IFT 6269: Probabilistic Graphical Models **Fall 2018**

Lecture 11 — October 12

Lecturer: Simon Lacoste-Julien *Scribe: Martin Weiss, Eeshan Gunesh Dhekane*

Disclaimer: These notes have only been lightly proofread.

11.1 Graph Theory review

11.1.1 Directed Graph

Definition 11.1 *A **Directed Graph** G consists of a set of **Nodes** or **Vertices** $V = \{1, \ldots, n\}$ and a set of **Edges** E such that $E \subseteq V \times V$, i.e, E is a set of ordered pairs of distinct vertices : $E = \{(i, j) \mid i, j \in V, i \neq j\}$.*

We will only consider graphs that do not have a self-loops, i.e., $(i, i) \notin E \; \forall \; i \in V$.

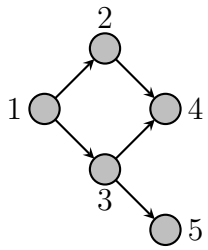

Figure 11.1: Directed graph G with $V = \{1, \ldots, 5\}$ and $E = \{(1, 2), (2, 4), (1, 3), (3, 4), (3, 5)\}$

Definition 11.2 *A **Directed Path** from vertex i to vertex j of directed graph G consists of an ordered sequence of vertices $(i, v_1, \ldots, v_k, j)$, where $k \geq 0$, such that $(i, v_1), (v_1, v_2), \ldots, (v_{k-1}, v_k)$, and $(v_k, j) \in E$. We denote this directed path from i to j by a squiggly arrow $i \rightsquigarrow j$.*

Equivalently, a directed path can also be viewed as sequence of edges mentioned above. The same path can be represented as ordered sequence of edges : $((i, v_1), (v_1, v_2), \ldots, (v_{k-1}, v_k), (v_k, j))$. The example given below shows a directed path P from 1 to 4 (Figure [11.2]).

Definition 11.3 *The set of **Parents** of a vertex i, denoted by π_i, is the set of vertices of G from which there is an edge to i, i.e., $\pi_i = \{j \mid j \in V, (j, i) \in E\}$. Analogously, the set of **Children** of a vertex k, denoted by $ch(k)$, is the set of vertices of G to which there is an edge from k, i.e., $ch(k) = \{\ell \mid \ell \in V, (k, \ell) \in E\}$.*

Figure [11.3] below shows the parent of 2, which is 1 and the children of 3, which are $4, 5$.

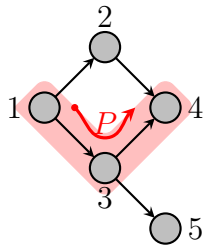

Figure 11.2: A directed path P from 1 to 4 with vertices $1, 3, 4$ and edges $(1, 3), (3, 4)$.

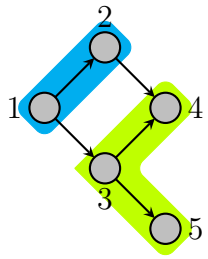

Figure 11.3: 1 is the parent of 2 and $4, 5$ are children of 3.

11.1.2 Undirected Graph

Definition 11.4 *An **Undirected Graph** G consists of a set of **Nodes** or **Vertices** $V = \{1, \ldots, n\}$ and a set of **Edges** E such that E is set of $2-$sets of V without any self-loops, i.e., $E = \{\{i, j\} \mid i, j \in V, i \neq j\}$.*

Thus, the edge $\{i, j\}$ is identical to the edge $\{j, i\}$. Since there are no self-loops, for any edge $e = \{i, j\} \in E$, we have $|e| = 2$. The Figure [11.4] shows an example of an undirected graph.

