1 Representing probabilistic independence in graphs

Graphical model = graphical representation of (conditional) independence relationships in a joint distribution = the distribution itself

graphical model - structure (a graph)
- parametrization (depends on the graph, parameters are “local”)

A graph is defined as $G = (V, E)$ where

- $V$ is the set of graph vertices (or nodes); each node represents a variable
- $E$ is the set of graph edges; edges encode the dependencies.

More precisely: a missing edge encodes an independence relationship.

Idea: Independence in the joint distribution $\iff$ Separation in graph

This mapping is not unique and not perfect. But even so, graphical representations are useful. Allowing for efficient computations is the main reason. Helping scientists understand a problem and communicate about it is another reason.

There are multiple “languages” for representing independence relations in graphs. The most popular ones are Markov random fields and Bayesian
Networks. Later on we will also study Decomposable models and junction tree representations. These graphical representations are tools for understanding and designing inference algorithms.

2 Markov Random Fields (Markov networks)

2.1 Encoding independencies in undirected graphs

An arbitrary undirected graph can be seen as encoding a set of independencies. The following rules states when two sets of variables $U_1, U_2 \subseteq V, U_1 \cap U_2 = \emptyset$ are separated in an undirected graph. We denote separation by $\perp$ and take it to mean “independence in the joint distribution over $V$”.

$$U_1 \perp U_2 \mid U_3 \iff \text{all paths between sets } U_1 \text{ and } U_2 \text{ pass through set } U_3$$

We say that $U_3$ blocks the paths between $X$ and $Y$; think of it as “blocking the flow of information”.

$n(A) = \text{the neighbors of variable } A$

A consequence of this rule is the following Markov property for MRFs also called the local Markov property:
A set of variables that separates node $A$ from the rest of the graph is called a Markov blanket for $A$. The set $n(A)$ is thus a Markov blanket, and it is the minimal Markov blanket of $A$. Adding a node to a Markov blanket preserves the Markov blanket property.

### 2.2 I-maps and perfect maps

Let us consider an undirected graph $G = (V, E)$ and a probability distribution $P$ over the set of variables $V$. If every separation relationship "$U_1$ separated from $U_2$ by $U_3$ from $G$ corresponds to a conditional independence $U_1 \perp U_2 \mid U_3$ under $P$, then we say that the $G$ is an I-map (independence map) of $P$.

One can think of a graph $G$ as a representant of the family of all probability distributions on $V$ for which $G$ is an I-map. Some of these distributions may have additional independencies, which do not appear in $G$. For example, any $G$ is an I-map for the distribution over $V$ in which all the variables are mutually independent.

If $G$ is an I-map of and $P$ has no independencies except for those represented by $G$, the we say that $G$ is a perfect map for $P$.

Any undirected graph $G$ has a perfect map (Geiger and Pearl, 88), but not any $P$ has a perfect map as an undirected graph. The distributions representable by graphs are a subclass of all distributions. An example of limitation imposed by the graph representation is that, in any graph (see for instance the graph above) $A \perp G \mid E$ and $B \perp G \mid E$ is equivalent to $\{A, B\} \perp G \mid E$. However, this is not always true in a distribution (example: the parity function).
2.3 Factorization

Now we will characterize the set of distributions for which a graph $G$ is an I-map. For this, we need a new definition. A **clique** of a graph $G$ is a set of nodes $C \subseteq V$ which are **fully connected** in $G$ (i.e. all possible edges between nodes in $C$ appear in $E$). A **maximal** clique is a clique which is not contained in any other clique of the graph.

For example, in figure 1, all the nodes are cliques of size one (but not maximal), all the edges are cliques of size two, and the triangles $CDE$, $EFG$ are cliques of size three. The maximal cliques are $AB$, $BD$, $AC$, $CDE$, $EFG$.

**Theorem 1** Let $G$ be a graph and assume $P$ can be factored in the following way

$$P = \prod_{C \text{ maximal clique in } G} \phi_C(x_c)$$  \hspace{1cm} (1)

where $\phi_C$ is a non-negative function depending only on the variables in $C$. Then, $G$ is an I-map of $P$.

We will illustrate this theorem by an example shortly. The converse is a more powerful result, and is known as the Hammersley-Clifford theorem.

**Theorem 2 (Hammersley-Clifford)** If $P > 0$ and $G$ is an I-map of $P$, then $P$ can be written as a product of functions defined over the cliques of $G$ as in (1).

