Royal Institute of Technology

MACHINE LEARNING 2 - UGM

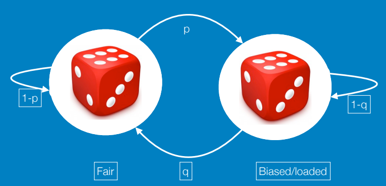
Lecture 5

At the end maximize $P(S^1, \dots, S^N | \Theta)$

1.5

WHAT AN HMM DOES

```
Rolls:    6641532161621152346532143566342616552
Die:      LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLLLFFF
```

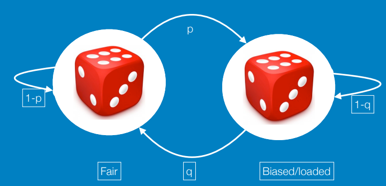


- Starts in the state z_1
- When in state z_t
  - outputs $p(x_t|z_t)$
  - moves to $p(z_{t+1}|z_t)$
- Stops after a fixed number of steps or when reaching a stop step

---

WHAT AN HMM DOES

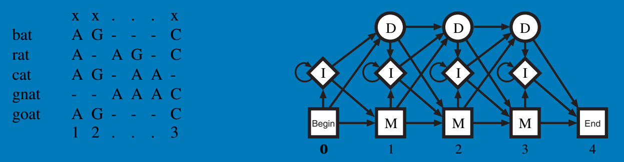
```
Rolls:    6641532161621152346532143566342616552
Die:      LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLLLFFF
```



- 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

The parameters

INFERENCE TYPES



- Filtering: $p(z_t|x_{1:t})$, online
- Smoothing (MAP): $p(z_t|x_{1:T})$ offline
- Fixed lag smoothing: $p(z_t|x_{1:t+l})$
- Prediction: $p(z_{t+h}|x_{1:t})$
- 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

SPECIAL CASE: HIDDEN MARKOV MODEL (HMM)

$$\begin{array}{ccccccc} z_1 & \longrightarrow & z_2 & \longrightarrow & z_3 & \cdots & z_V \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ x_1 & & x_2 & & x_3 & & x_V \end{array}$$

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

SPECIAL CASE: HIDDEN MARKOV MODEL (HMM)

$$\begin{array}{ccccccc} z_1 & \longrightarrow & z_2 & \longrightarrow & z_3 & \cdots & z_V \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ x_1 & & x_2 & & x_3 & & x_V \end{array}$$

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

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

SMOOTHING

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

- Smoothing, MAP state: $p(z_t|x_{1:T})$, offline

SMOOTHING

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

- Smoothing, MAP state: $p(z_t|x_{1:T})$, offline

SMOOTHING

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

- Smoothing, MAP state: $p(z_t|x_{1:T})$, offline

SMOOTHING

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

- Smoothing, MAP state: $p(z_t|x_{1:T})$, offline

SMOOTHING

$$p(\boldsymbol{Z}_t = k|\boldsymbol{x}_{1:T}) = \frac{p(\boldsymbol{x}_{1:T}, \boldsymbol{Z}_t = k)}{p(\boldsymbol{x}_{1:T})}$$
$$\propto p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)p(\boldsymbol{x}_{t:T}|\boldsymbol{Z}_t = k)$$
$$= f_t(k)\, \underbrace{p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}_{\text{emission}}\, b_t(k)$$

- Smoothing, MAP state: $p(z_t|x_{1:T})$, offline

MARGINAL OVER PAIRS OF STATES, AND PARIS OF STATE AND SYMBOL

$$p(\boldsymbol{Z}_{t-1} = k, \boldsymbol{Z}_t = l | \boldsymbol{X} = \boldsymbol{x}_n, \boldsymbol{\theta})$$

$$p(\boldsymbol{X}_t = s, \boldsymbol{Z}_t = k) | \boldsymbol{X} = \boldsymbol{x}_n, \boldsymbol{\theta})$$

