Structural Causal Models (SCMs) for Dummies

Part I

- Kaushal Paneri

Notes from *“Elements of Causal Inference” - Jonas Peters et al.*

Correlation doesn’t imply Causation!

We may at least infer the existence of causal links from statistical dependences

(Less well known fact)

Reichenbach's Common Cause Principle

- If two random variables ***X*** and ***Y*** are statistically dependent, then there exists a third variable ***Z*** that causally influences both.

- As a special case, ***Z*** may coincide with either ***X*** or ***Y***.

- Furthermore, ***Z*** screens ***X*** and ***Y*** from each other in the sense that given ***Z***, they become independent.

Example: MNIST
Model I

- We provide a sequence of class labels ***y*** to a human writer.

- ***X***: Handwritten digit image

- ***y***: class label

- ***X*** := f (***y***,$\boldsymbol{N_x}$), where $\boldsymbol{N_x}$ is the observational noise.

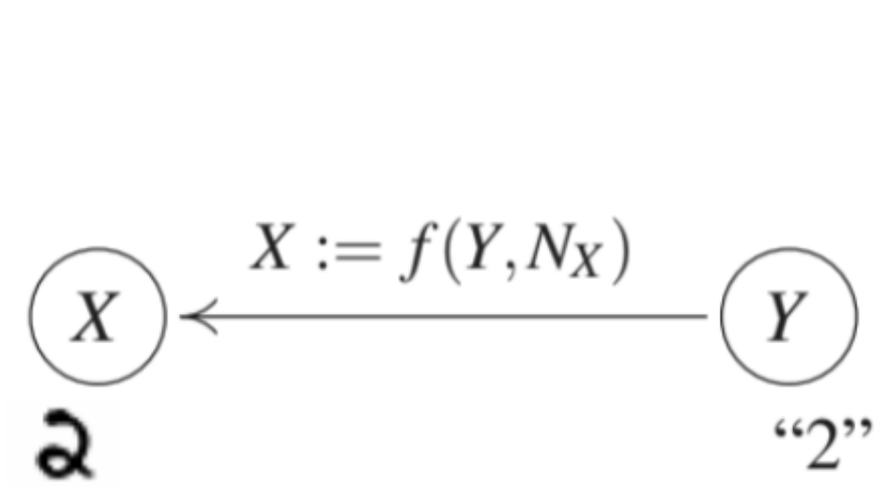
Example: MNIST
Model II

- We do not provide class labels to the writer. She decides herself what to write and she records the digits alongside.

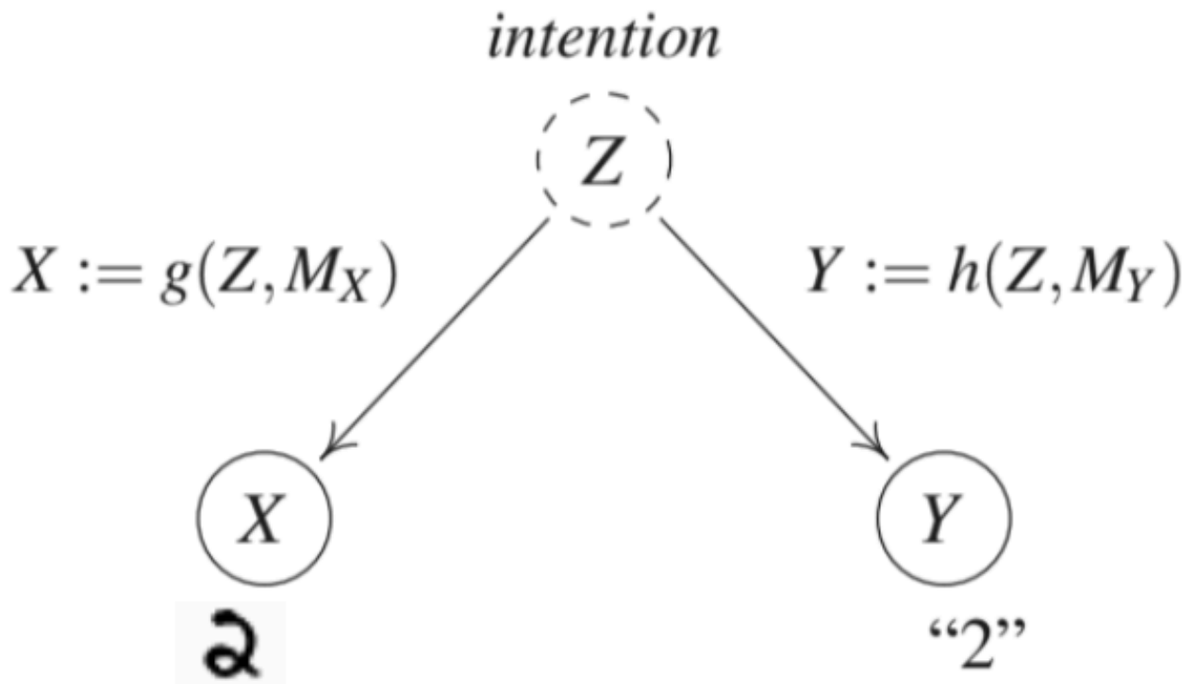
- In this case, both image ***X*** and label ***y*** are functions of the writer's intention (Call it ***Z***).

