COS 484/584

Expectation Maximization

Spring 2021

(some slides adapted from Regina Barzilay and Michael Collins)

Midterm

- Logistics announced on Canvas
  - March 10, 12pm ET - March 11, 12pm ET
- Please fill out the survey on your preferred time for taking the exam so we can better plan email support
- Midterm review: COS 484 precept this week (March 5)
  - TAs have posted a survey on Canvas - please fill it out if you'd like them to review specific topics

Expectation Maximization

- If we have **partially observable data**, x_i examples only, then

$$L(\theta) = \sum_i \log \sum_{y \in \mathcal{Y}} P(x_i, y \mid \theta)$$

- The EM (Expectation Maximization) algorithm is a method for finding

$$\theta_{MLE} = \arg\max_\theta L(\theta) = \arg\max_\theta \sum_i \log \sum_{y \in \mathcal{Y}} P(x_i, y \mid \theta)$$

The three coins example

The three coins example

- In the three coins example,

  $\mathcal{Y} = \{H, T\}$ (possible outcomes of coin 0)

  $\mathcal{X} = \{HHH, TTT, HTT, THH, HHT, TTH, HTH, THT\}$

  $\theta = \{\lambda, p_1, p_2\}$

The three coins example

- In the three coins example,

  $\mathcal{Y} = \{H, T\}$ (possible outcomes of coin 0)

  $\mathcal{X} = \{HHH, TTT, HTT, THH, HHT, TTH, HTH, THT\}$

  $\theta = \{\lambda, p_1, p_2\}$

(all possible observations of length 3)

The three coins example

- In the three coins example,

  $\mathcal{Y} = \{H, T\}$ (possible outcomes of coin 0)

  $\mathcal{X} = \{HHH, TTT, HTT, THH, HHT, TTH, HTH, THT\}$ (all possible observations of length 3)

  $\theta = \{\lambda, p_1, p_2\}$

- and $P(x, y \mid \theta) = P(y \mid \theta)\ P(x \mid y, \theta)$

  where

$$P(y \mid \theta) = \begin{cases} \lambda \text{ if } y = H \\ 1 - \lambda \text{ if } y = T \end{cases}$$

  and

$$P(x \mid y, \theta) = \begin{cases} p_1^h\ (1 - p_1)^t \text{ if } y = H \\ p_2^h\ (1 - p_2)^t \text{ if } y = T \end{cases}$$

The three coins example

The three coins example

- Calculating various probabilities:

$$P(x = THT, y = H \mid \theta) = \lambda p_1 (1 - p_1)^2$$

$$P(x = THT, y = T \mid \theta) = (1 - \lambda) p_2 (1 - p_2)^2$$

The three coins example

- Calculating various probabilities:

  $$P(x = THT, y = H \mid \theta) = \lambda p_1(1 - p_1)^2$$

  $$P(x = THT, y = T \mid \theta) = (1 - \lambda)p_2(1 - p_2)^2$$

$$\begin{aligned} P(x = THT \mid \theta) &= P(x = THT, y = H \mid \theta) + P(x = THT, y = T \mid \theta) \\ &= \lambda p_1(1 - p_1)^2 + (1 - \lambda)p_2(1 - p_2)^2 \end{aligned}$$

$$\begin{aligned} P(y = H \mid x = THT, \theta) &= \frac{P(x = THT, y = H \mid \theta)}{P(x = THT \mid \theta)} \\ &= \frac{\lambda p_1(1 - p_1)^2}{\lambda p_1(1 - p_1)^2 + (1 - \lambda)p_2(1 - p_2)^2} \end{aligned}$$

The three coins example

- New estimates:

$$(\langle \mathtt{HHH} \rangle, H) \quad P(y = \mathtt{H} \mid \mathtt{HHH}) = 0.0508$$
$$(\langle \mathtt{HHH} \rangle, T) \quad P(y = \mathtt{T} \mid \mathtt{HHH}) = 0.9492$$
$$(\langle \mathtt{TTT} \rangle, H) \quad P(y = \mathtt{H} \mid \mathtt{TTT}) = 0.6967$$
$$(\langle \mathtt{TTT} \rangle, T) \quad P(y = \mathtt{T} \mid \mathtt{TTT}) = 0.3033$$
$$(\langle \mathtt{HHH} \rangle, H) \quad P(y = \mathtt{H} \mid \mathtt{HHH}) = 0.0508$$
$$(\langle \mathtt{HHH} \rangle, T) \quad P(y = \mathtt{T} \mid \mathtt{HHH}) = 0.9492$$
$$(\langle \mathtt{TTT} \rangle, H) \quad P(y = \mathtt{H} \mid \mathtt{TTT}) = 0.6967$$
$$(\langle \mathtt{TTT} \rangle, T) \quad P(y = \mathtt{T} \mid \mathtt{TTT}) = 0.3033$$
$$(\langle \mathtt{HHH} \rangle, H) \quad P(y = \mathtt{H} \mid \mathtt{HHH}) = 0.0508$$
$$(\langle \mathtt{HHH} \rangle, T) \quad P(y = \mathtt{T} \mid \mathtt{HHH}) = 0.9492$$

$$\lambda = \frac{3 \times 0.0508 + 2 \times 0.6967}{5} = 0.3092$$

$$p_1 = \frac{3 \times 3 \times 0.0508 + 0 \times 2 \times 0.6967}{3 \times 3 \times 0.0508 + 3 \times 2 \times 0.6967} = 0.0987$$

$$p_2 = \frac{3 \times 3 \times 0.9492 + 0 \times 2 \times 0.3033}{3 \times 3 \times 0.9492 + 3 \times 2 \times 0.3033} = 0.8244$$

The three coins example

$$(\langle \text{HHH} \rangle, H) \quad P(y = \text{H} \mid \text{HHH}) = 0.0508$$
$$(\langle \text{HHH} \rangle, T) \quad P(y = \text{T} \mid \text{HHH}) = 0.9492$$
$$(\langle \text{TTT} \rangle, H) \quad P(y = \text{H} \mid \text{TTT}) = 0.6967$$
$$(\langle \text{TTT} \rangle, T) \quad P(y = \text{T} \mid \text{TTT}) = 0.3033$$
$$(\langle \text{HHH} \rangle, H) \quad P(y = \text{H} \mid \text{HHH}) = 0.0508$$
$$(\langle \text{HHH} \rangle, T) \quad P(y = \text{T} \mid \text{HHH}) = 0.9492$$
$$(\langle \text{TTT} \rangle, H) \quad P(y = \text{H} \mid \text{TTT}) = 0.6967$$
$$(\langle \text{TTT} \rangle, T) \quad P(y = \text{T} \mid \text{TTT}) = 0.3033$$
$$(\langle \text{HHH} \rangle, H) \quad P(y = \text{H} \mid \text{HHH}) = 0.0508$$
$$(\langle \text{HHH} \rangle, T) \quad P(y = \text{T} \mid \text{HHH}) = 0.9492$$

- New estimates:

$$\lambda = \frac{3 \times 0.0508 + 2 \times 0.6967}{5} = 0.3092$$

$$p_1 = \frac{3 \times 3 \times 0.0508 + 0 \times 2 \times 0.6967}{3 \times 3 \times 0.0508 + 3 \times 2 \times 0.6967} = 0.0987$$

$$p_2 = \frac{3 \times 3 \times 0.9492 + 0 \times 2 \times 0.3033}{3 \times 3 \times 0.9492 + 3 \times 2 \times 0.3033} = 0.8244$$

Summary

Summary

- Begin with parameters: $\lambda = 0.3,\ p_1 = 0.3,\ p_2 = 0.6$

Summary

- Begin with parameters: $\lambda = 0.3,\ p_1 = 0.3,\ p_2 = 0.6$

- Fill in hidden variables, using

  $P(y = H \,|\, x = \langle HHH \rangle) = 0.0508$

  $P(y = H \,|\, x = \langle TTT \rangle) = 0.6967$

Summary

- Begin with parameters: $\lambda = 0.3,\ p_1 = 0.3,\ p_2 = 0.6$

- Fill in hidden variables, using

  $P(y = H \mid x = \langle HHH \rangle) = 0.0508$

  $P(y = H \mid x = \langle TTT \rangle) = 0.6967$

- This gives us a pseudo-annotated dataset with **fractional** counts

Summary

- Begin with parameters: $\lambda = 0.3,\ p_1 = 0.3,\ p_2 = 0.6$

- Fill in hidden variables, using

  $P(y = H \mid x = \langle HHH \rangle) = 0.0508$

  $P(y = H \mid x = \langle TTT \rangle) = 0.6967$

- This gives us a pseudo-annotated dataset with **fractional** counts

- Re-estimate parameters to be

  $\lambda = 0.3092,\ p_1 = 0.0987,\ p_2 = 0.8244$

Summary

- Begin with parameters: $\lambda = 0.3,\ p_1 = 0.3,\ p_2 = 0.6$

- Fill in hidden variables, using

  $P(y = H \mid x = \langle HHH \rangle) = 0.0508$

  $P(y = H \mid x = \langle TTT \rangle) = 0.6967$

- This gives us a pseudo-annotated dataset with **fractional** counts

Repeat!

- Re-estimate parameters to be

  $\lambda = 0.3092,\ p_1 = 0.0987,\ p_2 = 0.8244$

EM iterations (example 1)

$P(y = H | x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967
1	0.3738	0.0680	0.7578	0.0004	0.9714	0.0004	0.9714
2	0.4859	0.0004	0.9722	0.0000	1.0000	0.0000	1.0000
3	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

The coin example for $x = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. The solution that EM reaches is intuitively correct: the coin tosser has two coins, one which always shows heads, and another which always shows tails, and is picking between them with equal probability ($\lambda = 0.5$) using coin 0.

EM iterations (example 1)

$P(y = H | x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967
1	0.3738	0.0680	0.7578	0.0004	0.9714	0.0004	0.9714
2	0.4859	0.0004	0.9722	0.0000	1.0000	0.0000	1.0000
3	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

The coin example for $x = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. The solution that EM reaches is intuitively correct: the coin tosser has two coins, one which always shows heads, and another which always shows tails, and is picking between them with equal probability ($\lambda = 0.5$) using coin 0.

EM iterations (example 1)

$P(y = H | x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967
1	0.3738	0.0680	0.7578	0.0004	0.9714	0.0004	0.9714
2	0.4859	0.0004	0.9722	0.0000	1.0000	0.0000	1.0000
3	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

The coin example for $x = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. The solution that EM reaches is intuitively correct: the coin tosser has two coins, one which always shows heads, and another which always shows tails, and is picking between them with equal probability ($\lambda = 0.5$) using coin 0.

Posterior probabilities $\bar{p}_i$ show that we are certain that coin 1 (tail-biased) generated x_2 and x_4, whereas coin 2 generated x_1 and x_3

EM iterations (example 2)

$P(y = H \mid x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$	$\tilde{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508

Coin example for $\{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle\}$

EM iterations (example 2)

$P(y = H \mid x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$	$\tilde{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508

Coin example for $\{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle\}$

EM iterations (example 2)

$P(y = H | x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$	$\tilde{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508

Coin example for $\{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle\}$

Which of these would you expect EM to converge to?

A) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

B) $\lambda = 0.5,\ p_1 = 1,\ p_2 = 0$

C) $\lambda = 0.4,\ p_1 = 0,\ p_2 = 1$

EM iterations (example 2)

$P(y = H | x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$	$\tilde{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508
1	0.3092	0.0987	0.8244	0.0008	0.9837	0.0008	0.9837	0.0008
2	0.3940	0.0012	0.9893	0.0000	1.0000	0.0000	1.0000	0.0000
3	0.4000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Coin example for $\{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle\}$

Which of these would you expect EM to converge to?

A) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

B) $\lambda = 0.5,\ p_1 = 1,\ p_2 = 0$

C) $\lambda = 0.4,\ p_1 = 0,\ p_2 = 1$

EM iterations (example 2)

$P(y = H | x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$	$\tilde{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508
1	0.3092	0.0987	0.8244	0.0008	0.9837	0.0008	0.9837	0.0008
2	0.3940	0.0012	0.9893	0.0000	1.0000	0.0000	1.0000	0.0000
3	0.4000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Coin example for $\{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle\}$

Which of these would you expect EM to converge to?

A) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

B) $\lambda = 0.5,\ p_1 = 1,\ p_2 = 0$

C) $\lambda = 0.4,\ p_1 = 0,\ p_2 = 1$

EM iterations (example 2)

$P(y = H | x_1)$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$	$\tilde{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508
1	0.3092	0.0987	0.8244	0.0008	0.9837	0.0008	0.9837	0.0008
2	0.3940	0.0012	0.9893	0.0000	1.0000	0.0000	1.0000	0.0000
3	0.4000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Coin example for $\{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle\}$

Which of these would you expect EM to converge to?

A) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

B) $\lambda = 0.5,\ p_1 = 1,\ p_2 = 0$

C) $\lambda = 0.4,\ p_1 = 0,\ p_2 = 1$

λ is now 0.4, indicating that coin 0 has a probability 0.4 of selecting the tail-biased coin 1

EM iterations (example 3)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967

Coin example for $x = \{\langle HHT \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$.

EM iterations (example 3)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967

Coin example for $x = \{\langle HHT \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$.

EM iterations (example 3)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967

Coin example for $x = \{\langle HHT\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$.

Which of these would you expect EM to converge to?

A) $\lambda = 0.49,\ p_1 = 0.12,\ p_2 = 0$

B) $\lambda = 0.49,\ p_1 = 0,\ p_2 = 0.82$

C) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

EM iterations (example 3)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967
1	0.4005	0.0974	0.6300	0.0375	0.9065	0.0025	0.9065
2	0.4632	0.0148	0.7635	0.0014	0.9842	0.0000	0.9842
3	0.4924	0.0005	0.8205	0.0000	0.9941	0.0000	0.9941
4	0.4970	0.0000	0.8284	0.0000	0.9949	0.0000	0.9949

Coin example for $x = \{\langle HHT \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$.

Which of these would you expect EM to converge to?

A) $\lambda = 0.49,\ p_1 = 0.12,\ p_2 = 0$

B) $\lambda = 0.49,\ p_1 = 0,\ p_2 = 0.82$

C) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

EM iterations (example 3)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967
1	0.4005	0.0974	0.6300	0.0375	0.9065	0.0025	0.9065
2	0.4632	0.0148	0.7635	0.0014	0.9842	0.0000	0.9842
3	0.4924	0.0005	0.8205	0.0000	0.9941	0.0000	0.9941
4	0.4970	0.0000	0.8284	0.0000	0.9949	0.0000	0.9949

Coin example for $x = \{\langle HHT\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$.

Which of these would you expect EM to converge to?

A) $\lambda = 0.49,\ p_1 = 0.12,\ p_2 = 0$

B) $\lambda = 0.49,\ p_1 = 0,\ p_2 = 0.82$

C) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

EM iterations (example 3)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967
1	0.4005	0.0974	0.6300	0.0375	0.9065	0.0025	0.9065
2	0.4632	0.0148	0.7635	0.0014	0.9842	0.0000	0.9842
3	0.4924	0.0005	0.8205	0.0000	0.9941	0.0000	0.9941
4	0.4970	0.0000	0.8284	0.0000	0.9949	0.0000	0.9949

Coin example for $x = \{\langle HHT\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$.

- EM selects a tails-only coin ($p_1 = 0$), and a coin which is heavily heads-biased ($p_2 = 0.8284$).
- It is certain that x_1 and x_3 were generated by coin 2 since they contain heads.
- x_2 and x_4 could have been generated by either coin but coin 1 (tail-biased) is far more likely.

EM iterations (example 4)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7000	0.7000	0.3000	0.3000	0.3000	0.3000

Coin example for $x = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$.

EM iterations (example 4)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7000	0.7000	0.3000	0.3000	0.3000	0.3000

Coin example for $x = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$.

Which of these would you expect EM to converge to?

A) $\lambda = 0.3,\ p_1 = 0.5,\ p_2 = 0.5$

B) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

C) $\lambda = 0.5,\ p_1 = 0,\ p_2 = 1$

EM iterations (example 4)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7000	0.7000	0.3000	0.3000	0.3000	0.3000

Coin example for $x = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$.

Which of these would you expect EM to converge to?

A) $\lambda = 0.3,\ p_1 = 0.5,\ p_2 = 0.5$

B) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

C) $\lambda = 0.5,\ p_1 = 0,\ p_2 = 1$

EM iterations (example 4)

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7000	0.7000	0.3000	0.3000	0.3000	0.3000
1	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
2	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
3	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
4	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
5	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
6	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000

Coin example for $x = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$.

Which of these would you expect EM to converge to?

A) $\lambda = 0.3,\ p_1 = 0.5,\ p_2 = 0.5$

B) $\lambda = 0.5,\ p_1 = 0.5,\ p_2 = 0.5$

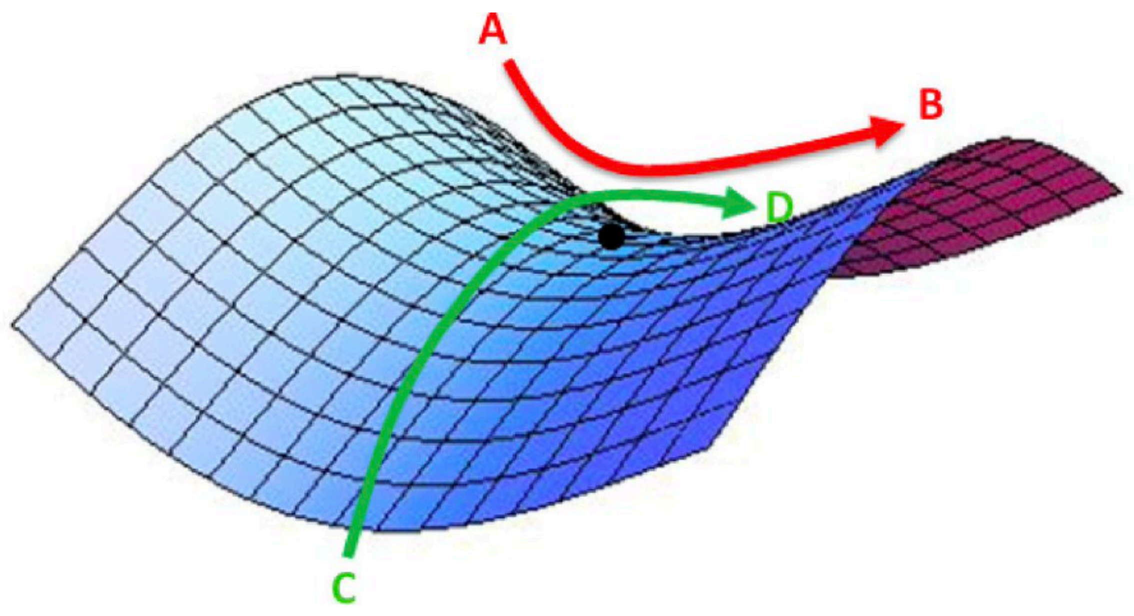
C) $\lambda = 0.5,\ p_1 = 0,\ p_2 = 1$

Initialization matters

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7000	0.7000	0.3000	0.3000	0.3000	0.3000
1	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
2	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
3	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
4	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
5	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
6	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000

Coin example for $x = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$.

In this case, EM is stuck at a **saddle point**.



Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7001	0.7000	0.3001	0.2998	0.3001	0.2998
1	0.2999	0.5003	0.4999	0.3004	0.2995	0.3004	0.2995
2	0.2999	0.5008	0.4997	0.3013	0.2986	0.3013	0.2986
3	0.2999	0.5023	0.4990	0.3040	0.2959	0.3040	0.2959
4	0.3000	0.5068	0.4971	0.3122	0.2879	0.3122	0.2879
5	0.3000	0.5202	0.4913	0.3373	0.2645	0.3373	0.2645
6	0.3009	0.5605	0.4740	0.4157	0.2007	0.4157	0.2007
7	0.3082	0.6744	0.4223	0.6447	0.0739	0.6447	0.0739
8	0.3593	0.8972	0.2773	0.9500	0.0016	0.9500	0.0016
9	0.4758	0.9983	0.0477	0.9999	0.0000	0.9999	0.0000
10	0.4999	1.0000	0.0001	1.0000	0.0000	1.0000	0.0000
11	0.5000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Coin example for $x = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$.

If we initialize p_1 and p_2 even a small amount away from the saddle point $p_1 = p_2$, EM diverges and eventually reaches the global maximum

The EM algorithm

The EM algorithm

- θ^t is the parameter vector at the t^{th} iteration

The EM algorithm

Superscript for iteration #

- θ^t is the parameter vector at the t^{th} iteration

The EM algorithm

Superscript for iteration #

- θ^t is the parameter vector at the t^{th} iteration

- Choose θ^0 at random (or using smart heuristics)

The EM algorithm

Superscript for iteration #

- θ^t is the parameter vector at the t^{th} iteration
- Choose θ^0 at random (or using smart heuristics)
- Iterative procedure defined as:

$$\theta^t = \arg\max_{\theta} Q(\theta, \theta^{t-1})$$

where

$$Q(\theta, \theta^{t-1}) = \sum_{i} \sum_{y \in \mathcal{Y}} P(y \mid x_i, \theta^{t-1}) \ \log P(x_i, y \mid \theta)$$

The EM algorithm

Superscript for iteration #

- θ^t is the parameter vector at the t^{th} iteration
- Choose θ^0 at random (or using smart heuristics)
- Iterative procedure defined as:

$$\theta^t = \arg\max_{\theta} Q(\theta, \theta^{t-1})$$

where

$$Q(\theta, \theta^{t-1}) = \sum_i \sum_{y \in \mathcal{Y}} P(y \mid x_i, \theta^{t-1}) \ \log P(x_i, y \mid \theta)$$

How did we get $\arg\max_{\theta} Q$ from $\arg\max_{\theta} \sum_i \log \sum_{y \in \mathcal{Y}} P(x_i, y \mid \theta)$? => Jensen's inequality

(advanced; see optional reading from Andrew Ng)

The EM algorithm

- θ^t is the parameter vector at the t^{th} iteration
- Choose θ^0 at random (or using smart heuristics)
- (E step): Compute *expected* counts for every parameter θ_r :

$$\overline{Count}(r) = \sum_{i=1}^{n} \sum_{y} P(y \mid x_i, \theta^{t-1})\ Count(x_i, y, r)$$

The EM algorithm

- θ^t is the parameter vector at the t^{th} iteration
- Choose θ^0 at random (or using smart heuristics)
- (E step): Compute *expected* counts for every parameter θ_r :

$$\overline{Count}(r) = \sum_{i=1}^{n} \sum_{y} P(y \mid x_i, \theta^{t-1}) \; Count(x_i, y, r)$$

- (M step): Re-estimate parameters using expected counts to **maximize likelihood (MLE estimate)**

e.g. $\theta_{DT \rightarrow NN} = \dfrac{\overline{Count}(DT \rightarrow NN)}{\sum_{\beta} \overline{Count}(DT \rightarrow \beta)}$

The EM algorithm

The EM algorithm

- Iterative procedure defined as $\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$ where

$$Q(\theta, \theta^{t-1}) = \sum_i \sum_{y \in \mathcal{Y}} P(y \mid x_i, \theta^{t-1}) \ \log P(x_i, y \mid \theta)$$

The EM algorithm

- Iterative procedure defined as $\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$ where

  $$Q(\theta, \theta^{t-1}) = \sum_i \sum_{y \in \mathcal{Y}} P(y \mid x_i, \theta^{t-1}) \ \log P(x_i, y \mid \theta)$$

- Key points:
  - Intuition: Fill in hidden variables y according to $P(y \mid x_i, \theta)$
  - Create a "pseudo-dataset" with fractional counts
  - EM is guaranteed to converge to a **local** maximum, or saddle-point, of the likelihood function
  - In general, if $\arg\max_\theta \sum_i \log P(x_i, y_i \mid \theta)$ has a simple analytic solution, then

    $$\arg\max_\theta \sum_i \sum_y P(y \mid x_i, \theta) \log P(x_i, y \mid \theta)$$ also has a simple solution.

