Notes on Markov Mixing Times by Peres et al.

1 Basics

Definition 1. Let Ω be a finite set we call our state space, and consider a sequence of $\Omega-$valued random variables: $(X_0, X_1, \ldots)$. Such a sequence is a *Markov chain with state space Ω and transition matrix P* where the rule that generates the sequence is:

$$\begin{aligned}\mathbf{P}\{X_{t+1} = y | X_t = x, X_{t-1} = z_1, X_{t-2} = z_2, \ldots\} &= \mathbf{P}\{X_{t+1} = y | X_t = x\} \\ &= P(x, y)\end{aligned}$$

Notice that for this to make sense, P must be a stochastic matrix:

$$\sum_{y \in \Omega} P(x, y) = 1 \text{ for any } x \in \Omega$$

Notice that if we have a vector row vector $\mu_t = (\mathbf{P}(X_t = \omega_1), \mathbf{P}(X_t = \omega_2), \ldots, \mathbf{P}(X_t = \omega_n))$ which stores the probability distribution of X_t, then the rule for updating μ_t at a later time is given by matrix-vector multiplication by P on the right:

$$\mu_{t+k} = \mu_t P^k$$

Of particular interest in our study will be so called *stationary distributions* of the Markov chain, which are invariant under P, that is to say they satisfy:

$$\pi = \pi P$$

Proposition 2. *(Random Mapping Representation)*

Any Markov chain can be thought about in the following way, known as the random mapping representation. Take i.i.d variables $Z_n \sim Unif([0,1])$ and define a determenistic function $f : \Omega \times [0,1] \to \Omega$ so that:

$$\mathbf{P}\{f(x, Z) = y\} = P(x, y)$$

(This can be done by making $f(x, z) = y$ for $z \in I_y \subset [0,1]$ with $|I_y| = P(x, y)$). With this set up, we realize the Markov chain recursivly by the rule:

$$X_{t+1} = f(X_t, Z_t)$$

One of the advantages of thinking about it this way,is that all of the randomness is wrapped up into the $Z_t's$, and f is a determenistic function.

Definition 3. A Markov chain is called *irreducible* if for any $x, y \in \Omega$ there exists some t such that $P^t(x, y) > 0$. If we think of a Markov chain as a directed graph where vertices are states $\omega \in \Omega$, and edges connect edges for which $P(x, y) > 0$, then saying the Markov chain is *irreducible* is the same as saying the graph is *connected.*

Definition 4. The *period* of a state $\omega \in \Omega$ is defined to be $\gcd\{t \geq 1 : P^t(x, x) > 0\}$.

Proposition 5. *For an irreducible Markov chain, the period of any two states are equal. For this reason, it makes sense to say "the period of the Markov chain"*

Proof. Say we have two states $x, y \in \Omega$. Since its irreducible, we can find a loop (i.e. a path of positive probability) starting at x, going through y, and returing to x. Say this loop is lenght m. Then $m \in \{t \geq 1 : P^t(x,x) > 0\}$ and $m \in \{t \geq 1 : P^t(y,y) > 0\}$. Any other loop starting and ending at y can be turned into a loop starting and ending at x by gluing on these paths of total length m. Hence $\{t \geq 1 : P^t(x,x) > 0\} + m \subset \{t \geq 1 : P^t(y,y) > 0\}$. Taking gcdhere are keeping in mind that $\gcd | m$ here, we conclude $\gcd\{t \geq 1 : P^t(x,x) > 0\} \geq \gcd\{t \geq 1 : P^t(y,y) > 0\}$. The opposite inequality holds by the same argument. □

Definition 6. A Markov chain of period 1 is called *aperiodic.*

Proposition 7. *(From Rosenthal) If a Markov chain is irreducible and aperiodic, then for each pair $(x,y) \in \Omega^2$, there is a number $n_0 = n_0(i,j)$ so that for all $n > n_0$ $P^n(x,y) > 0$.*

Proof. Let $T = \{t \geq 1 : P^t(x,x) > 0\}$. Since the Markov chain is aperiod, $\gcd T = 1$, so find integers $k_1, k_2, \ldots, k_m \in T$ and $b_1, b_2, \ldots b_m \in \mathbb{Z}$, so that $b_1k_1 + \ldots + b_mk_m = 1$. Additionaly, choose $a \in T$ and c such that $P^c(x,y) > 0$ (by irreducibility). Now let $M = |b_1|k_1 + |b_2|k_2 + \ldots + |b_m|k_m$, and let $n_0 = aM + c$. For any $n \geq n_0$, write $n = n_0 + qa + s$ where $0 \leq s < a$ comes from the division algorithm. Now consider:

$$\begin{aligned} n &= n_0 + qa + s \\ &= aM + c + qa + s \\ &= a\sum_{l=1}^{m} |b_l|k_l + c + qa + s\sum_{l=1}^{m} b_lk_l \\ &= qa + \sum_{l=1}^{m} \left(|b_l|a + b_ls\right)k_l + c \end{aligned}$$

Notice each term $|b_l|a + b_ls$ is positive since $s < a$. With this rewriting we have that:

$$P^n(x,y) \geq \left(P^a(x,x)\right)^q \left(\prod_{l=1}^{m} \left(P^{k_l}(x,x)\right)^{a|b_l| + sb_l}\right) P^c(x,y) > 0$$

□

Corollary 8. *If P is aperiodic and irreducibe, then there is an integer $r > 0$ so that $P^r(x,y) > 0$ for all $x, y \in \Omega$.*

Proof. Take the maximum over all the n_0's in the last proposition. □

Definition 9. A *lazy* version of a Markov chain P is one where we use the transition probability $Q = (P + I)/2$, so that with probability $\frac{1}{2}$, the chain feels lazy and stays in the same state. The advantage of doing this is that lazy chains are always aperiodic.

