Markov Chains

Sabrina Sixta

September 23, 2020

1 Foundations of Markov Chains

1.1 Definition

We first introduce Markov Chains in the most basic format, which comprises discrete time steps and state spaces.

Definition 1.1 (Markov Chain)**.** A Markov Chain is a sequence of random variables equipped with the following attributes:

- A state space S,
- An initial probability $\{v_i\}_{i\in S}$ where $v_i = P(X_0 = i)$,
- A transition probability $\{p_{ij}\}_{i,j\in S}$ where $p_{ij} = P(X_{n+1} = i|X_n = i)$.

The Markov chain existence theorem states that given the above three attributes a sequence of random variables can be generated. Since the p_{ij} is not a function of n, a Markov chain is time-homogeneous. The Markov property refers to the fact that $P(X_{n+1} = i| \cap_{l\leq n} \{X_l = i_l\}) = P(X_{n+1} = i|X_n = i_n)$.

Theorem 1.1 (Chapman-Kolmogomorov equation)**.**

$$p_{ij}^{(n+m)} = \sum_{k\in S} p_{ik}^{(n)} p_{kj}^{(m)}$$

Proof.

$$\begin{aligned}
p_{ij}^{(n+m)} &= P(X_{n+m} = j|X_0 = i) \\
&= \sum_{k\in S} P(X_{n+m} = j|X_n = k, X_0 = i)P(X_n = k|X_0 = i) \\
&\qquad\qquad\qquad \text{By the Law of total probability} \\
&= \sum_{k\in S} P(X_{n+m} = j|X_n = k)P(X_n = k|X_0 = i) \qquad \text{By the Markov property} \\
&= \sum_{k\in S} P(X_m = j|X_0 = k)P(X_n = k|X_0 = i) \qquad \text{Because of time invariance} \\
&= \sum_{k\in S} p_{ik}^{(n)} p_{kj}^{(m)}
\end{aligned}$$

$\square$

Corollary (Chapman-Kolmogomorov inequality)**.**

$$p_{ij}^{(n+m)} \geq p_{ik}^{(n)} p_{kj}^{(m)}$$

1.2 Transience and recurrence

The following are equivalent notations

- $p_{ij}^{(n)} = P_i(X_n = j) = P(X_n = j|X_0 = i)$
- $E_i[X_n = j] = E[X_n = j|X_0 = i]$
- $\tau_i = \inf\{n : X_n = i\}$
- $f_{ij}^{(n)} = P_i(\{X_n = j\} \bigcap_{l<n} \{X_l \neq j\}) = P_i(\tau_i = n)$
- $f_{ij} = P_i(\exists n : X_n = j) = \sum_{n \geq 1} f_{ij}^{(n)} = P_i(\tau_i < \infty) = p_{ij} + \sum_{k \in S, k \neq j} p_{ik} f_{kj}$

Definition 1.2 (Transience and recurrence)**.** State i is transient if $f_{ii} < 1$ and is recurrent if $f_{ii} = 1$

Theorem 1.2. *State i is recurrent: $f_{ii} = 1 \iff P_i(X_n = i \;\; i.o.) = 1 \iff \sum_{n \geq 1} p_{ii}^{(n)} = \infty \iff P_i(\tau_i < \infty) = 1$*

Proof. First $\iff$

$$\begin{aligned} P_i(X_n = i \;\; i.o.) &= P_i(\lim_{k \to \infty} |n : X_n = i| > k) \\ &= \lim_{k \to \infty} P_i(|n : X_n = i| > k) && \text{By continuity of probabilities} \\ &= \lim_{k \to \infty} (f_{ii})^k = 1 \iff f_{ii} = 1 \end{aligned}$$

Second $\iff$ (Since the events are not independent the Borel-Cantelli lemma can only be used in 1 direction although I suspect the time invariance could be used to prove the converse.)

$$\begin{aligned} \sum_{n \geq 1} p_{ii}^{(n)} &= \sum_{n \geq 1} P_i(X_n = i) \\ &= \sum_{n \geq 1} E_i[I(X_n = i)] \\ &= E_i[\sum_{n \geq 1} I(X_n = i)] && \text{By MCT} \\ &= E_i[|n : X_n = i|] \\ &\geq E_i[|n : X_n = i||X_n = i \;\; i.o.]P(X_n = i \;\; i.o.) \\ &\geq \infty * 1 = \infty \end{aligned}$$

Third $\Longleftrightarrow$

$$\begin{aligned} P_i(\tau_i < \infty) &= \lim_{k\to\infty} P_i(\tau_i \leq k) \\ &= \lim_{k\to\infty} P_i(\bigcup_{l=1}^{k} \tau_i = l) && \text{By continuity of probabilities} \\ &= \lim_{k\to\infty} \sum_{l=1}^{k} P_i(\tau_i = l) && \text{Since events are disjoint} \\ &= \lim_{k\to\infty} \sum_{l=1}^{k} f_{ii}^{(n)} = f_{ii} \end{aligned}$$

□

Note that the theorem refers only the probability of a Markov chain returning to its initial state. This is because the theorem is not true for a Markov chain transitioning from one state to another.

Definition 1.3 (State communication)**.** States i and j (denoted ij)communicate if $f_{ij} > 0$ and $f_{ji} > 0$. In other words, if there exists an n, m such that $p_{ij}^{n} > 0$ and $p_{ji}^{m} > 0$ then states i, j communicate.

The following are attributes that can be associated with a Markov chain .

Definition 1.4 (Irreducible)**.** A Markov chain is irreducible if all states communicate.

Definition 1.5 (Decomposable)**.** A Markov chain is decomposable if there exists a partition $S_1, S_2, \ldots$ such that $f_{ij} = 0$ if state i and j belong to different subsets.

If a Markov chain is decomposable, then it is not irreducible, but the converse is not true.

A state can be classified as recurrent, but a Markov chain can also be referred to as recurrent.

Definition 1.6. A Markov chain is recurrent if it is irreducible and for all $i, j \in S$, $f_{ij} = 1$.

Theorem 1.3. *If a Markov chain is recurrent then the following properties hold:*

1. *There exists $k \in S$ such that $f_{kk} = 1$*
2. *For all $i, j \in S$, $f_{ij} = 1$*
3. *There exists $i, j \in S$ such that $\sum_{n\geq 1} p_{ij}^{(n)} = \infty$*
4. *For all $i, j \in S$, $\sum_{n\geq 1} p_{ij}^{(n)} = \infty$*

Note the subtleties in this theorem. Bullet (1) states that there exists a state $k \in S$ such that $f_{kk} = 1$. Why not state instead that there exists states $i, j \in S$ such that $f_{ij} = 1$? Because this is not equivalent. Think of the random with $p > 1/2$, a non recurrent Markov chain . If $j > i$, $f_{ij} = 1$, but $f_{ji} < 1$ and $f_{ii} < 1$.

What about the bullet (3)? If bullet (1) is referring to a single state k why can bullet (3) be more general and refer to any state i, j? Again think of the random walk where $p > 1/2$ and

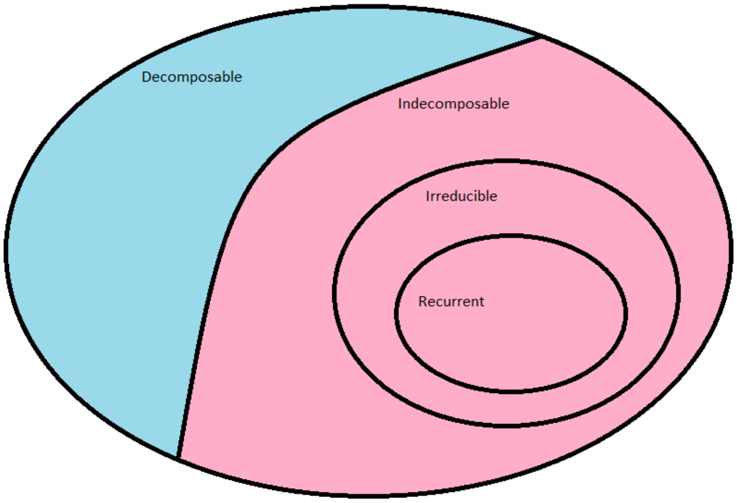

Figure 1:

$j > i$. So we know that $f_{ij} = 1$ even though the Markov chain is not recurrent, but assuming $j - i$ is even,

$$\sum_{n\geq 1} p_{ij}^{(n)} = \sum_{n\geq 1} \binom{2n}{n + (j-i)/2} p^{n+(j-i)/2}(1-p)^{n-(j-i)/2} \leq \left(\frac{p}{1-p}\right)^{(j-i)/2} \sum_{n\geq 1} (4p(1-p))^n < \infty$$

This shows that even though $f_{ij} = 1$, $\sum_{n\geq 1} p_{ij}^{(n)} < \infty$.

Proposition 1.4. *If state i and j communicate, then state i is recurrent if and only if state j is recurrent.*

Proof. Suppose state i is recurrent then $\sum_{n\geq 1} p_{ii}^{(n)} = \infty$ and there exists an s, t such that $p_{ji}^{(s)} > 0, p_{ij}^{(t)} > 0$ and so by the Kolmogomorov-Chapman inequality,

$$\sum_{n\geq 1} p_{jj}^{(n)} \geq \sum_{n\geq s+t} p_{ji}^{(s)} p_{ii}^{(n)} p_{ij}^{(t)} = \infty$$

This implies that state j is also recurrent. □

This tells us that either all states in an irreducible Markov chain are recurrent (called a recurrent Markov chain as above) or transient.

Some observations about Markov chain :

1. If $j \to i$ then there exists a positive probability of reaching state i from j without passing through j first.
2. If $j \to i$ and $f_{jj} = 1$ then $f_{ij} = 1$

Definition 1.7 (Mean recurrence time)**.** The mean recurrence time of a state i is $m_i = E_i[\tau_i]$

First note that $\inf\{\emptyset\} = \infty$. This means that all transient states have infinite mean recurrence time. Recurrent states that have infinite mean recurrence time are called null recurrent and recurrent states that have finite mean recurrence time are called positive recurrent. If a state space is finite and the Markov chain is irreducible then all recurrent states are positive recurrent.

Mean recurrence is closely linked to the stationary distribution of irreducible Markov chains. This will be shown in the next section, but for now a preliminary lemma is introduced.

Lemma 1.5. *Let* $G_n(i,j) = E_i[l : 1 \leq l \leq n \cap X_l = j] = \sum_{l=1}^{n} p_{ij}^{(l)}$*. For a recurrent Markov chain*

$$\lim_{n\to\infty} \frac{G_n(i,j)}{n} = \frac{1}{m_j}$$

Proof. Let T_j^r be the time of the rth hit of state j then,

$$T_j^r = T_j^1 + (T_j^2 - T_j^1) + \ldots + (T_j^r - T_j^{r-1})$$

and so for $s \geq 1$, $E_j[T_j^s - T_j^{s-1}] = E_j[\tau_j]$ which means that $E_i[T_j^r] = E_i[T_j^1] + rm_j$ and so assuming $E_i[T_j^1] < \infty$, by the strong law of large numbers $\lim_{r\to\infty} \frac{T_j^r}{r} = m_j$ with probability 1.

Let r_n represent the number of hits of state j by time n. By definition since r_n represents the maximum number of hits by time n which must have occurred before time n so, $T^{r_n} \leq n \leq T^{r_n+1}$ and,

$$\frac{T^{r_n}}{r_n} \leq \frac{n}{r_n} \leq \frac{T^{r_n+1}}{r_n}$$

Further, since state j is recurrent $\lim_{n\to\infty} r_n = \infty$ with probability 1. This means that $\lim_{r_n\to\infty} \frac{T_j^{r_n}}{r_n} = m_j$ with probability 1 and so by sandwich, $\lim_{r_n\to\infty} \frac{n}{r_n} = m_j$ or $\lim_{r_n\to\infty} \frac{r_n}{n} = \frac{1}{m_j}$ with probability 1. Lastly, since $\frac{r_n}{n} \leq 1$ by the bounded convergence theorem,

$$\lim_{n\to\infty} \frac{G_n(i,j)}{n} = \lim_{n\to\infty} \frac{E_i[r_n]}{n} = E_i\left[\lim_{n\to\infty} \frac{r_n}{n}\right] = \frac{1}{m_j}$$

□

Proposition 1.6. *If state i and j communicate, then state i is positive recurrent if and only if state j is positive recurrent.*

Proof. If state i is positive recurrent then, $\frac{\sum_{l=1}^{n} p_{ij}^{(l)}}{n} > 0$. By communication, there exists an s, t such that $p_{ji}^{(s)} > 0, p_{ij}^{(t)} > 0$ and so by the Kolmogomorov-Chapman inequality,

$$\frac{\sum_{n\geq 1} p_{jj}^{(n)}}{n} \geq \frac{\sum_{n\geq s+t} p_{ji}^{(s)} p_{ii}^{(n)} p_{ij}^{(t)}}{n} > 0$$

□

An irreducible Markov chain are either all transient, null recurrent or positive recurrent. An irreducible Markov chain on a finite space is always positive recurrent.

1.3 Period of a state

Definition 1.8 (Period of a state)**.** A state i is of period d if the greatest common denominator of $\{n : p_{ii}^{(n)} > 0\}$ is d.

