Probabilistic Graphical Models

Lecture 4: Factor Graphs, Gaussian Networks

Matthew Walter

TTI-Chicago

April 16, 2020

Gibbs Distribution (Revisited)

- A distribution P is a **Gibbs distribution** *parameterized* by a set of factors $\Phi = \{\phi_1(\boldsymbol{D_1}), \ldots, \phi_K(\boldsymbol{D_K})\}$ over sets $\boldsymbol{D}_k$ if

$$P(X_1, \ldots, X_n) = \frac{1}{Z}\phi_1(\boldsymbol{D_1}) \times \ldots \times \phi_K(\boldsymbol{D_K})$$

where Z is a normalizing constant known as the **partition function**

$$Z = \sum_{X_1, \ldots, X_n} \phi_1(\boldsymbol{D_1}) \times \ldots \times \phi_K(\boldsymbol{D_K})$$

Markov Networks (Revisited)

- A **Markov Network** is a pair $B = (P, H)$ for which P is a Gibbs distribution that factorizes over H

$$P(X_i, \ldots, X_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

  where Z is the **partition function** that normalizes the distribution

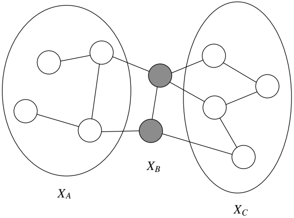
$$Z = \sum_{X_i, \ldots, X_n} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{X}_C)$$

- Provides an alternative representation for joint distributions
- Factors are not equivalent to conditional or marginal distributions over variables in their scope
- Also known as **Markov random fields** (MRFs)

Markov Network Independencies (Revisited)

Consider a Markov network structure H over $X_1, \ldots X_n$

- A path $X_i — \cdots — X_k$ in H is **active** given $\boldsymbol{Z}$ if none of the X_i's along the path are in $\boldsymbol{Z}$
- A set of nodes $\boldsymbol{Z}$ **separates** $\boldsymbol{X}$ and $\boldsymbol{Y}$ in H ($\text{sep}_H(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$) if there is no active path between any $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$



$$\text{sep}_H(X_A; X_C \mid X_B)$$

- **Global Markov independencies** of H are

$$I(H) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}) : \text{sep}_H(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})\}$$

Representation Theorem (Revisited)

- **Soundness**: $\text{sep}_H(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \Rightarrow P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid Z)$

Theorem

If P is a Gibbs distribution that factorizes over H, then

$$I(H) \subseteq I(P)$$

- **Completeness**: $\text{sep}_H(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z}) \Leftarrow P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid Z)$? Not in general

Theorem (Hammersley-Clifford)

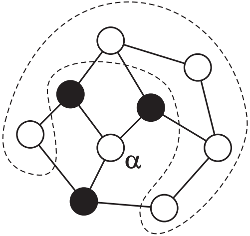
If P is a positive distribution and $I(H) \subseteq I(P)$, then P is a Gibbs distribution that factorizes over H

Local Markov Independencies (Revisited)

- **Pairwise Markov Independencies**: Any pair of non-neighboring variables are independent given everything else

$$I_p(H) = \{(X \perp Y \mid \boldsymbol{X} - \{X, Y\}) : X\text{—}Y \notin H\}$$

- For a Markov network H, the **Markov blanket** of a variable X $\text{MB}_H(X)$ is the neighbors of X



- **Local Markov Independencies**: A random variable X is independent of all other variables given its Markov blanket

$$I_l(H) = \{(X \perp \boldsymbol{X} - X - \text{MB}_H(X) \mid \text{MB}_H(X)) : X \in \boldsymbol{X}\}$$

Relationship Between Markov Independencies (Revisited)

For a *positive* distribution P, the following are equivalent

1. $P \models I_l(H)$
2. $P \models I_p(H)$
3. $P \models I(H)$

Non-positive distributions P (i.e., one or more events has 0 probability) satisfy some (weaker) but not all (stronger) properties

This equivalence is useful in constructing Markov networks that are minimal I-maps for a given (positive) distribution

From Distributions to Graphs

Pairwise independencies: $\{X, Y\} \notin H \Rightarrow P \vDash (X \perp Y \mid \boldsymbol{X} - \{X, Y\})$

From Distributions to Graphs

Pairwise independencies: $\{X, Y\} \notin H \Rightarrow P \models (X \perp Y \mid \boldsymbol{X} - \{X, Y\})$

Theorem

If P is positive and H is constructed by adding edges such that above applies, then H is a <u>minimal</u> I-map for P

