabilities, P_A and P_B, denoting the probabilities prevailing under actions A or B, respectively, there is no way we can deduce from this input the probability $P_{A \wedge B}$ corresponding to the joint action $A \wedge B$ or indeed any Boolean combination of the propositions A and B. This means that, in principle, the impact of all anticipated joint actions would need to be specified in advance – an insurmountable task.

The peculiar status of actions in probability theory can be seen most clearly in comparison to the status of observations. By specifying a probability function $P(s)$ on the possible states of the world, we automatically specify how probabilities should change with every conceivable observation e, since $P(s)$ permits us to compute (by conditioning on e) the posterior probabilities $P(E \mid e)$ for every pair of events E and e. However, specifying $P(s)$ tells us nothing about how probabilities should change in response to an external action $do(A)$. In general, if an action $do(A)$ is to be described as a function that takes $P(s)$ and transforms it to $P_A(s)$, then $P(s)$ tells us nothing about the nature of $P_A(s)$, even when A is an elementary event for which $P(A)$ is well-defined (e.g., "raise the temperature by 1 degree" or "turn the sprinkler on"). With the exception of the trivial requirement that $P_A(s)$ be zero if s implies $\neg A$, a requirement that applies uniformly to every $P(s)$, probability theory does not tell us how $P_A(s)$ should differ from $P'_A(s)$, where $P'(s)$ is some other preaction probability function. Conditioning on A is clearly inadequate for capturing this transformation, as we have seen in many examples in Chapters 1 and 3 (see e.g. Section 1.3.1), because conditioning represents passive observations in an unchanging world whereas actions change the world.

Drawing analogy to visual perception, we may say that the information contained in $P(s)$ is analogous to a precise description of a three-dimensional object; it is sufficient for predicting how that object will be viewed from any angle outside the object, but it is insufficient for predicting how the object will be viewed if manipulated and squeezed by external forces. Additional information about the physical properties of the object must be supplied for making such predictions. By analogy, the additional information required for describing the transformation from $P(s)$ to $P_A(s)$ should identify those elements of the world that remain invariant under the action $do(A)$. This extra information is provided by causal knowledge, and the $do(\cdot)$ operator enables us to capture the invariant elements (thus defining $P_A(s)$) by locally modifying the graph or the structural equations. The next section will compare this device to the way actions are handled in standard decision theory.

4.1.2 Actions in Decision Analysis

Instead of introducing new operators into probability calculus, the traditional approach has been to attribute the differences between seeing and doing to differences in the total evidence available. Consider the statements: "the barometer reading was observed to be x" and "the barometer reading was set to level x." The former helps us predict the weather, the latter does not. While the evidence described in the first statement is limited

to the reading of the barometer, the second statement also tells us that the barometer was manipulated by some agent, and conditioning on this additional evidence should render the barometer reading irrelevant to predicting the rain.

The practical aspects of this approach amount to embracing the acting agents as variables in the analysis, constructing an augmented distribution function including the decisions of those agents, and inferring the effect of actions by conditioning those decision variables to particular values. Thus, for example, the agent manipulating the barometer might enter the system as a decision variable "squeezing the barometer"; after incorporating this variable into the probability distribution, we could infer the impact of manipulating the barometer simply by conditioning the augmented distribution on the event "the barometer was squeezed by force y and has reached level x."

For this conditioning method to work properly in evaluating the effect of future actions, the manipulating agent must be treated as an ideal experimenter acting out of free will, and the associated decision variables must be treated as exogenous – causally unaffected by other variables in the system. For example, if the augmented probability function encodes the fact that the current owner of the barometer tends to squeeze the barometer each time she feels arthritis pain, we will be unable to use that function for evaluating the effects of deliberate squeezing of the barometer, even by the same owner. Recalling the difference between acts and actions, whenever we set out to calculate the effect of a pending action, we must ignore all mechanisms that constrained or triggered the execution of that action in the past. Accordingly, the event "The barometer was squeezed" must enter the augmented probability function as independent of all events that occurred prior to the time of manipulation, similar to the way action variable F entered the augmented network in Figure 3.2.

This solution corresponds precisely to the way actions are treated in decision analysis, as depicted in the literature on influence diagrams (IDs) (Howard and Matheson 1981; Shachter 1986; Pearl 1988b, chap. 6). Each decision variable is represented as exogenous variable (a parentless node in the diagram), and its impact on other variables is assessed and encoded in terms of conditional probabilities, similar to the impact of any other parent node in the diagram.[4]

The difficulty with this approach is that we need to anticipate in advance, and represent explicitly, all actions whose effects we might wish to evaluate in the future. This renders the modeling process unduly cumbersome, if not totally unmanageable. In circuit diagnosis, for example, it would be awkward to represent every conceivable act of component replacement (similarly, every conceivable connection to a voltage source, current source, etc.) as a node in the diagram. Instead, the effects of such replacements are implicit in the circuit diagram itself and can be deduced from the diagram, given its causal interpretation. In econometric modeling likewise, it would be awkward to represent every conceivable variant of policy intervention as a new variable in the economic equations. Instead, the effects of such interventions can be deduced from the structural

[4] The ID literature's insistence on divorcing the links in the ID from any causal interpretation (Howard and Matheson 1981; Howard 1990) is at odds with prevailing practice. The causal interpretation is what allows us to treat decision variables as root nodes, unassociated with all other nodes (except their descendants).

interpretation of those equations, if only we can tie the immediate effects of each policy to the corresponding variables and parameters in the equations. The compound action "raise taxes and lower interest rates," for example, need not be introduced as a new variable in the equations, because the effect of that action can be deduced if we have the quantities "taxation level" and "interest rates" already represented as (either exogenous or endogenous) variables in the equations.

The ability to predict the effect of interventions without enumerating those interventions in advance is one of the main advantages we draw from causal modeling and one of the main functions served by the notion of causation. Since the number of actions or action combinations is enormous, they cannot be represented explicitly in the model but rather must be indexed by the propositions that each action enforces directly. Indirect consequences of enforcing those propositions are then inferred from the causal relationships among the variables represented in the model. We will return to this theme in Chapter 7 (Section 7.2.4), where we further explore the invariance assumptions that must be met for this encoding scheme to work.

4.1.3 Actions and Counterfactuals

As an alternative to Bayesian conditioning, philosophers (Lewis 1976; Gardenfors 1988) have studied another probability transformation called "imaging," which was deemed useful in the analysis of subjunctive conditionals and which more adequately represents the transformations associated with actions. Whereas Bayes conditioning of $P(s \mid e)$ transfers the entire probability mass from states excluded by e to the remaining states (in proportion to their current probabilities, $P(s)$), imaging works differently: each excluded state s transfers its mass individually to a select set of states $S^*(s)$ that are considered to be "closest" to s (see Section 7.4.3). Although providing a more adequate and general framework for actions (Gibbard and Harper 1976), imaging leaves the precise specification of the selection function $S^*(s)$ almost unconstrained. Consequently, the problem of enumerating future actions is replaced by the problem of encoding distances among states in a way that would be both economical and respectful of common understanding of the causal laws that operate in the domain. The second requirement is not trivial, considering that indirect ramifications of actions often result in worlds that are quite dissimilar to the one from which we start (Fine 1975).

The difficulties associated with making the closest-world approach conform to causal laws will be further elaborated in Chapter 7 (Section 7.4). The structural approach pursued in this book escapes these difficulties by basing the notion of interventions directly on causal mechanisms and by capitalizing on the properties of invariance and autonomy that accompany these mechanisms. This mechanism-modification approach can be viewed as a special instance of the closest-world approach, where the closeness measure is crafted so as to respect the causal mechanisms in the domain; the selection function $S^*(s)$ that ensues is represented in (3.11) (see discussion that follows).

The operationality of this mechanism-modification semantics was demonstrated in Chapter 3 and led to the quantitative predictions of the effects of actions, including actions that were not contemplated during the model's construction. The *do* calculus that

emerged (Theorem 3.4.1) extends this prediction facility to cases where some of the variables are unobserved. In Chapter 7 we further use the mechanism-modification interpretation to provide semantics for counterfactual statements, as outlined in Section 1.1.4. In this chapter, we will extend the applications of the *do* calculus in several directions, as outlined in the Preface.

