STATS 235: Modern Data Analysis
Markov and Hidden Markov Models

Babak Shahbaba

Department of Statistics, UCI

Stochastic processes; random walk

- Stochastic processes are dynamic random variables. They are indexed (usually by time), $\{X_n\}$. A discrete time, time-homogeneous stochastic process is simply a sequence of random variables, $X_0, X_1, ..., X_n$ defined on the same probability space.

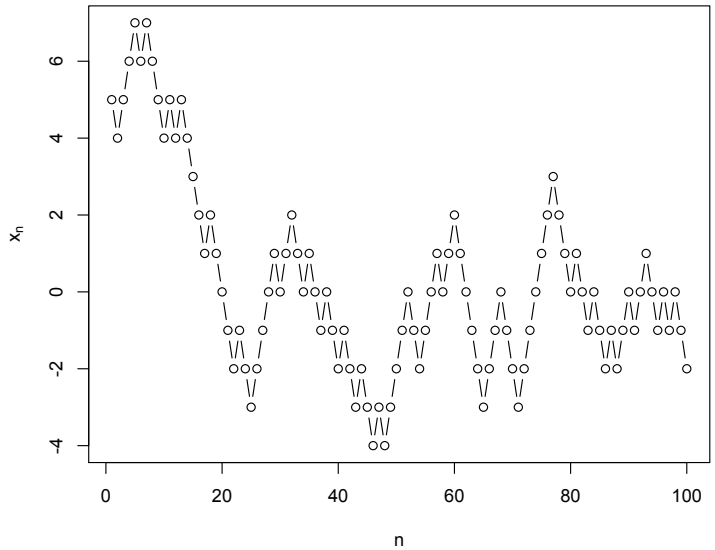
- One of the simplest stochastic processes (and one of the most useful ones) is the simple random walk.

- Consider a sequence of iid random variables $\{Z_i\}$ such that $P(Z_i = 1) = p$ and $P(Z_i = -1) = 1 - p = q$. The stochastic process represented by $\{X_n\}$ where $X_0 = a$ and $X_n = a + Z_1 + ... + Z_n$ is called a random walk process.

- This stochastic process could be interpreted as a gambler's fortune (in dollars) at time n, where the gambler starts with a dollars, makes \$1 bets repeatedly, and the probability of winning is p.

Random walk

- A random walk with $a = 5$ and $p = 0.4$.



- To obtain the distribution of such process note that $P(X_n = a + k) = 0$ unless $-n \leq k \leq n$ and $n + k$ is even. To get to $a + k$ after n steps, we need $\frac{n+k}{2}$ of $Z = 1$ and $\frac{n-k}{2}$ of $Z = -1$. Therefore:

$$P(X_n = a + k) = \binom{n}{\frac{n+k}{2}} p^{\frac{n+k}{2}} q^{\frac{n-k}{2}}$$

Discrete time, discrete space, time-homogenous Markov chains

- The above simple random walk is a special case of another well-known stochastic process called *Markov chains*.

- A Markov chain represent the stochastic movement of some particle from one state in the space S to another state in S. The particle initially starts from state i with probability $\pi_i^{(0)}$, and every time it is in state i, there is a p_{ij} chance it moves to state j in its next step.

- Therefore, a Markov chain has three main elements: 1- a state space S, which is a finite or countable set, 2- an initial distribution $\{\pi^{(0)}\}_i \in S$ which are non-negative numbers assigned to each state and they are summing to 1, and 3- transition probabilities $\{p_{ij}\}$ which are non-negative numbers representing the probability of going from state i to j, and $\sum_j p_{ij} = 1$.

Markov chains: An example

- For example, in the above simple random walk, we have
  - State space: $S = \mathcal{Z}$, the set of all integers
  - Initial distribution: $\pi^{(0)} = \delta_a$, i.e., a point mass at a such that $\pi_a = 1$ and $\pi_i = 0$, where $i \neq a$.
  - Transition probabilities: $p_{i,i+1} = p$, $p_{i,i-1} = 1 - p$ for all $i \in \mathcal{Z}$, and $p_{i,j} = 0$ when $j \neq i \pm 1$.

