Markov Chains

Concepts and Detailed Balance

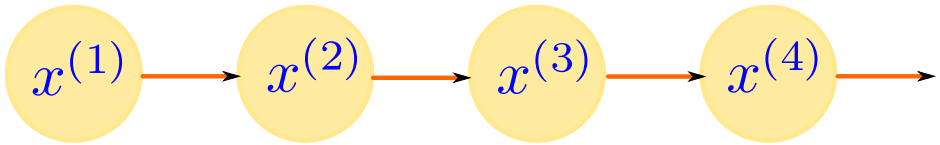
Sachin Shanbhag

Department of Scientific Computing
Florida State University,
Tallahassee, FL 32306.

Definitions

Typically, a stochastic process is a system which evolves in time while undergoing random fluctuations.

It is a non-deterministic sequence of random variables, $x^{(i)}$, for $x^{(i)} \in \mathbf{X}$.



A particular x_i is called a state.

$\mathbf{X}$ is called the state space or the phase space. It is the set or space of all possible $\{x_1, x_2, x_3, ..., \}$ (think *population*).

x_i may be scalar/vector, discrete/continuous, or more complicated.

Simple Examples

Trivial: series of independent coin tosses $[\mathbf{X} = \{H, T\}]$.

A particular realization may be,

$$H \to T \to T \to H \to T \to \cdots$$

Not a particularly interesting example because successive states are independent.

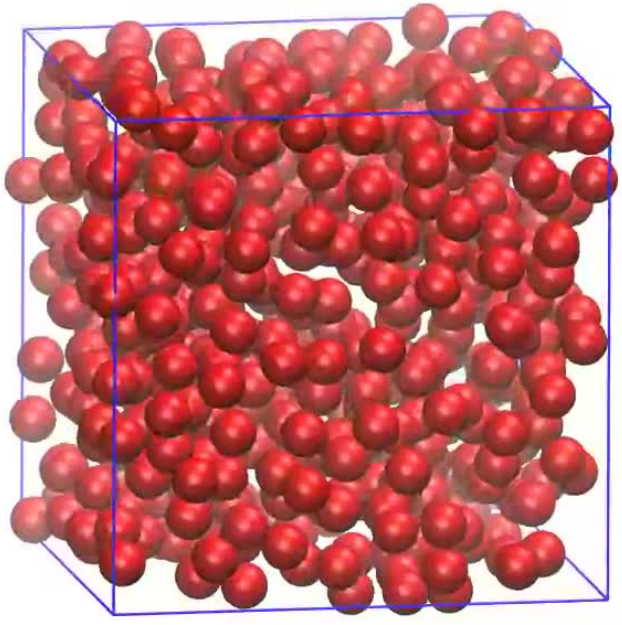
In more interesting examples, there is some correlation between successive states.

- number of customers in a store every hour
- stock price of Apple Inc. every second
- daily low and high temperatures in Tallahassee

What is the state space for each of these examples?

Complicated Example

Consider a molecular simulation with N similar particles



wikipedia.org

A state x_i (or configuration) may correspond to the positions and velocities of all the particles in the system.

Each particle has 6 "co-ordinates":

- position $((r_x, r_y, r_z))$
- velocity $((v_x, v_y, v_z))$

Thus x_i is a vector of size $6N$.

Here the superscript identifies the position of a state in a sequence, while the subscript identifies a particular state.

Definitions

A Markov process is a stochastic process in which future states are independent of past states given the present state.

