Royal Institute of Technology

MACHINE LEARNING 2 - CONT HMM. EM

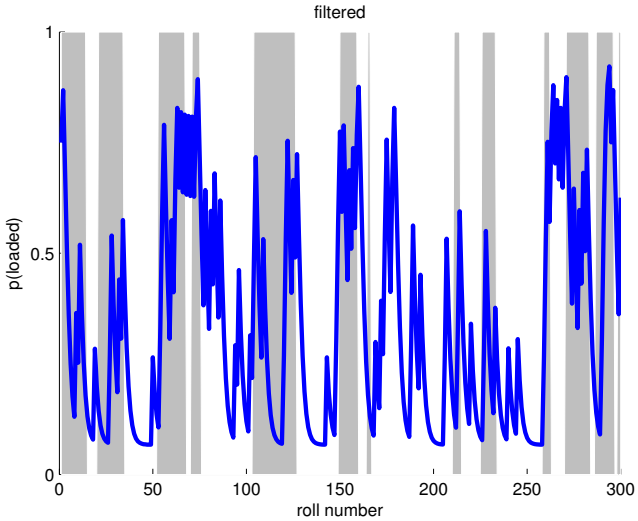
Lecture 8

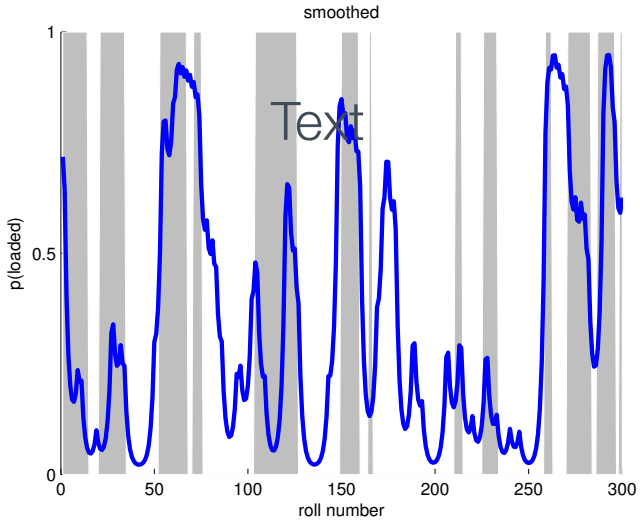
LAST LECTURE

- DGM semantics
- UGM
- De-noising
- HMMs
  - Applications (interesting probabilities)
  - DP for generation probability etc.

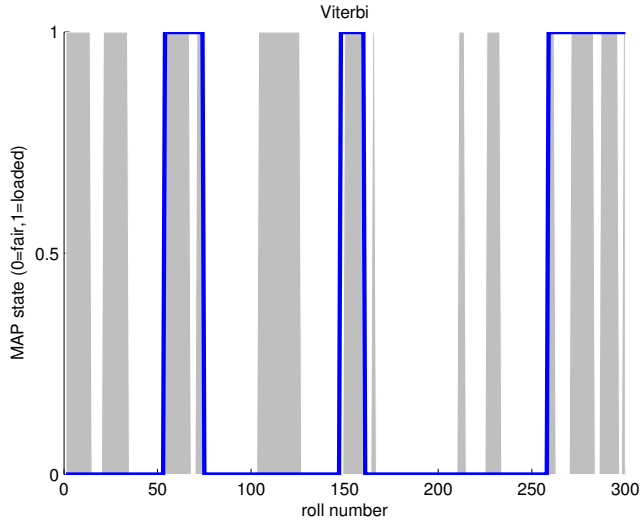
INFERENCE IN THE OCCASIONALLY DISHONEST CASINO

Grey regions are states corresponding to biased die





Text



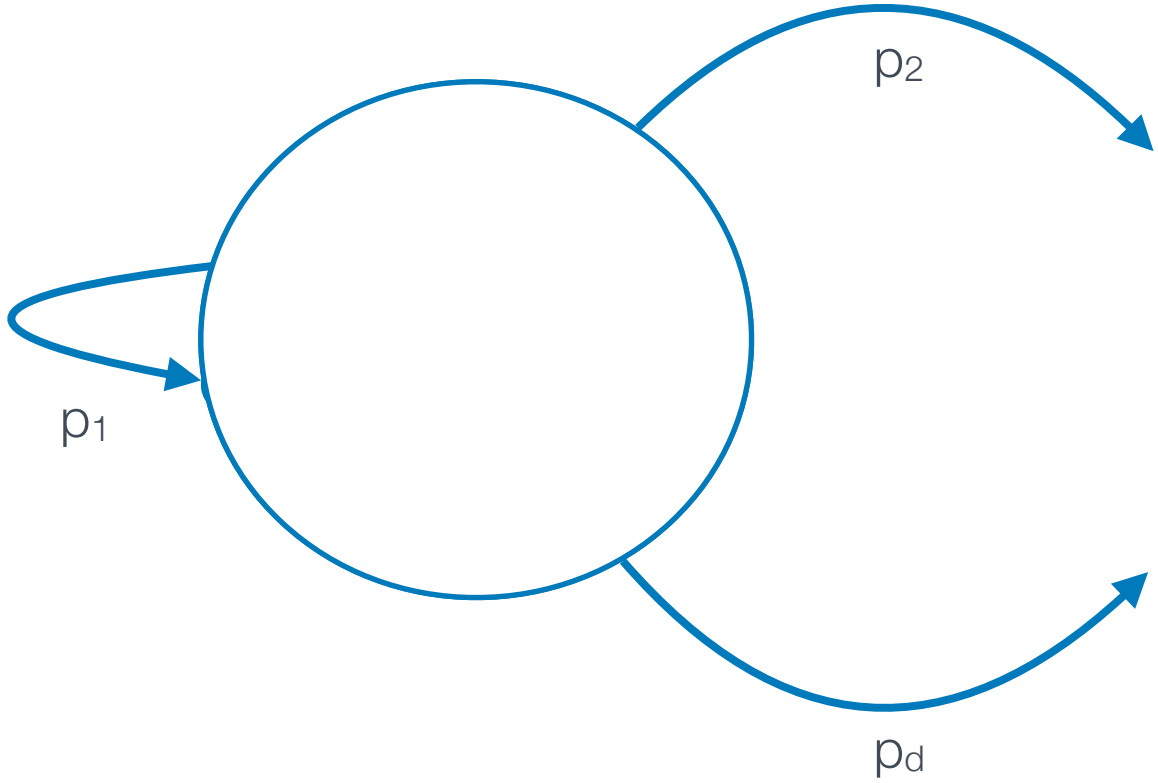
- Filtering: $p(z_t|x_{1:t})$, online
- Smoothing, MAP state: $p(z_t|x_{1:T})$ offline
- Viterbi, MAP path argmax $p(z_{1:T}|x_{1:T})$

