Probabilistic Graphical Models in Computer Vision (IN2329)

Csaba Domokos

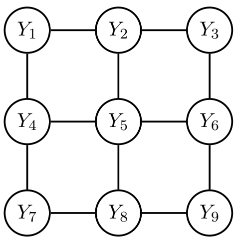
Summer Semester 2017

2. Graphical models

Agenda for today's lecture *

In the **previous lecture** we learnt about

- Discrete probability space
- Conditional probability
- Independence, conditional independence



Today we are going to learn about

1. Random variables $(Y_1, \ldots, Y_9)$
2. Probability distributions
   - Joint distribution $(p(y_1, \ldots, y_9))$
   - Marginal distribution $(p(y_1))$
   - Conditional distribution $(p(y \mid x))$
3. Graphical models

σ-algebra, measure, measure space *

Assume an arbitrary set Ω and $\mathcal{A} \subseteq \mathcal{P}(\Omega)$. The set $\mathcal{A}$ is a **σ-algebra over Ω** if the following conditions are satisfied:

1. $\varnothing \in \mathcal{A}$,
2. $A \in \mathcal{A} \Rightarrow \bar{A} \in \mathcal{A}$ (i.e. it is *closed under complementation*),
3. $A_i \in \mathcal{A}$ $(i \in \mathbb{N}) \Rightarrow \bigcup_{i=0}^{\infty} A_i \in \mathcal{A}$ (i.e. it is *closed under countable union*).

It is a consequence of this definition that $\Omega \in \mathcal{A}$ is also satisfied. (See exercise.)

Assume an *arbitrary set* Ω and a *σ-algebra* $\mathcal{A}$ over Ω. A function $P : \mathcal{A} \to [0, \infty]$ is called a **measure** if the following conditions are satisfied:

1. $P(\varnothing) = 0$,
2. P is σ-additive.

Let $\mathcal{A}$ be a *σ-algebra* over Ω and $P : \mathcal{A} \to [0, \infty]$ is a *measure*. $(\Omega, \mathcal{A})$ is said to be a **measurable space** and the triple $(\Omega, \mathcal{A}, P)$ is called a **measure space**.

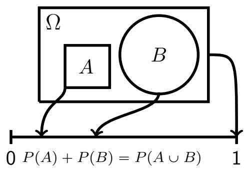
Probability space *

A **probability space** is a triple $(\Omega, \mathcal{A}, P)$, where $(\Omega, \mathcal{A})$ is a *measurable space*, and P is a *measure* such that $P(\Omega) = 1$, called a **probability measure**.

To summarize:
A triple $(\Omega, \mathcal{A}, P)$ is called **probability space**, if

- the **sample space** Ω is *not empty*,
- $\mathcal{A}$ is a **σ-algebra** over Ω, and
- $P : \mathcal{A} \to \mathbb{R}$ is a function with the following properties:
  1. $P(A) \geqslant 0$ for all $A \in \mathcal{A}$
  2. $P(\Omega) = 1$
  3. **σ-additive**: if $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$ and $A_i \cap A_j = \varnothing$ for $i \neq j$, then

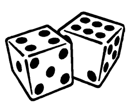
$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n) .$$



Random variables

Example: throwing two "fair" dice *

We have the *sample space* $\Omega = \{(i, j) : 1 \leqslant i, j \leqslant 6\}$ and the *(uniform) probability measure* $P(\{(i, j)\}) = \frac{1}{36}$, where $(\Omega, \mathcal{P}(\Omega), P)$ forms a *(discrete) probability space*.



In many cases it would be more natural to consider *attributes* of the outcomes. A **random variable** is a way of reporting an *attribute* of the *outcome*.

Le us consider the *sum of the numbers showing on the dice*, defined by the **mapping** $X : \Omega \to \Omega'$, $X(i, j) = i + j$, where $\Omega' = \{2, 3, \ldots, 12\}$.