An alternative definition of periodicity for a Markov chain is the following:

Definition 1.9 (Period of a state alternative definition)**.** A Markov chain is of period d if there exists a partition of $S = \dot{\bigcup}_{r\geq 1} S_r$ such that if $i \in S_r$ then $P_i(S_{r+1}) = 1$.

Theorem 1.7. *If state i and j communicate then state i and j have the same period.*

Proof. Since $f_{ij} > 0$ and $f_{ji} > 0$, there exists n, m such that $p_{ij}^{(n)} > 0$ and $p_{ji}^{(m)} > 0$. Let d_i, d_j be the periods for state i and j, respectively. So we know that $p_{ii}^{(n+d_ik+m)} \geq p_{ij}^{(n)} p_{jj}^{(d_ik)} p_{ji}^{(m)} > 0$ for all $k \geq 0$. This implies that $n + m$ is a multiple of d_i and by symmetry, $n + m$ is a multiple of d_j. So $n + d_ik + m$ is a multiple of d_j for all k. This implies that $d_j \in CD(\{n : p_{ii}^{(n)} > 0\})$ (where CD represents the common denominators) and so $d_j \leq GCD(\{n : p_{ii}^{(n)} > 0\}) = d_i$. By symmetry $d_j \leq d_i$ and so $d_i = d_j$. □

This tells us that all states in an irreducible Markov chain all have of the same period.

Definition 1.10 (Aperiodicity)**.** A Markov chain whose states are all equal to 1 is called aperiodic.

1.4 Stationarity

Definition 1.11 (Stationarity)**.** The distribution π is a stationary distribution of a transition probability P if $\pi P = \pi$ or in other words, for all $i \in S$, $\pi_i = \sum_{k\in S} \pi_k p_{ki}$.

1.4.1 The eigenvalue connection

Solving for the stationary distribution is closely tied to the eigenvalue and eigenvectors of the transition matrix P. Firstly, π is the left eigenvector of P when the eigenvalue is 1. Let $\lambda_0, \ldots, \lambda_{n-1}$ be the eigenvalues of P. There will always be one $\lambda_0 = 1$, furthermore, $|\lambda_i| \leq 1$ for $i \in \{0, \ldots, n-1\}$.

Let $\lambda_* = \max_{i\in\{1,\ldots,n-1\}} \lambda_i$. To have a $\lambda_* < 0$ means that π is unique and $vP^k \to \pi$ (we will see what this means later). If all entries of P are strictly positive then $\lambda_* < 1$. Furthermore, $\lambda_* < 1$ if and only if P is indecomposable and aperiodic.

The eigenvalues and eigenvectors can be calculated in two ways. The first way is to solve for $|P - \lambda I| = 0$ to get the eigenvectors and then calculate v such that $P - \lambda v = 0$ to get the eigenvalues.

The second approach is used when the Markov chain is a random walk (this means that the Markov chain is loation invariant, $P(x, y) = Q(y - x)$ for some distribution Q) on a finite abelian group, $\mathbb{Z}/(n)$. The characteristic function is denoted as follows,

$$\chi_m(x) = e^{2\pi i m x/n}$$

The eigenvalue is, $\lambda_m = E_Q[\chi_m]$ and the eigenvector is $\overline{\chi}_m$ where $\overline{\chi}_m(x) = \chi_m(-x)$. Thus, we get that $\overline{\chi}_m P = \lambda_m \overline{\chi}_m$.

Eigenvectors will return when we talk about convergence bounds.

Note that eigenvalues are the finite version of spectral norm (or spectral gap?). We call $\lambda_* = \max_{i\in\{1,\ldots,n-1\}\neq 0} |\lambda_i| = ||P_0||$ the operator norm of P without π.

1.4.2 Existence of stationarity

Definition 1.12 (Doubly stochastic)**.** A Markov chain is doubly stochastic if $\sum_{i\in S} p_{ij} = 1 \forall j \in S$ or in other words the columns all add to 1.

If a Markov chain is doubly stochastic, then $\pi_i = \frac{1}{|S|}$ is a stationary distribution.

Proof.

$$\sum_{i\in S} \frac{1}{|S|} p_{ij} = \frac{1}{|S|} * 1 = \pi_j$$

□

Definition 1.13 (Reversible)**.** A Markov chain is reversible with respect to π if $\pi_i p_{ij} = \pi_j p_{ji} \forall i, j \in S$.

Theorem 1.8. *If P is reversible with respect to π then*

$$\sum_{i\in S} \pi_i p_{ij} = \sum_{i\in S} \pi_j p_{ji} = \pi_j$$

and so π is a stationary distribution of P.

A reversible Markov chain can be constructed from undirected graphs with weighted edges (All Markov chain can be constructed from directed graphs with weighted edges).

For example a reversible can be constructed from the graph in figure 3 where, $p_{12} = \frac{5}{5+7}$ and $\pi_3 = \frac{7+6}{2(7+5+6)}$

Theorem 1.9. *An irreducible Markov chain that has all positive recurrent states has a unique stationary distribution π where $\pi_i = \frac{1}{m_i}$.*

Proof. The proof utilizes lemma 1.5. □

Corollary. *If a Markov chain has a stationary distribution π and state i is null recurrent then $\pi_i = 0$.*

Proof.

$$\begin{aligned}
\pi_i = \sum_{j\in S} \pi_j p_{ji}^{(n)} &= \lim_{n\to\infty} \frac{\sum_{l=1}^{n} \sum_{j\in S} \pi_j p_{ji}^{(l)}}{n} \\
&= \sum_{j\in S} \pi_j \lim_{n\to\infty} \frac{\sum_{l=1}^{n} p_{ji}^{(l)}}{n} && \text{Since sum is finite by M-test} \\
&= \sum_{j\in S} \pi_j \frac{1}{\infty} = 0 && \text{By 1.5 if state } i \text{ is null recurrent}
\end{aligned}$$

□

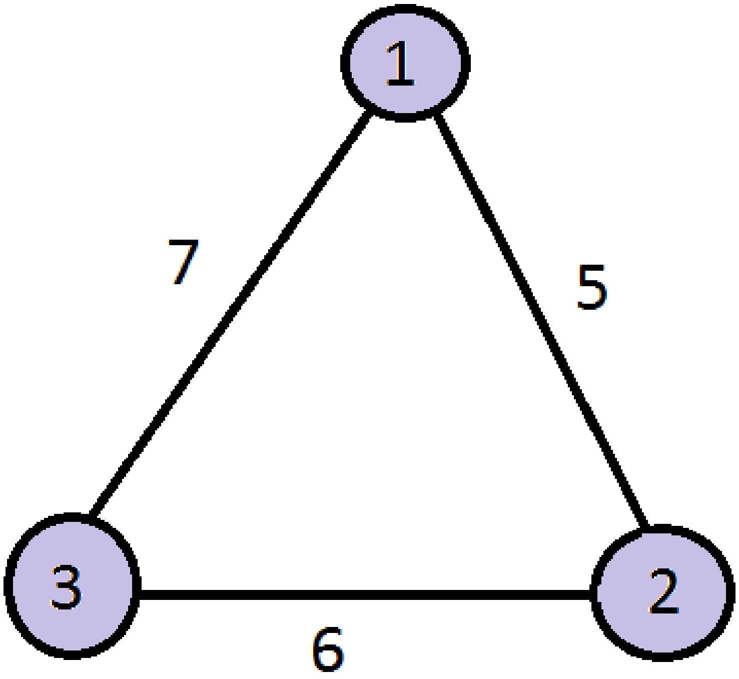

Figure 2: A reversible Markov chain can be constructed from an undirected graph with weighted edges.

By the above corollary a recurrent Markov chain is either all null or positive recurrent. Thus, for an irreducible Markov chain either all states are transitive or null recurrent and so no stationary distribution exists or all states are positive recurrent and the stationary distribution for any state j is $\pi_j = \frac{1}{m_j}$.

1.4.3 Markov chain convergence

Aperiodic Markov chains

Stationarity is not a sufficient condition to ensure that the Markov chain to converges to the stationary distribution. As stated above, convergence occurs when P is indecomposable and aperiodic. This is because the stationary distribution is unique. As figure 1 shows, a Markov chain that is irreducible is indecomposable and so convergence occurs when P is irreducible and aperiodic. A more formal proof will be shown later.

Before the fundamental lemma is proven, two results must be stated.

Lemma 1.10. *An irreducible Markov chain that has a stationary distribution must be positive recurrent.*

Proof. An irreducible Markov chain has states that are either all transitive, null recurrent or positive recurrent. If a stationary distribution exists, then the states cannot be transitive (exercise). If a stationary distribution is null recurrent then by corollary 1.4.2 each state i has stationary measure $\pi_i = 0$ and $\sum_{i \in S} \pi_i = 0$. This is a contradiction. Thus the Markov chain must have all positive recurrent states. □

Lemma 1.11. *A Markov chain is irreducible and aperiodic if and only if for each $i, j \in S$ there exists $n(i, j)$ such that for all $n \geq n(i, j)$, $p_{ij}^{(n)} > 0$. (I modified this to be an iff statement)*

Proof. Refer to textbook for $\Longleftarrow$. Beginning of personal proof: Let $A = \{n : p_{ii}^{(n)} > 0\}$. If $m, n \in A$ then $m + n \in A$. If $m \in A$ then for all $k \in \mathbb{N}$, $km \in A$. What type of set is A? Further, it must be proven that A^c is finite. I am having trouble proving this.

$\Longrightarrow$: Suppose that for each $i, j \in S$ there exists $n(i, j)$ such that for all $n \geq n(i, j)$, $p_{ij}^{(n)} > 0$. The Markov chain is irreducible, further there exists an $N = n(i, j) + n(j, i)$ such that for $n \geq N, p_{ii}^{(n)} > 0$ and so the Markov chain is aperiodic. □

Theorem 1.12 (Markov chain convergence theorem 1)**.** *If a Markov chain is irreducible and aperiodic and there exists a stationary distribution π then $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$.*

Proof. Let (X_n, Y_n) be a new Markov chain such that X_n and Y_n are two independent copies of the original irreducible and aperiodic Markov chain . This means that the transition probabilities are $p_{(ij)(kl)} = p_{ij}p_{kl}$ and $\pi_{ij} = \pi_i \pi_j$.

Since for each $i, j \in S$ there exists $n(i, j)$ such that for all $n \geq n(i, j)$, $p_{ij}^{(n)} > 0$, for the new Markov chain chain, there exists $n_0 = \max\{n(i, j), n(k, l)\}$ such that for all $n \geq n_0$, $p_{(ij)(kl))}^{(n)} > 0$ and so the Markov chain is irreducible and aperiodic by lemma 1.11.

Since the Markov chain is irreducible and has a stationary distribution, by lemma 1.10 the states are positive recurrent.

Let $i, j, k \in S$ and $i_0 \in S$ such that $\tau = inf\{n : X_n = Y_n = i_0\}$. We first want to show that $|p_{ik} - p_{jk}| \to 0$:

$$\begin{aligned}|p_{ik}^{(n)} - p_{jk}| &= |P_{(ij)}(X_n = k|\tau > n)P_{(ij)}(\tau > n) + P_{(ij)}(X_n = k|\tau \leq n)P_{(ij)}(\tau \leq n)\\ &\quad - P_{(ij)}(Y_n = k|\tau > n)P_{(ij)}(\tau > n) - P_{(ij)}(Y_n = k|\tau \leq n)P_{(ij)}(\tau \leq n)|\\ &= |P_{(ij)}(X_n = k|\tau > n)P_{(ij)}(\tau > n) - P_{(ij)}(Y_n = k|\tau > n)P_{(ij)}(\tau > n)|\\ &\qquad\qquad \text{Since } P_{(ij)}(X_n = k|\tau > n) = P_{(ij)}(Y_n = k|\tau > n)\\ &\leq P_{(ij)}(\tau > n)\end{aligned}$$

Since the Markov chain is irreducible and aperiodic, with probability 1, there exists an n such that $X_n = Y_n = i_0$ and so $P_{(ij)}(\tau > n) \to 0$.

Next we want to show that $|p_{ij} - \pi_j| \to 0$:

$$\begin{aligned}\lim_{n\to\infty} |p_{ij}^{(n)} - \pi_j| &= \lim_{n\to\infty} |\sum_{k\in S} \pi_k p_{ij}^{(n)} - \sum_{k\in S} \pi_k p_{kj}^{(n)}|\\ &= \lim_{n\to\infty} |\sum_{k\in S} \pi_k (p_{ij}^{(n)} - p_{kj}^{(n)})|\\ &\leq \sum_{k\in S} \pi_k \lim_{n\to\infty} |(p_{ij}^{(n)} - p_{kj}^{(n)})| \qquad \text{by the M-test}\\ &= \sum_{k\in S} \pi_k * 0 = 0\end{aligned}$$

□

The above theorem can be extended to any general initial distribution. Denote $P(X_n = j)$ as the probability of reaching point j at time n given transition probabilities P and initial distribution υ. Then since the summation is bounded by the M-test,

$$\lim_{n\to\infty} |P(X_n = j) - \pi_j| \leq \lim_{n\to\infty} \sum_{k\in S} \upsilon_k |p_{kj}^{(n-1)} - \pi_j| = \sum_{k\in S} \upsilon_k \lim_{n\to\infty} |p_{kj}^{(n-1)} - \pi_j| = 0$$

A more general statement can be shown through eigenvectors for Markov chain on finite state space S. It is more general, because irredicible Markov chain are indecomposable, but not all indecomposable Markov chain are irreducible. This statement also doesn't assume that a stationary distribution exists.