- We assume that not only the process of generating images is noisy but also the one recording the label will have noise which is independent of observational noise.

Example: MNIST

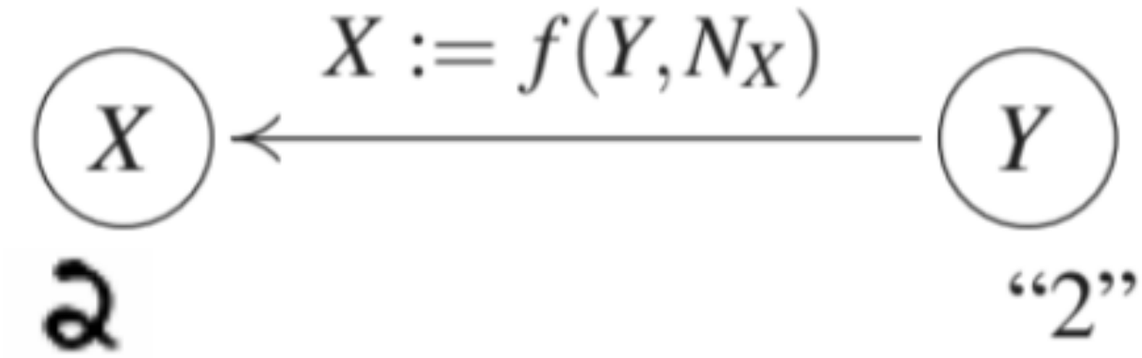


Model I



Model II

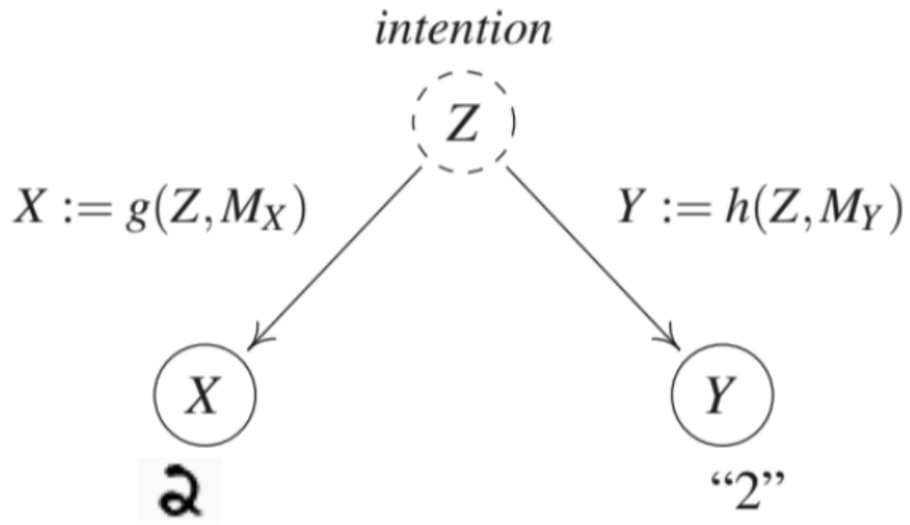
Model I
Possible Interventions



- ***X***: exchange it for another image after it has been produced. then this has no effect on the class labels that were provided to the writer.