$$P(x^{(i)}|x^{(i-1)}, x^{(i-2)}, ..., x^{(0)}) = P(x^{(i)}|x^{(i-1)})$$

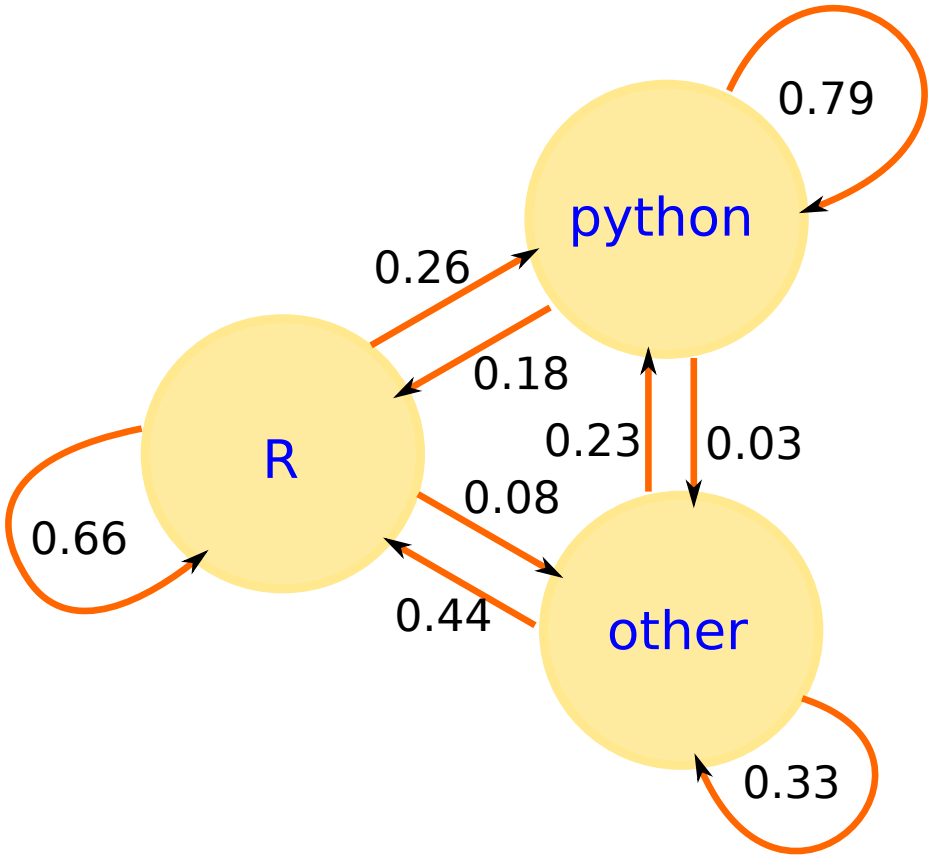
For a state space with n states, specifying the $n \times n$ matrix $W_{ij} = P(x_i|x_j)$ completely specifies the Markov process.

This matrix $\mathbf{W} = W_{ij}$ is called the transition probability matrix (TPM).

For typical problems the matrix $\mathbf{W}$ is huge (size of state space squared) and is never directly stored on the computer.

Motivating Example

I found this picture describing the traffic of programmers between R, python, and "other" languages for the year 2013.



Ex: over a cycle, the probability that an R user switches to python is 0.26, and remains with R is 0.66.

Motivating Example

I don't trust such numbers, but let us temporarily suspend our skepticism and take them seriously.

Let the state space $\mathbf{X} = \{r, p, o\}$ (R, Python, Other).

Let $\{\pi_r, \pi_p, \pi_o\}$ represent the fraction of programmers using R, python, and "other" as their primary programming language, respectively.

Particular Markov chains may look like:

- rrppprrpo...
- orrrppppp...

In this example, one may think of a particular Markov chain as the trajectory of a particular programmer over several cycles.

Transition Probability Matrix

Now let us encode this graphic into a TPM.

$$\mathbf{W} = \begin{bmatrix} 0.66 & 0.18 & 0.44 \\ 0.26 & 0.79 & 0.23 \\ 0.08 & 0.03 & 0.33 \end{bmatrix}$$

Convention: Order of the columns and rows is:

$$\begin{array}{ccc} \text{R} & \rightarrow & 1 \\ \text{Python} & \rightarrow & 2 \\ \text{Other} & \rightarrow & 3 \end{array}$$

Thus, $W_{32} = W_{3\leftarrow 2} = W_{o\leftarrow p} = W_{op} = 0.03$.

```
W = np.matrix([[0.66, 0.26, 0.08],[0.18, 0.79, 0.03],
               [0.44, 0.23, 0.33]]).T
```

Transition Probability Matrix

Given this information, we may be interested in questions like:

- after n cycles, what is the relationship of π_r, π_p, π_o to the initial values
- after infinite cycles, what is the eventual "marketshare"?

Let $\pi_r^{(n)}$ represent the fraction of R programmers after n cycles.

It is easy to see that after the first cycle:

$$\pi_r^{(2)} = \pi_r^{(1)} W_{rr} + \pi_p^{(1)} W_{rp} + \pi_o^{(1)} W_{ro}$$

Similarly for $\pi_p^{(2)}$ and $\pi_o^{(2)}$.

Knowing the rules for matrix multiplication, we can combine this to write:

$$\boldsymbol{\pi}^{(2)} = \begin{bmatrix} \pi_r^{(2)} \\ \pi_p^{(2)} \\ \pi_o^{(2)} \end{bmatrix} = \begin{bmatrix} W_{rr} & W_{rp} & W_{ro} \\ W_{pr} & W_{pp} & W_{po} \\ W_{or} & W_{op} & W_{oo} \end{bmatrix} \begin{bmatrix} \pi_r^{(1)} \\ \pi_p^{(1)} \\ \pi_o^{(1)} \end{bmatrix} = \mathbf{W}\boldsymbol{\pi}^{(1)}.$$

If this is not immediately clear, spend some time thinking about this equation, until you agree with it.