**Exercise** The theorem doesn’t always hold if $P(x) = 0$ for some $x$. Can you construct such a counterexample? (Hint: give $P$ lots of zeros.)

If a distribution $P$ can be written in the form (1) for some graph $G$ we say that $P$ **factors according to graph $G$**.

**Example** For the undirected $G$ in figure 1

$$P_{ABCDEFG} = \phi_{AB}(a,b)\phi_{AC}(a,c)\phi_{BD}(b,d)\phi_{CDE}(c,d,e)\phi_{EFG}(e,f,g)$$
The functions $\phi$ are called **clique potentials**. They are required to be non-negative (positive if $P > 0$). Clique potentials are not uniquely defined. One can obtain equivalent factorizations by dividing/multiplying with functions of variables that are common between cliques. For instance, we can rewrite the above joint distribution as

$$P_{ABCDEFG} =$$

$$= (2\phi_{AB}(a, b))(\phi_{AC}(a, c)/2)\phi_{BD}(b, d)\phi_{CDE}(c, d, e)\phi_{EFG}(e, f, g)$$

$$= (h(a)\phi_{AB}(a, b))(\phi_{AC}(a, c)/h(a))\phi_{BD}(b, d)\phi_{CDE}(c, d, e)\phi_{EFG}(e, f, g) \text{ for any } h(a) > 0$$

$$= \phi_{AB}(a, b)\phi_{AC}(a, c)\phi_{BD}(b, d)(\phi_{CDE}^\prime(c, d, e)h(c, d))\phi_{EFG}(e, f, g)$$

The last example shows why we only need to consider maximal cliques in the factorization of $P$. Because of the non-unicity of the $\phi$’s, the parameters of the clique potentials are hard to interpret. The potentials do not, in general, represent probability tables. However, there are some important special cases when the $\phi$’s have probabilistic interpretations – these will be the decomposable models we will study later. The Hidden Markov model you have already encountered is one of them.

**Where do the $\phi$ potentials come from?** Sometimes, they come from physical models, where $(-\log \phi)$ represents an energy. This is the case of the Ising model in lecture 1. Note that the potential energy is defined up an additive constant; this fits with the $\phi$ potential being defined up to multiplicative constants.

Other times, they are “made up” by e.g engineers who want to represent a problem. For example the lattice models representing images, are MRF’s where the graph and the potential functions are artificial (but useful) representations for images.

Sometimes, the potentials are obtained by a combination of scientific grounds, estimation from data, and convenience consideration. This is the case in the modeling of spatial processes. In such processes, the graph can represent: a grid of locations where weather measurements are taken (an irregular network), the states of the US, with edges between neighboring states (for the study of e.g ecological processes), a watershed (points and edges along rivers), a transportation network or a social contacts network (in epidemiology), etc.

In a factored representation the savings in terms of number of parameters
w.r.t the multidimensional table representation are significant. Assume that all variables are binary, and all potentials are represented by (unnormalized) tables. Then for the graph in figure 1 the total number of parameters is

\[ 3 \times 2^2 + 2 \times 2^3 = 28 \]

The size of a probability table over 7 binary variables is \(2^7 - 1 = 127\) thus in this example we save 99 parameters (almost 80%).

### 2.4 Factorization (1) implies the independencies prescribed by \(G\) – an example

We will consider the example in figure 1. Let’s prove that

\[ \{A, B\} \perp F \mid C, D \text{ in } P \]

if \(P\) factors according to the graph.

We will use the following fact about distributions:

**Lemma 3** \(X \perp Y \mid Z\) under \(P\) iff there exist functions \(h_1(x, z), h_2(y, z)\) such that \(P_{XYZ}(x, y, z) = h_1(x, z)h_2(y, z)\).