Example: EM for HMM

Example: EM for HMM

- We observe only word sequences $X_1, X_2, \ldots, X_n$ (no tags Y)

Example: EM for HMM

- We observe only word sequences $X_1, X_2, \ldots, X_n$ (no tags Y)

Each X and Y is a sequence on its own

Example: EM for HMM

- We observe only word sequences $X_1, X_2, \ldots, X_n$ (no tags Y)
- Let θ be the vector of all transition parameters (include initial state distribution as a special case, $\varnothing \to s$)

Each X and Y is a sequence on its own

Example: EM for HMM

- We observe only word sequences $X_1, X_2, \ldots, X_n$ (no tags Y)
- Let θ be the vector of all transition parameters (include initial state distribution as a special case, $\varnothing \rightarrow s$)
- Let ϕ be the vector of all emission parameters

Each X and Y is a sequence on its own

Example: EM for HMM

- We observe only word sequences $X_1, X_2, \ldots, X_n$ (no tags Y)
- Let θ be the vector of all transition parameters (include initial state distribution as a special case, $\varnothing \rightarrow s$)
- Let ϕ be the vector of all emission parameters
- Initialize parameters to some values θ^0 and ϕ^0

Each X and Y is a sequence on its own

Recap: Estimating HMM parameters

Recap: Estimating HMM parameters

Which of these is the correct MLE estimate for the transition parameter $\theta_{a \to b}$ of an HMM (where a, b, b' are states) ?

A) $\theta_{a \to b} = \dfrac{Count(a \to b)}{\sum_{a'} Count(a' \to b)}$

B) $\theta_{a \to b} = \dfrac{Count(a \to b)}{\sum_{b'} Count(a \to b')}$

C) $\theta_{a \to b} = \dfrac{Count(a \to b)}{\sum_{a'} \sum_{b'} Count(a' \to b')}$

Recap: Estimating HMM parameters

Which of these is the correct MLE estimate for the transition parameter $\theta_{a \to b}$ of an HMM (where a, b, b' are states) ?

A) $\theta_{a \to b} = \frac{Count(a \to b)}{\sum_{a'} Count(a' \to b)}$

B) $\theta_{a \to b} = \frac{Count(a \to b)}{\sum_{b'} Count(a \to b')}$

C) $\theta_{a \to b} = \frac{Count(a \to b)}{\sum_{a'} \sum_{b'} Count(a' \to b')}$

Recap: Estimating HMM parameters

Recap: Estimating HMM parameters

- Maximum likelihood estimates:

$$\theta_{a \to b} = \frac{Count(a \to b)}{\sum_{b'} Count(a \to b')}$$ (where a, b are states)

$$\phi_{a \to A} = \frac{Count(a \to A)}{\sum_{A'} Count(a \to A')}$$ (where A is an observation)

Recap: Estimating HMM parameters

- Maximum likelihood estimates:

$$\theta_{a \to b} = \frac{Count(a \to b)}{\sum_{b'} Count(a \to b')}$$ (where a, b are states)

$$\phi_{a \to A} = \frac{Count(a \to A)}{\sum_{A'} Count(a \to A')}$$ (where A is an observation)

- Here, counts are estimated by simply checking for occurrence of the transition/emission in every data sequence

e.g. $Count(a \to b) = \sum_{i=1}^{n} Count(X_i, Y_i, a \to b)$

(number of times the transition occurs in each data point)

Example: EM for HMM

Example: EM for HMM

- Initialize parameters θ^0 and ϕ^0

Example: EM for HMM

- Initialize parameters θ^0 and ϕ^0
- **(E-Step)** Compute **expected** counts

$$\overline{Count}(a \to b) = \sum_{i=1}^{n} \sum_{Y} P(Y|X_i, \theta^{t-1}, \phi^{t-1})\ Count(X_i, Y, a \to b)$$

$$= \sum_{i=1}^{n} \sum_{Y} P(Y|X_i, \theta^{t-1}, \phi^{t-1})\ Count(Y, a \to b)$$

$$\overline{Count}(a \to A) = \sum_{i=1}^{n} \sum_{Y} P(Y|X_i, \theta^{t-1}, \phi^{t-1})\ Count(X_i, Y, a \to A)$$

Example: EM for HMM

- Initialize parameters θ^0 and ϕ^0

- **(E-Step)** Compute **expected** counts

$$\overline{Count}(a \to b) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(X_i, Y, a \to b)$$

$$= \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(Y, a \to b)$$

$$\overline{Count}(a \to A) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(X_i, Y, a \to A)$$

Example: EM for HMM

previous parameters

- Initialize parameters θ^0 and ϕ^0

- **(E-Step)** Compute **expected** counts

$$\overline{Count}(a \rightarrow b) = \sum_{i=1}^{n} \sum_{Y} \boxed{P(Y | X_i, \theta^{t-1}, \phi^{t-1})} \; Count(X_i, Y, a \rightarrow b)$$

$$= \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(Y, a \rightarrow b)$$

$$\overline{Count}(a \rightarrow A) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(X_i, Y, a \rightarrow A)$$

Example: EM for HMM

- Initialize parameters θ^0 and ϕ^0

previous parameters

- **(E-Step)** Compute **expected** counts

$$\overline{Count}(a \rightarrow b) = \sum_{i=1}^{n} \sum_{Y} \boxed{P(Y|X_i, \theta^{t-1}, \phi^{t-1})} \; Count(X_i, Y, a \rightarrow b)$$

$$= \sum_{i=1}^{n} \sum_{Y} P(Y|X_i, \theta^{t-1}, \phi^{t-1}) \; Count(Y, a \rightarrow b)$$

(why?)

$$\overline{Count}(a \rightarrow A) = \sum_{i=1}^{n} \sum_{Y} P(Y|X_i, \theta^{t-1}, \phi^{t-1}) \; Count(X_i, Y, a \rightarrow A)$$

Example: EM for HMM

previous parameters

- Initialize parameters θ^0 and ϕ^0
- **(E-Step)** Compute **expected** counts

$$\overline{Count}(a \rightarrow b) = \sum_{i=1}^{n} \sum_{Y} P(Y \mid X_i, \theta^{t-1}, \phi^{t-1}) \; Count(X_i, Y, a \rightarrow b)$$

$$= \sum_{i=1}^{n} \sum_{Y} P(Y \mid X_i, \theta^{t-1}, \phi^{t-1}) \; Count(Y, a \rightarrow b)$$

(why?)

$$\overline{Count}(a \rightarrow A) = \sum_{i=1}^{n} \sum_{Y} P(Y \mid X_i, \theta^{t-1}, \phi^{t-1}) \; Count(X_i, Y, a \rightarrow A)$$

#times transition appears in Y

Example: EM for HMM

- **(M-Step)**

$$\theta^t_{a\to b} = \frac{\overline{Count}(a \to b)}{\sum_{a\to b'} \overline{Count}(a \to b')}$$

$$\phi^t_{a\to A} = \frac{\overline{Count}(a \to A)}{\sum_{a\to A'} \overline{Count}(a \to A')}$$

Example: EM for HMM

- **(M-Step)**

$$\theta^t_{a\to b} = \frac{\overline{Count}(a \to b)}{\sum_{a\to b'} \overline{Count}(a \to b')}$$

$$\phi^t_{a\to A} = \frac{\overline{Count}(a \to A)}{\sum_{a\to A'} \overline{Count}(a \to A')}$$

Very similar to the MLE update we saw for HMMs

- **(E-Step)** Compute **expected** counts

$$\overline{Count}(a \rightarrow b) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(X_i, Y, a \rightarrow b)$$

$$= \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(Y, a \rightarrow b)$$

$$\overline{Count}(a \rightarrow A) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(X_i, Y, a \rightarrow A)$$

- **(E-Step)** Compute **expected** counts

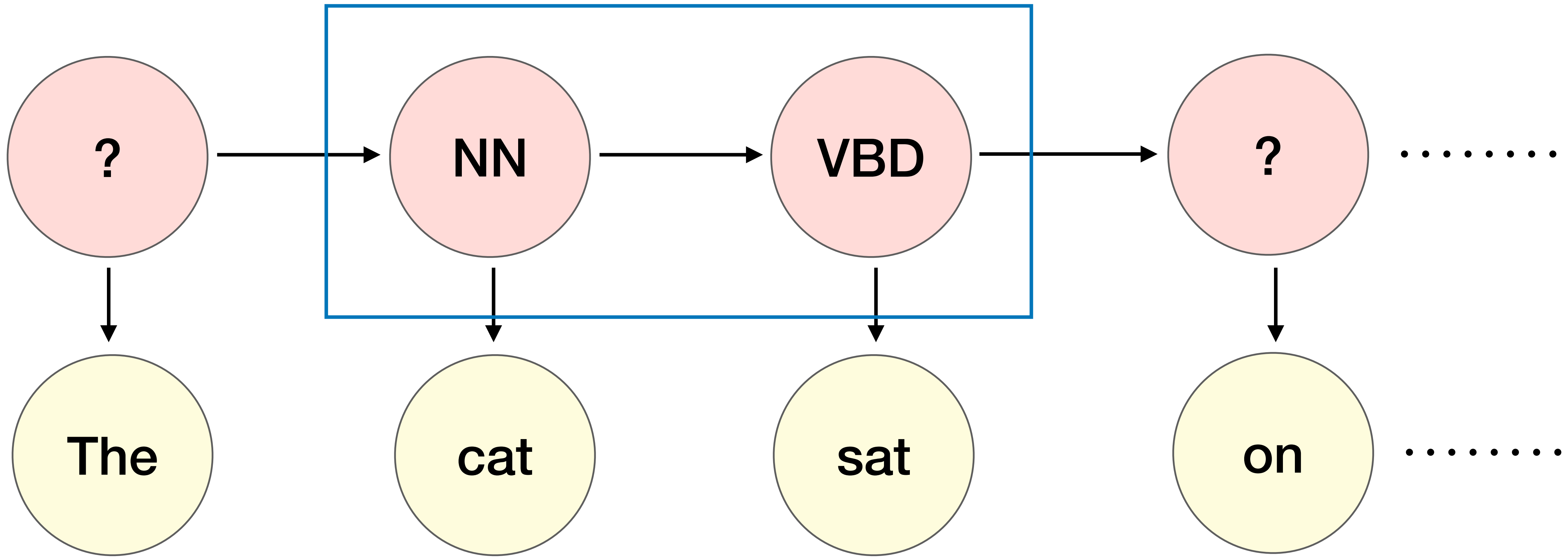
$$\overline{Count}(a \rightarrow b) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \ Count(X_i, Y, a \rightarrow b)$$

$$= \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \ Count(Y, a \rightarrow b)$$

$$\overline{Count}(a \rightarrow A) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \ Count(X_i, Y, a \rightarrow A)$$

Cannot enumerate all possible Y!

