# Probabilistic Graphical Models

Tianyu Du

June 5, 2020

# 1 Graphical Representations

## 1.1 Factors

**Definition 1.1.** Let $X_1, X_2, \cdots, X_k$ be a set of random variables, then a **factor** $\phi$ is a mapping from values of these random variables to $\mathbb{R}$.

$$\phi : Val(X_1, X_2, \cdots, X_k) \to \mathbb{R} \tag{1}$$

The set of random variables $\{X_1, X_2, \cdots, X_k\}$ is defined as the **scope** of $\phi$.

**Remark 1.1.** In principle, a factor can take any value in $\mathbb{R}$. However, in practice, we restrict our considerations to factors with positive ranges only.

**Definition 1.2.** Let $\phi_1$ and $\phi_2$ be two factors with scopes $\{A, B\}$ and $\{B, C\}$. Then the **factor product** $\phi_1 \times \phi_2$ is a factor with scope $\{A, B, C\}$ defined as

$$\phi_1 \cdot \phi_2(a, b, c) = \phi_1(a, b) \cdot \phi_2(b, c) \tag{2}$$

**Definition 1.3.** Let $\phi$ be a factor with scope $\{A, B, C\}$, then **marginalizing $C$ from $\phi$** results in a factor $\phi'$ with scope $\{A, B\}$ defined as the following:

$$\phi'(a, b) = \sum_{c \in Val(C)} \phi(a, b, c) \tag{3}$$

**Definition 1.4.** The **factor reduction** operation restricts $\phi(A, B, C)$ to take only a specific value

of $C = c$, and results in a factor $\phi'$ with scope $\{A, B\}$.

$$\phi'(a, b) = \phi(a, b, c) \tag{4}$$

## 1.2  Semantics and Factorization

**Definition 1.5.** A **Bayesian network** consists of (i) a directed acyclic graph (DAG) $G$ whose nodes correspond to random variables $X_1, \cdots, X_n$ (ii) and a conditional probability distribution $P(X_i | Par_G(X_i))$ for each node $X_i$. The joint distribution is defined as the factorization

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | \mathrm{Par}_G(X_i)) \tag{5}$$

**Definition 1.6.** Let $G$ be a graph over $X_1, \cdots, X_n$, then the joint probability $P$ **factorizes** over $G$ if and only if

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | \mathrm{Par}_G(X_i)) \tag{6}$$

## 1.3  Pass of Influences in Bayesian Networks

**Definition 1.7.** A path $X_1 - \cdots - X_k$ in Bayesian network $G$ is **active** if there is no explaining-away structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ in it.

**Definition 1.8.** Let $Z \subseteq V_G$ be a set of random variables in the Bayesian network, then a path $X_1 - \cdots - X_k$ in $G$ is **active conditioned on** $Z$ if

1. for all explaining-away structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ in the path, $X_i$ or some decedents of $X_i$ are in $Z$,

2. <u>and</u> no other node in the path is in $Z$.

**Definition 1.9.** Let $X, Y, Z \subseteq V_G$, if there is no path from $X$ to $Y$ is active conditioned on $Z$, then $X$ and $Y$ are **d-separated** by $Z$ in graph $G$ denoted as d-sep$_G(X, Y | Z)$.

## 1.4  Independencies and Factorizations

**Definition 1.10.** Let $X, Y, Z$ be random variables with distribution $P$, then $X \perp\!\!\!\perp Y$ if and only if $P(X, Y) = P(X)P(Y)$, $X \perp\!\!\!\perp Y | Z$ if and only if $P(X, Y | Z) = P(X|Z)P(Y|Z)$.

**Proposition 1.1.** Let $X, Y, Z$ be random variables with distribution $P$, then $X \perp\!\!\!\perp Y$ if and only if $P(X, Y)$ factorizes as the following

$$P(X, Y) \propto \phi_1(X)\phi_1(Y) \tag{7}$$

and $X \perp\!\!\!\perp Y | Z$ if and only if $P(X, Y, Z)$ factorizes as

$$P(X, Y, Z) \propto \phi_1(X, Z)\phi_1(Y, Z) \tag{8}$$

*Proof.* Relation (7) follows the definition immediately. Suppose $X \perp\!\!\!\perp Y | Z$, then

$$P(X, Y | Z) = P(X | Z)P(Y | Z) \tag{9}$$
$$\iff P(X, Y, Z) = P(X | Z)P(Y | Z)P(Z) \tag{10}$$
$$P(X, Y, Z) \propto P(X | Z)P(Z)P(Y | Z)P(Z) \tag{11}$$
$$= P(X, Z)P(Y, Z) \tag{12}$$
$$= \phi_1(X, Z)\phi_1(Y, Z) \tag{13}$$