Theorem 1.13. . *If $\lambda_* < 1$ then there is a unique stationary distribution on P.*

Theorem 1.14 (Markov chain convergence theorem 2)**.** *If a Markov chain is indecomposable, aperiodic and has finite state space, then there exists a unique stationary distribution such that* $\min_{n\to\infty} P_k(X_n = j) = \pi_j$

Sketch. The proof is divided into two parts. First we will show that if $\lambda_* < 1$ then the Markov chain converges. Next we will show that $\lambda_* < 1 \iff$ the Markov chain is indecomposable and aperiodic. □

Example 1.1. The following matrix is aperiodic and indecomposable but not irreducible. Further, $\lambda_* = 1/3$.

$$\begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Periodic Markov chains

If a Markov chain is of period $d > 1$, not all hope is lost. There exists other forms of convergence.

Theorem 1.15. *If a Markov chain is irreducible, a stationary distribution exists and the period is d then*

$$\lim_{n\to\infty} \frac{1}{d} \sum_{r=0}^{d-1} p_{ij}^{(n+r)} = \pi_j$$

Proof. For a Markov chain, there exists a partition of S, $S_1, \ldots, S_d$ such that $P(S_i, S_{i+1}) = 1$ where $i \in \mathbb{Z}/(d)$. By symmetry $\pi(S_i) = \frac{1}{d}$. Pick S_r and let $p'_{ij} = p_{ij}^{(d)}$ for $i, j \in S_r$. Since the original Markov chain is irreducible, this sub Markov chain is irreducible and aperiodic and there exists a stationary distribution $\pi' = (\pi * d)I_{S_r}$, which is the original stationary distribution restricted to the subset S_r and is defined on p'_{ij}. By irreducibility and aperiodicity, we get that $\lim_{n\to\infty} p_{ij}^{\prime(n)} = \pi'_j$ or in other words, $\lim_{n\to\infty} p_{ij}^{(dn)} = \pi * d$ when $i, j \in S_r$.

Further since $p_{ij}^{(dn+r)} = 0$ for $1 \leq r \leq d-1$, $\lim_{n\to\infty} \frac{1}{d}\sum_{r=0}^{d-1} p_{ij}^{(n+r)} = \pi_j$ when $i, j \in S_r$. This can be shown to generalize to any $i, j \in S$. □

Theorem 1.16. *If a Markov chain is of period d, it is irreducible, and a stationary distribution exists then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n} p_{ij}^{(m)} = \pi_j$$

.

Proof.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n} p_{ij}^{(m)} = \lim_{n\to\infty} \frac{1}{dn} \sum_{m=0}^{n} \sum_{r=0}^{d-1} p_{ij}^{(dn+r)} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n} \frac{1}{d} \sum_{r=0}^{d-1} p_{ij}^{(dn+r)} = \pi_j$$

Since $\lim_{n\to\infty} \frac{1}{d}\sum_{r=0}^{d-1} p_{ij}^{(dn+r)} = \pi_j$ by lemma 1.15 and so by the Cesaro principle, the sample average of the limit also converges to π_j. □

This tells us that any irreducible Markov chain has at most 1 stationary distribution.

1.5 Continuous-state space Markov chains

We are going to extend the above theory to continuous state spaces while keeping the time periods discrete. The original definitions must be generalized.

Definition 1.14 (Markov Chain where S is continuous)**.** A Markov Chain is a sequence of random variables $\{X_n\}_{n\geq 0}$ equipped with the following attributes:

- A continuous state space S,
- An initial probability υ where $\upsilon(A) = P(X_0 \in A)$,
- A transition probability P where $P_n(x, A) = P(X_{n+1} \in A | X_n = x)$.

So

$$\begin{aligned} P(X_0 \in A_0, X_1 \in A_1, \ldots, X_d \in A_d) &= \int_{A_0} \int_{A_1} \cdots \int_{A_d} P(x_{d-1}, dx_d) \ldots P(x_2, dx_1) P(x_0, dx_1) \upsilon(dx_0) \\ &= \int_{A_0} \upsilon(dx_0) \int_{A_1} P(x_0, dx_1) \ldots \int_{A_d} P(x_{d-1}, dx_d) \end{aligned}$$

Definition 1.15 (ϕ-irreducible Markov chain)**.** A Markov chain on a general state space is considered ϕ-irreducible if for all sets $A \subseteq S$ such that $\phi(A) > 0$ and for all $x \in S$, there exists an $n \in \mathbb{N}$ such that $P^n(x, A) > 0$

Definition 1.16 (Period)**.** A Markov chain is of period d if there exists a partition $S = \dot{\bigcup}_{i=1}^{d} S_i$ such that the probability of transition from any set to the next set is 1 that is, $P(S_i, S_{i+1}) = 1$ where $i \in \mathbb{Z}/(d)$. Further there is no partition that is greater.

Theorem 1.17. *If there exists a $\delta > 0$ such that for all $x \in S$ and [**for all subsets** $A \subseteq [x - \delta, x + \delta]$**,** $P(x, A) > 0$] then the Markov chain is λ-irreducible and aperiodic.*

Proof. **Irreducible:** Let $A \subseteq S$ be a set such that $\lambda(A) > 0$. There exists an $M \in \mathbb{R}$ such that $M_A = [-M, M] \cap A \neq \emptyset$ and $\lambda(M_A) > 0$. Next, lets choose an $n \in \mathbb{N}$ such that $x_n\delta > M > -M > x - n\delta$. And so, $P^n(x, [x - n\delta, x + n\delta]) > 0$. Further, since every subset of $[x - n\delta, x + n\delta]$ has positive transition probability and $M_A \subseteq [x - n\delta, x + n\delta]$, $P^n(x, M_A) > 0$.

Aperiodic: Suppose the period of the Markov chain is $d > 1$ and let $x \in S_i$. Since $d > 1$, this means that $P(x, S_{i+1}) = 1$. Further by assumption, all subsets of $[x - \delta, x + d\delta]$ have positive transition probability and so, $[x - \delta, x + \delta] \subset S_{i+1}$. This means that $x \in S_i$ and $x \in S_{i+1}$, a contradiction since S_i's are disjoint, so $d = 1$. □

Definition 1.17 (Stationarity)**.** A distribution, π, is a stationary distribution of P if for all $A \subseteq S$ that have positive measure on π,

$$\pi(A) = \int_S \pi(x) P(x, A) dx$$

For continuous space Markov chain the convergence theorem 1.12 still applies. Although one must be more careful with the definition and formalizing convergence.

Finally a generalization of the definition of reversibility is as follows,

Definition 1.18 (Reversible)**.** Let $L^2(\pi)$ be the set of all measurable, real-valued functions on S that are squared integrable with respect to $\pi(dx)$. A continuous Markov chain is reversible with respect to π if for all $f, g \in L^2(\pi)$,

$$\int_X g(x) P f(x) \pi(dx) = \int_X f(x) P g(x) \pi(dx)$$

Definition 1.19 (Non negative definite)**.** A continuous Markov chain is non-negative definite if for all $f \in L^2(\pi)$,

$$\int_X f(x)Pf(x)\pi(dx) \geq 0$$

1.6 Total variation distance

Total variation is metric that will be used for defining convergence and bounds.

Definition 1.20 (Total variation)**.** The total variation distance between two measures υ_1, υ_2 is,

$$||\upsilon_1 - \upsilon_2|| = \sup_{A \subseteq S} |\upsilon_1(A) - \upsilon_2(A)|$$

Total variation is a metric. (Fun exercise)

Proof. **Identity:** If $\upsilon_1 = \upsilon_2$ then for all $A \in \mathcal{F}$, $\upsilon_1(A) = \upsilon_2(A)$ which occurs $\iff$ $\sup_{A \subseteq S} |\upsilon_1(A) - \upsilon_2(A)| = 0$.

Symmetry: Obvious.

Subadditivity:

$$\begin{aligned} ||\upsilon_1 - \upsilon|| &= \sup_{A \subseteq S} |\upsilon_1(A) - \upsilon_2(A)| \\ &\leq \sup_{A \subseteq S} |\upsilon_1(A) - \upsilon_3(A)| + |\upsilon_3(A) - \upsilon_2(A)| \\ &\leq \sup_{A \subseteq S} |\upsilon_1(A) - \upsilon_3(A)| + \sup_{A \subseteq S} |\upsilon_3(A) - \upsilon_2(A)| \end{aligned}$$

□

If the distribution functions have continuous densities then the total variation in 3 is the area of A or the area of B, which are equal. Why? Because the areas of densities for u and v are 1, so their differences must also cancel out.

In more formal terms the same applies for finite S.

Proposition 1.18. *If S is finite, then*

$$||\upsilon_1 - \upsilon_2|| = \frac{1}{2} \sum_{x \in S} |\upsilon_1(x) - \upsilon_2(x)|$$

Properties 1.19. *The following are equivalent definitions for total variation. (Note that the proofs are sketches– they wouldn't be final answers in a homework assignment.)*

1. $||\upsilon_1 - \upsilon_2|| = \sup_{f: S \to [0,1]} |\int f d\upsilon_1 - \int f d\upsilon_2|$

   Proof. Let $\{A_n\}_{n \geq 1}$ be a sequence of nested sets such that $A_n \in \mathcal{F}$ and

$$\lim_{n \to \infty} |\upsilon_1(A_n) - \upsilon_2(A_n)| = \sup_{A \subseteq S} |\upsilon_1(A) - \upsilon_2(A)|$$

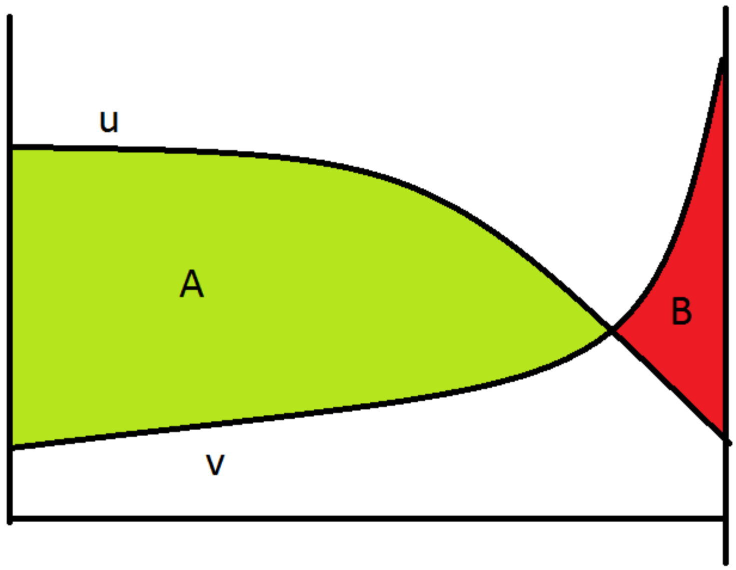

Figure 3:

Such a sequence exists, by definition of supremum and by continuity of probabilities. Further, let $f(x) = \lim_{n\to\infty} I_{x\in A_n}$. Then,

$$\begin{aligned}
\sup_{A\subseteq S} |\upsilon_1(A) - \upsilon_2(A)| &= \lim_{n\to\infty} |\upsilon_1(A_n) - \upsilon_2(A_n)| \\
&= \lim_{n\to\infty} |\int_{s\in S} I_{s\in A_n} d\upsilon_1 - \int_{s\in S} I_{s\in A_n} d\upsilon_2| \\
&\qquad\qquad\qquad\qquad I_{s\in A_n} \text{ is mesurable b/c } A_n \in \mathcal{F} \\
&\leq \sup_{f:S\to[0,1]} |\int f d\upsilon_1 - \int f d\upsilon_2| \qquad \text{Since it is true for each } I_{s\in A_n}
\end{aligned}$$

Further there exists an $f_n \to f'$ such that $|\int f' d\upsilon_1 - \int f' d\upsilon_2| = \sup_{f:S\to[0,1]} |\int f d\upsilon_1 - \int f d\upsilon_2|$. Let B_n be the set on which f_n is positive. Since f_n's are measurable, $B_n \in \mathcal{F}$ and so there exists an $A_{m(n)}$ such that

$$\begin{aligned}
|\int_{s\in S} f_n(s) d\upsilon_1 - \int_{s\in S} f_n(s) d\upsilon_2| &\leq |\int_{s\in S} I_{s\in B_n} d\upsilon_1 - \int_{s\in B_n} I_{s\in B_n} d\upsilon_2| \\
&\leq |\int_{s\in S} I_{s\in A_n} d\upsilon_1 - \int_{s\in A_n} I_{s\in B_n} d\upsilon_2|
\end{aligned}$$