- Can be computed from forward and backward similarly

FORWARD FILTERING, BACKWARDS SAMPLING

$$\boldsymbol{z}^s_{1:T} \sim p(\boldsymbol{Z}_{1:T} | \boldsymbol{x}_{1:T})$$

- Sample from posterior
- Sample in order $z_T, \ldots, z_1$
- Start somewhat differently

FORWARD FILTERING, BACKWARDS SAMPLING

$$\boldsymbol{z}^s_{1:T} \sim p(\boldsymbol{Z}_{1:T} | \boldsymbol{x}_{1:T})$$

$$\begin{aligned} p(\boldsymbol{Z}_t = k | \boldsymbol{Z}_{t+1} = l, \boldsymbol{x}_{1:t}) &= \frac{p(\boldsymbol{Z}_t = k, \boldsymbol{Z}_{t+1} = l, \boldsymbol{x}_{1:t})}{p(\boldsymbol{x}_{1:t}, \boldsymbol{Z}_{t+1} = l)} \\ &= \frac{p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k) p(\boldsymbol{Z}_{t+1} = l, \boldsymbol{x}_t | \boldsymbol{Z}_t = k)}{f_{t+1}(l)} \\ &= \frac{f_t(k) p(\boldsymbol{Z}_{t+1} = l | \boldsymbol{Z}_t = k) p(\boldsymbol{x}_t | \boldsymbol{Z}_t = k)}{f_{t+1}(l)} \end{aligned}$$

- Sample from posterior
- Sample in order $z_T, \ldots, z_1$
- Start somewhat differently

LEARNING TRANSITION AND EMISSION PARAMETERS - FULLY OBSERVED DATA

- Parameters
  - transition $A_{kl} = p(Z_t = l | Z_{t-1} = k)$
  - z_0 always q^* and A_{q^*k} is start probability
  - emission
    $$B_{ks} = p(X_t = s | Z_t = k)$$
- Likelihood

$$L(\boldsymbol{\theta}; \mathcal{D}) = \prod_{n=1}^{N} \prod_{t=1}^{T} \left[\prod_{k,s} B_{ks}^{I(x_{n,t}=s, z_{n,t}=k)} \prod_{k,l} A_{kl}^{I(z_{n,t-1}=k, z_{n,t}=l)} \right]$$

LEARNING TRANSITION AND EMISSION PARAMETERS - FULLY OBSERVED DATA

- Parameters
  - transition $A_{kl} = p(Z_t = l | Z_{t-1} = k)$
  - z_0 always q* and A_{q^*k} is start probability
  - emission

$$B_{ks} = p(X_t = s | Z_t = k)$$

- Loglikelihood

$$l(\boldsymbol{\theta}; \mathcal{D}) = \sum_{n=1}^{N} \sum_{t=1}^{T} [\sum_{k,s} I(x_{n,t} = s, z_{n,t} = k) \log B_{ks} + \sum_{k,l} I(z_{n,t-1} = k, z_{n,t} = l) \log A_{kl}]$$

MAXIMIZING LOG-LIKELIHOOD - COMPLETE DATA

$$l(\boldsymbol{\theta}; \mathcal{D}) = \sum_{n=1}^{N} \sum_{t=1}^{T} \left[\sum_{k,s} I(x_{n,t} = s, z_{n,t} = k) \log B_{ks} + \sum_{k,l} I(z_{n,t-1} = k, z_{n,t} = l) \log A_{kl} \right]$$

$$= \sum_{k,s} \left[\underbrace{\sum_{n=1}^{N} \sum_{t=1}^{T} I(x_{n,t} = s, z_{n,t} = k)}_{M_{k,s}} \right] \log B_{ks} + \sum_{k,l} \left[\underbrace{\sum_{n=1}^{N} \sum_{t=1}^{T} I(z_{n,t-1} = k, z_{n,t} = l)}_{N_{k,l}} \right] \log A_{kl}$$

- Maximized by

$B_{ks} = M_{k,s} / \sum_s M_{k,s} = M_{k,s}/N_k$ and $A_{kl} = N_{k,l} / \sum_l N_{k,l} = N_{k,l}/N_k$

