Actions and their Effects

Michael U. Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, The University of Edinburgh

Autumn Semester 2025

Recap

- Topic 1: Representation What reasonably weak assumptions can we make to efficiently represent $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$?
  - Directed and undirected graphical models
  - Factorisation and independencies
- Topic 2: Exact inference Can we further exploit the assumptions on $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ to efficiently compute the posterior probability or derived quantities?
  - Yes! Factorisation can be exploited by using the distributive law and by caching computations.
  - Variable elimination and message passing algorithms
  - Inference for hidden Markov models
- Issue 3: Thank you for the numbers. But what shall I best do?

  Topic 3: Actions and decision making How to predict the outcome of actions and choose optimal actions?

Kidney stone example

	Overall success rate	Small stones	Large stones
Treatment *a*	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment *b*	**83%** (289/350)	87% (234/270)	69% (55/80)

- A hospital collects the data above on the success rate of two surgery procedures to remove kidney stones (data were collected in 1986).
- Treatment *a*: open surgery, treatment *b*: minimally-invasive procedure (percutaneous nephrolithotomy)
- Overall, treatment *b* looks to be more effective than *a*
- When broken down for both small and large kidney stones, treatment *a* is more effective than *b*.
- Which treatment (action) is more effective when the size of the kidney stones is unknown?

Example 6.37 in Peters, Janzing and Schölkopf, 2017

Kidney stone example

	Overall success rate	Small stones	Large stones
Treatment *a*	78% (273/**350**)	93% (81/87)	73% (192/**263**)
Treatment *b*	83% (289/**350**)	87% (234/**270**)	69% (55/80)

- Treatment assignment is not random: Treatment *a* tends to be assigned for cases of large stones (more difficult to treat), and treatment *b* for small stones (easier to treat).
- Surgeons may expect treatment *a* to be better than treatment *b* and therefore assign the difficult cases to treatment a with higher probability.
- Having more often to deal with difficult problems explains why treatment *a* performs better per subpopulation, but not overall.
- An example of “Simpson’s paradox”, where a trend that holds in all subpopulations may not hold at the population level.
- Still: which treatment is more effective when the size of the kidney stones is unknown?

Program

1. Modelling actions as interventions in causal DAGs
2. Computing the effect of interventions

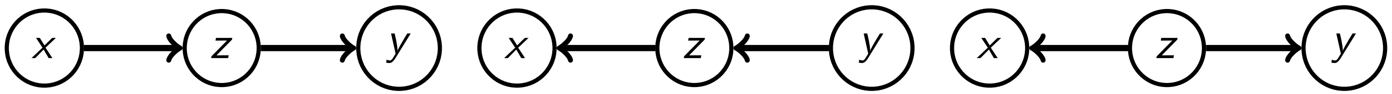
Program

1. Modelling actions as interventions in causal DAGs
   - Causal DAGs
   - Interventions change the data generating process
   - Interventions change the DAG locally

2. Computing the effect of interventions

Causal DAGs

- Causal DAGs are DAGs where the arrows are assumed to represent a causal direction.
- Causal DAGs represent nature's data-generating mechanism.
- Before: given $p(\mathbf{x})$, we drew a DAG based on the independencies and variable ordering chosen.
- In DGMs, the incoming arrows for x_i specified the parent set pa_i and hence what goes into the conditioning set in $p(x_i|\mathrm{pa}_i)$, but the arrows didn't have a mechanistic or causal meaning.
- This is different for causal DAGs.
- The following three graphs represent the same independencies but different causal mechanisms.



Actions as interventions in the data generating process

- As in DAGs, causal DAGs specify a data generating process via ancestral sampling.
- Different root nodes (nodes without parents) in the DAG represent independent root causes.
- Picking a topological ordering, we generate data according to $x_i \sim p(x_i|\mathrm{pa}_i)$ for all i. (for root nodes: $p(x_i|\mathrm{pa}_i) = p(x_i)$)
- We model an action on variable x_k as an intervention in the data generating process where x_k is not sampled from $p(x_k|\mathrm{pa}_k)$ but from a new distribution $p'(x_k)$.
- Intervention disconnects x_k from its parents and makes it a root variable (cause).
- When intervening on x_k, the data generating mechanisms of the other variables remain unchanged; we can change one mechanism without changing the others.