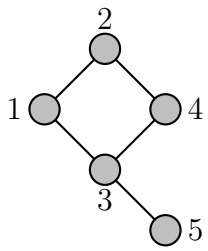

Figure 11.4: Undirected graph G, $V = \{1, \ldots, 5\}$, $E = \{\{1, 2\}, \{2, 4\}, \{1, 3\}, \{3, 4\}, \{3, 5\}\}$

Definition 11.5 *An **undirected Path** from vertex i to vertex j of directed path G consists of an ordered sequence of vertices $(i, v_1, \ldots, v_k, j)$, where $k \geq 0$, such that $\{i, v_1\}, \{v_1, v_2\}, \ldots, \{v_{k-1}, v_k\}$, and $\{v_k, j\} \in E$.*

Equivalently, an undirected path can also be viewed as sequence of edges mentioned above. The example given below shows a directed path P from 2 to 3 (Figure [11.5]).

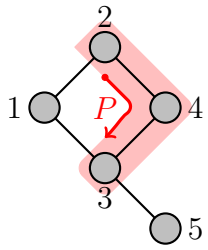

Figure 11.5: Undirected path P from 2 to 3.

Definition 11.6 *The set of* ***Neighbors*** *of a vertex i, denoted by $N(i)$, is the set of vertices that are connected with i through an edge, i.e., $N(i) = \{j \mid \{i, j\} \in E\}$.*

For an undirected graph, the neighbors replace the notions of sets of parent and children. Figure [11.6] shows the neighbors of vertex 4.

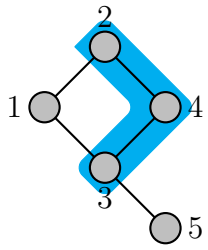

Figure 11.6: Vertex 4 with its neighbors $2, 3$.

11.1.3 Directed Acyclic Graph

Definition 11.7 *A* ***Cycle*** *in a (directed/undirected) graph G consists of an ordered sequence of nodes $v_1, \ldots, v_k, v_1$ such that $v_1 \neq v_k$, there exists an (directed/undirected) edge from v_i to v_{i+1} $\forall i \in \{1, \ldots, k-1\}$, there exists an (directed/undirected) edge from v_k to v_1 and $v_i \neq v_j$ for $i \neq j$.*

Equivalently, there exists a (directed/undirected) path in G from v to v for some vertex v. In the examples of directed and undirected graphs above, there is no cycle in the directed graph. However, there is a cycle in the undirected graph (namely, $1 - 3 - 4 - 2-$).

Definition 11.8 *A directed graph with no cycles is called a* ***Directed Acyclic Graph****.*

Note that the directed graph considered in [11.1] is indeed a directed acyclic graph (DAG).

Definition 11.9 *An ordering $I : V \to \{1, \ldots, n\}$ on the vertex set $V = \{1, \ldots, n\}$ of a directed graph G is said to be* ***Topological for*** *G if and only if: 1) I is bijective and 2) $a \in \pi_b$ implies that $I(a) < I(b)$.*

What this deinition implies is that if we order (in the increasing manner) the vertices based on the topological ordering, we will always have the parent of any node appearing before the node itself (and all the directed arrows would "point to the right", leaving no "back edges"). Observe that for the DAG from Figure [11.1], the ordering of the vertices is already a topological ordering, which is displayed in Figure [11.7].

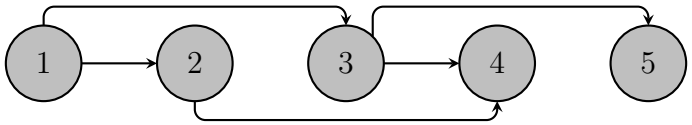

Figure 11.7: Example of Topological Ordering on DAG from Figure [11.1].

Theorem 11.10 (Characterization of DAGs using Topological Ordering) *A directed graph G is a DAG $\iff$ G has a topological ordering.*

Proof
($\Rightarrow$) If G is given to be a DAG, perform Depth-First Search algorithm on it. Number in descending order the nodes for which we run out of children while performing the DFS. Because there is no cycle, you will always find nodes with no children during this algorithm and thus this generate a topological ordering (in $\mathcal{O}(|V| + |E|)$ time).

