EM Part 2

The EM Algorithm

Input: initial model parameters $\mathbf{w}^{(0)}$, training data $\langle \tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{\tilde{N}} \rangle$

Output: learned parameters $\mathbf{w}$

$t \leftarrow 0$

repeat

E step:

for $i = 1$ to $\tilde{N}$ **do**

$$\forall \boldsymbol{y} \in \mathcal{Y}_{\tilde{\boldsymbol{x}}_i}, q_i^{(t)}(\boldsymbol{y}) \leftarrow p_{\mathbf{w}^{(t)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i) = \frac{p_{\mathbf{w}^{(t)}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}_{\tilde{\boldsymbol{x}}_i}} p_{\mathbf{w}^{(t)}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}')}$$

end for

M step: $\mathbf{w}^{(t+1)} \leftarrow \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}_{\tilde{\boldsymbol{x}}_i}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y})$

$t \leftarrow t + 1$

until $\mathbf{w}^{(t)} \approx \mathbf{w}^{(t-1)}$

$\mathbf{w} \leftarrow \mathbf{w}^{(t)}$

$$= \mathbb{E}_{q_i^{(t)}(\boldsymbol{y})} \left[\log p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}) \right]$$

What Do We Need?

- Start with the M step

$$\mathbf{w}^{(t+1)} \leftarrow \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{\tilde{N}} \sum_{y \in \mathcal{Y}_{\tilde{x}_i}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y})$$

- What values do we need? These are called **sufficient statistics**. They depend on your model.
- In an HMM, we need to know the number of times
  - You are in state s
  - You transition from state s to state t
  - You emit symbol x from state s

Helpful Recipe

- Think about the complete-data likelihood
- What are the various quantities you need to compute the MLE?
- Replace these quantities with their **expected values** under the q distribution
- Run MLE as normal
- Repeat

Why Does EM Work?

$$\Phi_{ML}(\mathbf{w}) = \sum_{i=1}^{\tilde{N}} \log \sum_{\boldsymbol{y} \in \mathcal{Y}_{\tilde{\boldsymbol{x}}}} p_{\mathbf{w}}(\tilde{\boldsymbol{x}}, \boldsymbol{y})$$

Theorem. At every step of the above algorithm the M step will find an $\mathbf{w}^{t+1}$ such that

$$\Phi_{ML}(\mathbf{w}^{(t+1)}) \geq \Phi(\mathbf{w}^{(t)})$$

Proof. We define the following quantity.

$$Q^{(t)}(\mathbf{w}) = \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}_{\tilde{\boldsymbol{x}}_i}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}^{(t)}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y})$$

where $q_i^{(t)}(\boldsymbol{y}) = p_{\mathbf{w}^{(i)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$

Consider the difference between the likelihood objective $\Phi(\mathbf{w})$ and $Q^{(t)}(\mathbf{w})$

$$
\begin{aligned}
&\Phi(\mathbf{w}) - Q^{(t)}(\mathbf{w}) \\
&\quad = \sum_{i=1}^{\tilde{N}} \log \sum_{\boldsymbol{y}' \in \mathcal{Y}_{\tilde{\boldsymbol{x}}_i}} p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}') - \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}) \\
&\quad = \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log \sum_{\boldsymbol{y}' \in \mathcal{Y}_{\tilde{\boldsymbol{x}}_i}} p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}') - \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y} \\
&\quad = \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log \frac{\sum_{\boldsymbol{y}' \in \mathcal{Y}_{\tilde{\boldsymbol{x}}_i}} p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}')}{p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y})} \\
&\quad = -\sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log \frac{p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}_{\tilde{\boldsymbol{x}}_i}} p_{\mathbf{w}}(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}')} \\
&\quad = -\sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)
\end{aligned}
$$

$$\Phi(\mathbf{w}) - Q^{(t)}(\mathbf{w}) = -\sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

$$\Phi(\mathbf{w}) = Q^{(t)}(\mathbf{w}) - \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

$$\Phi(\mathbf{w}) - Q^{(t)}(\mathbf{w}) = -\sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

$$\Phi(\mathbf{w}) = Q^{(t)}(\mathbf{w}) - \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

Recall that the M step does the following:

$$\mathbf{w}^{(t+1)} = \arg\max_{\mathbf{w}} Q^{(t)}(\mathbf{w})$$

$$\Phi(\mathbf{w}) - Q^{(t)}(\mathbf{w}) = -\sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

$$\Phi(\mathbf{w}) = Q^{(t)}(\mathbf{w}) - \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

Recall that the M step does the following:

$$\mathbf{w}^{(t+1)} = \arg\max_{\mathbf{w}} Q^{(t)}(\mathbf{w})$$

Therefore (part 1),

$$\max_{\mathbf{w}} Q^{(t)}(\mathbf{w}) = Q^{(t)}(\mathbf{w}^{(t+1)}) \geq Q^{(t)}(\mathbf{w}^{(t)})$$

$$\Phi(\mathbf{w}) = \boxed{Q^{(t)}(\mathbf{w})} \boxed{- \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)}$$

$$- \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}^{(t+1)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i) \geq - \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}^{(t)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

$$\Phi(\mathbf{w}) = Q^{(t)}(\mathbf{w}) - \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

$$\left(- \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}^{(t+1)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i) \right) - \left(\sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}^{(t)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i) \right)$$

$$= \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}^{(t)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i) - \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log p_{\mathbf{w}^{(t+1)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)$$

$$= \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log \frac{p_{\mathbf{w}^{(t)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)}{p_{\mathbf{w}^{(t+1)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)}$$

$$= \sum_{i=1}^{\tilde{N}} \sum_{\boldsymbol{y} \in \mathcal{Y}} q_i^{(t)}(\boldsymbol{y}) \log \frac{q_i^{(t)}(\boldsymbol{y})}{p_{\mathbf{w}^{(t+1)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i)}$$

$$= \sum_{i=1}^{\tilde{N}} D_{KL} \left(q_i^{(t)}(\boldsymbol{y}) \mid\mid p_{\mathbf{w}^{(t+1)}}(\boldsymbol{y} \mid \tilde{\boldsymbol{x}}_i) \right)$$

$$\geq 0$$

In the structured case

$$p(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(t)}) = \frac{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}^{(t)})}{\sum_{\boldsymbol{y} \in \mathcal{Y}_{\boldsymbol{x}}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}^{(t)})}$$

- Let us assume the latent variables (and probably the observations) are structured
- EM works just the same as always
-

Aggregate Bigram Model

- Process
  - Let x_0= <s>
  - Let i = 0
  - While $x_i \neq$</s> repeat:
    - $i \leftarrow i + 1$
    - Sample a class y_i from $p(Y = y_i | X = x_{i-1})$
    - Sample a word x_i from $p(X = x_i | Y = y_i)$

Saul & Pereira. (1997). "Word Classes"

Aggregate Bigram Model

The parameters of the model are:

$$\theta = \langle a, b \rangle$$

$a(x|y) =$ The probability of every word in the vocabulary following every class.

$b(y|x) =$ The probability of every class following every word in the vocabulary.

Aggregate Bigram Model

The parameters of the model are:

$$\theta = \langle a, b \rangle$$

$a(x|y) =$ The probability of every word in the vocabulary following every class.

$b(y|x) =$ The probability of every class following every word in the vocabulary.

How many are there in total?

Aggregate Bigram Model

The parameters of the model are:

$$\theta = \langle a, b \rangle$$

$a(x|y) =$ The probability of every word in the vocabulary following every class.

$b(y|x) =$ The probability of every class following every word in the vocabulary.

