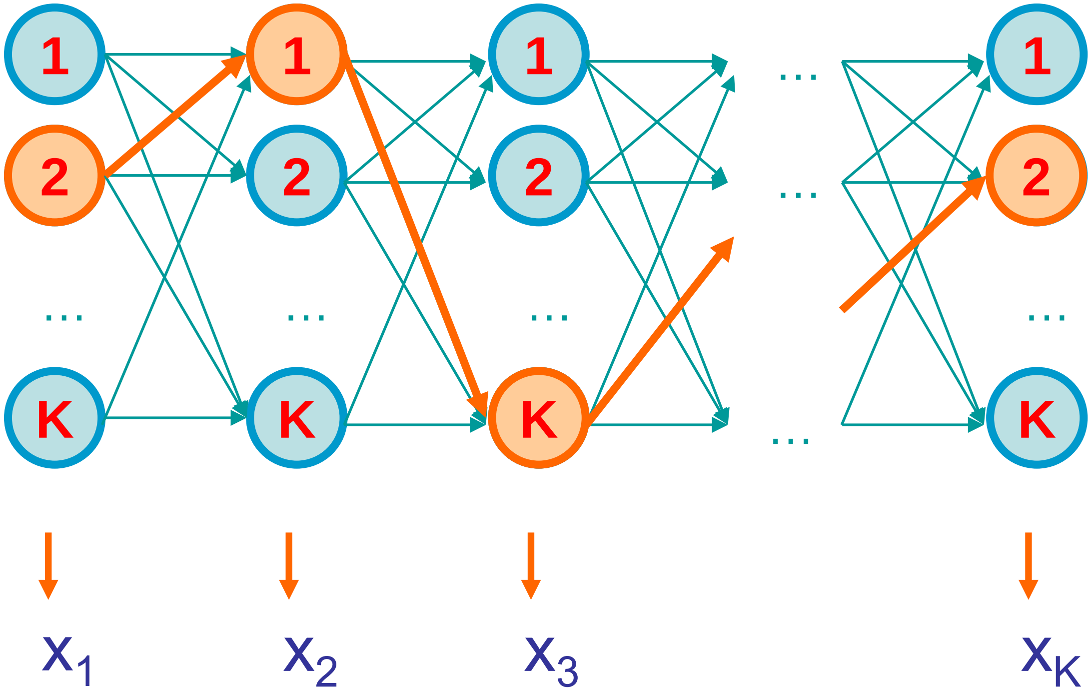
Hidden Markov Model

- Most pages of the slides are from lecture notes from Prof. Serafim Batzoglou's course in Stanford:
  - CS 262: Computational Genomics (Winter 2004)
  - http://ai.stanford.edu/~serafim/cs262/

Hidden Markov Models

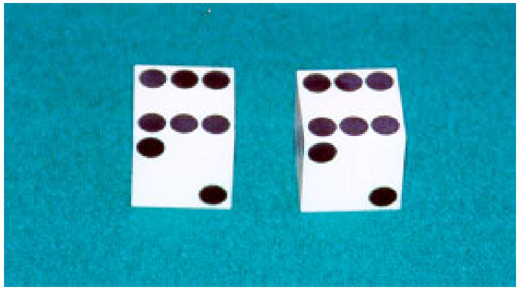


Applications of hidden Markov models

- HMMs can be applied in many fields where the goal is to recover a data sequence that is not immediately observable (but other data that depends on the sequence is).
  - Cryptanalysis
  - Speech recognition
  - Machine translation
  - Partial discharge
  - Gene prediction
  - Alignment of bio-sequences

* reference: wikipiedia.com

Example: The Dishonest Casino



A casino has two dice:

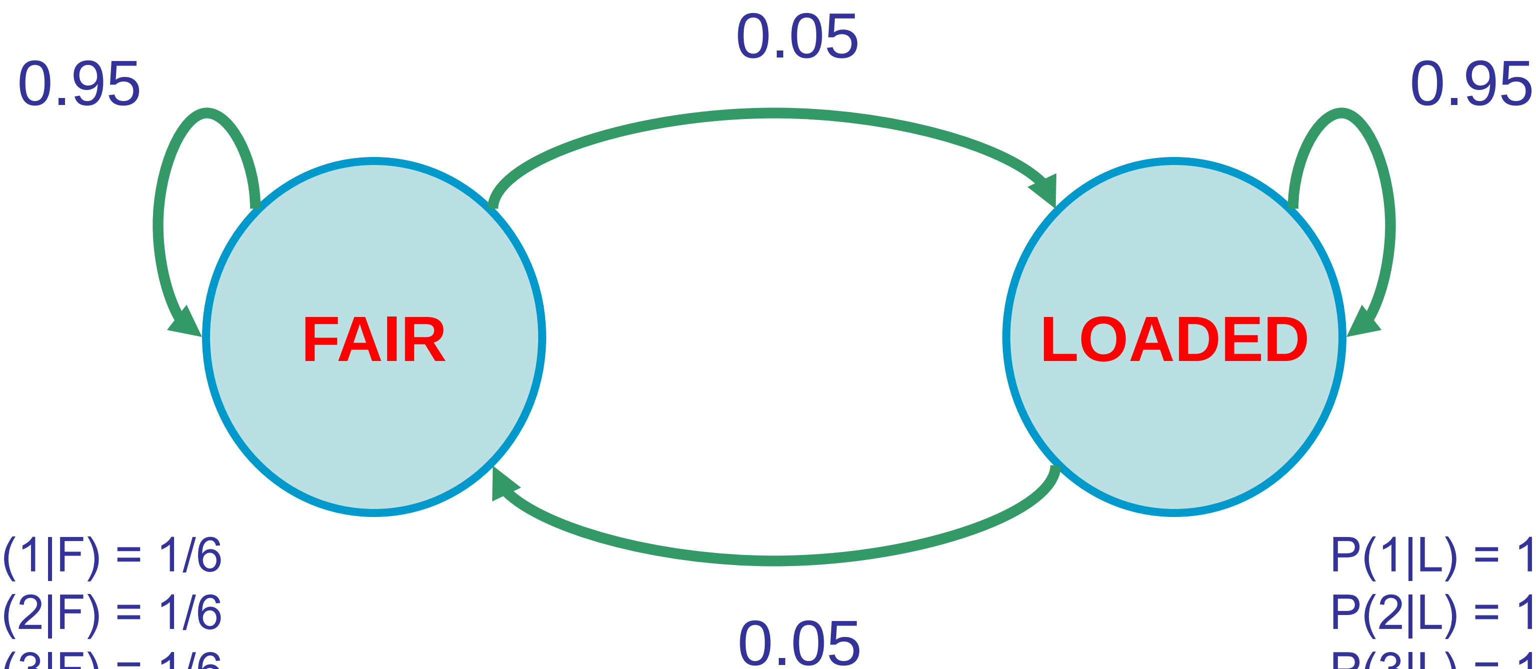
- Fair die
  P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6
- Loaded die
  P(1) = P(2) = P(3) = P(4) = P(5) = 1/10
  P(6) = 1/2

Casino player switches back-&-forth between fair and loaded die once roughly around every 20 turns

Game:

1. You bet $1
2. You roll (always with a fair die)
3. Casino player rolls (maybe with fair die, maybe with loaded die)
4. Highest number wins $2

The dishonest casino model



P(1|F) = 1/6
P(2|F) = 1/6
P(3|F) = 1/6
P(4|F) = 1/6
P(5|F) = 1/6
P(6|F) = 1/6

P(1|L) = 1/10
P(2|L) = 1/10
P(3|L) = 1/10
P(4|L) = 1/10
P(5|L) = 1/10
P(6|L) = 1/2

Question # 1 – Evaluation

GIVEN

A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

QUESTION

How likely is this sequence, given our model of how the casino works?

This is the **EVALUATION** problem in HMMs

Question # 2 – Decoding

GIVEN

A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

QUESTION

What portion of the sequence was generated with the fair die, and what portion with the loaded die?