4.2 CONDITIONAL ACTIONS AND STOCHASTIC POLICIES

The interventions considered in our analysis of identification (Sections 3.3–3.4) were limited to actions that merely force a variable or a group of variables X to take on some specified value x. In general (see the process control example in Section 3.2.3), interventions may involve complex policies in which a variable X is made to respond in a specified way to some set Z of other variables – say, through a functional relationship $x = g(z)$ or through a stochastic relationship whereby X is set to x with probability $P^*(x \mid z)$. We will show, based on Pearl (1994b), that identifying the effect of such policies is equivalent to identifying the expression $P(y \mid \hat{x}, z)$.

Let $P(y \mid do(X = g(z)))$ stand for the distribution (of Y) prevailing under the policy $do(X = g(z))$. To compute $P(y \mid do(X = g(z)))$, we condition on Z and write

$$\begin{aligned} P(y \mid do(X = g(z))) &= \sum_z P(y \mid do(X = g(z)), z) P(z \mid do(X = g(z))) \\ &= \sum_z P(y \mid \hat{x}, z)|_{x=g(z)} P(z) \\ &= E_z[P(y \mid \hat{x}, z)|_{x=g(z)}]. \end{aligned}$$

The equality

$$P(z \mid do(X = g(z))) = P(z)$$

stems, of course, from the fact that Z cannot be a descendant of X; hence, any control exerted on X can have no effect on the distribution of Z. Thus, we see that the causal effect of a policy $do(X = g(z))$ can be evaluated directly from the expression of $P(y \mid \hat{x}, z)$ simply by substituting $g(z)$ for x and taking the expectation over Z (using the observed distribution $P(z)$).

This identifiability criterion for conditional policy is somewhat stricter than that for unconditional intervention. Clearly, if a policy $do(X = g(z))$ is identifiable then the simple intervention $do(X = x)$ is identifiable as well, since we can always obtain the latter by setting $g(z) = x$. The converse does not hold, however, because conditioning on Z might create dependencies that will prevent the successful reduction of $P(y \mid \hat{x}, z)$ to a hat-free expression.

A stochastic policy, which imposes a new conditional distribution $P^*(x \mid z)$ for x, can be handled in a similar manner. We regard the stochastic intervention as a random process in which the unconditional intervention $do(X = x)$ is enforced with probability $P^*(x \mid z)$. Thus, given $Z = z$, the intervention $do(X = x)$ will occur with probability

$P^*(x \mid z)$ and will produce a causal effect given by $P(y \mid \hat{x}, z)$. Averaging over x and z gives the effect (on Y) of the stochastic policy $P^*(x \mid z)$:

$$P(y)|_{P^*(x|z)} = \sum_x \sum_z P(y \mid \hat{x}, z) P^*(x \mid z) P(z).$$

Because $P^*(x \mid z)$ is specified externally, we see again that the identifiability of $P(y \mid \hat{x}, z)$ is a necessary and sufficient condition for the identifiability of any stochastic policy that shapes the distribution of X by the outcome of Z.

Of special importance in planning is a STRIPS-like action (Fikes and Nilsson 1971) whose immediate effects $X = x$ depend on the satisfaction of some enabling precondition $C(w)$ on a set W of variables. To represent such actions, we let $Z = W \cup PA_X$ and set

$$P^*(x \mid z) = \begin{cases} P(x \mid pa_X) & \text{if } C(w) = \text{false}, \\ 1 & \text{if } C(w) = \text{true and } X = x, \\ 0 & \text{if } C(w) = \text{true and } X \neq x. \end{cases}$$

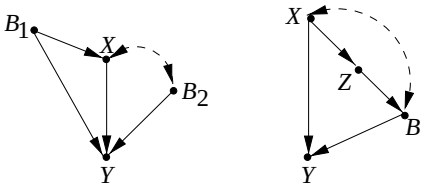
4.3 WHEN IS THE EFFECT OF AN ACTION IDENTIFIABLE?

In Chapter 3 we developed several graphical criteria for recognizing when the effect of one variable on another, $P(y \mid do(x))$, is identifiable in the presence of unmeasured variables. These criteria, like the back-door (Theorem 3.3.2) and front-door (Theorem 3.3.4), are special cases of a more general class of semi-Markovian models for which repeated application of the inference rules of *do* calculus (Theorem 3.4.1) will reduce $P(y \mid \hat{x})$ to a hat-free expression, thus rendering it identifiable. The semi-Markovian model of Figure 3.1 (or Figure 3.8(f)) is an example where direct application of either the back-door or front-door criterion would not be sufficient for identifying $P(y \mid \hat{x})$ and yet the expression is reducible (hence identifiable) by a sequence of inference rules of Theorem 3.4.1. In this section we establish a complete characterization of the class of models in which the causal effect $P(y \mid \hat{x})$ is identifiable in *do* calculus.

4.3.1 Graphical Conditions for Identification

Theorem 4.3.1 characterizes the class of "*do*-identifiable" models in the form of four graphical conditions, any one of which is sufficient for the identification of $P(y \mid \hat{x})$ when X and Y are singleton nodes in the graph. Theorem 4.3.2 then asserts the completeness (or necessity) of these four conditions; one of which must hold in the model for $P(y \mid \hat{x})$ to be identifiable in *do* calculus. Whether these four conditions are necessary in general (in accordance with the semantics of Definition 3.2.4) depends on whether the inference rules of *do* calculus are complete. This question, to the best of my knowledge, is still open.

Theorem 4.3.1 (Galles and Pearl 1995)
Let X and Y denote two singleton variables in a semi-Markovian model characterized by graph G. A sufficient condition for the identifiability of $P(y \mid \hat{x})$ is that G satisfy one of the following four conditions.



(a) (b)

Figure 4.1 Condition 3 of Theorem 4.3.1. In (a), the set $\{B_1, B_2\}$ blocks all back-door paths from X to Y, and $P(b_1, b_2 \mid \hat{x}) = P(b_1, b_2)$. In (b), the node B blocks all back-door paths from X to Y, and $P(b \mid \hat{x})$ is identifiable using Condition 4.

1. *There is no back-door path from X to Y in G; that is,* $(X \perp\!\!\!\perp Y)_{G_{\underline{X}}}$.
2. *There is no directed path from X to Y in G.*
3. *There exists a set of nodes B that blocks all back-door paths from X to Y so that* $P(b \mid \hat{x})$ *is identifiable.* (A special case of this condition occurs when B consists entirely of nondescendants of X, in which case $P(b \mid \hat{x})$ reduces immediately to $P(b)$.)
4. *There exist sets of nodes* Z_1 *and* Z_2 *such that:*
   (i) Z_1 *blocks every directed path from X to Y* (*i.e.,* $(Y \perp\!\!\!\perp X \mid Z_1)_{G_{\overline{Z_1}\overline{X}}}$);
   (ii) Z_2 *blocks all back-door paths between* Z_1 *and Y* (*i.e.,* $(Y \perp\!\!\!\perp Z_1 \mid Z_2)_{G_{\overline{X}\underline{Z_1}}}$);
   (iii) Z_2 *blocks all back-door paths between X and* Z_1 (*i.e.,* $(X \perp\!\!\!\perp Z_1 \mid Z_2)_{G_{\underline{X}}}$; *and*
   (iv) Z_2 *does not activate any back-door paths from X to Y* (*i.e.,* $(X \perp\!\!\!\perp Y \mid Z_1, Z_2)_{G_{\overline{Z_1}\overline{X(Z_2)}}}$). (This condition holds if (i)–(iii) are met and no member of Z_2 is a descendant of X.)

   (A special case of condition 4 occurs when $Z_2 = \emptyset$ and there is no back-door path from X to Z_1 or from Z_1 to Y.)

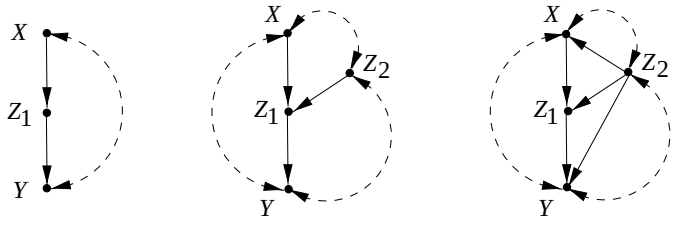
Proof

Condition 1. This condition follows directly from Rule 2 (see Theorem 3.4.1). If $(Y \perp\!\!\!\perp X)_{G_{\underline{X}}}$ then we can immediately change $P(y \mid \hat{x})$ to $P(y \mid x)$, so the query is identifiable.

