

## Lecture 12 — October 12

Lecturer: Simon Lacoste-Julien

Scribe: Philippe Beardsell

Based on the scribe notes from Jaime Roquero and JieYing Wu.

Proofread and quickly corrected by Simon Lacoste-Julien.

## General themes in this class

(A) Modeling high dimensional distributions

- Representation: how to represent a family of distributions. → Examples of convenient families are given by graphical models (DGM, UGM).
- Parametrization: how to parameterize the members of the family of distributions → an example for this that we will see is using the exponential family (but there are many others)

(B) Inference → how do we compute  $p(x_Q | x_E)$ , where  $Q$  is the query and  $E$  the evidence?

- Lecture 13 : elimination algorithm
- Lecture 14 : sum-product algorithm (belief propagation)

(C) Statistical estimation: how do we estimate the model from observations? → Examples of principles that we see: maximum likelihood estimators, maximum entropy, method of moments

## 12.1 Undirected Graphical Models (UGM)

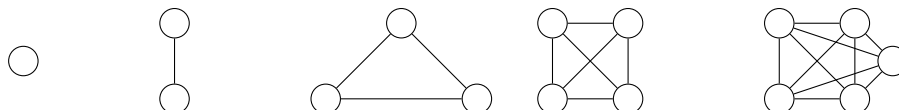
(a.k.a. Markov random fields or Markov networks)

let  $G = (V, E)$  be an undirected graph

and let  $\mathcal{C}$  be the set of **cliques** of  $G$ , where a clique is a fully connected set of nodes

(i.e.  $C \in \mathcal{C} \iff \forall i \neq j \in C, \{i, j\} \in E$ )

Examples of set of nodes which are cliques from size 1 to 5 :




### 12.1.1 UGM associated with G

$$\mathcal{L}(G) \triangleq \left\{ p : p(x_V) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \text{ for some "potentials" } \psi_C \text{ s.t. } \psi_C(x_C) \geq 0 \ \forall x_C \right\}$$

and where  $Z$  is the normalizing constant

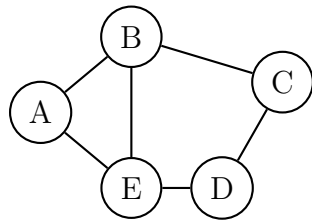
$$Z \triangleq \sum_x \left( \prod_{C \in \mathcal{C}} \psi_C(x_C) \right) \quad \text{"partition function"}$$

 The functions  $\psi_C$  are potential functions and are not probability distributions ! Unlike in a DGM, where we could think of  $C$  to be the node and its parents, which implies  $\psi_C(x_C) = p(x_i, x_{\pi_i})$ , in a UGM, the potential  $\psi_C(x_C)$  is not directly related to the probability distribution  $p(x_C)$ .

**Remark:**

- We can multiply any  $\psi_C(\cdot)$  by a constant without changing  $p$  (because we will re-normalize with a new  $Z$ )

Therefore, for some undirected graph  $G$  there are multiple ways to define the probability  $p(x_V)$ . For example, consider the following graph



we could write

$$P(A, B, C, D, E) \propto \psi(A, B, E)\psi(A, B)\psi(A, E)\psi(B, E) \cdot \psi(B, C)\psi(C, D)\psi(D, E)$$

but we could also write

$$P(A, B, C, D, E) \propto \psi'(A, B, E)\psi(B, C)\psi(C, D)\psi(D, E)$$

Note that in the second equation, we can rewrite  $\psi(A, B)\psi(A, E)\psi(B, E)$  to the simpler potential function  $\psi'(A, B, E)$ , as  $(A, B, E)$  form a clique of 3 nodes. The potential function  $\psi'(A, B, E)$  encompasses all the information about the dependencies between the nodes  $(A, B, E)$ , so there is no loss of generality in making that transformation. Therefore, it is sufficient to consider only  $\mathcal{C}_{\max}$ , the set of **maximal cliques**, where a maximal clique is a clique that cannot be extended by including an additional vertex. We can restrict ourselves to that case given that all cliques are subsets of one or more maximal cliques.