Proposition 10. *Let $\tau_x^+ := \min\{t \geq 1 : X_t = x\}$ be the first return time to x. Let $\mathbf{P}_x, \mathbf{E}_x$ refer to the probability space for the Markov chain with initial condition $X_0 = x$. Then for irreducible chains, we have that $\mathbf{E}_x\left(\tau_y^+\right) < \infty$.*

Proof. For each pair of states $(x,y) \in \Omega^2$, we know we can find an $n_{x,y}$ so that $\epsilon_{x,y} := P^{n_{x,y}}(x,y) > 0$. Let $n = \max_{x,y\in\Omega^2} n_{x,y}$ and $\epsilon = \min_{x,y\in\Omega^2} \epsilon_{x,y}$. Thus, the probability of hitting state y at a time between t and $t+n$ is at least ϵ (Since if we are at state z, the the probability to hit at exactly the time $t + n_{z,y}$ is $\epsilon_{z,y} \geq \epsilon$). Hence for $k > 0$ we have:

$$\mathbf{P}_x\left\{\tau_y^+ > kn\right\} \leq (1-\epsilon)\,\mathbf{P}_x\{\tau_y^+ > (k-1)n\}$$

By repeating this inequality we get:

$$\mathbf{P}_x\left\{\tau_y^+ > kn\right\} \leq (1-\epsilon)^k$$

Now use the identity for $\mathbf{E}$ to get:

$$\begin{aligned}\mathbf{E}_x\left(\tau_y^+\right) &= \sum_{t\geq 0}\mathbf{P}_x\{\tau_y^+ > t\} \\ &\leq \sum_{k\geq 0} n\mathbf{P}_x\{\tau_y^+ > kn\} \text{ since they are decreasing events} \\ &\leq \sum_{k\geq 0} n\left(1-\epsilon\right)^k < \infty\end{aligned}$$

□

Theorem 11. *Let P be an irreducibe Markov chain. Then there exists a unique stationary distributionπso that $\pi = \pi P$.*

Proof. Choose a fixed state $z \in \Omega$ and define:

$$\begin{aligned}\bar{\pi}(y) &= \mathbf{E}_z\left(\text{number of visits to } y\text{before returning to } z\right) \\ &= \sum_{t=0}^{\infty}\mathbf{P}_z\left\{X_t = y, \tau_z^+ > t\right\}\end{aligned}$$

Notice that $\bar{\pi}(y) \leq \mathbf{E}_z\left(\tau_z^+\right) < \infty$ by the last result, so this definition is ok. Now we verify that $\bar{\pi}$is stationary:

$$\begin{aligned}\left(\bar{\pi}P\right)(y) &= \sum_{x\in\Omega}\bar{\pi}(x)P(x,y) \\ &= \sum_{x\in\Omega}\sum_{t=0}^{\infty}\mathbf{P}_z\left\{X_t = x, \tau_z^+ > t\right\}P(x,y) \\ &= \sum_{t=0}^{\infty}\mathbf{P}_z\left\{X_{t+1} = y, \tau_z^+ > t\right\} \\ &= \sum_{t=1}^{\infty}\mathbf{P}_z\left\{X_t = y, \tau_z^+ \geq t\right\} \\ &= \sum_{t=1}^{\infty}\mathbf{P}_z\left\{X_t = y, \tau_z^+ > t\right\} + \sum_{t=1}^{\infty}\mathbf{P}_z\left\{X_t = y, \tau_z^+ = t\right\} \\ &= \bar{\pi}(y) - \mathbf{P}_z\left\{X_0 = y, \tau_z^+ > 0\right\} + \sum_{t=1}^{\infty}\mathbf{P}_z\left\{X_t = y, \tau_z^+ = t\right\} \\ &= \bar{\pi}(y) - \mathbf{P}_z\left\{X_0 = y\right\} + \mathbf{P}_z\left\{X_{\tau_z^+} = y\right\}\end{aligned}$$

The last two terms cancel out! If $y = z$ then both are 1, and if $y \neq z$ both are 0. Finally, we normalize $\bar{\pi}$ by $\sum_{x\in\Omega}\bar{\pi}(x) = \mathbf{E}_z\left\{\tau_z^+\right\} < \infty$ to get a probability distribution

$$\pi(x) = \frac{\bar{\pi}(x)}{\mathbf{E}_z(\tau_z^+)}$$

(This proof will work with τ_z^+ replaced by any stopping time τ with the property that $\mathbf{E}_z\left(\tau\right) < \infty$, and $\mathbf{P}_z\left\{X_\tau = z\right\} = 1$)

To see uniqueness, we show that the eigenspace of P of eigenvalue 1 is one dimensional. In particular, we show that solutions to $h = Ph$ must have $h(x) = cons't$, a 1 dimensional subspace. Say $h(x)$ is maximal value, so $h(x) \geq h(y)$ for all $y \in \Omega$. But then for the neighbours of x in the graph (those for which $P(x,y) > 0$, denoted here as $x \sim y$):

$$h(x) = \sum_{x\sim y}P(x,y)h(y) \leq \left(\sum_{y\in\Omega}P(x,y)\right)h(x) = h(x)$$

Since we have a sandwhich here, it must be that $h(y) = h(x)$ for all $y \sim x$. Since the graph is connected (the Markov chain is irreducible), repeating this argument until we span the whole graph gives h is constant.

Finally, by choosing the special node z to be the node we are on, and using uniqueness, we get the nice form:

$$\pi(z) = \frac{1}{\mathbf{E}_z(\tau_z^+)}$$

□

2 Introduction to Markov Chain Mixing

Definition 12. The *total variation distance* between two distributions μ and ν on a finite state space Ω is defined by:

$$|\mu - \nu|_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|$$

In other words "the maximum difference that $\mu$and ν assign to an event"

Proposition 13. *If $\mu_n \to \mu$ in the total variation sense, i.e. $|\mu_n - \mu|_{TV} \to 0$, then $\mu_n \Rightarrow \mu$i.e. weak convergence.*

Proof. For any $\mu-$continuity set E (in fact any set A at all)

$$\begin{aligned} |\mu_n(E) - \mu(E)| &\leq \max_{A \subset \Omega} |\mu_n(A) - \mu(A)| \\ &= |\mu_n - \mu|_{TV} \\ &\to 0 \end{aligned}$$

□

Here are some useful rephrasings of this distance:

Proposition 14.

$$\begin{aligned} |\mu - \nu|_{TV} &= \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| \\ &= \sum_{x \in \{\mu(x) \geq \nu(x)\}} |\mu(x) - \nu(x)| \end{aligned}$$

