2 points related to viterbi (max product)

1) $\arg\max_y \; w^T \underbrace{\phi(x, y)}$

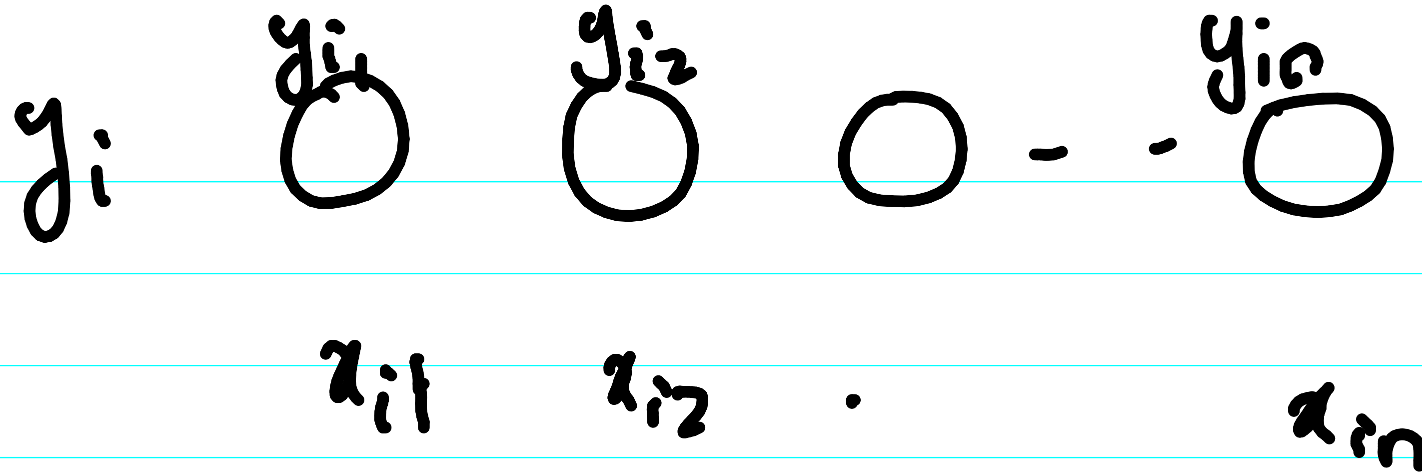
can be computed efficiently using max product

labels = m
length of sequence = n $\Big\} \; O(mn^2)$

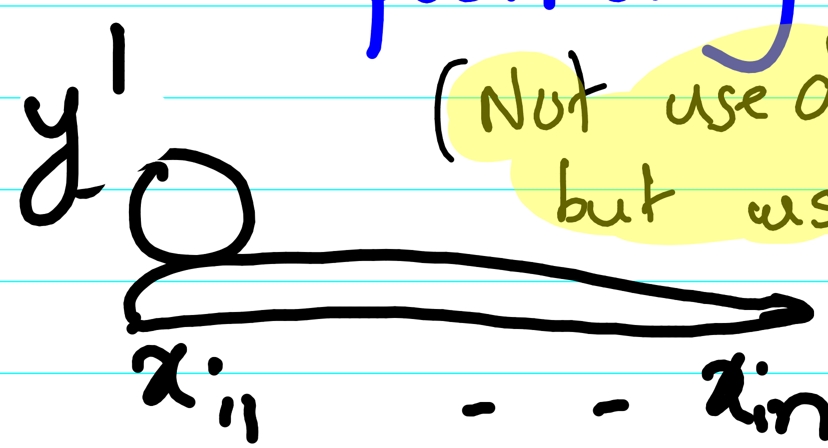
2) Can you use some dynamic programming even while computing $w^T \phi(x_i, y_i)$ across diff i's

$O(n)$

Both known



3) belief that label y_j appears in 1st position of sequence



(Not used in STRUCTSUM but used by models such as HMM & CRF & useful in general)

I am proposing:

$$\mu_i(y') = \sum_{y \text{ s.t. } y[i]=y'} w^T\phi(x_i, y)$$: H/W write DP steps

Hint: Just as

Max product : $\max_{a,b} g(a)f(a,b) = \max_a g(a) \underbrace{\max_b f(a,b)}_{\mu_b(a)}$

Sum product $\sum_{a,b} g(a)f(a,b) = \sum_a g(a) \underbrace{\sum_b f(a,b)}_{\nu_b(a)}$

We will now study more complex settings

http://www.cse.iitb.ac.in/~cs717/notes/classNotes/StructuredOutput/STRUCTSVMWITHKERNELS.pdf