Generalizing, after n steps.

$$\boldsymbol{\pi}^{(n+1)} = \mathbf{W}\boldsymbol{\pi}^{(n)} = \mathbf{W}^n\boldsymbol{\pi}^{(1)}.$$

Example

Using

$$\mathbf{W} = \begin{bmatrix} 0.66 & 0.18 & 0.44 \\ 0.26 & 0.79 & 0.23 \\ 0.08 & 0.03 & 0.33 \end{bmatrix} \text{ and } \boldsymbol{\pi}^{(1)} = \begin{bmatrix} 0.3333 \\ 0.3333 \\ 0.3333 \end{bmatrix},$$

We find,

$$\boldsymbol{\pi}^{(2)} = \mathbf{W}\boldsymbol{\pi}^{(1)} = \begin{bmatrix} 0.4229 \\ 0.4817 \\ 0.0953 \end{bmatrix}, \quad \boldsymbol{\pi}^{(5)} = \mathbf{W}^4\boldsymbol{\pi}^{(1)} = \begin{bmatrix} 0.3896 \\ 0.5383 \\ 0.0721 \end{bmatrix}$$

$$\boldsymbol{\pi}^{(10)} = \mathbf{W}^9\boldsymbol{\pi}^{(1)} = \begin{bmatrix} 0.3816 \\ 0.5483 \\ 0.0702 \end{bmatrix}, \quad \boldsymbol{\pi}^{(50)} = \mathbf{W}^{49}\boldsymbol{\pi}^{(1)} = \begin{bmatrix} 0.3812 \\ 0.5487 \\ 0.0701 \end{bmatrix}$$

We notice that after a reasonably large n, the distribution settles to

$$\boldsymbol{\pi}^{(\infty)} = \begin{bmatrix} 0.3812 \\ 0.5487 \\ 0.0701 \end{bmatrix}$$

Before we analyze this further, let us try a different initial starting state, say

$$\boldsymbol{\pi}^{(1)} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$$

We get the sequence:

$$\boldsymbol{\pi}^{(2)} = \begin{bmatrix} 0.477 \\ 0.372 \\ 0.151 \end{bmatrix}, \boldsymbol{\pi}^{(5)} = \begin{bmatrix} 0.4034 \\ 0.5209 \\ 0.0757 \end{bmatrix}$$

$$\boldsymbol{\pi}^{(10)} = \begin{bmatrix} 0.3822 \\ 0.5475 \\ 0.0703 \end{bmatrix}, \boldsymbol{\pi}^{(50)} = \begin{bmatrix} 0.3812 \\ 0.5487 \\ 0.0701 \end{bmatrix}$$

Stationary Distribution

End up with the *same* stationary or limiting distribution $\boldsymbol{\pi}^{(\infty)}$.

In fact, $\boldsymbol{\pi}^{(\infty)}$ is independent of $\boldsymbol{\pi}^{(1)}$.

After enough cycles, the Markov chain forgets where it started from, and the stationary distribution is only a function of $\mathbf{W}$.

Note an interesting property of $\boldsymbol{\pi}^{(\infty)}$.

$$\boldsymbol{\pi}^{(\infty)} = \mathbf{W}\boldsymbol{\pi}^{(\infty)}.$$

If you have seen some linear algebra, this might remind you of an eigenvalue problem: $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$.

Here $\lambda = 1$, $\mathbf{v} = \boldsymbol{\pi}^{(\infty)}$, and $\mathbf{A} = \mathbf{W}$.

Thus the stationary distribution is the eigenvector of $\mathbf{W}$ corresponding to eigenvalue 1.

Transition Probability Matrix

A Markov matrix $\mathbf{W}$ satisfies three properties:

- ergodicity
- normalization
- balance

As demonstrated before, its largest eigenvalue is 1. The corresponding eigenvector is the limiting distribution.

Notation Alert

In the rest of the lecture, I shall omit the superscript of (∞) and assume that π_i is the probability of state i in the stationary distribution.