How many are there in total?

$$|\theta| = |V| \times K \times 2$$

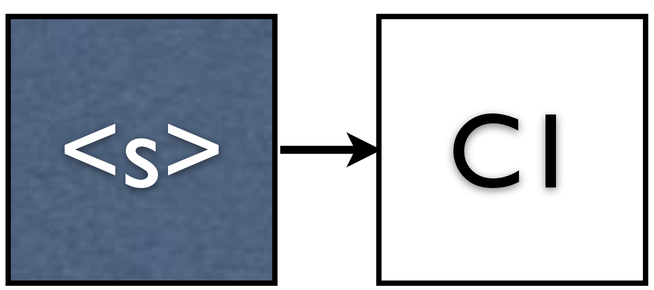
Aggregate Bigram Model

<s>

1

Aggregate Bigram Model

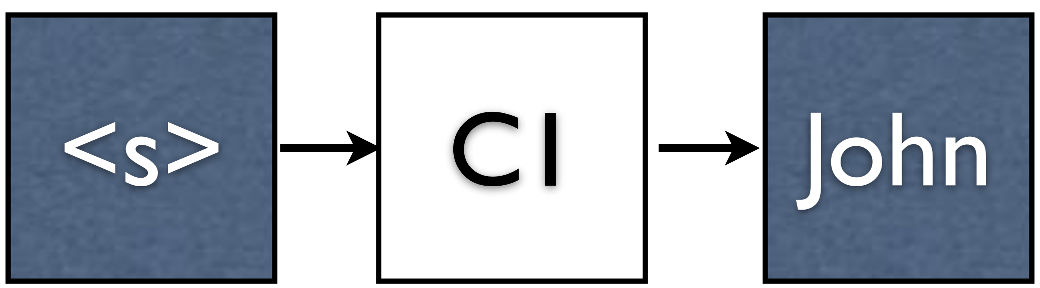
$b(C1|\langle \mathtt{s} \rangle)$



1

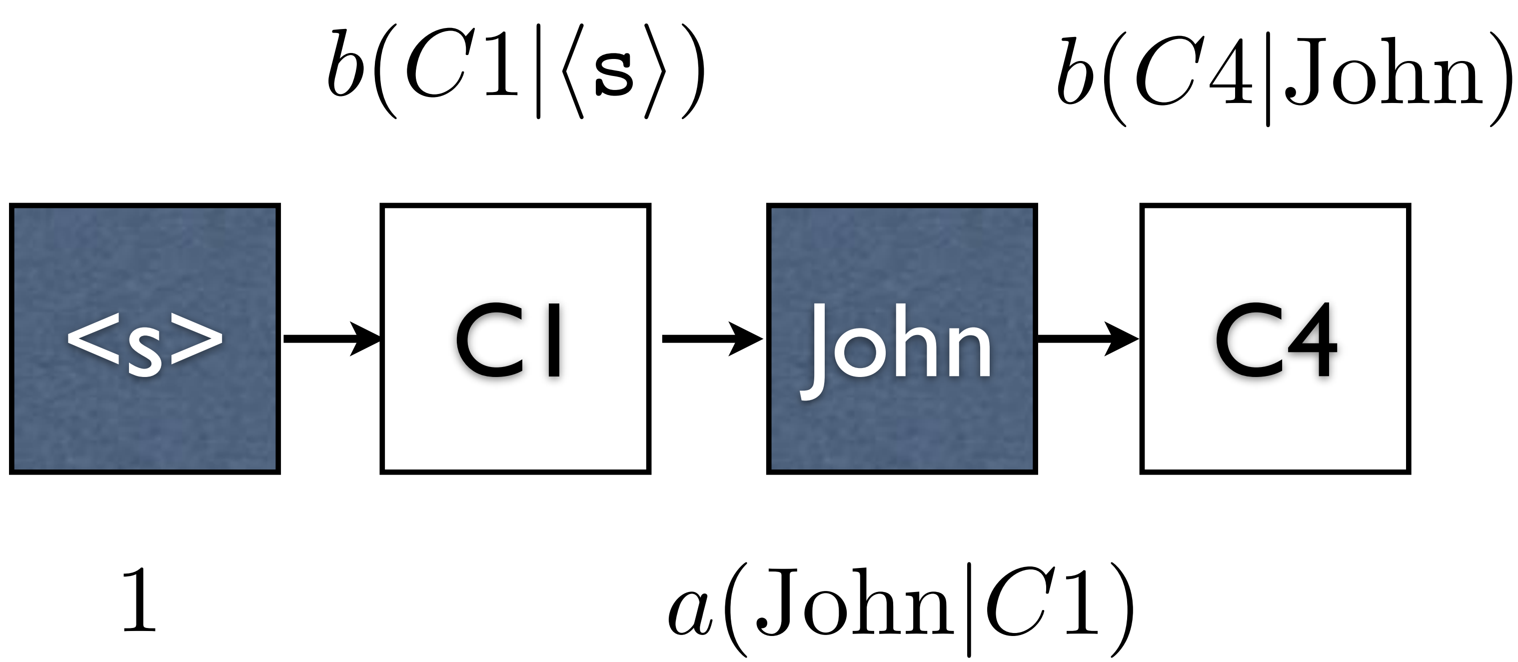
Aggregate Bigram Model

$b(C1|\langle \mathtt{s} \rangle)$

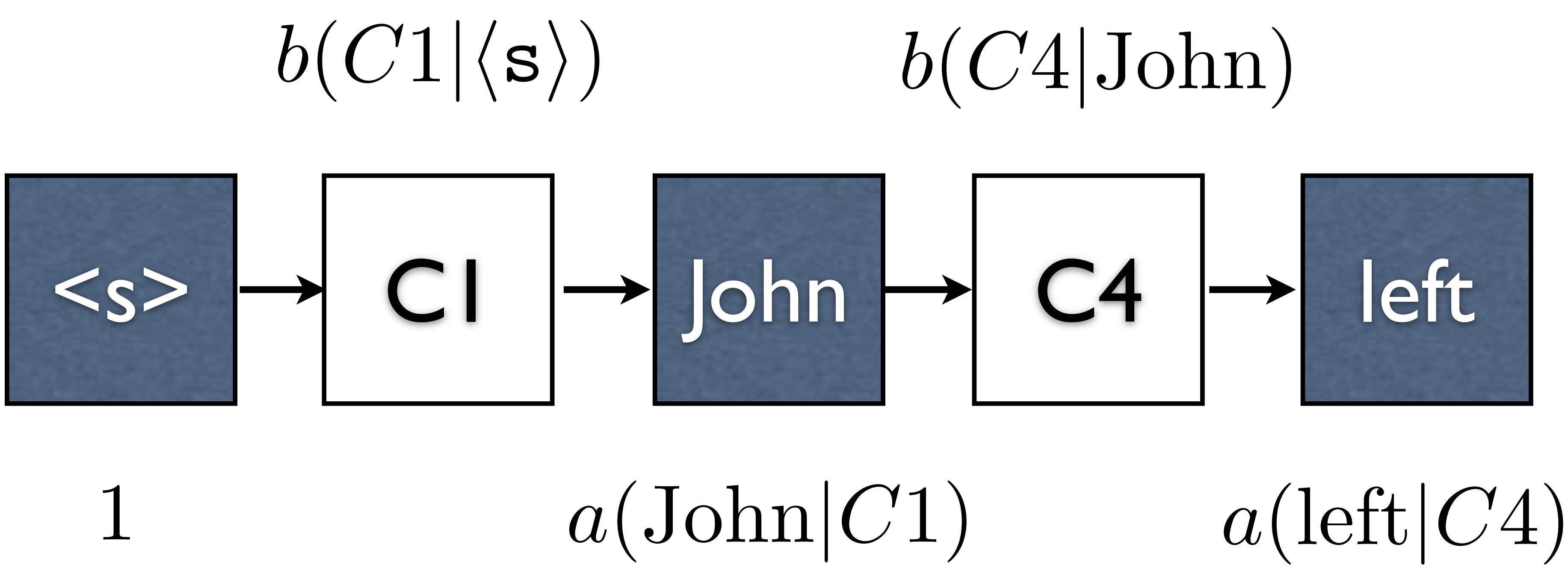


1 $a(\mathrm{John}|C1)$

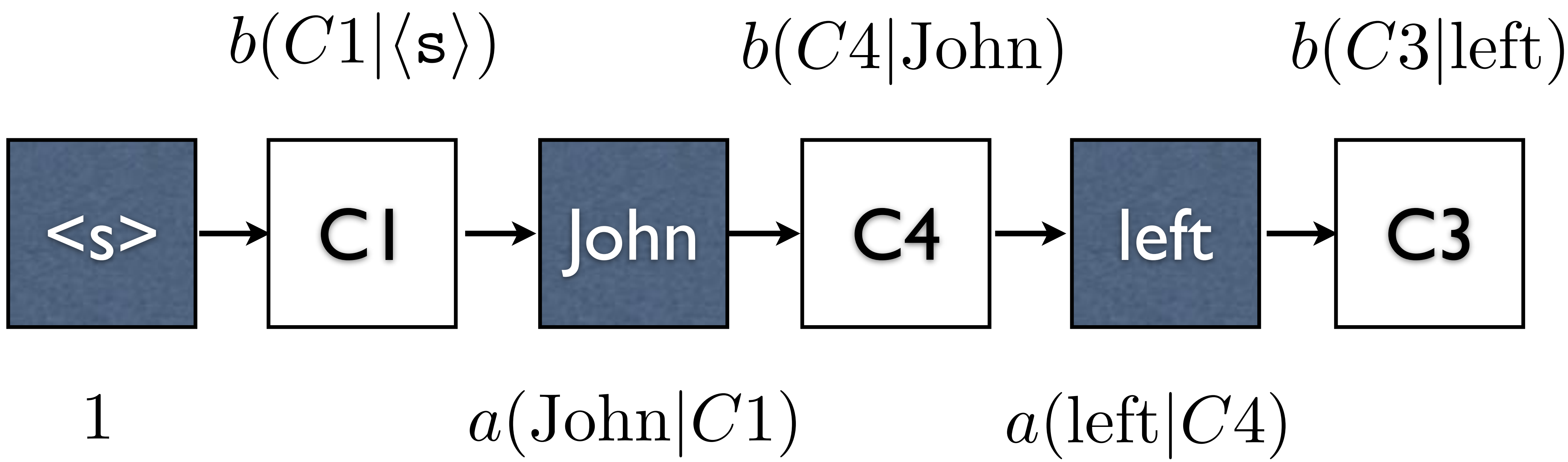
Aggregate Bigram Model



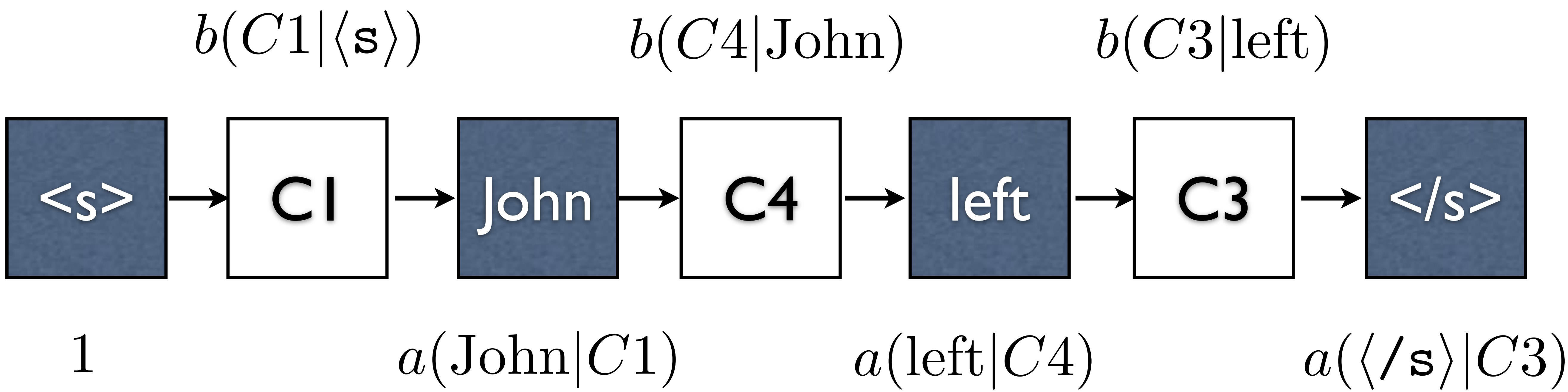
