Royal Institute of Technology

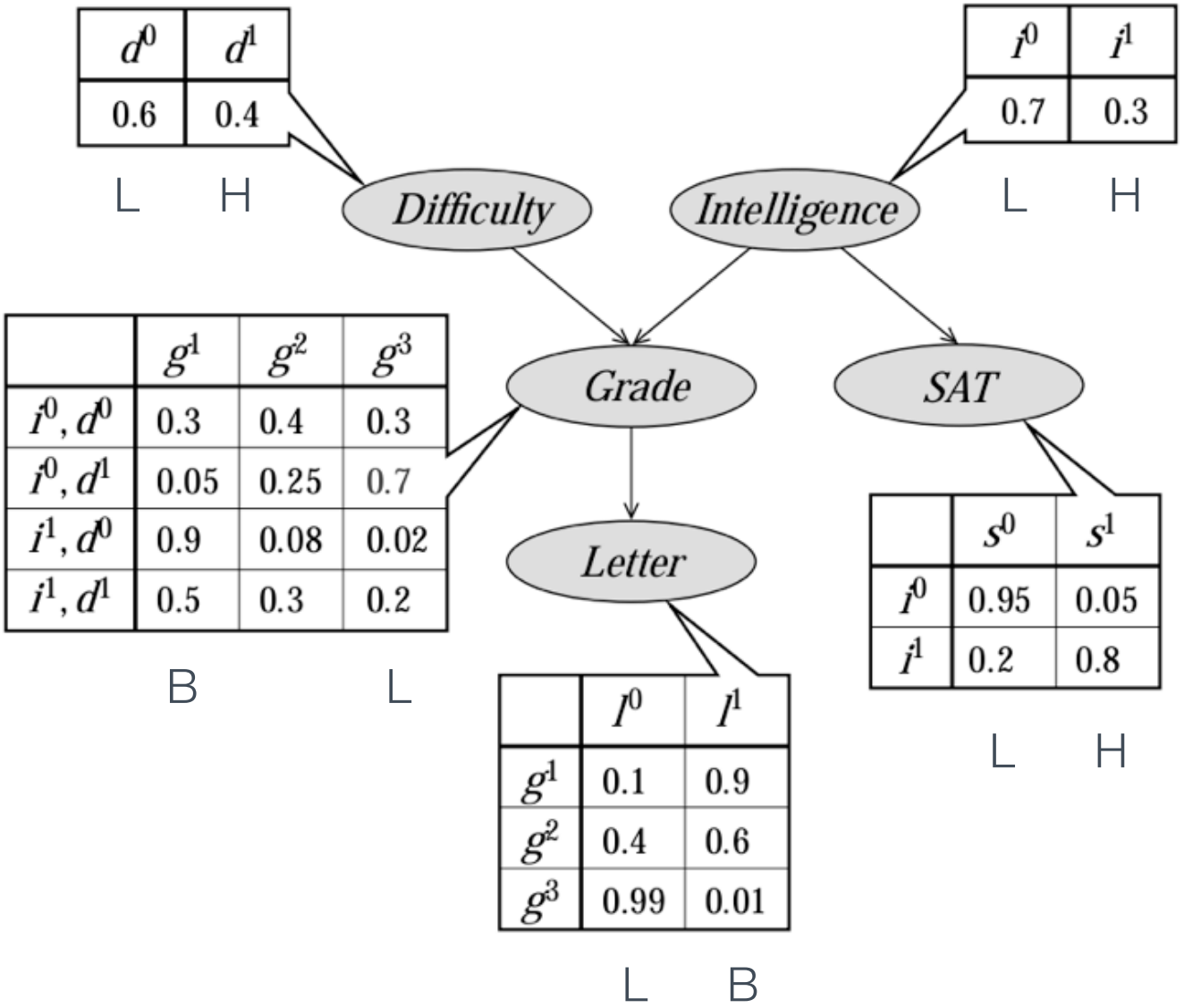
MACHINE LEARNING 2 – UGM,HMMS

Lecture 7

THIS LECTURE

- DGM semantics
- UGM
- De-noising
- HMMs
  - Applications (interesting probabilities)
  - DP for generation probability etc.
  - (later Baum-Welch)

EXTENDED STUDENT EXAMPLE

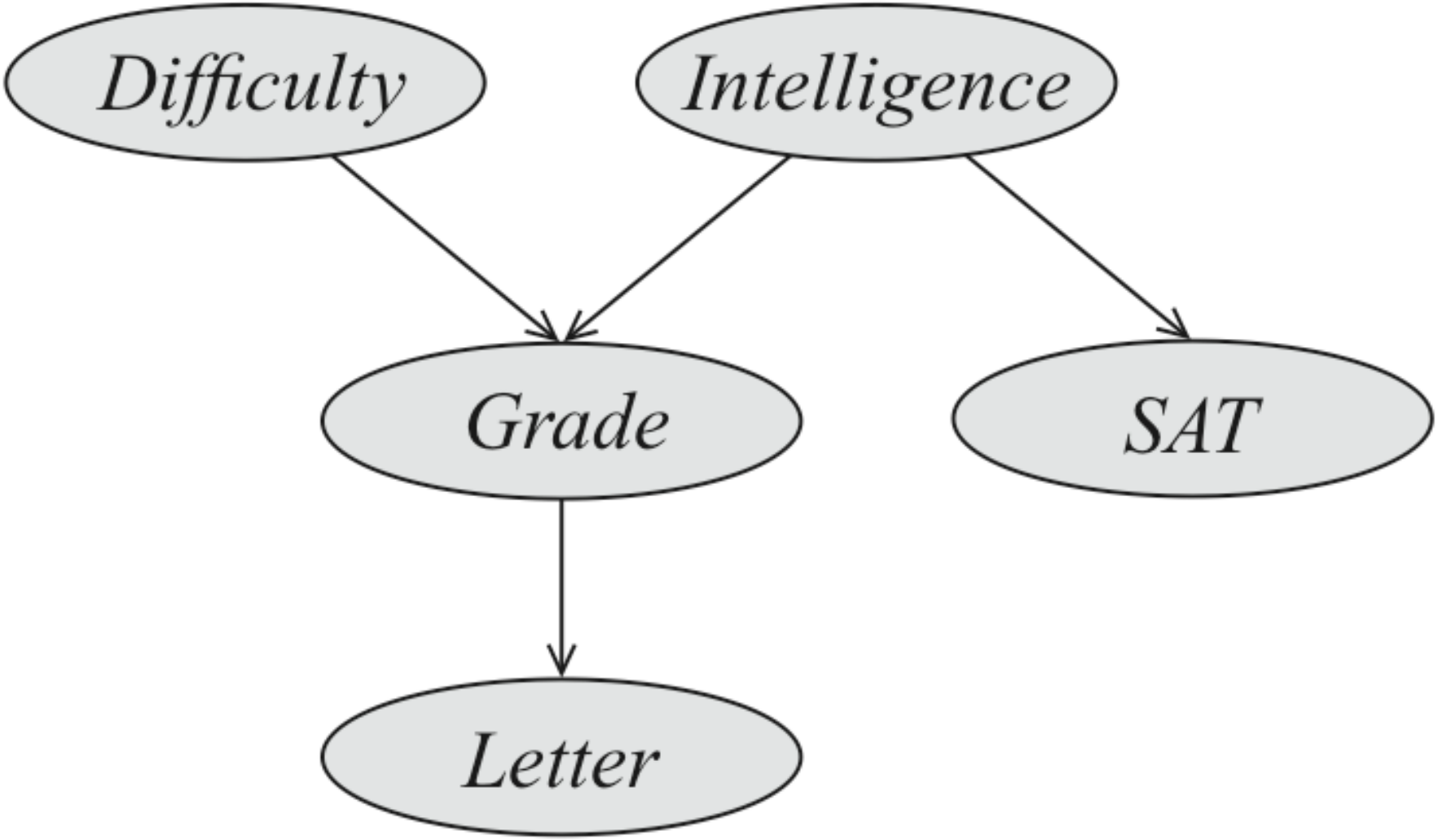


B - better

H - higher

L - less

EXTENDED STUDENT EXAMPLE



INDEPENDENCE I-MAP

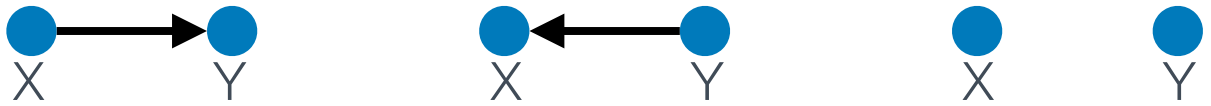
- I(G) (conditional) independences implied by G (not yet defined)
- I(P) (conditional) independences in the distribution P
- G I-map for P in I(G) ⊆ I(P)

p

X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

q

X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1



INDEPENDENCE I-MAP

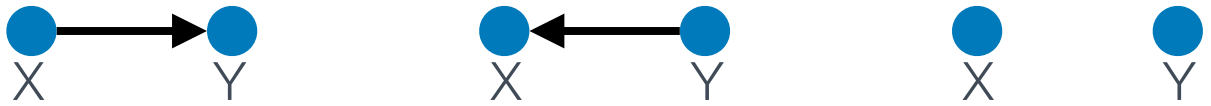
- I(G) independences implied by G (not yet defined)
- I(P) independences in the distribution P
- G I-map for P in I(G) ⊆ I(P)

p

X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

q

X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1



- p: X and Y ind. ex. p(X=1) = 0.48 + 0.12 =0.6, p(Y=1) = 0.8, and p(X=1,Y=1) = 0.48
- q: X and Y are dependent

INDEPENDENCE I-MAP

- I(G) independences implied by G (not yet defined)
- I(P) independences in the distribution P
- G I-map for P in I(G) ⊆ I(P)

p

X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

q

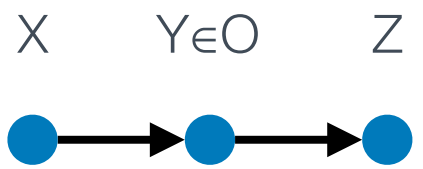
X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1

- All three graphs are I-maps for p
- G_1 and G_2 are I-maps for q, but G_3 is not

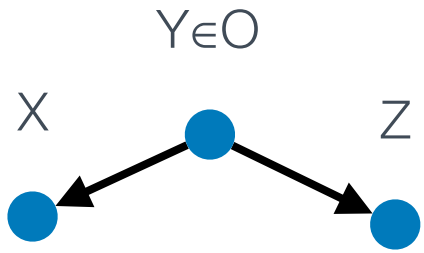
D-SEPARATION

★ A path is d-separated by O if it has

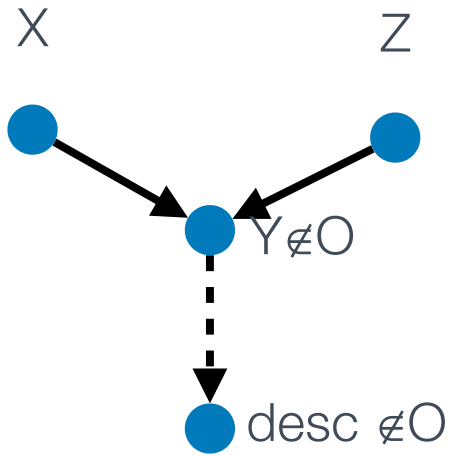
- a chain $X \rightarrow Y \rightarrow Z$ where $Y \in O$
- a fork $X \leftarrow Y \rightarrow Z$ where $Y \in O$
- a v-structure $X \rightarrow Y \leftarrow Z$ where $(Y \cup \text{desc}(Y)) \cap O = \varnothing$



