μ_t : distribution at time t (vector) $= [\Pr[X_t=0] \cdots \Pr[X_t=n]]$

$\mu_t(i) = \Pr[X_t=i]$ → (This must actually be conditioned on the initial distribution μ_0)

$\mu_t = \mu_0 P^t$

If π is a stationary distribution if $\pi P = \pi$

<u>If</u> $\lim_{t\to\infty} \mu_0 P^t$ exists and is equal to π, then it must be a stationary distribution:

$$\lim_{t\to\infty} \mu_t = \pi \Rightarrow \pi P = \pi$$

Does $\lim \mu_t$ exist? Does it depend on the initial distribution?

Brauer's Fixed Point Theorem

If K is compact (closed and bounded) and convex, and if $f: K \to K$ is continuous, then

$$\exists\, x \in K : f(x) = x$$

→ Transition matrix is a linear operator over the probability simplex, Thus we can conclude $\exists\, \pi \in P\text{-simplex}(S) : \pi P = \pi$.

→ **Result** Every Markov chain has <u>at least one</u> stationary distribution.

Will we eventually converge to it? Is it unique?

Some Properties and Definitions Regarding Markov Chains

① Irreducibility:

We say $x \in S$ communicates with $y \in S$ and write $x \to y$ if there is a non-zero probability of reaching y by starting from x:

$$x \to y \iff \exists t \geq 0 : P^t(x,y) > 0$$

A chain is irreducible if "all of its states communicate with one another":

$$\text{Irreducible} := \forall x,y \in S,\ \exists t\ P^t(x,y) > 0$$

1.A Partitioning the state space:

Define $x \leftrightarrow y$ as $x \to y$ and $y \to x$ (i.e., non-zero probability of reaching x from y and vice versa)

$\leftrightarrow$ is an equivalence relation and it partitions S into equivalence classes.

> Note: This is partitioning the graph of the chain into its strongly connected components.

The next theorem will show that an irreducible Markov chain has a unique stationary distribution.

Uniqueness of Stationary Distribution

Theorem: If a Markov chain is irreducible, it has a unique stationary distribution.

Proof: First, we prove the following lemma:

Lemma 1: For any vector $f \in \mathbb{R}^{|S|}$, if $Pf = f$ then f is a constant vector.

→ Let $f_i = M = \max_j f_j$ be the largest component of f. We have

$$M = f_i = (Pf)_i = \sum_j p(i,j) f_j \leq M \sum_j p(i,j) = M$$

($f_j \leq M$)

Thus, for any state s_j that $p(i,j) > 0$ we must have $f_j = M$ (otherwise the r.h.s would be less than the l.h.s)

So, every state s_j that is adjacent to state s_i, $f_j = M$. Following the same argument we can show that for every state s_j that is at distance 2 from s_i, $f_j = M$, and so on. Because the chain is irreducible, all of the states can be reached from s_i, hence, $f_i = M$ for all i.

Now we can prove the main theorem. From lemma 1, we deduce that $\dim \ker P-I \leq 1$ as: $f \in \ker P-I \Rightarrow (P-I)f = 0 \Rightarrow Pf = f \xrightarrow{\text{lemma 1}} f = c \cdot \mathbb{1}$

Using rank-nullity theorem and the fact that for any matrix column rank = row rank along with the fact that P is a square matrix, we can see that:

$$\dim \{u : uP-I) = 0\} = \dim \ker (P-I) \leq 1.$$

We know that there exists a stationary distribution π s.t. $\pi P = \pi$

Thus, if $uP = u$ then $u = c \cdot \pi$ and hence the uniqueness of π. ∎

Some Properties and Definitions Regarding Markov Chains (Cont.)

→ We saw how irreducibility implies the uniqueness of stationary distribution.

But! Will we converge to this distribution? (Meaning that $\lim_{t\to\infty} \mu_t$ exists)

(2) Period and Aperiodicity:

For any state $x \in S$ we define its period as

$$T(x) = \{t \geq 1 : P^t(x,x) > 0\}$$
$$\text{period of } x := \gcd T(x).$$

In irreducible chains, the period for all states are equal and we can define the period for the whole chain.

IDEA: Connection with coloring of directed graphs?

Why? You may ask. Because of the following lemma:

2.A $\quad x \leftrightarrow y \Rightarrow \gcd T(x) = \gcd T(y)$

Proof by some number theoretic manipulations!

→ A Markov chain is said to be aperiodic, if its period is 1.

(Obviously, it must be irreducible to even define its period)

The next theorem shows that an aperiodic chain will converge to its stationary distribution.

Convergence to Stationary Distribution

Theorem: If a Markov chain is irreducible and aperiodic then it has a unique stationary distribution π and $\lim_{t\to\infty} \mu_t = \pi$, irrespective of the initial distribution μ_0!

Proof: It's a bit long and involved!