- ***y***: change the class label provided to the writer. It will have strong effect on produced image.

Model II
Possible Interventions



- ***X***: exchange it for another image after it has been produced. then this has no effect on the class labels that were provided to the writer.

- ***y***: change the class label provided to the writer. Unlike before, this **will not affect** the image.

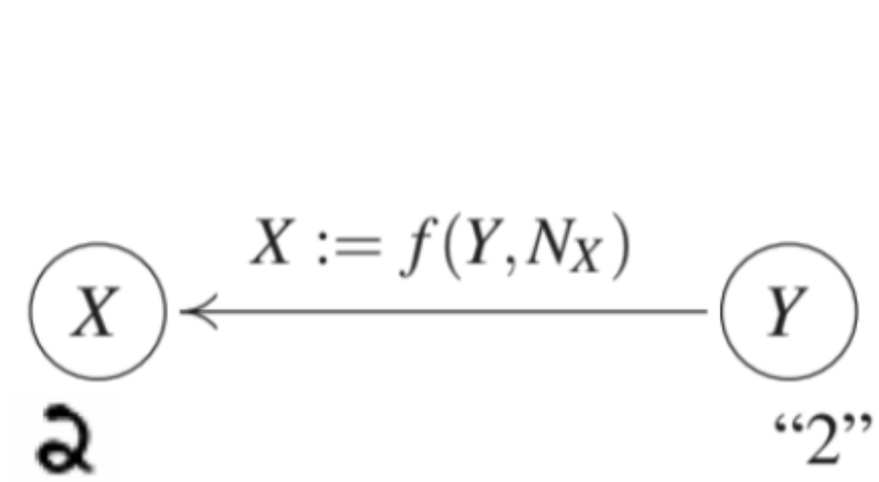
Example: MNIST

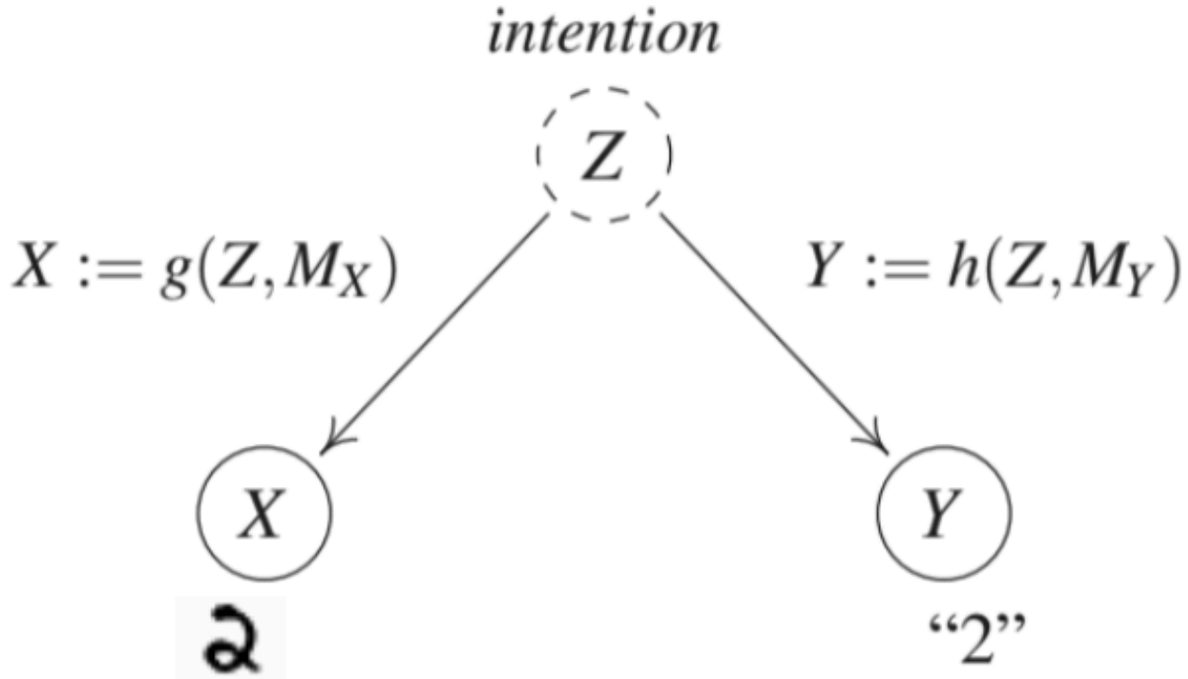
- Model I and II have same observational distribution over $\boldsymbol{X}$ and $\boldsymbol{y}$. but different intervention distributions.