Chain

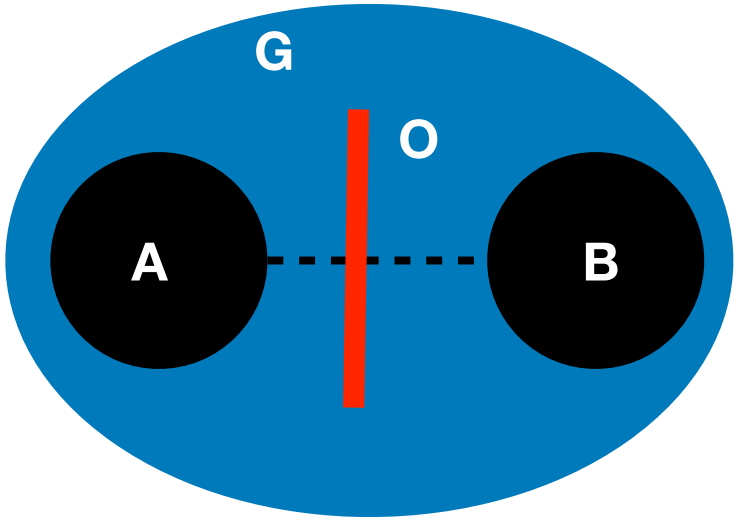


Fork



v-struct

D-SEPARATION SETS AND CI OF DAGS



★ A is d-separated from B given O if every undirected path between A and B is d-separated by O

★ Cond. ind rel. in DAG G,

$$\boldsymbol{x}_A \perp_G \boldsymbol{x}_B | \boldsymbol{x}_O$$

A is d-separated from B given O

FACTORIZATION OVER G

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} p(x_n | \boldsymbol{x}_{\mathrm{pa}(x_n)})$$

p can be factorized over G if it can be expressed as above

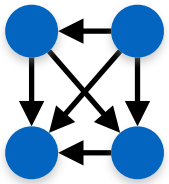
SOUNDNESS AND COMPLETENESS

- I(G) conditional independence relations implied by d-sep in G
- I(p) conditional independence relations satisfied by p

- Theorem

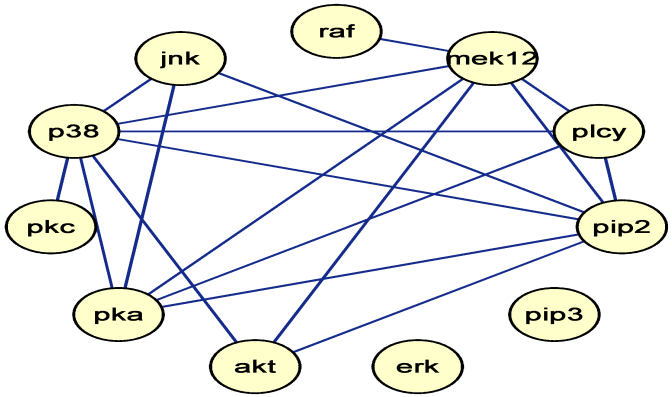
  A distribution P can be factorised over G iff $I(G) \subseteq I(p)$

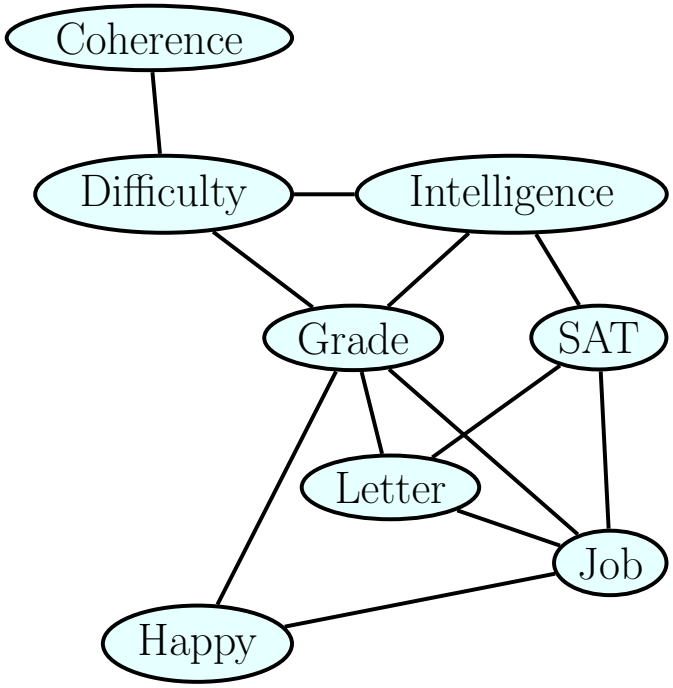
- "=" not possible to achieve, ex. clique and independent distribution



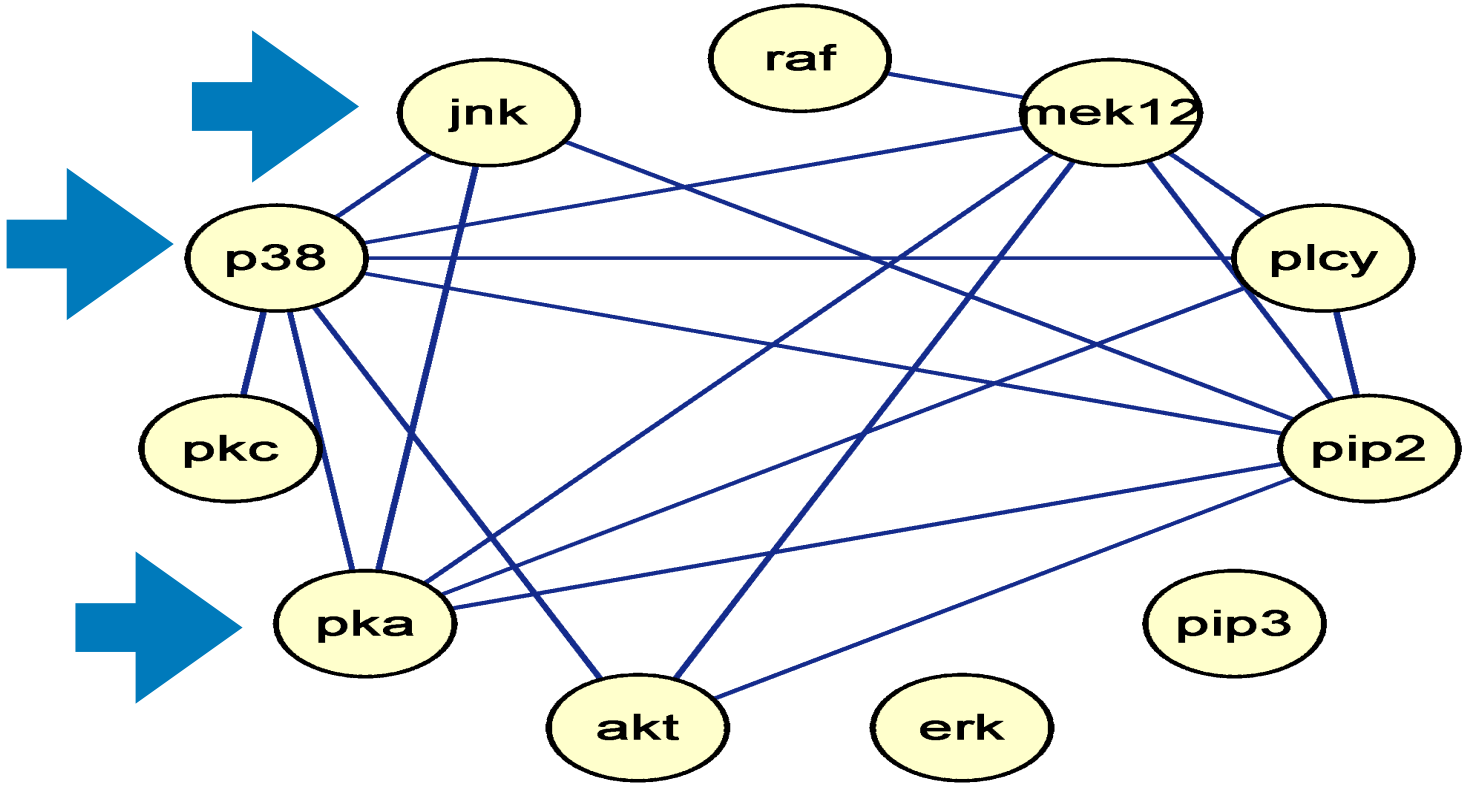
UGM

- UGMs - Undirected graphical models
- What is the direction between 2 pixels, 2 proteins?
- Probabilistic interpretation?
- p factorizes over G – can be expressed as normalized product over factors associated with cliques