This is the **DECODING** question in HMMs

Question # 3 – Learning

GIVEN

A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

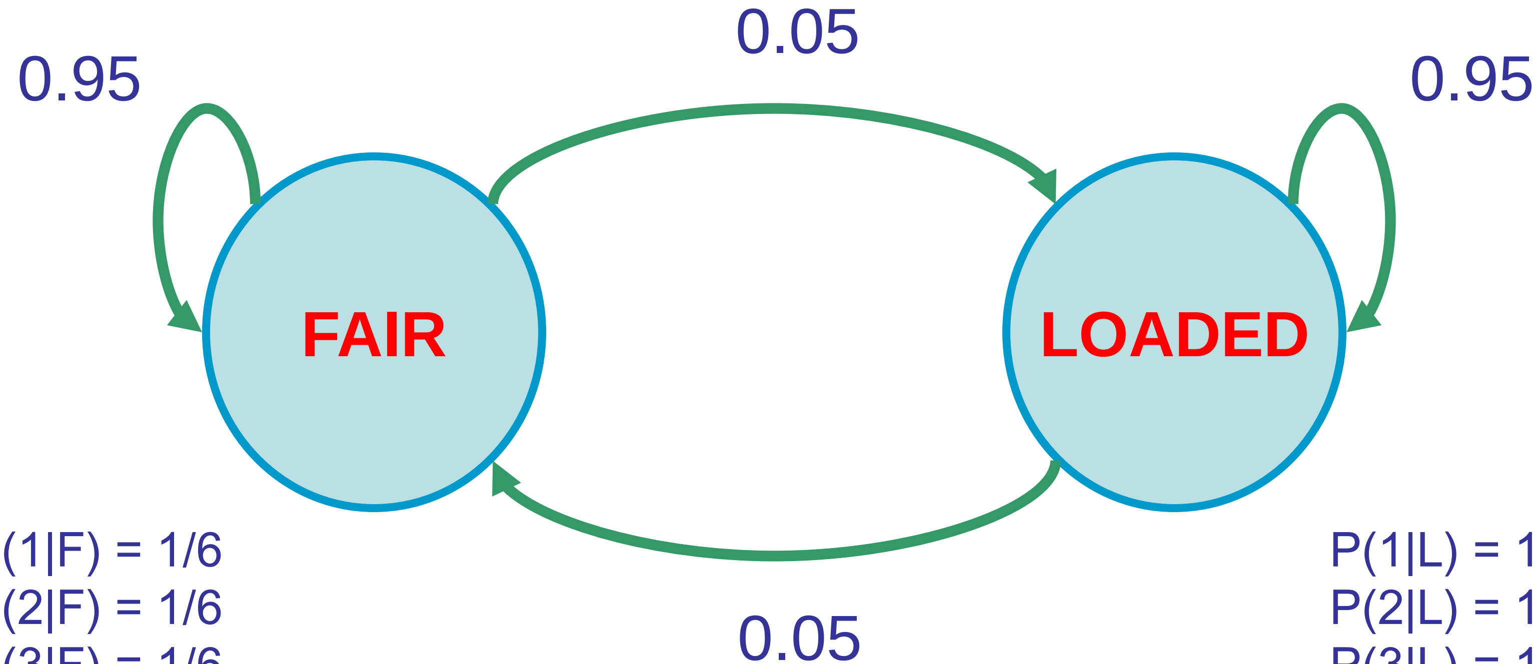
QUESTION

Model parameters:

How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded, and back?

This is the **LEARNING** question in HMMs (build model in training)

The dishonest casino model



P(1|F) = 1/6
P(2|F) = 1/6
P(3|F) = 1/6
P(4|F) = 1/6
P(5|F) = 1/6
P(6|F) = 1/6

P(1|L) = 1/10
P(2|L) = 1/10
P(3|L) = 1/10
P(4|L) = 1/10
P(5|L) = 1/10
P(6|L) = 1/2

Definition of a hidden Markov model

Definition: A hidden Markov model (HMM)

- Alphabet $\Sigma = \{ b_1, b_2, \ldots, b_M \}$ (observable symbols)
- Set of states $Q = \{ 1, \ldots, K \}$ (hidden states)
- Transition probabilities between any two states