I. STRUCTSVM with arbit kernels on i/p space

II. o/p Structure more complex such as DAG's, cyclic graphs, higher order relations (MLN, BLP)

III. Constrained Conditional models: Adding complexity to runtime constraints instead of adding a complexity to features

IV. Composing features to give "interpretable" composite features instead of a number as o/p by kernels

II, IV: Require a bit of Submodular optimisation

This formidable set of $Y \neq Y_i$ & $Y' \neq Y_j$ have jumped from constraints to the objective

$$\max_{\alpha} \sum_{i, Y \neq Y_i} \alpha_{iY} - \frac{1}{2} \underbrace{\sum_{i, Y \neq Y_i} \sum_{j, Y' \neq Y_j}}_{\text{2 summations}} \alpha_{iY} \alpha_{j,Y'} \langle \underbrace{\psi_i^{\delta}(Y), \psi_j^{\delta}(Y')}_{\text{1 summation}} \rangle$$

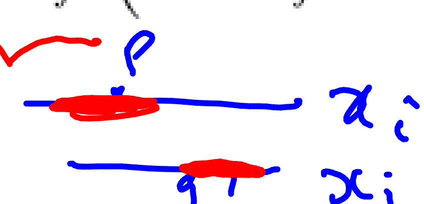
such that,

$$\forall i, \forall Y \neq Y_i : \quad \alpha_{iY} \geq 0$$
$$\forall i : \quad n \sum_{Y \neq Y_i} \frac{\alpha_{iY}}{\Delta(Y_i, Y)} \leq C$$

Recall: $\psi_i^{\delta}(Y) = \overbrace{\psi(X_i, Y_i)}^{\text{desired features}} - \overbrace{\psi(X_i Y)}^{\text{incidental features}}$

$$\langle \mathbf{f}, \psi_i^{\delta}(Y) \rangle = \langle \mathbf{f}, \psi(X_i, Y_i) \rangle - \langle \mathbf{f}, \psi(X_i, Y) \rangle$$

$$\kappa_E^{\delta}\left((X_i, Y_i, Y), (X_j, Y_j, Y')\right) = \sum_{p=1}^{l_i} \sum_{q=1}^{l_j} \kappa_E(x_i^p, x_j^q) \left(\Lambda(y_i^p, y_j^q) + \Lambda(y^p, y'^q) \right)$$

