

Recitation 2: Markov Networks

Teaching Assistant: Eitan Richardson

Recommended reading:

- PGM Book, Chapter 4 [1]

2.1 Reminder – Markov Networks

2.1.1 Motivation

A *Bayesian Network* assigns a directionality to each influence between random variables. In some cases (e.g. the *misconception example* in the PGM book – example 3.8), the directionality creates conflicts and an undirected model is more suitable.

The *Markov Network* is an undirected graph in which an edge indicates a direct influence between a pair of RVs. The joined probability (over all RVs) is factorized as a (normalized) product of *clique potentials* or *factors* defined over cliques in the graph (i.e. every pair of variables that are in a scope of the same factor, are connected by an edge).

Intuitively, each factor defines the *affinity* (or *compatibility*) between values of the variables in its scope. For example, if the factor is relatively high in a specific assignment of values to RVs in a clique RV, this combination is more likely to appear together.

☞ The factor values are not limited to $[0, 1]$

☞ A single factor does not represent the marginal joint distribution of its scope – it depends on other factors too.

2.1.2 Formal Definitions

Definition 2.1 Factor

A (non-negative) factor is a function $\phi(\mathbf{D}) : \text{Val}(\mathbf{D}) \mapsto \mathbb{R}^+$, where \mathbf{D} is a set of RVs (the factor's scope).

Definition 2.2 Factor Product

A product of two factors $\phi_1 \times \phi_2$ is a new factor ψ over the union of their scopes. The new factor ψ is computed by multiplying matching rows in ϕ_1 and ϕ_2 .

Definition 2.3 Gibbs Distribution

A Gibbs distribution over a set of RVs $\{X_1, \dots, X_n\}$, parametrized by a set of factors $\{\phi_1(\mathbf{D}_1), \dots, \phi_m(\mathbf{D}_m)\}$ is defined as:

$$P_{\Phi}(X_1, \dots, X_n) = \frac{1}{Z} \prod_{j=1}^m \phi_j(\mathbf{D}_j),$$

¹Original LaTeX template courtesy of UC Berkeley.

where the partition function Z is defined as:

$$Z = \sum_{X_1, \dots, X_n} \left[\prod_{j=1}^m \phi_j(\mathbf{D}_j) \right]$$

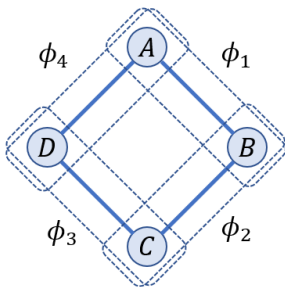
Definition 2.4 *Markov Network Factorization*

A Gibbs distribution with factors $\{\phi_1(\mathbf{D}_1), \dots, \phi_m(\mathbf{D}_m)\}$ factorizes over a Markov Network \mathcal{H} if each set \mathbf{D}_i is a clique in \mathcal{H} (i.e. the induced sub-graph is complete).

☞ The factorization is over cliques and not necessarily over maximal cliques.

2.1.3 Example – Multiplying Factors

We have the following MN (similar to the "misconception example"), with factors defined over pairs of RVs.



A	B	$\phi_1(A, B)$	B	C	$\phi_2(B, C)$	
0	0	100	0	0	1	
0	1	1	0	1	5	...
1	0	1	1	0	5	
1	1	10	1	1	0	

The joint probability for a specific assignment is:

$$P(A = 0, B = 1, C = 0, D = 1) = \frac{1}{Z} \phi_1(0, 1) \phi_2(1, 0) \phi_3(0, 1) \phi_4(1, 0) = \frac{1}{Z} 1 \times 5 \times \dots$$

Where Z normalizes the probability such that the sum over all combinations is 1.

If we wanted (for reasons we'll see later on in the course) to create a new factor $\psi = \phi_1 \times \phi_2$ we would get:

A	B	C	$\psi(A, B, C)$
0	0	0	100
0	0	1	500
0	1	0	5
0	1	1	0
1	0	0	1
1	0	1	5
1	1	0	50
1	1	1	0

2.2 Example – Image Denoising

Many Computer Vision problems can be posed as Markov Networks (also known as *Markov Random Fields*). Typically each pixel is defined as a RV.