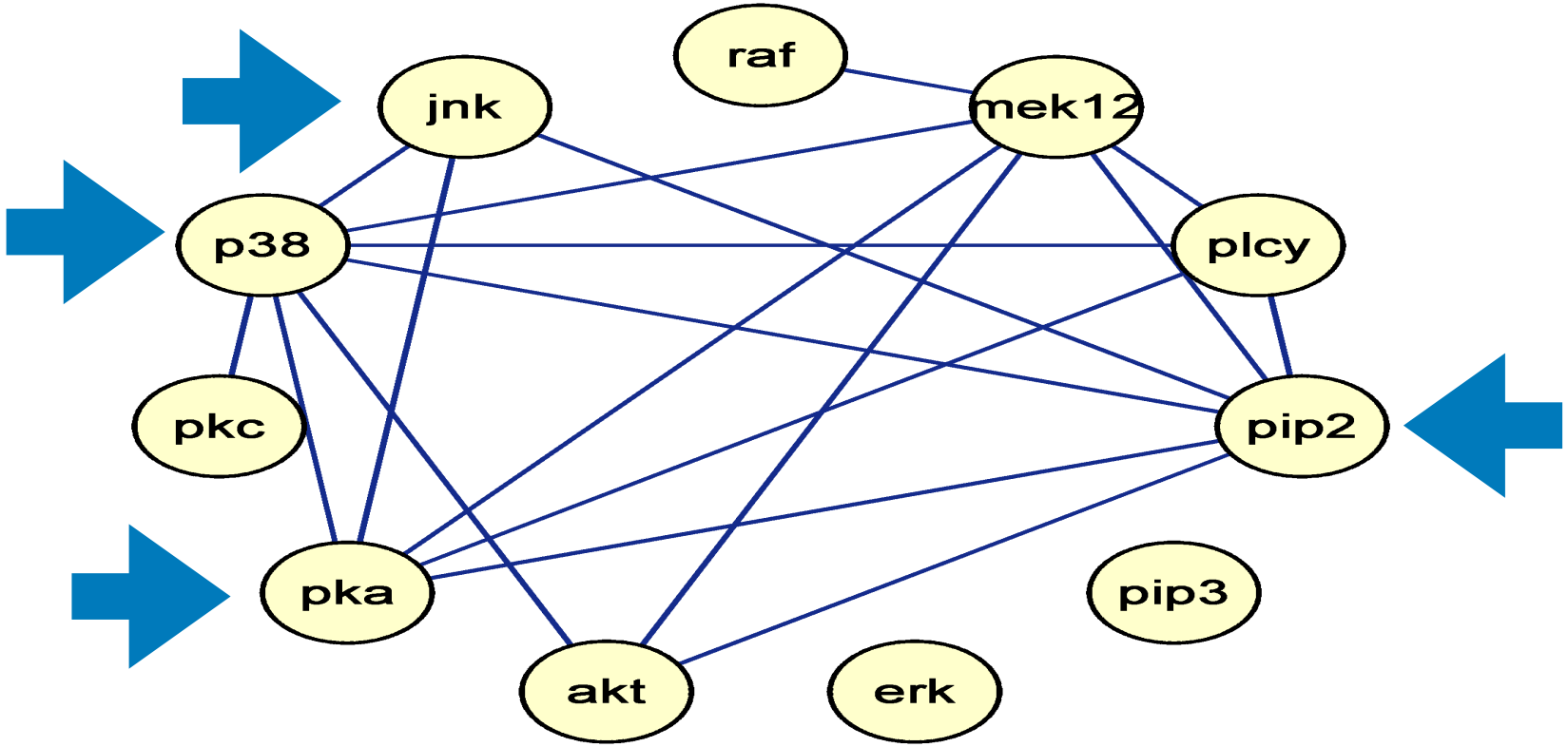




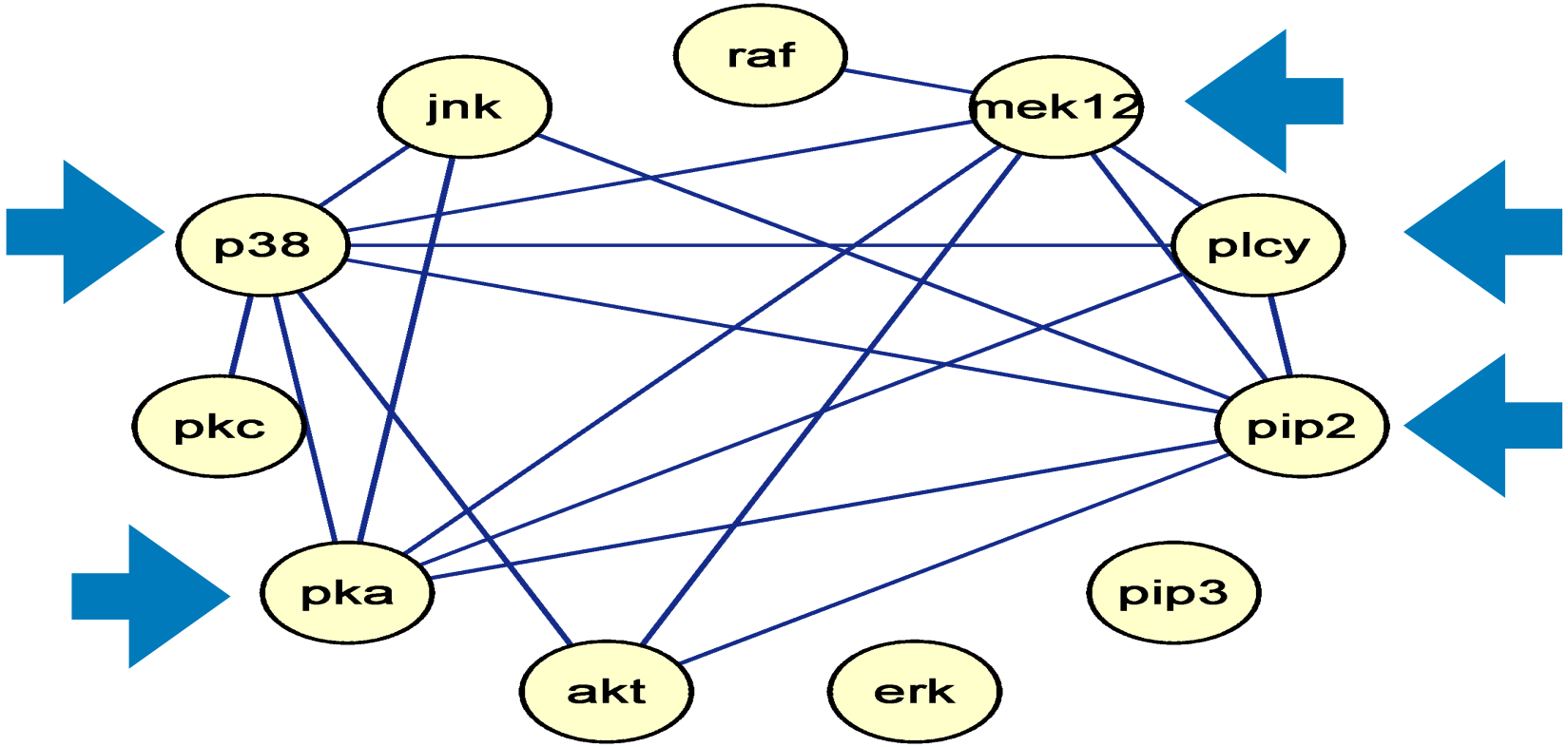
EXAMPLE CLIQUE



EXAMPLE MAXIMAL CLIQUE



EXAMPLE MAXIMUM CLIQUE

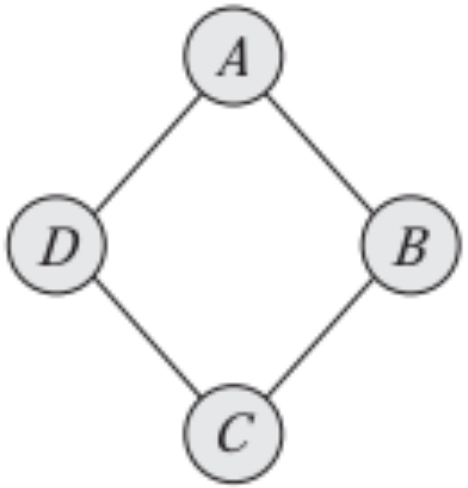


UGM

- An undirected graph G with so-called factors associated with its maximal cliques $C(G)$, for $C \in C(G)$ factor ψ_C
- ψ_C is a function from the clique's variables (the scope) to non-neg real numbers

$$p(x_1, \ldots, x_V) = \frac{1}{Z} \prod_{C \in C(G)} \psi_C(x_C)$$

$$Z = \sum_{x_1, \ldots, x_V} \prod_{C \in C(G)} \psi_C(x_C)$$



Scope A,B B,C C,D D,A

$\phi_1(A, B)$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

(a)

$\phi_2(B, C)$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

(b)

$\phi_3(C, D)$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

(c)

$\phi_4(D, A)$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

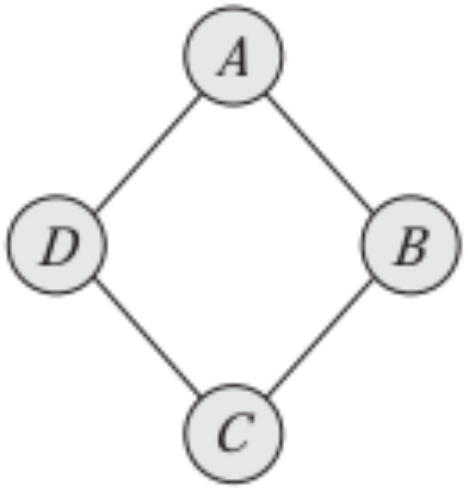
(d)

Factors – misconception example

- $$P(A, B, C, D) = \frac{1}{Z}\phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(D, A)$$

- $$Z = \sum_{a,b,c,d} \phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(d, a)$$

PROBABILISTIC INTERPRETATION



$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100
(a)			(b)			(c)			(d)		

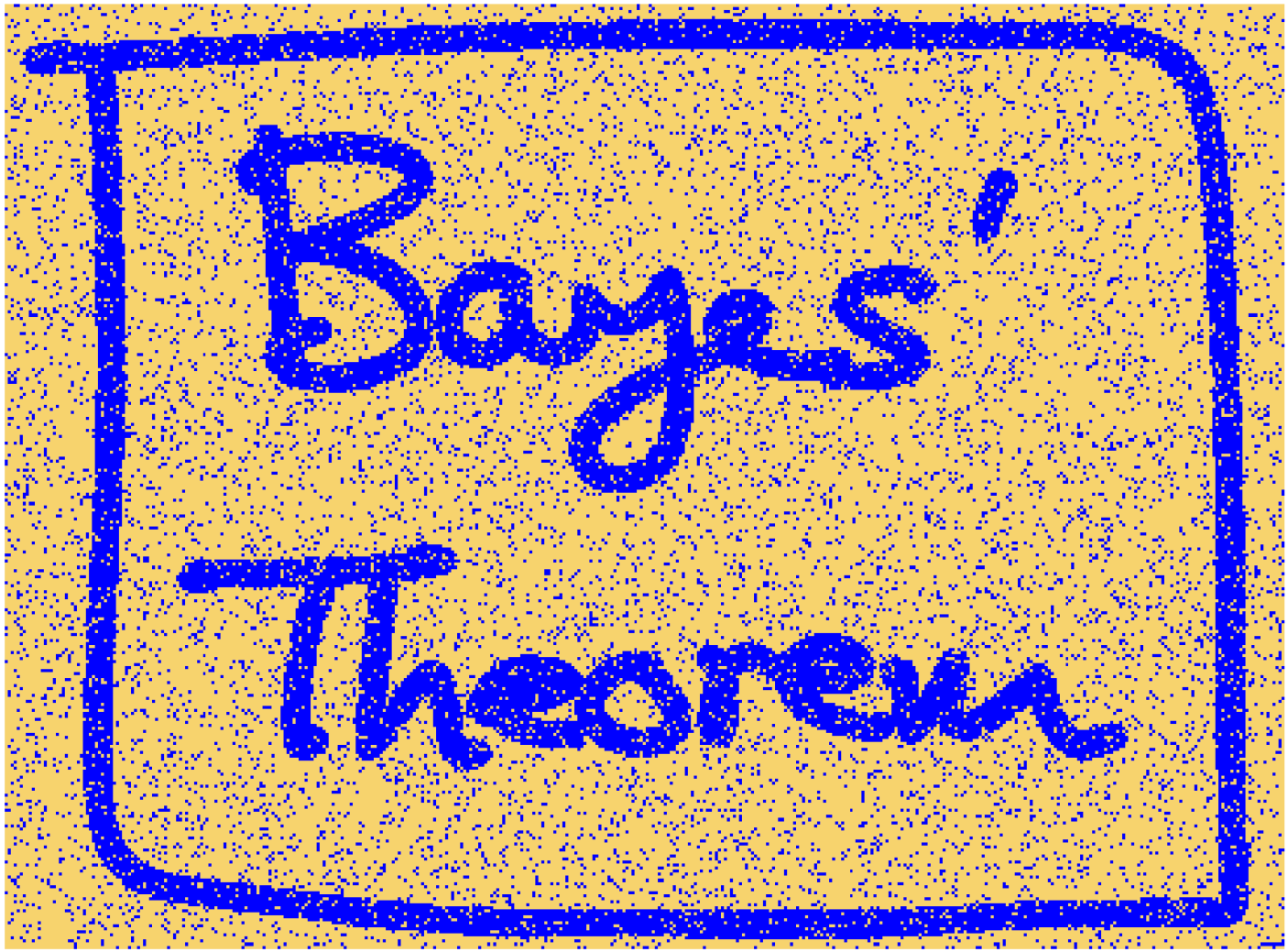
Misconception

$$\phi_1(A=1,B=1)\phi_2(B=1,C=0)\phi_3(C=0,D=1)\phi_4(D=1,A=1)$$
$$= 10 \cdot 1 \cdot 100 \cdot 100$$
$$= 100000$$