Actions as interventions in the data generating process

- This means each parent-child relationship in a causal DAG is thought to represent a stable and autonomous physical mechanism.
- Intervention defines a new model, the postinterventional distribution, that is denoted by $p(\mathbf{x}; do(x_k) \sim p')$ or $p(\mathbf{x}; do(x_k))$ for simplicity.
- Postinterventional distribution factorises as

$$p(\mathbf{x}; do(x_k) \sim p') = \prod_{i<k} p(x_i|\mathrm{pa}_i) \cdot p'(x_k) \cdot \prod_{i>k} p(x_i|\mathrm{pa}_i) \quad (1)$$

$$= \prod_{i \neq k} p(x_i|\mathrm{pa}_i) \cdot p'(x_k) \quad (2)$$

Atomic interventions

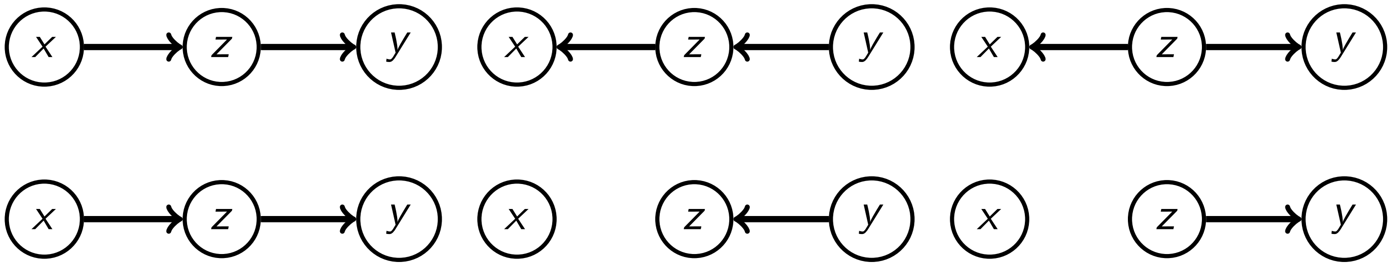
- Important special case is when the action/intervention sets a variable x_k to a specific value a.
- Called "atomic intervention" and corresponds to $p'(x_k) = \delta(x_k - a)$
- Postinterventional distribution is

$$p(\mathbf{x}; do(x_k) \sim \delta(x_k - a)) = \begin{cases} \prod_{i \neq k} p(x_i | \mathrm{pa}_i) & \text{if } x_k = a \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

- Notation: $p(\mathbf{x}; do(x_k) = a)$ or simply $p(\mathbf{x}; do(x_k))$ if clear from context.

Graph surgery

- Intervening on x_k makes it a root cause. Graphically, this means all incoming edges into x_k are removed.
- Resulting graph is denoted by $G_{\bar{x_k}}$ if G is the original graph.
- First row: original graph G. Second row: $G_{\bar{x}}$



Sprinkler example

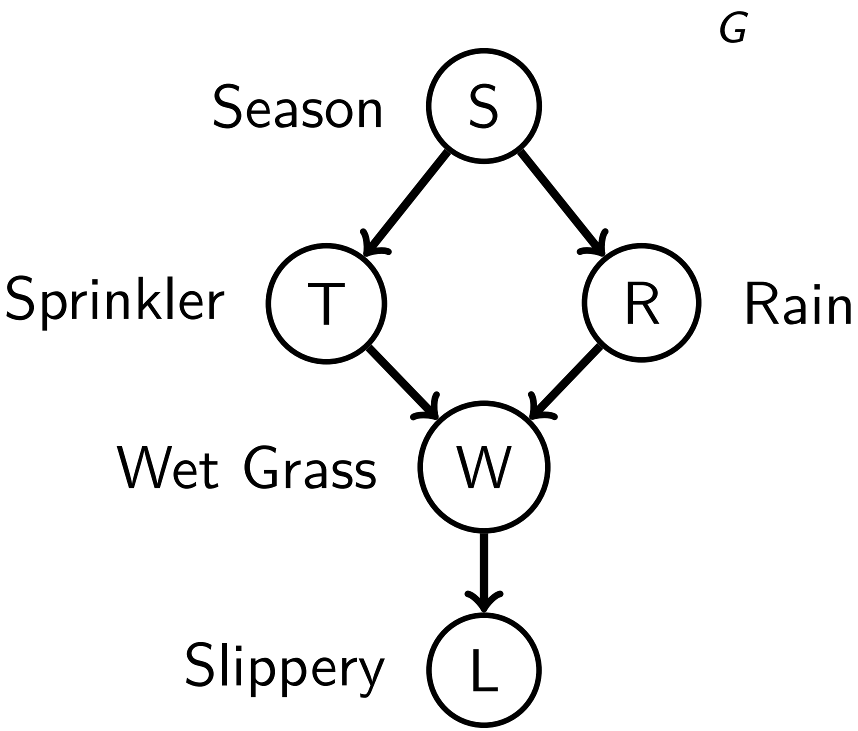
(based on Judea Pearl, Causality, 2nd ed, Ch1)

