Graphical Model & Gibbs Sampling

Hung-yi Lee

Structured Learning

We also know how to involve hidden information.

Problem 1: Evaluation

- What does F(x,y) look like? $F(x,y) = w \cdot \phi(x,y)$

Problem 2: Inference

- How to solve the “arg max” problem

$$y = \arg\max_{y \in Y} F(x, y)$$

Problem 3: Training

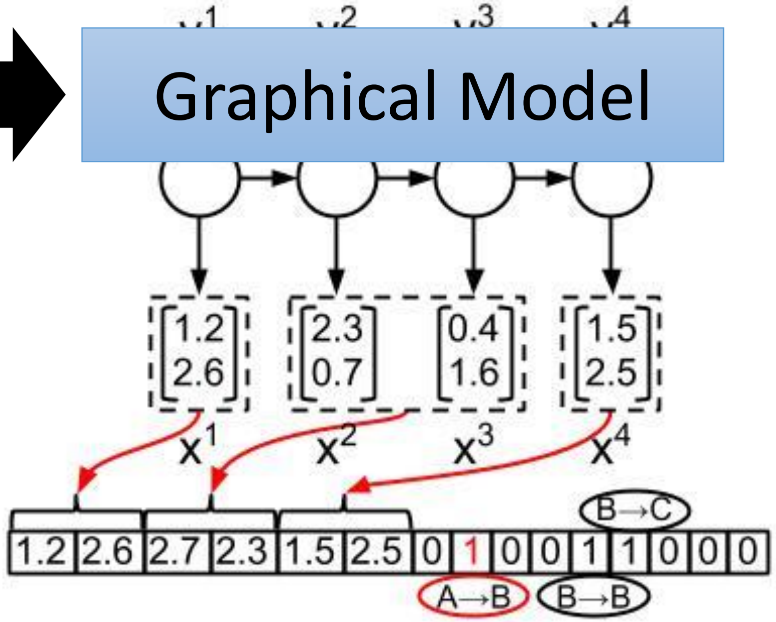
- Given training data, how to find F(x,y) Structured SVM, etc.

Difficulties

Difficulty 1. Evaluation

Graphical Model

$$F(x,y) = w \cdot \phi(x,y)$$



$\phi(x,y)$

Hard to figure out? Hard to interpret the meaning?

Difficulty 2. Inference

Gibbs Sampling

We can use Viterbi algorithm to deal with sequence labeling. How about other cases?

Graphical Model

A language which describes the evaluation function

Graphical Model

$F(x, y)$ ⟺ Graph

- Define and describe your evaluation function F(x,y) by a graph
- There are three kinds of graphical model.
  - *Factor graph*, *Markov Random Field* (MRF) and *Bayesian Network* (BN)
  - Only *factor graph* and *MRF* will be briefly mentioned today.

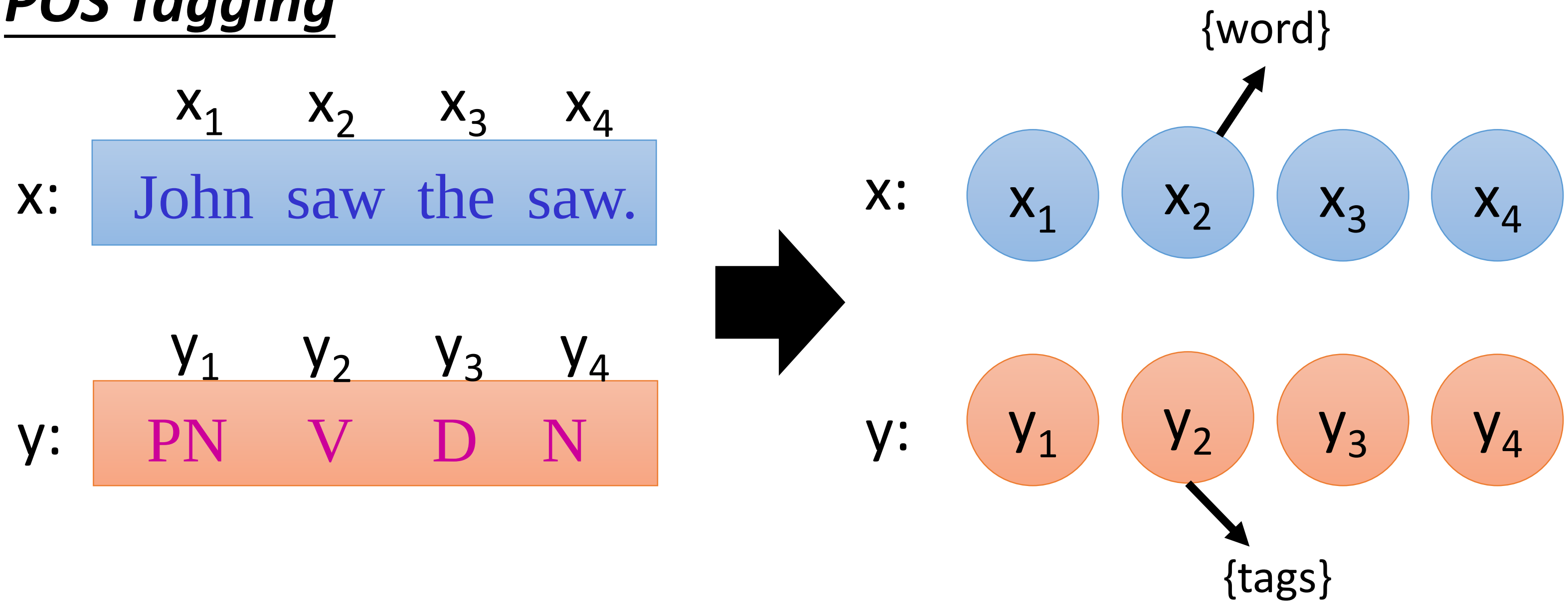
Decompose F(x,y)

- $F(x, y)$ is originally a ***global*** function
  - Define over the whole x and y
- Based on graphical model, $F(x, y)$ is the composition of some ***local*** functions
  - x and y are decomposed into smaller components
  - Each local function defines on only a few related components in x and y
  - Which components are related → defined by Graphical model

Decomposable x and y

- x and y are decomposed into smaller components

<u>POS Tagging</u>



x_1 x_2 x_3 x_4

x: John saw the saw.

y_1 y_2 y_3 y_4

y: PN V D N

{word}

x: x_1 x_2 x_3 x_4

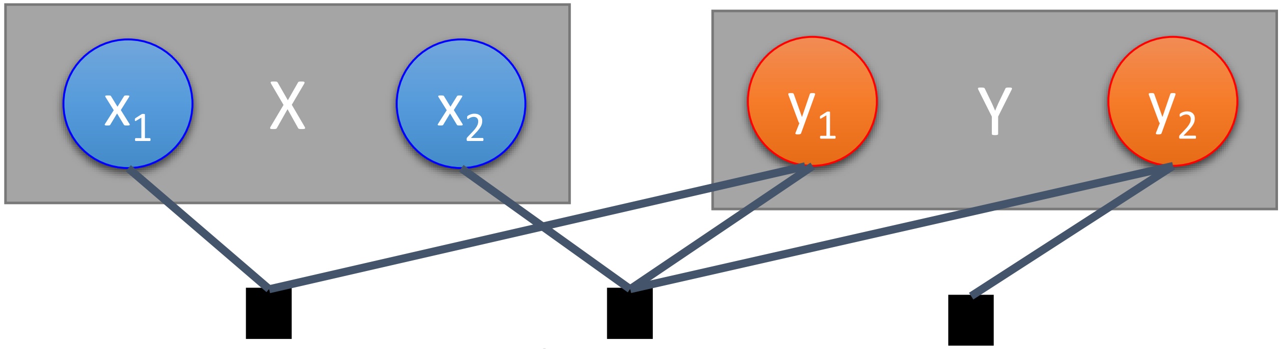
y: y_1 y_2 y_3 y_4

{tags}

Factor Graph

Each factor influences some components.

Each factor corresponds to a local function.



factor a $f_a(x_1, y_1)$

factor b $f_b(x_2, y_1, y_2)$

factor c $f_d(y_2)$

Larger value means more compatible.

$$F(x, y) = f_a(x_1, y_1) + f_b(x_2, y_1, y_2) + f_c(y_2)$$

You only have to define the factors.

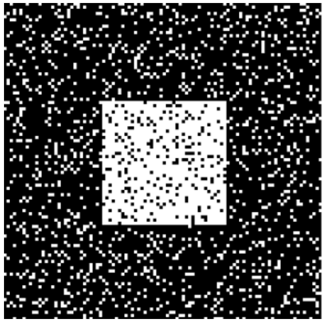
The local functions of the factors are learned from data.

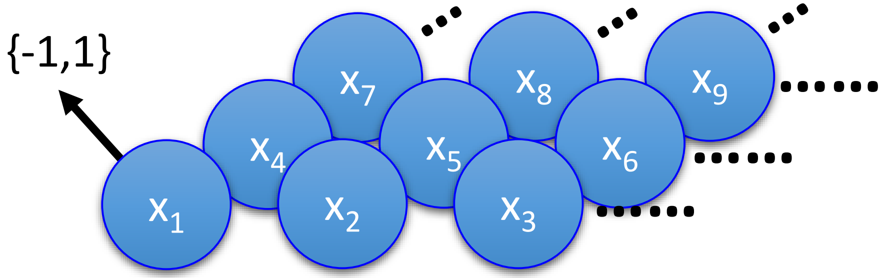
Factor Graph - Example

- ***Image De-noising***

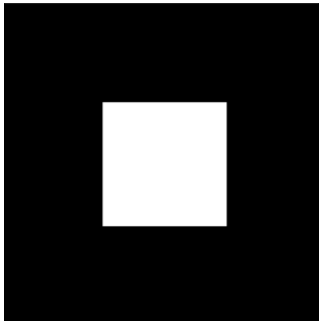
Each pixel is one component

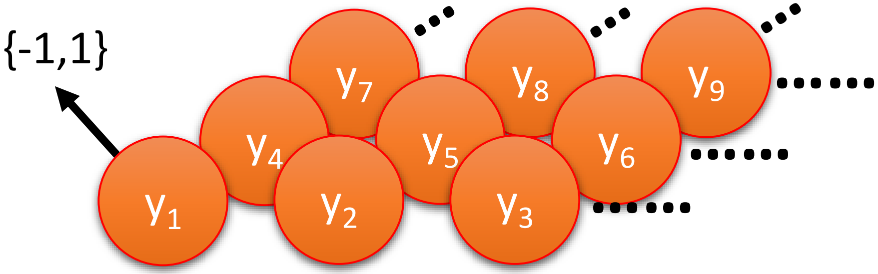
Noisy image
x





Clean image
y





http://cs.stanford.edu/people/karpathy/visml/ising_example.html

Factor Graph - Example

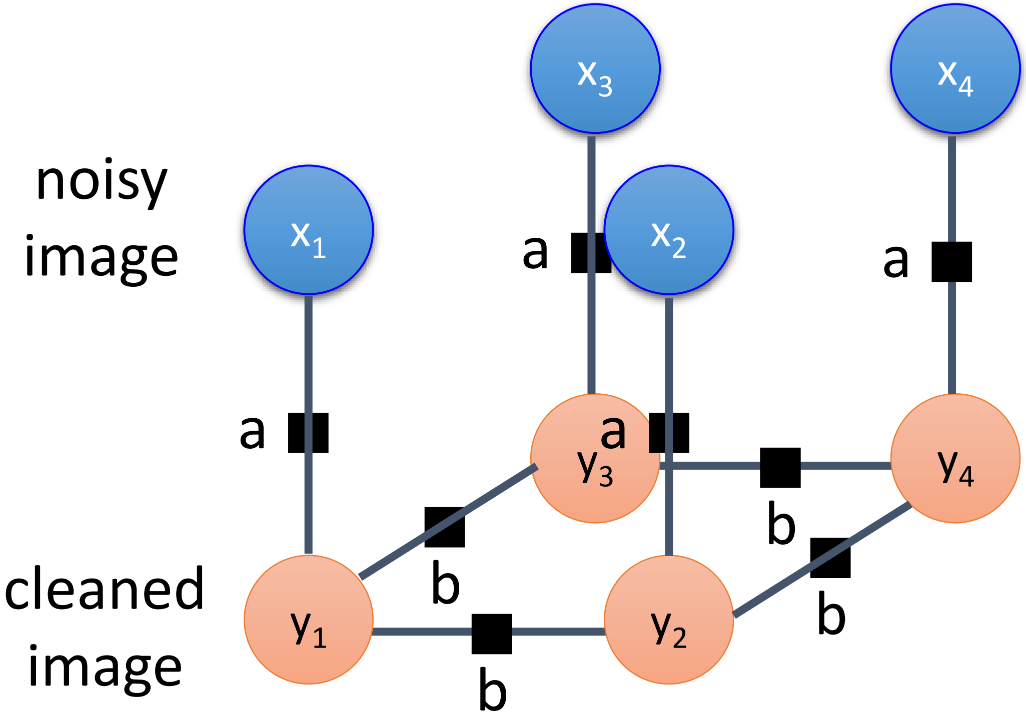
Factor:

Noisy and clean images are related

- **a**: the values of x_i and y_i

The colors in the clean image is smooth.