a_{ij} = transition prob from state i to state j

$a_{i1} + \ldots + a_{iK} = 1$, for all states $i = 1 \ldots K$

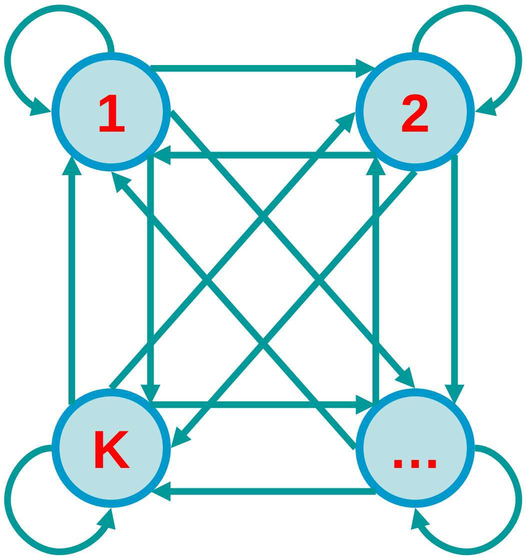
- Start probabilities a_{0i}

$a_{01} + \ldots + a_{0K} = 1$

- Emission probabilities within each state

$e_k(b_i) = P(x = b_i \mid \pi = k)$

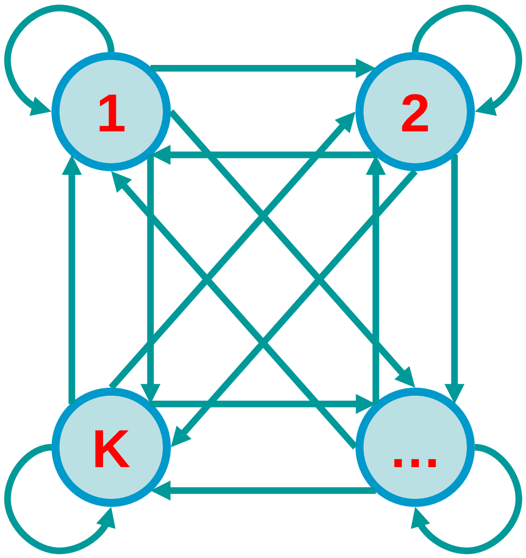
$e_i(b_1) + \ldots + e_i(b_M) = 1$, for all states $i = 1 \ldots K$



A Hidden Markov Model is memory-less

At each time step t,
the only thing that affects future states
is the current state π_t

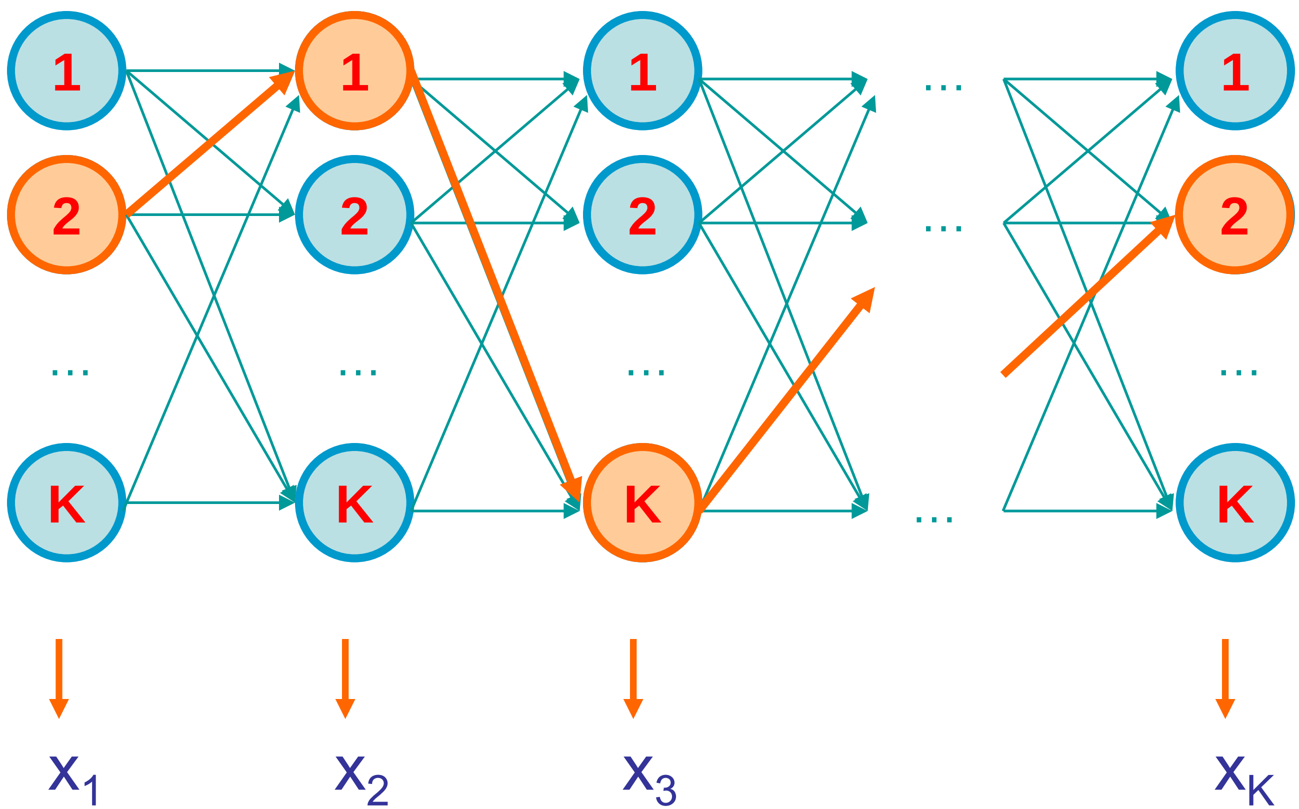
$P(\pi_{t+1} = k \mid \text{“whatever happened so far”}) =$
$P(\pi_{t+1} = k \mid \pi_1, \pi_2, \ldots, \pi_t, x_1, x_2, \ldots, x_t) =$
$P(\pi_{t+1} = k \mid \pi_t)$