Efficient EM



$$Y = \langle y_1, y_2, \ldots, y_m \rangle$$

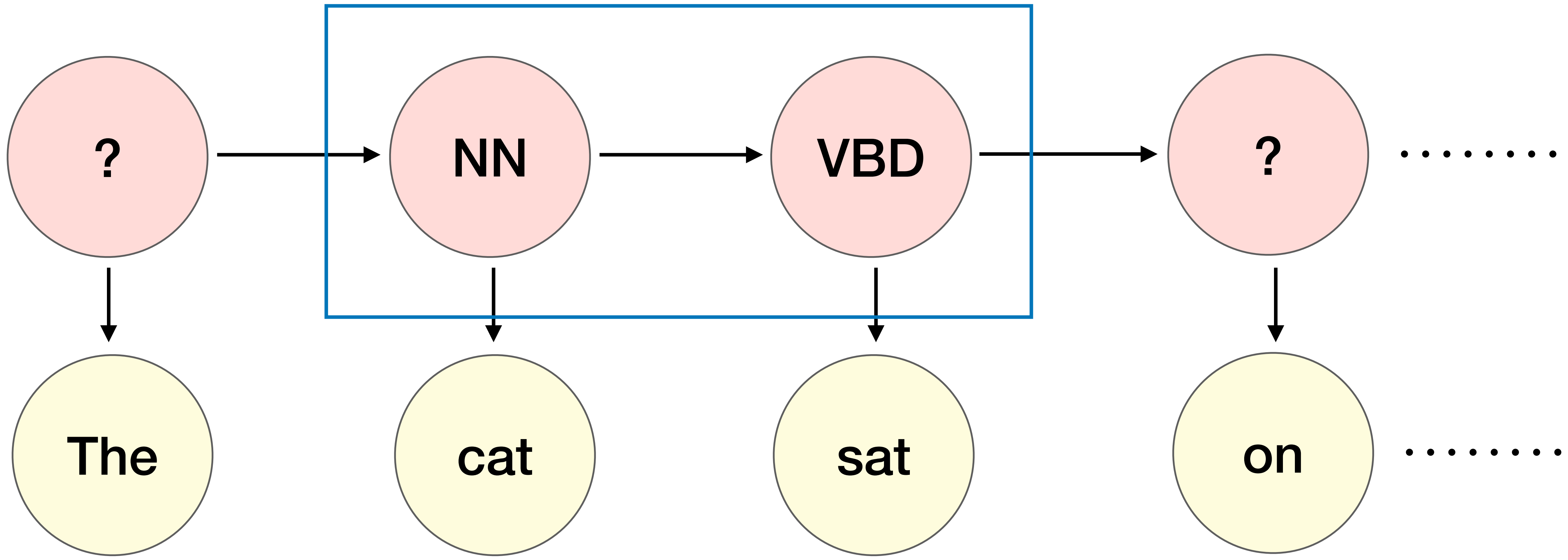
- **(E-Step)**

$$\overline{Count}(NN \rightarrow VBD) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(Y, NN \rightarrow VBD)$$

$$= \sum_{i} \sum_{j=1}^{m-1} P(y_j = NN, y_{j+1} = VBD \,|\, X_i, \theta^{t-1}, \phi^{t-1})$$

where m is the length of the sequence X_i

Efficient EM



$$Y = \langle y_1, y_2, \ldots, y_m \rangle$$

- **(E-Step)**

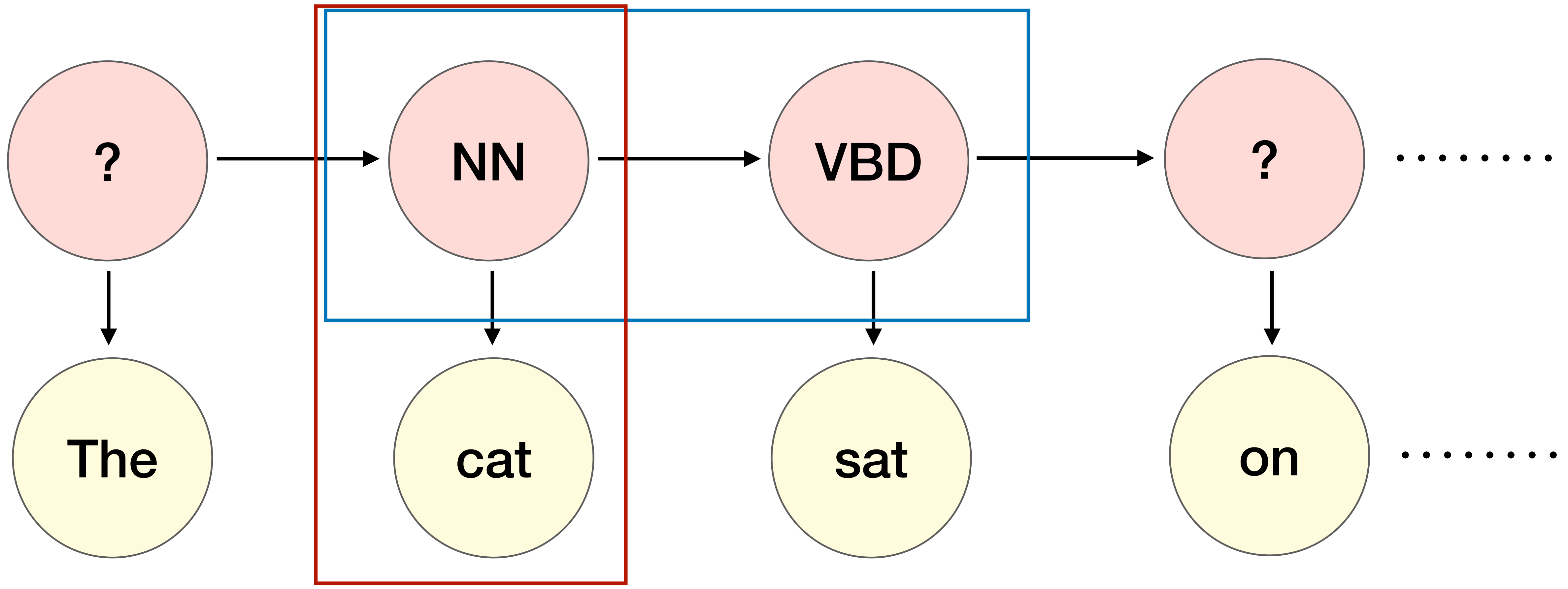
$$\overline{Count}(NN \rightarrow VBD) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1})\ Count(Y, NN \rightarrow VBD)$$

$$= \sum_{i} \sum_{j=1}^{m-1} P(y_j = NN, y_{j+1} = VBD | X_i, \theta^{t-1}, \phi^{t-1})$$

All other y variables marginalized out

where m is the length of the sequence X_i

Efficient EM



$$Y = \langle y_1, y_2, \ldots, y_m \rangle$$

- **(E-Step)**

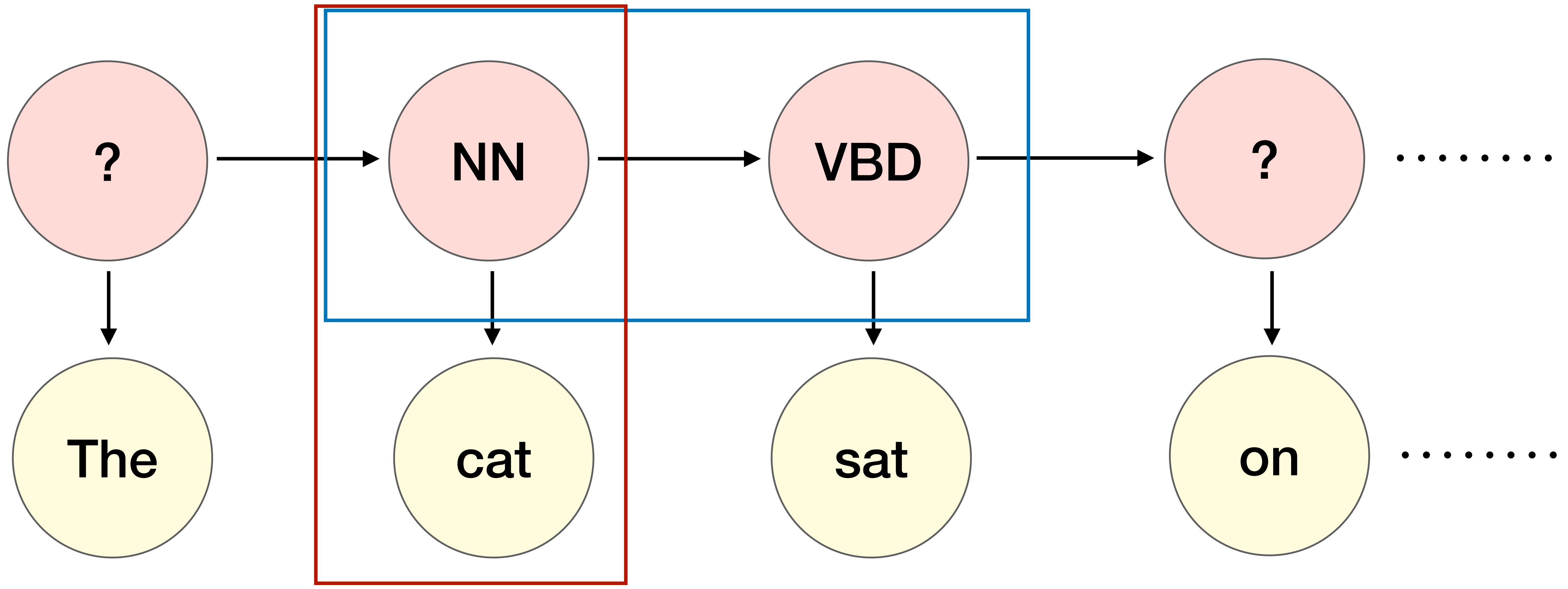
$$\overline{Count}(NN \rightarrow VBD) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \; Count(Y, \theta_k)$$

$$= \sum_{i} \sum_{j=1}^{m} P(y_j = NN, y_{j+1} = VBD | X_i, \theta^{t-1}, \phi^{t-1})$$

where m is the length of the sequence X_i

Similarly, $\overline{Count}(NN \rightarrow cat) = \sum_{i} \sum_{j:X_{ij} = cat} P(y_j = NN | X_i, \theta^{t-1}, \phi^{t-1})$

Efficient EM



$$Y = \langle y_1, y_2, \ldots, y_m \rangle$$

- **(E-Step)**

$$\overline{Count}(NN \rightarrow VBD) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1}) \ Count(Y, \theta_k)$$

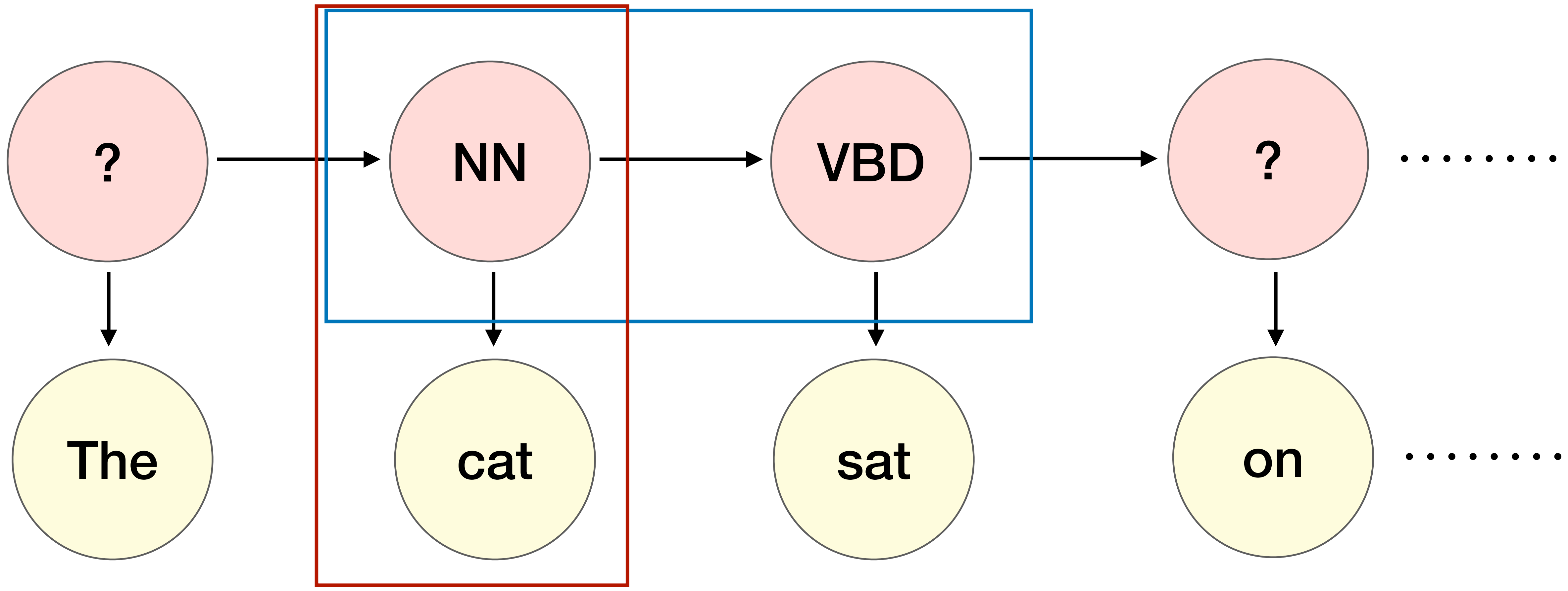
$$= \sum_{i} \sum_{j=1}^{m} P(y_j = NN, y_{j+1} = VBD | X_i, \theta^{t-1}, \phi^{t-1})$$

All other y variables marginalized out

where m is the length of the sequence X_i

Similarly, $\overline{Count}(NN \rightarrow cat) = \sum_{i} \sum_{j:X_{ij} = cat} P(y_j = NN | X_i, \theta^{t-1}, \phi^{t-1})$

Efficient EM



$$Y = \langle y_1, y_2, \ldots, y_m \rangle$$

- **(E-Step)**

$$\overline{Count}(NN \rightarrow VBD) = \sum_{i=1}^{n} \sum_{Y} P(Y | X_i, \theta^{t-1}, \phi^{t-1})\ Count(Y, \theta_k)$$

$$= \sum_{i} \sum_{j=1}^{m} P(y_j = NN, y_{j+1} = VBD | X_i, \theta^{t-1}, \phi^{t-1})$$

All other y variables marginalized out

where m is the length of the sequence X_i

Similarly, $\overline{Count}(NN \rightarrow cat) = \sum_{i} \sum_{j:X_{ij} = cat} P(y_j = NN | X_i, \theta^{t-1}, \phi^{t-1})$

only indices where the word is 'cat'

Forward-backward algorithm

Forward-backward algorithm

- Define:

  $\alpha_s(j) = P(x_1, \ldots, x_{j-1}, y_j = s \mid \theta, \phi)$ (forward probability)

  i.e. the marginal probability of seeing observations $x_1, \ldots, x_{j-1}$ and the particular state s in the j^{th} position.

