PROBABILISTIC GRAPHICAL MODELS.

Lecture 2: FACTOR GRAPHS; MARGINALISATION BY MESSAGE PASSING; SAMPLING (MCMC).

I. Factor graphs are a convenient detailled view of factorisation properties of a pdf. For us they represent mainly a very practical calculational tool for:

→ Marginalization, inference, (sampling).

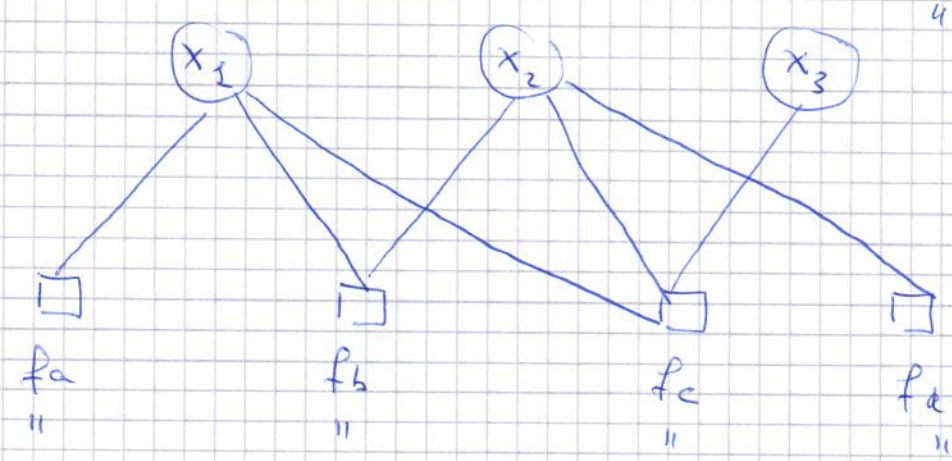
But they do not bring out clearly independence statements and should be viewed as complementing BN and MRF description. Also, when you do modelling you would typically not get the factor graph directly but rather the BN or MRF.

Definition. A distribution is said to have a factor graph representation if it can be written in the form

$$p(\underline{x}) = \frac{1}{Z} \prod_{S \subset V} f_S(x_S)$$

where $Z = \sum_{\underline{x} \in \mathcal{A}^V} \prod_{S \subset V} f_S(x_S)$, and the product runs over a certain number of subsets $S \subset V$ (not necessarily cliques here!) and $x_S \equiv \{x_i\}_{i \in S}$. Here $f_S(x_S) \geq 0$ are not normalized.

→ The factor graph is :

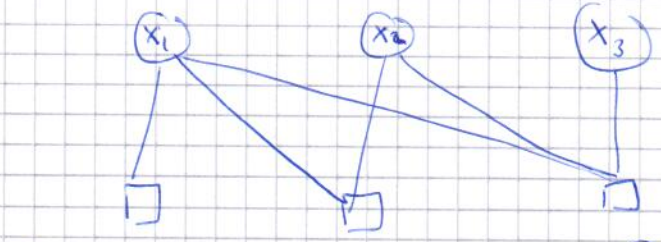


$f_a(x_1)$ $f_b(x_1, x_2)$ $f_c(x_1 x_2 x_3)$ $f_d(x_2)$.

$$p(\underline{x}) = \frac{1}{Z} f_a(x_1) f_b(x_1, x_2) f_c(x_1, x_2, x_3) f_d(x_2)$$

→ Non unique representation :

$$p(\underline{x}) = \frac{1}{Z} f_a(x_1) f_b(x_1, x_2) \underbrace{f_c(x_1, x_2, x_3) f_d(x_2)}_{\tilde{f}_c(x_1, x_2 x_3)}$$



→ Factor graph is Bipartite : Variable Nodes ∪ Constraint or factor Nodes.

II. BN & MRF conversion to Factor Graphs.

II.1 : MRF → factor graph conversion.