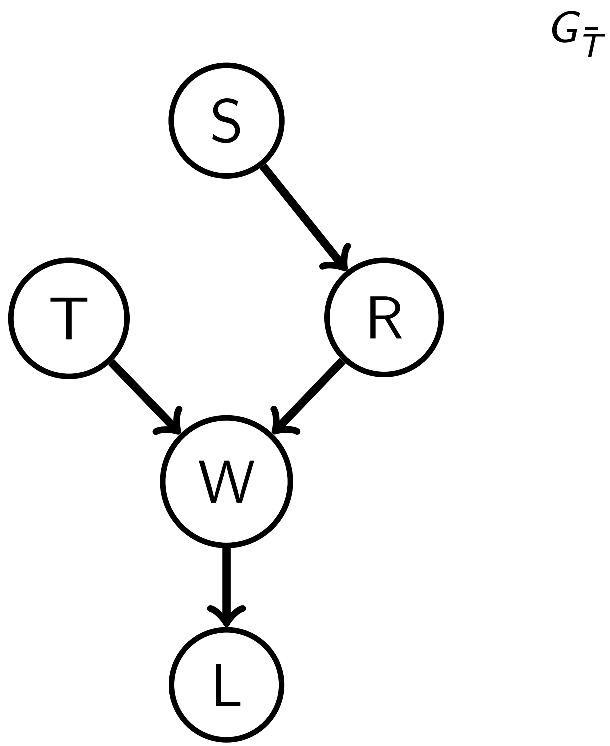
- Sprinklers tend to be on as a function of the season

$$p(S,T,R,W,L) = p(S)p(T|S)p(R|S)p(W|R,S)p(L|W)$$

- I can switch it on/off at any time, according to $p'(T)$

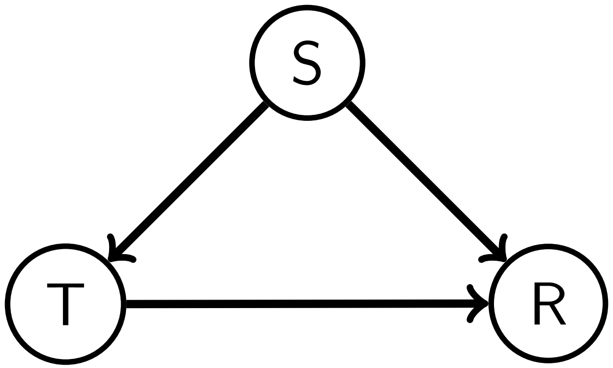
$$p(S,T,R,W,L; do(T)) = p(S)p'(T)p(R|S)p(W|R,S)p(L|W)$$



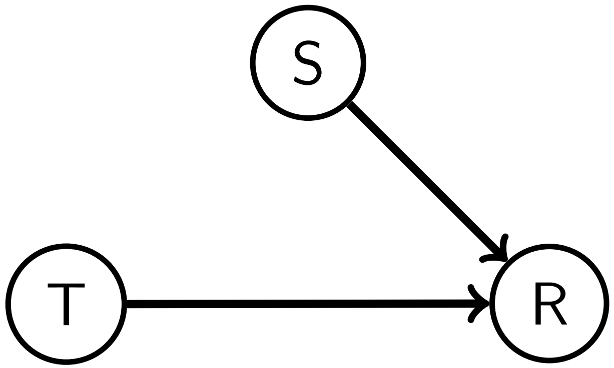


Kidney stone example

- In the kidney stone example, we had three binary variables: treatment T, stone size S, and the result R.
- Treatment is prescribed depending on stone size. Result also depends on the stone size (difficulty of surgery). This gives the DAG G.
- Variables such as S that are the common cause of other variables are called confounders.
- If we intervene on the treatment, we get the graph $G_{\bar{T}}$, disconnecting T from the confounder S.



G



$G_{\bar{T}}$

Program

1. Modelling actions as interventions in causal DAGs
   - Causal DAGs
   - Interventions change the data generating process
   - Interventions change the DAG locally
2. Computing the effect of interventions

Program

1. Modelling actions as interventions in causal DAGs

2. Computing the effect of interventions
   - Inverse probability weighting and adjustment for direct causes
   - Observing vs acting: the role of backdoors
   - Backdoor adjustment

How do we compute the effect of interventions (actions)?