On [Being and] Time!

① Hitting Time :

For a subset A of states, we are interested in when we arrive at (hit) them. To this end, for any $A \subseteq S$ we define the following notions representing hitting time:

$$\tau_A := \min \{ t \geq 0 : X_t \in A \},$$

$$\tau_A^+ := \min \{ t > 0 : X_t \in A \}.$$

→ Notice that τ_A, τ_A^+ are both <u>random variables</u> that depend on the initial distribution!

In many scenarios, we start from a fixed state x and would like to see how long will it take before we reach a state y. The expected value of this random quantity is denoted by:

$$\mathbb{E}_x [\tau_y] = \mathbb{E}[\tau_y \mid X_0 = x].$$

We may also be interested in the time that it takes to return to the starting state:

$$\mathbb{E}_x [\tau_x^+] \rightarrow \text{Return time}$$

On Being and Time (Cont.)

→ A Friendly Reminder for Calculating Conditional Expectations:

If X is a random variable and A, B are events:

$$\mathbb{E}[X|A] = \Pr[B|A]\mathbb{E}[X|A,B] + \Pr[B^c|A]\mathbb{E}[X|A,B^c]$$

→ Why did I write this reminder? Because when calculating the expected hitting time, conditioning on what happens at the beginning is often a good idea!

Theorem: [Irreducibility ⇒ Almost surely finite hitting time]

In any finite, irreducible Markov chain,

$$\forall x,y \in S : \Pr_x[\tau_y^+ < \infty] = 1.$$

Proof:

Because of irreducibility, there exists T s.t. $\Pr_z[\tau_y^+ \leq T] \geq \varepsilon > 0$, for all $z \in S$ (Let $T = \max\{\min\{t : P^t(z,y) > 0\} : z \in S\}, \ldots$)

Thus, $\Pr_x[\tau_y^+ > kT] \leq (\Pr_z[\tau_y^+ > T])^k < (1-\varepsilon)^k$.

By letting $k \to \infty$, we see that $\Pr_x[\tau_y^+ = \infty] = 0$. ∎

On Being and Time (Cont.)

Theorem: [Irreducibility => Finite expected hitting time]

In any finite, irreducible Markov chain,

$$\forall x, y \in S : \mathbb{E}_x[\tau_y^+] < \infty.$$

Proof:

Using the previous theorem and by some clever notational trickery, we can show that there exists $c \in \mathbb{R}$, $\rho \in [0,1)$ such that

$$\Pr_x[\tau_y^+ > t] \leq c\rho^t \quad \text{, for all } t = 0, 1, \ldots.$$

Now, we can write:

$$\mathbb{E}_x[\tau_y^+] = \sum_{t=1}^{\infty} \Pr_x[\tau_y^+ > t] \leq c \sum_{t=1}^{\infty} \rho^t < \infty. \quad \blacksquare$$

The two last theorems show that in an irreducible chain, regardless of the starting position, we will almost surely hit any state $s \in S$ and will do so in an expected finite time.

→ The next question that we'll investigate is "how to compute $\mathbb{E}_x[\tau_y^+]$, now that we know its finite?". We'll see that this quantity is actually related to the stationary distribution π!

Expected Hitting Time and The Stationary Distribution

Let's start by a definition: $N_t(x) = |\{r \in \{1,\dots,t\} : X_r = x\}|$ for all $x \in S$.

Notice that $N_t(x)$ is a random variable, denoting the state visitation count for state x until time t.

Intuitively, $\mathbb{E}_x[\tau_x^+] \approx \frac{t}{N_t(x)}$.

On the other hand, $\frac{N_t(x)}{t}$ shows us the fraction of time that was spend at state x.

So we expect this to approach the stationary distribution, i.e., $\frac{N_t(x)}{t} \approx \pi(x)$.

This hand-wavy argument results in the following relation:

$$\mathbb{E}_x[\tau_x^+] = \frac{1}{\pi(x)}$$

Our next task is to rigorously prove this equation!

Expected Hitting Time and Stationary Distribution

Theorem: For any finite, irreducible Markov chain,

$$\mathbb{E}_x[\tau_x^+] = \frac{1}{\pi(x)} \quad \text{for all } x \in S$$

Proof:

Let $x \in S$ be an arbitrary state, π be the unique stationary distribution of the chain, and τ_i $(1 \leq i \leq n)$ be a random variable, indicating the time between the $i-1$-th visit to x and the i-th one.

By the Markov property, $\tau_1, \tau_2, \ldots \tau_n, \tau_x^+$ are all i.i.d. variables.

So, $\mathbb{E}_x[\tau_i] = \mathbb{E}_x[\tau_x^+]$.