THIS LECTURE

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

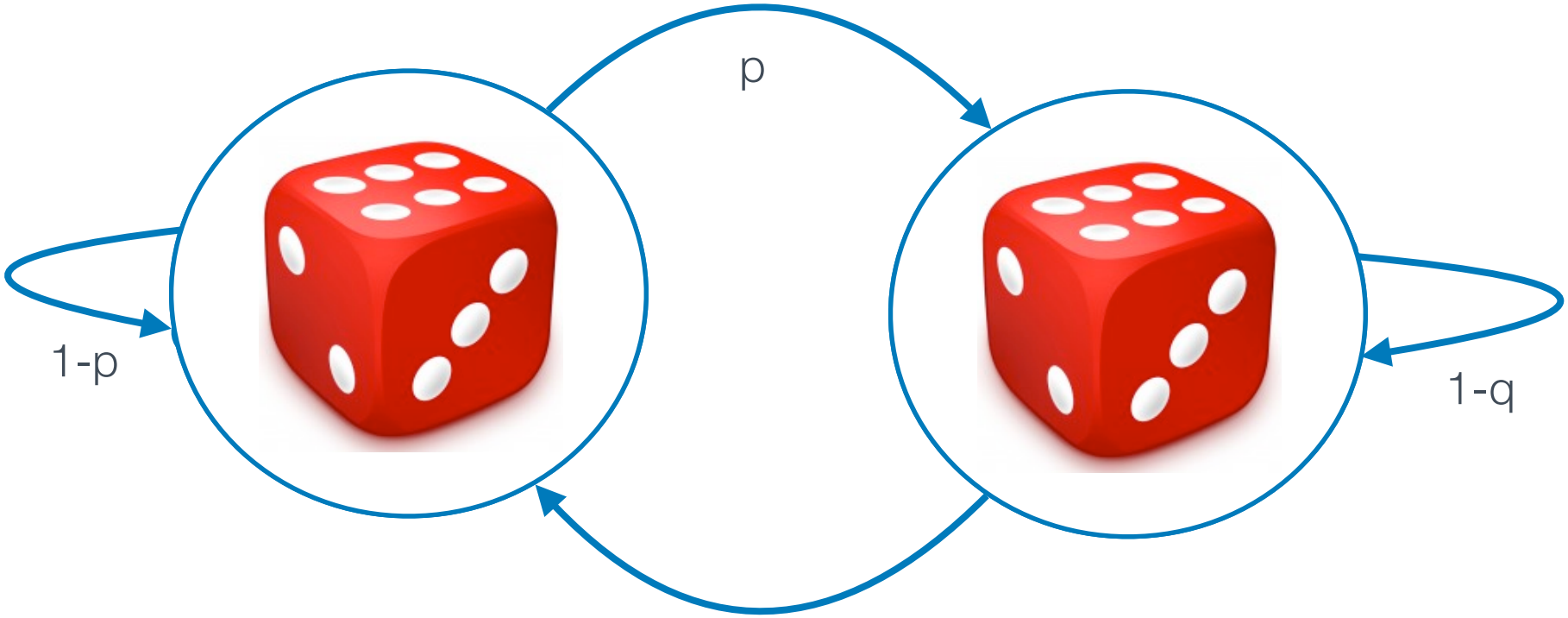
MARKOV CHAINS (DISCRETE)

