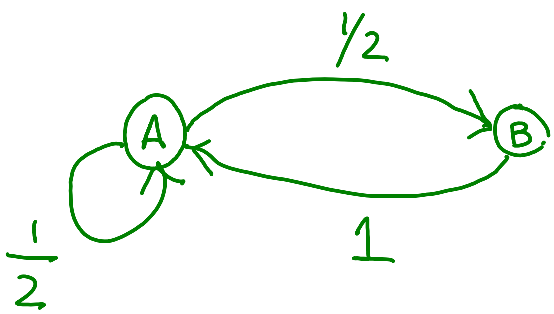
Markov chains (Lecture 1)

Consider the following weighted (multi)-graph[1]



Consider a particle that at time $t=0$ is at A. At time $t=1$, it remains at A with probability 1/2 and goes to b with probability 1/2.

In general if it is at state $X \in \{A, B\}$, at time t, then it goes to state Y at time $(t+1)$ with probability P_{XY}. Here,

$$P_{AB} = P_{AA} = \frac{1}{2}, \quad P_{BA} = 1, \quad P_{BB} = 0$$

– This is an example of a Markov chain. Note the following property:

– Given the state of the particle at time t, the probability distribution of its state at time $t+1$ is independent of its states at times $0, 1, 2, 3, 4, \ldots, t-1$.

[1] i.e., one which possibly has self-loops.

This is called the Markov property.

The Markov property is a rather ubiquitous characteristic of several disparate systems (or, at the very least, of reasonably good models of those systems). Many systems can be seen as a random walk on a graph, which is in some sense a canonical example of a Markov chain.

Random walks on graphs.

$G = (V, E)$ is an undirected graph. At time t, a particle is at some vertex $v \in V$. A probability distribution P_v over the neighbors of v (possibly including v), is provided. The particle samples a vertex u according to P_v and moves to u at time $(t+1)$.

Finite Markov chains: Formal definition

In this class, we will mostly be looking at finite state space Markov chains. Here, one has a state space Ω, and a transition matrix P of dimension $|\Omega| \times |\Omega|$, s.t.

$$P_{ij} = \Pr(X_{t+1} = j \mid X_t = i),$$ where

$\Omega = \{1, 2, \ldots, |\Omega|\}$, and $(X_0, X_1, X_2, \ldots X_t, \ldots)$ is the Markov chain

- In general, the starting state X_0 can also be a random variable, with distribution μ_0.
- Thinking of the distribution μ_t at time t as a row vector in $|\Omega|$ dimensions, we can write

$$\mu_{t+1} = \mu_t P, \quad \text{for all } t \geq 0, t \text{ integer.}$$

Questions about Markov chains

(1) How long does the chain take to reach a given state $\omega \in \Omega$? (Hitting time)

(2) How long does the chain take to explore the whole state space? (Cover time)

(3) Do the distributions μ_t converge to some fixed (i.e. independent of the starting state) distribution as $t \uparrow \infty$? If yes, how long does this take? (Mixing time)

- We will mostly be looking at this last question in this course.

(4) Learning Markov Chains

Setup: – (1) Model a system as a Markov Chain
(2) Observe the evolution of the system over time, possibly several times.
(3) Try to learn the transition matrix

Examples:

(1) Hidden Markov models :-

$$\begin{tikzcd}
\text{States of the system} & X_0 \arrow[r, "P"] \arrow[d] & X_1 \arrow[r, "P"] \arrow[d] & X_2 \arrow[r, "P"] \arrow[d] & X_3 \arrow[r, "P"] \arrow[d] & \cdots \arrow[r, "P"] & X_n \arrow[d] \\
\text{Observations} & Y_0 & Y_1 & Y_2 & Y_3 & & Y_n
\end{tikzcd}$$

Problem: Given observations of the Y_i, learn the transition matrix P.

– Used in audio processing, econometrics etc.
– The learnt matrix P can be used for making predictions.

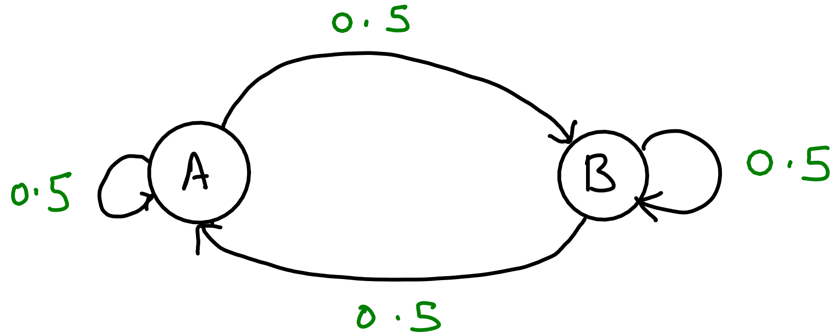
(2) Biology :- Several models of evolution can be formalized as Markov chains.

– E.g. Quasispecies [Eigen '1970s]

→ Again, "learning" the transition can be used to validate and/or make "predictions".

Convergence of Markov chains

An example



$$P = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

Suppose the starting distripution is $\mu_0 = \underbrace{(0.5, 0.5)}_{:=\pi}$. Then

$$\boxed{\mu_0 P = \mu_0}$$

"μ_0 is $\underline{\underline{a}}$ stationary distribution of P"

Consider a different starting distribution:-

$$\mu_0' = (0.1, \; 0.9)$$

Then $\mu_0' P = \pi := (0.5, 0.5)$.

