Royal Institute of Technology

MACHINE LEARNING 2 - THE EM ALGORITHM

Lecture 9

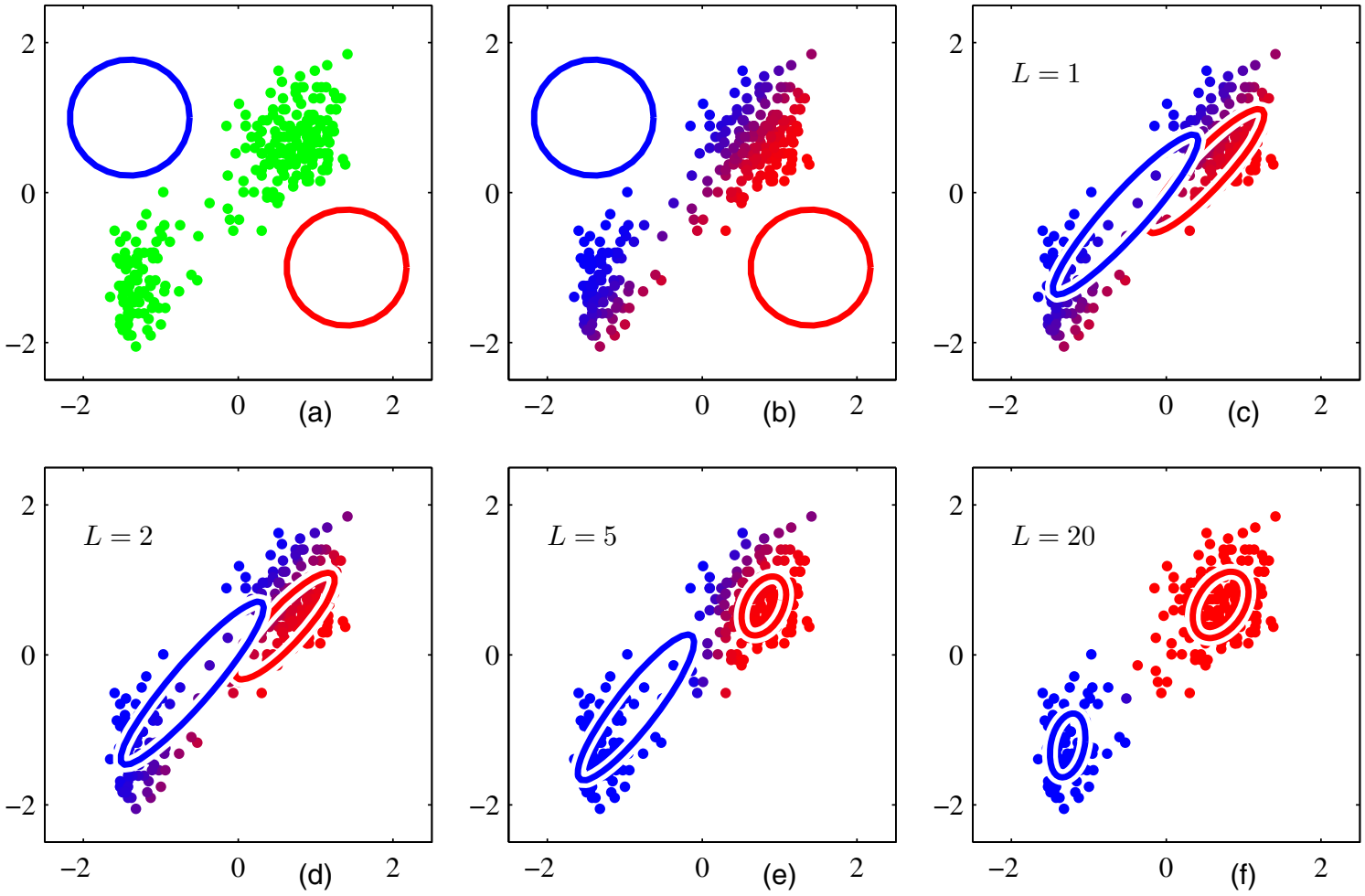
LAST LECTURE

- Backward
- Smoothing
- Sampling
- Viterbi
- K-means (inspiration)
- GMM (towards EM)

THIS LECTURE

- GMM
- EM
  - given “structure” and observations find parameters
- EM algorithm for GMM
- EM algorithm for BMM
- Baum-Welch - EM algorithm for training an HMM