- This difference is not visible in pure probabilistic description (where everything is derived from joint distribution $P_{\boldsymbol{X},\boldsymbol{y}}$).

- However, we were able to discuss it by incorporating structural knowledge about how $P_{\boldsymbol{X},\boldsymbol{y}.}$

Structural Causal Models



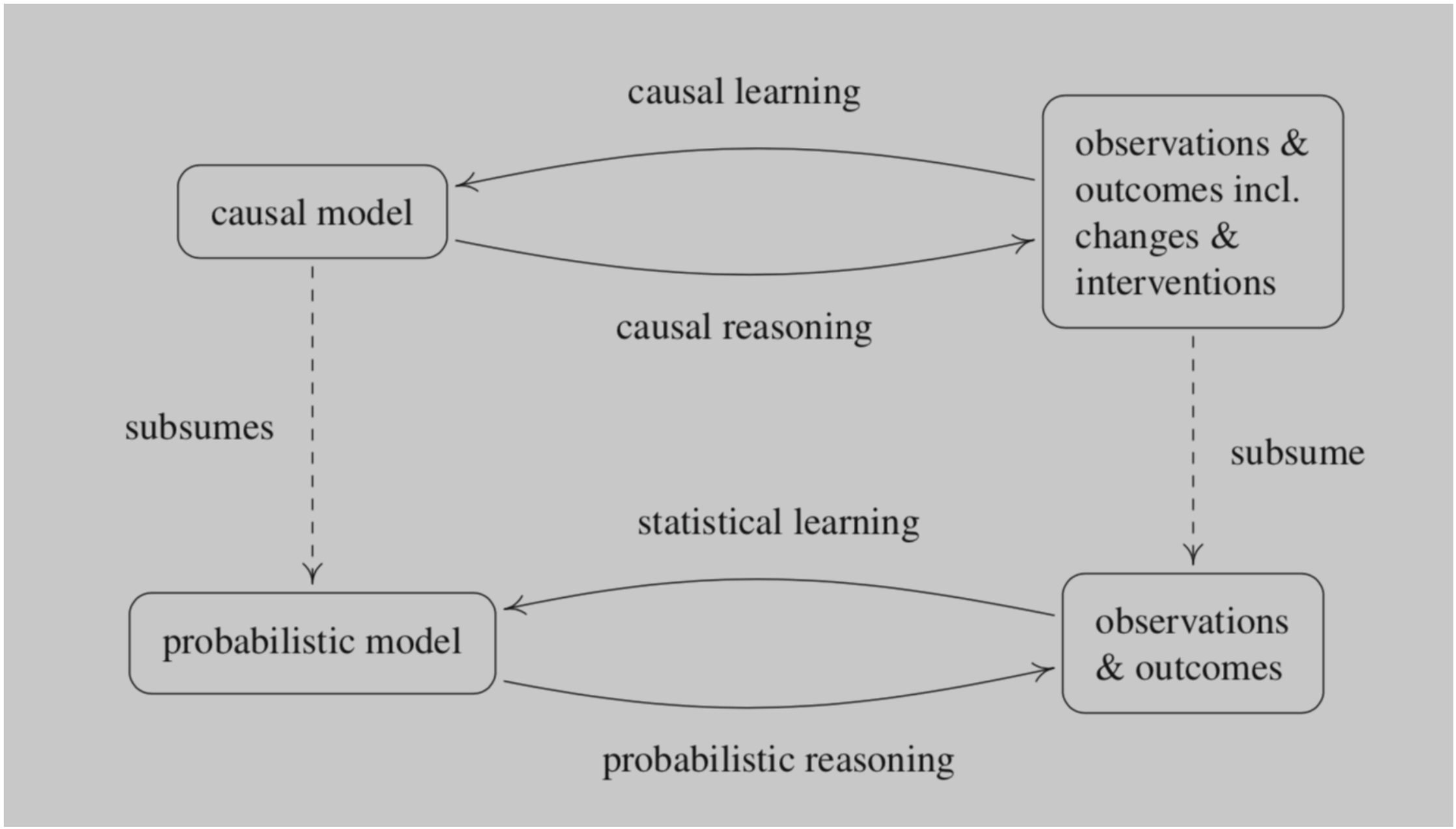


Beware of Similarity

SCMs	BayesNets
Entails a Joint distribution over all observations	Entails a Joint distribution over all observations
Same distribution can be generated by a different SCM	Same distribution can be generated by a different BayesNet

Transition from SCM to a probability model is possible, at the cost of loosing information about the effect of intervention.

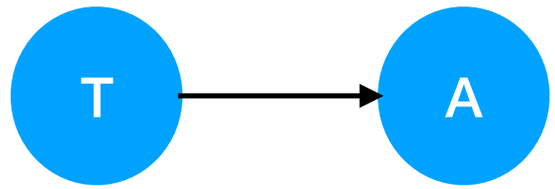
Causal Learning Vs. Statistical Learning



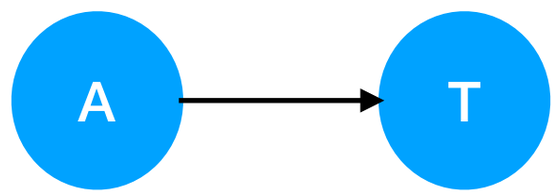
Which is the correct direction?

- Suppose we have estimated a joint density *p(a,t)*, of the altitude ***A*** and the average annual temperature ***T*** of a sample of cities in some country.

$$p(a,t) = p(a|t)\, p(t)$$



$$p(a,t) = p(t|a)\, p(a)$$



- How Can we decide which one is the causal structure?

Hypothetical Intervention

- Let’s perform an experiment where we change the altitude A of city by some mechanism that rises the grounds on which the city is built.
- We found out that temperature decreases.

- Now we do another intervention by building a massive heating system around the city that raises the average temperature by a few degrees.

Localized Intervention

- Note that if we change the altitude ***A***, then we assume that the physical mechanism *p(t|a)* responsible for producing an average temperature (e.g., chemical compositions of the atmosphere, laws of pressure, wind mechanics) is still in place and leads to a changed ***T***.

- So, the intuition is A->T is a correct causal structure if:

  - It is in principle possible to perform a localized intervention on A. i.e., to change *p(a)* without changing *p(t|a).*

  - *p(a)* and *p(t|a)* are autonomous or invariant mechanisms or objects in the world.

Principle of Independent Mechanisms

- *The causal generative process of system's variables is composed of autonomous modules that do to inform or influence each other.*

- *In a probabilistic case, this means that the conditional distribution of each variable given its causes (i.e. it's mechanisms) does not inform or influence the other conditional distributions. In case we have one two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.*

Bivariate SCM

Definition 3.1 (Structural causal models) *An SCM $\mathfrak{C}$ with graph $C \to E$ consists of two* ***assignments***

$$C := N_C, \tag{3.1}$$

$$E := f_E(C, N_E), \tag{3.2}$$

where $N_E \perp\!\!\!\perp N_C$, that is, N_E is independent of N_C.