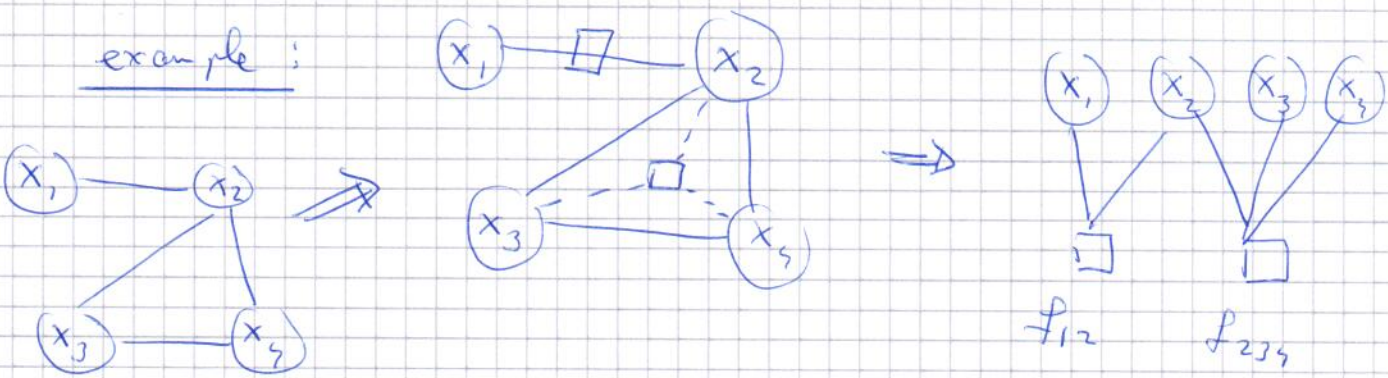
By Hammersley-Clifford thm a MRF can be written as :

$$p(\underline{x}) = \frac{1}{Z} \prod_{\substack{\text{max cliques} \\ C}} \psi_C(x_C) \quad .$$

associate ○ to variables $x_1 \dots x_N$.

associate □ to max cliques C.

example :

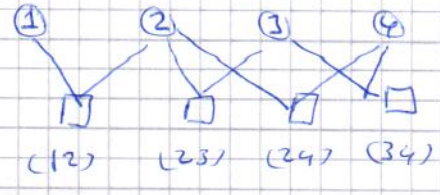


$$p(\underline{x}) = \frac{1}{Z} \underbrace{\psi_{12}(x_1, x_2)}_{f_{12}(x_1, x_2)} \underbrace{\psi_{234}(x_2, x_3, x_4)}_{f_{234}(x_2, x_3, x_4)} .$$

for example

If $f_{234}(x_2, x_3, x_4) = \psi_{23}(x_2, x_3)\psi_{24}(x_2, x_4)\psi_{34}(x_3, x_4)$

the fact graph is not unique.

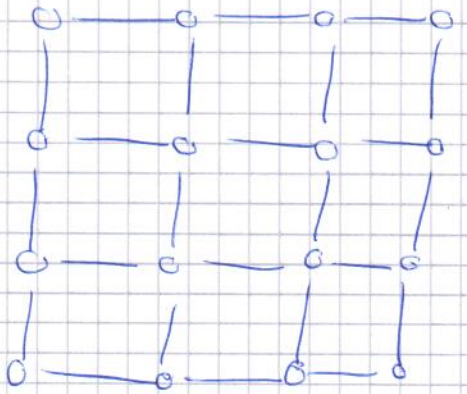


example of the Ising model on square lattice $\mathbb{Z}^2$;

$$p(\Sigma) = \prod_{(ij)\in E} e^{J s_i s_j} \prod_{i\in V} e^{h s_i}.$$

(written as Gibbs distr / not necessarily max cliques.).

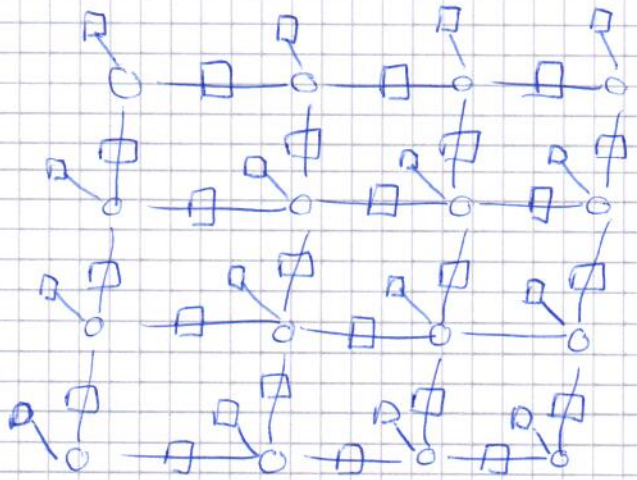
↳ MRF on the graph



max cliques are $(ij) \in E$.

(and $i \in V$ are not max cliques.)

↳ Factor graph



$\square \leftrightarrow e^{J s_i s_j}$, $e^{h s_i}$

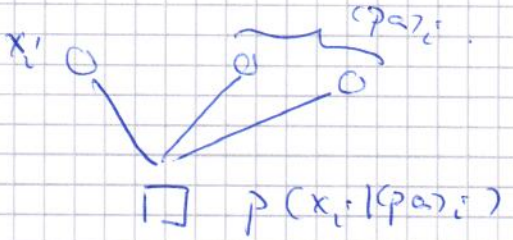
$\circ \leftrightarrow s_i$.

