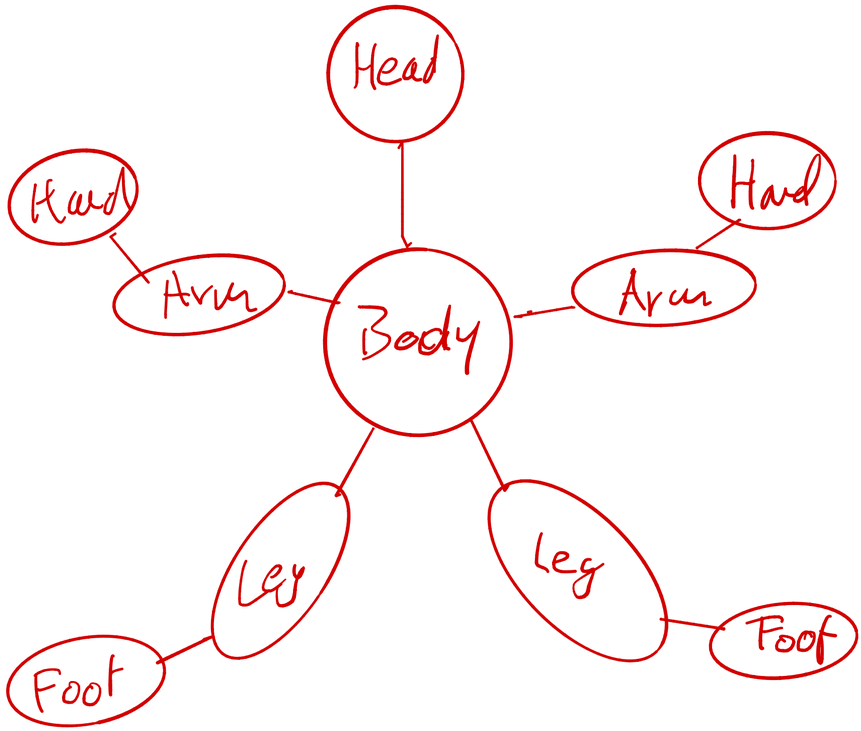
Machine Learning 2

Graphical Models
- Markov Random Fields
- Definition

Patrick Forré

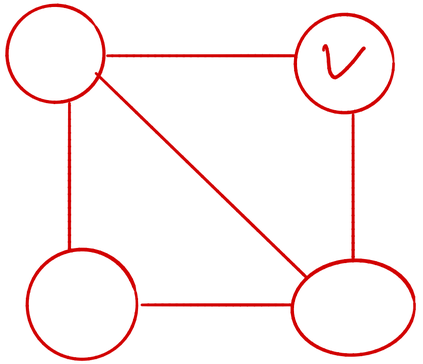
Markov Random Fields - Motivation

- Modelling sparse/local symmetric relationships/correlations



Undirected Graph (UG) - Definition

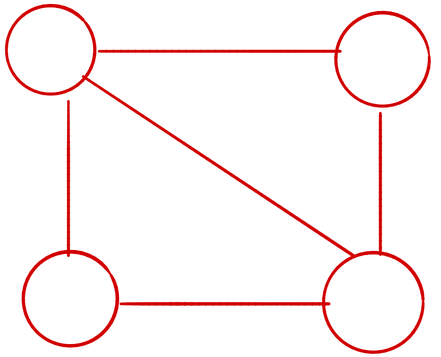
- An Undirected Graphs (UG) $G = (V, E)$ consists of:
  - set of nodes/vertices V,
  - set of undirected edges between different nodes E



- The set of neighbours of a node $v \in V$ in G:

$$\partial(v) := \{ w \in V \mid w - v \in E \}$$

Cliques - Definition



- Let $G = (V, E)$ be undirected graph.
- A Clique C of G is a subgraph of G where every pair of nodes in C has an edge (completely connected).

Addendum: We allow single node sets to be cliques as well: $\{v\}$.

- A Maximal Clique C of G is a Clique that does not lie in a bigger one.

Factorization Property

- Let V be an (index) set of variables $X_1, \ldots, X_M$ with a joint distribution $p(x_V)$.
- Let $G = (V, E)$ be an undirected graph (UG).
- We say that $p(x_V)$ factorizes over G if:
  - there is a set of cliques $\mathscr{C}$ of G and
  - non-negative real-valued functions $\psi_C(x_C) \geq 0$ for every $C \in \mathscr{C}$ (dependent on $x_C = (x_v)_{v \in C}$) such that:
  - $p(x_V) \propto \prod_{c \in \mathscr{C}} \psi_c(x_c)$

"factors"
"potentials" $[\pm \log \psi_c(x_c))$

- W.l.o.g. we can assume $\mathscr{C}$ to be the set of all maximal cliques $\mathscr{C}(G)$ of G.

Markov Random Fields (MRF) - Definition

- A Markov Random Field (G, p) by definition consists of:
  - an (index) ser V be of random variables $X_1, \ldots, X_M$ with a joint distribution $p(x_V)$,
  - a undirected graph (UG) structure $G = (V, E)$, such that:
  - $p(x_V)$ factorizes over G.
- A MRF (G, p) is called positive or Gibbs Random Field (GRF) if $p(x_V) > 0$ for all possible values x_V.

