HMMs for Acoustic Modeling

Live Session

CS 753, 2022

Instructor: Preethi Jyothi, IITB

Around the world

wer_are_we

WER are we? An attempt at tracking states of the art(s) and recent results on speech recognition. *Feel free to correct!* (Inspired by Are we there yet?)

WER

LibriSpeech

(Possibly trained on more data than LibriSpeech.)

WER test-clean	WER test-other	Paper	Published	Notes
5.83%	12.69%	*Humans* Deep Speech 2: End-to-End Speech Recognition in English and Mandarin	December 2015	*Humans*
1.8%	2.9%	HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units	June 2021	CNN-Transformer + Transformer LM (Self-Supervised, Libri-light-60K Unlabeled Data)
1.9%	3.9%	Conformer: Convolution-augmented Transformer for Speech Recognition	May 2020	Convolution-augmented-Transformer(Conformer) + 3-layer LSTM LM (data

Image from: https://github.com/syhw/wer_are_we

Role-Playing Seminars

ROLE	Description	Evaluation/Due	Score

Role-Playing Seminars

ROLE	Description	Evaluation/Due	Score
Talk Presentation	Give a 15-20 min talk (as the author) to clearly explain the main ideas in the paper	According to talk schedule	**8**

Role-Playing Seminars

ROLE	Description	Evaluation/Due	Score
Talk Presentation	Give a 15-20 min talk (as the author) to clearly explain the main ideas in the paper	According to talk schedule	**8**
Scientific Reporter	Write a technical article about the paper for a non-specialist but general CS audience	Due towards the end of the semester	**5**

Role-Playing Seminars

ROLE	Description	Evaluation/Due	Score
Talk Presentation	Give a 15-20 min talk (as the author) to clearly explain the main ideas in the paper	According to talk schedule	**8**
Scientific Reporter	Write a technical article about the paper for a non-specialist but general CS audience	Due towards the end of the semester	**5**
Peer Reviewer	Provide a full review of the paper	Reviews will be discussed in the same session as the talk	**4**

Role-Playing Seminars

ROLE	Description	Evaluation/Due	Score
Talk Presentation	Give a 15-20 min talk (as the author) to clearly explain the main ideas in the paper	According to talk schedule	**8**
Scientific Reporter	Write a technical article about the paper for a non-specialist but general CS audience	Due towards the end of the semester	**5**
Peer Reviewer	Provide a full review of the paper	Reviews will be discussed in the same session as the talk	**4**
Poster Presentation	Prepare a poster visualising the main highlights of the paper	GatherTown poster session at the end of the semester	**5**

Role-Playing Seminars

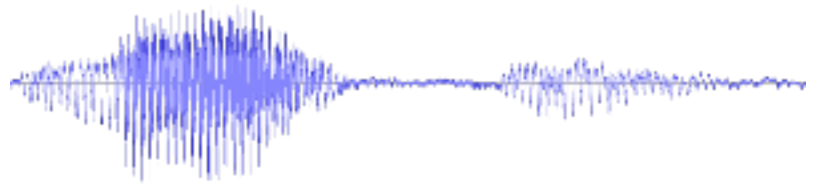
ROLE	Description	Evaluation/Due	Score
Talk Presentation	Give a 15-20 min talk (as the author) to clearly explain the main ideas in the paper	According to talk schedule	**8**
Scientific Reporter	Write a technical article about the paper for a non-specialist but general CS audience	Due towards the end of the semester	**5**
Peer Reviewer	Provide a full review of the paper	Reviews will be discussed in the same session as the talk	**4**
Poster Presentation	Prepare a poster visualising the main highlights of the paper	GatherTown poster session at the end of the semester	**5**
Hacker	Implement a small part/simplified version of the paper	Code/demo submission after the final exams	**8**

Role-Playing Seminars

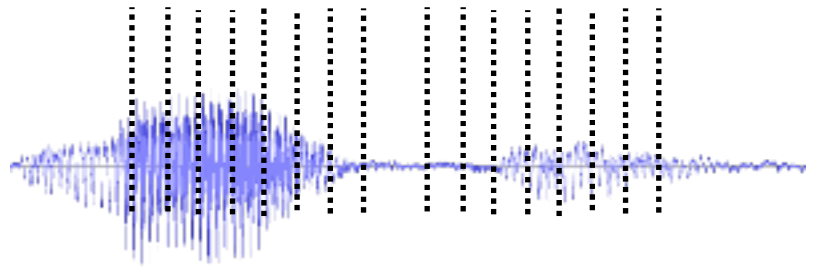
ROLE	Description	Evaluation/Due	Score
Talk Presentation	Give a 15-20 min talk (as the author) to clearly explain the main ideas in the paper	According to talk schedule	**8**
Scientific Reporter	Write a technical article about the paper for a non-specialist but general CS audience	Due towards the end of the semester	**5**
Peer Reviewer	Provide a full review of the paper	Reviews will be discussed in the same session as the talk	**4**
Poster Presentation	Prepare a poster visualising the main highlights of the paper	GatherTown poster session at the end of the semester	**5**
Hacker	Implement a small part/simplified version of the paper	Code/demo submission after the final exams	**8**