Proof. Let $B = \{x : \mu(x) - \nu(x) \geq 0\}$ and let $A \subset \Omega$ be any event. Then:

$$\begin{aligned} \mu(A) - \nu(A) &= (\mu(A \cap B) - \nu(A \cap B)) + (\mu(A \cap B^c) - \nu(A \cap B^c)) \\ &\leq \mu(A \cap B) - \nu(A \cap B) \\ &= (\mu(B) - \nu(B)) - (\mu(B \cap A^c) - \nu(B \cap A^c)) \\ &\leq \mu(B) - \nu(B) \end{aligned}$$

Similarly, one can see that $\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c)$. Hence we have:

$$\begin{aligned} |\mu(A) - \nu(A)| &\leq \frac{1}{2} (\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)) \\ &= \frac{1}{2} \sum_{x \in B} |\mu(x) - \nu(x)| + \frac{1}{2} \sum_{x \in B^c} |\mu(x) - \nu(x)| \\ &= \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| \end{aligned}$$

Finally, notice that the above inequalities are saturated when $A = B$, so this set is the set that achieves the maximum in the definition of TV. □

Remark 15. By the triangle inequality for real numbers, and the above proposition, it is clear we have a triangle ineqaulity for TV:

$$|\mu - \nu|_{TV} \leq |\mu - \eta|_{TV} + |\eta - \nu|_{TV}$$

Proposition 16. *Another formula for TV:*

$$\begin{aligned} |\mu - \nu|_{TV} &= \frac{1}{2} \sup \{\mathbf{E}_\mu(f) - \mathbf{E}_\nu(f) : |f| \leq 1\} \\ &= \frac{1}{2} \sup \left\{ \sum_{x \in \Omega} f(x)\mu(x) - \sum_{x \in \Omega} f(x)\nu(x) : |f| \leq 1 \right\} \end{aligned}$$

Proof. Have for $|f| \leq 1$:

$$\begin{aligned} \sum_{x \in \Omega} f(x)\mu(x) - \sum_{x \in \Omega} f(x)\nu(x) &\leq \sum_{x \in \Omega} |f(x)| \left(\mu(x) - \nu(x)\right) \\ &\leq \sum_{x \in \Omega} 1 \cdot |\mu(x) - \nu(x)| \end{aligned}$$

So by our last proposition, $|\mu - \nu|_{TV} \leq \frac{1}{2} \sup \{\mathbf{E}_\mu(f) - \mathbf{E}_\nu(f) : |f| \leq 1\}$. The other inequality follows by choosing $f = \text{sgn}\mu(x) - \nu(x)$. □

2.1 Coupling and Total Variation Distance

Definition 17. A *coupling* of two probability distributions $\mu$and ν is a pair of random variables X, Y defined on a single probability space such that the marginal distribution of X is μ and the marginal distribution for Y is ν. i.e. $\mathbf{P}\{X = x\} = \mu(x)$, $\mathbf{P}(Y = y) = \nu(y)$. Coupling can be a great tool!

Example 18. If X, Y are Bernoulli random variables (i.e. coinflips- 1 or 0 with probability $\frac{1}{2}$), then X, Y being indepenedent, or $X = 1 - Y$ or even $X = Y$ are all examples of couplings. Notice that in this last example, $\mathbf{P}(X \neq Y) = 0$. One can specify the coupling by specfying a distribution q on $\Omega \times \Omega$. e.g. if X, Y are independent, then $q(x, y) = \frac{1}{4}$ for every $x, y \in \Omega \times \Omega$ etc.

Problem 19. Let X, Y be couplings of μ,ν then:

$$|\mu - \nu|_{TV} = \inf \{\mathbf{P}(X \neq Y) : (X, Y) \text{ is a coupling of } \mu, \nu\}$$

And in fact this infimum is attained for some coupling, which we will call the optimal coupling.

Proof. For any event A we have:

$$\begin{aligned} \mu(A) - \nu(A) &= \mathbf{P}\{X \in A\} - \mathbf{P}\{Y \in A\} \\ &= \mathbf{P}\{X \in A, Y \notin A\} + \mathbf{P}\{X \in A, Y \in A\} \\ &\quad -\mathbf{P}\{X \in A, Y \in A\} - \mathbf{P}\{X \notin A, Y \in A\} \\ &\leq \mathbf{P}\{X \in A, Y \notin A\} \\ &\leq \mathbf{P}\{X \neq Y\} \end{aligned}$$

The last inequality follows since $\{X \in A, Y \notin A\} \subset \{X \neq Y\}$. From this inequality and the definition of TV it is clear that:

$$|\mu - \nu|_{TV} \leq \inf \{\mathbf{P}(X \neq Y) : (X, Y) \text{ is a coupling of } \mu, \nu\}$$

To see the reverse we will create a coupling which gives the opposite ineqaultiy. This inequality will try and make X and Y equal as often as they can be, to minimize $\mathbf{P}(X \neq Y)$. See the following image from Peres et al. (pg 51):

(Height here is probability, the $x-$axis represents Ω). Notice that for a single distribution this picture gives a way to make a random variable whose distribution is μ. *The uniform distribution by area* is the key. The probability space is everywhere under the dashed line, and the probability of any region is proportional

to the area of the region. The value of X is the $x-$coordinate of the point chosen. Now, in the case in the above figure, we have the distribution of μ and ν overplotted. If we restrict our attention to regions I and III, we recover the picture for μ and regions II and III give the picture for ν.

Let A_I, A_{II}, A_{III} be the area of regions I II and III respectivly. Notice $A_I = \sum_{x\in\Omega} \mu(x) \wedge \nu(x)$. Since $A_I + A_{III} = 1$, we know $A_{III} = 1 - A_I$. Similarly, $A_{II} = 1 - A_I$. Have then $A_{III} = A_{II}$. Consider the following distribution for the choice of two points, call them a and b, in the region under $\mu$and ν above.

With probilility A_{III} : $a = b$ is uniformly chosen in region III.

With probability $1 - A_{III}$: a is uniformly chosen in region I, and b is uniformly chosen in region II.

Notice that the marginal distribition for a is to be uniformly chosen anywhere in regions I and III, so that the $x-$coordinate of a gives back μ. Similarly, b gives the distribution for ν. Let $X = a_x$, $Y = b_x$ be random varibales, by these remarks (X, Y) is a coupling for μ,ν. This is the coupling we will use.