Markov chains: Formal definition

- A Markov chain is formally defined as a sequence of random variables, $X_0, X_1, ..., X_n$, representing states in the set S such that

$$P(X_0 = i_0, X_1 = i_1, ..., X_n = i_n) = \pi^{(0)}_{i_0} p_{i_0 i_1} ... p_{i_{n-1} i_n}.$$

- The above definition is based on the joint distribution, which can be used to derive the conditional distributions of the form

$$\begin{aligned} P(X_{k+1} = j | X_k = i) &= \frac{P(X_k = i, X_{k+1} = j)}{P(X_k = i)}, \quad P(X_k = i) > 0 \\ &= \frac{\sum_{i_0, i_1, ..., i_{k-1}} \pi^{(0)}_{i_0} p_{i_0 i_1} p_{i_1 i_2} ... p_{i_{k-1} i} p_{ij}}{\sum_{i_0, i_1, ..., i_{k-1}} \pi^{(0)}_{i_0} p_{i_0 i_1} p_{i_1 i_2} ... p_{i_{k-1} i}} \\ &= p_{ij} \end{aligned}$$

Markov property

- As we can see, this does not depend on k (i.e, it's time homogeneous).
- Also, it does not depend on the prior steps to get to $X_k = i$ (a.k.a the Markov property).
- This latter property of Markov process has been used as the formal definition of Markov process, however, we do not use this definition to avoid the complexity due to situations with $P(X_k = i) = 0$.

