RANDOM WALKS AND REPRESENTATIONS

JOHN PIKE

These lecture notes were originally written for MATH 484 at Boğaziçi University in the Summer semester of 2022. They are for personal educational use only and are not to be published or redistributed. Much of the material is taken directly from *Markov Chains and Mixing Times* by David Levin, Yuval Peres, and Elizabeth Wilmer and *Group Representations in Probability and Statistics* by Persi Diaconis. These notes likely contain typos and mistakes. All such errors are mine and corrections are greatly appreciated.

Solutions to the homework exercises appearing at the end of these notes are available upon request.

Day 1: Through Theorem 1.1

Day 2: Through Theorem 1.2

Day 3: Finished Subsection 1.4

Day 4: Through Example 1.9

Day 5: Through Example 2.1

Day 6: Through Example 2.2

Day 7: Finished Section 2

Day 8: Through Proposition 3.1

Day 9: Through Corollary 3.2

Day 10: Exam 1

Day 11: Review; Through Proposition 3.6

Day 12: Finished Section 3

Day 13: Finished Subsubsection 4.2.1

Day 14: To Example 4.3

Day 15: Finished Subsection 4.2.3 (some computations omitted)

Day 16: Through cutoff examples (skipped Subsubsection 4.2.4)

Day 17: Finished Section 5.2

Day 18: Finished Section 5

Day 19: Exam 2

1 Markov Chains

1.1 Definition and First Properties

Suppose that $X_0, X_1, X_2, \ldots$ is a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in a countable set $\mathcal{S}$.

We say that $\{X_k\}_{k=0}^{\infty}$ is a *Markov chain* with *state space* $\mathcal{S}$ if for every $n \in \mathbb{N}_0$ and every $x_0, x_1, \ldots, x_{n+1} \in \mathcal{S}$ with $\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) > 0$,

$$\mathbb{P}(X_{n+1} = x_{n+1} \,|X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} \,|X_n = x_n).$$

(Here we are using the standard notation $\mathbb{P}(A\,|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ for $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, and $\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$.)

In words, Markov chains are forgetful; the future depends on the past only through the present.

We take it as implicit in the definition that the time index is discrete. Also, we will typically restrict our attention to finite state spaces.
Finally, we will assume that our chains are *temporally homogeneous*, so that

$$\mathbb{P}(X_{n+1} = y \,|X_n = x) = \mathbb{P}(X_1 = y \,|X_0 = x).$$

for all $x, y \in \mathcal{S}$, $n \in \mathbb{N}_0$.

The dynamics of the Markov chain are thus governed by the *transition function* $p : \mathcal{S} \times \mathcal{S} \to [0, 1]$ given by $p(x, y) = \mathbb{P}(X_1 = y \,|X_0 = x)$.

The following result shows that the finite dimensional distributions of $\{X_k\}_{k=0}^{\infty}$ are completely determined by the *initial distribution* $\mu_0 = \mathscr{L}(X_0)$ and the transition function. For this reason, we often talk of the chain p without reference to any particular sequence of random variables.

Proposition 1.1. *For any $n \in \mathbb{N}$, $x_0, x_1, \ldots, x_n \in \mathcal{S}$,*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \mu_0(x_0)p(x_0, x_1)p(x_1, x_2)\cdots p(x_{n-1}, x_n).$$

Proof. We first observe that

$$\mathbb{P}(X_0 = x_0, X_1 = x_1) = \mathbb{P}(X_0 = x_0)\,\mathbb{P}(X_1 = x_1 \,|X_0 = x_0) = \mu_0(x_0)p(x_0, x_1).$$

Now suppose that $\mathbb{P}(X_0 = x_0, \ldots, X_k = x_k) = \mu_0(x_0)p(x_0, x_1)\cdots p(x_{k-1}, x_k)$. If this probability is 0, then

$$\mathbb{P}(X_0 = x_0, \ldots, X_k = x_k, X_{k+1} = x_{k+1}) = 0 = \mu_0(x_0)p(x_0, x_1)\cdots p(x_{k-1}, x_k)p(x_k, x_{k+1}).$$

Otherwise, the Markov property and induction hypothesis give

$$\begin{aligned}\mathbb{P}(X_0 = x_0, \ldots, X_{k+1} = x_{k+1}) &= \mathbb{P}(X_0 = x_0, \ldots, X_k = x_k)\,\mathbb{P}(X_{k+1} = x_{k+1} \,|X_0 = x_0, \ldots, X_k = x_k) \\ &= \mathbb{P}(X_0 = x_0, \ldots, X_k = x_k)\,\mathbb{P}(X_{k+1} = x_{k+1} \,|X_k = x_k) \\ &= \mu_0(x_0)p(x_0, x_1)\cdots p(x_{k-1}, x_k)p(x_k, x_{k+1}).\end{aligned}$$

□

An immediate consequence of this result is that

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \,|X_0 = x_0) = p(x_0, x_1)\cdots p(x_{n-1}, x_n).$$

Accordingly, if $\mathbb{P}(X_0 = x_0, \ldots, X_m = x_m) > 0$, then for any $n \in \mathbb{N}$, $x \in \mathcal{S}$,

$$\begin{aligned}
&\mathbb{P}(X_{m+n} = x \mid X_0 = x_0, \ldots, X_m = x_m) \\
&\quad = \sum_{\mathbf{y} \in \mathcal{S}^{n-1}} \mathbb{P}(X_{m+1} = y_1, \ldots, X_{m+n-1} = y_{n-1}, X_{m+n} = x \mid X_0 = x_0, \ldots, X_m = x_m) \\
&\quad = \sum_{\mathbf{y} \in \mathcal{S}^{n-1}} \frac{\mathbb{P}(X_0 = x_0, \ldots, X_{m+n-1} = y_{n-1}, X_{m+n} = x)}{\mathbb{P}(X_0 = x_0, \ldots, X_m = x_m)} \\
&\quad = \sum_{\mathbf{y} \in \mathcal{S}^{n-1}} \frac{\mu_0(x_0)p(x_0, x_1) \cdots p(x_m, y_1) \cdots p(y_{n-1}, x)}{\mu_0(x_0)p(x_0, x_1) \cdots p(x_{m-1}, x_m)} \\
&\quad = \sum_{\mathbf{y} \in \mathcal{S}^{n-1}} p(x_m, y_1) \cdots p(y_{n-1}, x) \\
&\quad = \sum_{\mathbf{y} \in \mathcal{S}^{n-1}} \mathbb{P}(X_1 = y_1, \ldots, X_{n-1} = y_{n-1}, X_n = x \mid X_0 = x_m) \\
&\quad = \mathbb{P}(X_n = x \mid X_0 = x_m).
\end{aligned}$$

This shows that the Markov property is not limited to a single time step and the process 'starts afresh' whenever we observe its current state.

When $\mathcal{S} = \{s_1, \ldots, s_N\}$, we can encode the transition function in the $N \times N$ matrix P with (i, j)-entry $P(i, j) = p(s_i, s_j)$.

By construction, P is a *stochastic matrix*—that is, $P(i, j) \geq 0$ and $\sum_{k=1}^{N} P(i, k) = 1$ for all $i, j \in [N]$.

We will often forego enumerating the state space and just index the rows and columns of the matrix by elements of $\mathcal{S}$ so that $P(x, y) = p(x, y)$.

If we represent the distribution of X_k with the row vector

$$\mu_k = \begin{bmatrix} \mu_k(1) & \cdots & \mu_k(N) \end{bmatrix} = \begin{bmatrix} \mathbb{P}(X_k = s_1) & \cdots & \mathbb{P}(X_k = s_N) \end{bmatrix},$$

then we see that X_{k+1} has distribution

$$\mu_{k+1}(j) = \mathbb{P}(X_{k+1} = s_j) = \sum_{i=1}^{N} \mathbb{P}(X_k = s_i)\, \mathbb{P}(X_{k+1} = s_j \mid X_k = s_i) = \sum_{i=1}^{N} \mu_k(i) P(i, j).$$

In vector form, we have $\mu_{k+1} = \mu_k P$, and thus by induction, $\mu_n = \mu_0 P^n$ for $n \in \mathbb{N}$.

Example 1.1. Any Markov chain on $\mathcal{S} = \{s_1, s_2\}$ is described by the transition matrix $P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$ for some $p, q \in [0, 1]$. Here p is the probability of transitioning from s_1 to s_2 in one time step and q is the probability of transitioning from s_2 to s_1. Assume $0 < p + q < 2$ to avoid trivialities.

If the initial state is $X_0 = s_2$, then the distribution of X_1 is

$$\mu_1 = \mu_0 P = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} = \begin{bmatrix} q & 1-q \end{bmatrix},$$

and the distribution of X_2 is

$$\mu_2 = \mu_0 P^2 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} (1-p)^2 + pq & (1-p)p + p(1-q) \\ q(1-p) + (1-q)q & qp + (1-q)^2 \end{bmatrix} = \begin{bmatrix} q(1-p) + (1-q)q & qp + (1-q)^2 \end{bmatrix}.$$

This makes sense because in order for the chain started at s_2 to be at s_1 after two time steps, say, it must have either transitioned from s_2 to s_1 to s_1, which happens with probability $q(1-p)$, or from s_2 to s_2 to s_1, which happens with probability $(1-q)q$.

To compute the distribution of X_n for an arbitrary initial distribution $\mu_0 = \begin{bmatrix} r & 1-r \end{bmatrix}$, $r \in [0,1]$, we observe that P has characteristic polynomial

$$\varphi(\lambda) = \det(\lambda I - P) = \lambda^2 + (p+q-2)\lambda + (1-p-q),$$

so the quadratic formula gives its eigenvalues as $\lambda_1 = 1$ and $\lambda_2 = 1-p-q$.

Solving the homogeneous equations $(\lambda_k I - P)\mathbf{x} = \mathbf{0}$ gives the corresponding eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -p \\ q \end{bmatrix}$, so

$$P^n = \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1-p-q \end{bmatrix}^n \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix}^{-1} = \frac{1}{p+q} \begin{bmatrix} q+p(1-p-q)^n & p-p(1-p-q)^n \\ q-q(1-p-q)^n & p+q(1-p-q)^n \end{bmatrix},$$

and thus

$$\mu_n = \begin{bmatrix} r & 1-r \end{bmatrix} P^n = \begin{bmatrix} \frac{q-[q-r(p+q)](1-p-q)^n}{p+q} & \frac{p+[q-r(p+q)](1-p-q)^n}{p+q} \end{bmatrix}.$$

In particular, $\lim_{n\to\infty} \mu_n = \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix}$, which is independent of r.

One readily checks that this limiting probability distribution, $\pi = \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix}$, satisfies $\pi P = \pi$, so π is a left eigenvector of P with eigenvalue 1. We say that π is the *stationary distribution* of P.

We are often interested in fixing a transition matrix P and varying the initial distribution, so it is convenient to write $\mathbb{P}_\mu$ for the law of the chain with $\mu_0 = \mu$:

$$\mathbb{P}_\mu(X_n = y) = \mathbb{P}(X_n = y \,|X_0 \sim \mu) = (\mu P^n)(y).$$

When the chain is started at a fixed state x, we abuse notation and write $\mathbb{P}_x$ for $\mathbb{P}_{\delta_x}$, so that

$$\mathbb{P}_x(X_n = y) = \mathbb{P}(X_n = y \,|X_0 = x) = (\delta_x P^n)(y) = P^n(x,y),$$

the entry in row x and column y of P^n.

Note that for any distribution μ, we have $\mathbb{P}_\mu(X_n = y) = \sum_{x\in\mathcal{S}} \mu(x)\mathbb{P}_x(X_n = y)$.

Whereas our construction thus far has necessitated that probability distributions on $\mathcal{S}$ be represented as row vectors with right multiplication by P corresponding to updating the distribution by one time step, we can think of column vectors as representing functions $f : \mathcal{S} \to \mathbb{R}$, the i^{th} coordinate being interpreted as $f(s_i)$.

In this view, we have

$$(Pf)(x) = \sum_{y\in\mathcal{S}} P(x,y)f(y) = \sum_{y\in\mathcal{S}} \mathbb{P}_x(X_1 = y)f(y) = \mathbb{E}[f(X_1) \,|X_0 = x].$$

As with our parameterization of probabilities by initial distributions, this conditional expectation will be denoted $\mathbb{E}_x[f(X_1)]$, and in general we write

$$\mathbb{E}_\mu[f(X_n)] = \sum_{y\in\mathcal{S}} \mathbb{P}_\mu(X_n = y)f(y) = \sum_{x,y\in\mathcal{S}} \mu(x)P^n(x,y)f(y) = \mu P^n f.$$

Remark 1.1. If the state space is countably infinite, we can no longer interpret the transition function as a matrix, but the multi-step transition probabilities are computed similarly. Namely, if we define $p^n(x,y) = \mathbb{P}(X_n = y \,|X_0 = x)$, then we have the *Chapman-Kolmogorov equation*

$$\begin{aligned} p^{m+n}(x,z) &= \sum_{y\in\mathcal{S}} \mathbb{P}(X_{m+n} = z, X_m = y \,|X_0 = x) \\ &= \sum_{y\in\mathcal{S}} \frac{\mathbb{P}(X_{m+n} = z, X_m = y, X_0 = x)}{\mathbb{P}(X_m = y, X_0 = x)} \cdot \frac{\mathbb{P}(X_m = y, X_0 = x)}{\mathbb{P}(X_0 = x)} \\ &= \sum_{y\in\mathcal{S}} \mathbb{P}(X_{m+n} = z \,|X_m = y, X_0 = x) \cdot \mathbb{P}(X_m = y \,|X_0 = x) \\ &= \sum_{y\in\mathcal{S}} \mathbb{P}(X_{m+n} = z \,|X_m = y) \cdot \mathbb{P}(X_m = y \,|X_0 = x) = \sum_{y\in\mathcal{S}} p^m(x,y)p^n(y,z). \end{aligned}$$

(We are implicitly summing over those $y \in \mathcal{S}$ for which $\mathbb{P}(X_{m+n} = z, X_m = y \,|X_0 = x) \neq 0$ so the conditional probabilities are defined.)

In particular, the update rule is $p^{n+1}(x,z) = \sum_{y\in\mathcal{S}} p^n(x,y)p(y,z)$, which is a sort of infinite matrix multiplication.

1.2 Random Walks

While transition functions/matrices are often convenient computational devices, the perspective they afford is somewhat lacking in terms of intuition. Markov chains typically arise from some kind of story that involves updating states on the basis of the present location and some random input. We are wandering around with no memory of the past, guided by constant injections of randomness.

Example 1.2. Simple symmetric random walk on the n-cycle is a Markov chain with state space $\mathcal{S} = \{0, 1, \ldots, n-1\}$ and transition matrix

$$P(i,j) = \begin{cases} \frac{1}{2}, & j \equiv i+1 \,(\mathrm{mod}\, n) \\ \frac{1}{2}, & j \equiv i-1 \,(\mathrm{mod}\, n) \\ 0, & \text{else} \end{cases}.$$

The idea is that n points are arranged in a circle and at each time step we flip a fair coin and move one step clockwise if the coin lands heads and one step counterclockwise if the coin lands tails.
This can be formalized by letting $Z_1, Z_2, \ldots$ be i.i.d. $\mathrm{Unif}(\{-1,1\})$ and defining $X_n = X_{n-1} + Z_n \,(\mathrm{mod}\, n)$.
(Viewing the states as elements of $\mathbb{Z}/n\mathbb{Z}$, we have $X_n = X_0 + \sum_{i=1}^n Z_n$ and the transition matrix is given by $P(i,j) = \mathbb{P}(Z_1 = j - i)$. We will elaborate on this perspective shortly.)

A *random mapping representation* of a transition matrix P on a state space $\mathcal{S}$ is a function $f : \mathcal{S} \times \Lambda \to [0,1]$ along with Λ-valued random variable Z satisfying $\mathbb{P}(f(x,Z) = y) = P(x,y)$.

If $Z_1, Z_2, \ldots$ is a sequence of i.i.d. random variables with $Z_1 =_d Z$, then the sequence $\{X_k\}_{k=0}^\infty$ defined by $X_0 \sim \mu$ and $X_n = f(X_{n-1}, Z_n)$ for $n \in \mathbb{N}$ is easily seen to be a Markov chain with initial distribution μ and transition matrix P.

In Example 1.2, a random mapping representation is given by taking Z to be uniform on $\Lambda = \{-1,1\}$ and $f(x,z) = x + z \,(\mathrm{mod}\, n)$. The following result gives another.

Theorem 1.1. *Every finite state space Markov chain has a random mapping representation.*

Proof. Suppose that P is the transition matrix of a Markov chain on $\mathcal{S} = \{s_1, \ldots, s_N\}$. Let Z be uniformly distributed on $\Lambda = [0,1]$ and set $F_{i,k} = \sum_{j=1}^{k} P(s_i, s_j)$. If we define $f(s_i, z) = s_k$ when $F_{i,k-1} < z \le F_{i,k}$, then

$$\mathbb{P}\big(f(s_i, Z) = s_k\big) = F_{i,k} - F_{i,k-1} = P(s_i, s_k),$$

so (f, Z) is a random mapping representation of P. □

Example 1.3. Example 1.2 can be generalized significantly: Any probability μ on a finite group G induces a random walk via $P(g,h) = \mu(hg^{-1})$. The Markov chain is defined by $X_{n+1} = g_{n+1}X_n$ where $g_1, g_2, \ldots$ are chosen independently from μ. We will refer to this random walk as (G, μ).

For instance, let $G = S_n$, and let $\mu(\tau) = \binom{n}{2}^{-1} 1\{\tau = (i\,j) \text{ for some distinct } i, j \in [n]\}$ be the uniform distribution on transpositions. We can think of permutations as representing arrangements of a deck of cards: $\sigma \in S_n$ corresponds to the ordering in which $\sigma(k)$ is the label of the k^{th} card from the top. (Equivalently, the card labeled ℓ is in position $\sigma^{-1}(\ell)$.)

Left-multiplying σ by $\tau = (i\,j)$ corresponds to interchanging card i and card j, so we can think of the random walk in terms of repeatedly shuffling the deck by randomly transposing pairs of cards.

In models of card shuffling, we typically multiply on the right—so $X_{n+1} = X_n\sigma_{n+1}$, $P(\sigma, \pi) = \mu(\sigma^{-1}\pi)$—as we want shuffles to act on positions rather than labels.

For the random transposition case, right-multiplying σ by $\tau = (i\,j)$ corresponds to interchanging the cards in positions i and j.

$$\tau \circ \sigma(k) = \begin{cases} \sigma(k), & \sigma(k) \notin \{i,j\} \\ i, & \sigma(k) = j \\ j, & \sigma(k) = i \end{cases}, \qquad \sigma \circ \tau(k) = \begin{cases} \sigma(k), & k \notin \{i,j\} \\ \sigma(j), & k = i \\ \sigma(i), & k = j \end{cases}.$$

In this example, the two conventions are essentially equivalent, but there is usually a distinction.

For instance, consider shuffling by removing the top card and inserting it in a random position. Here we need to multiply on the right by permutations distributed according to $\mu(\sigma) = \frac{1}{n} 1\{\sigma = (1 \cdots k) \text{ for some } k \in [n]\}$.

Left multiplication by $(1 \cdots k)$ would correspond to replacing the card labeled k with the card labeled 1 and the card labeled j with that labeled $j+1$ for $j < k$. This requires looking at the cards.

More concretely, if the deck is initially ordered as $\pi = 361524$ (in one-line notation), right multiplication by the cycle $\sigma = (123)$ gives the ordering $\pi\sigma = 613524$, whereas left multiplication gives $\sigma\pi = 162534$.

The inverse of this "top-to-random shuffle"—namely, placing a randomly chosen card on the top of the deck—corresponds to right multiplication by $(k \cdots 1) = (1 \cdots k)^{-1}$ with k chosen uniformly from $[n]$.

Note that the left-invariant walk (having transition matrix $P(x,y) = \mu(x^{-1}y)$) transforms into the right-invariant walk driven by $\check{\mu}(g) = \mu(g^{-1})$ under the anti-automorphism $x \mapsto x^{-1}$, so it suffices to stick with one convention for developing theory and then translate the results when a particular model is better suited to the other choice. Our random walks will generally be right-invariant (so that we multiply on the left).

The designation "right-invariant" stems from the fact that $P(xg, yg) = \mu\left(yg(xg)^{-1}\right) = \mu(yx^{-1}) = P(x,y)$.

Another important class of examples are random walks on graphs.

Example 1.4. Let $G = (V, E)$ be a simple, connected, and undirected graph with finite vertex set V and edge set E. For $u, v \in V$, write $u \sim v$ if $\{u, v\} \in E$ and define $\deg(u) = |\{v \in V : v \sim u\}|$. Simple random walk on G proceeds by moving from the present vertex to a neighbor chosen uniformly at random—that is, $P(u, v) = \dfrac{1}{\deg(u)} 1\{v \sim u\}$.

More generally, suppose that $G = (V, \overrightarrow{E})$ is a directed graph (possibly containing self-loops) and let $w : \overrightarrow{E} \to (0, \infty)$. One can define a random walk on V by $P(u, v) = \dfrac{w\big((u,v)\big)}{\sum_{x:(u,x)\in\overrightarrow{E}} w\big((u,x)\big)}$.

In fact, every Markov chain on a finite state space $\mathcal{S}$ can be interpreted as a random walk on the directed graph having vertices indexed by $\mathcal{S}$, edge set $\{(u, v) : P(u, v) > 0\}$, and edge weights $w\big((u, v)\big) = P(u, v)$.

For instance, the random walk on the group G driven by μ can be viewed as the random walk on the associated Cayley graph of G with $w(g, sg) = \mu(s)$.

Example 1.5. Let $G = (\mathbb{Z}/2\mathbb{Z})^d$, and let $\mu(x) = \frac{1}{d}$ if x has exactly one coordinate equal to one, $\mu(x) = 0$ otherwise. The random walk (G, μ) is equivalent to simple random walk on the d-dimensional hypercube.

If we define $\|x\| = |\{i \in [d] : x_i = 1\}|$, then one can verify that $Y_n = \|X_n\|$ is a Markov chain. Indeed, this is equivalent to the *Ehrenfest chain* in which d balls are distributed among two urns and at each time step a ball is chosen at random and moved to the opposite urn. Y_n records the number of balls in urn 1 at time n.

Remark 1.2. One can replace "finite" with "countable" in Theorem 1.1 and the exact same argument holds. Likewise, one can define random walks on countable groups and graphs (provided that the latter have all vertices of finite degree) just as in Examples 1.3 and 1.4. However, a fair amount of the theory we will develop in this course is specific to finite state spaces, so we will mostly resist the temptation to generalize going forward.

1.3 Irreducibility and Aperiodicity

Perhaps the most interesting thing about Markov chains is that the distribution of X_n often converges to a fixed stationary distribution that does not depend on μ_0. When the state space is finite, this happens precisely when the chain is irreducible and aperiodic, so our next order of business is to unpack these terms.

First, the chain with transition matrix P is *irreducible* if for any $x, y \in \mathcal{S}$, there is a $k \in \mathbb{N}$ such that $P^k(x, y) > 0$. In other words, it is possible to transition from any state to any other in finite time.
(If we build a directed graph with vertex set $\mathcal{S}$ and edge relation $(x, y) \in \overrightarrow{E}$ if $P(x, y) > 0$, irreducibility of P is equivalent to connectedness of the graph.)

We say that y is *accessible* from x if there is an $r > 0$ with $P^r(x, y) > 0$, and we say that x *communicates* with y if $x = y$ or if x is accessible from y and y is accessible from x. The relation "communicates with" is easily seen to be an equivalence relation and thus partitions the state space into *communicating classes*. Irreducibility is equivalent to having a single communicating class. A communicating class is *closed* if it contains all states accessible from any of its members. If a chain is not irreducible, but starts in a closed communicating class, then one can just restrict the state space to that class to obtain an irreducible chain.

As an illustration, consider random walk on a finite group G driven by a measure μ. We define the *support* of μ as $\Sigma_\mu = \{g \in G : \mu(g) > 0\}$, and we say that $S \subseteq G$ *generates* G if every element of G can be expressed as a product of elements of S. (This happens precisely when S is not contained in a proper subgroup of G.)

Proposition 1.2. *The random walk (G, μ) is irreducible if and only if Σ_μ generates G.*

Proof. Suppose that the random walk is irreducible and let g be an arbitrary element of G. Then there is some $k > 0$ such that $P^k(e, g) > 0$ where e is the identity element of G. In order for this to happen, there must be some sequence $s_1, \ldots, s_k \in \Sigma_\mu$ so that $s_k \cdots s_1 = s_k \cdots s_1 e = g$. As g was arbitrary, this shows that Σ_μ generates G.

Conversely, suppose that Σ_μ generates G, and let $g, h \in G$ be arbitrary. By assumption, there exist $t_1, \ldots, t_\ell \in \Sigma_\mu$ such that $hg^{-1} = t_1 \cdots t_\ell$, hence

$$P^\ell(g, h) \geq P(g, t_\ell g) P(t_\ell g, t_{\ell-1} t_\ell g) \cdots P\left(t_2 \cdots t_\ell g, (hg^{-1})g\right) = \mu(t_\ell) \cdots \mu(t_1) > 0,$$

so the random walk is irreducible. □

Next, for any element $x \in \mathcal{S}$, let $I_x = \left\{k > 0 : P^k(x, x) > 0\right\}$ be the set of positive times for which it is possible for the chain started at x to return to x, and define the *period* of x as $d_x = \gcd(I_x)$ with the convention that $\gcd(\emptyset) = \infty$. A state x is called *aperiodic* if $d_x = 1$.

Our first observation is that all states in an irreducible Markov chain have the same period.

Proposition 1.3. *If P is irreducible, then $d_x = d_y$ for all $x, y \in \mathcal{S}$.*

Proof. Irreducibility implies that there exist $r, \ell \in \mathbb{N}$ with $P^r(x, y), P^\ell(y, x) > 0$, so, setting $m = r + \ell$, we have $P^m(x, x) \geq P^r(x, y) P^\ell(y, x) > 0$ because one way to get from x to x in m steps is to travel to y in r steps and then back to x in ℓ steps. It follows that $m \in I_x$, so $d_x | m$.

Moreover, if $n \in I_y$, then $P^{m+n}(x, x) \geq P^r(x, y) P^n(y, y) P^\ell(y, x) > 0$, so $d_x | (m + n)$ and thus $d_x | n$. As $n \in I_y$ was arbitrary, we conclude that $d_x | d_y$.

Interchanging the roles of x and y in the preceding argument shows that $d_y | d_x$ as well, hence $d_x = d_y$. □

In light of this fact, we can define the *period* of an irreducible Markov chain to be the common period of all of its states, and we say that the chain is *aperiodic* if this value is 1.

The following example illustrates how periodicity can be an obstruction to convergence.

Example 1.6. Let m be even and consider the chain on $\mathbb{Z}/m\mathbb{Z}$ with transition probabilities $P(x, x + 1) = P(x, x - 1) = 1/2$. Then after an even number of steps, the chain is guaranteed to be at a state with the same parity (even/odd) as the initial state, and it is guaranteed to have opposite parity after an odd number of steps. Thus there is no possibility that $P^k(x, y)$ converges to a nonzero value.

Since P is irreducible with finite state space, we cannot get convergence to 0 either because if we set $r(x, y) = \min\{r \in \mathbb{N} : P^r(x, y) > 0\}$, $r = \max_{x,y \in S} r(x, y)$, and $q = \min_{x,y \in S} P^{r(x,y)}(x, y)$, then for all $n \in \mathbb{N}_0$ and all $x, y \in \mathcal{S}$, there is some $n < k \leq n + r$ with $P^k(x, y) \geq q/r$. To see this, note that wherever the chain started at x ends up after n steps, it has probability at least q of visiting y within the next r steps.

That is,

$$\begin{aligned}\mathbb{P}_x\big(X_k = y \text{ for some } n < k \le n+r\big) &= \sum_{z\in\mathcal{S}} \mathbb{P}_x\big(X_k = y \text{ for some } n < k \le n+r,\, X_n = z\big)\\ &\ge \sum_{z\in\mathcal{S}} P^n(x,z)P^{r(z,y)}(z,y) \ge \sum_{z\in\mathcal{S}} P^n(x,z)q = q.\end{aligned}$$

As $\mathbb{P}_x\big(X_k = y \text{ for some } n < k \le n+r\big) \le \sum_{k=n+1}^{n+r} \mathbb{P}_x(X_k = y)$, one of the summands must have value at least q/r.

The periodicity in the example above is also exhibited by the random transposition chain mentioned in Example 1.3. The issue is that after an even number of steps the state of the chain will have the same parity as the initial state, while after an odd number of steps it will have opposite parity. The random walk hops back and forth between the cosets of $A_n \triangleleft S_n$.

Proposition 1.4. *Let G be a finite group and μ a probability on G. If the random walk (G,μ) is irreducible (so Σ_μ is not contained in a proper subgroup of G), then it is aperiodic if and only if Σ_μ is not contained in a coset of a proper normal subgroup of G.*

Proof. Suppose that $\Sigma_\mu \subseteq gN$ for some proper normal subgroup $N \triangleleft G$, and denote the order of g by $o(g) = \min\big\{k \in \mathbb{N} : g^k = e\big\}$; irreducibility implies that $g \neq e$ and thus $o(g) > 1$. Then the random walk started at e transitions between the cosets of N like $N \to gN \to g^2N \to \cdots \to g^{o(g)}N = N \to \cdots$.
(If $X_n = g^n h_n$, then $X_{n+1} = ghg^n h_n = g(g^{-1}\widetilde{h}g)g^n h_n = \widetilde{h}g^{n+1}h_n = g^{n+1}\widehat{h}h_n = g^{n+1}h_{n+1}$.)
In particular, $P^n(e,e) = 0$ if $o(g) \nmid n$, so e, and thus every element of G, has period a multiple of $o(g)$.

Conversely, suppose e is periodic with period $d > 1$, and let $N = \{h \in G : P^{kd}(e,h) > 0 \text{ for some } k \in \mathbb{N}\}$. I claim that N is a proper normal subgroup of G.

First, since G is finite, we need only show that N is closed under products to see that it is a subgroup of G. To this end, let $x, y \in N$. Then there exist $s_1, \ldots, s_{kd}, t_1, \ldots, t_{\ell d} \in \Sigma_\mu$ with $x = s_{kd}\cdots s_1 e$ and $y = t_{\ell d}\cdots t_1 e$, so $xy = s_{kd}\cdots s_1 t_{\ell d}\cdots t_1 e$ and thus $P^{(k+\ell)d}(e,xy) \ge \mu(t_1)\cdots\mu(t_{\ell d})\mu(s_1)\cdots\mu(s_{kd}) > 0$. Accordingly, $xy \in N$, hence $N \le G$.

Also, for any $s \in \Sigma_\mu$, irreducibility implies that $P^k(s,e) > 0$ for some $k \in \mathbb{N}$. It follows that $P^{k+1}(e,e) \ge P(e,s)P^k(s,e) > 0$ and thus $d|(k+1)$ since $e \in N$. This precludes s from belonging to N as we would then have $P^{\ell d+k}(e,e) \ge P^{\ell d}(e,s)P^k(s,e) > 0$, giving the contradiction that $d|(\ell d + k)$ and thus $d > 1$ divides both k and $k+1$. This shows that N is a proper subgroup of G.

Next, we observe that for any $g \in G$, $k \in \mathbb{N}$, we have $P^k(g,e) = P^k(e,g^{-1})$ since $e = s_k\cdots s_1 g$ implies $g^{-1} = s_k\cdots s_1 e$. Now let $g \in G$, $x \in N$ be arbitrary. By irreducibility and the definition of N, there exist $r_1, \ldots, r_j, s_1, \ldots, s_{kd}, t_1, \ldots, t_\ell \in \Sigma_\mu$ such that $g = r_j\cdots r_1$, $x = s_{kd}\cdots s_1$, and $g^{-1} = t_\ell\cdots t_1$, hence $P^{\ell+kd+j}(e, gxg^{-1}) > 0$. Now $P^{j+\ell}(e,e) \ge P^j(e,g)P^\ell(g,e) = P^j(e,g)P^\ell(e,g^{-1}) > 0$, so it must be the case that $d|(j+\ell)$ and thus $d|(\ell + kd + j)$. This shows that $gxg^{-1} \in N$, establishing normality.

Finally, for any $s, t \in \Sigma_\mu$ we have $t \in sN$. This is because $P^k(e,s^{-1}) > 0$ for some $k \in \mathbb{N}$, so $P^{k+1}(e,e) \ge P(e,s)P^k(s,e) = P(e,s)P^k(e,s^{-1}) > 0$, so $k+1 = \ell d$ for some $\ell \in \mathbb{N}$. Consequently, $P^{\ell d}(e,s^{-1}t) = P^{k+1}(e,s^{-1}t) \ge P(e,t)P^k(t,s^{-1}t) = P(e,t)P^k(e,s^{-1}) > 0$, hence $s^{-1}t \in N$.

Σ_μ is thus contained in the coset sN and the proof is complete. □

The coset transitions exhibited in the previous example are generic in the following sense.

Proposition 1.5 (**Skipped**). *If P is an irreducible Markov chain with period d, then the state space can be partitioned as $\mathcal{S} = \bigsqcup_{k=0}^{d-1} C_k$ such that P^d is irreducible on each C_k and $P(x,y) > 0$ only if there is a k with $x \in C_k$ and $y \in C_{k+1}$ where the addition is taken modulo d.*

Proof. Fix $x_0 \in \mathcal{S}$ and define

$$C_k = \left\{ y \in \mathcal{S} : \, P^{md+k}(x_0, y) > 0 \text{ for some } m \in \mathbb{N}_0 \right\}, \quad k = 0, 1, \ldots, d-1.$$

Irreducibility guarantees that $\mathcal{S} = \bigcup_{k=1}^{d-1} C_k$. To see that the sets are disjoint, suppose that $P^{md+j}(x_0, y)$ and $P^{nd+k}(x_0, y)$ are positive for some $m, n, j, k \in \mathbb{N}_0$ with $j \leq k < d$. We know that there is an $\ell \in \mathbb{N}$ with $P^\ell(y, x_0) > 0$, so $P^{md+j+\ell}(x_0, x_0)$ and $P^{nd+k+\ell}(x_0, x_0)$ are positive, hence d divides both $md + j + \ell$ and $nd + k + \ell$, and therefore the difference $(nd + k + \ell) - (md + j + \ell) = (n - m)d + (k - j)$. As $k - j$ is between 0 and $d - 1$, we conclude that $j = k$. In sum, $C_j \cap C_k \neq \emptyset$ implies $j = k$.

To see that P^d is irreducible on C_k, suppose that $x, y \in C_k$. Then there exist $\ell, m, n \in \mathbb{N}_0$ such that $P^\ell(x, x_0) > 0$ and $P^{md+k}(x_0, x), P^{nd+k}(x_0, y) > 0$. Since $P^{md+k+\ell}(x_0, x_0) \geq P^{md+k}(x_0, x) P^\ell(x, x_0) > 0$, we have that $k + \ell = qd$ for some $q \in \mathbb{N}_0$. This shows that

$$P^{(n+q)d}(x, y) = P^{\ell+nd+k} \geq P^\ell(x, x_0) P^{nd+k}(x_0, y) > 0.$$

Finally, if $P(x, y) > 0$ and $x \in C_k$, then $P^{md+k}(x_0, x) > 0$ for some $m, k \in \mathbb{N}_0$, hence

$$P^{md+k+1}(x_0, y) \geq P^{md+k}(x_0, x) P(x, y) > 0,$$

showing that $y \in C_{k+1}$. □

Remark 1.3. It is worth observing that aperiodicity can always be circumvented by adding holding probabilities. The *ε-lazy* version of P is the chain $Q = \varepsilon I + (1 - \varepsilon)P$, which proceeds at each time step by flipping an ε-coin, then staying at the current state if it lands heads and transitioning according to P otherwise. Since $Q(x, x) \geq \varepsilon$ for all $s \in \mathcal{S}$, Q is aperiodic. The unqualified term "lazy" corresponds to $\varepsilon = 1/2$.