It can be seen that this mapping leads a *probability space* $(\Omega', \mathcal{P}(\Omega'), P')$, such that $P' : \mathcal{P}(\Omega') \to [0, 1]$ is defined as

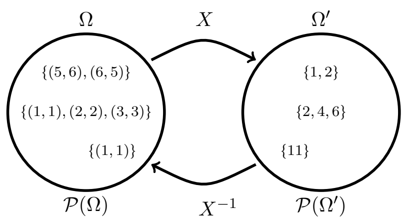
$$P'(A') = P(\{(i, j) : X(i, j) \in A'\}) .$$

<u>*Example*</u>: $P'(\{11\}) = P(\{(5, 6), (6, 5)\}) = \frac{2}{36}$.

Preimage mapping

Let $X : \Omega \to \Omega'$ be an arbitrary *mapping*. The **preimage mapping** $X^{-1} : \mathcal{P}(\Omega') \to \mathcal{P}(\Omega)$ is defined as

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\} .$$



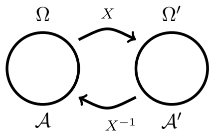
Random variable

Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ *measurable spaces*. A *mapping* $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ is called **random variable**, if

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\} \in \mathcal{A} \ .$$

Let $X : (\Omega, \mathcal{A}) \to (\Omega' \subseteq \mathbb{R}, \mathcal{A}')$ be a *random variable* and P a *measure* over $\mathcal{A}$. Then

$$P'(A') := P_X(A') \triangleq P(X^{-1}(A'))$$



defines a measure over $\mathcal{A}'$. P_X is called the **image measure** of P by X.

Specially, if P is a *probability measure* then P_X is a *probability measure* over $\mathcal{A}'$. (See Exercise.)

Example: throwing two "fair" dice *

We are given two *sample spaces* $\Omega = \{(i, j) : 1 \leqslant i, j \leqslant 6\}$ and $\Omega' = \{2, 3, \ldots, 12\}$. We assume the *(uniform) probability measure* P over $(\Omega, \mathcal{P}(\Omega))$. Let us define a mapping $X : (\Omega, \mathcal{P}(\Omega)) \to (\Omega', \mathcal{P}(\Omega'))$, where $X(i, j) = i + j$.

Question: Is X a random variable?

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\} \in \mathcal{P}(\Omega)$$

is satisfied, since for any $\omega' \in \Omega'$ one can find an $\omega \in \Omega$ such that $X(\omega) = \omega'$. Therefore X is a *random variable*. Moreover, P is a *probability measure*, hence the *image measure*

$$P_X(A') \triangleq P(X^{-1}(A'))$$

is a *probability measure* on $(\Omega', \mathcal{P}(\Omega'))$.

Example: $P_X(\{2, 4, 5\}) = P(X^{-1}(\{2, 4, 5\})) =$
$P(\{(1,1), (1,3), (2,2), (3,1), (1,4), (2,3), (3,2), (4,1)\}) = \frac{8}{36} = \frac{2}{9}$.

Labeling via random variables

In the last lecture we defined the *labeling* L providing a label, taken from a label set $\mathcal{L}$, for each pixel i on an image.

By applying a *random variable*

$$X : \{(r, g, b) \in \mathbb{Z}^3 \mid 0 \leqslant r, g, b \leqslant 255\} \to \mathcal{L}$$

we can model the probability of the labeling for a given pixel as

$$P_X(\text{the given pixel has the label } l) \ .$$

Probability distributions

Probability distribution

Note that a *random variable* is a (measurable) **mapping** from a probability space to a measure space. It is *neither a variable nor random*.

Let $X : (\Omega, \mathcal{A}, P) \to (\Omega' \subseteq \mathbb{R}, \mathcal{A}')$ be a *random variable*. Then the *image measure* P_X of P by X is called **probability distribution**.