A parse of a sequence

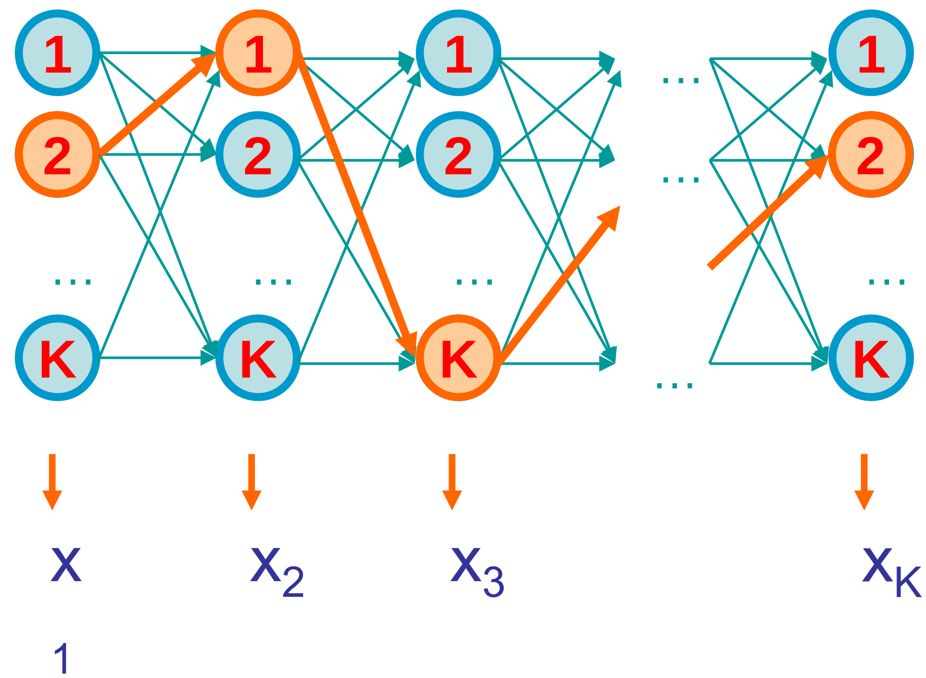
Given a sequence $x = x_1......x_N$,

A <u>parse</u> of x is a sequence of states $\pi = \pi_1, ..., \pi_N$



Likelihood of a parse

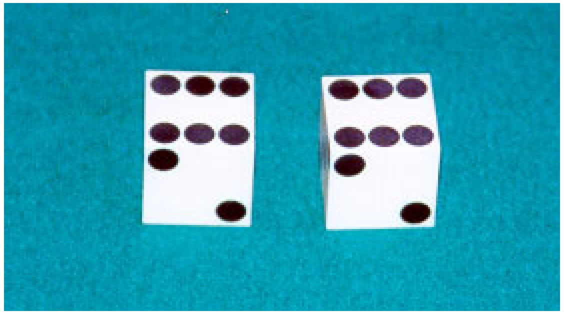
Given a sequence $x = x_1\ldots\ldots x_N$
and a parse $\pi = \pi_1, \ldots\ldots, \pi_N$,