From Distributions to Graphs

Pairwise independencies: $\{X, Y\} \notin H \Rightarrow P \models (X \perp Y \mid \boldsymbol{X} - \{X, Y\})$

Theorem

If P is positive and H is constructed by adding edges such that above applies, then H is a minimal I-map for P

The **Markov blanket** $\mathrm{MB}_P(X)$ of X in a distribution P is the minimal set $\boldsymbol{U}$ such that $X \notin \boldsymbol{U}$ and

$$(X \perp (\boldsymbol{X} - \{X\} - \boldsymbol{U}) \mid \boldsymbol{U}) \in I(P)$$

From Distributions to Graphs

Pairwise independencies: $\{X, Y\} \notin H \Rightarrow P \models (X \perp Y \mid \boldsymbol{X} - \{X, Y\})$

Theorem

If P is positive and H is constructed by adding edges such that above applies, then H is a minimal I-map for P

The **Markov blanket** $\mathrm{MB}_P(X)$ of X in a distribution P is the minimal set $\boldsymbol{U}$ such that $X \notin \boldsymbol{U}$ and

$$(X \perp (\boldsymbol{X} - \{X\} - \boldsymbol{U}) \mid \boldsymbol{U}) \in I(P)$$

Theorem

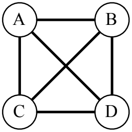
For a positive distribution P, if H is constructed by adding an edge $X—Y$ for all X and $Y \in MB_P(X)$, then H is the unique minimal I-map for P.

Higher-order Potentials

- Up till now, we have considered pairwise Markov networks that involve only unary $\phi_i(X_i)$ and pairwise potentials $\phi_{ij}(X_i, X_j)$
- It is often useful to have **higher-order** potentials, e.g.,

$$\phi(X, Y, Z) = 1[X + Y + Z \geq 1]$$

  where X, Y, and Z are binary, enforcing that at least one variable is 1
- Markov networks are useful, but they obscure lower level structure of the Gibbs parameterization

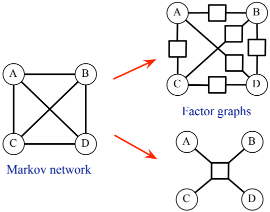


Factor Graphs

- The markov network H does not make explicit the structure of the distribution, i.e., maximum cliques vs. complete graph subsets

Factor Graphs

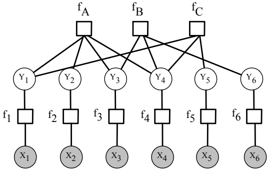
- The markov network H does not make explicit the structure of the distribution, i.e., maximum cliques vs. complete graph subsets
- A **factor graph** is a bipartite undirected graph with variable nodes (oval) and factor nodes (square). Edges exist only between variable nodes and factor nodes
- Each factor node is associated with a single potential, the scope of which is the variables that are the factor's neighbors

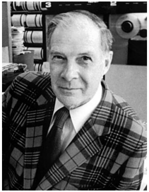


- Distribution is the same as an MRF, just a different data structure

Example: Low-Density Parity-Check Codes

- Error correcting codes for transmitting messages over noisy channels





Richard Hamming
(1915–1998)
UChicago (B.S. 1937)

- Each factor in the top row enforces even parity

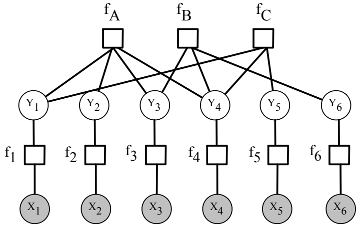
$$f_A(Y_1, Y_2, Y_3, Y_4) = 1 \text{ if } Y_1 \oplus Y_2 \oplus Y_3 \oplus Y_4 = 0$$

- Only assignments $\boldsymbol{Y}$ with non-zero probability are the following **codewords**:

$$000000,\ 011001,\ 110010,\ 101011,\ 111100,\ 100101,\ 001110,\ 010111$$

- $f_i(Y_i, X_i) = P(X_i \mid Y_i)$: the likelihood of a bit flip per noise model

Examle: Low-Density Parity-Check Codes



- *Decoding* problem is to infer **maximum a posteriori** (MAP) estimate

$$\boldsymbol{Y}^* = \arg\max_{\boldsymbol{Y}} P(\boldsymbol{Y} \mid \boldsymbol{X})$$

- Since Z and $P(\boldsymbol{X})$ are constants w.r.t. $\boldsymbol{Y}$, we can equivalently maximize $P(\boldsymbol{Y}, \boldsymbol{X})$