We use the notation $p(x)$ for $P(X = x)$, where

$$p(x) := P(X = x) \triangleq P(\{\omega \in \Omega : X(\omega) = x\}) \ .$$

Joint distribution

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be *discrete* random variables, where $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote the values of Y.

We introduce the notation

$$p_{ij} \triangleq P(X = x_i, Y = y_j) \quad i, j = 1, 2, \ldots$$

for the probability of the *events*

$$\{X = x_i, Y = y_j\} := \{\omega \in \Omega : X(\omega) = x_i \text{ and } Y(\omega) = y_j\} \ .$$

These probabilities p_{ij} form a *distribution*, called the **joint distribution** of X and Y.

Remark that

$$\sum_i \sum_j p_{ij} = 1 \ .$$

Marginal distributions

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be *discrete* random variables, where $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote the values of Y.

The *distributions* defined by the probabilities

$$p_i \triangleq P(X = x_i) \quad \text{and} \quad q_j \triangleq P(Y = y_j)$$

are called the **marginal distributions** of X and of Y, respectively.

Let us consider the *marginal distribution* of X. Then

$$p_i = P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij} \ .$$

Similarly, the *marginal distribution* of Y is given by

$$q_j = P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij} \ .$$

Example: marginal distribution *

Consider the problem of *binary segmentation*. Let us define a pixel to be "bright", if all its (RGB) intensities are at least 128, otherwise the given pixel is considered to be "dark".

Assume we are given the following table with probabilities:

	Dark	Bright	
Foreground	0.163	0.006	0.169
Background	0.116	0.715	0.831
	0.279	0.721	1

The marginal distributions of discrete random variables corresponding to the values of {foreground, background} and {dark, bright} are shown in the last column and last row, respectively.

The following also holds

$$\sum_i p_i = \sum_i P(X = x_i) = \sum_i \sum_j P(X = x_i, Y = y_i) = \sum_i \sum_j p_{ij} = 1 \ .$$

Conditional distribution

Let X and Y be *discrete random variables*, where $x_1, x_2, \dots$ denote the values of X and $y_1, y_2, \dots$ denote the values of Y.

The **conditional distribution** of X given Y is defined by

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{\sum_k p_{kj}} = \frac{p_{ij}}{q_j} .$$

Summary *

- A **random variable** $X : (\Omega, \mathcal{A}, P) \to (\Omega' \subseteq \mathbb{R}, \mathcal{A}', P_X)$ is a (measurable) **mapping** from a probability space to a measure space.
- The image measure P_X of P by X is called **probability distribution**.
- The function $F_X : \mathbb{R} \to \mathbb{R}$, $F_X(x) = P(x < X)$ is called **cumulative distribution function** of X.
- Probability distributions
  - Joint distribution
  - Marginal distribution
  - Conditional distribution

Graphical models

Graphical models

Probabilistic graphical models encode a joint $p(\mathbf{x}, \mathbf{y})$ or conditional $p(\mathbf{y} \mid \mathbf{x})$ probability distribution such that given some *observations* $\mathbf{x}$ we are provided with a full probability distribution over all feasible solutions.

The graphical models allow us to encode relationships between a set of random variables using a concise language, by means of a **graph**.

We will use the following notations

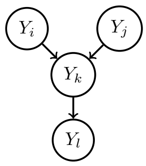
- $\mathcal{V}$ denotes a **set of output variables** (e.g., for pixels) and the corresponding random variables are denoted by Y_i for all $i \in \mathcal{V}$.
- The **output domain** $\mathcal{Y}$ is given by the product of individual variable domains $\mathcal{Y}_i$ (e.g., a single label set $\mathcal{L}$), that is $\mathcal{Y} = \times_{i\in\mathcal{V}} \mathcal{Y}_i$.
- The **input domain** $\mathcal{X}$ is application dependent (e.g., $\mathcal{X}$ is a set of images).
- The **realization** $\mathbf{Y} = \mathbf{y}$ means that $Y_i = y_i$ for all $i \in \mathcal{V}$.
- $G = (\mathcal{V}, \mathcal{E})$ is an (un)directed graph, which encodes the **conditional independence assumption**.