Forward-backward algorithm

- Define:

  $\alpha_s(j) = P(x_1, \ldots, x_{j-1}, y_j = s \,|\, \theta, \phi)$ (forward probability)

  i.e. the marginal probability of seeing observations $x_1, \ldots, x_{j-1}$ and the particular state s in the j^{th} position.

- $\beta_s(j) = P(x_j, \ldots, x_m \,|\, y_j = s, \theta, \phi)$ (backward probability)

  i.e. the marginal probability of seeing observations $x_j, \ldots, x_m$ given $y_j = s$

Forward-backward algorithm

- Define:

  $\alpha_s(j) = P(x_1, \ldots, x_{j-1}, y_j = s \mid \theta, \phi)$ (forward probability)

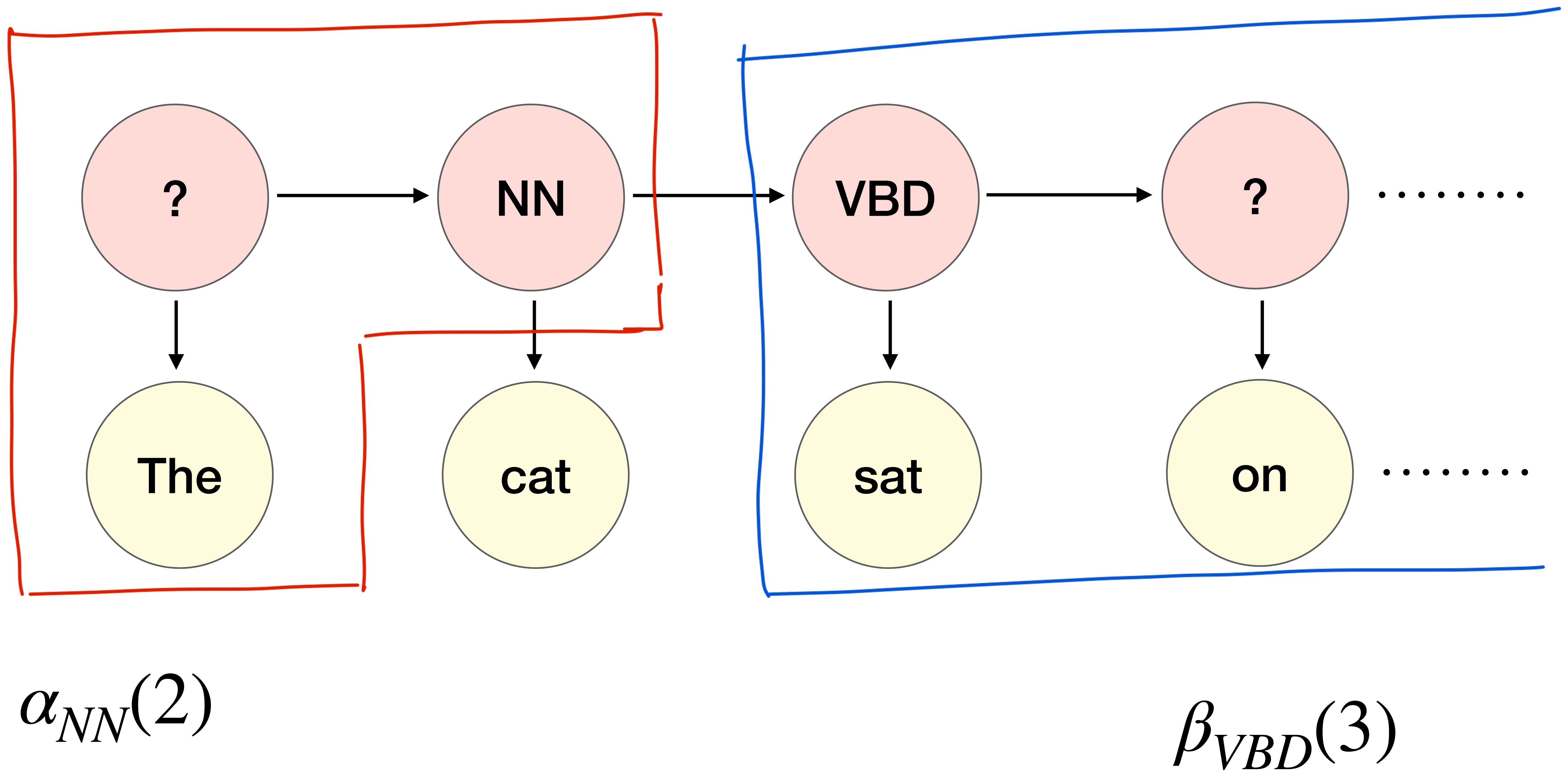
  i.e. the marginal probability of seeing observations $x_1, \ldots, x_{j-1}$ and the particular state s in the j^{th} position.

- $\beta_s(j) = P(x_j, \ldots, x_m \mid y_j = s, \theta, \phi)$ (backward probability)

  i.e. the marginal probability of seeing observations $x_j, \ldots, x_m$ given $y_j = s$

- Let us now try to express expected counts in terms of α, β

Forward-backward algorithm



$$\alpha_s(j) = P(x_1, \ldots, x_{j-1}, y_j = s \mid \theta, \phi)$$

$$\beta_s(j) = P(x_j, \ldots, x_m \mid y_j = s, \theta, \phi)$$

Forward-backward algorithm

- Observation likelihood,

$$
\begin{aligned}
Z = P(x_1, x_2, \ldots, x_m \mid \theta, \phi) &= \sum_s P(x_1, x_2, \ldots, x_{j-1}, y_j = s, x_j, \ldots x_m \mid \theta, \phi) \\
&= \sum_s P(x_1, x_2, \ldots, x_{j-1}, y_j = s \mid \theta, \phi) P(x_j, \ldots x_m \mid y_j = s, \theta, \phi) \\
&= \sum_s \alpha_s(j) \beta_s(j)
\end{aligned}
$$

for any $j \in 1,\ldots,m$

Forward-backward algorithm

Forward-backward algorithm

- Observation likelihood,

  $Z = P(x_1, x_2, \ldots, x_m \mid \theta, \phi) = \sum_s \alpha_s(j)\beta_s(j)$ for any $j \in 1,\ldots,m$

Forward-backward algorithm

- Observation likelihood,

$$Z = P(x_1, x_2, \ldots, x_m \mid \theta, \phi) = \sum_s \alpha_s(j)\beta_s(j) \quad \text{for any } j \in 1,\ldots,m$$

- Now, we can compute the following in terms of α, β :

$$P(y_j = s \mid X, \theta, \phi) = \frac{P(X, y_j = s \mid \theta, \phi)}{P(X \mid \theta, \phi)} = \frac{P(x_1, \ldots, x_{j-1}, y_j = s \mid \theta, \phi) P(x_j, \ldots, x_m \mid y_j = s, \theta, \phi)}{Z} = \frac{\alpha_s(j)\beta_s(j)}{Z}$$

Forward-backward algorithm

- Observation likelihood,

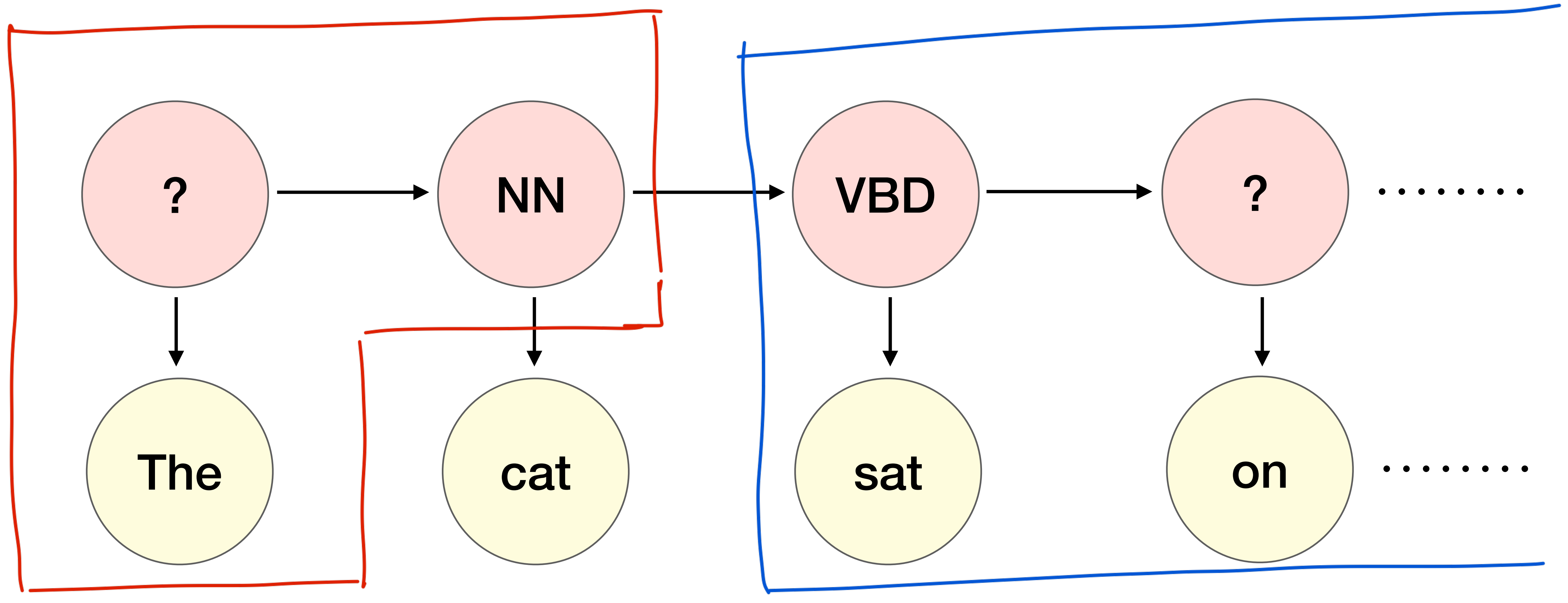
$$Z = P(x_1, x_2, \ldots, x_m | \theta, \phi) = \sum_s \alpha_s(j)\beta_s(j) \quad \text{for any } j \in 1,\ldots,m$$

- Now, we can compute the following in terms of α, β :

$$P(y_j = s | X, \theta, \phi) = \frac{P(X, y_j = s | \theta, \phi)}{P(X | \theta, \phi)} = \frac{P(x_1, \ldots, x_{j-1}, y_j = s | \theta, \phi) P(x_j, \ldots, x_m | y_j = s, \theta, \phi)}{Z} = \frac{\alpha_s(j)\beta_s(j)}{Z}$$

- and $P(y_j = s, y_{j+1} = s' | X, \theta, \phi) = \dfrac{\alpha_s(j)\ \theta_{s \to s'}\ \phi_{s \to x_j}\ \beta_{s'}\ (j+1)}{Z}$