II, 2. BN → Factor graph conversion.

$$p(\underline{x}) = \prod_{i=1}^{N} p(x_i \mid (pa)_i) \quad \text{a BN.}$$

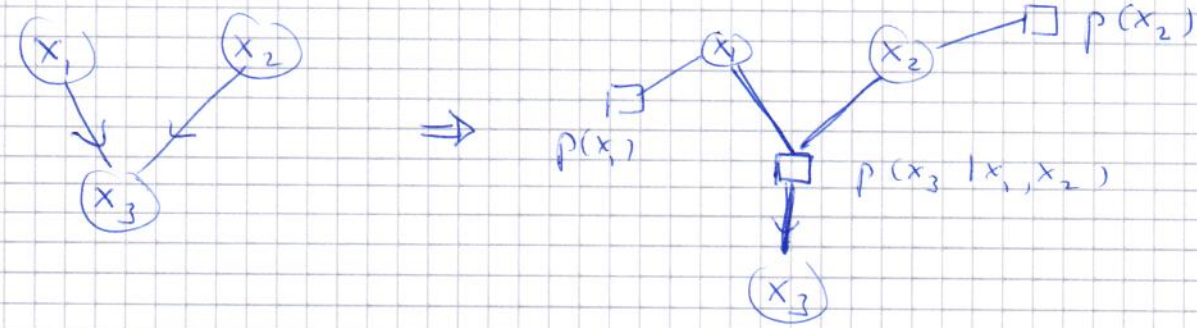
associate ○ to $x_1 \dots x_N$

associate □ to $p(x_i \mid (pa)_i)$.

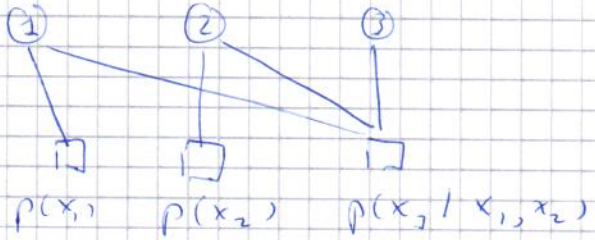


Example.

$$p(\underline{x}) = p(x_1)\, p(x_2)\, p(x_3 \mid x_1, x_2)$$



or



II, 3. Special Case of Trees.

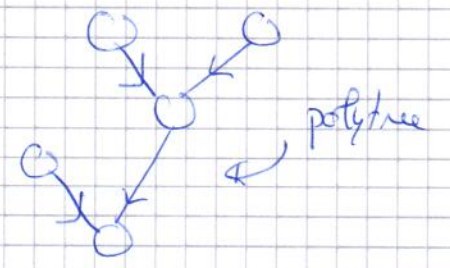
Factor graphs are useful devices to perform marginalization. It turns out marginalization by message passing is exact on trees (see later). So it is interesting to look more precisely when is the factor graph a tree ?.

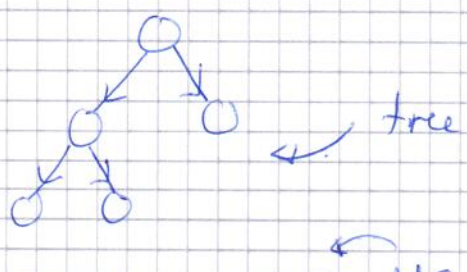
Definition: Fact graph is a tree iff there are no loops in its bipartite graph of variable & factor nodes.

Definition: A MRF has a tree structure if its undirected graph has no loops (basically same definition as for factor graphs).

Definition: A BN is a polytree if the graph (without taking directions of edges into account) is a tree.

If furthermore there is a unique node with no parents (the ancestor of everybody!) + all nodes have a unique parent Then we call the BN a Tree.