Finally, notice that $\mathbf{P}\{X \neq Y\} = 1 - A_{III} = A_I = A_{II}$. Notice that by the definitions here that $A_I + A_{III} = \sum_{x\in\Omega} |\mu(x) - \nu(x)|$ so since they are equal we have $A_I = A_{III} = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)|$ which is exactly $|\mu - \nu|_{TV}$. □

2.2 The Convergence Theorem

Theorem 20. *(Convergence theorem)*

If P is an irreducible aperiodic Markov chain, with stationary distribution π. Then there exists constant $\alpha \in (0, 1)$ and $C > 0$ so that (here δ_x is the dirac distribution at x)

$$\max_{x\in\Omega} \left|\delta_x P^t - \pi\right|_{TV} \to 0$$

Proof to follow later....there is an algebraic proof that we can do now that is technical and (as far as I can see) non-intuitive. There is a much nicer proof using coupling that will follow later.

Later we will see that we can improve the result to:

$$\max_{x\in\Omega} \left|\delta_x P^t - \pi\right|_{TV} \leq C\alpha^t$$

For some α,C constants.

2.3 Stadardizing Distance from Stationarity.

Definition 21. Let $d(t)$ be the distance from stationarity in the total variation sense:

$$d(t) = \max_{x\in\Omega} \left|\delta_x P^t - \pi\right|_{TV}$$

We know this is going to zero as $t \to \infty$ by the convergence theorem. The following will be handy too:

$$\bar{d}(t) = \max_{x,y\in\Omega} \left|\delta_x P^t - \delta_y P^t\right|_{TV}$$

Proposition 22. *The following inequality relates d and $\bar{d}$:*

$$d(t) \leq \bar{d}(t) \leq 2d(t)$$

Proof. $\bar{d}(t) \leq 2d(t)$ follows immediatly by the triangle inequality for TV, inserting π as the intermediate

point. To see $d(t) \leq \bar{d}(t)$, we use the identity $\pi(A) = \sum_{y\in\Omega} \pi(y)P^t(y, A)$:

$$\begin{aligned}
|\delta_x P^t - \pi|_{TV} &= \max_{A\subset\Omega} \left|P^t(x, A) - \pi(A)\right| \\
&= \max_{A\subset\Omega} \left| \left(\sum_{y\in\Omega} \pi(y)\right) P^t(x, A) - \sum_{y\in\Omega} \pi(y)P^t(y, A)\right| \\
&= \max_{A\subset\Omega} \left|\sum_{y\in\Omega} \pi(y) \left(P^t(x, A) - P^t(y, A)\right)\right| \\
&\leq \max_{A\subset\Omega} \sum_{y\in\Omega} \pi(y) \left|P^t(x, A) - P^t(y, A)\right| \\
&\leq \sum_{y\in\Omega} \pi(y) \max_{A\subset\Omega} \left|P^t(x, A) - P^t(y, A)\right| \\
&= \sum_{y\in\Omega} \pi(y) \left|\delta_x P^t - \delta_y P^t\right| \\
&\leq \max_{y\in\Omega} \left|\delta_x P^t - \delta_y P^t\right|
\end{aligned}$$

Taking max over all $x \in \Omega$ now gives the result. □

Proposition 23. *The function $\bar{d}$ is submultiplicative:*

$$\bar{d}(t + s) \leq \bar{d}(s)\bar{d}(t)$$

Proof. Fix $x, y \in \Omega$. Let X_s,Y_s be the optimal coupling of $\delta_x P^s$ and $\delta_y P^s$. Since this is the optimal coupling, we have that:

$$|\delta_x P^s - \delta_y P^s|_{TV} = \mathbf{P}\left\{X_s \neq Y_s\right\}$$

Now since, $P^{s+t} = P^s P^t$, we have that:

$$P^{s+t}(x, w) = \sum_z \mathbf{P}\{X_s = z\}P^t(z, w) = \mathbf{E}\left(P^t(X_s, w)\right)$$

Similarly, $P(y, w) == \mathbf{E}\left(P^t(Y_s, w)\right)$ so we combine these to get:

$$P^{s+t}(x, w) - P^{s+t}(y, w) = \mathbf{E}\left(P^t(X_s, w) - P^t(Y_s, w)\right)$$

Summing over $w \in \Omega$ gives us the TV:

$$\begin{aligned}
\left|P^{s+t}(x, \cdot) - P^{s+t}(y, \cdot)\right|_{TV} &= \frac{1}{2} \sum_{w\in\Omega} \left|\mathbf{E}\left(P^t(X_s, w) - P^t(Y_s, w)\right)\right| \\
&\leq \mathbf{E}\left(\frac{1}{2} \sum_{w\in\Omega} \left|P^t(X_s, w) - P^t(Y_s, w)\right|\right) \\
&= \mathbf{E}\left(\left|P^t(X_s, \cdot) - P^t(Y_s, \cdot)\right|_{TV}\right) \\
&= \mathbf{E}\left(\left|P^t(X_s, \cdot) - P^t(Y_s, \cdot)\right|_{TV} \mathtt{1}_{\{X_s\neq Y_s\}}\right) \\
&\leq \mathbf{E}\left(\bar{d}(t)\mathtt{1}_{\{X_s\neq Y_s\}}\right) \\
&= \bar{d}(t)\mathbf{P}\left\{X_s \neq Y_s\right\} \\
&= \bar{d}(t)\bar{d}(s)
\end{aligned}$$

Maximizing over x, y now completes the proof. □

Corollary 24. *For non-negative integers c,*

$$d(ct) \leq \bar{d}(ct) = \bar{d}(t + t + \ldots + t) \leq \bar{d}(t)^c \leq 2^c d(t)^c$$

This shows that if $d(t) \to 0$, then it must be that $d(t) \leq C\alpha^t$ for some C, α. This is why the technical proof of the convergence theorem is not nessasary, the nice proof is enough!