ASSIGNING POINTS TO MULTIPLE MEANS



K-MEANS AS GMM

- Fixed variance, a Gaussian and mean per cluster, i.e., $\theta_c = (\mu_c, \sigma^2)$
- Idea: each point can belong to several means (clusters)
- Use responsibilities to find means

$$r_{nc} = p(z_n = c|\boldsymbol{x}_n, \boldsymbol{\theta}) = \frac{p(z_n = c|\boldsymbol{\theta})p(\boldsymbol{x}_n|z_n = c, \boldsymbol{\theta})}{\sum_{c=1}^{C} p(z_n = c|\boldsymbol{\theta})p(\boldsymbol{x}_n|z_n = c, \boldsymbol{\theta})}$$

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_n r_{nc} \boldsymbol{x}_n, \qquad \text{where } N_c = \sum_n r_{nc}$$

1-DIM GAUSSIAN MIXTURE MODELS

$\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$

$Z \sim \text{Cat}(\boldsymbol{\pi})$

$p(\boldsymbol{X}|Z = c) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}_c, \sigma_c)$

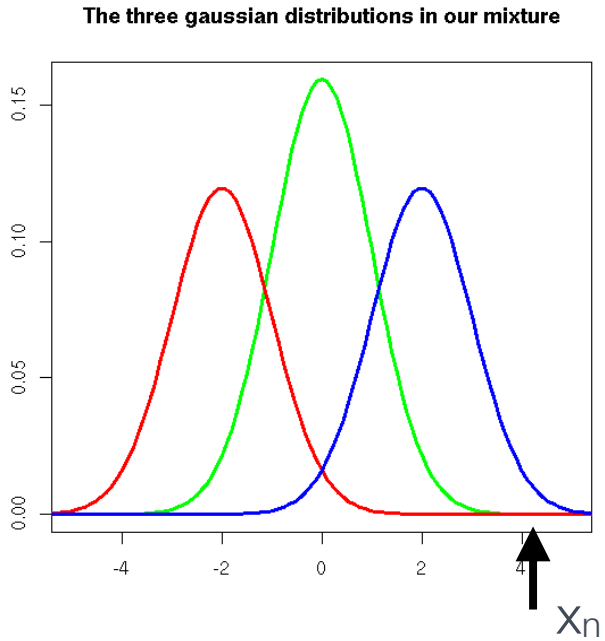
$\boldsymbol{\theta}_c = (\boldsymbol{\mu}_c, \sigma_c)$

Z hidden $\sim \text{Cat}(\boldsymbol{\pi})$

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_c, \sigma_c)$

EXAMPLE

z_n is red with probability 1/2, green with probability 3/10, blue with probability 1/5



z_n = blue

x_n is generated from the Gaussian indicated by z_n

We get $x_1,\ldots,x_N$

GMM

So,

$$p(x_n, z_n) = p(z_n)p(x_n|z_n)$$

and

$$p(x_n) = \sum_{c=1}^{C} p(z_n = c)p(x_n|z_n = c) = \sum_{c=1}^{C} \pi_c p(x_n|z_n = c)$$

and

$$r_{nc} = p(z_n = c|x_n) = \frac{p(z_n = c, x_n)}{p(x_n)} = \frac{\pi_c p(x_n|z_n = c)}{\sum_{c=1}^{C} \pi_c p(x_n|z_n = c)}$$

COMPLETE LOG LIKELIHOOD GMM

All parameters

$$
\begin{aligned}
l(\theta'; \mathcal{D}) &= \log \prod_n p(\boldsymbol{x}_n, z_n | \theta') \\
&= \sum_n \log \prod_c (\pi'_c p(\boldsymbol{x}_n | Z_n = c, \theta')^{I(z_n = c)} \\
&= \sum_n \sum_c I(z_n = c) \log(\pi'_c p(\boldsymbol{x}_n | \theta'_c)) \\
&= \sum_n \sum_c I(z_n = c) \log \pi'_c + \sum_n \sum_c I(z_n = c) \log p(\boldsymbol{x}_n | \theta'_c) \\
&= \sum_c \sum_{n: I(z_n = c)} \log \pi'_c + \sum_c \sum_{n: I(z_n = c)} \log p(\boldsymbol{x}_n | \theta'_c) \\
&= \sum_c N_c \log \pi'_c + \sum_c \sum_{n: I(z_n = c)} \log p(\boldsymbol{x}_n | \theta'_c)
\end{aligned}
$$

$$
N_c = \sum_n I(z_n = c)
$$

MLE -COMPLETE DATA FOR GAUSSIAN

$$\mathcal{D} = (z_1, \boldsymbol{x}_1), \ldots, (z_N, \boldsymbol{x}_N))$$

- Maximizing the complete log likelihood

$$l(\theta'; \mathcal{D}) = \sum_c N_c \log \pi'_c + \sum_c \sum_{n:I(z_n=c)} \log p(\boldsymbol{x}_n|\theta'_c)$$