And so,

$$\begin{aligned}\lim_{n\to\infty} |\int_{s\in S} f_n(s)d\upsilon_1 - \int_{s\in S} f_n(s)d\upsilon_2| &\leq |\int_{s\in S} I_{s\in A_n} d\upsilon_1 - \int_{s\in A_n} I_{s\in B_n} d\upsilon_2| \\ &= \sup_{A\subseteq S} |\upsilon_1(A) - \upsilon_2(A)|\end{aligned}$$

Since both inequalities hold, we get an equality. □

2. $||\upsilon_1 - \upsilon_2|| = \frac{1}{b-a} \sup_{f:X\to[a,b]} |\int f d\upsilon_1 - \int f d\upsilon_2|$

Proof. If $g_n : S \to [a,b]$ then $g_n = a + (b-a)f_n$ where $f_n : S \to [0,1]$,

$$\begin{aligned}\frac{1}{b-a} \sup_{g:X\to[a,b]} |\int g d\upsilon_1 - \int g d\upsilon_2| &= \frac{1}{b-a} \sup_{f:X\to[0,1]} |\int a + (b-a) f d\upsilon_1 - \int a + (b-a) f d\upsilon_2| \\ &= \sup_{f:X\to[0,1]} |\int f d\upsilon_1 - \int f d\upsilon_2| \qquad \text{By linearity} \\ &= \sup_{A\subseteq S} |\upsilon_1(A) - \upsilon_2(A)|\end{aligned}$$

□

3. If π is a stationary distribution for a Markov chain then $||P^n(x,\cdot) - \pi(\cdot)|| \leq ||P^{n-1}(x,\cdot) - \pi(\cdot)||$

Proof.

$$\begin{aligned}|P^{n+1}(x,A) - \pi(A)| &= |\int_{y\in S} P^n(x,dy)P(y,A) - \int_{y\in S} \pi(dy)P(y,A)| \\ &= |\int_{y\in S} P^n(x,dy)P(y,A) - \int_{y\in S} \pi(dy)P(y,A)| \\ &= |\int_{y\in S} f(y)P^n(x,dy) - \int_{y\in S} f(y)\pi(dy)| \quad \text{Where } f(y) = P(y,A) \\ &\leq \sup_{f:S\to[0,1]} |\int_{y\in S} f(y)P^n(x,dy) - \int_{y\in S} f(y)\pi(dy)| \\ &= ||P^n(x,\cdot) - \pi(\cdot)|| \qquad \text{By property A}\end{aligned}$$

□

4. Let $t(n) = 2\sup_{x\subseteq S} ||P^n(x,\cdot) - \pi(\cdot)||$, where $\pi(\cdot)$ is a stationary distribution, then $t(m + n) \leq t(m)t(n)$ or

$$\sup_{x\subseteq S} ||P^{n+m}(x,\cdot) - \pi(\cdot)|| \leq 2 \sup_{x\subseteq S} ||P^n(x,\cdot) - \pi(\cdot)|| \times \sup_{x\subseteq S} ||P^m(x,\cdot) - \pi(\cdot)||$$

5. *If $P^n(x,\cdot)$ and $\pi(\cdot)$ have densities f_n and g with respect to some σ-finite measure ρ and $M = \max\{f_n, g\}$, $m = \min\{f_n, g\}$ then,*

$$||P^n(x,\cdot) - \pi(\cdot)|| = \frac{1}{2}\int_S M - m d\rho = 1 - \int_S m d\rho$$

   In figure 3, this equation represents the average area of A and B.

6. *There exists joint defined random variables X_n and Y such that $L(X_n) = P^n(x,\cdot)$ and $L(Y) = \pi(\cdot)$ and*

$$||P^n(x,\cdot) - \pi(\cdot)|| = 1 - P(X_n = Y)$$

Ideas:

Suppose the transition probability P has a density $f(x,y)$.

Then $\int_{y\in S} f(x,y)dy = 1$ but $\int_{x\in S} f(x,y)dx \geq 0$ and is not bounded. If P is ϕ-irreducible then for any $A \subseteq S$ where $\phi(A) > 0$ there exists a $B \subseteq S$ such that $\phi(B) > 0$ and $P(B, A) > 0$.

I'd be really interested in showing the following (conjecture)

- If π is the stationary distribution and the Markov chain is irreducible and aperiodic. Further then for all $n \geq 1$

$$0 < ||P^n(x,\cdot) - \pi(\cdot)|| < ||P^{n-1}(x,\cdot) - \pi(\cdot)||$$

  That is, the convergence is asymptotic.

  Proof. (Attempt) □

- If there exists a $\delta > 0$ such that the transition proability $P(x,\cdot)$ has strictly positive density over $[x-\delta, x+\delta]$ could be shown that the transition distribution is absolutely continuous that is $P^n >> P^{n-1}$

Theorem 1.20. *If a Markov chain is ϕ-irreducible and aperiodic and there exists a stationary distribution π where [$\phi >> \pi$] then for π-a.e and $x \in X$.*

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

.

2 MCMC

Problem: Given an unnormalised (or improper) π_u where $0 < \int \pi_u(x)dx < \infty$ and π non-negative a.e. we want to estimate $E_\pi[f]$ where π is the normalised distribution.

Markov Chain Monte Carlo is an algorithm to simulate random variables from complicated distributions that cannot be directly sampled. Its most prominent use is in Bayesian statistics when the density function is improper or unnormalised.

Lets break down the definition into parts.

Definition 2.1 (Monte Carlo simulation)**.** Monte Carlo simulation is about using i.i.d. samples to estimate the $E_\pi[f]$ through the sample mean. That is,

$$E_\pi[f] \approx \frac{1}{N}\sum_{n=1}^{N} f(X_n)$$

By the weak law of large numbers if the X_is has the finite first and second moment, then the average converges and so the Monte Carlo simulation will converge in probability to $E_\pi[f]$.

In the turn of the century Russia, a big debate was going on. Was pairwise independence a necessary condition for the weak law of large numbers to hold? Nekrasov thought so, but Markov wanted to prove him wrong. This is how the concept of Markov chains came about: it was Markov's counter-example. (p. 244-245 of Statisticians of the Centuries). Regardless of how Markov did it, it is with this theoretical underpinning that allows us to use a sequence of non-independent random variables whose sample mean still converges to the sample mean in probability. Further, if the samples cannot be estimated because the normalising constant for π_u is unknown then we need another solution that uses a Markov chain who mean converges in probability to the sample mean. This is the foundation of Markov Chain Monte Carlo (MCMC).

MCMC differs from Monte Carlo simulation in that it doesn't sample i.i.d. random variables. Instead a Markov chain is simulated whose unconditional distribution converges to π. That is $L(X_n) \to \pi$ as $n \to \infty$.

Definition 2.2 (Markov Chain Monte Carlo (MCMC) algorithm)**.** MCMC is a Monte Carlo algorithm where the samples are generated from a random a Markov chain.

Suppose $X_1, X_2, \ldots X_N$ are simulated values from a Markov chain. For a sufficiently large $N_0 < N$, and for $N_0 \leq n \leq N$ $L(X_n) \approx \pi$ and so $E_\pi[f]$ is estimated as follows.

$$E_\pi[f] \approx \frac{1}{N-N_0}\sum_{n=N_0}^{N} f(X_n)$$

The next question that comes to mind is how will we get a Markov chain whose limiting distribution is π?

If we can construct a transition probability P that is reversible with respect to π then π is also a stationary distribution and combining this with the Markov chain convergence theorem 1 1.12, the Markov chain will converge in distribution to π regardless of the initial distribution and so we obtain what we are looking for and the initial distribution is irrelevant.

The Metropolis Hastings Algorithm generates a Markov chain that converges to a stationary distribution π.

Definition 2.3 (Metropolis Algorithm)**.** Given an unnormalized distribution π_u and a state space S a Markov chain can be generated as follows.

Choose a proposal density $q(\cdot,\cdot)$ that is symmetric ($q(x,y) = q(y,x)$) and define the density of P as follows if $y \neq x$,

$$dP(x,y) = q(x,y)min\{1, \frac{\pi(y)}{\pi(x)}\}$$

And $P(x,\{x\}) = 1 - \int_S q(x,y)min\{1, \frac{\pi(y)}{\pi(x)}\}dx$.

Theorem 2.1. *The transition probability generated by the Metropolis algorithm is reversible. If $y \neq x$*

$$\begin{aligned}\pi(x)P(x,y) &= \pi(x)q(x,y)\min\{1, \frac{\pi(y)}{\pi(x)}\} \\ &= q(x,y)\min\{\pi(x),\pi(y)\} \\ &= q(y,x)\min\{\pi(x),\pi(y)\} \\ &= \pi(y)P(y,x)\end{aligned}$$

Theorem 2.2 (MCMC convergence theorem)**.** *If S is a compact set and the density of π [where $\pi << \lambda$] is strictly positive over S and there exists a $\delta > 0$ such that for all $x \in S$ and $y \in B_\delta(x)$, $q(x,y) > 0$ then the corresponding P generated from the Metropolis Hastings algorithm is π-irreducible and aperiodic and so*

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

Proof. This is a direct result of theorem 1.17 and 1.20. □

The Metropolis algorithm can be generalized to the Metropolis Hastings algorithm.

Definition 2.4 (Metropolis-Hastings algorithm)**.** Given an unnormalized distribution π_u and a state space S a Markov chain can be generated as follows.

Choose a proposal density $q(\cdot,\cdot)$ and define the density of P as follows if $y \neq x$,

$$dP(x,y) = q(x,y)min\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\}$$

And $P(x,\{x\}) = 1 - \int_S q(x,y)min\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\}dx$.

The Metropolis-Hastings algorithm is reversible and the MCMC convergence theorem still holds.

3 Markov Chain convergence

3.1 Foundations of the minorization and drift conditions

Now that we are provided with a selection of requirements for a Markov chain to converge to stationarity (ie that Markov chain is ergodic), the next question is at what rate does the law of the Markov chain converge to stationarity?

Definition 3.1 (Uniformly ergodic (section 3.3 of [2]))**.** A Markov chain with stationary distribution π is uniformly ergodic if there exists a $1 > \rho > 0$ and $M < \infty$ such that

$$||P(x,\cdot) - \pi(\cdot)|| \leq M\rho^n$$

Theorem 3.1 (Proposition 7 of [2])**.** *A Markov chain is uniformly ergodic $\iff$ there exists an $n_0 \in \mathbb{N}$ such that,*

$$\sup_{x\in S} ||P^{n_0}(x,\cdot) - \pi(\cdot)|| < 1/2$$

Proof. $\Longrightarrow$: Obvious

$\Longleftarrow$: Suppose that $\sup_{x\in S} ||P^{n_0}(x,\cdot)-\pi(\cdot)|| < 1/2$. Let

$$\beta = t(n_0) = 2\sup_{x\in S} ||P^{n_0}(x,\cdot)-\pi(\cdot)||$$

So, $\beta < 1$ and by

Let $m \in \mathbb{N}$ and by proposition 1.19 part 4, $t(jn) \leq t(n)^j$ and so,

$$\begin{aligned} ||P^m(x,\cdot)-\pi(\cdot)|| &\leq ||P^{\lfloor \frac{m}{n_0}\rfloor n_0}(x,\cdot)-\pi(\cdot)|| && \text{By prop 1.19 part 3} \\ &= t(\lfloor \frac{m}{n_0}\rfloor n_0)/2 \\ &\leq \beta^{\lfloor \frac{m}{n_0}\rfloor}/2 && \text{By prop 1.19 part 4} \\ &\leq \beta^{-1}/2 \times \beta^{\frac{m}{n_0}} \\ &= M\rho^m && \text{where } M = \beta^{-1}/2 \text{ and } \rho = \beta^{\frac{1}{n_0}} \end{aligned}$$

□

Definition 3.2 (Minorisation condition and small set (equation 8 in [2])). The minorisation condition on a small set $C \subseteq S$ holds when there exists an $\epsilon > 0$, a distribution υ and an $n_0 \in \mathbb{N}$ such that, for all $A \subseteq S$ and $x \in C$

$$P^{n_0}(x,A) \geq \epsilon\upsilon(A) \tag{1}$$

Theorem 3.2 (Minorization condition and uniform ergodicity). *If a Markov chain has a stationary distribution π and the minorization condition holds with the small set equal to the state space, $C = S$, then the Markov chain is uniformly ergodic.*

Proof. See presentation. □

The special thing about the small set is that there is positive probability of coupling happening. That is if two random variables belong to the small set, they can be constructed so that there is a positive probability that they will become equal.