For random walks on groups, we can avoid periodicity by having $\mu(e) > 0$.

Our next results involve a bit of elementary number theory. We begin by recalling the following generalization of the Euclidean algorithm.

Lemma 1.1 (Bézout's lemma). *The greatest common divisor of positive integers $b_1, \ldots, b_k$ is the smallest positive integer which can be expressed as an integral linear combination of $b_1, \ldots, b_k$.*

Let $d = \alpha_1 b_1 + \ldots + \alpha_k b_k$ be the smallest positive integer which can be so expressed. (Such a number exists by the well-ordering of $\mathbb{N}$.) If any b_j is divided by d, then its remainder $0 \leq r_j < d$ is of the form $r_j = \alpha_1' b_1 + \ldots + \alpha_k' b_k$ since it is obtained by subtracting a multiple of d from b_j. As d is the smallest positive integer of this form, it must be the case that $r_j = 0$, hence d is a divisor of each b_j. If c is any other common divisor of $b_1, \ldots, b_k$, then c divides d as well, hence d is the greatest common divisor.

We can now state what is essentially a special case of the fact that finitely generated subsemigroups of $\mathbb{N}$ are eventually arithmetic progressions with difference the greatest common divisor of the generators.

Lemma 1.2. *If $d_x = 1$, then there is an $m_x \in \mathbb{N}$ such that $P^m(x,x) > 0$ for all $m \geq m_x$.*

Proof. We first observe that there is a finite $F_x \subset I_x$ such that $\gcd(F_x) = \gcd(I_x) = 1$. To see that this is so, note that $d(n) = \gcd(I_x \cap [1,n])$ is a nonincreasing $\mathbb{N}$-valued function and thus can only decrease a finite number of times. Let $N = \max\{n \in \mathbb{N} : d(n) < d(n-1)\}$ and set $F_x = I_x \cap [1,N] = \{b_1, ..., b_n\}$.

Bézout's lemma shows that there are integers $\alpha_1, ..., \alpha_n$ with $\sum_{i=1}^n \alpha_i b_i = 1$.

Set $\alpha = \max_i |\alpha_i|$, $b = \sum_i b_i$. For any $m \in \mathbb{N}$, there exist $q, r \in \mathbb{N}_0$ with $r < b$ such that

$$m = qb + r = q\sum_i b_i + r\sum_i \alpha_i b_i = \sum_i (q + r\alpha_i) b_i.$$

If $q \geq \alpha b$, then $q + r\alpha_i \geq 0$ for all i. In other words, every integer greater than or equal to αb^2 can be written as a sum of elements in $F_x \subset I_x$. Since I_x is closed under addition, this means it contains every integer greater than or equal to $m_x = \alpha b^2$. □

With the preceding in hand, it is easy to establish the following fundamental result.

Theorem 1.2. *If P is an irreducible and aperiodic Markov chain with finite state space $\mathcal{S}$, then there is an $N \in \mathbb{N}$ such that $n \geq N$ implies $P^n(x,y) > 0$ for all $x, y \in \mathcal{S}$.*

Proof. First note that for any $x, y \in \mathcal{S}$, there is an $N(x,y) > 0$ such that $P^n(x,y) > 0$ for all $n \geq N(x,y)$. Specifically, $N(x,y) = m_x + r(x,y)$ with m_x as in Lemma 1.2 and $r(x,y) = \min\{r \in \mathbb{N} : P^r(x,y) > 0\}$. Indeed, irreducibility implies $r(x,y) < \infty$ and if $n \geq N(x,y)$, then $P^n(x,y) \geq P^{n-r(x,y)}(x,x)P^{r(x,y)}(x,y) > 0$ since $n - r(x,y) \geq m_x$.

If $\mathcal{S}$ is finite, then we can take $N = \max_{x,y \in S} N(x,y) < \infty$. □

1.4 Stationary Distributions

Recall the generic two-state chain from Example 1.1. Clearly this chain is irreducible precisely when $p, q > 0$, and it is aperiodic if $\min\{p,q\} < 1$ as well because then there's a positive holding probability. In this case, the matrix P^n is seen to have strictly positive entries for all large n, and we saw that there was a probability distribution π such that $\pi P = \pi$ and $\lim_{n\to\infty} \mu_0 P^n = \pi$.

In general, we say that a probability distribution π on $\mathcal{S}$ is a *stationary distribution* for a Markov chain with transition function p if for all $y \in \mathcal{S}$,

$$\pi(y) = \sum_{x \in X} \pi(x) p(x,y).$$

When $\mathcal{S}$ is finite, this is just the statement that π is a left eigenvector of P with eigenvalue 1. By construction, if $X_n \sim \pi$, then $X_{n+1} \sim \pi P = \pi$ and thus $X_m \sim \pi$ for all $m \geq n$. Once the chain is stationary, it remains so ever after.

Our next order of business is to show that an irreducible and aperiodic Markov chain with finite state space always has a unique stationary distribution which satisfies $\lim_{n\to\infty} P^n(x,y) = \pi(y)$ for all $x \in \mathcal{S}$.

First we observe that since the rows of P sum to one, the column vector $\mathbf{1}$ having all entries equal to one satisfies $P\mathbf{1} = \mathbf{1}$, so 1 is indeed an eigenvalue of P. The following simple argument shows that 1 is the largest eigenvalue of P in modulus.

Proposition 1.6. *If P is the transition matrix of a Markov chain on N states, then the eigenvalues of P satisfy $1 = \lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$.*

Proof. We know that 1 is an eigenvalue of P. Now suppose that $Pf = \lambda f$ for some $f \neq 0$ and let x be such that $|f(x)| \geq |f(y)|$ for all y. Then

$$|\lambda|\,|f(x)| = |\lambda f(x)| = \left|\sum_y P(x,y)f(y)\right| \leq \sum_y P(x,y)\,|f(y)| \leq \sum_y P(x,y)\,|f(x)| = |f(x)|\,. \qquad \square$$

A similar argument establishes the existence of stationary distributions for irreducible finite state space Markov chains.

Proposition 1.7. *If P is irreducible, then any left eigenvector ν with eigenvalue 1 has all entries strictly positive (after being multiplied by a suitable scalar).*

Proof. Choose y so that $|\nu(y)| \geq |\nu(x)|$ for all x. Replacing ν by $-\nu$ if necessary, we can suppose that $\nu(y) > 0$ (as eigenvectors cannot be identically 0 and can be assumed real if the associated eigenvalue is). Now suppose that $\nu(z) \leq 0$ for some z and choose $n \in \mathbb{N}$ so that $P^n(z,y) > 0$. Then

$$\begin{aligned}\nu(y) = \left(\nu P^n\right)(y) &= \sum_x \nu(x)P^n(x,y) \leq \sum_{x\neq z} \nu(x)P^n(x,y) \\ &\leq \sum_{x\neq z} \nu(y)P^n(x,y) < \sum_x \nu(y)P^n(x,y) = \nu(y),\end{aligned}$$

a contradiction. $\square$

As we have already established the existence of right and thus left eigenvectors corresponding to 1, this shows that irreducible Markov chains have stationary distributions—just take $\pi(y) = \nu(y)/\sum_x \nu(x)$.

The next proposition shows that this stationary distribution is unique.

Proposition 1.8. *If P is irreducible, then there is exactly one probability vector π satisfying $\pi P = \pi$.*

Proof. Suppose that h is a right eigenvector with eigenvalue 1 and let x be such that $h(x) \geq h(y)$ for all y. If there were some w with $h(w) < h(x)$, choosing n so that $P^n(x,w) > 0$ gives the contradiction that

$$\begin{aligned}h(x) = \left(P^n h\right)(x) &= \sum_y P^n(x,y)h(y) \\ &= P^n(x,w)h(w) + \sum_{y\neq w} P^n(x,y)h(y) \\ &\leq P^n(x,w)h(w) + \sum_{y\neq w} P^n(x,y)h(x) \\ &< P^n(x,w)h(x) + \sum_{y\neq w} P^n(x,y)h(x) \\ &= \sum_y P^n(x,y)h(x) = h(x).\end{aligned}$$

This shows that all right eigenvectors with eigenvalue 1 are constant, so the kernel of $I - P$ has dimension 1. By rank-nullity, the column space and thus the row space of $I - P$ has codimension 1, so the left null space of $I - P$ has dimension 1. Accordingly, every left eigenvector of P with eigenvalue 1 is a multiple of π.

The proof is now complete since any multiple other than 1 is not a probability distribution. $\square$

Remark 1.4. The proof of Proposition 1.8 showed that 1 is an eigenvalue of P with geometric multiplicity 1. If its algebraic multiplicity were higher, there would be a generalized eigenvector f satisfying $(I-P)f = \mathbf{1}$. But this gives the contradiction $0 = (\pi - \pi)f = \pi(I-P)f = \pi\mathbf{1} = 1$. Thus 1 is a simple eigenvalue of P.

The only real obstacle to convergence at this point is that some of the eigenvalues other than 1 might lie on the unit circle. There is where aperiodicity comes into play.

Proposition 1.9. *Assume that P is irreducible and aperiodic. Then all eigenvalues of P other than 1 have modulus less than one.*

Proof. We know there is an $N \in \mathbb{N}$ such that all entries of P^n are positive whenever $n \geq N$. Suppose that P has an eigenvalue $\lambda \neq 1$ with $|\lambda| = 1$, and take $m \geq N$ so that $\mathrm{Re}(\lambda^m) < 0$. Then $Q = P^m$ is a positive stochastic matrix and we can take $\varepsilon > 0$ to be half the smallest diagonal entry of Q so that $A = Q - \varepsilon I$ is a positive matrix. If $\mathbf{v}$ is an eigenvector of P with eigenvalue λ, then $A\mathbf{v} = \lambda^m\mathbf{v} - \varepsilon\mathbf{v}$, hence $\lambda^m - \varepsilon$ is an eigenvalue of A which lies outside the unit circle. But this is impossible since A is positive with rows summing to $1-\varepsilon$, so the argument from Proposition 1.6 shows its eigenvalues have modulus at most $1-\varepsilon$. □

Remark 1.5 (Skipped). Necessarily, finite irreducible Markov chains have finite periods. The following result relates the eigenvalues of periodic chains to roots of unity.

Proposition 1.10. *If P is irreducible and ω is a primitive n^{th} root of unity, then for any (and thus all) $x \in \mathcal{S}$, $I_x \subseteq n\mathbb{Z}$ if and only if ω is an eigenvalue of P.*

Proof. If $I_x \subseteq n\mathbb{Z}$, then x has period $d = kn$ for some $k \in \mathbb{N}$, so if $\omega \in \sqrt[n]{1}$, $\omega^d = \omega^{kn} = 1$. Now let $C_0, C_1, \ldots, C_{d-1}$ be as in Proposition 1.5. Then there is a function $j : \mathcal{S} \to \{0, 1, \ldots, d-1\}$ such that $x \in C_{j(x)}$ and $P(x,y) > 0$ implies $j(y) = j(x) + 1$ where the addition is mod d. Define $f(x) = \omega^{j(x)}$. Then

$$Pf(x) = \sum_y P(x,y)\omega^{j(y)} = \sum_{y \in C_{j(x)+1}} P(x,y)\omega^{j(x)+1} = \omega^{j(x)+1} = \omega f(x)$$

since $P(x,\cdot)$ lives on $C_{j(x)+1}$. Consequently, f is an eigenfunction with eigenvalue ω.

Conversely, suppose that $Pf = \omega f$ for some function f and $\omega \in \sqrt[n]{1}$ with $\omega^r \neq 1$ for $0 < r < n$. Choosing x so that $|f(x)| \geq |f(y)|$ for all y, we see that

$$|f(x)| = |\omega f(x)| \leq \textstyle\sum_y P(x,y)\,|f(y)| \leq \sum_y P(x,y)\,|f(x)| = |f(x)|\,.$$

Since $P(x,\cdot)$ is a probability and $\sum_y P(x,y)\,|f(y)| = |f(x)|$, we have that $|f(y)| = |f(x)|$ for all y with $P(x,y) > 0$. In fact, since the average of complex numbers all having modulus c has modulus c if and only if all numbers have the same argument, f must be constant on $B_1 = \{y : P(x,y) > 0\}$, and this constant value must be $\sum_y P(x,y)f(y) = \omega f(x)$.

Of course, since any $y \in B_1$ has $|f(y)| = |f(x)| \geq |f(z)|$ for all z, we can repeat this argument to find that $f(z) = \omega f(y) = \omega^2 f(x)$ for all $z \in B_2 = \{z : P(y,z) > 0\}$. As P is irreducible and ω is a primitive n^{th} root of unity, we see that $\mathcal{S} = \bigsqcup_{k=1}^n B_k$ with $B_k = \{z : f(z) = \omega^k f(x)\}$. As the chain clearly transitions as $B_1 \to B_2 \to \cdots \to B_n \to B_1$, we must have $I_x \subseteq n\mathbb{Z}$. □

At this point we have established that an irreducible and aperiodic N-state Markov chain P has eigenvalues $\lambda_1 = 1 > |\lambda_2| \geq \cdots \geq |\lambda_N|$.

We can thus express P in the Jordan normal form $P = SJS^{-1}$ where $J = \text{diag}(J_1, \ldots, J_k)$ is the block diagonal matrix with $J_1 = [1]$ and each J_i with $i > 1$ having some lesser eigenvalue along the main diagonal, 1's or 0's along the super diagonal, and 0's elsewhere. The matrix S has as its columns the corresponding generalized eigenvectors of P.

Since each Jordan block other than J_1 converges to a zero matrix and we can take S to have first column $\mathbf{1}$,

$$P^n = SJ^nS^{-1} \to \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \\ 1 & 0 & \cdots & 0 \end{bmatrix} S^{-1} = \begin{bmatrix} \ell_1 & \cdots & \ell_N \\ & \vdots & \\ \ell_1 & \cdots & \ell_N \end{bmatrix}$$

with $[\ell_1 \quad \cdots \quad \ell_N]$ the first row of S^{-1}.

Now the choice of $\mathbf{1}$ for the first column of S necessitates that $\ell_k = \pi(s_k)$ since π^T is the only vector which is orthogonal to the generalized eigenspaces for the other eigenvalues and has inner product 1 with $\mathbf{1}$.
(If f is a generalized eigenvector for $\lambda \neq 1$, then $0 = \pi(\lambda I - P)^m f = (\lambda - 1)^m \pi f$.)

It follows that $\lim_{n\to\infty} P^n(x, y) = \pi(y)$ for all $x, y \in S$, and that the exponential rate of convergence is given by the modulus of the subdominant eigenvalue, λ_2.

Essentially, we just proved the relevant portions of the Perron-Frobenius theorem in the context of Markov chains. If you're willing to start with this result, then one can establish the 'Fundamental Theorem of Finite Markov Chains' much more directly. We'll give another proof in the next section.

Theorem 1.3. *If P is the transition matrix for an irreducible and aperiodic Markov chain on a finite state space $\mathcal{S}$, then there is a unique and strictly positive probability π on $\mathcal{S}$ such that $\pi(y) = \sum_{x\in\mathcal{S}} \pi(x)P(x, y)$ for all $y \in \mathcal{S}$. Moreover, $\lim_{n\to\infty} P^n(x, y) = \pi(y)$ for all $x, y \in \mathcal{S}$.*

Proof. Theorem 1.2 shows that there is an $N \in \mathbb{N}$ such that P^N is strictly positive. Also, Proposition 1.6 shows that P and thus P^N has spectral radius 1.

The Perron-Frobenius theorem thus implies that 1 is a simple eigenvalue of P^N (and all other eigenvalues have strictly lesser moduli).

It also tells us that the left and right eigenvectors associated with this eigenvalue can be chosen to have all components strictly positive (and no other eigenvectors have this property). This shows there is a unique probability vector π satisfying $\pi P^N = \pi$.

Moreover, since $P^N\mathbf{1} = \mathbf{1}$ and this right eigenvector satisfies $\pi\mathbf{1} = 1$, Perron-Frobenius guarantees that $\lim_{m\to\infty} \left(P^N\right)^m = \mathbf{1}\pi$, the matrix with all rows equal to π.

Now the eigenvalues of P^N are all N^{th} powers of the eigenvalues of P, so P also has 1 as a simple eigenvalue and we must have that $\pi P = \pi$ since any left eigenvector of P corresponding to 1 would also be such a left eigenvector of P^N. This shows that P has a unique and strictly positive stationary distribution π.

Finally, since $\pi P = \pi$ and $\lim_{m\to\infty} P^{mN} = \mathbf{1}\pi$, we see that $\lim_{m\to\infty} P^{mN+k} = \mathbf{1}\pi P^k = \mathbf{1}\pi$ for all $0 \leq k < N$.
This allows us to conclude that $\lim_{n\to\infty} P^n = \mathbf{1}\pi$, or $\lim_{n\to\infty} P^n(x, y) = \pi(y)$ for all $x, y \in \mathcal{S}$. □

Note that the preceding implies that for any initial distribution μ_0,

$$\lim_{k\to\infty} (\mu_0 P^k)(y) = \lim_{k\to\infty} \sum_x \mu_0(x)P^k(x, y) = \sum_x \mu_0(x) \lim_{k\to\infty} P^k(x, y) = \sum_x \mu_0(x)\pi(y) = \pi(y)$$

where we can pass the limit through the sum since the state space is finite.

1.5 Return Times and Reversibility

In the next section we will discuss obtaining quantitative rates of convergence in the setting of Theorem 1.3, but first we take a moment to examine stationary distributions in greater detail.

To begin, suppose that $\{X_k\}_{k\in\mathbb{N}_0}$ is an irreducible Markov chain with state space $\mathcal{S}$. Given $A \subseteq \mathcal{S}$, define the *hitting time*

$$\tau_A = \inf\{k \geq 0 : X_k \in A\}.$$

Also, for $x \in \mathcal{S}$, set

$$\tau_x^+ = \inf\{k > 0 : X_k = x\}.$$

On $\{X_0 \neq x\}$, we have $\tau_x^+ = \tau_{\{x\}}$. If $X_0 = x$, we call τ_x^+ the *first return time*.

Remark 1.6. Note that for all $n \in \mathbb{N}_0$, the events $\{\tau_A \leq n\}$ and $\{\tau_x^+ \leq n\}$ are completely determined by $X_0, \ldots, X_n$. Such $\mathbb{N}_0$-valued random variables are called *stopping times* with respect to the Markov chain.

The *strong Markov property* asserts that if T is a stopping time for $\{X_k\}_{k=0}^\infty$, then conditional on $T < \infty$ and $X_T = x$, $\{X_{T+k}\}_{k=0}^\infty$ is a Markov chain with initial state x and is independent of $X_0, \ldots, X_{T-1}$.
In discrete time, this is a consequence of the ordinary Markov property.

Proposition 1.11. *For any states $x, y \in \mathcal{S}$ of an irreducible finite state space Markov chain, $\mathbb{E}_x[\tau_y^+] < \infty$.*

Proof. Since the chain is irreducible and $|\mathcal{S}| < \infty$, there is an integer r and an $\varepsilon \in (0,1)$ such that for all $n \in \mathbb{N}_0$ and all $x, y \in \mathcal{S}$, there is some $n < k \leq n + r$ with $P^k(x,y) > \varepsilon$. In particular, the probability that the chain fails to visit y in any time interval of length r is at most $1 - \varepsilon$. (See Example 1.6 for elaboration.) It follows that for all $n \in \mathbb{N}$,

$$\mathbb{P}_x\left(\tau_y^+ > nr\right) \leq (1-\varepsilon)\mathbb{P}_x\left(\tau_y^+ > (n-1)r\right).$$

(In order for the chain to have failed to visit y in the first nr steps, it must not have visited in the first $(n-1)r$ steps and then stayed away from y for the next r steps as well.)

Repeated application of this observation shows that $\mathbb{P}_x\left(\tau_y^+ > nr\right) \leq (1-\varepsilon)^n$ for all $n \in \mathbb{N}_0$.

Also, since $\mathbb{P}_x\left(\tau_y^+ > k\right)$ is a nonincreasing function of k, we have $\mathbb{P}_x\left(\tau_y^+ > k\right) \leq (1-\varepsilon)^n$ for all $nr \leq k < (n+1)r$.

We conclude that

$$\mathbb{E}_x[\tau_y^+] = \sum_{k=0}^\infty \mathbb{P}_x\left(\tau_y^+ > k\right) \leq \sum_{n=0}^\infty r\mathbb{P}_x\left(\tau_y^+ > nr\right) \leq r\sum_{n=0}^\infty (1-\varepsilon)^n = \frac{r}{\varepsilon} < \infty. \qquad \square$$

An irreducible Markov chain is called *recurrent* if $\mathbb{P}_x(\tau_x^+ < \infty) = 1$. In other words, the chain is recurrent if it is guaranteed to return to its initial location. A chain which is not recurrent is called *transient*.

Recurrent chains are further classified according to whether the expected return time is finite: If $\mathbb{E}_x[\tau_x^+] < \infty$, we say the chain is *positive recurrent*, and if $\mathbb{E}_x[\tau_x^+] = \infty$, we call it *null recurrent*. One can show that simple random walk on $\mathbb{Z}^d$ is transient for $d \geq 3$ and null recurrent for $d = 1, 2$.

Proposition 1.11 shows that irreducible finite state space chains are positive recurrent.

For countably infinite state space chains, Theorem 1.3 holds if we add the assumption of positive recurrence. The following result suggests why this is important.

Theorem 1.4. *Let p be the transition function of an irreducible and positive recurrent Markov chain $\{X_k\}_{k=0}^{\infty}$ with countable state space $\mathcal{S}$, and define $\pi(x) = 1/\mathbb{E}_x[\tau_x^+]$. Then π is a stationary distribution for p.*

Proof. Fix $z \in \mathcal{S}$ and define μ_z by

$$\mu_z(y) = \mathbb{E}_z\left[\sum_{k=0}^{\infty} 1\left\{X_k = y,\, \tau_z^+ > k\right\}\right] = \sum_{k=0}^{\infty} \mathbb{P}_z\left(X_k = y,\, \tau_z^+ > k\right),$$

the expected number of visits to y before returning to z. By construction, $\mu_z(z) = 1$.

Since the event $\{\tau_z^+ > k\}$ is determined by $X_0, X_1, \ldots, X_k$, we have

$$\begin{aligned}
\sum_{x\in\mathcal{S}} \mu_z(x)p(x,y) &= \sum_{x\in\mathcal{S}}\sum_{k=0}^{\infty} \mathbb{P}_z\left(X_k = x,\, \tau_z^+ > k\right) p(x,y) \\
&= \sum_{x\in\mathcal{S}}\sum_{k=0}^{\infty} \mathbb{P}_z\left(X_k = x,\, X_{k+1} = y,\, \tau_z^+ > k\right) \\
&= \sum_{k=0}^{\infty}\sum_{x\in\mathcal{S}} \mathbb{P}_z\left(X_k = x,\, X_{k+1} = y,\, \tau_z^+ > k\right) \\
&= \sum_{k=0}^{\infty} \mathbb{P}_z\left(X_{k+1} = y,\, \tau_z^+ \geq k+1\right) \\
&= \sum_{k=1}^{\infty} \mathbb{P}_z\left(X_k = y,\, \tau_z^+ \geq k\right) \\
&= \sum_{k=1}^{\infty} \mathbb{P}_z\left(X_k = y,\, \tau_z^+ > k\right) + \sum_{k=1}^{\infty} \mathbb{P}_z\left(X_k = y,\, \tau_z^+ = k\right) \\
&= \sum_{k=0}^{\infty} \mathbb{P}_z\left(X_k = y,\, \tau_z^+ > k\right) - \mathbb{P}_z\left(X_0 = y,\, \tau_z^+ > 0\right) + \mathbb{P}_z\left(X_{\tau_z^+} = y\right) \\
&= \mu_z(y) + \mathbb{P}_z\left(X_{\tau_z^+} = y\right) - \mathbb{P}_z\left(X_0 = y,\, \tau_z^+ > 0\right).
\end{aligned}$$

Now $\mathbb{P}_z\left(X_{\tau_z^+} = y\right)$ and $\mathbb{P}_z\left(X_0 = y,\, \tau_z^+ > 0\right)$ are both 0 if $y \neq z$ and are both 1 if $y = z$.

In either case, we see that $\sum_{x\in\mathcal{S}} \mu_z(x)p(x,y) = \mu_z(y)$, hence μ_z is stationary for p.

Since we are also assuming that the chain is positive recurrent,

$$\begin{aligned}
\sum_{y\in\mathcal{S}} \mu_z(y) &= \sum_{y\in\mathcal{S}}\sum_{k=0}^{\infty} \mathbb{P}_z\left(X_k = y,\, \tau_z^+ > k\right) \\
&= \sum_{k=0}^{\infty}\sum_{y\in\mathcal{S}} \mathbb{P}_z\left(X_k = y,\, \tau_z^+ > k\right) \\
&= \sum_{k=0}^{\infty} \mathbb{P}_z\left(\tau_z^+ > k\right) = \mathbb{E}_z\left[\tau_z^+\right]
\end{aligned}$$

is finite, so we can normalize to get the stationary distribution

$$\pi(x) = \frac{\mu_z(x)}{\sum_{y\in\mathcal{S}} \mu_z(y)} = \frac{\mu_z(x)}{\mathbb{E}_z\left[\tau_z^+\right]} = \frac{1}{\mathbb{E}_x[\tau_x^+]}$$

where the last equality is obtained by choosing $z = x$. □

This more probabilistic construction of the stationary distribution says that the stationary probability of a state is the long term fraction of the time spent at that state: On average, the state x gets visited once in a time interval of length $\mathbb{E}_x[\tau_x^+]$.

Indeed, irreducibility and positive recurrence imply that from any initial state, the chain will reach x in finite time and then can be broken up into a series of i.i.d. excursions between successive visits to x. By the strong law of large numbers, the average number of visits to x in n steps converges to the reciprocal of the average time between visits.

Example 1.7. Consider the random walk (G, μ) on the finite group G with Σ_μ generating G. As an irreducible Markov chain on a finite state space, there is a unique stationary distribution π.

In fact, $\pi(g) = \frac{1}{|G|}$, regardless of any features of G and μ beyond being finite and generating, respectively!

To see that this is so, note that for any $h \in G$,

$$\sum_{g \in G} \frac{1}{|G|} P(g, h) = \frac{1}{|G|} \sum_{g \in G} \mu(hg^{-1}) = \frac{1}{|G|} \sum_{k \in G} \mu(k) = \frac{1}{|G|}.$$

Thus if μ is not concentrated on a coset of a proper normal subgroup, the chain will be approximately uniformly distributed on G after a sufficiently large number of steps.

Since irreducible random walks on finite groups have uniform stationary distributions, their transition matrices have $\mathbf{1}^T$ as a left eigenvector with eigenvalue 1. In other words, both the rows and columns of the transition matrices sum to one. A matrix of nonnegative reals with rows and columns summing to one is called *doubly stochastic*, and the set of doubly stochastic $n \times n$ matrices is precisely the convex hull of the set of $n \times n$ permutation matrices (which have a one in each row and column and zeros elsewhere).

Sometimes we can show that a probability π is stationary for a Markov chain p by checking that it satisfies the *detailed balance equations*

$$\pi(x)p(x, y) = \pi(y)p(y, x) \quad \text{for all } x, y.$$

Indeed, this implies that for every state y,

$$\sum_x \pi(x)p(x, y) = \sum_x \pi(y)p(y, x) = \pi(y) \sum_x p(y, x) = \pi(y).$$

If p has a stationary distribution π satisfying the detailed balance equations, we say the chain is *reversible*.

The name derives from the fact that for any $n \in \mathbb{N}$ and any sequence of states $x_0, x_1, \ldots, x_n$, we have

$$\begin{aligned}
\mathbb{P}_\pi(X_0 = x_0, \ldots, X_n = x_n) &= \pi(x_0)p(x_0, x_1)p(x_1, x_2) \cdots p(x_{n-1}, x_n) \\
&= p(x_1, x_0)\pi(x_1)p(x_1, x_2) \cdots p(x_{n-1}, x_n) \\
&= p(x_1, x_0)p(x_2, x_1)\pi(x_2) \cdots p(x_{n-1}, x_n) \\
&= \ldots = p(x_1, x_0)p(x_2, x_1) \cdots p(x_n, x_{n-1})\pi(x_n) \\
&= \pi(x_n)p(x_n, x_{n-1}) \cdots p(x_2, x_1)p(x_1, x_0) \\
&= \mathbb{P}_\pi(X_0 = x_n, \ldots, X_n = x_0).
\end{aligned}$$

That is, in stationarity $(X_0, \ldots, X_n) =_d (X_n, \ldots X_0)$. The chain looks the same run forwards or backwards!

Example 1.8. Recall that simple random walk on a finite, connected, and undirected graph $G = (V, E)$ defines an irreducible Markov chain with transition matrix $P(u, v) = \frac{1}{\deg(u)} 1\{u \sim v\}$.

The chain is reversible with respect to $\pi(v) = \deg(v)/2|E|$ since

$$\pi(u)P(u,v) = \frac{\deg(u)}{2|E|} \cdot \frac{1\{u \sim v\}}{\deg(u)} = \frac{1\{u \sim v\}}{2|E|} = \frac{\deg(v)}{2|E|} \cdot \frac{1\{v \sim u\}}{\deg(v)} = \pi(v)P(v,u).$$

π is a probability measure on V by the 'handshaking lemma,' $\sum_{v \in V} \deg(v) = 2|E|$, which follows by observing that if we orient the edges, then $|E| = \sum_v \deg^+(v) = \sum_v \deg^-(v)$ (where $\deg^+$ and $\deg^-$ denote the outdegree and indegree, respectively), hence $2|E| = \sum_v \left(\deg^+(v) + \deg^-(v)\right) = \sum_v \deg(v)$.

Example 1.9. The random walk (G, μ) is reversible precisely when $\mu(g) = \mu(g^{-1})$ for all $g \in G$. This is because the stationary distribution is uniform, so $\pi(g)P(g,h) = \pi(h)P(h,g)$ if and only if

$$\mu(hg^{-1}) = P(g,h) = P(h,g) = \mu(gh^{-1}).$$

Symmetry of the driving measure is sufficient for reversibility since $gh^{-1} = (hg^{-1})^{-1}$, and necessity follows by taking $h = e$ in the equation above.

If P is irreducible and reversible with respect to a probability π on $\mathcal{S} = \{s_1, \ldots, s_N\}$ and we define an inner product on $\mathbb{R}^{\mathcal{S}}$ by $\langle f, g \rangle_\pi = \sum_{x \in S} f(x)g(x)\pi(x)$, then

$$\begin{aligned}
\langle Pf, g \rangle_\pi &= \sum_{x \in \mathcal{S}} (Pf)(x)g(x)\pi(x) \\
&= \sum_{x \in \mathcal{S}} \sum_{y \in \mathcal{S}} P(x,y)f(y)g(x)\pi(x) \\
&= \sum_{y \in \mathcal{S}} \sum_{x \in \mathcal{S}} f(y)\pi(x)P(x,y)g(x) \\
&= \sum_{y \in \mathcal{S}} \sum_{x \in \mathcal{S}} f(y)P(y,x)g(x)\pi(y) \\
&= \sum_{y \in \mathcal{S}} f(y)(Pg)(y)\pi(y) = \langle f, Pg \rangle_\pi ,
\end{aligned}$$

hence P is self-adjoint. The spectral theorem thus implies that all eigenvalues of P are real and that $\mathbb{R}^{\mathcal{S}}$ has a basis of eigenfunctions of P which are orthonormal with respect to $\langle \cdot, \cdot \rangle_\pi$.

Alternatively, if D is the diagonal matrix with $D(x,x) = \sqrt{\pi(x)}$, then the matrix $A = DPD^{-1}$ satisfies

$$A(x,y) = \sqrt{\frac{\pi(x)}{\pi(y)}} P(x,y) = \frac{\pi(x)P(x,y)}{\sqrt{\pi(x)\pi(y)}} = \frac{\pi(y)P(y,x)}{\sqrt{\pi(x)\pi(y)}} = \sqrt{\frac{\pi(y)}{\pi(x)}} P(y,x) = A(y,x)$$

and thus is diagonalizable over $\mathbb{R}$. If φ is an eigenfunction of A with eigenvalue λ, then the function $f(x) = \varphi(x)/\sqrt{\pi(x)}$ satisfies $f = D^{-1}\varphi$ and thus

$$Pf = D^{-1}ADf = D^{-1}A\varphi = D^{-1}\lambda\varphi = \lambda f.$$

Since A has real eigenvalues and an eigenbasis $\{\varphi_k\}_{k=1}^N$ that's orthonormal with respect to the standard inner product, P has the same real eigenvalues and a basis of eigenfunctions $f_k = D^{-1}\varphi_k$ which are orthonormal with respect to $\langle \cdot, \cdot \rangle_\pi$.

Assuming that P is also aperiodic, we can write the eigenvalues as $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N > -1$. Letting $f_1, \ldots, f_N$ be a corresponding orthonormal basis of eigenfunctions, we see that any $g : \mathcal{S} \to \mathbb{R}$ can be decomposed as $g(y) = \sum_{k=1}^N \alpha_k f_k(y)$ with $\alpha_k = \langle f_k, g \rangle_\pi$. The expected value of $g(X_n)$ given that $X_0 = x$ is thus

$$\begin{aligned}\mathbb{E}_x\left[g(X_n)\right] &= \sum_{y\in\mathcal{S}} P^n(x,y)g(y) = \sum_{y\in\mathcal{S}} P^n(x,y)\sum_{k=1}^N \alpha_k f_k(y) \\ &= \sum_{k=1}^N \alpha_k \sum_{y\in\mathcal{S}} P^n(x,y) f_k(y) = \sum_{k=1}^N \alpha_k \lambda_k^n f_k(x).\end{aligned}$$

Expanding the point mass δ_y in the eigenbasis $\{f_k\}_{k=1}^N$ gives $\delta_y(x) = \sum_{k=1}^N \alpha_k f_k(x)$ with $\alpha_k = \langle f_k, \delta_y \rangle_\pi = f_k(y)\pi(y)$. It follows that

$$P^n(x,y) = \left(P^n \delta_y\right)(x) = \sum_{k=1}^N f_k(y)\pi(y)\left(P^n f_k\right)(x) = \sum_{k=1}^N \lambda_k^n f_k(x) f_k(y)\pi(y).$$

Since $f_1 \equiv 1$, we have

$$P^n(x,y) = \pi(y) + \sum_{k=2}^N \lambda_k^n f_k(x) f_k(y)\pi(y), \tag{1.1}$$

giving another proof of Theorem 1.3 if one also assumes reversibility.

Remark 1.7. In the following sections, we will be interested in computing the distance to stationarity after n steps. One way to measure this is in terms of the $L^2(\pi)$ norm

$$\left\|P^n(x,\cdot) - \pi\right\|_{L^2(\pi)} = \left[\sum_{y\in\mathcal{S}} \left|P^n(x,y) - \pi(y)\right|^2 \pi(y)^{-1}\right]^{\frac{1}{2}}.$$

(Given a finite signed measure μ, write $f_\mu(x) = \frac{\mu(x)}{\pi(x)}$ for its Radon-Nikodym derivative with respect to π. Then $\|\mu\|_{L^2(\pi)}^2 = \sum_x \mu(x)^2\pi(x)^{-1} = \sum_x f_\mu(x)^2\pi(x) = \langle f_\mu, f_\mu \rangle_\pi$.)