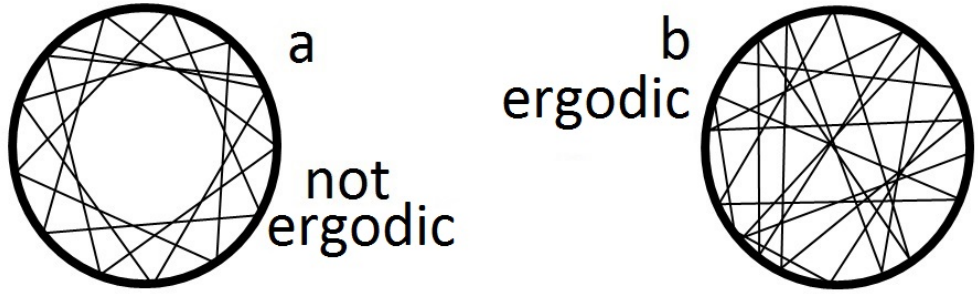
$$\pi_i^{(\infty)} \rightarrow \pi_i$$

Furthermore, I will use the shorthand, $\boldsymbol{\pi} = [\pi_1 \;\; \pi_2 \;\; ...\pi_N]^T$.

Ergodicity

This condition ensures that it is theoretically possible for the Markov chain to visit all points in the state space (if run sufficiently long).

For complicated problems, this is not an easy property to prove.



wikipedia.org

Normalization

The columns of $\mathbf{W}$ have to sum to 1, since they represent transition probabilities.

$$\sum_i W_{ij} = W_{1j} + W_{2j} + ... + W_{jj} + ... + W_{Nj} = 1.$$

In the next step, the Markov chain has to travel to some state.

This can include staying put.

Balance

This condition ensures that the equilibrium state corresponds to eigenvalue 1.

$$\mathbf{W}\boldsymbol{\pi} = \boldsymbol{\pi}.$$

The Design Problem

If $\mathbf{W}$ is given to us, finding $\boldsymbol{\pi}$ is relatively easy. It is uniquely determined if $\mathbf{W}$ satisfies the constraints above.

However, in a typical application, our goal is to sample

$$x_i \sim \boldsymbol{\pi}(x).$$

We therefore **design** or **construct** the matrix $\mathbf{W}$ with a target $\boldsymbol{\pi}(x)$.

Given a particular $\boldsymbol{\pi}(x)$ there are many possible choices of $\mathbf{W}$.

Mathematically, this is easy to comprehend.

Given a matrix, one can find it principal eigenvector; but given a principal eigenvector, one cannot "find" a unique matrix.

The Design Problem

Practically, this is a good thing! It means we have flexibility in setting up $\mathbf{W}$.

Different MCMC algorithms make different choices for $\mathbf{W}$.

Usually, the last condition of balance, is replaced with a much stricter condition called detailed balance, since it greatly facilitates algorithm design.

Note that the imposition of detailed balance still does not completely specify $\mathbf{W}$.

It eliminates many choices, but still leaves many open.

Detailed Balance

This is a sufficient condition, which is stricter than the balance condition.

It states that the net traffic between any two states (in "equilibrium") is zero.

$$W_{ij}\pi_j = W_{ji}\pi_i.$$

While more restrictive, it is easier to specify and enforce.

We have sufficient latitude in specifying $\mathbf{W}$ anyway.

If you sum over "j" then, using the normalization condition, you have:

$$\sum_j W_{ij}\pi_j = \sum_j W_{ji}\pi_i \quad \text{detailed balance}$$

$$= \pi_i \sum_j W_{ji} \quad \text{normalization}$$

$$= \pi_i.$$

Thus detailed balance $\implies$ balance (which is required),

But balance $\not\Rightarrow$ detailed balance.

Summary

A Markov process is a stochastic process in which *next* state depends only on the *current* state

$$\pi(x^{(k+1)}|x^{(k)}) \quad \text{(transition probability)}$$

The transition probability matrix describes the transition probabilities between all pairs of states in the state space.

$$W_{ij} = W_{i\leftarrow j} = \pi(x_i^{(k+1)}|x_j^{(k)})$$

The TPM has a useful feature

$$\mathbf{W}\pi^{(\infty)} = \pi^{(\infty)}.$$

We can exploit this by designing a $\mathbf{W}$ that helps us sample from $\pi(x)$.

The TPM has to satisfy: (i) ergodicity, (ii) normalization, and (iii) balance conditions.

In practice the balance condition,

$$\sum_j W_{ij}\pi_j = \pi_i, \qquad i = 1, 2, ..., N$$

is replaced by detailed balance.

$$W_{ij}\pi_j = W_{ji}\pi_i.$$

Even so, there are many $\mathbf{W}$ that correspond to a particular target $\pi(x)$.

These different TPMs correspond to different MCMC methods