Expected complete: of all data

$$\log p(\mathcal{D}|\boldsymbol{\theta}')$$

$$= \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\theta}')$$

$$= \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n | \boldsymbol{x}_n, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\theta}')}{p(\boldsymbol{z}_n | \boldsymbol{x}_n, \boldsymbol{\theta})}$$

$$\geq \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n | \boldsymbol{x}_n, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\theta}')}{p(\boldsymbol{z}_n | \boldsymbol{x}_n, \boldsymbol{\theta})}$$

$$= \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n | \boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\theta}') - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n | \boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n | \boldsymbol{x}_n$$

$$= \underbrace{\sum_n Q_n(\boldsymbol{\theta}'; \boldsymbol{\theta})}_{Q(\boldsymbol{\theta}'; \boldsymbol{\theta})} - \underbrace{\sum_n R_n(\boldsymbol{\theta}; \boldsymbol{\theta})}_{R_n(\boldsymbol{\theta}; \boldsymbol{\theta})}$$

HIDDEN VARIABLES - ONE EM-STEP MAXIMIZING Q

$$\sum_{n=1}^{N} E_{p(\boldsymbol{Z}|\boldsymbol{x}_n, \boldsymbol{\theta})}[l(\theta'; \boldsymbol{Z}, \boldsymbol{x}_n)]$$

HIDDEN VARIABLES - ONE EM-STEP MAXIMIZING Q

$$\sum_{n=1}^N E_{p(\boldsymbol{Z}|\boldsymbol{x}_n,\boldsymbol{\theta})}[l(\theta';\boldsymbol{Z},\boldsymbol{x}_n)]$$
$$= \sum_{n=1}^N E_{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}[l(\theta';\boldsymbol{Z},\boldsymbol{X})]$$

HIDDEN VARIABLES - ONE EM-STEP MAXIMIZING Q

$$\sum_{n=1}^N E_{p(\boldsymbol{Z}|\boldsymbol{x}_n,\boldsymbol{\theta})}[l(\theta';\boldsymbol{Z},\boldsymbol{x}_n)]$$
$$= \sum_{n=1}^N E_{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}[l(\theta';\boldsymbol{Z},\boldsymbol{X})]$$
$$= \sum_{n=1}^N E_{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}\left[\sum_{t=1}^T\left[\sum_{k,s} I(\boldsymbol{X}_t=s,\boldsymbol{Z}_t=k)\log B'_{ks}+\sum_{k,l} I(\boldsymbol{Z}_{t-1}=k,\boldsymbol{Z}_t=l)\log A'_{kl}\right]\right]$$

HIDDEN VARIABLES - ONE EM-STEP MAXIMIZING Q

$$\sum_{n=1}^N E_{p(\boldsymbol{Z}|\boldsymbol{x}_n,\boldsymbol{\theta})}[l(\theta';\boldsymbol{Z},\boldsymbol{x}_n)]$$
$$= \sum_{n=1}^N E_{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}[l(\theta';\boldsymbol{Z},\boldsymbol{X})]$$
$$= \sum_{n=1}^N E_{p(\boldsymbol{Z},\boldsymbol{X}|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}\left[\sum_{t=1}^T\left[\sum_{k,s} I(\boldsymbol{X}_t=s,\boldsymbol{Z}_t=k)\log B'_{ks}+\sum_{k,l} I(\boldsymbol{Z}_{t-1}=k,\boldsymbol{Z}_t=l)\log A'_{kl}\right]\right]$$
$$= \sum_{k,s}\left[\underbrace{\sum_{n=1}^N\sum_{t=1}^T p(\boldsymbol{X}_t=s,\boldsymbol{Z}_t=k)|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}_{\overline{M}_{k,s}}\right]\log B'_{ks}$$
$$+\sum_{k,l}\left[\underbrace{\sum_{n=1}^N\sum_{t=1}^T p(\boldsymbol{Z}_{t-1}=k,\boldsymbol{Z}_t=l|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}_{\overline{N}_{k,l}}\right]\log A'_{kl}$$