Hard Intervention

- Suppose we may be interested in a situation in which variable ***E*** is set to the value 4 (irrespective of the value of ***C***). i.e., we replace assignment in *3.2* by E:=4.

$$\boldsymbol{C} := \boldsymbol{N_C}$$

$$\boldsymbol{E} := 4$$

- This is called a hard intervention and is denoted by *do(**E** := 4)*.

- The distribution over C is denoted by $P_C^{C';do(\boldsymbol{E}:=4)}$, or simply $P^{do(\boldsymbol{E}:=4)}$.

Soft Intervention

- Intervention $do(\mathbf{E} := g_E(\mathbf{C}) + N'_E)$ keeps a functional dependence on C but changes the noise distribution.

- This is an example of a Soft Intervention.

Example

Example 3.2 (Cause-effect interventions) Suppose that the distribution $P_{C,E}$ is entailed by an SCM $\mathfrak{C}$

$$\begin{aligned} C &:= N_C \\ E &:= 4 \cdot C + N_E, \end{aligned} \tag{3.3}$$

with $N_C, N_E \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and graph $C \to E$. Then,

$$\begin{aligned} P_E^{\mathfrak{C}} = \mathcal{N}(0,17) &\neq \mathcal{N}(8,1) = P_E^{\mathfrak{C};do(C:=2)} = P_{E|C=2}^{\mathfrak{C}} \\ &\neq \mathcal{N}(12,1) = P_E^{\mathfrak{C};do(C:=3)} = P_{E|C=3}^{\mathfrak{C}}. \end{aligned}$$

Example

$$C := N_C$$
$$E := 4 \cdot C + N_E,$$

$$P_C^{\mathfrak{C};do(E:=2)} = \mathcal{N}(0,1) = P_C^{\mathfrak{C}} = P_C^{\mathfrak{C};do(E:=314159265)} \left(\neq P_{C|E=2}^{\mathfrak{C}} \right).$$

No matter how strongly you intervene on E, the distribution of C remains what it was before.

This behavior corresponds well to our intuition of C “Causing” E.

Counterfactual: Example

Suppose there exists a rather effective treatment T for curing an eye disease.

For, 99% of Patients (P): It works & they get cured, (T=1) (B=0)
otherwise without treatment (T=0), they would have turned Blind (B=1) within a day.

For 1% of Patient (P): The opposite effect
T = 1, they go blind (B=1)
T = 0, they remain healthy (B=0)

(Which of these categories a patient belong is controlled by a rare condition N_B (unknown to the doctor.)
The decision of whether to give treatment or not is independent to N_B.
Suppose that decision is the noise variable N_T

Counterfactual: Example

Assume underlying causal model:

$$\mathcal{C} = \begin{cases} T := N_T \\ B := T N_B + (1-T)(1-N_B) \end{cases}$$

$N_B \sim \text{Ber}(0.01)$

Corresponding Causal Graph

Counterfactual: Example

Now imagine a specefic patient with poor eyesight comes to the hospital & goes blind after receiving the treatment.

We can now ask

What would have happened had the doctor didn't give the treatment ? ($T=0$)

The observation $B = T = 1$ implies that for a given patient, we had $N_B = 1$.

Counterfactual: Example

$\phi \mid B=1, T=1 \quad : \quad T := 1$

$B := T \cdot 1 + (1-T) \cdot (1-1) = T$

- Note that we only update the noise distributions; conditioning does not change the structure of the assignments themselves. The physical mechanism is unchanged that is: what leads to blindness.
- We have gained knowledge about the previously unknown noise variables for a given patient.
- Let's calculate the effect of *do(**T** = 0)* for this patient.

$\phi \mid B=1, T=1; do(T:=0) \quad : \quad T := 0$

$B := T$

Counterfactual: Example

$\phi \mid B=1, T=1; do(T:=0) \quad (B=0) = 1$

- However, we can still argue that doctor acted optimally (according to available knowledge).