Aggregate Bigram Model



Aggregate Bigram Model



Aggregate Bigram Model



Aggregate Bigram Model

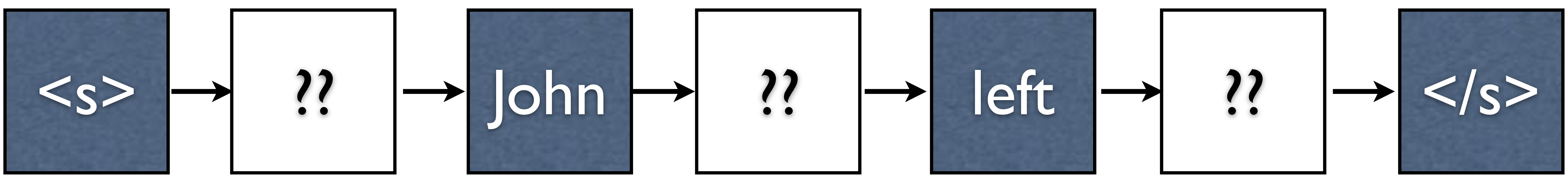
- How do we learn parameters?
  - We have a joint probability model
  - We have some observable data (the words)
  - We have some hidden data (the classes)

Aggregate Bigram Model

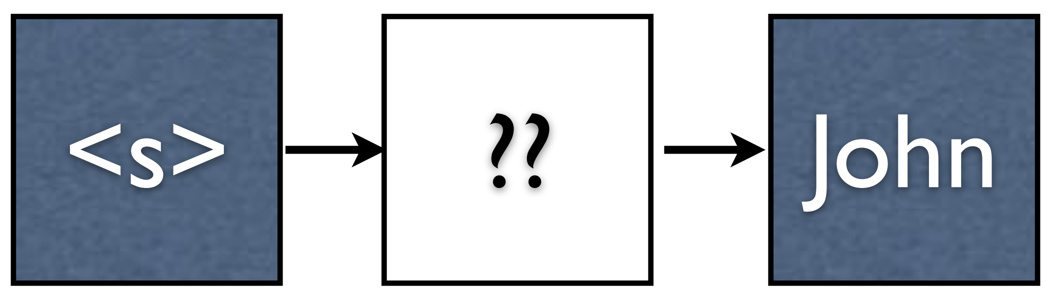
- How do we learn parameters?
  - We have a joint probability model
  - We have some observable data (the words)
  - We have some hidden data (the classes)