- **b**: the values of the neighboring y_i



$$f_a(x_i, y_i) = \begin{cases} 1 & x_i = y_i \\ -1 & x_i \neq y_i \end{cases}$$

$$f_b(y_i, y_j) = \begin{cases} 2 & y_i = y_j \\ -2 & y_i \neq y_j \end{cases}$$

The weights can be learned from data.

Factor Graph - Example

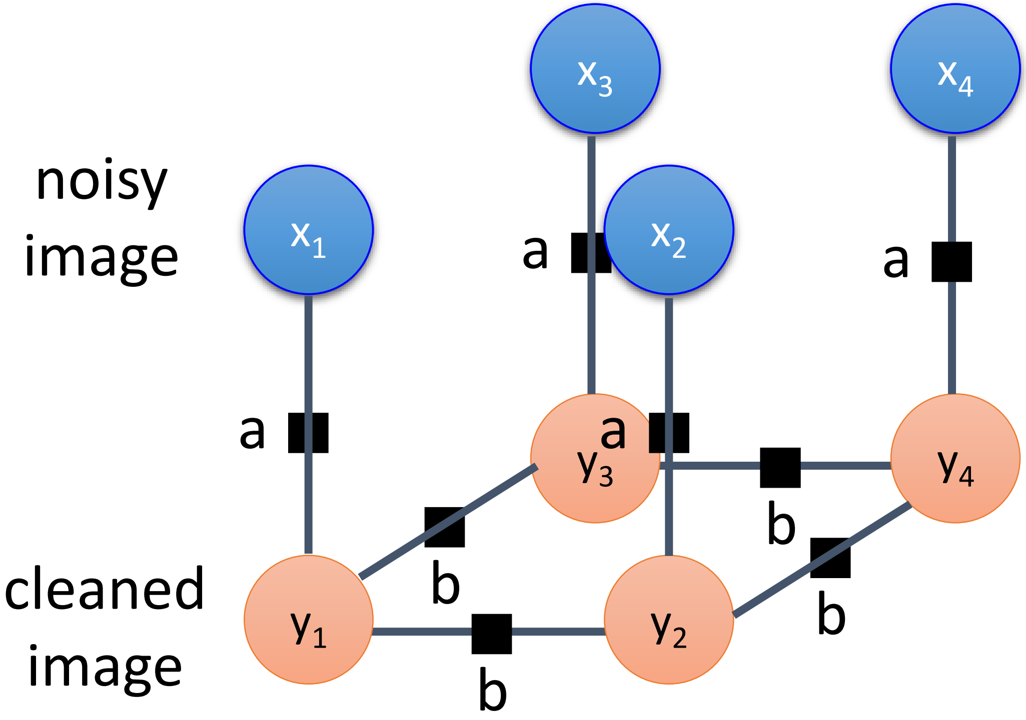
Factor:

Noisy and clean images are related

- **a**: the values of x_i and y_i

The colors in the clean image is smooth.

- **b**: the values of the neighboring y_i



Realize $F(x, y)$ easily from the factor graph

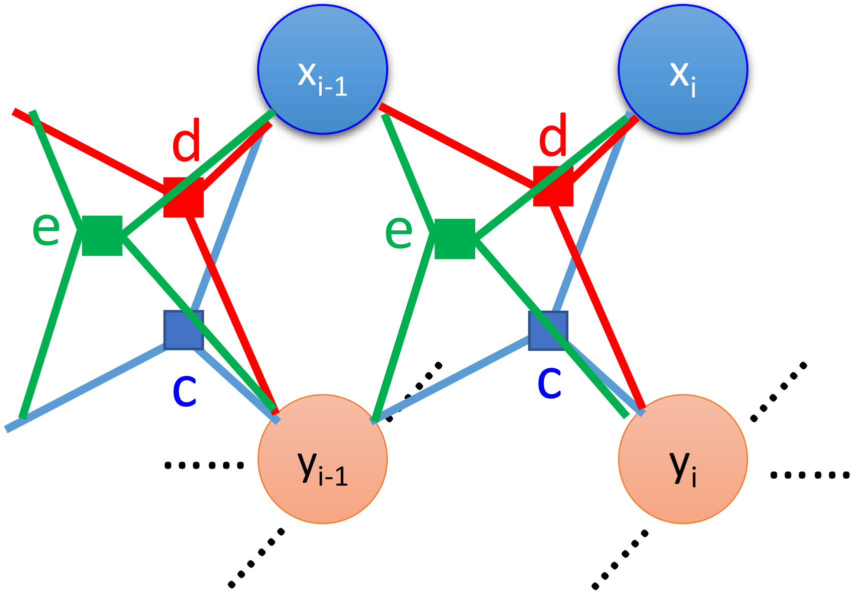
$$F(x,y) = \sum_{i=1}^{4} f_a(x_i, y_i)$$

$$+f_b(y_1, y_2) + f_b(y_1, y_3)$$
$$+f_b(y_2, y_4) + f_b(y_3, y_4)$$

Factor Graph - Example

Factor:

- **c**: the values of x_i and the values of the neighboring y_i
- **d**: the values of the neighboring x_i and the values of y_i



$$f_c(x_i, y_i, y_{i-1})$$

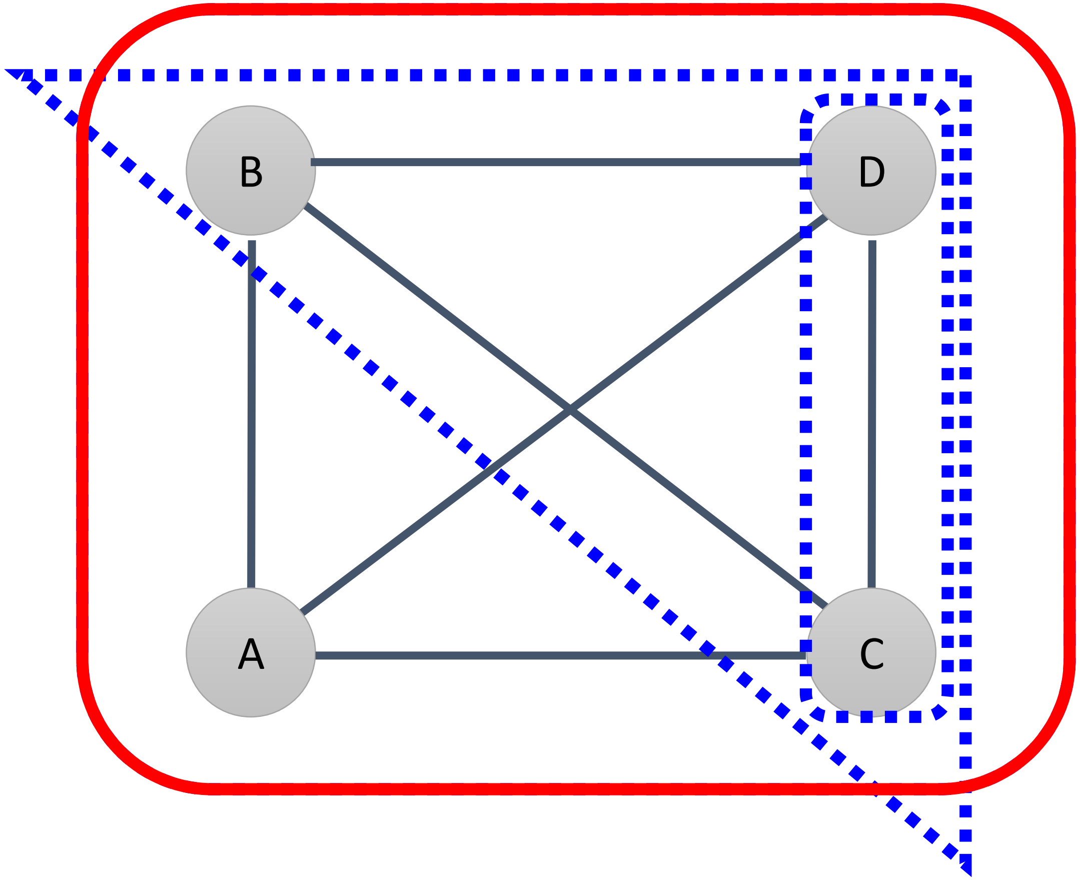
$$f_d(x_i, x_{i-1}, y_i)$$

$$f_e(x_i, x_{i-1}, y_i, y_{i-1})$$

Markov Random Field (MRF)

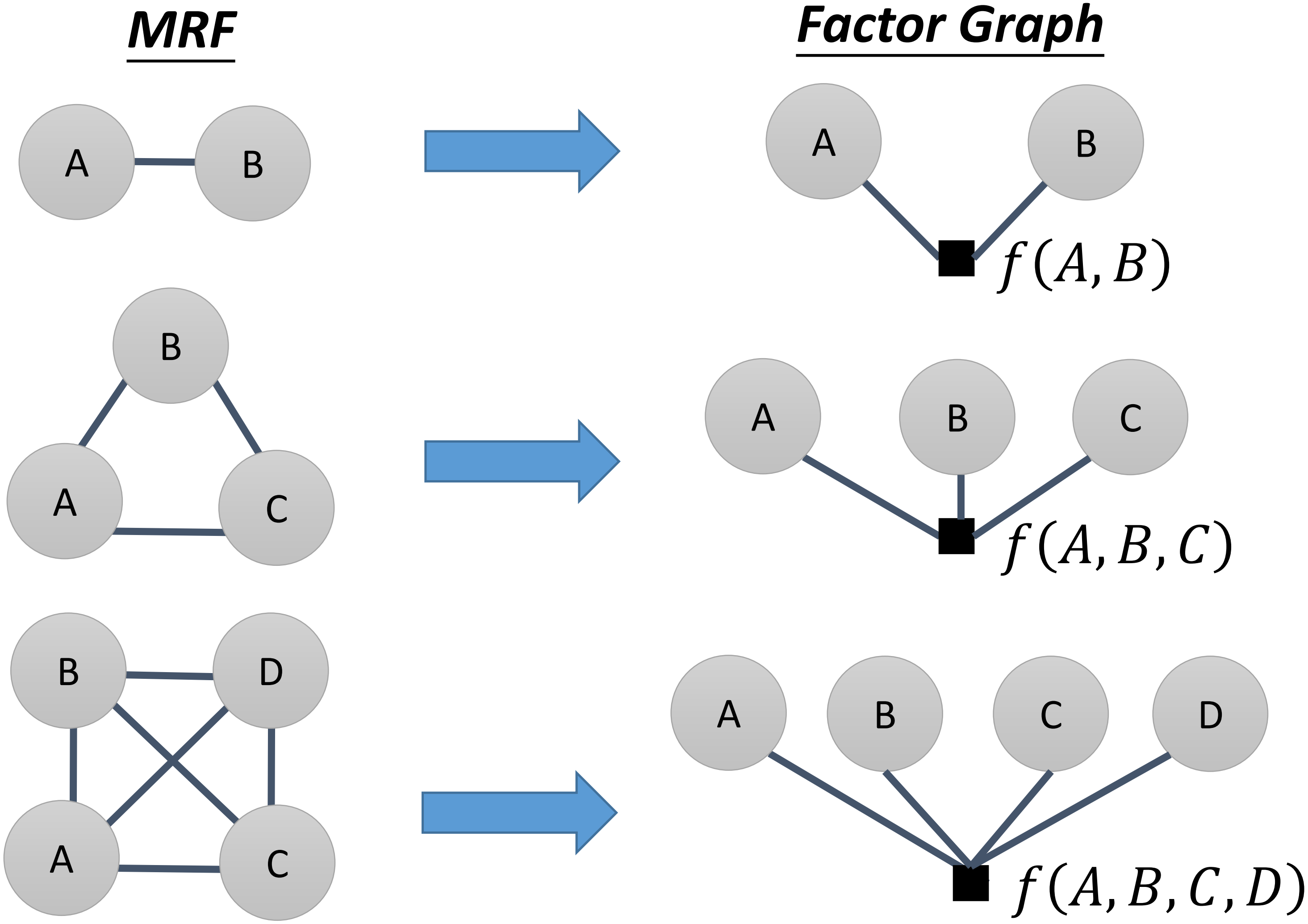
Clique: a set of components connecting to each other