In image denoising, the task is to find the most probable clean image given an observed noisy one. We define the Markov Network as a regular graph with a node for each pixel and edges between neighboring pixels. We parameterize the model as a *pairwise Markov Network* with *node potentials* – factors defined over single RVs, and *edge potentials* – factors defined over pairs of neighboring RVs.

The node factors ϕ_i encourages the target pixel value Y_i to be close to the observed pixel value x_i . The edge factors ϕ_{ij} encourages neighboring pixels to have a similar value (smoothness term).

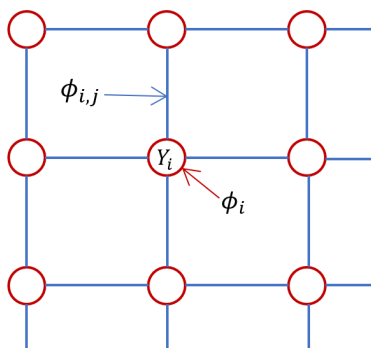


Figure 2.1: Pairwise Markov Network (Markov Random Field) for Image Denoising

Possible parameterization:

$$\begin{aligned} \phi_i &= e^{-(Y_i - x_i)^2} \\ \phi_{ij} &= e^{-\min(d_{\max}^2, (Y_i - Y_j)^2)} \\ P_{\Phi}(Y_1, \dots, Y_n) &= \frac{1}{Z} \prod_{i=1}^n \phi_i(Y_i) \prod_{(i,j) \in \mathcal{E}} \phi_{ij}(Y_i, Y_j) \end{aligned}$$

It is common to use negative log-space parameterization:

$$\begin{aligned} \epsilon_i &= (Y_i - x_i)^2 \\ \epsilon_{ij} &= \min(d_{\max}^2, (Y_i - Y_j)^2) \end{aligned}$$

The *energy function* is defined as:

$$E(Y_1, \dots, Y_n) = \sum_{i=1}^n \epsilon_i(Y_i) + \sum_{(i,j) \in \mathcal{E}} \epsilon_{ij}(Y_i, Y_j)$$

So that:

$$P_{\Phi}(Y_1, \dots, Y_n) = \frac{1}{Z} e^{-E(Y_1, \dots, Y_n)}$$

The task of maximizing P_{Φ} is equivalent to minimizing the energy function:

$$\arg \min_{y_1, \dots, y_n} E(y_1, \dots, y_n)$$

☞ We don't care about the partition function, since we only want to find the arg min.

2.2.1 Conditional Random Fields

A *Conditional Random Field* encodes a *conditional distribution* $P(\mathbf{Y} \mid \mathbf{X})$ of target variables \mathbf{Y} given (a disjoint set of) *observed variables* \mathbf{X} .

The CRF mechanism is similar to MRF (MN), except that the partition function Z is calculating by summing only over the target variables \mathbf{Y} – the value of Z depends on the assignment to set of observed variables \mathbf{X} .

The image denoising problem can be described by the following CRF (with the same solution formulation we defined above):

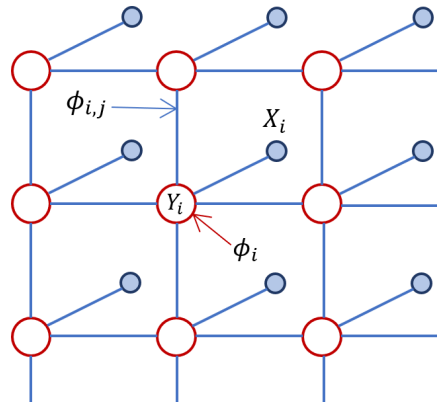


Figure 2.2: A Conditional Random Field (CRF)

2.3 Independence in Markov Networks

Like in BN, the MN graph structure encodes different *conditional independencies* between sets of RVs.

2.3.1 Global Independences

Definition 2.5 Active Path

A path between two RVs is active if no variable along the path is observed.