By the law of large numbers, as $n \to \infty$, $\frac{\overbrace{\tau_1 + \cdots + \tau_n}^{\to t}}{\underbrace{n}_{N_t(x)}} = \mathbb{E}_x[\tau_x^+]$

So we have $\frac{1}{\mathbb{E}_x[\tau_x^+]} = \frac{\mathbb{E}_{\mu_0}[N_t(x)]}{t}$, as $t \to \infty$. (I)

On the other hand, we have:

$$N_t(x) = \sum_{r=1}^{t} \mathbb{1}(X_r = x) \Rightarrow \mathbb{E}_\pi[N_t(x)] = \sum_{r=1}^{t} \underbrace{\Pr_\pi[X_r = x]}_{\pi(x)} = t\,\pi(x) \quad \text{(II)}$$

Combining (I), (II) gives us the desired result. ☒

Countably Infinite State Spaces

So far, we've only considered finite Markov chains. But what about chains with infinitely many states? Do they have a stationary distribution? Is it unique? Will we converge to it?

Remember that we defined the hitting time as

$$\tau_A^+ = \min\{t > 0 : X_t \in A\},$$

to represent the time that it takes to reach a [set of] state[s].

Definition: We define $\boxed{\rho_{xy} = \Pr_x[\tau_y^+ < \infty]}$ to represent the probability of reaching y, in finite time, given that we've started from x.

Definition: We call a state $x \in S$ "recurrent" if starting from x, we eventually return to it with probability 1. Otherwise, that state is called "transitive":

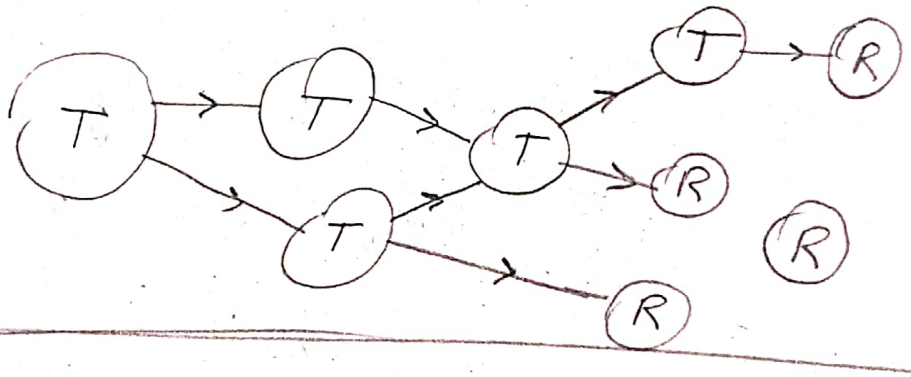
$x \in S$ is recurrent $\rightarrow \rho_{xx} = \Pr_x[\tau_x^+ < \infty] = 1$,

$x \in S$ is transitive $\rightarrow \rho_{xx} = \Pr_x[\tau_x^+ < \infty] < 1$.

$\rightsquigarrow$ We have seen that in irreducible, finite chains, all of the states are recurrent.

Countably Infinite State Spaces (Cont.)

An Aside: In finite chains, if we consider the topologically sorted version of the graph of strongly connected components of the state-transition graph, every state in a component with no outgoing edge is a recursive state and all others are transitive.



Similar to the finite chains, we can see that the "mutual communication" relation ($\leftrightarrow$: existance of a finite! cycle between two states) is again an equivalence relation that partitions the state space into equivalence classes.

We will further devide recurrent states into "positive" and null recurrent, and see that the states in each equivalence class of $\leftrightarrow$ relation are either null recurrent, positive recurrent, or transitive.

Keeping the topologically sorted classes in mind is again helpful, although now each class may contain infinitely many states!

Number of Returns

Before proceeding to present the state classification theorems, we need one more definition.

Definition: For a state $x \in S$, let $N(x)$ be a random variable denoting the number of times x is visited:

$$N(x) = |\{t > 0 : X_t = x\}|$$

Theorem: In a Markov chain, a state $x \in S$ is recursive, if and only if $N(x)$ is almost surely infinite: $P_{xx} = 1 \iff Pr_x[N(x) = \infty] = 1$

Proof:

a) [$\Rightarrow$] Clearly, $Pr_x[N(x) \geq 1] = 1$, because $P_{xx} = 1$ means that we will return to x (at least once) with probability 1. Inductively assume $Pr_x[N(x) \geq n] = 1$. Then:

$$Pr_x[N(x) \geq n+1] = \underbrace{Pr[N(x) \geq n]}_{1} \underbrace{Pr_x[N(x) \geq n+1 \mid N(x) \geq n]}_{\text{recursiveness of } x \text{ + strong Markov property} \rightarrow = 1}$$

$$= 1.$$

b) [$\Leftarrow$] If $Pr_x[N(x) = \infty] = 1$ then $Pr_x[N(x) \geq 1] = 1$. This means that almost surely there exists $t > 0$ such that $X_t = x$. Thus,

$$Pr_x[\tau_x^+ < \infty] = 1 = P_{xx}$$

If we almost surely return to x, then we almost surely return to it infinitely many times! ☒

Number of Returns (Cont.)