Squaring both sides and invoking Equation (1.1) gives

$$\begin{aligned}\left\|P^n(x,\cdot) - \pi\right\|_{L^2(\pi)}^2 &= \sum_{y\in\mathcal{S}} \left(\sum_{k=2}^N \lambda_k^n f_k(x) f_k(y)\pi(y)\right)^2 \pi(y)^{-1} \\ &= \sum_{y\in\mathcal{S}} \pi(y) \left(\sum_{k=2}^N \lambda_k^{2n} f_k(x)^2 f_k(y)^2 + 2\sum_{2\leq k<\ell\leq N} \lambda_k^n f_k(x) f_k(y) \lambda_\ell^n f_\ell(x) f_\ell(y)\right) \\ &= \sum_{k=2}^N \lambda_k^{2n} f_k(x)^2 \sum_{y\in\mathcal{S}} f_k(y)^2\pi(y) + 2\sum_{2\leq k<\ell\leq N} \lambda_k^n \lambda_\ell^n f_k(x) f_\ell(x) \sum_{y\in\mathcal{S}} f_k(y) f_\ell(y)\pi(y) \\ &= \sum_{k=2}^N \lambda_k^{2n} f_k(x)^2 \langle f_k, f_k \rangle_\pi + 2\sum_{2\leq k<\ell\leq N} \lambda_k^n \lambda_\ell^n f_k(x) f_\ell(x) \langle f_k, f_\ell \rangle_\pi = \sum_{k=2}^N \lambda_k^{2n} f_k(x)^2.\end{aligned}$$

Now we say that a chain is *transitive* if for any $x, y \in \mathcal{S}$, there is a bijection $\varphi_{xy} : \mathcal{S} \to \mathcal{S}$ with $\varphi_{xy}(x) = y$ and $P\big(\varphi_{xy}(w), \varphi_{xy}(z)\big) = P(w,z)$ for all $w, z \in \mathcal{S}$.

An easy induction argument shows that this implies $P^n\big(\varphi_{xy}(w), \varphi_{xy}(z)\big) = P^n(w,z)$ for all $n \in \mathbb{N}$.

Thus transitivity basically means that the chain 'looks the same' from any state, and random walks on groups are clearly transitive since one can take $\varphi_{xy}(w) = wx^{-1}y$.

Necessarily, transitive chains have a uniform stationary distribution because for any $x, y \in \mathcal{S}$,

$$\sum_z U(z)P(z,x) = \sum_z U\big(\varphi_{xy}(z)\big)P(\varphi_{xy}(z),y) = \sum_w U(w)P(w,y),$$

hence

$$1 = \sum_x \sum_z U(z)P(z,x) = \sum_x \sum_w U(w)P(w,y) = |\mathcal{S}| \sum_w U(w)P(w,y)$$

or $\sum_w U(w)P(w,y) = 1/|\mathcal{S}| = U(y)$.

Moreover, for any $x, y \in \mathcal{S}$, we have

$$\begin{aligned}\|P^n(x,\cdot) - U\|^2_{L^2(U)} &= |\mathcal{S}| \sum_{z\in\mathcal{S}} |P^n(x,z) - U(z)|^2 \\ &= |\mathcal{S}| \sum_{z\in\mathcal{S}} \big|P^n\big(y,\varphi_{xy}(z)\big) - U(\varphi_{xy}(z))\big|^2 = \|P^n(y,\cdot) - U\|^2_{L^2(U)}\,.\end{aligned}$$

Thus if P is both reversible and transitive, then for any $x, y \in \mathcal{S}$,

$$\|P^n(x,\cdot) - U\|^2_{L^2(U)} = \|P^n(y,\cdot) - U\|^2_{L^2(U)} = \sum_{k=2}^N \lambda_k^{2n} f_k(y)^2,$$

so

$$|\mathcal{S}|\,\|P^n(x,\cdot) - U\|^2_{L^2(U)} = \sum_y \sum_{k=2}^N \lambda_k^{2n} f_k(y)^2.$$

As $\langle f_k, f_k\rangle_U = 1$, we get

$$\|P^n(x,\cdot) - U\|^2_{L^2(U)} = \sum_{k=2}^N \lambda_k^{2n} \sum_y f_k(y)^2 U(y) = \sum_{k=2}^N \lambda_k^{2n}. \tag{1.2}$$

We know that the k-step distribution of an irreducible and aperiodic Markov chain on a finite state space converges to the chain's stationary distribution as $k \to \infty$, regardless of the initial distribution. However, in practice one would like to know how well the stationary distribution approximates the distribution of X_k for fixed values of k.

2.1 Total Variation

To make the preceding question mathematically rigorous, we need a metric on the space of probability measures on $\mathcal{S}$. We saw one such example in Remark 1.7, but a choice that is arguably more natural is the *total variation distance* defined by

$$\|\mu - \nu\|_{TV} = \max_{A \subseteq \mathcal{S}} |\mu(A) - \nu(A)| .$$

We will see that the maximum is actually attained by some event $B \subseteq \mathcal{S}$, so we don't need to speak of suprema even when $\mathcal{S}$ is countably infinite.

Also, for any $A \subseteq \mathcal{S}$,

$$\mu(A^C) - \nu(A^C) = \big(1 - \mu(A)\big) - \big(1 - \nu(A)\big) = \nu(A) - \mu(A).$$

This shows that the absolute value can be dropped from the definition if we so desire.

Finally, we observe that total variation does indeed define a metric on the space of probabilities on $\mathcal{S}$ since $\|\mu - \mu\|_{TV} = 0$, $\|\mu - \nu\|_{TV} = \|\nu - \mu\|_{TV}$, and if η is another probability on $\mathcal{S}$, then

$$\begin{aligned}\|\mu - \nu\|_{TV} &= \max_{A \subseteq \mathcal{S}} |\mu(A) - \nu(A)| \le \max_{A \subseteq \mathcal{S}} \big(|\mu(A) - \eta(A)| + |\eta(A) - \nu(A)| \big) \\ &\le \max_{A \subseteq \mathcal{S}} |\mu(A) - \eta(A)| + \max_{A \subseteq \mathcal{S}} |\eta(A) - \nu(A)| = \|\mu - \eta\|_{TV} + \|\eta - \nu\|_{TV} .\end{aligned}$$

As our definition of the total variation distance involves taking a maximum over $2^{\mathcal{S}}$, it can be difficult to work with at times. Fortunately, there are other equivalent characterizations that we can appeal to if need be. One of the most computationally useful constructions realizes total variation as half the L^1 distance between the associated mass functions.

Proposition 2.1. *Suppose that μ and ν are probabilities on a countable set $\mathcal{S}$. Then*

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \mathcal{S}} |\mu(x) - \nu(x)| .$$

Proof. Define $B = \{x \in \mathcal{S} : \mu(x) \ge \nu(x)\}$. Then for any $A \subseteq \mathcal{S}$,

$$\begin{aligned}\mu(A) - \nu(A) &= \sum_{x \in A \cap B} [\mu(x) - \nu(x)] + \sum_{x \in A \setminus B} [\mu(x) - \nu(x)] \\ &\le \sum_{x \in A \cap B} [\mu(x) - \nu(x)] \\ &\le \sum_{x \in B \cap A} [\mu(x) - \nu(x)] + \sum_{x \in B \setminus A} [\mu(x) - \nu(x)] = \mu(B) - \nu(B).\end{aligned}$$

(The first inequality uses $\mu(x) - \nu(x) < 0$ for $x \in A \setminus B$ and the second uses $\mu(x) - \nu(x) \ge 0$ for $x \in B \setminus A$.)

This shows that

$$\|\mu-\nu\|_{TV} = \max_{A\subseteq\mathcal{S}} \big[\mu(A)-\nu(A)\big] = \mu(B)-\nu(B).$$

Thus, since $\nu(B^C)-\mu(B^C) = \mu(B)-\nu(B)$, we have

$$\begin{aligned} 2\,\|\mu-\nu\|_{TV} &= \big[\mu(B)-\nu(B)\big] + \big[\nu(B^C)-\mu(B^C)\big] \\ &= \sum_{x\in B}[\mu(x)-\nu(x)] + \sum_{x\in B^C}[\nu(x)-\mu(x)] \\ &= \sum_{x\in B}|\mu(x)-\nu(x)| + \sum_{x\in B^C}|\mu(x)-\nu(x)| = \sum_{x\in\mathcal{S}}|\mu(x)-\nu(x)|\,. \end{aligned}$$

□

Corollary 2.1. *For any probabilities μ and ν on $\mathcal{S}$, $\|\mu-\nu\|_{TV} \le 1$.*

Proof.

$$\|\mu-\nu\|_{TV} = \frac{1}{2}\sum_{x\in\mathcal{S}}|\mu(x)-\nu(x)| \le \frac{1}{2}\sum_{x\in\mathcal{S}}|\mu(x)| + \frac{1}{2}\sum_{x\in\mathcal{S}}|\nu(x)| = 1.$$

□

Note that the proof of Proposition 2.1 also gives the description

$$\|\mu-\nu\|_{TV} = \sum_{x:\mu(x)\ge\nu(x)} \big[\mu(x)-\nu(x)\big]. \tag{2.1}$$

Another important characterization of the total variation distance involves coupling.

A *coupling* of distributions μ and ν on a countable set $\mathcal{S}$ is a pair of $\mathcal{S}$-valued random variables (X,Y), defined on a common probability space, satisfying $\mathbb{P}\big(X=x\big)=\mu(x)$ and $\mathbb{P}\big(Y=y\big)=\nu(y)$ for all $x,y\in\mathcal{S}$. (One sometimes also describes such a coupling as a probability q on $\mathcal{S}\times\mathcal{S}$ with marginals μ and ν; think of q as the joint distribution of X and Y.)

Example 2.1. Suppose μ and ν are probabilities on $\{0,1\}$ with $\mu(1)=p$ and $\nu(1)=q$, $0<p<q<1$.

One way to couple μ and ν is to take $X\sim\mu$ and $Y\sim\nu$ independent.

Another possibility is to let $X\sim\mu$ and define Z as follows: If $X=1$, then $Z=1$; if $X=0$, flip a coin with heads probability $\frac{q-p}{1-p}$ and set $Z=1$ if the coin lands heads and $Z=0$ otherwise. Then Z is a $\{0,1\}$-valued random variable with $\mathbb{P}\big(Z=1\big)=\mathbb{P}\big(Z=1,X=1\big)+\mathbb{P}\big(Z=1,X=0\big)=p+(1-p)\frac{q-p}{1-p}=q$, so (X,Z) is a coupling of μ and ν.

Note that $\mathbb{P}\big(Z\ge X\big)=1$ while $\mathbb{P}\big(Y\ge X\big)=1-p(1-q)<1$. Also, one easily checks that

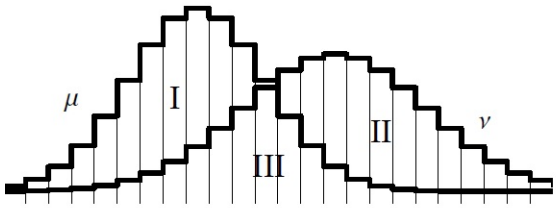
$$\|\mu-\nu\|_{TV} = q-p = \mathbb{P}\big(X\ne Z\big).$$

Proposition 2.2. *Let μ and ν be probability distributions on $\mathcal{S}$. Then*

$$\|\mu-\nu\|_{TV} = \min\big\{\mathbb{P}\big(X\ne Y\big) : (X,Y) \text{ is a coupling of } \mu \text{ and } \nu\big\}.$$

Proof. First observe that if (X,Y) is a coupling of μ and ν, then for any event A,

$$\begin{aligned} \mu(A)-\nu(A) &= \mathbb{P}\big(X\in A\big)-\mathbb{P}\big(Y\in A\big) \\ &= \mathbb{P}\big(X\in A, Y\in A\big)+\mathbb{P}\big(X\in A, Y\notin A\big)-\mathbb{P}\big(Y\in A, X\in A\big)-\mathbb{P}\big(Y\in A, X\notin A\big) \\ &\le \mathbb{P}\big(X\in A, Y\notin A\big) \le \mathbb{P}\big(X\ne Y\big). \end{aligned}$$



Taking a maximum over $A \subseteq \mathcal{S}$ shows that

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

so it remains only to exhibit a coupling (X, Y) of μ and ν for which $\mathbb{P}\big(X \neq Y\big) = \|\mu - \nu\|_{TV}$.

To this end, set

$$p = \sum_{x \in \mathcal{S}} \big[\mu(x) \wedge \nu(x)\big],$$

and note that

$$\begin{aligned}
\sum_{x \in \mathcal{S}} \big[\mu(x) \wedge \nu(x)\big] &= \sum_{x:\mu(x)<\nu(x)} \mu(x) + \sum_{x:\mu(x)\geq\nu(x)} \nu(x) \\
&= \sum_{x:\mu(x)<\nu(x)} \mu(x) + \sum_{x:\mu(x)\geq\nu(x)} \mu(x) - \sum_{x:\mu(x)\geq\nu(x)} \mu(x) + \sum_{x:\mu(x)\geq\nu(x)} \nu(x) \\
&= 1 - \sum_{x:\mu(x)\geq\nu(x)} \big[\mu(x) - \nu(x)\big] \\
&= 1 - \|\mu - \nu\|_{TV}\,.
\end{aligned}$$

We will construct X and Y as follows: Flip a coin with heads probability p.

If the coin lands heads, choose a value Z according to the distribution with mass function

$$\varphi_{\mathrm{III}}(x) = \frac{\mu(x) \wedge \nu(x)}{1 - \|\mu - \nu\|_{TV}}$$

and set $X = Y = Z$.

If the coin lands tails, choose X according to

$$\varphi_{\mathrm{I}}(x) = \frac{\mu(x) - \nu(x)}{\|\mu - \nu\|_{TV}} 1\big\{\mu(x) > \nu(x)\big\}$$

and choose Y independently according to

$$\varphi_{\mathrm{II}}(x) = \frac{\nu(x) - \mu(x)}{\|\mu - \nu\|_{TV}} 1\big\{\nu(x) > \mu(x)\big\}.$$

(Equation (2.1) shows that these are indeed mass functions.)

Then (X, Y) is a coupling of μ and ν since consideration of whether or not $\mu(x) \leq \nu(x)$ shows that

$$\begin{aligned}
p\varphi_{\mathrm{III}}(x) + (1-p)\varphi_{\mathrm{I}}(x) &= \mu(x), \\
p\varphi_{\mathrm{III}}(x) + (1-p)\varphi_{\mathrm{II}}(x) &= \nu(x).
\end{aligned}$$

Moreover, φ_{I} and φ_2 have disjoint support, so $X \neq Y$ precisely when the coin lands tails, hence

$$\mathbb{P}\big(X \neq Y\big) = 1 - p = \|\mu - \nu\|_{TV}\,. \qquad \square$$

2.2 Mixing Time

We will explore coupling in greater detail shortly, but first let's return to the question that motivated our interest in total variation to begin with.

The setup here is a Markov chain $\{X_k\}_{k=0}^{\infty}$ with transition matrix P, finite state space $\mathcal{S}$, and stationary distribution π. If the chain starts at $X_0 = x$, then $\mathbb{P}\big(X_t = y\big) = P^t(x, y)$, so it is natural to define

$$d(t) := \max_x \left\| P^t(x, \cdot) - \pi \right\|_{TV},$$

the worst case distance to stationarity after t steps.

This definition actually maximizes over all initial distributions since for any probability μ on $\mathcal{S}$,

$$\begin{aligned}
\left\| \mu P^t - \pi \right\|_{TV} &= \frac{1}{2} \sum_{y \in \mathcal{S}} \left| \left(\mu P^t\right)(y) - \pi(y) \right| \\
&= \frac{1}{2} \sum_{y \in \mathcal{S}} \left| \left(\sum_{x \in \mathcal{S}} \mu(x) P^t(x, y) \right) - \pi(y) \right| \\
&= \frac{1}{2} \sum_{y \in \mathcal{S}} \left| \sum_{x \in \mathcal{S}} \mu(x) \left(P^t(x, y) - \pi(y) \right) \right| \\
&\leq \frac{1}{2} \sum_{y \in \mathcal{S}} \sum_{x \in \mathcal{S}} \mu(x) \left| P^t(x, y) - \pi(y) \right| \\
&= \frac{1}{2} \sum_{x \in \mathcal{S}} \sum_{y \in \mathcal{S}} \mu(x) \left| P^t(x, y) - \pi(y) \right| \\
&= \sum_{x \in \mathcal{S}} \mu(x) \left\| P^t(x, \cdot) - \pi \right\|_{TV} \leq d(t).
\end{aligned}$$

Another important observation about $d(t)$ is that it is nonincreasing.

Proposition 2.3. *$d(t) \geq d(t+1)$ for every $t \in \mathbb{N}$.*

Proof. For any $t \in \mathbb{N}$, $x \in \mathcal{S}$, we have

$$\begin{aligned}
\sum_y \left| P^{t+1}(x, y) - \pi(y) \right| &= \sum_y \left| \sum_z P(x, z) P^t(z, y) - \pi(y) \right| \\
&= \sum_y \left| \sum_z P(x, z) \big(P^t(z, y) - \pi(y) \big) \right| \\
&\leq \sum_y \sum_z P(x, z) \left| P^t(z, y) - \pi(y) \right| \\
&= \sum_z P(x, z) \sum_y \left| P^t(z, y) - \pi(y) \right|.
\end{aligned}$$

Multiplying through by $1/2$ and noting that the average cannot exceed the maximum gives

$$\left\| P^{t+1}(x, \cdot) - \pi \right\|_{TV} \leq \sum_z P(x, z) \left\| P^t(z, \cdot) - \pi \right\|_{TV} \leq d(t),$$

and the result follows since x was arbitrary. □

It can also be useful to consider the worst case difference between the t-step distributions of the chain started at different states,

$$\overline{d}(t) := \max_{x,y\in\mathcal{S}} \left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV}.$$

Proposition 2.4. *$d(t) \le \overline{d}(t) \le 2d(t)$ for every $t \in \mathbb{N}$.*

Proof. Since $\pi(z) = \sum_y \pi(y)P^t(y,z)$, we have

$$\begin{aligned}
\left\|P^t(x,\cdot) - \pi\right\|_{TV} &= \frac{1}{2}\sum_z \left|P^t(x,z) - \pi(z)\right| \\
&= \frac{1}{2}\sum_z \left|\sum_y \pi(y)\left(P^t(x,z) - P^t(y,z)\right)\right| \\
&\le \frac{1}{2}\sum_z \sum_y \pi(y)\left|P^t(x,z) - P^t(y,z)\right| \\
&= \sum_y \pi(y)\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV} \le \max_y \left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV}.
\end{aligned}$$

Taking a maximum over x shows that $d(t) \le \overline{d}(t)$.

For the second inequality, we have

$$\begin{aligned}
\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{TV} &= \frac{1}{2}\sum_z \left|P^t(x,z) - P^t(y,z)\right| \\
&\le \frac{1}{2}\sum_z \left(\left|P^t(x,z) - \pi(z)\right| + \left|\pi(z) - P^t(y,z)\right|\right) \\
&= \left\|P^t(x,\cdot) - \pi\right\|_{TV} + \left\|P^t(y,\cdot) - \pi\right\|_{TV}.
\end{aligned}$$

□

The other big fact about $\overline{d}$ is that it's submultiplicative.

Proposition 2.5. *$\overline{d}(s+t) \le \overline{d}(s)\overline{d}(t)$ for every $s,t \in \mathbb{N}$.*

Proof. Fix $x, y \in \mathcal{S}$ and let (X_s, Y_s) be the optimal coupling of $P^s(x,\cdot)$ and $P^s(y,\cdot)$ from Proposition 2.2. Since $P^{s+t} = P^sP^t$ and $X_s \sim P^s(x,\cdot)$, we have that

$$P^{s+t}(x,w) = \sum_z P^s(x,z)P^t(z,w) = \sum_z \mathbb{P}\big(X_s = z\big)P^t(z,w) = \mathbb{E}\left[P^t(X_s,w)\right].$$

The exact same reasoning gives $P^{s+t}(y,w) = \mathbb{E}\left[P^t(Y_s,w)\right]$, so

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

and thus

$$\begin{aligned}
\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{TV} &= \frac{1}{2}\sum_w \left|P^{s+t}(x,w) - P^{s+t}(y,w)\right| \\
&= \frac{1}{2}\sum_w \left|\mathbb{E}\left[P^t(X_s,w) - P^t(Y_s,w)\right]\right| \\
&\le \frac{1}{2}\sum_w \mathbb{E}\left|P^t(X_s,w) - P^t(Y_s,w)\right| = \mathbb{E}\left[\frac{1}{2}\sum_w \left|P^t(X_s,w) - P^t(Y_s,w)\right|\right].
\end{aligned}$$

Now $\frac{1}{2}\sum_w |P^t(X_s,w) - P^t(Y_s,w)| = \|P^t(X_s,\cdot) - P^t(Y_s,\cdot)\|_{TV}$ is 0 when $X_s = Y_s$ and it is always bounded above by $\overline{d}(t)$, so

$$\begin{aligned}\left\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\right\|_{TV} &\le \mathbb{E}\left[\frac{1}{2}\sum_w \left|P^t(X_s,w) - P^t(Y_s,w)\right|\right] \\ &\le \mathbb{E}\left[\overline{d}(t)1\{X_s \neq Y_s\}\right] \\ &= \overline{d}(t)\mathbb{P}\big(X_s \neq Y_s\big) = \overline{d}(t)\overline{d}(s)\end{aligned}$$

since (X_s, Y_s) is an optimal coupling. Maximizing over x, y yields the assertion. □

While $d(t)$ does not have this nice submultiplicative property in general, Propositions 2.4 and 2.5 show that for any $k, t \in \mathbb{N}$,

$$d(kt) \le \overline{d}(kt) \le \overline{d}(t)^k \le 2^k d(t)^k.$$

Much of our focus going forward will be in estimating the *ε-mixing time*

$$t_{\text{mix}}(\varepsilon) := \min\{t : \ d(t) \le \varepsilon\},$$

which tells us how long we have to run the chain to ensure that the total variation distance to stationarity is at most ε, regardless of initial state.

It is often convenient to take the (arbitrary but standard) choice of $\varepsilon = 1/4$ and define the *mixing time* $t_{\text{mix}} = t_{\text{mix}}(1/4)$. Observe that for any $\ell \in \mathbb{N}$, we have

$$d\big(\ell t_{\text{mix}}(\varepsilon)\big) \le 2^\ell d\big(t_{\text{mix}}(\varepsilon)\big)^\ell \le (2\varepsilon)^\ell,$$

hence $d(\ell t_{\text{mix}}) \le 2^{-\ell}$.

As $2^{-\ell} \le \varepsilon$ when $\ell \ge \log_2(\varepsilon^{-1})$, we have

$$t_{\text{mix}}(\varepsilon) \le \left\lceil \log_2(\varepsilon^{-1}) \right\rceil t_{\text{mix}}.$$

Example 2.2. If (G, μ) is an irreducible random walk on a finite group, then Example 1.7 shows that the uniform distribution $U(g) = \frac{1}{|G|}$ is stationary. Since the transition matrix $P(x,y) = \mu(yx^{-1})$ satisfies $P(x,y) = P(xg, yg)$, we see that for any initial state $g \in G$, we have

$$\begin{aligned}\left\|P^t(g,\cdot) - U\right\|_{TV} &= \frac{1}{2}\sum_{h\in G}\left|P^t(g,h) - U(h)\right| \\ &= \frac{1}{2}\sum_{h\in G}\left|P^t(id, hg^{-1}) - \frac{1}{|G|}\right| \\ &= \frac{1}{2}\sum_{k\in G}\left|P^t(id,k) - U(k)\right| = \left\|P^t(id,\cdot) - U\right\|_{TV}.\end{aligned}$$

As such, $d(t) = \|P^t(id,\cdot) - U\|_{TV}$ and we can always just assume that the random walk starts at the identity. (Remark 1.7 shows that an analogous result holds for the $L^2(\pi)$ distance.)

In this case, the distribution of the chain after t steps is given by the t-fold *convolution* of μ with itself, defined by $\mu^{*1} = \mu$ and $\mu^{*k} = \mu * \mu^{*(k-1)}$ where $\mu * \nu(s) = \sum_{t\in G}\mu(st^{-1})\nu(t)$ means 'choose t from ν and then, independently, choose r from μ and form the product $s = rt$.'

Also, if we define the *time reversal* of P by

$$\widehat{P}(x,y) = \frac{U(y)}{U(x)}P(y,x) = P(y,x) = \mu(xy^{-1}) = \check{\mu}(yx^{-1})$$

where $\check{\mu}(g) = \mu(g^{-1})$ is the measure driving the inverse walk, then

$$\begin{aligned}\left\|\widehat{P}^t(id,\cdot) - U\right\|_{TV} &= \frac{1}{2}\sum_{g\in G}\left|\check{\mu}^{*t}(g) - U(g)\right| = \frac{1}{2}\sum_{g\in G}\left|\mu^{*t}(g^{-1}) - U(g)\right| \\ &= \frac{1}{2}\sum_{g\in G}\left|\mu^{*t}(g^{-1}) - U(g^{-1})\right| = \frac{1}{2}\sum_{h\in G}\left|\mu^{*t}(h) - U(h)\right| = \left\|P^t(id,\cdot) - U\right\|_{TV}.\end{aligned}$$

In particular, the inverse walk has the same mixing time as the original. (This also shows that the mixing behavior doesn't depend on whether we multiply on the left or right; see Example 1.3.)

Remark 2.1. While we will mostly be focused on total variation in this class, there are several other useful metrics and corresponding notions of mixing times.

For instance, if π is a strictly positive probability on $\mathcal{S}$ (e.g. the stationary distribution of an irreducible Markov chain) and $1 \le p < \infty$, one could consider

$$\|\mu - \nu\|_{L^p(\pi)} = \left[\sum_{x\in\mathcal{S}}\left|\frac{\mu(x)}{\pi(x)} - \frac{\nu(x)}{\pi(x)}\right|^p \pi(x)\right]^{\frac{1}{p}}.$$

An important case which often arises when using spectral methods to analyze Markov chain convergence is $p = 2$ (see Remark 1.7), and Cauchy-Schwarz shows that

$$\begin{aligned}2\|\mu - \nu\|_{TV} = \|\mu - \nu\|_{L^1(\pi)} &= \sum_x \sqrt{\pi(x)}\left|\frac{\mu(x)}{\pi(x)} - \frac{\nu(x)}{\pi(x)}\right| \cdot \sqrt{\pi(x)} \\ &\le \left(\sum_x \left|\frac{\mu(x)}{\pi(x)} - \frac{\nu(x)}{\pi(x)}\right|^2 \pi(x)\right)^{\frac{1}{2}} \cdot \left(\sum_x \pi(x)\right)^{\frac{1}{2}} = \|\mu - \nu\|_{L^2(\pi)}.\end{aligned}$$

Also common in the Markov chain literature is the *separation distance* (which is not actually a metric) given by $s(\mu,\nu) = \max_x\left(1 - \frac{\mu(x)}{\nu(x)}\right)$. This upper-bounds total variation and is especially useful for arguments involving strong stationary times.

2.3 Coupling

The notion of coupling provides a powerful means of bounding Markov chain mixing times.

The general idea is to extend the concept from a static to a dynamic perspective by defining a *coupling of a transition probability* p on a countable state space $\mathcal{S}$ to be an $\mathcal{S}\times\mathcal{S}$-valued process $\{(X_k, Y_k)\}_{k\ge 0}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the coordinate processes $\{X_k\}_{k\ge 0}$ and $\{Y_k\}_{k\ge 0}$ are each Markov chains with transition probability p.

The following result illustrates the utility of this construction.

Lemma 2.1. *Suppose that $\{(X_k, Y_k)\}_{k\geq 0}$ is a coupling of p with $X_0 \sim \mu$, $Y_0 \sim \nu$. If T is a random time such that $X_k = Y_k$ on $\{T \leq k\}$, then*

$$\|\mu p^t - \nu p^t\|_{TV} \leq \mathbb{P}(T > t).$$

Proof. By construction, (X_t, Y_t) is a coupling of μp^t and νp^t. Since $\{X_t \neq Y_t\} \subseteq \{T > t\}$, Proposition 2.2 implies

$$\|\mu p^t - \nu p^t\|_{TV} \leq \mathbb{P}(X_t \neq Y_t) \leq \mathbb{P}(T > t). \qquad \square$$

When $\nu = \pi$, Lemma 2.1 gives the bound $\|\mu p^t - \pi\|_{TV} \leq \mathbb{P}(T > t)$.

It can also be convenient to take $\mu = \delta_x$ and $\nu = \delta_y$ to get $\|p^t(x, \cdot) - p^t(y, \cdot)\|_{TV} \leq \mathbb{P}_{x,y}(T > t)$.

If $\mathcal{S}$ is finite, one can then bound the distance to stationarity using

$$d(t) \leq \bar{d}(t) \leq \max_{x,y} \mathbb{P}_{x,y}(T > t).$$

Our next example uses these ideas to give another proof of the Markov chain convergence theorem.

Example 2.3. Suppose that P is the transition matrix of an irreducible and aperiodic Markov chain with finite state space $\mathcal{S}$. Define a new chain on $\mathcal{S} \times \mathcal{S}$ by

$$Q\big((x_1, y_1), (x_2, y_2)\big) = P(x_1, x_2)P(y_1, y_2).$$

We first observe that Q is irreducible. Indeed, given any $(x, y), (x', y') \in \mathcal{S} \times \mathcal{S}$, irreducibility of P implies the existence of $j, k \in \mathbb{N}$ such that $P^j(x, x'), P^k(y, y') > 0$ and aperiodicity implies the existence of $m \in \mathbb{N}$ such that $P^n(x, x), P^n(y, y) > 0$ for all $n > m$. Consequently,

$$\begin{aligned} Q^{j+k+m}\big((x, y), (x', y')\big) &= P^{j+k+m}(x, x')P^{j+k+m}(y, y') \\ &\geq P^{k+m}(x, x)P^j(x, x')P^{j+m}(y, y)P^k(y, y') > 0. \end{aligned}$$

Now let $\{(X_k, Y_k)\}_{k\geq 0}$ be a Markov chain on $\mathcal{S} \times \mathcal{S}$ with transition matrix Q and initial state (x, y), and set $T = \min\{t : X_t = Y_t\}$.
Note that for any $z \in \mathcal{S}$, the random time $T_z = \min\{t : X_t, Y_t = z\}$ satisfies $\mathbb{P}(T_z \geq T) = 1$ (since $\{X_t, Y_t = z\} \subseteq \{X_t = Y_t\}$) and $\mathbb{P}(T_z < \infty) = 1$ (since the expected hitting time of (z, z) for the chain started at (x, y) is finite by Proposition 1.11). It follows that $\mathbb{P}(T < \infty) = 1$ and thus $\lim_{t\to\infty} \mathbb{P}(T > t) = 0$.
Next, define

$$Z_k = \begin{cases} Y_k, & k \leq T \\ X_k, & k > T \end{cases}$$

and consider the process $(X_k, Z_k)_{k\geq 0}$. Since $X_0 = x$, $Z_0 = y$, and each marginally evolves according to P, (X_t, Z_t) is a coupling of $P^t(x, \cdot)$ and $P^t(y, \cdot)$. As $X_k = Z_k$ on $\{T \leq k\}$, Lemma 2.1 gives

$$\|P^t(x, \cdot) - P^t(y, \cdot)\|_{TV} \leq \mathbb{P}(T > t) \to 0 \text{ as } t \to \infty.$$

Taking a maximum over (x, y) shows that $\bar{d}(t)$ and thus $d(t)$ goes to 0 as $t \to \infty$. This in turn implies Theorem 1.3.

Remark 2.2 (Skipped). The above argument can be modified to show that if p is an irreducible, aperiodic, and positive recurrent Markov chain on a countable state space $\mathcal{S}$, then for every $x \in \mathcal{S}$, $p^t(x,\cdot)$ converges in total variation to the stationary distribution as $t \to \infty$. (When $\mathcal{S}$ is infinite, this is a stronger statement than pointwise convergence.)

The main issue is that a new proof that $\mathbb{P}\big(T < \infty\big)$ is needed. Basically, one notes that p has a stationary distribution π by Theorem 1.4 and then an easy calculation shows that $\widetilde{\pi}(x,y) = \pi(x)\pi(y)$ is stationary for $q\big((x_1,y_1),(x_2,y_2)\big) = p(x_1,x_2)p(y_1,y_2)$. As q is irreducible (by the same argument as before) and possesses a stationary distribution, one can infer that it is positive recurrent—this step takes a little bit of work.

Together with irreducibility, this shows that the hitting time of any state from any other is finite with full probability, so proceeding as in the $|\mathcal{S}| < \infty$ case gives $\lim_{t\to\infty} \|p^t(x,\cdot) - p^t(y,\cdot)\|_{TV} = 0$ for all $x, y \in \mathcal{S}$.

One may then observe that

$$\begin{aligned}\big\|p^t(x,\cdot) - \pi\big\|_{TV} &= \frac{1}{2}\sum_z \big|p^t(x,z) - \pi(z)\big| = \frac{1}{2}\sum_z \left|\sum_y \pi(y)\big(p^t(x,z) - p^t(y,z)\big)\right| \\ &\le \frac{1}{2}\sum_z \sum_y \pi(y)\big|p^t(x,z) - p^t(y,z)\big| = \sum_y \pi(y)\big\|p^t(x,\cdot) - p^t(y,\cdot)\big\|_{TV}.\end{aligned}$$

Since $\|p^t(x,\cdot) - p^t(y,\cdot)\|_{TV} \le 1$ and $\sum_y \pi(y) = 1$, we can pass a limit through the sum to get

$$\lim_{t\to\infty} \big\|p^t(x,\cdot) - \pi\big\|_{TV} \le \lim_{t\to\infty} \sum_y \pi(y)\big\|p^t(x,\cdot) - p^t(y,\cdot)\big\|_{TV} = 0.$$

The trick in which one starts with a coupling $\big\{(X_k,Y_k)\big\}_{k\ge 0}$ and then defines a new coupling $\big\{(X_k,Z_k)\big\}_{k\ge 0}$ by forcing the trajectories to merge upon first colliding is common, but it can fail in certain cases.

Given any coupling $\big\{(X_k,Y_k)\big\}_{k\ge 0}$ with *coupling time* $\tau = \inf\big\{t : X_t = Y_t\big\}$, we can define the process $Z_k = Y_k 1\{\tau \ge k\} + X_k 1\{\tau < k\}$ having initial distribution $\mathscr{L}(Y_0)$, but the problem is that it may not be a Markov chain with transition function p.

A sufficient condition for this *sticky* construction to work is that $\big\{(X_k,Y_k)\big\}_{k\ge 0}$ is a *faithful coupling* of p, which means that it defines a Markov chain on $\mathcal{S}\times\mathcal{S}$ with transition probability q satisfying

(1) $\sum_{y'\in S} q\left((x,y),(x',y')\right) = p(x,x')$ for all $x, y, x' \in S$.
(2) $\sum_{x'\in S} q\left((x,y),(x',y')\right) = p(y,y')$ for all $x, y, y' \in S$.

This can be proved by conditioning on the value of τ to compute

$$P(Z_{k+1} = z, Z_0 = z_0, \ldots, , Z_{k-1} = z_{k-1}) = p(z_{k-1}, z)P(Z_0 = z_0, \ldots, , Z_{k-1} = z_{k-1}).$$

Most couplings one 'naturally' encounters are faithful, and it's typically a pretty easy condition to verify. Taking the coordinates to be independent copies of the chain (as in Example 2.3) will always yield a faithful coupling.

Also, we usually don't bother explicitly invoking $\{Z_k\}_{k\ge 0}$, appealing instead to the following immediate corollary of Lemma 2.1.