($\Leftarrow$) (trivial) If there is a topological ordering of G, then G can not have any back edges and hence, it can not have any cycles. Thus, G is a DAG. ■

11.2 Notation for Graphical Models

- Given n random variables $X_1, \ldots, X_n$. We assume that X_i are discrete random variables for simplicity for this part of the class. This is because defining conditional distribution on continuous random variables is challenging. (Please refer to [**Borel-Kolmogorov Paradox**] to see the challenges in defining conditional distributions.)
- Given a graph $G = (V, E)$ such that $V = \{1, \ldots, n\}$. We associate one random variable per node of G and letting random variable X_i associate with node i.
- For any subset $A \subseteq V$ of vertices, $p(X_A)$ is defined as : $p(x_A) = P\{X_i = x_i \mid i \in A\}$. It is easy to see that $p(x_A) = \sum_{x_{A^C}} p(x_A, x_{A^C})$, where $\sum_{x_{A^C}}$ denotes summing over all possible values of $\{x_i\}_{i \in V \setminus A}$. For instance, $x_{1,2,4}$ represents $\{x_1, x_2, x_4\}$.
- The joint probability is given by : $p(X_1 = x_1, \ldots, X_n = x_n) = p(x_1, \ldots, x_n) = p(x_V)$.

11.3 About Graphical Models

A **Graphical Model** is essentially a graph that models the dependencies between a set of random variables. Graphical models lie at the intersection of probability theory and computer science, in that they use graphs to model distributions over random variables. Graphs are highly efficient data structures for storing information related to dependencies and thus, they are extremely useful in the case of modeling distributions. For instance, consider 100 random variables $\{X_i\}_{1\leq i\leq 100} \in \{0,1\}$. Then, in order to represent the distribution in table format, we would require 2^{100} variables, which is intractable to represent explicitly in a computer. In contrast, we can use graphical models (with certain assumptions) to keep the problem tractable.

11.4 Conditional Independence Revisited

Let $A, B, C \subseteq V$ be three subsets of vertices.

- We say that $X_A \perp\!\!\!\perp X_B \mid X_C \iff p(x_A, x_B \mid x_C) = p(x_A \mid x_C)p(x_B \mid x_C)$ $\forall x_A, x_B, x_C,$ s.t. $p(x_C) > 0$. This is the Factorization forumulation (F).
- An equivalent Conditional formulation (C) states that $X_A \perp\!\!\!\perp X_B \mid X_C$ $\iff p(x_A \mid x_B, x_C) = p(x_A \mid x_C)$ $\forall x_A, x_B, x_C$ s.t. $p(x_B, x_C) > 0$.
- We can state the "marginal independence" of X_A, X_B as $X_A \perp\!\!\!\perp X_B \mid \phi$.

11.5 Two Facts About Conditional Independence

1. **Can repeat variables:** you are allowed to repeat variables in a conditional statement (for convenience). For example, $X \perp\!\!\!\perp Y, Z | Z, W$ is fine to say. It is actually equivalent to $X \perp\!\!\!\perp Y | Z, W$ (the second Z on the left does not do anything). This will be useful when writing generic theorems about conditional statements from a graphical model (to avoid excluding the repition cases).
2. **Decomposition:** $X \perp\!\!\!\perp Y, Z | W$ implies both $X \perp\!\!\!\perp Y | W$ and $X \perp\!\!\!\perp Z | W$ (it decomposes in two conditional independence statements).

11.6 Directed Graphical Models

Definition 11.11 *Let $G = (V, E)$ be a DAG with $V = \{1, \ldots, n\}$. A directed graphical model (DGM) (associated with G), also known as a Bayesian network, is a **family** of distributions*

over X_V defined as follows:

$$\mathcal{L}(G) \triangleq \{p \text{ is a distribution over } X_V : \exists \textbf{\textit{legal}} \text{ factors } f_i\text{'s} \tag{11.1}$$

$$\text{s.t. } p(x_V) = \prod_{i=1}^{n} f_i(x_i|x_{\pi_i}) \, \forall x_V\} \tag{11.2}$$

In the definition above, the legal factors are functions $f_i : \Omega_{X_i} \times \Omega_{X_{\pi_i}} \to [0,1]$ s.t. $\sum_{x_i} f(x_i, x_{\pi_i}) = 1 \; \forall x_{\pi_i}$ (and thus f_i is like a conditional probability table (CPT) – it could be used to define a conditional distribution on X_i given the values of its parents X_{π_i}).