Probabilities on outgoing edges sum to one



$$\sum_{i \in [d]} p_i = 1$$

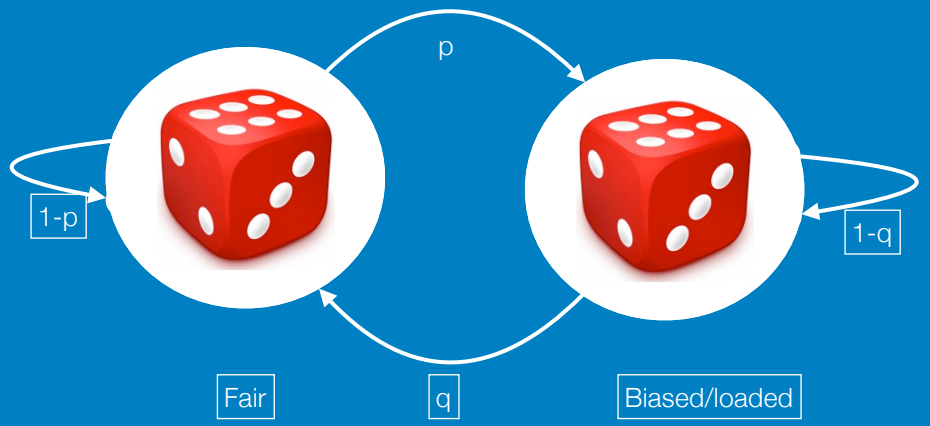
THE OCCASIONALLY DISHONEST CASINO



★ We observe the sequence of dice outcomes of visited vertices

```
Rolls:    6641532161621152346532143566342616552342323151424641566632 46
Die:      LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFLLLLLLL
```

```
Rolls:    664153216162115234653214356634261655 2
Die:      LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFF
```



WHAT AN HMM DOES

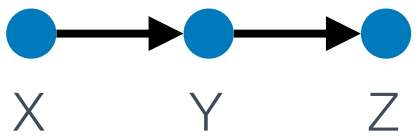
- ★ Starts in the state z_1
- ★ When in state z_t
  - outputs $p(x_t|z_t)$ B_{x_t,t_t}
  - moves to $p(z_{t+1}|z_t)$ A_{z_{t+1},z_t}
- ★ Stops after a fixed number of steps or when reaching a stop step

The parameters

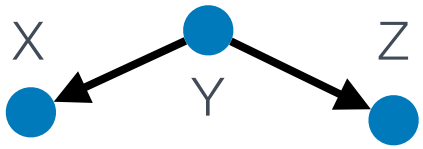
AN HMM CAN BE SEEN AS A DGM

$$Z_1 \rightarrow Z_2 \rightarrow Z_3 \rightarrow \cdots \rightarrow Z_T \rightarrow Z_{T+1}$$
$$\downarrow \quad \downarrow \quad \downarrow \quad \quad \downarrow$$
$$x_1 \quad x_2 \quad x_3 \quad \quad x_T$$

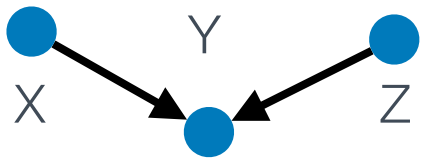
Chain



Fork

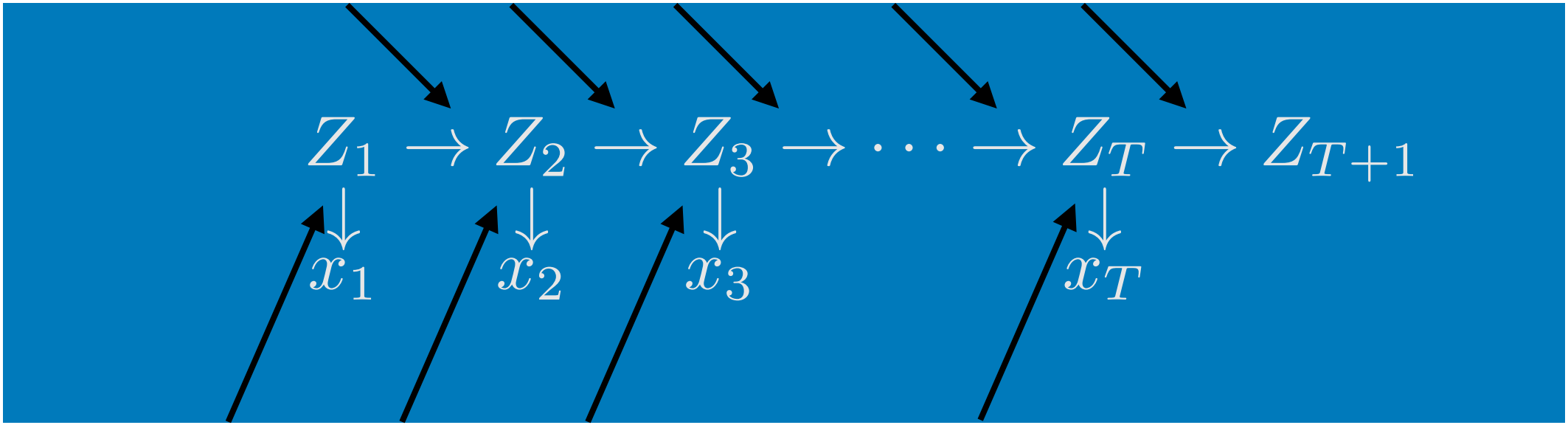


v-struct



SPECIAL CASE: HIDDEN MARKOV MODEL (HMM)