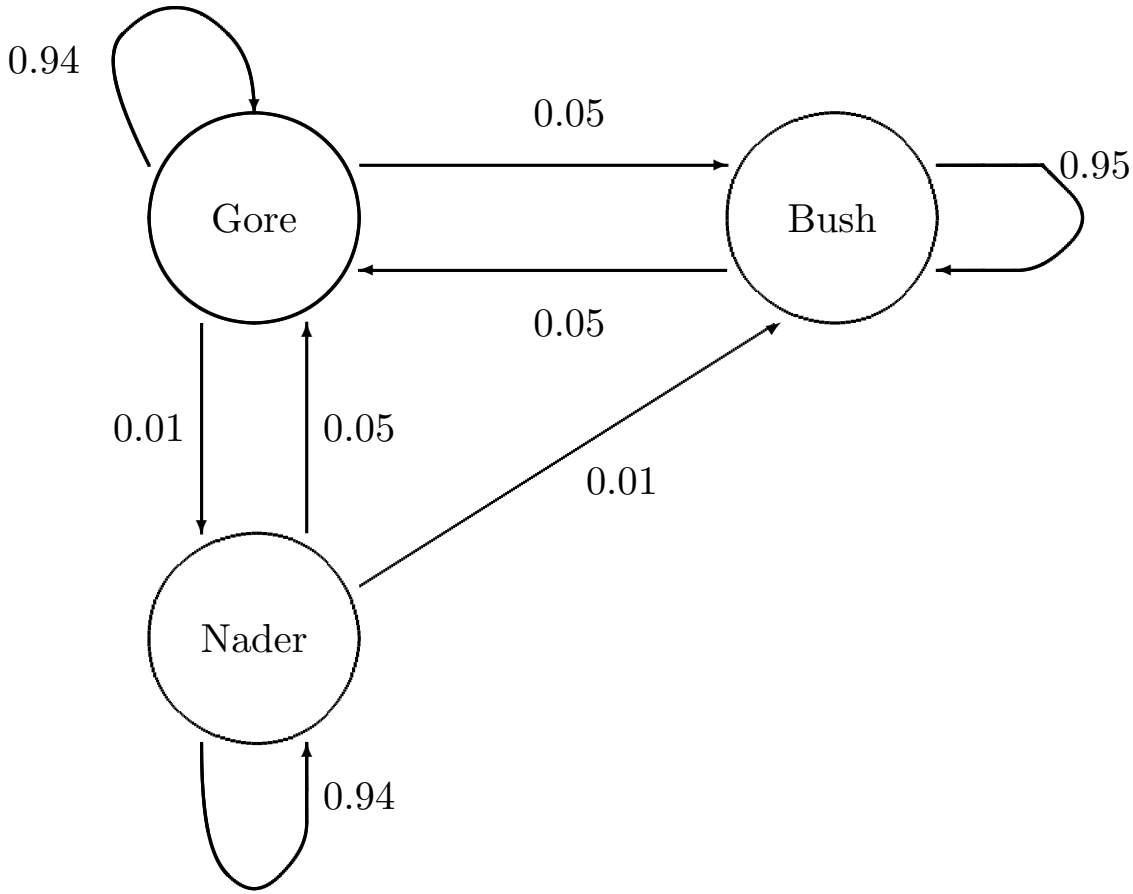
Example: 2000 presidential election

- Consider the 2000 US presidential election with three candidates: Gore, Bush and Nader (note that this is just an illustrative example and does not reflect the reality of that election).
- We assume that the initial distribution of votes (i.e., probability of winning) was $\pi = (0.49, 0.45, 0.06)$ for Gore, Bush and Nader respectively.
- Further, we assume the following transition probability matrix:

	Gore	*Bush*	*Nader*
Gore	0.94	0.05	0.01
Bush	0.05	0.95	0
Nader	0.05	0.01	0.94

Example: 2000 presidential election

- We can present the above Markov chain schematically as follows:



Example: 2000 presidential election

- If we represent the initial distribution $\pi^{(0)}$ by a row vector, and the transition probability by a square matrix P such that the element in row i and column j is p_{ij}, we can obtain the distribution of states in step n, $\pi^{(n)}$, as follows

$$\begin{aligned} \pi^{(1)} &= \pi^{(0)} P \\ \pi^{(2)} &= \pi^{(1)} P = \pi^{(0)} P^2 \\ &\cdots \\ \pi^{(n)} &= \pi^{(0)} P^n \end{aligned}$$

- For the above example, we have

$$\begin{aligned} \pi^{(0)} &= (0.4900, 0.4500, 0.0600) \\ \pi^{(10)} &= (0.4656, 0.4655, 0.0689) \\ \pi^{(100)} &= (0.4545, 0.4697, 0.0758) \\ \pi^{(200)} &= (0.4545, 0.4697, 0.0758) \end{aligned}$$

Stationary distribution

- As we can see last, after some steps, the above Markov chain converges to a distribution, (0.4545, 0.4697, 0.0758), and does not moved afterwards.
- The chain would have reached this distribution (eventually) regardless of what initial distribution $\pi^{(0)}$ we chose. Therefor, $\pi = (0.4545, 0.4697, 0.0758)$ is the *stationary distribution* for the above Markov chain.
- Stationary distribution: A distribution of Markov chain states is called to be stationary if it remain the same in the next time step(s). That is, if the current distribution is $\{\pi_i\}$, we have $[\pi][p] = [\pi]$, where $[p]$ is the transition probability matrix. By induction, this also means $[\pi][p]^{(n)} = [\pi]$. In other words $\sum_i \pi_i p_{ij}^{(n)} = \pi_j$ for any $n \in \mathcal{N}$.

Stationary distribution

- How can we find out whether such distribution exists?
- Also, how do we know whether the chain would converge to this distribution?
- To find out the answer, we briefly discuss some properties of Markov chains.
- Finding the stationary distribution is an interesting problem on its own, but as we will see later, this would not be our concern in this course.

Recurrent vs. transient

- Recurrent states: a state i is called *recurrent* (or persistent) if starting from i, we will eventually visit i again with probability 1. All the states in the following MC are recurrent.

	Gore	*Bush*	*Nader*
Gore	0.94	0.05	0.01
Bush	0.05	0.95	0
Nader	0.05	0.01	0.94

- Transient state: a state i is called transient if starting from i, the probability of re-visiting it is less than 1. Nader in the following MC is a transient state:

	Gore	*Bush*	*Nader*
Gore	0.95	0.05	0
Bush	0.05	0.95	0
Nader	0.05	0.01	0.94

Irreducibility

- Irreducible: A Markov chain is irreducible if the chain can move from any state to another state. For example, the simple random walk is irreducible. The following chain is however reducible since Nader does not communicate with the other two states (Gore and Bush).

	Gore	*Bush*	*Nader*
Gore	0.95	0.05	0
Bush	0.05	0.95	0
Nader	0	0	1

Aperiodicity

- Period: the period of a state i is the greatest common divisor of the times at which it is possible to move from i to i. All the states in the following MC have period 3.

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

- Aperiodic: a Markov chain is said to be aperiodic if the period of each state is 1, otherwise the chain is periodic.