Exercise 25. Show that for μ, ν distributions on a Markov chain P that:

$$|\mu P - \nu P|_{TV} \leq |\mu - \nu|_{TV}$$

Proof. Let X_0, Y_0 be the optimal coupling for μ, ν. Use the random mapping representation and a $Z \sim Unif[0,1]$ to define:

$$\begin{aligned} X_1 &= f(X_0, Z) \\ Y_1 &= f(Y_0, Z) \end{aligned}$$

By the definition of the random mapping represnetation X_1, Y_1 are distributed like μP and νP repsectivly and so are a coupling of these distributions. Moreover, if $X_0 = Y_0$, then $X_1 = Y_1$ too i.e. $\{X_0 = Y_0\} \subset \{X_1 = Y_1\}$. Hence $\{X_0 \neq Y_0\} \supset \{X_1 \neq Y_1\}$

$$\begin{aligned} |\mu - \nu|_{TV} &= \mathbf{P}(X_0 \neq Y_0) \\ &\geq \mathbf{P}(X_1 \neq Y_1) \\ &\geq \inf\{\mathbf{P}(X \neq Y) : X, Y \text{ a coupling of } \mu P, \nu P\} \\ &= |\mu P - \nu P|_{TV} \end{aligned}$$

This shows that $d(s)$, and $\bar{d}(s)$ are strictly non-increasing function of s. □

2.4 Mixing Time

Definition 26. For a given Markov chain, let $t_{mix}(\epsilon) = \min\{t : d(t) \leq \epsilon\}$ and, for shorthand, $t_{mix} = t_{mix}(\frac{1}{4})$. We know these exists since $d(t) \to 0$ as $t \to \infty$ by the convergence theorem.

Notice that, by our above work:

$$d(ct_{mix}(\epsilon)) \leq (2d(t_{mix}(\epsilon))^c \leq (2\epsilon)^c$$

So for $\epsilon = \frac{1}{4}$, $d(ct_{mix}) \leq 2^{-c}$. If we do a change of variable $t = ct_{mix}$,$c = \frac{t}{t_{mix}}$ then this is saying:

$$d(t) \leq \left(2^{\frac{1}{t_{mix}}}\right)^{-t}$$

3 Coupling

In the last chapter we coupled random variables to get some information about them, here we will couple entire Markov chains to get more information.

Example 27. Consider a simple random walk on the state space $\{0, 1, \ldots n\}$. At each step the walker moves to the left (minus 1) or two the right (plus 1) with probability $\frac{1}{2}$, and at the endpoint states 0 and n, it will stay put with probability $\frac{1}{2}$ and move in the middle with probability $\frac{1}{2}$.

One way to couple to Markov chains $(X_0, X_1, \ldots)$and $(Y_0, Y_1, \ldots)$ with initial conditions $X_0 = x$,$Y_0 = y$ is to select iid variables Δ_k which are ± 1 with probability $\frac{1}{2}$ each and let X_n, Y_n move up and down with the delta's (unless we try to go over an endpoint):

$$\begin{aligned} X_n &= (X_{n-1} + \Delta_n) \wedge n \vee 0 \\ Y_n &= (Y_{n-1} + \Delta_n) \wedge n \vee 0 \end{aligned}$$

Notice that with this formulation that if $X_k = Y_k$ then $X_n = Y_n$ for all $n \geq k$; once they meet they will be equal forevermore.

Here we can get some cute information. Notice that if $x < y$ then $X_k < Y_k$ for all n, so if $X_k = n$ then it must be that $Y_k = n$ too. Hence:

$$P^t(x,n) = \mathbf{P}\{X_t = n\} \leq \mathbf{P}\{Y_t = n\} = P^t(y,n)$$

This kind of intuitive thing is often made clear with coupling, e.g. percolation.

Definition 28. A coupling of Markov chains with transition P is a sequence of random variables $(X_n, Y_n)_{n=1}^{\infty}$ so that both $X_n$and Y_n are Markov chains with transition P. Any such coupling can be modified so that it has the property $X_k = Y_k \Rightarrow X_n = Y_n \forall n \geq k$, by simply leaving X alone and demanding equality of Y for such $n \geq k$ (since both X, Y have the same transition probabilities, this leaves the distribution of Y unharmed!) We will call such a coupling a *glued coupling. An* example of this is the random mapping representation. We will often use $\mathbf{P}_{x,y}$ to denote the probability space for such a coupling with $X_0 = x$,$Y_0 = y$.

Theorem 29. *Let (X_k, Y_k) be a glued coupling (so they stick together if they ever meet) with $X_0 = x, Y_0 = y$. Let τ_{couple} be the first time the chains meet:*

$$\tau_{couple} = \min\{t : X_t = Y_t\}$$

Then:

$$\left|P^t(x,\cdot) - P^t(y,\cdot)\right|_{TV} \leq \mathbf{P}_{x,y}\{\tau_{couple} > t\}$$

Proof. For any fixed t, $X_t, Y_t$are a coupling for the distributions $P^t(x,\cdot), P^t(y,\cdot)$. So by our last section:

$$\begin{aligned} \left|P^t(x,\cdot) - P^t(y,\cdot)\right|_{TV} &\leq \mathbf{P}_{x,y}\{X_t \neq Y_t\} \\ &\leq \mathbf{P}_{x,y}\{\tau_{couple} > t\} \end{aligned}$$

Where the last inequality follows since $\{X_t \neq Y_t\} \subset \{\tau_{couple} > t\}$ by the definition of τ. □

Corollary 30. *The above shows:*

$$d(t) \leq \bar{d}(t) \leq \max_{x,y \in \Omega} \mathbf{P}_{x,y}\{\tau_{couple} > t\}$$

Example 31. Consider the lazy random walk on the $n-$cycle $\mathbb{Z}_n$. A nice coupling to consider is one in which we first flip a coin, and if heads then X_t moves, and if tails Y_t moves (that way they both fullfill the laziness criteria). Another coinflip will be nessasary to see which direction the active walker moves. We glue them after they first collide. This method ensures that the two cannot "walk over" each other.