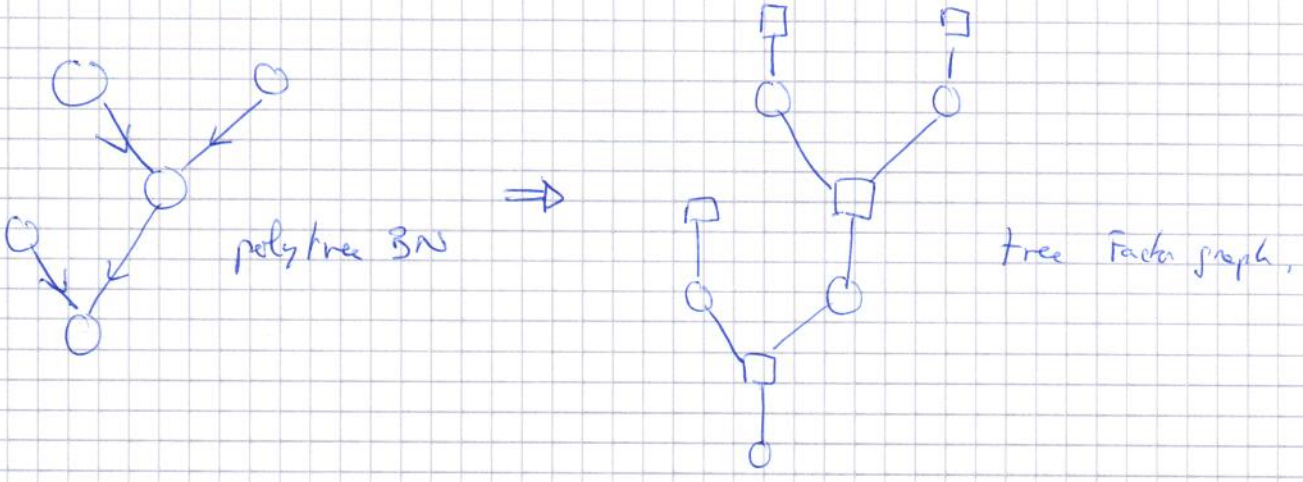




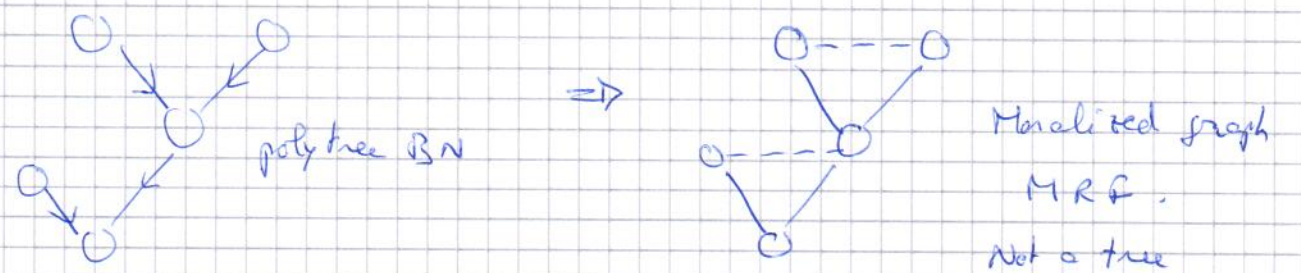
Properties of conversions for trees

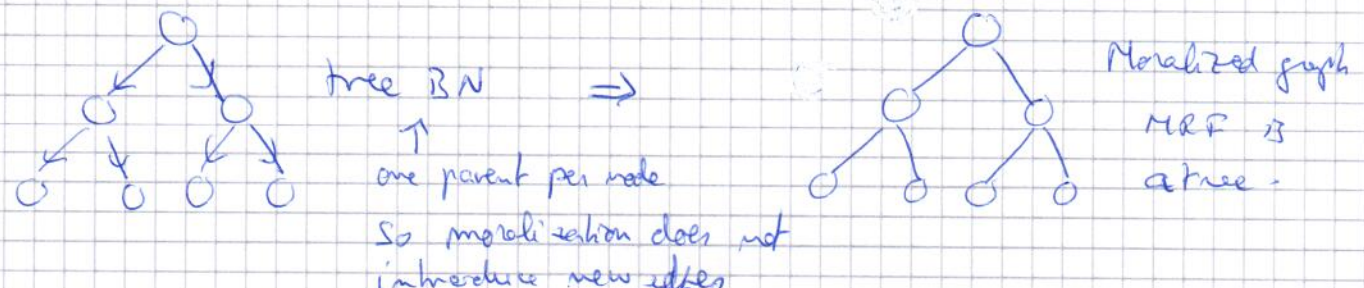
* A tree MRF converted to a Factor graph (say by the canonical conversion above) yields a tree factor graph.

* A polytree BN converted to a Factor graph (say by the canonical conversion above) yields a tree factor graph.



* Remark that by Moralization we converted BN to MRF. But a polytree BN does not become a tree MRF:





III. MARGINALISATION BY MESSAGE PASSING.

Important tasks for learning are:

- Compute marginals $p(x_i)$ & $p(x_i, x_j)$ e.g. in order to fit them to empirical frequencies to learn parameters.
- Marginalize over hidden variables e.g. in RBM in order to learn parameters from samples.
- Learn the underlying graph itself.

The main method for efficient marginalization is by "message passing" / "belief propagation" / "sum product algorithm" (synonyms).

