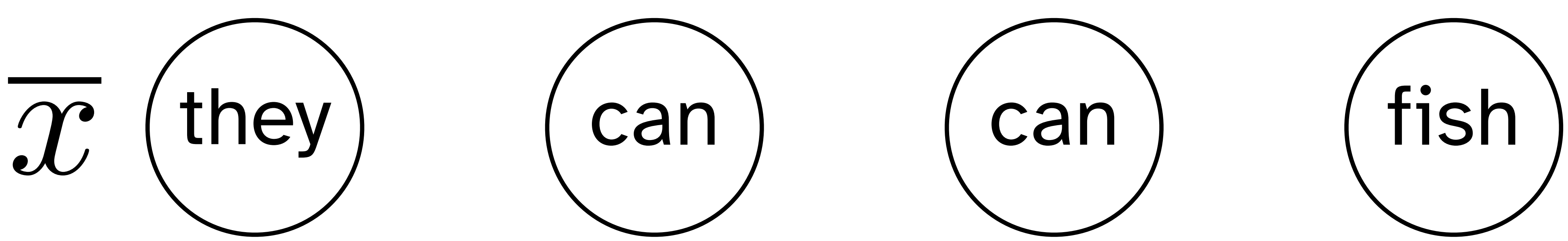
Structured Prediction II

Berkeley NLP

EECS 183/283a: Natural Language Processing

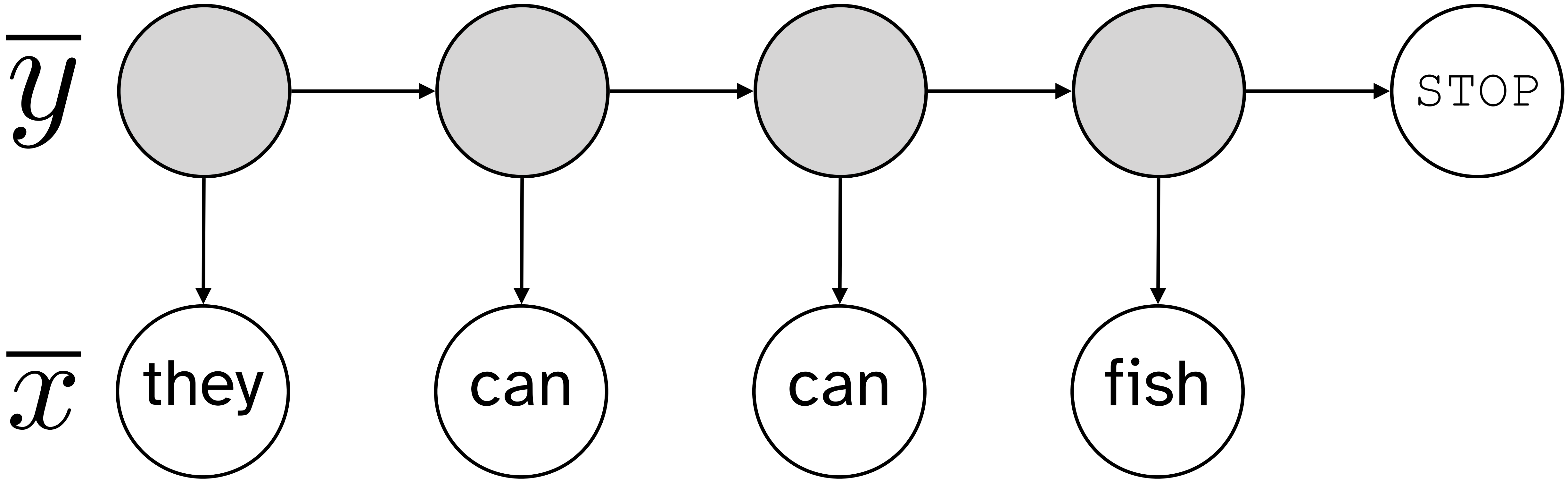
HMM Inference



HMM Inference

Generative story:

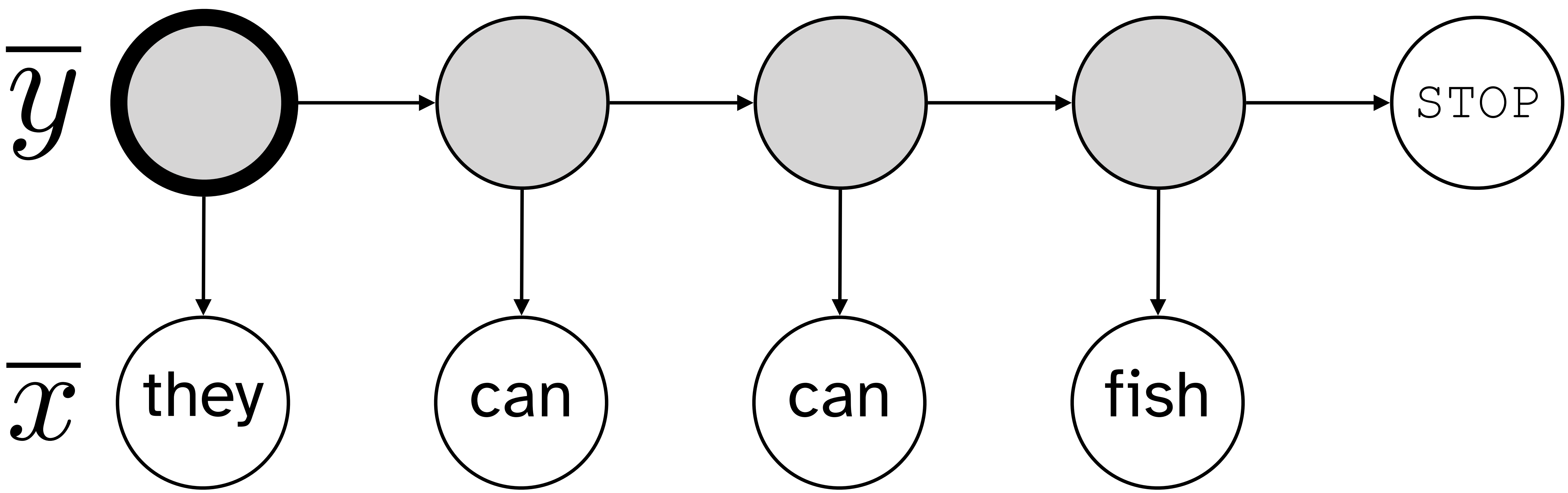
- Observed words are generated from hidden POS tags
- Each hidden POS tag is generated from a previous hidden POS tag
- The first POS tag is generated from some prior



HMM Inference

start logprobs

N	-1
V	-1



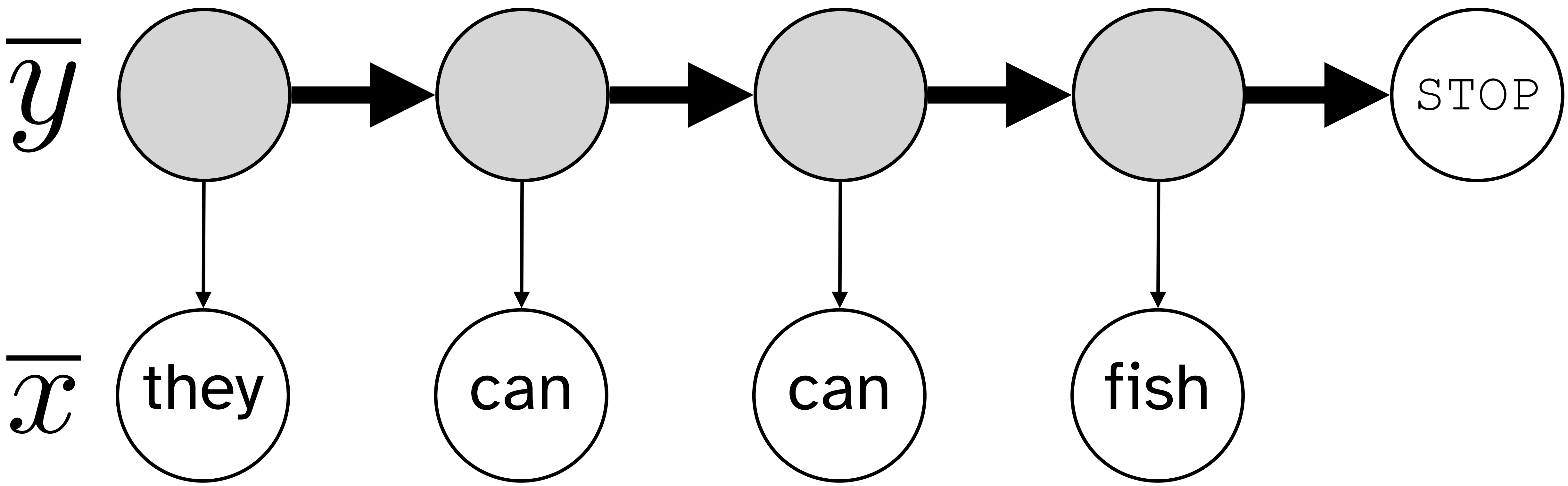
HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2