Bayesian networks

Assume a **directed, acyclic** graphical model $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.

The factorization is given as

$$p(\mathbf{Y} = \mathbf{y}) = \prod_{i\in\mathcal{V}} p(y_i \mid \mathbf{y}_{\mathrm{pa}_G(i)}) ,$$



where $p(y_i \mid \mathbf{y}_{\mathrm{pa}_G(i)})$, assuming that $p(y_i \mid \emptyset) \equiv p(y_i)$, is a conditional probability distribution on the parents of node $i \in \mathcal{V}$, denoted by $\mathrm{pa}_G(i)$.

The *conditional independence assumption* is encoded by G that is a variable is conditionally independent of its non-descendants given its parents.

Example:

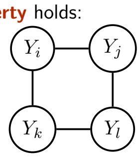
$$\begin{aligned} p(\mathbf{y}) &= p(y_l \mid y_k)\ p(y_k \mid y_i, y_j)\ p(y_i)\ p(y_j) \\ &= p(y_l \mid y_k)\ p(y_k \mid y_i, y_j)\ p(y_i, y_j) = p(y_l \mid y_k)\ p(y_i, y_j, y_k) \\ &= p(y_l \mid y_i, y_j, y_k)\ p(y_i, y_j, y_k) = p(y_i, y_j, y_k, y_l) . \end{aligned}$$

MRF

Markov random field

An *undirected graphical model* $G = (\mathcal{V}, \mathcal{E})$ is called **Markov Random Field** (MRF) if two nodes are *conditionally independent* whenever they are *not connected*. In other words, for any node i in the graph, the **local Markov property** holds:

$$p(Y_i \mid Y_{\mathcal{V}\setminus\{i\}}) = p(Y_i \mid Y_{N(i)}) ,$$



where $N(i)$ is denotes the neighbors of node i in the graph. Alternatively, we can use the following equivalent notation:

$$Y_i \perp\!\!\!\perp Y_{\mathcal{V}\setminus \mathrm{cl}(i)} \mid Y_{N(i)} ,$$

where $\mathrm{cl}(i) = N(i) \cup \{i\}$ is the *closed neighborhood* of i.

Example: $Y_i \perp\!\!\!\perp Y_l \mid Y_j, Y_k \quad \Rightarrow \quad p(y_i \mid y_j, y_k, y_l) = p(y_i \mid y_j, y_k)$, or

$$p(y_l \mid y_i, y_j, y_k) = p(y_l \mid y_j, y_k) .$$

Gibbs distribution

A *probability distribution* $p(\mathbf{y})$ on an *undirected graphical model* $G = (\mathcal{V}, \mathcal{E})$ is called **Gibbs distribution** if it can be factorized into *potential functions*

$$\psi_c(\mathbf{y}_c) > 0$$

defined on *cliques* (i.e. fully connected subgraph) that cover all nodes and edges of G. That is,

$$p(\mathbf{y}) = \frac{1}{Z} \prod_{c\in\mathcal{C}_G} \psi_c(\mathbf{y}_c) ,$$

where $\mathcal{C}_G$ denotes the set of all (maximal) cliques in G and

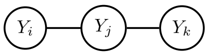
$$Z = \sum_{\mathbf{y}\in\mathcal{Y}} \prod_{c\in\mathcal{C}_G} \psi_c(\mathbf{y}_c) .$$

is the *normalization constant*. Z is also known as **partition function**.