Corollary 2.2. *Suppose that* $\big\{(X_k,Y_k)\big\}_{k\ge 0}$ *is a faithful coupling of* p *with* $X_0 \sim \mu$, $Y_0 \sim \nu$, *and let* $\tau = \inf\big\{t : X_t = Y_t\big\}$. *Then*

$$\big\|\mu p^t - \nu p^t\big\|_{TV} \le \mathbb{P}\big(\tau > t\big).$$

In the above setting, it is common to start one copy of the chain in a specified distribution, let the other copy begin in stationarity, and then let them evolve according to the same transition mechanism until they meet and proceed simultaneously forever after. As the second chain was stationary to begin with, it remains so for all time, thus the first chain must have equilibrated by the time they couple.

Though the preceding story captures the intuition, it is not strictly correct as it overlooks a subtle point: $Y_t \sim \pi$ for all t does not guarantee that $Y_T \sim \pi$ for a random time T.

For example, consider the chain with state space $\{x, y\}$ and transition probabilities $P(x, y) = 1$, $P(y, x) = P(y, y) = \frac{1}{2}$. It is easy to see that $\pi(x) = \frac{1}{3}$, $\pi(y) = \frac{2}{3}$ is stationary for P. If we let $\{X_t\}$ be a copy of the chain started at y, let $\{Y_t\}$ be another copy of the chain with initial distribution π, and let T be the coupling time of $\{X_t\}$ and $\{Y_t\}$, then we necessarily have that $Y_T = y$ since $W_t = x$ implies that $W_{t-1} = y$ for any chain $\{W_t\}$ having transition probability P.

The theory of *strong stationary times* addresses this issue and provides another means of estimating the distance to stationarity. In the interest of time, we will leave this topic to independent pursuit.

We conclude with two examples illustrating the use of coupling in obtaining quantitative bounds on mixing.

Example 2.4 (Lazy Random Walk on the Hypercube)**.** In this example, $\mathcal{S} = (\mathbb{Z}/2\mathbb{Z})^d$, $P(x, x) = \frac{1}{2}$, $P(x, y) = \frac{1}{2d}$ if x and y differ in exactly one coordinate, and $P(x, z) = 0$ otherwise.

That is, at each step we flip a fair coin. If it comes up heads, we stay put. If it comes up tails, we choose one of our d neighbors uniformly at random and move there.

As an irreducible random walk on a finite group, the stationary distribution is uniform. The $\frac{1}{2}$ holding probabilities ensure aperiodicity, so the convergence theorem implies that the distribution of the position at time t will approach the uniform distribution as $t \to \infty$. We want to know how fast this happens.

To this end, we let $X_0 = x$ and let Y_0 have the uniform distribution on $\mathcal{S}$. Let $U_1, U_2, ...$ be i.i.d. uniform on $\{1, ..., d\}$ and let $V_1, V_2, ...$ be i.i.d. uniform on $\{0, 1\}$. All random variables are taken to be independent. At time t, the U_t^{th} coordinate of each chain is set to V_t and the others remain as they were.

In other words, at every time step we pick a coordinate at random and set its value (in both chains) to 0 or 1 according to the toss of a fair coin.

It is easy to see that $\{X_t\}$ and $\{Y_t\}$ are each evolving according to P. Moreover, the two chains agree in coordinate U_t from time t onward, so they have coupled by time $T = \inf\{t : \{U_1, ..., U_t\} = \{1, ..., d\}\}$. Since T does not depend on the initial state, Lemma 2.1 shows that

$$d(t) = \max_x \left\| P^t(x, \cdot) - \pi \right\|_{TV} \leq \mathbb{P}(T > t).$$

Thus the variation bound reduces to a 'coupon collector' problem:

If we let $A_k^t = \{k \notin \{U_1, ..., U_t\}\}$ for $k = 1, ..., d$, then

$$\mathbb{P}(T > t) = \mathbb{P}\left(\bigcup_{k=1}^{d} A_k^t \right) \leq dP(A_1^t) = d\left(1 - \frac{1}{d}\right)^t \leq de^{-\frac{t}{d}}.$$

Therefore, if $t = d\log(d) + cd$ where $c > 0$ is chosen so that $t \in \mathbb{N}$, then $d(t) \leq e^{-c}$. It follows that $t_{\text{mix}}(\varepsilon) \leq d\log(d) - d\log(\varepsilon)$. (We will see later that the correct rate is $\frac{1}{2}d\log(d)$, independent of ε.)

To help with our analysis of the final example of this subsection, we pause to consider the famous 'gambler's ruin' problem. The idea is that a player with an initial fortune of $\$k$ repeatedly wagers \$1 on a game of even odds until she wins $\$N$ or goes broke. Here $0 \leq k \leq N$.

We can represent the gambler's fortune over time as a Markov chain $\{X_t\}_{t\geq 0}$ with state space $\{0, 1, \ldots, N\}$, initial state $X_0 = k$, and transition matrix $P(0,0) = P(N,N) = 1$ and $P(j, j+1) = P(j, j-1) = 1/2$ for $0 < j < N$. (This is a *simple symmetric random walk* with *absorbing barriers* 0 and N.)

Let $T = \inf\{t : X_t \in \{0, N\}\}$ be the total number of games she plays.

Proposition 2.6. *In the gambler's ruin problem described above,*

$$\mathbb{P}_k(X_T = N) = \frac{k}{N} \quad \textit{and} \quad \mathbb{E}_k[T] = k(N-k).$$

Proof. Write $p_k = \mathbb{P}_k(X_T = N)$ for the probability that the gambler walks away victorious. Clearly $p_0 = 0$ and $p_N = 1$. By conditioning on the outcome of the first game, we see that for all $0 < k < N$,

$$p_k = \frac{1}{2}p_{k-1} + \frac{1}{2}p_{k+1}.$$

Subtracting $\frac{1}{2}(p_k + p_{k-1})$ from both sides shows that $p_k - p_{k-1} = p_{k+1} - p_k$. That is, the difference between successive values of p_k is constant. Since $p_0 = 0$, we see that $p_k = kp_1$. Substituting $p_N = 1$ into this identity gives $p_1 = 1/N$. It follows that $p_k = k/N$ as claimed.

Similarly, let $e_k = \mathbb{E}_k[T]$. Then $e_0 = e_N = 0$ and for $0 < k < N$,

$$e_k = \frac{1}{2}(1 + e_{k-1}) + \frac{1}{2}(1 + e_{k+1}).$$

(If the first game is a loss, which happens with probability 1/2, then one turn has been used up and the gambler starts over with a fortune of $\$(k-1)$. The second term is also an application of the Markov property, but corresponds to winning the first game.)

Subtracting $\frac{1}{2}(e_k + e_{k-1})$ from both sides shows that the increments $\Delta_k = e_k - e_{k-1}$ satisfy $\Delta_k = \Delta_{k+1} + 2$ and thus

$$\Delta_k = \Delta_1 - 2(k-1)$$

for $k = 1, \ldots, N$.

Taking the boundary conditions into account gives

$$\begin{aligned} 0 = e_N &= \sum_{k=1}^{N}(e_k - e_{k-1}) + e_0 = \sum_{k=1}^{N}\Delta_k \\ &= \sum_{k=1}^{N}\big(\Delta_1 - 2(k-1)\big) = N\Delta_1 - 2\sum_{j=0}^{N-1} j = N\Delta_1 - N(N-1). \end{aligned}$$

Therefore, $\Delta_1 = N - 1$ and $\Delta_j = (N-1) - 2(j-1)$, hence

$$\begin{aligned} e_k &= \sum_{j=1}^{k}\Delta_j = \sum_{j=1}^{k}[(N-1) - 2(j-1)] \\ &= (N-1)k - 2\sum_{i=0}^{k-1} i = (N-1)k - k(k-1) = k(N-k). \end{aligned}$$

□

Example 2.5 (Lazy Random Walk on the n-cycle)**.** Here we have $\mathcal{S} = \mathbb{Z}/n\mathbb{Z}$, $P(x,x) = \frac{1}{2}$, $P(x,y) = \frac{1}{4}$ if $y = x \pm 1 \,(\mathrm{mod}\, n)$, and $P(x,z) = 0$ otherwise.

We will construct a coupling consisting of two copies of the chain started at different states by flipping a fair coin to decide which of the chains to move according to simple symmetric random walk on the cycle.

More formally, set $X_0 = x$ and $Y_0 = y$ for some $x, y \in \mathcal{S}$, and let $U_1, U_2, \ldots, V_1, V_2, \ldots$ be i.i.d. uniform on $\{-1, 1\}$. If $U_t = 1$, then $X_{t+1} = X_t + V_t \,(\mathrm{mod}\, n)$ and $Y_{t+1} = Y_t$. If $U_t = -1$, then $X_{t+1} = X_t$ and $Y_{t+1} = Y_t + V_t \,(\mathrm{mod}\, n)$.

It is not hard to see that $\big\{(X_t, Y_t)\big\}_{t \geq 0}$ is a faithful coupling of P with coupling time $\tau = \inf\big\{t : X_t = Y_t\big\}$, so Corollary 2.2 gives

$$\big\|P^t(x,\cdot) - P^t(y,\cdot)\big\|_{TV} \leq \mathbb{P}_{x,y}\big(\tau > t\big).$$

If we let D_t denote the clockwise distance from X_t to Y_t, then D_t is a simple symmetric random walk on $\{0, 1, \ldots, n\}$ and $\big\{X_t = Y_t\big\} = \big\{D_t \in \{0, n\}\big\}$, so $\tau_{x,y}$ is the length of play in the gambler's ruin problem with initial fortune $k := y - x \,(\mathrm{mod}\, n)$. It follows from Proposition 2.6 and Chebychev's inequality that

$$\big\|P^t(x,\cdot) - P^t(y,\cdot)\big\|_{TV} \leq \mathbb{P}_{x,y}\big(\tau > t\big) \leq \frac{\mathbb{E}_{x,y}[\tau]}{t} = \frac{k(n-k)}{t}.$$

Maximizing over k shows that

$$d(t) \leq \bar{d}(t) \leq \frac{n^2}{4t},$$

so $t_{\mathrm{mix}} \leq n^2$.

3 Linear Representations of Finite Groups

One of the most powerful methods for obtaining sharp mixing time bounds for random walks on finite groups involves the use of representation theory, so we will take a bit of time to establish some background in the subject in preparation for the Fourier analysis to come.

3.1 Basic Notions

The general idea of representation theory is that one can study a group by letting it act linearly on a vector space. In this course, we will work exclusively with finite dimensional vector spaces over $\mathbb{C}$.

Formally, a *representation* of a finite group G is a pair (ρ, V) where V is a vector space and ρ is a homomorphism from G to $GL(V)$, the group of automorphisms of V.

Thus for every $s, t \in G$, we have $\rho(st) = \rho(s)\rho(t)$. Writing I for the identity map on V, this implies that $\rho(id) = I$ and $\rho(s^{-1}) = \rho(s)^{-1}$.

Since the codomain is part of the definition of a function, we will often just speak of the representation ρ. We call V the *representation space* and say that $d_\rho = \dim(V)$ is the *degree* or *dimension* of ρ.
Also, we will occasionally find it convenient to employ the notation $\rho_s := \rho(s)$.

When $V \cong \mathbb{C}^n$ comes equipped with a basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, we can represent $a \in GL(V)$ as the $n \times n$ matrix having j^{th} column $a(\mathbf{e}_j)$. (Linearity implies $a(c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_n) = c_1 a(\mathbf{e}_1) + \cdots + c_n a(\mathbf{e}_n)$.) In this view, a representation of G is a rule that associates an invertible matrix to each group element in a manner that respects the underlying structure.

We won't be doing anything too fancy in this class, so we can generally just think of the representation space as $\mathbb{C}^n$ with the standard basis and treat our representations as matrices.

Of course, the choice of basis is arbitrary, so let's say that representations (ρ, V) and (ρ', V') of G are *equivalent* if there is a linear bijection $\tau : V \to V'$ which satisfies

$$\tau \circ \rho_s = \rho'_s \circ \tau \quad \text{for all } s \in G.$$

Example 3.1. We always have the *trivial representation* $\rho_0(s) = 1$ for all $s \in G$

When $G = S_n$, another one-dimensional representation is the *sign representation* $\rho_\pm(\sigma) = \operatorname{sgn}(\sigma)$ where $\operatorname{sgn}(\sigma) = (-1)^m$ if σ can be expressed as a product of m transpositions (or if σ has m inversions).

Example 3.2. Suppose that $W = \operatorname{span}(\mathbf{w})$ is a one-dimensional subspace of $\mathbb{C}^d$ and let $\eta_0(s) = I_d$, the $d \times d$ identity matrix, for all $s \in G$. The map $\tau : W \to \mathbb{C}$ defined by $\tau(c\mathbf{w}) = c$ is a linear bijection satisfying

$$\rho_0(s)\tau(c\mathbf{w}) = \rho_0(s)c = c = \tau(c\mathbf{w}) = \tau\big(\eta_0(s)c\mathbf{w}\big),$$

so (η_0, W) is equivalent to $(\rho_0, \mathbb{C})$. Similarly, $\eta_\pm(\sigma) = \operatorname{sgn}(\sigma) I_d$ is equivalent to $\rho_\pm$.

In general, $\rho(s)\mathbf{v} = \mathbf{v}$ and $\rho(\sigma)\mathbf{v} = \operatorname{sgn}(\sigma)\mathbf{v}$ are valid representations for any vector space V. However, when $\dim(V) > 1$, these are direct sums of trivial/sign representations; see below.

Example 3.3. If $|G| = m$, the *left regular representation* is (λ, V) where V is an m-dimensional vector space with basis $\{\mathbf{e}_g\}_{g \in G}$ and λ satisfies $\lambda(g)\mathbf{e}_h = \mathbf{e}_{gh}$ for all $g, h \in G$. The *right regular representation* on V is given by $\rho(g)\mathbf{e}_h = \mathbf{e}_{hg^{-1}}$. The map defined by $\tau(\mathbf{e}_g) = \mathbf{e}_{g^{-1}}$ shows that λ and ρ are equivalent.

Example 3.4. If $G = S_n$, the *permutation representation* is defined by taking V to be a vector space with basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ and letting $\rho : S_n \to V$ be given by $\rho(\sigma)\mathbf{e}_k = \mathbf{e}_{\sigma(k)}$.
This associates to each $\sigma \in S_n$ the permutation matrix R_σ having (i,j)-entry $1\big\{\sigma(j) = i\big\}$. For any $\mathbf{x} \in \mathbb{C}^n$, $R_\sigma \mathbf{x}$ has k^{th} coordinate $x_{\sigma^{-1}(k)}$.

More generally if G acts on a set $X = \{x_1, \ldots, x_n\}$ (so there is a map $\varphi : G \times X \to X$ satisfying $\varphi(id, x) = x$ and $\varphi(gh, x) = \varphi\big(g, \varphi(h, x)\big)$ for all $x \in X$, $g, h \in G$), and V is a vector space with basis $\{\mathbf{e}_x\}_{x \in X}$, the associated permutation representation (ρ, V) is defined by $\rho(g)\mathbf{e}_x = \mathbf{e}_{\varphi(g,x)}$.

The (left) regular representation is the special case $X = G$, $\varphi(g, h) = gh$.

If (ρ, V) is a representation and W is a subspace of V which is *stable* under ρ (so $\rho(s)\mathbf{w} \in W$ for every $s \in G$, $\mathbf{w} \in W$), the restriction of ρ to W gives a *subrepresentation*. We always have the subrepresentations corresponding to $W = V$ and $W = \{\mathbf{0}\}$. If ρ admits no other subrepresentations, we say that it is *irreducible*.

Now recall that V is said to be the *direct sum* of $W_1, W_2 \leq V$ (written $V = W_1 \oplus W_2$) if every $\mathbf{v} \in V$ can be uniquely expressed as $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$ with $\mathbf{w}_1 \in W_1$ and $\mathbf{w}_2 \in W_2$.
This is equivalent to requiring that $W_1 \cap W_2 = \{\mathbf{0}\}$ and $\dim(V) = \dim(W_1) + \dim(W_2)$.

(We can also form the *external direct sum* of vector spaces U and V as the vector space consisting of ordered pairs in $U \times V$ with all operations performed componentwise.)

The *direct sum of representations* (ρ^1, W_1) and (ρ^2, W_2) is the representation $(\rho^1 \oplus \rho^2, W_1 \oplus W_2)$ defined by $(\rho^1 \oplus \rho^2)_s(\mathbf{w}_1 + \mathbf{w}_2) = \rho^1_s(\mathbf{w}_1) + \rho^2_s(\mathbf{w}_2)$.
(For external direct sums, the analogous definition is $(\rho^1 \oplus \rho^2)_s\big(\mathbf{w}_1, \mathbf{w}_2\big) = \big(\rho^1_s(\mathbf{w}_1), \rho^2_s(\mathbf{w}_2)\big)$.)
By construction, $\rho^1 \oplus \rho^2$ has degree $d_{\rho^1 \oplus \rho^2} = d_{\rho^1} + d_{\rho^2}$.

The direct sum of more than two representations is defined by $\rho^1 \oplus \cdots \oplus \rho^{k+1} = (\rho^1 \oplus \cdots \oplus \rho^k) \oplus \rho^{k+1}$.
If we think of $\rho^1(s), \ldots, \rho^k(s)$ as matrices, then we can express the direct sum as the block diagonal matrix

$$\rho^1 \oplus \cdots \oplus \rho^k(s) = \begin{bmatrix} \rho^1(s) & & O \\ & \ddots & \\ O & & \rho^k(s) \end{bmatrix}.$$

Here we are assuming that a basis of $\bigoplus_{i=1}^k W_i$ is given by $\{\mathbf{e}^1_1, \ldots, \mathbf{e}^1_{d_1}, \ldots, \mathbf{e}^k_1, \ldots, \mathbf{e}^k_{d_k}\}$ with $\{\mathbf{e}^i_1, \ldots, \mathbf{e}^i_{d_i}\}$ the corresponding basis for W_i.

Example 3.5. Let $G = S_3$ and $W = \big\{\mathbf{x} \in \mathbb{C}^3 : x_1 + x_2 + x_3 = 0\big\}$. A basis for W is given by $\mathbf{w}_1 = \mathbf{e}_1 - \mathbf{e}_2$ and $\mathbf{w}_2 = \mathbf{e}_2 - \mathbf{e}_3$ where $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are the standard basis vectors in $\mathbb{C}^3$. W is stable under the permutation representation $(\rho, \mathbb{C}^3)$ since permuting the coordinates of a vector does not change their sum.

I claim that the *standard representation* (ρ, W) is irreducible. Indeed every nontrivial subspace of W is of the form $W' = \text{span}(\mathbf{w})$ for some nonzero $\mathbf{w} = (x, y, z)$ in W.

Without loss of generality, assume that $x \neq 0$ so that $(1, y', z') \in W'$. If W' were stable under ρ, then we would also have $(y', 1, z')$ and thus $(1 - y', y' - 1, 0)$ in W'.

If $y' \neq 1$, this implies that $\mathbf{e}_1 - \mathbf{e}_2$ and thus $\mathbf{e}_2 - \mathbf{e}_3$ are in W', hence $W' = W$.
If $y' = 1$, we would have $(1, 1, -2) \in W'$ (as the coordinates must sum to 0), so $(1, -2, 1)$ and thus $(0, 3, -3)$ are in W', which implies that $\mathbf{e}_2 - \mathbf{e}_3$ and thus $\mathbf{e}_1 - \mathbf{e}_2$ are in W'.

We can express $\rho(\pi)$ in matrix form by computing

π	$\rho(\pi)\mathbf{w}_1$	$\rho(\pi)\mathbf{w}_2$	$\rho(\pi)$
id	$(1,-1,0) = \mathbf{w_1}$	$(0,1,-1) = \mathbf{w_2}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$(1\,2)$	$(-1,1,0) = -\mathbf{w}_1$	$(1,0,-1) = \mathbf{w}_1 + \mathbf{w}_2$	$\begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}$
$(1\,3)$	$(0,-1,1) = -\mathbf{w}_2$	$(-1,1,0) = -\mathbf{w}_1$	$\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$
$(2\,3)$	$(1,0,-1) = \mathbf{w}_1 + \mathbf{w}_2$	$(0,-1,1) = -\mathbf{w}_2$	$\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$
$(1\,2\,3)$	$(0,1,-1) = \mathbf{w}_2$	$(-1,0,1) = -\mathbf{w}_1 - \mathbf{w_2}$	$\begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}$
$(3\,2\,1)$	$(-1,0,1) = -\mathbf{w}_1 - \mathbf{w_2}$	$(1,-1,0) = \mathbf{w_1}$	$\begin{bmatrix} -1 & 1 \\ -1 & 0 \end{bmatrix}$

Observe that the orthogonal complement of W in $\mathbb{C}^3$ is $W^\perp = \text{span}(\mathbf{1})$. This one-dimensional subspace carries the trivial representation and we can form the direct sum $\rho_0 \oplus \rho$.

Relative to the basis $\mathcal{B} = \{\mathbf{1}, \mathbf{w}_1, \mathbf{w}_2\}$, this has matrix form

$$\rho_0 \oplus \rho\big((1\,3)\big) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \; \rho_0 \oplus \rho\big((3\,2\,1)\big) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & -1 & 0 \end{bmatrix}, \ldots$$

To express these in the standard basis $\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, we must conjugate with the change of basis matrix $P_{\mathcal{E}\leftarrow\mathcal{B}}$ whose j^{th} column is the standard coordinates of the j^{th} vector in $\mathcal{B}$.

This gives the equivalent matrix representations

$$\begin{aligned}
(\rho_0 \oplus \rho)'\big((1\,3)\big) &= \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \\
(\rho_0 \oplus \rho)'\big((3\,2\,1)\big) &= \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \\
&\vdots
\end{aligned}$$

which we recognize as the permutation representation!

Our next order of business is to show that the irreducible representations are the building blocks of all others in the sense that every representation is a direct sum of irreducible representations.

To this end, we record the following proposition.

Proposition 3.1. *Let $\rho : G \to GL(V)$ be a representation and suppose that $W \leq V$ is stable under ρ. Then there exists a complement $W' \leq V$ such that $V = W \oplus W'$ and W' is stable under ρ.*

Proof. Let $\langle \cdot, \cdot \rangle$ be an inner product on V and define a new inner product $\langle \cdot, \cdot \rangle_\rho$ by

$$\langle \mathbf{x}, \mathbf{y} \rangle_\rho = \sum_{s \in G} \langle \rho(s)\mathbf{x}, \rho(s)\mathbf{y} \rangle .$$

This is indeed conjugate-symmetric, linear in the first argument, and positive-definite since $\langle \cdot, \cdot \rangle$ is an inner product and $\rho(s)$ is invertible. Moreover, it is invariant under ρ in the sense that

$$\begin{aligned} \langle \rho(t)\mathbf{x}, \rho(t)\mathbf{y} \rangle_\rho &= \sum_{s \in G} \langle \rho(s)\rho(t)\mathbf{x}, \rho(s)\rho(t)\mathbf{y} \rangle = \sum_{s \in G} \langle \rho(st)\mathbf{x}, \rho(st)\mathbf{y} \rangle \\ &= \sum_{u \in G} \langle \rho(u)\mathbf{x}, \rho(u)\mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle_\rho . \end{aligned}$$

Let W' be the orthogonal complement of W with respect to this inner product. Then $V = W \oplus W'$ and W' is stable under ρ because for any $\mathbf{x} \in W'$, $\mathbf{y} \in W$, $t \in G$, we have $\mathbf{z} = \rho(t^{-1})\mathbf{y} \in W$ and thus

$$\begin{aligned} \langle \rho(t)\mathbf{x}, \mathbf{y} \rangle_\rho &= \sum_{s \in G} \langle \rho(s)\rho(t)\mathbf{x}, \rho(s)\mathbf{y} \rangle = \sum_{s \in G} \left\langle \rho(st)\mathbf{x}, \rho(st)\rho(t^{-1})\mathbf{y} \right\rangle \\ &= \sum_{u \in G} \langle \rho(u)\mathbf{x}, \rho(u)\mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle_\rho = 0. \end{aligned}$$

□

Remark 3.1. Note that the invariance of $\langle \cdot, \cdot \rangle_\rho$ means that if $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ is an orthonormal basis of V with respect to $\langle \cdot, \cdot \rangle_\rho$, then $\langle \rho(s)\mathbf{f}_i, \rho(s)\mathbf{f}_j \rangle_\rho = \delta_{ij}$ for all $s \in G$, $i, j \in [n]$.

Also, if $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is an orthonormal basis of V with respect to $\langle \cdot, \cdot \rangle$ and M is the linear transformation defined by $M\mathbf{e}_i = \mathbf{f}_i$, then $\langle M\mathbf{e}_i, M\mathbf{e}_j \rangle_\rho = \langle \mathbf{f}_i, \mathbf{f}_j \rangle_\rho = \delta_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle$, hence $\langle M\mathbf{u}, M\mathbf{v} \rangle_\rho = \langle \mathbf{u}, \mathbf{v} \rangle$ by linearity.

It follows that the equivalent representation $\tau = M^{-1}\rho M$ satisfies

$$\langle \tau(s)\mathbf{e}_i, \tau(s)\mathbf{e}_j \rangle = \langle M\tau(s)\mathbf{e}_i, M\tau(s)\mathbf{e}_j \rangle_\rho = \langle \rho(s)M\mathbf{e}_i, \rho(s)M\mathbf{e}_j \rangle_\rho = \langle \rho(s)\mathbf{f}_i, \rho(s)\mathbf{f}_j \rangle_\rho = \delta_{ij}$$

and thus is unitary with respect to $\langle \cdot, \cdot \rangle$.

As such, we can always assume that our representations are unitary.

We are now able to prove the following extremely powerful result which enables us to study representations by breaking them up into their irreducible components.

Theorem 3.1 (Maschke's Theorem)**.** *Every representation is a direct sum of irreducible representations.*

Proof. If $d_\rho = 1$, then ρ is irreducible since V has no nontrivial subspaces. Now assume that the result holds for all representations of degree at most k and let $d_\rho = k + 1$. If (ρ, V) is irreducible, then we are done. Otherwise, there is a stable subspace $W < V$ and Proposition 3.1 gives $V = W \oplus W'$ with $\dim(W), \dim(W') \leq k$. The induction hypothesis shows that W and W' are direct sums of irreps and the result follows by the principle of induction. □

The direct sum construction gives us a means of constructing new representations of G from old ones. The other main way of doing this is by taking tensor products.

The *tensor product* of vector spaces U and V is the space $U \otimes V$ consisting of formal linear combinations of symbols of the form $\mathbf{u} \otimes \mathbf{v}$ (with $\mathbf{u} \in U$, $\mathbf{v} \in V$) subject to the relations

$$(\alpha \mathbf{u}_1 + \beta \mathbf{u}_2) \otimes \mathbf{v} = \alpha \mathbf{u}_1 \otimes \mathbf{v} + \beta \mathbf{u_2} \otimes \mathbf{v},$$
$$\mathbf{u} \otimes (\alpha \mathbf{w}_1 + \beta \mathbf{w}_2) = \alpha \mathbf{u} \otimes \mathbf{w}_1 + \beta \mathbf{u} \otimes \mathbf{w}_2.$$

If $\{\mathbf{e}_i\}_{i \in [m]}$ and $\{\mathbf{f}_j\}_{j \in [n]}$ are bases for U and V, then a basis for $U \otimes V$ is given by $\{\mathbf{e}_i \otimes \mathbf{f}_j\}_{i \in [m], j \in [n]}$.

The *tensor product of representations* (ρ, U) and (η, V) is the representation $(\rho \otimes \eta, U \otimes V)$ defined by $(\rho \otimes \eta)_s(\mathbf{u} \otimes \mathbf{v}) = \rho_s(\mathbf{u}) \otimes \eta_s(\mathbf{v})$, having degree $d_{\rho \otimes \eta} = d_\rho d_\eta$.
If $\rho(s)$ and $\eta(s)$ are in matrix form relative to $\{\mathbf{e}_i\}_{i \in [m]}$ and $\{\mathbf{f}_j\}_{j \in [n]}$, then relative to $\{\mathbf{e}_i \otimes \mathbf{f}_j\}_{i \in [m], j \in [n]}$, their tensor product has (block) matrix form

$$\rho \otimes \eta(s) = \begin{bmatrix} \rho(s)_{11}\eta(s) & \rho(s)_{12}\eta(s) & \cdots & \rho(s)_{1d_\rho}\eta(s) \\ \rho(s)_{21}\eta(s) & \rho(s)_{22}\eta(s) & \cdots & \rho(s)_{2d_\rho}\eta(s) \\ \vdots & & \ddots & \vdots \\ \rho(s)_{d_\rho 1}\eta(s) & \rho(s)_{d_\rho 2}\eta(s) & \cdots & \rho(s)_{d_\rho d_\rho}\eta(s) \end{bmatrix}.$$

3.2 Characters

Recall that if V is a vector space with basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, then $a \in GL(V)$ can be represented as the matrix $[a_{i,j}]$ having j^{th} column $a(\mathbf{e}_j)$. This enables us to define the *trace* of a as $\text{Tr}(a) = \sum_{k=1}^n a_{k,k}$.
It's not hard to see that the trace does not depend on the basis chosen. In fact, it is equal to the sum of the eigenvalues (with multiplicity) of a.

We define the *character* of a representation $\rho : G \to GL(V)$ as the function $\chi_\rho : G \to \mathbb{C}$ given by

$$\chi_\rho(s) = \text{Tr}\big(\rho(s)\big).$$

If ρ is an irreducible representation, we call χ_ρ an *irreducible character*. If $d_\rho = 1$, then $\chi_\rho = \rho$ is called a *linear character*.

Characters are extremely useful objects. One reason for this is that they retain a lot of information about the associated representation even though they are scalar- rather than matrix-valued.

Example 3.6. If ρ is the n-dimensional permutation representation of S_n from Example 3.4, then

$$\chi_\rho(\sigma) = \sum_{k=1}^n \rho(\sigma)_{k,k} = \sum_{k=1}^n 1\{\sigma(k) = k\}$$

gives the number of fixed points of σ.

Proposition 3.2. *If χ is the character of a representation ρ having degree d, then*

(1) $\chi(id) = d$
(2) $\chi(s^{-1}) = \overline{\chi(s)}$
(3) $\chi(sts^{-1}) = \chi(t)$

Proof. The first assertion follows from the fact that $\rho(id) = I_d$.

For the second, if $o(s) = m$, then $\rho(s)^m = \rho(s^m) = \rho(id) = I_d$, hence the eigenvalues of $\rho(s)$ must be m^{th} roots of unity. (This is also a consequence of the fact that we can choose a basis in which our representations are unitary.) It follows that

$$\begin{aligned}\overline{\chi(s)} = \operatorname{Tr}\big(\overline{\rho(s)}\big) &= \sum_{k=1}^{d} \overline{\lambda_k} = \sum_{k=1}^{d} \lambda_k^{-1} \\ &= \operatorname{Tr}\big(\rho(s)^{-1}\big) = \operatorname{Tr}\big(\rho(s^{-1})\big) = \chi(s^{-1}).\end{aligned}$$

Finally since $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$,

$$\begin{aligned}\chi(sts^{-1}) = \operatorname{Tr}\big(\rho(sts^{-1})\big) &= \operatorname{Tr}\big(\rho(s)\rho(t)\rho(s)^{-1}\big) \\ &= \operatorname{Tr}\big(\rho(s)^{-1}\rho(s)\rho(t)\big) = \operatorname{Tr}\big(\rho(t)\big) = \chi(t).\end{aligned}$$

□

For our next results about characters, we need to establish the following surprisingly useful fact known as *Schur's Lemma*.

Lemma 3.1. *Let (ρ^1, V_1) and (ρ^2, V_2) be irreducible representations of G, and suppose that $f : V_1 \to V_2$ is a linear map that satisfies*

$$f \circ \rho_s^1 = \rho_s^2 \circ f \text{ for all } s \in G.$$

(1) If ρ^1 and ρ^2 are not equivalent, then $f \equiv \mathbf{0}$.

(2) If $V_1 = V_2$ and $\rho^1 = \rho^2$, then f is a scalar multiple of the identity.

Proof. Note that $\ker(f) = \big\{\mathbf{v} \in V_1 : f(\mathbf{v}) = \mathbf{0}\big\}$ is stable under ρ^1 since $\mathbf{v} \in \ker(f)$ implies

$$f\big(\rho_s^1(\mathbf{v})\big) = \rho_s^2\big(f(\mathbf{v})\big) = \rho_s^2(\mathbf{0}) = \mathbf{0}.$$

Similarly, $\operatorname{Im}(f) = \big\{\mathbf{w} \in V_2 : \mathbf{w} = f(\mathbf{v}) \text{ for some } \mathbf{v} \in V_1\big\}$ is stable under ρ^2 since $\mathbf{w} = f(\mathbf{v})$ implies that

$$\rho_s^2(\mathbf{w}) = \rho_s^2\big(f(\mathbf{v})\big) = f\big(\rho_s^1(\mathbf{v})\big).$$

Thus by irreducibility, $\ker(f)$ is either $\{\mathbf{0}\}$ or V_1 and $\operatorname{Im}(f)$ is either $\{\mathbf{0}\}$ or V_2.

It follows that if $f \not\equiv \mathbf{0}$, then $\ker(f) = \{\mathbf{0}\}$ and $\operatorname{Im}(f) = V_2$, so f is a bijection and the representations are equivalent.

Now suppose that $V_1 = V_2$ and $\rho^1 = \rho^2$. The claim certainly holds if $f \equiv \mathbf{0}$. Otherwise, f has a nonzero eigenvalue λ. In this case, the map $f_\lambda = f - \lambda I$ has a nontrivial kernel and satisfies $f_\lambda \circ \rho_s^1 = \rho_s^2 \circ f_\lambda$, hence $f_\lambda \equiv \mathbf{0}$. (Observe that it is important here that we are working over an algebraically closed field.) □

Corollary 3.1. *Let (ρ^1, V_1) and (ρ^2, V_2) be irreducible representations of G, write $d = \dim(V_1)$, and let $h : V_1 \to V_2$ be a linear map. Define*

$$\widetilde{h} = \frac{1}{|G|} \sum_{s \in G} (\rho_s^2)^{-1} \circ h \circ \rho_s^1. \tag{3.1}$$

(1) If ρ^1 and ρ^2 are not equivalent, then $\widetilde{h} \equiv \mathbf{0}$.

(2) If $V_1 = V_2$ and $\rho^1 = \rho^2$, then $\widetilde{h} = \lambda I$ with $\lambda = \operatorname{Tr}(h)/d$.

Proof. For any $t \in G$,

$$\begin{aligned}(\rho_t^2)^{-1} \circ \widetilde{h} \circ \rho_t^1 &= \frac{1}{|G|} \sum_{s \in G} (\rho_t^2)^{-1} (\rho_s^2)^{-1} \circ h \circ \rho_s^1 \rho_t^1 \\ &= \frac{1}{|G|} \sum_{s \in G} (\rho_{st}^2)^{-1} \circ h \circ \rho_{st}^1 = \widetilde{h},\end{aligned}$$

hence $\widetilde{h} \circ \rho_t^1 = \rho_t^2 \circ \widetilde{h}$.

If ρ^1 and ρ^2 are not equivalent, then Schur's lemma implies that $\widetilde{h} \equiv \mathbf{0}$.