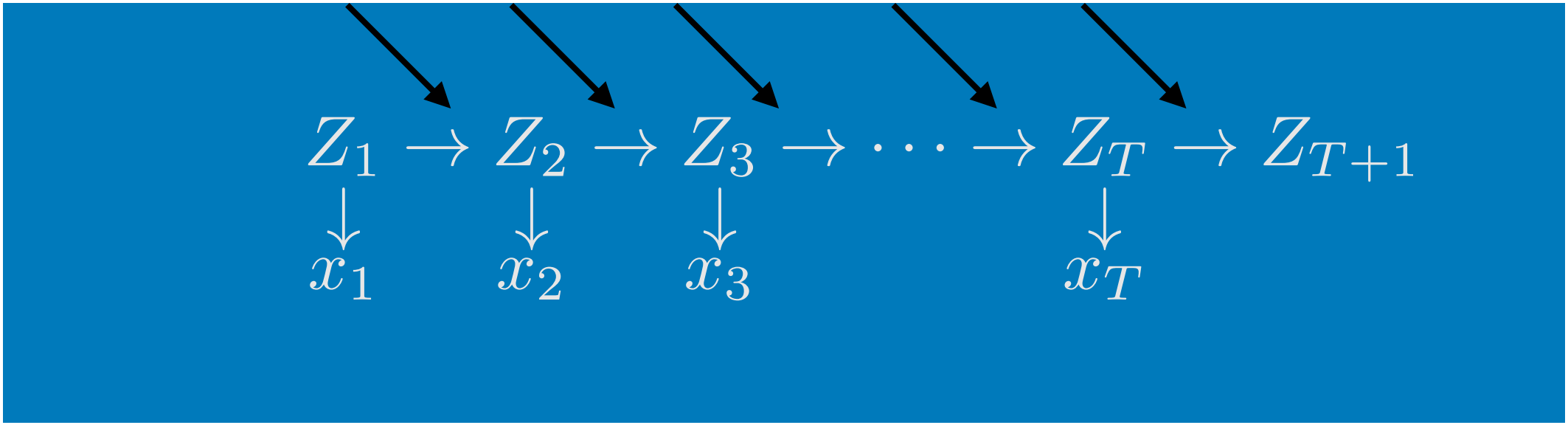
Combinations of the transition distributions



Combinations of emission the emission distribution

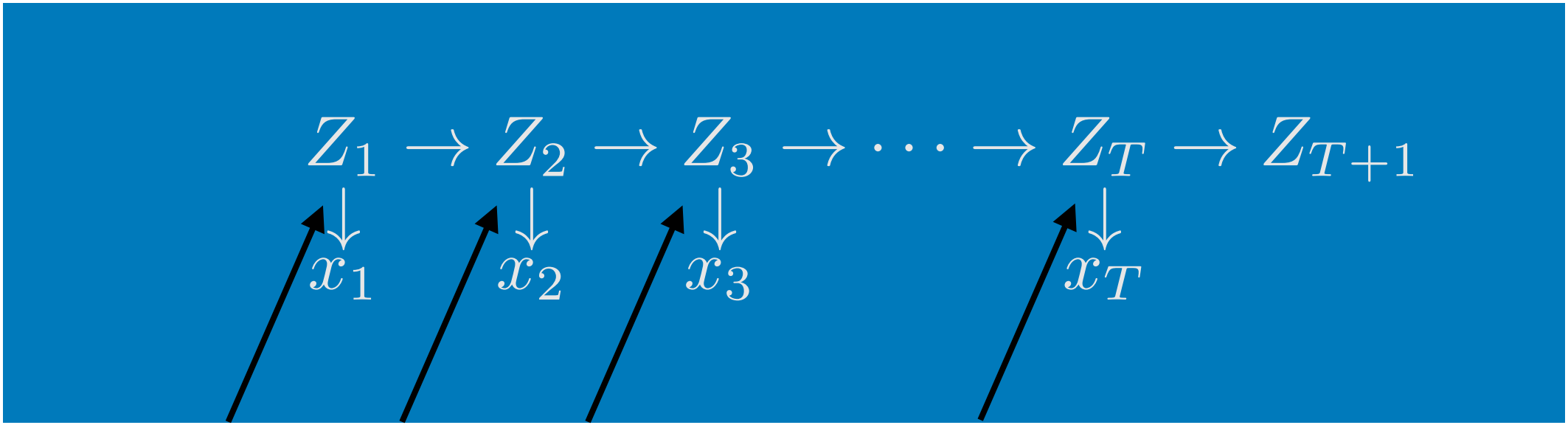
All the same

Stat	1	2	3	4
1	A_{11}	A_{21}	A_3	A_{41}
2	A_{12}	A_{22}	A_3	A_{42}
3	A_{13}	A_{23}	A_3	A_{43}
4	A_{14}	A_{24}	A_3	A_{44}



TRANSITION PROBABILITIES FOR 4 STATES HMM

EMISSION PROBABILITIES - HMM WITH 4 STATES & 3 SYMBOLS



All the same

State	1	2	3
1	B_{11}	B_{21}	B_{31}
2	B_{12}	B_{22}	B_{32}
3	B_{13}	B_{23}	B_{33}
4	B_{14}	B_{24}	B_{34}

Sum rule gives Z_{t+1}

Defined by

$$b_t(k) := p(x_{t+1:T}|Z_t = k) = \sum_{l \in [K]} p(x_{t+1:T}, Z_{t+1} = l|Z_t = k)$$

Each term in the sum is a probability of an event

$$? \rightarrow Z_t = k \rightarrow Z_{t+1} = l \rightarrow \cdots ?$$
$$\downarrow \qquad \downarrow \qquad \downarrow$$
$$? \qquad x_{t+1} \qquad x_T$$

"which is an AND of"

$$Z_t = k \rightarrow Z_{t+1} = l \qquad Z_{t+1} = l \qquad Z_{t+1} = l \rightarrow ? \rightarrow \cdots ?$$
$$\downarrow \qquad\qquad \downarrow \qquad\qquad\qquad\qquad \downarrow \qquad\qquad \downarrow$$
$$x_{t+1} \qquad\qquad x_{t+1} \qquad\qquad\qquad\qquad x_{t+2} \qquad\qquad x_T$$