In this way, the distance between the two particles does a simple random walk on $\{0, 1, \ldots \frac{n}{2}\}$, where 0 is absorbing and the end state $\frac{n}{2}$ moves to the left with probability 1. This can be thought of as a projection of the random walk on $\{0, 1, \ldots n\}$ with absorbing endpoints at 0 and at n. τ is the time until we are absorbed. The theory of martingales here (or linear systems) shows us that $\mathbf{E}_{x,y}(\tau) = k(n-k)$ where k is the starting distance. Hence:

$$d(t) \leq \max_{x,y \in \mathbb{Z}_{\ltimes}} \mathbf{P}_{x,y}\{\tau > t\} \leq \max_{x,y \in \mathbb{Z}_{\ltimes}} \frac{\mathbf{E}_{x,y}\{\tau\}}{t} = \frac{(n/2)^2}{t} = \frac{n^2}{4t}$$

Hence for this system, this calculation shows us that, $t_{mix} \leq n^2$. Some calculations later on will give us a lower bound, $t_{mix} \geq Cn^2$. A similar type of calculation will work for the torus.

Example 32. Random Walk on the Hypercube.

Consider two lazy random walks on the hypercube $\{0,1\}^n$. A convenient way to represent this walk is to pick a bit at random, and *refresh* the bit with either 0 or 1 chosen at random. (Half the time, the bit will agree with what was there before (laziness) and half the time it will change). Our coupling will refresh the bits from X and Y with the same bit. In this way, two bits agree if they have ever been refreshed. This is exactly the classic coupon collecting problem! By summing many geometric random variables we can understand this distribution. For example, its easy to see that $\mathbf{E}(\tau) = n\sum_{k=1}^{n} \frac{1}{k} \approx n \log n$

Here is a slick deviation estimate Let $A_i$be the event that the $i-$th coupon flavor does not appear in the first $n\log n+cn$ coupon draws Then:

$$\mathbf{P}\left(\tau > n\log n + cn\right) = \mathbf{P}\left(\bigcup A_i\right) \leq \sum_{i=1}^{n} \mathbf{P}\left(A_i\right)$$

But each coupon draw has a probaility of $(1-\frac{1}{n})$ to not give the $i-$th flavour, so since there are $n\log n+cn$ such draws, we have $\mathbf{P}\left(A_i\right) = \left(1-\frac{1}{n}\right)^{n\log n+cn}$, so we have:

$$\begin{aligned}
\mathbf{P}\left(\tau > n\log n + cn\right) &\leq \sum_{i=1}^{n}\left(1-\frac{1}{n}\right)^{n\log n+cn} \\
&= n\left(1-\frac{1}{n}\right)^{n\log n+cn} \\
&\leq n\exp\left(\frac{-1}{n}\right)^{n\log n+cn} \\
&= n\exp\left(-\log n\right)\exp\left(-c\right) \\
&= \exp(-c)
\end{aligned}$$

Since this coupling is the same as the coupon collecting problem:

$$d(n\log n + cn) \leq \mathbf{P}\left\{\tau > n\log n + cn\right\} \leq \exp(-c)$$

Hence:

$$\begin{aligned}
t_{mix}(\epsilon) &\leq n\log n + \log\left(\frac{1}{\epsilon}\right)n \\
t_{mix} &= O(n\log n) \text{ as } n\to\infty
\end{aligned}$$

Now that we have done a little bit of legwork and set up some notation, here is a nice proof of the convergence theorem using coupling.

Theorem 33. *(Convergence theorem) If P is an irreducible aperiodic Markov chain, with stationary distribution π. Then there exists constant $\alpha\in(0,1)$ and $C>0$ so that (here δ_x is the dirac distribution at x) Then:*

$$\max_{x\in\Omega}\left|\delta_x P^t - \pi\right|_{TV} \to 0$$

i.e. $d(t)\to 0$. (This proof of the convergence theorem is an exercise in the book)

Proof. We will show that $\bar{d}(t)\to 0$ instead. Since $d(t)\leq\bar{d}(t)$, as we've shown independently, this will be sufficient. Recall $\bar{d}(t) = \max_{x,y\in\Omega}\left|\delta_x P^t - \delta_y P^t\right|_{TV}$. Let $x,y\in\Omega$ be arbitary, and let X_t, Y_t be a coupling for the markov chain starting at x,y respectivly. Assume also that X_t, Y_t have the gluing property; just make it so that once they meet they are equal forevermore as we have previously described. Let $\tau = \inf\{t : X_t = Y_t\}$ be the random variable which measures the time when they meet. Notice $\{X_t = Y_t\} = \{t\geq\tau\}$ i.e. $\{X_t\neq Y_t\} = \{\tau > t\}$. Then:

$$\begin{aligned}
\left|\delta_x P^t - \delta_y P^t\right|_{TV} &\leq \mathbf{P}\left\{X_t\neq Y_t\right\} \\
&= \mathbf{P}\left\{\tau > t\right\}
\end{aligned}$$

So to prove $\left|\delta_x P^t - \delta_y P^t\right|_{TV}\to 0$, it suffices to show that $\mathbf{P}\left\{\tau>t\right\}\to 0$. If we take X_t, Y_t to be the *indepedent coupling* (i.e. that is the choice of $X_{t+1}|X_t$ is independent of $Y_{t+1}|Y_t$ at every time step, until they meet at which point they glue). Now, before time τ, think of (X_t, Y_t) as a Markov chain on $\Omega\times\Omega$, we see that all the states are *recurrent* (meaning that they will all be visited infitetly often) since it is irreducible with stationary distribution $\pi\times\pi$(see next lemma). In particular the states (z,z) for $z\in\Omega$ will eventually be visited almost surely. When this happens, $X_t = Y_t = z$ and we will have gluing! In other words, $\tau<\infty$ almost surely, so $\mathbf{P}\left\{\tau>t\right\}\to 0$ as $t\to\infty$. □

Lemma 34. *If a Markov chain is irreducible with stationary distribution π then all the states are reccurant.*

Proof. Since the Markov chain is irreducible, it is easy to see that all the states are either recurrant or all the states are transient. Supposing by contradiction that all the states are transient, we have $\sum_t P^t(x,y) < \infty$ for every $x, y \in \Omega$. In particular, $P^t(x,y) \to 0$ for every $x, y \in \Omega$. Now write, for any t, $\pi(z) = \sum_x \pi(x)P^t(x,z)$. Since there are only finetly many states and $P^t(x,y) \to 0$, have then $\pi(z) = \sum_x \pi(x)P^t(x,z) \to 0$. Hence $\pi(z) = 0$ for every $z \in \Omega$ which is a contradiction. □

3.1 Stationary Times

Example 35. (Top-to-random shuffle) Here is one way to shuffle a deck of cards (shuffling a deck of cards is a Markov process on the symetric group....but its easier to think about as cards) Take the top card and put it somewhere, chosen uniformly, in the middle of the deck. Repeat! One can see that by the time the card which is originally on the bottom reaches the top the deck is now uniformly distributed (as an element of the symetric group, I will call this "uniformly shuffled"). At any given step, the cards underneath this card all uniformly distributed on the symetric group (see this by induction, each additional card is equally likely to be placed in any of the slots under), so by the time this card reaches the top, the whole deck is uniformly shuffled.