Theorem: For any two states $x, y \in S$ of a Markov chain,

$$\mathbb{E}_x[N(y)] = \frac{P_{xy}}{1-P_{yy}}.$$

Proof:

$$\mathbb{E}_x[N(y)] = \sum_{k=1}^{\infty} \Pr_x[N(y) \geq k]$$

$$= \sum_{k=1}^{\infty} \underbrace{\Pr_x[\tau_y^+ < \infty]}_{x \to y} \underbrace{\Pr_y[N(y) \geq k-1]}_{(k-1)\times\ y \to y} = P_{xy} \sum_{k=1}^{\infty} P_{yy}^{k-1}$$

(strong Markov property)

$$= \frac{P_{xy}}{1-P_{yy}} \quad \blacksquare$$

Corollary: $\mathbb{E}_x[N(x)] = \frac{P_{xx}}{1-P_{xx}}$. Thus, if x is recursive ($P_{xx}=1$) then $\mathbb{E}_x[N(x)] = \infty$, AND for $\mathbb{E}_x[N(x)]$ to be ∞, it must be that $P_{xx}=1$, meaning that x is recursive. Hence,

$$P_{xx} = 1 \iff \mathbb{E}_x[N(x)] = \infty$$

Recursive States

The two last theorems show us three equivalent ways of characterizing recursive states:

① $\rho_{xx} = 1$: [By definition] a state x is recursive if we almost surely return to it (in finite time):

$$\rho_{xx} = \Pr_x[\tau_x^+ < \infty] = 1$$

② $\Pr_x[N(x) = \infty] = 1$: [By the theorem p.13] a state x is recursive if we almost surely visit it infinitely many times:

$$\Pr_x[N(x) = \infty] = 1$$

③ $\mathbb{E}_x[N(x)] = \infty$: [By the theorem p.14] a state x is recursive if we expect to visit it infinitely many times:

$$\mathbb{E}_x[N(x)] = \infty$$

Classification of States

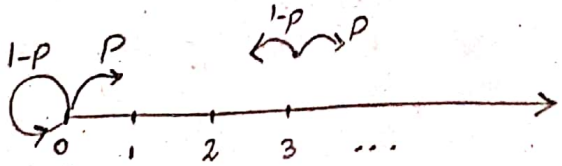
In a finite Markov chain, a state is of one of the following classes:

① Transitive : $\rho_{xx} = \Pr_x[\tau_x^+ < \infty] < 1$

② Recurrent : $\rho_{xx} = \Pr_x[\tau_x^+ < \infty] = 1$

- → (2.1) Positive Recurrent : $\mathbb{E}_x[\tau_x^+] < \infty$.
- → (2.2) Null Recurrent : $\mathbb{E}_x[\tau_x^+] = \infty$

→ Pay attention to null recurrent states' definition. If x is null recurrent, we will almost surely return to it in finite time (even more, we will almost surely return to it infinitely many times!) HOWEVER, the expected return time is not finite! How can this be? Consider a random walk on $\mathbb{N}$:



You can show that all of the states are
- Positive Recurrent if $p < \frac{1}{2}$
- Transitive if $p > \frac{1}{2}$
- Null Recurrent if $p = \frac{1}{2}$

Notice that if $\mathbb{E}_x[\tau_x^+] < \infty$, then it immediately follows that $\Pr_x[\tau_x^+ < \infty] = 1$.

Classification of States (Cont.)

Theorem: If x is recursive and $x \to y$, then y is also recursive.

$$\rho_{xx} = 1,\ x \to y \implies \rho_{yy} = 1$$

Proof:

First, notice that we have $\mathbb{E}_x[N(y)] = \sum_{t=1}^{\infty} p^t(x,y)$ because

$$N(y) = \sum_{t=1}^{\infty} \mathbb{1}(X_t = y) \Rightarrow \mathbb{E}_x[N(y)] = \sum_{t \geq 1} p^t(x,y).$$

Now to show that y is recurrent, it suffices to show that $\mathbb{E}_y[N(y)] = \infty$. Because $x \to y$ and x is recurrent, it must be that $y \to x$ (otherwise $\rho_{xx} \leq 1 - \rho_{xy} < 1$). So we must have $a, b \in \mathbb{N}$ such that

$$p^a(x,y) > 0\ ,\ p^b(y,x) > 0.$$

Let r be any natural number. From Kolmogorov-Chapman equality we have:

$$p^{a+r+b}(y,y) \geq p^b(y,x)\, p^r(x,x)\, p^a(x,y).$$

Hence

$$\mathbb{E}_y[N(y)] = \sum_{t \geq 1} p^t(y,y) \geq \sum_{r \geq 1} p^{a+r+b}(y,y)$$

$$\geq \underbrace{p^b(y,x)\, p^a(x,y)}_{>0} \underbrace{\sum_{r \geq 1} p^r(x,x)}_{\mathbb{E}_x[N(x)] = \infty} = \infty$$

□

Classification of States (Cont.)