Definition 3.3 (Geometrically ergodic (section 3.4 of [2])). A Markov chain with stationary distribution π is geometrically ergodic if there exists a $1 > \rho > 0$ and $M : S \mapsto R_+ - \{\infty\}$ such that

$$||P(x,\cdot)-\pi(\cdot)|| \leq M(x)\rho^n \tag{2}$$

Definition 3.4 (Univariate drift condition (Equation 10 in [2])). A Markov chain satisfies the diffusion condition when there exists a function $V : S \mapsto [1,\infty]$ and two constants, $0 < \lambda < 1$ and $b < \infty$ such that,

$$E[V(X_{n+1})|X_n] = \int_{y\in S} V(y)P(X_n,dy) \leq \lambda V(X_n) + bI_C(X_n) \tag{3}$$

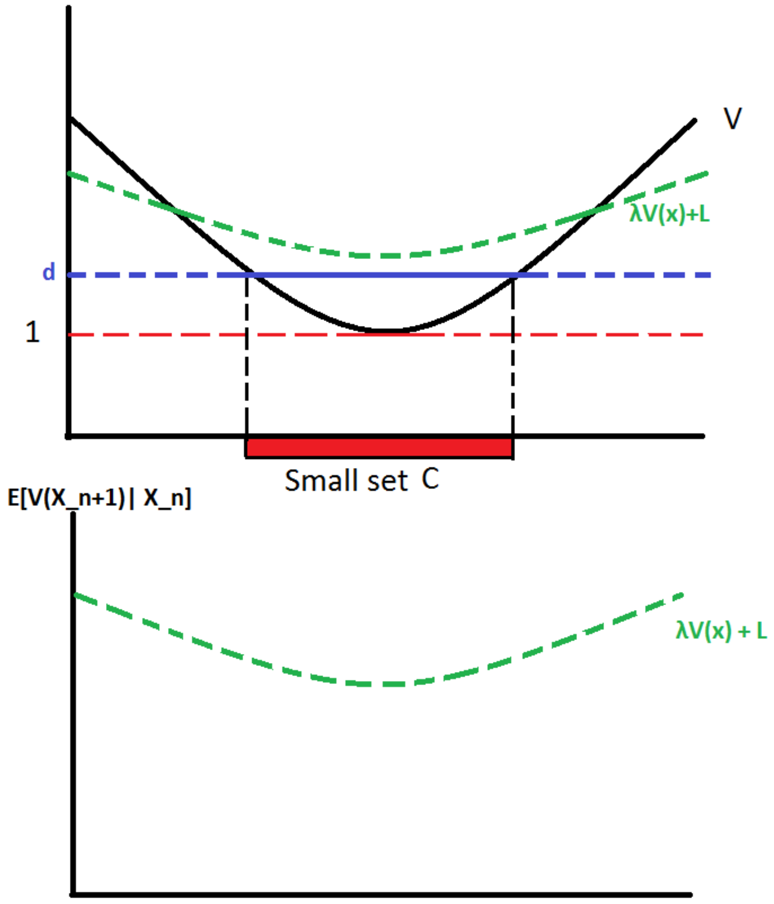

Figure 4:

Definition 3.5 (Bivariate drift condition (Equation 11 in [2]))**.** A bivariate condition holds on a Markov chain if there exists an $\alpha > 1$ and a bivariate function $h(x, y) : S \times S \mapsto [1, \infty]$ such that for all $(X_n, Y_n) \notin C \times C$,

$$E[h(X_{n+1}, Y_{n+1})|X_n, Y_n] = \int_{s\in S}\int_{t\in S} h(s, t)P(X_n, ds)P(Y_n, dt) \leq h(X_n, Y_n)/\alpha \tag{4}$$

Definition 3.6 (Petite set)**.** A subset $C' \subseteq S$ is a petite set relative to a Markov chain if there exists a positive integer n_0, $\epsilon > 0$ and a probability measure $\upsilon(\cdot)$ on S such that for all $x \in C'$ and $A \subseteq S$

$$\sum_{1\leq i\leq n_0} P^i(x, A) \geq \epsilon\upsilon(A) \tag{5}$$

The main purpose of this note is to prove the following theorem.

Theorem 3.3 (Minorization condition with drift and geometric ergodicity)**.** *Suppose a Markov chain has the following attributes. It is*

- *ϕ-irreducible and aperiodic,*
- *a minorisation condition holds (def. 3.2),*
- *a diffusion condition holds (def. 3.4).*

Then the Markov chain is geometrically ergodic (def. 3.3).

To prove this we will first state the following lemmas.

Lemma 3.4 (Lemma 18 of [2])**.** *Suppose a Markov chain has a drift (def. 3.4) and minorization condition (def. 3.2) Let $C' = C \cup S_d$ where C is a small set (def. 3.2) and $S_d = \{x \in S : V(x) \leq d\}$, where*

$$d > b/(1 - \lambda) - 1 \tag{6}$$

then C' is petite for some $n_0 \in \mathbb{N}$, $\epsilon > 0$ and distribution υ.

Lemma 3.5 (Lemma 17 of [2])**.** *If a Markov chain is ϕ-irreducible and aperiodic, then a petite set is a small set.*

Lemma 3.6 (Proposition 11 of [2])**.** *Suppose that the univariate drift condition (def. 3.4) holds on a Markov chain and let $d = \inf_{x\in C^c} V(x)$. If equation 6 holds then the bivariate drift condition holds (def. 3.5) with $h(x, y) = \frac{1}{2}(V(x) + V(y))$ and $\alpha^{-1} = \lambda + b/(d + 1)$.*

Lemma 3.7 (Theorem 12 of [2])**.** *Suppose that a minorization condition (def. 3.2) and a bivariate drift condition (def. 3.5) holds on a Markov chain. Let*

$$B_{n_0} = \max\{1, \alpha^{n_0}(1 - \epsilon) \sup_{(x,y)\in C\times C} \bar{R}h(x, y)\}$$

where for $x, y \in C \times C$

$$\bar{R}h(x, y) = \frac{1}{(1 - \epsilon)^2}\int_X\int_X h(z, w)(P^{n_0}(x, dz) - \epsilon\upsilon(dz))(P^{n_0}(y, dw) - \epsilon\upsilon(dw))$$

Further, let $\{X_n\}_{n\geq1}$ and $\{Y_n\}_{n\geq1}$ be two copies of the Markov chain with initial distribution $L(X_0, Y_0)$. Then for $k \in \mathbb{N}$ and integer $1 \leq j \leq k$

$$||L(X_k) - L(Y_k)|| \leq (1 - \epsilon)^j + \alpha^{-k}B_{n_0}^{j-1}E[h(X_0, Y_0)]$$

It is clear to see that the above lemmas follow a given sequence to prove theorem 3.3. First a petite set C' is proposed that satisfies equation 6. Next, we show that C' is small. Then because equation 6 and a minorization condition hold, a bivariate drift condition holds. Since a bivariate drift condition holds, there is a geometric ergodic bound.

Proof of lemma 3.4. The minorization condition holds, so given the notation defined in def. 3.2, define $N \in \mathbb{N}$ $1 - \lambda^N d > 0$ and define $r = 1 - \lambda^N d > 0$. We will show that for any $x \in C'$ and $A \subseteq S$

$$\sum_{i=1}^{N+n_0} P^i(x, A) \geq r\epsilon v(A) \tag{7}$$

First define the stopping time,

$$\tau_0 = \inf\{n \in \mathbb{N} : X_n \in C\}$$

Step 1: Let $W_n = Z_{\min\{n,\tau_0\}}$ and $Z_n = \lambda^{-n}V(X_n)$. Show that W_n is a supermartingale.

If $\tau_0 \leq n$ then $W_n = Z_{\tau_0}$ and $E[W_{n+1}|X_n] = W_n$

If $\tau_0 > n$ then $W_n = \lambda^{-n}V(X_n)$ and $X_1, X_2, \ldots X_n \notin C$, the drift condition is $E[V_{n+1}|X_n] \leq \lambda V(X_n)$ and

$$E[W_{n+1}|X_n] = E[\lambda^{-(n+1)}V(X_{n+1})|X_n] \leq \lambda^{-(n+1)}\lambda V(X_n) = \lambda^{-n}V(X_n) = W_n$$

So for all n, $E[W_{n+1}|X_n] \leq W_n$ and so W_n is a supermartingale.

Step 2: Show that $P_{X_0}(\tau_0 < N) \geq 1 - \lambda^N d = r$ for any $X_0 \in C'$.

$$\begin{aligned}
P_{X_0}(\tau_0 > N) &= P_x(\lambda^{-\tau_0} > \lambda^{-N}) \\
&\leq \frac{E_{X_0}[\lambda^{-\tau_0}]}{\lambda^{-N}} && \text{By Markov's inequality} \\
&\leq \lambda^N E_{X_0}[\lambda^{-\tau_0}V(X_{\tau_0})] && \text{Since } V(\cdot) \geq 1 \\
&\leq \lambda^N E_{X_0}[\lambda^0 V(X_0)] = \lambda^N V(X_0) && \text{Since } W_n \text{ is a supermartingale by step 1} \\
&\leq \lambda^N d && \text{Since } X_0 \in S
\end{aligned}$$

This shows that $P_{X_0}(\tau_0 \leq N) \geq 1 - \lambda^N d = r$.

Step 3: Show that $\sum_{i=1}^{N+n_0} P^i(x, A) \geq r\epsilon v(A)$ for any $x \in C'$.

$$\begin{aligned}
\sum_{i=1}^{N+n_0} P^i(x, A) &\geq \sum_{i=1}^{N} P^{i+n_0}(x, A) \\
&\geq \sum_{i=1}^{N} \int_{y \in C} P^i(x, dy) P^{n_0}(y, A) && \text{By the Chapman-Kolmogomorov inequality} \\
&\geq \sum_{i=1}^{N} \int_{y \in C} P^i(x, dy) \epsilon v(A) && \text{By the minorization condition} \\
&= \epsilon v(A) \sum_{i=1}^{N} P^i(x, C) \\
&\geq \epsilon v(A) P(\cup_{i=1}^{N} X_i \in C | X_0 = x) && \text{By subadditivity} \\
&= \epsilon v(A) P_{X_0}(\tau_0 \leq N) \\
&\geq r\epsilon v(A) && \text{Since } P_{X_0}(\tau_0 \leq N) \geq r \text{ by step 2}
\end{aligned}$$

And so C' is a petite set. □

Proof of 3.6. Suppose $(X_n, Y_n) \notin C \times C$. This means that at least one X_n or Y_n is in C^c.

$$\begin{aligned}
E[h(X_{n+1}, Y_{n+1}) | X_n, Y_n] &= \frac{E[V(X_{n+1})|X_n] + E[V(Y_{n+1})|Y_n]}{2} \\
&\leq \frac{\lambda V(X_n) + \lambda V(Y_n) + b}{2} && \text{By the minorization condition} \\
&\leq \frac{\lambda(V(X_n) + V(Y_n))}{2} + \frac{b}{2} \frac{V(X_n) + V(Y_n)}{d+1} \\
& && \text{Since } V(\cdot) \geq 1 \text{ and } V(x) \geq 1 \forall x \in C \\
&= \frac{V(X_n) + V(Y_n)}{2} \left(\lambda + \frac{b}{d+1} \right) \\
&= \alpha^{-1} \frac{V(X_n) + V(Y_n)}{2}
\end{aligned}$$

Further,

$$\alpha^{-1} < 1 \iff \lambda + \frac{b}{d+1} < 1 \iff \frac{b}{1-\lambda} - 1 < d$$

which is true by assumption. □

Proof of 3.7. Let $N_k = |\{m : 0 \leq m \leq k \text{ and } X_m = Y_m\}|$ and fix $1 \leq j \leq k$. First,

$$\begin{aligned}
||L(X_k) - L(Y_k)|| &\leq P(X_k \neq Y_k) \\
&= P(X_k \neq Y_k \cap N_k \geq j) + P(X_k \neq Y_k \cap N_k < j)
\end{aligned}$$

Each term can be bounded. Since $N_k \geq j$ in the first term, at least j times the Markov chain is in the small set and so by coupling we get that

$$P(X_k \neq Y_k \cap N_k \geq j) \leq (1-\epsilon)^j$$

The second term is more labour intensive.

First it must be shown that the term $M_k = \alpha^k h(X_k, Y_k) I(X_k \neq Y_k) B^{-N_{k-1}}$ is a supermartingale. At the core of the proof is the bivariate drift condition.

$$E[M_k | X_k, Y_k] = E[\alpha^k h(X_k, Y_k) I(X_k \neq Y_k) B^{-N_{k-1}} | X_k, Y_k]$$

Case 1: If $X_k = Y_k$, then $X_{k+1} = Y_{k+1}$ and so $M_{k+1} = M_k = 0$. Inequality holds.