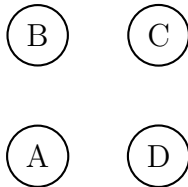
e.g.  $C' \subseteq C$ , then redefine  $\psi_C^{new}(x_C) = \psi_C^{old}(x_C)\psi_{C'}^{old}(x_{C'})$

[Note: we will see later that it is sometimes convenient to consider the "over-parametrization" of trees using both  $\psi_i(x_i)$  and  $\psi_{ij}(x_i, x_j)$ ]

**Property 12.1.1** *as before*,  $E \subseteq E' \implies \mathcal{L}(G) \subseteq \mathcal{L}(G')$

Trivial graphs

- consider  $G = (V, E)$  with  $E = \emptyset$

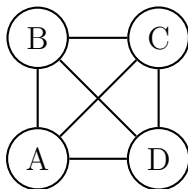


For  $p \in \mathcal{L}(G)$ , we get:

$$p(x_V) = \prod_{i=1}^n \psi_i(x_i) \text{ as } \mathcal{C} = \{\{i\} \in V\}$$

This gives us that  $\mathcal{L}(G)$  is the fully factorized set and that  $X_1, \dots, X_n$  are all mutually independent.

- consider  $G = (V, E)$  with  $\forall i, j \in V, \{i, j\} \in E$  (i.e.  $G$  is one big clique)



For  $p \in \mathcal{L}(G)$ , we get:

$$p(x) = \frac{1}{Z} \psi_V(x_V) \text{ as } \mathcal{C} \text{ is reduced to a single set } V$$

We make no conditional independence assumptions between any of the  $x_i$ ; and **any** distribution is in  $\mathcal{L}(G)$ .

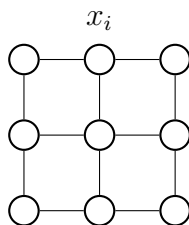
Property 12.1.2

- if  $\psi_C(x_C) > 0 \quad \forall x_C$

we can then see that  $p$  is in an *exponential family*:

$$p(x_V) = \exp \left\{ \underbrace{\sum_{C \in \mathcal{C}} \overbrace{\log \psi_C(x_C)}^{<\theta_C, T_C(x_C)>}}_{\text{negative energy function}} - \log Z \right\}$$

Example: Ising model in physics :  $x_i \in \{0, 1\}$



node potentials  $\rightarrow E_i = \psi_i(x_i = 1)$   
 edge potentials  $\rightarrow E_{i,j} = \psi_{ij}(x_i = 1, x_j = 1)$

Another example could be social network modeling.

### 12.1.2 Conditional independence for UGM

As for the directed graphical models, we can view the undirected graphical models as encoding a set of independence assumptions in their structure.

**Definition 12.1** We say that  $p$  satisfies the *global Markov property* (with respect to an undirected graph  $G$ ) if and only if

$\forall$  disjoint  $A, B, S \subseteq V$  s.t.  $S$  separates  $A$  from  $B$  in  $G$ , then we have:  $X_A \perp\!\!\!\perp X_B \mid X_S$ .

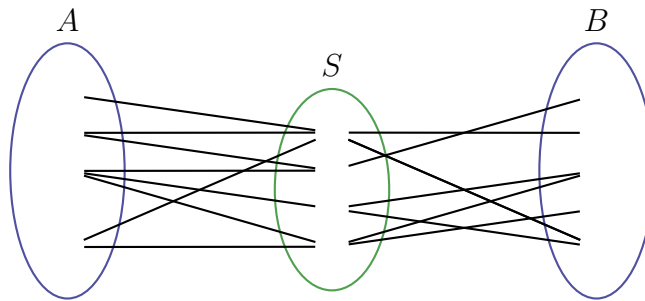


Figure 12.1: The set  $S$  separates  $A$  from  $B$ . All paths from  $A$  to  $B$  must pass through  $S$ .

#### Proposition 12.2

$p \in \mathcal{L}(G) \implies p$  satisfies the global Markov property for  $G$

##### Proof:

Without loss of generality, we can assume  $A \cup B \cup S = V$ .