30

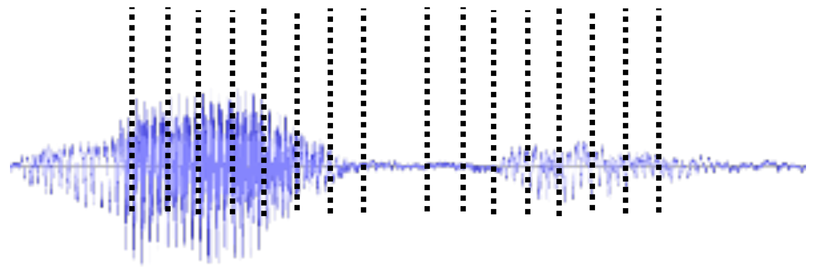
Summary: Training



Summary: Training

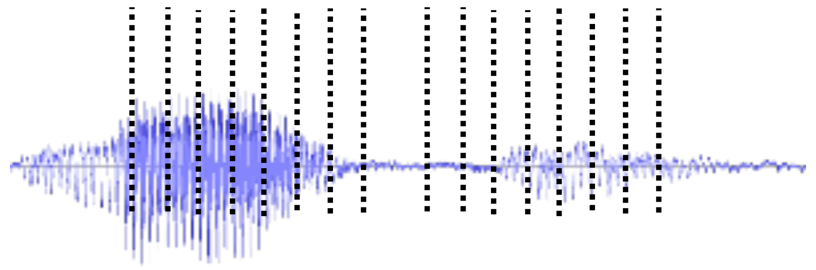


Summary: Training



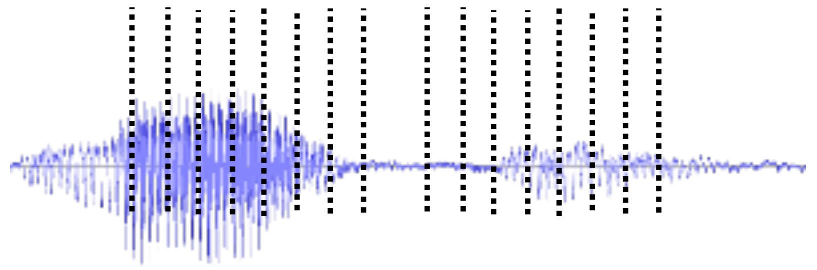
$\mathbf{O} = O_1, \ldots, O_T$

Summary: Training



$\mathbf{O} = O_1, \ldots, O_T$ **and** $\mathbf{w} = w_1, \ldots, w_M$

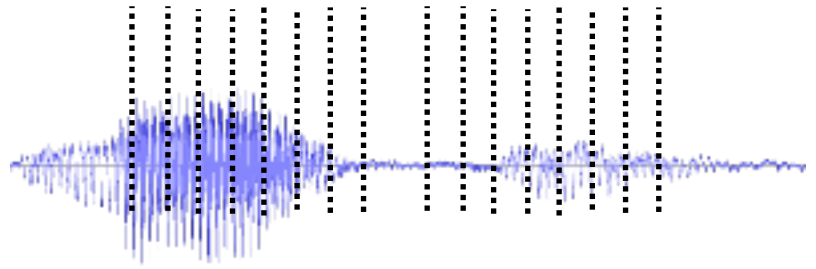
Summary: Training



$\mathbf{O} = O_1, \ldots, O_T$ **and** $\mathbf{w} = w_1, \ldots, w_M$

$\boxed{\mathbf{w}}$

Summary: Training



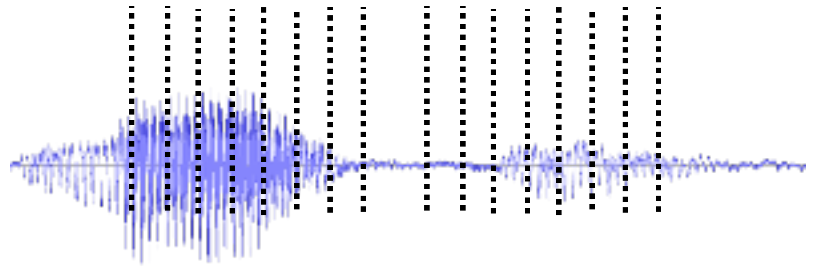
$\mathbf{O} = O_1, \ldots, O_T$ **and** $\mathbf{w} = w_1, \ldots, w_M$