Corollary: if $x \leftrightarrow y$ then x is recursive if and only if y is:

$$x \leftrightarrow y \implies (\rho_{xx} = 1 \iff \rho_{yy} = 1)$$

This means that in any equivalence class of $\leftrightarrow$ relation (strongly connected components of the state-transition graph) either all of the states are transitive or they are all recursive.

Corollary: If $x \to y$ and $\rho_{yx} < 1$, then x is transitive.

Proof: Because $x \to y$, there exists $t \geq 1$ s.t. $P^t(x,y) > 0$.

Also, we have: $\Pr_x[\tau_x^+ = \infty] \geq \Pr_x[\tau_x^+ = \infty \mid X_t = y] \Pr_x[X_t = y]$

$\Pr_x[\tau_x^+ = \infty] = 1 - \rho_{xx}$; $\Pr_x[\tau_x^+ = \infty \mid X_t = y] = (1 - \rho_{yx}) > 0$; $\Pr_x[X_t = y] = P^t(x,y) > 0$

$\Rightarrow \rho_{xx} < 1$. So x is transitive.

This corollary shows us how we can identify a transitive state. If there is a state y that is reachable from x, but that won't almost surely return to x, then x is transitive, as there is a strictly positive probability of going to y and never returning to x.

Intuitively, it's just saying that if the scc of x is not a leaf in the topologically sorted state transition graph, then it must be a transitive state.

Classification of States (Cont.)

Theorem: Assume that $y \in S$ is a transitive state. For every $x \in S$

$$\lim_{t\to\infty} P^t(x,y) = 0$$

(as an immediate result of which $\lim_{t\to\infty} \mu_t(y) = 0$)

Proof: Because y is transitive

$$E_x[N(y)] < \infty \Rightarrow \sum_{t=1}^{\infty} P^t(x,y) < \infty.$$

But from analysis, we know that if $\sum a_i < \infty$ converges, then it's a Cauchy sequence and therefore $\lim_{i\to\infty} a_i = 0$, Proving that $\lim_{t\to\infty} P^t(x,y) = 0$. ∎

Question: How do we know that $E_x[N(y)]$ converges?

Answer: Remember the theorem on P.14: $E_x[N(y)] = \frac{P_{xy}}{1-P_{yy}}$.

Stationary Distribution

Theorem: For any irreducible Markov chain, the following propositions are equivalent: (↳ $(\forall x,y\ x \to y)$)

i) There exists a positive recurrent state,

ii) There exists a stationary distribution,

iii) Every state is a positive recurrent state.

Proof:

a) (ii → iii) Assume that π is a stationary distribution for the chain.

Recall that by the theorem on p.10, $\mathbb{E}_x[\tau_x^+] = \frac{1}{\pi(x)}$ for every state x.

(The proof on p.10 works for the infinite state space as well).

Thus, if we show that $\forall x \in S,\ \pi(x) > 0$, then it follows that $\mathbb{E}_x[\tau_x^+] < \infty$ for every $x \in S$, proving that every state is positive recurrent.

Let y be a state with non-zero stationary probability: $\pi(y) > 0$.

For any state x, $y \to x$, thus, there exists t such that $P^t(y,x) > 0$.

But now we have

$$\pi(x) = \Pr_\pi[X_t = x] \geq \underbrace{\Pr_\pi[X_0 = y]}_{\pi(y)} P^t(y,x) > 0 .$$

Showing that indeed every state has a non zero probability in the stationary distribution.

Stationary Distribution (Cont.)

b) ($i \to ii$) We now focus on proving the existance of a stationary distribution in any irreducible chain with a positive recurrent state.

Let $N^z(x)$ be a random variable denoting the number of visits to x before arriving at z. We will prove the following lemma, which will automatically give us the desired result:

Lemma: Let $z \in S$ be a positive recurrent state in an irreducible Markov chain. If

$$\boxed{\mu_z(x) = \mathbb{E}_z[N^z(x)],}$$

then we have an stationary distribution by setting $\pi(x) = \dfrac{\mu_z(x)}{\sum_{y \in S} \mu_z(y)}$.

Proof: To show that π is a stationary distribution, we must prove first.

$$\mu^T P = \mu \iff \forall y \in S \quad \sum_x \mu_z(x) P(x,y) = \mu_z(y).$$

Recall (P.6) that τ_z^+ is the time it takes to return to z. We can write:

$$\mu_z(x) = \mathbb{E}_z[N^z(x)] = \mathbb{E}_z\Big[\sum_{t \geq 0} \mathbb{1}(X_t = x, t < \tau_z^+)\Big]$$

$$= \sum_{t \geq 0} \underbrace{\Pr_z[X_t = x, t < \tau_z^+]}_{:= \hat{P}_t(z,x)} = \sum_{t \geq 0} \hat{P}_t(z,x)$$

$\to \hat{P}_t(z,x)$ is the probability of visiting x at time t before the first visit to z.