Examples *

$\mathcal{C}_{G_1} = \{\{i\},\{j\},\{k\},\{i,j\},\{j,k\}\}$, hence

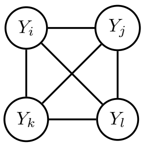
$$p(\mathbf{y}) = \frac{1}{Z}\psi_i(y_i)\psi_j(y_j)\psi_k(y_k)\psi_{ij}(y_i,y_j)\psi_{jk}(y_j,y_k)$$



G_1

$\mathcal{C}_{G_2} = 2^{\{i,j,k,l\}}\backslash\varnothing$ (i.e. all nonempty subsets of $\mathcal{V}_2$)

$$p(\mathbf{y}) = \frac{1}{Z}\prod_{c\in 2^{\{i,j,k,l\}}\backslash\varnothing}\psi_c(\mathbf{y}_c)$$



G_2

$$\begin{aligned}\mathcal{C}_{G_2} = &\{\{i\},\{j\},\{k\},\{l\},\\ &\{i,j\},\{i,k\},\{i,l\},\{j,k\},\{j,l\},\\ &\{i,j,k\},\{i,j,l\},\{i,k,l\},\{j,k,l\},\\ &\{i,j,k,l\}\}\end{aligned}$$

Hammersley-Clifford theorem

Let $G = (\mathcal{V},\mathcal{E})$ be an *undirected graphical model*. The Hammersley-Clifford theorem tells us that the followings are equivalent:

- G is an MRF model.
- The joint probability distribution $p(\mathbf{y})$ on G is a Gibbs-distribution.

An MRF defines a family of **joint probability distributions** by means of an *undirected* graph $G = (\mathcal{V},\mathcal{E})$, $\mathcal{E}\subset\mathcal{V}\times\mathcal{V}$ (there are no self-edges), where the graph encodes *conditional independence assumptions* between the random variables corresponding to $\mathcal{V}$.

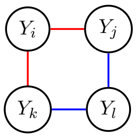
Proof of the Hammersley-Clifford theorem (backward direction) *

Let $\mathrm{cl}(i) = N_i \cup \{i\}$ and assume that $p(\mathbf{y})$ follows *Gibbs-distribution*.

$$p(y_i \mid \mathbf{y}_{N_i}) = \frac{p(y_i,\mathbf{y}_{N_i})}{p(\mathbf{y}_{N_i})} = \frac{\sum_{\mathcal{V}\backslash\mathrm{cl}(i)} p(\mathbf{y})}{\sum_{y_i}\sum_{\mathcal{V}\backslash\mathrm{cl}(i)} p(\mathbf{y})} = \frac{\sum_{\mathcal{V}\backslash\mathrm{cl}(i)}\frac{1}{Z}\prod_{c\in\mathcal{C}_G}\psi_c(\mathbf{y}_c)}{\sum_{y_i}\sum_{\mathcal{V}\backslash\mathrm{cl}(i)}\frac{1}{Z}\prod_{c\in\mathcal{C}_G}\psi_c(\mathbf{y}_c)}\,.$$

Let us define two sets: $\mathcal{C}_i := \{c\in\mathcal{C}_G : i\in c\}$ and $\mathcal{R}_i := \{c\in\mathcal{C}_G : i\notin c\}$. Obviously, $\mathcal{C}_G = \mathcal{C}_i\cup\mathcal{R}_i$ for all $i\in\mathcal{V}$.

Example:



$C_i = \{(i,j),(i,k)\}$
$R_i = \{(j,l),(k,l)\}$

$$\begin{aligned}p(y_i \mid \mathbf{y}_{N_i}) &= \frac{\sum_{\mathcal{V}\backslash\mathrm{cl}(i)}\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)\prod_{d\in\mathcal{R}_i}\psi_d(\mathbf{y}_d)}{\sum_{y_i}\sum_{\mathcal{V}\backslash\mathrm{cl}(i)}\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)\prod_{d\in\mathcal{R}_i}\psi_d(\mathbf{y}_d)}\\ &= \frac{\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)\cdot\sum_{\mathcal{V}\backslash\mathrm{cl}(i)}\prod_{d\in\mathcal{R}_i}\psi_d(\mathbf{y}_d)}{\sum_{y_i}\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)\cdot\sum_{\mathcal{V}\backslash\mathrm{cl}(i)}\prod_{d\in\mathcal{R}_i}\psi_d(\mathbf{y}_d)}\\ &= \frac{\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)}{\sum_{y_i}\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)}\end{aligned}$$