Maximum Clique: a clique that is not included by other cliques



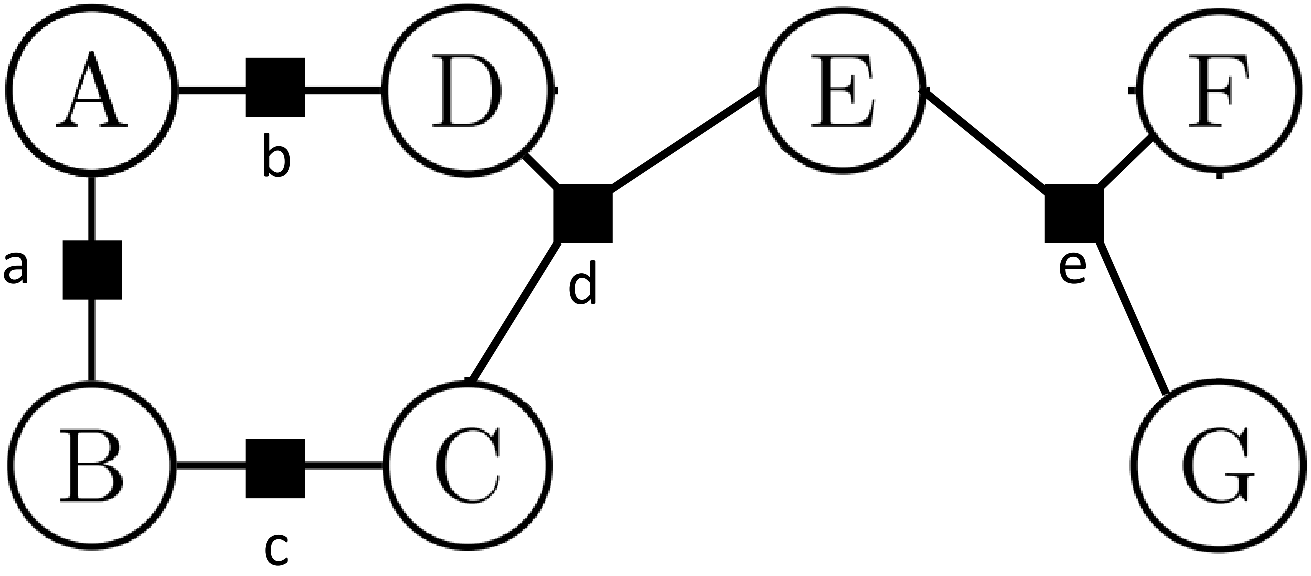
MRF

Each maximum clique on the graph corresponds to a factor



MRF

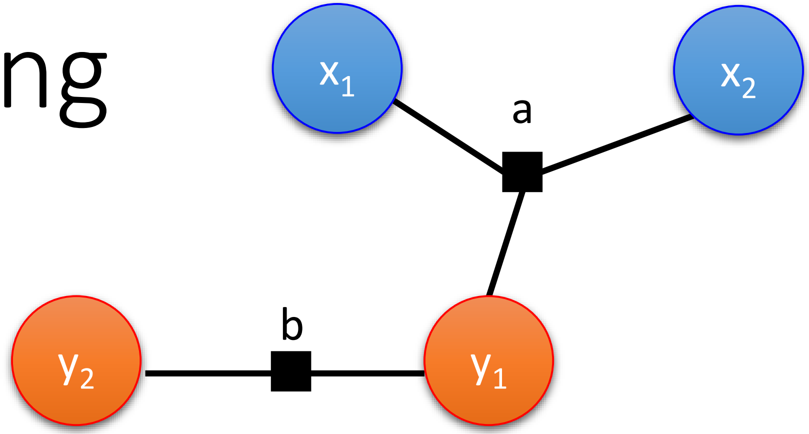
MRF → ***Factor Graph***



Evaluation Function

$$f_a(A,B) + f_b\ (A,D) + f_c(B,C) + f_d(C,D,E) + f_e(E,F,G)$$

Training



$$F(x,y) = f_a(x_1,x_2,y_1) + f_b(y_1,y_2)$$

$$= w_a \cdot \phi_a(x_1,x_2,y_1) + w_b \cdot \phi_b(y_1,y_2)$$

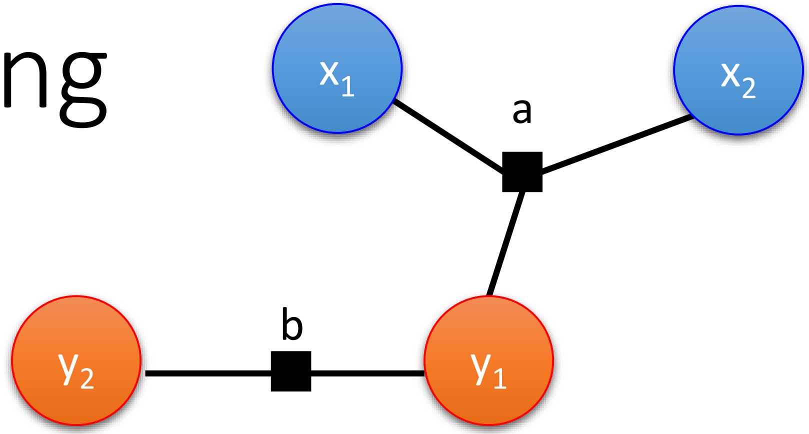
$$= \begin{bmatrix} w_a \\ w_b \end{bmatrix} \begin{bmatrix} \phi_a(x_1,x_2,y_1) \\ \phi_b(y_1,y_2) \end{bmatrix}$$

$$= w \cdot \phi(x,y)$$

Simply training by
structured perceptron
or structured SVM

Max-Margin Markov Networks (M3N)

Training



$$F(x,y) = f_a(x_1,x_2,y_1) + \underline{f_b(y_1,y_2)}$$

$$= w_a \cdot \phi_a(x_1,x_2,y_1) + \underline{w_b \cdot \phi_b(y_1,y_2)}$$

$$y_1, y_2 \epsilon \{+1,-1\}$$

y_1	y_2	$f_b(y_1,y_2)$
+1	+1	w_1
+1	-1	w_2
-1	+1	w_3
-1	-1	w_4

$$w_b = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}$$

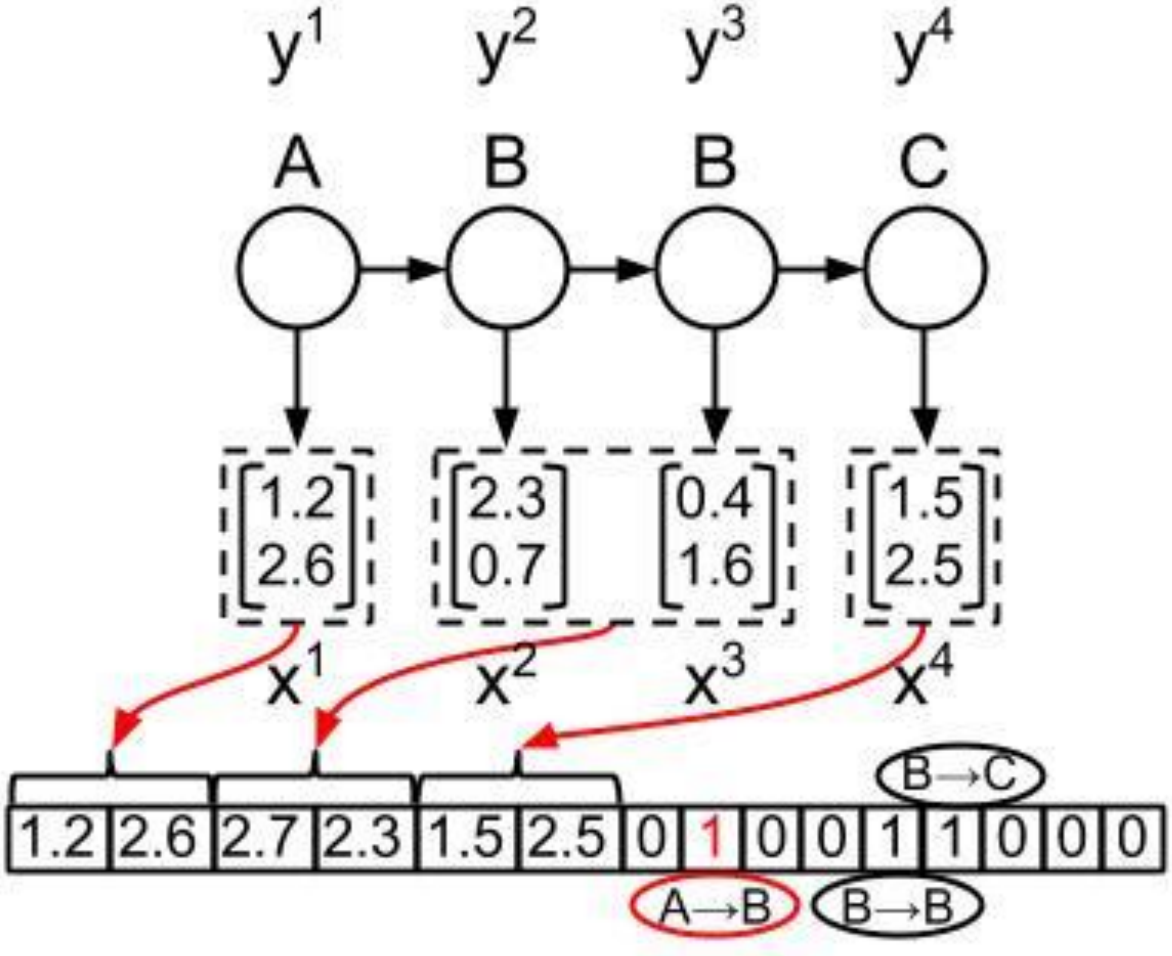
$$\phi_b(+1,+1) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\phi_b(+1,-1) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\phi_b(-1,+1) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\phi_b(-1,-1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Now can you interpret this?