Backward recursion

$$b_t(k) := p(\boldsymbol{x}_{t+1:T} | \boldsymbol{Z}_t = k)$$

- DP also for the backward variable b_t

$$b_t(k) = \sum_l \underbrace{p(\boldsymbol{Z}_{t+1} = l | \boldsymbol{Z}_t = k)}_{\text{transition}} \underbrace{b_{t+1}(l)}_{\text{"smaller"}} \underbrace{p(\boldsymbol{x}_{t+1} | \boldsymbol{Z}_{t+1} = l)}_{\text{emission}}$$

- Implementation analogous, complexity same

OFF-LINE SMOOTHING

$$p(\boldsymbol{Z}_t = k|\boldsymbol{x}_{1:T}) \propto f_t(k)\, \underbrace{p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}_{\text{emission}}\, b_t(k)$$

OFF-LINE SMOOTHING

$$p(\boldsymbol{Z}_t = k|\boldsymbol{x}_{1:T}) \propto f_t(k)\, \underbrace{p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}_{\text{emission}}\, b_t(k)$$

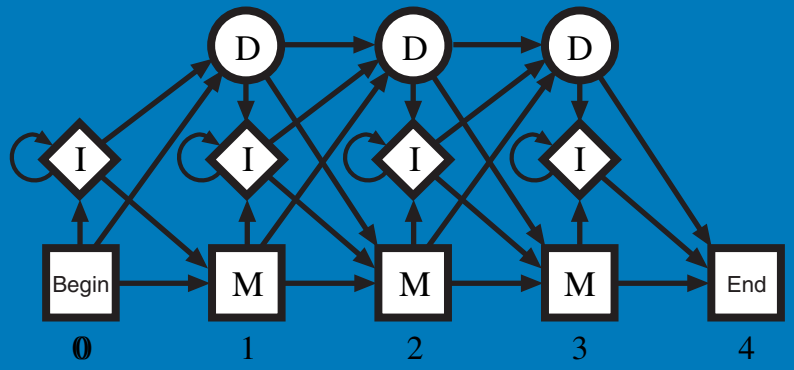
Up to a constant

TWO SLICED SMOOTH MARGINALS - MARGINAL OVER PAIRS OF STATES

$$p(\boldsymbol{Z}_t = k, \boldsymbol{Z}_{t+1} = l | \boldsymbol{x}_{1:T})$$

- Can be computed from forward and backward similarly

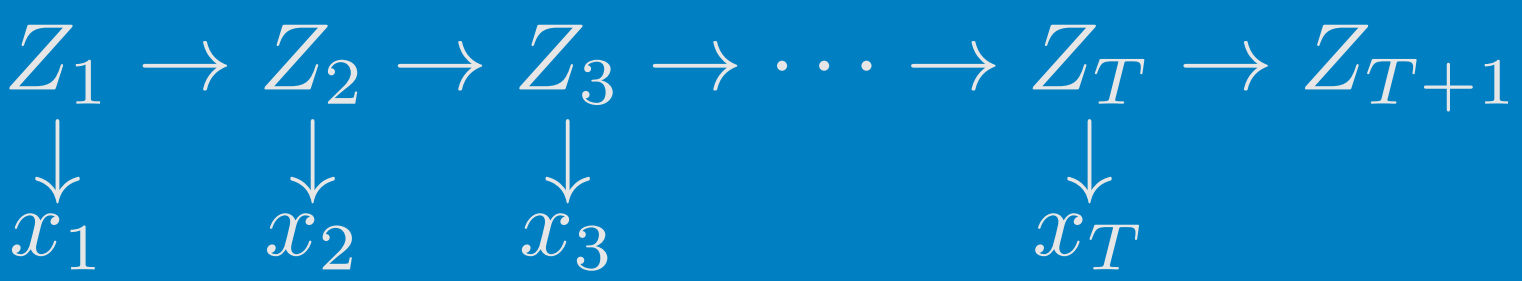
SAMPLING FROM POSTERIOR



How much did each previous state contribute to the probability mass of the present state?

Forward recursion

$$f_t(k) = \sum_l \underbrace{f_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1}=l)}_{\text{emission } B_{x_{t-1},l}} \underbrace{p(\boldsymbol{Z}_t=k|\boldsymbol{Z}_{t-1}=l)}_{\text{transition } A_{lk}}$$



BACKWARDS SAMPLING OF POSTERIOR

Sample $z^s_{1:T+1} \sim p(\boldsymbol{Z}_{1:T+1} = k|\boldsymbol{x}_{1:T})$ by

(i) $$z^s_{T+1} \sim p(\boldsymbol{Z}_{T+1} = k|\boldsymbol{x}_{1:T}) \propto p(\boldsymbol{Z}_{T+1} = k, \boldsymbol{x}_{1:T})$$

(ii) $$z^s_t \sim p(\boldsymbol{Z}_t = k|\boldsymbol{Z}_{t+1} = l, \boldsymbol{x}_{1:t})$$
$$\propto f_t(k)p(\boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k)p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)$$

- Sample from posterior
- Sample in order $z_T,\ldots,z_1$
- Start somewhat differently

We want

$\text{argmax}_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$

Not!