Boltzmann Distribution

- We can rewrite a factor $\phi(\boldsymbol{D})$: $\text{Val}(\boldsymbol{D}) \rightarrow \mathbb{R}^+$ as

$$\phi(\boldsymbol{D}) = \exp(-\psi(\boldsymbol{D}))$$

where $\psi(\boldsymbol{D}) = -\log \phi(\boldsymbol{D})$ is the **energy function** (not surprisingly, derived from statistical physics)

Boltzmann Distribution

- We can rewrite a factor $\phi(\boldsymbol{D}) : \mathrm{Val}(\boldsymbol{D}) \rightarrow \mathbb{R}^+$ as

$$\phi(\boldsymbol{D}) = \exp(-\psi(\boldsymbol{D}))$$

  where $\psi(\boldsymbol{D}) = -\log \phi(\boldsymbol{D})$ is the **energy function** (not surprisingly, derived from statistical physics)
- The factorized distribution then becomes (Boltzmann distribution)

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{k=1}^{K} \exp\left(-\psi_k(\boldsymbol{D}_k)\right) = \frac{1}{Z} \exp\left(-\sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)\right)$$

- $\sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)$ is referred to as the "free energy"

Boltzmann Distribution

- We can rewrite a factor $\phi(\boldsymbol{D}) : \mathsf{Val}(\boldsymbol{D}) \rightarrow \mathbb{R}^{+}$ as

$$\phi(\boldsymbol{D}) = \exp(-\psi(\boldsymbol{D}))$$

  where $\psi(\boldsymbol{D}) = -\log \phi(\boldsymbol{D})$ is the **energy function** (not surprisingly, derived from statistical physics)

- The factorized distribution then becomes (Boltzmann distribution)

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{k=1}^{K} \exp\left(-\psi_k(\boldsymbol{D}_k)\right) = \frac{1}{Z} \exp\left(-\sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)\right)$$

- $\sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)$ is referred to as the "free energy"
- Gives rise to interpretation as energy minimization

$$\arg\max\, P(X_1, \ldots, X_n) = \arg\min\, \sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)$$

Log-Linear Markov Networks with Features

- A **feature** is a function $f : \text{Val}(\boldsymbol{D}_i) \rightarrow \mathbb{R}$

[1] We can have multiple features over the same variables

Log-Linear Markov Networks with Features

- A **feature** is a function $f : \mathrm{Val}(\boldsymbol{D}_i) \rightarrow \mathbb{R}$
- A distribution P is a **log-linear model** over a Markov network H if it is associated with
  - A set of features $\boldsymbol{F} = \{f_1(\boldsymbol{D}_1), \ldots, f_K(\boldsymbol{D}_M)\}$[1] where $\boldsymbol{D}_i$ is a complete subgraph in H
  - A set of weights $\{w_1, \ldots, w_M\}$

  such that

$$P(X_1, \ldots, X_n) \propto \exp\left(-\sum_{i=1}^{M} w_i f_i(\boldsymbol{D}_i)\right)$$

[1] We can have multiple features over the same variables

Log-Linear Markov Networks with Features

- A **feature** is a function $f : \mathrm{Val}(\boldsymbol{D}_i) \rightarrow \mathbb{R}$
- A distribution P is a **log-linear model** over a Markov network H if it is associated with
  - A set of features $\boldsymbol{F} = \{f_1(\boldsymbol{D}_1), \ldots, f_K(\boldsymbol{D}_M)\}$[1] where $\boldsymbol{D}_i$ is a complete subgraph in H
  - A set of weights $\{w_1, \ldots, w_M\}$

  such that

$$P(X_1, \ldots, X_n) \propto \exp\left(-\sum_{i=1}^{M} w_i f_i(\boldsymbol{D}_i)\right)$$

- Features and weights can be reused for different factors

[1] We can have multiple features over the same variables

Log-Linear Markov Networks with Features

- A **feature** is a function $f : \text{Val}(\boldsymbol{D}_i) \rightarrow \mathbb{R}$
- A distribution P is a **log-linear model** over a Markov network H if it is associated with
  - A set of features $\boldsymbol{F} = \{f_1(\boldsymbol{D}_1), \ldots, f_K(\boldsymbol{D}_M)\}$[1] where $\boldsymbol{D}_i$ is a complete subgraph in H
  - A set of weights $\{w_1, \ldots, w_M\}$

  such that

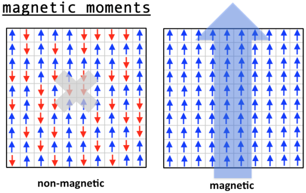
$$P(X_1, \ldots, X_n) \propto \exp\left(-\sum_{i=1}^{M} w_i f_i(\boldsymbol{D}_i)\right)$$