$\blacksquare$

**Theorem 1.1** (Factorization $\implies$ Independence)**.** If $P$ factorizes over $G$, and d-sep$_G(X, Y | Z)$ then $P$ satisfies $(X \perp\!\!\!\perp Y | Z)$.

**Theorem 1.2** (Causal Markov Condition)**.** For any random variable $X_i$ in the Bayesian network, $X_i$ is d-separated from all its non-descendants by $\text{Par}_G(X_i)$.

**Corollary 1.1.** If $P$ factorizes over $G$, then in $P$, any variable is independent of its non-descendants given its parents.

**Definition 1.11.** Let $\mathcal{I}(G)$ denote the collection of independencies implicitly encoded by d-separations in graph $G$,

$$\mathcal{I}(G) := \{(X \perp\!\!\!\perp Y | Z) : X, Y, Z \in V \ s.t. \ \text{d-sep}_G(X, Y | Z)\} \tag{14}$$

If a distribution $P$ over $V$ satisfies all independencies in $\mathcal{I}(G)$, then we say that $G$ is an **I-map** (independency map) of $P$.

That is, the I-map of distribution $P$ is a graphical representation of all (and probably more) independencies of $P$.

**Example 1.1.** Let $P$ be a probability distribution and let $G$ be an I-map for $P$. Let $\mathcal{I}(P)$ and $\mathcal{I}(G)$ denote sets of independencies in $P$ and $G$. Suppose $G$ is a I-map of $P$, then all independencies encoded in $G$ are satisfied by $P$, therefore,

$$\mathcal{I}(G) \subseteq \mathcal{I}(P) \tag{15}$$

**Example 1.2.** The I-map can be used for two graphs as well. $G_1$ is a I-map of $G_1$ if $\mathcal{I}(G_1) \subseteq \mathcal{I}(G_2)$. That is, $G_1$ is an I-map of $G_2$ if it does not make independence assumptions that are not true in $G_2$.

**Theorem 1.3** (Independence $\implies$ Factorization)**.** If $G$ is an I-map for $P$, that is, $P$ adheres all independencies encoded in $G$, then $P$ factorizes over $G$.

## 1.5 Template Models

**Definition 1.12.** A **template variable** $X(U_1, \cdots, U_k)$ is instantiated (duplicated) multiple times in a graph. **Template models** are languages that specify how ground variables (i.e., instantiations of template variables) inherit dependency model from template.

**Notation 1.1.** Let $X^{(t)}$ denote the variable at time $t\Delta$, where $\Delta$ is the time granularity in the discrete timeline. Let $X^{(t:t')} = \{X^{(t)}, {}^{(t+1)}, \cdots, X^{(t')}\}$ denote the set of variables over a period of time.

**Definition 1.13.** A Bayesian network is said to satisfy the **Markov assumption** if

$$X^{(t+1)} \perp\!\!\!\perp X^{(0:t-1)} | X^{(t)} \tag{16}$$

When Markov assumption holds, we may express the joint distribution of all $X$ as

$$P(X^{(0:T)}) = P(X^{(0)}) \prod_{t=0}^{T-1} P(X^{(t+1)} | X^{(t)}) \tag{17}$$

**Definition 1.14.** A series of random variables $X^{(0)}, X^{(1)}, \cdots, X^{(T)}$ satisfies the **time invariance**

4

assumption if there exists a template probability model $P(X'|X)$ such that for all $t$,

$$P(X^{(t+1)}|X^{(t)}) = P(X'|X) \tag{18}$$

**Definition 1.15.** A **2-time-slice Bayesian network** (2TNB) over $X_1, \cdots, X_n$ (that is, $n$ random variables for each time step) is specified as a Bayesian network fragment such that

- The nodes include $X_1', \cdots, X_n'$ and a subset of $X_1, \cdots, X_n$,

- and only the nodes $X_n', \cdots, X_n'$ have parents and a conditional probability distribution.