- Boils down to maximizing

$$\sum_{n:z_n=c} \log p(\boldsymbol{x}_n|\theta'_c)$$

that is

$$\sum_{n:z_n=c} \log \frac{1}{\sqrt{2\pi\sigma'^2_c}} \exp\left(-\frac{1}{2\sigma'^2_c}(x_n - \mu'_c)^2\right)$$

EM & EXPECTED LOG LIKELIHOOD (Q-TERM)

- Iteratively maximizing the expected log likelihood (expected sufficient statistics).
- Iteratively maximizing the expected log likelihood in practice always leads to a local maxima
- The expectation is over hidden variables given data and current parameters
- We maximize the expression by choosing new parameters.

RELATIONS BETWEEN LOG-LIKELIHOODS AND Q-TERMS

log-likelihood

Q-term or expected complete log-likelihood

$$Q(\boldsymbol{\theta}', \boldsymbol{\theta}) = \sum_n E_{p(Z_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \left[l(\boldsymbol{\theta}'; Z_n, \boldsymbol{x}_n) \right]$$

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

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

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

The Q-term may not increase in every step!

LOG LIKELIHOOD & EXPECTED LOG LIKELIHOOD (Q-TERM)

Complete log likelihood

$$l(\theta'; \mathcal{D}) = \sum_n \sum_c I(z_n = c) \log \pi'_c + \sum_n \sum_c I(z_n = c) \log p(\boldsymbol{x}_n | \theta'_c)$$

Expected complete log likelihood a.k.a the Q term

$$\sum_n E_{p(Z_n | \boldsymbol{x}_n, \boldsymbol{\theta})} \left[l(\theta'; Z_n) \right] = \sum_n \sum_c r_{nc} \log \pi'_c + \sum_n \sum_c r_{nc} \log p(\boldsymbol{x}_n | \theta'_c)$$

MAXIMIZATION

★ We want $\text{argmax}_{\theta'} E_{p(z_n|\boldsymbol{x}_n,\boldsymbol{\theta})}\left[l(\theta';\mathcal{D})\right]$

★ The 2 sums $\sum_c \left(\sum_n r_{nc}\right) \log \pi'_c$ & $\sum_c \sum_n r_{nc} \log p(\boldsymbol{x}_n|\theta'_c)$ are independent

★ So, $\pi'_c = \sum_n r_{nc}/N = r_c/N$

★ In the second, different c indices are independent

★ So, we want to maximize each

$$\sum_n r_{nc} \log \frac{1}{\sqrt{2\pi\sigma'^2_c}} \exp\left(-\frac{1}{2\sigma'^2_c}(x_n - \mu_c)^2\right)$$

GMM EM-ALGORITHM

- E-step: compute $r_{nc} = p(Z_n = c | x_n, \boldsymbol{\theta})$
- M-Step: maximize (1) mixture coefficients and (2) each