$$\boxed{\mathbf{w}}$$

$$\downarrow$$

$$\boxed{w_1, \ldots, w_M}$$

Summary: Training



$\mathbf{O} = O_1, \ldots, O_T$ **and** $\mathbf{w} = w_1, \ldots, w_M$

$$\mathbf{w}$$

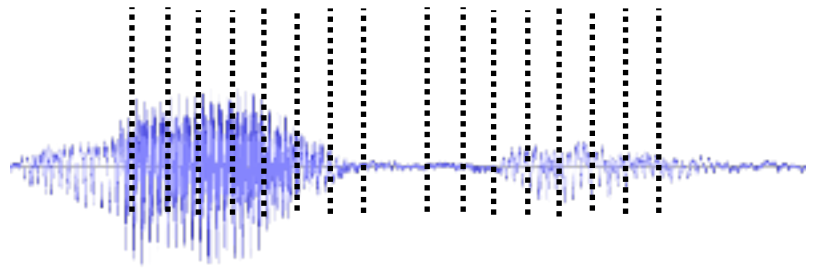
$$\downarrow$$

$$w_1, \ldots, w_M$$

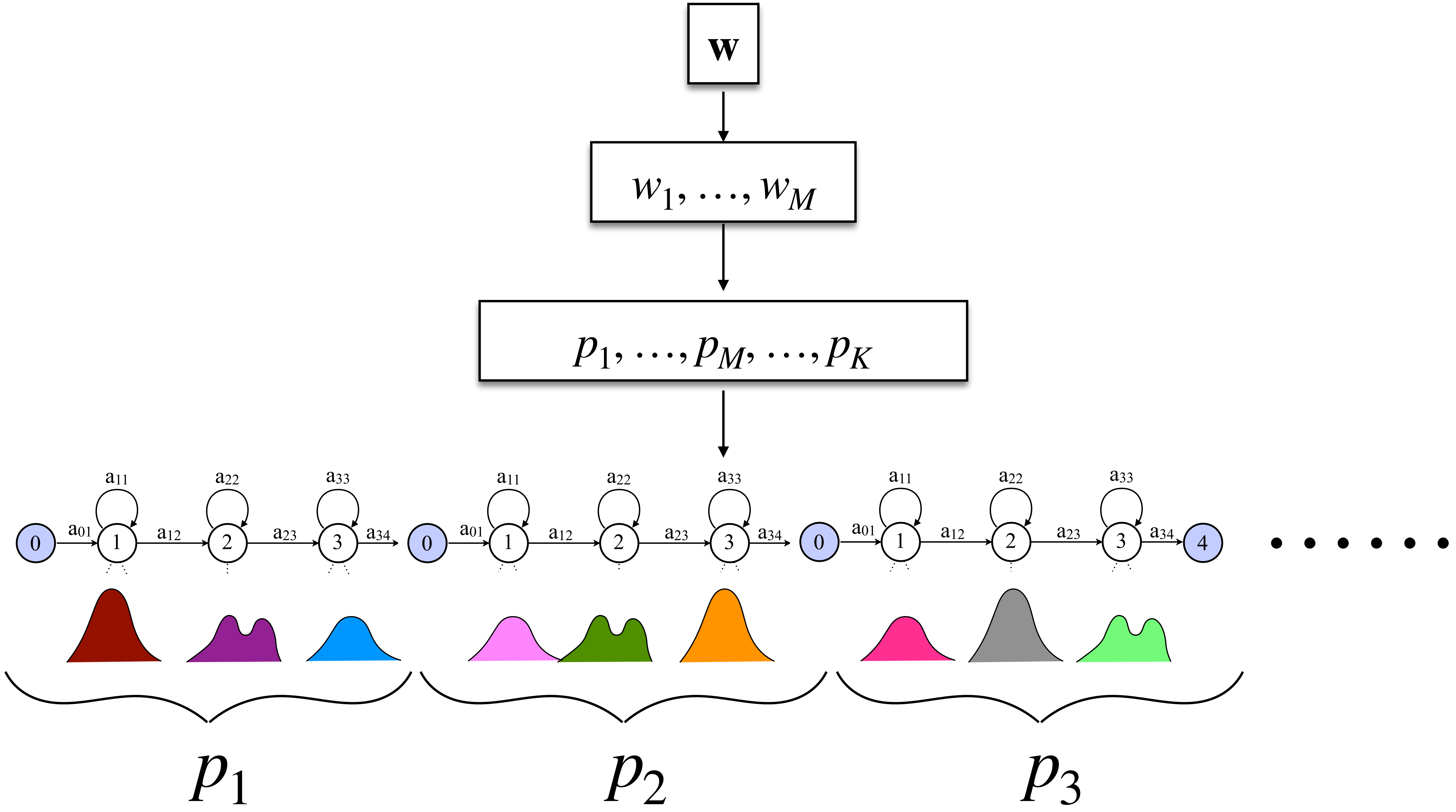
$$\downarrow$$

$$p_1, \ldots, p_M, \ldots, p_K$$

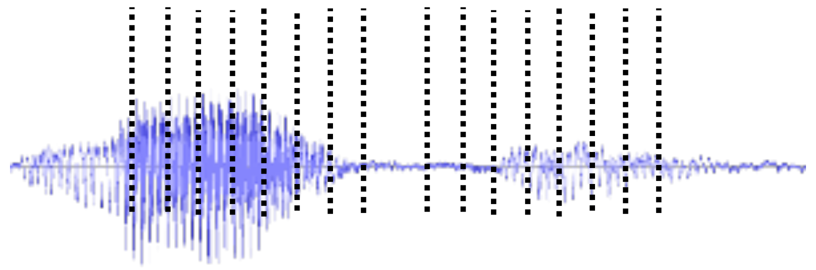
Summary: Training



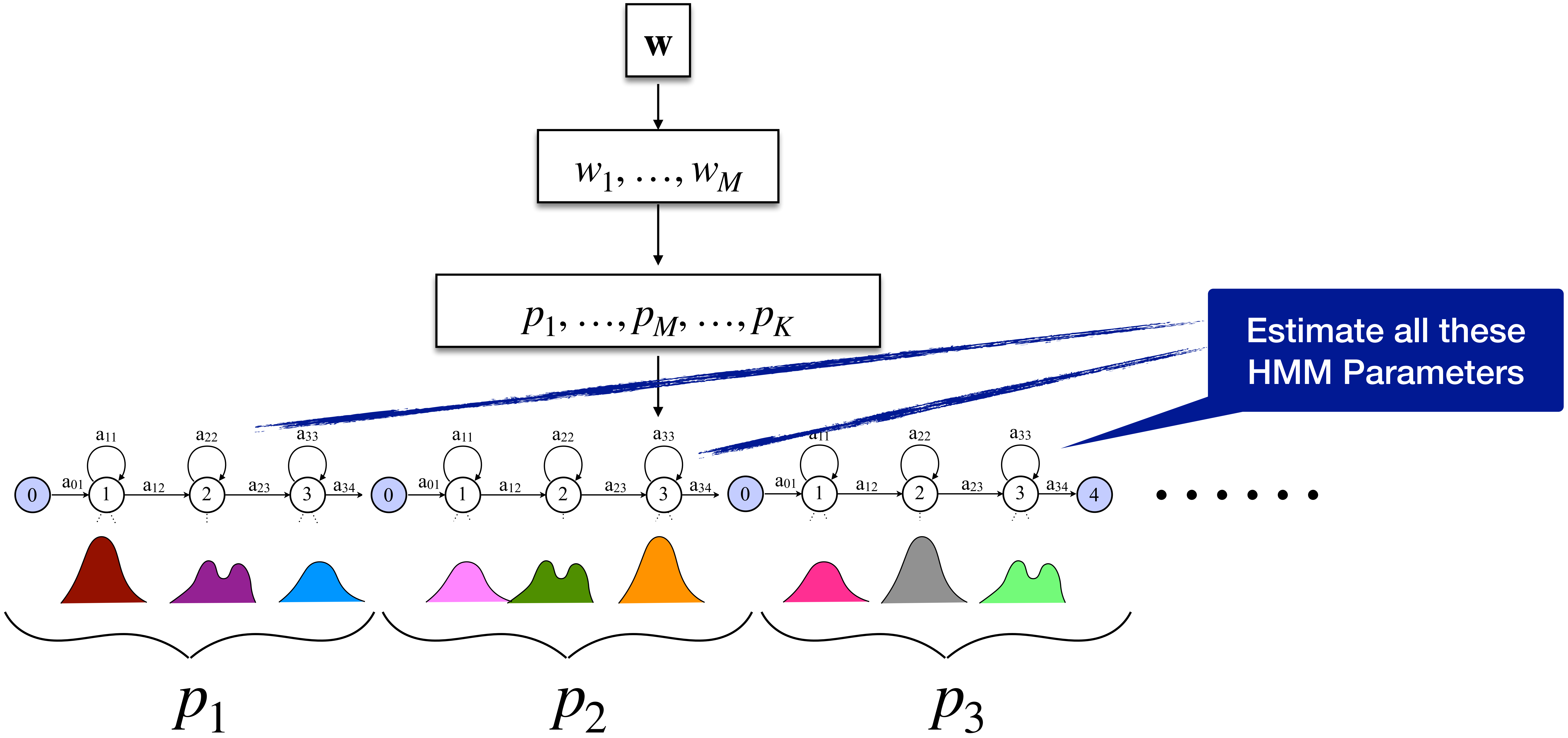
$\mathbf{O} = O_1, \ldots, O_T$ **and** $\mathbf{w} = w_1, \ldots, w_M$