**Exercise** Prove the lemma.

To use Lemma 3 in our proof, we write \(P_{ABCDF}\) in product form.

\[
P_{ABCDF}(a, b, c, d, e, f, g) =
\]

\[
= \sum_{e,g} P_{ABCDEFG}
\]

\[
= \sum_{e,g} \phi_{AB}(a, b)\phi_{AC}(a, c)\phi_{BD}(b, d)\phi_{CDE}(c, d, e)\phi_{EFG}(e, f, g)
\]

\[
= \phi_{AB}(a, b)\phi_{AC}(a, c)\phi_{BD}(b, d) \sum_{e} \phi_{CDE}(c, d, e) \sum_{g} \phi_{EFG}(e, f, g)
\]

\[ \underline{\psi_{EF}(e,f)} \]
\[ \phi_{AB}(a, b)\phi_{AC}(a, c)\phi_{BD}(b, d) \sum_{e} \phi_{CDE}(c, d, e)\psi_{EF}(e, f) \quad (6) \]

\[ \sum_{e} \phi_{CDE}(c, d, e)\psi_{CDF}(c, d, f) \quad (7) \]

From the last form of the marginal \( P_{ABCD} \) we can conclude that the relationship \( \{A, B\} \perp F | C, D \) is true under \( P \).

### 2.5 Gaussian Markov fields

A multivariate normal distribution over has the form

\[ P_V(x_V) \propto e^{-\frac{1}{2}(x_V - \mu_V)^T \Sigma^{-1}(x_V - \mu_V)} \quad (8) \]

If we denote the inverse covariance by \( D = \Sigma^{-1} \) and if, for simplicity, we assume \( \mu_V \equiv 0 \), then the multivariate normal over \( V \) can be written as

\[ P_V(x_V) \propto e^{-\frac{1}{2}x_V^TDx_V} \]

\[ = e^{-\frac{1}{2} \sum_{i \in V} D_{ii}x_i^2 - \sum_{i<j} D_{ij}x_ix_j} \quad (9) \]

This is equivalent with a MRF that has an edge for every non-zero \( D_{ij} \). Conversely, in multivariate Gaussian, the zeros in the inverse covariance matrix encode the conditional independencies.

### 2.6 Markov chains

The joint distribution

\[ P_{X_1X_2X_3X_4X_5} = (P_{X_1}P_{X_2|X_1})P_{X_3|X_2}P_{X_4|X_3}P_{X_5|X_4} \]

is a product of conditional distributions involving \( X_{t+1}, X_t \). \( X_{t+1}, X_t \) are neighbors in the chain. In this case the clique potentials, one for each edge,
are $\phi_{1,2} = P_{X_1}P_{X_2|X_1}$, $\phi_{i,i+1} = P_{X_{i+1}|X_i}$, for $i > 1$, and have a probabilistic interpretation.

The same can be said about the Hidden Markov Model (lecture 2).

## 2.7 Trees

![Tree Diagram]

**Tree** = connected graph with no cycles (we also call it **spanning** tree). If disconnected and no cycles, we call it a **forest**. Sometimes we use the term tree to mean either a spanning tree or a forest.

Property: between every two variables in a spanning tree there is exactly one path (at most one path for forests).

All cliques in a spanning tree have size 2.

## 3 Bayesian Networks

### 3.1 A limitation of Markov networks

Suppose that a variable called “burglar alarm (A)” becoming true can have two causes: a burglary (B) or an earthquake (E). A reasonable assumption is that burglaries and earthquakes occur independently ($B \perp E$). But, given that the alarm sounds $A = 1$, and hearing that an earthquake as taken place ($E = 1$), most people will believe that the burglary is less likely to have
taken place than if there had been no earthquake ($E = 0$), hence $E$ carries information about $B$ when $A = 1$ and therefore $B, E$ are dependent given $A$. In other words, it is reasonable to assume that

\begin{align}
B & \perp E \\
B & \not\perp E|A.
\end{align}

Markov nets cannot represent this situation.

**Exercise** Find other examples of the kind (i) $X \perp Y$ and $X \not\perp Y|Z$, i.e two independent causes which can produce the same effect. Find examples that fit the other three possible combinations of marginal and conditional (in)dependence, i.e. (ii) $X \perp Y, X \perp Y|Z$, (iii) $X \not\perp Y, X \perp Y|Z$, (iv) $X \not\perp Y, X \not\perp Y|Z$. Can you describe them in words?

The case illustrated above is important enough to warrant the introduction of another formalism for encoding independencies in graphs, called **D-separation** and based on **directed graphs**.

### 3.2 Directed Acyclic Graphs (DAG’s)

A **Directed Acyclic Graph (DAGs)** is a directed graph $G = (V, \vec{E})$ which contains no directed cycles.