- Recall the postinterventional distribution

$$p(\mathbf{x}; do(x_k) \sim p') = \prod_{i \neq k} p(x_i|\mathrm{pa}_i) \cdot p'(x_k) \quad (4)$$

- If all terms in the factorisation are known, we can compute marginals or conditionals using the inference techniques that we have seen so far (variable elimination, message passing if applicable etc).
- We can use the model to predict the effect/outcome of an intervention, e.g. compute $p(x_i|do(x_k))$ for some i, without performing the action.
- But computation may not always be (computationally) feasible. Limitation discussed on the inference slides apply.
- Let us leverage the connection between $p(\mathbf{x}; do(x_k) \sim p')$ and $p(\mathbf{x})$ to obtain alternatives.

Relation between pre and postinterventional distribution

$$p(\mathbf{x}; do(x_k) \sim p') = \prod_{i \neq k} p(x_i|\mathrm{pa}_i) \cdot p'(x_k)$$

- With $p(\mathbf{x}) = \prod_i p(x_i|\mathrm{pa}_i)$ prior to the intervention, it follows that

$$p(\mathbf{x}; do(x_k) \sim p') = \frac{p(\mathbf{x})}{p(x_k|\mathrm{pa}_k)} p'(x_k) \quad (5)$$

- With $p(x_k|\mathrm{pa}_k) = p(x_k, \mathrm{pa}_k)/p(\mathrm{pa}_k)$, we have

$$p(\mathbf{x}; do(x_k) \sim p') = \frac{p(\mathbf{x})}{p(x_k, \mathrm{pa}_k)} p(\mathrm{pa}_k) p'(x_k) \quad (6)$$

$$= p(\tilde{\mathbf{x}}_k|x_k, \mathrm{pa}_k) p(\mathrm{pa}_k) p'(x_k) \quad (7)$$

  where $\tilde{\mathbf{x}}_k$ denotes all variables but x_k, pa_k.

- Gives rise to two methods: inverse probability weighting and adjustment for direct causes.

Inverse probability weighting

$$p(\mathbf{x}; do(x_k) \sim p') = \frac{p(\mathbf{x})}{p(x_k|\mathrm{pa}_k)} p'(x_k)$$

- Assume we have n samples $\mathbf{x}^{(i)} \sim p(\mathbf{x})$ available and that evaluating $p(x_k|\mathrm{pa}_k)$ is possible.
- We can use them to compute expectations with respect to $p(\mathbf{x}; do(x_k) \sim p')$ by computing a weighted average.
- Let $g(\mathbf{x})$ be an arbitrary function, then:

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x}; do(x_k) \sim p')}[g(\mathbf{x})] &= \int p(\mathbf{x}; do(x_k) \sim p') g(\mathbf{x}) \mathrm{d}\mathbf{x} && (8) \\
&= \int \frac{p(\mathbf{x})}{p(x_k|\mathrm{pa}_k)} p'(x_k) g(\mathbf{x}) \mathrm{d}\mathbf{x} && (9) \\
&= \int p(\mathbf{x}) \frac{p'(x_k)}{p(x_k|\mathrm{pa}_k)} g(\mathbf{x}) \mathrm{d}\mathbf{x} && (10) \\
&= \mathbb{E}_{p(\mathbf{x})}\left[\frac{p'(x_k)}{p(x_k|\mathrm{pa}_k)} g(\mathbf{x})\right] && (11)
\end{aligned}$$

which we approximate as a sample average.

Inverse probability weighting

- We have

$$\mathbb{E}_{p(\mathbf{x};do(x_k)\sim p')}[g(\mathbf{x})] = \mathbb{E}_{p(\mathbf{x})}\left[\frac{p'(x_k)}{p(x_k|\mathrm{pa}_k)}g(\mathbf{x})\right] \tag{12}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} w_i g(\mathbf{x}^{(i)}), \quad \mathbf{x}^{(i)} \sim p(\mathbf{x}) \tag{13}$$

  with $w^{(i)} = \frac{p'(x_k^{(i)})}{p(x_k^{(i)}|\mathrm{pa}_k^{(i)})}$
- The term $p(x_k|\mathrm{pa}_k)$ is called the propensity score.
- The effect of an intervention on x_k can be computed from observational data, i.e. the samples $\mathbf{x}_i \sim p(\mathbf{x})$.
- Practical use depends on n and the effective sample size (see lectures on sampling).