Machine Learning 2

Graphical Models
- Markov Random Fields
- Global Markov Property for MRFs

Patrick Forré

Separation in Undirected Graphs

- Let $G = (V, E)$ be an undirected graph (UG) and $C \subseteq V$ a subset.
  - A path $v = v_0 - v_1 - \cdots - v_n = w$ between nodes $v, w \in V$ in G is called C-blocked or blocked by C if any of the nodes lies in C.
  - Otherwise the path is called C-open.
- Let $A, B, C \subseteq V$ be any subsets of the nodes of G.
  - We then say that A and B are separated by C if every path from a node $v \in A$ to a node $w \in B$ in G is blocked by C.
  - In symbols: $A \perp B \mid C$

$(A \perp B :\Leftrightarrow A \perp B \mid \emptyset)$

Exercise: Separoid Axioms for Separation

- Let $G = (V, E)$ be an UG and $A, B, C, F \subseteq V$ be subset of nodes.
- Redundancy: $A \perp B \,|\, A$ always holds.
- Symmetry: $A \perp B \,|\, C \iff B \perp A \,|\, C$
- Decomposition: $A \perp B \cup F \,|\, C \implies A \perp B \,|\, C$
- Weak Union: $A \perp B \cup F \,|\, C \implies A \perp F \,|\, B \cup C$
- Contraction: $A \perp F \,|\, B \cup C \;\wedge\; A \perp B \,|\, C \implies A \perp B \cup F \,|\, C$
- So we have the equivalence:

$$A \perp B \cup F \,|\, C \iff A \perp F \,|\, B \cup C \;\wedge\; A \perp B \,|\, C$$

Theorem - Global Markov Property for MRF

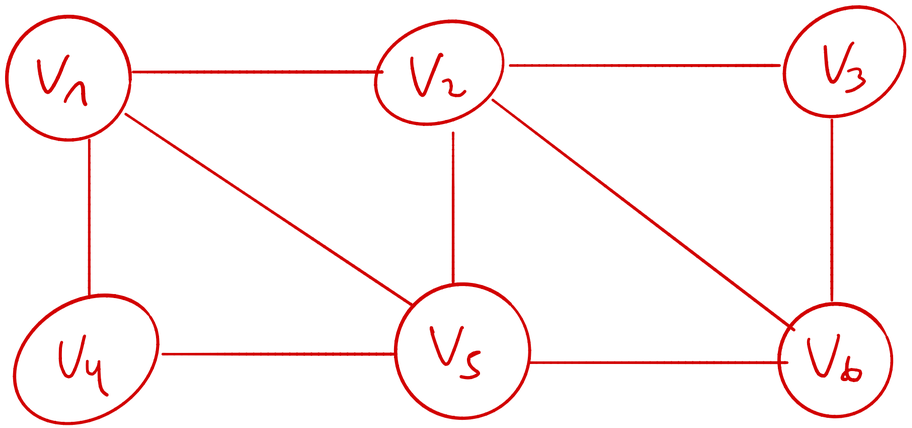
- Let (G, p) be a Markov Random Field (MRF) with undirected graph $G = (V, E)$ and random variables $X_1, \ldots, X_M$ with joint distribution $p(x_V)$. For a set of nodes $F \subseteq V$ we write $X_F = (X_v)_{v \in F}$.
- For any three subsets $A, B, C \subseteq V$ of nodes we have the implication:

$$A \perp B \,|\, C \implies X_A \perp\!\!\!\perp X_B \,|\, X_C.$$

  i.e. separation implies corresponding conditional independence.

- The reverse implication is NOT true in general.

Examples: Separation in MRF

- 

$\{V_1\} \not\perp \{V_i\} \quad \forall i \neq 1$

$\{V_1\} \perp \{V_6\} \mid \{V_2, V_5\}$

$\Rightarrow X_1 \perp\!\!\!\perp X_6 \mid X_2, X_5$

$\{V_3\} \perp \{V_5\} \mid \{V_6, V_2\}$

$\Rightarrow X_3 \perp\!\!\!\perp X_5 \mid X_6, X_2$

Hammersley-Clifford Theorem

- Let $G = (V, E)$ be an undirected graph and $p(x_V) > 0$ a strictly positive joint distribution such that:
  - (Pairwise Markov Property): For all non-adjacent nodes $v, w \in V$ in G the conditional independences hold: $X_v \perp\!\!\!\perp X_w \,|\, X_{V \setminus \{v,w\}}$.
- Then $p(x_V)$ factorized over G.
  - Thus (G, p) is a (positive) Gibbs/Markov Random Field.

- In short: MRF => Global MP => Pairwise MP; + reverse if $p(x_V) > 0$.