HMM Inference

start logprobs

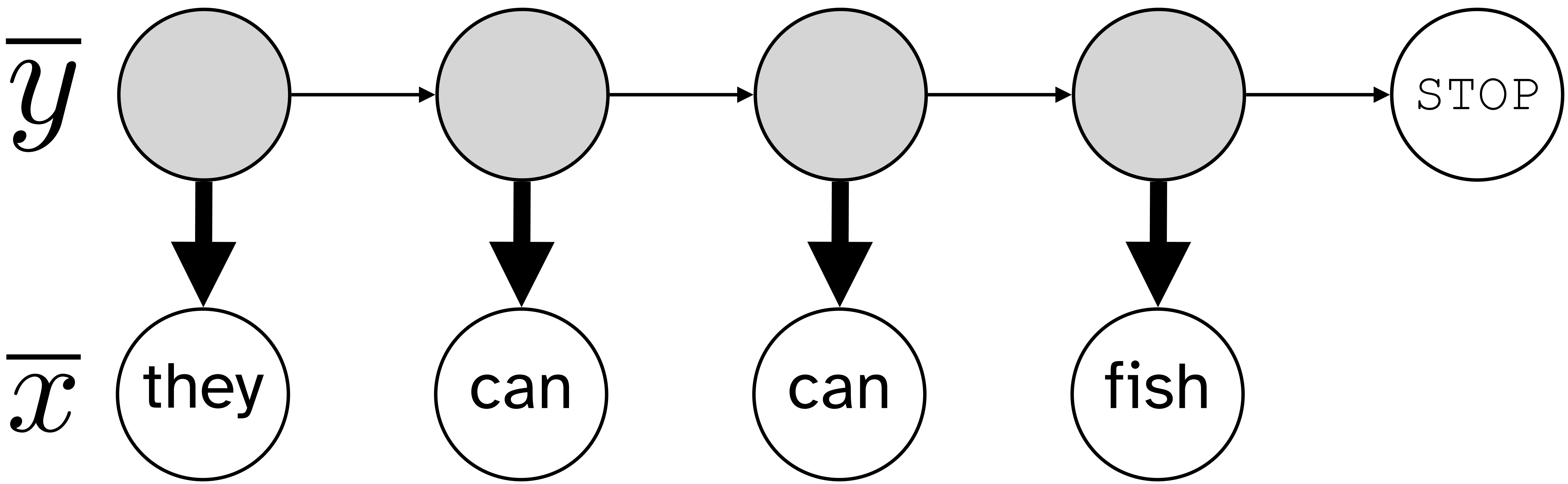
N	-1
V	-1

transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1



HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

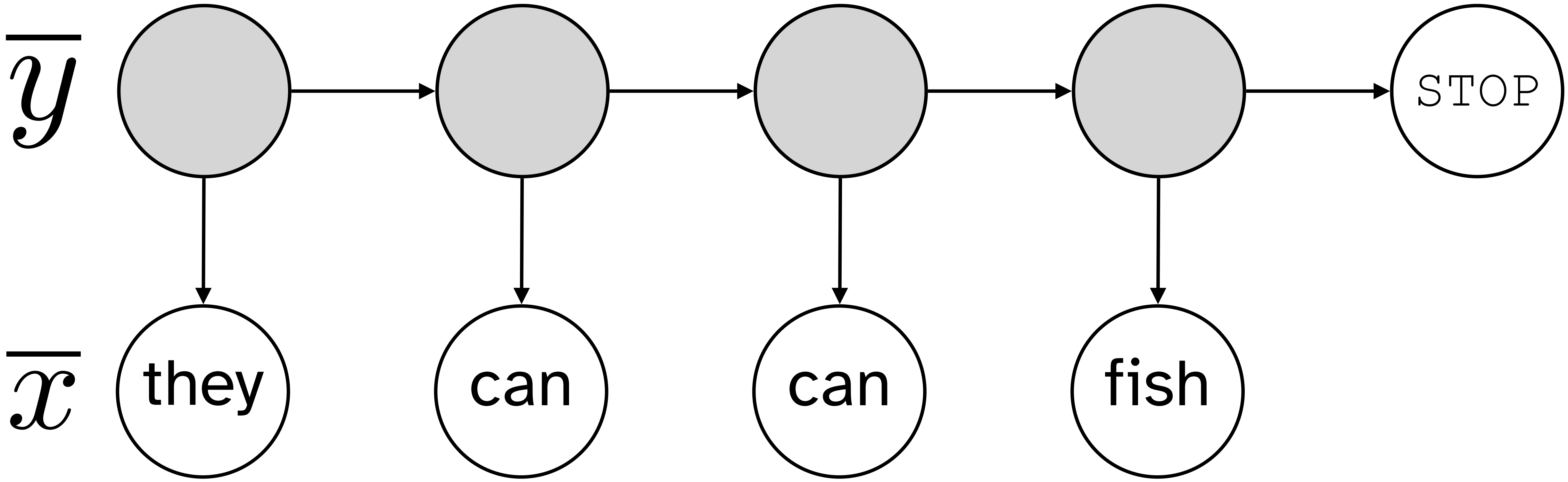
	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_1(\tilde{y}) = \log P(x_1 \mid \tilde{y}) + \log P(\tilde{y})$$

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}})\right)$$



HMM Inference

start logprobs

N	-1
V	-1

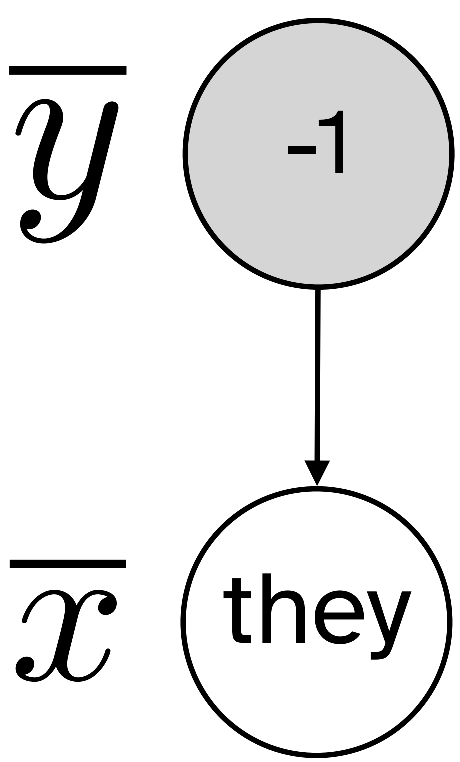
transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_1(\tilde{y}) = \log P(x_1 \mid \tilde{y}) + \log P(\tilde{y})$$



$$i = 1$$
$$\tilde{y} = \mathrm{N}$$

HMM Inference

start logprobs

N	-1
V	-1

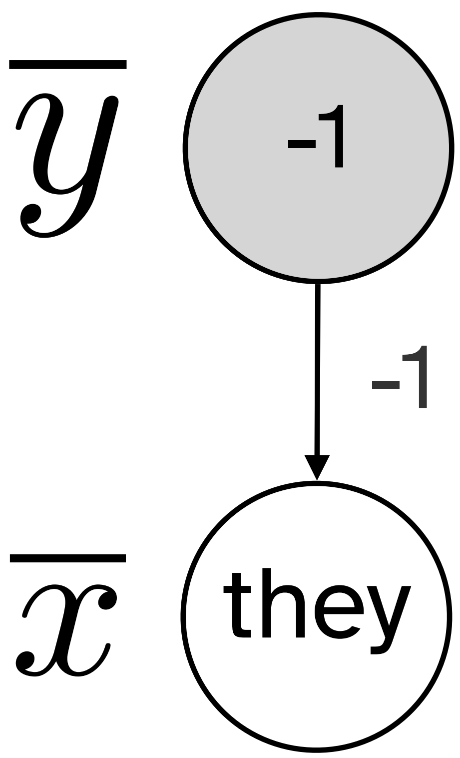
transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_1(\tilde{y}) = \log P(x_1 \mid \tilde{y}) + \log P(\tilde{y})$$