Proof of the Hammersley-Clifford theorem (backward direction) *

$$\begin{aligned}p(y_i \mid \mathbf{y}_{N_i}) &= \frac{\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)}{\sum_{y_i}\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)}\\ &= \frac{\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)}{\sum_{y_i}\prod_{c\in\mathcal{C}_i}\psi_c(\mathbf{y}_c)}\cdot\frac{\prod_{c\in\mathcal{R}_i}\psi_c(\mathbf{y}_c)}{\prod_{c\in\mathcal{R}_i}\psi_c(\mathbf{y}_c)}\\ &= \frac{\prod_{c\in\mathcal{C}_G}\psi_c(\mathbf{y}_c)}{\sum_{y_i}\prod_{c\in\mathcal{C}_G}\psi_c(\mathbf{y}_c)}\\ &= \frac{p(\mathbf{y})}{p(\mathbf{y}_{\mathcal{V}\backslash\{i\}})} = \frac{p(y_i,\mathbf{y}_{\mathcal{V}\backslash\{i\}})}{p(\mathbf{y}_{\mathcal{V}\backslash\{i\}})}\\ &= p(y_i \mid \mathbf{y}_{\mathcal{V}\backslash\{i\}})\,.\end{aligned}$$

Therefore the *local Markov property* holds for any node $i\in\mathcal{V}$.

Binomial theorem *

Reminder: Let $x,y\in\mathbb{R}$ and $n\in\mathbb{N}$, then

$$(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^{(n-k)}y^k\,,$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

We will use the following identity

$$0 = (1-1)^n = \sum_{k=0}^{n}(-1)^k\binom{n}{k}\,.$$

Reminder: A k-**combination** of a set $\mathcal{S}$ is a subset of k distinct elements of $\mathcal{S}$. If $|\mathcal{S}| = n$, then number of k-combinations is equal to $\binom{n}{k}$.

Proof of the Clifford-Hammersley theorem (forward direction) *

We define a *candidate* potential function for any subset $s\subseteq\mathcal{V}$ as follows:

$$f_s(\mathbf{Y}_s = \mathbf{y}_s) = \prod_{z\subseteq s} p(\mathbf{y}_z,\mathbf{y}^*_{\bar z})^{(-1^{|s|-|z|})}$$

where $p(\mathbf{y}_z,\mathbf{y}^*_{\bar z})$ is a strictly positive distribution and $\mathbf{y}^*_{\bar z}$ means an (arbitrary but fixed) *default realization* of the variables $\mathbf{Y}_{\bar z}$ for the set $\bar z = \mathcal{V}\backslash\{z\}$. We will use the following notation: $q(\mathbf{y}_z) := p(\mathbf{y}_z,\mathbf{y}^*_{\bar z})$.

Assume that the *local Markov property* holds for any node $i\in\mathcal{V}$.
First, we show that, if s is not a clique, then $f_s(\mathbf{y}_s) = 1$. For this sake, let us assume that s is **not** a clique, therefore there exist $a,b\in s$ that are not connected to each other. Hence

$$f_s(\mathbf{Y}_s = \mathbf{y}_s) = \prod_{z\subseteq s} q(\mathbf{y}_z)^{(-1^{|s|-|z|})} = \prod_{w\subseteq s\backslash\{a,b\}}\left(\frac{q(\mathbf{y}_w)\;q(\mathbf{y}_{w\cup\{a,b\}})}{q(\mathbf{y}_{w\cup\{a\}})\;q(\mathbf{y}_{w\cup\{b\}})}\right)^{(-1^*)}\,,$$

where -1^* meaning either 1 or -1 is not important at all.