In general this method is approximate and there are no convergence garantees. But on tree factor graphs it is exact. Here we derive the message passing algorithm for trees (and discuss the implementation on general factor graphs after).

[See typed notes on Web page -].

IV SAMPLING FROM PROBABILISTIC GRAPHICAL MODELS.

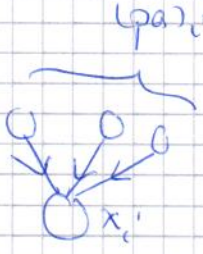
For high-dimensional pdf the elementary sampling methods (rejection sampling, importance sampling) fail or are not well suited or reliable. Our main aim is to introduce the Markov Chain Monte Carlo (MCMC) method and in particular the popular easily applicable sub-case of Gibbs sampling also known as heat bath dynamics or Glauber dynamics.

Before introducing MCMC we briefly discuss the easier topic of sampling from Belief Networks with ancestral sampling.

IV.1 Belief Networks and ancestral sampling.

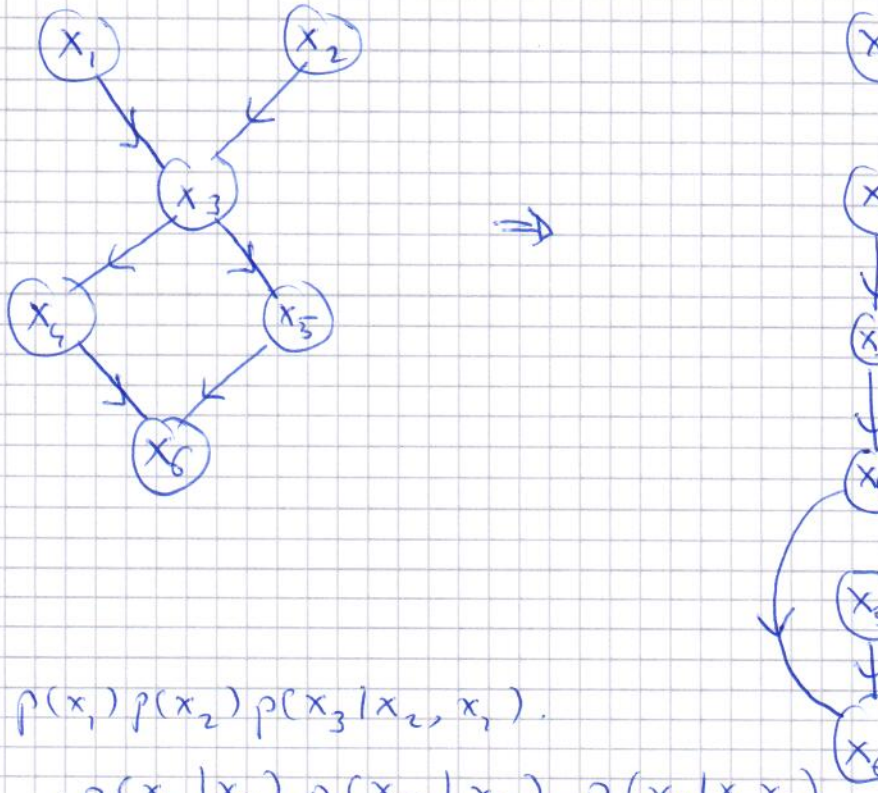
Recall for a BN:

$$p(\underline{x}) = \prod_{i=1}^{K} p(x_i | (pa)_i).$$



We can order variables such that parental variables always come before children (order generations). One can show that this is always possible for a Directed Acyclic Graph (DAG): $(pa)_i < i$ (partial order).

For example :



$$p(\underline{x}) = p(x_1)\, p(x_2)\, p(x_3 | x_2, x_1) \cdot p(x_4 | x_3)\, p(x_5 | x_3)\, p(x_6 | x_4, x_5)$$

We sample in the order $x_1, x_2, x_3, x_4, x_5, x_6$ i.e from ancestors to later generations :

Forward or Ancestral sampling algorithm

$\hat{x}_1 \sim p(x_1)$

$\hat{x}_2 \sim p(x_2)$

$\hat{x}_3 \sim p(x_3 | \hat{x}_2, \hat{x}_1)$

$\hat{x}_4 \sim p(x_4 | \hat{x}_3)$

$\hat{x}_5 \sim p(x_5 | \hat{x}_4)$

$\hat{x}_6 \sim p(x_6 | \hat{x}_4, \hat{x}_5)$

This is an exact procedure. Moreover if we run algorithm m times we have m samples $\underline{x}^{(1)} \dots \underline{x}^{(m)}$ which are all independent.