If $V_1 = V_2$ and $\rho^1 = \rho^2$, Schur's lemma ensures that $\widetilde{h} = \lambda I$, and taking the trace of both sides in Equation (3.1) shows that $\lambda = \mathrm{Tr}(h)/d$. □

Let us now suppose that our representations are given in the matrix form $\rho_s^1 = [r_{i,j}(s)]$, $\rho_s^2 = [q_{i,j}(s)]$. Writing the linear maps from Corollary 3.1 as $h = [x_{i,j}]$, $\widetilde{h} = [\widetilde{x}_{i,j}]$, Equation (3.1) can be expressed entrywise as

$$\widetilde{x}_{i,\ell} = \frac{1}{|G|} \sum_{s,j,k} q_{i,j}(s^{-1}) x_{j,k} r_{k,\ell}(s). \tag{3.2}$$

In the first case, $\widetilde{x}$ is the zero matrix for every choice of x—such as those with a single entry equal to 1 and all others 0—so we must have

$$\frac{1}{|G|} \sum_{s} q_{i,j}(s^{-1}) r_{k,\ell}(s) = 0.$$

In the second case, $\widetilde{x}_{i,\ell} = \lambda \delta_{i\ell}$ with $\lambda = \frac{1}{d} \sum_{j,k} x_{j,k} \delta_{jk}$. Substituting this into Equation (3.2) yields

$$\frac{1}{d} \sum_{j,k} x_{j,k} \delta_{jk} \delta_{i\ell} = \frac{1}{|G|} \sum_{s,j,k} r_{i,j}(s^{-1}) x_{j,k} r_{k,\ell}(s).$$

As this holds for all choices of x, we can equate coefficients to obtain

$$\frac{1}{|G|} \sum_{s} r_{i,j}(s^{-1}) r_{k,\ell}(s) = \frac{1}{d} \delta_{jk} \delta_{i\ell}.$$

Recalling that we can choose bases so that our representations are unitary and thus satisfy $r_{i,j}(s^{-1}) = \overline{r_{j,i}(s)}$ (and employing the reindexing $s \mapsto s^{-1}$, $k \leftrightarrow \ell$ for the sake of aesthetics), we record the foregoing as

Corollary 3.2. *Let (ρ^1, V_1) and (ρ^2, V_2) be irreducible representations of G having (unitary) matrix form $\rho_s^1 = [r_{i,j}(s)]$, $\rho_s^2 = [q_{i,j}(s)]$, and write $d = \dim(V_1)$. Then for all valid indices i, j, k, ℓ*

(1) If ρ^1 and ρ^2 are not equivalent,

$$\frac{1}{|G|} \sum_{s} q_{i,j}(s) \overline{r_{k,\ell}(s)} = 0.$$

(2) If $V_1 = V_2$ and $\rho^1 = \rho^2$,

$$\frac{1}{|G|} \sum_{s} r_{i,j}(s) \overline{r_{k,\ell}(s)} = \begin{cases} \frac{1}{d}, & i = k \text{ and } j = \ell \\ 0, & \text{otherwise} \end{cases}.$$

That is, the matrix entries of the irreducible representations are orthogonal with respect to the inner product

$$(f \,|\, g) = \frac{1}{|G|} \sum_{s \in G} f(s) \overline{g(s)}, \quad f, g : G \to \mathbb{C}.$$

One immediate consequence of this observation is that there are only finitely many irreducible representations of a finite group G since $\dim\left(\mathbb{C}^G\right) = |G|$. (We will say more about this shortly.)

Another is the *first orthogonality relation* given below.

Theorem 3.2. *The irreducible characters are orthonormal with respect to* $(\cdot\,|\,\cdot)$.

Proof. Let ρ be an irreducible representation of degree d with $[r_{i,j}(t)] = \rho(t)$ a unitary matrix. The associated character is $\chi_\rho(t) = \sum_{k=1}^{d} r_{k,k}(t)$, and Corollary 3.2 implies

$$\begin{aligned}(\chi_\rho\,|\,\chi_\rho) &= \frac{1}{|G|}\sum_{s\in G}\chi_\rho(s)\overline{\chi_\rho(s)} \\ &= \frac{1}{|G|}\sum_{s\in G}\sum_{k=1}^{d} r_{k,k}(s)\sum_{\ell=1}^{d}\overline{r_{\ell,\ell}(s)} \\ &= \sum_{k,\ell}(r_{k,k}\,|\,r_{\ell,\ell}) = \sum_{k,\ell}\delta_{k\ell}/d = 1.\end{aligned}$$

Similarly, if η is an inequivalent irrep with $[q_{i,j}(t)] = \eta(t)$ a unitary matrix, then

$$(\chi_\rho\,|\,\chi_\eta) = \sum_{k,\ell}(r_{k,k}\,|\,q_{\ell,\ell}) = 0. \qquad \square$$

We can now say a bit more about the direct sum decomposition from Theorem 3.1.

Proposition 3.3. *Let (ρ, V) be a representation of G, and suppose that $V = W_1\oplus\cdots\oplus W_k$ is a decomposition of V into irreducible components. If (η, W) is an irreducible representation of G, then the number of W_i which are equivalent to W is $(\chi_\rho\,|\,\chi_\eta)$.*

Proof. Since the character of a direct sum is the sum of the constituent characters (see Homework 4),

$$(\chi_\rho\,|\,\chi_\eta) = (\chi_1\,|\,\chi_\eta) + \cdots + (\chi_k\,|\,\chi_\eta)$$

with χ_i the character of W_i. The result follows since $(\chi_i\,|\,\chi_\eta)$ is 1 if $W_i \cong W$ and 0 otherwise. $\square$

An upshot of this result is that the multiplicity of W in V does not depend on the chosen decomposition.

Corollary 3.3. *Representations with the same character are equivalent.*

Proof. Both contain the same irreps with the same multiplicity. $\square$

Corollary 3.4. *For any representation (ρ, V), $(\chi_\rho\,|\,\chi_\rho)$ is a positive integer which equals 1 iff ρ is irreducible.*

Proof. Let $V = n_1V_1 \oplus \cdots \oplus n_mV_m$ be a direct sum decomposition of V into irreducible components. Here $V_1, \ldots, V_m$ is a complete list of the irreps and $n_i \in \mathbb{N}_0$ is the number of copies of V_i in V.

Writing χ_i for the character corresponding to V_i, we have

$$(\chi_\rho\,|\,\chi_\rho) = \left(\sum_{i=1}^{m} n_i\chi_i \,\middle|\, \sum_{j=1}^{m} n_j\chi_j\right) = \sum_{i,j} n_in_j\,(\chi_i\,|\,\chi_j) = \sum_{i=1}^{m} n_i^2.$$

This is a positive integer that equals 1 if and only if some n_i is 1 and the rest are 0. $\square$

Example 3.7. Let ρ be the n-dimensional permutation representation of S_n. Arguing as in Example 3.5, the subspaces $W = \{\mathbf{x} \in \mathbb{C}^n : x_1 + \cdots + x_n = 0\}$ and $W^\perp = \text{span}(\mathbf{1})$ are stable under ρ, so ρ is the direct sum of the standard representation and the trivial representation. The latter is irreducible since it is one-dimensional. If we are able to show that $(\chi_\rho \,|\, \chi_\rho) = 2$, then we can conclude that the standard representation is irreducible as well.

To this end, let $X = \sum_{i=1}^n 1\{\sigma(i) = i\}$ be the random variable that records the number of fixed points in a permutation σ drawn uniformly from S_n. We saw in Example 3.6 that $\chi_\rho(\sigma)$ gives the number of fixed points of σ. Since σ and σ^{-1} have the same number of fixed points,

$$(\chi_\rho \,|\, \chi_\rho) = \frac{1}{n!} \sum_{\sigma \in S_n} \chi_\rho(\sigma)\chi_\rho(\sigma^{-1}) = \mathbb{E}[X^2].$$

The desired result follows since $X^2 = \sum_{i=1}^n 1\{\sigma(i) = i\} + \sum_{i \neq j} 1\{\sigma(i) = i, \sigma(j) = j\}$ has expectation

$$\begin{aligned} E[X^2] &= \sum_{i=1}^n \mathbb{P}\big(\sigma(i) = i\big) + \sum_{i \neq j} \mathbb{P}\big(\sigma(i) = i, \sigma(j) = j\big) \\ &= n\frac{(n-1)!}{n!} + n(n-1)\frac{(n-2)!}{n!} = 2. \end{aligned}$$

It turns out to be quite instructive to play the same sort of game with the regular representation of an arbitrary finite group G.

Recall from Example 3.3 that this is the representation λ defined by $\lambda(g)\mathbf{e}_h = \mathbf{e}_{gh}$ for $\{\mathbf{e}_s\}_{s \in G}$ a basis of the representation space. It has degree $d_\lambda = |G|$ and matrix form (with respect to $\{\mathbf{e}_s\}_{s \in G}$) $\ell_{g,h}(s) = 1\{sh = g\}$. Since $sg \neq g$ for $s \neq id$, the character of the regular representation is

$$\chi_\lambda(s) = \sum_{g \in G} \ell_{g,g}(s) = \begin{cases} |G|\,, & s = id \\ 0, & \text{otherwise} \end{cases}.$$

Proposition 3.4. *Every irreducible representation is contained in the regular representation with multiplicity equal to its degree.*

Proof. Let λ be the regular representation of G and let ρ be an irreducible representation. Then

$$(\chi_\lambda \,|\, \chi_\rho) = \frac{1}{|G|} \sum_{s \in G} \chi_\lambda(s)\chi_\rho(s^{-1}) = \frac{1}{|G|}\chi_\lambda(id)\chi_\rho(id) = \frac{1}{|G|}\,|G|\,d_\rho = d_\rho. \qquad \square$$

Corollary 3.5. *Let $\rho_1, \ldots, \rho_m$ be a complete list of the irreducible representations of G with ρ_k having character χ_k and degree d_k.*

(1) $\sum_{k=1}^m d_k^2 = |G|$

(2) For $s \neq id$, $\sum_{k=1}^m d_k\chi_k(s) = 0$

Proof. Proposition 3.4 shows that the regular representation has character $\chi(s) = \sum_{k=1}^m d_k\chi_k(s)$.

Taking $s = id$ gives (1), and taking $s \neq id$ gives (2). $\square$

The preceding results give upper bounds on the degrees and number of irreducible representations of G, as well as a criterion for checking that one has an exhaustive list of the irreps.

Also, since we know that the entries of the irreducible representations in unitary matrix form are orthogonal with the entries of ρ_k having norm d_k^{-1}, the fact that there are $\sum_{k=1}^m d_k^2 = |G|$ of them gives

Proposition 3.5. *Let $[r_{i,j}^k]$ be the matrix form of the irreducible representation ρ_k with respect to a basis that makes it unitary. Then an orthonormal basis for $\mathbb{C}^G$ is given by $\left\{\sqrt{d_k} r_{i,j}^k : k \in [m],\, i,j \in [d_k]\right\}$.*

3.3 Further Results

At this point, it is natural to define the *Fourier transform* of $f : G \to \mathbb{C}$ at the representation ρ as the $d_\rho \times d_\rho$ matrix

$$\widehat{f}(\rho) = \sum_{s \in G} f(s)\rho(s).$$

This is a generalization of the discrete Fourier transform ($G = \mathbb{Z}/n\mathbb{Z}$) and, suitably interpreted, the standard Fourier transform for functions on $\mathbb{R}$. As such, it possesses many familiar properties.

Proposition 3.6. *The convolution of functions $f, g : G \to \mathbb{C}$ has Fourier transform*

$$\widehat{f * g}(\rho) = \widehat{f}(\rho)\widehat{g}(\rho).$$

Proof. Multiplying by $\rho(t^{-1})\rho(t)$ and making the change of variables $u = st^{-1}$ gives

$$\begin{aligned}
\widehat{f * g}(\rho) &= \sum_{s \in G} (f * g)(s)\rho(s) = \sum_{s \in G} \sum_{t \in G} f(st^{-1})g(t)\rho(s) \\
&= \sum_{s \in G} \sum_{t \in G} f(st^{-1})g(t)\rho(st^{-1})\rho(t) \\
&= \sum_{u \in G} f(u)\rho(u) \sum_{t \in G} g(t)\rho(t) = \widehat{f}(\rho)\widehat{g}(\rho).
\end{aligned}$$

□

The fact that Fourier transforms take convolutions into products is immensely useful in practice. Together with our next result, it forms much of the basis for the spectral analysis in the following section.

The first part shows that a function is completely determined by its Fourier transforms at irreps and gives a rule for recovering the function from its transforms.
The second can be thought of as relating 'inner products' in the time and frequency domains.

Theorem 3.3. *Let $\rho_1, \ldots, \rho_m$ be the irreducible representations of G with $d_1, \ldots, d_m$ the corresponding degrees. Then for any $f, g : G \to \mathbb{C}$, we have*

Fourier Inversion Formula:

$$f(s) = \frac{1}{|G|} \sum_{i=1}^m d_i \mathrm{Tr}\big(\rho_i(s^{-1})\widehat{f}(\rho_i)\big)$$

Plancherel Formula:

$$\sum_{s \in G} f(s)g(s^{-1}) = \frac{1}{|G|} \sum_{i=1}^m d_i \mathrm{Tr}\big(\widehat{f}(\rho_i)\widehat{g}(\rho_i)\big)$$

Proof. Since both sides of the above equations are linear in f, it suffices to prove the result for $f(s) = \delta_{st}$, in which case $\widehat{f}(\rho_i) = \rho_i(t)$.

For the inversion formula, the right-hand side is then

$$\frac{1}{|G|}\sum_{i=1}^{m} d_i \mathrm{Tr}\big(\rho_i(s^{-1})\rho_i(t)\big) = \frac{1}{|G|}\sum_{i=1}^{m} d_i \chi_i(s^{-1}t) = \delta_{st}$$

by Corollary 3.5.

For the Plancherel formula, we must show that

$$g(t^{-1}) = \frac{1}{|G|}\sum_{i=1}^{m} d_i \mathrm{Tr}\big(\rho_i(t)\widehat{g}(\rho_i)\big),$$

and this follows immediately from the inversion formula. □

Now recall that $u, t \in G$ are said to be *conjugate* if $u = sts^{-1}$ for some $s \in G$. It is easy to check that conjugacy is an equivalence relation, so the group is partitioned into conjugacy classes.

We say that $f : G \to \mathbb{C}$ is a *class function* if it is constant on conjugacy classes—that is, $f(sts^{-1}) = f(t)$ for all $s, t \in G$.

Proposition 3.7. *If f is a class function on G, then its Fourier transform at an irreducible representation ρ is given by $\widehat{f}(\rho) = \lambda I$ with*

$$\lambda = \frac{1}{d_\rho}\sum_{s\in G} f(s)\chi_\rho(s) = \frac{|G|}{d_\rho}\left(f \,|\, \overline{\chi_\rho}\right).$$

Proof. Observe that

$$\begin{aligned}\rho(s)\widehat{f}(\rho)\rho(s)^{-1} &= \sum_t f(t)\rho(s)\rho(t)\rho(s)^{-1} = \sum_t f(t)\rho(sts^{-1}) \\ &= \sum_u f(s^{-1}us)\rho(u) = \sum_u f(u)\rho(u) = \widehat{f}(\rho),\end{aligned}$$

so Schur's lemma shows that $\widehat{f}(\rho) = \lambda I$.

Taking traces of both sides gives

$$d_\rho \lambda = \mathrm{Tr}\Big(\sum_t f(t)\rho(t)\Big) = \sum_t f(t)\chi_\rho(t).$$ □

Remark 3.2. To give an idea of where all this is going, recall that the distribution after k steps of the random walk (G, μ) started from the identity is given by the k-fold convolution μ^{*k}. Proposition 3.6 shows that the Fourier transform is given by $\widehat{\mu^{*k}}(\rho_i) = \widehat{\mu}(\rho_i)^k$.

If μ is constant on the conjugacy classes of G (which happens in many natural examples), then Proposition 3.7 tells us that $\widehat{\mu}(\rho_i) = \lambda_i I$ and thus $\widehat{\mu^{*k}}(\rho_i) = \lambda_i^k I$ with $\lambda_i = \sum_{s\in G}\mu(s)\frac{\chi_i(s)}{d_i}$, the expectation of the associated *character ratio* under μ.

Applying the inversion formula then yields

$$\mu^{*k}(s) = \frac{1}{|G|}\sum_{i=1}^{m} d_i \mathrm{Tr}\Big(\rho_i(s^{-1})\widehat{\mu^{*k}}(\rho_i)\Big) = \frac{1}{|G|}\sum_{i=1}^{m} d_i \lambda_i^k \overline{\chi_i(s)}.$$

Theorem 3.4. *The irreducible characters form an orthonormal basis for the space of class functions.*

Proof. The first orthogonality relation tells us that the irreducible characters are orthonormal with respect to $(\cdot \,|\, \cdot)$, so it remains only to show that there are enough of them. In particular, the result will follow upon demonstrating that any class function which is orthogonal to the conjugates of each irreducible character must be identically 0.

Suppose that f is such a function. Then for any irrep ρ_i, Proposition 3.7 shows that $\widehat{f}(\rho_i) = \lambda_i I$ with $\lambda_i = |G|\, d_i^{-1}\, (f \,|\, \overline{\chi_i}) = 0$. Fourier inversion then implies that $f \equiv 0$. □

Corollary 3.6. *The number of irreducible representations is equal to the number of conjugacy classes.*

Proof. We know that the irreducible characters form a basis for the space of class functions. Another basis is $\{1_{C_k}\}_{k=1}^{r}$ where $C_1, \ldots, C_r$ are the distinct conjugacy classes of G. □

Another consequence of Theorem 3.4 is the *second orthogonality relation*.

Theorem 3.5. *Let $\chi_1, \ldots, \chi_m$ be the irreducible characters of G. For any $s \in G$, write $\mathrm{cl}(s)$ for the conjugacy class of s. Then*

$$\frac{1}{|G|} \sum_{i=1}^{m} \chi_i(s)\overline{\chi_i(t)} = \frac{1}{|\mathrm{cl}(s)|} 1\{t \in \mathrm{cl}(s)\}.$$

Proof. Set $f_s(t) = 1\{t \in \mathrm{cl}(s)\}$. Then f_s is a class function, so Theorem 3.4 implies $f_s(t) = \sum_{i=1}^{m} \alpha_i \chi_i(t)$ where

$$\alpha_i = (f_s \,|\, \chi_i) = \frac{1}{|G|} \sum_t f_s(t)\overline{\chi_i(t)} = \frac{|\mathrm{cl}(s)|}{|G|}\overline{\chi_i(s)}.$$

Taking conjugates of

$$1\{t \in \mathrm{cl}(s)\} = \sum_{i=1}^{m} \frac{|\mathrm{cl}(s)|}{|G|}\overline{\chi_i(s)}\chi_i(t)$$

and dividing by $|\mathrm{cl}(s)|$ yields the assertion. □

Much more remains to be said about group representations, even without considering infinite groups or ground fields other than $\mathbb{C}$. In particular, Serre's text contains an excellent discussion of topics such as induction/restriction and semidirect products, and Sagan's text is a wonderful source for information about the representation theory of S_n (which we will also touch on briefly in the following section).

However, in the interest of getting back to Markov chains, we conclude with a few remarks about abelian groups.

Proposition 3.8. *The irreducible representations of a finite abelian group G are all one-dimensional.*

Proof. The conjugacy classes of an abelian group all have size one, so Corollary 3.6 implies that there are $|G|$ of them. Since the sum of the squared degrees is also equal to $|G|$, they all have degree one. □

In your homework you are asked to show that the irreducible representations of $\mathbb{Z}/n\mathbb{Z}$ are all of the form $\omega(k) = \omega^k$ as ω ranges over the n^{th} roots of unity.

As finite abelian groups are products of cyclic groups, knowing how to compute representations of products will tell us (in principle) all about the representation theory of these groups.

Recall that if G_1 and G_2 are groups, their product is the group $G_1 \times G_2$ with $(s_1, t_1)(s_2, t_2) = (s_1 s_2, t_1 t_2)$. Given representations (ρ^1, V_1) of G_1 and (ρ^2, V_2) of G_2, we can define the representation $(\rho^1 \otimes \rho^2, V_1 \otimes V_2)$ of $G_1 \times G_2$ by

$$(\rho^1 \otimes \rho^2)_{(s,t)}(\mathbf{v}_1 \otimes \mathbf{v}_2) = \rho^1_s(\mathbf{v}_1) \otimes \rho^2_t(\mathbf{v}_2).$$

The associated character is $\chi_{\rho^1 \otimes \rho^2}\big((s,t)\big) = \chi_{\rho_1}(s)\chi_{\rho^2}(t)$ and the degree is $d_{\rho^1 \otimes \rho^2} = d_{\rho^1} d_{\rho^2}$.

This follows by thinking about $(\rho^1 \otimes \rho^2)_{(s,t)}$ as a block diagonal matrix as we did when discussing the tensor product of two representations of a single group.

(When $G_1 = G_2 = G$, the restriction of the representation $\rho^1 \otimes \rho^2$ of $G \times G$ to the diagonal gives the representation $\rho^1 \otimes \rho^2$ of G.)

Proposition 3.9. *Let G_1 and G_2 be finite groups. Every irreducible representation of $G_1 \times G_2$ is equivalent to $\rho^1 \otimes \rho^2$ with $\rho^i \in Irr(G_i)$.*

Proof. Let ρ^1 and ρ^2 be irreducible representations of G_1 and G_2, respectively. Then

$$\begin{aligned}
\left(\chi_{\rho^1 \otimes \rho^2} \,|\, \chi_{\rho^1 \otimes \rho^2}\right) &= \frac{1}{|G_1 \times G_2|} \sum_{(s,t) \in G_1 \times G_2} \chi_{\rho^1 \otimes \rho^2}(s,t)\overline{\chi_{\rho^1 \otimes \rho^2}(s,t)} \\
&= \frac{1}{|G_1|\,|G_2|} \sum_{s \in G_1} \sum_{t \in G_2} \chi_1(s)\chi_2(t)\overline{\chi_1(s)\chi_2(t)} \\
&= \frac{1}{|G_1|} \sum_{s \in G_1} \chi_1(s)\overline{\chi_1(s)} \cdot \frac{1}{|G_2|} \sum_{t \in G_2} \chi_2(t)\overline{\chi_2(t)} \\
&= (\chi_1 \,|\, \chi_1)\,(\chi_2 \,|\, \chi_2) = 1 \cdot 1 = 1,
\end{aligned}$$

hence $\rho^1 \otimes \rho^2$ is irreducible by Corollary 3.4.

To see that this accounts for all of them, observe that, in the obvious notation,

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} d^2_{\rho_i \otimes \rho_j} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} d_i^2 d_j^2 = \sum_{i=1}^{m_1} d_i^2 \sum_{j=1}^{m_2} d_j^2 = |G_1|\,|G_2| = |G_1 \times G_2|\,. \qquad \square$$

The basic idea of this section has already been hinted at in Remark 3.2: The probability that the random walk on G, driven by μ and started at the identity, is at state s after k steps is $\mu^{*k}(s)$. Taking Fourier transforms turns the convolution into a product, and then Fourier inversion gives a closed-form expression for the k-step distribution (which is especially nice when μ is a class function).

We will begin with a theorem that condenses this line of reasoning into a serviceable bound on the mixing time and then apply it in several examples. While the general result is elegant and often quite powerful, we'll see that many of the details must be worked out on a case by case basis and can get rather involved.

4.1 General Theory

We assume throughout that G is a finite group and μ is a probability on G whose support is generating. Let P be the transition matrix for the random walk (G, μ) defined by $P(x, y) = \mu(yx^{-1})$. Example 1.7 shows that the uniform distribution, $U(g) = \frac{1}{|G|}$, is stationary for P, and Example 2.2 shows that the total variation distance between $P^k(g, \cdot)$ and $U(\cdot)$ does not depend on g. As such, we will always start our chain at the identity, and we employ the notation $P^k(s) := P^k(id, s) = \mu^{*k}(s)$.

We denote the Hermitian of a matrix M by $M^\dagger$ so that $M^\dagger(i, j) = \overline{M(j, i)}$.

Observe that if g is a real-valued function on G, $h(s) = g(s^{-1})$, and ρ is a unitary representation, then

$$\widehat{g}(\rho)^\dagger = \sum_{s\in G} g(s)\rho(s)^\dagger = \sum_{s\in G} g(s)\rho(s^{-1}) = \sum_{s\in G} g(s^{-1})\rho(s) = \widehat{h}(\rho),$$

so the Plancherel formula can be written as

$$\sum_{s\in G} f(s)g(s) = \sum_{s\in G} f(s)h(s^{-1}) = \frac{1}{|G|}\sum_{\rho\in Irr(G)} d_\rho \mathrm{Tr}\left(\widehat{f}(\rho)\widehat{h}(\rho)\right) = \frac{1}{|G|}\sum_{\rho\in Irr(G)} d_\rho \mathrm{Tr}\left(\widehat{f}(\rho)\widehat{g}(\rho)^\dagger\right). \tag{4.1}$$

Finally, recall that the $L^2(U)$ distance between probabilities μ and ν on G is defined by

$$\|\mu - \nu\|_2 = \left(\sum_{s\in G}\left|\frac{\mu(s)}{U(s)} - \frac{\nu(s)}{U(s)}\right|^2 U(s)\right)^{\frac{1}{2}} = |G|^{\frac{1}{2}}\sqrt{\sum_{s\in G}|\mu(s) - \nu(s)|^2}.$$

Theorem 4.1. *For any probability Q on G, we have*

$$4\|Q - U\|_{TV}^2 \le \|Q - U\|_2^2 = \sum_{\rho\in Irr^*(G)} d_\rho \mathrm{Tr}\left(\widehat{Q}(\rho)\widehat{Q}(\rho)^\dagger\right)$$

where $Irr^*(G) = Irr(G) \setminus \{\rho_0\}$ is the set of nontrivial irreducible representations of G.

Proof. The inequality is Cauchy-Schwarz, as detailed in Remark 2.1.

For the equality, Equation (4.1) gives

$$\sum_{s\in G}|Q(s) - U(s)|^2 = \frac{1}{|G|}\sum_{\rho\in Irr(G)} d_\rho \mathrm{Tr}\left(\left(\widehat{Q}(\rho) - \widehat{U}(\rho)\right)\left(\widehat{Q}(\rho) - \widehat{U}(\rho)\right)^\dagger\right).$$

As it was shown in the homework that $\widehat{U}(\rho_0) = 1 = \widehat{Q}(\rho_0)$ and $\widehat{U}(\rho) = 0$ if $\rho \in Irr^*(G)$, this simplifies to

$$|G|\sum_{s\in G}|Q(s) - U(s)|^2 = \sum_{\rho\in Irr^*(G)} d_\rho \mathrm{Tr}\left(\widehat{Q}(\rho)\widehat{Q}(\rho)^\dagger\right). \qquad \square$$

Example 4.1. Suppose that G is an abelian group. I claim that the irreducible characters of G form an orthonormal basis of eigenfunctions for the transition matrix $P(r,s) = \mu(sr^{-1})$, where the eigenvalue associated with χ is given by the Fourier transform $\widehat{\mu}(\chi)$.

Indeed, the irreducible characters are orthonormal w.r.t. $(\cdot\,|\,\cdot)$, and for any such χ, we have

$$\begin{aligned}(P\chi)(s) &= \sum_{t\in G} P(s,t)\chi(t) = \sum_{t\in G}\mu(ts^{-1})\chi(t)\\ &= \sum_{r\in G}\mu(r)\chi(rs) = \chi(s)\sum_{r\in G}\mu(r)\chi(r) = \chi(s)\widehat{\mu}(\chi).\end{aligned}$$

Now these characters are precisely the irreps of G (which justifies the penultimate equality) and all have degree 1, so Theorem 4.1 gives

$$\left\|P^k - U\right\|_2^2 = \sum_{\chi\neq 1}\mathrm{Tr}\left(\widehat{\mu}(\chi)^k\left(\widehat{\mu}(\chi)^k\right)^\dagger\right) = \sum_{\lambda\neq 1}|\lambda|^{2k}$$

where the sum runs over the nontrivial eigenvalues $\lambda = \widehat{\mu}(\chi)$.

Even though the random walk is not assumed to be reversible, we get the same basic expression as Equation (1.2) except that we are taking moduli of the eigenvalues as they may not be real.

Example 4.2. More generally, let G be an arbitrary finite group with μ constant on the conjugacy classes of G. Then Proposition 3.7 shows that $\widehat{\mu}(\rho) = \lambda_\rho I$ with $\lambda_\rho = \frac{1}{d_\rho}\sum_{s\in G}\mu(s)\chi_\rho(s)$, and Theorem 4.1 gives

$$\begin{aligned}\left\|P^k - U\right\|_2^2 &= \sum_{\rho\in Irr^*(G)} d_\rho \mathrm{Tr}\left(\widehat{\mu}(\rho)^k\left(\widehat{\mu}(\rho)^k\right)^\dagger\right)\\ &= \sum_{\rho\in Irr^*(G)} d_\rho \mathrm{Tr}\left(|\lambda_\rho|^{2k} I_{d_\rho}\right) = \sum_{\rho\in Irr^*(G)} d_\rho^2\,|\lambda_\rho|^{2k}.\end{aligned}$$

If μ is symmetric as well, then $\overline{\lambda_\rho} = \frac{1}{d_\rho}\sum_{s\in G}\mu(s)\chi_\rho(s^{-1}) = \frac{1}{d_\rho}\sum_{s\in G}\mu(s^{-1})\chi_\rho(s^{-1}) = \lambda_\rho$, so the modulus may be dropped.

We will mainly work in the total variation distance, so it should be noted that the inequality in Theorem 4.1 is generally not that bad since we usually apply the result when Q and U are 'close,' hence Cauchy-Schwarz is not giving up too much. In fact, we always have

$$\begin{aligned}4\left\|\mu-\nu\right\|_{TV}^2 &= \sum_{s\in G}|\mu(s)-\nu(s)|^2 + \sum_{s\in G}|\mu(s)-\nu(s)|\sum_{t\in G\setminus\{s\}}|\mu(t)-\nu(t)|\\ &\geq \sum_{s\in G}|\mu(s)-\nu(s)|^2 + \sum_{s\in G}|\mu(s)-\nu(s)|\left|\sum_{t\in G\setminus\{s\}}(\mu(t)-\nu(t))\right|\\ &= \sum_{s\in G}|\mu(s)-\nu(s)|^2 + \sum_{s\in G}|\mu(s)-\nu(s)|\left|\big(1-\mu(s)\big)-\big(1-\nu(s)\big)\right| = 2\sum_{s\in G}|\mu(s)-\nu(s)|^2,\end{aligned}$$

which gives the lower bound $\|Q-U\|_{TV}^2 \geq \dfrac{1}{2\,|G|}\displaystyle\sum_{\rho\in Irr^*(G)} d_\rho \mathrm{Tr}\left(\widehat{Q}(\rho)\widehat{Q}(\rho)^\dagger\right)$.

You may recognize $\mathrm{Tr}\left(\widehat{Q}(\rho)\widehat{Q}(\rho)^\dagger\right)$ as the sum of squared singular values of $\widehat{Q}(\rho)$, or equivalently, the squared *Frobenius norm* of $\widehat{Q}(\rho)$ where the Frobenius norm of $A = [a_{ij}]_{i,j=1}^n$ is $\|A\|_F = \sqrt{\sum_{i,j}|a_{ij}|^2}$.

Unfortunately, I'm not aware of any useful way to leverage these characterizations for $\widehat{Q}(\rho) = \widehat{\mu}(\rho)^k$.

4.2 Concrete Examples

4.2.1 n-Cycle

Let $G = \mathbb{Z}/n\mathbb{Z}$ for an odd integer $n \geq 5$ and consider the random walk on G driven by $\mu(\pm 1) = \frac{1}{2}$.

(This is the random walk from Example 2.5 without holding probabilities. We assume that n is odd to avoid the periodicity observed in Example 1.6.)

We will show that a little more that n^2 steps are required for the chain to equilibrate. To facilitate the proof, we first establish some bounds on cosine.

Lemma 4.1. $\cos(x) \leq e^{-\frac{x^2}{2}}$ *for* $x \in [0, \pi/2]$ *and* $\cos(x) \geq e^{-\frac{x^2}{2} - \frac{2x^4}{3}}$ *for* $x \in [0, \pi/4]$.

Proof. The function $g(x) = \log\left(e^{\frac{x^2}{2}}\cos(x)\right)$ has derivative $g'(x) = x - \tan(x) \leq 0$ for $x \in [0, \pi/2)$, and the upper bound follows upon exponentiating $g(x) \leq g(0) = 0$.

For the lower bound, observe that $h(x) = \log(\cos(x))$ has as its first few derivatives $h'(x) = -\tan(x)$, $h''(x) = -\sec^2(x)$, $h^{(3)}(x) = -2\sec^2(x)\tan(x)$, $h^{(4)}(x) = -4\sec^2(x)\tan^2(x) - 2\sec^4(x)$, so a third order Taylor expansion about 0 gives $h(x) = -\frac{x^2}{2} + R_3(x)$ where $|R_3(x)| \leq \frac{x^4}{4!}\sup_{t\in[0,\pi/4]}\left|h^{(4)}(t)\right|$.

Since secant and tangent are increasing on $[0, \pi/4]$, we see that

$$\log(\cos(x)) \geq -\frac{x^2}{2} - \frac{x^4}{4!}\left|h^{(4)}(\pi/4)\right| = -\frac{x^2}{2} - \frac{2x^4}{3}. \qquad \square$$

Theorem 4.2. *Let P be simple random walk on the n-cycle with $n \geq 5$ odd, and set $\alpha = \pi^2/2$, $\beta = 2\pi^4/3$. For any $k \geq n^2$, we have*

$$\frac{1}{2}e^{-\alpha k/n^2 - \beta k/n^4} \leq \left\|P^k - U\right\|_{TV} \leq \frac{3}{4}e^{-\alpha k/n^2}.$$

Proof. The irreducible representations of G are given by $\chi_j(x) = e^{\frac{2\pi i j x}{n}}$, $j = 0, 1, \ldots, n-1$, and the Fourier transform of μ at χ_j is

$$\widehat{\mu}(j) = \frac{1}{2}\left(e^{\frac{2\pi i j}{n}} + e^{-\frac{2\pi i j}{n}}\right) = \cos\left(\frac{2\pi j}{n}\right).$$

Thus by Theorem 4.1, we have

$$\left\|P^k - U\right\|_{TV}^2 \leq \frac{1}{4}\sum_{j=1}^{n-1}\cos^{2k}\left(\frac{2\pi j}{n}\right) = \frac{1}{2}\sum_{j=1}^{(n-1)/2}\cos^{2k}\left(\frac{\pi j}{n}\right).$$

(The equality comes from rearranging the sum and using symmetries of cosine. Details are left as homework.)

Invoking Lemma 4.1 yields

$$\begin{aligned}\left\|P^k - U\right\|_{TV}^2 &\leq \frac{1}{2}\sum_{j=1}^{(n-1)/2}\cos^{2k}\left(\frac{\pi j}{n}\right) \leq \frac{1}{2}\sum_{j=1}^{(n-1)/2} e^{-\pi^2 j^2 k/n^2} \\ &\leq \frac{1}{2}e^{-\pi^2 k/n^2}\sum_{j=1}^{\infty} e^{-\pi^2(j+1)(j-1)k/n^2} \leq \frac{1}{2}e^{-\pi^2 k/n^2}\sum_{j=0}^{\infty} e^{-2\pi^2 jk/n^2} = \frac{e^{-\pi^2 k/n^2}}{2\left(1 - e^{-2\pi^2 k/n^2}\right)}.\end{aligned}$$

As $\left[2\left(1 - e^{-2\pi^2 k/n^2}\right)\right]^{-1}$ is decreasing in k and $\left[2\left(1 - e^{-2\pi^2}\right)\right]^{-\frac{1}{2}} \approx 0.7071$, the asserted upper bound for $k \geq n^2$ follows upon taking square roots.

For the lower bound, observe that the sum $\sum_{j=1}^{n-1} \cos^{2k}\left(\frac{2\pi j}{n}\right)$ is dominated by the $m = (n-1)/2$ term.