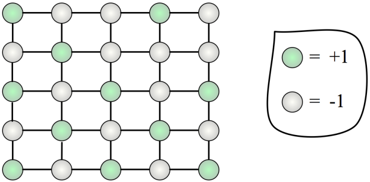
- Features and weights can be reused for different factors
- Historically, features are designed by hand & weights learned from data

[1] We can have multiple features over the same variables

Example: Ising Model

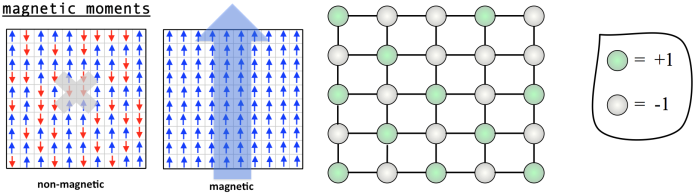
- Statistical mechanics model of ferromagnetism
- An atom's spin is influenced by the spins of nearby atoms





Example: Ising Model

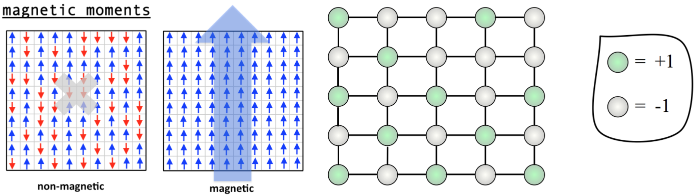
- Statistical mechanics model of ferromagnetism
- An atom's spin is influenced by the spins of nearby atoms



- Each atom $X_i \in \{-1, +1\}$ indicating direction of spin

Example: Ising Model

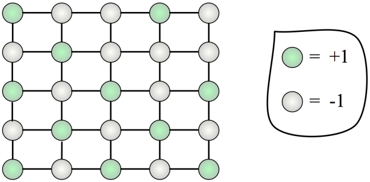
- Statistical mechanics model of ferromagnetism
- An atom's spin is influenced by the spins of nearby atoms



- Each atom $X_i \in \{-1, +1\}$ indicating direction of spin
- Inference: If the spin at position i is -1, what is the probability that the spin at position j is $+1$?
- Are phase transitions possible?
- Invented by physicist Wilhelm Lenz (1920), who gave it as a problem to his student Ernst Ising

Example: Ising Model

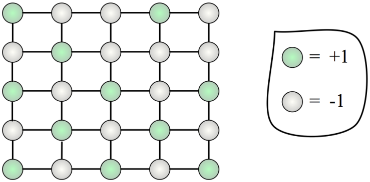
- Each atom $X_i \in \{-1, +1\}$ indicating direction of spin
- An atom's spin is influenced by the spins of nearby atoms



$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{i,j} \phi_{ij}(x_i, x_j) \cdot \prod_i \phi_i(x_i)$$

Example: Ising Model

- Each atom $X_i \in \{-1, +1\}$ indicating direction of spin
- An atom's spin is influenced by the spins of nearby atoms

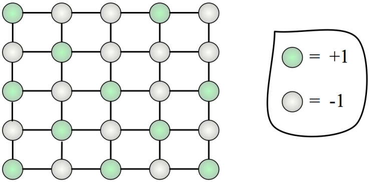


$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{i,j} \phi_{ij}(x_i, x_j) \cdot \prod_i \phi_i(x_i)$$

$$= \frac{1}{Z} \Big(\prod_{\{i,j\} \in H} \exp(w_{i,j} x_i x_j) \Big) \cdot \Big(\prod_i \exp(u_i x_i) \Big)$$

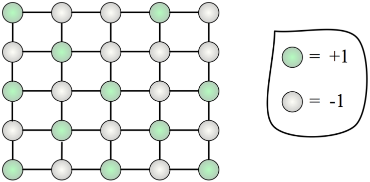
Example: Ising Model

- Each atom $X_i \in \{-1, +1\}$ indicating direction of spin
- An atom's spin is influenced by the spins of nearby atoms



$$
\begin{aligned}
P(x_1, \ldots, x_n) &= \frac{1}{Z} \prod_{i,j} \phi_{ij}(x_i, x_j) \cdot \prod_i \phi_i(x_i) \\
&= \frac{1}{Z} \Big(\prod_{\{i,j\} \in H} \exp(w_{i,j} x_i x_j) \Big) \cdot \Big(\prod_i \exp(u_i x_i) \Big) \\
&= \frac{1}{Z} \exp\Big(\sum_{\{i,j\} \in H} w_{i,j} x_i x_j + \sum_i u_i x_i \Big)
\end{aligned}
$$