Summary: Training



$\mathbf{O} = O_1, \ldots, O_T$ **and** $\mathbf{w} = w_1, \ldots, w_M$



Summary: Test

Summary: Test

$$\mathbf{O} = O_1, \ldots, O_T$$

Summary: Test

$\mathbf{O} = O_1, \ldots, O_T$

$$Q^* = \arg\max_Q P(Q \mid O)$$

Summary: Test

$$\mathbf{O} = O_1, \ldots, O_T$$

$$Q^* = \arg\max_{Q} P(Q \mid O)$$

Compute using Viterbi algorithm!

Monophone HMMs — Good enough?

Monophone HMMs — Good enough?

- A phone is affected by its phonetic context.
  - E.g. Coarticulation: Production of a speech sound is affected by adjacent speech sounds. “soon” vs. “seat”. “ten” vs. “tenth”.

Monophone HMMs — Good enough?

- A phone is affected by its phonetic context.
  - E.g. Coarticulation: Production of a speech sound is affected by adjacent speech sounds. “soon” vs. “seat”. “ten” vs. “tenth”.
- Use phones in context instead of monophones. E.g. diphones or triphones.

Monophone HMMs — Good enough?

- A phone is affected by its phonetic context.
  - E.g. Coarticulation: Production of a speech sound is affected by adjacent speech sounds. “soon” vs. “seat”. “ten” vs. “tenth”.

- Use phones in context instead of monophones. E.g. diphones or triphones.

- Triphones are commonly used in ASR systems. Phone p with left context l and right context r is written as “l-p+r”
  - “hello world” → sil-h+eh h-eh+l eh-l+ow l-ow+w ow-w-er w-er+l er-l+d l-d+sil

Triphone HMM Models

- Number of observed triphones that appear in training data ≈ 10,000s

Triphone HMM Models

- Number of observed triphones that appear in training data ≈ 10,000s
- If each triphone HMM has 3 states and each state generates m-component GMMs ($m \approx 64$), for d-dimensional acoustic feature vectors ($d \approx 40$) with Σ having d parameters (assuming diagonal covariance matrices)
    - Hundreds of millions of parameters!

Triphone HMM Models

- Number of observed triphones that appear in training data ≈ 10,000s
- If each triphone HMM has 3 states and each state generates m-component GMMs ($m \approx 64$), for d-dimensional acoustic feature vectors ($d \approx 40$) with Σ having d parameters (assuming diagonal covariance matrices)
  - Hundreds of millions of parameters!
- Insufficient data to learn all triphone models reliably. What do we do? Share parameters across triphone models.

Triphone HMM Models

- Number of observed triphones that appear in training data ≈ 10,000s
- If each triphone HMM has 3 states and each state generates m-component GMMs ($m \approx 64$), for d-dimensional acoustic feature vectors ($d \approx 40$) with Σ having d parameters (assuming diagonal covariance matrices)
  - Hundreds of millions of parameters!
- Insufficient data to learn all triphone models reliably. What do we do? Share parameters across triphone models.

More in this week's lecture on tied-state HMMs!

Recall ASR Decoding

Recall ASR Decoding

$$\mathbf{O} = O_1, \ldots, O_T$$

Recall ASR Decoding

$\mathbf{O} = O_1, \ldots, O_T$

$$Q^* = \arg\max_{Q} P(Q \mid O)$$

Recall ASR Decoding

$$\mathbf{O} = O_1, \ldots, O_T$$

$$Q^* = \arg\max_{Q} P(Q \mid O)$$

How do we go from the best state sequence to the best word sequence?