Machine Learning 2

Graphical Models
- Markov Random Fields
- Examples of MRFs

Patrick Forré

Example: Gaussian MRF

- Consider an M-dimensional Gaussian $p(x \mid \mu, \Omega^{-1})$, where:
  - $\Omega = \Sigma^{-1}$ is the precision matrix.
- Let $G = (V, E)$ be an undirected graph with:
  - $V = \{1, \ldots, M\}$ and $\{v - w \mid v \neq w, \Omega_{v,w} \neq 0\} \subseteq E$.
- Then $(G, p(x \mid \mu, \Omega^{-1}))$ is a positive MRF:

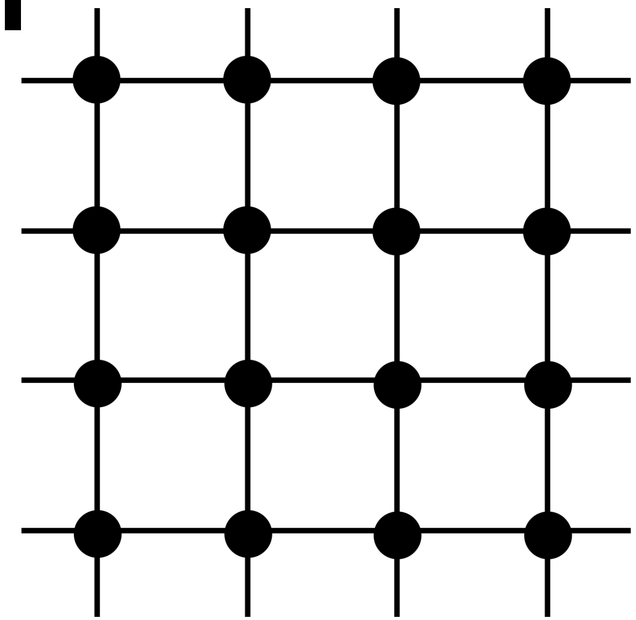
$$p(x \mid \mu, \Omega^{-1}) \propto \exp\left[-\tfrac{1}{2}(x-\mu)^T \Omega (x-\mu)\right]$$

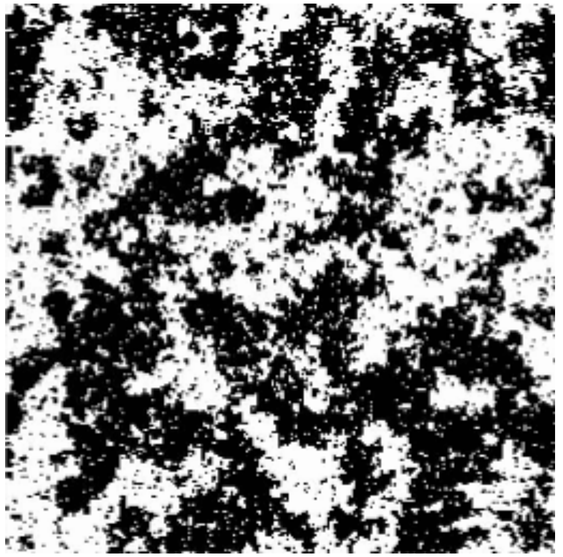
$$= \prod_{v \in V} \exp\left(-\tfrac{1}{2}(x_v-\mu_v)^2 \cdot \Omega_{v,v}\right) \cdot \prod_{v-w \in E} \exp\left(-(x_v-\mu_v)\cdot(x_w-\mu_w)\cdot\Omega_{v,w}\right)$$

- It follows: $\Omega_{v,w} = 0$ implies $X_v \perp\!\!\!\perp X_w \mid X_{V \setminus \{v,w\}}$.

Example: Ising Model

- Let $G = (V, E)$ be a lattice.
- Binary random variables $X_v \in \{\pm 1\}$ with:
- $p(x_V) \propto \exp\left(\sum_{v\text{-}w \in E} J_{v,w} \cdot x_v \cdot x_w + \sum_{v \in W} h_v \cdot x_v \right)$





Positive discrete MRFs as exponential family

- Let $G = (V, E)$ be an undirected graph and for $v \in V$ fix a finite state space $\mathcal{X}_v$. Put $\mathcal{X}_A = \times_{v \in A} \mathcal{X}_v$, which is also finite.
- Consider all positive MRFs on $\mathcal{X}_V$ over G:

$$p(x_V \mid \Psi) = \frac{1}{Z(\Psi)} \prod_{C \in \mathscr{C}} \psi_C(x_C), \quad \text{with factors } \Psi = (\psi_C)_{C \in \mathscr{C}}.$$

$$p(x_V|\eta) = \exp\left[\sum_{C \in \mathscr{C}} \eta_C^T \cdot T_C(x_V) - A(\eta)\right]$$

- These form an exponential family with:
  - natural parameters: $\eta_C = (\log \psi_C(z_C))_{z_C \in \mathcal{X}_C}$ for $C \in \mathscr{C}$,
  - sufficient statistics: $T_C(x_V) = (\mathbb{1}_{z_C}(x_C))_{z_C \in \mathcal{X}_C}$ (one-hot encoding all values z_C)
  - $A(\eta) = \log Z(\Psi(\eta)),\ h(x_V) = 1$

Machine Learning 2

Graphical Models
- Markov Random Fields
vs Bayesian Networks

Patrick Forré

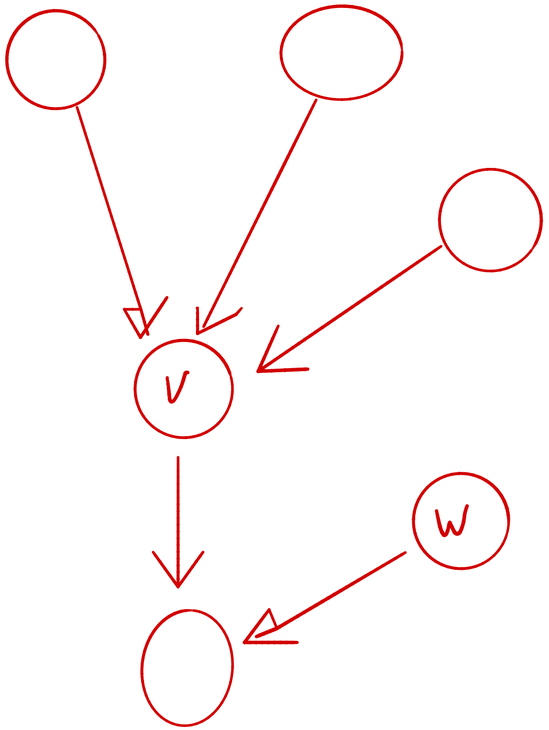
Converting BNs into MRFs via Moralization