Adjustment for direct causes

$$p(\mathbf{x}; do(x_k) \sim p') = p(\tilde{\mathbf{x}}_k | x_k, \mathrm{pa}_k) p(\mathrm{pa}_k) p'(x_k)$$

- Assume we would like to compute $p(x_i; do(x_k) \sim p')$, $i \neq k$
- Marginalising over all variables but x_i, x_k, pa_k, we have

$$p(x_i, x_k, \mathrm{pa}_k; do(x_k) \sim p') = p(x_i | x_k, \mathrm{pa}_k) p(\mathrm{pa}_k) p'(x_k) \quad (14)$$

- Marginalising out the parent variables gives

$$p(x_i, x_k; do(x_k) \sim p') = \mathbb{E}_{p(\mathrm{pa}_k)} \left[p(x_i | x_k, \mathrm{pa}_k) \right] p'(x_k) \quad (15)$$

- Further marginalising out $x_k \sim p'(x_k)$ gives

$$p(x_i; do(x_k) \sim p') = \mathbb{E}_{p(\mathrm{pa}_k) p'(x_k)} \left[p(x_i | x_k, \mathrm{pa}_k) \right] \quad (16)$$

- For atomic interventions where $p'(x_k) = \delta(x_k - a)$ we obtain

$$p(x_i; do(x_k) = a) = \mathbb{E}_{p(\mathrm{pa}_k)} \left[p(x_i | x_k = a, \mathrm{pa}_k) \right] \quad (17)$$

Adjustment for direct causes

$$p(x_i; do(x_k) = a) = \mathbb{E}_{p(\mathrm{pa}_k)}\left[p(x_i | x_k = a, \mathrm{pa}_k)\right]$$

- When computing the causal effect of setting $x_k = a$ on x_i, we
  - compute $p(x_i | x_k = a, \mathrm{pa}_k)$ for each value of the parents pa_k
  - average with respect to their marginal distribution $p(\mathrm{pa}_k)$.
- This is called adjusting for the direct causes / the parents
- For discrete-valued pa_i, this corresponds to computing the effect $p(x_i | x_k = a, \mathrm{pa}_k)$ for each subpopulation/stratum separately, and then averaging them together, weighted by the probability of each subpopulation/stratum.
- In case of $p(x_i; do(x_k) \sim p')$, we vary x_k and average over $p'(x)$ too.

Connection to graph surgery

- When computing

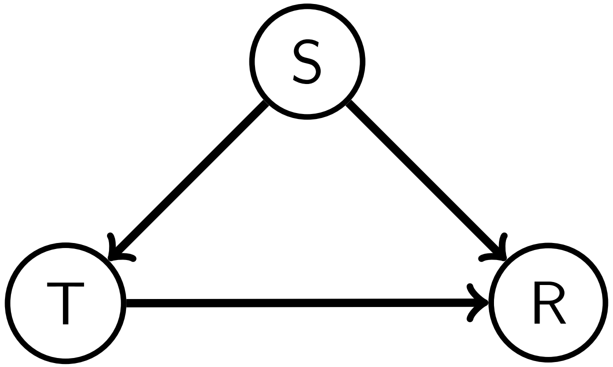
$$p(x_i; do(x_k) = a) = \mathbb{E}_{p(\mathrm{pa}_k)} \left[p(x_i | x_k = a, \mathrm{pa}_k) \right] \tag{18}$$

or, more generally,

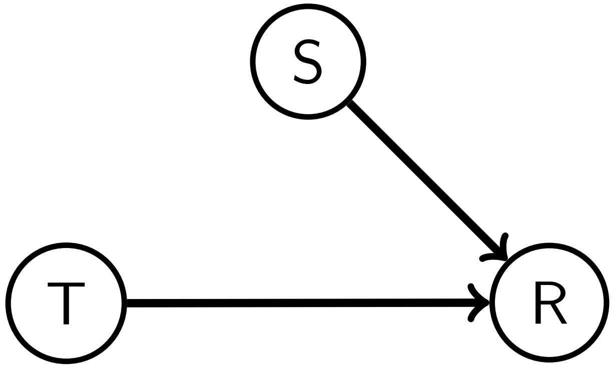
$$p(x_i; do(x_k) \sim p') = \mathbb{E}_{p(\mathrm{pa}_k)p'(x_k)} \left[p(x_i | x_k, \mathrm{pa}_k) \right] \tag{19}$$

the intervened-on variable x_k and its parents pa_k are root variables with distributions $p'(x_k)$ and $p(\mathrm{pa}_k)$.