To see why, consider the case where  $A \cup B \cup S \subset V$ . Then, let

$$\tilde{A} \triangleq A \cup \{a \in V : a \text{ and } A \text{ are not separated by } S\}$$

$$\text{and } \tilde{B} \triangleq V \setminus \{S \cup \tilde{A}\}$$

By definition, we have the disjoint union  $\tilde{A} \cup \tilde{B} \cup S = V$ , and we now show that  $\tilde{A}$  and  $\tilde{B}$  are separated by  $S$ . By contradiction, suppose there is an  $a \in \tilde{A}$  and  $b \in \tilde{B}$  which are **not** separated by  $S$ , i.e. there exists a path from  $a$  to  $b$  not passing through  $S$ . Then by definition,  $b$  would be in  $\tilde{A}$ , contradicting the definition of  $\tilde{B}$  (as  $b$  cannot be in  $\tilde{A}$  and  $\tilde{B}$  at the same time). We also have that  $B \subseteq \tilde{B}$  as the original  $B$  was separated from  $A$  by  $S$ . Thus we have  $\tilde{A} \cup \tilde{B} \cup S = V$  and  $\tilde{A}$  and  $\tilde{B}$  are separated by  $S$ . If we can show that  $X_{\tilde{A}} \perp\!\!\!\perp X_{\tilde{B}} \mid X_S$ , then by the decomposition property, this implies  $X_A \perp\!\!\!\perp X_B \mid X_S$  for any subsets  $A$  of  $\tilde{A}$  and  $B$  of  $\tilde{B}$ , giving the required general case. We thus continue the proof with  $A \cup B \cup S = V$ .

Let  $C \in \mathcal{C}$ . We cannot have  $C \cap A \neq \emptyset$  and  $C \cap B \neq \emptyset$ , i.e. the clique  $C$  can't intersect both  $A$  and  $B$  at the same time (otherwise, part of  $B$  would be connected to  $A$  by direct edges from this clique). Thus,

$$p(x) = \frac{1}{Z} \prod_{\substack{C \in \mathcal{C} \\ C \subseteq A \cup S}} \psi_C(x_C) \prod_{\substack{C' \in \mathcal{C} \\ C' \not\subseteq A \cup S \\ \Rightarrow C' \subseteq B \cup S \\ C' \not\subseteq S}} \psi_{C'}(x_{C'}) = f(x_{A \cup S})g(x_{B \cup S})$$

$$\begin{aligned} p(x_A | x_S) &\propto p(\underbrace{x_A, x_S}_{x_{A \cup S}}) = \sum_{x_B} f(x_{A \cup S})g(x_{B \cup S}) \\ &= f(x_{A \cup S}) \underbrace{\sum_{x_B} g(x_{B \cup S})}_{\text{cst w.r.t } x_A} \\ \Rightarrow p(x_A | x_S) &= \frac{f(x_A, x_S)}{\sum_{x'_A} f(x'_A, x_S)} \end{aligned}$$

Similarly,

$$p(x_B | x_S) = \frac{g(x_B, x_S)}{\sum_{x'_B} g(x'_B, x_S)}$$

Thus,

$$p(x_A | x_S)p(x_B | x_S) = \frac{f(x_{A \cup S})g(x_{B \cup S})}{\sum_{x'_A} \sum_{x'_B} f(x'_A, x_S)g(x'_B, x_S)} = \frac{p(x_V)}{p(x_S)} = p(x_A, x_B | x_S)$$

This proves  $X_A \perp\!\!\!\perp X_B \mid X_S$ .  $\square$

To converse of the above theorem is not always true (see assignment 3), but if we assume that the probability is strictly positive, it holds as given in the following (deep) theorem.

**Theorem 12.3** (*Hammersley-Clifford*)

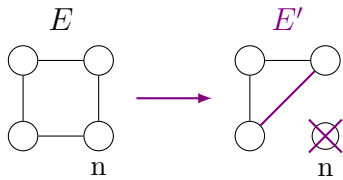
$$\text{if } p(x_V) > 0 \ \forall x_V$$

then,  $p \in \mathcal{L}(G) \iff p$  satisfies the global Markov property.