Proof of the Clifford-Hammersley theorem (forward direction) *

We have

$$f_s(\mathbf{Y}_s = \mathbf{y}_s) = \prod_{w\subseteq s\backslash\{a,b\}}\left(\frac{q(\mathbf{y}_w)\;q(\mathbf{y}_{w\cup\{a,b\}})}{q(\mathbf{y}_{w\cup\{a\}})\;q(\mathbf{y}_{w\cup\{b\}})}\right)^{(-1^*)}\,.$$

$$\begin{aligned}\frac{q(\mathbf{y}_w)}{q(\mathbf{y}_{w\cup\{a\}})} &\overset{\Delta}{=} \frac{p(\mathbf{y}_w,y^*_a,y^*_b,y^*_{\bar w\backslash\{a,b\}})}{p(y_a,\mathbf{y}_w,y^*_b,y^*_{\bar w\backslash\{a,b\}})} = \frac{p(y^*_a \mid \mathbf{y}_w,y^*_b,y^*_{\bar w\backslash\{a,b\}})}{p(y_a \mid \mathbf{y}_w,y^*_b,y^*_{\bar w\backslash\{a,b\}})}\\ &\overset{a\perp\!\!\!\perp b}{=} \frac{p(y^*_a \mid \mathbf{y}_w,y_b,y^*_{\bar w\backslash\{a,b\}})}{p(y_a \mid \mathbf{y}_w,y_b,y^*_{\bar w\backslash\{a,b\}})} = \frac{p(\mathbf{y}_w,y_b,y^*_{\bar w\backslash\{b\}})}{p(\mathbf{y}_w,y_a,y_b,y^*_{\bar w\backslash\{a,b\}})} \overset{\Delta}{=} \frac{q(\mathbf{y}_{w\cup\{b\}})}{q(\mathbf{y}_{w\cup\{a,b\}})}\,.\end{aligned}$$

Therefore

$$f_s(\mathbf{Y}_s = \mathbf{y}_s) = \prod_{w\subseteq s\backslash\{a,b\}} 1^{(-1^*)} = 1 \quad\text{for all } s\notin\mathcal{C}_G\,.$$

Proof of the Clifford-Hammersley theorem (forward direction) *

We also show that $\prod_{s\subseteq\mathcal{V}} f_s(\mathbf{y}_s) = p(\mathbf{y})$. Consider any $z\subset\mathcal{V}$ and the corresponding factor $q(\mathbf{y}_z)$. Let $n = |\mathcal{V}|-|z|$.

- $q(\mathbf{y}_z)$ occurs in $f_z(\mathbf{y}_z)$ as $q(\mathbf{y}_z)^{(-1^0)} = q(\mathbf{y}_z)$.
- $q(\mathbf{y}_z)$ also occurs in the functions $f_s(\mathbf{y}_s)$ for $s\subseteq\mathcal{V}$, where $|s| = |z|+1$. The number of such factors is $\binom{n}{1}$. The exponent of those factors is $-1^{|s|-|z|} = -1^1 = -1$.
- $q(\mathbf{y}_z)$ occurs in the functions $f_s(\mathbf{y}_s)$ for $s\subseteq\mathcal{V}$, where $|s| = |z|+2$. The number of such factors is $\binom{n}{2}$ and their exponent is $-1^{|s|-|z|} = 1$.