- The arrow $\mathrm{pa}_k \to x_k$ is removed from the graph, in line with graph surgery.



G



$G_{\bar{T}}$

Kidney stone example

$$p(x_i; do(x_k) = a) = \mathbb{E}_{p(\mathrm{pa}_k)}\left[p(x_i | x_k = a, \mathrm{pa}_k)\right]$$

	Overall success rate	Small stones	Large stones
Treatment a	78% (273/350)	93% (81/87)	73% (192/263)
Treatment b	83% (289/350)	87% (234/270)	69% (55/80)

- Which treatment is more effective when the size of the kidney stones is unknown?
- We compute $p(R = 1 | do(T) = a)$ and $p(R = 1 | do(T) = b)$
- The parent variable of T is S,
  $p(S = \text{small}) = (87 + 270)/700 = 0.510$,
  $p(S = \text{large}) = (263 + 80)/700 = 0.490$
- $p(R = 1 | T = a, S = \text{small}) = 0.931$ and
  $p(R = 1 | T = a, S = \text{large}) = 0.730$, hence

$$p(R = 1 | do(T) = a) = 0.931 \cdot 0.510 + 0.730 \cdot 0.490 = 0.833$$

Kidney stone example

$$p(x_i; do(x_k) = a) = \mathbb{E}_{p(\mathrm{pa}_k)} [p(x_i | x_k = a, \mathrm{pa}_k)]$$

	Overall success rate	Small stones	Large stones
Treatment a	78% (273/350)	93% (81/87)	73% (192/263)
Treatment b	83% (289/350)	87% (234/270)	69% (55/80)

- $p(R = 1 | T = b, S = \text{small}) = 0.867$ and $p(R = 1 | T = b, S = \text{large}) = 0.688$, hence

$$p(R = 1 | do(T) = b) = 0.867 \cdot 0.510 + 0.688 \cdot 0.490 = 0.779$$

- We see that $p(R = 1 | do(T) = a) > p(R = 1 | do(T) = b)$. Treatment a is more effective.
- But when choosing a treatment, success rate may only be one criterion. Others may be recovery time, duration of the procedure, etc.

Difference between conditioning and intervening

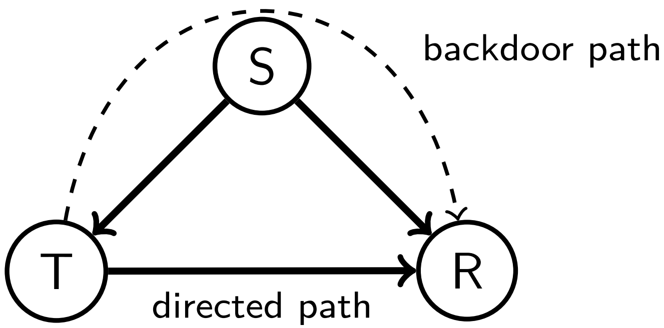
- In the example, we found that the postinterventional and conditional distributions are not the same

$$p(R = 1|do(T) = a) = 0.833 \neq p(R = 1|T = a) = 0.780$$
$$p(R = 1|do(T) = b) = 0.779 \neq p(R = 1|T = b) = 0.826$$

- What is the reason for this?
- Conditioning corresponds to a filtering process where we take all outcomes from the data generating process, keep those in line with the observed values (the conditioning set), and re-normalise.
- Interventions (actions) are different: we locally change the data generating process and depending where and how we intervene, the distribution of downstream variables changes.

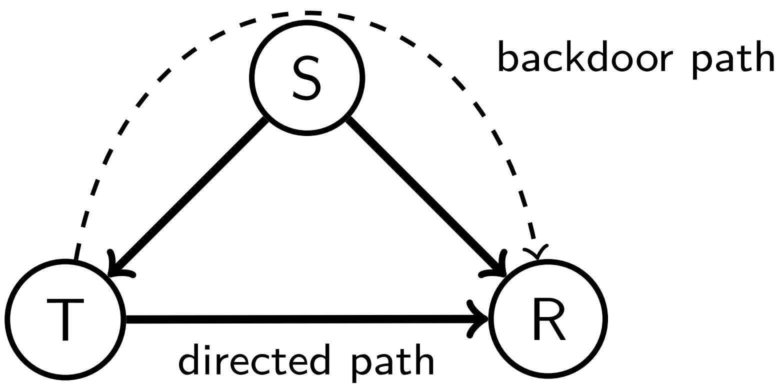
Directed and backdoor paths

- To better understand the difference between conditioning and intervening, consider how probability mass/information can flow between two nodes.
- Consider all the paths (trails) from a node x_k to x_i. We can distinguish between those that start with
  - arrows going out of x_k: directed (causal) paths
  - arrows going into x_k: backdoor (associative) paths
- For d-separation (independencies, conditioning), both types of paths matter; causal and associative effect are mixed.
- For interventions, only directed paths matter; backdoor paths are cut in the graph surgery