$$i = 1$$

$$\tilde{y} = \mathrm{N}$$

HMM Inference

start logprobs

N	-1
V	-1

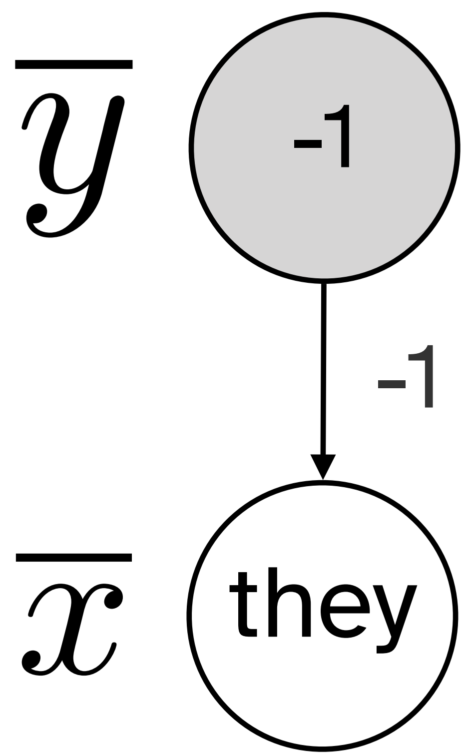
transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_1(\tilde{y}) = \log P(x_1 \mid \tilde{y}) + \log P(\tilde{y})$$
$$v_1(\mathrm{N}) = -2$$



$$i = 1$$
$$\tilde{y} = \mathrm{N}$$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

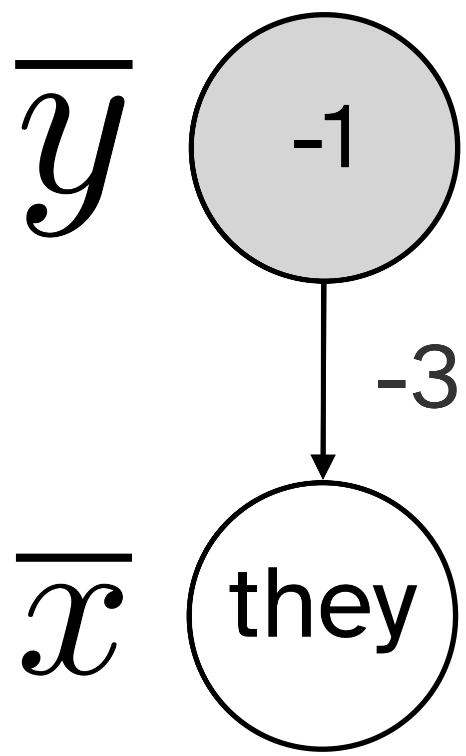
emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_1(\tilde{y}) = \log P(x_1 \mid \tilde{y}) + \log P(\tilde{y})$$

$$v_1(\mathrm{N}) = -2$$

$$v_1(V) = -4$$



$$i = 1$$

$$\tilde{y} = V$$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

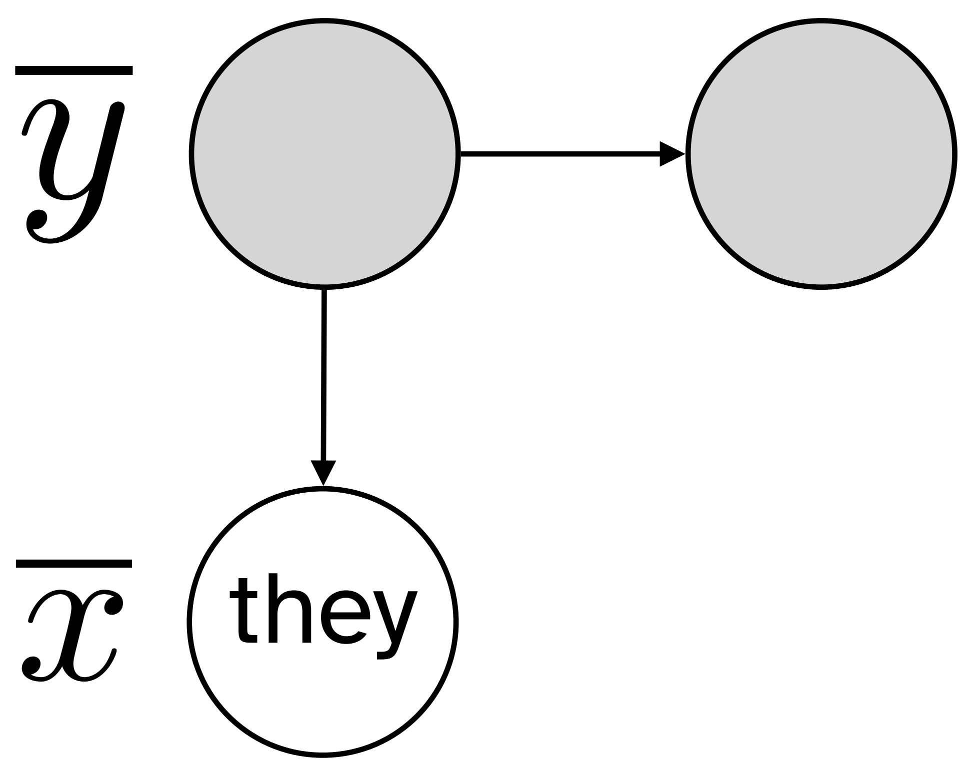
emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}})\right)$$

$$v_1(\mathrm{N}) = -2$$

$$v_1(V) = -4$$



$$i = 2$$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

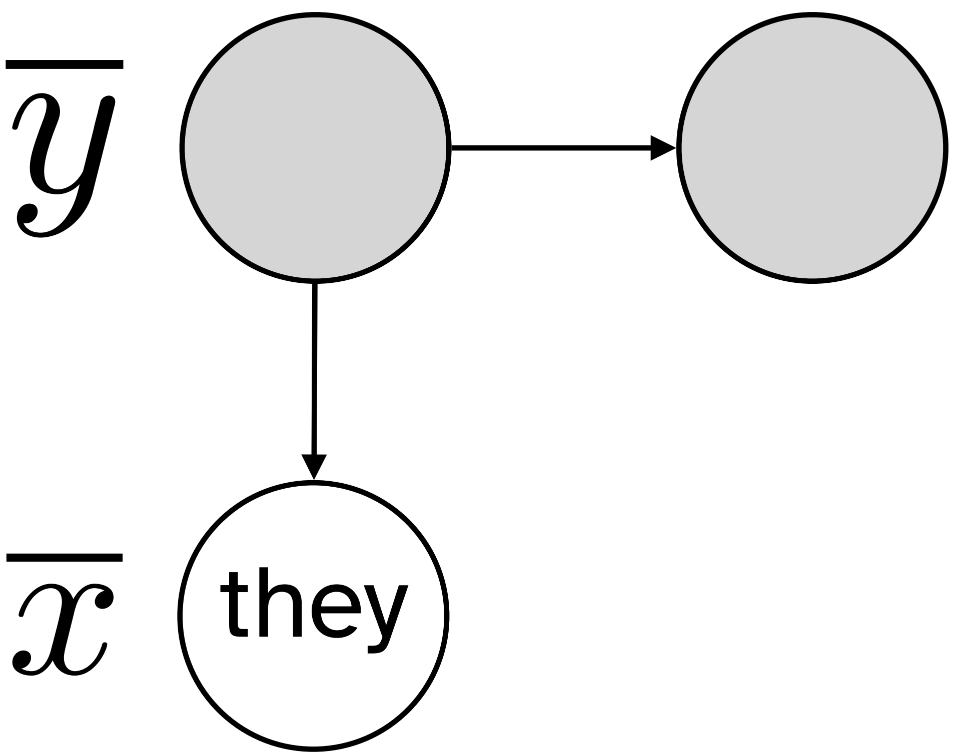
	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}}) \right)$$