Condition 2. If there is no directed path from X to Y in G, then $(Y \perp\!\!\!\perp X)_{G_{\overline{X}}}$. Hence, by Rule 3, $P(y \mid \hat{x}) = P(y)$ and so the query is identifiable.

Condition 3. If there is a set of nodes B that blocks all back-door paths from X to Y (i.e., $(Y \perp\!\!\!\perp X \mid B)_{G_{\underline{X}}}$), then we can expand $P(y \mid \hat{x})$ as $\sum_b P(y \mid \hat{x}, b) P(b \mid \hat{x})$ and, by Rule 2, rewrite $P(y \mid \hat{x}, b)$ as $P(y \mid x, b)$. If the query $(b \mid \hat{x})$ is identifiable, then the original query must also be identifiable. See examples in Figure 4.1.

Condition 4. If there is a set of nodes Z_1 that block all directed paths from X to Y and a set of nodes Z_2 that block all back-door paths between Y and Z_1 in $G_{\overline{X}}$, then we expand $P(y \mid \hat{x}) = \sum_{z_1, z_2} P(y \mid \hat{x}, z_1, z_2) P(z_1, z_2 \mid \hat{x})$ and rewrite $P(y \mid \hat{x}, z_1, z_2)$ as $P(y \mid \hat{x}, \hat{z}_1, z_2)$ using Rule 2, since all back-door paths between Z_1 and Y are blocked by Z_2 in $G_{\overline{X}}$. We can reduce $P(y \mid \hat{x}, \hat{z}_1, z_2)$ to $P(y \mid \hat{z}_1, z_2)$ using Rule 3, since $(Y \perp\!\!\!\perp X \mid Z_1, Z_2)_{G_{\overline{Z_1}\overline{X(Z_2)}}}$. We can rewrite $P(y \mid \hat{z}_1, z_2)$ as $P(y \mid z_1, z_2)$ if $(Y \perp\!\!\!\perp Z_1 \mid Z_2)_{G_{\underline{Z_1}}}$. The only way that this independence cannot hold is if there is a path from Y to Z_1 through X, since $(Y \perp\!\!\!\perp Z_1 \mid Z_2)_{G_{\overline{X}\underline{Z_1}}}$. However, we can block this path by conditioning and



(a) (b) (c)

Figure 4.2 Condition 4 of Theorem 4.3.1. In (a), Z_1 blocks all directed paths from X to Y, and the empty set blocks all back-door paths from Z_1 to Y in $G_{\overline{X}}$ and all back-door paths from X to Z_1 in G. In (b) and (c), Z_1 blocks all directed paths from X to Y, and Z_2 blocks all back-door paths from Z_1 to Y in $G_{\overline{X}}$ and all back-door paths from X to Z_1 in G.

summing over X and so derive $\sum_{x'} P(y \mid \hat{z}_1, z_2, x')P(x' \mid \hat{z}_1, z_2)$. Now we can rewrite $P(y \mid \hat{z}_1, z_2, x')$ as $P(y \mid z_1, z_2, x')$ using Rule 2. The $P(x' \mid \hat{z}_1, z_2)$ term can be rewritten as $P(x' \mid z_2)$ using Rule 3, since Z_1 is a child of X and the graph is acyclic. The query can therefore be rewritten as $\sum_{z_1,z_2} \sum_{x'} P(y \mid z_1, z_2, x')P(x' \mid z_2)P(z_1, z_2 \mid \hat{x})$, and we have $P(z_1, z_2 \mid \hat{x}) = P(z_2 \mid \hat{x})P(z_1 \mid \hat{x}, z_2)$. Since Z_2 consists of nondescendants of X, we can rewrite $P(z_2 \mid \hat{x})$ as $P(z_2)$ using Rule 3. Since Z_2 blocks all back-door paths from X to Z_1, we can rewrite $P(z_1 \mid \hat{x}, z_2)$ as $P(z_1 \mid x, z_2)$ using Rule 2. The entire query can thus be rewritten as $\sum_{z_1,z_2} \sum_{x'} P(y \mid z_1, z_2, x')P(x' \mid z_2)P(z_1 \mid x, z_2)P(z_2)$. See examples in Figure 4.2. □

Theorem 4.3.2

The four conditions of Theorem 4.3.1 are necessary for identifiability in do *calculus. That is, if all four conditions of Theorem 4.3.1 fail in a graph G, then there exists no finite sequence of inference rules that reduces $P(y \mid \hat{x})$ to a hat-free expression.*

A proof of Theorem 4.3.2 is given in Galles and Pearl (1995).

4.3.2 Remarks on Efficiency

In implementing Theorem 4.3.1 as a systematic method for determining identifiability, Conditions 3 and 4 would seem to require exhaustive search. In order to prove that Condition 3 does not hold, for instance, we need to prove that no such blocking set B can exist. Fortunately, the following theorems allow us to significantly prune the search space so as to render the test tractable.

Theorem 4.3.3

If $P(b_i \mid \hat{x})$ is identifiable for one minimal set B_i, then $P(b_j \mid \hat{x})$ is identifiable for any other minimal set B_j.

Theorem 4.3.3 allows us to test Condition 3 with a single minimal blocking set B. If B meets the requirements of Condition 3 then the query is identifiable; otherwise, Condition 3 cannot be satisfied. In proving this theorem, we use the following lemma.

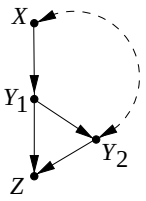


Figure 4.3 Theorem 4.3.1 ensures a reducing sequence for $P(y_2 \mid \hat{x}, y_1)$ and $P(y_1 \mid \hat{x})$, although none exists for $P(y_1 \mid \hat{x}, y_2)$.

Lemma 4.3.4
If the query $P(y \mid \hat{x})$ is identifiable and if a set of nodes Z lies on a directed path from X to Y, then the query $P(z \mid \hat{x})$ is identifiable.

Theorem 4.3.5
Let Y_1 and Y_2 be two subsets of nodes such that either (i) *no nodes Y_1 are descendants of X or* (ii) *all nodes Y_1 and Y_2 are descendants of X and all nodes Y_1 are nondescendants of Y_2. A reducing sequence for $P(y_1, y_2 \mid \hat{x})$ exists* (*per Corollary 3.4.2*) *if and only if there are reducing sequences for both $P(y_1 \mid \hat{x})$ and $P(y_2 \mid \hat{x}, y_1)$.*

The probability $P(y_1, y_2 \mid \hat{x})$ might pass the test in Theorem 4.3.1 if we apply the procedure to both $P(y_2 \mid \hat{x}, y_1)$ and $P(y_1 \mid \hat{x})$, but if we try to apply the test to $P(y_1 \mid \hat{x}, y_2)$ then we will not find a reducing sequence of rules. Figure 4.3 shows just such an example. Theorem 4.3.5 guarantees that, if there is a reducing sequence for $P(y_1, y_2 \mid \hat{x})$, then we should always be able to find such a sequence for both $P(y_1 \mid \hat{x})$ and $P(y_2 \mid \hat{x}, y_1)$ by proper choice of Y_1.

Theorem 4.3.6
If there exists a set Z_1 that meets all of the requirements for Z_1 in Condition 4, then the set consisting of the children of X intersected with the ancestors of Y will also meet all of the requirements for Z_1 in Condition 4.

Theorem 4.3.6 removes the need to search for Z_1 in Condition 4 of Theorem 4.3.1. Proofs of Theorems 4.3.3–4.3.6 are given in Galles and Pearl (1995).

4.3.3 Deriving a Closed-Form Expression for Control Queries

The algorithm defined by Theorem 4.3.1 not only determines the identifiability of a control query but also provides a closed-form expression for $P(y \mid \hat{x})$ in terms of the observed probability distribution (when such a closed form exists) as follows.