Directed and backdoor paths

- Unblocked/open backdoor paths lead to dependencies (associations) between two variables, but there is no causal connection.
- Such associations between variables without a causal origin are said to be "spurious".
- Non-descendants of a variable x_k cannot be changed by an intervention on x_k (as there is a topological ordering of the variables, for which they have been generated prior to x_k)
- Hence causal effects only travel along directed paths, not backdoor paths.



When is intervening and conditioning the same?

- It follows that the existence of open backdoor paths leads to a difference between conditional and postinterventional distributions.
- In other words, if the only active trails between x_k and x_i given $\mathbf{z}$ are directed paths, i.e. no open backdoor path exists, then $p(x_i|\mathbf{z}; do(x_k) = a) = p(x_i|\mathbf{z}, x_k = a)$.
- We can use d-separation in a modified graph to check whether all backdoor paths are closed:
  1. Remove all outgoing arrows from x_k, call the resulting graph $G_{\underline{x_k}}$ (this removes possible directed paths from the graph)
  2. Check whether $x_i \perp\!\!\!\perp x_k|\mathbf{z}$ in $G_{\underline{x_k}}$ (if so, all backdoor paths are closed)
- This leads to the following result on action/observation exchange (Pearl, Biometrika 82 (4), 1995, slightly simplified version)

If $x_i \perp\!\!\!\perp x_k|\mathbf{z}$ in $G_{\underline{x_k}}$ then $p(x_i|\mathbf{z}; do(x_k) = a) = p(y|\mathbf{z}, x_k = a)$

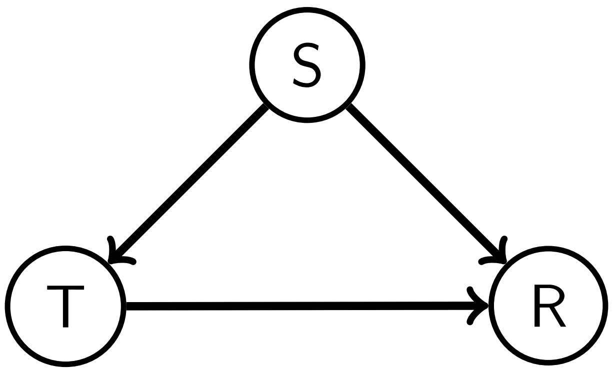
Kidney stone example

	Overall success rate	Small stones	Large stones
Treatment *a*	78% (273/350)	93% (81/87)	73% (192/263)
Treatment *b*	83% (289/350)	87% (234/270)	69% (55/80)

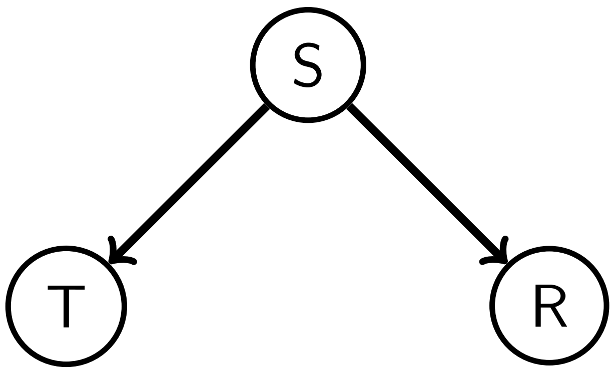
- Assume we now know the size of the stone S (e.g. through CT scans).
- Since $T \perp\!\!\!\perp R | S$ in $G_{\underline{T}}$, S blocks all backdoor paths
- Interventional and conditional distribution are the same:

$$p(R = 1 | S; do(T)) = p(R = 1 | S, T) \tag{20}$$

- Values can be read out directly from the table.