- Let $G = (V, E)$ be a DAG and (G, p) be a Bayesian Network (BN):

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{Pa}^G(v)})$$

- To turn this factorization into one for a Markov Random Field (MRF) we define the Moralization of G as the undirected graph $\tilde{G} = (V, \tilde{E})$ that turns the sets $\mathrm{Pa}^G(v) \cup \{v\}$ into cliques:
  - For each $v \in V$ connect all parents $w, w' \in \mathrm{Pa}^G(v)$ with undirected edges.
  - Replace all directed edges with undirected ones.
- Thus: (G, p) BN implies $(\tilde{G}, p)$ MRF.

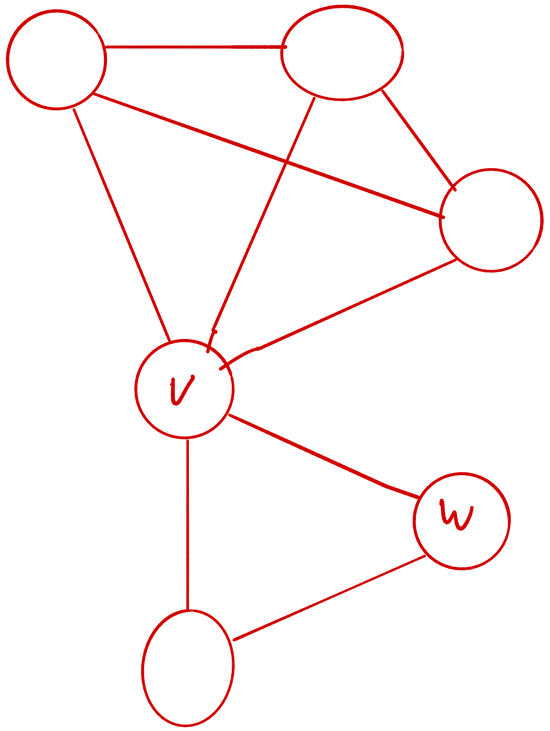
Examples: Moralization



$\{v\} \perp \{w\}$

$\Downarrow$

$X_v \perp\!\!\!\perp X_w$



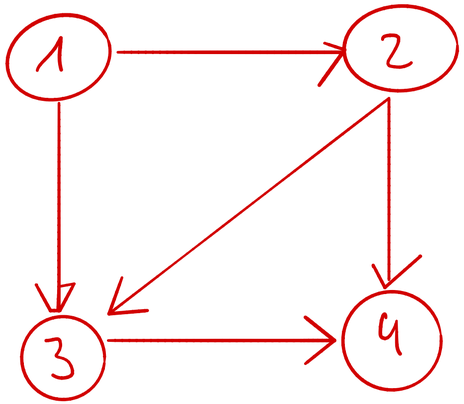
$\{v\} \not\perp \{w\}$

$\not\Downarrow$

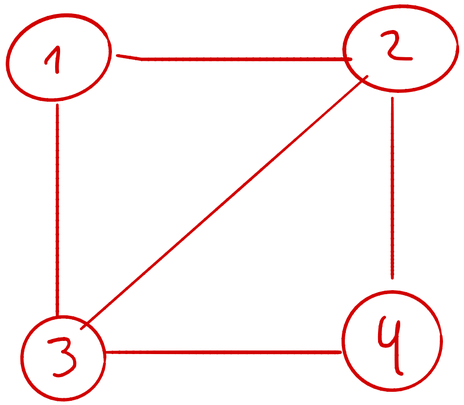
$X_v \perp\!\!\!\perp X_w$?

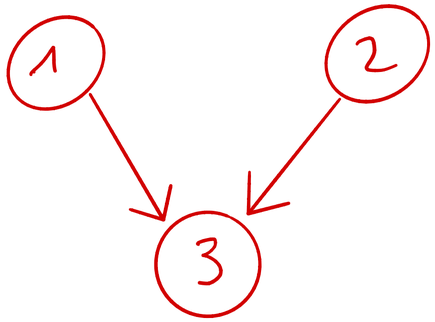
- Note, separation in $\tilde{G}$ may detect less conditional independencies than d-separation in G.