Function: ClosedForm$(P(y \mid \hat{x}))$.

Input: Control query of the form $P(y \mid \hat{x})$.

Output: Either a closed-form expression for $P(y \mid \hat{x})$, in terms of observed variables only, or FAIL when the query is not identifiable.

1. If $(X \perp\!\!\!\perp Y)_{G_{\overline{X}}}$ then return $P(y)$.
2. Otherwise, if $(X \perp\!\!\!\perp Y)_{G_{\underline{X}}}$ then return $P(y \mid x)$.

3. Otherwise, let $B = \text{BlockingSet}(X, Y)$ and $Pb = \text{ClosedForm}(b \mid \hat{x})$; if $Pb \neq$ FAIL then return $\sum_b P(y \mid b, x) * Pb$.
4. Otherwise, let $Z_1 = \text{Children}(X) \cap (Y \cup \text{Ancestors}(Y))$, $Z_3 = \text{BlockingSet}(X, Z_1)$, $Z_4 = \text{BlockingSet}(Z_1, Y)$, and $Z_2 = Z_3 \cup Z_4$; if $Y \notin Z_1$ and $X \notin Z_2$ then return $\sum_{z_1, z_2} \sum_{x'} P(y \mid z_1, z_2, x') P(x' \mid z_2) P(z_1 \mid x, z_2) P(z_2)$.
5. Otherwise, return FAIL.

Steps 3 and 4 invoke the function $\text{BlockingSet}(X, Y)$, which selects a set of nodes Z that d-separate X from Y. Such sets can be found in polynomial time (Tian et al. 1998). Step 3 contains a recursive call to the algorithm $\text{ClosedForm}(b \mid \hat{x})$ itself, in order to obtain an expression for causal effect $P(b \mid \hat{x})$.

4.3.4 Summary

The conditions of Theorem 4.3.1 sharply delineate the boundary between the class of identifying models (such as those depicted in Figure 3.8) and nonidentifying models (Figure 3.9). These conditions lead to an effective algorithm for determining the identifiability of control queries of the type $P(y \mid \hat{x})$, where X is a single variable. Such queries are identifiable in *do* calculus if and only if they meet the conditions of Theorem 4.3.1. The algorithm further gives a closed-form expression for the causal effect $P(y \mid \hat{x})$ in terms of estimable probabilities.

Applications to causal analysis of nonexperimental data in the social and medical sciences are discussed in Chapter 3 and further elaborated in Chapters 5 and 6. In Chapter 9 (Corollary 9.2.17) we will apply these results to problems of *causal attribution*, that is, to estimate the probability that a specific observation (e.g., a disease case) is causally attributable to a given event (e.g., exposure).

4.4 THE IDENTIFICATION OF PLANS

This section, based on Pearl and Robins (1995), concerns the probabilistic evaluation of plans in the presence of unmeasured variables, where each plan consists of several concurrent or sequential actions and each action may be influenced by its predecessors in the plan. We establish a graphical criterion for recognizing when the effects of a given plan can be predicted from passive observations on measured variables only. When the criterion is satisfied, a closed-form expression is provided for the probability that the plan will achieve a specified goal.

4.4.1 Motivation

To motivate the discussion, consider an example discussed in Robins (1993, apx. 2), as depicted in Figure 4.4. The variables X_1 and X_2 stand for treatments that physicians prescribe to a patient at two different times, Z represents observations that the second physician consults to determine X_2, and Y represents the patient's survival. The hidden variables U_1 and U_2 represent, respectively, part of the patient's history and the patient's disposition

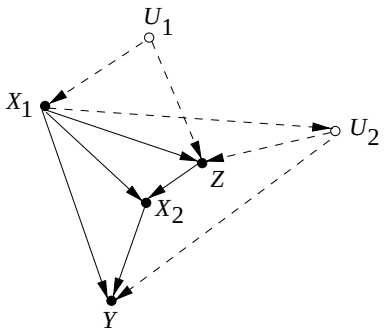


Figure 4.4 The problem of evaluating the effect of the plan $(do(x_1), do(x_2))$ on Y, from nonexperimental data taken on X_1, Z, X_2, and Y.

to recover. A simple realization of such structure could be found among AIDS patients, where Z represents episodes of PCP. This is a common opportunistic infection of AIDS patients that (as the diagram shows) does not have a direct effect on survival Y because it can be treated effectively, but it is an indicator of the patient's underlying immune status (U_2), which can cause death. The terms X_1 and X_2 stand for bactrim, a drug that prevents PCP (Z) and may also prevent death by other mechanisms. Doctors used the patient's earlier PCP history (U_1) to prescribe X_1, but its value was not recorded for data analysis.

The problem we face is as follows. Assume we have collected a large amount of data on the behavior of many patients and physicians, which is summarized in the form of (an estimated) joint distribution P of the observed four variables (X_1, Z, X_2, Y). A new patient comes in, and we wish to determine the impact of the (unconditional) plan $(do(x_1), do(x_2))$ on survival, where x_1 and x_2 are two predetermined dosages of bactrim to be administered at two prespecified times.

In general, our problem amounts to that of evaluating a new plan by watching the performance of other planners whose decision strategies are indiscernible. Physicians do not provide a description of all inputs that prompted them to prescribe a given treatment; all they communicate to us is that U_1 was consulted in determining X_1 and that Z and X_1 were consulted in determining X_2. But U_1, unfortunately, was not recorded. In epidemiology, the plan evaluation problem is known as "time-varying treatment with time-varying confounders" (Robins 1993). In artificial intelligence applications, the evaluation of such plans enables one agent to learn to act by observing the performance of another agent, even in cases where the actions of the other agent are predicated on factors that are not visible to the learner. If the learner is permitted to act as well as observe, then the task becomes much easier: the topology of the causal diagram could also be inferred (at least partially), and the effects of some previously unidentifiable actions could be determined.

As in the identification of actions (Section 4.3), the main problem in plan identification is the control of "confounders," that is, unobserved factors that trigger actions and simultaneously affect the response. However, unlike the problem treated in Section 4.3, plan identification is further complicated by the fact that some of the confounders (e.g. Z) are affected by control variables. As remarked in Chapter 3, one of the deadliest sins in the design of statistical experiments (Cox 1958, p. 48) is to adjust for such variables, because the adjustment would simulate holding a variable constant; holding constant a variable that stands between an action and its consequence interferes with the very quantity we wish estimate – the total effect of that action.

Two other features of Figure 4.4 are worth noting. First, the quantity $P(y \mid \hat{x}_1, \hat{x}_2)$ cannot be computed if we treat the control variables X_1 and X_2 as a single compound variable X. The graph corresponding to such compounding would depict X as connected to Y by both an arrow and a curved arc (through U) and thus would form a bow pattern (see Figure 3.9), which is indicative of nonidentifiability. Second, the causal effect $P(y \mid \hat{x}_1)$ in isolation is not identifiable because U_1 creates a bow pattern around the link $X \rightarrow Z$, which lies on a directed path from X to Y (see the discussion in Section 3.5).

The feature that facilitates the identifiability of $P(y \mid \hat{x}_1, \hat{x}_2)$ is the identifiability of $P(y \mid x_1, z, \hat{x}_2)$ – the causal effect of the action $do(X_2 = x_2)$ alone, conditioned on the observations available at the time of this action. This can be verified using the back-door criterion, observing that $\{X_1, Z\}$ blocks all back-door paths between X_2 and Y. Thus, the identifiability of $P(y \mid \hat{x}_1, \hat{x}_2)$ can be readily proven by writing

$$P(y \mid \hat{x}_1, \hat{x}_2) = P(y \mid x_1, \hat{x}_2) \tag{4.1}$$

$$= \sum_z P(y \mid z, x_1, \hat{x}_2) P(z \mid x_1) \tag{4.2}$$

$$= \sum_z P(y \mid z, x_1, x_2) P(z \mid x_1), \tag{4.3}$$

where (4.1) and (4.3) follow from Rule 2, and (4.2) follows from Rule 3. The subgraphs that permit the application of these rules are shown in Figure 4.5 (in Section 4.4.3).