- Ergodic: a state is *ergodic* if it is aperiodic and recurrent. A Markov chain is said to be ergodic if all of its states are ergodic.

Reversibility (very important)

- Reversibility: a Markov chain is said to be *reversible* with respect to a probability distribution $\{\pi_i\}$ if $\pi_i p_{ij} = \pi_j p_{ji}$.

- We can show that if a Markov chain is reversible with respect to $\{\pi_i\}$, then $\{\pi_i\}$ is a stationary distribution:

$$\begin{aligned} \sum_i \pi_i p_{ij} &= \sum_i \pi_j p_{ji} \\ &= \pi_j \sum_i p_{ji} \\ &= \pi_j \end{aligned}$$

  since for all Markov chains $\sum_i p_{ji} = 1$.

- The above condition is also called *detailed balance*.

Discrete time, general space, time-homogeneous Markov chains

- We can define a Markov chain on a general state space $\mathcal{X}$ with initial distribution $\pi^{(0)}$ and transition probabilities $P(x, A)$ interpreted as when at point $x \in \mathcal{X}$, this is the probability of jumping to the subset A.

- As before, a Markov chain $X_0, X_1, ...$ is defined by

$$P(X_0 \in A_0, X_1 \in A_1, ..., X_n \in A_n) =$$
$$\int_{A_0} \pi^{(0)}(dx_0) \int_{A_1} P(x_0, dx_1) ... \int_{A_n} P(x_{n-1}, dx_n)$$

- As an example, consider a Markov chain the real line as its state space, $N(1, 1)$ as its initial distribution, and $P(x, .) = N(\frac{x}{2}, \frac{3}{4})$ as its transition probability.

Discrete time, general space, time-homogeneous Markov chains

- In this case, although we cannot talk about irreducibility and period as we discussed for discrete space since on a continuous the probability of visiting the same point is zero, we can still talk about irreducibility for sets A with non-zero measure, ϕ, on $\mathcal{X}$.

- *A chain in ϕ-irreducible if for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$ and for all $x \in \mathcal{X}$, there exists a positive integer n such that $P^n(x, A) > 0$.*

- Similarly, we need to modify our definition of period.

Discrete time, general state space, time-homogeneous Markov chains

- A distribution π is a stationary distribution if

$$\pi(A) = \int_{\mathcal{A}} \pi(dx)P(x, A)$$

- As for the discrete case, a continuous space is reversible with respect to π if

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$$

Discrete time, general state space, time-homogeneous Markov chains

- Also, if the chain is reversible with respect to π then, π is its stationary distribution:

$$\begin{aligned}
\int_{x\in\mathcal{X}} \pi(dx)P(x,dy) &= \\
\int_{x\in\mathcal{X}} \pi(dy)P(y,dx) &= \\
\pi(dy)\int_{x\in\mathcal{X}} P(y,dx) &= \\
\pi(dy)
\end{aligned}$$

- The above Markov chain with $N(1,1)$ initial distribution and $P(x,.) = N(\frac{x}{2},\frac{3}{4})$ transition probability converges to N(0, 1).

Main convergence theorem of Markov chains

- The main convergence theorem: If a Markov chain is ϕ-irreducible and aperiodic and has a stationary distribution π, then for all measurable $A \subseteq \mathcal{X}$, we have $\lim_{n\to\infty} P^n(x, A) = \pi(A)$.
- That is, regardless of the initial state, the chain converges to its stationary distribution.
- The convergence is in terms of the "total variation distance" defined between two probability measures as follows:

$$\|\nu_1(.) - \nu_2(.)\| = \sup_A |\nu_1(A) - \nu_2(A)|$$

- Therefore, the main convergence theorem indicates that

$$\lim_{n\to\infty} \|P^n(x, .) = \pi(.)\| = 0$$

Finding the stationary distribution

- Recall that for a stationary distribution π, we have

$$\pi P = \pi$$

  where P is the transition matrix.

- Therefore, π is a left eigenvector for the matrix p with the corresponding eigenvalue of 1.

MLE for Markov models

- Consider a discrete Markov chain with M states.
- The probability of each observed sequence $x = (x_1, \ldots, x_T)$ is

$$P(X = x|\theta) = P(X_1 = x_1|\theta) \prod_{t=2}^{T} P(X_t = x_t|X_{t-1} = x_{t-1}, \theta)$$

  where θ includes the initial and transition probabilities.