If we multiply **all** those factors, we get

$$\begin{aligned}q(\mathbf{y}_z)^1\;q(\mathbf{y}_z)^{-\binom{n}{1}}\;q(\mathbf{y}_z)^{\binom{n}{2}}\;\ldots\;q(\mathbf{y}_z)^{(-1^n)\binom{n}{n}} &= q(\mathbf{y}_z)^{\binom{n}{0}-\binom{n}{1}+\binom{n}{2}+\cdots+(-1)^n\binom{n}{n}}\\ &= q(\mathbf{y}_z)^0 = 1\,.\end{aligned}$$

So all factors cancel themselves out except of $q(\mathbf{y})$, that is $p(\mathbf{y}) = \prod_{c\subseteq\mathcal{C}_G} f_c(\mathbf{y}_c)$. □

Factor graph

Factor graph

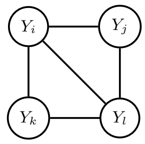
Factor graphs are *undirected graphical models* that **make the factorization explicit** of the probability function.
A factor graph $G = (\mathcal{V}, \mathcal{F}, \mathcal{E}')$ consists of

- variable nodes V (◯) and factor nodes $\mathcal{F}$ (■),
- edges $\mathcal{E}' \subseteq V \times \mathcal{F}$ between variable and factor nodes
- $N : \mathcal{F} \to 2^V$ is the *scope of a factor*, defined as the **set of neighboring variables**, i.e. $N(F) = \{i \in V : (i, F) \in \mathcal{E}\}$.

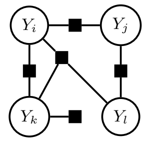
A family of distribution is defined that factorizes as:

$$p(\mathbf{y}) = \frac{1}{Z} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}_{N(F)}) \quad \text{with} \quad Z = \sum_{\mathbf{y} \in \mathcal{Y}} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}_{N(F)}) .$$

Each factor $F \in \mathcal{F}$ connects a subset of nodes, hence we write $\mathbf{y}_F = \mathbf{y}_{N(F)} = (y_{v_1}, \ldots, y_{v_{|F|}})$.

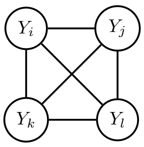


MRF

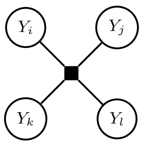


Factor graph

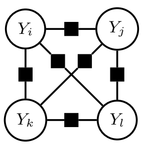
Examples *



An exemplar MRF



$p_1(\mathbf{y}) = \frac{1}{Z_1} \psi_{ijkl}(y_i, y_j, y_k, y_l)$



$$p_2(\mathbf{y}) = \frac{1}{Z_2} \psi_{ij}(y_i, y_j) \cdot \psi_{ik}(y_i, y_k) \cdot \psi_{il}(y_i, y_l) \cdot \psi_{jk}(y_j, y_k) \cdot \psi_{jl}(y_j, y_l) \cdot \psi_{kl}(y_k, y_l)$$

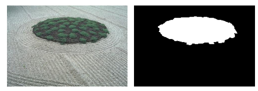
Factor graphs are universal, explicit about the factorization, hence it is easier to work with them.

Summary *

- A **graphical models** allow us to *encode relationships between a set of random variables* using a concise language, by means of a graph.
- A **Bayesian network** is a *directed acyclic* graphical model $G = (\mathcal{V}, \mathcal{E})$, where *conditional independence assumption* is encoded by G that is a variable is conditionally independent of its non-descendants given its parents.
- An **MRF** defines a family of **joint probability distributions** by means of an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where the graph encodes conditional independence assumptions between the random variables.
- **Factor graphs** are universal, explicit about the factorization, hence it is easier to work with them.

In the **next lecture** we will learn about

- Conditional random field (CRF)
- Inference for graphical models
- Binary image segmentation
- `EM algorithm`


Source: Berkeley Segmentation Dataset

Literature *

Probability theory

1. Marek Capiński and Ekkerhard Kopp. *Measure, Integral and Probability*. Springer, 1998
2. Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009

Graphical models

3. Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009
4. Sebastian Nowozin and Christoph H. Lampert. Structured prediction and learning in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3–4), 2010
5. J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished, 1971
6. Samson Cheung. Proof of Hammersley-Clifford theorem. Unpublished, February 2008