ECE521 Lecture 19

HMM cont.

Inference in HMM

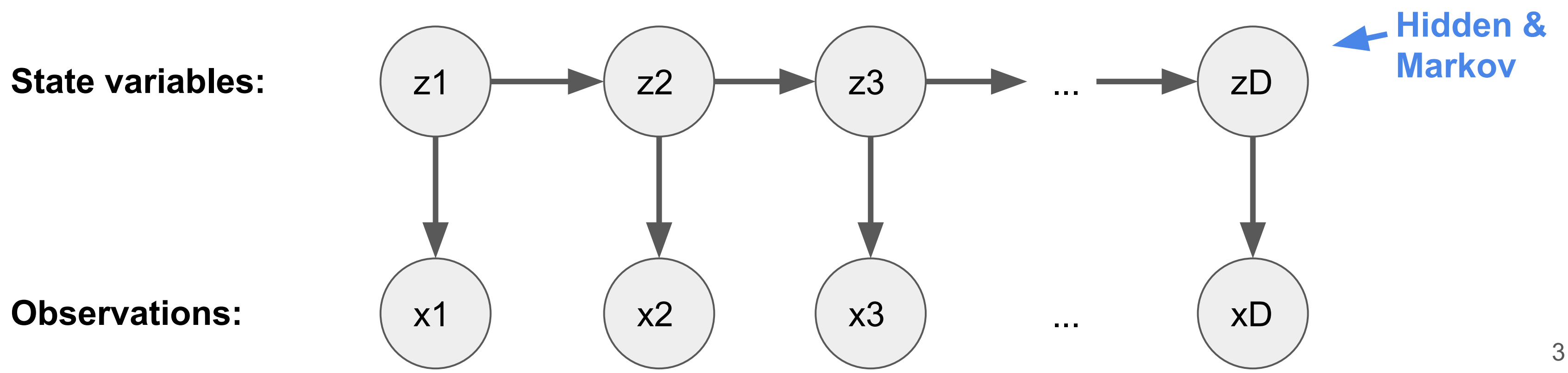
UNIVERSITY OF TORONTO

Outline

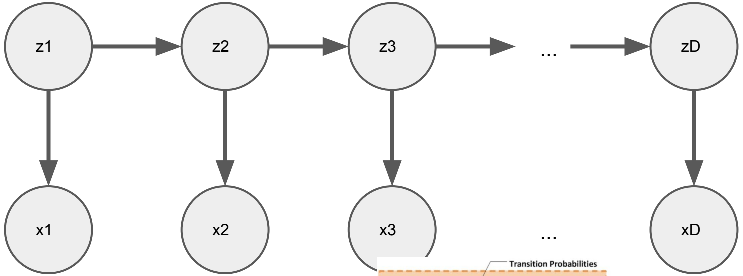
- **Hidden Markov models**
  - Model definitions and notations
  - Inference in HMMs
  - Learning in HMMs

Hidden Markov models

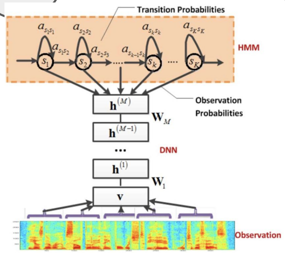
- Formally, a **hidden Markov model** defines a generative process on a sequence of observed random variables $\{x_1, \ldots, x_D\}$ through its corresponding latent sequence $\{z_1, \ldots, z_D\}$. HMMs are **generative models**.
    - **state**: z_d **observation/input**: x_d

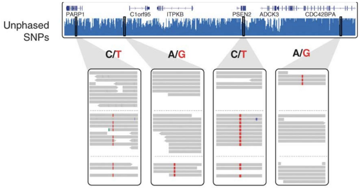


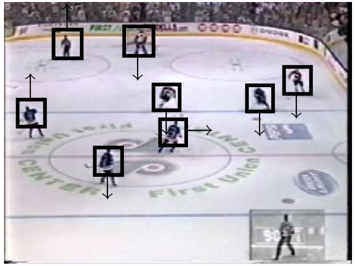
Hidden Markov models



- Applications of HMMs:
  - Speech recognition: Cortana, Siri, Google Assistant (before 2016)
    - inputs: observed Fourier coeff.   states: utterances
  - Bioinformatics: GeneMark and its variance
    - inputs: raw DNA sequence   states: protein-coding region or not
  - Communication: coding theory, error-correction-codes
    - inputs: raw bit stream   states: corrected bit stream
  - Tracking: Kalman filtering, particle filtering
    - inputs: raw sensory data   states: location, velocity, pose...