LEARNING TRANSITION AND EMISSION PARAMETERS - HIDDEN VARIABLE

$$\sum_{n=1}^N E_{p(\boldsymbol{Z}|\boldsymbol{x}_n,\boldsymbol{\theta})}[l(\theta';\boldsymbol{Z},\boldsymbol{x}_n)]$$
$$= \sum_{k,s}\left[\underbrace{\sum_{n=1}^N\sum_{t=1}^T p(\boldsymbol{X}_t=s,\boldsymbol{Z}_t=k)|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}_{\overline{M}_{k,s}}\right]\log B'_{ks}$$

jl

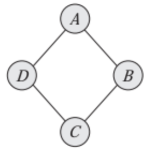
$$+\sum_{k,l}\left[\underbrace{\sum_{n=1}^N\sum_{t=1}^T p(\boldsymbol{Z}_{t-1}=k,\boldsymbol{Z}_t=l|\boldsymbol{X}=\boldsymbol{x}_n,\boldsymbol{\theta})}_{\overline{N}_{k,l}}\right]\log A'_{kl}$$

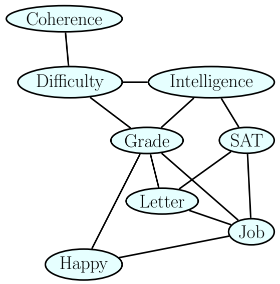
- Maximized by

  $B'_{ks}=M'_{k,s}/\sum_s M'_{k,s}=M'_{k,s}/N'_k$ and $A'_{kl}=N'_{k,l}/\sum_l N'_{k,l}=N'_{k,l}/N'_k$

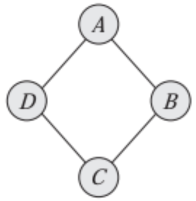
UGM

- 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





PROBABILISTIC INTERPRETATION



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

(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

(b) $\phi_2(B,C)$

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

(c) $\phi_3(C,D)$

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

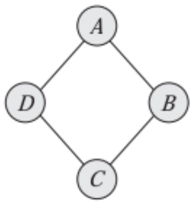
(d) $\phi_4(D,A)$

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

Factors – misconception example

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

A FACTOR PRODUCT



(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

(b) $\phi_2(B,C)$

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

(c) $\phi_3(C,D)$

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

(d) $\phi_4(D,A)$

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

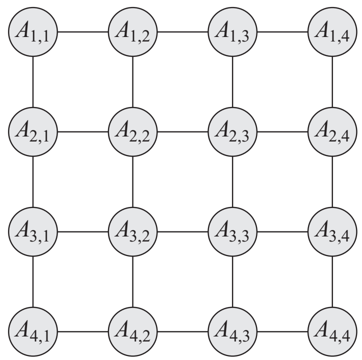
Misconception

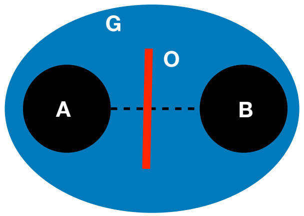
- $$\begin{aligned}\phi_1(A=1,B=1)\phi_2(B=1,C=0)\phi_3(C=0,D=1)\phi_4(D=1,A=1) \\ = 10\cdot 1\cdot 100\cdot 100 \\ = 100000\end{aligned}$$

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

A PAIRWISE UGM

- Can be useful for images
- Only option in grids





SEPARATION AND CI OF UGM

★ A is separated from B given O in G if there is no path between A and B in G\O

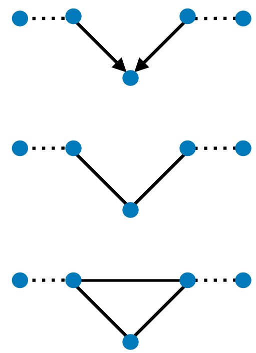
★ In a graph G,

⇕

A is separated from B given O

DEF I-MAP

- G is an I-map for p if all independence relation in G hold for p, i.e., I(G)⊆I(p)
- Moralize add edge between any two parents
- We can moralize a DGM and get a UGM having no more independence relations
- Each family has a cliques in the moralized UGM