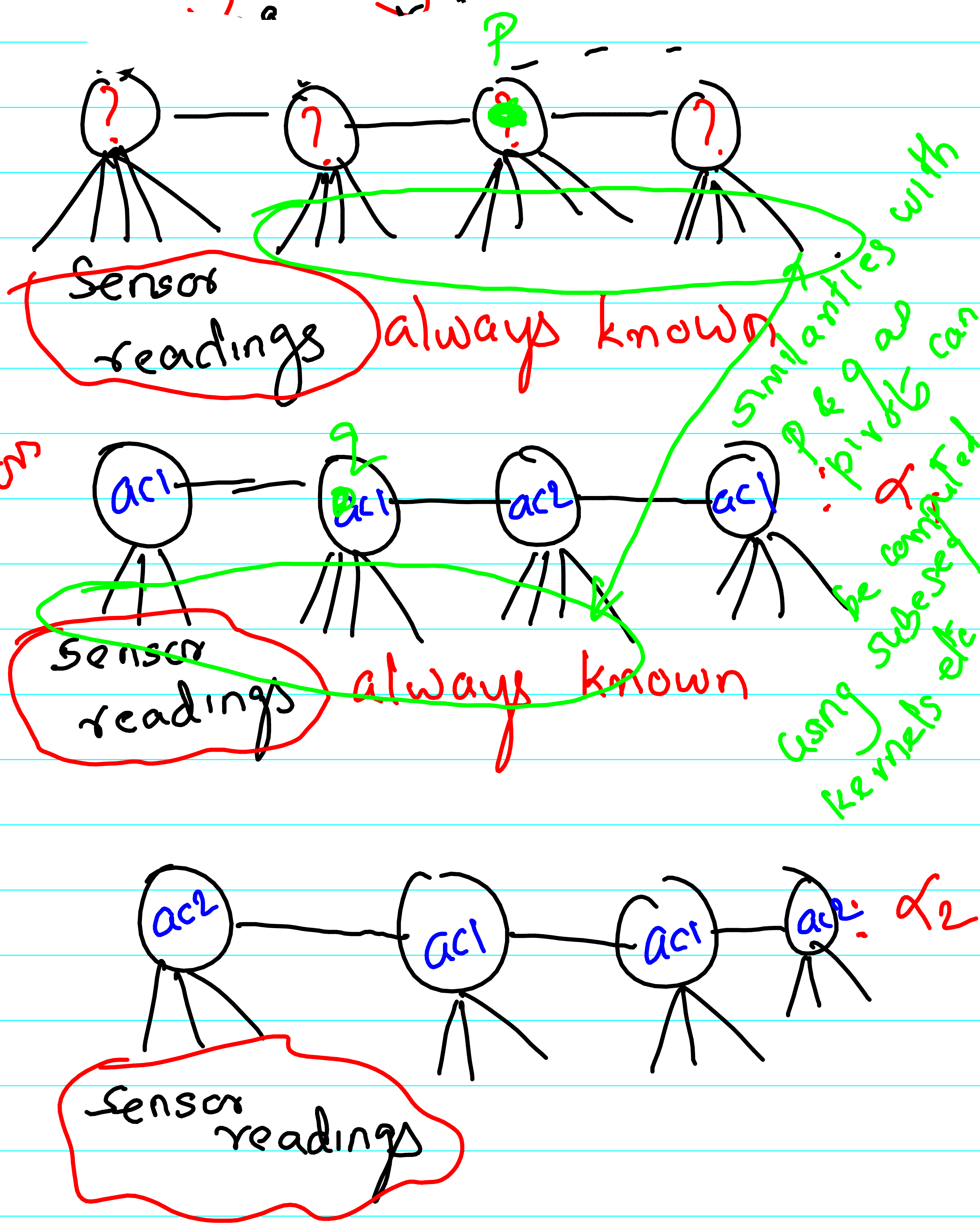


The kernel $\kappa_E(x_i^p, x_j^q)$ can be defined as a Set-Sequence (String) kernel), where we may be considering some window time steps q, with p and q as pivots.

Eg: Activity recognition

test instance

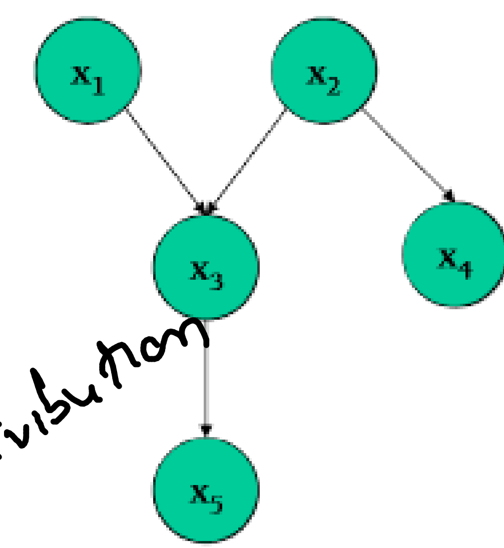
training instances that are support vectors



Definition 1 *Let $\mathcal{R} = \{X_1, X_2, \ldots, X_n\}$ be a set of random variables, with each X_i $(1 \leq i \leq n)$ assuming values $x_i \in \mathcal{X}_i$. Let $\mathcal{X}_S = \{X_i \mid i \in S\}$ where $S \subseteq \{1, 2, \ldots, n\}$. Let $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ be a directed acyclic graph with vertices $\mathcal{V} = \{1, 2, \ldots, n\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ such that each edge $e = (i, j) \in \mathcal{E}$ is a directed edge. We will assume a one to one correspondence between the set of variables $\mathcal{R}$ and the vertex set $\mathcal{V}$; vertex i will correspond to random variable X_i. Let π_i be the set of vertices from which there is edge incident on vertex i. That is, $\pi_i = \{j \mid j \in \mathcal{V}, (j, i) \in \mathcal{E}\}$. Then, the family $\mathcal{F}(\mathcal{G})$ of joint distributions associated with the DAG[2] $\mathcal{G}$ is specified by the factorization induced by $\mathcal{G}$ as follows:*

$$F(G) = \left\{ p(x) \;\middle|\; p(x) = \prod_i p(x_i \mid x_{\pi_i}) \right\}$$

a filter based on factorisation requirement on joint distribution



$\pi_{x_4} = \{x_2\}$

$\pi_{x_3} = \{x_1, x_2\}$

Figure 1.1: A directed graphical model.

$$P(x_1, x_2 \ldots x_5) = P(x_5 \mid x_3)\, P(x_3 \mid x_1, x_2)\, P(x_4 / x_2)\, P(x_1)\, P(x_2)$$