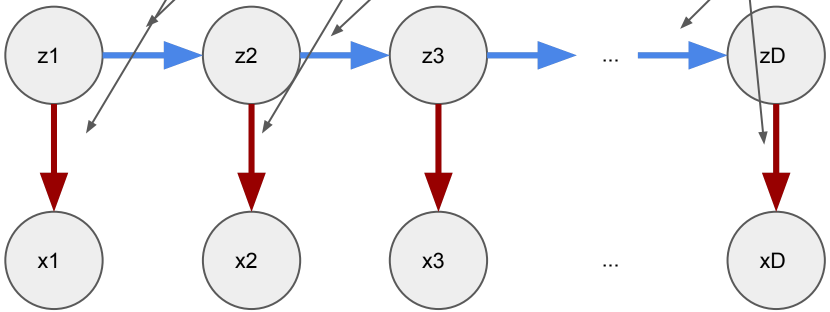




Source: Dahl et al. 2012; Kitzman et al 2010; Lu et al. 2007

Hidden Markov models

- Definition:
    - **Observations**: $\{x_1, \ldots, x_D\}$
    - **States**: $\{z_1, \ldots, z_D\}$
    - **State transition probabilities**: $\{p(z_2|z_1), p(z_3|z_2), \ldots, p(z_D|z_{D-1})\}$
    - **Emission probabilities:** $\{p(x_1|z_1), p(x_2|z_2), \ldots, p(x_D|z_D)\}$

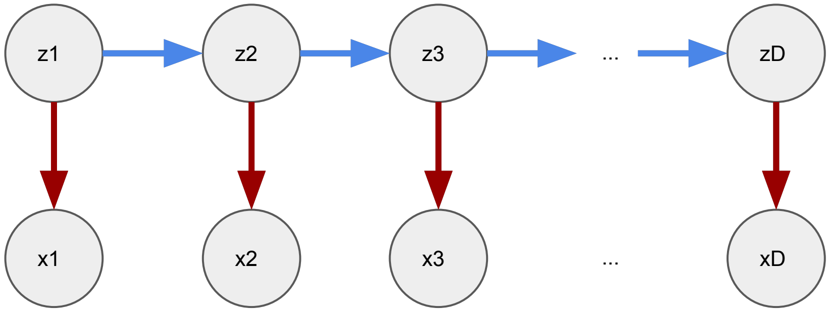


Hidden Markov models

- Definition:
  - **Observations**: $\{x_1, \ldots, x_D\}$
  - **States**: $\{z_1, \ldots, z_D\}$
  - **State transition probabilities:** $p(z_t | z_{t-1})$ (shared across the sequence/timesteps)
  - **Emission probabilities:** $p(x_t | z_t)$ (shared across the sequence/timesteps)

Sequence modelling assumption:
sharing conditional probability distribution.

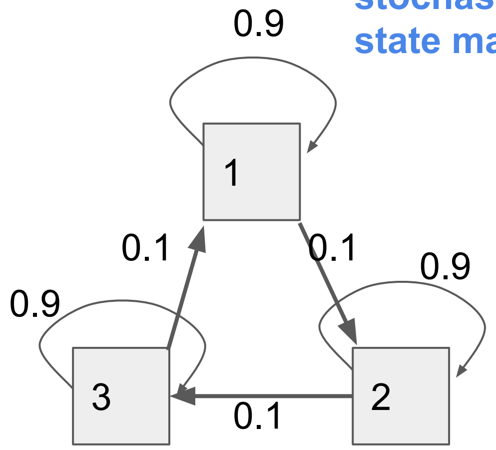
We gain computational efficiency:
state transition memory requirement
O(D*|Z|^2) vs. O(|Z|^2)



Hidden Markov models

- Example1:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - We have the following transition probabilities

$p(z_t \mid z_{t-1})$	$z_t = 1$	2	3
$z_{t-1} = 1$	0.9	0.1	0
2	0	0.9	0.1
3	0.1	0	0.9

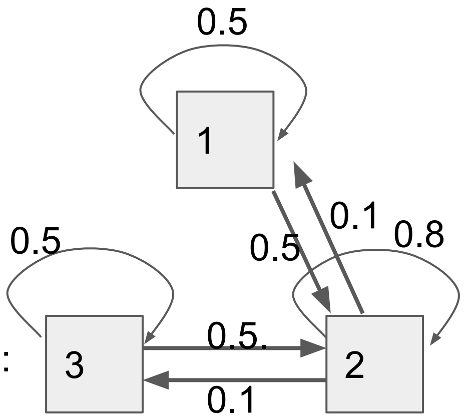


stochastic finite state machine

Hidden Markov models

- Example2:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - We have the following transition probabilities