v	1			
N	-2			
V	-4			



$i = 2$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

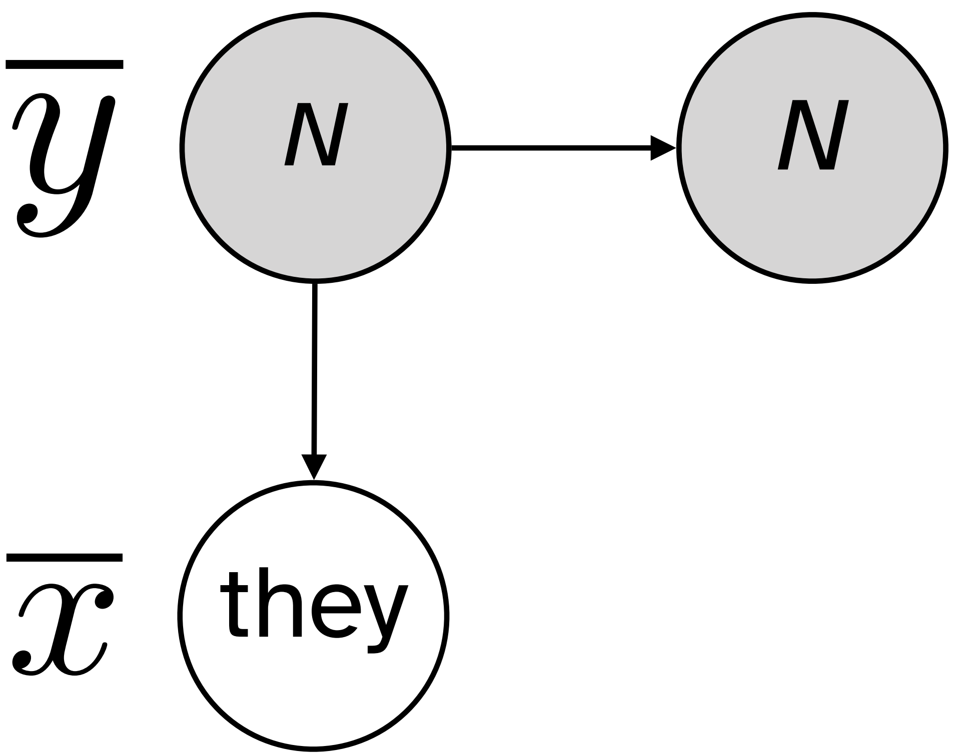
	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + \underbrace{v_{i-1}(\tilde{y}_{\text{prev}})}_{-2}\right)$$

v	1			
N	-2			
V	-4			



$$\tilde{y}_{\text{prev}} = \text{N}$$
$$i = 2$$
$$\tilde{y} = \text{N}$$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

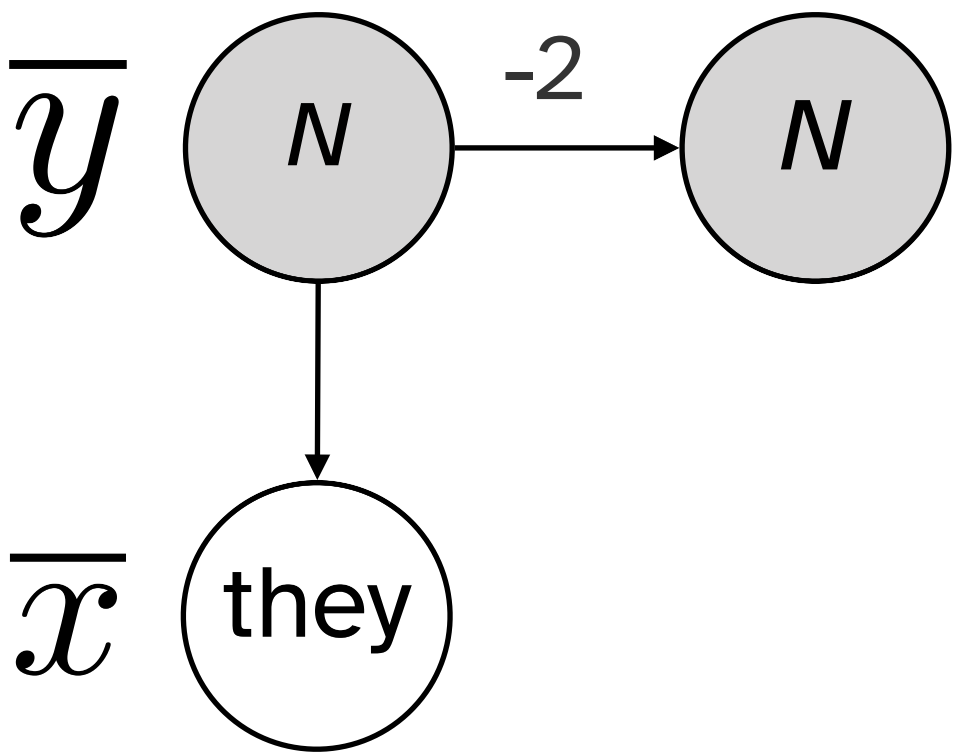
emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}}) \right)$$

-2

v	1			
N	-2			
V	-4			



-4 $\tilde{y}_{\text{prev}} = \text{N}$

$i = 2$

$\tilde{y} = \text{N}$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

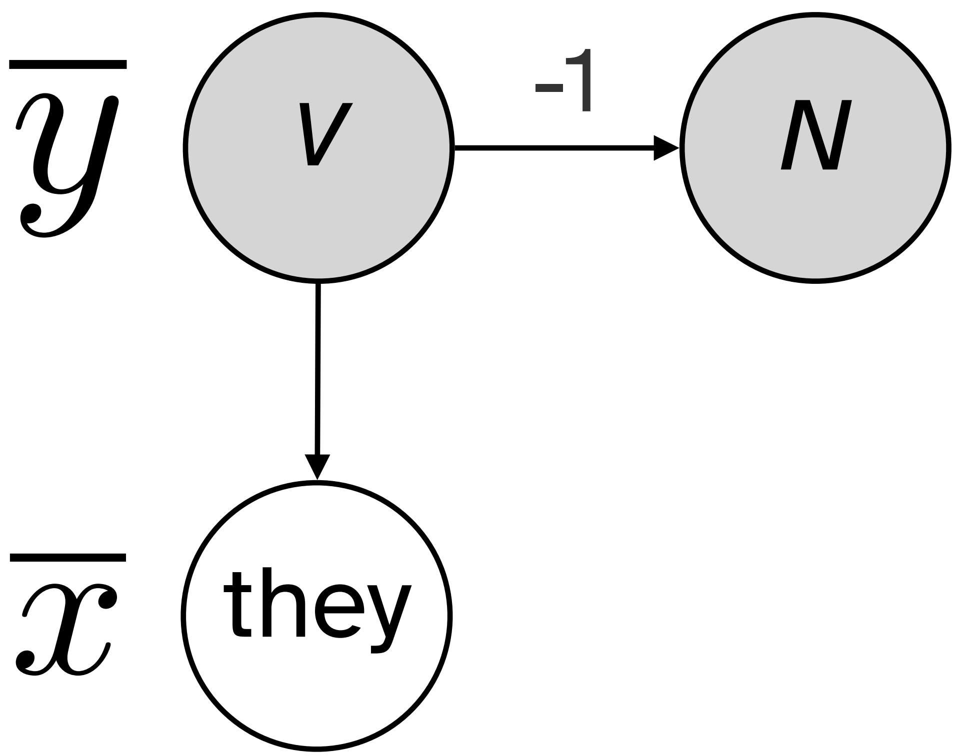
	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + \underbrace{v_{i-1}(\tilde{y}_{\text{prev}})}_{-4} \right)$$