**Proof:** see chapter 16 of Michael I. Jordan's book

**Property 12.1.3** *Closure with respect to marginalization*

As for directed graphical models, we also have a marginalization notion in undirected graphs, but it is slightly different. If  $p(x)$  factorizes in  $G$ , then  $p(x_1, \dots, x_{n-1})$  factorizes in the graph where the node  $n$  is removed **and all neighbors are connected**.



let  $V' = V \setminus \{n\}$   
 $E' = \text{edges in } G \setminus \{n\} + \text{connect all neighbors of } n \text{ in } G \text{ together (new clique)}$

$$\{\text{marginal on } x_{1:n-1} \text{ for } p \in \mathcal{L}(G)\} = \mathcal{L}(G') .$$

### 12.1.3 DGM vs UGM

**Definition 12.4** *Markov blanket* The Markov blanket for a node  $i$  is the smallest set of nodes  $M$  such that the node  $X_i$  is conditionally independent of all the other nodes ( $X_V$ ) given  $X_M$ :

$$X_i \perp\!\!\!\perp X_V \mid X_M .$$

- for an UGM:  $M = \{j : \{i, j\} \in E\} = \text{set of neighbors of } i$
- for a DGM: the Markov blanket of node  $i$  include its parents, its children and the parents of all its children, i.e.

$$M = \pi_i \cup \text{children}(i) \cup \bigcup_{j \in \text{children}(i)} \pi_j .$$

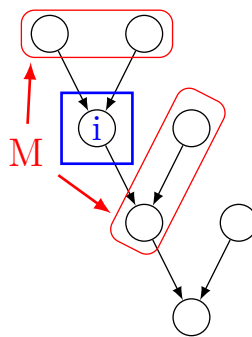
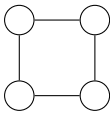
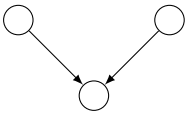


Table 12.1 summarizes the differences between DGM and UGM.

### 12.1.4 Moralization

Let  $G$  be a DAG; when can we transform  $G$  to an undirected graph  $\bar{G}$  such that the DGM from  $G$  is the same as the UGM on  $\bar{G}$ ? Before answering this question, we first define the undirected graph  $\bar{G}$  so that  $\mathcal{L}(G) \subseteq \mathcal{L}(\bar{G})$ .

Table 12.1: Summary of the main differences between DGM and UGM

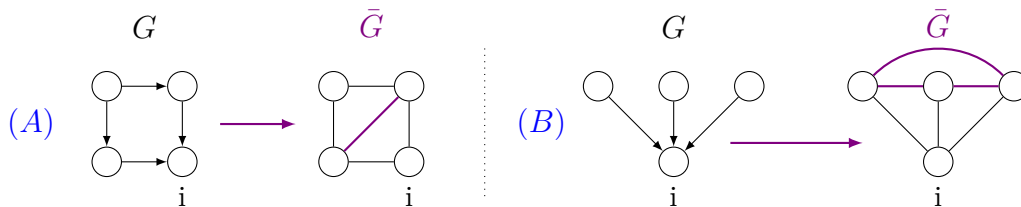
	Directed graphical model	Undirected graphical model
Factorization	$p(x) = \prod_{i=1}^n p(x_i   x_{\pi_i})$	$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$
Conditional independence	d-separation $[X_i \perp\!\!\!\perp X_{nd(i)} \mid X_{\pi_i}]$ and many more!	separation $[X_A \perp\!\!\!\perp X_B \mid X_S]$
Marginalization	not closed in general, only when marginalizing leaf nodes	closed
cannot exactly capture some families	grid 	v-structure 

**Definition 12.5** for  $G$  a DAG, we call  $\bar{G}$  the *moralized graph* of  $G$

where  $\bar{G}$  is an undirected graph with the same set of vertices  $V$

$$\text{and } \bar{E} = \underbrace{\{ \{i, j\} : (i, j) \in E \}}_{\text{undirected version of } E} \cup \underbrace{\{ \{k, l\} : k \neq l \in \pi_i \text{ for some } i \}}_{\text{"moralization"}}$$

That is, the moralization<sup>1</sup> can be explained less formally as **connecting all the parents of  $i$  ( $\pi_i$ ) with  $i$  in a big clique**. Note that we only need to add edges when  $|\pi_i| > 1$ , i.e. when there is a v-structure. Here are two examples of this transformation :



Note that in the conversion process from a Bayesian network to a Markov random field, we lose the marginal independence of the parents.

We are now in position to answer the original question of when a DGM yields the same as a UGM.

<sup>1</sup>Note that the terminology “moralization” come from the fact that we are “marrying” all the parents (by adding edges between them), and thus from a traditional Christian point of view, we are making the “family moral”.