$p(z_t \mid z_{t-1})$	$z_t = 1$	2	3
$z_{t-1} = 1$	0.5	0.5	0
2	0.1	0.8	0.1
3	0	0.5	0.5



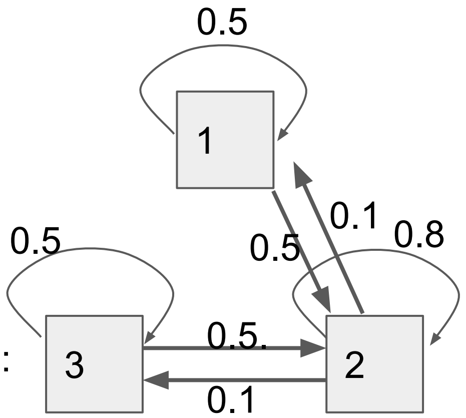
The transition probability defines a degree 2 Markov model.
Generate states: we can generate a sequence of states as the following:

- Consider an initial state z1 = 2
- $p(z_2 \mid z_1 = 2) = [0.1, 0.8, 0.1]^T$, we can sample z2 and repeat the process for the next timestep.

Hidden Markov models

- Example2:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - We have the following transition probabilities

$p(z_t \mid z_{t-1})$	$z_t = 1$	2	3
$z_{t-1} = 1$	0.5	0.5	0
2	0.1	0.8	0.1
3	0	0.5	0.5



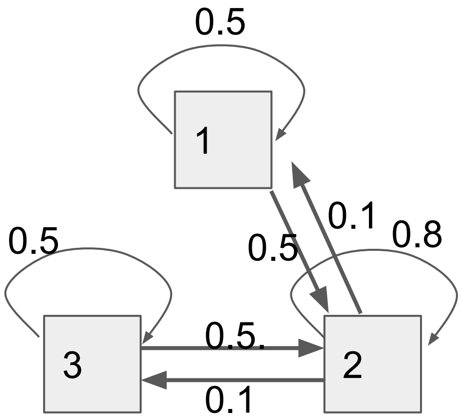
The transition probability defines a degree 2 Markov model.
Generate states: we can generate a sequence of states as the following:

- Consider an initial state z1 = 2
- The samples of a sequence hidden states: 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2, ...

Hidden Markov models

- Example2:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - We have the following transition probabilities

$p(z_t \mid z_{t-1})$	$z_t = 1$	2	3
$z_{t-1} = 1$	0.5	0.5	0
2	0.1	0.8	0.1
3	0	0.5	0.5



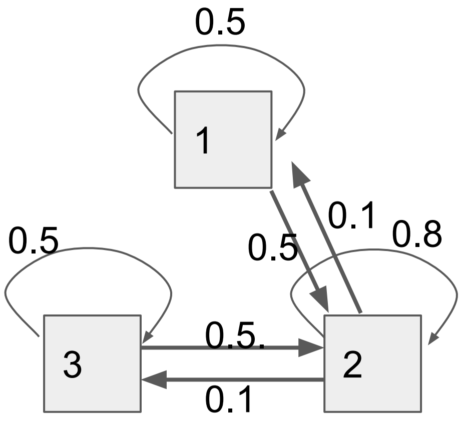
Prediction: we can even predict the future states given the current state using the transition probability

- Consider the following state sequence: 2, 2, 2, 2, 2, 3, 3, 2, 2, ...
- What is the next states in the sequence? $p(z_{10} | z_9 = 2) = [0.1, 0.8, 0.1]^T$

Hidden Markov models

- Example2:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - We have the following transition probabilities

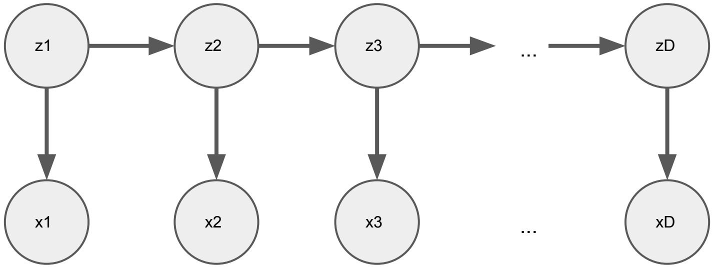
$$P = \begin{array}{r|rrr} p(z_t|z_{t-1}) & z_t = 1 & 2 & 3 \\ \hline z_{t-1} = 1 & 0.5 & 0.5 & 0 \\ 2 & 0.1 & 0.8 & 0.1 \\ 3 & 0 & 0.5 & 0.5 \end{array}$$