$\phi(x,y)$

Probability Point of View

- $F(x, y)$ can be any real number
- If you like probability

Between 0 and 1

$$P(x, y) = \frac{e^{F(x,y)}}{\sum_{x',y'} e^{F(x',y')}}$$

To be positive

normalization

$$P(y|x) = \frac{P(x, y)}{P(x)}$$

$$= \frac{P(x, y)}{\sum_{y''} P(x, y'')} = \frac{\frac{e^{F(x,y)}}{\cancel{\sum_{x',y'} e^{F(x',y')}}}}{\boxed{\sum_{y''}} \frac{e^{F(x,y'')}}{\cancel{\sum_{x',y'} e^{F(x',y')}}}} = \frac{e^{F(x,y)}}{\sum_{y''} e^{F(x,y'')}}$$

Gibbs Sampling

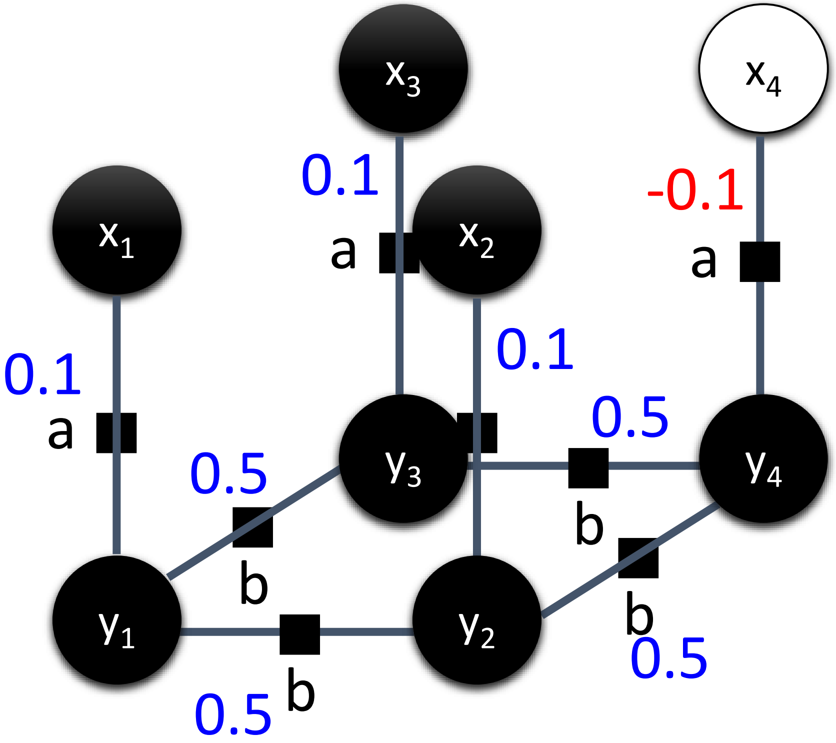
Inference for the dumb

$$f_a(x_i, y_i) = \begin{cases} 0.1 & x_i = y_i \\ -0.1 & x_i \neq y_i \end{cases}$$

$$f_b(y_i, y_j) = \begin{cases} 0.5 & y_i = y_j \\ -0.5 & y_i \neq y_j \end{cases}$$

Given input noisy image x

$$x_1, x_2, x_3, x_4 = -1, -1, -1, 1$$



Inference:

$$\tilde{y} = \arg\max_{y} \mathrm{F}(x, y)$$

$$y_1, y_2, y_3, y_4 = -1, -1, -1, -1$$

$$\Rightarrow F(x, y) = 2.2$$ max

$$y_1, y_2, y_3, y_4 = 1,1,1,1$$

$$\Rightarrow F(x, y) = 1.8$$

$$y_1, y_2, y_3, y_4 = -1,1,1,-1$$

$$\Rightarrow F(x, y) = -2.2$$

⋮

Enumerate all possible **y**

Design an efficient algorithm to do that is not always easy.

Sampling?

Probability point of view:

$$P(x, y) = \frac{e^{F(x,y)}}{\sum_{x',y'} e^{F(x',y')}}$$

$$\boxed{P(y|x) = \frac{e^{F(x,y)}}{\underline{\sum_{y''} e^{F(x,y'')}}} \quad \boxed{\text{Independent of y}}} \propto F(x, y)$$

$$\tilde{y} = \arg\max_{y} \mathrm{F}(x,y)$$

=

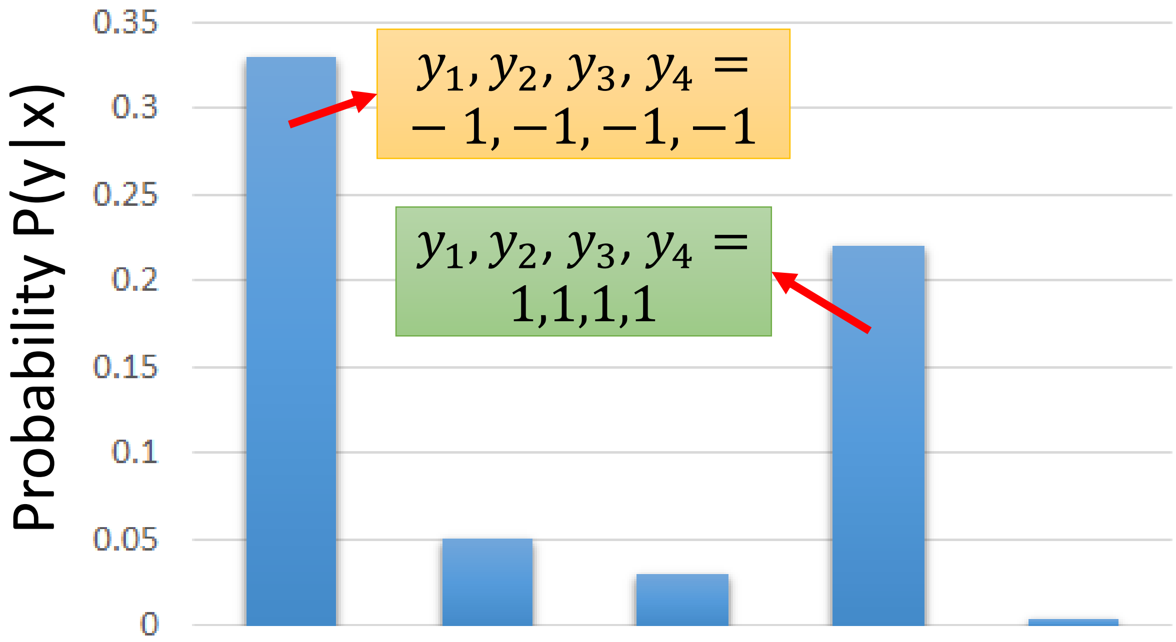
$$\tilde{y} = \arg\max_{y} P(y|x)$$

Sampling?

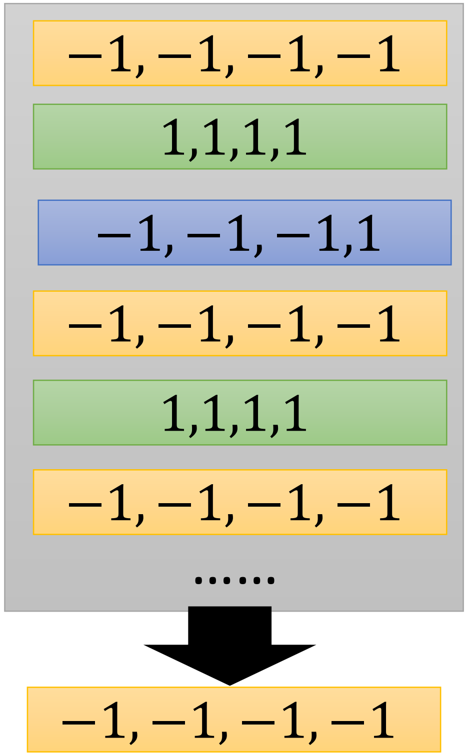
$$\tilde{y} = \arg\max_{y} P(y|x)$$

- $P(y|x)$ is a distribution

Given $x_1, x_2, x_3, x_4 = -1, -1, -1, 1$



Sample from the distribution ……



Max probability

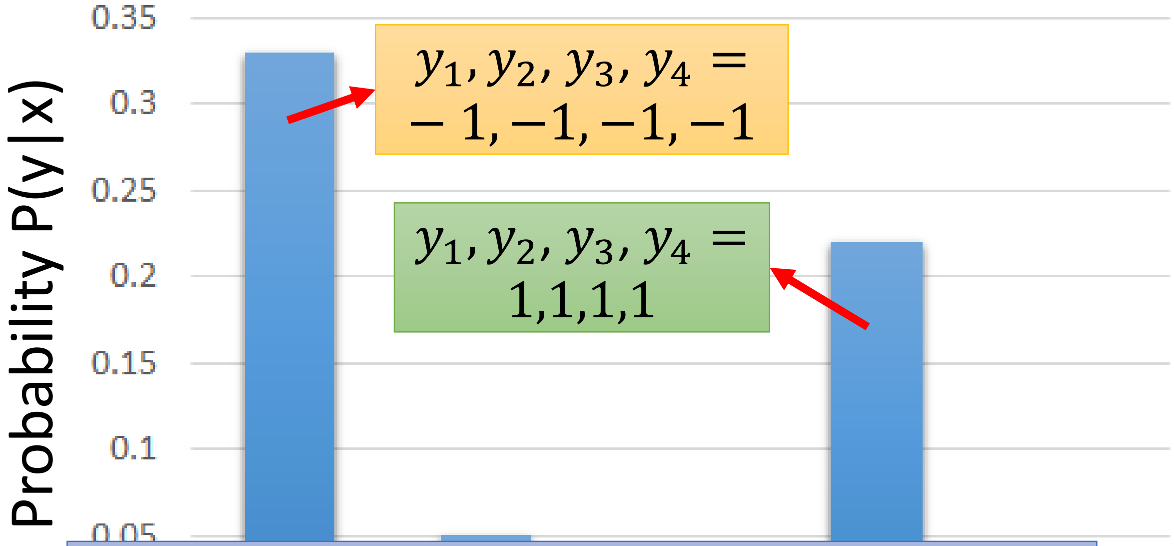
Inference result

Sampling?

$$P(y|x) = \frac{e^{F(x,y)}}{\sum_{y''} e^{F(x,y'')}}$$

- $P(y|x)$ is a distribution

Given $x_1, x_2, x_3, x_4 = -1, -1, -1, 1$



$y_1, y_2, y_3, y_4 =$
$-1, -1, -1, -1$

$y_1, y_2, y_3, y_4 =$
$1, 1, 1, 1$

- ➢ It is hard to know the distribution.
- ➢ If we know the distribution, why bother with the sampling?

Sample from the distribution ……

$-1, -1, -1, -1$

$1,1,1,1$

$-1, -1, -1, 1$

$-1, -1, -1, -1$

$1,1,1,1$

$-1, -1, -1, -1$

……

$-1, -1, -1, -1$

Max probability

Inference result

Gibbs Sampling

- There is a probability distribution $P(\mathbf{y}|\mathbf{x})$
  - $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$
- We want to sample from $P(\mathbf{y}|\mathbf{x})$, but it is too complex to do that
- However, $P(y_i|y_1, y_2, \ldots, \underline{y_{i-1}, y_{i+1}}, \ldots, y_N,\mathbf{x})$ can be computed
- We can sample from $P(\mathbf{y}|\mathbf{x})$ by Gibbs sampling

Gibbs Sampling

$\boldsymbol{y}^{\mathbf{0}} = \{y_1^0, y_2^0, \cdots, y_N^0\}$ Initialization

For t = 1 to T: T samples

$y_1^t \sim P(y_1 | y_2 = y_2^{t-1}, y_3 = y_3^{t-1}, y_4 = y_4^{t-1}, \cdots, y_N = y_N^{t-1}, \boldsymbol{x})$

$y_2^t \sim P(y_2 | y_1 = y_1^t, y_3 = y_3^{t-1}, y_4 = y_4^{t-1}, \cdots, y_N = y_N^{t-1}, \boldsymbol{x})$

$y_3^t \sim P(y_3 | y_1 = y_1^t, y_2 = y_2^t, y_4 = y_4^{t-1}, \cdots, y_N = y_N^{t-1}, \boldsymbol{x})$

$\vdots$

$y_N^t \sim P(y_N | y_1 = y_1^t, y_2 = y_2^t, y_3 = y_3^t, \cdots, y_{N-1} = y_{N-1}^t, \boldsymbol{x})$

Get a sample: $\boldsymbol{y}^t = \{y_1^t, y_2^t, \cdots, y_N^t\}$

$\boldsymbol{y}^1, \boldsymbol{y}^2, \boldsymbol{y}^3, \ldots, \boldsymbol{y}^T$

As sampling from P(**y**|**x**)

Gibbs Sampling

- Is $P(y_i|y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N, \mathbf{x})$ easy to be computed?