This derivation also highlights how conditional plans can be evaluated. Assume we wish to evaluate the effect of the plan $\{do(X_1 = x_1), do(X_2 = g(x_1, z))\}$. Following the analysis of Section 4.2, we write

$$\begin{aligned} P(y \mid do(X_1 = x_1), do(X_2 = g(x_1, z))) &= P(y \mid x_1, do(X_2 = g(x_1, z))) \\ &= \sum_z P(y \mid z, x_1, do(X_2 = g(x_1, z))) P(z \mid x_1) \\ &= \sum_z P(y \mid z, x_1, x_2) P(z \mid x_1)|_{x_2 = g(x_1, z)}. \end{aligned}$$

Again, the identifiability of this conditional plan rests on the identifiability of the expression $P(y \mid z, x_1, \hat{x}_2)$, which reduces to $P(y \mid z, x_1, x_2)$ because $\{X_1, Z\}$ blocks all back-door paths between X_2 and Y.

The criterion developed in the next section will enable us to recognize in general, by graphical means, whether a proposed plan can be evaluated from the joint distribution on the observables and, if so, to identify which covariates should be measured and how they should be adjusted.

4.4.2 Plan Identification: Notation and Assumptions

Our starting point is a knowledge specification scheme in the form of a causal diagram, like the one shown in Figure 4.4, that provides a qualitative summary of the analyst's understanding of the relevant data-generating processes.[5]

[5] An alternative specification scheme using counterfactual statements was developed by Robins (1986, 1987), as described in Section 3.6.4.

Notation

A *control problem* consists of a directed acyclic graph (DAG) G with vertex set V, partitioned into four disjoint sets $V = \{X, Z, U, Y\}$, where

X = the set of control variables (exposures, interventions, treatments, etc.);

Z = the set of observed variables, often called *covariates*;

U = the set of unobserved (latent) variables; and

Y = an outcome variable.

We let the control variables be ordered $X = X_1, X_2, \ldots, X_n$ such that every X_k is a nondescendant of X_{k+j} ($j > 0$) in G, and we let the outcome Y be a descendant of X_n. Let N_k stand for the set of observed nodes that are nondescendants of any element in the set $\{X_k, X_{k+1}, \ldots, X_n\}$. A *plan* is an ordered sequence $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ of value assignments to the control variables, where $\hat{x}_k$ means "X_k is set to x_k." A *conditional plan* is an ordered sequence $(\hat{g}_1(z_1), \hat{g}_2(z_2), \ldots, \hat{g}_n(z_n))$, where each g_k is a function from a set Z_k to X_k and where $\hat{g}_k(z_k)$ stands for the statement "set X_k to $g_k(z_k)$ whenever Z_k attains the value z_k." The support Z_k of each $g_k(z_k)$ function must not contain any variables that are descendants of X_k in G.

Our problem is to *evaluate* an unconditional plan[6] by computing $P(y \mid \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, which represents the impact of the plan $(\hat{x}_1, \ldots, \hat{x}_n)$ on the outcome variable Y. The expression $P(y \mid \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ is said to be *identifiable* in G if, for every assignment $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, the expression can be determined uniquely from the joint distribution of the observables $\{X, Y, Z\}$. A control problem is identifiable whenever $P(y \mid \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ is identifiable.

Our main identifiability criteria are presented in Theorems 4.41 and 4.4.6. These invoke d-separation tests on various subgraphs of G, defined in the same manner as in Section 4.3. We denote by $G_{\overline{X}}$ (and $G_{\underline{X}}$, respectively) the graphs obtained by deleting from G all arrows pointing to (emerging from) nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$. Finally, the expression $P(y \mid \hat{x}, z) \triangleq P(y, z \mid \hat{x})/P(z \mid \hat{x})$ stands for the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

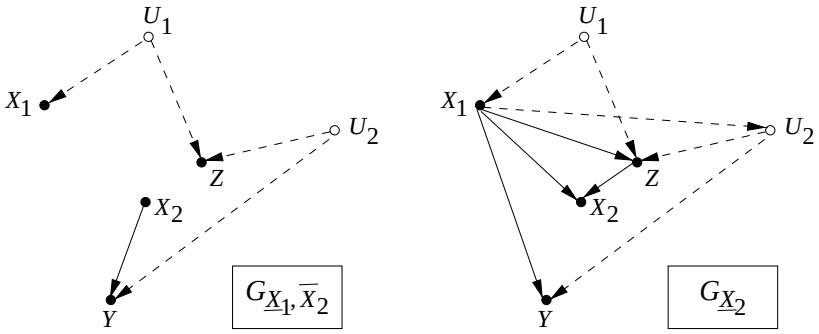
4.4.3 Plan Identification: A General Criterion

Theorem 4.4.1 (Pearl and Robins 1995)
The probability $P(y \mid \hat{x}_1, \ldots, \hat{x}_n)$ is identifiable if, for every $1 \leq k \leq n$, there exists a set Z_k of covariates satisfying

$$Z_k \subseteq N_k, \tag{4.4}$$

(*i.e., Z_k consists of nondescendants of $\{X_k, X_{k+1}, \ldots, X_n\}$*) *and*

[6] Identification of conditional plans can be obtained from Theorem 4.4.1 using the method described in Section 4.2 and exemplified in Section 4.4.1.



(a) (b)

Figure 4.5 The two subgraphs of G used in testing the identifiability of the plan $(\hat{x}_1, \hat{x}_2)$ in Figure 4.4.

$$(Y \perp\!\!\!\perp X_k \mid X_1, \ldots, X_{k-1}, Z_1, Z_2, \ldots, Z_k)_{G_{\underline{X}_k, \overline{X}_{k+1}, \ldots, \overline{X}_n}}. \tag{4.5}$$

When these conditions are satisfied, the effect of the plan is given by

$$P(y \mid \hat{x}_1, \ldots, \hat{x}_n) = \sum_{z_1, \ldots, z_n} P(y \mid z_1, \ldots, z_n, x_1, \ldots, x_n) \times \prod_{k=1}^{n} P(z_k \mid z_1, \ldots, z_{k-1}, x_1, \ldots, x_{k-1}). \tag{4.6}$$

Before presenting its proof, let us demonstrate how Theorem 4.4.1 can be used to test the identifiability of the control problem shown in Figure 4.4. First, we will show that $P(y \mid \hat{x}_1, \hat{x}_2)$ cannot be identified without measuring Z; in other words, that the sequence $Z_1 = \emptyset$, $Z_2 = \emptyset$ would not satisfy conditions (4.4)–(4.5). The two d-separation tests encoded in (4.5) are

$$(Y \perp\!\!\!\perp X_1)_{G_{\underline{X}_1, \overline{X}_2}} \quad \text{and} \quad (Y \perp\!\!\!\perp X_2 \mid X_1)_{G_{\underline{X}_2}}.$$

The two subgraphs associated with these tests are shown in Figure 4.5. We see that $(Y \perp\!\!\!\perp X_1)$ holds in $G_{\underline{X}_1, \overline{X}_2}$ but that $(Y \perp\!\!\!\perp X_2 \mid X_1)$ fails to hold in $G_{\underline{X}_2}$. Thus, in order to pass the test, we must have either $Z_1 = \{Z\}$ or $Z_2 = \{Z\}$; since Z is a descendant of X_1, only the second alternative satisfies (4.4). The tests applicable to the sequence $Z_1 = \emptyset$, $Z_2 = \{Z\}$ are $(Y \perp\!\!\!\perp X_1)_{G_{\underline{X}_1, \overline{X}_2}}$ and $(Y \perp\!\!\!\perp X_2 \mid X_1, Z)_{G_{\underline{X}_2}}$. Figure 4.5 shows that both tests are now satisfied, because $\{X_1, Z\}$ d-separates Y from X_2 in $G_{\underline{X}_2}$. Having satisfied conditions (4.4)–(4.5), equation (4.6) provides a formula for the effect of plan $(\hat{x}_1, \hat{x}_2)$ on Y:

$$P(y \mid \hat{x}_1, \hat{x}_2) = \sum_{z} P(y \mid z, x_1, x_2) P(z \mid x_1), \tag{4.7}$$

which coincides with (4.3).

The question naturally arises of whether the sequence $Z_1 = \emptyset$, $Z_2 = \{Z\}$ can be identified without exhaustive search. This question will be answered in Corollary 4.4.5 and Theorem 4.4.6.

