par

MATH 7710
Mathematics for AI Safety

Lecture Notes

Math for AI Safety — Lionel Levine

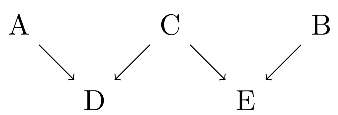
G-Markov Distributions

September 11, 2024 — *Notes by Arkar Oak Soe*

1 Applications and examples of the d-separation theorem

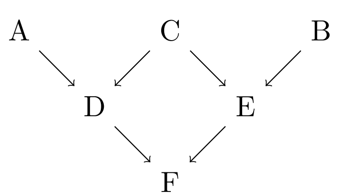
We begin by verifying some conditional independence statements about a given G-Markov distribution.

Example 1



1. $A \perp\!\!\!\perp B \,|\, \varnothing$ as there is only one path from A to B and it is blocked by the collider at D. (It's also blocked by the collider at E, but one blocking vertex is enough to block a path!)
2. $A \perp\!\!\!\perp B \,|\, D$ as the path is blocked by the collider at E
3. $A \not\perp\!\!\!\perp B \,|\, (D, E)$ as the path is unblocked.
4. $A \perp\!\!\!\perp B|(C, D, E)$ as the path is blocked by C.

Example 2



1. $A \not\perp\!\!\!\perp B \,|\, F$ as both paths from A to B are unblocked. Note that D is a collider on the upper path but not on the lower path. D does not block the upper path because it has a descendant in the conditioning set, namely F.
2. $A \not\perp\!\!\!\perp B \,|\, (D, F)$ because the upper path from A to B is unblocked.

2 How G-Markov distributions get their name

G-Markov distributions can be considered an extension of Markov chains. To see this, we'll prove a theorem that characterizes G-Markov distributions by their conditional independence relations.

Let $G = (V, E)$ be a directed acyclic graph with vertex set $V = \{X_1, X_2, \ldots X_n\}$. We say that G is *properly labeled* if $(X_i, X_j) \in E$ implies $i < j$. Write $\overline{\text{des}}(X_i) = \text{des}(X_i) \cup \{X_i\}$ for the set of descendants of X_i including itself.

Given a tuple of outcomes $(x_1, \ldots, x_n)$ we'll use the notation $\mathbf{pa}_i = (x_k)_{X_k \in \text{par}(X_i)}$ for the sub-tuple of outcomes of the parents of X_i.

Theorem 1 *The distribution $(X_1, X_2, \ldots, X_n)$ is G-Markov if and only if*

$$X_i \perp\!\!\!\perp X_j \,|\, \text{par}(X_i) \qquad \forall i, j \text{ such that } X_j \notin \overline{\text{des}}(X_i). \tag{1}$$

Proof: By re-indexing the random variables we may assume G is properly labeled. If (1) holds, then

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \,|\, x_1, \ldots, x_{i-1}) = \prod_{i=1}^{n} p(x_i \,|\mathbf{pa}_i)$$

so the distribution is G-Markov. In the second equality we've used (1) to drop the conditioning on all X_j that are not parents of X_i. We were able to do this because G is properly labeled, so the set $\{X_1, \ldots, X_{i-1}\}$ contains $\text{par}(X_i)$ and is disjoint from $\text{des}(X_i)$.

For the converse, let $\gamma = (X_i, Y, \ldots, X_j)$ be any path in G from X_i to X_j. There are two cases depending whether the first vertex is a parent or child of X_i.

Case 1: If $Y \in \text{par}(X_i)$ then γ is blocked by Y.

Case 2: If Y is a child of X_i, then since $X_j \notin \text{des}(X_i)$, the path must have a collider. The *first* collider, call it Z, is a descendant of X_i. So $\text{des}(Z) \subset \text{des}(X_i)$ is disjoint from $\text{par}(X_i)$ (by acyclicity). So γ is blocked by Z.

Since all paths are blocked, (1) follows from the d-separation theorem. □

To see how discrete-time Markov chains are a special case, consider the graph

$$G : X_1 \to X_2 \to \cdots \to X_n.$$

From the theorem above, we obtain the Markov property:

$$X_{t+1} \perp\!\!\!\perp X_s \,|\, X_t \qquad \text{for all } 1 \le s < t < n.$$

Equivalently,

$$p(x_{t+1}|x_1,\dots,x_t) = p(x_{t+1}|x_t) \qquad \text{for all } 1 \le t < n.$$

Thinking of t as a discrete time parameter, the Markov property says "the future is conditionally independent of the past, given the present." One way to think about (1) is that G-Markov distributions have a "time" index indexed by a DAG instead of by the natural numbers: In this interpretation, X_i is the "future" and $\text{par}(X_i)$ is the "present". Non-parent ancestors of X_i are the "past". According to (1), not only is X_i conditionally independent of this "past", it is conditionally independent of all its non-descendants.