Further, the 2TBN defines a conditional distribution

$$P(X'|X) = \prod_{i=1}^{n} P(X_i'|\mathrm{Par}(X_i')) \tag{19}$$

**Definition 1.16.** A **dynamic Bayesian network** (DNB) over $X_1, \cdots, X_n$ is defined by

- a 2TNB, $\mathrm{BN}_\rightarrow$, over $X_1, \cdots, X_n$,

- and a Bayesian network, $\mathrm{BN}^{(0)}$, over $X_1^{(0)}, \cdots, X_n^{(0)}$.

**Definition 1.17.** For a trajectory over $0, \cdots, T$, the **ground (unrolled) network** of a DNB is a model such that

- the dependency model for $X_1^{(0)}, \cdots, X_n^{(0)}$ is copied from $\mathrm{BN}^{(0)}$,

- and the dependency model for $X_1^{(t)}, \cdots, X_n^{(t)}$ is copied from $\mathrm{BN}_\rightarrow$.

## 1.6 Plate Models

We can use plate models to represent repetitions.

**Example 1.3.** Let $O_t \overset{i.i.d.}{\sim} \mathrm{Bernoulli}(\theta)$, and $\{O_t\}_t$ is simply a set of coin tosses outcomes. Such scenario can be modelled as in Figure 1. Figure 2 illustrates an equivalent representation of this plate model.
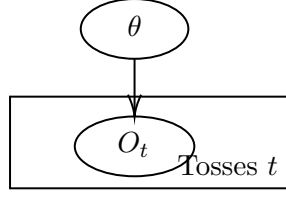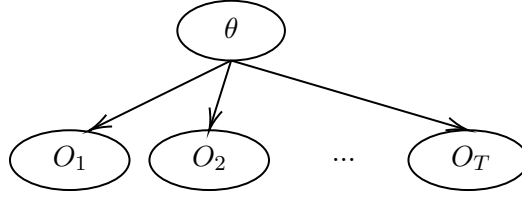
Figure 1: A plate model for Bernoulli trails



Figure 2: A plate model for Bernoulli trails



**Remark 1.2.** Parameters outside the plate are often omitted.

**Example 1.4** (Nested Plate)**.**

**Example 1.5** (Overlapping Plate)**.**

**Example 1.6** (Collective Inference)**.**

**Definition 1.18.** A **plate dependency model** consists of a template variable $A(U_1, \cdots, U_k)$ and template parents $B_1(\mathbf{U}_1), \cdots, B_m(\mathbf{U}_m)$, where $\mathbf{U}_k \subseteq \{U_1, \cdots, U_k\}$. The conditional probability distribution in this model is $P(A|B_1, \cdots, B_m)$.

Figure 3: Plate dependency model
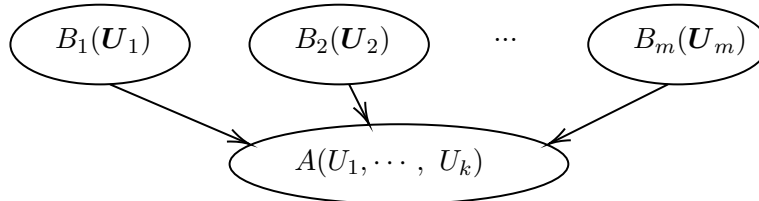


**Definition 1.19.** The concrete instantiation (**ground network**) of a plate dependency model consists of instantiations $u_1, \cdots, u_k$ of $U_1, \cdots, U_k$.

## 1.7 Local Structures

**Definition 1.20.** A **general conditional probability distribution** $P(X|Y_1, \cdots, Y_k)$ specifies distribution over $X$ for each realization of $Y_1, \cdots, Y_k$. Any factor $\phi$ with scope $\{X, Y_1, \cdots, Y_K\}$

satisfying

$$\sum_x \phi(x, y_1, \cdots, y_k) = 1 \quad \forall y_1, \cdots, y_k \tag{20}$$

defines a valid general CPD.