G



$G_{\underline{T}}$

Back-door adjustment

- By adjusting for the parents/direct causes, we can compute postinterventional distributions from the conditional $p(x_i | x_k, \mathrm{pa}_k)$. In case of atomic interventions, we had

$$p(x_i; do(x_k) = a) = \mathbb{E}_{p(\mathrm{pa}_k)} \left[p(x_i | x_k = a, \mathrm{pa}_k) \right] \tag{21}$$

- Expectation can be approximated as sampled average if we can observe the parents of the intervened-on variable x_k.
- We here derive a more general result that can be used when the parents are unobserved.
- We start with the sum-rule applied to $p(x_i; do(x_k) = a)$ (working with discrete variables for clarity)

$$\begin{aligned} p(x_i; do(x_k) = a) &= \sum_{\mathbf{z}} p(x_i, \mathbf{z}; do(x_k) = a) && (22) \\ &= \sum_{\mathbf{z}} p(x_i | \mathbf{z}; do(x_k) = a) p(\mathbf{z}; do(x_k) = a) \end{aligned}$$

Back-door adjustment

$$p(x_i; do(x_k) = a) = \sum_{\mathbf{z}} p(x_i|\mathbf{z}; do(x_k) = a) p(\mathbf{z}; do(x_k) = a)$$

- If (1) $\mathbf{z}$ blocks all backdoor paths from x_k to x_i, i.e. $x_i \perp\!\!\!\perp x_k | \mathbf{z}$ in $G_{\underline{x_k}}$. Then $p(x_i|\mathbf{z}; do(x_k) = a) = p(x_i|\mathbf{z}, x_k = a)$ and

$$p(x_i; do(x_k) = a) = \sum_{\mathbf{z}} p(x_i|\mathbf{z}, x_k = a) p(\mathbf{z}; do(x_k) = a) \quad (23)$$

- If (2) no component of $\mathbf{z}$ is a descendant of x_k. Then $p(\mathbf{z}; do(x_k) = a) = p(\mathbf{z})$ (non-descendants are not affected by actions on x_k) and

$$p(x_i; do(x_k) = a) = \sum_{\mathbf{z}} p(x_i|\mathbf{z}, x_k = a) p(\mathbf{z}) \quad (24)$$

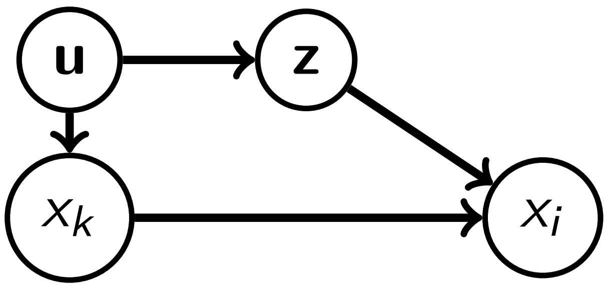
$$= \mathbb{E}_{p(\mathbf{z})} \left[p(x_i|\mathbf{z}, x_k = a) \right] \quad (25)$$

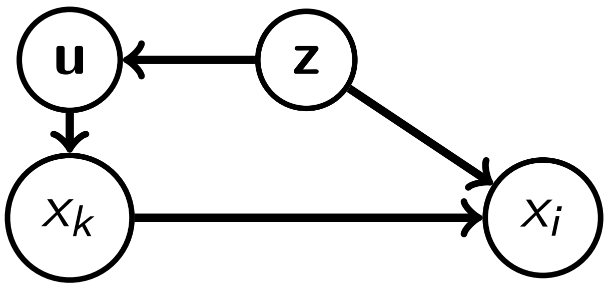
- This is called the back-door adjustment to compute the causal effect of $do(x_k) = a$ on x_i.
- $\mathbf{z} = \mathrm{pa}_k$ gives the adjustment formula for direct causes.

Back-door adjustment

- Example configurations where $\mathbf{z}$ satisfies the two conditions are shown below.
- The parents $\mathbf{u}$ of x_k are assumed unobserved.
- Observing $\mathbf{z}$ is sufficient to compute $p(x_i; do(x_k) = a)$ from $p(x_i|\mathbf{z}, x_k = a)$ via

$$p(x_i; do(x_k) = a) = \mathbb{E}_{p(\mathbf{z})}\left[p(x_i|\mathbf{z}, x_k = a)\right] \qquad (26)$$





Program recap

1. Modelling actions as interventions in causal DAGs
   - Causal DAGs
   - Interventions change the data generating process
   - Interventions change the DAG locally

2. Computing the effect of interventions
   - Inverse probability weighting and adjustment for direct causes
   - Observing vs acting: the role of backdoors
   - Backdoor adjustment