☞ Unlike in BN, the separation definition in MN is simple and contains only one case.

Definition 2.6 Separation

Two sets of variables \mathbf{X} and \mathbf{Y} are separated given a third set \mathbf{Z} , denoted $\text{Sep}_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$ if there is no active path between any $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given \mathbf{Z} .

Definition 2.7 Global Independencies in a Markov Network

The global independence associated with a Markov Network \mathcal{H} is defined as:

$$\mathcal{I}(\mathcal{H}) = \{(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) : \text{Sep}_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\}$$


☞ Unlike in BN, the global independence is monotonic in \mathbf{Z} i.e. as we observe more variables (increase \mathbf{Z}), the set of independencies can only grow.

2.3.1.1 Soundness of Global Independence

Theorem 2.8 *Soundness of MN Global Independence*

P is a Gibbs distribution that factorizes over $\mathcal{H} \implies \mathcal{H}$ is an I-map for P .

(In other words, the global independencies hold in any Gibbs distribution that factorizes over our MN.)

 Proof shown in class

The opposite direction is only true for *positive* distributions:

Theorem 2.9 *Hammersley-Clifford*

\mathcal{H} is an I-map for a positive distribution $P \implies P$ is a Gibbs distribution that factorizes over \mathcal{H} .

Theorem 2.10 *Completeness*

X and Y are not separated given \mathbf{Z} in $\mathcal{H} \implies X$ and Y are conditionally dependent given \mathbf{Z} in some Gibbs distribution P that factorizes over \mathcal{H} .

Proof: We need to construct such a distribution:

An active (unblocked by \mathbf{Z}) trail must exist in the graph between X and Y . We choose some minimal trail (without cycles):

$$(X = U_1) - U_2 - \dots - U_{k-1} - (U_k = Y) \quad \forall i, U_i \notin \mathbf{Z}$$

Every edge along the trail is part of some clique:

$$\forall i \exists C_i \text{ s.t. } \{U_i, U_{i+1}\} \subseteq C_i$$

Lets define the clique potentials ϕ_i :

$$\phi_i = \begin{cases} W \gg 1, & U_i = U_{i+1} \\ 1, & \text{otherwise} \end{cases}$$

(Other variables in the clique don't effect the factor value).

The above construction is valid since no two edges along our trail belong to the same clique (otherwise, there would be a shortcut edge and the path will not be minimal).

We define all other factors in \mathcal{H} to be uniform 1.

We can easily see (left as an exercise) that in this setting, $P(U_1, \dots, U_k) = \prod_{i=1}^{k-1} \phi_i$ and that $X = U_1$ and $Y = U_k$ are dependent in this joint distribution. ■

2.4 Additional Markov Network Independencies

In addition to the global independencies $\mathcal{I}(\mathcal{H})$, the Markov Network structure encodes more local independencies, which we'll discuss now.

Definition 2.11 *Pairwise Independencies*

Two variables that are not directly connected by an edge, are conditionally independent given all other variables in the MN:

$$\mathcal{I}_P(\mathcal{H}) = \{(X \perp Y \mid \mathcal{X} - \{X, Y\}) : X - Y \notin \mathcal{H}\}$$

Definition 2.12 *Markov Blanket*

$MB_{\mathcal{H}}(X)$ is the set of all neighbors of X in \mathcal{H} (also denoted $NE_{\mathcal{H}}(X)$)

Definition 2.13 *Local (Markov Blanket) Independencies*

$$\mathcal{I}_{LM}(\mathcal{H}) = \{(X \perp \mathcal{X} - \{X\} - MB_{\mathcal{H}}(X) \mid MB_{\mathcal{H}}(X)) : X \in \mathcal{X}\}$$

How are the three sets of independencies we define related?