EQUIVALENCE I-MAP AND FACTORIZATION

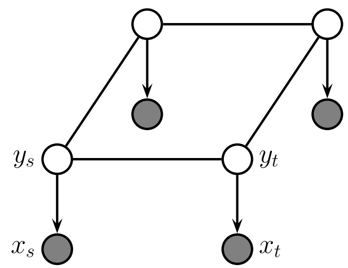
- For positive distributions p (i.e., ∀y, p(y)>0),

  I(G) ⊆ I(p) ⇔ p can be expressed as a normalised product over factors of G (as below)

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$$

THE POTTS MODEL

p(x | y) ex Gaussian



Values

$$y_t \in \{1, 2, 3\}$$

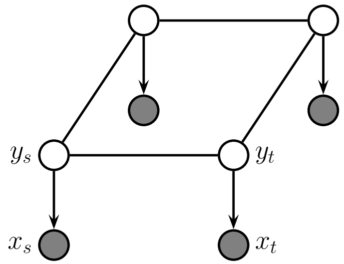
Factors of form

$$\varphi(y_s, y_t) = \begin{pmatrix} e^{w_{st}} & e^0 & e^0 \\ e^0 & e^{w_{st}} & e^0 \\ e^0 & e^0 & e^{w_{st}} \end{pmatrix}$$

Tied weights

$$w_{st} = J$$

THE POTTS MODEL



Values

$$y_t \in \{1,2,3\}$$

Factors of form

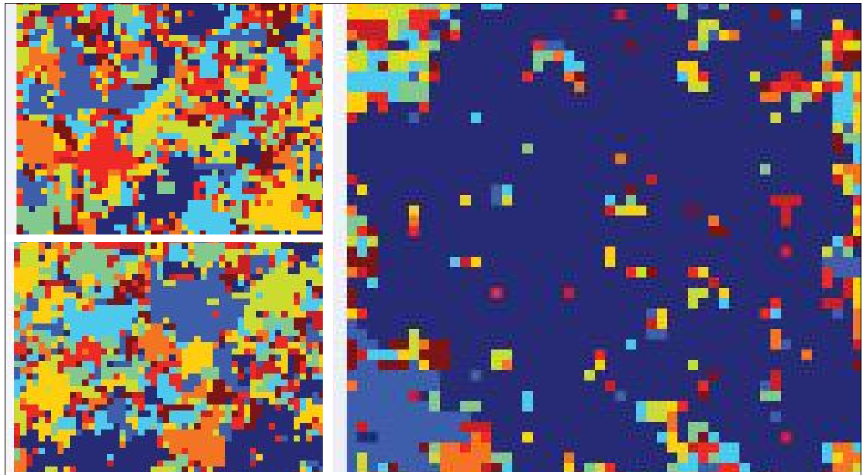
$$\psi(y_s, y_t) = \begin{pmatrix} e^{w_{st}} & e^0 & e^0 \\ e^0 & e^{w_{st}} & e^0 \\ e^0 & e^0 & e^{w_{st}} \end{pmatrix}$$

Tied weights

$$w_{st} = J$$

Likelihood

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{y}|J)\prod_t p(x_t|y_t, \boldsymbol{\theta}) = \left[\frac{1}{Z(J)}\prod_{s\sim t}\psi(y_s, y_t; J)\right]\prod_t p(x_t|y_t, \boldsymbol{\theta})$$



RESULTS J=1.42, 1.44, 1.46

THE END

RELATIONS BETWEEN LOG-LIKELIHOODS AND Q-TERMS

Theorem: for $\boldsymbol{\theta}' = \text{argmax}_{\boldsymbol{\theta}''} Q(\boldsymbol{\theta}'', \boldsymbol{\theta})$

$$\log p(\mathcal{D}|\boldsymbol{\theta}') \geq Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) = \log p(\mathcal{D}|\boldsymbol{\theta})$$

So by maximizing Q-term (through ESS) we monotonically increase the likelihood.

The Q-term may not increase in every step!