Taking this as inspiration, let's consider the test function $\varphi(x) = \cos\left(\frac{2\pi m x}{n}\right)$ in regard to the characterization $\|\mu - \nu\|_{TV} = \frac{1}{2}\max_{\|f\|_\infty \le 1} \left|\mu(f) - \nu(f)\right|$.

(As discussed in Example 4.1, this is an eigenfunction corresponding to the top nontrivial eigenvalue of the chain. We will see that such functions often form useful *distinguishing statistics* for lower-bounding the distance to stationarity, so this is not as contrived as it may first appear.)

Clearly $\|\varphi\|_\infty = 1$, and its expectation under the uniform distribution is

$$U(\varphi) = \frac{1}{n}\sum_{j=0}^{n-1} \cos\left(\frac{2\pi m j}{n}\right) = 0$$

since, writing $\omega = e^{\frac{2\pi i m}{n}}$, the sum is the real part of $\sum_{j=0}^{n-1} \omega^j = \frac{1-\omega^n}{1-\omega} = 0$.

As the expected value of φ under P^k is

$$\begin{aligned} P^k(\varphi) &= \sum_{j=0}^{n-1} P^k(j)\cos\left(\frac{2\pi m j}{n}\right) = \mathrm{Re}\left(\sum_{j=0}^{n-1} \mu^{*k}(j)\chi_m(j)\right) \\ &= \widehat{\mu}(m)^k = \cos^k\left(\frac{2\pi m}{n}\right) = \cos^k\left(\pi - \frac{\pi}{n}\right) = (-1)^k \cos^k\left(\frac{\pi}{n}\right), \end{aligned}$$

Lemma 4.1 gives

$$\|\mu - \nu\|_{TV} \ge \frac{1}{2}\left|\cos^k\left(\frac{\pi}{n}\right)\right| \ge \frac{1}{2}\exp\left(-\frac{\pi^2 k}{2n^2} - \frac{2\pi^4 k}{3n^4}\right). \qquad \square$$

4.2.2 Hypercube

Recall the random walk (G, μ) from Example 2.4: $G = (\mathbb{Z}/2\mathbb{Z})^d$, $\mu(\mathbf{0}) = \frac{1}{2}$, and $\mu(\mathbf{e}_1) = \cdots = \mu(\mathbf{e}_d) = \frac{1}{2d}$ where $\mathbf{0}$ is the vector of all zeros and $\mathbf{e}_i$ is the vector with a one in coordinate i and zeros elsewhere.

By Proposition 3.9, the irreducible representations of G are given by $\rho_{\mathbf{x}}(\mathbf{y}) = \prod_{i=1}^d \rho_{x_i}(y_i)$ where $\rho_0(y) = 1$ and $\rho_1(y) = (-1)^y$ for $y \in \{0, 1\}$. More concisely, $\rho_{\mathbf{x}}(\mathbf{y}) = (-1)^{\mathbf{x}\cdot\mathbf{y}}$, and in particular, $\rho_{\mathbf{x}}(\mathbf{e}_i) = (-1)^{x_i}$.

Thus, letting $\omega(\mathbf{x})$ denote the number of coordinates in $\mathbf{x}$ that are equal to 1, we have

$$\widehat{\mu}(\rho_{\mathbf{x}}) = \frac{1}{2}\rho_{\mathbf{x}}(\mathbf{0}) + \frac{1}{2d}\sum_{i=1}^d (-1)^{x_i} = \frac{1}{2} + \frac{1}{2d}\left[(d - \omega(\mathbf{x})) - \omega(\mathbf{x})\right] = 1 - \frac{\omega(\mathbf{x})}{d},$$

whence

$$\begin{aligned} \left\|P^k - U\right\|_{TV}^2 &\le \frac{1}{4}\sum_{\mathbf{x}\neq\mathbf{0}} \left(1 - \frac{\omega(\mathbf{x})}{d}\right)^{2k} = \frac{1}{4}\sum_{j=1}^d \binom{d}{j}\left(1 - \frac{j}{d}\right)^{2k} \\ &\le \frac{1}{4}\sum_{j=1}^d \frac{d^j}{j!}\exp\left(-\frac{2k}{d}\right)^j \le \frac{1}{4}\left[\exp\left(de^{-\frac{2k}{d}}\right) - 1\right]. \end{aligned}$$

Taking $k \ge \frac{d}{2}\left[\log(d) + c\right]$ for some $c > 0$ gives

$$\left\|P^k - U\right\|_{TV}^2 \le \frac{1}{4}\left[\exp\left(de^{-\log(d)-c}\right) - 1\right] = \frac{1}{4}\left(e^{e^{-c}} - 1\right).$$

To lower-bound the variation distance, consider the random variable $Z(\mathbf{x}) = d - 2\omega(\mathbf{x}) = \sum_{i=1}^{d} Y_i(\mathbf{x})$ where $Y_i(\mathbf{x}) = (-1)^{x_i} = \rho_{\mathbf{e}_i}(\mathbf{x})$.

Under the uniform distribution on G, $Y_1, \dots Y_d$ are i.i.d. Unif(± 1), so $\mathbb{E}_U[Z] = 0$ and $\mathrm{Var}_U(Z) = d$.

Under P^k, we have

$$\mathbb{E}_k[Y_i] = \sum_{\mathbf{x}\in G} \mu^{*k}(\mathbf{x})\rho_{\mathbf{e}_i}(\mathbf{x}) = \widehat{\mu}(\rho_{\mathbf{e}_i})^k = \left(1 - \frac{1}{d}\right)^k,$$

hence $\mathbb{E}_k[Z] = d\left(1 - \frac{1}{d}\right)^k$.

Also, since $\rho_{\mathbf{y}}(\mathbf{x})\rho_{\mathbf{z}}(\mathbf{x}) = \rho_{\mathbf{y}+\mathbf{z}}(\mathbf{x})$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in G$, we see that for any $i \neq j$,

$$\mathbb{E}_k[Y_iY_j] = \sum_{\mathbf{x}\in G} \mu^{*k}(\mathbf{x})\rho_{\mathbf{e}_i}(\mathbf{x})\rho_{\mathbf{e}_j}(\mathbf{x}) = \sum_{\mathbf{x}\in G} \mu^{*k}(\mathbf{x})\rho_{\mathbf{e}_i+\mathbf{e}_j}(\mathbf{x}) = \widehat{\mu}(\rho_{\mathbf{e}_i+\mathbf{e}_j})^k = \left(1 - \frac{2}{d}\right)^k.$$

As $Y_i^2 \equiv 1$, we conclude that

$$\mathbb{E}_k[Z^2] = \sum_{i=1}^{d} \mathbb{E}_k[Y_i^2] + \sum_{i\neq j} \mathbb{E}_k[Y_iY_j] = d + d(d-1)\left(1 - \frac{2}{d}\right)^k,$$

and thus

$$\mathrm{Var}_k(Z) = \mathbb{E}_k[Z^2] - \mathbb{E}_k[Z]^2 = d + d(d-1)\left(1 - \frac{2}{d}\right)^k - d^2\left(1 - \frac{1}{d}\right)^{2k}.$$

Because

$$\begin{aligned} d(d-1)\left(1 - \frac{2}{d}\right)^k - d^2\left(1 - \frac{1}{d}\right)^{2k} &\leq d^2\left(\frac{d-2}{d}\right)^k - d^2\left(\frac{d-1}{d}\right)^{2k} \\ &= d^{2-2k}\left[(d^2-2d)^k - (d-1)^{2k}\right] < 0, \end{aligned}$$

we have the more manageable upper bound $\mathrm{Var}_k(Z) \leq d$.

Now consider the event $A_k = \left\{Z < \frac{1}{2}\mathbb{E}_k[Z]\right\}$. Chebychev gives

$$U(A_k) = 1 - \mathbb{P}_U\left(Z \geq \frac{1}{2}\mathbb{E}_k[Z]\right) \geq 1 - \frac{4d}{\mathbb{E}_k[Z]^2}$$

and

$$P^k(A_k) \leq \mathbb{P}_k\left(|Z - \mathbb{E}_k[Z]| \geq \frac{1}{2}\mathbb{E}_k[Z]\right) \leq \frac{4\mathrm{Var}_k(Z)}{\mathbb{E}_k[Z]^2}$$

so that

$$\left\|P^k - U\right\|_{TV} = \max_{B\subseteq G}\left[U(B) - P^k(B)\right] \geq 1 - \frac{4d}{\mathbb{E}_k[Z]^2} - \frac{4\mathrm{Var}_k(Z)}{\mathbb{E}_k[Z]^2} \geq 1 - \frac{8d}{\mathbb{E}_k[Z]^2}.$$

If we assume that $d \geq 4$ and use the fact that

$$\log(1-x) = -\sum_{k=1}^{\infty}\frac{x^k}{k} \geq -x - \frac{1}{2}\sum_{k=2}^{\infty}x^k = -x - \frac{x^2}{2(1-x)} \geq -x - x^2$$

for $0 \leq x \leq \frac{1}{2}$, then we see that

$$\mathbb{E}_k[Z] = d\left(1 - \frac{1}{d}\right)^k = de^{k\log(1-1/d)} \geq de^{-\frac{k}{d} - \frac{k}{d^2}}.$$

Taking $k = \frac{d}{2}[\log(d) - c]$ and observing that $f(d) = d^{-\frac{1}{2d}}$ is increasing on $[4, \infty]$ with $f(4) \geq \sqrt{2/3}$ gives

$$\begin{aligned}\mathbb{E}_k[Z] &\geq de^{-\frac{k}{d} - \frac{k}{d^2}} = de^{-\frac{1}{2}\log(d) + \frac{c}{2} - \frac{1}{2d}\log(d) + \frac{c}{2d}} \\ &\geq \sqrt{d} \cdot e^{\frac{c}{2}} \cdot d^{-\frac{1}{2d}} \geq e^{\frac{c}{2}}\sqrt{\frac{2d}{3}}.\end{aligned}$$

It follows that

$$\left\|P^k - U\right\|_{TV} \geq 1 - \frac{8d}{\mathbb{E}_k[Z]^2} \geq 1 - 12e^{-c}.$$

To recap,

Theorem 4.3. *For lazy simple random walk on the hypercube $(\mathbb{Z}/2\mathbb{Z})^d$, $d \geq 4$, we have*

(1) If $k \geq \frac{d}{2}[\log(d) + c]$ with $c > 0$, then

$$\left\|P^k - U\right\|_{TV} \leq \frac{1}{2}\left(e^{e^{-c}} - 1\right)^{\frac{1}{2}}.$$

(2) If $k \leq \frac{d}{2}[\log(d) - c]$ with $c > 0$, then

$$\left\|P^k - U\right\|_{TV} \geq 1 - 12e^{-c}.$$

4.2.3 Random Transpositions

Theorem 4.1 was introduced by Diaconis and Shahshahani in 1981 to study shuffling by random transpositions. This is the random walk on $G = S_n$ driven by the measure μ defined by $\mu(id) = \frac{1}{n}$, $\mu(\sigma) = \frac{2}{n^2}$ for σ a transposition. (S_n is generated by transpositions, so the chain is irreducible, and the positive holding probability prevents periodicity.)

The image is that we have a deck of cards labeled $1, \ldots, n$ and at each time step we we choose a pair $(i, j) \in [n]^2$ uniformly at random and swap the cards in positions i and j. When $i = j$, this corresponds to doing nothing. (While this image suggests that the random walk should be left-invariant, the discussion in Examples 1.3 and 2.2 shows that we may treat it as though we are updating by left-multiplication so that no modification of the theory is needed.) Ergodicity and transitivity imply $\mu^{*k} \to U$, so repeated shuffles do indeed 'mix up' the deck. The question is "How many are needed?"

The fact that the underlying group is nonabelian (unlike the other examples of this subsection) means that we will have to deal with representations having degree greater than 1. However, the driving measure is constant on conjugacy classes and this simplifies the analysis significantly. Still, a completely rigorous and self-contained derivation of the mixing behavior of this chain would take us too far afield into the representation theory of the symmetric group, so we'll accept a few facts on faith in the discussion that follows (though we will meander a bit to take in some of the flavor of this fascinating topic).

To begin, Example 4.2 shows that the Fourier transform of μ at the irrep ρ is $\widehat{\mu}(\rho) = \lambda_\rho I$ where

$$\lambda_\rho = \frac{1}{d_\rho}\sum_{s \in G}\mu(s)\chi_\rho(s) = \frac{1}{d_\rho}\left(\frac{1}{n}d_\rho + \binom{n}{2}\frac{2}{n^2}\chi_\rho(\tau)\right) = \frac{1}{n} + \frac{n-1}{n}r_\rho(\tau).$$

Here τ is a generic transposition and $r_\rho(\tau) = \chi_\rho(\tau)/d_\rho$ is the character ratio corresponding to ρ, which is $\mathbb{R}$-valued since $\overline{\chi_\rho(\tau)} = \chi_\rho(\tau^{-1}) = \chi_\rho(\tau)$.

As such, we have the total variation bound

$$\left\|P^k - U\right\|_{TV}^2 \leq \frac{1}{4} \sum_{\rho \in Irr^*(G)} d_\rho^2 \lambda_\rho^{2k} = \frac{1}{4} \sum_{\rho \in Irr^*(G)} d_\rho^2 \left(\frac{1}{n} + \frac{n-1}{n} r_\rho(\tau)\right)^{2k}.$$

The general idea is to choose k large enough to kill the 'slowest' representation and then show that the contributions from the rest are negligible.

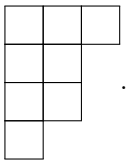
This slow term corresponds to the $(n-1)$-dimensional standard representation, whose character is $\chi_{\text{std}}(\sigma) = (\#\text{ fixed points of } \sigma) - 1$. It follows that $\frac{n-1}{n} r_{\text{std}}(\tau) = \frac{n-3}{n}$, so the corresponding summand in the variation bound is

$$\begin{aligned} d_{\text{std}}^2 \left(\frac{1}{n} + \frac{n-1}{n} r_{\text{std}}(\tau)\right)^{2k} &= (n-1)^2 \left(1 - \frac{2}{n}\right)^{2k} \\ &\leq n^2 \exp\left(2k \log\left(1 - \frac{2}{n}\right)\right) \leq n^2 e^{-\frac{4k}{n}}, \end{aligned}$$

which equals e^{-4c} for $k = \frac{1}{2} n \log(n) + cn$. This turns out to be the correct rate, and the bulk of the proof lies in dealing with the other irreps.

For this, we need some facts about the irreducible representations of S_n. First, a *partition* λ of n (denoted $\lambda \vdash n$) is an indexed collection of positive integers $\lambda = (\lambda_1, \ldots \lambda_k)$ where $\lambda_1 \geq \cdots \geq \lambda_k$ and $\lambda_1 + \cdots + \lambda_k = n$. The partitions of n index the conjugacy classes of S_n by specifying cycle type, and they can be represented by *Young diagrams*, which are arrays of boxes with λ_i in the i^{th} row.

For instance the partition $(3, 2, 2, 1) \vdash 8$ has Young diagram



.

If the boxes are populated by distinct elements of $[n]$, then the resulting object is called a *Young tableau* of *shape* λ, and we declare two tableaux equivalent if they have the same elements in each row, e.g.

$$\begin{array}{|c|c|c|}\hline 3 & 1 & 5 \\ \hline 2 & 4 \\ \cline{1-2} 8 & 6 \\ \cline{1-2} 7 \\ \cline{1-1} \end{array} \quad \sim \quad \begin{array}{|c|c|c|}\hline 1 & 3 & 5 \\ \hline 2 & 4 \\ \cline{1-2} 6 & 8 \\ \cline{1-2} 7 \\ \cline{1-1} \end{array} \quad .$$

The tableau on the right is called *standard* because the entries increase along rows and down columns.

Equivalence classes of tableaux are called *tabloids* where the tabloid associated with tableau t is denoted $\{t\}$. The number of tabloids of shape $\lambda = (\lambda_1, \ldots \lambda_k)$ is clearly $n!/\lambda_1! \cdots \lambda_k!$, and one can show that the irreducible representations of S_n correspond via tabloids to partitions of n.

Specifically, for a fixed partition λ, define the *permutation module* M^λ to be the vector space with basis vectors $\mathbf{e}_{\{t\}}$ as $\{t\}$ runs over the tabloids of shape λ, and consider the representation $(\rho_\lambda, M^\lambda)$ by $\rho_\lambda(\pi)\mathbf{e}_{\{t\}} = \mathbf{e}_{\{\pi t\}}$ where permutations act on tableaux in the obvious way. For instance,

$$\pi : \begin{array}{|c|c|c|}\hline 3 & 1 & 5 \\ \hline 2 & 4 \\ \cline{1-2} 8 & 6 \\ \cline{1-2} 7 \\ \cline{1-1}\end{array} \mapsto \begin{array}{|c|c|c|}\hline \pi(3) & \pi(1) & \pi(5) \\ \hline \pi(2) & \pi(4) \\ \cline{1-2} \pi(8) & \pi(6) \\ \cline{1-2} \pi(7) \\ \cline{1-1}\end{array} .$$

Example 4.3. For the trivial partition (n), there is a single tabloid, so $\rho_{(n)}$ is the trivial representation.

At the other extreme, each tabloid $\{t\}$ of shape $(1^n) := (1, \ldots, 1)$ contains only the tableau t, which we can represent as the permutation σ with $\sigma(i)$ the element in row i of t. $\rho_{(1^n)}$ is thus the regular representation $\rho_{(1^n)}(\pi)\mathbf{e}_\sigma = \mathbf{e}_{\pi\sigma}$.

The tabloids of shape $(n-1, 1)$ are determined by the content of the single box in the second row, so $M^{(n-1,1)}$ has basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ with $\rho_{(n-1,1)}(\pi)\mathbf{e}_k = \mathbf{e}_{\pi(k)}$. This is the n-dimensional permutation representation, which splits as the direct sum of the trivial representation and the standard representation.

Now define a partial order on partitions by $(\lambda_1, \ldots, \lambda_k) \trianglerighteq (\mu_1, \ldots, \mu_\ell)$ if $\lambda_1 + \cdots + \lambda_i \geq \mu_1 + \cdots + \mu_i$ for all $i \geq 1$ (with, for example, $\lambda_i = 0$ when $i > k$). Thus, $(4,2) \trianglerighteq (3,3), (4,1,1)$, but $(3,3)$ and $(4,1,1)$ are not comparable. More colorfully, $\lambda \trianglerighteq \mu$ if one can obtain the diagram of λ from the diagram of μ by successively removing a box from the end of one row and appending it to a row above so that each intermittent stage represents a valid diagram.

It turns out that each M^λ contains an invariant subspace S^λ that corresponds to a unique irreducible representation, and any other irreps appearing in the direct sum decomposition of M^λ correspond to partitions μ with $\mu \trianglerighteq \lambda$. We will not go into all the grisly details, but we observe that the *Specht module* S^λ can be constructed as follows.

For a tableau t of shape λ, define its *column stabilizer* C_t to be the set of all permutations π such that t and πt have the same elements in each column, and consider the *polytabloid* $\mathbf{f}_t = \sum_{\pi \in C_t} \operatorname{sgn}(\pi)\mathbf{e}_{\{\pi t\}} \in M^\lambda$.

You are asked in the homework to show that $\rho_\lambda(\sigma)\mathbf{f}_t = \mathbf{f}_{\sigma t}$ for all tableau t of shape λ and all $\sigma \in S_n$.

Thus if we define S^λ to be the span of the polytabloids for tableaux of shape λ, then S^λ is an invariant subspace of M^λ. In fact, one can show that $(\rho_\lambda, S^\lambda)$ is irreducible and not equivalent to (ρ_μ, S^μ) if $\mu \neq \lambda$. As the number of partitions of n equals the number of conjugacy classes of S_n, this accounts for all irreps.

Example 4.4. We saw in Example 4.3 that each tabloid t of shape (1^n) corresponds to some $\sigma \in S_n$. Clearly its column stabilizer is $C_t = S_n$, so the associated polytabloid is $\mathbf{f}_t = \sum_{\pi \in S_n} \operatorname{sgn}(\pi)\mathbf{e}_{\pi\sigma}$. Since $\operatorname{sgn} : S_n \to \{\pm 1\}$ is a homomorphism,

$$\rho_{(1^n)}(\alpha)\mathbf{f}_t = \mathbf{f}_{\alpha t} = \sum_{\pi \in S_n} \operatorname{sgn}(\pi)\mathbf{e}_{\pi\alpha\sigma} = \operatorname{sgn}(\alpha^{-1}) \sum_{\pi \in S_n} \operatorname{sgn}(\pi\alpha)\mathbf{e}_{\pi\alpha\sigma} = \operatorname{sgn}(\alpha)\mathbf{f}_t,$$

hence $S^{(1^n)}$ is the sign representation.

Of course, the polytabloids are not linearly independent in general, but it is known that a basis for S^λ is given by $\{\mathbf{f}_t : t \text{ is a standard tableau of shape } \lambda\}$. The size of this basis (corresponding to the degree of the associated irrep) can be computed using the famous hook-length formula. We record these observations as

Fact 4.1. *The irreducible representations are in one-to-one correspondence with the partitions of n, and the irrep associated with $\lambda \vdash n$ has degree equal to the number of standard Young tableaux of shape λ.*

Corollary 4.1. *The dimension of the irrep corresponding to the partition $\lambda = (\lambda_1, \dots, \lambda_k) \vdash n$ satisfies $d_\lambda \leq \binom{n}{\lambda_1} d_{\lambda^*}$ where $\lambda^* = (\lambda_2, \dots, \lambda_k) \vdash (n - \lambda_1)$.*

Proof. There are $\binom{n}{\lambda_1}$ ways to populate the first row of a standard Young tableau of shape λ, and there are d_{λ^*} standard Young tableau of shape λ^* consisting of the remaining elements that can be appended below it. The inequality is because some of the tableau so constructed will not increase down all columns. □

The only other fact that we need for the upper bound is the following special case of a result due to Frobenius.

Fact 4.2. *For a partition $\lambda = (\lambda_1, \dots, \lambda_k) \vdash n$, let $r_\lambda(\tau) = \frac{\chi_\lambda(\tau)}{d_\lambda}$ be the character ratio for the irrep associated with λ evaluated at a transposition. Writing $\lambda' = (\lambda'_1, \dots, \lambda'_\ell)$ for the transpose of λ—that is, the partition of n whose Young diagram has λ_j boxes in column j—we have*

$$r_\lambda(\tau) = \frac{1}{\binom{n}{2}} \sum_j \left[\binom{\lambda_j}{2} - \binom{\lambda'_j}{2} \right] = \frac{1}{n(n-1)} \sum_j \left[\lambda_j^2 - (2j-1)\lambda_j \right].$$

It is also worth observing at this point that we always have $d_\lambda = d_{\lambda'}$ since swapping rows and columns in a standard tableau of shape λ gives a standard tableau of shape λ'.

The next ingredient is a simple monotonicity result for character ratios.

Lemma 4.2. *If $\lambda \trianglerighteq \mu$, then $r_\lambda(\tau) \geq r_\mu(\tau)$.*

Proof. If $\lambda \trianglerighteq \mu$, then the diagram of λ can be obtained from that of μ by successively removing a box and appending it to a row above, so it suffices to prove the result when $\lambda_a = \mu_a + 1$, $\lambda_b = \mu_b - 1$ and $\lambda_j = \mu_j$ if $j \neq a, b$ for some $a < b$. In this case, Fact 4.2 gives

$$\begin{aligned} r_\lambda(\tau) - r_\mu(\tau) &= \frac{1}{n(n-1)} \big[\lambda_a^2 - (2a-1)\lambda_a - (\lambda_a - 1)^2 + (2a-1)(\lambda_a - 1) \\ &\qquad + \lambda_b^2 - (2b-1)\lambda_b - (\lambda_b + 1)^2 + (2b-1)(\lambda_b + 1) \big] \\ &= \frac{1}{n(n-1)} \big[\lambda_a^2 - 2a\lambda_a + \lambda_a - \lambda_a^2 + 2\lambda_a - 1 + 2a\lambda_a - 2a - \lambda_a + 1 \\ &\qquad + \lambda_b^2 - 2b\lambda_b + \lambda_b - \lambda_b^2 - 2\lambda_b - 1 + 2b\lambda_b - \lambda_b + 2b - 1 \big] \\ &= \frac{1}{n(n-1)} \left[2(\lambda_a - \lambda_b) + 2(b - a - 1) \right] \geq \binom{n}{2}^{-1} \end{aligned}$$

since $\lambda_a \geq \lambda_b$ and $b \geq a + 1$. □

The basic idea at this point is to split up the irreps according to the length of the first row of their Young diagrams, and the following Lemma enables us to bound the character ratios in these regimes.

Lemma 4.3. *Let $\lambda = (\lambda_1, \ldots, \lambda_m)$ be a partition of n. Then we have the bounds*

(1) $r_\lambda(\tau) \le \dfrac{\lambda_1 - 1}{n-1}$.

(2) If $\lambda_1 \ge n/2$, *then* $r_\lambda(\tau) \le 1 - \dfrac{2(\lambda_1+1)(n-\lambda_1)}{n(n-1)}$.

Proof.

(1) Fact 4.2 gives

$$r_\lambda(\tau) = \frac{1}{n(n-1)} \sum_{j=1}^{m} \left[\lambda_j^2 - (2j-1)\lambda_j\right] \le \frac{1}{n(n-1)} \sum_{j=1}^{m} \left[\lambda_j(\lambda_j - 1)\right] \le \frac{\lambda_1 - 1}{n(n-1)} \sum_{j=1}^{m} \lambda_j = \frac{\lambda_1 - 1}{n-1}.$$

(2) For $\lambda_1 \ge n/2$, we have $\lambda \trianglelefteq (\lambda_1, n-\lambda_1)$, which is a partition since $\lambda_1 \ge n - \lambda_1$, so

$$\begin{aligned} r_\lambda(\tau) \le r_{(\lambda_1, n-\lambda_1)}(\tau) &= \frac{1}{n(n-1)} \left[\left(\lambda_1^2 - \lambda_1\right) + \left((n-\lambda_1)^2 - 3(n-\lambda_1)\right)\right] \\ &= \frac{1}{n(n-1)} \left[n^2 - n - 2(\lambda_1 n - \lambda_1^2 + n - \lambda_1)\right] = 1 - \frac{2(\lambda_1+1)(n-\lambda_1)}{n(n-1)}. \end{aligned}$$

□

Corollary 4.2. *Let λ be such that $r_\lambda(\tau) \ge 0$. Then*

(1) $\left|\dfrac{1}{n} + \dfrac{n-1}{n} r_\lambda(\tau)\right| \le \dfrac{\lambda_1}{n}$.

(2) If $\lambda_1 \ge n/2$, *then* $\left|\dfrac{1}{n} + \dfrac{n-1}{n} r_\lambda(\tau)\right| \le 1 - \dfrac{2(\lambda_1+1)(n-\lambda_1)}{n^2}$.

Now observe that for every partition λ, we have $r_\lambda(\tau) < 0$ iff $r_{\lambda'}(\tau) = -r_\lambda(\tau) > 0$ by the first equality in Fact 4.2. Thus we may write

$$\begin{aligned} \sum_{\rho \in Irr^*(G)} d_\rho^2 \left(\frac{1}{n} + \frac{n-1}{n} r_\rho(\tau)\right)^{2k} &\le \sum_{\substack{\lambda \ne (n): \\ r_\lambda(\tau) \ge 0}} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r_\lambda(\tau)\right)^{2k} + \sum_{\substack{\lambda \ne (n): \\ r_\lambda(\tau) < 0}} d_\lambda^2 \left(\frac{1}{n} - \frac{n-1}{n} r_\lambda(\tau)\right)^{2k} \\ &= \sum_{\substack{\lambda \ne (n): \\ r_\lambda(\tau) \ge 0}} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r_\lambda(\tau)\right)^{2k} + \sum_{\substack{\lambda' \ne (1^n): \\ r_{\lambda'}(\tau) > 0}} d_{\lambda'}^2 \left(\frac{1}{n} + \frac{n-1}{n} r_{\lambda'}(\tau)\right)^{2k} \\ &\le 2 \sum_{\substack{\lambda \ne (n): \\ r_\lambda(\tau) \ge 0}} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r_\lambda(\tau)\right)^{2k}. \end{aligned}$$

We will bound this sum by breaking it up according to whether or not λ_1 is less than $m := \lceil (1-\alpha)n \rceil$ for some $0 < \alpha < 1/2$ to be determined:

$$\begin{aligned} \sum_{\substack{\lambda \ne (n): \\ r_\lambda(\tau) \ge 0}} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r_\lambda(\tau)\right)^{2k} &= \sum_{j=1}^{m-1} \sum_{\substack{\lambda : r_\lambda(\tau) \ge 0, \\ \lambda_1 = j}} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r_\lambda(\tau)\right)^{2k} + \sum_{j=m}^{n-1} \sum_{\substack{\lambda : r_\lambda(\tau) \ge 0, \\ \lambda_1 = j}} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r_\lambda(\tau)\right)^{2k} \\ &\le \sum_{j=1}^{m-1} \sum_{\substack{\lambda : r_\lambda(\tau) \ge 0, \\ \lambda_1 = j}} d_\lambda^2 \left(\frac{j}{n}\right)^{2k} + \sum_{j=m}^{n-1} \sum_{\substack{\lambda : r_\lambda(\tau) \ge 0, \\ \lambda_1 = j}} d_\lambda^2 \left(1 - \frac{2(j+1)(n-j)}{n^2}\right)^{2k}. \end{aligned}$$

Since Corollary 4.1 implies

$$\sum_{\substack{\lambda: r_\lambda(\tau)\geq 0,\\ \lambda_1=j}} d_\lambda^2 \leq \sum_{\lambda:\lambda_1=j} \binom{n}{j}^2 d_{\lambda^*}^2 \leq \binom{n}{j}^2 \sum_{\lambda\vdash(n-j)} d_\lambda^2 = \binom{n}{j}^2 (n-j)! = \binom{n}{j}\frac{n!}{j!},$$

we obtain

$$\begin{aligned}\sum_{\substack{\lambda\neq(n):\\ r_\lambda(\tau)\geq 0}} d_\lambda^2 \left(\frac{1}{n}+\frac{n-1}{n}r_\lambda(\tau)\right)^{2k} &\leq \sum_{j=1}^{m-1}\binom{n}{j}\frac{n!}{j!}\left(\frac{j}{n}\right)^{2k} + \sum_{j=m}^{n-1}\binom{n}{j}\frac{n!}{j!}\left(1-\frac{2j(n-j+1)}{n^2}\right)^{2k}\\ &\leq \sum_{j=1}^{n-m}\binom{n}{j}\frac{n!}{(n-j)!}\left(1-\frac{2j(n-j+1)}{n^2}\right)^{2k} + \sum_{j=n-m+1}^{n-1}\binom{n}{j}\frac{n!}{(n-j)!}\left(1-\frac{j}{n}\right)^{2k}\end{aligned}$$

where the final inequality is reindexing by $j\mapsto n-j$ and noting that $(n-j)(j+1)\geq j(n-j+1)$ for $j\leq n/2$.

If $k=\frac{1}{2}n\log(n)+cn$, then, observing that $\frac{j(n-j+1)}{n}$ is minimal for $j=1$, we have

$$\left(1-\frac{2j(n-j+1)}{n^2}\right)^{2k} \leq e^{-\frac{2j(n-j+1)}{n}\log(n)}e^{-4c\frac{j(n-j+1)}{n}} \leq e^{-4c}n^{-2j}n^{\frac{2j(j-1)}{n}}.$$

As $\binom{n}{j}\frac{n!}{(n-j)!}\leq\frac{n^{2j}}{j!}$, the first sum is at most $e^{-4c}\sum_{j=1}^{n-m}\frac{1}{j!}n^{\frac{2j(j-1)}{n}}$. The ratio of successive terms is

$$\frac{j!}{(j+1)!}n^{\frac{2j(j+1)}{n}-\frac{2j(j-1)}{n}} = \frac{1}{j+1}n^{\frac{4j}{n}},$$

which is decreasing for $j<\frac{n}{4\log(n)}-1$ and increasing for $j>\frac{n}{4\log(n)}-1$. This ratio is thus largest for $j=1$ and $j=n-m-1$, and if we can show that both of these extreme values are at most $q<1$, the sum can be bounded by $\frac{1}{1-q}$ (as the $j=1$ summand is 1). For $j=1$ we get $\frac{1}{2}n^{\frac{4}{n}}$ which is less than 1 when $n\geq 17$, and for $j=n-m-1$ we get

$$\frac{1}{n-m}n^{\frac{4(n-m)}{n}} \approx \frac{1}{\alpha n}n^{4\alpha} = \frac{1}{\alpha}n^{4\alpha-1}.$$

Taking $\alpha=1/5$ guarantees that this will be smaller than 1 for $n>5^5$. It follows that if $m=\lceil\frac{4}{5}n\rceil$, then there is a constant A that does not depend on n such that

$$\sum_{j=1}^{n-m}\binom{n}{j}\frac{n!}{(n-j)!}\left(1-\frac{2j(n-j+1)}{n^2}\right)^{2k} \leq Ae^{-4c}$$

for all n.

For the second sum, we have

$$\left(1-\frac{j}{n}\right)^{2k} \leq \left(1-\frac{n-m+1}{n}\right)^{2cn}\left(1-\frac{j}{n}\right)^{n\log(n)} \leq (1-\alpha)^{2cn}\left(1-\frac{j}{n}\right)^{n\log(n)}.$$

When $\alpha=\frac{1}{5}$, $(1-\alpha)^{2n}\leq e^{-\frac{2n}{5}}\leq e^{-4}$ once $n\geq 10$, so if we can show that $\sum_{j=n-m+1}^{n-1}\binom{n}{j}\frac{n!}{(n-j)!}\left(1-\frac{j}{n}\right)^{n\log(n)}$ is bounded uniformly in n, then we will obtain

$$\left\|P^k-U\right\|_{TV}^2 \leq \frac{1}{4}\sum_{\rho\in Irr^*(G)} d_\rho^2\left(\frac{1}{n}+\frac{n-1}{n}r_\rho(\tau)\right)^{2k} \leq \frac{1}{2}\sum_{\substack{\lambda\neq(n):\\ r_\lambda(\tau)\geq 0}} d_\lambda^2\left(\frac{1}{n}+\frac{n-1}{n}r_\lambda(\tau)\right)^{2k} \leq Ce^{-4c}$$

for some universal constant C whenever $k=\frac{n}{2}\log(n)+cn$. This remaining task is left for homework.

To wrap up our analysis, we need a matching lower bound, which will be a walk in the park compared to what we just did. As before, the idea is to construct a statistic related to the slow term in the Fourier bound and compare its behavior under P^k and U.

For this, let $X(\sigma) = \sum_{j=1}^{n} 1\{\sigma(j) = j\}$ be the random variable on S_n which records the number of fixed points of a chosen permutation, and fix some $m \in [n]$.

For each $S \subseteq [n]$ with $|S| = m$, the event $A_S = \{\sigma(j) = j \text{ for all } j \in S\}$ has $\mathbb{P}_U(A_S) = \frac{(n-m)!}{n!}$. Since $\{X \geq m\}$ is equal to the union of all $\binom{n}{m}$ such subsets, countable subadditivity implies

$$\mathbb{P}_U(X \geq m) \leq \binom{n}{m} \frac{(n-m)!}{n!} = \frac{1}{m!}.$$

Now recall that we can sample from P^k by successively swapping pairs of cards chosen uniformly at random, say $(L_1, R_1), \ldots, (L_k, R_k)$, and there will certainly be at least m cards in their original positions if $|\{L_1, \ldots, L_k, R_1, \ldots R_k\}| \leq n - m$.

$\mathbb{P}_k(X \geq m)$ is thus at least the probability that it takes more than $2k$ rounds of coupon collecting to obtain $\ell := n - m + 1$ distinct types.