Case 2: If $X_k \neq Y_k$ and $(X_k, Y_k) \in C \times C$. Then $N_k = N_{k-1} + 1$ and $B^{-N_k - 1} \leq B^{-N_{k-1}}$, which is counter to what we want. This is where the definition of B comes in and the application of coupling is used.

$$\begin{aligned}
E[M_{k+1} | X_k, Y_k] &= E[\alpha^{k+1} h(X_{k+1}, Y_{k+1}) I(X_{k+1} \neq Y_{k+1}) B^{-N_k} | X_k, Y_k] \\
&= \alpha^k B^{-N_{k-1}} \frac{\alpha}{B} E[h(X_{k+1}, Y_{k+1}) I(X_{k+1} \neq Y_{k+1}) | X_k, Y_k] \\
&= \alpha^k B^{-N_{k-1}} \frac{\alpha}{B} \left((1-\epsilon) \int_{z \in S} \int_{w \in S} h(w,z) \left(\frac{P^{n_0}(X_k, dw) - \epsilon \upsilon(dw)}{1-\epsilon} \right) \left(\frac{P^{n_0}(Y_k, dw) - \epsilon \upsilon(dw)}{1-\epsilon} \right) \right) \\
&= \alpha^k B^{-N_{k-1}} \left(\frac{\alpha(1-\epsilon) \bar{R} h(x,y)}{\max\{1, \alpha^{n_0}(1-\epsilon) \sup_{(x,y) \in C \times C} \bar{R} h(x,y)\}} \right) \\
&\leq \alpha^k B^{-N_{k-1}} \qquad \text{Since term is less than or equal to 1} \\
&\leq \alpha^k B^{-N_{k-1}} h(X_k, Y_k) I(X_k \neq Y_k) \qquad \text{Since } h(X_k, Y_k) I(X_k \neq Y_k) \geq 1 \\
&= M_k
\end{aligned}$$

Case 3: If $X_k \neq Y_k$ and $(X_k, Y_k) \notin C \times C$. If $(X_k, Y_k) \notin C \times C$ then $N_k = N_{k-1}$ and so,

$$\begin{aligned}
E[M_{k+1} | X_k, Y_k] &= E[\alpha^{k+1} h(X_{k+1}, Y_{k+1}) I(X_{k+1} \neq Y_{k+1}) B^{-N_k} | X_k, Y_k] \\
&= \alpha^{k+1} B^{-N_{k-1}} E[h(X_{k+1}, Y_{k+1}) I(X_{k+1} \neq Y_{k+1}) | X_k, Y_k] \\
&\leq \alpha^{k+1} B^{-N_{k-1}} I(X_k \neq Y_k) \frac{h(X_{k+1}, Y_{k+1})}{\alpha} \\
&\text{By the bivariate drift condition and } I(X_{k+1} \neq Y_{k+1}) \leq I(X_k \neq Y_k) \\
&= \alpha^k B^{-N_{k-1}} I(X_k \neq Y_k) h(X_{k+1}, Y_{k+1}) \\
&= M_k
\end{aligned}$$

Thus, $M_k = \alpha^k h(X_k, Y_k) I(X_k \neq Y_k) B^{-N_{k-1}}$ is a supermartingale. Next, the inequality is shown.

Note that in the 4th equality the term $I(X_k \neq Y_k) B^{-N_{k-1}}$ is a supermartingale, however B^{j-1} is an increasing function and for $P(X_k \neq Y_k \cap N_k \geq j)$ to converge to 0, we need $j \to \infty$ as $k \to \infty$ and so $\frac{\alpha^k}{\alpha^k}$ is added to the equation. To bound the numerator the term $h(X_k, Y_k)$ is added because it is greater that 1 and combined with α^k provides a supermartingale property.

$$
\begin{aligned}
P(X_k \neq Y_k \cap N_{k-1} < j) &= P(X_k \neq Y_k \cap N_{k-1} \leq j-1) \\
&= P(X_k \neq Y_k \cap B^{-N_{k-1}} \geq B^{-(j-1)}) \\
&= P(I(X_k \neq Y_k)B^{-N_{k-1}} \geq B^{-(j-1)}) && \text{Because } B^{-(j-1)} > 0 \\
&= B^{j-1}E[I(X_k \neq Y_k)B^{-N_{k-1}}] && \text{By Markov's inequality} \\
&= \frac{B^{j-1}}{\alpha^k}E[\alpha^k h(X_k, Y_k)I(X_k \neq Y_k)B^{-N_{k-1}}] && \text{Since } h(X_k, Y_k) \geq 1 \\
&\leq \frac{B^{j-1}}{\alpha^k}E[\alpha^0 h(X_0, Y_0)I(X_0 \neq Y_0)B^{-N_{-1}}] \\
&\qquad \text{Since } \alpha^k h(X_k, Y_k)I(X_k \neq Y_k)B^{-N_{k-1}} \text{ is a supermartingale} \\
&= \frac{B^{j-1}}{\alpha^k}E[h(X_0, Y_0)]
\end{aligned}
$$

□

Putting all of the lemmas together, we have a proof of theorem 3.3.

3.2 Discussion of bounds related to the minorization and drift conditions

Definition 3.7 (Convergence rate)**.** The convergence rate of a Markov chain is denoted as

$$\rho_* = \inf\{\rho : ||P^n(x, \cdot) - \pi(\cdot)|| \leq M(x)\rho^n\}$$

or is the minimum ρ that satisfies definition 3.3.

This discussion will focus on the convergence rate of geometrically ergodic Markov chains. We want to study how well certain famous bounds that apply the minorization and drift conditions estimate the convergence rate.

The goal is to better understand the limits of a bound that utilizes the drift and minorization condition. To do this, famous bounds that use the drift and minorization condition are studied.

3.2.1 Rosenthal's bound

The results come from [3] and [5].

Theorem 3.8 (Original: Rosenthal's bound, see theorem 12 in [1]. This version is taken from theorem 4 in [3])**.** *Suppose that*

1. *there exists* $\lambda < 1$ *and* $L < \infty$ *and a function* $V : S \to [0, \infty)]$ *such that for all* $x \in S$

$$E[V(X_{n+1})|X_n] \leq \lambda V(X_n) + L$$

2. *there exists* $d > 2L/(1-\lambda)$ *and a probability measure* $\upsilon : \mathcal{F} \to [0, 1]$ *such that for every* $x \in C$ *where*

$$C = \{x \in S : V(x) \leq d\}$$

and for all $A \in \mathcal{F}$, $P(x, A) \geq (1-\gamma)\upsilon(A)$. *(ie. the minorization condition holds)*

Then for all $x \in S$, $m \geq 0$ and $a \in (0,1)$,

$$||P^m(x,\cdot) - \pi(\cdot)|| \leq \gamma^{am} + \left(1 + \frac{L}{1-\lambda} + V(x)\right)\left[\left(\frac{1+2L+\lambda d}{1+d}\right)^{1-a}[1+2(\lambda d + L)]^a\right]^m$$

Proof. See theorem 3.3 for a proof of a similar theorem. Some differences are that b is replaced by $2b$ in lemma 3.6 to take into account the drift condition.

Regardless of the original minorization condition, in the proof a new minorization condition is constructed where the small set C is modified to be a set that can be defined as follows for $2L/(1-\lambda) \leq d < \infty$.

$$C = \{x \in S : V(x) \leq d\}$$

The small set can be shrunk to ensure that $d < \infty$ (see lemma 14 of probsurv). Also intuitively it makes sense given the small set is concentrated around the smallest values of $V(\cdot)$.

The small set is also expanded so that the new set, C' is defined as follows, $C' = C \cup \{x \in S : V(x) \leq d\}$.

Expanding the small set means that the original γ in the minorization condition may be larger than the constant found in the bound. □

Theorem 3.8 implies that a upper bound on the convergence rate, which we will call Rosenthal's bound, denoted ρ_{rose}, is as follows.

$$\rho_* \leq \rho_{rose} = \max\{\gamma^a, \left(\frac{1+2L+\lambda d}{1+d}\right)^{1-a}[1+2(\lambda d+L)]^a\}$$

Note that the requirement that $\inf_{x\in C^c} V(x) = d > 2L/(1-\lambda)$ ensures that the chain is aperiodic (it is a more strict requirement than aperiodicity) and that $\pi(C) \geq 1/2$.

Proposition 3.9 (Proposition 2.16 in [6])**.** *For C in theorem 3.8, $\pi(C) \geq 1/2$.*

Proof. Let notation be as in theorem 3.8.

$$\begin{aligned}
\frac{L}{1-\lambda} &\geq E_\pi[V(X_n)] && \text{By cutoff argument}\\
&\geq d(1-\pi(C)) && \text{Since } V(X) \geq dI_{C^C}(x)\\
&\geq \frac{2L(1-\pi(C))}{1-\lambda} && \text{Since } d \geq \tfrac{2L}{1-\lambda}
\end{aligned}$$

This implies that,

$$1/2 \geq 1-\pi(C) \iff \pi(C) \leq 1/2$$

□

The following is a general result for when a Markov chain satisfies a minorization condition.

Lemma 3.10 (Lemma 5 in [5], which was left as an exercise to the reader)**.** *Let $k(x,\cdot)$ be the density of $K(x,\cdot)$ with respect to some measure μ. If assumption 2 in 3.8 holds, then for $x, y \in C$,*

$$\frac{1}{2}\int_S |k(x,x') - k(y,x')|dx \leq \gamma$$

Proof. By assumption A2, for $x, y \in C$

$$|K(x,\cdot) - K(y,\cdot)| \leq 1 - (1-\gamma)\upsilon(\cdot) = (1 - \upsilon(\cdot)) + \gamma\upsilon(\cdot)$$

and so,

$$\frac{1}{2}\int_S |k(x,s) - k(y,s)|\mu(ds) \leq \int_S \gamma\upsilon(ds) = \gamma$$

□

Lemma 5 is then used in 3.1 to prove that Rosenthal's bound in the autoregressive normal process converges to 1 exponentially fast as the dimension p increases.

Example 3.1 (Autoregressive normal process). **Setup:** Let $\{\vec{X}_n\}_{n\geq 1}$ be a p dimensional Markov chain such that

$$L(\vec{X}_{n+1}|\vec{X}_n) = N\left(\frac{\vec{X}_n}{2}, \frac{3}{4}I_p\right)$$

The chain is reversible with respect to the distribution $\pi(\cdot) = N(0, I_p)$ since $P(x,dy)\pi(dx)$ is symmetric with respect to x, y.

$$\begin{aligned} P(x,dy)\pi(dx) &= Ce^{-\frac{2}{3}||x/2-y||^2}e^{-||x||^2} \\ &= C\exp(\frac{2}{3}\sum_{i=1}^{p}\left(\frac{x_i}{2} - y_i)^2 + \frac{1}{2}\sum_{i=1}^{p}x_i^2\right) \\ &= C\exp\left(\sum_{i=1}^{p}\frac{2}{3}x_i^2 - \frac{2}{3}x_iy_i + \frac{2}{3}y_i^2\right) \end{aligned}$$

Thus, π is a stationary distribution of P. Further, it is known that the convergence rate is $\rho_* = 1/2$. That is there is a function $M(x) < \infty$ such that,

$$||L(\vec{X}_n) - \pi(\cdot)|| \leq M(x)(1/2)^n$$

and there is no value smaller than 1/2.

Any drift and minorization condition that satisfies the requirements in theorem 3.8 has $\rho_{mnd} \to 1$ as $p \to \infty$ exponentially.

Proof. Let C be the small set. In order for $\pi(C) \geq 1/2$ and be the smallest set, C must be a hypersphere with radius of $\sqrt{m_p}$ where m_p is the median of the χ^2_p distribution. Let X_n be the Markov chain . The invariant stationary distribution of $\pi(\cdot)$ is $N(0, I_p)$. So we solve for r,

$$\begin{aligned} \frac{1}{2} &\leq P(X_n \in B_r(0_p)) \\ &= P(\sum_{i=1}^{p} X_i^2 \leq r^2) \\ &= P(\chi^2_p \leq r^2) \end{aligned}$$

For the above to hold, $r = \sqrt{m_p}$ where m_p is the median of the χ^2_p-distribution and so $diam(C) = 2\sqrt{m_p}$

Let $k(x, \cdot)$ be the density associated with $K(x, \cdot) \sim N(x/2, 3I_p/4)$

$$\begin{aligned}
&\sup_{x,y\in C} \int_{\mathbb{R}^p} |k(x, x') - k(y, x')|dx' \\
&\geq \int_{\mathbb{R}^p} |k((-\sqrt{m_p}, 0, \ldots, 0), x) - k((\sqrt{m_p}, 0, \ldots, 0), x)|d\vec{x}_p \\
&= \int_{\mathbb{R}^p} \left| \left(\sqrt{\frac{2}{3\pi}}\right)^p e^{-\frac{2}{3}((x_1-m_p/2)^2+\sum_{x=2}^p x_i^2)} - \left(\sqrt{\frac{2}{3\pi}}\right)^p e^{-\frac{2}{3}((x_1+m_p/2)^2+\sum_{x=2}^p x_i^2)} \right| d\vec{x}_p \\
&= \int_{\mathbb{R}} \left| \sqrt{\frac{2}{3\pi}} e^{-\frac{2}{3}((x_1-m_p/2)^2} - \sqrt{\frac{2}{3\pi}} e^{-\frac{2}{3}((x_1+m_p/2)^2} \right| \int_{\mathbb{R}^{p-1}} \left(\sqrt{\frac{2}{3\pi}}\right)^{p-1} e^{-\frac{2}{3}\sum_{x=2}^p x_i^2} d\vec{x}_{p-1}dx_1 \\
&= \sqrt{\frac{2}{3\pi}} \int_{\mathbb{R}} \left| e^{-\frac{2}{3}((x_1-m_p/2)^2} - e^{-\frac{2}{3}((x_1+m_p/2)^2} \right| dx_1 \\
&= 2\sqrt{\frac{2}{3\pi}} \int_{\mathbb{R}_+} e^{-\frac{2}{3}((x_1-m_p/2)^2} - e^{-\frac{2}{3}((x_1+m_p/2)^2} dx_1 \\
&= 2(P(X \geq 0) - P(Y \geq 0)) \qquad \text{where } X \sim N(\tfrac{\sqrt{m_p}}{2}, \tfrac{3}{4}) \text{ and } Y \sim N(\tfrac{-\sqrt{m_p}}{2}, \tfrac{3}{4}) \\
&= 2\left(\Phi\left(\sqrt{\frac{m_p}{3}}\right) - \Phi\left(-\sqrt{\frac{m_p}{3}}\right)\right) \\
&= 2\left(1 - 2\Phi\left(-\sqrt{\frac{m_p}{3}}\right)\right) \\
&= 2 - 4\Phi\left(-\sqrt{\frac{m_p}{3}}\right)
\end{aligned}$$

Since m_p is of order p (see figure 5), $\Phi\left(-\sqrt{\frac{m_p}{3}}\right)$ is of order e^n and so

$$\begin{aligned}
\gamma &\geq \frac{1}{2} \sup_{x,y\in C} \int_S |k(x, s) - k(y, s)|\mu(ds) && \text{By lemma 3.10} \\
&\geq \frac{1}{2}\left(2 - 4\Phi\left(-\sqrt{\frac{m_p}{3}}\right)\right) \\
&= \left(1 - 2\Phi\left(-\sqrt{\frac{m_p}{3}}\right)\right)
\end{aligned}$$

And so as $p \to \infty$, $\gamma \to 1$ at a rate e^{-n} □

In [3], Qin and Hobert try to uncover whether the Rosenthal bound (and later the Baxendale bound) is tight given the assumptions. If the bound is tight with respect to the assumptions while it is not tight with respect to ρ_*, then the bound has limitations.