Two notes: recall that π_i are the parents are node i. In the definition above, the factors do not have to be unique (i.e. we do not rule out the possibility that the same distribution could have two expansions with different factors). But it turns out that we can actually prove that the factors are unique (as we will see when we show that $p(x_i|x_{\pi_i}) = f(x_i, x_{\pi_i})$ below, and thus the factors are uniquely specified by the distribution).

Terminology: if we can write $p(x_V) = \prod_{i=1}^n f_i(x_i|x_{\pi_i})$ where f_i's are legal factors and π_i's are determined from a DAG G, then we say that p factorizes according to G, and we denote this by $p \in \mathcal{L}(G)$ (i.e. p is also a member of the DGM for G). We will also sometimes write $p(x_V) \in \mathcal{L}(G)$ if we want to make which variables are considered for the distribution explicit (see notation in the proofs below).

To give one example, see the three nodes graph from Figure 11.10. Then $p \in \mathcal{L}(G)$ for this graph if and only if there exists some legal factors f_x, f_y and f_z s.t. $p(x,y,z) = f_x(x) f_y(y) f_z(z|x,y)$.

11.7 Leaf-Plucking Property

We first show a fundamental property of DGM which is used in a lot of proofs:

Proposition 11.12 ("Leaf-plucking" property) *Let n be a **leaf** in the DAG G (i.e. n is not the parent of anything) and suppose $p(x_V) \in \mathcal{L}(G)$.*
a) then $p(x_{1:(n-1)}) \in \mathcal{L}(G - \{n\})$
b) if $p(x_{1:n}) = \prod_{i=1}^n f_i(x_i|x_{\pi_i})$, then $p(x_{1:n}) = \prod_{i=1}^{n-1} f_i(x_i|x_{\pi_i})$.

Proof

$$p(x_n, x_{1:(n-1)}) = f_n(x_n|x_{\pi_n}) \prod_{j \neq n} f_j(x_j| \underbrace{x_{\pi_j}}_{\text{no } n \text{ in any } \pi_j})$$

$$p(x_{1:(n-1)}) = \sum_{x_n} p(x_n, x_{1:(n-1)}) = (\underbrace{\sum_{x_n} f_n(x_n|x_{\pi_n})}_{\text{1 by definition}}) \left(\prod_{j \neq n} \underbrace{f_j(x_j|x_{\pi_j})}_{\text{no } x_n \text{ there}} \right)$$

■

We now use this property to show the important fact that the factors are the same as conditional probabilities defined from the joint in a DGM G (and thus the factors are the correct conditionals).

Proposition 11.13 *If $p(x) \in \mathcal{L}(G)$ then, for all $i \in \{1, \ldots, n\}$, $p(x_i|x_{\pi_i}) = f_i(x_i, x_{\pi_i})$.*

Proof We prove this by induction on $n = |V|$, the cardinality of the set V. Since G is a DAG, there exists a leaf, i.e. a node with no children. Without loss of generality, we can assume that the leaf is labeled by n (if not, then just relabel the nodes so that it is true). We first notice:

$$\begin{aligned}
\forall x, p(x_1, \ldots, x_{n-1}) &= \sum_{x_n} p(x_1, \ldots, x_n) \\
&= \sum_{x_n} \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}) \\
&= \sum_{x_n} f_n(x_n, x_{\pi_n}) \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \\
&= \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \sum_{x_n} f_n(x_n, x_{\pi_n}) \ (*) \\
&= \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) \\
&= g(x_1, \ldots, x_{n-1}) \ (**)
\end{aligned} \tag{11.3}$$