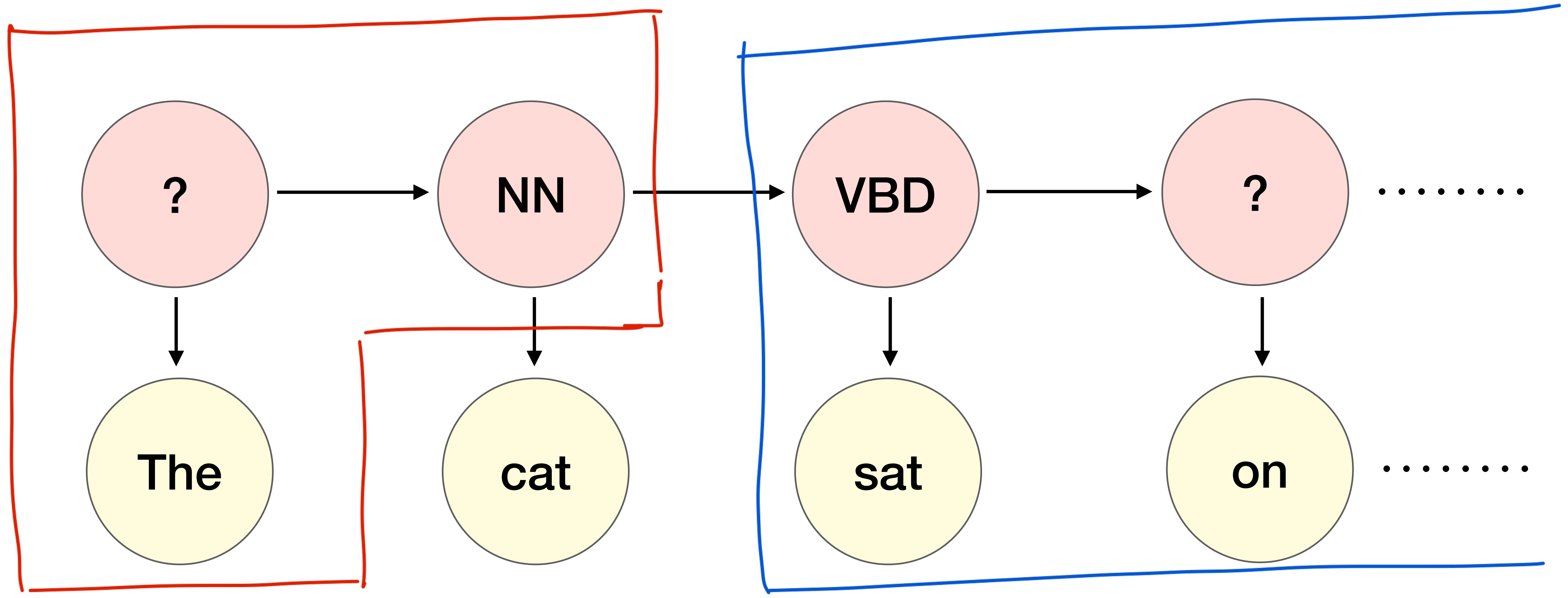
Forward-backward algorithm



$\alpha_{NN}(2)$

$\beta_{VBD}(3)$

Forward-backward algorithm

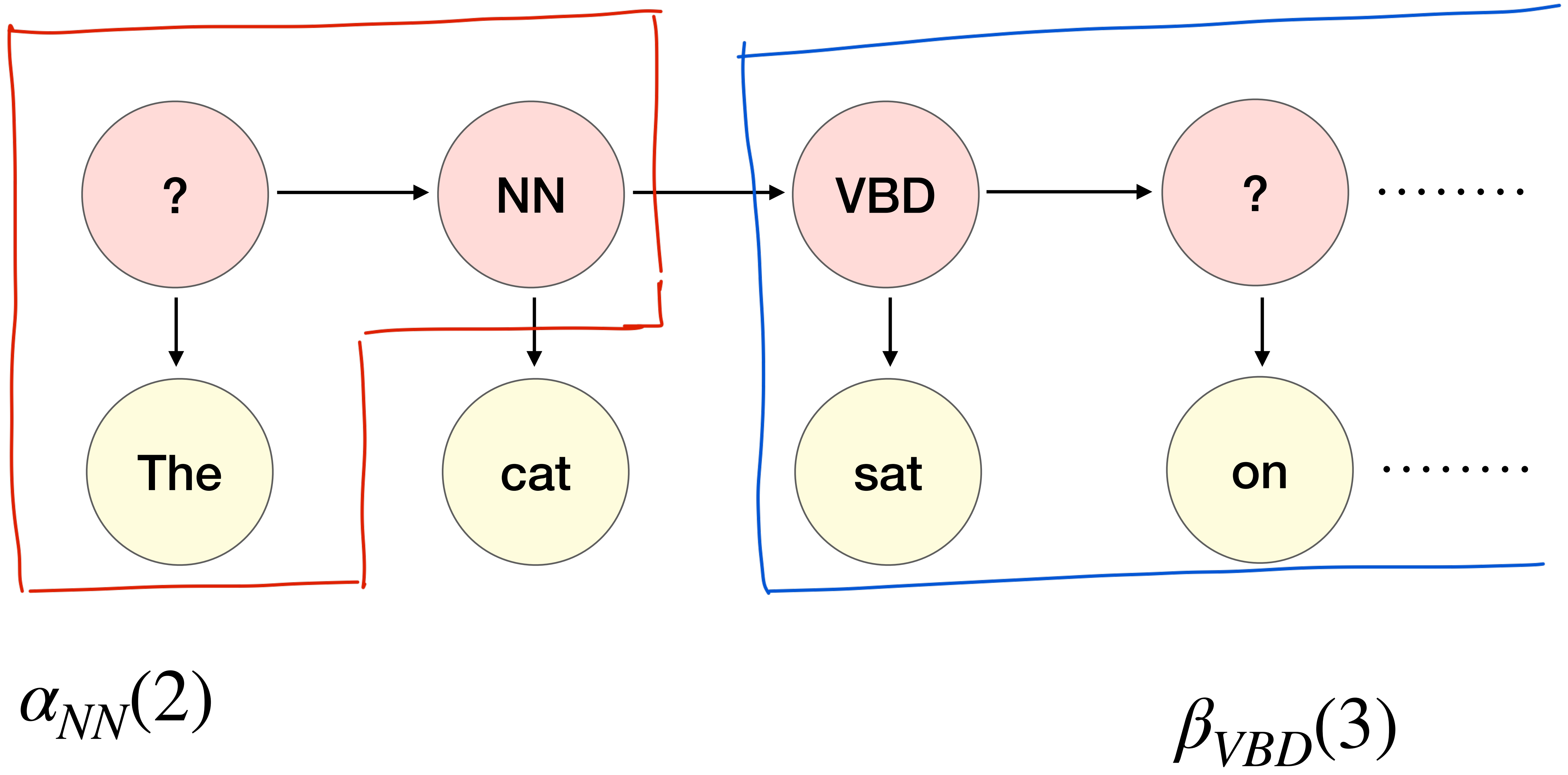


$\alpha_{NN}(2)$

$\beta_{VBD}(3)$

- $P(y_j = NN, y_{j+1} = VBD \,|\, X, \theta, \phi) = \dfrac{\alpha_{NN}(2)\ \theta_{NN \to VBD}\ \phi_{NN \to cat}\ \beta_{VBD}\ (3)}{Z}$

Forward-backward algorithm



- $P(y_j = NN, y_{j+1} = VBD \mid X, \theta, \phi) = \dfrac{\alpha_{NN}(2)\ \theta_{NN \rightarrow VBD}\ \phi_{NN \rightarrow cat}\ \beta_{VBD}\ (3)}{Z}$

- $P(y_j = NN \mid X, \theta, \phi) = \dfrac{\alpha_{NN}(2)\beta_{NN}(2)}{Z}$

Forward-backward algorithm

- $P(y_j = s \mid X, \theta, \phi) = \dfrac{\alpha_s(j)\beta_s(j)}{Z}$

  $P(y_j = s, y_{j+1} = s' \mid X, \theta, \phi) = \dfrac{\alpha_s(j)\ \theta_{s \to s'}\ \phi_{s \to x_j}\ \beta_{s'}\ (j+1)}{Z}$

- Given these, we can now estimate the expected counts:

  $\overline{Count}(s \to s') = \sum_i \sum_{j=1}^{m} P(y_j = s, y_{j+1} = s' \mid X_i, \theta, \phi)$

  $\overline{Count}(s \to o) = \sum_i \sum_{j:X_{ij} = o} P(y_j = s \mid X_i, \theta, \phi)$

  for all s, s', o

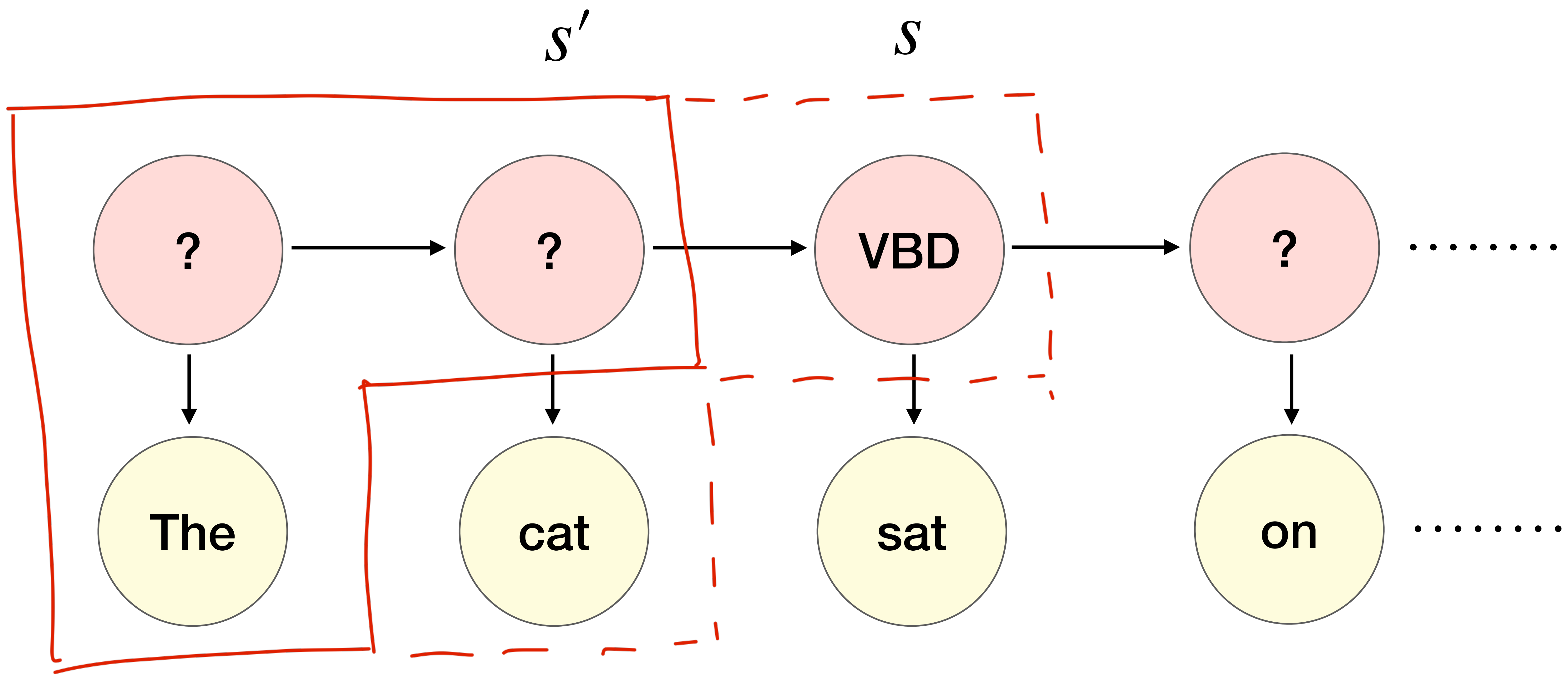
Dynamic programming

$$\alpha_s(j) = P(y_j = s, x_1, \ldots, x_{j-1})$$

$$= \sum_{s'} P(y_{j-1} = s', x_1, \ldots, x_{j-2})\ P(x_{j-1} \,|\, y_{j-1} = s')\ P(y_j = s \,|\, y_{j-1} = s')$$

$$= \sum_{s'} \alpha_{s'}\ (j-1)\ \phi_{s' \to x_{j-1}}\ \theta_{s' \to s}$$

α and β can be computed very efficiently!