EM

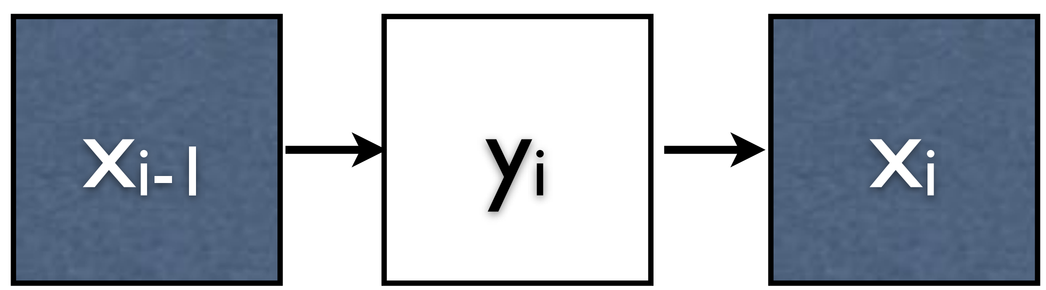
Aggregate Bigram Model



Aggregate Bigram Model

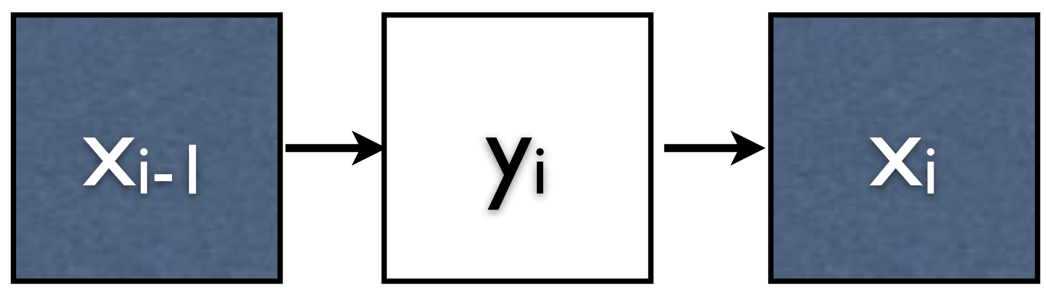


Aggregate Bigram Model



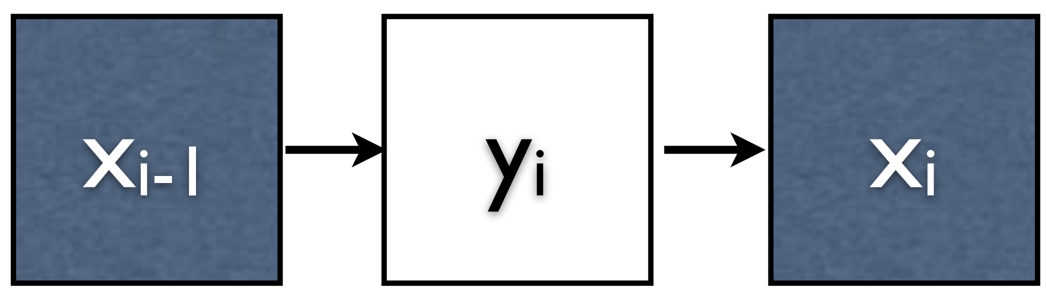
Aggregate Bigram Model

$b(y_i|x_{i-1})$



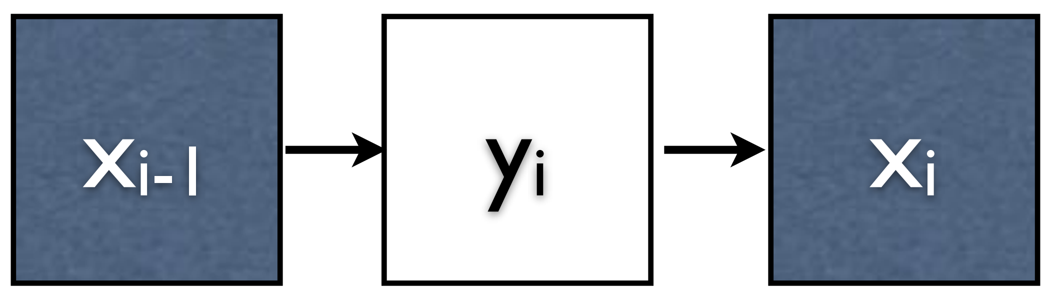
Aggregate Bigram Model

$$b(y_i|x_{i-1}) \ \times a(x_i|y_i)$$



Aggregate Bigram Model

$$b(y_i|x_{i-1}) \ \times a(x_i|y_i)$$

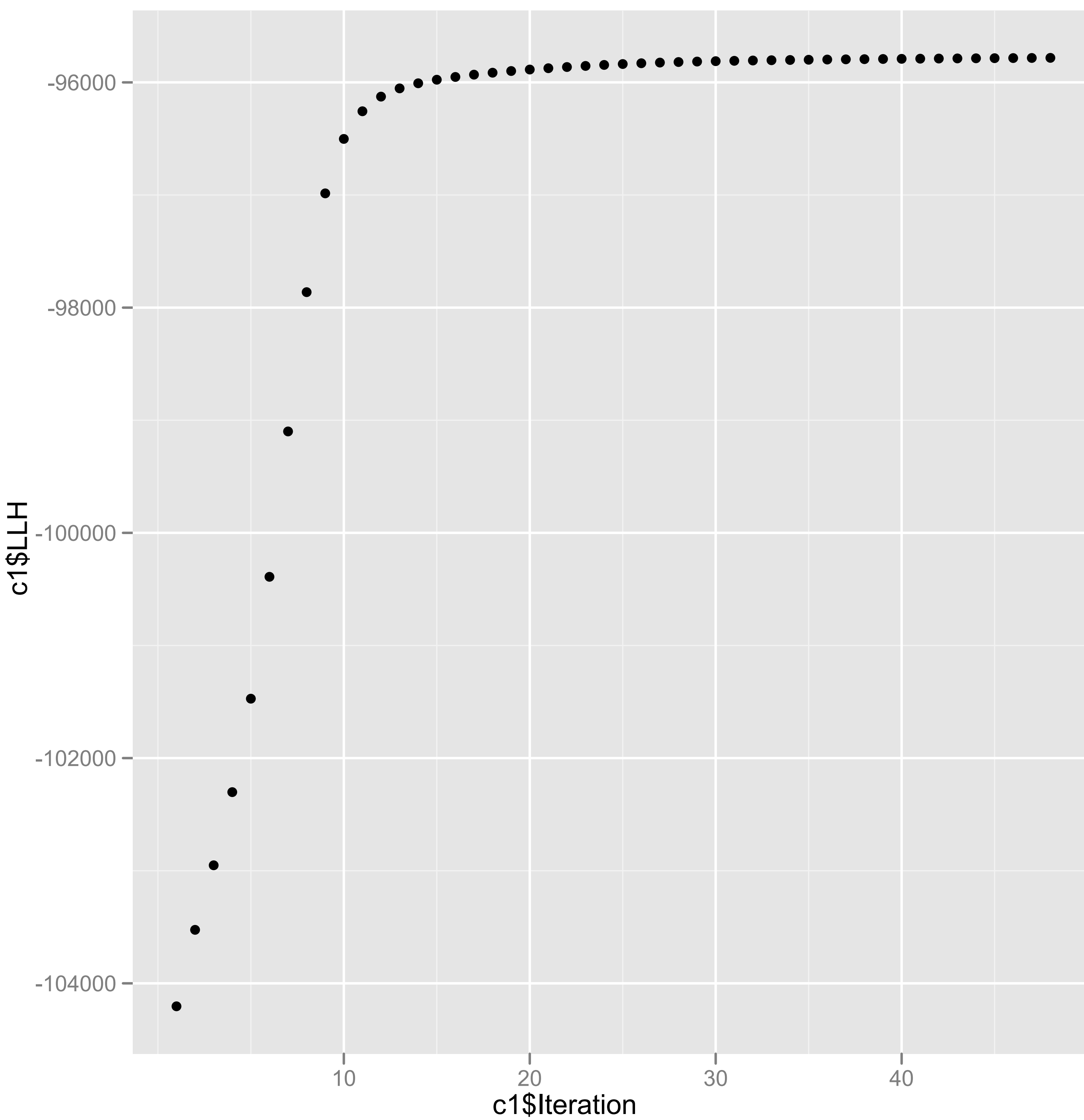


$$p(y_i|x_{i-1}, x_i) = \frac{b(y_i|x_{i-1}) \times a(x_i|y_i)}{\sum_{y'=1}^{K} b(y'|x_{i-1}) \times a(x_i|y')}$$

Example

- Let’s treat the **letters** in English words as the “words” in our language
  - Output: clustering over letters
- For this example, we assume K=2

Likelihood



-96000
-98000
-100000
-102000
-104000
c1$LLH
10
20
30
40
c1$Iteration

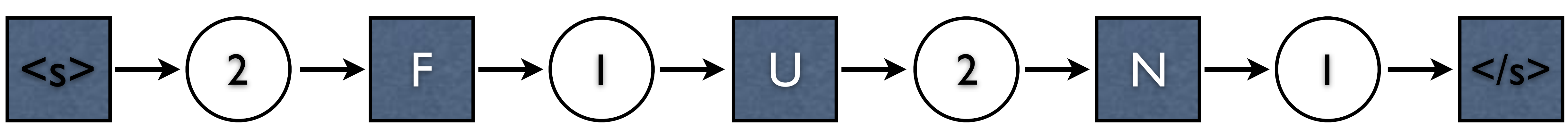
What was learned?