We also have: $\sum_x \mu_z(x) P(x,y) = \sum_{t \geq 0} \sum_x \hat{P}_t(z,x) P(x,y)$.

$$\sum_x \hat{P}_t(z,x) P(x,y) = \sum_x \Pr_z[X_t = x, X_{t+1} = y, t < \tau_z^+] = \Pr_z[X_{t+1} = y, t < \tau_z^+]$$

$$= \hat{P}_{t+1}(z,y)$$

IF $y \neq z$! +
Some notational abuse!

Stationary Distribution (Cont.)

Putting all of these together, we must prove the following:

$$\sum_{t\geq 0} \hat{P}_t(z,x) = \sum_{t\geq 0} \hat{P}_{t+1}(z,x) = \sum_{t\geq 1} \hat{P}_t(z,x).$$

($t\geq 0$: $t = 0, \dots, \tau_z^+ - 1$; $t\geq 1$: $t = 1, \dots, \tau_z^+$)

Consider two cases:

i) $x \neq z$: $\hat{P}_0(z,x) = \Pr_z[X_0 = x, \dots] = 0$, $\hat{P}_{\tau_z}(z,x) = 0$ ✓

ii) $x = z$: $\hat{P}_0(z,x) = 1 = \hat{P}_{\tau_z}(z,x)$ ✓

Now that we've shown that $\mu_z(x)$ is invariant under the transition transformation, we can devide μ_z's by their sum to get a probability distribution π, that is also P-invariant. However, one final thing is left, proving that the denominator of π, is finite:

$$\sum_x \mu_z(x) = \sum_x \mathbb{E}_z[N^z(x)] = \mathbb{E}_z\left[\sum_x N^z(x)\right] = \mathbb{E}_z[\tau_z^+]$$

$$< \infty.$$

→ This last derivation is intuitively simple: By summing the total number of visits to each state prior to the first return to z, we are effectively computing the time it takes to return to z, which is finite because z is positive recursive.

□

Uniqueness of Stationary Distribution

We saw that an irreducible Markov chain has a stationary distribution, if and only if it was positive recurrent. The next question that naturally follows is

"When is the stationary distribution unique?"

Remark: Remember that in finite chains, irreducibility implies existance AND uniqueness of the stationary distribution. Similarly, we shall see that irreducibility + positive recurrence implies existance AND uniqueness of the stationary distribution for infinite chains.

Let's start by a definition:

Definition: For any state x of a Markov chain, let

$$N_t(x) := \left|\{0 \leq t' \leq t : X_{t'} = x\}\right|$$

be a random variable denoting the number of times x was encountered in the first t steps.

Uniqueness of Stationary Distribution (Cont.)

Theorem: For any recursive state x in an irreducible Markov chain,

$$\lim_{t\to\infty} \frac{N_t(x)}{t} = \frac{1}{E_x[\tau_x^+]}, \quad \text{almost surely.}$$

Proof: Let τ_i $(i \geq 1)$ be a random variable representing the time between the i-th and $i+1$-th visit to x.

It can be seen that $\tau_1, \tau_2, \cdots, \tau_x^+$ are all i.i.d. random variables.

Thus, the law of large numbers gives us (almost surely)

$$\lim_{k\to\infty} \frac{\tau_1 + \cdots + \tau_k}{k} = E_x[\tau_x^+].$$

Let $T(k) = \tau_1 + \cdots + \tau_k$. For any given t, $T(N_t(x)) \leq t \leq T(N_t(x)+1)$. ($N_t(x)$: x's visitation count until time t.)

Now we can write:

$$\frac{T(N_t(x))}{N_t(x)} \leq \frac{t}{N_t(x)} \leq \frac{T(N_t(x)+1)}{N_t(x)+1} \frac{N_t(x)+1}{N_t(x)}$$

By letting $t\to\infty$, all three expressions will converg to $E_x[\tau_x^+]$, proving that almost surely $\frac{N_t(x)}{t} \to \frac{1}{E_x[\tau_x^+]}$, almost surely.

□

Uniqueness of Stationary Distribution (Cont.)