3 Parameter counts

Consider the joint probability distribution of n random variables $X_1,\dots,X_n$, where each X_i takes values in $\{0,1\}$.

- There are 2^n parameters to specify for the joint distribution

  $$p(x_1,\dots,x_n) = P(X_1 = x_1,\dots,X_n = x_n)$$

  corresponding to each possible n-tuple of outcomes $(x_1,\dots,x_n) \in \{0,1\}^n$.
- However, since probabilities sum to 1, there are only $2^n - 1$ free parameters.

We can also write the joint distribution as a product of conditional distributions:

$$p(x_1,\dots,x_n) = \prod_{i=1}^{n} p(x_i|x_1,\dots,x_{i-1}).$$

The number of free parameters in the conditional distribution $p(x_i|x_1,\dots,x_{i-1})$ is 2^{i-1}, since for each i and $x_1,\dots,x_{i-1}$ we have

$$p(0|x_1,\dots,x_{i-1}) + p(1|x_1,\dots,x_{i-1}) = 1.$$

Thus, the total number of free parameters in the joint distribution is:

$$\sum_{i=1}^{n} 2^{i-1}$$

which equals $2^n - 1$, consistent with our earlier count.

For a G-Markov distribution, the joint distribution $p(x_1,\dots,x_n)$ can be factored according to the conditional independence structure:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | \mathbf{pa}_i),$$

where $\mathbf{pa}_i$ represents a tuple of outcomes for the parents of X_i in G. The number of free parameters for each conditional distribution is determined by the number of parent variables for each X_i. The total number of free parameters is

$$\sum_{i=1}^{n} 2^{|\mathrm{Pa}(X_i)|},$$

where $|\mathrm{Pa}(X_i)|$ is the number of parents of X_i.

Consider a G-Markov distribution where each X_i has at most k parents, i.e., $|\mathrm{Pa}(X_i)| \leq k \, \forall i$. In this case, the number of free parameters is

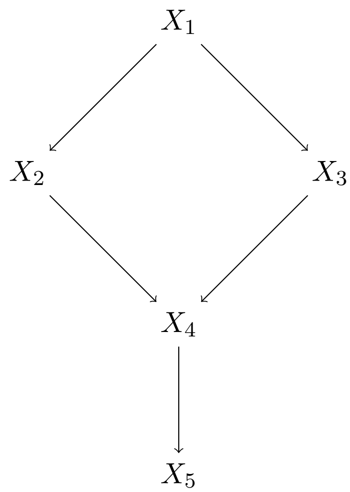
$$\sum_{i=1}^{n} 2^{|\mathrm{Pa}(X_i)|} \leq 2^k n.$$

which is often *much* smaller than the $2^n - 1$ free parameters required for an unrestricted joint distribution. For example if $n = 1000$ and $k = 10$, then $2^k n$ is around a million (tractable!) whereas $2^n - 1$ is more than the number of atoms in the universe ($\approx 2^{270}$).

Most joint distributions on 1000 variables are completely intractable: they have high Kolmogorov complexity. You'd need a universe much bigger than ours just to write down a complete description of $p(x_1, \ldots, x_{1000})$! But the distributions we care about predicting are the ones actually arising in our universe, and *those* distributions have a lot of structure. The G-Markov condition is a flexible way of imposing structure on a joint distribution to make it tractable.

4 Functional causal models

Next time: In addition to the distribution, we can also model which variables X_i are functions of which other variables. For example,



$$
\begin{aligned}
X_1 &= f(U_1) && \text{(Season)} \\
X_2 &= f(X_1, U_2) && \text{(Rain)} \\
X_3 &= f(X_1, U_3) && \text{(Sprinkler is on or off)} \\
X_4 &= f(X_2, X_3, U_4) && \text{(Pavement is wet)} \\
X_5 &= f(X_4, U_5) && \text{(Pavement is slippery)}
\end{aligned}
$$

Where:

- $U_1, U_2, \ldots, U_n$ are background variables or disturbances that are jointly independent.
- $f_1, f_2, \ldots, f_n$ are deterministic functions that describe the dependencies between the X_i.