Remarks :

1) If we have some variable which is evidential, say X_6 is observed to a value $\bar{X}_6$ and we want to sample from $p(x_1 x_2 x_3 x_4 x_5 \mid \bar{x}_6)$. Then we cannot use ancestral sampling as such because

$$p(x_1 x_2 x_3 x_4 x_5 \mid \bar{x}_6) = \frac{p(x_1)\, p(x_2)\, p(x_3 \mid x_2 x_1)\, p(x_4 \mid x_5)\, p(\bar{x}_6 \mid x_4 x_5)}{\sum_{x_1 x_2 x_3 x_4 x_5} (\text{Numerator})}$$

So a priori one would have to compute the marginal $p(\bar{x}_6)$ in the denominator and ancestral sampling is not enough for that. One could still use ancestral sampling combined with rejections each time $\hat{x}_6 \neq \bar{x}_6$. But this in practice turns out to be very costly.

2) In Nice cases it could be that the marginal above in the denominator equals 1. (For example this is the case for $p(x_5) = 1$; exercise!). Then one can use ancestral sampling again for $p(x_1 x_2 x_3 x_4 x_6 \mid \bar{x}_5)$.

3) In a Nutshell if we have a denominator (Normalizing factor) which is Non trivial then it is difficult to use ancestral sampling; and MCMC is very useful even for BN.

IV.2 MCMC Method

As said before this is useful for sampling MRF, Factor graph, and conditional probabilities of BN's. Because of the Markov properties of MRF's it can be often implemented nicely (heat bath dynamics in next paragraph). It is however not an exact method contrary to ancestral sampling and the samples are not strictly iid.

The general idea is to construct a Markov chain s.t $p(\underline{x})$ is the stationary distribution of this MC. Then one runs the MC starting from an initial state: $\underline{x}^{(0)} \to \underline{x}^{(1)} \to \dots \to \underline{x}^{(t)} \to \dots \to \underline{x}^{(T)}$

We hope that for T large enough $\underline{x}^{(T)}$ is approximately distributed according to the stationary distribution $p(\underline{x})$ [This is strictly true for $T = +\infty$].

In practice because $\underline{x}^{(T)}, \underline{x}^{(T+1)}, \underline{x}^{(T+2)}, \dots$ are not independent we take samples every M-th iteration: $\underline{x}^{(T)}, \underline{x}^{(T+M)}, \underline{x}^{(T+2M)}, \dots$ for M large. Or we run the chain many times from many initial conditions starting.

This is costly ---

IV.2.a) Recapitulation on Markov Chains.

Take a finite discrete state space $S = \{1, 2, \ldots, |S|\}$. A MC is a stochastic process $\{\underline{x}^{(t)}, t = 0, 1, 2, \ldots\}$ with $\underline{x}^{(t)} \in S$ satisfying

$$q(\underline{x}^{(t)} / \underline{x}^{(t-1)} \ldots \underline{x}^{(0)}) = q(\underline{x}^{(t)} / \underline{x}^{t-1})$$

We consider homogeneous (in time) chains for which

$$q(\underline{x}^{t} = j / \underline{x}^{t-1} = i) = Q_{i \to j} = Q_{ij} \quad \text{indep of } t.$$

Q_{ij} is a $|S| \times |S|$ matrix called the transition matrix.

Note that $\sum_{j=1}^{|S|} Q_{ij} = 1$. [for i you go somewhere with probability one].

Let $\Pi_t(\underline{x}^t)$ be the probability of the state at time t.

It is easy to show (the Chapman-Kolmogorov equation)

$$\Pi_t(\underline{x}^t) = \sum_{\underline{x}^{t-1} \in S} q(\underline{x}^t / \underline{x}^{t-1}) \Pi_{t-1}(\underline{x}^{t-1})$$

Definition: A stationary distribution is a distribution satisfying

$$\Pi_{stat}(\underline{x}) = \sum_{\underline{x}' \in S} q(\underline{x} / \underline{x}') \Pi_{stat}(\underline{x}').$$

It may not always exist. If it exists it may not be unique.

Definition: A MC is called irreducible if the transition matrix Q connects any state i & j in finite time and in both directions: $\exists\, n_0, n_1$ s.t $(Q^{n_0})_{i\to j} \neq 0$ and $(Q^{n_1})_{j\to i} \neq 0$.

Theorem: For an irreducible MC on a finite state space there always exist a unique stationary distribution.

Definition: We say that an irreducible MC on a finite state space is ergodic if for all $\underline{x} \in S$ we have

$$\pi_t(\underline{x}) \to \pi_{stat}(\underline{x}) \quad \text{as} \quad t \to +\infty,$$

Theorem: an irreducible and aperiodic MC on a finite state space is ergodic. [A criterion for aperiodicity is that $\exists$ self loops or Q has not the eigenvalue -1 in spectrum].