– In fact for this Markov chain,

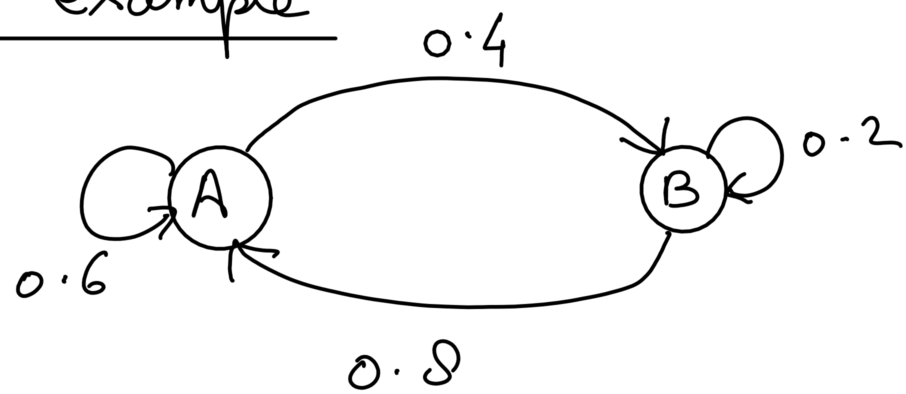
(i) there is <u>only one</u> stationary distribution

(ii) For <u>every</u> starting distribution μ_0, the distribution at time $t = 1$ is the stationary distribution π.

– The chain converges in one step!

– The convergence is exact!

Another example



$$P = \begin{bmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \end{bmatrix}$$

Consider starting distribution $\mu_0 = \left(\frac{2}{3}, \frac{1}{3}\right) = \pi$.

Then $\boxed{\mu_0 P = \mu_0}$ "μ_0 is a stationary distribution"

But now consider a starting distribution $\mu_0' = (0.1, 0.9)$. Then $\mu_1' \neq \mu_0$.

– In fact no starting distribution except μ_0 ever becomes equal to π:

$$\boxed{\forall t \geq 0, \mu' \neq \pi, \ \mu' P^t \neq \pi \quad (\text{Why?})}$$

But, $\forall \mu'$, $\mu' P^t$ does <u>converge</u> to π:

$$\forall \mu', \ \lim_{t \to \infty} \| \mu' P^t - \pi \|_1 = 0$$

[See jupyter notebook]

<u>Note on convergence</u>: For most of the course, we will measure convergen in terms of the

<u>Total variation distance</u>

Let p and q be probability distributions on the same finite state space Ω. Then the total variation distance or statistical distance between p and q, denoted $d_{TV}(p,q)$ is the <u>maximum possible</u> difference in the probability of an event under the two distributions. Thus

$$d_{TV}(p,q) := \max_{S \subseteq \Omega} \left| \sum_{\omega \in S} \big(p(\omega) - q(\omega)\big) \right|$$

<u>Exercise</u>: Show that

(*) $$d_{TV}(p,q) = \sum_{\substack{\omega \in \Omega \\ p(\omega) > q(\omega)}} \big(p(\omega) - q(\omega)\big)$$

(*) $$d_{TV}(p,q) = \frac{1}{2} \|p - q\|_1 \text{ where } \|p-q\|_1 := \sum_{\omega \in \Omega} |p(\omega) - q(\omega)|$$

(*) $$d_{TV}(p,q) = \sup_{f : \Omega \to [0,1]} \left(\mathop{E}_{x \sim p}[f(x)] - \mathop{E}_{x \sim q}[f(x)] \right)$$

Some questions:

(Q) Does every Markov chain have a stationary distribution?

(Q) Is the stationary distribution unique?

(Q) Does the chain "converge" to this stationary distribution?

---

(1) Every Markov chain has a stationary distribution

Proof:- Since P is stochastic,

$$P\vec{1} = \vec{1}$$

$\Rightarrow$ 1 is a (right) eigenvalue of P with eigenvector $\vec{1}$ $\Rightarrow$ 1 is also a (left) eigenvalue of P with eigenvector v (say), $v \neq 0$.

Now, we claim that u, s.t.

$$u_i = |v_i|$$

is also a left eigenvector with eigenvalue 1.

To see this, note that $(uP)_j = \sum_i |v_i| P_{ij} \geq \left|\sum v_i P_{ij}\right|$
$= |v_j| = u_j$

Thus, we have

$$(uP)_j \geq u_j \quad \forall j \quad \text{—} \quad \circledast$$

On the other hand

$$\sum_j (uP)_j = \sum_j \sum_i u_i P_{ij}$$

$$= \sum_i u_i \sum_j P_{ij}$$

$$= \sum_i u_i \ , \quad \text{because} \sum_j P_{ij} = 1 \ \forall i$$

(P is stochastic)

Thus, none of the inequalities in $\circledast$ can be strict, and hence

$$uP = u .$$

By construction, u is a non-zero vector with non-negative entries. It follows that π, defined so that

$$\pi_i = \frac{u_i}{\sum u_j}$$

is a stationary distribution of P. ∎