**Proposition 12.6** for a DAG  $G$  with no  $v$ -structure [forest]

$$\text{then } \mathcal{L}(G) = \mathcal{L}(\bar{G})$$

but in general, we can only say that  $\mathcal{L}(G) \subseteq \mathcal{L}(\bar{G})$

(note that  $\bar{G}$  is the minimal undirected graph such that  $\mathcal{L}(G) \subseteq \mathcal{L}(\bar{G})$ )

**Proof:** This will be done in assignment!

**Proposition 12.7 (Flipping a covered edge in a DGM)** Let  $G = (V, E)$  be a DAG. We say that a directed edge  $(i, j) \in E$  is a *covered edge* if and only if  $\pi_j = \pi_i \cup \{i\}$ . Suppose the edge  $(i, j) \in E$  is covered and define  $G' = (V, E')$ , with  $E' = (E \setminus \{(i, j)\}) \cup \{(j, i)\}$ . Prove that  $\mathcal{L}(G) = \mathcal{L}(G')$ .

**Proof.** Note that in order to identify the factors of the decomposition of the joint distribution provided by  $G'$  with conditional distributions, we need to show that  $G'$  is indeed a DAG! We know that  $G$  is a DAG, but must prove that flipping  $(i, j)$  did not introduce any cycles for  $G'$ .

**$G'$  is a DAG.** Recall that a graph is a DAG if and only if it has a topological order. WLOG, assume that the vertices of the original graph  $G$  are indexed with such a topological ordering  $(1, \dots, i, \dots, j, \dots, n)$  and so  $j = i + k$  (for some  $k \in \mathbb{N}$ ).

Now, the sequence  $(1, \dots, i, j, i + 1, \dots, i + k - 1, j, i + k + 1, \dots, n)$  is also a topological ordering of  $G$  since  $\pi_j \subset \{1, \dots, i\}$  and (b)  $\forall m > 0$ , if  $\pi_{i+m} \subset \{1, \dots, i + m - 1\}$  then  $\pi_{i+m} \subset \{1, \dots, i + m - 1\} \cup \{j\}$ .

Then,  $(1, \dots, j, i, \dots, n)$  is a topological ordering of  $G'$  since everyone's ancestors are to their left. Therefore,  $G'$  is a DAG.

$\mathcal{L}(G) \subseteq \mathcal{L}(G')$ . Let  $p \in \mathcal{L}(G)$ . We thus have  $p(x) = \prod_{k=1}^n p(x_k | x_{\pi_k})$ , where  $\pi_k$  denotes the parents of  $k$  in  $G$ . Consider any  $x_i, x_j, x_{\pi_i}$  such that  $p(x_i, x_j, x_{\pi_i}) \neq 0$ . Then by the chain rule (valid for any distribution), we have

$$p(x_i | x_{\pi_i})p(x_j | x_i, x_{\pi_i}) = p(x_i, x_j | x_{\pi_i}) = p(x_j | x_{\pi_i})p(x_i | x_j, x_{\pi_i}). \quad (12.1)$$

As  $(i, j)$  is a covered edge, we have  $\pi_j = \pi_i \cup \{i\}$ . Moreover, by definition of  $E'$ , we have  $\pi'_j = \pi_i$  and  $\pi'_i = \pi_j \cup \{j\}$  with  $\pi'_i$  the parents of  $i$  in  $G'$ . So note that equation (12.1) can be interpreted as:

$$p(x_i | x_{\pi_i})p(x_j | x_{\pi_j}) = p(x_j | x_{\pi'_j})p(x_i | x_{\pi'_i}).$$

As  $\pi'_k = \pi_k$  for any  $k \neq i, j$ , we can simply swap the two terms for  $i$  and  $j$  in the product factorization of  $p$ :

$$p(x) = p(x_i | x_{\pi_i})p(x_j | x_{\pi_j}) \prod_{k \neq i, j} p(x_k | x_{\pi_k}) = p(x_j | x_{\pi'_j})p(x_i | x_{\pi'_i}) \prod_{k \neq i, j} p(x_k | x_{\pi'_k}).$$



If  $p(x_i, x_j, x_{\pi_i}) = 0$ , then both the LHS and RHS above are equal to zero and so are still equal. We thus have  $p \in \mathcal{L}(G')$ . By symmetry, we can reverse the argument, and thus  $\mathcal{L}(G) = \mathcal{L}(G')$ .  $\square$