Proof of Theorem 4.4.1

The proof given here is based on the inference rules of *do* calculus (Theorem 3.4.1), which facilitate the reduction of causal effect formulas to hat-free expressions. An alternative proof, using latent variable elimination, is given in Pearl and Robins (1995).

Step 1. The condition $Z_k \subseteq N_k$ implies $Z_k \subseteq N_j$ for all $j \geq k$. Therefore, we have

$$P(z_k \mid z_1, \ldots, z_{k-1}, x_1, \ldots, x_{k-1}, \hat{x}_k, \hat{x}_{k+1}, \ldots, \hat{x}_n) = P(z_k \mid z_1, \ldots, z_{k-1}, x_1, \ldots, x_{k-1}).$$

This is so because no node in $\{Z_1, \ldots, Z_k, X_1, \ldots, X_{k-1}\}$ can be a descendant of any node in $\{X_k, \ldots, X_n\}$. Hence, Rule 3 allows us to delete the hat variables from the expression.

Step 2. The condition in (4.5) permits us to invoke Rule 2 and write:

$$P(y \mid z_1, \ldots, z_k, x_1, \ldots, x_{k-1}, \hat{x}_k, \hat{x}_{k+1}, \ldots, \hat{x}_n) = P(y \mid z_1, \ldots, z_k, x_1, \ldots, x_{k-1}, x_k, \hat{x}_{k+1}, \ldots, \hat{x}_n).$$

Thus, we have

$$\begin{aligned}
&P(y \mid \hat{x}_1, \ldots, \hat{x}_n) \\
&= \sum_{z_1} P(y \mid z_1, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) P(z_1 \mid \hat{x}_1, \ldots, \hat{x}_n) \\
&= \sum_{z_1} P(y \mid z_1, x_1, \hat{x}_2, \ldots, \hat{x}_n) P(z_1) \\
&= \sum_{z_2} \sum_{z_1} P(y \mid z_1, z_2, x_1, \hat{x}_2, \ldots, \hat{x}_n) P(z_1) P(z_2 \mid z_1, x_1, \hat{x}_2, \ldots, \hat{x}_n) \\
&= \sum_{z_2} \sum_{z_1} P(y \mid z_1, z_2, x_1, x_2, \hat{x}_3, \ldots, \hat{x}_n) P(z_1) P(z_2 \mid z_1, x_1) \\
&\vdots \\
&= \sum_{z_n} \cdots \sum_{z_2} \sum_{z_1} P(y \mid z_1, \ldots, z_n, x_1, \ldots, x_n) \\
&\quad \times P(z_1) P(z_2 \mid z_1, x_1) \cdots P(z_n \mid z_1, x_1, z_2, x_2, \ldots, z_{n-1}, x_{n-1}) \\
&= \sum_{z_1, \ldots, z_n} P(y \mid z_1, \ldots, z_n, x_1, \ldots, x_n) \prod_{k=1}^{n} P(z_k \mid z_1, \ldots, z_{k-1}, x_1, \ldots, x_{k-1}). \qquad \square
\end{aligned}$$

Definition 4.4.2 (Admissible Sequence and *G*-Identifiability)

Any sequence $Z_1, \ldots, Z_n$ of covariates satisfying the conditions in (4.4)–(4.5) *will be called* admissible, *and any expression $P(y \mid \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ that is identifiable by the criterion of Theorem 4.4.1 will be called G-*identifiable.[7]

[7] The term "*G*-admissibility" was used in Pearl and Robins (1995) to evoke two associations: (1) Robins's *G-estimation* formula (equation (3.63)), which coincides with (4.6) when G is complete and contains no unobserved confounders; and (2) the *graphical* nature of the conditions in (4.4)–(4.5).

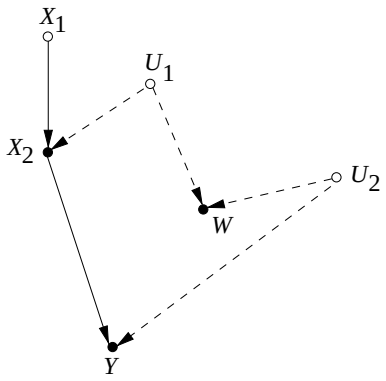


Figure 4.6 An admissible choice $Z_1 = W$ that rules out any admissible choice for Z_2. The choice $Z_1 = \emptyset$ would permit the construction of an admissible sequence ($Z_1 = \emptyset$, $Z_2 = \emptyset$).

The following corollary is immediate.

Corollary 4.4.3
A control problem is G-identifiable if and only if it has an admissible sequence.

The property of G-identifiability is sufficient but not necessary for general plan identifiability as defined in Section 4.4.2. The reasons are twofold. First, the completeness of the three inference rules of *do* calculus is still a pending conjecture. Second, the kth step in the reduction of (4.6) refrains from conditioning on variables Z_k that are descendants of X_k – namely, variables that may be affected by the action $do(X_k = x_k)$. In certain causal structures, the identifiability of causal effects requires that we condition on such variables, as demonstrated by the front-door criterion (Theorem 3.3.4).

4.4.4 Plan Identification: A Procedure

Theorem 4.4.1 provides a declarative condition for plan identifiability. It can be used to ratify that a proposed formula is valid for a given plan, but it does not provide an effective procedure for deriving such formulas because the choice of each Z_k is not spelled out procedurally. The possibility exists that some unfortunate choice of Z_k satisfying (4.4) and (4.5) might prevent us from continuing the reduction process even though another reduction sequence is feasible.

This is illustrated in Figure 4.6. Here W is an admissible choice for Z_1, but if we make this choice then we will not be able to complete the reduction, since no set Z_2 can be found that satisfies condition (4.5): $(Y \perp\!\!\!\perp X_2 \mid X_1, W, Z_2)_{G_{\underline{X}_2}}$. In this example it would be wiser to choose $Z_1 = Z_2 = \emptyset$, which satisfies both $(Y \perp\!\!\!\perp X_1 \mid \emptyset)_{G_{\underline{X}_1, \overline{X}_2}}$ and $(Y \perp\!\!\!\perp X_2 \mid X_1, \emptyset)_{G_{\underline{X}_2}}$.

The obvious way to avoid bad choices of covariates, like the one illustrated in Figure 4.6, is to insist on always choosing a "minimal" Z_k, namely, a set of covariates satisfying (4.5) that has no proper subset satisfying (4.5). However, since there are usually many such minimal sets (see Figure 4.7), the question remains of whether every choice of a minimal Z_k is "safe": Can we be sure that no choice of a minimal subsequence $Z_1, \ldots, Z_k$ will ever prevent us from finding an admissible Z_{k+1} when some admissible sequence $Z_1^*, \ldots, Z_n^*$ exists?

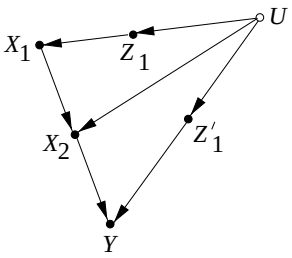


Figure 4.7 Nonuniqueness of minimal admissible sets: Z_1 and Z_1' are each minimal and admissible, since $(Y \perp\!\!\!\perp X_1 \mid Z_1)$ and $(Y \perp\!\!\!\perp X_1 \mid Z_1')$ both hold in $G_{\underline{X}_1, \overline{X}_2}$.

The next result guarantees the safety of every minimal subsequence $Z_1, \ldots, Z_k$ and hence provides an effective test for G-identifiability.

Theorem 4.4.4
If there exists an admissible sequence $Z_1^*, \ldots, Z_n^*$ then, for every minimally admissible subsequence $Z_1, \ldots, Z_{k-1}$ of covariates, there is an admissible set Z_k.

A proof is given in Pearl and Robins (1995).

Theorem 4.4.4 now yields an effective decision procedure for testing G-identifiability as follows.

Corollary 4.4.5
A control problem is G-identifiable if and only if the following algorithm exits with success.

1. *Set* $k = 1$.
2. *Choose any minimal* $Z_k \subseteq N_k$ *satisfying* (4.5).
3. *If no such* Z_k *exists then exit with failure; else set* $k = k + 1$.
4. *If* $k = n + 1$ *then exit with success; else return to step* 2.