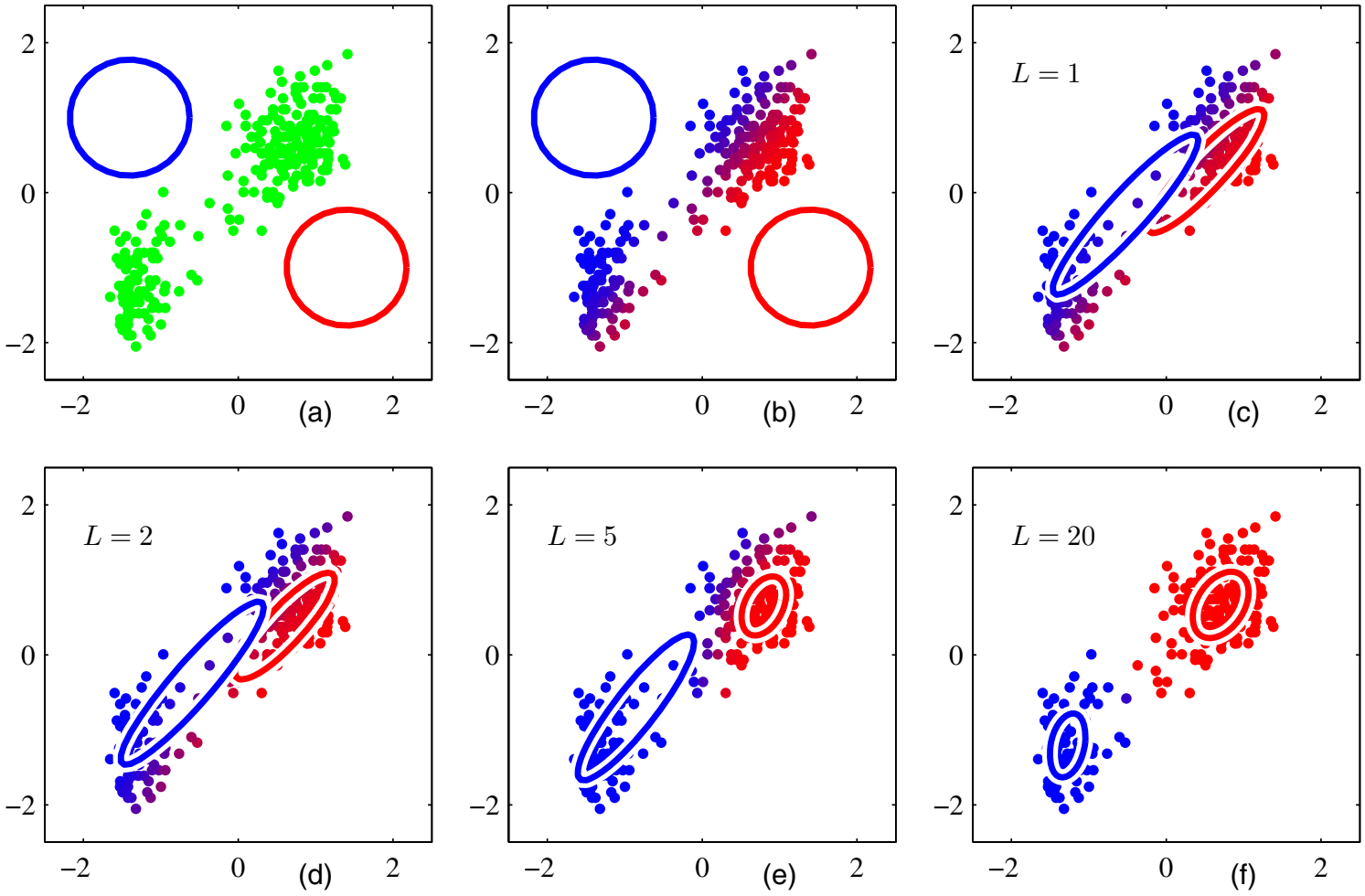
$$\sum_n r_{nc} \log \frac{1}{\sqrt{2\pi\sigma_c'^2}} \exp\left(-\frac{1}{2\sigma_c'^2}(x_n - \mu_c)^2\right)$$

by setting

$$\mu_c' = \frac{\sum_n r_{nc} x_n}{r_c} \quad \text{and} \quad \sigma_c'^2 = \frac{1}{\alpha_c'^2} = \sum_n r_{nc}(x_n - \mu_c')^2 / r_c$$

- set $\boldsymbol{\theta} = \boldsymbol{\theta}'$
- Stop when solution or likelihood hardly change otherwise repeat

E AND M STEPS



EM-ALGORITHM IN GENERAL

- E-step: compute $E_{p(Z_n|\boldsymbol{x}_n,\boldsymbol{\theta})}\left[l(\theta';Z_n,\boldsymbol{x}_n)\right]$
- M-Step:

$$\theta' = \text{argmax}_{\theta'} \sum_n E_{p(Z_n|\boldsymbol{x}_n,\boldsymbol{\theta})}\left[l(\theta';Z_n,\boldsymbol{x}_n)\right]$$

- set $\boldsymbol{\theta} = \boldsymbol{\theta}'$
- Stop when solution or likelihood hardly change otherwise repeat

- ★ Starting points
- ★ Number of starting points
- ★ Sieving starting points
- ★ The competition
  - The first iterations of EM show huge improvement in the likelihood. These are then followed by many iterations that slowly increase the likelihood. Conjugate gradient shows the opposite behaviour.