Definition: A MC is said to satisfy the detailed balance condition. If its stationary distribution satisfies

$$q(\underline{x}|\underline{x}')\,\pi_{stat}(\underline{x}') = q(\underline{x}'|\underline{x})\,\pi_{stat}(\underline{x}).$$

[i.e Mass transfer $\underline{x}' \to \underline{x}$ = Mass transfer $\underline{x}' \to \underline{x}$].

Remark : any distribution that satisfies the detailed balance condition must be stationary. Indeed :

$$q(\underline{x}|\underline{x}')\pi(\underline{x}') = q(\underline{x}'|\underline{x})\pi(\underline{x})$$

$$\Rightarrow \underbrace{\sum_{\underline{x}} q(\underline{x}|\underline{x}')}_{=1}\pi(\underline{x}') = \sum_{\underline{x}} q(\underline{x}'|\underline{x})\pi(\underline{x})$$

$$\Rightarrow \pi(\underline{x}') = \sum_{\underline{x}} q(\underline{x}'|\underline{x})\pi(\underline{x})$$

which is the definition of stationarity.

IV.2.b The Metropolis-Hasting MCMC method.

Let $p(\underline{x})$ a distr from which we want to sample. We construct $q(\underline{x}|\underline{x}')$ s.t the MC is ergodic and has stat distr $p(\underline{x})$.

Metropolis-Hastings Method :

Take a "base chain" or "proposal chain" $\tilde{q}(\underline{x}|\underline{x}')$ over state space S. We take $\tilde{q}$ st this is an irreducible chain.

- Start chain at $\underline{x}^{(0)}$ at time $t=0$.
- At time t generate $\underline{x}^{t+1} \sim \tilde{q}(\cdot|\underline{x}^t)$ a "proposal"
- Accept $\underline{x}^{t+1}$ with probability $A(\underline{x}^{t+1}, \underline{x}^t)$ where :

$$A(\underline{x}^{t+1}, \underline{x}^{t}) = \min\left[1, \frac{\tilde{q}(\underline{x}^{t}/\underline{x}^{t+1})\, p(\underline{x}^{t+1})}{\tilde{q}(\underline{x}^{t+1}/\underline{x}^{t})\, p(\underline{x}^{t})}\right]$$

- If $\underline{x}^{t+1}$ is accepted it forms the new state. If it is not accepted keep the state $\underline{x}^{t}$.
- iterate.

Properties of the Metropolis-Hastings chain (Not difficult to show).

(i) It is irreducible (because $\tilde{q}$ is)

(ii) It is aperiodic (because rejection probability introduces self-loops).

(iii) On a finite state space by the previous thm $\exists$ unique stationary distn and the chain is ergodic i.e

$$\pi_t(\underline{x}) \to \pi_{stat}(\underline{x}) = p(\underline{x}) \quad \forall\ \underline{x} \in S.$$

(iv) The chain satisfies detailed balance condition.

Special Case of Metropolis-Hastings (called Metropolis).

For a symmetric proposal chain $\tilde{q}(\underline{x}|\underline{x}') = \tilde{q}(\underline{x}'|\underline{x})$ we have $A(\underline{x}', \underline{x}) = \min\left(1, \frac{p(\underline{x}')}{p(\underline{x})}\right)$.

The new sample $\underline{x}'$ is accepted with probability 1 if it is more probable: $p(\underline{x}') \geq p(\underline{x})$. It is accepted with lower probability $\frac{p(\underline{x}')}{p(\underline{x})}$ if it is less probable $p(\underline{x}') < p(\underline{x})$.

Proof of detailed balance condition for the Metropolis-Hastings chain

We should show $q(\underline{x}'|\underline{x})\, p(\underline{x}) = q(\underline{x}|\underline{x}')\, p(\underline{x}')$

For $\underline{x} = \underline{x}'$ this is trivial.

For $\underline{x} \neq \underline{x}'$: $q(\underline{x}'|\underline{x}) = \underbrace{A(\underline{x}', \underline{x})}_{\text{acceptance prob}} \underbrace{\tilde{q}(\underline{x}'|\underline{x})}_{\text{proposal prob}}$

So we should check if

$$A(\underline{x}', \underline{x})\, \tilde{q}(\underline{x}'|\underline{x})\, p(\underline{x}) \overset{?}{=} A(\underline{x}, \underline{x}')\, \tilde{q}(\underline{x}|\underline{x}')\, p(\underline{x}')$$

i.e. $\min\left(\tilde{q}(\underline{x}'|\underline{x})p(\underline{x}); \tilde{q}(\underline{x}|\underline{x}')p(\underline{x}')\right) \overset{?}{=} \min\left(\tilde{q}(\underline{x}|\underline{x}')p(\underline{x}'); \tilde{q}(\underline{x}'|\underline{x})p(\underline{x})\right)$

which is obviously true. ∎

A.3 Gibbs Sampling.