To find how likely is the parse:
(given our HMM)

$$
\begin{aligned}
P(x, \pi) = P(x_1, \ldots, x_N, \pi_1, \ldots\ldots, \pi_N) = \\
& P(x_N, \pi_N \mid \pi_{N-1})\, P(x_{N-1}, \pi_{N-1} \mid \pi_{N-2}) \ldots\ldots P(x_2, \pi_2 \mid \pi_1)\, P(x_1, \pi_1) = \\
& P(x_N \mid \pi_N)\, P(\pi_N \mid \pi_{N-1}) \ldots\ldots P(x_2 \mid \pi_2)\, P(\pi_2 \mid \pi_1)\, P(x_1 \mid \pi_1)\, P(\pi_1) = \\
& a_{0\pi 1}\, a_{\pi 1 \pi 2} \ldots\ldots a_{\pi N-1 \pi N}\, e_{\pi 1}(x_1) \ldots\ldots e_{\pi N}(x_N)
\end{aligned}
$$

Example: the dishonest casino



Let the sequence of rolls be:

x = 1, 2, 1, 5, 6, 2, 1, 6, 2, 4

Then, what is the likelihood of

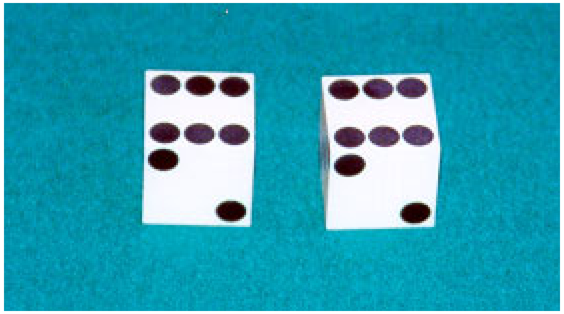
π = Fair, Fair, Fair, Fair, Fair, Fair, Fair, Fair, Fair, Fair?

(say initial probs $a_{0Fair} = ½$, $a_{oLoaded} = ½$)

$½ \times P(1 \mid Fair)\ P(Fair \mid Fair)\ P(2 \mid Fair)\ P(Fair \mid Fair) \ldots P(4 \mid Fair) =$

$½ \times (1/6)^{10} \times (0.95)^{9} = .00000000521158647211 = 5.21 \times 10^{-9}$

Example: the dishonest casino



So, the likelihood the die is fair in all this run is just 5.21×10^{-9}

OK, but what is the likelihood of

π = Loaded, Loaded, Loaded, Loaded, Loaded, Loaded, Loaded, Loaded, Loaded, Loaded?

$\frac{1}{2} \times P(1 \mid \text{Loaded})\, P(\text{Loaded} \mid \text{Loaded}) \ldots P(4 \mid \text{Loaded}) =$

$\frac{1}{2} \times (1/10)^8 \times (1/2)^2\, (0.95)^9 = .00000000078781176215 = 7.9 \times 10^{-10}$

Therefore, it is after all 6.59 times more likely that the die is fair all the way, than that it is loaded all the way.

Example: the dishonest casino

Let the sequence of rolls be:

x = 1, 6, 6, 5, 6, 2, 6, 6, 3, 6

Now, what is the likelihood π = F, F, …, F?

$\frac{1}{2} \times (1/6)^{10} \times (0.95)^{9} = 0.5 \times 10^{-9}$, same as before

What is the likelihood

π = L, L, …, L?

$\frac{1}{2} \times (1/10)^{4} \times (1/2)^{6} (0.95)^{9} = .00000049238235134735 = 0.5 \times 10^{-7}$

So, it is 100 times more likely the die is loaded

The three main questions on HMMs

1. Evaluation

   GIVEN a HMM M, and a sequence x,
   FIND Prob[x | M]

2. Decoding

   GIVEN a HMM M, and a sequence x,
   FIND the sequence π of states that maximizes P[x, π | M]

3. Learning

   GIVEN a HMM M, with unspecified transition/emission probs.,
   and a sequence x,

   FIND parameters $\theta = (e_i(.), a_{ij})$ that maximize P[x | θ]