Dynamic programming

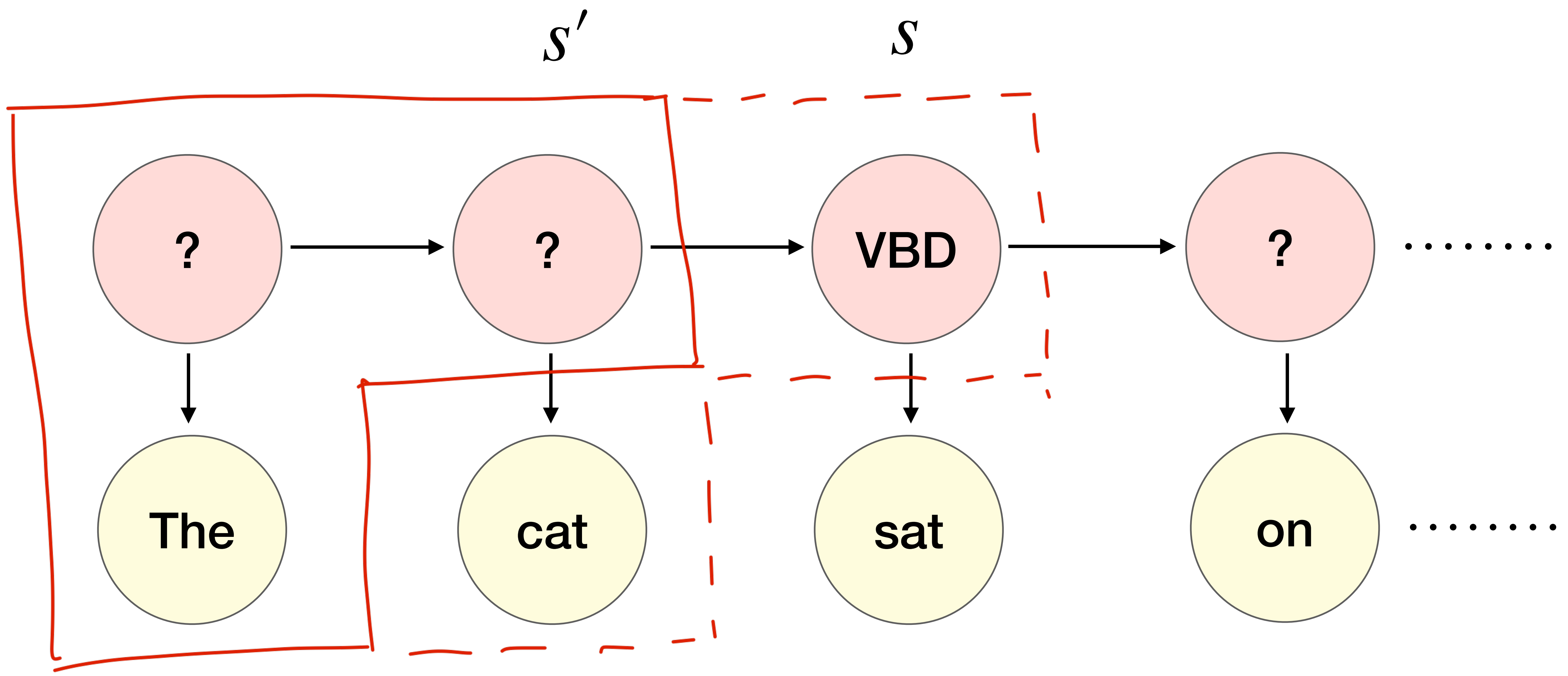
$$\alpha_s(j) = P(y_j = s, x_1, \ldots, x_{j-1})$$

$$= \sum_{s'} P(y_{j-1} = s', x_1, \ldots, x_{j-2}) \; P(x_{j-1} \,|\, y_{j-1} = s') \; P(y_j = s \,|\, y_{j-1} = s')$$

$$= \sum_{s'} \alpha_{s'}(j-1) \; \phi_{s' \to x_{j-1}} \; \theta_{s' \to s}$$

$$\alpha_s(1) = \theta_{\varnothing \to s}$$

α and β can be computed very efficiently!



Dynamic programming

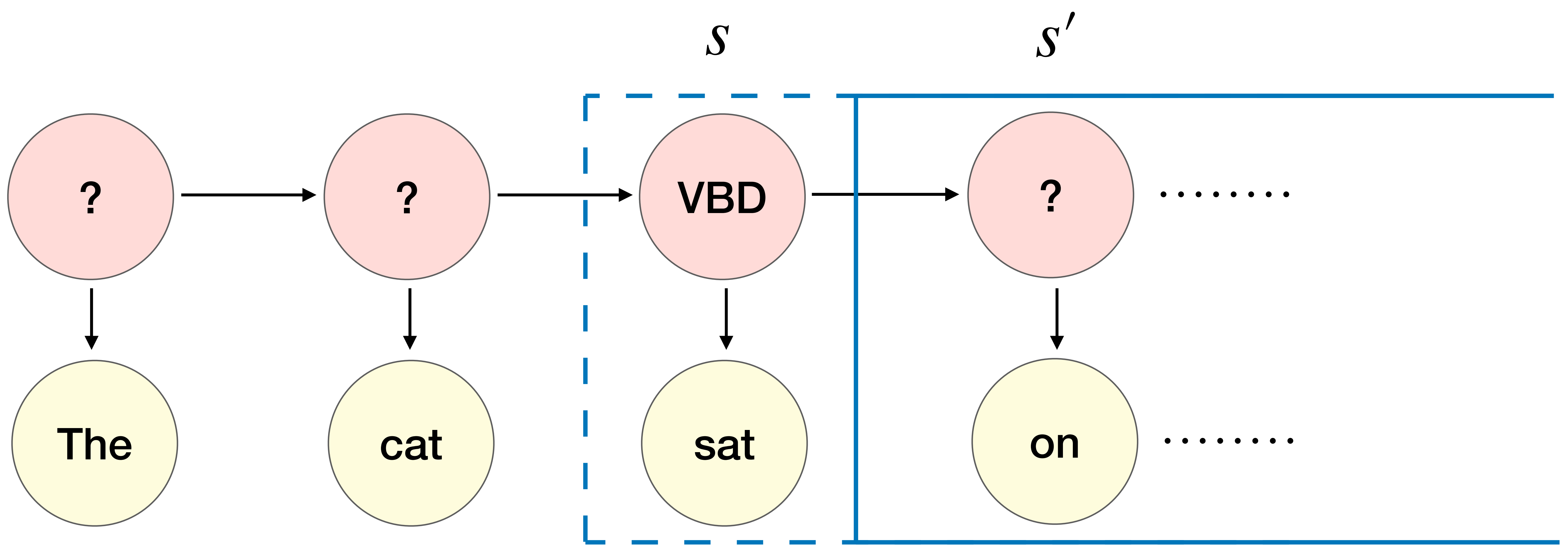
- Similarly,

$$\beta_s(j) = P(x_j, \ldots, x_m \mid y_j = s)$$

$$= \sum_{s'} P(x_{j+1}, \ldots, x_m \mid y_{j+1} = s')\ P(y_{j+1} = s' \mid y_j = s)\ P(x_j \mid y_j = s)$$

$$= \phi_{s \to x_j}\ \sum_{s'} \beta_{s'}\ (j+1)\ \ \theta_{s \to s'}$$

α and β can be computed very efficiently!



Dynamic programming

- Similarly,

$$\beta_s(j) = P(x_j, \ldots, x_m \mid y_j = s)$$

$$= \sum_{s'} P(x_{j+1}, \ldots, x_m \mid y_{j+1} = s')\ P(y_{j+1} = s' \mid y_j = s)\ P(x_j \mid y_j = s)$$

$$= \phi_{s \to x_j} \sum_{s'} \beta_{s'}(j+1)\ \ \theta_{s \to s'}$$

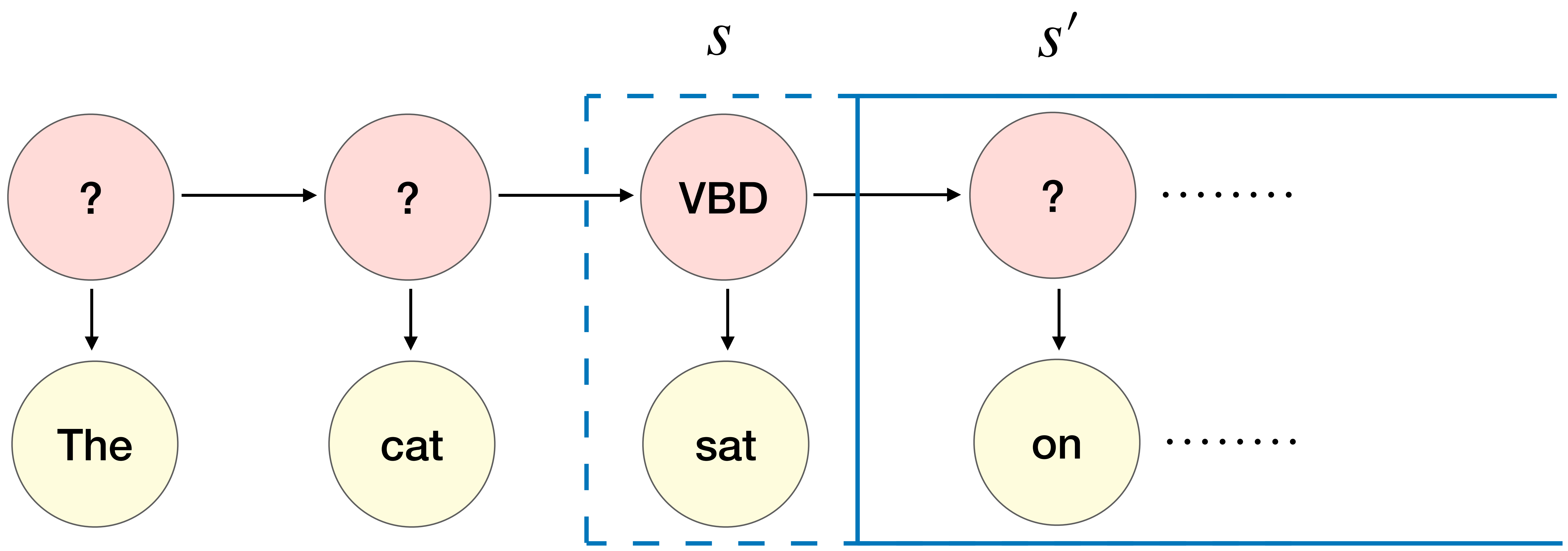
What is the base case?

A) $\beta_s(m) = \phi_{s \to x_m}$

B) $\beta_s(m) = 1$

C) $\beta_s(m) = \theta_{\varnothing \to s}$

α and β can be computed very efficiently!