(In physics goes back to Glauber 60's ; in computer science goes back to Geman & Geman 84 ; also known as heat-bath dynamics).

Algorithm for Gibbs sampling.

- Take vertex i at random in $\{1, \dots, N\}$ and make the move $\underline{x} \to \underline{x}'$ where

$$\begin{cases} x_i' \sim p(x_i' \mid \{x_j\}_{j \neq i}) \\ x_j' = x_j \quad \text{for } j \neq i. \end{cases}$$

One can see that this corresponds to a Metropolis-Hastings chain with proposal or bare chain

$$\tilde{q}(\underline{x}' \mid \underline{x}) = \begin{cases} \frac{1}{N} p(x_i' \mid \{x_j\}_{j \neq i}) & \text{if } \underline{x}' \text{ s.t.} \\ & x_j' = x_j \text{ for all } j \neq i \\ & \text{for some } i \\ 0 & \text{otherwise} \end{cases}$$

It is also possible to see that $A(x', x) = 1$.

11.3.a. Application to MRF and factor graphs.

Let $p(\underline{x}) = \frac{1}{Z} \prod_c \psi_c(x_c)$

where c runs over max cliques or factor nodes of a factor graph. Note that $Z = \sum_{\underline{x} \in \mathcal{S}^N} \prod_c \psi_c(x_c)$ is intractable. But note also that it simplifies in ratios $\frac{p(\underline{x}')}{p(\underline{x})}$. Therefore with the Metropolis-Hastings (and Gibbs sampling) methods we do not need to compute Z. Moreover in Gibbs sampling we change one node at a time and the MRF property becomes very handy:

$$\underline{x} = (x_1 \dots x_N) \longrightarrow \underline{x}' = (x_1' \dots x_N') = (x_1 \dots x_{i-1}\ x_i'\ x_{i+1} \dots x_N)$$

where i is drawn at random in $\{1 \dots N\}$ and

$$p(x_i' \mid \{x_j\}_{j \neq i}) = p(x_i' \mid MB(i))$$

The r.h.s can be written more explicitly as

$$\frac{\prod_{c:\, c \ni i} \psi_c(x_c')}{\int dx_i' \prod_{c:\, c \ni i} \psi_c(x_c')} \leftarrow x_c' = \{x_i', x_{c \setminus i}\}.$$

for which we sample by low dim-methods.

Example of the Ising model (or Boltzman machine)

$$p_{Ising}(\underline{s}) = \frac{\exp\left\{\sum_{(i,j)\in E} J_{ij} s_i s_j + \sum_{i\in V} h_i s_i\right\}}{Z}$$

where $s_i \in \{-1, +1\}$. The MC makes the move

$\underline{s} = (s_1 \dots s_N) \to \underline{s}' = (s_1 \dots s_{i-1}\ s_i'\ s_{i+1} \dots s_N)$ for some random $i \in \{1 \dots N\}$ with probability

$$p(s_i' \mid \{s_j\}_{j\neq i}) = p(s_i' \mid (MB)(i))$$

$$= \frac{\exp\left\{\sum_{j\sim i} J_{ij} s_i' s_j + h_i s_i'\right\}}{\sum_{s_i'} \exp\left\{\sum_{j\sim i} J_{ij} s_i' s_j + h_i s_i'\right\}}$$

$$= \frac{\exp\left\{s_i'\left(\sum_{j\sim i} J_{ij} s_j + h_i\right)\right\}}{2\cosh\left(\sum_{j\sim i} J_{ij} s_j + h_i\right)}$$

$$= \frac{1}{2}\left\{1 + s_i' \tanh\left(\sum_{j\sim i} J_{ij} s_j + h_i\right)\right\}$$

Summary of Algorithm:

- Pick i uniformly at random in $\{1 \dots N\}$
- Set $s_i' = \pm 1$ with Prob $= \frac{1}{2}\left\{1 \pm \tanh\left(\sum_{j\sim i} J_{ij} s_j + h_i\right)\right\}$
- Iterate

This yields $\underline{s}^{(0)}, \underline{s}^{(1)}, \dots, \underline{s}^{(T)}$. For $T \to +\infty$ we have $\underline{s}^{(T)} \sim p_{Ising}(\underline{s})$.