$(\text{argmax}_{\boldsymbol{z}_1} p(\boldsymbol{z}_1|\boldsymbol{x}_{t+1:T}), \ldots, \text{argmax}_{\boldsymbol{z}_T} p(\boldsymbol{z}_T|\boldsymbol{x}_{t+1:T}))$

Viterbi variable

$$v_t(k) := \max_{\boldsymbol{z}_{1:t-1}} p(\boldsymbol{z}_{1:t-1}, \boldsymbol{Z}_t = k, \boldsymbol{x}_{1:t})$$

It gives what we want

$$\max_k v_T(k)$$

VITERBI

- MAP path
- Viterbi learning: used, as approximation, to speed up parameter learning
- Again DP now with Viterbi variable
- For the path, use back pointers

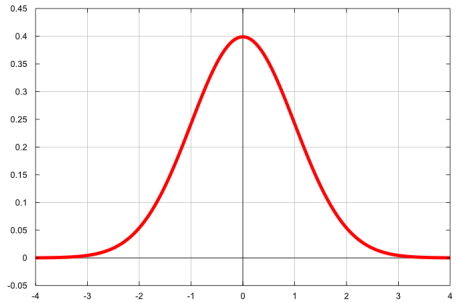
Chapter 3

EXPECTATION-MAXIMIZATION THEORY

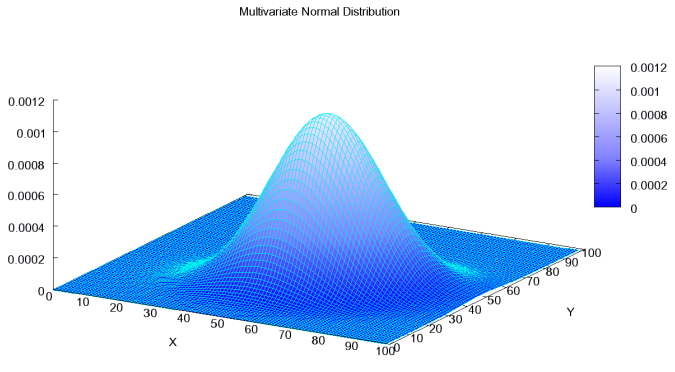
3.1 Introduction

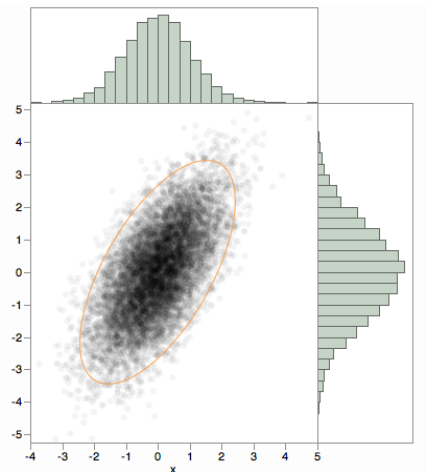
Learning networks are commonly categorized in terms of supervised and unsupervised networks. In unsupervised learning, the training set consists of input training patterns only. In contrast, in supervised learning networks, the training data consist

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

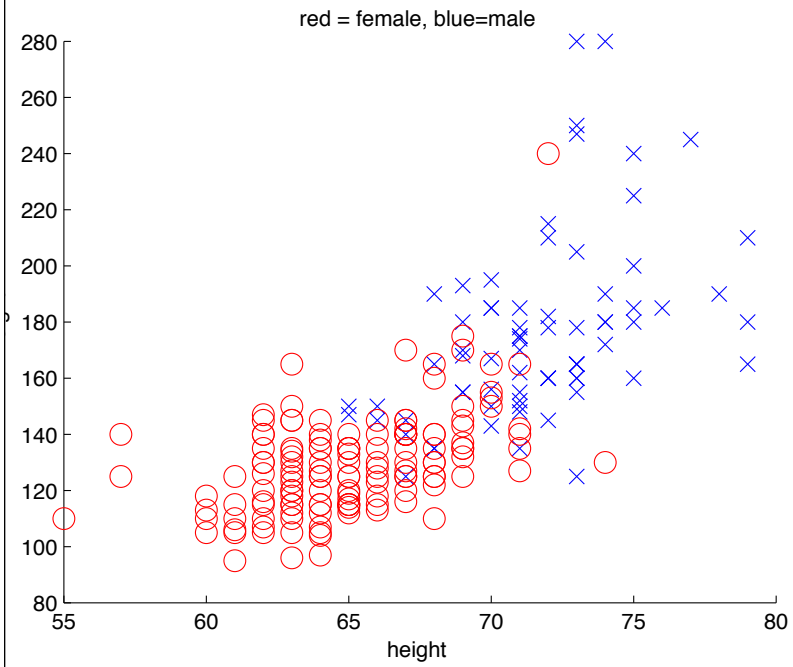


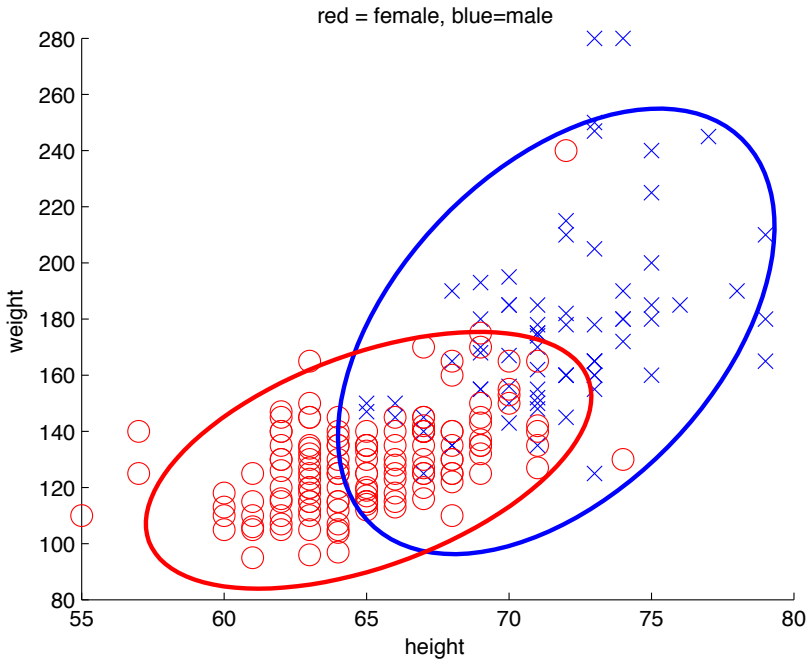
$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$