Example: Ising Model



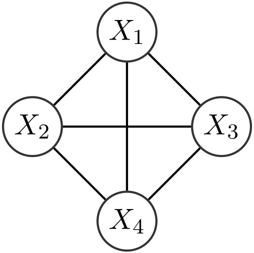
$$P(x_1, \ldots, x_n) = \frac{1}{Z} \exp(-E(\boldsymbol{x}))$$

where

$$E(\boldsymbol{x}) = -\Big(\sum_{\{i,j\} \in H} w_{i,j} x_i x_j + \sum_i u_i x_i \Big)$$

- $w_{i,j} > 0$ encourages neighboring atoms to have the same spin (ferromagnetic), whereas $w_{i,j} < 0$ encourages $x_i \neq x_j$
- Unary node potentials $\exp(u_i x_i)$ encode a bias on individual atoms
- Scaling the weight parameters $w_{i,j}$ and u_i makes the distribution more or less spiky

Example: Boltzmann Machine

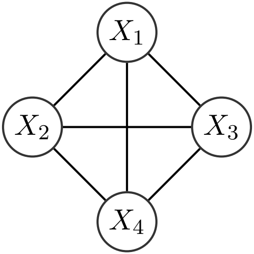


- A fully-connected graph with pairwise potentials over binary-valued variables $X_i \in \{0, 1\}$ is called a **Boltzmann machine**
- The joint distribution takes the same form as the Ising model

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \exp\Big(\sum_{\{i,j\} \in H} w_{i,j} x_i x_j + \sum_i u_i x_i \Big)$$

- Proposed by Hinton and Senjowski (1985) as one of the first neural networks for representation learning

Example: Boltzmann Machine



- A fully-connected graph with pairwise potentials over binary-valued variables $X_i \in \{0, 1\}$ is called a **Boltzmann machine**
- Conditional distribution over X_i given its neighbors $\boldsymbol{U} = \mathsf{MB}_H(X_i)$

$$P(x_i \,|\, \boldsymbol{U}) = \frac{1}{1 + e^{-z}} \quad \text{where} \quad z = -\left(\sum_{j \in \boldsymbol{U}} w_{i,j} x_j\right) - w_i$$

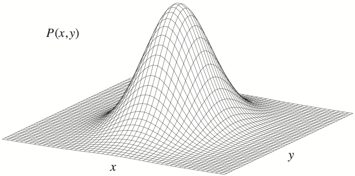
Gaussian Markov Random Fields

- Let's briefly return to continuous random variables
- Suppose we have a multivariate Gaussian density p over $X_1, \ldots, X_n$
- We denote this as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the **mean vector** and $\Sigma \in \mathbb{R}^{n \times n}$ is the **covariance matrix**

Gaussian Markov Random Fields

- Let's briefly return to continuous random variables
- Suppose we have a multivariate Gaussian density p over $X_1, \ldots, X_n$
- We denote this as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the **mean vector** and $\Sigma \in \mathbb{R}^{n \times n}$ is the **covariance matrix**
- The density function is defined as

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right)$$



Gaussian Markov Random Fields

- The term in the exponential can be expressed as

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Lambda(\boldsymbol{x}-\boldsymbol{\mu})$$

Gaussian Markov Random Fields

- The term in the exponential can be expressed as

$$\begin{aligned}-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\\ &= -\frac{1}{2}(\boldsymbol{x}^\top\Lambda\boldsymbol{x}-2\boldsymbol{x}^\top\Lambda\boldsymbol{\mu}+\boldsymbol{\mu}^\top\Lambda\boldsymbol{\mu})\end{aligned}$$

Gaussian Markov Random Fields

- The term in the exponential can be expressed as

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Lambda(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= -\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \Lambda \boldsymbol{\mu} + \boldsymbol{\mu}^\top \Lambda \boldsymbol{\mu}) \end{aligned}$$

- This is referred to as the **information (canonical) form** $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$ where $\Lambda = \Sigma^{-1} > 0$ is the **information matrix** and $\boldsymbol{\eta} = \Lambda \boldsymbol{\mu}$ is the **information (potential) vector**