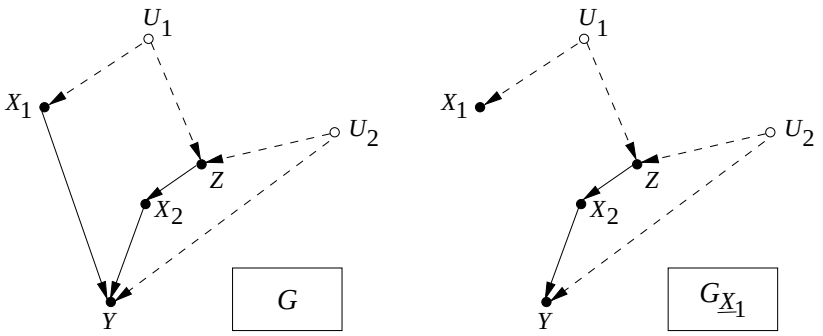
A further variant of Theorem 4.4.4 can be stated that avoids the search for minimal sets Z_k. This follows from the realization that, if an admissible sequence exists, we can rewrite Theorem 4.4.1 in terms of an explicit sequence of covariates $W_1, W_2, \ldots, W_n$ that can easily be identified in G.

Theorem 4.4.6
The probability $P(y \mid \hat{x}_1, \ldots, \hat{x}_n)$ is G-identifiable if and only if the following condition holds for every $1 \leq k \leq n$:

$$(Y \perp\!\!\!\perp X_k \mid X_1, \ldots, X_{k-1}, W_1, W_2, \ldots, W_k)_{G_{\underline{X}_k, \overline{X}_{k+1}, \ldots, \overline{X}_n}},$$

where W_k is the set of all covariates in G that are both nondescendants of $\{X_k, X_{k+1}, \ldots, X_n\}$ and have either Y or X_k as descendant in $G_{\underline{X}_k, \overline{X}_{k+1}, \ldots, \overline{X}_n}$. Moreover, if this condition is satisfied then the plan evaluates as

$$P(y \mid \hat{x}_1, \ldots, \hat{x}_n) = \sum_{w_1, \ldots, w_n} P(y \mid w_1, \ldots, w_n, x_1, \ldots, x_n) \times \prod_{k=1}^{n} P(w_k \mid w_1, \ldots, w_{k-1}, x_1, \ldots, x_{k-1}). \tag{4.8}$$



(a) (b)

Figure 4.8 Causal diagram G in which proper ordering of the control variables X_1 and X_2 is important.

A proof of Theorem 4.4.6, together with several generalizations, can be found in Pearl and Robins (1995) and Robins (1997). Extensions to G-identifiability are reported in Kuroki and Miyakawa (1999).

The reader should note that, although Corollary 4.4.5 and Theorem 4.4.6 are procedural in the sense of offering systematic tests for plan identifiability, they are still *order-dependent*. It is quite possible that an admissible sequence exists in one ordering of the control variables and not in another when both orderings are consistent with the arrows in G. The graph G in Figure 4.8 illustrates such a case. It is obtained from Figure 4.4 by deleting the arrows $X_1 \rightarrow X_2$ and $X_1 \rightarrow Z$, so that the two control variables (X_1 and X_2) can be ordered arbitrarily. The ordering (X_1, X_2) would still admit the admissible sequence $(\emptyset, Z)$ as before, but no admissible sequence can be found for the ordering (X_2, X_1). This can be seen immediately from the graph $G_{\underline{X}_1}$, in which (according to (4.5) with $k = 1$) we need to find a set Z such that $\{X_2, Z\}$ d-separates Y from X_1. No such set exists.

The implication of this order sensitivity is that, whenever G permits several orderings of the control variables, all orderings need be examined before we can be sure that a plan is not G-identifiable. Whether an effective search exists through the space of such orderings remains an open question.

4.5 DIRECT EFFECTS AND THEIR IDENTIFICATION

4.5.1 Direct versus Total Effects

The causal effect we have analyzed so far, $P(y \mid \hat{x})$, measures the *total* effect of a variable (or a set of variables) X on a response variable Y. In many cases, this quantity does not adequately represent the target of investigation and attention is focused instead on the direct effect of X on Y. The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects (see Hesslow 1976; Cartwright 1989) tells the story of a birth-control pill that is suspect of producing thrombosis in women and, at the same time, has a negative indirect effect on thrombosis by reducing the rate of pregnancies (pregnancy is known to encourage thrombosis). In this example, interest is focused on the direct effect of the pill because it represents a stable biological relationship that, unlike the total effect, is invariant to marital status and other social factors that may affect women's chances of getting pregnant or of sustaining pregnancy.

Another class of examples involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

In all these examples, the requirement of holding the mediating variables fixed must be interpreted as (hypothetically) setting these variables to constants by physical intervention, not by analytical means such as selection, conditioning, or adjustment. For example, it will not be sufficient to measure the association between the birth-control pill and thrombosis separately among pregnant and nonpregnant women and then aggregate the results. Instead, we must perform the study among women who became pregnant before the use of the pill and among women who prevented pregnancy by means other than the drug. The reason is that, by conditioning on an intermediate variable (pregnancy in the example), we may create spurious associations between X and Y even when there is no direct effect of X on Y. This can easily be illustrated in the model $X \rightarrow Z \leftarrow U \rightarrow Y$, where X has no direct effect on Y. Physically holding Z constant would permit no association between X and Y, as can be seen by deleting all arrows entering Z. But if we were to condition on Z, a spurious association would be created through U (unobserved) that might be construed as a direct effect of X on Y.

4.5.2 Direct Effects, Definition, and Identification

Controlling all variables in a problem is obviously a major undertaking, if not an impossibility. The analysis of identification tells us under what conditions direct effects can be estimated from nonexperimental data even without such control. Using our $do(x)$ notation (or $\hat{x}$ for short), we can express the direct effect as follows.

Definition 4.5.1 (Direct Effect)
The direct effect of X on Y is given by $P(y \mid \hat{x}, \hat{s}_{XY})$, where S_{XY} is the set of all endogenous variables except X and Y in the system.

We see that the measurement of direct effects is ascribed to an ideal laboratory; the scientist controls for all possible conditions S_{XY} and need not be aware of the structure of the diagram or of which variables are truly intermediaries between X and Y. Much of the experimental control can be eliminated, however, if we know the structure of the diagram. For one thing, there is no need to actually hold *all* other variables constant; holding constant the direct parents of Y (excluding X) should suffice. Thus, we obtain the following equivalent definition of a direct effect.

Corollary 4.5.2
The direct effect of X on Y is given by $P(y \mid \hat{x}, \widehat{pa}_{Y\setminus X})$, *where* $pa_{Y\setminus X}$ *stands for any realization of the parents of Y, excluding X.*

Clearly, if X does not appear in the equation for Y (equivalently, if X is not a parent of Y), then $P(y \mid \hat{x}, \widehat{pa}_{Y\setminus X})$ defines a constant distribution on Y that is independent of x, thus matching our understanding of "having no direct effect." In general, assuming that X is a parent of Y, Corollary 4.5.2 implies that the direct effect of X on Y is identifiable whenever $P(y \mid \widehat{pa}_Y)$ is identifiable. Moreover, since the conditioning part of this expression corresponds to a plan in which the parents of Y are the control variables, we conclude that a direct effect is identifiable whenever the effect of the corresponding parents' plan is identifiable. We can now use the analysis of Section 4.4 and apply the graphical criteria of Theorems 4.4.1 and 4.4.6 to the analysis of direct effects. In particular, we can state our next theorem.

Theorem 4.5.3
Let $PA_Y = \{X_1, \ldots, X_k, \ldots, X_m\}$. *The direct effect of any* X_k *on Y is identifiable whenever the conditions of Corollary 4.4.5 hold for the plan* $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$ *in some admissible ordering of the variables. The direct effect is then given by* (4.8).

Theorem 4.5.3 implies that if the effect of one parent of Y is identifiable then the effect of every parent of Y is identifiable as well. Of course, the magnitude of the effect would differ from parent to parent, as seen in (4.8).

The following corollary is immediate.