Definition 2 *Let $\mathcal{R} = \{X_1, X_2, \ldots, X_n\}$ be a set of random variables, with each X_i $(1 \le i \le n)$ assuming values $x_i \in \mathcal{X}_i$. Let $\mathbf{X}_\mathcal{S} = \{X_i \mid i \in \mathcal{S}\}$ where $\mathcal{S} \subseteq \{1, 2, \ldots, n\}$. Let $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ be a directed acyclic graph with vertices $\mathcal{V} = \{1, 2, \ldots, n\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ such that each edge $e = (i, j)$ is a directed edge. Let $\pi_i = \{j \mid j \in \mathcal{V},\ (j, i) \in \mathcal{E}\}$. Then, the family $\mathcal{C}(\mathcal{G})$ of joint distributions associated with the DAG $\mathcal{G}$ is specified by the conditional independence induced by $\mathcal{G}$ as follows:*

family of distributions specified by ~~the~~ the single graph G

$$\mathcal{C}(\mathcal{G}) = \left\{ p(\mathbf{x}) \,\middle|\, X_i \perp \mathbf{X}_{\mu_{i-1}} \mid \mathbf{X}_{\pi_i}\ \forall\, 1 \le i \le n,\ \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \right\} \tag{1.9}$$

$X_{\mu_{i-1}}$ = Set of all non-descendents of X_i

Claim: $F(G) = C(G)$

Simple eg:

1

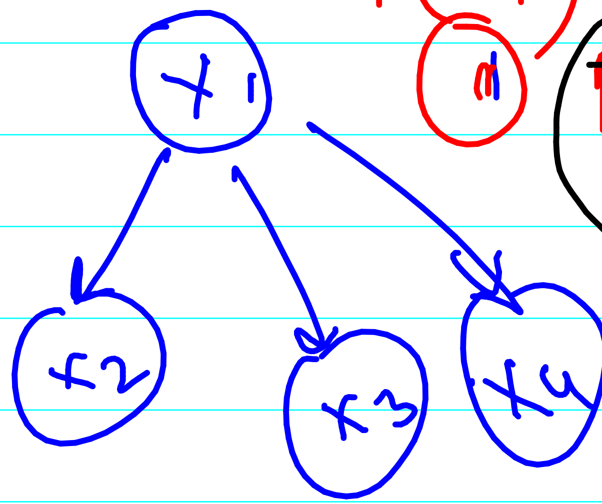
X_1 X_2 X_3

$F(G)$: 1 $p(X_1, X_2, X_3) = p(X_1)\,p(X_2)\,p(X_3)$

$C(G)$: 2 $X_1 \perp\!\!\!\perp X_2$, $X_1 \perp\!\!\!\perp X_3$, $X_2 \perp\!\!\!\perp X_3$

1 & 2 are equivalent? Prove.

②



F(G) ① $P(x_1 x_2 x_3 x_4) = P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_1)\, P(x_1)$

C(G) ②
$X_2 \perp\!\!\!\perp X_3 \mid X_1$
$X_2 \perp\!\!\!\perp X_4 \mid X_1$
$X_3 \perp\!\!\!\perp X_4 \mid X_1$

Are ① & ② equivalent?

H/W: Prove equivalence

eg ① ⇒ ②

$$P(x_1\, x_2\, x_3\, x_4) = \overset{P(x_2|x_1)}{P(x_2 \cancel{|x_1, x_3\, x_4})}\; \overset{P(x_3|x_1)}{P(\cancel{x_3}|x_1, x_4)}\; P(x_4|x_1)\, P(x_1)$$

To Prove in general that

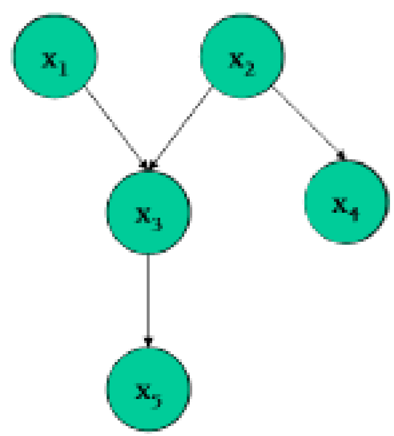
$$F(G) = C(G),$$

Prove that

for any P, $P \in F(G)$ iff $P \in C(G)$

Please read proof to Theorem 1 from

http://www.cse.iitb.ac.in/~cs717/notes/classNotes/graphicalModelsReading.pdf



H/W: verify that the following 2 are equivalent (& in fact are imposed by 2 different interpretations of the above graph)