$$Z = \sum_{a,b,c,d} \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)$$

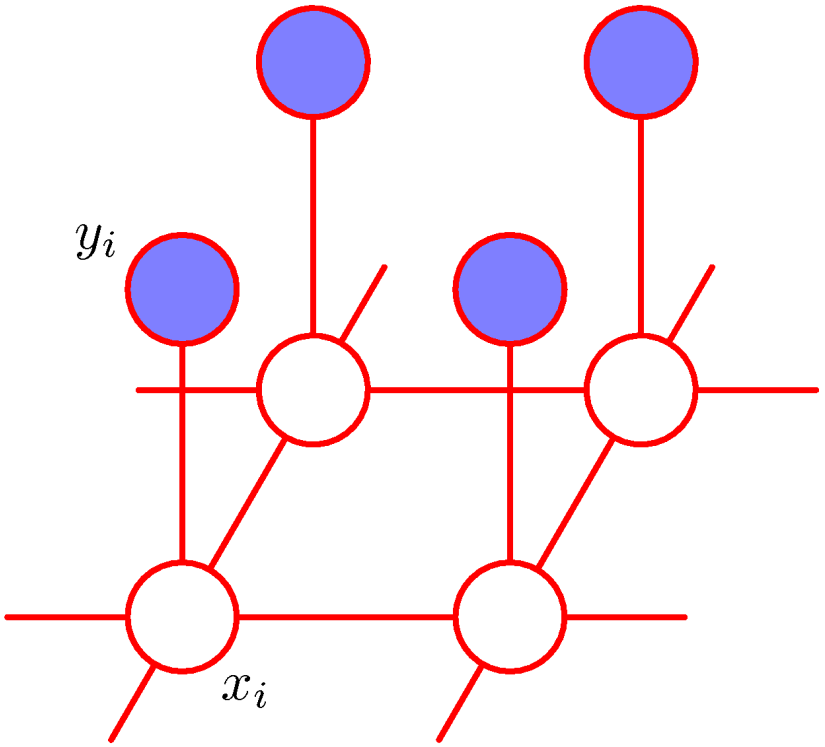
A FACTOR PRODUCT

DE-NOISING



ISING MODEL-DE-NOISING

p(y | x) ex Gaussian



Values -1,1

Factors of form

$$e^{\beta x_i x_j}$$

and

$$e^{\eta x_i y_i}$$

ISING MODEL-DE-NOISING

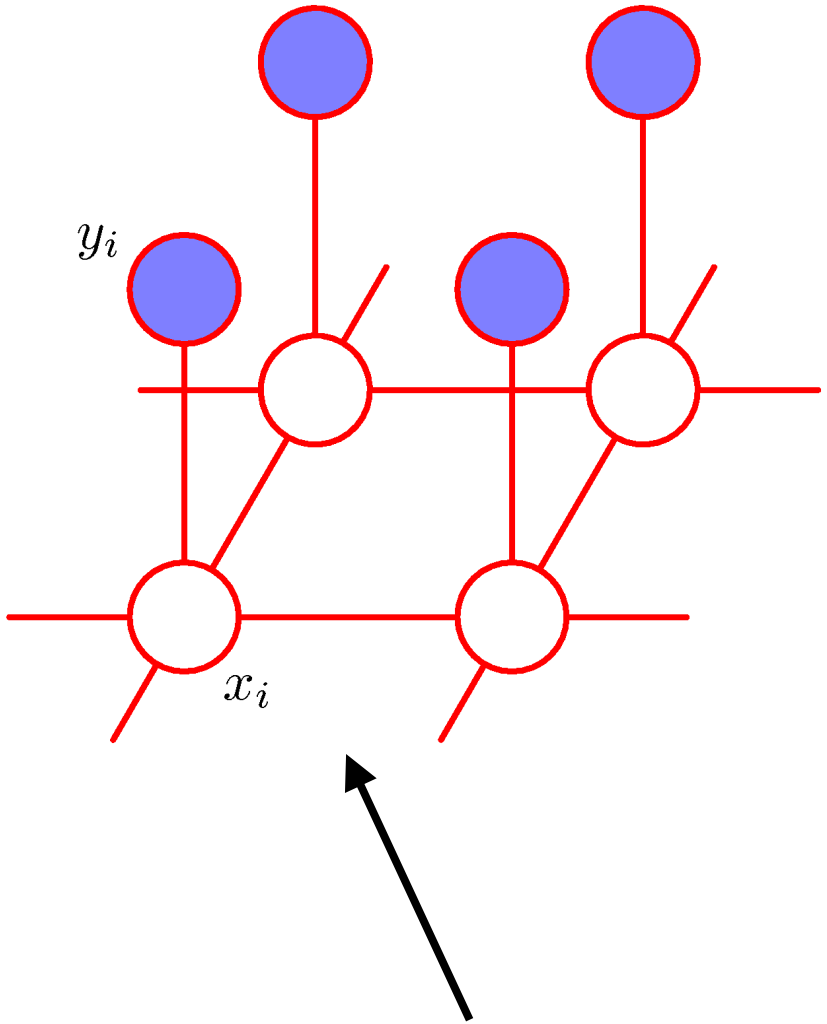
Values -1,1

Factors of form

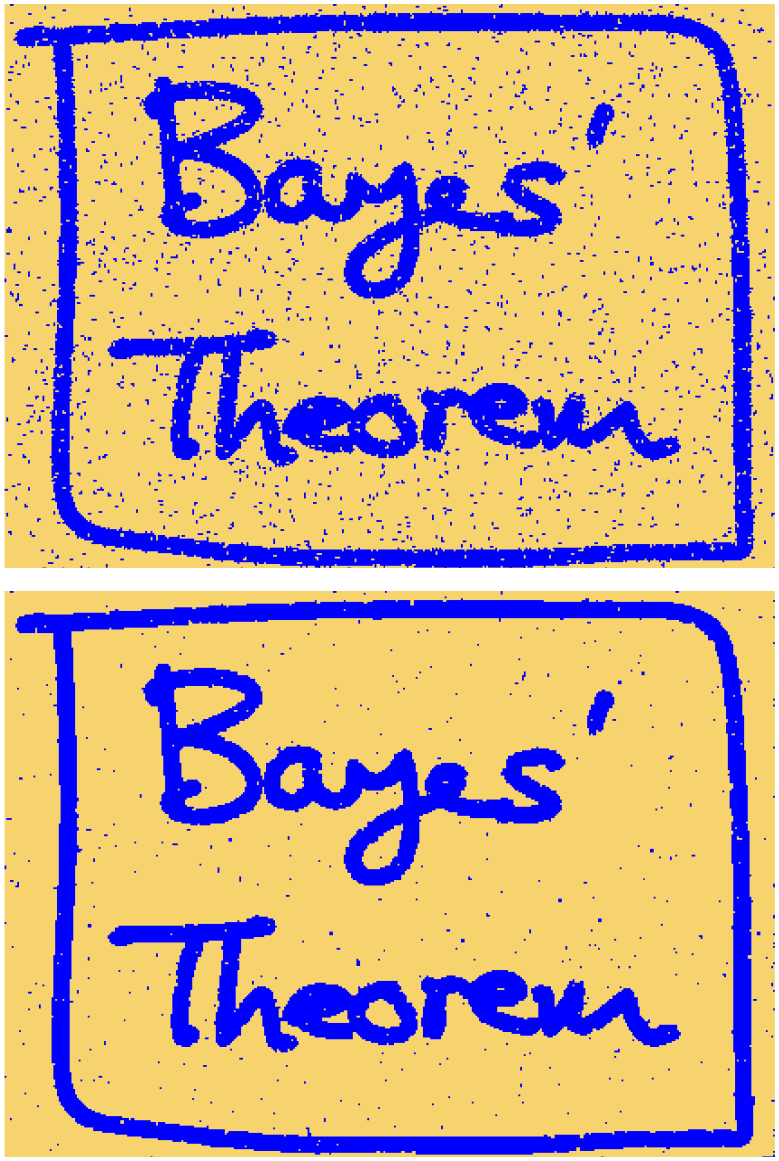
$$e^{\beta x_i x_j}$$

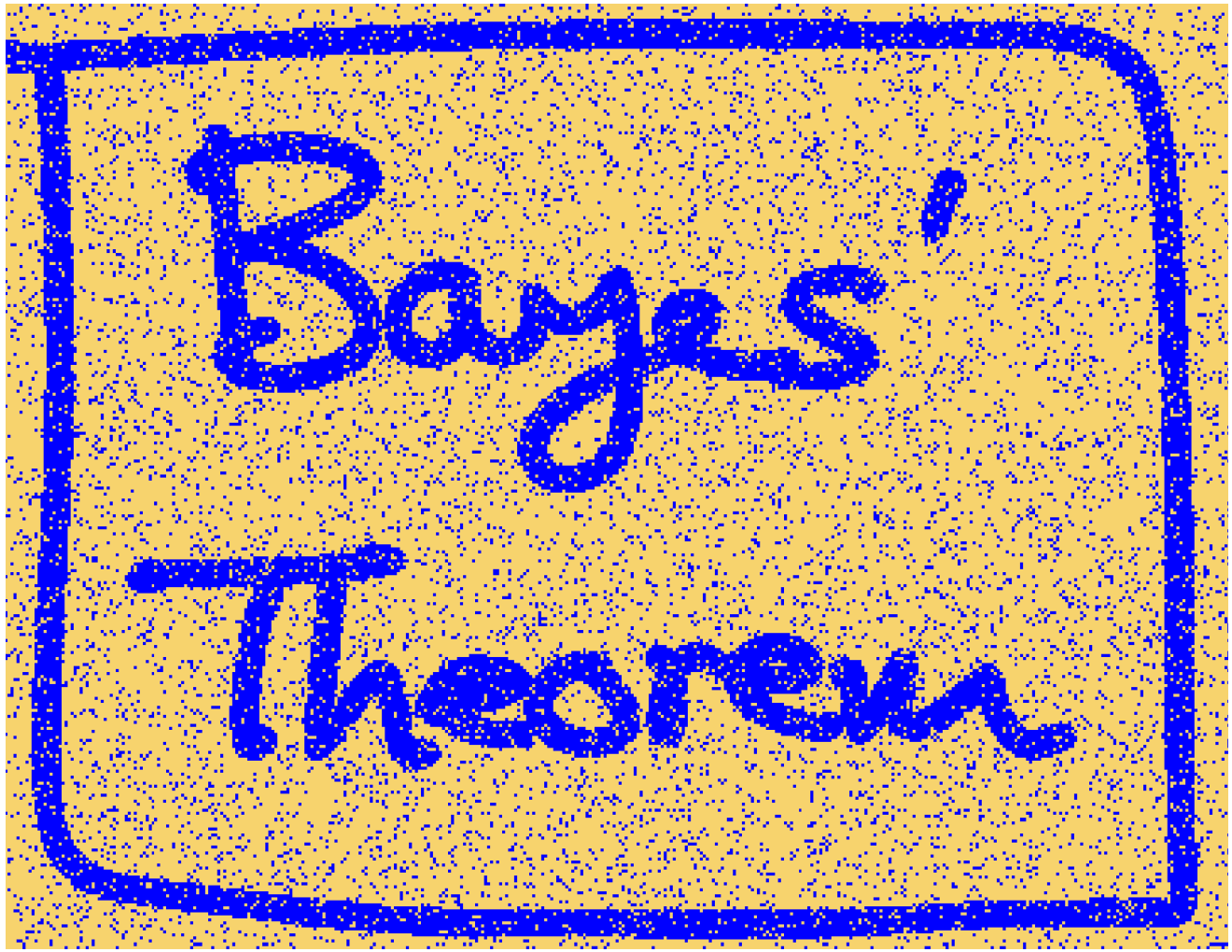
and

$$e^{\eta x_i y_i}$$



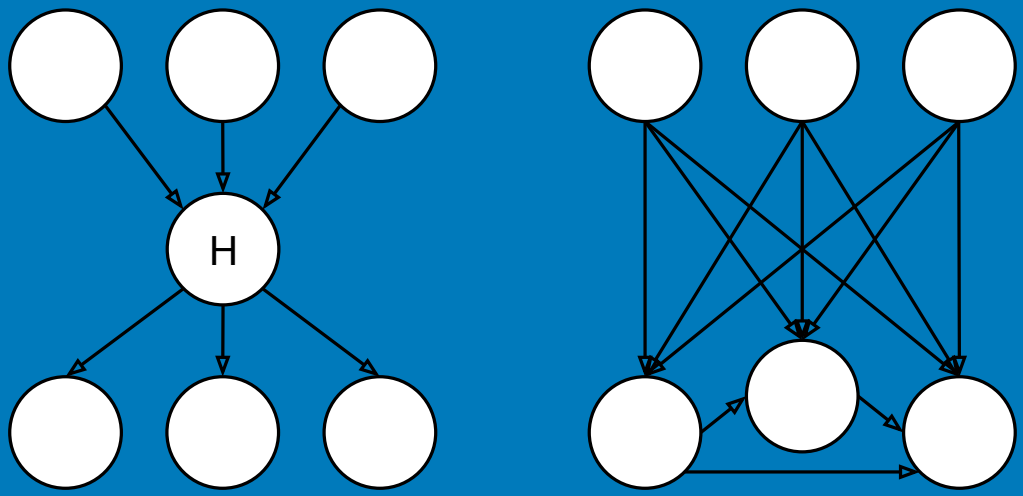
- Bipartite graph
- Suggests iterative procedu