Recall ASR Decoding

$$\mathbf{O} = o_1, \ldots, o_T$$

$$Q^* = \arg\max_{Q} P(Q \mid O)$$

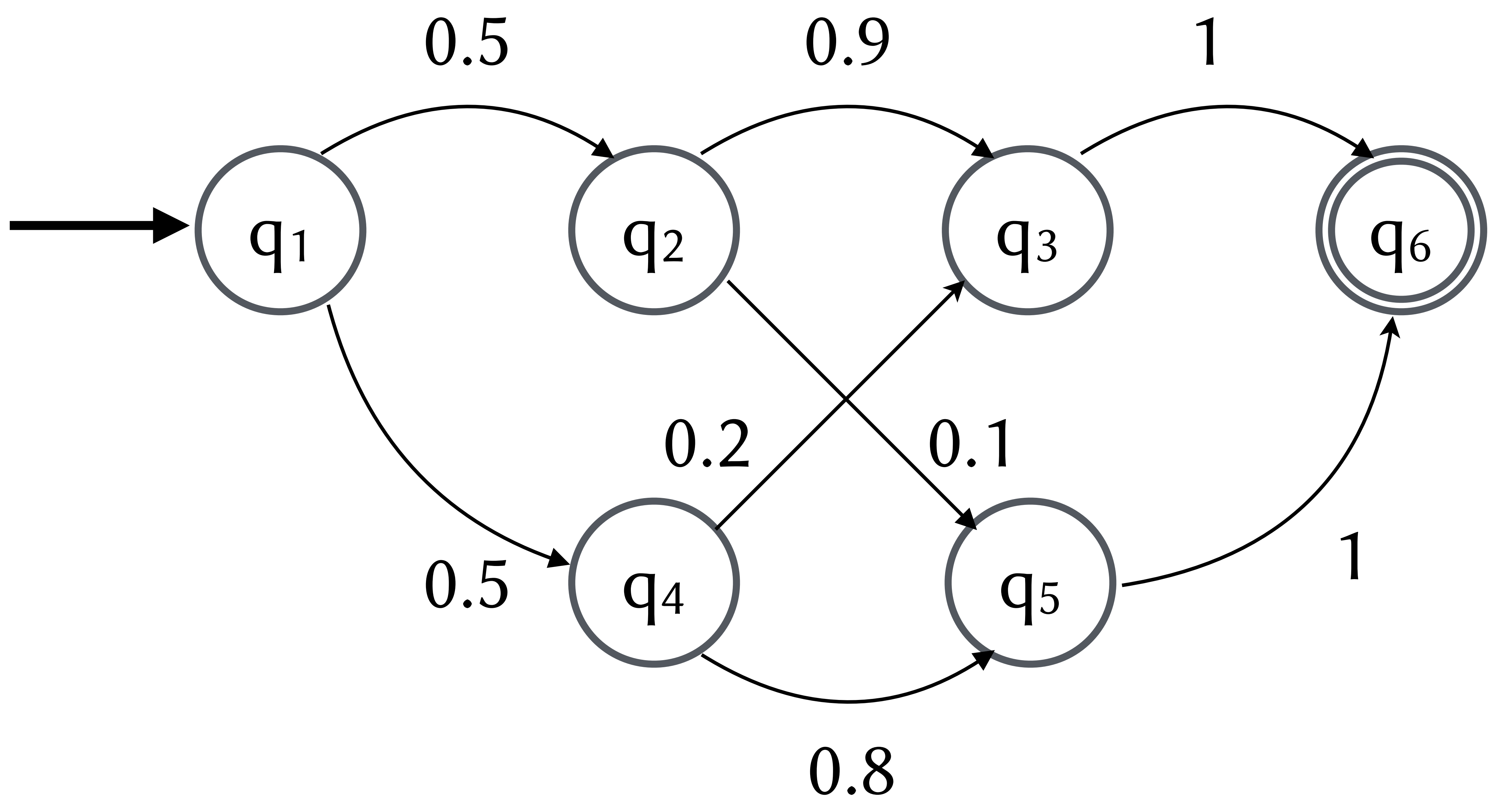
How do we go from the best state sequence to the best word sequence?

Weighted
Finite State
Transducers

Recap: HMM Forward/Viterbi Algorithms

HMMs (I)

This HMM generates hidden state sequences ($S_i \in \{q_1, \ldots, q_6\}$) and observation sequences ($O_i \in \{a, b, c\}$) of length 4.