Gaussian Markov Random Fields

- The information form parametrization $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$, where $\Lambda = \Sigma^{-1}$ and $\boldsymbol{\eta} = \Lambda\boldsymbol{\mu}$ can be alternatively expressed as:

$$p(\boldsymbol{x}) \propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top\Lambda\boldsymbol{x} - 2\boldsymbol{x}^\top\Lambda\boldsymbol{\mu} + \boldsymbol{\mu}^\top\Lambda\boldsymbol{\mu})\Big)$$

Gaussian Markov Random Fields

- The information form parametrization $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$, where $\Lambda = \Sigma^{-1}$ and $\boldsymbol{\eta} = \Lambda\boldsymbol{\mu}$ can be alternatively expressed as:

$$
\begin{aligned}
p(\boldsymbol{x}) &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \Lambda \boldsymbol{\mu} + \boldsymbol{\mu}^\top \Lambda \boldsymbol{\mu})\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{\eta})\Big)
\end{aligned}
$$

Gaussian Markov Random Fields

- The information form parametrization $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$, where $\Lambda = \Sigma^{-1}$ and $\boldsymbol{\eta} = \Lambda\boldsymbol{\mu}$ can be alternatively expressed as:

$$\begin{aligned}
p(\boldsymbol{x}) &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \Lambda \boldsymbol{\mu} + \boldsymbol{\mu}^\top \Lambda \boldsymbol{\mu})\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{\eta})\Big) \\
&= \exp\Big(-\frac{1}{2}\sum_i (\Lambda_{ii} x_i^2 + 2\eta_i x_i) - \frac{1}{2}\sum_{i,j:i\neq j} (\Lambda_{ij} x_i x_j + \Lambda_{j,i} x_j x_i)\Big)
\end{aligned}$$

Gaussian Markov Random Fields

- The information form parametrization $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$, where $\Lambda = \Sigma^{-1}$ and $\boldsymbol{\eta} = \Lambda\boldsymbol{\mu}$ can be alternatively expressed as:

$$\begin{aligned}
p(\boldsymbol{x}) &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \Lambda \boldsymbol{\mu} + \boldsymbol{\mu}^\top \Lambda \boldsymbol{\mu})\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{\eta})\Big) \\
&= \exp\Big(-\frac{1}{2}\sum_i (\Lambda_{ii} x_i^2 + 2\eta_i x_i) - \frac{1}{2}\sum_{i,j:i\neq j} (\Lambda_{ij} x_i x_j + \Lambda_{j,i} x_j x_i)\Big) \\
&= \exp\Big(-\frac{1}{2}\sum_i (\Lambda_{ii} x_i^2 + 2\eta_i x_i) - \sum_{i,j:i\neq j} \Lambda_{ij} x_i x_j\Big)
\end{aligned}$$

Gaussian Markov Random Fields

- The information form parametrization $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$, where $\Lambda = \Sigma^{-1}$ and $\boldsymbol{\eta} = \Lambda \boldsymbol{\mu}$ can be alternatively expressed as:

$$
\begin{aligned}
p(\boldsymbol{x}) &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \Lambda \boldsymbol{\mu} + \boldsymbol{\mu}^\top \Lambda \boldsymbol{\mu})\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{\eta})\Big) \\
&= \exp\Big(-\frac{1}{2}\sum_i (\Lambda_{ii} x_i^2 + 2\eta_i x_i) - \frac{1}{2}\sum_{i,j:i\neq j} (\Lambda_{ij} x_i x_j + \Lambda_{j,i} x_j x_i)\Big) \\
&= \exp\Big(-\frac{1}{2}\sum_i (\Lambda_{ii} x_i^2 + 2\eta_i x_i) - \sum_{i,j:i\neq j} \Lambda_{ij} x_i x_j\Big) \\
&= \exp\Big(-\frac{1}{2}\sum_i (\Lambda_{ii} x_i^2 + 2\eta_i x_i)\Big) \cdot \exp\Big(-\sum_{i,j:i\neq j} \Lambda_{ij} x_i x_j\Big)
\end{aligned}
$$

Gaussian Markov Random Fields

- The information form parametrization $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$, where $\Lambda = \Sigma^{-1}$ and $\boldsymbol{\eta} = \Lambda\boldsymbol{\mu}$ can be alternatively expressed as (continued):

$$\begin{aligned} p(\boldsymbol{x}) &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \Lambda \boldsymbol{\mu} + \boldsymbol{\mu}^\top \Lambda \boldsymbol{\mu})\Big) \\ &\propto \exp\Big(-\frac{1}{2}\sum_i (\Lambda_{ii} x_i^2 + 2\eta_i x_i)\Big) \cdot \exp\Big(-\sum_{i,j:i \neq j} \Lambda_{ij} x_i x_j\Big) \end{aligned}$$