$a(X = \cdot \mid Y = 1)$

</s>	0.23
E	0.19
A	0.11
I	0.11
O	0.09
T	0.04
U	0.04
H	0.04
S	0.03
L	0.03

$a(X = \cdot \mid Y = 2)$

N	0.23
S	0.19
R	0.11
T	0.11
C	0.09
D	0.04
L	0.04
G	0.04
M	0.03
P	0.03



Word Alignment

das Haus	ein Buch	das Buch
the house	a book	the book

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences

$$\mathbf{e} = \langle e_1, e_2, \ldots, e_m \rangle \qquad \mathbf{f} = \langle f_1, f_2, \ldots, f_n \rangle$$

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences
- Lexical translation makes the following ***assumptions***:
  - Each word in e_i in $\mathbf{e}$ is generated from exactly one word in $\mathbf{f}$
  - Thus, we have an *alignment* a_i that indicates which word e_i "came from", specifically it came from f_{a_i}
  - Given the alignments $\mathbf{a}$, translation decisions are conditionally independent of each other and depend *only* on the aligned source word f_{a_i}

IBM Model 1

- Simplest possible lexical translation model
- Additional assumptions
  - The m alignment decisions are independent
  - The alignment distribution for each a_i is uniform over all source words and NULL

$$
\begin{aligned}
&\text{for each } i \in [1, 2, \ldots, m] \\
&\qquad a_i \sim \text{Uniform}(0, 1, 2, \ldots, n) \\
&\qquad e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})
\end{aligned}
$$

IBM Model 1

$$\boxed{\text{for each } i \in [1, 2, \ldots, m]}$$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m}$$

IBM Model 1

$$\text{for each } i \in [1, 2, \ldots, m]$$

$$\boxed{a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)}$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n}$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$\boxed{e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})}$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n} \, p(e_i \mid f_{a_i})$$

IBM Model 1

for each $i \in [1, 2, \ldots, m]$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$\boxed{e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})}$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n} \, p(e_i \mid f_{a_i})$$

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i, a_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

$$p(a, b, c, d) = p(a)p(b)p(c)p(d)$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$\begin{aligned} p(\mathbf{e} \mid \mathbf{f}, m) &= \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m) \\ &= \prod_{i=1}^{m} \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i}) \end{aligned}$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$\begin{aligned} p(\mathbf{e} \mid \mathbf{f}, m) &= \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m) \\ &= \prod_{i=1}^{m} \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i}) \\ &= \frac{1}{(1+n)^m} \prod_{i=1}^{m} \sum_{a_i=0}^{n} p(e_i \mid f_{a_i}) \end{aligned}$$