To bound this probability, set $T_0 = 0$ and $T_j = \inf\{t : j \text{ distinct coupons have been collected by time } t\}$ for $j \in [n]$. Then $T_\ell = \sum_{j=1}^{\ell} R_j$ where $R_j = T_j - T_{j-1}$.

By construction, these summands are independent with $R_j \sim \text{Geom}(\frac{n-j+1}{n})$, so for any $s, t > 0$,

$$\mathbb{P}(T_\ell \leq t) = \mathbb{P}\left(e^{-sT_\ell} \geq e^{-st}\right) \leq e^{st}\mathbb{E}\left[e^{-sT_\ell}\right] = e^{st} \prod_{j=1}^{\ell} \mathbb{E}\left[e^{-sR_j}\right].$$

As the moment generating function for $R \sim \text{Geom}(p)$ is

$$\mathbb{E}\left[e^{sR}\right] = \sum_{j=1}^{\infty} p(1-p)^{j-1}e^{sj} = pe^s \sum_{j=0}^{\infty} \left[e^s(1-p)\right]^j = \frac{pe^s}{1 - e^s(1-p)} = \frac{p}{e^{-s} - 1 + p},$$

we have

$$\mathbb{P}(T_\ell \leq t) \leq e^{st} \prod_{j=1}^{\ell} \frac{\frac{n-j+1}{n}}{e^s - 1 + \frac{n-j+1}{n}} = e^{st} \prod_{j=m}^{n} \frac{\frac{j}{n}}{e^s - 1 + \frac{j}{n}}.$$

Taking $s = \frac{1}{n}$ and $t = n\log(n) - 2cn$ for some $c > 0$, we have $e^{st} = ne^{-2c}$ and $e^s \geq 1 + \frac{1}{n}$, hence

$$\mathbb{P}(T_\ell \leq n\log(n) - 2cn) \leq ne^{-2c} \prod_{j=m}^{n} \frac{j}{j+1} = ne^{-2c}\frac{m}{n+1} < me^{-2c}.$$

Thus for $k = \frac{n}{2}\log(n) - cn$,

$$\mathbb{P}_k(X \geq m) \geq \mathbb{P}(T_\ell > 2k) = 1 - \mathbb{P}(T_\ell \leq n\log(n) - 2cn) > 1 - me^{-2c}.$$

Putting everything together yields

$$\left\|P^k - U\right\|_{TV} \geq \mathbb{P}_k(X \geq m) - \mathbb{P}_U(X \geq m) > 1 - me^{-2c} - \frac{1}{m!}.$$

Remark 4.1. It is advantageous to allow m to vary with n as we will soon be interested in what happens as $n \to \infty$, but fixing $m = 2$ shows that the distance to stationarity is always greater than $\frac{1}{2} - 2e^{-2c}$ after $\frac{1}{2}n\log(n) - cn$ steps, which is strictly positive whenever $c > \log(2)$.

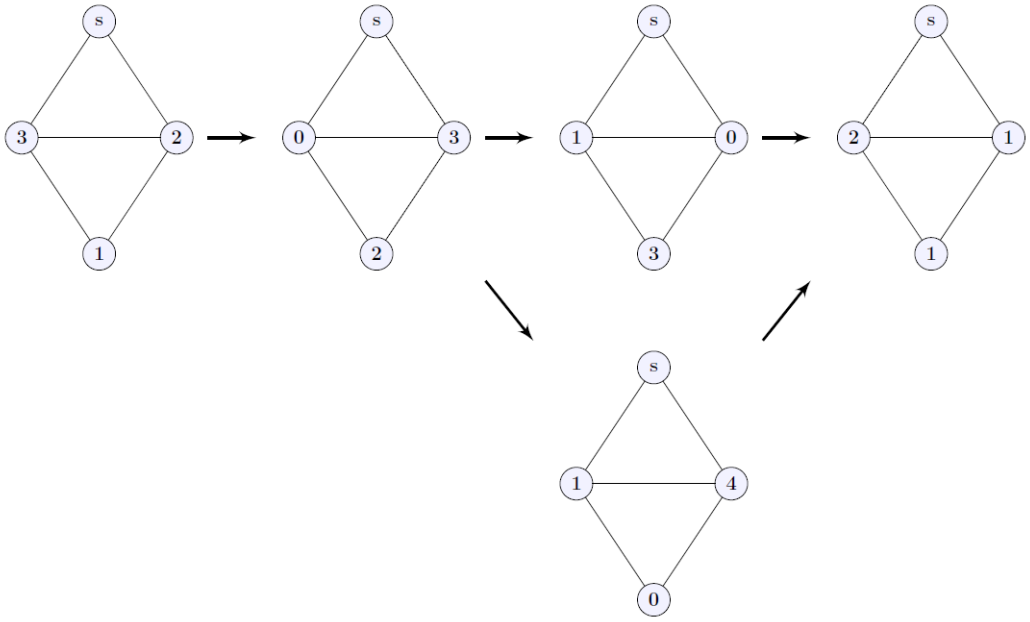
4.2.4 Abelian sandpile

Let $G = (V, E)$ be a finite, simple, connected, and undirected graph, and label the vertices $v_1, \dots, v_n$ however you choose, with the understanding that $s := v_n$ is a distinguished vertex called the *sink*.

A *sandpile* is a configuration of 'chips' on the nonsink vertices $\widetilde{V} = V \setminus \{s\}$, which we can express as the function $\eta : \widetilde{V} \to \mathbb{N}_0$ with $\eta(v)$ the number of chips at vertex v. We say that η is *stable* if $\eta(v) < \deg(v)$ for all $v \in \widetilde{V}$. If η is unstable, then we can successively *topple* vertices with $\eta(v) \geq \deg(v)$ by sending a chip to each of their neighbors, where chips falling into the sink exit the system forever. Thus toppling v results in the new configuration η' with

$$\eta'(u) = \eta(u) + 1\{u \sim v\} - \deg(v)1\{u = v\}.$$

The presence of the sink ensures that any configuration can be stabilized with finitely many topplings, and one can show that the final stable configuration, η°, does not depend on the order in which vertices are toppled.



Write $\mathcal{S}$ for the set of stable configurations on G and define the binary operation of pointwise addition followed by stabilization, $\eta \oplus \sigma = (\eta + \sigma)^\circ$. This makes $\mathcal{S}$ into a commutative monoid, so we can consider the random walk defined by $\eta_t = \eta_{t-1} \oplus \sigma_t$ where $\sigma_1, \sigma_2, \dots$ are sampled independently from some probability μ on $\mathcal{S}$. We will focus on the case where μ is uniform on V so that at each time step we drop a chip on a random vertex and then stabilize. This is clearly irreducible, and it is aperiodic since dropping a chip on the sink corresponds to holding. (In fact, the chain driven by the uniform distribution on $\widetilde{V}$ is irreducible and aperiodic as well since one can show that 1 is a simple eigenvalue and all others have strictly smaller modulus.)

It is easy to see that the 'saturated configuration' $\eta_*(v) = \deg(v) - 1$ is accessible from every stable configuration, so the chain will eventually be absorbed in the recurrent communicating class $\mathcal{G} = \eta_* \oplus \mathcal{S}$. It turns out that this set of *recurrent configurations* is the minimal ideal of $\mathcal{S}$, and this implies that $\mathcal{G}$ is an abelian group under $\oplus$. As the chain will eventually end up in $\mathcal{G}$ anyway, we can just restrict the state space to $\mathcal{G}$ to begin with so that we are back in familiar territory.

Note that the group identity $id \in \mathcal{G}$ is generally distinct from the monoid identity $\iota \equiv 0$ in $\mathcal{S}$. Specifically, id is the unique recurrent configuration from which ι can be obtained by a sequence of topplings. This is given by $\big(2\eta_* - (2\eta_*)^\circ\big)^\circ$ because $(2\eta_*)^\circ \leq \eta_*$, so $2\eta_* - (2\eta_*)^\circ \geq \eta_*$ and thus is accessible from η_*, and the sequence of topplings that stabilizes $2\eta_*$ takes $2\eta_* - (2\eta_*)^\circ$ to ι.

Since id belongs to the ideal $\mathcal{G}$, each stable configuration $\sigma \in \mathcal{S}$ has a 'recurrent representative' $\sigma \oplus id \in \mathcal{G}$ such that for each $\eta \in \mathcal{G}$, $(\sigma \oplus id) \oplus \eta = \sigma \oplus (id \oplus \eta) = \sigma \oplus \eta$. The measure driving the walk on $\mathcal{G}$ is thus the pushforward of the uniform distribution on $\big\{\delta_v : v \in V\big\}$ under the map $\sigma \mapsto \sigma \oplus id$ where $\delta_v(u) = 1\{u = v\}$ for $v \in \widetilde{V}$ and $\delta_s \equiv 0$.

While the foregoing is great for explaining the dynamics of our Markov chain, it will be helpful to have a more concrete description of the *sandpile group* $\mathcal{G}$. For this, define the *Laplacian* of G to be the $n \times n$ matrix

$$\overline{\Delta}_{i,j} = \begin{cases} \deg(v_i), & j = i \\ -1, & v_j \sim v_i \\ 0, & \text{otherwise} \end{cases},$$

and define the *reduced Laplacian*, Δ, to be the $(n-1) \times (n-1)$ submatrix of $\overline{\Delta}$ formed by deleting the n^{th} row and column (corresponding to the sink).

One can show that $\mathcal{G} \cong \mathbb{Z}^{n-1}/\Delta\mathbb{Z}^{n-1}$, the integer cokernel of the reduced Laplacian, via the isomorphism $\eta \mapsto \big(\eta(v_1), \ldots, \eta(v_{n-1})\big) + \Delta\mathbb{Z}^{n-1}$. The idea is that $\mathbf{x} \in \mathbb{Z}^{n-1}$ corresponds to the configuration with x_i chips at vertex v_i where we are now allowing sites to have a negative number chips, and $\mathbf{x} + \Delta\mathbf{y}$ is the configuration obtained from $\mathbf{x}$ by performing $-y_i$ topplings at vertex v_i for $i = 1, \ldots, n-1$. (A negative toppling means taking a chip from each neighbor, with the sink an infinite reservoir). This characterization shows that $|\mathcal{G}| = \det(\Delta)$, which is equal to the number of spanning trees in G by the matrix tree theorem.

It also allows us to compute the irreducible characters of $\mathcal{G}$. To begin, let's say that a function h from V to $\mathbb{T} = \big\{z \in \mathbb{C} : |z| = 1\big\}$ is *multiplicative harmonic* if $h(s) = 1$ and $h(v)^{\deg(v)} = \prod_{u \sim v} h(u)$ for all $v \in V$. Write $\mathcal{H}$ for the set of all multiplicative harmonic functions on G. $h \equiv 1$ is clearly multiplicative harmonic, so $\mathcal{H}$ is nonempty, and it is easy to see that it forms an abelian group under pointwise multiplication.

Now given $h \in \mathcal{H}$, define $\chi'_h : \mathbb{Z}^{n-1} \to \mathbb{T}$ by $\chi'_h(\mathbf{z}) = \prod_{j=1}^n h(v_j)^{z_j}$, and define $\chi_h : \mathbb{Z}^{n-1}/\Delta\mathbb{Z}^{n-1}$ by $\chi_h(\mathbf{z} + \Delta\mathbb{Z}^{n-1}) = \chi'_h(\mathbf{z})$. Clearly, $\chi'_h(\mathbf{w} + \mathbf{z}) = \chi'_h(\mathbf{w})\chi'_h(\mathbf{z})$ for all $\mathbf{w}, \mathbf{z} \in \mathbb{Z}^{n-1}$, so χ'_h is a homomorphism. Moreover, for each standard basis vector $\mathbf{e}_j$,

$$\chi'_h(\Delta\mathbf{e}_j) = \chi'_h\big(\deg(v_j)\mathbf{e}_j - \sum_{v_k \sim v_j} \mathbf{e}_k\big) = h(v_j)^{\deg(v_j)} \prod_{v_k \sim v_j} h(v_k)^{-1} = 1,$$

so χ'_h is constant on cosets of $\Delta\mathbb{Z}^{n-1}$. It follows that χ'_h descends to a character χ_h of $\mathcal{G}$.

Conversely, every irreducible character χ of $\mathbb{Z}^{n-1}/\Delta\mathbb{Z}^{n-1}$ lifts to a character χ' of $\mathbb{Z}^{n-1}$ that is constant on cosets of $\Delta\mathbb{Z}^{n-1}$. For such χ', define $h : V \to \mathbb{T}$ by $h(v_j) = \chi'(\mathbf{e}_j)$ for $1 \leq j \leq n-1$ and $h(s) = 1$. Since χ' is is a homomorphism, we have

$$\chi'(\mathbf{z}) = \chi'\Big(\sum_{j=1}^{n-1} z_j\mathbf{e}_j\Big) = \prod_{j=1}^{n-1} \chi'(\mathbf{e}_j)^{z_j} = \prod_{j=1}^{n-1} h(v_j)^{z_j}.$$

Moreover, $\chi' \equiv 1$ on $\Delta\mathbb{Z}^{n-1}$, so for each $j \in [n-1]$,

$$1 = \chi'(\Delta\mathbf{e}_j) = \chi'(v_j)^{\deg(v_j)} \prod_{v_k \sim v_j} \chi'(v_k)^{-1} = h(v_j)^{\deg(v_j)} \prod_{v_k \sim v_j} h(v_k)^{-1}.$$

This shows that $\chi = \chi_h$ with $h \in \mathcal{H}$, and we conclude that the irreducible characters of $\mathbb{Z}^{n-1}/\Delta\mathbb{Z}^{n-1}$ are precisely $\{\chi_h : h \in \mathcal{H}\}$.

In the language of sandpile configurations, the characters of $\mathcal{G}$ are $\{f_h : h \in \mathcal{H}\}$ with $f_h : \mathcal{G} \to \mathbb{T}$ given by $f_h(\eta) = \prod_{v \in V} h(v)^{\eta(v)}$. Example 4.1 shows that these characters form an orthonormal basis of eigenfunctions of the sandpile chain with f_h corresponding to the eigenvalue $\lambda_h = \frac{1}{n}\sum_{v \in V} f_h(\delta_v) = \frac{1}{n}\sum_{v \in V} h(v)$. Moreover, we have the mixing bound

$$\left\|P^k - U\right\|_{TV}^2 \leq \frac{1}{4} \sum_{h \in \mathcal{H} \setminus \{1\}} |\lambda_h|^{2k}.$$

Example 4.5. If $G = C_n$ is the n-cycle with vertices successively labeled $v_1, \ldots, v_n$, then for each n^{th} root of unity ω, the function $h(v_k) = \omega^k$ is multiplicative harmonic since

$$h(v_k)^{\deg(v_k)} = \omega^{2k} = \omega^{k-1}\omega^{k+1} = \prod_{v_j \sim v_k} h(v_j).$$

As there are n spanning trees on the n-cycle (each corresponding to a different deleted edge), this accounts for all of $\mathcal{H}$.

The eigenvalues are thus $\lambda_n = \frac{1}{n}\sum_{k=1}^{n} 1^k = 1$ and $\lambda_j = \frac{1}{n}\sum_{k=1}^{n} \omega^k = \frac{1}{n} \cdot \frac{\omega^{n+1}-\omega}{1-\omega} = 0$ for $1 \leq j \leq n-1$, so we see that $\left\|P^k - U\right\|_{TV}^2 \leq \frac{1}{4}\sum_{j=1}^{n-1} |\lambda_j|^{2k} = 0$ for each $k \in \mathbb{N}$; the random walk mixes completely after a single step.

Example 4.6. Let $G = K_n$ be the complete graph on $n \geq 3$ vertices. Set

$$M = \left\{z \in (\mathbb{Z}/n\mathbb{Z})^n : \textstyle\sum_{j=1}^{n-1} z_j = z_n = 0\right\}$$

with addition taken modulo n, and write $\omega = e^{\frac{2\pi i}{n}}$.

For each $z \in M$, the function $h_z(v_j) = \omega^{z_j}$ is multiplicative harmonic since $\prod_{j=1}^{n} h_z(v_j) = \omega^{\sum_{j=1}^{n} z_j} = 1$, so

$$\prod_{j \neq k} h_z(v_j) = \omega^{-z_k} = h_z(v_k)^{-1} = h_z(v_k)^{n-1}$$

for all $k = 1 \ldots, n$. Since each element of M is uniquely determined by specifying the first $n-2$ coordinates, $|M| = n^{n-2} = (\#$ spanning trees of $K_n)$, hence $\mathcal{H} = \{h_z\}_{z \in M}$.

The eigenvalues are thus $\lambda_z = \frac{1}{n}\sum_{j=1}^{n} \omega^{z_j}$ as z ranges over M. As no $z \in M \setminus \{0\}$ can have all coordinates in $\{1, 0\}$ or $\{-1, 0\}$ (because then the coordinates would not sum to 0), we see that $z_* = (1, -1, 0, \ldots, 0)$ gives a nontrivial eigenvalue of maximum modulus. (Any permutation of the first $n-1$ coordinates of z_* or $\pm(2, 1, \ldots, 1, 0)$ also gives a nontrivial eigenvalue of maximum modulus.) The top eigenvalue is thus

$$\lambda_{z_*} = \frac{1}{n}\left(\omega + \omega^{-1} + n - 2\right) = 1 - \frac{2}{n}\left(1 - \cos(2\pi/n)\right) \approx 1 - \frac{4\pi^2}{n^3}.$$

The corresponding eigenfunction is $f_*(\eta) = \prod_{v \in V} h_{z_*}(v)^{\eta(v)} = \omega^{\eta(v_1) - \eta(v_2)}$, which records information about the difference $\bmod n$ of chips at v_1 and v_2.

For the chain to equilibrate, it must run long enough for this chip difference to randomize, and since v_1 and v_2 have the same neighbors, it changes only when a chip is dropped at v_1 or v_2. This happens about once every $n/2$ steps and each such chip addition has the effect on $\eta(v_1) - \eta(v_2) \pmod n$ of a step in the random walk on the n-cycle, which takes order n^2 steps to reach stationarity. It follows that the sandpile chain on K_n needs at least order n^3 steps.

On the other hand, we have

$$\left\|P^k - U\right\|_{TV}^2 \leq \frac{1}{4} \sum_{z \in M \setminus \{0\}} |\lambda_z|^{2k} \leq n^{n-2} |\lambda_{z_*}|^{2k} \approx n^{n-2} \left(1 - \frac{4\pi^2}{n^3}\right)^{2k} \leq n^n e^{-\frac{8\pi^2 k}{n^3}},$$

so if $k = \frac{n^4}{8\pi^2}\log(n) + cn^3$, then $\left\|P^k - U\right\|_{TV}^2 \leq n^n e^{-n\log(n)} e^{-8\pi^2 c} = e^{-8\pi^2 c}$, hence order $n^4 \log(n)$ steps suffice.

One can show (in a precise sense to be discussed presently) that $\frac{1}{4\pi^2} n^3 \log(n)$ steps are necessary and sufficient for convergence.

Note that simple random walk on K_n (with $1/n$ holding) mixes immediately whereas the sandpile chain on K_n takes order $n^3 \log(n)$ steps. In contrast, the sandpile chain on C_n mixes immediately, but the simple random walk takes order n^2 steps. It has been shown that this inverse behavior is typical: If the random walk mixes sufficiently rapidly, then the sandpile chain mixes slowly!

Much more can be said about the sandpile chain both in general and for specific graphs, but we will conclude with a simple universal bound on its eigenvalues in terms of the number of vertices and maximum degree of the underlying graph. (We will obtain general mixing bounds based on the top nontrivial eigenvalue in the next section.)

The general idea is that large eigenvalues correspond to 'nearly constant' multiplicative harmonic functions, so one can show that the eigenvalues are small if the functions in $\mathcal{H}$ have large ranges. (In fact, this line of reasoning can be pursued much further to get sharp eigenvalue bounds in terms of the norms of vectors in an appropriate 'dual Laplacian lattice.')

To facilitate the argument, we give yet another cosine bound.

Lemma 4.4. *If $0 < x \leq \pi/2$, then $1 - \cos(x) \geq \frac{4x^2}{\pi^2}$.*

Proof. Define $f : \mathbb{R}^+ \to \mathbb{R}$ by $f(x) = \frac{1-\cos(x)}{x^2}$. Then $f'(x) = \frac{g(x)}{x^3}$ with $g(x) = x\sin(x) - 2\big(1 - \cos(x)\big)$, so if we can show that $g(x) \leq 0$ for $0 < x \leq \pi/2$, then it will follow that $1 - \cos(x) = x^2 f(x) \geq x^2 f(\pi/2) = \frac{4x^2}{\pi^2}$. For this, observe that $g'(x) = x\cos(x) - \sin(x)$ and $g''(x) = -x\sin(x) \leq 0$ for $x \in (0, \pi/2]$, hence $g'(x)$ is decreasing on $(0, \pi/2]$ with $\lim_{x\to 0^+} g'(x) = 0$. This shows that g is decreasing on $(0, \pi/2]$ with $\lim_{x\to 0^+} g(x) = 0$, and the proof is complete. □

Theorem 4.4. *If G is a graph on n vertices with maximum degree d, then every nontrivial eigenvalue of the sandpile chain satisfies $|\lambda| \leq 1 - \frac{8}{d^2 n}$.*

Proof. Any $h \in \mathcal{H}$ can be written as $h(v) = e^{2\pi i g(v)}$ for some function $g : V \to [0, 1]$.

Writing $\overline{\Delta}$ for the full Laplacian of G, we see that $h \in \mathcal{H}$ implies

$$e^{2\pi i \deg(v) g(v)} = h(v)^{\deg(v)} = \prod_{u \sim v} h(v) = e^{2\pi i \sum_{u\sim v} g(u)},$$

so $\exp\big(2\pi i\big(\overline{\Delta} g\big)(v)\big) = \exp\left(2\pi i \left(\deg(v) g(v) - \sum_{u\sim v} g(u)\right)\right) = 1$ and thus $\overline{\Delta} g \in \mathbb{Z}$.

If the range of g is $[a, b]$ with $0 < b - a < \frac{1}{d}$, then

$$\left|\overline{\Delta} g(v)\right| = \left|\sum_{u\sim v} g(u) - g(v)\right| \leq \sum_{u\sim v} |g(u) - g(v)| < \frac{\deg(v)}{d} \leq 1,$$

so it must be the case that $\overline{\Delta} g \equiv 0$. Uniqueness of harmonic extensions implies $g \equiv g(s)$, so $h \equiv 1$.

Therefore, no nontrivial multiplicative harmonic function maps V to a segment of the unit circle with arc length less than $\frac{2\pi}{d}$ and thus every nontrivial eigenvalue satisfies

$$|\lambda_h| = \left|\frac{1}{n}\sum_{v\in V} h(v)\right| \leq \frac{1}{n}\left(n - 2 + 2\cos(\pi/d)\right) = 1 - \frac{2}{n}\left(1 - \cos(\pi/d)\right) \leq 1 - \frac{8}{d^2 n}.$$

The first inequality is because two of the values of h are at distance at least $2\pi/d$, and subject to this constraint, the sum $\sum_{v\in V} h(v)$ has minimum modulus when two vertices get mapped to points on the circle at distance $2\pi/d$ apart and the rest get mapped to the midpoint.

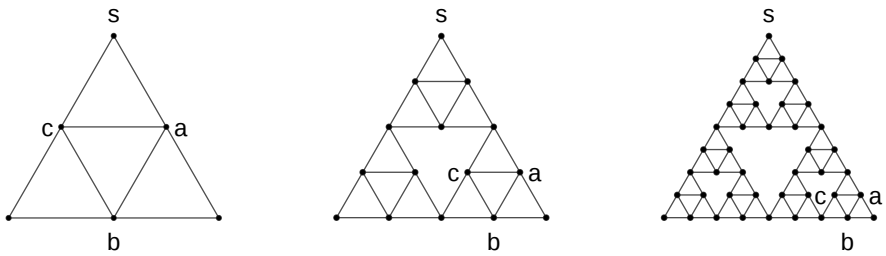
The final inequality is Lemma 4.4 and $d \geq 2$ for $n \geq 3$. (When $n \leq 2$, $\mathcal{G}$ is trivial and the assertion is vacuously true.) □

Remark 4.2. If h is multiplicative harmonic for G with sink $s = v$, then $h(u)^{-1}h$ is multiplicative harmonic for G with sink $s = u$, so the structure of the character group of $\mathcal{G}$, and thus of $\mathcal{G}$ itself, does not depend on the choice of sink vertex.

As such, there is no loss in placing the sink at a vertex of maximal degree, and examination of the proof of Theorem 4.4 shows that d may be taken to be the maximum degree of the nonsink vertices. It follows that d can be replaced by the penultimate term in the nondecreasing degree sequence of G if desired.

Example 4.7. The m-fold Sierpinski Gasket graph $SG_2(m)$ can be constructed recursively as follows: $SG_2(0)$ is the triangle K_3, and for $m \geq 1$, $SG_2(m)$ is obtained by gluing three copies of $SG_2(m-1)$ together to form a triangle with center cut out.

This recursive construction shows that $SG_2(m)$ has $|V(m)| = \frac{1}{2}\left(3^{m+1} + 3\right)$ vertices and maximal degree $d = 4$. We will take the sink to be the 'topmost' vertex.



It is known that the number of spanning trees of $SG_2(m)$ (and thus the order of its sandpile group) is $2^{\alpha(m)}3^{\beta(m)}5^{\gamma(m)}$ with $\alpha(m) = \frac{1}{2}(3^m - 1)$, $\beta(m) = \frac{1}{4}\left(3^{m+1} + 2m + 1\right)$, and $\gamma(m) = \frac{1}{4}\left(3^m - 2m - 1\right)$, though the invariant factor decomposition remains an open problem.

Referring to the figure above, define the function $h : V(m) \to \mathbb{T}$ by $h(a) = h(b) = h(c) = -1$ and $h(v) = 1$ otherwise. h is clearly multiplicative harmonic and the associated eigenvalue satisfies $|\lambda_h| = 1 - \frac{6}{|V(m)|}$. Also, Theorem 4.4 shows that the top nontrivial eigenvalue has modulus at most $1 - \frac{1}{2|V(m)|}$.

Thus even though we don't know the structure of the sandpile group of $SG_2(m)$, we can determine the modulus of the top eigenvalue of the sandpile chain with remarkable precision.

4.3 Spectral Decomposition of the Transition Matrix

If $G = \{s_1, ..., s_N\}$ is a finite group and μ a probability on G, then the transition matrix for the random walk (G, μ) is given by $Q(s_i, s_j) = \mu(s_j s_i^{-1})$.

Let $\rho_1, ..., \rho_m$ denote the irreducible representations of G with respect to bases such that $\rho_k(s_i)$ is unitary for each i, k. For $k = 1, ..., m$, define the $d_k^2 \times d_k^2$ block diagonal matrix

$$B_k = I_{d_k} \otimes \widehat{\mu}(\rho_k) = \begin{bmatrix} \widehat{\mu}(\rho_k) & & 0 \\ & \ddots & \\ 0 & & \widehat{\mu}(\rho_k) \end{bmatrix},$$

and define the $N \times N$ block diagonal matrix

$$B = \begin{bmatrix} B_1 & & & 0 \\ & B_2 & & \\ & & \ddots & \\ 0 & & & B_m \end{bmatrix}.$$

(The dimensions check out since the sum of the squared degrees of the irreps equals the order of the group.)

Let

$$\psi_k(s) = \sqrt{\frac{d_k}{N}} \operatorname{vec}\big(\rho_k(s)\big)$$

where the *vectorization operator* takes a matrix in $\mathbb{C}^{p \times q}$ to the vector in $\mathbb{C}^{pq}$ obtained by stacking its columns:

$$\operatorname{vec}(M) = \begin{bmatrix} M_{1,1} & \cdots & M_{p,1} & M_{1,2} & \cdots & M_{p,2} & \cdots & M_{1,q} & \cdots & M_{p,q} \end{bmatrix}^T.$$

Straightforward computations yield the identities $\operatorname{Tr}(A^\dagger B) = \operatorname{vec}(A)^\dagger \operatorname{vec}(B)$ and $\operatorname{vec}(AB) = (I \otimes A)\operatorname{vec}(B)$, and thus

$$\operatorname{Tr}(X^\dagger Y Z) = \operatorname{vec}(X)^\dagger (I \otimes Y) \operatorname{vec}(Z).$$

Let

$$\psi(s) = \begin{bmatrix} \psi_1(s) \\ \vdots \\ \psi_m(s) \end{bmatrix}$$

be the vector in $\mathbb{C}^N$ obtained by likewise concatenating the $\psi_k(s)$'s, and define the $N \times N$ matrix Φ by

$$\Phi = \begin{bmatrix} \psi(s_1)^T \\ \psi(s_2)^T \\ \vdots \\ \psi(s_N)^T \end{bmatrix} = \begin{bmatrix} \psi_1(s_1)^T & \psi_2(s_1)^T & \cdots & \psi_m(s_1)^T \\ \psi_1(s_2)^T & \psi_2(s_2)^T & \cdots & \psi_m(s_2)^T \\ \vdots & \vdots & & \vdots \\ \psi_1(s_N)^T & \psi_2(s_N)^T & \cdots & \psi_m(s_N)^T \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{\frac{d_1}{N}}\rho_1(s_1)_{1,1} & \sqrt{\frac{d_1}{N}}\rho_1(s_1)_{2,1} & \cdots & \sqrt{\frac{d_1}{N}}\rho_1(s_1)_{d_1,d_1} & \cdots & \sqrt{\frac{d_m}{N}}\rho_m(s_1)_{1,1} & \cdots & \sqrt{\frac{d_m}{N}}\rho_m(s_1)_{d_m,d_m} \\ \sqrt{\frac{d_1}{N}}\rho_1(s_2)_{1,1} & \sqrt{\frac{d_1}{N}}\rho_1(s_2)_{2,1} & \cdots & \sqrt{\frac{d_1}{N}}\rho_1(s_2)_{d_1,d_1} & \cdots & \sqrt{\frac{d_m}{N}}\rho_m(s_2)_{1,1} & \cdots & \sqrt{\frac{d_m}{N}}\rho_m(s_2)_{d_m,d_m} \\ \vdots & & & \vdots & & & & \vdots \\ \sqrt{\frac{d_1}{N}}\rho_1(s_N)_{1,1} & \sqrt{\frac{d_1}{N}}\rho_1(s_N)_{2,1} & \cdots & \sqrt{\frac{d_1}{N}}\rho_1(s_N)_{d_1,d_1} & \cdots & \sqrt{\frac{d_m}{N}}\rho_m(s_N)_{1,1} & \cdots & \sqrt{\frac{d_m}{N}}\rho_m(s_N)_{d_m,d_m} \end{bmatrix}.$$

Now recall that if we define an inner product on $\mathbb{C}^G$ by $(f \,|\, g) = \frac{1}{N}\sum_{s\in G} f(s)\overline{g(s)}$, then it follows from the Schur orthogonality relations that the matrix entries of the unitary irreducible representations satisfy $((\rho_a)_{i,j} \,|\, (\rho_b)_{k,l}) = 0$ for all i,j,k,l if $a \neq b$, and $((\rho_a)_{i,j} \,|\, (\rho_a)_{k,l}) = \frac{1}{d_a}\delta_{ik}\delta_{jl}$.

Consequently, the columns of Φ are orthonormal with respect to the standard inner product on $\mathbb{C}^N$, so Φ is a unitary matrix.

I claim that we can write $Q = \Phi B^T \Phi^{-1}$. To see that this is so, note that by Fourier inversion we have

$$\begin{aligned}
Q(s_i, s_j) = \mu(s_j s_i^{-1}) &= \frac{1}{N}\sum_{k=1}^{m} d_k \mathrm{Tr}\left(\rho_k(s_i s_j^{-1})\widehat{\mu}(\rho_k)\right) \\
&= \frac{1}{N}\sum_{k=1}^{m} d_k \mathrm{Tr}\left(\rho_k(s_i)\rho_k(s_j)^\dagger \widehat{\mu}(\rho_k)\right) \\
&= \frac{1}{N}\sum_{k=1}^{m} d_k \mathrm{Tr}\left(\rho_k(s_j)^\dagger \widehat{\mu}(\rho_k)\rho_k(s_i)\right) \\
&= \sum_{k=1}^{m} \mathrm{vec}\left(\sqrt{\frac{d_k}{N}}\rho_k(s_j)\right)^\dagger (I \otimes \widehat{\mu}(\rho_k))\,\mathrm{vec}\left(\sqrt{\frac{d_k}{N}}\rho_k(s_i)\right) \\
&= \sum_{k=1}^{m} \psi_k(s_j)^\dagger B_k \psi_k(s_i) = \left(\overline{\Phi} B \Phi^T\right)(s_j, s_i) \\
&= \left(\overline{\Phi} B \Phi^T\right)^T (s_i, s_j) = \Phi B^T \Phi^\dagger(s_i, s_j).
\end{aligned}$$

The utility of this decomposition becomes especially evident when the measure μ driving the random walk is constant on conjugacy classes since Fourier transforms of class functions are homotheties.

Specifically, if $f(s) = f(t)$ whenever s and t are conjugate, then, for ρ a unitary irreducible representation of G, $\widehat{f}(\rho) = \lambda I$ with I the $d_\rho \times d_\rho$ identity matrix and $\lambda = \frac{1}{d_\rho}\sum_{s\in G} f(s)\chi_\rho(s)$.

It follows that Q decomposes as $Q = \Phi B \Phi^{-1}$ where $B = B^T$ is a diagonal matrix and Φ is a unitary matrix. As such, the columns of Φ give an orthonormal basis of eigenvectors and the diagonal entries of B give the corresponding eigenvalues.

We record this result as

Theorem 4.5. *Suppose that Q is the transition matrix for the random walk on a finite group G which is driven by a measure μ that is constant on the conjugacy classes of G.*

If $\rho_1, ..., \rho_m$ are the unitary irreducible representations of G, then for each $k = 1, ..., m$, Q has an eigenvalue $\lambda_k = \frac{1}{d_k}\sum_{s\in G}\mu(s)\chi_k(s)$ occurring with multiplicity d_k^2.

An orthonormal basis of eigenfunctions for the eigenspace corresponding to λ_k is given by the (normalized) matrix entries of ρ_k, $\left\{\sqrt{\frac{d_k}{|G|}}\rho_k(s)_{i,j}\right\}_{i,j=1}^{d_k}$.

In particular, the eigenfunctions of Q depend only on the irreducible representations of G, so every random walk on G driven by a measure which is constant on conjugacy classes has the same set of eigenspaces (though the eigenvalues will vary with the measure defining the walk). Thus if we have determined an eigenbasis corresponding to the representation ρ_k for one such Markov chain, then we have found that eigenbasis for all of them!

A remarkable feature of Markov chain mixing is that in many natural examples there is a sharp transition to equilibrium in the sense that $\sup_\mu \left\|\mu P^k - \pi\right\|_{TV}$ stays close to 1 for some time and then abruptly drops and tends rapidly to 0. Roughly, it takes a long time for the chain to forget some crucial distinguishing feature of its initial distribution, but once this obstacle is overcome, the remaining mixing takes place over a much shorter time period.

5.1 Definitions and Examples

The behavior described above is known as the *cutoff phenomenon* and can be formalized as follows:

Suppose we are given a sequence of ergodic Markov chains $P_1, P_2, \ldots$ having finite state spaces $\Omega_1, \Omega_2, \ldots$ and stationary distributions $\pi_1, \pi_2, \ldots$, and let $t_{\text{mix}}^{(n)}(\varepsilon) = \min\left\{t : \max_{x\in\Omega_n} \|\delta_x P_n^t - \pi_n\|_{TV} < \varepsilon\right\}$ denote the ε-mixing time of P_n.