- Large is the noisy image; upper, UGM de-noised; and lower, graph cut de-noised

LATENT = HIDDEN

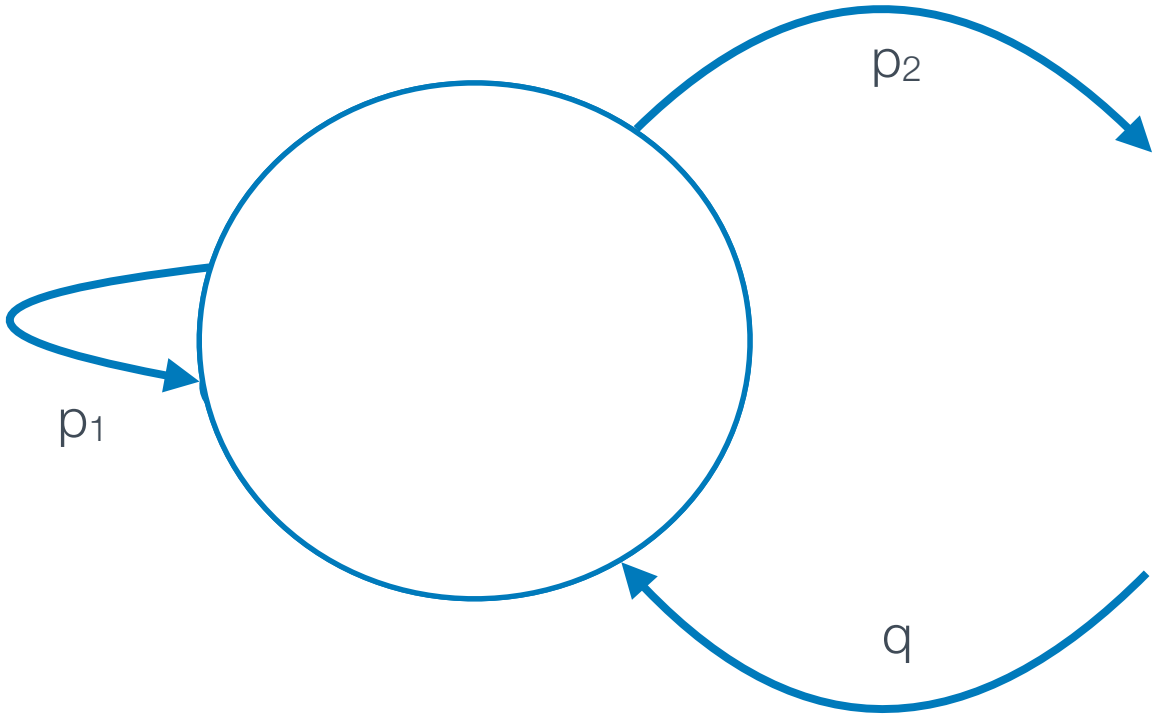


17 parameters

59 parameters

- ★ Can reduce #parameters
- ★ Can represent common causes

MARKOV CHAINS (DISCRETE)

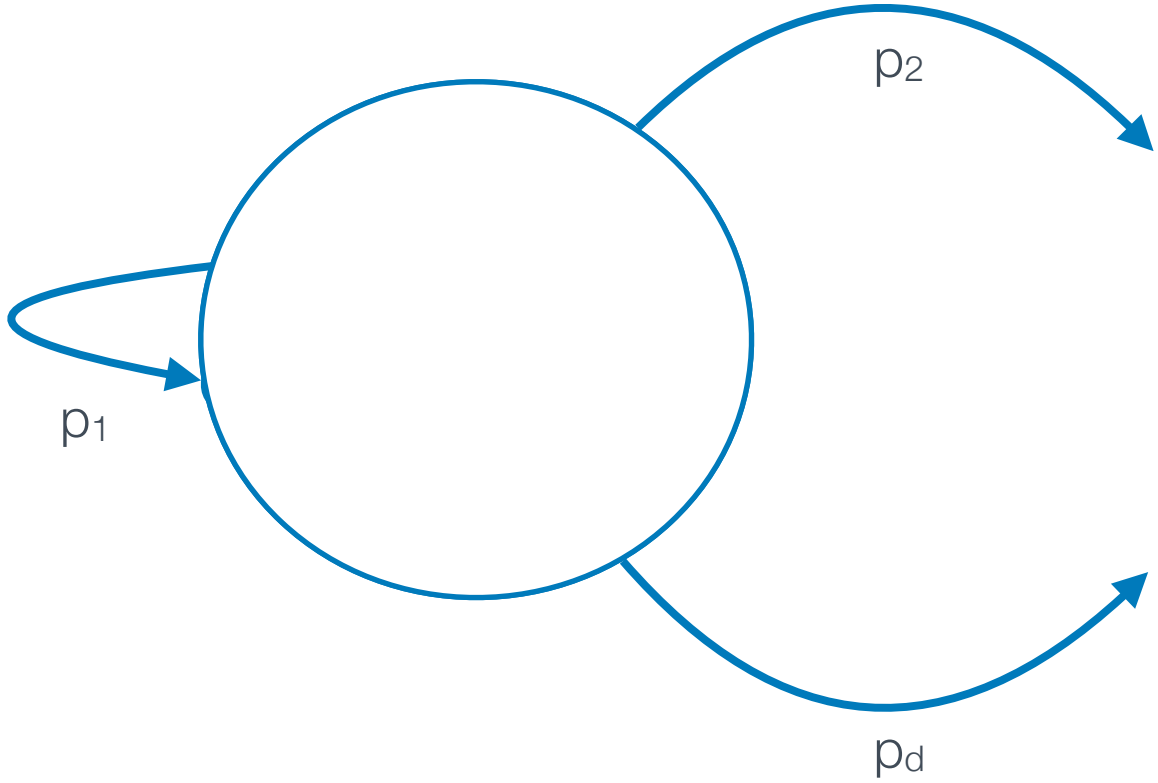


★Directed graph with transition probabilities

★ We observe the sequence of visited vertices

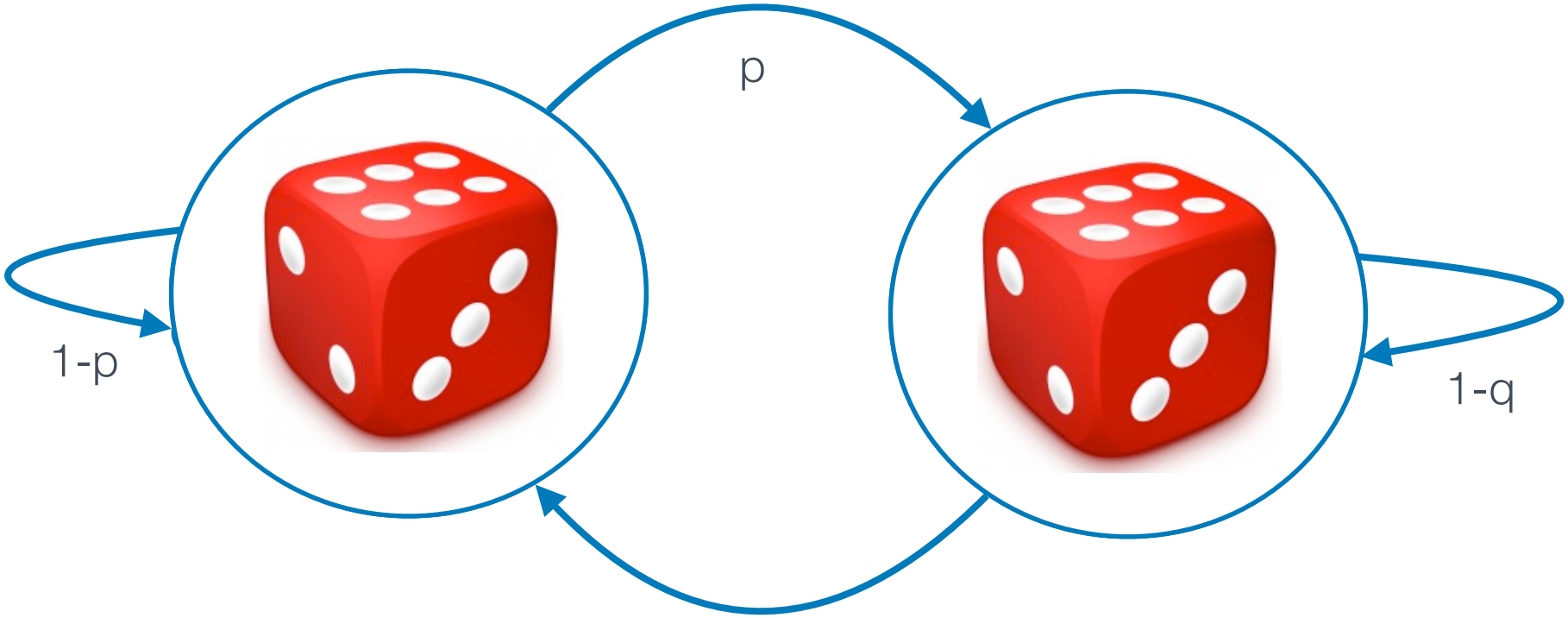
MARKOV CHAINS (DISCRETE)

Probabilities on outgoing edges sum to one



$$\sum_{i \in [d]} p_i = 1$$

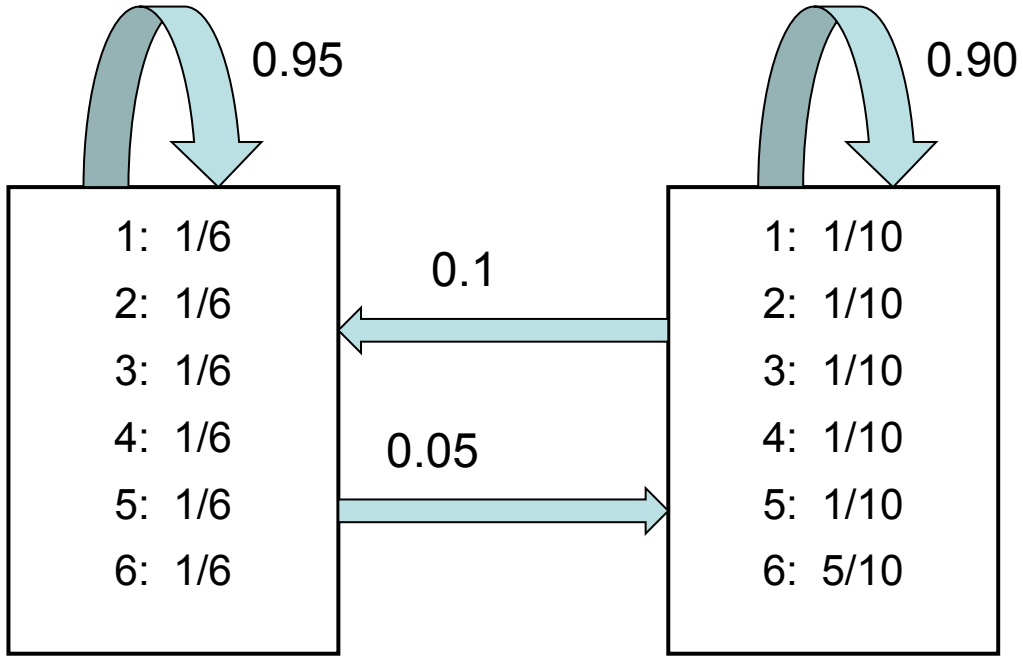
THE OCCASIONALLY DISHONEST CASINO



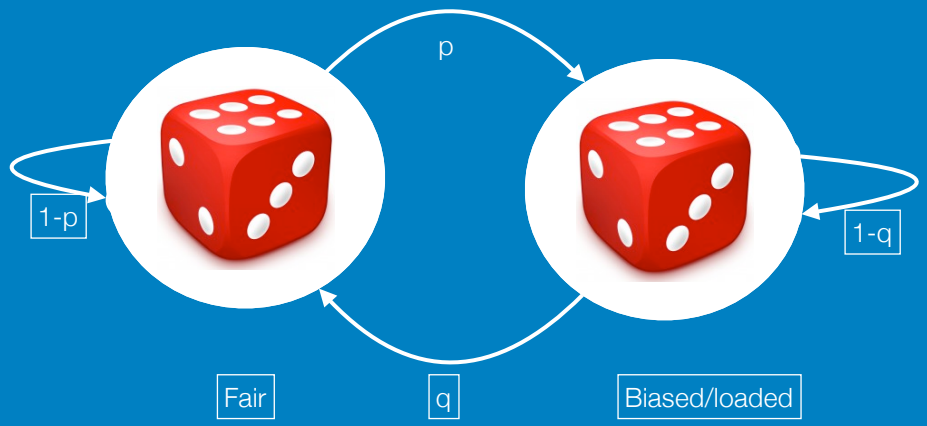
★ We observe the sequence of dice outcomes of visited vertices

```
Rolls:    6641532161621152346532143566342616552342323151424641566632 46
Die:      LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFLLLLLLL
```