v	1			
N	-2			
V	-4			



-5 $\tilde{y}_{\text{prev}} = \text{V}$

-4 $\tilde{y}_{\text{prev}} = \text{N}$

$i = 2$

$\tilde{y} = \text{N}$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

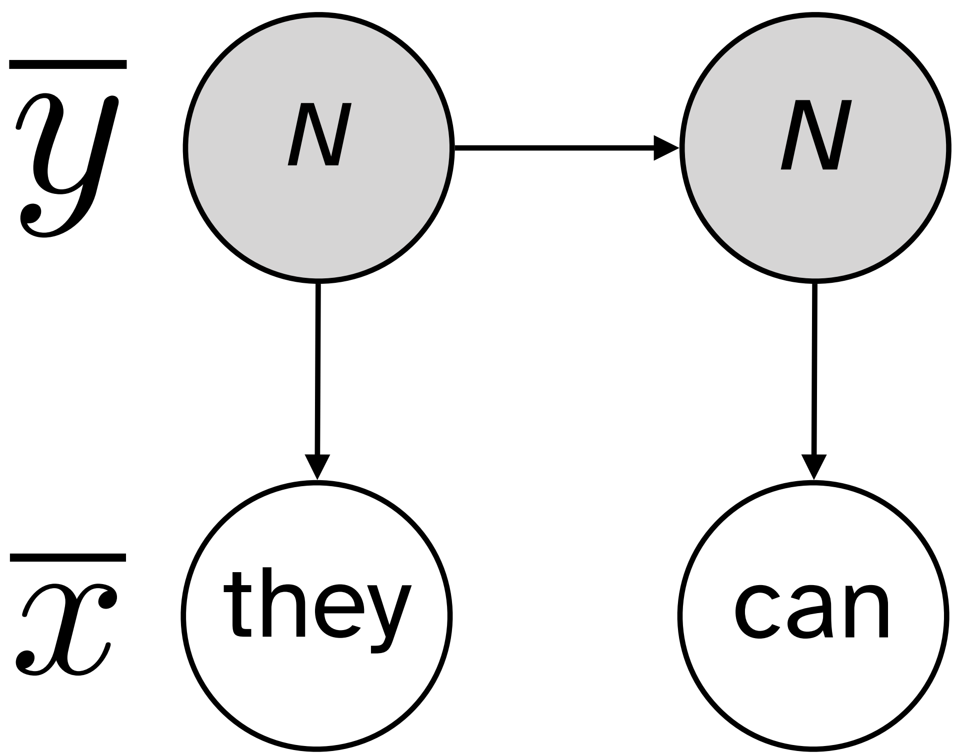
emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}}) \right)$$ -4

$\tilde{y}_{\text{prev}} = \text{N}$

v	1	2		
N	-2			
V	-4			



$i = 2$

$\tilde{y} = \text{N}$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

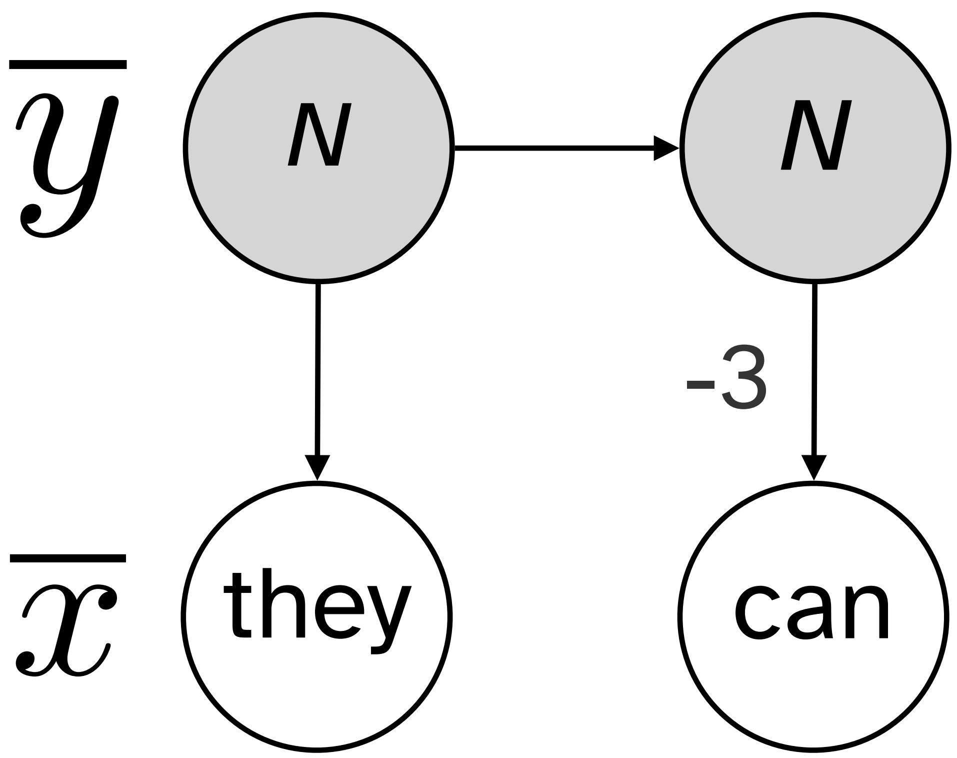
emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}})\right)$$ -4

$\tilde{y}_{\text{prev}} = \text{N}$

v	1	2		
N	-2			
V	-4			



$i = 2$

$\tilde{y} = \text{N}$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

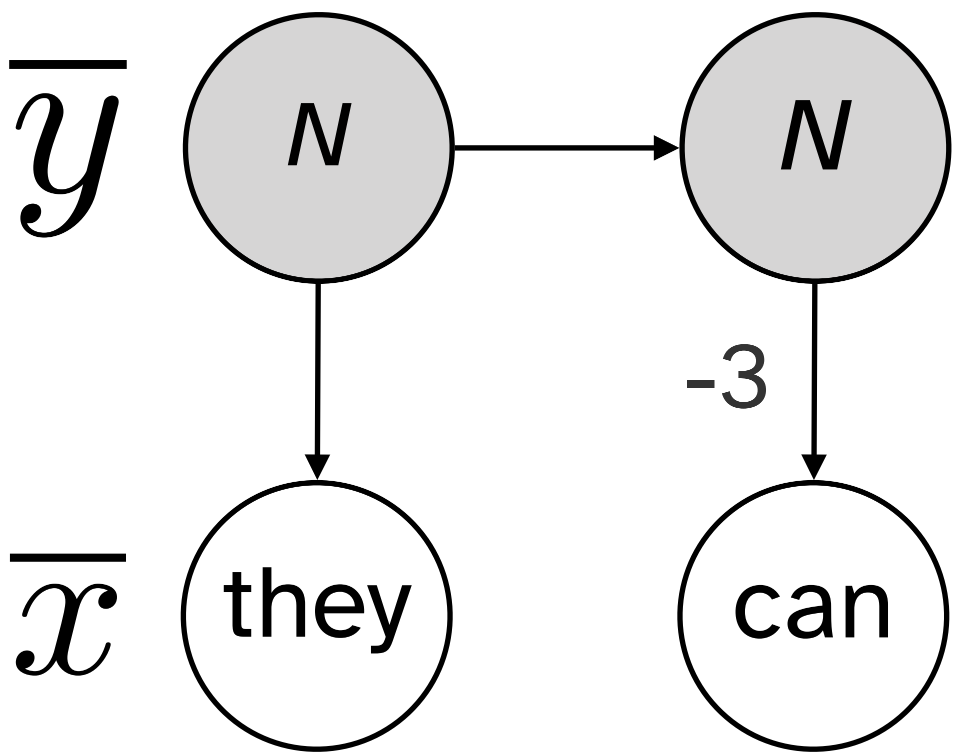
emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}})\right)$$ -4

$\tilde{y}_{\text{prev}} = \text{N}$

v	1	2		
N	-2 ←	-7		
V	-4			



$i = 2$

$\tilde{y} = \text{N}$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

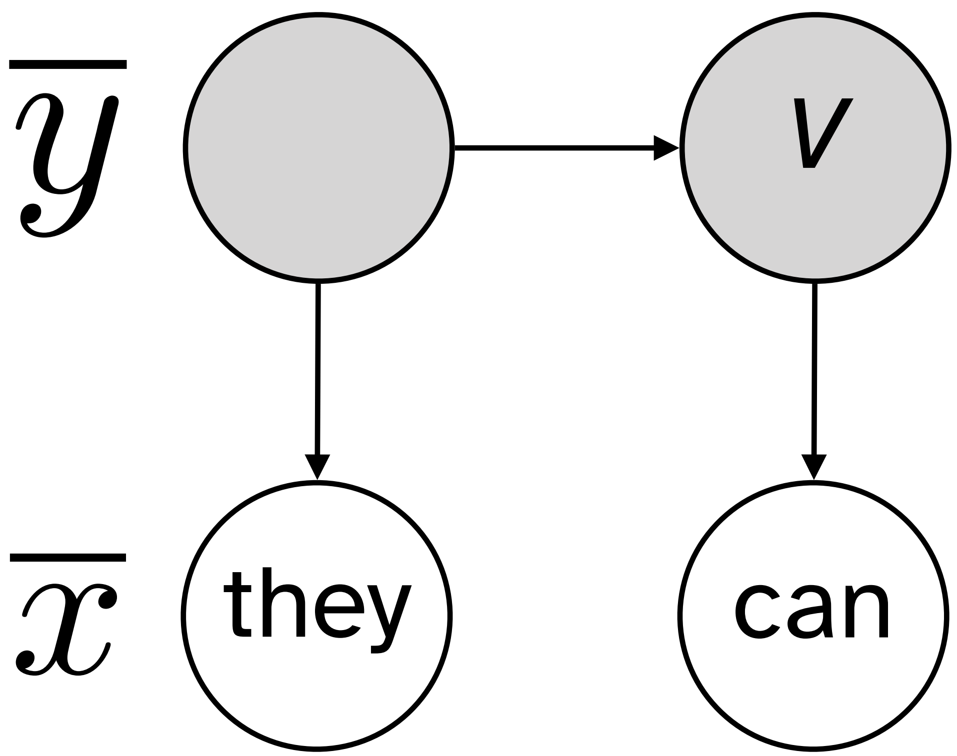
	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}})\right)$$