**Definition 1.21** (Context-Specific Independence)**.** Let $P \in \Delta(\mathcal{X})$, let $X, Y \in \mathcal{X}$ and $\mathbf{Z}, \mathbf{C} \subseteq \mathcal{X}$. Let $\mathbf{c}$ be a realization of random variables $\mathbf{C}$. Then $X$ is said to be independent from $Y$ given $Y$ in the context $c$, denoted as $(X \perp\!\!\!\perp_c Y | \mathbf{Z}, \mathbf{c})$, if

$$P(X, Y | \mathbf{Z}, \mathbf{c}) = P(X | \mathbf{Z}, \mathbf{c}) P(Y | \mathbf{Z}, \mathbf{c}) \tag{21}$$

## 1.8    Pairwise Markov Networks

**Definition 1.22.** A **pairwise Markov network** is a undirected graph whose nodes are random variables $X_1, \cdots, X_n$ and each edge $(X_i, X_j)$ is associated with a factor $\phi_{ij}(X_i, X_j)$.

Figure 4: A simple pairwise Markov network
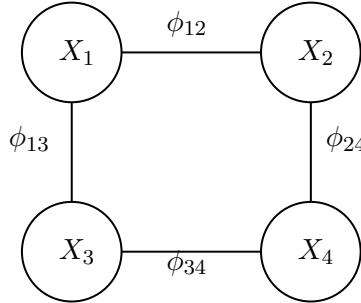


## 1.9    General Gibbs Distributions

**Definition 1.23.** A **Gibbs distribution** over random variables $X_1, \cdots, X_n$ is specified by a set of general factors, $\Phi = \{\phi_i(D_i)\}_{i=1}^k$, where each $D_i \subseteq \{X_1, \cdots, X_n\}$. The corresponding unnormalized probability and partition function are

$$\tilde{P}_\Phi(X_1, \cdots, X_n) = \prod_{i=1}^k \phi_i(D_i) \tag{22}$$

$$Z_\Phi = \sum_{X_1, \cdots, X_n} \tilde{P}_\Phi(X_1, \cdots, X_n) \tag{23}$$

7

The probability distribution is

$$P_\Phi(X_1, \cdots, X_n) = \frac{\tilde{P}_\Phi(X_1, \cdots, X_n)}{Z_\Phi} \tag{24}$$

**Definition 1.24.** The **induced Markov network** of a set of factors $\Phi = \{\phi_i(D_i)\}_{i=1}^k$, where $D_i \subseteq \{X_1, \cdots, X_n\}$, denoted as $H_\Phi$, is a network in which there is an edge between $X_i$ and $X_j$ whenever $\exists \, m \in [k] \; s.t. \; X_i, X_j \in D_m$.

**Definition 1.25.** A probability distribution $P$ **factorizes** over a Markov network $H$ if there exists $\Phi = \{\phi_1(D_1), \cdots, \phi_k(D_k)\}$ such that $P = P_\Phi$ and $H = H_\Phi$.

**Remark 1.3.** There can be multiple factorizations of a given Markov network.

**Definition 1.26.** A trail (path) $X_1 - \cdots - X_n$ in a Markov network is **active** given a set of nodes $\mathbf{Z}$ if no $X_i$ is in $\mathbf{Z}$.

## 1.10 Conditional Random Fields

**Definition 1.27** (CRF). A **CRF representation** over random variables $X \cup Y$ consists of a set of factors $\Phi = \{\phi_1(D_1), \cdots, \phi_k(D_k)\}$, where $D_i \subseteq X \cup Y$.

The unnormalized probability and partition function are defined as

$$\tilde{P}(X, Y) = \prod_{i=1}^k \phi_i(D_i) \tag{25}$$

$$Z(X) = \sum_Y \tilde{P}(X, Y) \tag{26}$$

The conditional probability is therefore

$$P(Y|X) = \frac{1}{Z(X)} \tilde{P}(X, Y) \tag{27}$$

**Remark 1.4.** A CRF is parameterized the same as a Gibbs distribution, both of them are defined using factors. However, they are normalized differently, the partition function in CRF, $Z(X)$, depends on the particular realization of $X$, but $Z_\Phi$ depends on factors only.