Factorisation interpretation: $F(G)$

$$P(x_1 \ldots x_5) = P(x_5 | x_3)\, P(x_3 | x_1 x_2)\, P(x_4 | x_2)\, P(x_1)\, P(x_2)$$

verify that one follows from other

Conditional independence interpretation $C(G)$

$$X_1 \perp\!\!\!\perp X_2,\quad X_1 \perp\!\!\!\perp X_4,\quad X_3 \perp\!\!\!\perp X_4 | X_1, X_2,\quad X_5 \perp\!\!\!\perp X_4 | X_3$$

Theorem 1 *The sets $\mathcal{F}(\mathcal{G})$ and $\mathcal{C}(\mathcal{G})$ are equal. That is $p \in \mathcal{F}(\mathcal{G})$ iff $p \in \mathcal{C}(\mathcal{G})$*

Proof: $\Leftarrow$: We will first prove that $\mathcal{F}(\mathcal{G}) \subseteq \mathcal{C}(\mathcal{G})$. Let $p \in \mathcal{F}(\mathcal{G})$. We will prove that $p \in \mathcal{C}(\mathcal{G})$, that is, $p(x_i \mid \mathrm{x}_{\mu_{i-1}}, \mathrm{x}_{\pi_i}) = p(x_i \mid \mathrm{x}_{\pi_i})$. This trivially holds for $i = 1$, since $\mathrm{x}_{\pi_i} = \emptyset$. For $i = 2$:

$$p(x_1, x_2) = p(x_1)p(x_1 \mid x_2) = p(x_1)p(x_1 \mid \mathrm{x}_{\pi_2})$$

where, the first equality follows by chain rule, whereas the second equality follows by virtue of (1.8). Consequently,

$$p(x_1 \mid x_2) = p(x_1 \mid \mathrm{x}_{\pi_2})$$

By induction

Assume that $p(x_i \mid \mathrm{x}_{\mu_{i-1}}) = p(x_i \mid \mathrm{x}_{\pi_i})$ for $i \leq k$. For $i = k+1$, it follows from chain rule and from (1.8) that

By chain rule

Since $p \in F(G)$

$$p(\mathrm{x}_{\mu_{k+1}}) = \prod_{i=1}^{k+1} p(x_i \mid \mathrm{x}_{\mu_{i-1}}) = \prod_{i=1}^{k+1} p(x_i \mid \mathrm{x}_{\pi_i})$$

Making use of the induction assumption for $i \leq k$ in the equation above, we can derive that

$$p(x_k \mid \mathrm{x}_{\mu_{k-1}}) = p(x_k \mid \mathrm{x}_{\pi_k})$$

equating whatever remain after using equality from induction

By induction on i, we obtain that $p(x_i \mid \mathrm{x}_{\mu_{i-1}}) = p(x_k \mid \mathrm{x}_{\pi_i})$ for all i. That is, $p \in \mathcal{C}(\mathcal{G})$. Since this holds for any $p \in \mathcal{F}(\mathcal{G})$, we must have that $\mathcal{F}(\mathcal{G}) \subseteq \mathcal{C}(\mathcal{G})$.

$\Rightarrow$: Next we prove that $\mathcal{C}(\mathcal{G}) \subseteq \mathcal{F}(\mathcal{G})$. Let $p' \in \mathcal{C}(\mathcal{G})$ satisfy the conditional independence assertions. That is, for any $1 \leq i \leq n$, $p'(x_i \mid \mathrm{x}_{\mu_{i-1}}) = p'(x_i \mid \mathrm{x}_{\pi_i})$. Then by chain rule, we must have:

$$p'(\mathrm{x}_{\mu_n}) = \prod_{i=1}^{n} p'(x_i \mid \mathrm{x}_{\mu_{i-1}}) = \prod_{i=1}^{k+1} p'(x_i \mid \mathrm{x}_{\pi_i})$$

Trivially true

Definition 4 *The set of probability distributions $\mathcal{D}(\mathcal{G})$ for a DAG $\mathcal{G}$ is defined as follows:*

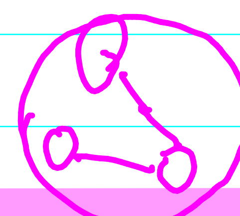
$$\mathcal{D}(\mathcal{G}) = \{p(\mathbf{x}) | \mathbf{X}_{\mathcal{A}} \perp \mathbf{X}_{\mathcal{B}} \mid \mathbf{X}_{\mathcal{C}}, \text{ whenever } \mathcal{A} \text{ and } \mathcal{B} \text{ are } d-\text{separated by } \mathcal{C}\} \tag{1.10}$$