Prediction: we can even predict the future states given the current state using the transition probability

- Consider the following state sequence: 2, 2, 2, 2, 2, 3, 3, 2, 2, ...
- What is the 15th hidden state in the sequence? $p(z_{15}|z_9 = 2) = (P^T)^6[0, 1, 0]^T = [0.14, 0.72, 0.14]^T$

Hidden Markov models



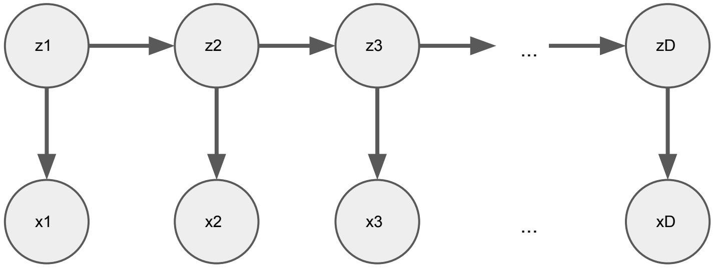
- Example3:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - The observations are also discrete random variables: $x_t \in \{A, C, G, T\}$
  - We have the following emission probabilities:

$p(x_t \mid z_t)$	$x_t =$ A	C	G	T
$z_t = 1$	0.3	0.1	0	0.6
2	0	0.8	0.2	0
3	0	0.1	0.9	0

Generate observations: let the state be fixed at z=2. A sequence of observations can be generated(emitted) from the emission probability distribution by sampling x: $x_t \sim p(x_t \mid z_t = 2)$

a particular realization:	C	G	C	C	C	C	C	C	C	G	C	C	C	C...
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14} $\ldots$

Hidden Markov models



- Example3:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - The observations are also discrete random variables: $x_t \in \{A, C, G, T\}$
  - We have the following emission probabilities:

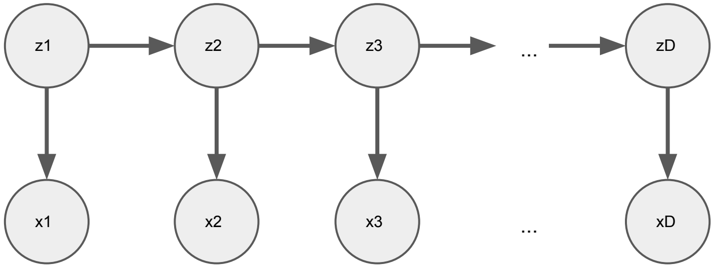
$p(x_t \mid z_t)$	$x_t =$ A	C	G	T
$z_t = 1$	0.3	0.1	0	0.6
2	0	0.8	0.2	0
3	0	0.1	0.9	0

Generate observations: let the state be fixed at z=1. A sequence of observations can be generated(emitted) from the emission probability distribution by sampling x: $x_t \sim p(x_t \mid z_t = 1)$

a particular realization:

T	T	T	A	T	T	A	A	T	T	T	A	T	T...	
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	$\ldots$

Hidden Markov models



- Example4:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - The observations are also discrete random variables: $x_t \in \{A, C, G, T\}$
  - We have the following transition and emission probabilities:

$p(z_t \mid z_{t-1})$	$z_t = 1$	2	3
$z_{t-1} = 1$	0.5	0.5	0
2	0.1	0.8	0.1
3	0	0.5	0.5

$p(x_t \mid z_t)$	$x_t =$ A	C	G	T
$z_t = 1$	0.3	0.1	0	0.6
2	0	0.8	0.2	0
3	0	0.1	0.9	0

Generate states and observations:

ancestral sampling: $z_t \sim p(z_t \mid z_{t-1})$
$x_t \sim p(x_t \mid z_t)$

a particular realization:

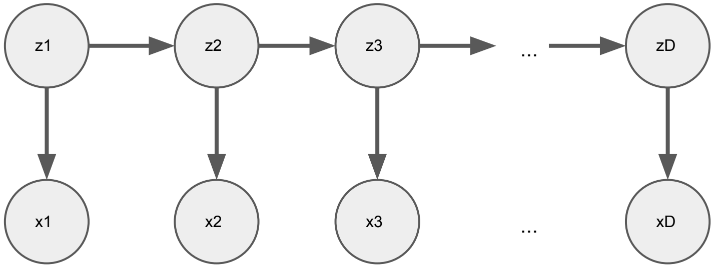
2 2 2 2 2 3 3 2 2 2 2 3 2 2 ...