v	1	2		
N	-2 ←	-7		
V	-4			



$i = 2$

$\tilde{y} = V$

HMM Inference

start logprobs

N	-1
V	-1

transition logprobs

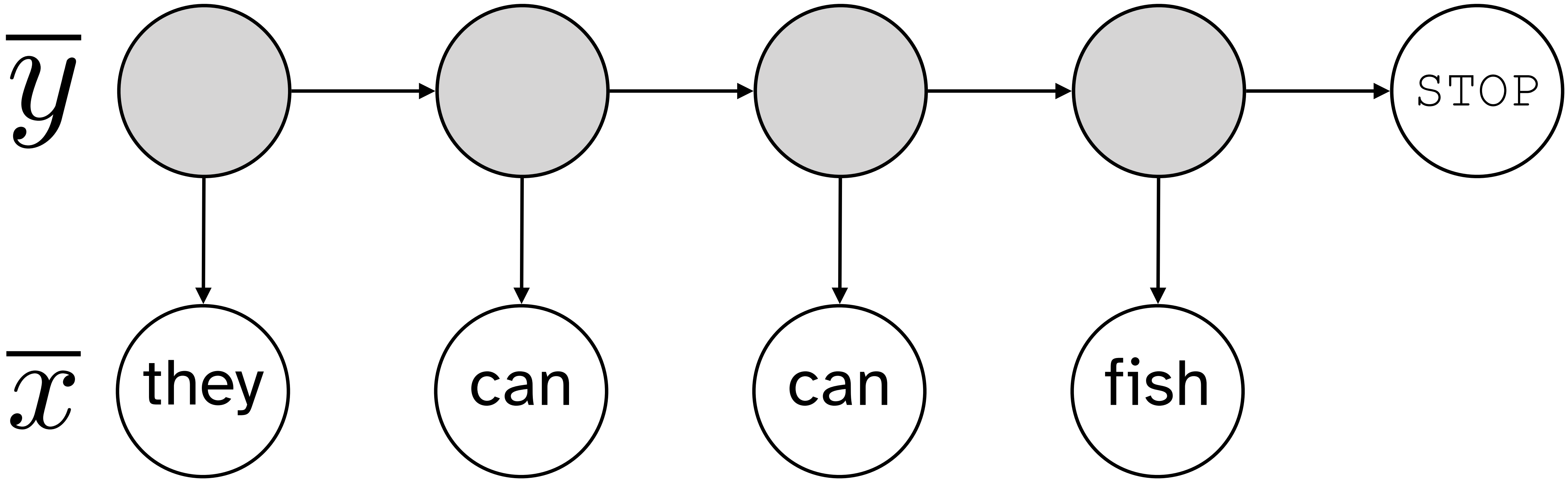
	N	V	**STOP**
N	-2	-1	-1
V	-1	-1	-2

emission logprobs

	they	can	fish
N	-1	-3	-1
V	-3	-1	-1

$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}})\right)$$

v	1	2		
N	-2	-7		
V	-4			



Unsupervised Tagging

- Our parameters were determined by counting over a labeled dataset:
  - How often the first token has a particular POS tag
  - How often one POS tag follows another (including STOP)
  - How often each POS tag “generates” each wordtype
- But what if we don’t have labeled data?

Motivating Example

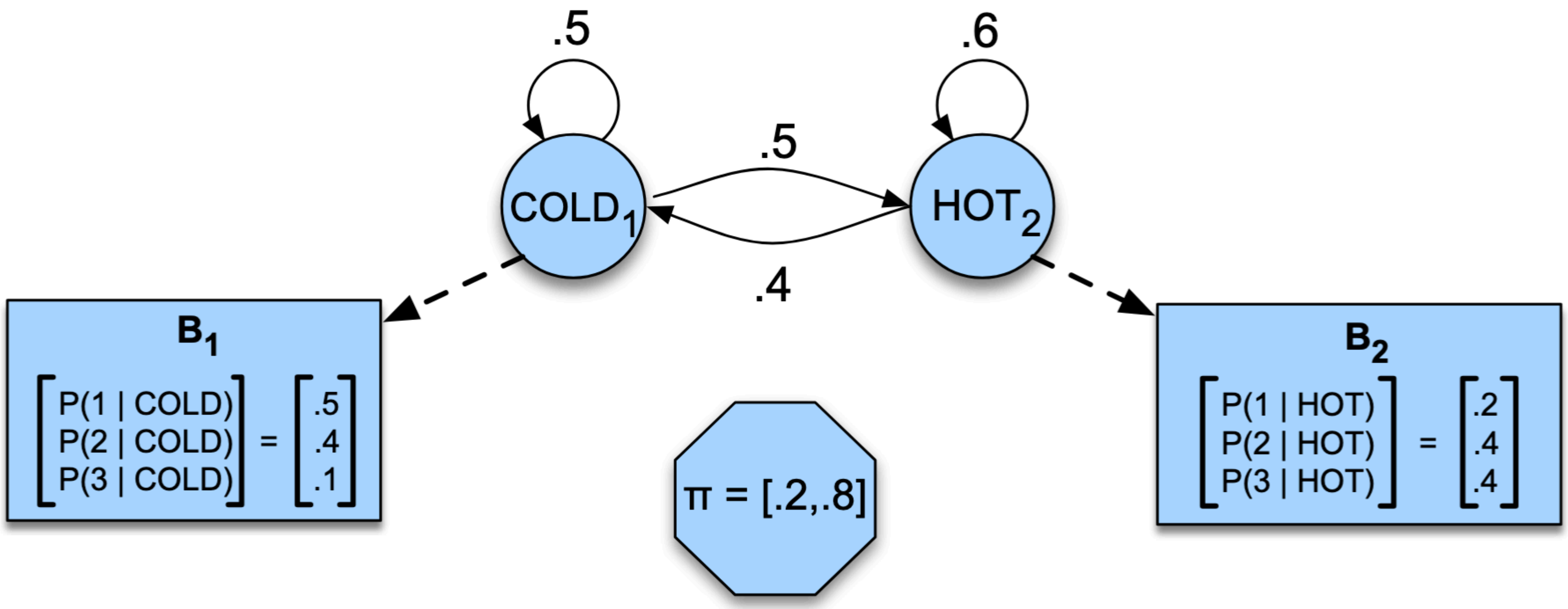
- You are a climatologist in the year 2799 studying the history of global warming
- You cannot find any weather records for Baltimore, MD in 2020 but you do have Jason Eisner's diary, which lists how many ice creams he ate per day
- Can we estimate the daily temperature using just this information?