Dynamic programming

- Similarly,

$$\beta_s(j) = P(x_j, \ldots, x_m \mid y_j = s)$$
$$= \sum_{s'} P(x_{j+1}, \ldots, x_m \mid y_{j+1} = s')\ P(y_{j+1} = s' \mid y_j = s)\ P(x_j \mid y_j = s)$$
$$= \phi_{s \to x_j} \sum_{s'} \beta_{s'}(j+1)\ \ \theta_{s \to s'}$$

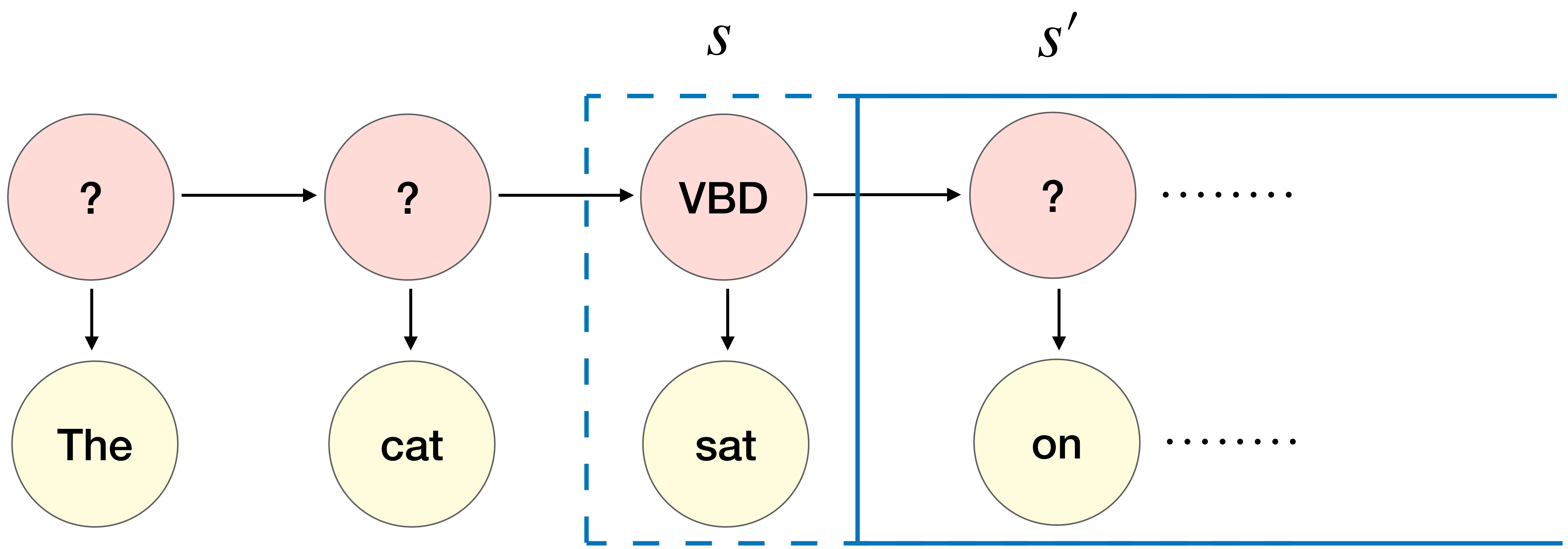
What is the base case?

A) $\beta_s(m) = \phi_{s \to x_m}$

B) $\beta_s(m) = 1$

C) $\beta_s(m) = \theta_{\varnothing \to s}$

α and β can be computed very efficiently!



Dynamic programming

- $\alpha_s(j) = \sum_{s'} \alpha_{s'}(j-1)\ \phi_{s' \to x_{j-1}}\ \theta_{s' \to s}$
- $\beta_s(j) = \phi_{s \to x_j} \sum_{s'} \beta_{s'}(j+1)\ \ \theta_{s \to s'}$
- Compute for all $s \in S, j \in [1,m]$

Dynamic programming

- $\alpha_s(j) = \sum_{s'} \alpha_{s'}(j-1)\ \phi_{s' \to x_{j-1}}\ \theta_{s' \to s}$
- $\beta_s(j) = \phi_{s \to x_j} \sum_{s'} \beta_{s'}(j+1)\ \ \theta_{s \to s'}$
- Compute for all $s \in S, j \in [1,m]$

What is the runtime of this dynamic programming algorithm?

A) $O(|S| \cdot m)$

B) $O(|S| \cdot m^2)$

C) $O(|S|^2 \cdot m)$

Dynamic programming

- $\alpha_s(j) = \sum_{s'} \alpha_{s'}(j-1)\ \phi_{s' \to x_{j-1}}\ \theta_{s' \to s}$
- $\beta_s(j) = \phi_{s \to x_j} \sum_{s'} \beta_{s'}(j+1)\ \ \theta_{s \to s'}$
- Compute for all $s \in S, j \in [1,m]$

What is the runtime of this dynamic programming algorithm?

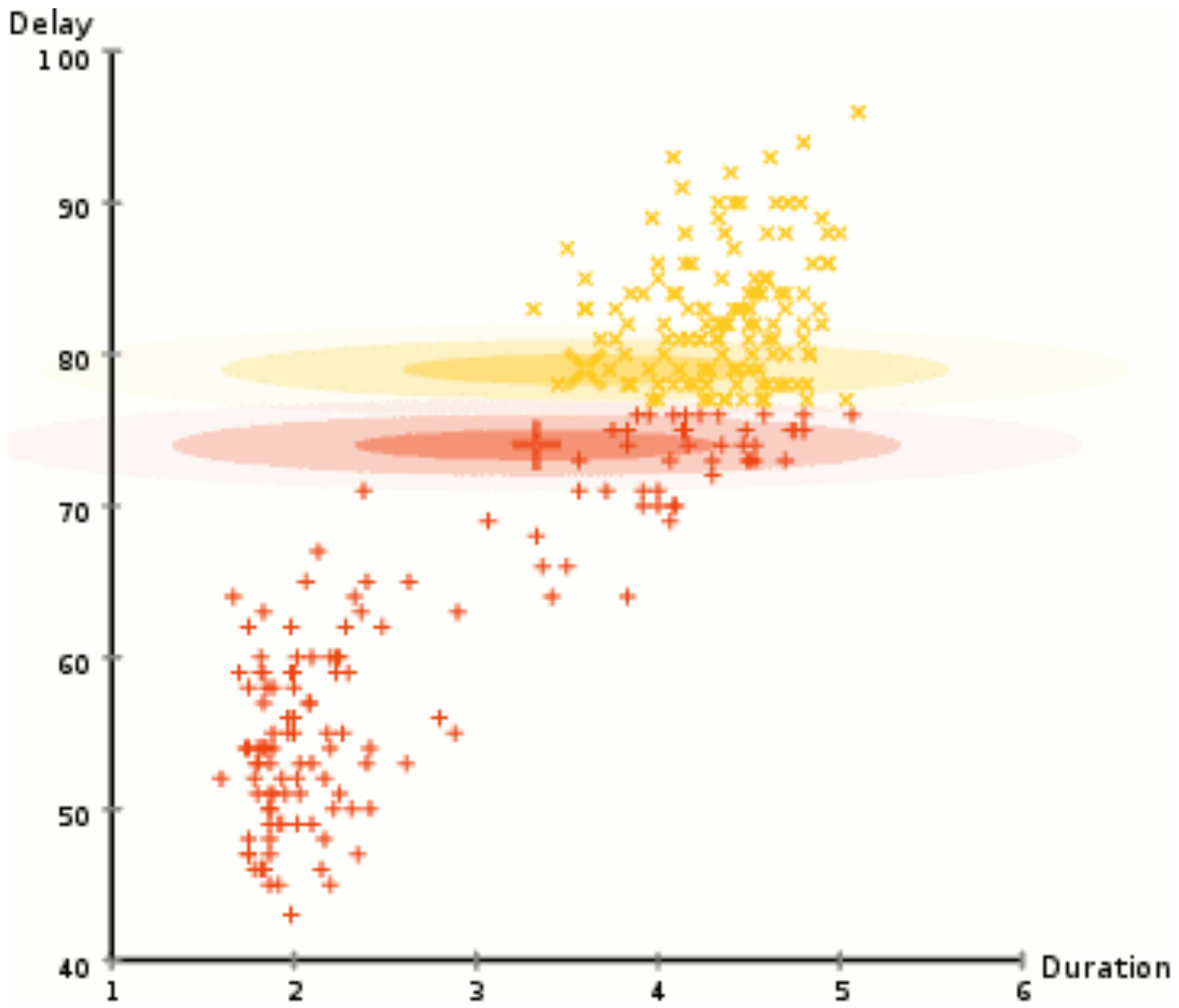
A) $O(|S| \cdot m)$

B) $O(|S| \cdot m^2)$

C) $O(|S|^2 \cdot m)$

EM applications

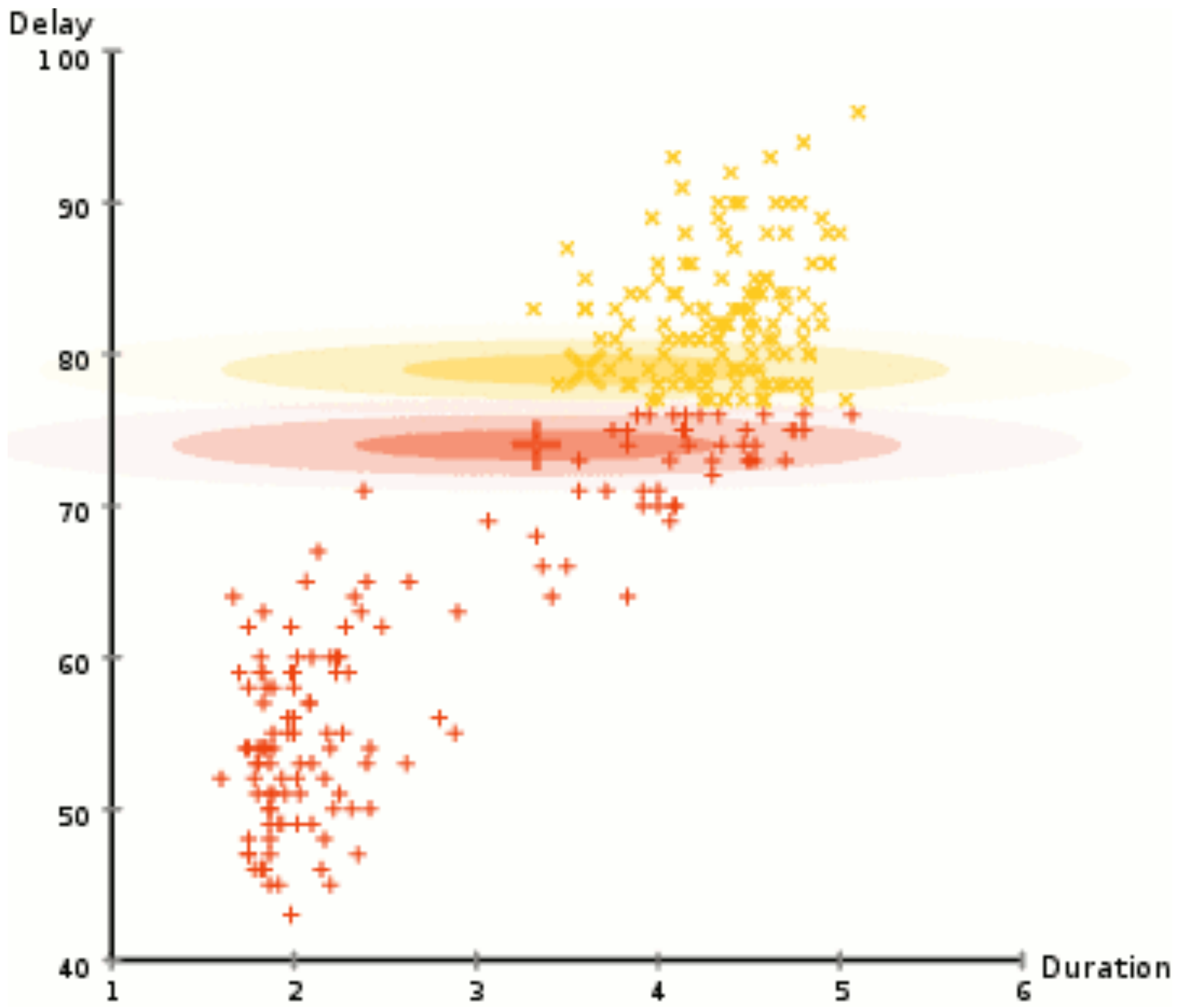
- Any task with unobserved latent variables
- In NLP:
  - Sequence modeling
  - Syntactic parsing (inside-outside algorithm)
- Clustering (cluster IDs = hidden variables)
- Computer vision (segmentation, activity recognition)
- Quantitative genetics, psychometrics, medical image reconstruction, structural engineering …



Clustering

EM applications

- Any task with unobserved latent variables
- In NLP:
  - Sequence modeling
  - Syntactic parsing (inside-outside algorithm)
- Clustering (cluster IDs = hidden variables)
- Computer vision (segmentation, activity recognition)
- Quantitative genetics, psychometrics, medical image reconstruction, structural engineering …



Clustering