Gaussian Markov Random Fields

- The information form parametrization $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$, where $\Lambda = \Sigma^{-1}$ and $\boldsymbol{\eta} = \Lambda\boldsymbol{\mu}$ can be alternatively expressed as (continued):

$$
\begin{aligned}
p(\boldsymbol{x}) &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top\Lambda\boldsymbol{x} - 2\boldsymbol{x}^\top\Lambda\boldsymbol{\mu} + \boldsymbol{\mu}^\top\Lambda\boldsymbol{\mu})\Big) \\
&\propto \exp\Big(-\frac{1}{2}\sum_i(\Lambda_{ii}x_i^2 + 2\eta_i x_i)\Big) \cdot \exp\Big(-\sum_{i,j:i\neq j}\Lambda_{ij}x_i x_j\Big) \\
&= \prod_i \exp\Big(-\frac{1}{2}(\Lambda_{ii}x_i^2 + 2\eta_i x_i)\Big) \cdot \prod_{i,j:i\neq j}\exp\Big(-\Lambda_{ij}x_i x_j\Big)
\end{aligned}
$$

Gaussian Markov Random Fields

- The information form parametrization $\boldsymbol{x} \sim \mathcal{N}^{-1}(\boldsymbol{\eta}, \Lambda)$, where $\Lambda = \Sigma^{-1}$ and $\boldsymbol{\eta} = \Lambda\boldsymbol{\mu}$ can be alternatively expressed as (continued):

$$
\begin{aligned}
p(\boldsymbol{x}) &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{x}^\top \Lambda \boldsymbol{x} - 2\boldsymbol{x}^\top \Lambda \boldsymbol{\mu} + \boldsymbol{\mu}^\top \Lambda \boldsymbol{\mu})\Big) \\
&\propto \exp\Big(-\frac{1}{2}\sum_i (\Lambda_{ii} x_i^2 + 2\eta_i x_i)\Big) \cdot \exp\Big(-\sum_{i,j:i\neq j} \Lambda_{ij} x_i x_j\Big) \\
&= \prod_i \exp\Big(-\frac{1}{2}(\Lambda_{ii} x_i^2 + 2\eta_i x_i)\Big) \cdot \prod_{i,j:i\neq j} \exp\Big(-\Lambda_{ij} x_i x_j\Big) \\
&= \prod_i \phi_i(x_i) \cdot \prod_{i,j:i\neq j} \phi_{ij}(x_i, x_j)
\end{aligned}
$$

Gaussian Markov Random Fields

$$p(\boldsymbol{x}) \propto \prod_i \exp\Big(-\frac{1}{2}(\Lambda_{ii}x_i^2 + 2\eta_i x_i)\Big) \cdot \prod_{i,j:i\neq j} \exp\Big(-\Lambda_{ij}x_i x_j\Big)$$
$$= \prod_i \phi_i(x_i) \cdot \prod_{i,j:i\neq j} \phi_{ij}(x_i, x_j)$$

- Any Gaussian distribution can be represented by a pairwise Markov network with quadratic node and edge potentials
- The Markov network is referred to as a **Gaussian Markov random field (GMRF)**
- Two nodes x_i and x_j have an edge in the GMRF only if $\Lambda_{ij} \neq 0$
- The structure of the information matrix Λ directly encodes the Markov network graph structure

Converting Bayesian Networks to Markov Networks

- Many inference algorithms are more convenient for MRFs

Converting Bayesian Networks to Markov Networks

- Many inference algorithms are more convenient for MRFs
- **Moralization** converts a Bayesian network to a Markov network
- The **moral graph** $\mathcal{M}[G]$ of a BN structure is an undirected graph over V that contains an edge between X_i and X_j if
    1. There is a direct edge between them (either direction)
    2. X_i and X_j are both parents of the same node

- Historically, derives from the expression "marrying the parents"

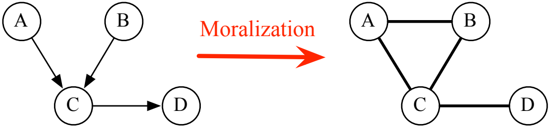
Converting Bayesian Networks to Markov Networks

- Many inference algorithms are more convenient for MRFs
- **Moralization** converts a Bayesian network to a Markov network
- The **moral graph** $\mathcal{M}[G]$ of a BN structure is an undirected graph over V that contains an edge between X_i and X_j if
  1. There is a direct edge between them (either direction)
  2. X_i and X_j are both parents of the same node