Proposition 2.14

$$P \models \mathcal{I}(\mathcal{H}) \implies P \models \mathcal{I}_{LM}(\mathcal{H}) \implies P \models \mathcal{I}_P(\mathcal{H})$$

Sometimes written as: $\mathcal{I}_P(\mathcal{H}) \subseteq \mathcal{I}_{LM}(\mathcal{H}) \subseteq \mathcal{I}(\mathcal{H})$


Proof: $P \models \mathcal{I}(\mathcal{H}) \implies P \models \mathcal{I}_{LM}(\mathcal{H})$

It is sufficient to show that any node X is *separated* from all other nodes given its neighbors:

$$\forall X, Y \in \mathcal{X}, X - Y \notin \mathcal{H} : sep_{\mathcal{H}}(X; Y \mid NE_{\mathcal{H}}(X))$$

This is trivially correct since any path between X and Y must pass through one of X 's neighbors. ■

Theorem 2.15 *For positive distributions, the three types of independencies are equivalent.*

 *Proof shown in class*

2.5 From Bayesian to Markov Network

We've seen a set of independencies that can't be expressed by a BN and require the undirected MN representation. Similarly, there are independencies that can't be expressed by a MN. For example, a v-structure: $X \rightarrow Z \leftarrow Y$. The BN graph structure defines: $P_{\mathcal{B}} \models (X \perp Y)$, but $P_{\mathcal{B}} \not\models (X \perp Y \mid Z)$ (the path between X and Y is active given Z). The same path is blocked (given Z) in the undirected graph $X-Y-Z$, so it encodes $(X \perp Y \mid Z)$, an independency that does not hold in the distribution. This means that to make the MN an I-map, we need to add an edge between X and Y , encoding no independencies.

2.5.1 Gibbs distribution induced by a BN

A distribution factorized over a BN is a Gibbs distribution, in which the factors ϕ_i are the locals CDPs $P_i(X_i \mid Pa_{X_i}^{\mathcal{G}})$ with the scopes $X_i, Pa_{X_i}^{\mathcal{G}}$.

The partition function for the above Gibbs distribution is 1, since the product of factors is already normalized.

2.5.2 The Moralized Graph

Definition 2.16 *Moralized Graph*

The Moral Graph $\mathcal{M}[\mathcal{G}]$ of a BN \mathcal{G} over \mathcal{X} is the underlying undirected graph of \mathcal{G} with added edges between all pairs of nodes with a common child (the parents must be married...)

Theorem 2.17 For any distribution $P_{\mathcal{B}}$ factorizing over a BN \mathcal{G} , $\mathcal{M}[\mathcal{G}]$ is an I-map of $P_{\mathcal{B}}$.

Theorem 2.18 $\mathcal{M}[\mathcal{G}]$ is a minimal I-map for the BN \mathcal{B}

The first theorem claims that the Moralized Graph is an I-map for any distribution that factorizes over the BN. The second claim refers instead to the set of independencies encoded by the BN structure and claims that the Moralized Graph is a minimal I-map for this set.

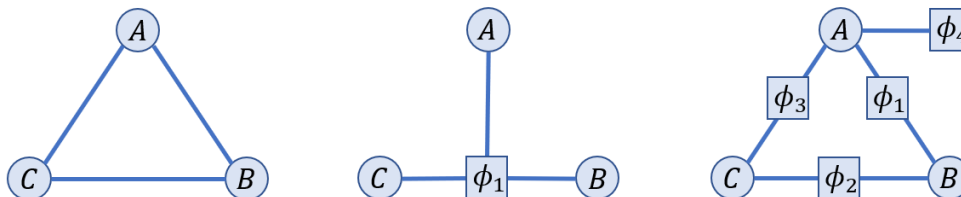
Proof: (Proof sketch)

The MN is constructed as a minimal I-map based on Markov Blankets in the BN \mathcal{G} (as defined in PS1). ■

2.6 Factor Graphs

The Markov Networks graph structure does not define over which cliques the factors are defined. To overcome this ambiguity an alternative representation called *Factor Graph* has explicit nodes for the factors. Edges exist only between factor nodes to RV nodes (which are part of the factor).

The example below shows two possible Factor Graphs for the MN on the left:



The Factor Graph in the middle represents the case in which there is one factor over the maximal clique – the entire graph.

On the right we have pairwise factors plus a single node factor for the RV A .

References

[1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.