![Diagram of DAGs](image)

this is a DAG  this is not a DAG

Below is a somewhat more complicated textbook example (the “chest-clinic” example). In this example, the arrows coincide with “causal links” between variables (i.e “Smoking causes Lung cancer”). This is not completely accidental. Bayesian networks are particularly fit for representing domains where
there are causal relationships. It is therefore useful to think of causal relationships when we try to build a Bayes net that represents a problem. But the formalism we study here is not necessarily tied to causality. One does not have to interpret the arrows as cause-effect relations and we will not do so. **Terminology:**

- **parent** Asia is parent of Tuberculosis
- **pa(variable)** the set of parents of a variable
  
  \[
  pa(\text{X-ray}) = \{\text{Lung cancer, Tuberculosis}\}
  \]
- **child** Lung cancer is child of Smoker
- **ancestor** Smoker is ancestor of Dyspnoea
- **descendent** Dyspnoea is descendent of Smoker
- **family** a node and its parents
  
  \{\text{Dyspnoea, Tuberculosis, Lung cancer, Bronchitis}\} are a family

But perhaps the most important concept in DAG’s is the **V-structure**, which denotes a variable having two parents which are *not connected* by an edge. The Burglar-Earthquake-Alarm example of the previous section is the V-structure.
Burglar (B) ─ Earthquake (E) ─ Landslide (L) ─ Earthquake (E) ─ Alarm (A)

a V-structure

not a V-structure

In figure 2, \((T, X, L), (T, D, L), (T, D, B)\) are V-structures.

### 3.3 D-separation

In a DAG, independence is encoded by the relation of d-separation, defined below.

\[
A \perp B \mid C \iff A \text{ d-separated from } B \text{ by } C
\]

**D-separation**: \(A\) is d-separated from \(B\) by \(C\) if all the paths between sets \(A\) and \(B\) are blocked by elements of \(C\). The three cases of d-separation:

1. \(X \rightarrow Z \rightarrow Y\)
   - if \(Z \in C\) the path is blocked, otherwise open

2. \(X \rightarrow Z \leftarrow Y\)
   - if \(Z \in C\) the path is blocked, otherwise open

3. \(X \rightarrow Z \rightarrow Y\)
   - if \(Z\) or one of its descendents \(\in C\) the path is open, otherwise blocked

**The directed Markov property**: \(X \perp \text{its non-descendants} \mid pa(X)\)
3.4 Equivalent DAG’s

Two DAG’s are said to be (likelihood) equivalent if they encode the same set of independencies. For example, the two graphs below are equivalent (encoding no independencies).

A \longrightarrow B \quad A \leftarrow B

If the arrows represented causal links, then the two above graphs would not be equivalent!

Another example of equivalence between DAGs was seen in lecture 2: HMM’s can be represented by a directed graph with “forward” arrows, as well as by one with “backward” arrows. (Note, in passing, that an HMM can also be represented as an undirected graph, demonstrating equivalence between a DAG and a MRF).

Yet another example is below. On the left, the ”chest clinic” DAG. In the middle, an equivalent DAG. On the right, another DAG which is not equivalent with the chest clinic example. [Exercise: verify that the graphs are/are not equivalent by looking at what independence relationships hold in the three graphs.]

---

Theorem 4 (Chickering) Two DAG’s are equivalent iff they have the same undirected skeleton and the same V-structures.

Consequently, we can invert an arrow in a DAG and preserve the same independencies, only if that arrow is not part of a V-structure. It can be proved
that one can traverse the class of all equivalent DAG’s by successive arrow reversals.

### 3.5 D-separation as separation in an undirected graph

Here we show that D-separation in a DAG is equivalent to separation in an undirected graph obtained from the DAG and the variables we are interested in. First two definitions, whose meaning will become clear shortly.

**Moralization** is the graph operation of connecting the parents of a V-structure. A DAG is **moralized** if all nodes that share a child have been connected. After a graph is moralized, all edges, be they original edges or new edges added by moralization, are considered as undirected. If $G$ is a DAG the graph obtained by moralizing $G$ is denoted by $G^m$ and is called the **moral graph** of $G$.

![Moralization Graph](image)

For any variable $X$ the set $\text{an}(X)$ denotes the ancestors of $X$ (including $X$ itself). Similarly, if $A$ is a set of nodes, $\text{an}(A)$ denotes the set of all ancestors of variables in $A$.

$$\text{an}(A) = \bigcup_{X \in A} \text{an}(X)$$

The **ancestral graph** of a set of nodes $A \subseteq V$ is the graph $G_{\text{an}(A)} = (\text{an}(A), E_A)$ obtained from $G$ by removing all nodes not in $\text{an}(A)$.