Theorem 2 *For any directed acyclic graph $\mathcal{G}$, $\mathcal{D}(\mathcal{G}) = \mathcal{C}(\mathcal{G}) = \mathcal{F}(\mathcal{G})$.*

d-separation based independence

Conditional independence

For digraphs, cliques are sets containing node & its parents i.e. C =



$\phi_c(x_c) = p(x_i | x_{\pi_i})$

$\phi_c(x_c) \geq 0$

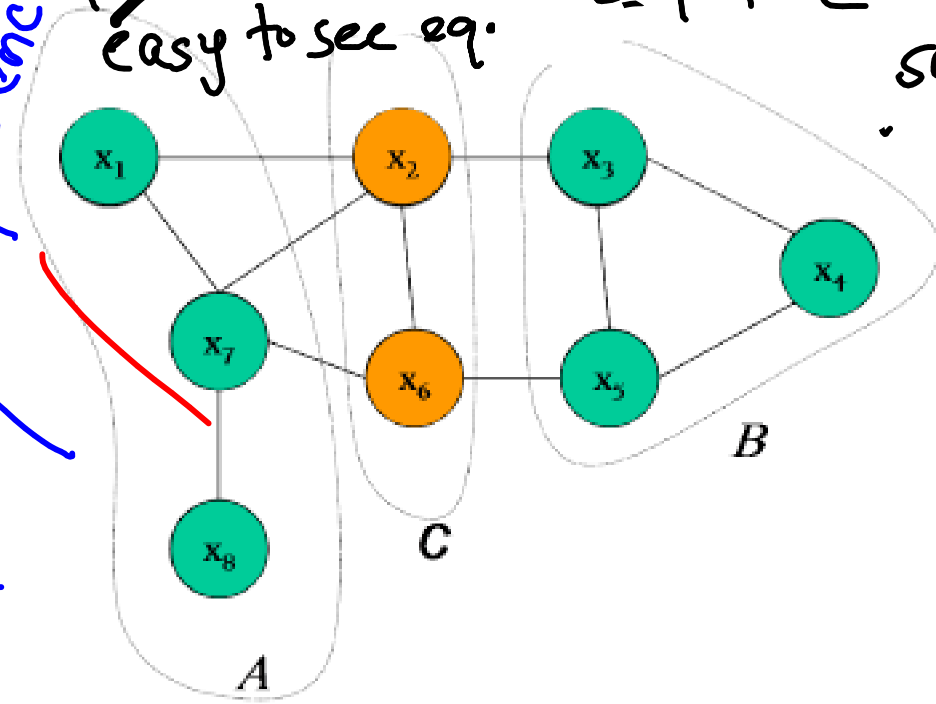
We can similarly discuss $F(G)$ & $C(G)$ for undirected graphs G

$\frac{1}{Z}\prod_{c \in \mathcal{C}} \phi_c(x_c)$

$x \perp$ all other nodes given its nbrs

$\iff X_A \perp\!\!\!\perp X_B \mid X_C$ & so on

easy to see eg.



$X_6 \perp\!\!\!\perp \{X_8, X_1, X_4, X_3\} \mid X_7, X_5, X_2$

M(G) $X_7 \perp\!\!\!\perp \ldots$

$X_8 \perp\!\!\!\perp \ldots$

conditional / Markov independence

F(G) Factorisation definition

$$P(x_1, \ldots x_8) = \phi_{X_1X_2X_7}(x_1, x_2 x_7)\, \phi_{X_7X_6X_2}(x_7 x_6 x_2)\, \phi_{X_7X_8}(x_7 x_8)\, \phi_{X_5X_6}(x_5 x_6)\, \phi_{X_2X_3}(x_2 x_3)\, \phi_{X_5X_3X_4}(x_5 x_3 x_4)$$