EMISSION DISTRIBUTIONS



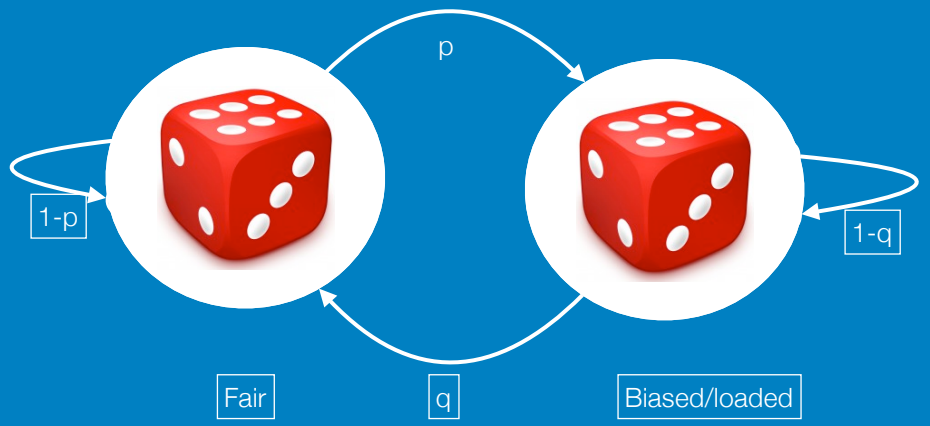
```
Rolls:    664153216162115234653214356634261655 2
Die:      LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFF
```



WHAT AN HMM DOES

- ★ Starts in the state z_1
- ★ When in state z_t
  - outputs $p(x_t|z_t)$
  - moves to $p(z_{t+1}|z_t)$
- ★ Stops after a fixed number of steps or when reaching a stop step

```
Rolls:    664153216162115234653214356634261655 2
Die:      LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFF
```



WHAT AN HMM DOES

- ★ Starts in the state z_1
- ★ When in state z_t
  - outputs $p(x_t|z_t)$ B_{x_t,t_t}
  - moves to $p(z_{t+1}|z_t)$ A_{z_{t+1},z_t}
- ★ Stops after a fixed number of steps or when reaching a stop step

The parameters

THE JOINT DISTRIBUTION

$$
\begin{aligned}
p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T+1}) & \\
&= p(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})p(\boldsymbol{z}_{1:T+1}) \\
&= p(\boldsymbol{z}_1)\left(\prod_{t=1}^{T} p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t)\right)\left(\prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{z}_t)\right)
\end{aligned}
$$

Categorial or Gaussian

- ★ Starts in the state z_1
- ★ When in state z_t
  - emits $p(x_t|z_t)$
  - transits to $p(z_{t+1}|z_t)$
- ★ Stops after a fixed number of steps or when reaching a stop step

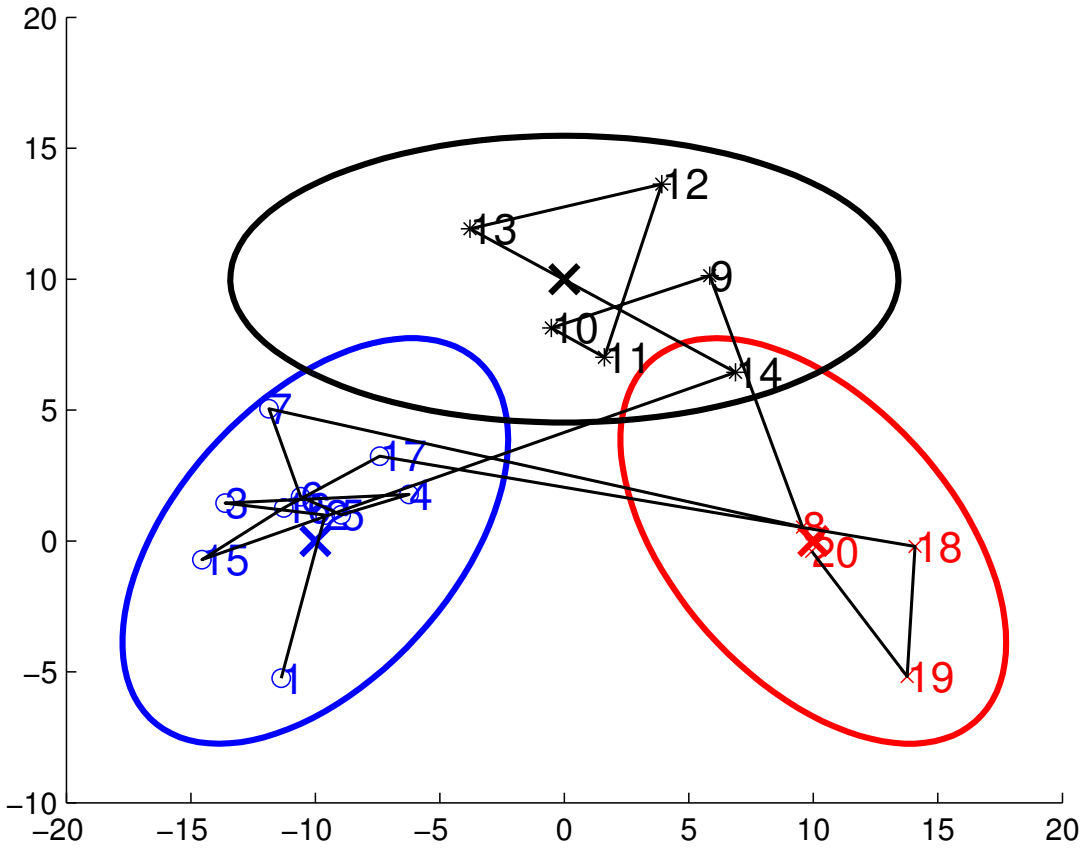
THE JOINT DISTRIBUTION

$$
\begin{aligned}
p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T+1}) & \\
&= p(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})p(\boldsymbol{z}_{1:T+1}) \\
&= p(\boldsymbol{z}_1)\left(\prod_{t=1}^{T} p(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t)\right)\left(\prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{z}_t)\right)
\end{aligned}
$$

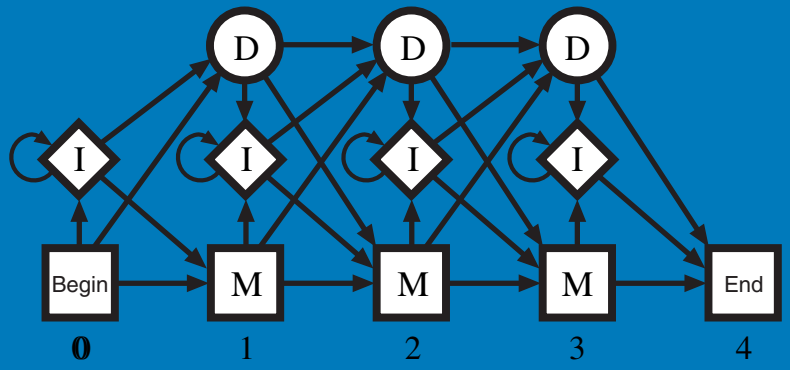
Categorial or Gaussian

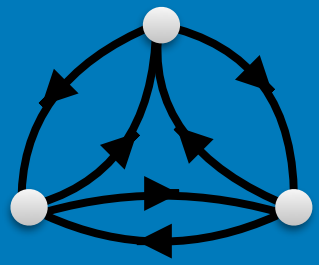
- ★ Starts in the state z_1
- ★ When in state z_t
  - emits $p(x_t|z_t)$
  - transits to $p(z_{t+1}|z_t)$
- ★ Stops after a fixed number of steps or when reaching a stop step

GAUSSIAN EMISSIONS AND HIDDEN STATES



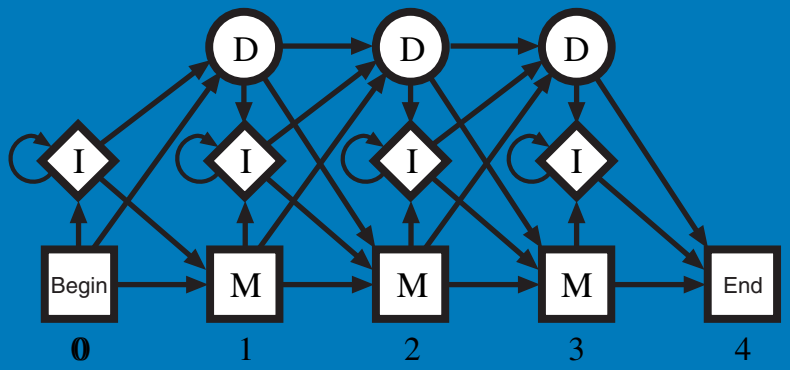
LAYERED OR NOT





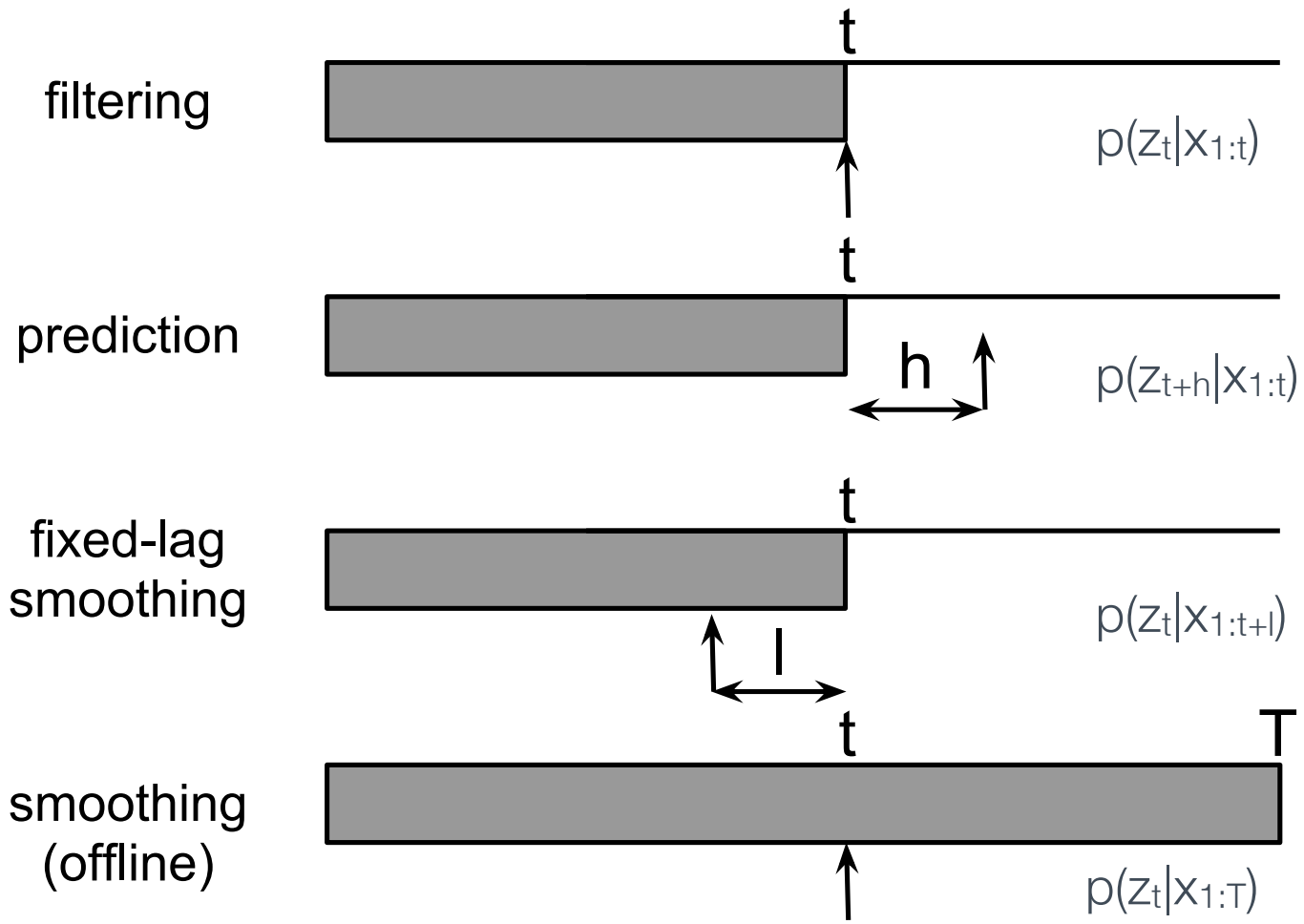
APPLICATIONS OF HMMS

```
      x  x  .  .  .  x
bat   A  G  -  -  -  C
rat   A  -  A  G  -  C
cat   A  G  -  A  A  -
gnat  -  -  A  A  A  C
goat  A  G  -  -  -  C
      1  2  .  .  .  3
```