$z_1 \; z_2 \; z_3 \; z_4 \; z_5 \; z_6 \; z_7 \; z_8 \; z_9 \; z_{10} \; z_{11} \; z_{12} \; z_{13} \; z_{14} \; \ldots$

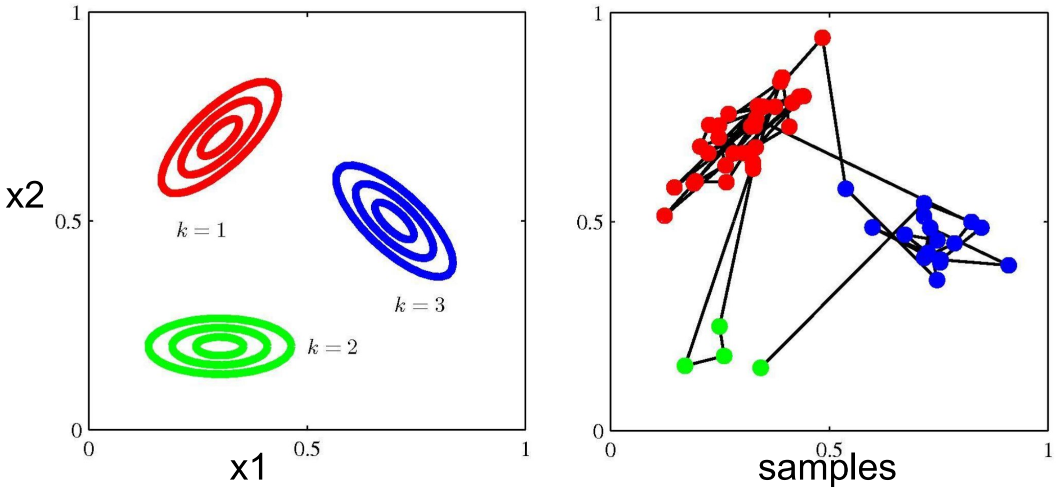
C G C C C G G C C G C G C C ...

$x_1 \; x_2 \; x_3 \; x_4 \; x_5 \; x_6 \; x_7 \; x_8 \; x_9 \; x_{10} \; x_{11} \; x_{12} \; x_{13} \; x_{14} \; \ldots$

Hidden Markov models

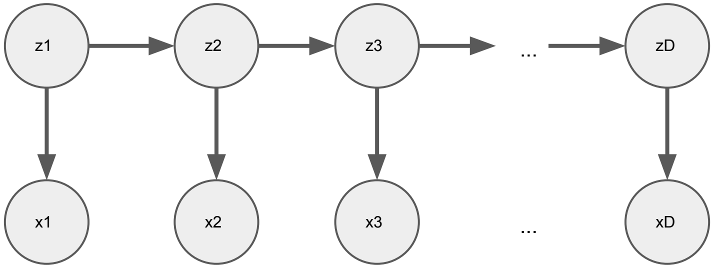


- Example5:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - The observations are 2D Gaussians: $x_t \in \mathbb{R}^2$



Source: Bishop 2008

Hidden Markov models



- Example6:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - The observations are also discrete random variables: $x_t \in \{A, C, G, T\}$
  - We have the following transition and emission probabilities:

$p(z_t \mid z_{t-1})$	$z_t = 1$	2	3
$z_{t-1} = 1$	0.5	0.5	0
2	0.1	0.8	0.1
3	0	0.5	0.5

$p(x_t \mid z_t)$	$x_t = \mathrm{A}$	C	G	T
$z_t = 1$	0.3	0.1	0	0.6
2	0	0.8	0.2	0
3	0	0.1	0.9	0

Inference:
What are latent states that generate the given observations?

Given a sequence of observations:

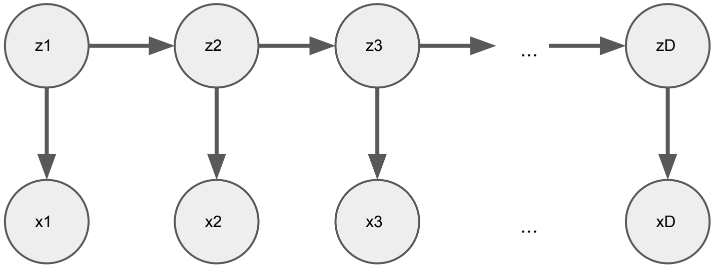
? ? ? ? ? ? ? ? ? ? ? ? ? ? ...