Claim: M(G) = F(G) for undirected graphs G

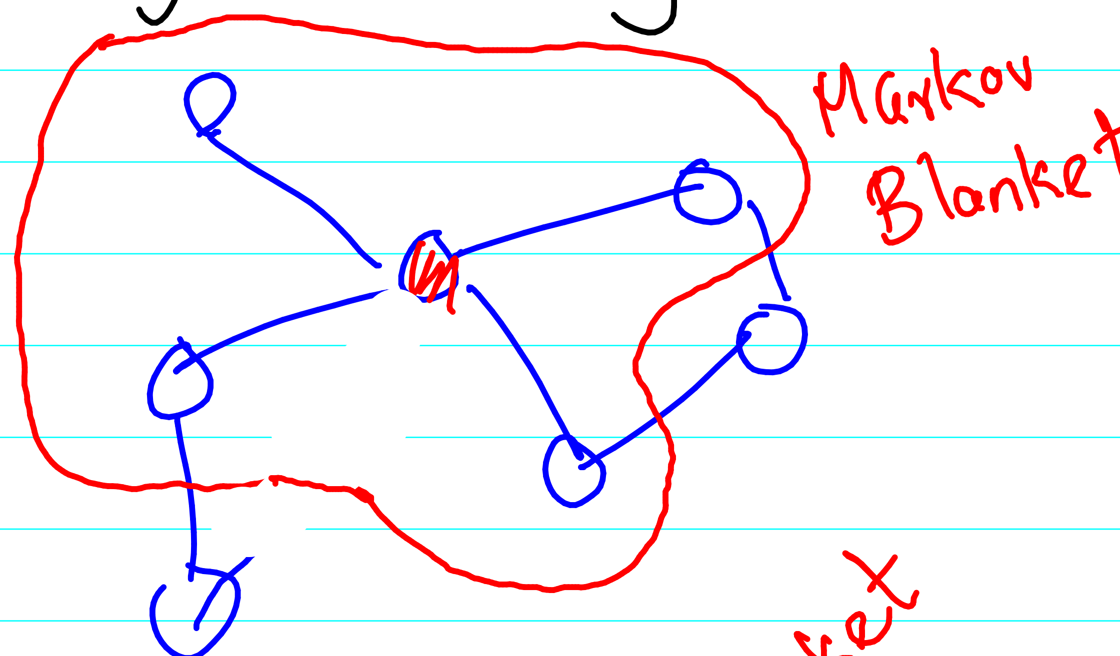
Refer to notes & slides for details

http://www.cse.iitb.ac.in/~cs717/notes/classNotes/graphicalModelsReading.pdf

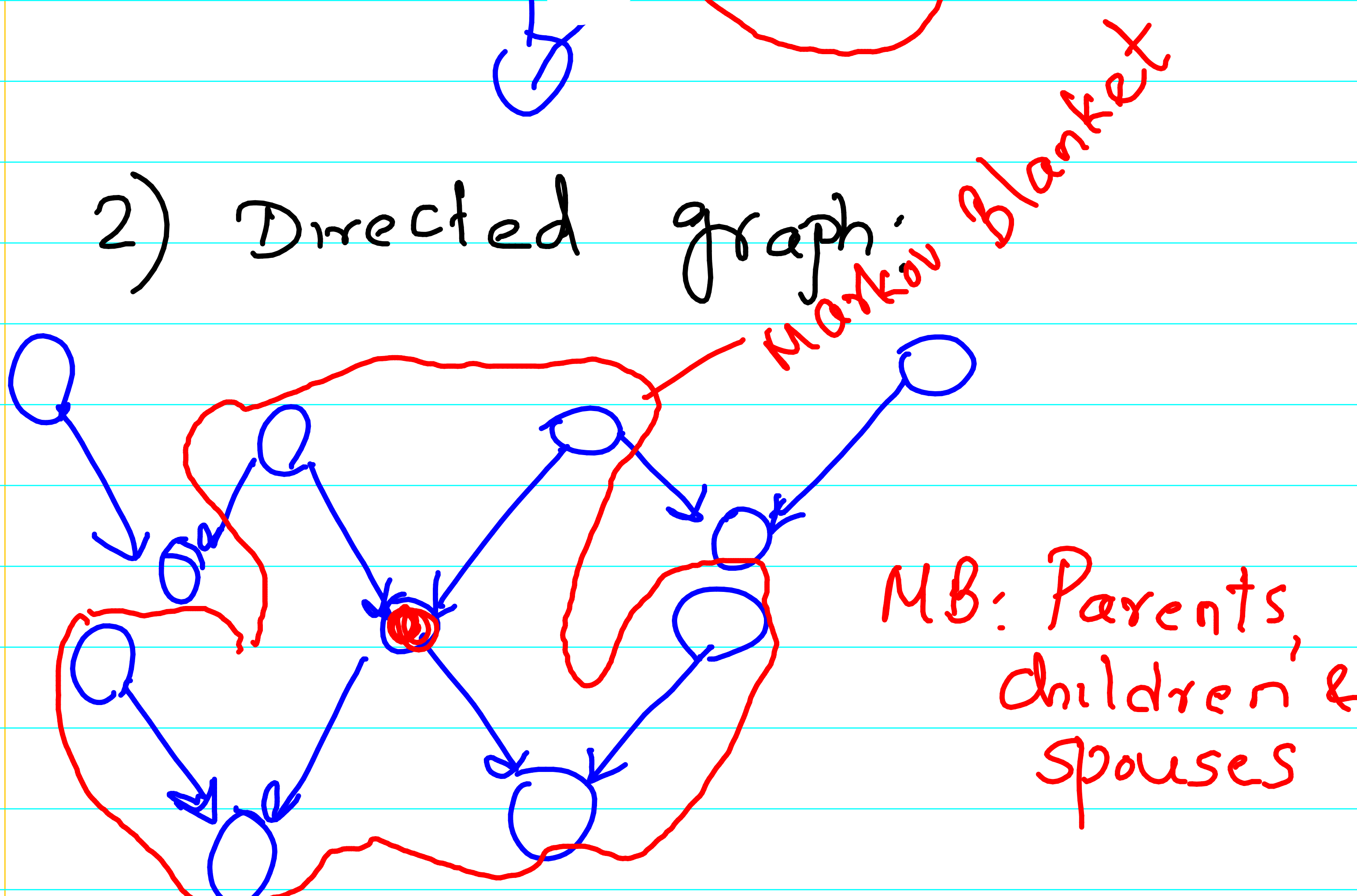
Markov Blankets:

1) Undirected graph:

Set of nbrs of a node



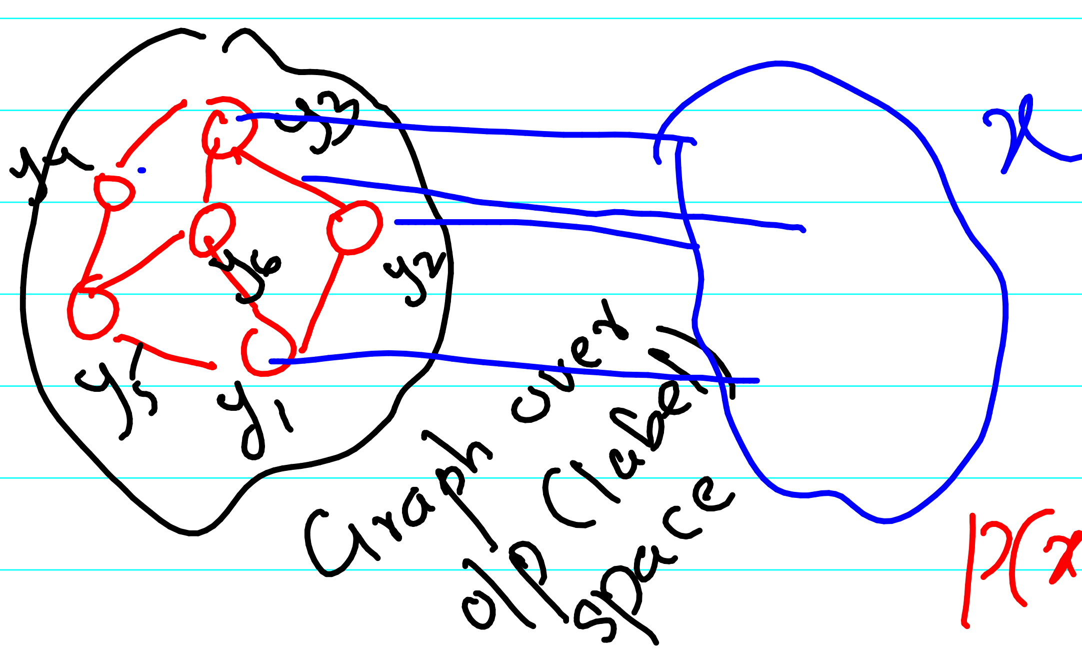
2) Directed graph:



Note: For linear & tree graphs, directed & undirected representations are equivalent

(eg: HMM can be treated as directed or undirected)

Markov Random fields



x is complex input

Defined only on o/p

Features on i/p & o/p

$$P(x,y) = \prod_c \phi_c(y_c) \times \prod_i \psi_i(y_i, x)$$

More on markov networks
in next lecture

http://www.cse.iitb.ac.in/~cs717/notes/classNotes/StructuredOutput/mmmn.pdf