The step $(*)$ is justified by the fact that n is a leaf and thus it never appears in any of the π_i for $i \in \{1, \ldots, n-1\}$. Step $(**)$ is also justified by the same kind of reasoning: since n is a leaf it cannot appear in any of the π_i explaining why it is only a function, say g, of $x_1, \ldots, x_{n-1}$. From this result, we can use an induction reasoning noticing that $G - \{n\}$ is still a DAG. To conclude this proof, we simply need to show that, indeed, $f_n(x_n, x_{\pi_n}) = p(x_n|x_{\pi_n})$—this property will automatically propagates by induction. We have:

$$p(x_n, x_{\pi_n}) = \sum_{x_i, i \notin \{n\} \cup \pi_n} p(x) = \left(\sum_{x_i, i \notin \{n\} \cup \pi_n} g(x_1, \ldots, g_{n-1}) \right) f_n(x_n, x_{\pi_n}). \tag{11.4}$$

Noticing that $\sum_{x_i, i \notin \{n\} \cup \pi_n} g(x_1, \ldots, x_{n-1})$ is a function of only x_{π_n}, say $h(x_{\pi_n})$, we can derive:

$$p(x_n|x_{\pi_n}) = \frac{p(x_n, x_{\pi_n})}{\sum_{x'_n} p(x'_n, x_{\pi_n})} = \frac{h(x_{\pi_n}) f_n(x_n, x_{\pi_n})}{h(x_{\pi_n})} = f_n(x_n, x_{\pi_n}). \tag{11.5}$$

■

Hence we can give an equivalent definition for a DAG to the notion of factorization:

Definition 11.14 (Equivalent definition of a DGM) *A DGM on G is the set of distributions $p(x)$ that factorizes according to G, denoted $p(x) \in \mathcal{L}(G)$ iff:*

$$\forall x, p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}). \tag{11.6}$$

Why didn't we start with the above definition for a DGM? The reason is that without the proof above, we would not know whether our definition makes sense, as this definition is **circular**. Indeed, the conditional $p(x_i|x_{\pi_i})$ is defined **from** the joint $p(x)$. So we are not allowed (normally) to define a joint by multiplying its conditionals (as you might get no distribution that satisfies this property).

Remark 11.7.1 *Adding edges $\implies$ more distributions i.e. G=(V, E) and G'=(V, E') with E subset of E then L(G) subset L(G')*

11.8 DGM Examples

11.8.1 Trivial Graphs

Example 11.8.1

- *(Trivial graph with empty edge set) Assume $E = \emptyset$, i.e. there is no edges. Then the DGM on this graph contains only fully independent distributions (i.e. $p(x) = \prod_{i=1}^{n} p(x_i)$). (this is the "smallest" DGM).*
- *(Complete digraph) Assume now we have a complete graph (thus with $n(n-1)/2$ edges as we need acyclic for it to be a DAG), we have: $p(x) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1})$, the so-called 'chain rule' which is always true. Thus all distributions on x_V belongs to the DGM on the complete graph (this is the "biggest" DGM).*

11.8.2 Graphs with three nodes

We give an insight of the different possible behaviors of a graph by thoroughly enumerating the possibilities for a 3-node graph.

- The two first options are the empty graph, leading to independence, and the complete graph that gives no further information than the chain rule.
- (Markov chain) A Markov chain is a certain type of DAG showed in Fig. 11.8. In this configuration we show that we have:

$$p(x, y, z) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Z \mid Y \tag{11.7}$$

  I.e. we have that the "future" Y is conditionally independent on the "past" X given the "present" Z (assuming the arrow would represent time). On the other, there are

some distributions $p \in \mathcal{G}(G)$ for which X is **not** marginally independent of Y (the "dependence" flows through Z).

To show the conditional independence statement, we have:

$$p(z|y,x) = \frac{p(x,y,z)}{p(x,y)} = \frac{p(x,y,z)}{\sum_{z'} p(z',x,y)} = \frac{p(x)p(y|x)p(z|y)}{\sum_{z'} p(x)p(y|x)p(z'|y)} = p(z|y)$$