The amount of time it takes for this card to reach the top is a random variable, lets call it τ_{top}, and we know that the markov chain (the deck of cards) will reach its equilibrium (the uniform distribution) at time τ_{top}. Also of note is that τ_{top} and $X_{\tau_{top}}$ are independent! We will see some examples later where this is not the case. We will see here how we can use τ_{top} to bound t_{mix}.

Definition 36. For a Markov chain X_t, we know what a stopping time is, a random variable τ such that τ does not "look into the future": $\{\tau \leq n\} \subset \sigma(X_1, \ldots, X_n)$. A slightly more general notion is a *randomized stopping time.* This is a random variable τ that might not be $X_1, X_2, \ldots, X_n$ measurable, but instead depends on the randomness that generated the $X's$. To be more precise, let $Z_1, Z_2, \ldots Z_n$ be random variabels that determine $X_1, X_2, \ldots X_n$ throught the random mapping representation $X_n = f(X_{n-1}, Z_n)$ τ is called a *randomized stopping time* if $\{\tau \leq n\} \subset \sigma(Z_1, \ldots, Z_n)$. Notice $\sigma(X_1, \ldots, X_n) \subset \sigma(Z_1, \ldots Z_n)$ because of the function f in the random mapping representation, so every randomized stopping time is a stopping time too; they are more general.

Notice we can build in some extra randomness into our randomized stopping times that the Markov chains don't even see, because the function f can ignore part of the information in Z if we want. (i.e. something like "Choose τ to be uniformly distributed between 1 and 6 is a randomized stopping time, because we can build that into the random variable Z_1) We will see a few examples of this below.

Example 37. If we consider the lazy random walk on the hypercube, we know already that a nice way to represent it is to choose a bit at random, and then refresh the bit. Let $\tau_{refresh}$ be the first time that each bit has been refreshed at least once (e.g. the coupon collecting problem). Notice one cannot determine $\tau_{refresh}$ just by looking at the states of the markov chain $X_1, X_2, \ldots X_n$ because sometimes during a refresh there is *no change* in the refreshed bit (it goes from a 0 to a 0 say). $\tau_{refresh}$ instead depends on the history of the randomness that drives the chain $Z_1, Z_2, \ldots Z_n$. In this example $Z \in \{1, 2, \ldots n\} \times \{0, 1\}$ chosen uniformly tells you which bit to refresh and what to refresh it to.

This is useful here because, like the top-to-bottom shuffling example, we know that at $\tau_{refresh}$ the Markov chain has reached its stationary distribution. $\tau_{refresh}$and τ_{top} are both examples of *stationary times* which we define below:

Definition 38. A *stationary time* τ is a random stopping time with the property that at time τ, the Markov chain has reached its stationary distribution:

$$\mathbf{P}_x\{X_\tau = y\} = \pi(y)$$

Example 39. Let X_t be the random walk on the $n-$cycle, $\mathbb{Z}_n$ and let σ be the first time that every vertex in the cycle has been visited. One can show that X_σ is uniformly distributed over the $n-1$ vertices of the n-cycle which are not the starting vertex (I wrote this up in a seperate document, its a bit long but should

be clear). Hence if we define a new stopping time τ by:

$$\tau = \begin{cases} 0 & \text{with probability } \frac{1}{n} \\ \sigma & \text{with probability } \frac{n-1}{n} \end{cases}$$

Then $X_\tau$will be uniformly distributed on the n-cycle. An important thing of note in this example is that, although X_τ is uniformly distributed, it is not independent of τ for example, if $\tau = 0$, then $X_\tau = x_0$a.s. is the starting node.

Example 40. Here is another example that has the property that τ and $X_\tau$are not independent, even though X_τ has the stationary distribution. Let ξ be a random variable chosen with the stationary distribution π, and let τ_ξ be the hitting time for X to hit ξ. Then τ is a randomized stopping time, and $X_{\tau_\xi} = \xi$ has the distribution π, so again this is a stationary time. Like the last example though, X_{τ_ξ} is not independent of τ. For example, if $\tau = 0$, then $X_\tau = x_0$a.s. is the starting node.

Definition 41. A *strong stationary time* for a Markov chain X_t with stationary distribution π is a randomized stopping time τ, possibly depending on the starting position x, is a stationary time so that τ and $X_\tau$are independent. i.e:

$$\mathbf{P}_x \{\tau = t, X_\tau = y\} = \mathbf{P}_x \{\tau = t\} \pi(y)$$

Example 42. The top to random shuffle, τ_{top} and the refreshing time on the hypercube $\tau_{refresh}$ are like this, but the example on the n-cycle and the cheapo $\xi \sim \pi$ examle are not!

Strong stationary times can be used to bound t_{mix}, as we shall see. First a lemma!