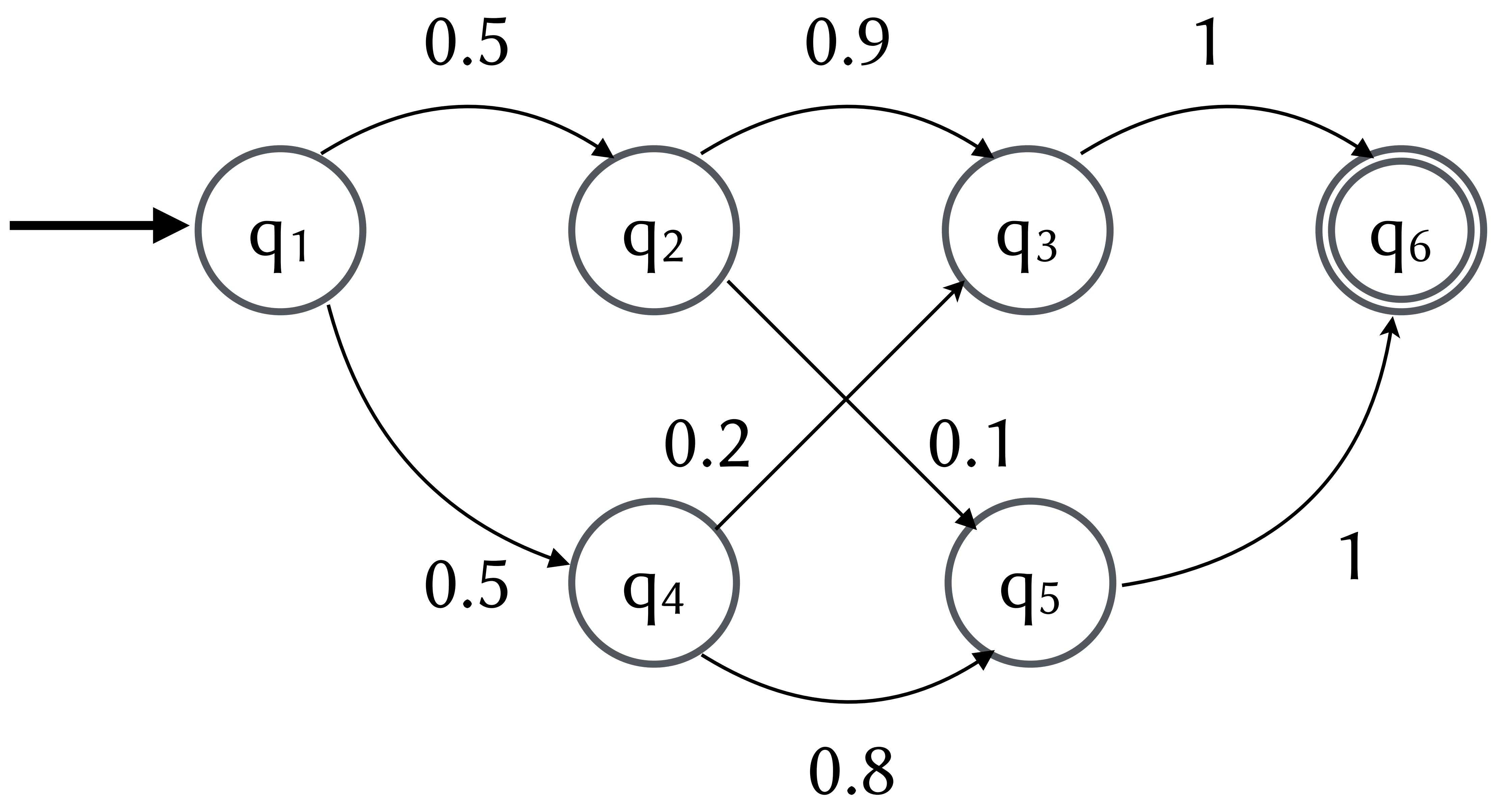


	a	b	c
q_1	0.5	0.3	0.2
q_2	0.3	0.4	0.3
q_3	0.2	0.1	0.7
q_4	0.4	0.5	0.1
q_5	0.3	0.3	0.4
q_6	0.9	0	0.1

True or False? $\Pr(O = bbca, S_1 = q_1, S_4 = q_6) = \Pr(O = bbca | S_1 = q_1, S_4 = q_6)$

HMMs (II)

This HMM generates hidden state sequences ($S_i \in \{q_1, \ldots, q_6\}$) and observation sequences ($O_i \in \{a, b, c\}$) of length 4.