$$P(y|x) = \frac{e^{F(x,y)}}{\boxed{\sum_{y''} e^{F(x,y'')}}}$$

$y_i \epsilon \{+1, -1\}$ ➡ 2^N possible y

Enumerate all possible y may not be tractable

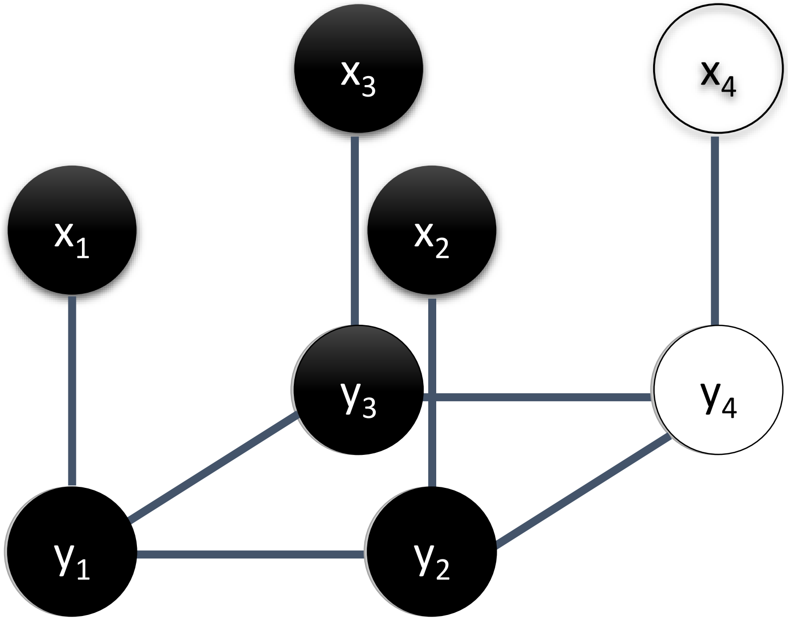
$$P(y_i|y_1, y_2, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N, \boldsymbol{x})$$

$$= \frac{e^{F(x,y_{-i},y_i)}}{\boxed{\sum_{y_i'} e^{F(x,y_{-i},y_i')}}}$$

$y_i \epsilon \{+1, -1\}$ ➡ 2 possible y_i

Enumerate all possible y_i may be tractable

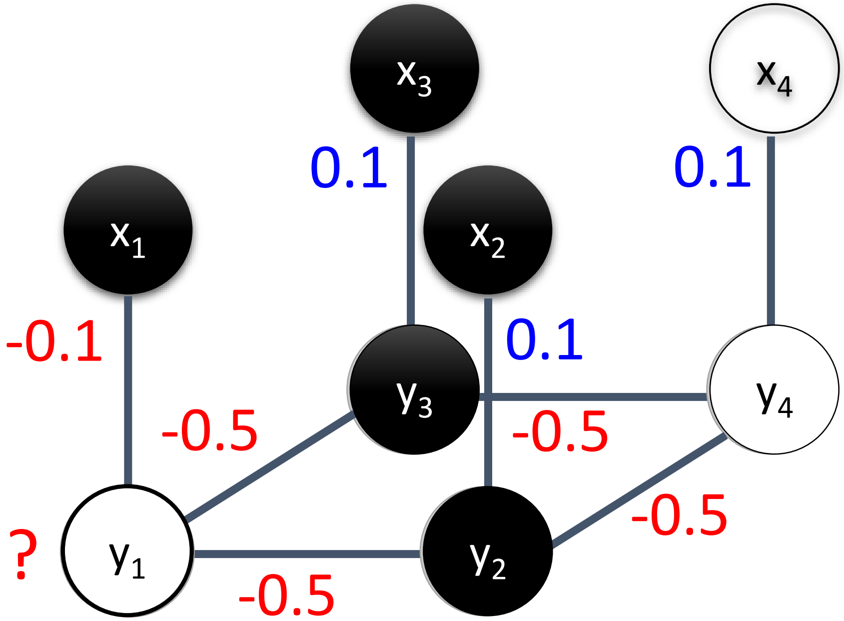
Initialization $y_1, y_2, y_3, y_4 = -1, -1, -1, 1$



Sample from P(**y**|**x**)
by Gibbs sampling

Sample y_1 given all the other variables

$$y_1 \sim P(y_1 | \boldsymbol{y_{-1}}, \boldsymbol{x}) \quad \boldsymbol{y_{-1}} = \{y_2, y_3, y_4\}$$



$$P(x, y) = \frac{e^{F(x,y)}}{\sum_{x',y'} e^{F(x',y')}}$$

Compute $P(y_1 = 1 | \boldsymbol{y_{-1}}, \boldsymbol{x})$ and $P(y_1 = -1 | \boldsymbol{y_{-1}}, \boldsymbol{x})$

$$P(y_1 = 1 | \boldsymbol{y_{-1}}, \boldsymbol{x}) = \frac{P(\boldsymbol{x}, y_1 = 1, \boldsymbol{y_{-1}})}{P(\boldsymbol{x}, y_1 = 1, \boldsymbol{y_{-1}}) + P(\boldsymbol{x}, y_1 = -1, \boldsymbol{y_{-1}})}$$

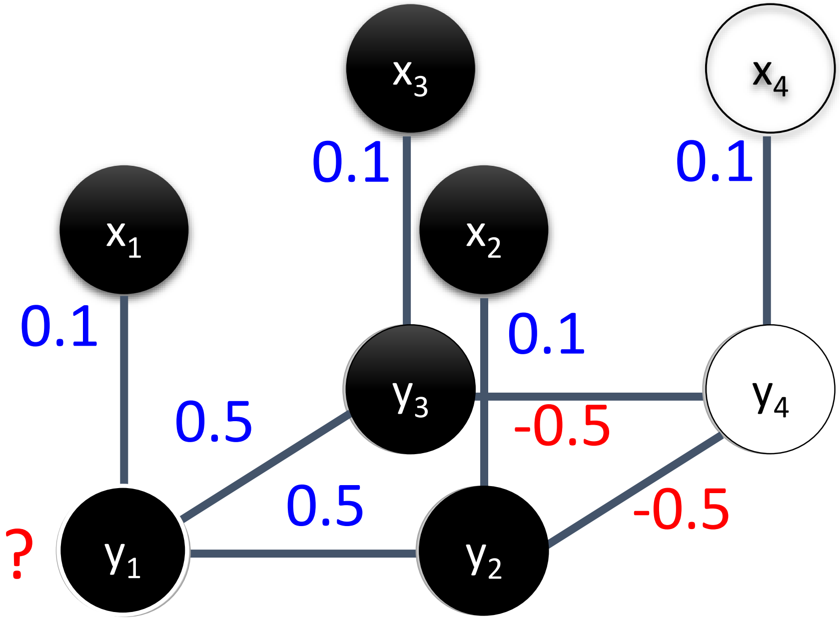
$$= \frac{e^{\boxed{F(x, y_1 = 1, \boldsymbol{y_{-1}})}}}{e^{\boxed{F(\boldsymbol{x}, y_1 = 1, \boldsymbol{y_{-1}})}} + e^{F(\boldsymbol{x}, y_1 = -1, \boldsymbol{y_{-1}})}}$$

= -1.8

= -1.8

Sample y_1 given all the other variables

$$y_1 \sim P(y_1 | \boldsymbol{y_{-1}}, \boldsymbol{x}) \qquad \boldsymbol{y_{-1}} = \{y_2, y_3, y_4\}$$



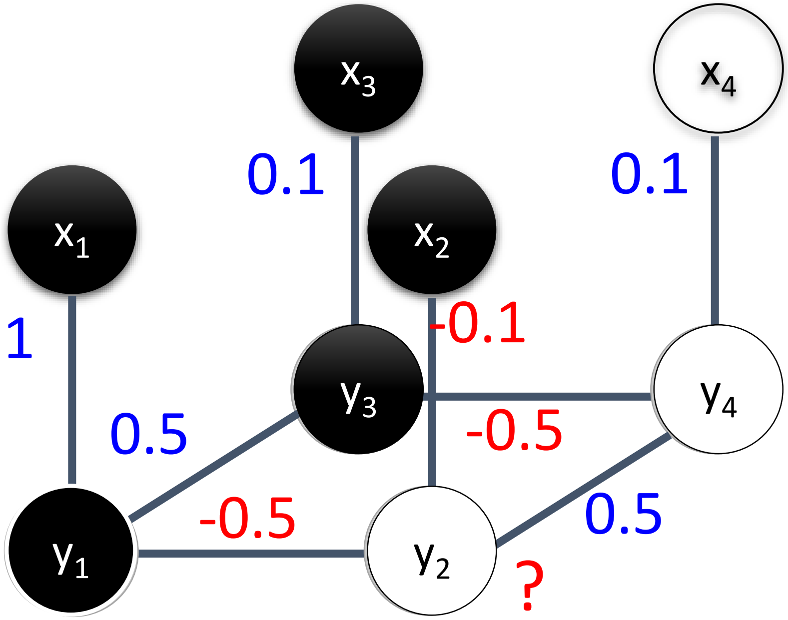
Compute $P(y_1 = 1 | \boldsymbol{y_{-1}}, \boldsymbol{x})$ and $P(y_1 = -1 | \boldsymbol{y_{-1}}, \boldsymbol{x})$

$$P(y_1 = 1 | \boldsymbol{y_{-1}}, \boldsymbol{x}) = \frac{P(\boldsymbol{x}, y_1 = 1, \boldsymbol{y_{-1}})}{P(\boldsymbol{x}, y_1 = 1, \boldsymbol{y_{-1}}) + P(\boldsymbol{x}, y_1 = -1, \boldsymbol{y_{-1}})}$$

$$= \frac{e^{\boxed{F(\boldsymbol{x}, y_1 = 1, \boldsymbol{y_{-1}})}} \quad = -1.8}{e^{\underbrace{\boxed{F(\boldsymbol{x}, y_1 = 1, \boldsymbol{y_{-1}})}}_{= -1.8}} + e^{\underbrace{\boxed{F(\boldsymbol{x}, y_1 = -1, \boldsymbol{y_{-1}})}}_{= 0.4}}} = 0.10 \xrightarrow[\text{Random sample}]{} y_1 = -1$$

Sample y_2 given all the other variables

$$y_2 \sim P(y_2|\boldsymbol{y_{-2}}, \boldsymbol{x})$$

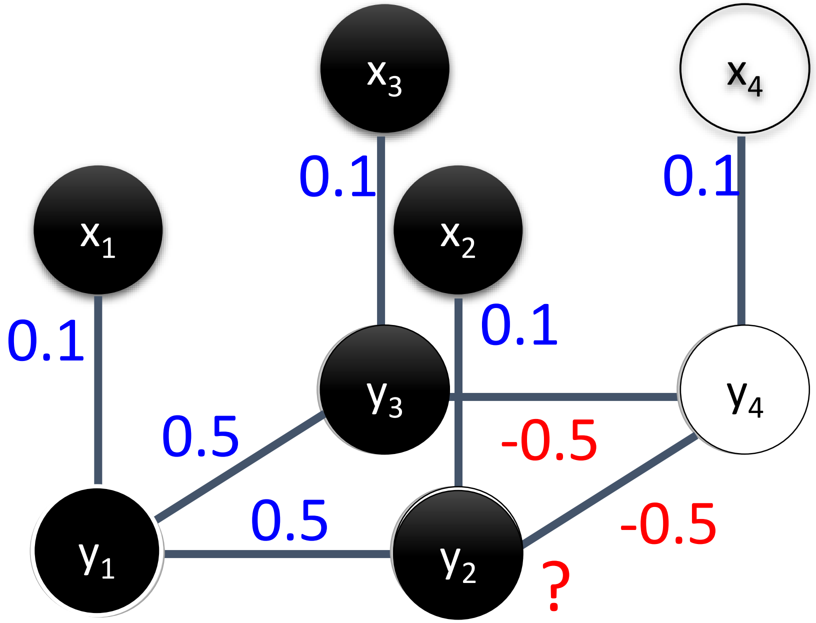


$$P(y_2 = 1|\boldsymbol{y_{-2}}, \boldsymbol{x}) = \frac{P(\boldsymbol{x}, y_2 = 1, \boldsymbol{y_{-2}})}{P(\boldsymbol{x}, y_2 = 1, \boldsymbol{y_{-2}}) + P(\boldsymbol{x}, y_2 = -1, \boldsymbol{y_{-2}})}$$

$$= \frac{e^{\boxed{F(\boldsymbol{x}, y_2=1, \boldsymbol{y_{-2}})}} \text{=0.2}}{\underbrace{e^{\boxed{F(\boldsymbol{x}, y_2=1, \boldsymbol{y_{-2}})}}}_{=0.2} + e^{F(\boldsymbol{x}, y_2=-1, \boldsymbol{y_{-2}})}}$$