Lemma 43. *Let X_t be an irreducible Markov chain with stationary distribution π. If τ is a strong stationary time for X_t, then for all $t \geq 0$:*

$$\mathbf{P}_x \{\tau \leq t, X_t = y\} = \mathbf{P}_x \{\tau \leq t\} \pi(y)$$

Proof. Let $Z_1, Z_2, \ldots$be the randomness in the Markov chain through the random mapping representation. Fix some $s \leq t$ and write:

$$\mathbf{P}_x \{\tau = s, X_t = y\} = \sum_{z \in \Omega} \mathbf{P}_x \{X_t = y | \tau = s, X_s = z\} \mathbf{P}_x \{\tau = s, X_s = z\}$$

Now, $\{\tau = s, X_s = z\} \in \sigma(Z_1, \ldots Z_n)$, while $\{X_t = y\} = \{f(X_s, (Z_{s+1}, Z_{s+2}, \ldots Z_t) = y\}$ depends only on X_s and $Z_{s+1}, \ldots Z_n$. Since the $Z's$ are independent then:

$$\begin{aligned} \mathbf{P}_x \{X_t = y | \tau = s, X_s = z\} &= \mathbf{P}_x \{f(X_s, (Z_{s+1}, Z_{s+2}, \ldots Z_t) = y | \tau = s, X_s = z\} \\ &= \mathbf{P}_x \{f(z, (Z_{s+1}, Z_{s+2}, \ldots Z_t) = y | \tau = s, X_s = z\} \\ &= \mathbf{P}_x \{f(z, (Z_{s+1}, Z_{s+2}, \ldots Z_t) = y\} \\ &= P^{t-s}(z, y) \end{aligned}$$

Using that along with $\mathbf{P}_x \{\tau = t, X_\tau = z\} = \mathbf{P}_x \{\tau = t\} \pi(z)$, (definition of a strong stationary time) we have:

$$\begin{aligned} \mathbf{P}_x \{\tau = s, X_t = y\} &= \sum_{z \in \Omega} \mathbf{P}_x \{X_t = y | \tau = s, X_s = z\} \mathbf{P}_x \{\tau = s, X_s = z\} \\ &= \left(\sum_{z \in \Omega} P^{t-s}(z, y) \pi(z) \right) \mathbf{P}_x \{\tau = t\} \\ &= \pi(y) \mathbf{P}_x \{\tau = t\} \end{aligned}$$

The last equality holds since π is the stationary distribution. Summing over s now gives the desired result. □

Proposition 44. *If τ is a strong stationary time, then*

$$d(t) = \max_{x\in\Omega} \left|P^t(x,\cdot) - \pi\right|_{TV} \leq \max_{x\in\Omega} \mathbf{P}_x\left\{\tau > t\right\}$$

We break the proof of this proposition up into two lemmas. We will use the notation:

$$s_x(t) = \max_{y\in\Omega}\left[1 - \frac{P^t(x,y)}{\pi(y)}\right]$$

And we will show that $d_x(t) \leq s_x(t)$ and $s_x(t) \leq \mathbf{P}_x\left\{\tau > t\right\}$, combining these inequalites and maximizing over x will prove the proposition. This is called the *seperation distance* between π and $P^t(x,\cdot)$.

Lemma 45. *If τ is a strong stationary time, then:*

$$s_x(t) \leq \mathbf{P}_x\left\{\tau > t\right\}$$

Proof. Fix $x \in \Omega$. Observe that for any $y \in \Omega$ that:

$$\begin{aligned}
1 - \frac{P^t(x,y)}{\pi(y)} &= 1 - \frac{\mathbf{P}_x\left\{X_t = y\right\}}{\pi(y)} \\
&\leq 1 - \frac{\mathbf{P}_x\left\{X_t = y, \tau \leq t\right\}}{\pi(y)} \\
&= 1 - \frac{\mathbf{P}\left\{\tau \leq t\right\}\pi(y)}{\pi(y)} \text{ by previous lemma} \\
&= 1 - \mathbf{P}\left\{\tau \leq t\right\} \\
&= \mathbf{P}\left\{\tau > t\right\}
\end{aligned}$$

Maximizing over s gives the result. □

Remark 46. The inequality above is an equality if and only if there is a node y so that $\mathbf{P}_x\left\{X_t = y\right\} = \mathbf{P}_x\left\{X_t = y, \tau \leq t\right\}$. This happens if and only if $\left\{X_t = y\right\} \subset \left\{\tau \leq t\right\}$. Such a node y is called a *halting state*, it is a node for which when we get there we know that the markov chain has reached its equilibrium distribution. An example is if we start with the state $\{x_1, x_2, \ldots x_n\}$ of the hypercube, the state $\{\overline{x_1}, \overline{x_2}, \ldots \overline{x_n}\}$ is a halting state, because we can only reach it after every bit has been refreshed. If there is one halting state starting at x, then we will have equality $s_x(t) = \mathbf{P}_x\left\{\tau > t\right\}$.

Lemma 47. *The seperation distance bounds the TV distance:*

$$\left|P^t(x,\cdot) - \pi\right|_{TV} \leq s_x(t)$$

Proof. Have:

$$\begin{aligned}
\left|P^t(x,\cdot) - \pi\right|_{TV} &= \sum_{y:P^t(x,y)<\pi(y)} \pi(y) - P^t(x,y) \\
&= \sum_{y:P^t(x,y)<\pi(y)} \pi(y)\left(1 - \frac{P^t(x,y)}{\pi(y)}\right) \\
&\leq \sum_{y:P^t(x,y)<\pi(y)} \pi(y) s_x(t) = s_x(t)
\end{aligned}$$

□

Example 48. Conisder the simple random walk on the folowing graph. G be the graph which is obtained by taking two copies of the complete graph on n vertices and identifying one special vertex in each of them, which we count as the same vertex, so that the Markov chain may pass from one of the copies to the other through this special vertex, call it v^*. Notice $\deg v^* = 2n-2$, and the degree of all the other vertices is

$n-1$. Let G' be G where we make the chain lazy by adding a single loop to v^* and n loops to each other vertex, so that G' is regular of degree $2n-1$ and has the uniform distribution as its stationary distribution.

Let $\tau = \tau_{v^\star} + 1$ be one step further than the first time we hit $v^\star$. Since $v^\star$has exactly one edge to every vertex on the graph, X_τ is uniformly distributed on $G^{'}$, and morevoer indepenD of τ. Hence τ is a strong stationary time. Notice that no matter where we are in the graph the probability of moving to $v^\star$ in the next step is $\frac{1}{2n-1}$, so the $\tau_{v^\star}$is geometrically distributed with $p = \frac{1}{2n-1}$, with $\mathbf{E}(\tau_{v^\star}) = 2n-1$, hence $\mathbf{E}(\tau) = 2n$ and by Markov inequality, $\mathbf{P}\{\tau > t\} \leq \frac{2n}{t}$. So our result shows that:

$$t_{mix} \leq 8n$$

Later on we will show that t_{mix} is precisly order n for this problem.

There a few more examples in the notes.