Examples: BNs vs. MRFs

- 

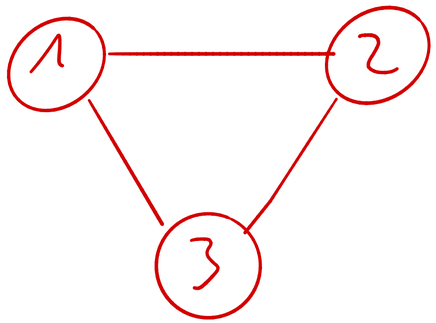
$\{v_1\} \perp \{v_4\} \mid \{v_2, v_3\}$

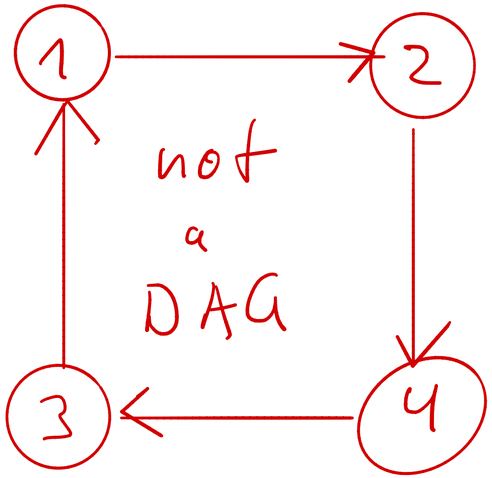




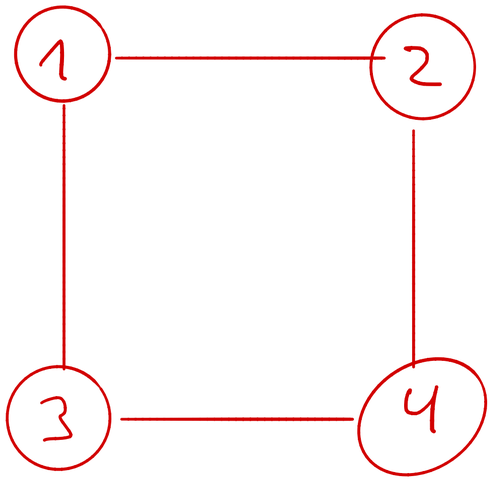
$\{v_1\} \perp \{v_2\}$

$\{v_1\} \not\perp \{v_2\}$





?



From MRFs to BNs via Perfect Elimination Orderings

- Let $G = (V, E)$ be an undirected graph and (G, p) a MRF:
  $p(x_V) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c)$
- IF G has a Perfect Elimination Ordering, i.e. $V = \{v_1, \ldots, v_M\}$ s.t.:
  - for all m we have: $\partial(v_m) \cap \{v_1, \ldots, v_{m-1}\}$ is a clique of G.
- Then we can define the DAG $G' = (V, E')$ by putting:
  - $\mathrm{Pa}^{G'}(v_m) := \partial(v_m) \cap \{v_1, \ldots, v_{m-1}\}$, i.e. draw directed edges to v_m.
- Then one can check that (G', p) is a BN and $\tilde{G}' = G$.

Machine Learning 2

Graphical Models
- Markov Random Fields
- Summary

Patrick Forré

Markov Random Fields - Summary

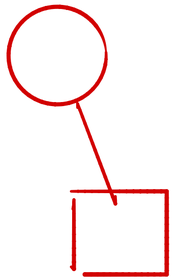
- MRF consists of an UG $G = (V, E)$ and a distribution $p(x_V)$ that factorizes according to the (maximal) cliques of G: $p(x_V) \propto \prod_{C \in \mathscr{C}} \psi_C(x_C)$.
- Separation: $A \perp B \,|\, C$ iff every path from A to B contains a node in C.
- Global Markov Property for MRFs: $A \perp B \,|\, C \implies X_A \perp\!\!\!\perp X_B \,|\, X_C$.
- Hammersley-Clifford: Pairwise MP + $p > 0 \implies (G, p)$ MRF.
- Moralization (connecting parents) turns BNs into MRFs.
- Perfect Elimination Orderings turn MRFs into BNs (in case of existence).
- BNs and MRFs in general model different conditional independence relations. Examples: Collider + 4-Cycle.

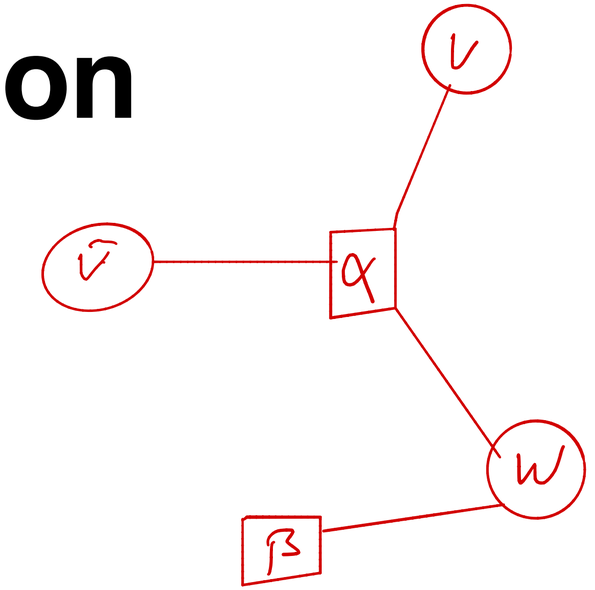
Machine Learning 2

Graphical Models
- Factor Graphs

Patrick Forré

Bipartite Graphs - Definition





- A Bipartite Graph $G = (V, F, E)$ consists of:
  - a set of nodes/vertices V,
  - another set of nodes/factors F,
  - a set of undirected edges E that connect nodes from V with nodes from F.
- The Neighbours of $v \in V$ is the set $\partial(v) = \{\alpha \in F \mid v - \alpha \in E\}$
- The Neighbours of $\alpha \in F$ is the set $\partial(\alpha) = \{v \in V \mid \alpha - v \in E\}$

Factor Graphs - Definition

- A Factor Graph (FG) consists of a bipartite graph $G = (V, F, E)$ and a probability distribution that factorizes as follows:

$$p(x_V) \propto \prod_{\alpha \in F} \psi_\alpha(x_\alpha)$$

where $x_\alpha := x_{\partial(\alpha)} := (x_v)_{v \in \partial(\alpha)}$ for $\alpha \in F$.