Sample y_2 given all the other variables

$$y_2 \sim P(y_2 | \boldsymbol{y}_{-2}, \boldsymbol{x})$$

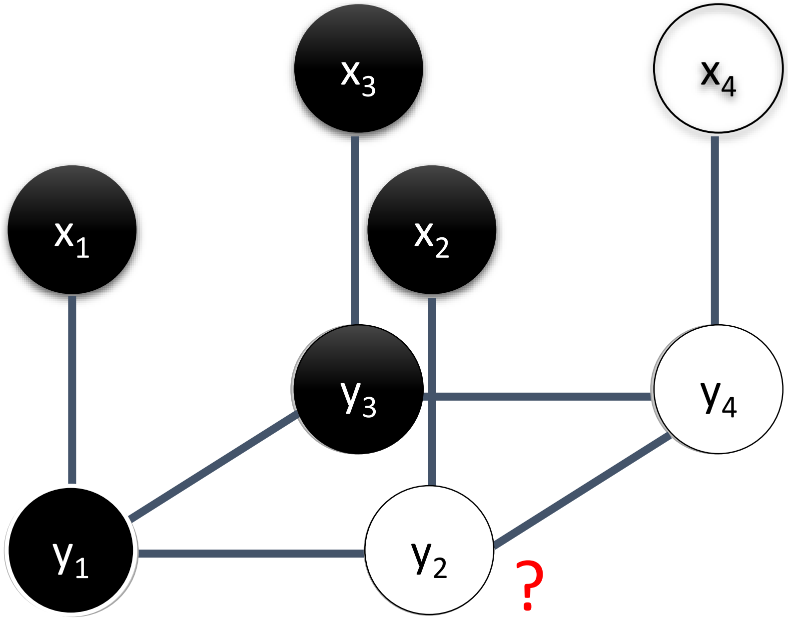


$$P(y_2 = 1 | \boldsymbol{y}_{-2}, \boldsymbol{x}) = \frac{P(\boldsymbol{x}, y_2 = 1, \boldsymbol{y}_{-2})}{P(\boldsymbol{x}, y_2 = 1, \boldsymbol{y}_{-2}) + P(\boldsymbol{x}, y_2 = -1, \boldsymbol{y}_{-2})}$$

$$= \frac{e^{\boxed{F(\boldsymbol{x}, y_2 = 1, \boldsymbol{y}_{-2})} = 0.2}}{e^{\underset{=0.2}{\boxed{F(\boldsymbol{x}, y_2 = 1, \boldsymbol{y}_{-2})}}} + e^{\underset{=0.4}{\boxed{F(\boldsymbol{x}, y_2 = -1, \boldsymbol{y}_{-2})}}}}$$

Sample y_2 given all the other variables

$$y_2 \sim P(y_2 | \boldsymbol{y}_{-2}, \boldsymbol{x})$$

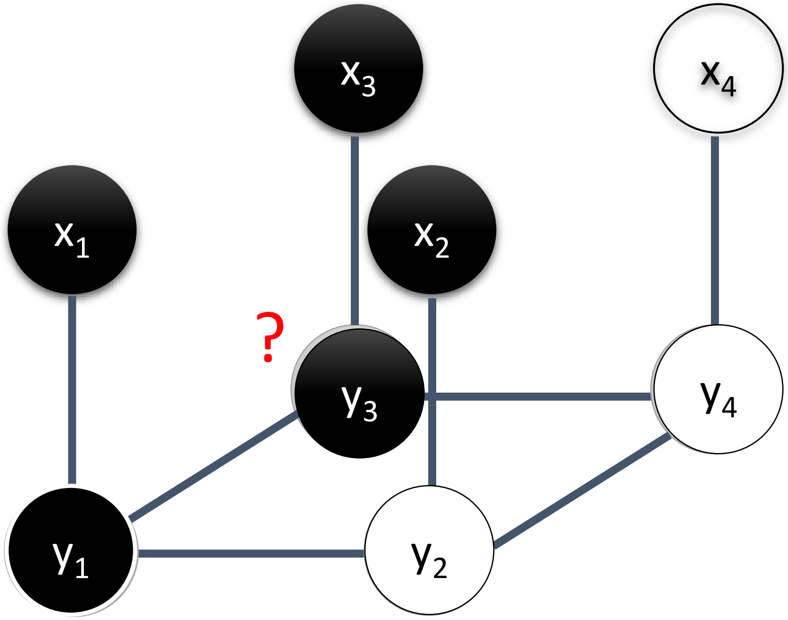


$$P(y_2 = 1 | \boldsymbol{y}_{-2}, \boldsymbol{x}) = \frac{P(\boldsymbol{x}, y_2 = 1, \boldsymbol{y}_{-2})}{P(\boldsymbol{x}, y_2 = 1, \boldsymbol{y}_{-2}) + P(\boldsymbol{x}, y_2 = -1, \boldsymbol{y}_{-2})}$$

$$= \frac{e^{\boxed{F(\boldsymbol{x}, y_2 = 1, \boldsymbol{y}_{-2})}}\ \text{=0.2}}{e^{\underset{\text{=0.2}}{\boxed{F(\boldsymbol{x}, y_2 = 1, \boldsymbol{y}_{-2})}}} + e^{\underset{\text{=0.4}}{\boxed{F(\boldsymbol{x}, y_2 = -1, \boldsymbol{y}_{-2})}}}} = 0.45 \xrightarrow[\text{Random sample}]{} y_2 = 1$$

Sample y_3 given all the other variables

$$y_3 \sim P(y_3 | \boldsymbol{y_{-3}}, \boldsymbol{x})$$



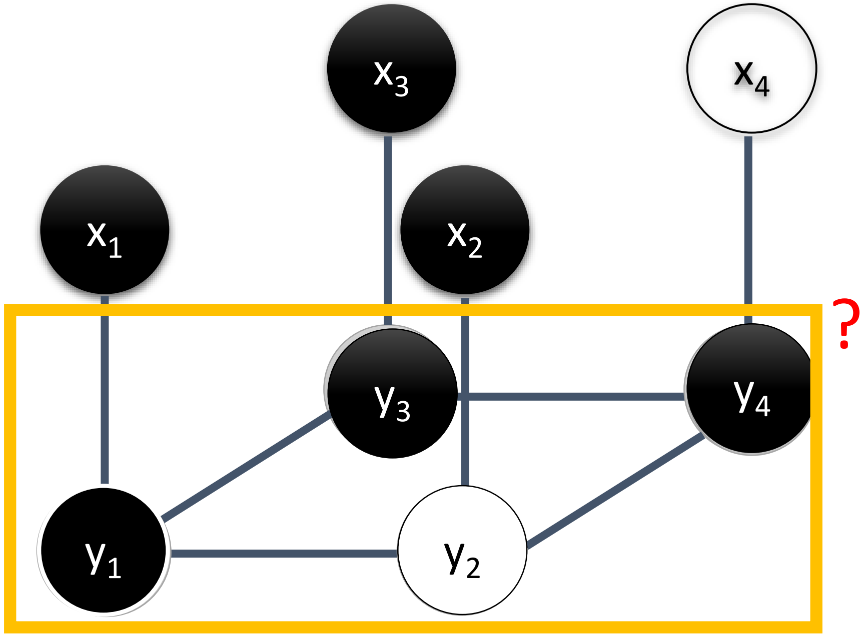
$$P(y_3 = 1 | \boldsymbol{y_{-3}}, \boldsymbol{x}) = ? \quad 0.45$$

$$\xrightarrow{\hspace{4cm}} y_3 = -1$$

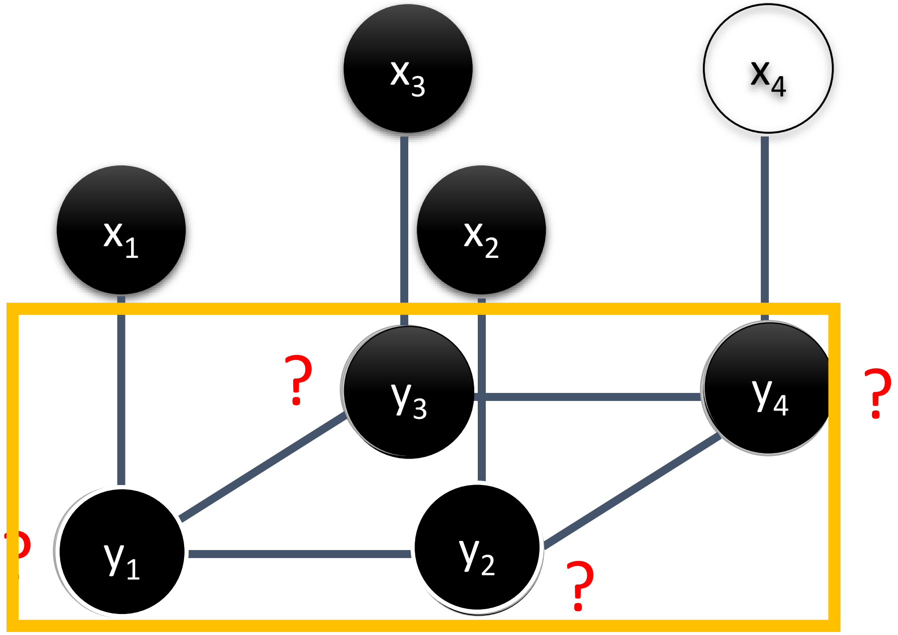
Random sample

Sample y_4 given all the other variables

$$y_4 \sim P(y_4 | \boldsymbol{y}_{-4}, \boldsymbol{x})$$



Get ***1-st*** sample y_1=-1, y_2=1, y_3=-1, y_4=-1



Get ***1-st*** sample y_1=-1, y_2=1, y_3=-1, y_4=-1

Get ***2-nd*** sample y_1=-1, y_2=-1, y_3=-1, y_4=-1

Get ***1-st*** sample $y_1=-1$, $y_2=1$, $y_3=-1$, $y_4=-1$

Get ***2-nd*** sample $y_1=-1$, $y_2=-1$, $y_3=-1$, $y_4=-1$

Get ***3-rd*** sample $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=1$

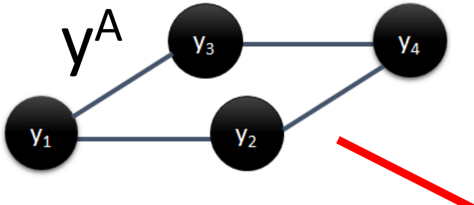
Get ***4-th*** sample $y_1=-1$, $y_2=1$, $y_3=-1$, $y_4=1$

Get ***5-th*** sample $y_1=1$, $y_2=1$, $y_3=1$, $y_4=1$

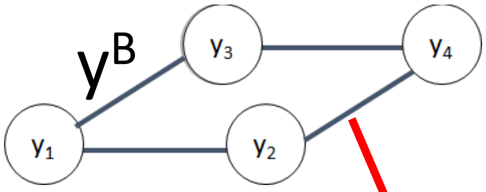
⋮

Until you want to stop

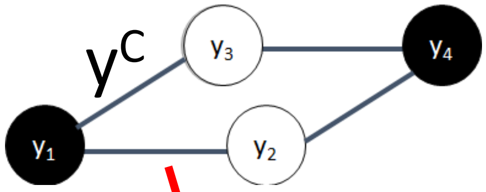
$P(y^A|x)\approx0.33$



$P(y^B|x)\approx0.22$



$P(y^C|x)\approx0.004$



No. of samples	y^A	y^B	y^C
10	2	3	0
100	23	37	0
1000	315	230	8
10000	3307	2225	40
100000	32637	22129	422

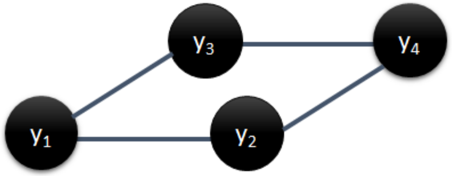
$P(y^A|x)\approx0.33$ $P(y^B|x)\approx0.22$ $P(y^C|x)\approx0.004$

From sampling: y^A would be the results of inference.