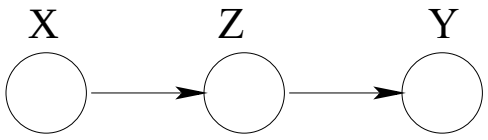

Figure 11.8: Markov Chain

- (Latent cause) It is the type of DAG given in Fig. 11.9. We show that:

$$p(x) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \mid Z \tag{11.8}$$

Indeed:

$$p(x,y|z)\frac{p(x,y,z)}{p(z)} = \frac{p(z)p(y|z)p(x|z)}{p(z)} = p(x|z)p(y|z)$$

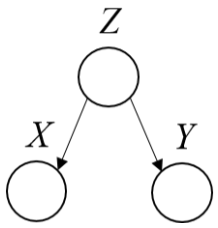

Figure 11.9: Latent cause

- (Explaining away) Represented in Fig.(11.10), we can show for this type of graph:

$$p(x) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \tag{11.9}$$

It basically stems from:

$$p(x,y) = \sum_z p(x,y,z) = p(x)p(y)\sum_z p(z) = p(x)p(y)$$

On the other hand, in general we do not have that X is conditionally independent on Y given Z (unlike for both the latent cause model and the Markov chain DGM). So here X is marginally independent on Y, but observing Z induces some dependence between

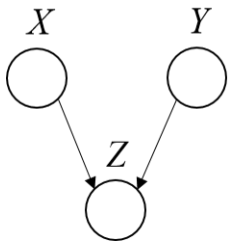

Figure 11.10: Explaining away, or V-structure

X and Y. From this graphical model, we can get the so-called non-motononic property of conditioning. For example, let X be "I'm abducted by alien", Y be "my watch is broken", and Z be "I am late". The v-structure explains this situation as there are competing explanation for why "I am late": I might have been abducted by aliens, or my watch could be broken and I did not notice the time... In this example, a meaningful distribution could yield that $p(\text{alien})$ is tiny; but then $p(\text{alien}|\text{late}) > p(\text{alien})$ (because knowing that I'm late give some evidence that perhaps I have been adbucted by alien). But $p(\text{alien}|\text{late}, \text{broken watch}) < p(\text{alien}|\text{late})$ (because now that I know that my watch is broken, it gets unlikely again that I have been abducted by alien, as it's more likely that I'm late because of the watch). Thus conditioning on more things can increase or decrease the probability of an event (hence the word "non-monotone").

Remark 11.8.1 *The use of 'cause' is not advised since observational statistics provide with correlations and no causality notion. Note also that in the 'explaining away' graph, in general $X \perp\!\!\!\perp Y|Z$ is not true. Lastly, it is important to remember that not every relationship can be expressed in terms of graphical models. As a counter-example take the XOR function where $Z = X \oplus Y$. The three random variables are pairwise independent, but not mutually independent.*

11.9 Conditional Independence Statements in DGMs

Definition 11.15 *Let $nd(i) \triangleq \{j : \text{no path from } i \text{ to } j\}$. Then j is said to be a non-descendent of i.*

Proposition 11.16 *If G is a DAG, then:*

$$p(x) \in \mathcal{L}(G) \Leftrightarrow X_i \perp\!\!\!\perp X_{nd(i)\setminus\pi_i} | X_{\pi_i} \tag{11.10}$$

Proof We will only prove the forward implication. Assume $(1, \ldots, n)$ is a topological order then:

$$\begin{cases} p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}) & : \text{because } p(x) \in \mathcal{L}(G) \\ p(x) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}) & : \text{chain rule, always true} \end{cases}$$

As we chose a topological order, we have $\pi_i \subset \{1, \dots, i-1\}$, and we show by induction that:

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

This directly implies that $X_i \perp\!\!\!\perp X_{\{1,\dots,i-1\}\setminus\pi_i}|X_{\pi_i}$. The key idea now is to notice that for all i, there exist a topological order such that $\text{nd}(i) = \{1, \dots, i-1\}$. ■

11.10 D-separation

We want to answer queries such as, given A, B and C three subsets, is $X_A \perp\!\!\!\perp X_B|X_C$ true? To answer those issues we need the d-separation notion, or directed separation. Indeed it is easy to see that the notion of separation is not enough in a directed graph and needs to be generalized.