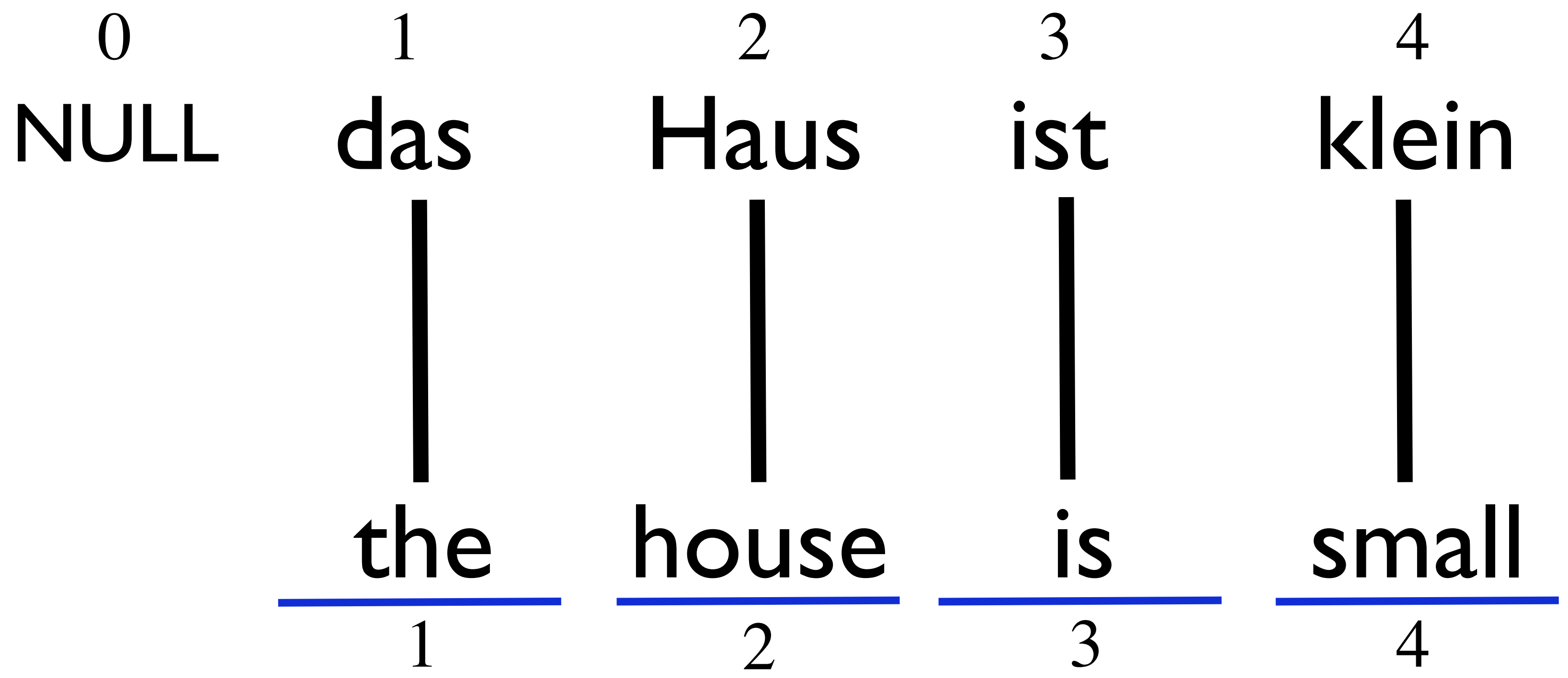
Example

0	1	2	3	4
NULL	das	Haus	ist	klein

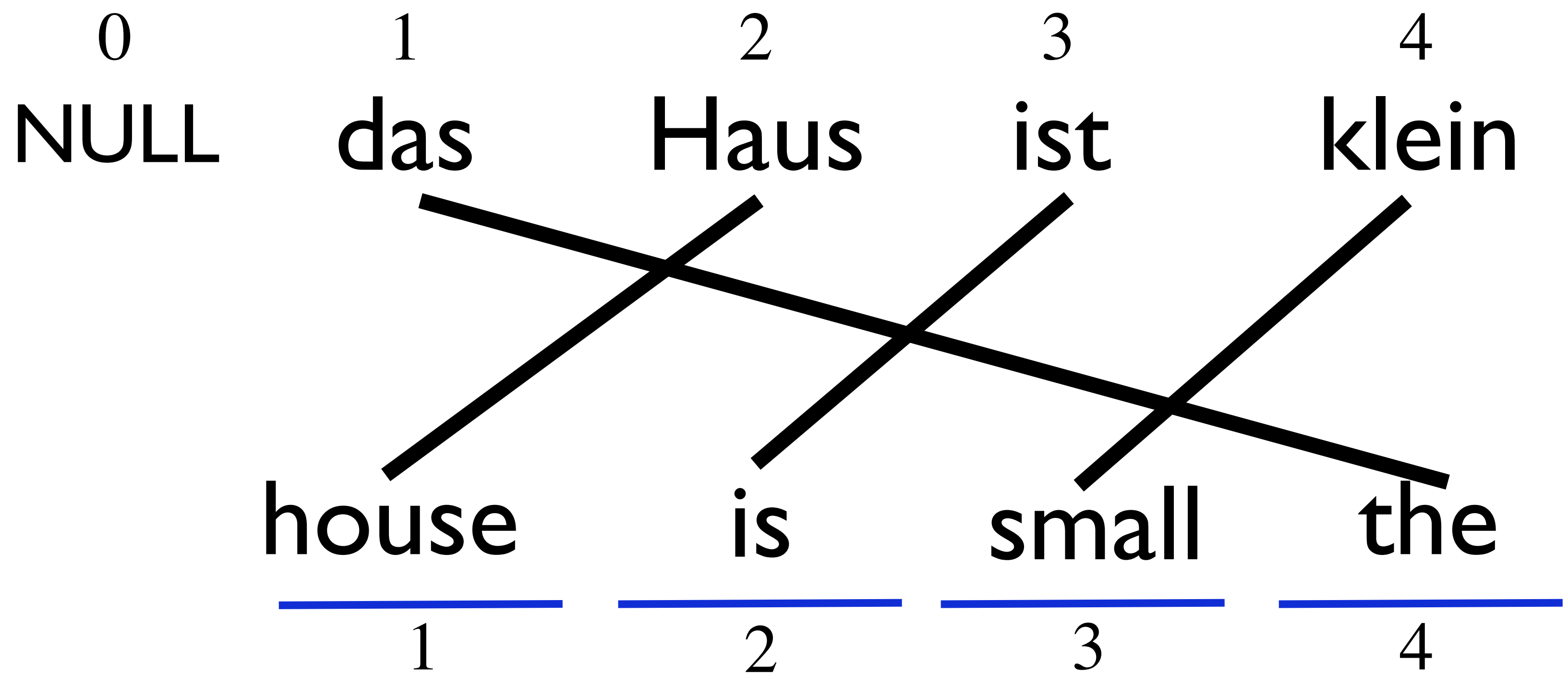
___	___	___	___
1	2	3	4

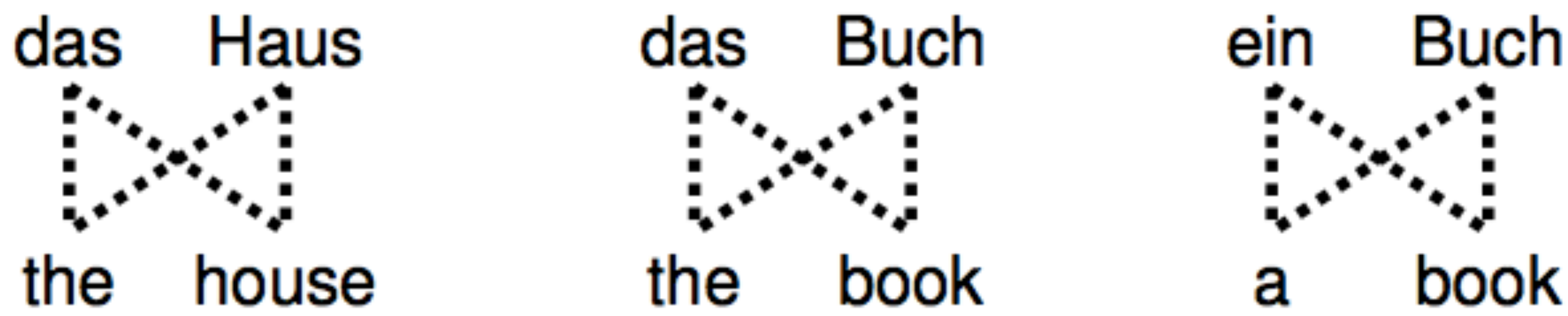
Start with a foreign sentence and a target length.

Example



Example





e	f	initial
the	das	0.25
book	das	0.25
house	das	0.25
the	buch	0.25
book	buch	0.25
a	buch	0.25
book	ein	0.25
a	ein	0.25
the	haus	0.25
house	haus	0.25

$$\textit{freq}(\text{Buch}, \text{book}) = ?$$

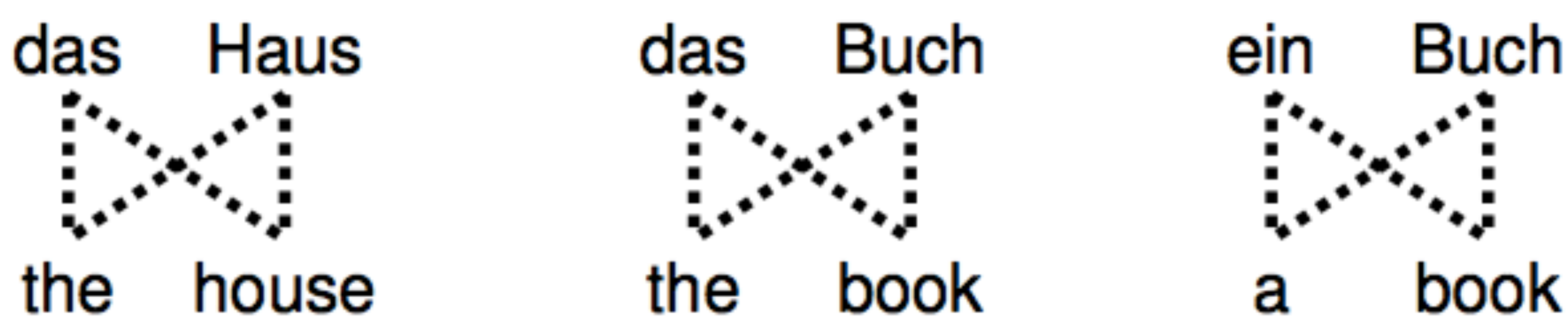
$$\textit{freq}(\text{das}, \text{book}) = ?$$

$$\textit{freq}(\text{ein}, \text{book}) = ?$$

$$\textit{freq}(\text{Buch}, \text{book}) = \sum_i \mathbb{I}(\tilde{e}_i = \text{book}, \tilde{f}_{a_i} = \text{Buch})$$

$$\mathbb{E}_{p_{\mathbf{w}^{(1)}}(\mathbf{a} | \mathbf{f} = \text{das Buch}, \mathbf{e} = \text{the book})} \sum_i \mathbb{I}[e_i = \text{book}, f_{a_i} = \text{Buch}]$$

Convergence



e	f	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

Evaluation

- Since we have a probabilistic model, we can evaluate **perplexity**.

$$\mathrm{PPL} = 2^{-\frac{1}{\sum_{(\mathbf{e},\mathbf{f})\in\mathcal{D}} |\mathbf{e}|} \log \prod_{(\mathbf{e},\mathbf{f})\in\mathcal{D}} p(\mathbf{e}|\mathbf{f})}$$