How about starting from different initialization?

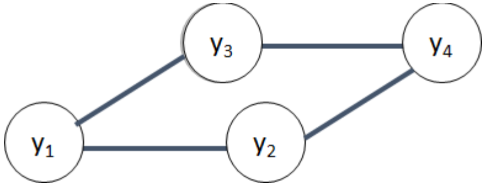
Not really change the final results.

Starting from …



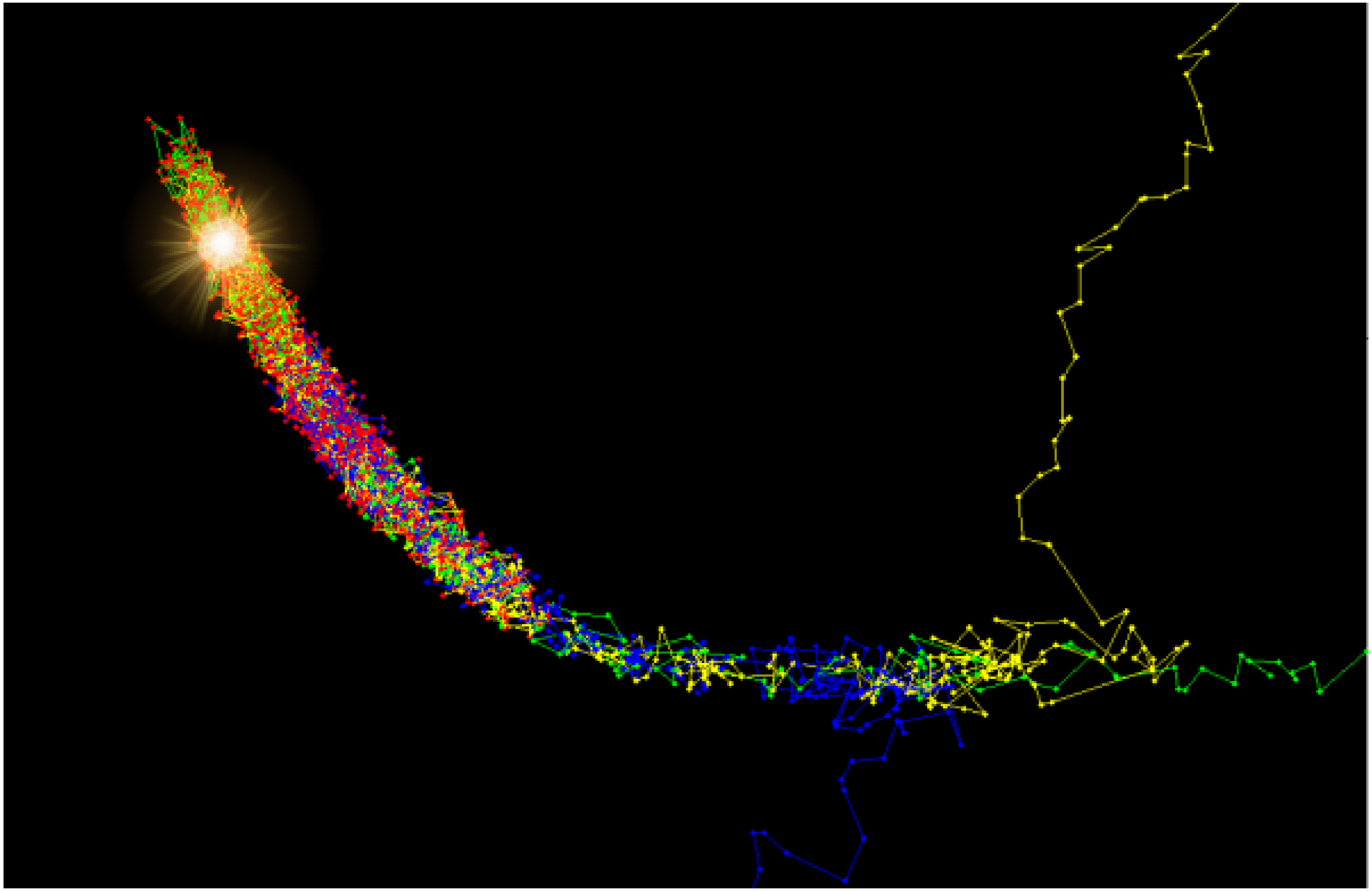
No. of samples	A	B	C
10	3	1	0
100	40	11	1
1000	331	237	2
10000	3251	2176	31
100000	32911	21845	385

Starting from …



No. of samples	A	B	C
10	0	3	0
100	28	31	0
1000	318	226	2
10000	3277	2169	46
100000	32319	21751	393

All rivers run into the sea.



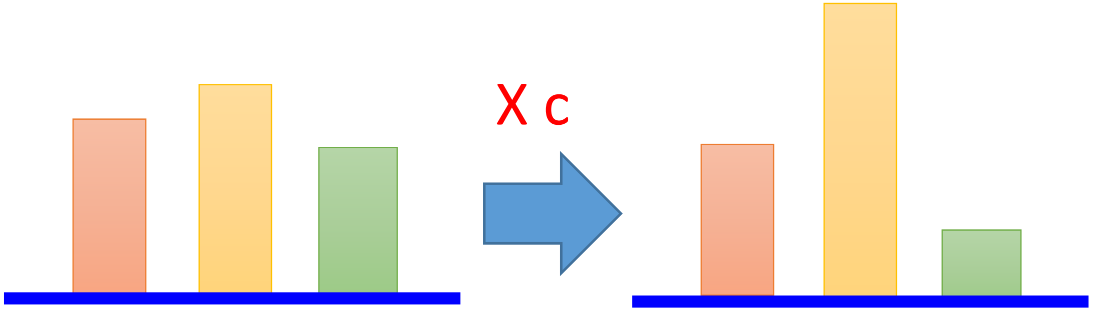
http://www.juergenwiki.de/work/wiki/doku.php?id=public:mcmc

Practical Suggestion

- "burn-in"
  - "burn-in" period: The first few of samples would be influenced by the initialization
  - Discard the samples in the "burn-in" period
- Modify the sampling distribution

c > 1

$$P(y_i|y_1, y_2, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N, \boldsymbol{x}) = \frac{e^{F(x,y_{-i},y_i) \text{ X c}}}{\sum_{y_i'} e^{F(x,y_{-i},y_i') \text{ X c}}}$$



Increase c after each interaction

Gibbs Sampling

A little bit of theory

Gibbs Sampling

Gibbs sampling from a distribution P(z) ($z = \{z_1,\ldots,z_N\}$)

```
$\boldsymbol{z}^{\mathbf{0}} = \{z_1^0, z_2^0, \cdots, z_N^0\}$
For t = 1 to T:
    $z_1^t \sim P(z_1 | z_2 = z_2^{t-1}, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
    $z_2^t \sim P(z_2 | z_1 = z_1^t, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
    $z_3^t \sim P(z_3 | z_1 = z_1^t, z_2 = z_2^t, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
      ⋮
    $z_N^t \sim P(z_N | z_1 = z_1^t, z_2 = z_2^t, z_3 = z_3^t, \cdots, z_{N-1} = z_{N-1}^t)$
    Output: $\boldsymbol{z}^t = \{z_1^t, z_2^t, \cdots, z_N^t\}$
```

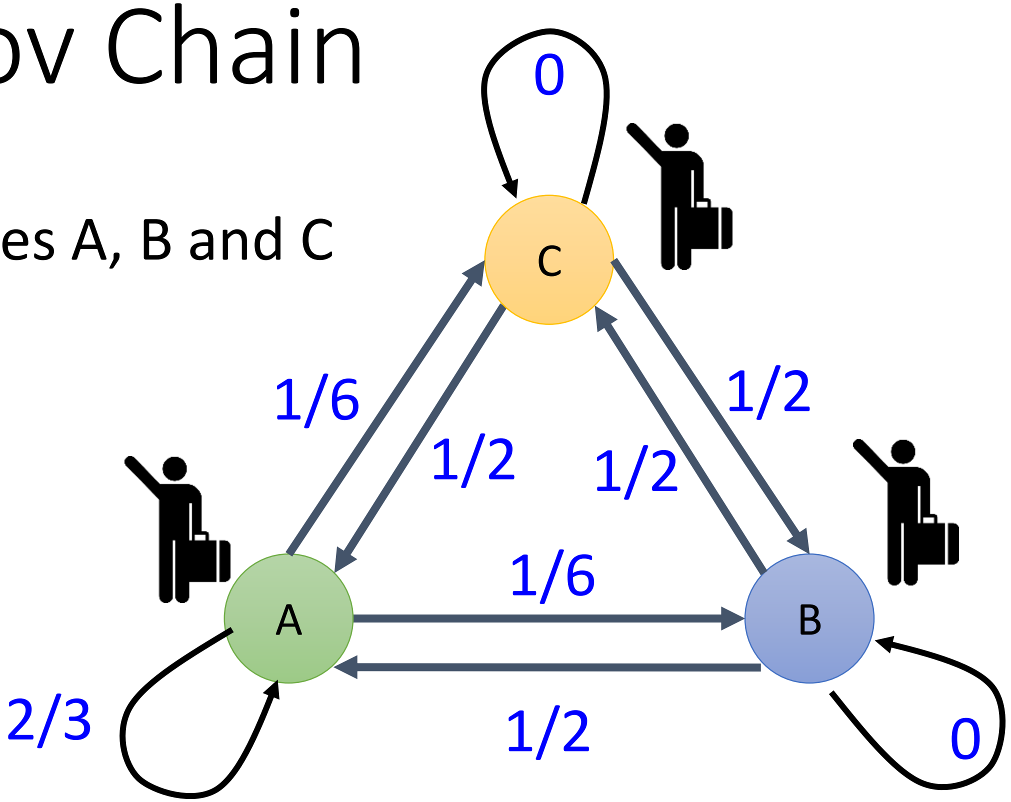
$\boldsymbol{z}^1, \boldsymbol{z}^2, \boldsymbol{z}^3, \ldots, \boldsymbol{z}^T$

As sampling from P(z)

Why?

Markov Chain

Three cities A, B and C



The traveler recorded the cities he visited each day.