Example from Jason Eisner, in Jurafsky and Martin

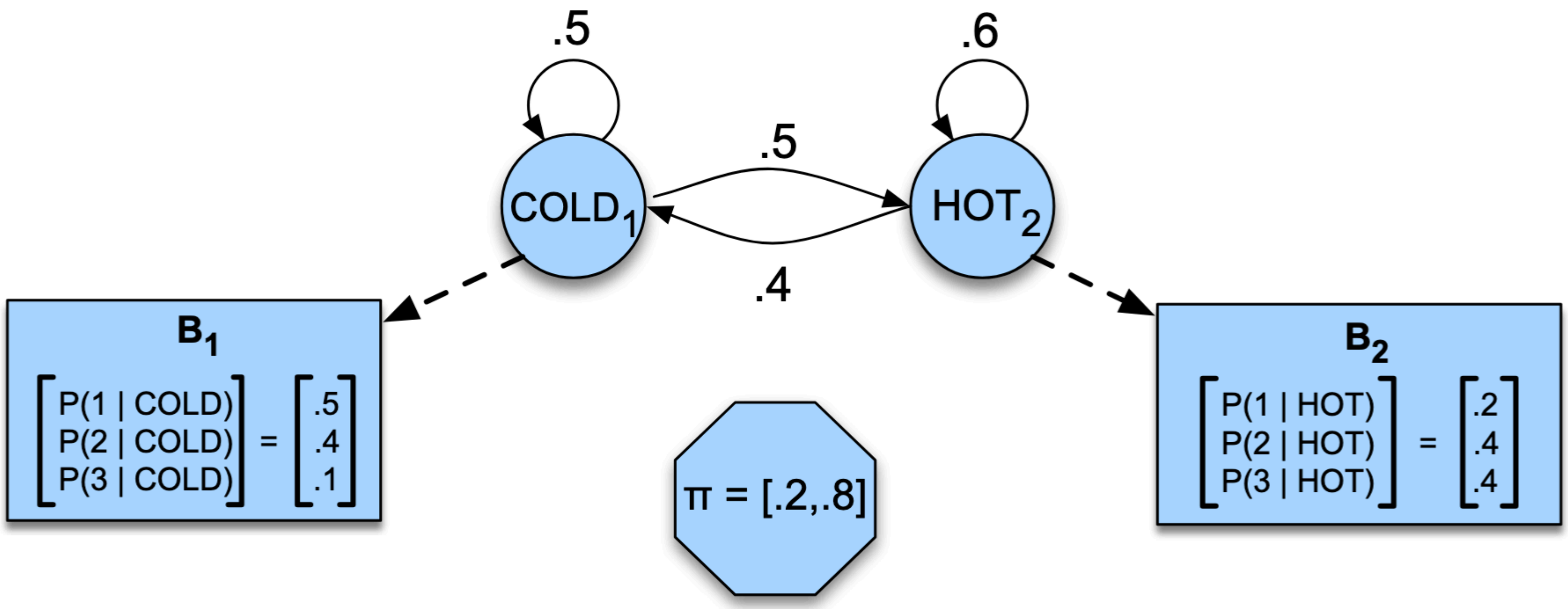
HMM Formulation

- **States:** set of possible underlying states (**hot** or **cold**)
- **Observations:** set of possible outcomes (**1**, **2**, or **3** ice creams)
- HMM is parameterized by:
  - **Initial distribution:** probability distribution over first state's value
  - **Transition probabilities:** probability of moving from one underlying state to another
  - **Emission probabilities:** probability of observing a particular outcome given an underlying state

HMM Formulation

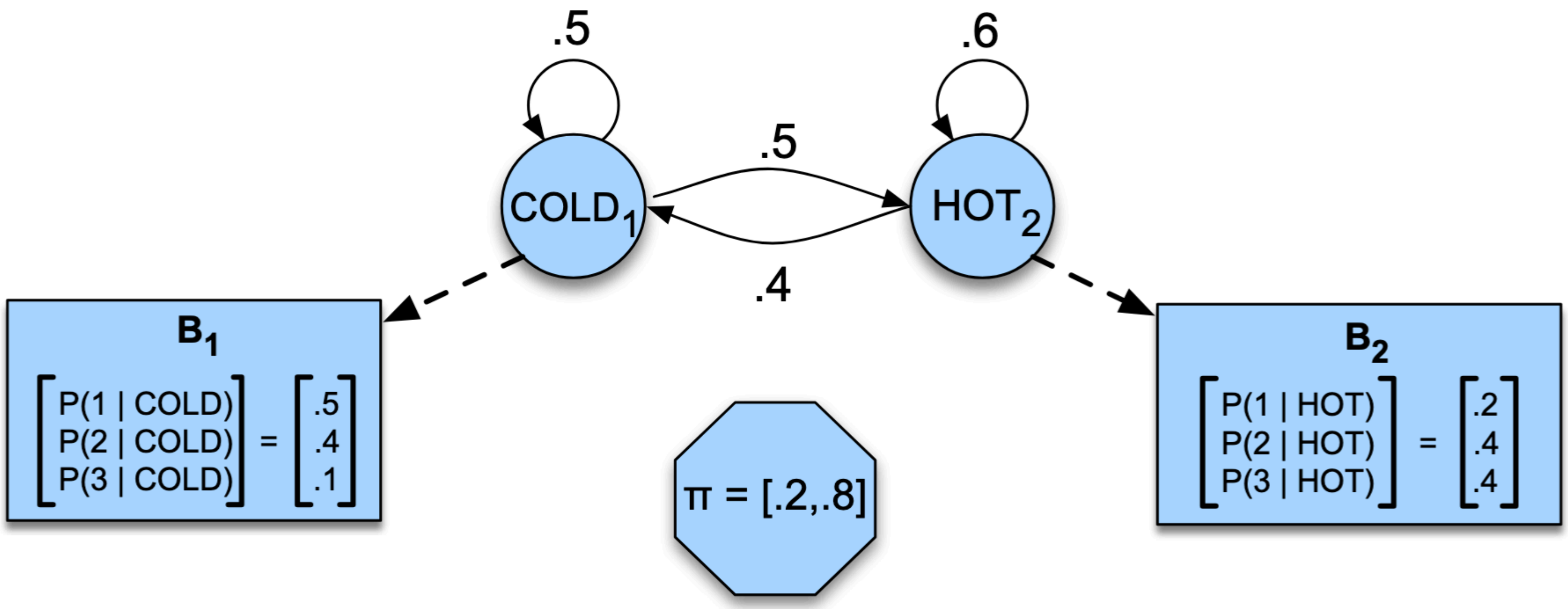


HMM Problems



1. **Likelihood:** given a sequence of observations, how likely is that sequence according to the HMM?
2. **Decoding:** given a sequence of observations, what's the most likely sequence of hidden states?
3. **Learning:** given a sequence of observations and the set of possible states, what are the HMM parameters that maximize the probability of the sequence?

With Labels?



1. **Likelihood:** given a sequence of observations, how likely is that sequence according to the HMM?
   Easy — just compute using our decomposition of P(x, y)
2. **Decoding:** given a sequence of observations, what's the most likely sequence of hidden states?
   Easy — use Viterbi
3. **Learning:** given a sequence of observations and the set of possible states, what are the HMM parameters that maximize the probability of the sequence?
   Easy — compute using counts in data

Example from Jason Eisner, in Jurafsky and Martin

Computing Likelihood

Given a sequence of observations $\overline{x}$, we want to compute $p(\overline{x})$

Computing Likelihood

Given a sequence of observations $\overline{x}$, we want to compute $p(\overline{x})$

- If we had access to the sequence of labels $\overline{y}$, this is just the product of the emission probabilities:

$$p(\overline{x}) = \prod_{i=1}^{|\overline{x}|} p_e(x_i \mid y_i)$$

Computing Likelihood

Given a sequence of observations $\overline{x}$, we want to compute $p(\overline{x})$

- If we had access to the sequence of labels $\overline{y}$, this is just the product of the emission probabilities:

$$p(\overline{x}) = \prod_{i=1}^{|\overline{x}|} p_e(x_i \mid y_i)$$

- But we don't actually know the labels! So we need to consider all possible latent label sequences $\overline{y}'$:

$$p(\overline{x}, \overline{y}') = p(\overline{x} \mid \overline{y}')p(\overline{y}') = \prod_{i=1}^{|\overline{x}|} p_e(x_i \mid y_i) \prod_{i=1}^{|\overline{x}|} p_t(y_i' \mid y_{i-1}')$$

$p_t(y_1' \mid y_0') = p_i(y')$

Computing Likelihood

Given a sequence of observations $\overline{x}$, we want to compute $p(\overline{x})$

- If we had access to the sequence of labels $\overline{y}$, this is just the product of the emission probabilities:

$$p(\overline{x}) = \prod_{i=1}^{|\overline{x}|} p_e(x_i \mid y_i)$$

- But we don't actually know the labels! So we need to consider all possible latent label sequences $\overline{y}'$:

$$p(\overline{x}, \overline{y}') = p(\overline{x} \mid \overline{y}')p(\overline{y}') = \prod_{i=1}^{|\overline{x}|} p_e(x_i \mid y_i) \prod_{i=1}^{|\overline{x}|} p_t(y_i' \mid y_{i-1}')$$

$$p_t(y_1' \mid y_0') = p_i(y')$$

- Then, $p(\overline{x}) = \sum_{\overline{y}' \in \mathcal{Y}} p(\overline{x}, \overline{y}') = \sum_{\overline{y}' \in \mathcal{Y}} p(\overline{x} \mid \overline{y}')p(\overline{y}')$

Exponential...