$z_1\ z_2\ z_3\ z_4\ z_5\ z_6\ z_7\ z_8\ z_9\ z_{10}\ z_{11}\ z_{12}\ z_{13}\ z_{14}\ \ldots$

C C C C G C C T T A C G C C ...

$x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{10}\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ \ldots$

Hidden Markov models



- Example6:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - The observations are also discrete random variables: $x_t \in \{A, C, G, T\}$
  - We have the following transition and emission probabilities:

$p(z_t \mid z_{t-1})$	$z_t = 1$	2	3
$z_{t-1} = 1$	0.5	0.5	0
2	0.1	0.8	0.1
3	0	0.5	0.5

$p(x_t \mid z_t)$	$x_t = $ A	C	G	T
$z_t = 1$	0.3	0.1	0	0.6
2	0	0.8	0.2	0
3	0	0.1	0.9	0

Inference:
What are latent states that generate the given observations?

Given a sequence of observations:

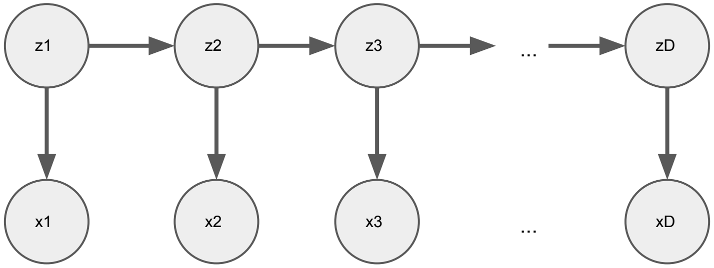
? ? ? ? ? ? ? ? ? ? ? ? ? ? ...

z_1 z_2 z_3 z_4 z_5 z_6 z_7 z_8 z_9 z_{10} z_{11} z_{12} z_{13} z_{14} $\ldots$

C C C C G C C T T A C G C C ...

x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12} x_{13} x_{14} $\ldots$

Hidden Markov models



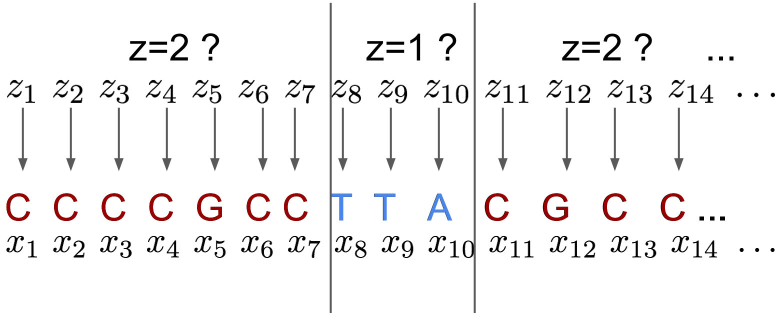
- Example6:
  - Consider an HMM with three discrete states: $z_t \in \{1, 2, 3\}$
  - The observations are also discrete random variables: $x_t \in \{A, C, G, T\}$
  - We have the following transition and emission probabilities:

$p(z_t \mid z_{t-1})$	$z_t = 1$	2	3
$z_{t-1} = 1$	0.5	0.5	0
2	0.1	0.8	0.1
3	0	0.5	0.5

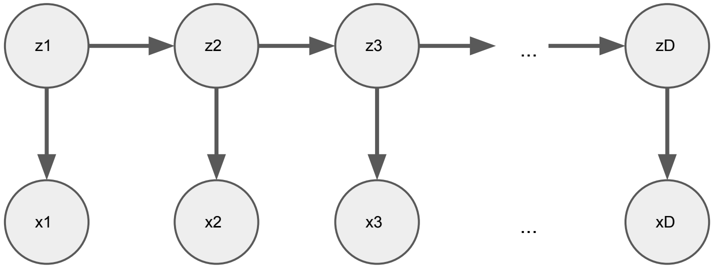
$p(x_t \mid z_t)$	$x_t = \mathrm{A}$	C	G	T
$z_t = 1$	0.3	0.1	0	0.6
2	0	0.8	0.2	0
3	0	0.1	0.9	0

Inference:
What are latent states that generate the given observations?