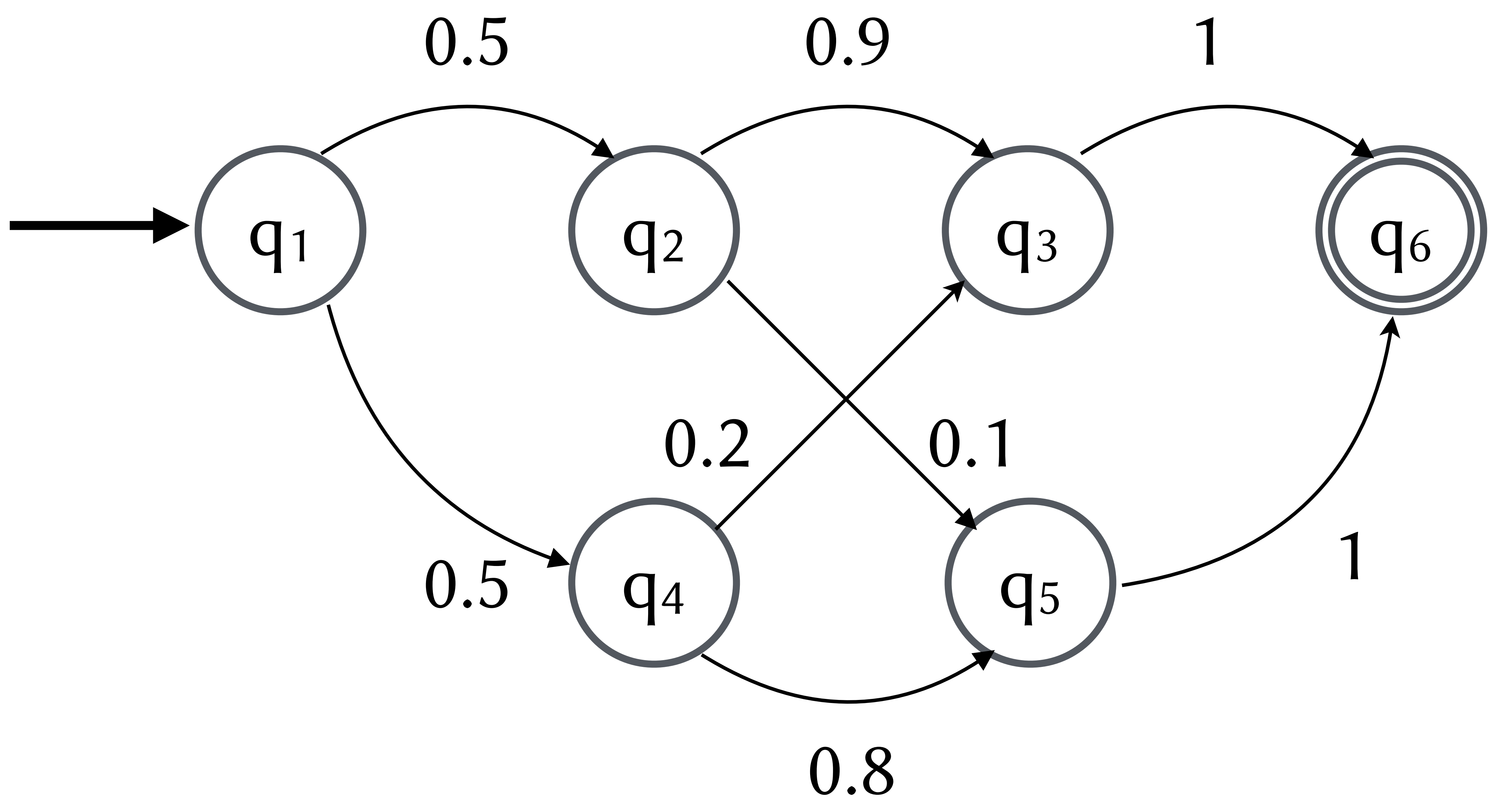


	a	b	c
q_1	0.5	0.3	0.2
q_2	0.3	0.4	0.3
q_3	0.2	0.1	0.7
q_4	0.4	0.5	0.1
q_5	0.3	0.3	0.4
q_6	0.9	0	0.1

True or False? $\Pr(O = acac, S_2 = q_2, S_3 = q_5) > \Pr(O = acac, S_2 = q_4, S_3 = q_3)$

HMMs (III)

This HMM generates hidden state sequences ($S_i \in \{q_1, \ldots, q_6\}$) and observation sequences ($O_i \in \{a, b, c\}$) of length 4.



	a	b	c
q_1	0.5	0.3	0.2
q_2	0.3	0.4	0.3
q_3	0.2	0.1	0.7
q_4	0.4	0.5	0.1
q_5	0.3	0.3	0.4
q_6	0.9	0	0.1

True or False? $\Pr(O = cbcb | S_2 = q_2, S_3 = q_5) = \Pr(O = baac, S_2 = q_4, S_3 = q_5)$

Recap: HMM Baum-Welch (EM) Algorithm

Expectation Maximization (EM) Algorithm

EM is an iterative algorithm used to compute Maximum Likelihood (ML) (or Maximum A posteriori MAP) estimates of observed data (denoted by x) in the presence of missing or hidden data (denoted by z). E.g., EM is used to compute:

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \sum_{i} \log \sum_{z} P(x_i, z|\theta)$$

Expectation Maximization (EM) Algorithm

EM is an iterative algorithm used to compute Maximum Likelihood (ML) (or Maximum A posteriori MAP) estimates of observed data (denoted by x) in the presence of missing or hidden data (denoted by z). E.g., EM is used to compute:

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \sum_{i} \log \sum_{z} P(x_i, z|\theta)$$

Two step iterative algorithm:

E step: Estimate hidden variables given the observations and current estimates of the model parameters.

M step: Estimating model parameters by maximising an auxiliary function (lower bound of the likelihood function) using estimates of the hidden data from the E step.

EM Algorithm: Fitting Parameters to Data

Observed data: i.i.d samples x_i, i=1, …, N Hidden data: Denoted by z

Goal: Find $\arg\max_\theta \mathcal{L}(\theta)$ where $\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta)$

Initial parameters: θ^0 (x is observed and z is hidden)

Iteratively compute θ^ℓ as follows that optimises an auxiliary function $Q(\theta, \theta^{\ell-1})$:

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \Pr(z|x_i; \theta^{\ell-1}) \log \Pr(x_i, z; \theta)$$

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

EM Algorithm: Fitting Parameters to Data

Observed data: i.i.d samples x_i, i=1, …, N Hidden data: Denoted by z

Goal: Find $\arg\max_\theta \mathcal{L}(\theta)$ where $\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta)$

Initial parameters: θ^0 (x is observed and z is hidden)

Iteratively compute θ^ℓ as follows that optimises an auxiliary function $Q(\theta, \theta^{\ell-1})$:

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \Pr(z|x_i; \theta^{\ell-1}) \log \Pr(x_i, z; \theta)$$

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

EM is guaranteed to converge to a local optimum or saddle points [Wu83]

Coin example to illustrate EM

Coin example to illustrate EM

$\rho_1 = \Pr(\mathrm{H})$

$\rho_2 = \Pr(\mathrm{H})$

$\rho_3 = \Pr(\mathrm{H})$

Coin example to illustrate EM

$\rho_1 = \Pr(\mathrm{H})$ $\rho_2 = \Pr(\mathrm{H})$ $\rho_3 = \Pr(\mathrm{H})$

Repeat:

Toss Coin 1 privately
if it shows H:
Toss Coin 2 twice
else
Toss Coin 3 twice

Coin example to illustrate EM

$\rho_1 = \Pr(\mathrm{H})$ $\rho_2 = \Pr(\mathrm{H})$ $\rho_3 = \Pr(\mathrm{H})$

Repeat:

- Toss Coin 1 privately
- if it shows H:
  - Toss Coin 2 twice
- else
  - Toss Coin 3 twice

The following sequence is observed: “HH, TT, HH, TT, HH”

Coin example to illustrate EM

$\rho_1 = \Pr(\mathrm{H})$ $\rho_2 = \Pr(\mathrm{H})$ $\rho_3 = \Pr(\mathrm{H})$

Repeat:
    Toss Coin 1 privately
    if it shows H:
        Toss Coin 2 twice
    else
        Toss Coin 3 twice

The following sequence is observed: "HH, TT, HH, TT, HH"

How do you estimate ρ_1, ρ_2 and ρ_3?