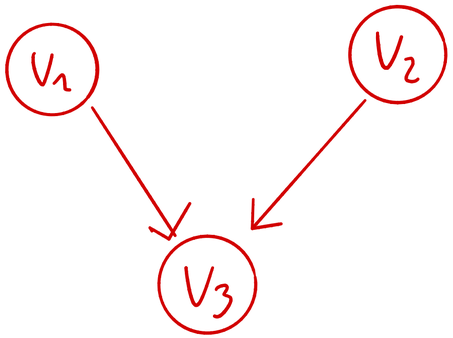
Converting BN into MRF into FG

- Let G be a DAG or UG and (G, p) be a BN or a MRF, resp., with a factorization:

$$p(x_V) \propto \begin{cases} \prod_{v \in V} p(x_v \mid x_{pa(v)}) & \psi_\alpha \\ \prod_{c \in \mathcal{C}} \psi_c(x_c) & \psi_\alpha \end{cases}$$

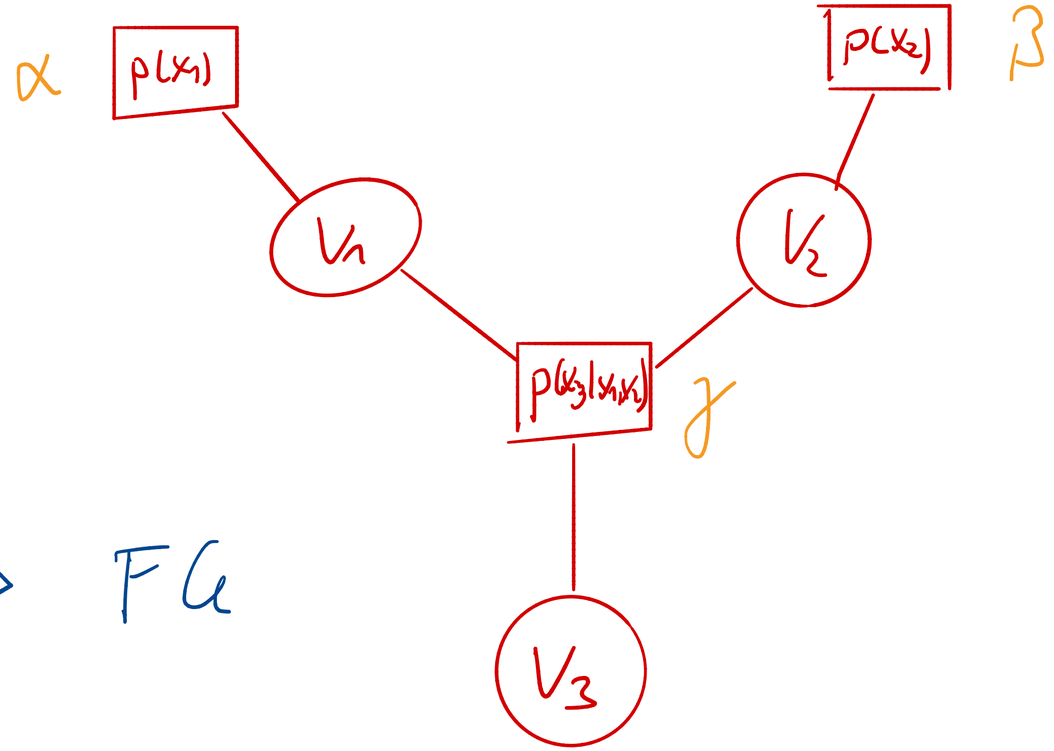
- Define a Factor Graph $G^\bullet = (V, F, E)$ with
  - $F = \{ \psi_c \mid c \in \mathcal{C} \}$ or $\{ p(x_v \mid x_{pa(v)}) \mid v \in V \}$
  - Define E via: connect $v \in V$ with $\alpha \in F$ iff x_v is an argument for factor α.