	Iter 1	Iter 2	Iter 3	Iter 4	...	Iter ∞
-log likelihood	-	7.66	7.21	6.84	...	-6
perplexity	-	2.42	2.3	2.21	...	2

Hidden Markov Models

- GMMs, the aggregate bigram model, and Model 1 don't have conditional dependencies between random variables
- Let's consider an example of a model where this is not the case

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}_{\boldsymbol{x}}} \eta(y_{|\boldsymbol{x}|} \rightarrow \text{STOP}) \prod_{i=1}^{|\boldsymbol{x}|} \eta(y_{i-1} \rightarrow y_i) \times \gamma(y_i \downarrow x_i)$$

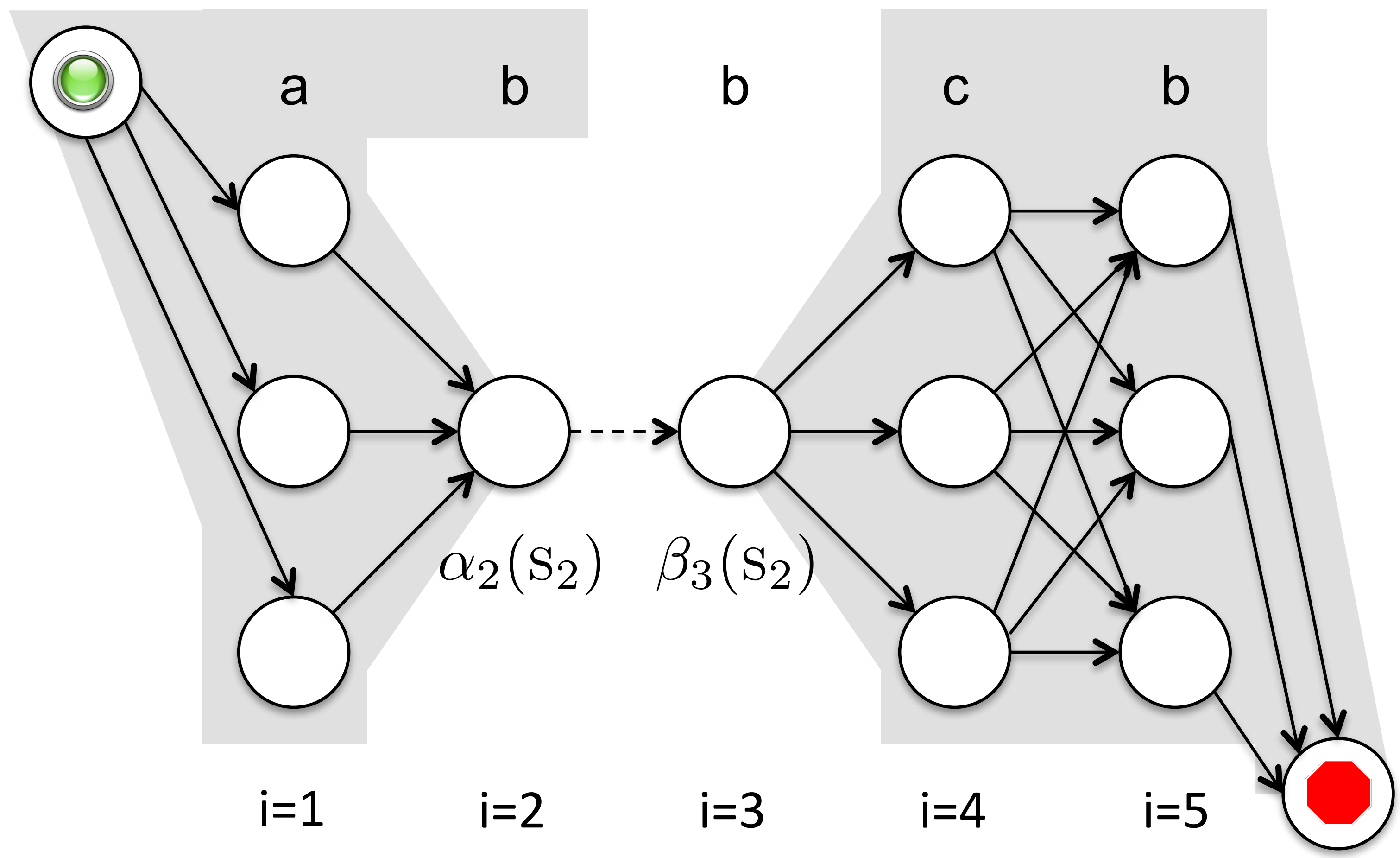
EM for HMMs

- What statistics are sufficient to determine the parameter values?

$$\begin{array}{rl} \mathit{freq}(q \downarrow x) & \text{How often does } q \text{ emit } x\text{?} \\ \mathit{freq}(q \to r) & \text{How often does } q \text{ transition to } r\text{?} \\ \mathit{freq}(q) & \text{How often do we visit } q\text{?} \end{array}$$

And of course...

$$\mathit{freq}(q) = \sum_{r \in \mathcal{Q}} \mathit{freq}(q \to r)$$



$$p(y_2 = q, y_3 = r \mid \boldsymbol{x}) \propto p(y_2 = q, y_3 = r, \boldsymbol{x})$$

$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \rightarrow r) \times \eta(r \downarrow x_3)}{\sum_{q', r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \rightarrow r') \times \eta(r' \downarrow x_3)}$$

$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \rightarrow r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\mathrm{STOP})}$$

$$
\begin{aligned}
p(y_2 = q, y_3 = r \mid \boldsymbol{x}) \propto p(y_2 = q, y_3 = r, \boldsymbol{x}) \\
&= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{\sum_{q',r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \to r') \times \eta(r' \downarrow x_3)} \\
&= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\text{STOP})}
\end{aligned}
$$

The expectation over the full structure is then

$$
\mathbb{E}[\mathit{freq}(q \to r)] = \sum_{i=1}^{|\boldsymbol{x}|} p(y_i = q, y_{i+1} = r \mid \boldsymbol{x})
$$