Given a sequence of observations:



Hidden Markov models



- We can mathematically reason about the inference problems using the joint probability of the HMM and the Bayes rule.
  - Joint distributions of the observations and the latent states in an HMM:

$$p(x_1, \dots, x_D, z_1, \dots, z_D) = p(z_1) \underbrace{\prod_{t=2}^{D} p(z_t | z_{t-1})}_{\text{transition}} \underbrace{\prod_{t=1}^{D} p(x_t | z_t)}_{\text{emission}}$$

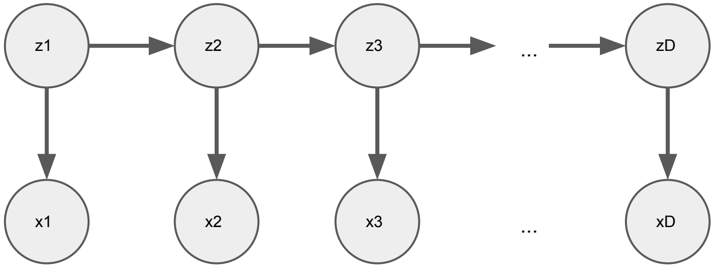
  - **Inference**: the conditional distribution (posterior) over the latent states given the observations.

$$p(z_1, \dots, z_D | x_1, \dots, x_D) = \frac{p(x_1, \dots, x_{t+1}, z_1, \dots, z_{t+1})}{\sum_{z_1, \dots, z_D} p(x_1, \dots, x_{t+1}, z_1, \dots, z_{t+1})}$$

  - **Prediction**: the marginal distribution of the next observation given the previous observations.

$$p(x_{t+1} | x_1, \dots, x_t) = \frac{\sum_{z_1, \dots, z_{t+1}} p(x_1, \dots, x_{t+1}, z_1, \dots, z_{t+1})}{\sum_{x_{t+1}, z_1, \dots, z_{t+1}} p(x_1, \dots, x_{t+1}, z_1, \dots, z_{t+1})}$$

Inference in HMMs



- **Inference** and **marginalization** (and **generation**) are the fundamental operations performed on a graphical model. They are the two sides of the same coin. We have to perform marginalization for inference.
  - E.g. given the joint dist. $p(x_1, \ldots, x_D, z_1, \ldots, z_D) = p(z_1) \prod_{t=2}^{D} p(z_t | z_{t-1}) \prod_{t=1}^{D} p(x_t | z_t)$
  - Computing the marginal distribution of the observations requires sum over the latent states:

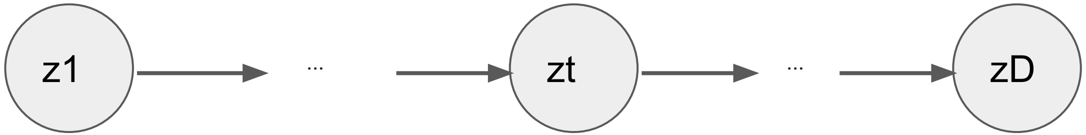
$$p(x_1, \ldots, x_D) = \sum_{z_1, \ldots, z_D} p(x_1, \ldots, x_{t+1}, z_1, \ldots, z_{t+1})$$
$$= \sum_{z_1, \ldots, z_D} p(z_1) \prod_{t=2}^{D} p(z_t | z_{t-1}) \prod_{t=1}^{D} p(x_t | z_t)$$

  - The marginal dist. is used for normalizing the posterior inference:

$$p(z_1, \ldots, z_D | x_1, \ldots, x_D) = \frac{p(x_1, \ldots, x_D, z_1, \ldots, z_D)}{p(x_1, \ldots, x_D)}$$

Sidetracking: Inference in HMMs

- Let us consider a simpler marginalization problem first. Recall the degree 2 Markov model that has a chain structure:



  - Joint distribution: $p(z_1, \ldots, z_D) = p(z_1) \prod_{t=2}^{D} p(z_t|z_{t-1})$
  - Suppose we would like to obtain the marginal distribution of zt that is

$$p(z_t) = \sum_{z1,\ldots,z_{t-1},z_{t+1},\ldots,z_D} p(z_1, \ldots, z_D)$$

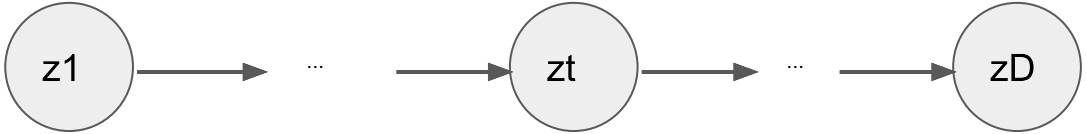
$$= \sum_{z1,\ldots,z_{t-1},z_{t+1},\ldots,z_D} p(z_1) \prod_{t=2}^{D} p(z_t|z_{t-1})$$

Is there any shortcut to perform this summation?