Definition 11.17 *Let $a, b \in V$, a chain from a to b is a sequence of nodes, say $(v_1, \dots, v_n)$ such that $v_1 = a$ and $v_n = b$ and $\forall j, (v_j, v_{j+1}) \in E$ or $(v_j, v_{j+1}) \in E$.*

We can notice that a chain is hence a path in the symmetrized graph, i.e. in the graph where if the relation $\to$ is true then $\leftrightarrow$ is true as well. Assume C is a set that is observed. We want to define a notion of being 'blocked' by this set C in order to answer the underlying question above.

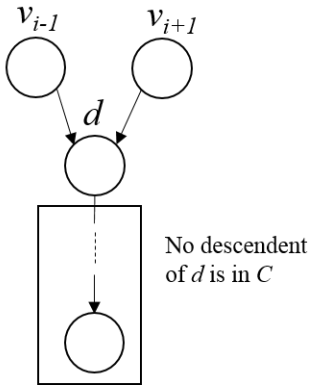

Figure 11.11: D-separation

Definition 11.18 (d-separation)

1. *A chain from a to b is blocked at node d "given C" if:*
   - *either $d \in C$ and (v_{i-1}, d, v_{i+1}) is not a V-structure;*

- *or $d \notin C$ and (v_{i-1}, d, v_{i+1}) is a V-structure and no descendants of d is in C.*

2. *A chain from a to b is considered blocked if it is blocked at some of the node d along it.*
3. *A and B are said to be d-separated by C if and only if all chains that go from $a \in A$ to $b \in B$ are blocked by the rules above.*

Example 11.10.1

- *(Markov chain) If you try to prove that any set of the future is independent to the past given the present with Markov theory, it might be difficult but the d-separation notion gives the results directly.*

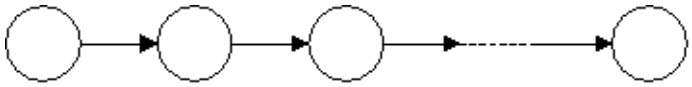

Figure 11.12: Markov chain

- *(Hidden Markov Model) Often used because we only observe a noisy observation of the random process.*

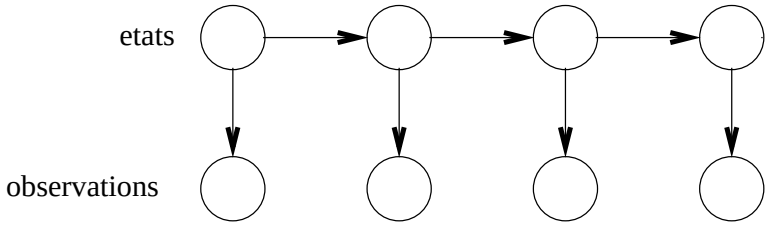

Figure 11.13: Hidden Markov Model

Proposition 11.19 (All conditional independence statements in a DGM) *$p \in \mathcal{L}(G)$ iff $X_A \perp\!\!\!\perp X_B | X_C$ $\forall A, B, C$ such that A and B are d-separated by C in G.*

11.11 "Bayes-Ball" Algorithm

This is an intuitive "reacheability" algorithm to determine all the conditional independence statements in a DAG (via d-seperation). Suppose we want to determine if X is conditionally independent from Z given Y. Place a ball on each of the nodes in X and let them bounce around according to some rules (described below) and see if any reaches Z. $X \perp\!\!\!\perp Z|Y$ is true if none reached Z, but not otherwise (the balls implement the path rules from d-separation, and are blocked accordingly).

The rules are as follows for the three canonical graph structures. Note that the balls are allowed to travel in either direction along the edges of the graph.

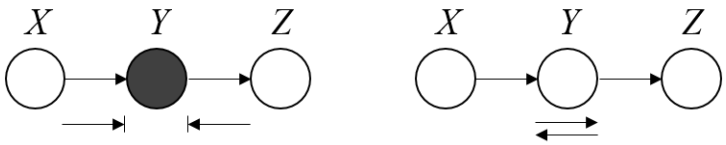

Figure 11.14: Markov chain rule: When Y is observed, balls are blocked (left). When Y is not observed, balls pass through (right)

1. **Markov chain:** Balls pass through when we do not observe Y, but are blocked otherwise.
2. **Two children:** Balls pass through when we do not observe Y, but are blocked otherwise.