A B C A A ……

This is a ***Markov chain***

state

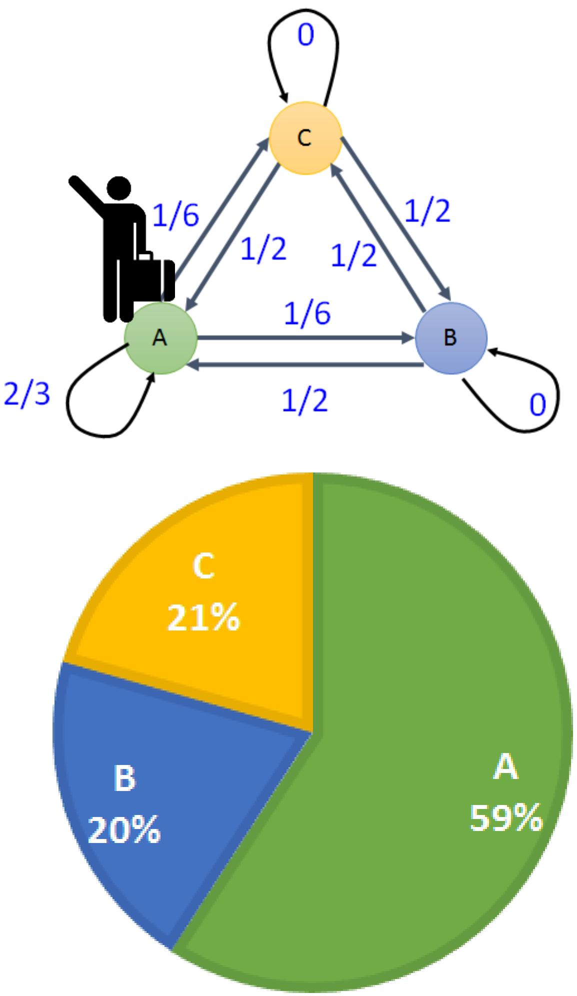
Markov Chain

With sufficient samples ……

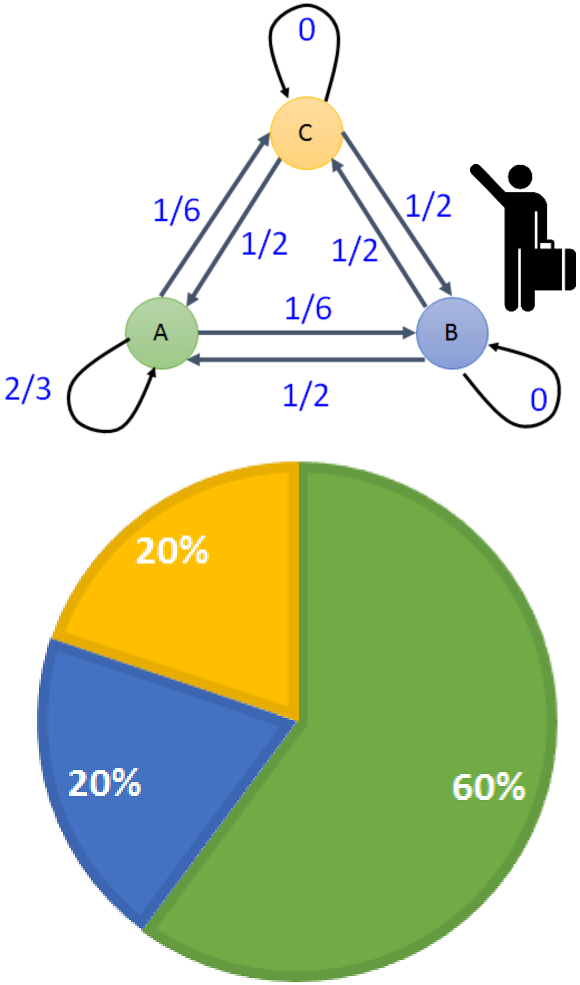
A : B : C = 0.6 : 0.2 : 0.2

(independent of the starting city)

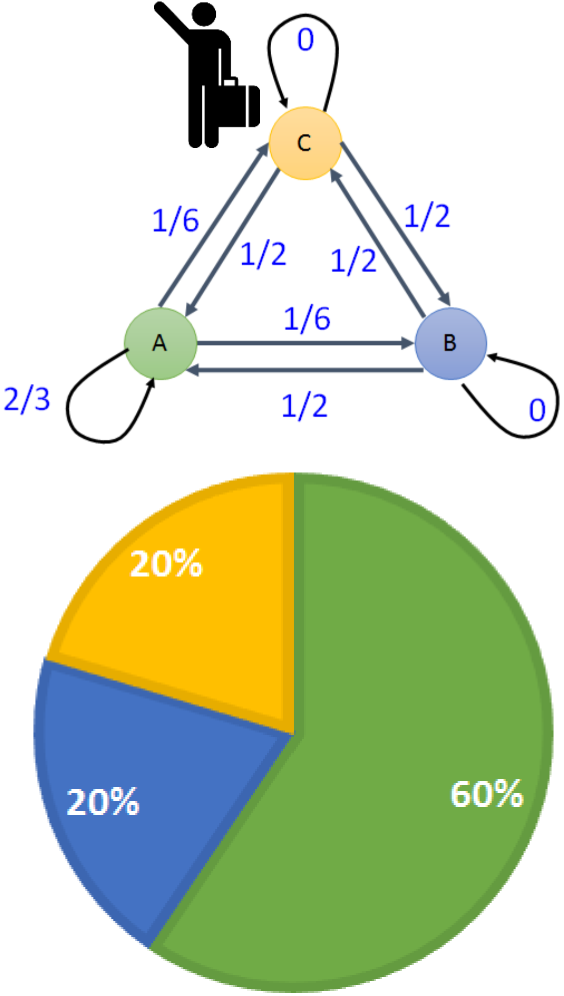
10000 days



10000 days



10000 days



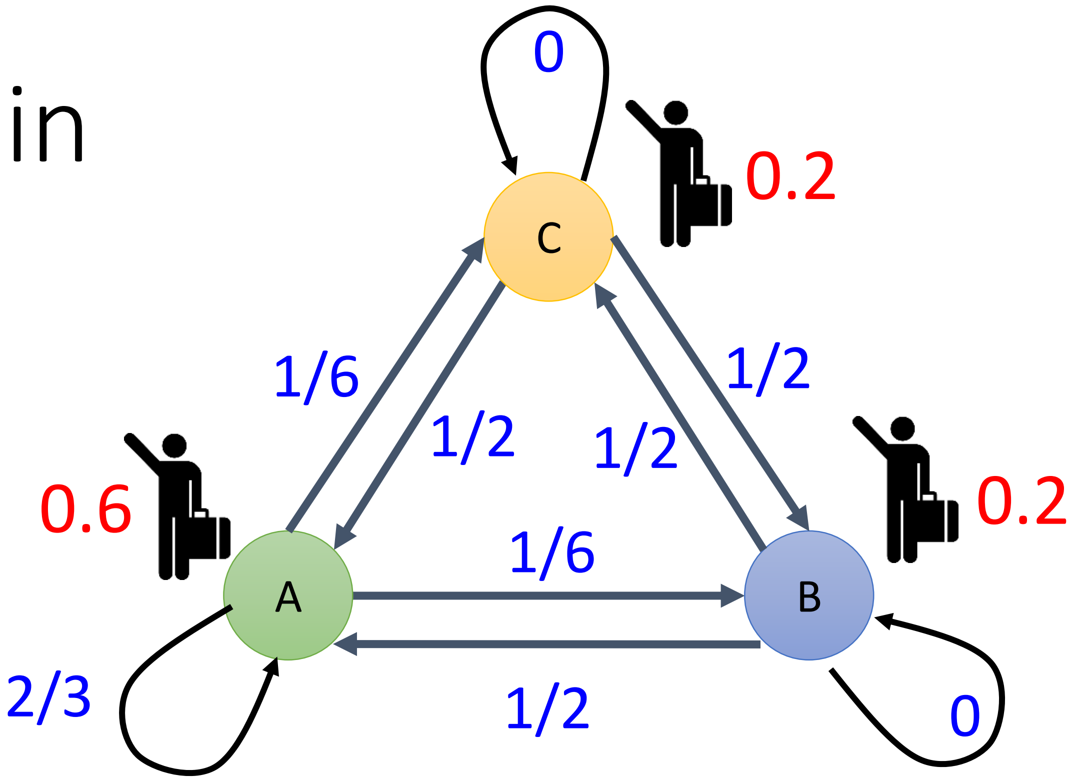
Markov Chain

P(A)=0.6

P(B)=0.2

P(C)=0.2

Stationary Distribution



$$\overset{2/3}{P_T(A|A)}\overset{0.6}{P(A)} + \overset{1/2}{P_T(A|B)}\overset{0.2}{P(B)} + \overset{1/2}{P_T(A|C)}\overset{0.2}{P(C)} = \overset{0.6}{P(A)}$$

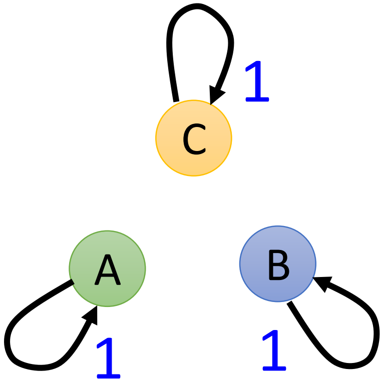
$$P_T(B|A)P(A) + P_T(B|B)P(B) + P_T(B|C)P(C) = P(B)$$

$$P_T(C|A)P(A) + P_T(C|B)P(B) + P_T(C|C)P(C) = P(C)$$

The distribution will not change.

Markov Chain

A Markov Chain can have multiple stationary distributions.



- Reaching which stationary distribution depends on starting state

The Markov Chain fulfill some conditions will have unique stationary distribution.

$P_T(s'|s)$ for any states s and s' is not zero

(*sufficient* but not *necessary* condition)

Unique stationary distribution

Markov Chain
from Gibbs Sampling

Gibbs sampling from a distribution P(z) ($z = \{z_1,\ldots,z_N\}$)

```
$\boldsymbol{z}^0 = \{z_1^0, z_2^0, \cdots, z_N^0\}$
For t = 1 to T:
    $z_1^t \sim P(z_1 | z_2 = z_2^{t-1}, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
    $z_2^t \sim P(z_2 | z_1 = z_1^t, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
    $z_3^t \sim P(z_3 | z_1 = z_1^t, z_2 = z_2^t, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
    ⋮
    $z_N^t \sim P(z_N | z_1 = z_1^t, z_2 = z_2^t, z_3 = z_3^t, \cdots, z_{N-1} = z_{N-1}^t)$
    Output: $\boldsymbol{z}^t = \{z_1^t, z_2^t, \cdots, z_N^t\}$
```

$\boldsymbol{z}^1, \boldsymbol{z}^2, \boldsymbol{z}^3, \ldots, \boldsymbol{z}^T$

state

- This is a ***Markov Chain***
  - z^t only depend on z^{t-1}

Markov Chain
from Gibbs Sampling

Gibbs sampling from a distribution P(z) ($z = \{z_1,\ldots,z_N\}$)

```
$\boldsymbol{z}^{\boldsymbol{0}} = \{z_1^0, z_2^0, \cdots, z_N^0\}$
For t = 1 to T:
    $z_1^t \sim P(z_1|z_2 = z_2^{t-1}, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
    $z_2^t \sim P(z_2|z_1 = z_1^t, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
    $z_3^t \sim P(z_3|z_1 = z_1^t, z_2 = z_2^t, z_4 = z_4^{t-1}, \cdots, z_N = z_N^{t-1})$
      ⋮
    $z_N^t \sim P(z_N|z_1 = z_1^t, z_2 = z_2^t, z_3 = z_3^t, \cdots, z_{N-1} = z_{N-1}^t)$
    Output: $\boldsymbol{z}^{\boldsymbol{t}} = \{z_1^t, z_2^t, \cdots, z_N^t\}$
```

$\boldsymbol{z}^1, \boldsymbol{z}^2, \boldsymbol{z}^3, \ldots, \boldsymbol{z}^T$

Proof that the Markov chain has unique stationary distribution which is P(z).

Markov Chain
from Gibbs Sampling

- Markov chain from Gibbs sampling has unique stationary distribution? Yes
  - $P_T(z'|z) > 0$, for any z and z'

$$\left.\begin{aligned}
z_1^t &\sim P(z_1|z_2 = z_2^{t-1}, z_3 = z_3^{t-1}, \cdots, z_N = z_N^{t-1}) \\
z_2^t &\sim P(z_2|z_1 = z_1^t, z_3 = z_3^{t-1}, \cdots, z_N = z_N^{t-1}) \\
z_3^t &\sim P(z_3|z_1 = z_1^t, z_2 = z_2^t, \cdots, z_N = z_N^{t-1}) \\
&\vdots \\
z_N^t &\sim P(z_N|z_1 = z_1^t, z_2 = z_2^t, \cdots, z_{N-1} = z_{N-1}^t)
\end{aligned}\right\}$$

None of the conditional probability is zero

can be any z^t

Markov Chain from Gibbs Sampling

- Show that P(z) is a stationary distribution

$$\sum_{z} P_T(z'|z)\, P(z) = P(z')$$

$$\begin{aligned} P_T(z'|z) = {} & P(z_1'|z_2, z_3, z_4, \cdots, z_N) \\ & \times P(z_2'|z_1', z_3, z_4, \cdots, z_N) \\ & \times P(z_3'|z_1', z_2', z_4, \cdots, z_N) \\ & \vdots \\ & \times P(z_N'|z_1', z_2', z_3', \cdots, z_{N-1}') \end{aligned}$$

Please do the math yourself

There is only one stationary distribution for Gibbs sampling, so we are done.

Thank you for your attention!

Acknowledgement

- 感謝 陳乃群 同學於上課時糾正投影片 上的錯誤