Theorem: [Uniqueness] Any irreducible, positive recurrent Markov chain has a unique stationary distribution defined by

$$\pi(x) = \frac{1}{\mathbb{E}_x[\tau_x^+]} \quad \forall x \in S$$

Proof:

The previous theorem implies that as $t \to \infty$, $\frac{N_t(x)}{t} \to \frac{1}{\mathbb{E}_x[\tau_x^+]}$, regardless of the initial distribution, X_0. In particular, if π is any stationary distribution for the chain (we know that at least one π exists) then

$$\frac{\mathbb{E}_\pi[N_t(x)]}{t} \xrightarrow{a.s.} \frac{1}{\mathbb{E}_x[\tau_x^+]}.$$

But also,

$$N_t(x) = \sum_{r=1}^{t} \mathbb{1}(X_r = x) \Rightarrow \mathbb{E}_\pi[N_t(x)] = t\pi.$$

Thus,

$$\lim_{t\to\infty} \frac{t\pi}{t} = \pi = \frac{1}{\mathbb{E}_x[\tau_x^+]}.$$

□

Fun Question: Let's say $\{X_t\}_{t=0}^{\infty}$ is a sequence of random variables that almost surely converge. Why can we take the expected value of the limit? That is,

$$X_t \xrightarrow{a.s.} X \Rightarrow \mathbb{E}[X_t] \to \mathbb{E}[X]$$

The Ergodic Theorem

Theorem: For any irreducible, positive recurrent Markov chain with the unique stationary distribution π, and a function $f: S \to \mathbb{R}$ with $\mathbb{E}_{X\sim\pi}[|f(X)|] < \infty$, we have:

$$\frac{1}{t}\sum_{s=0}^{t} f(X_s) \xrightarrow{a.s.} \mathbb{E}_\pi[f(X_0)].$$

→ What is the significance of this theorem?

Basically, it's saying that if we assign a number $f(x)$ to each state, then the time and space averages are equal!

→ To see an immediate result, let $f(x) = 1$ for some $x \in S$ and 0 elsewhere.

Then the right hand side is just $\mathbb{E}_\pi[\mathbb{1}(X=x)] = \pi_x$.

The left hand side is the normalized visitation count of state x.

So if $N_t(x)$ is the number of visits to x in the first t steps, the ergodic theorem implies that $\frac{N_t(x)}{t} \xrightarrow{a.s.} \pi_x$!

→ Another implication is that the long term behaviour of system ($\frac{1}{t}\sum f(X_s)$) is almost always the same, irrespective of the initial starting distribution.

Mixing Time and Convergence

We saw that for finite chains, irreducibility and aperiodicity implied convergence to the (unique) stationary distribution, i.e., $\lim_{t\to\infty} \mu_t = \pi$. We would like to study the convergence properties of infinite chains as well.

However, in the infinite state space setting we must thread carefully and define what we mean by "converge". Because unlike the finite case, here we are dealing with infinite dimensional entities ($\mu_t : \mathbb{N} \to [0,1]$ for instance) and different notions of convergence can be considered (point-wise vs. uniform)

To this end, we will define a distance function between distributions and study convergence under this metric.

But before that, we will study "mixing time", a random variable indicating the time it takes to get to the stationary distribution, exactly.

Let $f_1, f_2, \ldots$ be a sequence of functions from S to $\mathbb{R}$ and $f: S \to \mathbb{R}$. There are multiple ways of defining $\lim_{t\to\infty} f_t = f$.

① Point-wise convergence : $\forall x \in S \quad \lim_{t\to\infty} f_t(x) = f(x)$.

② Uniform convergence : $\lim_{t\to\infty} \sup_{x\in S} |f_t(x) - f(x)| = 0$

Uniform convergence is stronger than point-wise convergence. If S is finite, they are equivalent: ① $\Rightarrow$ ② $\rightarrow$ Only when $|S| < \infty$

② $\Rightarrow$ ① $\rightarrow$ Always.

Stationary Time

To formalize the notion of "the time it takes for the chain to be mixed" we define a stopping time:

Definition: Given any state $x \in S$, the stationary time τ is a random variable such that $\Pr_x[X_\tau = y] = \pi(y)$ for all $y \in S$.

τ is a stopping time, that is, its defined as the minimum time it takes to reach π. Formally:

$$\tau = \min\{t \geq 0 : \Pr_x[X_t = y] = \pi(y) \ \forall y \in S\}$$

→ It can be seen that the stationary time τ is dependent on the initial distribution.

→ We can use $E_x[\tau]$ to understand the mixing time of the chain.

→ Calculating $E_x[\tau]$ is often not easy!

→ Calculating the distribution of τ is even harder!

→ What can we do?

Mixing Time and Total Variation Distance

Instead of working out the time it takes to reach the stationary distribution exactly, we consider the time it takes to get "close" to it!

Definition: Let d be a distance function and $\varepsilon > 0$. The ε-mixing time is defined as

$$t_{mix}(\varepsilon) = \min\{t : d(\mu_t, \pi) < \varepsilon\}.$$

→ Notice that $t_{mix}(\varepsilon)$ is <u>not</u> a random variable! It's much easier to work with it compared to stationary time, τ.