Assume that $\alpha = 1$ going forward.

Definition 3.8 (Simple bound)**.** A bound is considered a simple bound if it is solely a function of the parameters in the assumption.

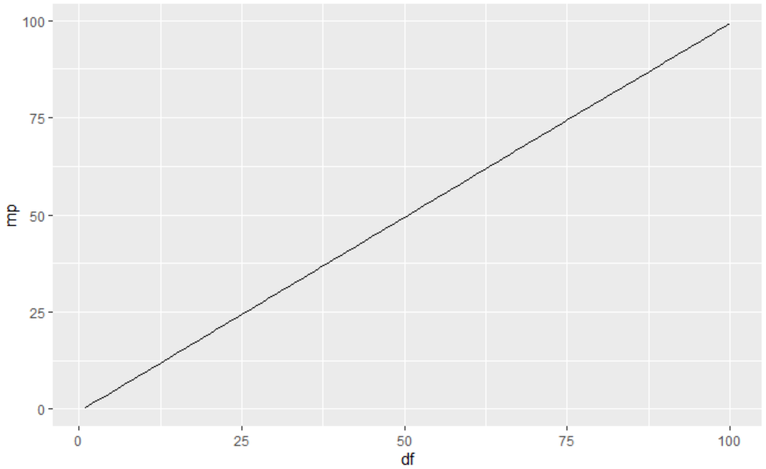

Figure 5: m_p is of order p

Rosenthal's bound is considered a simple bound, because it is a function of (γ, λ, L, d). Define

$$S_{(\gamma,\lambda,L,d)} = \{P : \text{The corresponding Markov chain has the parameters } (\gamma, \lambda, L, d)\}$$

and

$$\rho_{opt}(\gamma, \lambda, L, d) = \sup_{P \in S_{(\gamma,\lambda,L,d)}} \rho_*(P)$$

Then Rosenthal's bound is tight with respect to to the assumptions if, $\rho_{opt}(\gamma, \lambda, L, d)$ is close to $\rho_{rose}(\gamma, \lambda, L, d)$.

Naturally,

$$\rho_{rose}(\gamma, \lambda, L, d) \leq \rho_{opt}(\gamma, \lambda, L, d)$$

, but we want to know how well does $\rho_{opt}(\gamma, \lambda, L, d)$ estimate $\rho_{rose}(\gamma, \lambda, L, d)$.

Proposition 3.11 (Proposition 16 in [3])**.** *For a given* (γ, λ, L, d) *in Rosenthal's bound,*

$$\rho_{opt} \geq \gamma$$

Proof. A proof is constructed by creating a Markov chain that converges as slow as possible given the constraints of belonging to $S_{(\gamma,\lambda,L,d)}$. □

Note that $\alpha = 1$ and the authors note that the limitations shown here do not apply for when $\alpha > 1$ (or growing with respect to p).

Also note that a Markov chain can have varying (γ, λ, L, d) and so the sets of Markov chain , $S_{(\gamma,\lambda,L,d)}$, are not disjoint.

3.2.2 Baxendale's bound

The results come from [3].

This section is to show that Baxendale's bound is tight with respect to the theorems assumptions. This implies that any gap in the bound is due to variation within the Markov chain that satisfy the assumption parameters and not the bound itself.

Theorem 3.12 (Baxendale's bound, found in [3] as theorem 1)**.** *Suppose that*

1. *there exists $\lambda \in (0,1)$, $L < \infty$, a function $V : S \to [1,\infty)]$, and a small set C such that for all $x \in S$,*
$$E[V(X_{n+1})|X_n] \leq \lambda V(X_n) I_{C^c}(X_n) + L I_C(X_n)$$
2. *there exists $\epsilon \in [0,1)$ and a probability measure $\upsilon : \mathcal{F} \to [0,1]$ such that for every $x \in C$ and for all $A \in \mathcal{F}$,*
$$P(x,A) \geq \epsilon \upsilon(A)$$
. (ie. the minorization condition holds)
3. *There exists a β such that $\upsilon(C) \geq \beta$*
4. *The chain is reversible*
5. *The chain is non-negative definite*

Then for all $x \in S$, $m \geq 0$ and $a \in (0,1)$,

$$\rho_*(P) \leq \max\{\lambda, (1-\epsilon)^{1/\alpha_*}\} I_{\epsilon<1} + \lambda I_{\epsilon=1}$$

where

$$\alpha_* = \frac{\log[(L-\epsilon)/(1-\epsilon)] + \log \lambda^{-1}}{\log \lambda^{-1}}$$

Baxendale's bound is also a simple bound. Define

$$S_{(\epsilon,\lambda,L,\beta)} = \{P : \text{The corresponding Markov chain has the parameters } (\epsilon,\lambda,L,\beta)\}$$

and

$$\rho_{opt}(\epsilon,\lambda,L,\beta) = \sup_{P \in S_{(\epsilon,\lambda,L,\beta)}} \rho_*(P)$$

Then Baxendale's bound is tight with respect to to the assumptions as shown in the following proposition.

Proposition 3.13 (Theorem 8 in [3])**.** *For a given $(\epsilon,\lambda,L,\beta)$ in Baxendale's bound,*

$$\rho_{opt} \geq \max\{\lambda, (1-\epsilon)^{1/\alpha}\} I_{\epsilon<1} + \lambda I_{\epsilon=1}$$

where $\alpha = Int\{\alpha_*\}$

Proof. Constructed a Markov chain , P, that belongs to $S_{(\epsilon,\lambda,L,\beta)}$ with $\rho_*(P) = \max\{\lambda, (1-\epsilon)^{1/\alpha}\} I_{\epsilon<1} + \lambda I_{\epsilon=1}$. □

This indicates that the Baxendale bound is good given the assumptions. This shows the limitation of only using $(\epsilon,\lambda,L,\beta)$ in the bound.

Qin and Hobert also propose and optimal bound that is only a function of ϵ and C indicating the limitations of the minorization condition (see corollary 12 in [3]).

3.3 Bounds using Wasserstein distance

Definition 3.9 (Wasserstein distance)**.** Let $\mathcal{P}(S)$ be the set of all probability measures in S. The Wasserstein distance with respect to a measure ψ between two probability measures, $\mu, \upsilon \in \mathcal{P}(S)$ is,

$$W_\psi(\mu, \upsilon) = \inf_{\xi \in \tau(\mu,\upsilon)} \int_{S \times S} \psi(x, y)\xi(dx, dy)$$

Where $\tau(\mu, \upsilon)$ represents the set of all couplings. That is,

$$\tau(\mu, \upsilon) = \{\xi \in \mathcal{P}(S \times S) : \forall A \in \mathcal{F}\xi(A, S) = \mu(A) \text{ and } \xi(S, A) = \upsilon(A)\}$$

or

$$\tau(\mu, \upsilon) = \{L(X, Y) : L(X) = \mu, L(Y) = \upsilon\}$$

An alternative definition of the Wasserstein distance [**?**]: Suppose that $L(X) = \mu, L(Y) = \upsilon$ then,

$$W_\psi(\mu, \upsilon) = \inf E[\psi(X', Y')] \text{ where } (X', Y') \text{ is a coupling of } (X, Y)$$

The best way to interpret the Wasserstein distance when $\psi(x, y) = |x - y|$ is via the optimal transport approach, which addresses the question: what is the most efficient way of moving 1 unit of dirt from one distribution to another?

For example in figure 6 to get μ from υ, the area in region B needs to be moved to the area in region C. Suppose that the area in region B is filled with little stickers that aren't overlapping and we want to move these little stickers to cover region C in the most efficient way possible ie, the minimum distance spent carrying one sticker from region B to C. How can this be done? The sticker located at the tail ends of the sets B will be moved to the tail ends of C and the sticker located at the centers of B will be moved to the tip of C. All of the area in the region of A will not move. In this example, the Wasserstein distance is the average distance to move the area from B to the area of C.

The areas of A,B, and C can be defined with random variables, X_A, X_B, X_C where $L(X_i)$ is the normalized region of i. Then there exists $a, b \in [0, 1]$ such that $X = aX_A + bX_B$, $Y = aX_A + bX_C$ and $L(X) = \mu, L(Y) = \upsilon$. The question can then be asked as follows: what is the relationship between X_B and X_C?

Lemma 3.14. *If* $\lim_{n\to\infty} W_{|\cdot|}(\mu_n, \mu) = 0$ *then* $\mu_n \to \mu$ *in distribution.*

Wasserstein distance measures distance according to the probabilistic distance of the elements in S. The focus is as much on the distance between the set over which the measures are defined as the measures themselves. In contrast, total variation distance, doesn't seem to care about where the set of each distribution that doesn't overlap is located.

In figure 7 the measurement of the Wasserstein and TV is compared. Example 1 represents two uniform random variables, $U(0, 1)$, $U(1.1, 2.1)$. the area over which the two distributions are positive does not overlap. The Wasserstein distance is 1.1 (Let X be a r.v. s.t. $L(X) = U(0, 1)$ and $Y = X + 1.1$, then $E[|X - Y|] = 1.1$) while the TV is 1 ($|\mu(B) - \upsilon(B)| = |0 - 1| = 1$). The distances are close.

Example 2 represents two other distributions that overlap except that υ can be constructed so that the tip of density μ is clipped off and moved at least 100 units away. The Wasserstein distance is greater than $100\mu(B)$ while the TV distance is $\mu(B)$.

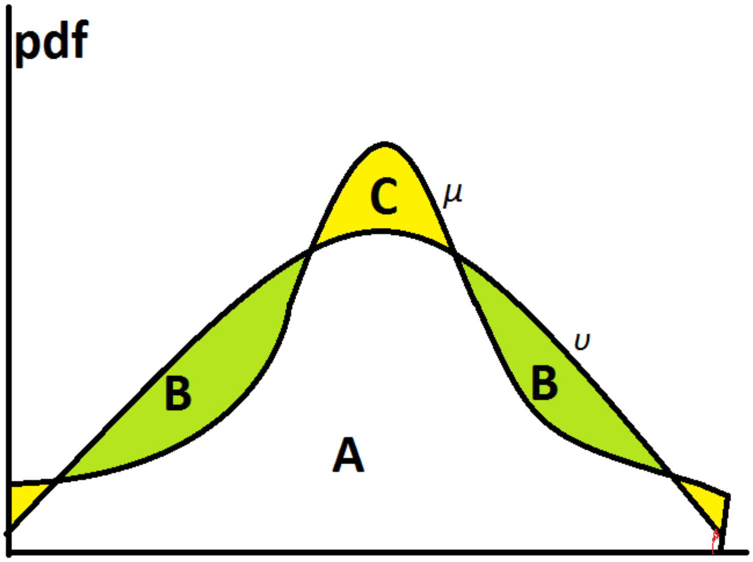

Figure 6: Visual of the Wasserstein distance between μ and υ

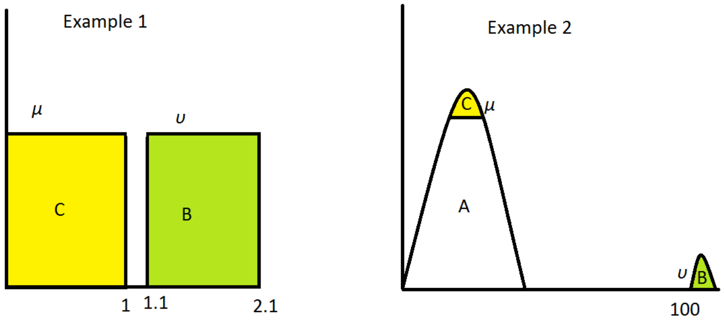

Figure 7: Contrast between two distribution distances

In order for the Wasserstein distance to bound the TV distance we must therefore take into account the relative difference between the sets that do not overlap.