Now we can state the main result.
**Theorem 5** Let $A, B, S \subseteq V$ be three disjoint sets of nodes in a DAG $G$. Then $A, B$ are $D$-separated by $S$ in $G$ iff they are separated by $S$ in the moral ancestral graph of $A, B, S$.

$$A \perp B \mid S \text{ in } G \quad \text{iff} \quad A \perp B \mid S \text{ in } (G_{an(A \cup B \cup S)})^m$$

The intuition is that observing/conditioning on a variable creates a dependence between its parents (if it has any). Moralization represents this link. Now why the ancestral graph? Note that an unobserved descendent cannot produce dependencies between its ancestors (ie cannot open a path in a directed graph). So we can safely remove all descendents of $A, B$ that are not in $S$. The descendents of $S$ itself that are not in $A, B$, and all the nodes that are not ancestors of $A, B, S$ can be removed by a similar reasoning. Hence, first the graph is pruned, then dependencies between parents are added by moralization. Now directions on edges can be removed, because DAG’s are just like undirected graphs if it weren’t for the V-structures, and we have already dealt with those.

The Theorem immediately suggests an algorithm for testing $D$-separation using undirected graph separation.

1. remove all nodes not in $an(A \cup B \cup S)$ to get $G_{an(A \cup B \cup S)}$
2. moralize the remaining graph to get $(G_{an(A \cup B \cup S)})^m$
3. remove all nodes in $S$ from $(G_{an(A \cup B \cup S)})^m$ to get $G'$
4. test if there is a path between $A$ and $B$ in $G'$

For example, test if $S \perp B \mid D$ in the chest clinic DAG.
3.6 Factorization

Now we construct joint probability distributions that have the independencies specified by a given DAG. Assume the set of discrete variables is \( V = \{X_1, X_2, \ldots, X_n\} \) and that we are given a DAG \( G = (V, E) \). The goal is to construct the family of distributions that are represented by the graph. This family is given by

\[
P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{pa}(X_i))
\]

In the above \( P(X_i|\text{pa}(X_i)) \) represents the conditional distribution of variable \( X_i \) given its parents. Because the factors \( P(X_i|\text{pa}(X_i)) \) involve a variable
Figure 3: The “chest clinic” DAG, with shorter variable names.

and its parents, that is, nodes closely connected in the graph structure, we often call them local probability tables (or local distributions).

Note that the parameters of each local table are (functionally) independent of the parameters in the other tables. We can choose them separately, and the set of all parameters for all conditional probability distributions form the family of distributions for which the DAG $G$ is an I-map.

If a distribution can be written in the form (13) we say that the distribution factors according to the graph $G$. A joint distributions that factors according to some DAG $G$ is called a Bayes net.

Note that any distribution is a Bayes net in a trivial way: by taking $G$ to be the complete graph, with no missing edges. In general, we want a Bayes net to be as sparse as possible, because representing independences explicitly has many computational advantages.

The Bayes net described by this graph is

$$P(A, S, T, L, B, X, D) = P(A)P(S)P(T|A)P(L|S)P(B)P(X|T, L)P(D|T, L, B)$$

A way of obtaining this decomposition starting from the graph is

1. Construct a topological ordering of the variables. A topological or-
**dering** is an ordering of the variables where the parents of each variable are always before the variable itself in the ordering. 

*A, S, T, L, B, X, D* is a topological ordering for the graph above.

2. Apply the chain rule following the topological ordering.

\[
\]
\[
P(X|A, S, T, L, B)P(D|A, S, T, L, B, X'S)
\]

3. Use the directed Markov property to simplify the factors

\[
P(S|A) = P(S)
\]
\[
P(T|A, S) = P(T|A)
\]
\[
P(L|A, S, T) = P(L|S)
\]
\[
P(B|A, S, T, L) = P(B), \text{ etc.}
\]

Let us now look at the number of parameters in such a model. Assume that in the example above all variables are binary. Then the number of unconstrained parameters in the model is

\[
1 + 1 + 2 + 2 + 1 + 4 + 8 = 19
\]

The number of parameters in a 7 way contingency table is \(2^7 - 1 = 127\) so we are saving 118 parameters. As we shall see, there are also other computational advantages to joint distribution representations of this form.