How should we define the distance function d? There are a number of options:

→ (Analysis) L_p norm : $\|\cdot\|_p$

→ (Measure Theory) Wasserstein distance : W_p

→ (Probability Theory) KL-divergence, total-variation distance

$\downarrow$ D_{KL} $\quad$ $\downarrow$ d_{tv} or δ

For our purposes, the total variation distance is very convenient.
Next, we will study some properties of this distance.

Total Variation Distance

Definition: Let μ, ν be two probability distributions over S. The total variation distance between them is defined as

$$d_{tv}(\mu,\nu) := \sup_{A \subseteq S} |\mu(A) - \nu(A)|$$

Theorem: Alternatively, we can find the total variation distance, by only considering the individual points in S, using the following equality:

$$d_{tv}(\mu,\nu) = \frac{1}{2} \sum_{x \in S} |\mu(x) - \nu(x)| = \frac{1}{2} \|\mu - \nu\|_1$$

As an immediate result, d_{tv} is a metric. (Hence, the triangle inequality!)

Probabilistic View of d_{tv}:

Let's say X and Y are random variables that have distributions μ, ν respectively. Is there any relation between $X, Y, d_{tv}(\mu,\nu)$? Yes!

Theorem: Let $X \sim \mu_X$, $Y \sim \mu_Y$ be random variables. We have:

$$d_{tv}(\mu_X, \mu_Y) = \min_{(X,Y)} \Pr[X \neq Y],$$

where the minimum is taken over all possible couplings of X, Y.

Coupling

In the last theorem on total variation distance, we used the word coupling.

Intuitively, a coupling for two random variables X, Y is determining the dependence of them on one another, or determining a joint distribution for them.

Definition: Let μ and ν be distributions over S, R, respectively.

A coupling of μ and ν is a (joint) distribution over $S \times R$ such that

$$\forall x \in S : \sum_{y \in R} \lambda(x,y) = \mu(x),$$

$$\forall y \in R : \sum_{x \in S} \lambda(x,y) = \nu(y).$$

→ Thus, to provide a coupling for random variables X, Y, we must determine how they are related to each other, or provide a joint distribution for them that gives us X, Y if we marginalize it.

→ Let's now revisit that last theorem. Given two distributions μ_X, μ_Y, in order to find $d_{tv}(\mu_X, \mu_Y)$, we can consider random variables $X \sim \mu_X$, $Y \sim \mu_Y$, couple them in such a way that $\Pr[X \neq Y]$ is minimized (i.e., X, Y agree with one another in most cases) and report this probability as d_{tv}.

→ As an immediate result, for any coupling of X, Y, $d_{tv}(\mu_X, \mu_Y) \leq \Pr[X \neq Y]$

Bounding the Mixing Time

Let's say (your) want to bound the mixing time for a chain, assuming we have started from the initial distribution μ_0. Recall that the mixing time is defined as

$$t_{mix}(\varepsilon) = \min\{t : d_{tv}(\mu_t, \pi) < \varepsilon\}.$$

Conventionally, $t_{mix} := t_{mix}(\frac{1}{4})$.

We proceed in the following manner: Consider two walks on this chain X_t, Y_t with $X_0 = \mu_0$, $Y_0 = \pi$. Because π is the stationary distribution, $Y_t = \pi \ \forall t$.
Also, note that $X_t \sim \mu_t$.

Consider any coupling for X_t, Y_t. From the last theorem on page 30, we have

$$d_{tv}(\mu_t, \pi) \leq \Pr[X_t \neq Y_t].$$

If we consider a coupling in which after X_t, Y_t collide for the first time, they "stick together" and move just like one another (i.e., $X_t = Y_t$, after collision), then we have

$$\Pr[X_t \neq Y_t] = \Pr[\text{No collision until time } t]$$
$$= \Pr[\tau > t] \quad \boxed{\leq \frac{\mathbb{E}[\tau]}{t}} \rightarrow \text{Markov inequality}$$

Where $\tau = \min\{t \geq 0 : X_t = Y_t\}$ is the time of first collision.

★ Thus, by providing a coupling in which X_t, Y_t collide quickly, we can get a small $\Pr[\tau > t]$ as an upper bound on $d_{tv}(\mu_t, \pi)$.
Finally, by solving $\Pr[\tau > t] < \varepsilon$ for t, we can get a bound on $t_{mix}(\varepsilon)$. (32)

Convergence Theorem

Theorem: For any irreducible, aperiodic, and positive recurrent Markov chain,

$$\lim_{t \to \infty} d_{tv}(\mu_t, \pi) = 0$$

That is, we converge to the stationary distribution in the sense of d_{tv} (uniform). For infinite state spaces, this is strictly stronger than pointwise convergence:

$$\lim_{t \to \infty} \mu_t(x) = \pi(x) \quad \forall x \in S.$$