GAUSSIAN – MVN





TWO DIMENSIONAL NORMAL

K-MEANS

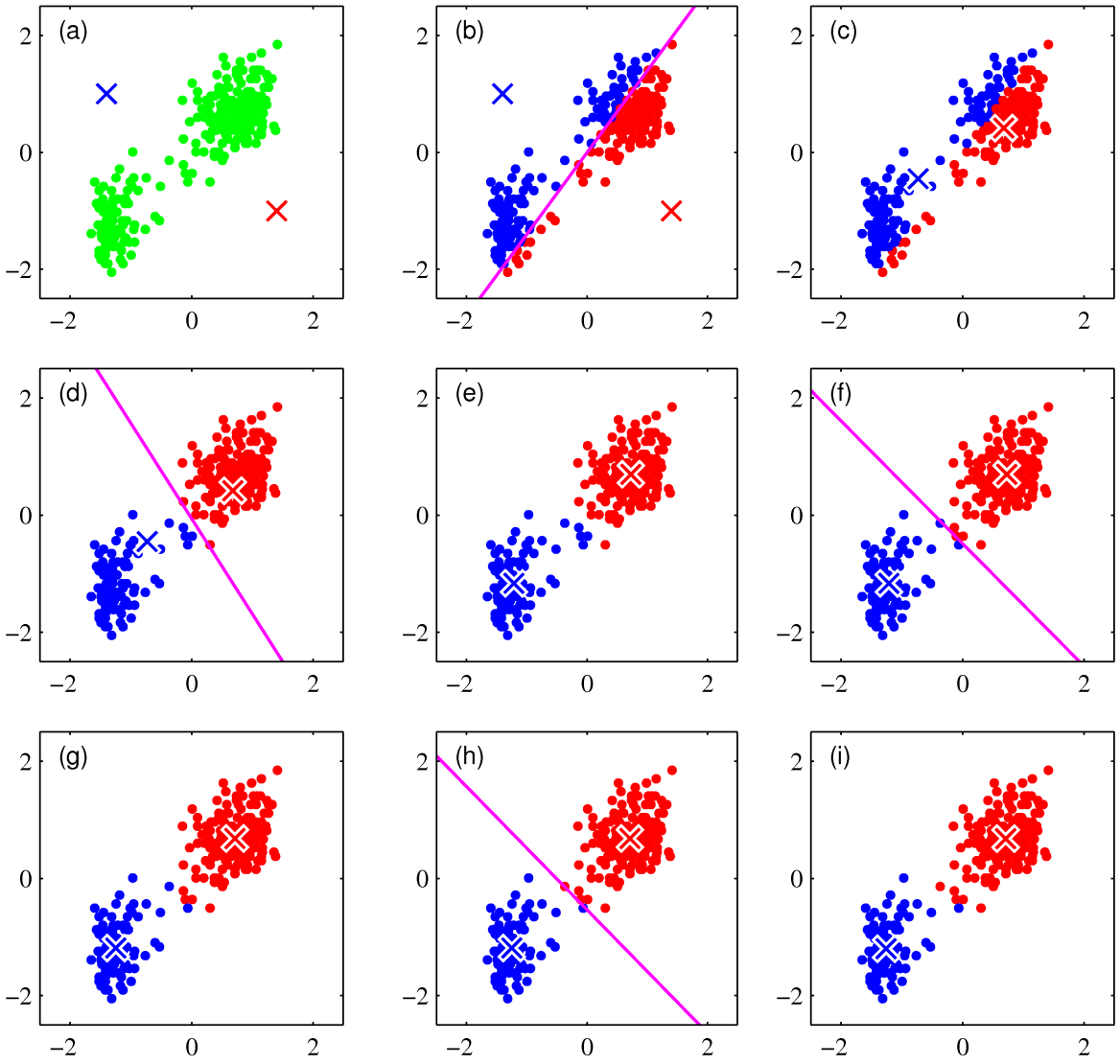
- ★ Data vectors $D=\{x_1,\ldots,x_N\}$
- ★ Randomly selected classes $z_1,\ldots,z_N$
- ★ Iteratively do

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n:z_n=c} \boldsymbol{x}_n, \qquad \text{where } N_c = |\{n : z_n = c\}|$$