PRACTICAL ISSUES

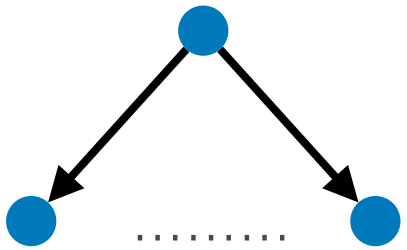
MIXTURE OF BERNOULLI

$\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$

$Z \sim \text{Cat}(\boldsymbol{\pi})$

$p(X_d|Z = c) = \text{Ber}(X|\theta_{cd})$

Z hidden $\sim \text{Cat}(\boldsymbol{\pi})$



$X_1 \sim \text{Ber}(\mu_{Z1})$ $X_D \sim \text{Ber}(\mu_{ZD})$

MIXTURE OF BERNOULLI - BASE

- Basic model D binary variables $x_1, \ldots, x_D$ and $\boldsymbol{x} = x_1, \ldots, x_D$
- Parameters $\boldsymbol{\mu} = \mu_1, \ldots, \mu_D$ where $p(x_i = 1|\boldsymbol{\mu}) = \mu_i$
- So $E[\boldsymbol{x}|\boldsymbol{\mu}] = \boldsymbol{\mu}$ and $\mathrm{cov}[\boldsymbol{x}|\boldsymbol{\mu}] = \mathrm{diag}\{\mu_i(1 - \mu_i)\}$

FULL MIXTURE MODEL - MEAN AND COVARIANCE

$$E[\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\pi}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \qquad E[x_i|\boldsymbol{\mu},\boldsymbol{\pi}] = \sum_{k=1}^{K} \pi_k \mu_{ki}$$

$$\text{cov}[x_i, x_j|\boldsymbol{\mu},\boldsymbol{\pi}] = \sum_{k=1}^{K} \pi_k \mu_{ki}\mu_{kj} - \sum_{k=1}^{K}\sum_{'k=1}^{K} \pi_k \pi_{k'} \mu_{kj}\mu_{k'j} \neq 0$$

MIXTURE OF BERNOULLI

- Class variable $z \in [K]$ and D binary variables $\boldsymbol{x} = x_1, \ldots, x_D$
- Parameters $\boldsymbol{\mu}_k = \mu_{k1}, \ldots, \mu_{kD}$ and $\boldsymbol{\pi} = \pi_1, \ldots, \pi_K$
- Mixture with likelihood

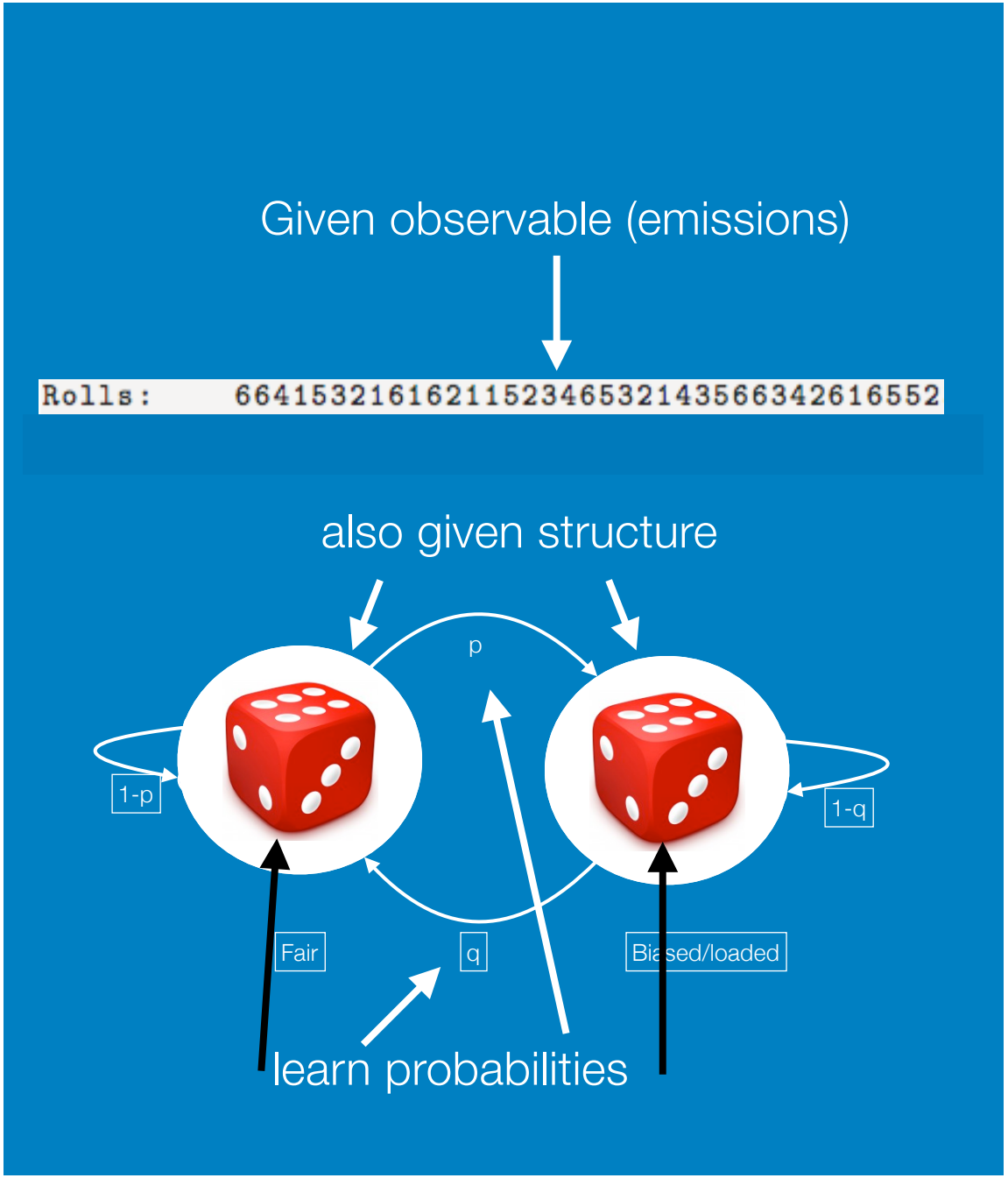
$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k)$$