The expectation over state occupancy is

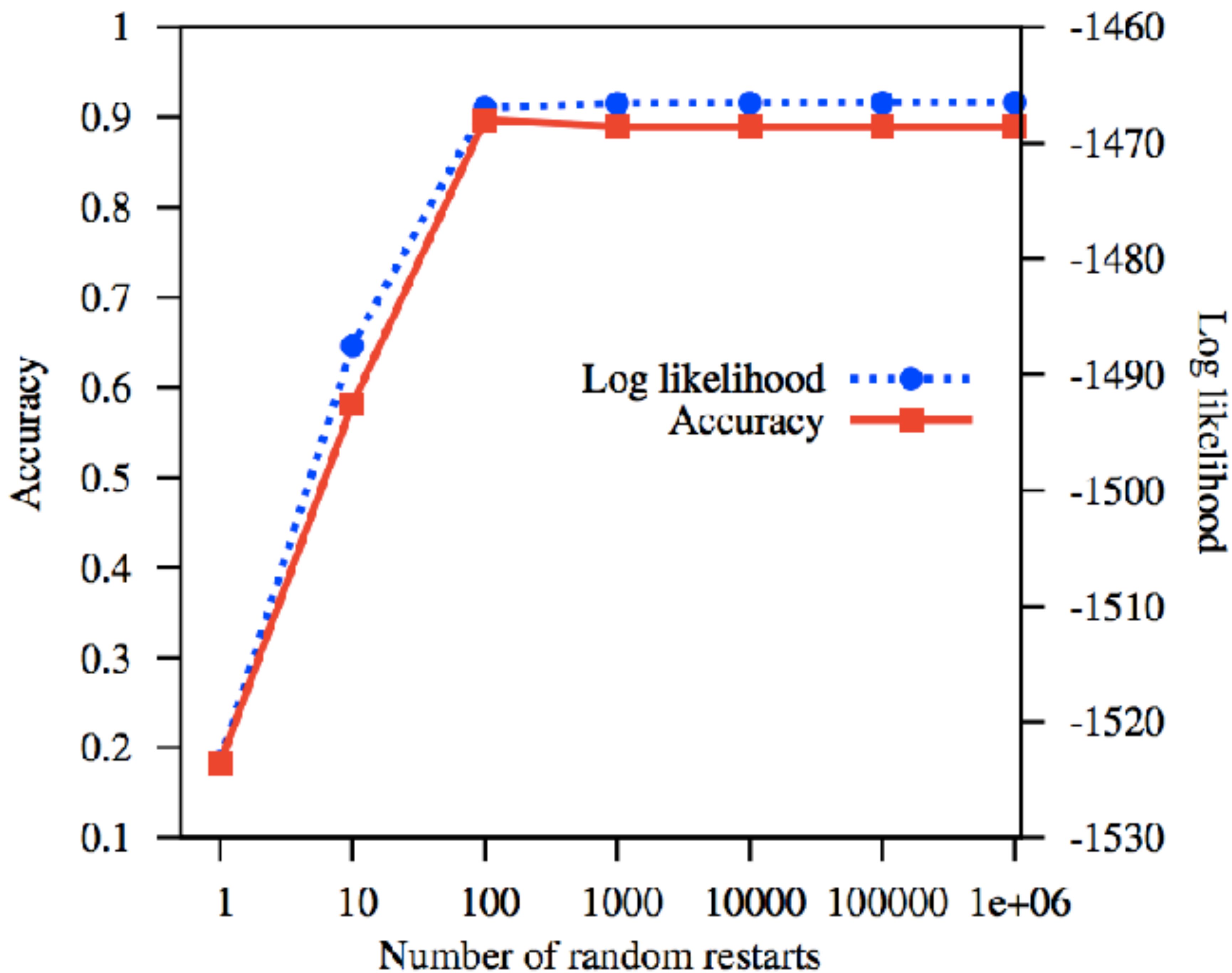
$$
\mathbb{E}[\mathit{freq}(q)] = \sum_{r \in \mathcal{Q}} \mathbb{E}[\mathit{freq}(q \to r)]
$$

What is $\mathbb{E}[\mathit{freq}(q \downarrow x)]$?

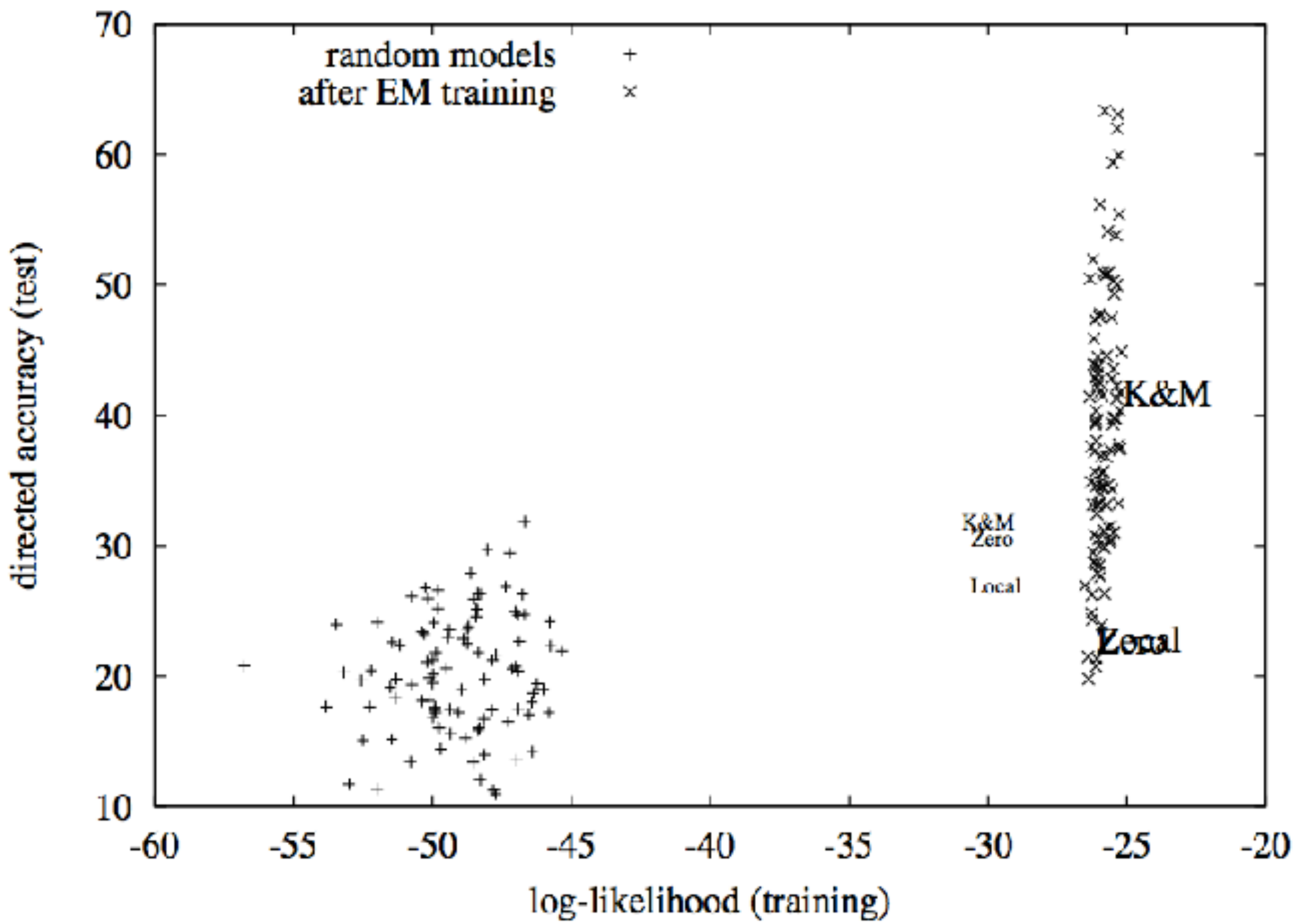
Random Restarts

- Non-convex optimization only finds a local solution
- Several strategies
  - Random restarts
  - Simulated annealing

Decipherment



Grammar Induction



Inductive Bias

- A model can learn nothing without inductive bias ... whence inductive bias?
  - Model structure
  - Priors (next week)
  - Posterior regularization (Google it)
- **Features** provide a very flexible means to bias a model

EM with Features

- Let's replace the multinomials with log linear distributions

$$\eta(q \to r) = \theta_{q,r}$$

$$= \frac{\exp \mathbf{w}^\top \boldsymbol{f}(q,r)}{\sum_{q' \in \mathcal{Q}} \exp \mathbf{w}^\top \boldsymbol{f}(q',r)}$$

How will the likelihood of this model compare to the likelihood of the previous model?

Learning Algorithm I

- E step
  - given model parameters, compute posterior distribution over transitions (states, etc)
  - compute $\mathbb{E}_{q(\boldsymbol{y})} \sum_{q,r} \boldsymbol{f}(q, r)$
  - These are your "empirical" expectations

Learning Algorithm 1

- M step
  - The gradient of the expected log likelihood of **x**,**y** under q(**y**) is

$$\nabla \mathbb{E}_{q(\boldsymbol{y})} \log p(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_q \sum_{q,r} \boldsymbol{f}(q,r) - \sum_{q,r} \mathbb{E}_q[freq(q)] \mathbb{E}_{p(r|q;\boldsymbol{w})} \boldsymbol{f}(q,r)$$

  - Use LBFGS or gradient descent to solve