**Example 1.7** (Logistic model as a CRF). Let $X_1, \cdots, X_k$ denote the $k$ random variables serve as features to generate the target variable $Y$. Figure 5 illustrates the logistic model as a Bayesian

network. For simplicity, assume $X_i$ and $Y$ are all binary. Define factors

$$\phi_i(X_i, Y) := \exp(w_i \mathbb{1}\{X_i = 1 \wedge Y = 1\}) = \exp(w_i X_i Y) \tag{28}$$

Therefore, the unnormalized probabilities are

$$\tilde{P}(Y = 0, X_i) = \prod_{i=1}^{k} \exp(w_i X_i 0) = 1 \tag{29}$$

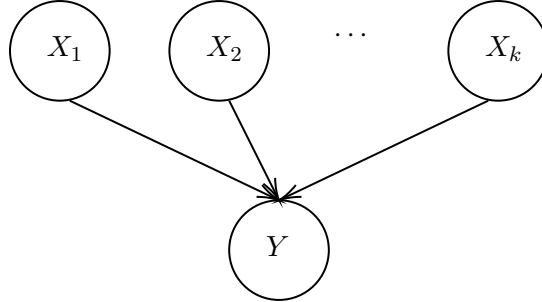$$\tilde{P}(Y = 1, X_i) = \prod_{i=1}^{k} \exp(w_i X_i) = \exp\left(\sum_{i=1}^{k} w_i X_i\right) \tag{30}$$

Then, the CPD can be expressed as

$$P(Y = 1 | X_1, \cdots, X_k) = \frac{\tilde{P}(Y = 1, X_i)}{\tilde{P}(Y = 0, X_i) + \tilde{P}(Y = 1, X_i)} \tag{31}$$

$$= \frac{\exp\left(\sum_{i=1}^{k} w_i X_i\right)}{1 + \exp\left(\sum_{i=1}^{k} w_i X_i\right)} \tag{32}$$

$$= \frac{1}{1 + \exp\left(-\sum_{i=1}^{k} w_i X_i\right)} = \sigma\left(\sum_{i=1}^{k} w_i X_i\right) \tag{33}$$

Figure 5: Logistic model as a Bayesian network



## 1.11   Independencies in Markov Networks

**Definition 1.28.** Two nodes $X$ and $Y$ in a Markov network $H$ are **separated** given set of nodes $Z$ if there is no active trail (path) in $H$ between $X$ and $Y$ given $Z$. Denoted as $\text{sep}_H(X, Y | Z)$

**Theorem 1.4.** If distribution $P$ factorizes over Markov network $H$, then $P$ satisfies

$$\text{sep}_H(X, Y|Z) \implies X \perp\!\!\!\perp Y|Z \tag{34}$$

*Proof.* ∎

**Definition 1.29.** Let $I(H)$ denote the collection of independencies induced by $H$:

$$\mathcal{I}(H) := \{(X \perp\!\!\!\perp Y|Z : \text{sep}_H(X, Y|Z)\} \tag{35}$$

If $P$ satisfies $\mathcal{I}(H)$, then $H$ is an **independency map** (I-map) of $P$.

**Theorem 1.5.** If $P$ factorizes over $H$, then $H$ is an I-map of $P$.

*Proof.* ∎

**Theorem 1.6** (Hammersley Clifford)**.** For a <u>positive</u> distribution $P$ (i.e., $P(x) > 0$ for all $x$), if $H$ is an I-map for $P$, then $P$ factorizes over $H$.

*Proof.* ∎

## 2   Inference

## 3   Learning

### 3.1   Learning Network Structure

**Motivation**   We have discussed methods to estimate parameters given a network structure and dataset, however, ways to choose the structure remains uncovered. Specifically, given a dataset $D$, we have to choose one graph $G$ as the skeleton of our model.

Compared with the graph representing the true data generating process, if the specified graph $G$ contains less edges, $G$ imposes more independence assumptions than necessary. In contrast, if the specified graph $G$ has extra edges, then the redundant dependencies would require more parameters to be fitted using the limited dataset and leads to bad generalization.

To select the best structure $G$ given dataset $D$, we can define a **scoring function** taking both

$G$ and $D$ as arguments. Then the model selection problem turns into an optimization problem:

$$\text{Optimal } G^* := \operatorname*{argmax}_{G \in \text{all models}} \text{score}(G, D) \tag{36}$$

### 3.1.1 Likelihood Structure Scores

**Definition 3.1. Likelihood structure score** defines the compatibleness between a graph $G$ and dataset $D$ as

$$\text{score}_L(G; D) := \ell(G_{\hat{\theta}}; D) \tag{37}$$

Where $\hat{\theta}$ is the maximum likelihood estimations from of parameters in graph $G$ from dataset $D$, and $\ell$ is the log-likelihood function.