Insight: each term in the product depends only on a subset of the marginalized variables!

Sidetracking: Inference in HMMs

- The shortcut idea is **distribute** the summations into the product:



  - First, we make the summation over each random variable explicit

$$p(z_t) = \sum_{z_1,\ldots,z_{t-1},z_{t+1},\ldots,z_D} p(z_1) \prod_{t=2}^{D} p(z_t|z_{t-1})$$

$$= \sum_{z_1} \sum_{z_2} \cdots \sum_{z_{t-1}} \sum_{z_{t+1}} \cdots \sum_{z_D} p(z_1)p(z_2|z_1)\ldots p(z_D|z_{D-1})$$

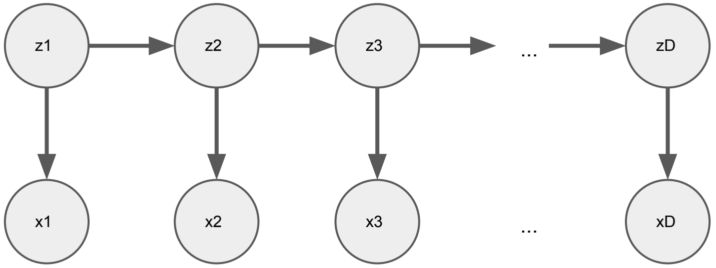
This is the intuition behind the **message-passing algorithms**

  - Then, perform **local summation** by distributing each sum into the products:

Can we simplify this further?

$$= \sum_{z_D} \Big\{ \sum_{z_{D-1}} \cdots \Big\{ \sum_{z_{t+1}} \Big\{ \sum_{z_{t-1}} \cdots \Big\{ \sum_{z_2} \Big\{ \sum_{z_1} p(z_1)p(z_2|z_1) \Big\} p(z_3|z_2) \Big\} \ldots p(z_{t-1}|z_{t-2}) \Big\} p(z_{t+1}|z_t) \Big\} \ldots p(z_D|z_{D-1})$$

Inference in HMMs



- We can then perform the same distribution trick to obtain the marginals in HMMs:

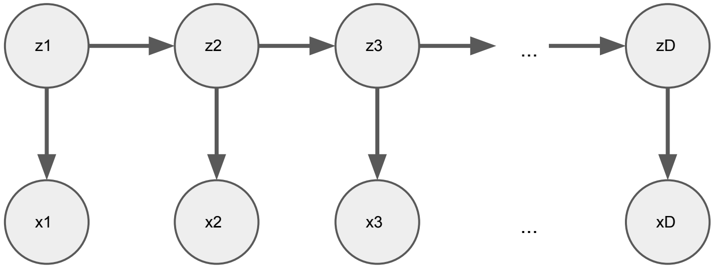
$$p(x_1, \ldots, x_D) = \sum_{z_1, \ldots, z_D} p(z_1) \prod_{t=2}^{D} p(z_t | z_{t-1}) \prod_{t=1}^{D} p(x_t | z_t)$$

  - Then, perform **local summation** by distributing each sum into the products:

$$= \sum_{z_D} \left\{ \ldots \left\{ \sum_{z_2} \left\{ \sum_{z_1} p(z_1) p(x_1|z_1) p(z_2|z_1) \right\} p(x_2|z_2) p(z_3|z_2) \right\} \ldots p(z_D|z_{D-1}) \right\} p(x_D|z_D)$$

Additional emission distribution comparing to the plain Markov model

Learning in HMMs



- Learning in HMMs is the same with all the other graphical models. For HMMs, we would like to like to ***adapt*** the parameters in the ***transition*** and the ***emission*** probability distributions.
    - First, obtain the marginal likelihood of the data/observations by performing marginalization over the latent state variables

$$p(x_1, \ldots, x_D) = \sum_{z_1, \ldots, z_D} p(x_1, \ldots, x_{t+1}, z_1, \ldots, z_{t+1})$$

    - Then, perform maximum likelihood estimation (MLE) through **gradient descent** by following the gradient of the log marginal likelihood:

$$\theta \leftarrow \theta + \eta \frac{\partial \log p(x_1, \ldots, x_D)}{\partial \theta}$$