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z; \theta)$$

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z; \theta)$$

where, for the coin example:

- each observation $x_i \in \mathcal{X} = \{\text{HH,HT,TH,TT}\}$

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z; \theta)$$

where, for the coin example:

- each observation $x_i \in \mathcal{X} = \{\text{HH,HT,TH,TT}\}$
- the hidden variable $z \in \mathcal{Z} = \{\text{H,T}\}$

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z; \theta)$$

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i;\theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z;\theta)$$

$$\Pr(x, z;\theta) = \Pr(x|z;\theta)\Pr(z;\theta)$$

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z; \theta)$$

$$\Pr(x, z; \theta) = \Pr(x|z; \theta) \Pr(z; \theta)$$

$\rho_1 = \Pr(\mathrm{H})$

$\rho_2 = \Pr(\mathrm{H})$

$\rho_3 = \Pr(\mathrm{H})$

where $\Pr(z; \theta) = \begin{cases} \rho_1 & \text{if } z = \mathrm{H} \\ 1 - \rho_1 & \text{if } z = \mathrm{T} \end{cases}$

Coin example to illustrate EM

Recall, for partially observed data, the log likelihood is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i;\theta) = \sum_{i=1}^{N} \log \sum_{z} \Pr(x_i, z; \theta)$$

$$\Pr(x, z; \theta) = \Pr(x|z;\theta)\,\Pr(z;\theta)$$

ρ_1 = Pr(H)

ρ_2 =Pr(H)

ρ_3 = Pr(H)

where $\Pr(z;\theta) = \begin{cases} \rho_1 & \text{if } z = \mathrm{H} \\ 1-\rho_1 & \text{if } z = \mathrm{T} \end{cases}$

$$\Pr(x|z;\theta) = \begin{cases} \rho_2^h (1-\rho_2)^t & \text{if } z = \mathrm{H} \\ \rho_3^h (1-\rho_3)^t & \text{if } z = \mathrm{T} \end{cases}$$

h : number of heads, t : number of tails

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, E-step]

Compute quantities involved in

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

where $\gamma(z, x) = \Pr(z \mid x; \theta^{\ell-1})$

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, E-step]

Compute quantities involved in

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

where $\gamma(z, x) = \Pr(z \mid x; \theta^{\ell-1})$

i.e., compute $\gamma(z, x_i)$ for all z and all i

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, E-step]

Compute quantities involved in

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

where $\gamma(z, x) = \Pr(z \mid x ; \theta^{\ell-1})$

i.e., compute $\gamma(z, x_i)$ for all z and all i

Suppose $\theta^{\ell-1}$ is $\rho_1 = 0.3$, $\rho_2 = 0.4$, $\rho_3 = 0.6$:

What is $\gamma(\mathrm{H}, \mathrm{HH})$?

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, E-step]

Compute quantities involved in

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

where $\gamma(z, x) = \Pr(z \mid x ; \theta^{\ell-1})$

i.e., compute $\gamma(z, x_i)$ for all z and all i

Suppose $\theta^{\ell-1}$ is $\rho_1 = 0.3$, $\rho_2 = 0.4$, $\rho_3 = 0.6$:

What is $\gamma(\mathrm{H}, \mathrm{HH})$? **= 0.16**

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, E-step]

Compute quantities involved in

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

where $\gamma(z, x) = \Pr(z \mid x; \theta^{\ell-1})$

i.e., compute $\gamma(z, x_i)$ for all z and all i

Suppose $\theta^{\ell-1}$ is $\rho_1 = 0.3$, $\rho_2 = 0.4$, $\rho_3 = 0.6$:

What is $\gamma(\mathrm{H}, \mathrm{HH})$? **= 0.16**

What is $\gamma(\mathrm{H}, \mathrm{TT})$? **= 0.49**

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

Find θ which maximises $Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, M-step]

Find θ which maximises

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

$$\rho_1 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i)}{N}$$

$$\rho_2 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i) h_i}{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i)(h_i + t_i)}$$

$$\rho_3 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{T}, x_i) h_i}{\sum_{i=1}^{N} \gamma(\mathrm{T}, x_i)(h_i + t_i)}$$

Baum-Welch Algorithm

Observed data: *N sequences,* x_i, $i=1\ldots N$ where $x_i \in \mathrm{V}$

Parameters θ of an HMM: transition matrix A, observation probabilities B

[EM Iteration, E-step]

Compute quantities involved in $Q(\theta,\theta^{\ell-1})$

$\gamma_{i,t}(j) = \Pr(z_t = j \mid x_i ;\theta^{\ell-1})$

$\xi_{i,t}(j,k) = \Pr(z_t = j,\, z_{t+1} = k \mid x_i ;\theta^{\ell-1})$

Baum Welch Algorithm for GMM-HMMs

Baum Welch Algorithm for GMM-HMMs

[Every EM Iteration]

Compute $\theta = \{ A_{jk}, (\mu_{jm}, \Sigma_{jm}, c_{jm}) \}$ for all j,k,m

Baum Welch Algorithm for GMM-HMMs

[Every EM Iteration]

Compute $\theta = \{ A_{jk}, (\mu_{jm}, \Sigma_{jm}, c_{jm}) \}$ for all j,k,m

$$A_{j,k} = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T_i} \xi_{i,t}(j,k)}{\sum_{i=1}^{N} \sum_{t=2}^{T_i} \sum_{k'} \xi_{i,t}(j,k')}$$

Baum Welch Algorithm for GMM-HMMs

[Every EM Iteration]

Compute $\theta = \{ A_{jk}, (\mu_{jm}, \Sigma_{jm}, c_{jm}) \}$ for all j,k,m

$$A_{j,k} = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T_i} \xi_{i,t}(j,k)}{\sum_{i=1}^{N} \sum_{t=2}^{T_i} \sum_{k'} \xi_{i,t}(j,k')}$$

$$\mu_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m) x_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}$$

$$\Sigma_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)(x_{it} - \mu_{jm})(x_{it} - \mu_{jm})^T}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}$$

$$c_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{m'=1}^{M} \gamma_{i,t}(j,m')}$$