- Historically, derives from the expression "marrying the parents"
- The reverse direction (MRF to BN) is far more difficult

Converting Bayesian Networks to Markov Networks

1. Moralize the directed graph to obtain undirected graph



2. Introduce one potential for each CPD (a factor over X_i and $\mathrm{Pa}_{X_i}^G$)

$$\phi_i(X_i, \mathrm{Pa}_{X_i}^G) = P(X_i \mid \mathrm{Pa}_{X_i}^G)$$

Converting Bayesian Networks to Markov Networks

1. Moralize the directed graph to obtain undirected graph

2. Introduce one potential for each CPD (a factor over X_i and $\mathrm{Pa}_{X_i}^G$)

$$\phi_i(X_i, \mathrm{Pa}_{X_i}^G) = P(X_i \,|\, \mathrm{Pa}_{X_i}^G)$$

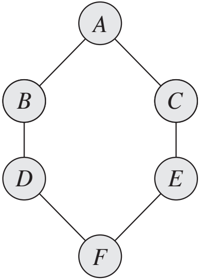
- Given a Bayesian network graph G, $\mathcal{M}[G]$ is a minimal I-map for G
- The addition of moralizing edges obfuscates some independencies, e.g., $(A \perp B) \in I(G)$
- If G is moral, then $\mathcal{M}[G]$ is a perfect I-map for G (no obfuscation of independencies)

Converting Markov Networks to Bayesian Networks

- Converting a Markov network H to a Bayesian network G is typically harder, and often involves adding many edges

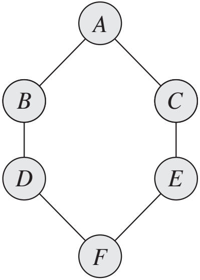
Converting Markov Networks to Bayesian Networks

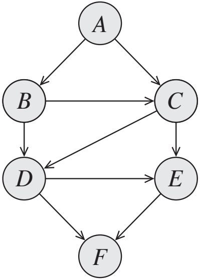
- Converting a Markov network H to a Bayesian network G is typically harder, and often involves adding many edges
- Consider the following MRF:



Converting Markov Networks to Bayesian Networks

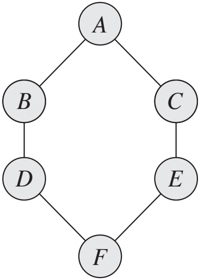
- Converting a Markov network H to a Bayesian network G is typically harder, and often involves adding many edges
- Consider the following MRF:

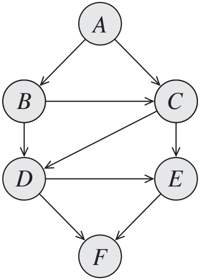




Converting Markov Networks to Bayesian Networks

- Converting a Markov network H to a Bayesian network G is typically harder, and often involves adding many edges
- Consider the following MRF:





- A Bayesian network G is a minimal I-map for a Markov network H iff G has no immoralities

Converting Markov Networks to Bayesian Networks

- An undirected graph is **chordal** (triangulated) if every cycle of length ≥ 3 has a shortcut (a "chord"), i.e., for any $X_1 - X_2 - \cdots - X_k - X_1$ ($k \geq 3$), there is an edge between nonconsecutive nodes X_i and X_j
- A directed graph is chordal if its undirected graph is chordal

Converting Markov Networks to Bayesian Networks

- An undirected graph is **chordal** (triangulated) if every cycle of length ≥ 3 has a shortcut (a "chord"), i.e., for any $X_1 - X_2 - \cdots - X_k - X_1$ $(k \geq 3)$, there is an edge between nonconsecutive nodes X_i and X_j
- A directed graph is chordal if its undirected graph is chordal
- Any nontriangulated cycle of length ≥ 3 contains an immorality
- Thus, if G is a minimal I-map for a Markov network H, then G must be chordal

Converting Markov Networks to Bayesian Networks

- An undirected graph is **chordal** (triangulated) if every cycle of length ≥ 3 has a shortcut (a "chord"), i.e., for any $X_1 - X_2 - \cdots - X_k - X_1$ $(k \geq 3)$, there is an edge between nonconsecutive nodes X_i and X_j
- A directed graph is chordal if its undirected graph is chordal
- Any nontriangulated cycle of length ≥ 3 contains an immorality
- Thus, if G is a minimal I-map for a Markov network H, then G must be chordal
- Generating a Bayesian network for a Markov network involves **triangulating** the graph by adding edges to make the graph chordal
- Triangulation results in a loss of independence relations present in H