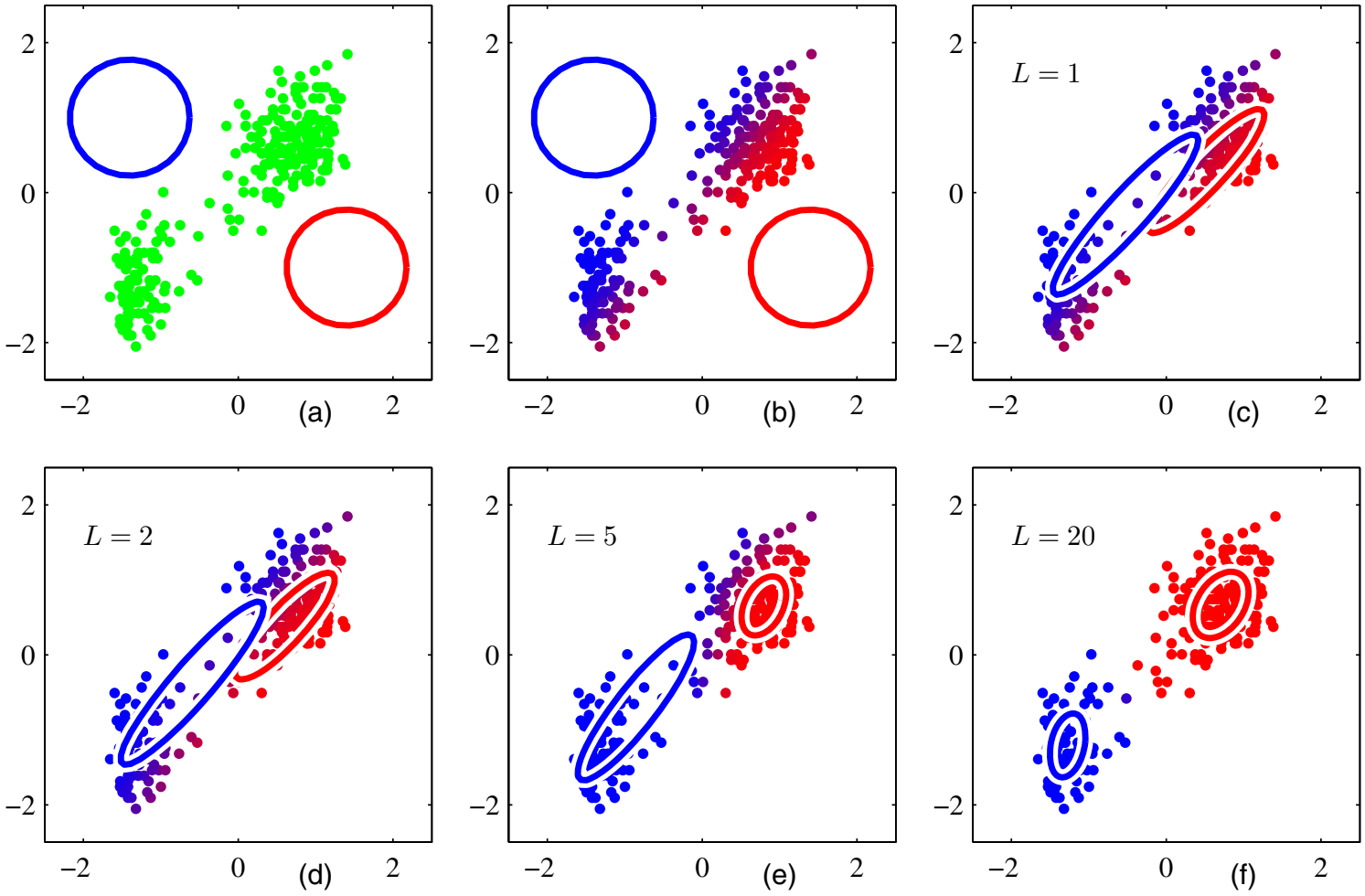
$$z_n = \text{argmin}_c ||\boldsymbol{x}_n - \boldsymbol{\mu}_c||_2$$

- ★ One step O(NKD), can be improved

ASSIGN POINT TO MEANS



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

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

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

IMAGE SEGMENTATION WITH K-MEANS

$K = 2$

$K = 3$

$K = 10$

Original image

GAUSSIAN MIXTURE MODELS (GMM)

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

$\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nD})$

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

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

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

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

1-DIM GAUSSIAN MIXTURE MODELS

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

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

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

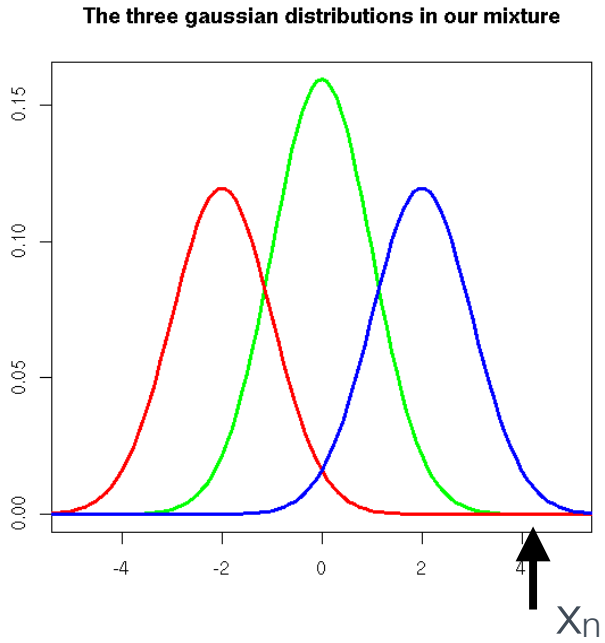
$\boldsymbol{\theta}_c = (\boldsymbol{\mu}_c, \sigma_c)$

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

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

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

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

MLE FOR GAUSSIAN

…and $\sum_{n:I(z_n=c)} \log \frac{1}{\sqrt{2\pi\sigma_c'^2}} \exp\left(-\frac{1}{2\sigma_c'^2}(x_n - \mu_c')^2\right)$

is maximized by

by $\mu_c' = \frac{\sum_{n:I(z_n=c)} x_n}{N_c}$ where $N_c = \sum_n I(z_n = c)$

EM & EXPECTED LOG LIKELIHOOD (Q-TERM)

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

THE END