Corollary 4.5.4
Let X_j *be a parent of Y. The direct effect of* X_j *on Y is, in general, nonidentifiable if there exists a confounding arc that embraces any link* $X_k \rightarrow Y$.

4.5.3 Example: Sex Discrimination in College Admission

To illustrate the use of this result, consider the study of Berkeley's alleged sex bias in graduate admission (Bickel et al. 1975), where data showed a higher rate of admission for male applicants overall but, when broken down by departments, a slight bias toward female applicants. The explanation was that female applicants tend to apply to the more competitive departments, where rejection rates are high; based on this finding, Berkeley was exonerated from charges of discrimination. The philosophical aspects of such reversals, known as Simpson's paradox, will be discussed more fully in Chapter 6. Here we focus on the question of whether adjustment for department is appropriate for assessing sex discrimination in college admission. Conventional wisdom has it that such adjustment is appropriate because "We know that applying to a popular department (one with considerably more applicants than positions) is just the kind of thing that causes rejection" (Cartwright 1983, p. 38), but we will soon see that additional factors should be considered.

Let us assume that the relevant factors in the Berkeley example are configured as in Figure 4.9, with the following interpretation of the variables:

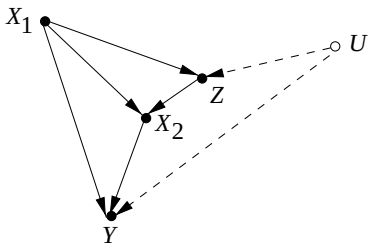


Figure 4.9 Causal relationships relevant to Berkeley's sex discrimination study. Adjusting for department choice (X_2) or career objective (Z) (or both) would be inappropriate in estimating the direct effect of gender on admission. The appropriate adjustment is given in (4.10).

X_1 = applicant's gender;

X_2 = applicant's choice of department;

Z = applicant's career objectives;

Y = admission outcome (accept/reject);

U = applicant's aptitude (unrecorded).

Note that U affects applicant's career objective and also the admission outcome Y (say, through verbal skills (unrecorded)).

Adjusting for department choice amounts to computing the following expression:

$$E_{x_2} P(y \mid \hat{x}_1, x_2) = \sum_{x_2} P(y \mid x_1, x_2) P(x_2). \tag{4.9}$$

In contrast, the direct effect of X_1 on Y, as given by (4.7), reads

$$P(y \mid \hat{x}_1, \hat{x}_2) = \sum_{z} P(y \mid z, x_1, x_2) P(z \mid x_1). \tag{4.10}$$

It is clear that the two expressions may differ substantially. The first measures the (average) effect of sex on admission among applicants to a given department, a quantity that is sensitive to the fact that some gender–department combinations may be associated with high admission rates merely because such combinations are indicative of certain aptitude (U) that was unrecorded. The second expression eliminates such spurious associations by separately adjusting for career objectives (Z) in each of the two genders.

To verify that (4.9) does not properly measure the direct effect of X_1 on Y, we note that the expression depends on the value of X_1 even in cases where the arrow between X_1 and Y is absent. Equation (4.10), on the other hand, becomes insensitive to x_1 in such cases – an exercise that we leave for the reader to verify.[8]

To cast this analysis in a concrete numerical setting, let us imagine a college consisting of two departments, A and B, both admitting students on the basis of qualification, Q, alone. Let us further assume (i) that the applicant pool consists of 100 males and 100 females and (ii) that 50 applicants in each gender have high qualifications (hence are admitted) and 50 have low qualifications (hence are rejected). Clearly, this college cannot be accused of sex discrimination.

[8] *Hint:* Factorize $P(y, u, z \mid \hat{x}_1, \hat{x}_2)$ using the independencies in the graph and eliminate u as in the derivation of (3.27).

Table 4.1. ***Admission Rate among Males and Females in Each Department***

	Males		Females		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Dept. *A*	50	50	0	0	50	50
Dept. *B*	0	50	50	100	50	150
Unadjusted	50%		50%		50%	
Adjusted	25%		37.5%			

A different result would surface, however, if we adjust for departments while ignoring qualifications, which amounts to using (4.9) to estimate the effect of gender on admission. Assume that the nature of the departments is such that *all and only* qualified male applicants apply to department A, while all females apply to department B (see Table 4.1).

We see from the table that adjusting for department would falsely indicate a bias of $37.5 : 25 (= 3 : 2)$ in favor of female applicants. An unadjusted (sometimes called "crude") analysis happens to give the correct result in this example – 50% admission rate for males and females alike – thus exonerating the school from charges of sex discrimination.

Our analysis is not meant to imply that the Berkeley study of Bickel et al. (1975) is defective, or that adjustment for department was not justified in that study. The purpose is to emphasize that no adjustment is guaranteed to give an unbiased estimate of causal effects, direct or indirect, absent a careful examination of the causal assumptions that ensure identification. Theorem 4.5.3 provides us with the understanding of those assumptions and with a mathematical means of expressing them. We note that if applicants' qualifications were not recorded in the data, then the direct effect of gender on admission will not be identifiable unless we can measure some proxy variable that stands in the same relation to Q as Z stands to U in Figure 4.9.

4.5.4 Average Direct Effects

Readers versed in structural equation models (SEMs) will note that, in linear systems, the direct effect $E(Y \mid \hat{x}, \widehat{pa}_{Y\setminus X})$ is fully specified by the path coefficient attached to the link from X to Y (see (5.24) for mathematical definition); therefore, the direct effect is independent of the values $pa_{Y\setminus X}$ at which we hold the other parents of Y. In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. For example, the direct effect of a pill on thrombosis would most likely be different for pregnant and nonpregnant women. Epidemiologists call such differences "effect modification" and insist on separately reporting the effect in each subpopulation.

Although the direct effect is sensitive to the levels at which we hold the parents of the outcome variable, it is sometimes meaningful to average the direct effect over those levels. For example, if we wish to assess the degree of discrimination in a given school without reference to specific departments, we should replace the controlled difference

$$P(\text{admission} \mid \widehat{\text{male}}, \widehat{\text{dept}}) - P(\text{admission} \mid \widehat{\text{female}}, \widehat{\text{dept}})$$

with some average of this difference over all departments. This average should measure the increase in admission rate in a hypothetical experiment in which we instruct all female candidates to retain their department preferences but change their gender identification (on the application form) from female to male.

In general, the average direct effect is defined as the expected change in Y induced by changing X from x to x' while keeping the other parents of Y constant at whatever value they obtain under $do(x)$. This hypothetical change is what lawmakers instruct us to consider in race or sex discrimination cases: "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (In Carson versus Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996)).

The formal expression for this hypothetical change involves probabilities of (nested) counterfactuals (see Section 7.1 for semantics and computation) that cannot be written in terms of the $do(x)$ operator.[9] Therefore, the average direct effect cannot in general be identified, even from data obtained under randomized control of all variables. However, if certain assumptions of "no confounding" are deemed valid,[10] then the average direct effect can be reduced to

$$\Delta_{x,x'}(Y) = \sum_{pa_{Y\setminus X}} [E(Y \mid \hat{x'}, \widehat{pa}_{Y\setminus X}) - E(Y \mid \hat{x}, \widehat{pa}_{Y\setminus X})] P(pa_{Y\setminus X} \mid \hat{x}), \tag{4.11}$$

and the techniques developed in Section 4.4 for identifying control-specific plans, $P(y \mid \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, become applicable.

Acknowledgment

Sections 4.3 and 4.4 are based, respectively, on collaborative works with David Galles and James Robins.

[9] Using the counterfactual notation of Section 7.1, the general expression for the average direct effect is

$$\Delta_{x,x'}(Y) = E(Y_{x'Z_x}) - E(Y_x),$$

where $Z = pa_{Y\setminus X}$. The subscript $x'Z_x$ represents the operation of setting X to x' and, simultaneously, setting Z to whatever value it would have obtained under the setting $X = x$. This general expression reduces to (4.11) if $Z_x \perp\!\!\!\perp Y_{x'z}$ holds for all z. Likewise, the average *indirect* effect is defined as $E(Y_{xZ_{x'}}) - E(Y_x)$.

[10] See details in Technical Report R-273 posted on www.cs.ucla.edu/~judea/.