Examples: Converting BN into MRF into FG

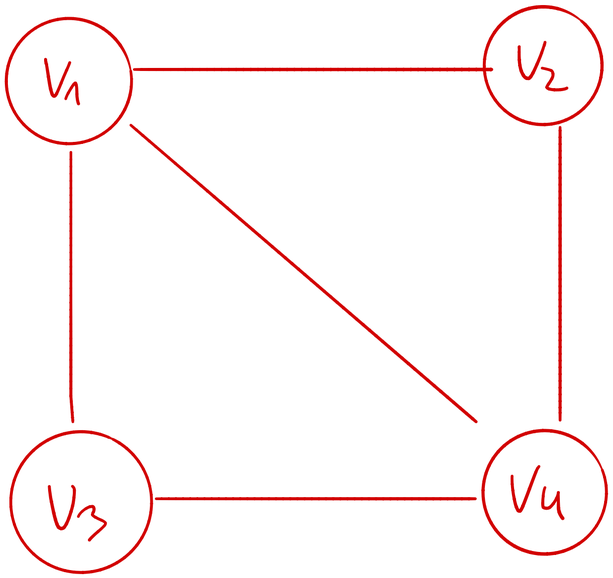
- 

$$p(X_V) = p(x_1) \cdot p(x_2) \cdot p(x_3 | x_1, x_2)$$

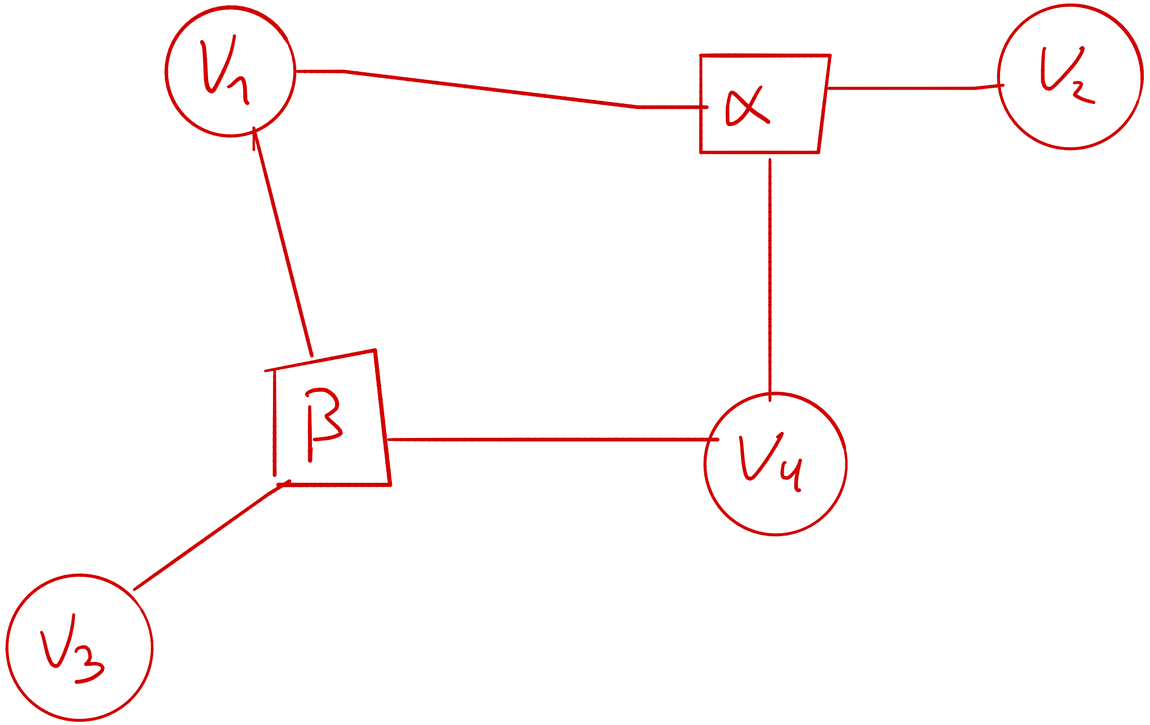
BN $\longrightarrow$ FG



Examples: Converting BN into MRF into FG

- 

MRF ⟶ FG



Machine Learning 2

Graphical Models
- Learning in (discrete) Bayesian Networks

Patrick Forré

MLE in (discrete) Bayesian Networks

- Let (G, p) be a BN with finite state space $\mathcal{X}_V$ and $D = \{x^{(1)}, \ldots, x^{(N)}\}$ i.i.d. data drawn from p.
- We consider G known, but want to learn all $p(x_v \mid x_{\mathrm{Pa}_G(v)})$ from D.
- Parameters: $\theta_v(z_v, z_{\mathrm{Pa}(v)})$, $v \in V$, $z_v \in \mathcal{X}_v$, $z_{\mathrm{Pa}(v)} \in \mathcal{X}_{\mathrm{Pa}(v)}$
- Likelihood:

$$p(D \mid \theta) = \prod_{n=1}^{N} \prod_{v \in V} \left(\prod_{z_v} \prod_{z_{\mathrm{Pa}(v)}} \theta_v(z_v, z_{\mathrm{Pa}(v)}) \right)^{\left[\mathbb{1}_{z_v}(x_v^{(n)}) \, \mathbb{1}_{z_{\mathrm{Pa}(v)}}(x_{\mathrm{Pa}(v)}^{(n)}) \right]}$$