- We can rewrite this probability in terms of the states and model parameters,

$$P(X = x|\theta) = \prod_{j=1}^{M} (\pi_j)^{I(x_1=j)} \prod_{t=2}^{T} \prod_{j=1}^{M} \prod_{k=1}^{M} (p_{jk})^{I(x_t=k, x_{t-1}=j)}$$

MLE for Markov models

- Given a sample of n IID sequence data $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where each $\mathbf{x}_i = (x_{i1}, \ldots, x_{iT_i})$, we can write the likelihood function in terms of θ as follows:

$$f(\theta) = \prod_{j=1}^{M} (\pi_j)^{n_{0j}} \prod_{j=1}^{M} \prod_{k=1}^{M} (p_{jk})^{n_{jk}}$$

  where n_{0m} is the number of times the chain starts at state m, and n_{jk} is the number of times we observe the chain moves from state i to j.

- To estimate θ, we can maximize the following log-likelihood function,

$$\mathcal{L}(\theta) = \sum_{j=1}^{M} n_{0j} \log \pi_j + \sum_{j=1}^{M} \sum_{k=1}^{M} n_{jk} \log p_{jk}$$

  with the constraints $\sum_j p_{ij} = 1$.

- In this case, MLE is simply the normalized counts:

$$\hat{\pi}_j = \frac{n_{0j}}{\sum_{j'} n_{0j'}} \qquad \hat{p}_{jk} = \frac{n_{jk}}{\sum_{k'} n_{jk'}}$$

Example: Language modeling

- Markov models are commonly used is language modeling, where we are interested in probability distributions over word sequences.
- In this case, the states are all possible words in a specific language.
- The marginal probability of each work, $P(X_t = j)$ is called *unigram*, and the first-order Markov model $P(X_t = k|X_{t-1} = j)$ is called a *bigram* model.
- We can also have second-order Markov models,

$$P(X_t = k|X_{t-1} = j, X_{t-2} = l)$$

  called a *trigram* model, or in general *n-gram* models.
- Such models can be used for automatic sentence completion like the one used in google search.

Hidden Markov Models (HMM)

- Sometimes, we don't observe the states directly. The observed data are instead assumed to be generated by a set of unobserved (hidden) states.
- We can still use a Markov model for transition among states. We refer to such models as *Hidden Markov Models (HMM)*.
- For example, in language modeling, instead of treating the words as states, we could consider the *part-of-speach (POS)* tags as states representing the categories assigned to each word with respect to its grammatical (e.g., nouns, verbs, adjectives, adverbs) or contextual role.

Hidden Markov Models (HMM)

- The joint distribution of hidden states z (e.g., POS) and observed data x (e.g., words) can be written as follows:

$$P(z,x) = P(z)P(x|z) = P(z_1)\prod_{t=1}^{T} P(z_t|z_{t-1}) \cdot \prod_{t=1}^{T} P(x_t|z_t)$$

- HMM models can be presented as **Directed Graphical Models**, which are commonly known as **Bayesian Networks**.

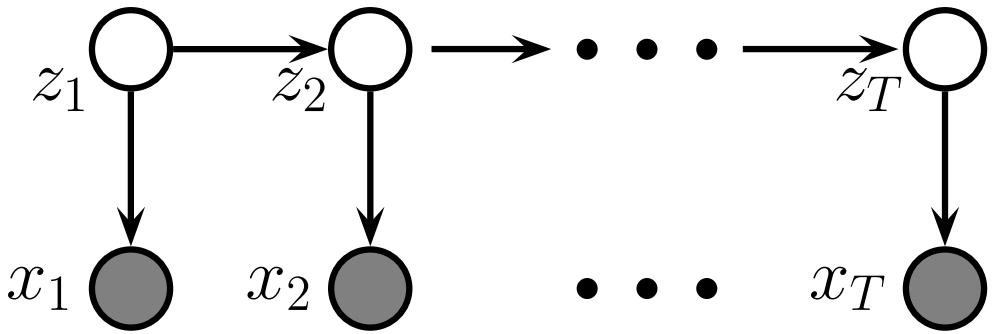


Fig10.4 in Murphy (2012): A first-order HMM.