where

$$p(x_i|\boldsymbol{\mu}_k) = \mu_{ki}^{x_i}(1 - \mu_{ki})^{1-x_i}$$

and

$$p(\boldsymbol{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1 - \mu_{ki})^{1-x_i}$$



BAUM-WELCH: LEARNING HMM PARAMETERS

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

The parameters we want to learn

LEARNING TRANSITION AND EMISSION PARAMETERS - FULLY OBSERVED DATA

- Parameters
  - transition $A_{lk} = p(Z_t = l | Z_{t-1} = k)$
  - emission $B_{sk} = p(X_t = s | Z_t = k)$
- Data

$$\mathcal{D} = \{(x^1_{1:T}, z^1_{1:T+1}), \ldots, (x^N_{1:T}, z^N_{1:T+1})\}$$

- Likelihood

$$L(\boldsymbol{\theta}; \mathcal{D}) = \prod_{n=1}^{N} \prod_{t=1}^{T} \left[\prod_{k,s} B_{sk}^{I(x_t^n = s, z_t^n = k)} \prod_{k,l} A_{lk}^{I(z_t^n = k, z_{t+1}^n = l)} \right]$$

MAXIMIZING LOG-LIKELIHOOD - COMPLETE DATA

$$l(\boldsymbol{\theta}; \mathcal{D}) = \sum_{k,s} M_{k,s} \log B_{ks} + \sum_{k,l} N_{k,l} \log A_{kl}$$

where

$$M_{s,k} = |\{(n,t) : x_t^n = s, z_t^n = k\}|$$

$$N_{k,l} = |\{(n,t) : z_t^n = l, z_{t+1}^n = k\}|$$

Maximized by

$$B_{sk} = M_{s,k} / \sum_s M_{s,k} = M_{s,k}/N_k \quad \& \quad A_{kl} = N_{k,l} / \sum_l N_{k,l} = N_{k,l}/N_k$$

where $N_k = |\{(n,t) : z_t^n = k\}|$

EM FOR HIDDEN DATA

$$\sum_{n=1}^{N} E_{p(Z_{1:T+1}|x^n_{1:T},\boldsymbol{\theta})}[l(\theta'; Z_{1:T+1}, x^n_{1:T})]$$

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

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

Maximized by

$$B_{sk} = \overline{M}_{s,k} / \sum_s \overline{M}_{s,k} = \overline{M}_{s,k} / \overline{N}_k \quad \& \quad A'_{kl} = \overline{N}_{k,l} / \sum_l \overline{N}_{k,l} = \overline{N}_{k,l} / \overline{N}_k$$

Both obtainable from forward and backward (smoothing like)

THE END