Theorem 3.15 (Theorem 2 in [7], proposition 11 in [5])**.** *Let the density function of the transition probability $K(x, \cdot)$ be $k(x, \cdot)$ with dominating measure μ. If there exists a $C < \infty$ such that for all $x, y \in S$,*

$$\int_S |k(x, x') - k(y, x')| d\mu x' \leq C\psi(x, y) \tag{8}$$

Then,

$$||K^m(x, \cdot) - \pi||_{TV} \leq \frac{C}{2} W_\psi(K_x^{m-1}, \pi)$$

We would further like to apply a geometric bound to the Wasserstein distance.

Theorem 3.16 (Proposition 1 in [7], proposition 6 in [5])**.** *If $c(x) = \int_S \psi(x, y)K(x, dy) < \infty \forall x \in S$ and for a random mapping $f(x) \sim K(x, \cdot)$ there exists a $\gamma < 1$ such that $\forall x, y \in S$,*

$$E[\psi(f(x), f(y))|] \leq \gamma\psi(x, y) \tag{9}$$

then for each $x \in S, m \in \mathbb{N}$,

$$W_\psi(K^m(x, \cdot), \pi) \leq \frac{c(x)}{1 - \gamma}\gamma^m$$

That is, the Markov chain is geometrically ergodic.

We can put theorem 3.15 and 3.16 together by finding a measure ψ that satisfies both theorems in which case,

$$||K^m(x, \cdot) - \pi||_{TV} \leq \frac{C}{2} W_\psi(K_x^{m-1}, \pi) \leq \frac{C}{2} \frac{c(x)}{1 - \gamma}\gamma^m$$

The following lemma will help find an upper bound for equation 9 in theorem 3.16, which can be difficult to find in practise.

Lemma 3.17. *If S is a convex subset of a Euclidean space and $\psi(x, y) = ||x - y||$ where $|| \cdot ||$ is a norm and f is a differentiable function then,*

$$E[||f(x) - f(y)||] \leq \sup_{t \in [0,1]} E\left[\frac{df(x + t(y - x))}{dt}\right]$$

and so, the requirement, equation 9, in theorem 3.16 can be replaced by

$$\sup_{t \in [0,1]} E\left[\frac{df(x + t(y - x))}{dt}\right] \leq \gamma||x - y||$$

Proof. Given the assumptions above,

$$\begin{aligned} E[||f(x)-f(y)||] &= E\left[||\int_0^1 \frac{df(x+t(y-x))}{dt}dt||\right] \\ &\leq E\left[\int_0^1 ||\frac{df(x+t(y-x))}{dt}||dt\right] \\ &\leq \int_0^1 E\left[||\frac{df(x+t(y-x))}{dt}||\right]dt \qquad \text{since } \frac{df(x+t(y-x))}{dt} \text{ is bounded} \\ &\leq \sup_{t\in[0,1]} E\left[||\frac{df(x+t(y-x))}{dt}||\right] \end{aligned}$$

□

Example 3.2 (Continuation of 3.1)**.** The example was one where is it known that $\rho_* = 1/2$, but for any drift and minorization condition that satisifies the requirements in the theorem 3.8, $\rho_{mnd} \to 1$ as $p \to \infty$.

Applying 3.15 and 3.16 will lead to the exact bound, $\rho_{was} = \rho_* = 1/2$.

Let the metric be $\psi(x,y) = ||x-y||_2$ and the random mapping be $f(\vec{X}_n) = \vec{X}_n/2 + 3/4\vec{Z}_p$. Then,

$$c(\vec{x}) = \int_{\vec{y}\in\mathbb{R}^p} ||\vec{x}-\vec{y}||K(x,dy) < \infty \forall x \in \mathbb{R} \tag{10}$$

since the normal has finite first and second moments and

$$E\left[||f(\vec{X}_n) - f(\vec{Y}_n)||\right] = E\left[||\vec{X}_n/2 + 3/4\vec{Z}_p - \vec{Y}_n/2 - 3/4\vec{Z}_p||\right] \tag{11}$$

$$= \frac{1}{2}E\left[||\vec{X}_n - \vec{Y}_n||\right] \tag{12}$$

Thus, by 3.16

$$d_{W,\psi}(K^m(\vec{X}_n,\cdot),\pi(\cdot)) \leq \frac{c(\vec{X}_n)}{1-\gamma}\gamma^m \tag{13}$$

Further, by lemma 3.15 since,

$$\begin{aligned} \int_S |k(x,t)-k(y,t)|dt &= 2\left|P(X \leq \frac{a(X+Y)}{2}) - P(Y \leq \frac{a(X+Y)}{2})\right| \\ &\qquad \text{Where } X \sim N(x/2, 3/4) \text{ and } Y \sim N(y/2, 3/4) \\ &= 2\left|\Phi\left(\frac{x-y}{\sqrt{3}}\right) - \Phi\left(\frac{y-x}{\sqrt{3}}\right)\right| \end{aligned}$$

This is indeed bounded by a constant multiple of $||x-y||_2$ for all p. This is known with computer calculations. The following ratio was calculated to see whether it was bounded by a constant.

$$Ratio = \frac{\int_S |k(x,t)-k(y,t)|dt}{||x-y||_2}$$

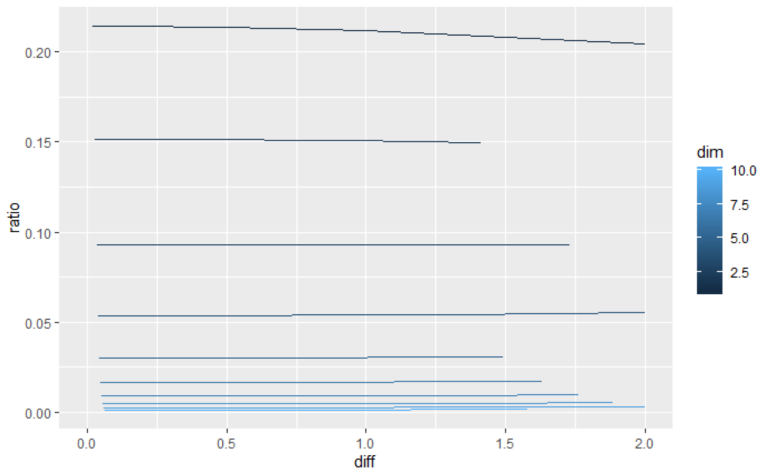

Figure 8:

It is only necessary to find the ratio when $||x-y||_2 \leq 2$ since $\int_S |k(x,t)-k(y,t)|dt \leq 2$. Figure 8 shows that the ratio is indeed bounded for $||x-y||_2 \leq 2$ and so equation 8 is indeed bounded in theorem 3.15 for any value p (in the figure, $p = dim$).

So, bounding the total variation distance with the Wasserstein distance, which is then bounded by a function results in the following bound,

$$K^m(x,\cdot) - \pi||_{TV} \leq \frac{C}{2}\frac{c(x)}{1-\gamma}\gamma^m$$

We can further simplify this. We know from equation 11 that $\gamma = 1/2$ and from equation 10 that $c(x)$ while increase at a rate p when the dimension increases and by the figure 8, C decreases at a rate of e^{-p} (this is all sort of winging it. May not be true) and so since γ remains constant, using this approach, we know that $\rho_* \leq 1/2$. Since $\rho_* = 1/2$ the upper bound estimates the convergence rate perfectly.

Next a novel bound for the Wasserstein distance stated by Hobert and Qin [4] is shown. It uses a drift and contraction assumption.

Theorem 3.18 (Theorem 2.5 of [4])**.** *Let ψ be a metric over which we want the estimate the Wasserstein distance. Assume the following for a Markov chain with transition kernel P on a Polish metric space $(S, \psi, \mathcal{B})$.*

1. *(Generalized drift condition) There exists a measurable function $V : S \to \mathbb{R}_+$ and an $a \in (0,\infty)$ such that $PV(x) < \infty$ for all $x \in S$ and*

$$\frac{\psi(x,y)}{a} \leq V(x) + V(y) + 1, (x,y) \in S \times S \tag{14}$$

2. *(Generalized contraction condition) There exists a measurable function $\Gamma : S \times S \to \mathbb{R}_+$ such that for each $(x,y) \in S \times S$,*

$$W_\psi(\delta_x P, \delta_y P) \leq \Gamma(x,y)\psi(x,y) \tag{15}$$

where $\delta_x = x$ with probability 1.

3. *(Restriction on Γ) There exists a $\Lambda : S \times S \to \mathbb{R}_+$ such that*

$$\Lambda(x, y) \geq \frac{PV(x) + PV(y) + 1}{V(x) + V(y) + 1} \tag{16}$$

and there exists a $r \in (0, 1)$ such that

$$\rho_r = \sup_{(x,y)\in S\times S} \Gamma(x, y)^r \Lambda(x, y)^{1-r} < 1 \tag{17}$$

Then there exists a stationary distribution π such that for every r that satisfies 17 and $x \in S$

$$W_\psi(\delta_x P^n, \pi) \leq a \left(\frac{PV(x) + V(x) + 1}{1 - \rho_r} \rho_r^n \right) \tag{18}$$

Proof. The proof shows that the sequence $\{\delta_x P^n\}_{n\geq 1}$ is a Cauchy sequence with an upper bound as in 18 in the Polish metric space $(S, \psi, \mathcal{B})$. Since P is defined on a Polish metric space it is assumed to be complete and thus there exists a π such that $W_\psi(\delta_x P^n, \pi)$ is bounded by the function in 18. □

3.4 Shift coupling

Another method to study is shift coupling.

Theorem 3.19 (Proposition 1 in [8])**.** *Suppose that X_k, X_k' are two Markov chains with the same transition probability. Let T, T' be the corresponding stopping times where $X_T = X_{T'}'$ We have that,*

$$|| \sum_{k=1}^{N} P(X_k \in A) - \sum_{k=1}^{N} P(X_k' \in A)|_{TV} \leq E[\min\{\max\{T, T'\}, N\}]$$

The proof was written by me, but probably resembles the original proof.

$$||\sum_{k=1}^{N} P(X_k \in A) - \sum_{k=1}^{N} P(X'_k \in A)|\_TV$$

$$= \sup_A \left| \sum_{k=1}^{N} P(X_k \in A) - \sum_{k=1}^{N} P(X'_k \in A) \right|$$

$$= \sup_A \left| \sum_{k=1}^{N} \left(\sum_{n=1}^{\infty} P(X_k \in A \cap \max\{T, T'\} = n) - \sum_{n=1}^{\infty} P(X'_k \in A \cap \max\{T, T'\} = n) \right) \right|$$

If $k > n$ then $P(X_k \in A) = P(X_{k+T'-T} \in A)$

$$= \sup_A \left| \sum_{n=1}^{N} \left(\sum_{k \leq n} P(X_k \in A \cap \max\{T, T'\} = n) - \sum_{k \leq n+T'-T} P(X'_{k+T-T'} \in A \cap \max\{T, T'\} = n) \right) \right|$$

$$= \sup_A \left| \sum_{k=1}^{N} P(X_k \in A \cap \max\{T, T'\} \geq k) - P(X'_{k+T-T'} \in A \cap \max\{T, T'\} \geq k + T' - T) \right|$$

WLOG $T' \leq T$ (the alternative can easily be proven by symmetry)

$$\leq \sup_A \left| \sum_{k=1}^{N} P(X_k \in A \cap \max\{T, T'\} \geq k) - P(X'_{k+T-T'} \in A \cap \max\{T, T'\} \geq k) \right|$$

$$\leq \sum_{k=1}^{N} P(\max\{T, T'\} \geq k)$$

$$= \sum_{k=1}^{\infty} P(\min\{\max\{T, T'\}, N\} \geq k)$$

$$= E[\min\{\max\{T, T'\}, N\}]$$

□

References

[1] J. S. Rosenthal, ”Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo,” *Journal of the American Statistical Association,* vol. 90, no. 430, pp. 558-566, Jun. 1995, doi: 10.2307/2291067.

[2] G. O. Roberts and J. S. Rosenthal, “General State Space Markov Chains and MCMC Algorithms,” *Probability Surveys,* vol. 1, no. 1, pp. 20–71, Nov. 2004, doi: 10.1214/154957804100000024.

[3] Q. Qin and J. P. Hobert, ”On the Limitations of Single Step Drift and Minorization on Markov Chain Convergence Analysis,” 2020, *arXiv:2003.09555.*

[4] Q. Qin and J. Hobert, ”Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions,” 2019, *arXiv:1902.02964.*

[5] Q. Qin and J. P. Hobert, "Wasserstein-Based Methods for Convergence Complexity Analysis of MCMC with Applications," 2018, *arXiv:1810.08826*.

[6] D. Jerison, "The Drift and Minorization Method for Reversible Markov Chains," PhD thesis, Department of Mathematics, Stanford University, Stanford, California, United States, 2016.

[7] B. Davis and J. Hobert, "On the convergence complexity of Gibbs samplers for a family of simply Bayesian random effects models," 2020, *arXiv:2004.14330*.

[8] G. O. Roberts, J. S. Rosenthal, "Shift-Coupling and Convergence Rates of Ergodic Averages," *Communications in Statistics. Stochastic Models,* vol. 13, no. 1, pp. 147-165, 1997, doi: 10.1080/15326349708807418.