Inference

- Given a sample of observed sequences, $x = (x_1, \ldots, x_T)$, we want to infer the hidden states at $z = (z_1, \ldots, z_T)$.
- There are several algorithms for this purpose:
  - Viterbi algorithm: we find the most probable state sequence: $\arg\max_z P(z|x)$.
  - The forwards algorithm: We recursively compute the *filtered* marginals $P(z_t|x_{1:t})$ using the observed data up to time t only (online inference).
  - The forwards-backwards algorithm: we compute the *smoothed* marginals $P(z_t|x_{1:T})$ using the observed data at all time points (offline inference).
- For inference regarding model parameters θ, which includes initial and transition probabilities as well as the parameters for the observed data $P(x_t|z_t)$, we could use the EM approach (similar to mixture models) called Baum-Welch algorithm.
- The **E** step uses the forwards-backwards algorithm, and the **M** step is very similar to what we discussed for Markov models.
- Alternatively, we could use Bayesian inference.

Example: occasionally dishonest casino

- This is based on a an example from Durbin et al. 1998 that occasionally switches to a loaded dice with a higher probability of 6.

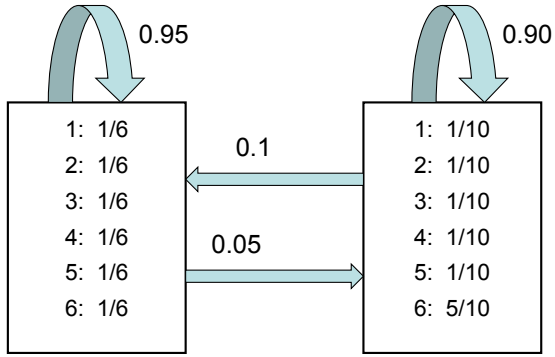


Fig17.9 in Murphy (2012): An HMM for occasionally dishonest casino.

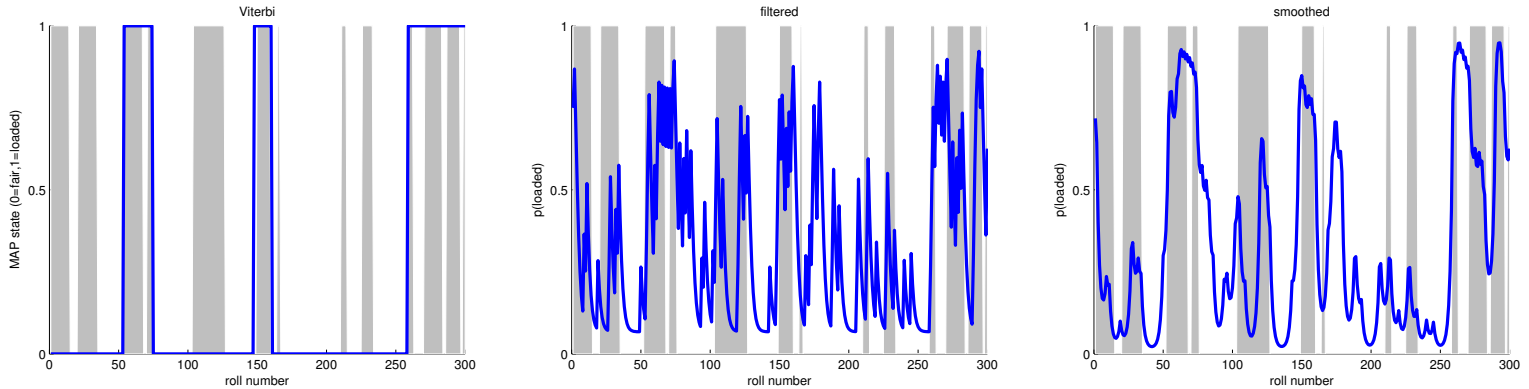


Fig17.10 in Murphy (2012): Viterbi Algorithm (left), forwards algorithm (middle), and forwards-backwards algorithm (right).

HMM package in R

- In R, you can use the HMM package to fit an HMM model
- Using the `dishonestCasino()` function, you can also simulate data for the dishonest casino example, and use the Viterbi algorithm to find the most probably path.