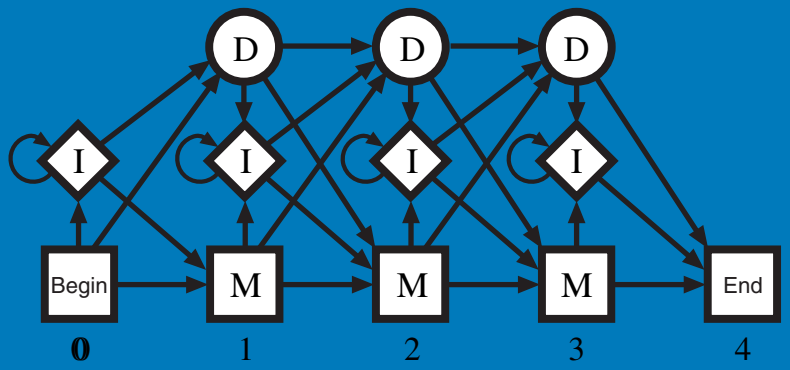
- Automatic speech recognition
- Part of speech tagging
- Gene finding
- Gene family characterization
- Secondary structure prediction

TERMINOLOGY X ABOVE Z BELOW



MORE INFERENCE TYPES

	x	x	.	.	.	x
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3



- Viterbi (MAP)

  argmax $p(z_{1:T}|x_{1:T})$
- Posterior samples:

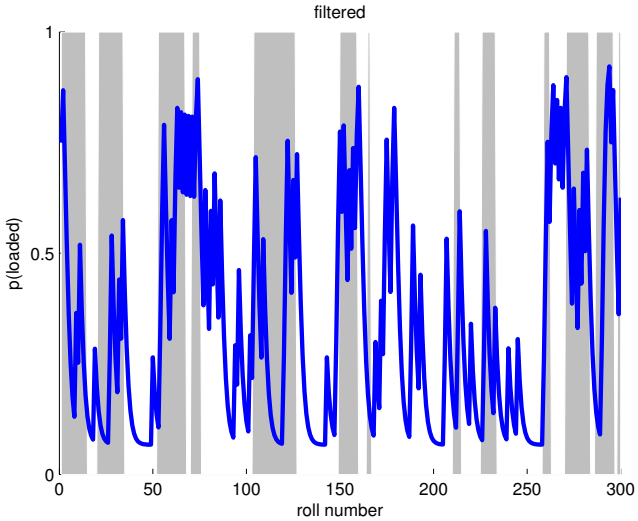
  $\sim p(z_{1:T}|x_{1:T})$
- Probability of data: $p(x_{1:T})$
- Parameters:

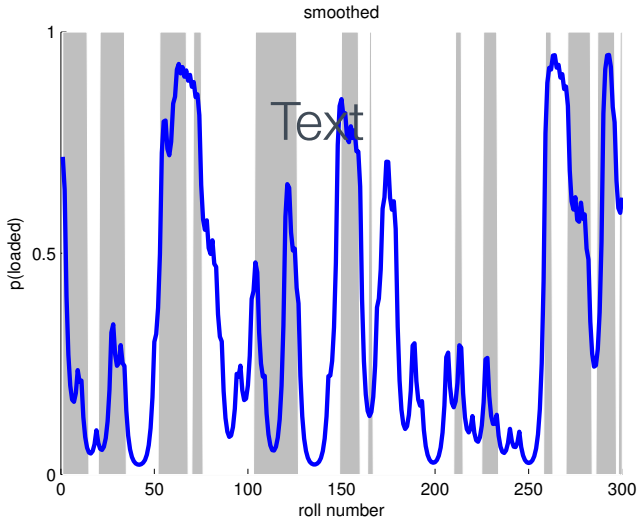
  given D & struct.
- Structure and param.:

  given D

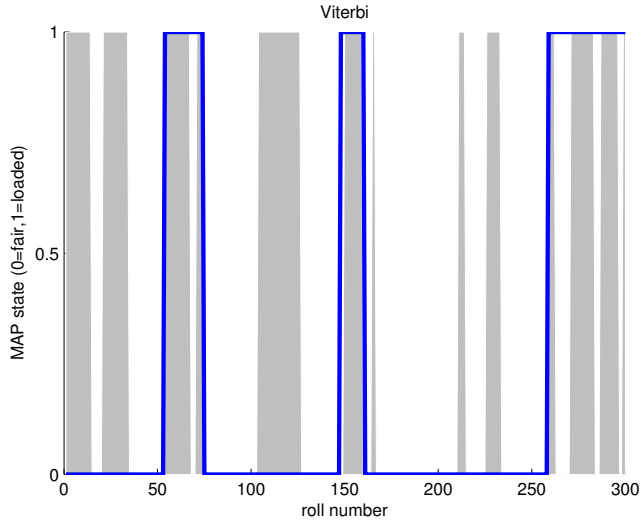
INFERENCE IN THE OCCASIONALLY DISHONEST CASINO

Grey regions are states corresponding to biased die

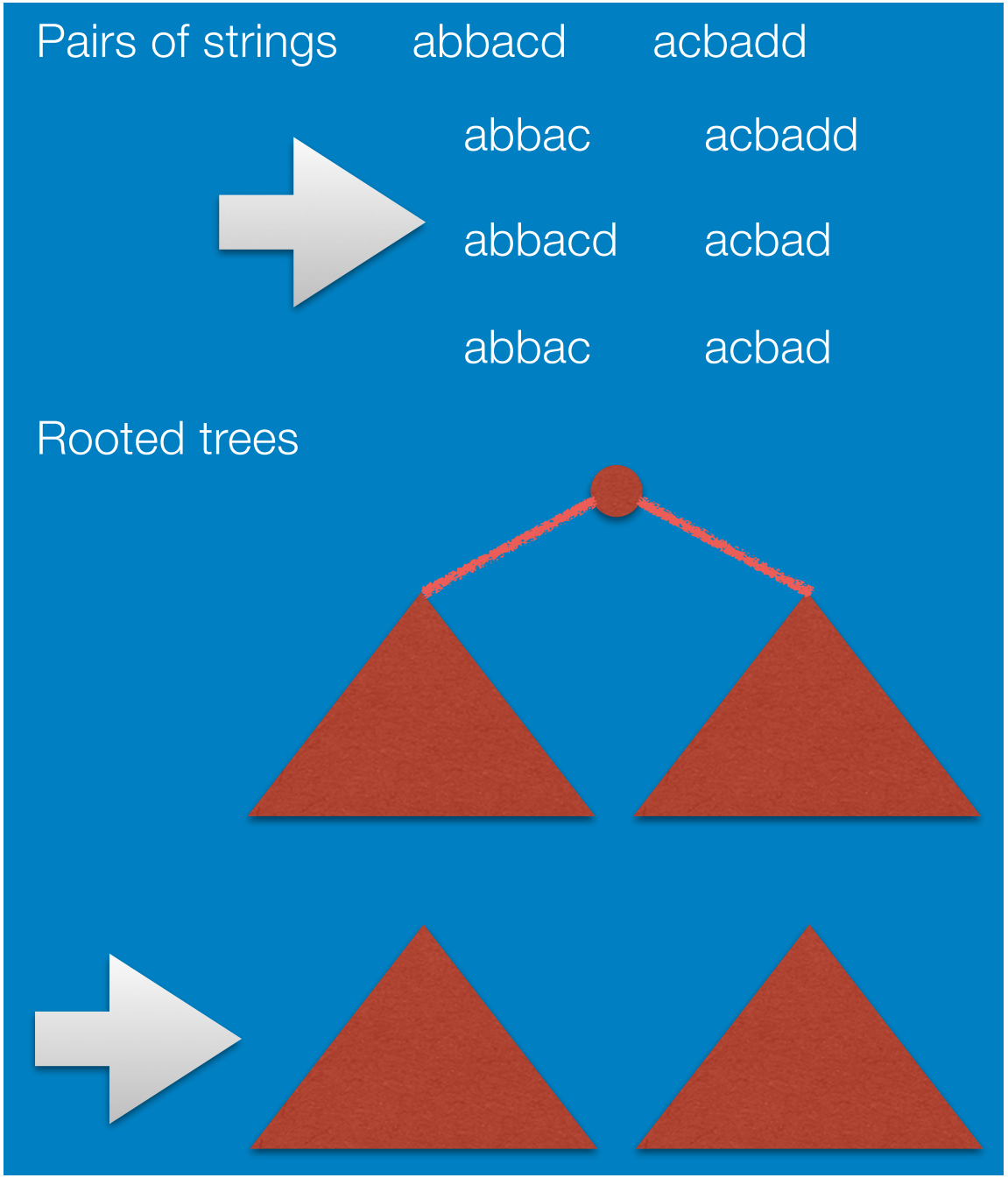




Text



- Filtering: $p(z_t|x_{1:t})$, online
- Smoothing, MAP state: $p(z_t|x_{1:T})$ offline
- Viterbi, MAP path argmax $p(z_{1:T}|x_{1:T})$



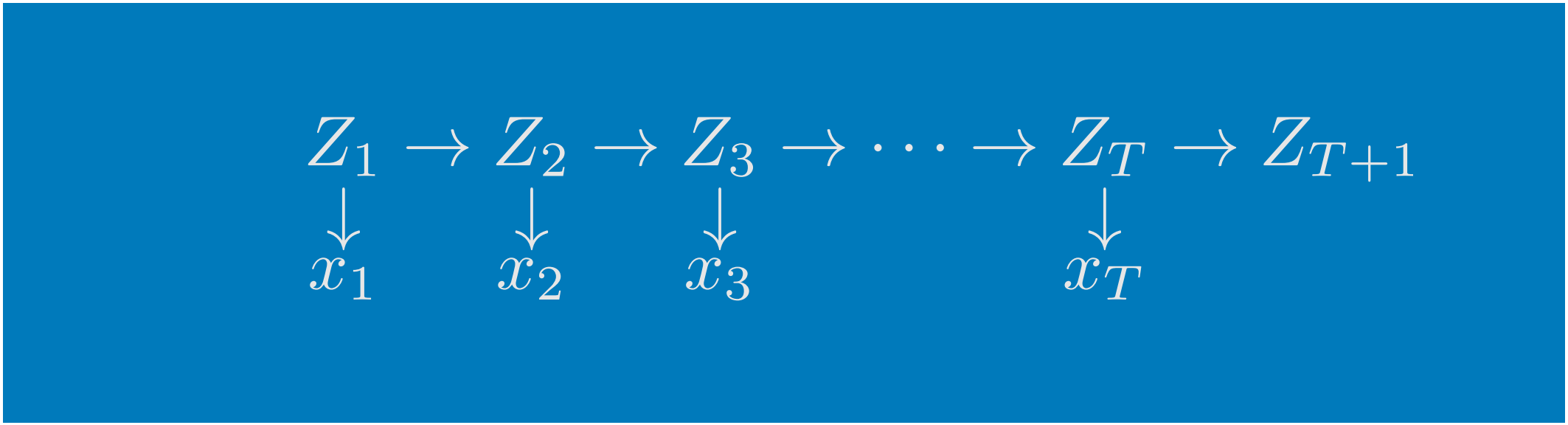
DP

- What is a subproblem?
- What is a subsolution?
- How do we decompose into smaller subproblems?
- How do we combine subsolutions into larger?
- How do we enumerate?
- How many and what time?

DP

Polynomial many	• What is a subproblem?
	• What is a subsolution?
Polynomial time	• How do we decompose into smaller subproblems?
Polynomial time	• How do we combine subsolutions into larger?
Polynomial time	• How do we enumerate?
Polynomial time overall	• How many and what time?