- Lagrangian:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \sum_{v \in V} \sum_{z_v} \sum_{z_{\mathrm{Pa}(v)}} \mathbb{1}_{z_v}(x_v^{(n)}) \cdot \mathbb{1}_{z_{\mathrm{Pa}(v)}}(x_{\mathrm{Pa}(v)}^{(n)}) \cdot \log \theta_v(z_v, z_{\mathrm{Pa}(v)}) - \sum_{v \in V} \sum_{z_{\mathrm{Pa}(v)}} \lambda_{v, z_{\mathrm{Pa}(v)}} \left[\sum_{z_v} \theta_v(z_v, z_{\mathrm{Pa}(v)}) - 1 \right]$$

constraint $\sum_{z_v \in \mathcal{X}_v} \theta_v(z_v, z_{\mathrm{Pa}(v)}) = 1$

$\theta_v(x_v^{(n)}, x_{\mathrm{Pa}(v)}^{(n)})$

MLE in (discrete) Bayesian Networks

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \sum_{v \in V} \sum_{z_v} \sum_{z_{Pa(v)}} \mathbb{1}_{z_v}(x_v^{(n)}) \cdot \mathbb{1}_{z_{Pa(v)}}(x_{Pa(v)}^{(n)}) \cdot \log \theta_v(z_v, z_{Pa(v)}) - \sum_{v \in V} \sum_{z_{Pa(v)}} \lambda_{v, z_{Pa(v)}} \left[\sum_{z_v} \theta_v(z_v, z_{Pa(v)}) - 1 \right]$$

MLE $\rightsquigarrow$ $\theta_{MLE}^{*} \in \arg\max_{\theta} \mathcal{L}(\theta)$

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial \theta_v(z_v, z_{Pa(v)})} = \frac{N(z_v, z_{Pa(v)})}{\theta_v(z_v, z_{Pa(v)})} - \lambda_{v, z_{Pa(v)}}$$

$$N(z_{Pa(v)}) = \sum_{z_v} N(z_v, z_{Pa(v)}) = \lambda_{v, z_{Pa(v)}} \cdot \underbrace{\sum_{z_v} \theta_v(z_v, z_{Pa(v)})}_{=1} = \lambda_{v, z_{Pa(v)}}$$

$$\Rightarrow \theta_v(z_v, z_{Pa(v)}) = \frac{N(z_v, z_{Pa(v)})}{N(z_{Pa(v)})}$$

← as expected from simple frequentist view.

Machine Learning 2

Graphical Models
- Learning in (positive discrete) Markov Random Fields

Patrick Forré

Positive discrete MRFs as exponential family

- Let $G = (V, E)$ be an undirected graph and for $v \in V$ fix a finite state space $\mathcal{X}_v$. Put $\mathcal{X}_A = \times_{v \in A} \mathcal{X}_v$, which is also finite.
- Consider all positive MRFs on $\mathcal{X}_V$ over G:

$$p(x_V \mid \Psi) = \frac{1}{Z(\Psi)} \prod_{C \in \mathscr{C}} \psi_C(x_C), \quad \text{with factors } \Psi = (\psi_C)_{C \in \mathscr{C}}.$$

- These form an exponential family: $p(x_V \mid \eta) = \exp\left(\sum_{C \in \mathscr{C}} \eta_C^\top T_C(x_V) - A(\eta) \right)$
  - natural parameters: $\eta_C = (\log \psi_C(z_C))_{z_C \in \mathcal{X}_C}$ for $C \in \mathscr{C}$,
  - sufficient statistics: $T_C(x_V) = (\mathbb{1}_{z_C}(x_C))_{z_C \in \mathcal{X}_C}$ (one-hot encoding for all values z_C)
  - $A(\eta) = \log Z(\Psi(\eta)), \quad h(x_V) = 1$

$\psi_C \mapsto (\psi_C(z_C))_{z_C \in \mathcal{X}_C}$

MLE in (discrete positive) Markov Random Field

- Let (G, p) be a MRF with finite state space $\mathcal{X}_V$ and $D = \{x^{(1)}, \ldots, x^{(N)}\}$ i.i.d. data drawn from strictly positive $p > 0$.
- We consider G known, but want to learn all η_{C,z_C} from D.
- Parameters: $\eta = (\eta_{C,z_C} \mid C \in \mathcal{C},\ z_C \in \mathcal{X}_C)$
- Likelihood: $p(D|\eta) = \prod_{n=1}^{N} \exp\left[\sum_{C \in \mathcal{C}} \eta_C^T T_C(x^{(n)}) - A(\eta)\right]$
- Lagrangian:

$$\mathcal{L}(\eta) = \log p(D|\eta) = \sum_{n=1}^{N} \sum_{C \in \mathcal{C}} \sum_{z_C \in \mathcal{X}_C} \eta_{C,z_C} \cdot \mathbb{1}_{z_C}(x_C^{(n)}) - N \cdot A(\eta)$$

MLE in (discrete positive) Markov Random Field

$$\mathcal{L}(\eta) = \log p(D|\eta) = \sum_{n=1}^{N} \sum_{c \in C} \sum_{z_c \in \mathcal{X}_c} \eta_{c,z_c} \cdot \mathbb{1}_{z_c}(x_c^{(n)}) - N \cdot A(\eta)$$

$$\eta_{MLE} = \arg\max_{\eta} \mathcal{L}(\eta)$$

$$\mathbb{E}_{\eta}[\mathbb{1}_{z_c}(X_c)] = p(z_c|\eta)$$

$$\Rightarrow 0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial \eta_{c,z_c}} = N(z_c) - N \cdot \frac{\partial A}{\partial \eta_{c,z_c}} = N(z_c) - N \cdot p(z_c|\eta)$$

$$\approx N(z_c) - N \cdot \frac{N_\eta(z_c)}{N_\eta} \quad ?$$

Instead:

Update rule:

$$\eta_{c,z_c}^{new} = \eta_{c,z_c} + \alpha \cdot \left(\frac{N(z_c)}{N} - \frac{N_\eta(z_c)}{N_\eta} \right)$$