Forward Algorithm

- Also dynamic programming — similar to Viterbi but we are trying to find the sum over possible latent sequences rather than the maximum

$$v_1(\tilde{y}) = \log P(x_1 \mid \tilde{y}) + \log P(\tilde{y})$$
$$v_i(\tilde{y}) = \log P(x_i \mid \tilde{y}) + \max_{\tilde{y}_{\text{prev}}} \left(\log P(\tilde{y} \mid \tilde{y}_{\text{prev}}) + v_{i-1}(\tilde{y}_{\text{prev}})\right)$$

Forward Algorithm

- Also dynamic programming — similar to Viterbi but we are trying to find the sum over possible latent sequences rather than the maximum

$$\alpha \in \mathbb{R}_{0:1}^{|\overline{x}| \times |\mathcal{S}|}$$

$$\alpha_1(s) = p_i(s) p(x_1 \mid s)$$

$$\alpha_i(s) = \sum_{s' \in \mathcal{S}} \alpha_{i-1}(s') p_t(s \mid s') p_e(x_i \mid s)$$

Forward Algorithm

- Also dynamic programming — similar to Viterbi but we are trying to find the sum over possible latent sequences rather than the maximum

To compute $p(\overline{x})$:

```
alpha = $\mathbf{0}^{|\overline{x}| \times |\mathcal{S}|}$
for i = 1 ... $|\overline{x}|$
  for s in states:
    if i == 1:
      alpha[i, s] = $p_i(s)p(x_1 \mid s)$
    else:
      for s' in states:
        alpha[i, s] += alpha[i-1, s'] $p_t(s \mid s')p_e(x_i \mid s)$

return sum([alpha[$|\overline{x}|$, s] for s in states])
```

Forward-Backward and Viterbi

1. **Likelihood:** given a sequence of observations, how likely is that sequence according to the HMM?
   Use forward algorithm

2. **Decoding:** given a sequence of observations, what's the most likely sequence of hidden states?
   Use Viterbi

3. **Learning:** given a sequence of observations and the set of possible states, what are the HMM parameters that maximize the probability of the sequence?
   ?

Unsupervised Learning

- **Learning algorithm:** forward-backward, aka Baum-Welch
- **Intuition:**
  - Parameters are computed using counts of co-occurrences of hidden states and observations

$$p_t(s \mid s') = \frac{C(s \to s')}{C(s)} \qquad p_e(x \mid s) = \frac{C(s, x)}{C(s)}$$

  - But, we don't have actual counts
- Expectation-maximization (EM) approximation:
  - **E-step:** let's assume our current transition and emission probabilities are correct. If they were, what would our counts be?
  - **M-step:** let's assume our current counts are correct. Then let's compute the empirical transition and emission probabilities as above.

Backward Algorithm

Helper function: backward algorithm for computing $p(\overline{x})$

- Also dynamic programming, where we are building up probability trellis $\beta \in \mathbb{R}_{0:1}^{|\overline{x}| \times |\mathcal{S}|}$
- Base case: $\beta_{|\overline{x}|}(s) = 1$
- Recurrence relation: $\beta_i(s) = \sum_{s' \in \mathcal{S}} p_t(s \mid s') p_e(x_i \mid s) \beta_{i+1}(s')$
- Termination: $p(\overline{x}) = \sum_{s \in \mathcal{S}} p_i(s) p_e(x_1 \mid s) \beta_1(s)$

Estimating Transition Probabilities

- Recall our original estimation of transition probabilities from counts of occurrences:

$$p_t(s \mid s') = \frac{C(s \to s')}{C(s)}$$

- First, let's try to estimate $C(s \to s')$

- Let's say we actually know the probability of a particular transition occurring in some timestep *i*

- Then, we could sum over all timesteps to estimate the total count for this transition

Estimating Transition Probabilities

- The probability of a particular transition occurring in some timestep *i*:

$$\xi_i(s, s') = p(s_t = s, s_{t+1} = s' \mid \overline{x})$$

- Joint probability with the observation sequence:

$$\hat{\xi}_i(s, s') = p(s_t = s, s_{t+1} = s', \overline{x})$$

- How to compute this? Use forward-backward algorithms

$$\hat{\xi}_i(s, s') = p_t(s, s') p_e(x_i \mid s) \alpha_i(s) \beta_{i+1}(s')$$

probability of the transition

probability of the observation

probability of being in current state given all possible previous latent states

probability of being in next state given all possible future latent states

Estimating Transition Probabilities

$$\xi_i(s, s') = p(s_t = s, s_{t+1} = s' \mid \overline{x})$$

$$\hat{\xi}_i(s, s') = p(s_t = s, s_{t+1} = s', \overline{x})$$

$$\hat{\xi}_i(s, s') = p_t(s, s')p_e(x_i \mid s)\alpha_i(s)\beta_{i+1}(s')$$

$$\xi_i(s, s') = \frac{p(s_t = s, s_{t+1} = s', \overline{x})}{p(\overline{x})} = \frac{\hat{\xi}_i(s, s')}{p(\overline{x})}$$

$$= \frac{\hat{\xi}_i(s, s')}{\sum_{\tilde{s} \in \mathcal{S}} \sum_{\tilde{s}' \in \mathcal{S}} \hat{\xi}_i(\tilde{s}, \tilde{s}')}$$

sum over all other possible transitions at this timestep

$$= \frac{p_t(s, s')p_e(x_i \mid s)\alpha_i(s)\beta_{i+1}(s')}{\sum_{\tilde{s} \in \mathcal{S}} \sum_{\tilde{s}' \in \mathcal{S}} p_t(\tilde{s}, \tilde{s}')p_e(x_i \mid \tilde{s})\alpha_i(\tilde{s})\beta_{i+1}(\tilde{s}')}$$

Estimating Transition Probabilities

$$C(s \to s') \approx \sum_{i=1}^{|\overline{x}|} \xi_i(s, s')$$

$$C(s) \approx \sum_{s \in \mathcal{S}} \sum_{i=1}^{|\overline{x}|} \xi_i(s, s')$$

$$\hat{p}_t(s \mid s') = \frac{\sum_{i=1}^{|\overline{x}|} \xi_i(s, s')}{\sum_{s \in \mathcal{S}} \sum_{i=1}^{|\overline{x}|} \xi_i(s, s')}$$

Estimating Emission Probabilities

- Recall our original estimation of transition probabilities from counts of occurrences:

$$p_e(x \mid s) = \frac{C(s, x)}{C(s)}$$

- Let's say we actually know the probability of being in state *s* at time *i*:

$$\gamma_i(s) = p(s_i = s \mid \overline{x}) = \frac{p(s_i = s, \overline{x})}{p(\overline{x})} = \frac{\alpha_i(s)\beta_i(s)}{p(\overline{x})}$$

Estimating Emission Probabilities

$$p_e(x \mid s) = \frac{C(s, x)}{C(s)}$$

$$\gamma_i(s) = p(s_i = s \mid \overline{x}) = \frac{p(s_i = s, \overline{x})}{p(\overline{x})} = \frac{\alpha_i(s)\beta_i(s)}{p(\overline{x})}$$

$$\hat{p}_e(x \mid s) = \frac{\sum_{i=1, x_i = x}^{|\overline{x}|} \gamma_i(s)}{\sum_{i=1}^{|\overline{x}|} \gamma_i(s)}$$

Putting it Together via Expectation-Maximization

initialize $\hat{p}_t(s \mid s')$ and $\hat{p}_e(x \mid s)$

until convergence, iterate over examples $\overline{x}$:

E-step: *assuming probabilities are correct, compute pseudocounts*

compute α and β for $\overline{x}$ using current probs

compute temporary counts

$$\gamma_i(s) = \frac{\alpha_i(s)\beta_i(s)}{\sum_{s' \in \mathcal{S}} \alpha_i(s)\beta_i(s')}$$

$$\xi_i(s, s') = \frac{\hat{p}_t(s, s')\hat{p}_e(x_i \mid s)\alpha_i(s)\beta_{i+1}(s')}{\sum_{\tilde{s} \in \mathcal{S}} \sum_{\tilde{s}' \in \mathcal{S}} \hat{p}_t(\tilde{s}, \tilde{s}')\hat{p}_e(x_i \mid \tilde{s})\alpha_i(\tilde{s})\beta_{i+1}(\tilde{s}')}$$

M-step: *assuming counts are correct, recompute probabilities*

$$\hat{p}_t(s \mid s') = \frac{\sum_{i=1}^{|\overline{x}|} \xi_i(s, s')}{\sum_{s \in \mathcal{S}} \sum_{i=1}^{|\overline{x}|} \xi_i(s, s')} \qquad \hat{p}_e(x \mid s) = \frac{\sum_{i=1, x_i = x}^{|\overline{x}|} \gamma_i(s)}{\sum_{i=1}^{|\overline{x}|} \gamma_i(s)}$$