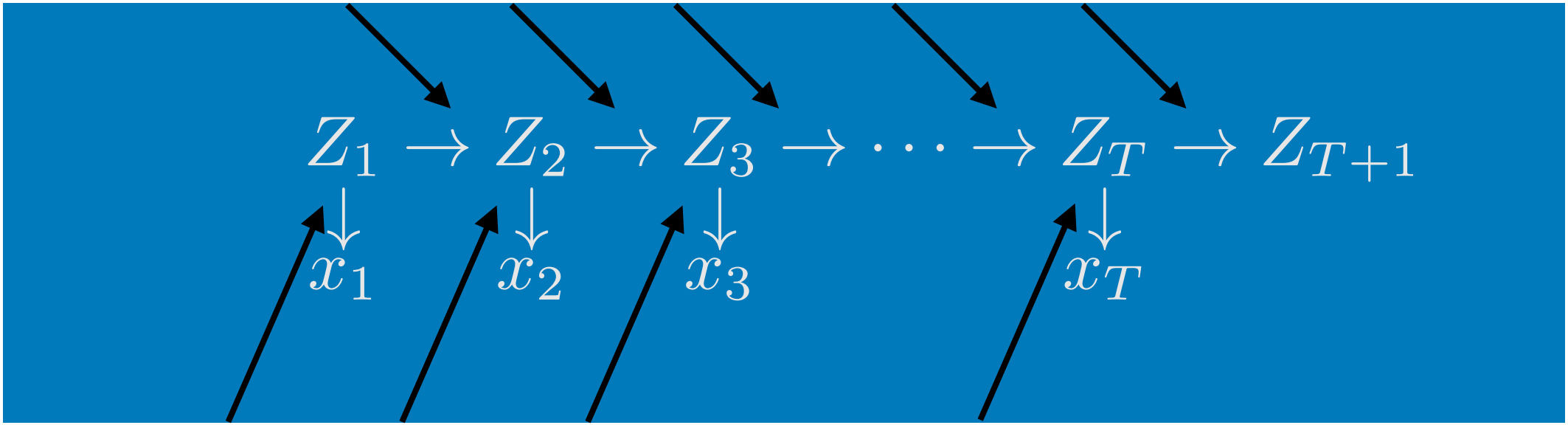
AN HMM CAN BE SEEN AS A DGM



- Z_i hidden
- X_i observable
- Hidden often not observable when training, never when applying

SPECIAL CASE: HIDDEN MARKOV MODEL (HMM)

Combinations of the transition distributions



Combinations of emission the emission distribution

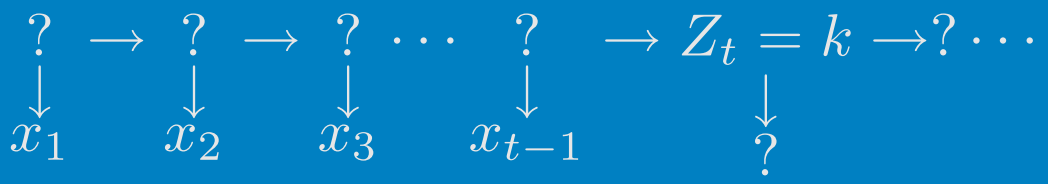
- Z_i hidden
- X_i observable
- Hidden often not observable when training, never when applying

JOINT &FORWARD VARIABLE

$$p(\boldsymbol{x}_{1:T}) = \sum_{\boldsymbol{z}_{1:T}} p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T})$$

$$f_t(k) := p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)$$

"Graphical model"



- Joint is easy to express
- The sum has exponentially many terms
- The forward variable, f_t, can be computed with DP

$Z_{t-1}=k'$ gives smaller

"Graphical model"

$$\begin{array}{ccccccccccc} ? & \to & ? & \to & ? & \cdots & ? & \to & Z_t = k & \to ? \cdots \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\ x_1 & & x_2 & & x_3 & & x_{t-1} & & ? & \end{array}$$

$$f_t(k) := p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)$$

Knowing also Z_{t-1} breaks it into smaller, i.e., the event

$$\begin{array}{ccccccccc} ? & \to & ? & \to & ? & \cdots Z_{t-1} = k' & \to & Z_t = k & \to ? \cdots \\ \downarrow & & \downarrow & & \downarrow & \downarrow & & \downarrow & \\ x_1 & & x_2 & & x_3 & x_{t-1} & & ? & \end{array}$$

"is the AND of the events"

$$\begin{array}{c} Z_{t-1} = k' \\ \downarrow \\ x_{t-1} \end{array} \qquad Z_{t-1} = k' \to Z_t = k$$

$$\begin{array}{ccccccc} ? & \to & ? & \to & ? & \cdots Z_{t-1} = k' & \to ? \\ \downarrow & & \downarrow & & \downarrow & \downarrow & \\ x_1 & & x_2 & & x_3 & ? & \end{array}$$

Applying sum rule

Notice, by the sum rule,

$$f_t(k) = p(x_{1:t-1}, Z_t = k) = \sum_{k' \in [K]} p(x_{1:t-1}, Z_{t-1} = k', Z_t = k)$$

The set of states

each term in the sum is a probability of an event

$$\begin{array}{ccccccccccc} ? & \to & ? & \to & ? & \cdots & Z_{t-1} = k' & \to & Z_t = k & \to ? \cdots \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\ x_1 & & x_2 & & x_3 & & x_{t-1} & & ? & \end{array}$$

which, as noted, can be broken into smaller

Forward recursion

$$f_t(k) = \sum_l \underbrace{f_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1}=l)}_{\text{emission}} \underbrace{p(\boldsymbol{Z}_t=k|\boldsymbol{Z}_{t-1}=l)}_{\text{transition}}$$

Forward recursion

$$f_t(k) = \sum_l \underbrace{f_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1}=l)}_{\text{emission } B_{x_{t-1},l}} \underbrace{p(\boldsymbol{Z}_t=k|\boldsymbol{Z}_{t-1}=l)}_{\text{transition } A_{lk}}$$

Forward recursion

$$f_t(k) = \sum_l \underbrace{f_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1}=l)}_{\text{emission } B_{x_{t-1},l}} \underbrace{p(\boldsymbol{Z}_t=k|\boldsymbol{Z}_{t-1}=l)}_{\text{transition } A_{lk}}$$

Given $x_1, \ldots, x_T$

Forward Algorithm

For the start state k*

$$s(0, k^*) := 1$$

For all other states k

$$s(0, k) := 0$$

For t=1 to T

For k=1 to K

$$s(t, k) := \sum_{l \in [K]} s(t-1, l) B_{x_{t-1,l}} A_{lk}$$

Time

Forward Algorithm

For the start state k^*

$$s(0, k^*) := 1$$

For all other states k

$$s(0, k) := 0$$

} constant time

For t=1 to T

O(K^2) {

For k=1 to K

$$s(t, k) := \sum_{l \in [K]} s(t-1, l) B_{x_{t-1,l}} A_{lk}$$

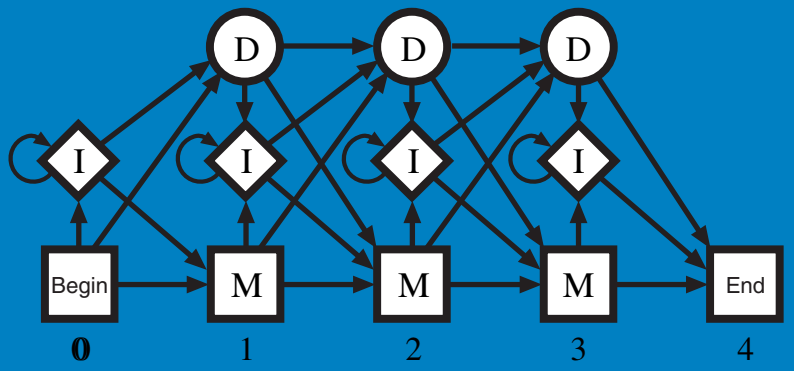
} O(TK^2)

O(K)

So in total time O(TK^2)

If layered

If layered, total time O(TK)



Forward Algorithm

For the start state k^*

$$s(0, k^*) := 1$$

For all other states k

$$s(0, k) := 0$$

} constant time

For t=1 to T

O(K) {

For k=1 to K

$$s(t, k) := \sum_{l \in [K]} s(t-1, l) B_{x_{t-1}, l} A_{lk}$$

} O(TK)

Replace by sum over constant number of states in previous layer

OBSERVATION PROBABILITY

$$p(\boldsymbol{x}_{1:T}) = \sum_k p(\boldsymbol{x}_{1:T}, z_{T+1} = k)$$

$$f_t(k) = p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)$$

In general, (e.g. t=T)

$$p(\boldsymbol{x}_{1:t}) = \sum_k f_{t+1}(k) = \sum_k p(\boldsymbol{x}_{1:t}, \boldsymbol{Z}_{t+1} = k)$$

since

$$\begin{array}{ccccccccc} Z_1 & \rightarrow & Z_2 & \rightarrow & Z_3 & \rightarrow \cdots \rightarrow & Z_T & \rightarrow & Z_{T+1} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\ x_1 & & x_2 & & x_3 & & x_T & & \end{array}$$

- The final probability is easily obtained

FILTERING

$$p(\boldsymbol{Z}_t = k | \boldsymbol{x}_{1:t})$$

- Filtering: $p(z_t|x_{1:t})$, online

FILTERING

$$p(\boldsymbol{Z}_t = k|\boldsymbol{x}_{1:t}) = \frac{p(\boldsymbol{x}_{1:t}, \boldsymbol{Z}_t = k)}{p(\boldsymbol{x}_{1:t})}$$
$$= \frac{p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}{p(\boldsymbol{x}_{1:t})}$$
$$= \frac{f_t(k)p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}{p(\boldsymbol{x}_{1:t})}$$

emission

data probability

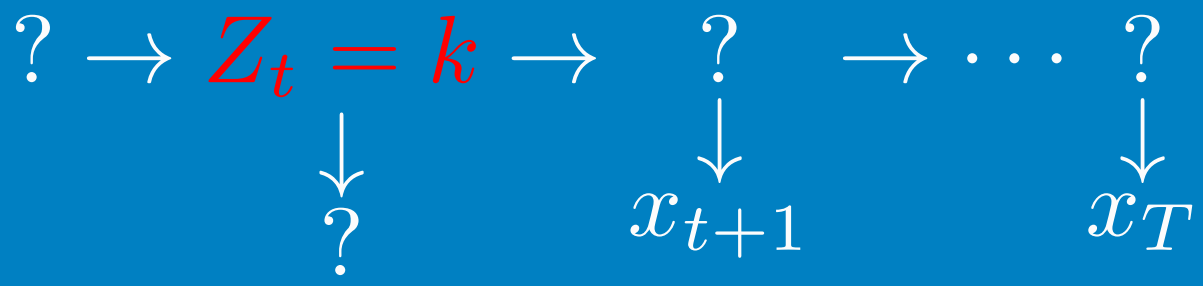
- Filtering: $p(z_t|x_{1:t})$, online

Backward variable

Defined by

$$b_t(k) := p(\boldsymbol{x}_{t+1:T} | \boldsymbol{Z}_t = k)$$

"Graphical model"



Sum rule gives Z_{t+1}

Defined by

$$b_t(k) := p(x_{t+1:T}|Z_t = k) = \sum_{l \in [K]} p(x_{t+1:T}, Z_{t+1} = l|Z_t = k)$$

Each term in the sum is a probability of an event

$$? \rightarrow Z_t = k \rightarrow Z_{t+1} = l \rightarrow \cdots ?$$
$$\downarrow \qquad \downarrow \qquad \downarrow$$
$$? \qquad x_{t+1} \qquad x_T$$

"which is an AND of"

$$Z_t = k \rightarrow Z_{t+1} = l \qquad Z_{t+1} = l \qquad Z_{t+1} = l \rightarrow ? \rightarrow \cdots ?$$
$$\downarrow \qquad\qquad \downarrow \qquad\qquad\qquad\qquad \downarrow \qquad\qquad \downarrow$$
$$x_{t+1} \qquad\qquad x_{t+1} \qquad\qquad\qquad\qquad x_{t+2} \qquad\qquad x_T$$