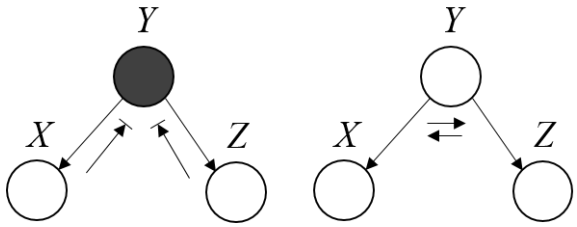

Figure 11.15: Rule when X and Z are Y's children: When Y is observed, balls are blocked (left). When Y is not observed, balls pass through (right)

3. **V-structure:** Balls pass through when we observe Y, but are blocked otherwise.

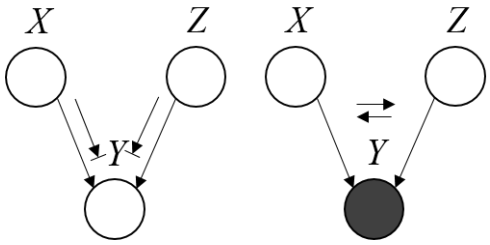

Figure 11.16: V-structure rule: When Y is not observed, balls are blocked (left). When Y is observed, balls pass through (right).

11.12 Properties: Inclusion, Reversal, Marginalization

Inclusion property. Here is a quite intuitive proposition about included graphs and their factorization.

Proposition 11.20 *If $G = (V, E)$ and $G' = (V, E')$ then:*

$$E \subset E' \Leftrightarrow \mathcal{L}(G) \subset \mathcal{L}(G') \tag{11.11}$$

Proof We have $p(x) = \prod_{i=1}^{n} p(x_i, x_{\pi_i(G)})$. As $E \subset E'$ it is obvious that $\pi_i(G) \subset \pi_i(G')$. Therefore, going back to the definition of graphical models through potential $f_i(x_i, x_{\pi_i})$ we get the result. ■

Reversal property. We also have some reversal properties. Let us first define the notion of V-structure.

Definition 11.21 *We say there is a V-structure (figure 11.10) in $i \in V$ if $|\pi_i| \geq 2$, i.e. has two or more parents.*

Proposition 11.22 *(Markov equivalence) If $G = (V, E)$ is a DAG and if for $(i, j) \in E, |\pi_i| = 0$ and $|\pi_j| \leq 1$, then (i, j) may be reversed, i.e. if $p(x)$ factorizes in G then it factorizes in $G' = (V, E')$ with $E' = (E - \{(i, j)\}) \cup \{(j, i)\}$.*

In terms of 3-nodes graph, this property ensures us that the Markov chain and latent cause are equivalent. Also, applying the reversal property multiplle times, we conclude that all directed trees built from an undirected tree give the same DGM.

On the other hand the V-structure lead to a different class of graph compared to the two others.

Definition 11.23 *An edge (i, j) is said to be covered if $\pi_j = \{i\} \cup \pi_i$.*

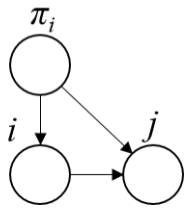

Figure 11.17: Edge (i, j) is covered

By reversing (i, j) we might not get a DAG as it might break the acyclic property. We have the following result:

Proposition 11.24 *Let $G = (V, E)$ be a graph and $(i, j) \in E$ a covered edge. Let $G' = (V, E')$ with $E' = (E - \{(i, j)\}) \cup \{(j, i)\}$, then if G' is a DAG, $\mathcal{L}(G) = \mathcal{L}(G')$.*

Marginalization. The underlying question is to know whether the marginalization of all distributions in a DGM yield another DGM. One can show that marginalizing the leaf node in a DGM yield a DGM on the smaller graph, but marginalizing internal nodes might yield a set of distributions which is not representable by a DGM.

/