We say that $\{P_n\}$ exhibits cutoff if for all $\varepsilon \in (0,1)$,

$$\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} = 1. \tag{5.1}$$

Now Proposition 2.3 implies that if $\delta_1 < \delta_2$, then $t_{\text{mix}}^{(n)}(\delta_1) \geq t_{\text{mix}}^{(n)}(\delta_2)$, so if $\{P_n\}$ exhibits cutoff, then for any $0 < \varepsilon_1 < \varepsilon_2 < 1$, taking $\varepsilon = \min\left\{\varepsilon_1, 1-\varepsilon_2\right\}$ gives

$$1 \leq \frac{t_{\text{mix}}^{(n)}(\varepsilon_1)}{t_{\text{mix}}^{(n)}(\varepsilon_2)} \leq \frac{t_{\text{mix}}(\varepsilon)}{t_{\text{mix}}(1-\varepsilon)} \to 1.$$

(Taking reciprocals shows that $t_{\text{mix}}^{(n)}(\varepsilon_2)/t_{\text{mix}}^{(n)}(\varepsilon_1) \to 1$ as well.)

Thus cutoff means that the ε-mixing time is asymptotically independent of ε, so it makes sense to just refer to the arbitrary value $t_{\text{mix}}^{(n)} = t_{\text{mix}}^{(n)}(1/4)$ when discussing the mixing behavior.

With this in mind, we have the equivalent formulation that $\{P_n\}$ exhibits cutoff at time $t_n = t_{\text{mix}}^{(n)}$ if $d_n(t) = \max_{x\in\Omega_n} \|\delta_x P_n^t - \pi_n\|_{TV}$ satisfies

$$\lim_{n\to\infty} d_n(ct_n) = \begin{cases} 1, & c < 1 \\ 0, & c > 1 \end{cases}.$$

(We take it as implicit that arguments of d_n are to be rounded to an appropriate nearest nonnegative integer.)

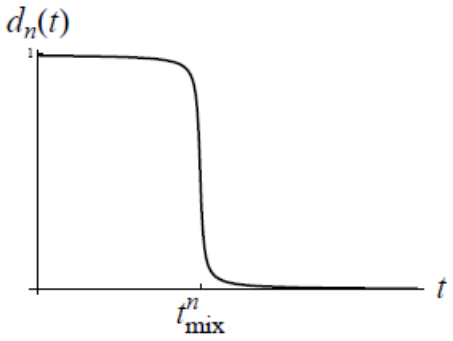
Indeed, if d_n converges to a step function when time is rescaled by t_n, then for any $\varepsilon, \gamma \in (0, 1/2)$, we have that for sufficiently large n, $t_{\text{mix}}^{(n)}(\varepsilon) \leq (1+\gamma)t_n$ and $t_{\text{mix}}^{(n)}(1-\varepsilon) \geq (1-\gamma)t_n$, hence $\frac{t^{(n)}(\varepsilon)}{t^{(n)}(1-\varepsilon)} \leq \frac{1+\gamma}{1-\gamma}$.

Letting $\gamma \searrow 0$ shows that the chain exhibits cutoff.

Conversely, suppose that $\{P_n\}$ satisfies Equation (5.1), and fix $\varepsilon \in (0,1)$, $0 < c_1 < 1 < c_2$. Then our previous observation implies $\frac{t_n}{t_{\text{mix}}^{(n)}(1-\varepsilon)}, \frac{t_n}{t_{\text{mix}}^{(n)}(\varepsilon)} \to 1$, so for n sufficiently large, we have $c_1 t_n < t_{\text{mix}}^{(n)}(1-\varepsilon)$ and $c_2 t_n > t_{\text{mix}}^{(n)}(\varepsilon)$, hence $d_n(c_1 t_n) \geq 1-\varepsilon$ and $d_n(c_2 t_n) \leq \varepsilon$.

As $\varepsilon \in (0,1)$ was arbitrary, we see that $\lim_{n\to\infty} d_n(c_1 t_n) = 1$ and $\lim_{n\to\infty} d_n(c_2 t_n) = 0$.

This characterization shows that if the sequence exhibits cutoff, then $t_{\text{mix}}^{(n)}$ steps are necessary and sufficient for the chain to mix.



Remark 5.1. Though we have phrased the cutoff phenomenon in terms of the total variation mixing time, one can similarly define cutoff with respect to L^2 or separation or some other distance.

Also, it is sometimes convenient to quantify the speed with which the distance drops from 1 to 0. For example, one says that the sequence has *cutoff window* $w_n \in o(t_n)$ if $\lim_{\alpha\to-\infty} \liminf_{n\to\infty} d_n(t_n + \alpha w_n) = 1$ and $\lim_{\alpha\to\infty} \limsup_{n\to\infty} d_n(t_n + \alpha w_n) = 1$.

We have actually established cutoff in a number of examples already. For instance, Theorem 4.3 shows that lazy SRW on the hypercube $(\mathbb{Z}/2\mathbb{Z})^n$ satisfies

$$\lim_{n\to\infty} d_n\left((1+\gamma)\cdot\frac{n}{2}\log(n)\right) \le \lim_{n\to\infty} \frac{1}{2}\left(e^{e^{-\gamma\log(n)}} - 1\right)^{\frac{1}{2}} = 0$$

and

$$\lim_{n\to\infty} d_n\left((1-\gamma)\cdot\frac{n}{2}\log(n)\right) \ge \lim_{n\to\infty} \left(1 - 12e^{-\gamma\log(n)}\right) = 1$$

for any $\gamma \in (0,1)$.

Similarly, the analysis in Subsubsection 4.2.3 shows that the random transposition walk on S_n satisfies

$$\lim_{n\to\infty} d_n\left((1+\gamma)\cdot\frac{n}{2}\log(n)\right) \le \lim_{n\to\infty} Ce^{-2\gamma\log(n)} = 0$$

and, taking $m(n) = \lfloor n^{\gamma/2} \rfloor$,

$$\lim_{n\to\infty} d_n\left((1-\gamma)\cdot\frac{n}{2}\log(n)\right) \ge \lim_{n\to\infty} 1 - m(n)e^{-\gamma\log(n)} - \frac{1}{m(n)!} = 1.$$

Of course, not all chains exhibit cutoff. For instance, Theorem 4.2 shows that for n odd and $k \ge n^2$, SRW on the n-cycle satisfies

$$\frac{1}{2}\exp\left(-\frac{(3\pi^2n^2+4\pi^4)k}{6n^4}\right) \le \left\|P^k - U\right\|_{TV} \le \frac{3}{4}\exp\left(-\frac{\pi^2 k}{2n^2}\right),$$

so a bit of arithmetic gives $t^{(n)}_{\text{mix}}(1/3) \ge \frac{6\log(3/2)n^4}{3\pi^2n^2+4\pi^4}$ and $t^{(n)}_{\text{mix}}(2/3) \le \frac{2\log(9/8)n^2}{\pi^2}$, hence

$$\frac{t^{(n)}_{\text{mix}}(1/3)}{t^{(n)}_{\text{mix}}(2/3)} \ge \frac{6\log(3/2)\,n^4}{3\pi^2n^2+4\pi^4}\cdot\frac{\pi^2}{2\log(9/8)\,n^2} \to \frac{\log(3/2)}{\log(9/8)} > 1.$$

(It is worth noting that a judicious choice of ε was required to make this calculation work...)

In this example, it makes sense to say that it takes order n^2 steps for the chain to mix, but the leading constant for necessity is strictly less than that for sufficiency.

The preceding arguments relied on precise upper and lower bounds that were made possible by Fourier analysis, but this is not always necessary for establishing cutoff for random walks on groups.

One of the most famous examples is the 'riffle shuffle,' which models the way that most of us actually mix cards. In this case, one can obtain a simple and exact formula for the probability that an n-card deck is in a particular order after k shuffles (based on the number of rising sequences in the associated permutation), leading to a demonstration of cutoff at time $\frac{3}{2}\log_2(n)$. In particular, it's not always coupon-collecting!

Similarly, you used a coupling argument in your homework to prove that the random-to-top shuffle on a deck of n cards satisfies $d_n\big(n\log(n)+cn\big) \le e^{-c}$.

This is the inverse of the top-to-random shuffle, which mixes the deck by successively removing the top card and inserting it in a position chosen uniformly at random, so the discussion from Example 2.2 shows that the same bound holds for top-to-random.

We can think of this as the random walk on S_n driven by $\mu(\sigma) = \frac{1}{n}1\{\sigma = (1\cdots k) \text{ for some } k \in [n]\}$, but the fact that the μ is not constant on conjugacy classes makes it difficult to analyze using representation theory. Nonetheless, we can prove cutoff by obtaining a matching lower bound as follows.

Let A_j be the event that the bottom j cards retain their original order for some fixed $j \ge 2$.

Clearly, for $n \ge j$, the uniform distribution on S_n satisfies $U(A_j) = \frac{1}{j!}$ since each permutation of cards $n-j+1,\dots,n$ is equally likely.

Now let T_k be the first time that k cards are below the one that was originally j^{th} from the bottom—the card labeled $n-j+1$ if the deck starts out ordered $1,2,\dots,n$—so that $T_{j-1}=0$, T_j is the first time a card is inserted below card $n-j+1$, and so forth. Setting $\tau_k = T_k - T_{k-1}$, we see that $\tau_j, \tau_{j+1}, \dots, \tau_{n-1}$ are independent with $\tau_k \sim \text{Geom}\big(\frac{k}{n}\big)$, and that card $n-j+1$ ascends to the top of the deck at time $T_{n-1} = \sum_{k=j}^{n-1}\tau_k$. As the bottom j cards retain their original order at least up to this time, we have

$$P^t(id, A_j) \ge \mathbb{P}\big(T_{n-1} \ge t\big).$$

Now if $X \sim \text{Geom}(p)$, then $\mathbb{E}[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2} \le \frac{1}{p^2}$, so

$$\mathbb{E}\big[T_{n-1}\big] = \sum_{k=j}^{n-1}\mathbb{E}[\tau_k] = \sum_{k=j}^{n-1}\frac{n}{k} = n\log(n) + \epsilon_n$$

and

$$\text{Var}\big(T_{n-1}\big) = \sum_{k=j}^{n-1}\text{Var}(\tau_k) \le \sum_{k=j}^{n-1}\frac{n^2}{k^2} \le Cn^2$$

where $\epsilon_n > -n\log(j)$ and $C < 1$.

These estimates follow by considering the upper Darboux sum

$$\sum_{k=j}^{n-1}\frac{1}{k} > \int_j^n \frac{dx}{x} = \log(n) - \log(j)$$

and the p-series

$$\sum_{k=j}^{n-1}\frac{1}{k^2} \le \sum_{k=2}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6} - 1 < 1.$$

Chebychev's inequality shows that for any $c > \log(j)$

$$\begin{aligned}\mathbb{P}\Big(T_{n-1} < n\log(n) - cn\Big) &\le \mathbb{P}\Big(T_{n-1} - \mathbb{E}\big[T_{n-1}\big] \le -cn - \epsilon_n\Big)\\ &\le \mathbb{P}\Big(\,\big|T_{n-1} - \mathbb{E}\big[T_{n-1}\big]\big| \ge cn + \epsilon_n\Big)\\ &\le \mathbb{P}\Big(\,\big|T_{n-1} - \mathbb{E}\big[T_{n-1}\big]\big| \ge \big(c - \log(j)\big)n\Big)\\ &\le \frac{\mathrm{Var}\big(T_{n-1}\big)}{(c-\log(j))^2 n^2} \le \frac{1}{(c-\log(j))^2}.\end{aligned}$$

Thus when $c > \log(j)$ is chosen so that $t = n\log(n) - cn \in \mathbb{N}$, we have

$$d\big(t\big) \ge P^t(id, A_j) - U(A_j) \ge 1 - \mathbb{P}\big(T_{n-1} < t\big) - \frac{1}{j!} \ge 1 - \frac{1}{(c-\log(j))^2} - \frac{1}{j!},$$

which can be made arbitrarily close to 1 by choosing c and j (and thus n) large enough.

5.2 Lower Bounds

As the illustrated in the foregoing examples, demonstrating cutoff generally amounts to finding tight upper and lower bounds on the variation distance that match up to second order terms in time, and such precise estimates often require detailed analyses specific to the chain in question. Still, there are some general techniques that prove useful time and again, such as the coupling and Fourier upper-bounds.

On the other hand, the definition of total variation distance gives an obvious class of lower bounds:

For any $B \subseteq \mathcal{S}$,

$$\big\|P_x^k - \pi\big\|_{TV} = \max_{A\subseteq\mathcal{S}} \big|P_x^k(A) - \pi(A)\big| \ge \big|P_x^k(B) - \pi(B)\big|,$$

and in fact, equality is achieved for the right choice of B. However, the identity of this optimal event is generally unclear, and it may well vary with k.

That said, experience has shown that a good choice for B is often related to some *distinguishing statistic* associated with a large nontrivial eigenvalue of the transition matrix.

The intuition is that if P is a transition matrix with stationary distribution π and $f : \mathcal{S} \to \mathbb{R}$ is an eigenfunction of P corresponding to a nontrivial eigenvalue λ, then the expectation of f with respect to P_x^k is $\delta_x P^k f = \lambda^k f(x)$, whereas its expectation under π is 0.

(The latter claim follows from the fact that π is a left eigenfunction of P with eigenvalue $1 \ne \lambda$, hence $\pi f = (\pi P)f = \pi(Pf) = \lambda\pi f$, which can happen only if $\pi f = 0$.)

Thus one suspects that for a suitable value of α (and normalization of f), the event $\big\{f \ge \alpha\big\}$ should have relatively high probability under P_x^k and relatively low probability under π. This can be made quantitative using Chebychev bounds.

Our next theorem automates this kind of moment estimate and often furnishes a correct rate. To facilitate its proof, we first observe that for any probabilities μ, ν on $\mathcal{S}$ and function $\varphi : \mathcal{S} \to \Lambda$, we have

$$\|\mu - \nu\|_{TV} = \max_{A\subseteq\mathcal{S}} |\mu(A) - \nu(A)| \ge \max_{E\subseteq\Lambda} \big|\mu\big(\varphi^{-1}(E)\big) - \nu\big(\varphi^{-1}(E)\big)\big|,$$

where the rightmost term is the variation distance between the pushforward measures $\mu\circ\varphi^{-1}$ and $\nu\circ\varphi^{-1}$ on Λ.

Theorem 5.1. *Let μ and ν be probabilities on a finite space $\mathcal{S}$ and suppose that $f : \mathcal{S} \to \mathbb{R}$ satisfies*

$$|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]| \geq r\sigma$$

with $\sigma^2 = \frac{1}{2}\left[\mathrm{Var}_\mu(f) + \mathrm{Var}_\nu(f)\right]$. Then

$$\|\mu - \nu\|_{TV} \geq 1 - \frac{4}{4 + r^2}.$$

Proof. Define the measures $\alpha = \mu \circ f^{-1}$ and $\beta = \nu \circ f^{-1}$ on $f(\mathcal{S}) \subseteq \mathbb{R}$, and denote their means by m_α and m_β, so that

$$m_\alpha = \sum_{y \in f(\mathcal{S})} y\mu\big(f^{-1}(y)\big) = \sum_{x \in \mathcal{S}} f(x)\mu(x) = \mathbb{E}_\mu[f]$$

and analogously for m_β. A nearly identical second moment computation shows that α and β have variances $\mathrm{Var}_\mu(f)$ and $\mathrm{Var}_\nu(f)$.

Replacing f with $-f$ if need be, we can assume that $M := \frac{1}{2}(m_\alpha - m_\beta) > 0$, and replacing f with $f - m_\beta - M$, we can take $m_\alpha = M$ and $m_\beta = -M$.

(The assumptions only depend on f through $|m_\alpha - m_\beta|$, $\mathrm{Var}(\alpha)$, and $\mathrm{Var}(\beta)$, all of which are invariant under these operations.)

Now define

$$\eta(y) = \frac{1}{2}\left(\alpha(y) + \beta(y)\right), \quad r(y) = \frac{\alpha(y)}{\eta(y)}, \quad s(y) = \frac{\beta(y)}{\eta(y)}$$

with the understanding that $r(y) = s(y) = 0$ if $\eta(y) = 0$.

Then Cauchy-Schwarz gives

$$4M^2 = \left[\sum_y y\alpha(y) - \sum_y y\beta(y)\right]^2 = \left[\sum_y y\big(r(y) - s(y)\big)\eta(y)\right]^2 \leq \sum_y y^2\eta(y) \cdot \sum_y \big(r(y) - s(y)\big)^2\eta(y).$$

Since

$$\sum_x y^2\eta(y) = \frac{1}{2}\left[\sum_y y^2\alpha(y) + \sum_y y^2\beta(y)\right] = \frac{\mathrm{Var}(\alpha) + m_\alpha^2 + \mathrm{Var}(\beta) + m_\alpha^2}{2} = \sigma^2 + M^2$$

and

$$|r(y) - s(y)| = 2\frac{|\alpha(y) - \beta(y)|}{\alpha(y) + \beta(y)} \leq 2,$$

hence

$$\sum_y \big(r(y) - s(y)\big)^2\eta(y) \leq 2\sum_y |r(y) - s(y)|\,\eta(y) = 2\sum_y |\alpha(y) - \beta(y)| = 4\,\|\alpha - \beta\|_{TV},$$

we see that $M^2 \leq \left(\sigma^2 + M^2\right)\|\alpha - \beta\|_{TV}$.

The observation preceding the theorem and our assumption that $2M \geq r\sigma$ thus yields

$$\begin{aligned}\|\mu - \nu\|_{TV} \geq \|\alpha - \beta\|_{TV} &\geq \frac{M^2}{\sigma^2 + M^2} \\ &= 1 - \frac{\sigma^2}{\sigma^2 + M^2} \geq 1 - \frac{4}{4 + r^2}.\end{aligned}$$

□

As noted previously, if f is an eigenfunction of P with eigenvalue λ, then $\mathbb{E}_{P_x^k}[f] = \lambda f(x)$ and $\mathbb{E}_\pi[f] = 0$. If λ has high multiplicity, we can choose f to be an average over a basis of eigenfunctions to reduce its variance and thus improve our bound.

For random walks on groups, the eigenfunctions often have nice algebraic structure—for instance, they're irreducible characters if the group is abelian—which facilitates estimation of the variances. For reversible chains, there are also variance bounds arising from the variational characterization of eigenvalues for Hermitian operators. (We will touch on this briefly in the following subsection. Had we more time, the geometric and functional arguments that ensue would be next on our list of topics!) We will content ourselves here with the following general result that goes by the name of *Wilson's method*.

Theorem 5.2. *Let X_k be an irreducible and aperiodic Markov chain with stationary distribution π, and suppose that Φ is an eigenfunction for the transition matrix with eigenvalue $\frac{1}{2} < \lambda < 1$.*

If $R > 0$ satisfies $\mathbb{E}_x\left[\big(\Phi(X_1) - \Phi(x)\big)^2\right] \le R$ for all $x \in \mathcal{S}$, then

$$\mathrm{Var}_{P_x^k}(\Phi), \mathrm{Var}_\pi(\Phi) \le \frac{R}{2(1-\lambda)}.$$

Proof. Since Φ is an eigenfunction corresponding to λ, we have $\mathbb{E}_x\big[\Phi(X_k)\big] = \delta_x P^k \Phi = \lambda^k \Phi(x)$.

Similarly, letting $\Delta_k = \Phi(X_{k+1}) - \Phi(X_k)$ denote the forward difference process, we have

$$\mathbb{E}_x\big[\Delta_k \,|X_k = z\,\big] = (\lambda - 1)\Phi(z),$$

and, by assumption,

$$\mathbb{E}_x\big[\Delta_k^2 \,|X_k = z\,\big] = \mathbb{E}_z\left[\big(\Phi(X_1) - \Phi(z)\big)^2\right] \le R.$$

It follows that

$$\begin{aligned}
\mathbb{E}_x\big[\Phi(X_{k+1})^2 \,|X_k = z\,\big] &= \mathbb{E}_x\big[\big(\Phi(X_k) + \Delta_k\big)^2 \,|X_k = z\,\big] \\
&= \Phi(z)^2 + 2\Phi(z)\mathbb{E}_x\big[\Delta_k \,|X_k = z\,\big] + \mathbb{E}_x\big[\Delta_k^2 \,|X_k = z\,\big] \\
&\le (2\lambda - 1)\Phi(z)^2 + R,
\end{aligned}$$

and thus

$$\mathbb{E}_x\big[\Phi(X_{k+1})^2\big] = \sum_z \mathbb{E}_x\big[\Phi(X_{k+1})^2 \,|X_k = z\,\big] P^k(x,z) \le (2\lambda - 1)\mathbb{E}_x\big[\Phi(X_k)^2\big] + R.$$

Since $\mathbb{E}_x\big[\Phi(X_0)^2\big] = \Phi(x)^2$ and $2\lambda - 1 > 0$, we see by induction that

$$\begin{aligned}
\mathbb{E}_x\big[\Phi(X_k)^2\big] &\le (2\lambda-1)^k \Phi(x)^2 + R\sum_{j=0}^{k-1}(2\lambda-1)^j \\
&= (2\lambda-1)^k\Phi(x)^2 + \frac{1-(2\lambda-1)^k}{2(1-\lambda)}R < (2\lambda-1)^k\Phi(x)^2 + \frac{R}{2(1-\lambda)},
\end{aligned}$$

thus

$$\mathrm{Var}_{P_x^k}\big(\Phi\big) = \mathbb{E}_x\big[\Phi(X_k)^2\big] - \mathbb{E}_x\big[\Phi(X_k)\big]^2 < \big[(2\lambda-1)^k - \lambda^{2k}\big]\Phi(x)^2 + \frac{R}{2(1-\lambda)} < \frac{R}{2(1-\lambda)}$$

where the final inequality used $\lambda > \frac{1}{2}$ and $0 < (\lambda-1)^2 = \lambda^2 - 2\lambda + 1$, thus $0 < 2\lambda - 1 < \lambda^2$.

To complete the proof, note that ergodicity implies

$$\mathrm{Var}_\pi(\Phi) = \lim_{k\to\infty} \mathrm{Var}_x\big(\Phi(X_k)\big) \le \frac{R}{2(1-\lambda)}.$$ □

5.3 Relaxation Time and the Spectral Gap

We have seen that the mixing time is often controlled by the top nontrivial eigenvalue of the transition matrix.

Indeed, the results from Example 4.2 and Theorem 4.5 show that if P is is the transition matrix for the random walk (G, μ) with μ constant on conjugacy classes (and not supported on a proper subgroup), then $\left\|P^k - \pi\right\|_{TV}^2 \leq \sum_{\lambda \neq 1} |\lambda|^{2k}$.

We also derived a similar result for reversible Markov chains, and that argument extends more generally to any chain that commutes with its time reversal, $\widehat{P}(x,y) = \frac{\pi(y)}{\pi(x)} P(y,x)$. This is because $P\widehat{P} = \widehat{P}P$ implies that the matrix $A = DPD^{-1}$ with $D = \operatorname{diag}\left(\sqrt{\pi(x_1)}, \ldots, \sqrt{\pi(x_n)}\right)$ satisfies $AA^T = A^T A$, thus one can apply the spectral theorem as before.

In fact, there are still other diagonalizable Markov chains (like random walks on hyperplane arrangements) for which the mixing time can be bounded by sums of powers of eigenvalues...

In these situations, one generally needs to take k large enough to kill the term corresponding to the subdominant eigenvalue in order to have any hope of nearing equilibrium.

Of course, if other large eigenvalues have sufficiently high multiplicity, they can dominate the upper bound. For instance, if $\lambda_1 = 1 - \frac{1}{n}$ is a simple eigenvalue and $\lambda_2 = \cdots = \lambda_m = 1 - \frac{2}{n}$ with $m \approx n/2$, then k must be at least large enough to make $(1 - \frac{1}{n})^k + m(1 - \frac{2}{n})^k \approx e^{-\frac{k}{n}} + \frac{n}{2} e^{-\frac{2k}{n}}$ small. Whereas $k = cn$ with $c > 1$ suffices to make the first term small, one needs k on the order of $n \log(n)$ to drive the second term to zero.

(This is not just hypothetical as there are explicit procedures for generating symmetric stochastic matrices with any spectrum of the form $1 = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{n-1} \geq 0$. Reversible Markov chains with uniform stationary distributions all have symmetric transition matrices, and by replacing P with $\frac{1}{2}(P + I)$, one can ensure that their eigenvalues are nonnegative.)

Moreover, as the random transposition walk illustrates, even when the true rate comes from the top eigenvalue, care must still be taken with the others to establish sharp mixing bounds.

These caveats aside, it makes sense to study the top eigenvalue in greater detail. While many of the arguments can be generalized, we will mostly restrict our attention to the case of an ergodic and reversible Markov chain P with eigenvalues $1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{n-1} > -1$ and stationary distribution π. Write $f_0, \ldots, f_{n-1}$ for the corresponding eigenfunctions, chosen to be orthonormal with respect to the inner product $\langle f, g \rangle_\pi = \sum_x f(x) g(x) \pi(x)$. The discussion in Subsection 1.5 guarantees this is possible, and the normalization ensures that $f_0 \equiv 1$.

Write $\lambda_* = \max\{|\lambda_1|, |\lambda_{n-1}|\}$ for the largest magnitude of the nontrivial eigenvalues and define the *absolute spectral gap* by $\gamma_* = 1 - \lambda_*$. (By adding laziness, one can ensure all eigenvalues are positive so that this coincides with the usual spectral gap $\gamma = 1 - \lambda_1$. See the remark below for another characterization of γ.)

For a general irreducible and aperiodic Markov chain having (possibly complex) eigenvalues $1 = \beta_0 > |\beta_1| \geq \cdots \geq |\beta_{n-1}|$, we will also call the quantity $\gamma_* = 1 - |\beta_1|$ the absolute spectral gap.

Define the *relaxation time* by $t_{\text{rel}} = 1/\gamma_*$.

Remark 5.2 (Skipped). Given any functions $f, g : \mathcal{S} \to \mathbb{R}$, we define the *Dirichlet form* $\mathcal{E}(f,g) = \langle (I-P)f, g\rangle_\pi$ and the *energy* $\mathcal{E}(f) := \frac{1}{2}\sum_{x,y\in\mathcal{S}} \big[f(x)-f(y)\big]^2 P(x,y)\pi(x)$.

Observe that reversibility gives

$$\begin{aligned}\mathcal{E}(f,g) &= \sum_x [f(x)-(Pf)(x)]\, g(x)\pi(x) = \sum_x \pi(x)\sum_y [f(x)P(x,y)-f(y)P(x,y)]\, g(x)\\ &= \sum_{x,y}\pi(x)P(x,y)\big(f(x)-f(y)\big)g(x) = \sum_{x,y}\pi(y)P(y,x)\big(f(x)-f(y)\big)g(x).\end{aligned}$$

Interchanging the roles of x and y shows that an equivalent expression is

$$\mathcal{E}(f,g) = \sum_{x,y}\pi(x)P(x,y)\big(f(y)-f(x)\big)g(y) = \sum_{x,y}\pi(y)P(y,x)\big(f(y)-f(x)\big)g(y).$$

Averaging the two then yields

$$\mathcal{E}(f,g) = \frac{1}{2}\sum_{x,y}\pi(y)P(y,x)\big(f(x)-f(y)\big)\big(g(x)-g(y)\big) = \frac{1}{2}\sum_{x,y}\big(f(x)-f(y)\big)\big(g(x)-g(y)\big)P(x,y)\pi(x).$$

In particular, $\mathcal{E}(f) = \mathcal{E}(f,f)$.

Now any function $h \in \mathbb{R}^{\mathcal{S}}$ can be expressed as $h = \sum_{j=0}^{n-1}\alpha_j f_j$ with $\alpha_j = \langle h, f_j\rangle_\pi$, so its expectation under π is given by

$$\pi(h) = \sum_x h(x)\pi(x) = \sum_{j=0}^{n-1}\alpha_j \sum_x f_j(x)\pi(x) = \alpha_0$$

since $\sum_x f_j(x)\pi(x) = \langle f_j, f_0\rangle_\pi = \delta_{j0}$.

Similarly, the second moment under π is

$$\pi(h^2) = \langle h,h\rangle_\pi = \left\langle \sum_{j=0}^{n-1}\alpha_j f_j, \sum_{j=0}^{n-1}\alpha_j f_j\right\rangle_\pi = \sum_{i,j}\alpha_i\alpha_j\langle f_i, f_j\rangle = \sum_{j=0}^{n-1}\alpha_j^2.$$

As $Ph = \sum_{j=0}^{n-1}\alpha_j P f_j = \sum_{j=0}^{n-1}\alpha_j\lambda_j f_j$, a completely parallel computation gives $\langle Ph, h\rangle_\pi = \sum_{j=0}^{n-1}\lambda_j\alpha_j^2$.

Thus if h is a function normalized to have $\pi(h)=0$ and $\pi(h^2)=1$, then $h = \sum_{j=1}^{n-1}\alpha_j f_j$ with $\sum_{j=1}^{n-1}\alpha_j^2 = 1$.

Its energy is therefore

$$\mathcal{E}(h) = \langle (I-P)h, h\rangle_\pi = \langle h,h\rangle_\pi - \langle Ph,h\rangle_\pi = 1 - \sum_{j=0}^{n-1}\lambda_j\alpha_j^2 \ge 1-\lambda_1.$$

Since equality is attained for the function $h = f_1$, we have the 'Courant-Fischer characterization'

$$\gamma = \min_{\substack{h\in\mathbb{R}^{\mathcal{S}}:\\ \pi(h)=0,\,\pi(h^2)=1}} \mathcal{E}(h) = \min_{\substack{h\in\mathbb{R}^{\mathcal{S}}:\\ \pi(h)=0,\,h\not\equiv 0}} \frac{\mathcal{E}(h)}{\mathrm{Var}_\pi(h)}.$$

Theorem 5.3. *Suppose P is irreducible and aperiodic with stationary distribution π. Then for any $\varepsilon \in (0,1)$, we have*

$$t_{\mathrm{mix}}(\varepsilon) \ge (t_{\mathrm{rel}}-1)\log\left(\frac{1}{2\varepsilon}\right).$$

Proof. Irreducibility and aperiodicity ensure that P has eigenvalues $1 = \lambda_0 > |\lambda_1| \ge \cdots \ge |\lambda_{n-1}|$. Assume that $\lambda = |\lambda_1| \ne 0$ as the inequality is trivial in this case.

Let φ be a corresponding eigenfunction normalized so that $\|\varphi\|_\infty = 1$ and write r for the state with $|\varphi(r)| = 1$.

Since π is a left eigenfunction with eigenvalue 1, $\langle \varphi, \mathbf{1} \rangle_\pi = \sum_y \varphi(y)\pi(y) = \langle \varphi, \pi^T \rangle = 0$.

It follows that

$$\begin{aligned}\lambda^t &= \left|(P^t\varphi)(r)\right| = \max_{x\in\mathcal{S}} \left|(P^t\varphi)(x)\right| = \max_{x\in\mathcal{S}} \left|(P^t\varphi)(x) - \langle \varphi, \mathbf{1}\rangle_\pi\right| \\ &= \max_{x\in\mathcal{S}} \left|\sum_{y\in\mathcal{S}} \left[P^t(x,y)\varphi(y) - \pi(y)\varphi(y)\right]\right| \le \max_{x\in\mathcal{S}} \sum_{y\in\mathcal{S}} \left|\left[P^t(x,y) - \pi(y)\right]\varphi(y)\right| \\ &\le \|\varphi\|_\infty \max_{x\in\mathcal{S}} \sum_{y\in\mathcal{S}} \left|P^t(x,y) - \pi(y)\right| = 2\max_{x\in\mathcal{S}} \left\|P_x^t - \pi\right\|_{TV} = 2d(t).\end{aligned}$$

Setting $t = t_{\text{mix}}(\varepsilon)$ and using $\log(x) \le x - 1$ for $x \ge 1$, we have that $\lambda^{t_{\text{mix}}(\varepsilon)} \le 2d\big(t_{\text{mix}}(\varepsilon)\big) \le 2\varepsilon$, so

$$t_{\text{mix}}(\varepsilon)\left(\frac{1}{\lambda} - 1\right) \ge t_{\text{mix}}(\varepsilon)\log\left(\frac{1}{\lambda}\right) \ge \log\left(\frac{1}{2\varepsilon}\right),$$

and thus

$$t_{\text{mix}}(\varepsilon) \ge \frac{\lambda}{1-\lambda}\log\left(\frac{1}{2\varepsilon}\right) = (t_{\text{rel}} - 1)\log\left(\frac{1}{2\varepsilon}\right). \qquad \square$$

Theorem 5.4. *If P is also reversible and $\pi_{\min} = \min_x \pi(x)$, then*

$$t_{\text{mix}}(\varepsilon) \le t_{\text{rel}}\log\left(\frac{1}{2\varepsilon\sqrt{\pi_{\min}}}\right).$$

Proof. Let $\{f_k\}_{k=0}^{n-1}$ be an orthonormal eigenbasis in $L^2(\pi)$. Remarks 2.1 and 1.7 show that

$$4\left\|P^t(x,\cdot) - \pi\right\|_{TV}^2 \le \left\|P^t(x,\cdot) - \pi\right\|_{L^2(\pi)}^2 = \sum_{k=1}^{n-1} \lambda_k^{2t} f_k(x)^2 \le \lambda_*^{2t} \sum_{k=1}^{n-1} f_k(x)^2$$

Since

$$\sum_{k=0}^{n-1} f_k(x)^2\pi(x) = \sum_{k=0}^{n-1}\left(\sum_{y\in\mathcal{S}} \delta_x f_k(y)\pi(y)\right) f_k(x) = \sum_{k=0}^{n-1} \langle \delta_x, f_k\rangle_\pi f_k(x) = \delta_x(x) = 1,$$

we obtain

$$\begin{aligned}4\left\|P^t(x,\cdot) - \pi\right\|_{TV}^2 &\le \lambda_*^{2t}\sum_{k=1}^{n-1} f_k(x)^2 = \lambda_*^{2t}\frac{1-\pi(x)}{\pi(x)} \\ &\le \frac{\lambda_*^{2t}}{\pi_{\min}} = \frac{(1-\gamma_*)^{2t}}{\pi_{\min}} \le \frac{1}{\pi_{\min}} e^{-2\gamma_* t},\end{aligned}$$

and thus

$$d(t) \le \frac{1}{2\sqrt{\pi_{\min}}} e^{-\gamma_* t}.$$

It follows that $d(t) \le \varepsilon$ once $t \ge t_{\text{rel}}\log\left(\frac{1}{2\varepsilon\sqrt{\pi_{\min}}}\right)$. $\square$

Taken together, Theorems 5.3 and 5.4 show that the mixing time of a reversible and ergodic Markov chain is essentially the relaxation time up to a factor of about $-\log(\pi_{\min})$. Of course, $\pi_{\min} \le 1/|\mathcal{S}|$, so this is not insignificant.

Still, Theorem 5.3, which applies even in the nonreversible case, provides an intriguing necessary condition for cutoff.

Corollary 5.1. *A sequence of ergodic Markov chains with mixing times* $\{t_{\text{mix}}^{(n)}\}$ *and relaxation times* $\{t_{\text{rel}}^{(n)}\}$ *presents cutoff only if*

$$\lim_{n\to\infty} \frac{t_{\text{mix}}^{(n)}}{t_{\text{rel}}^{(n)} - 1} = \infty.$$

Proof. For any $\varepsilon \leq 3/4$, we have

$$\frac{t_{\text{mix}}(\varepsilon)}{t_{\text{mix}}(1-\varepsilon)} \geq \frac{t_{\text{mix}}(\varepsilon)}{t_{\text{mix}}} \geq \frac{t_{\text{rel}} - 1}{t_{\text{mix}}} \log\left(\frac{1}{2\varepsilon}\right).$$

Thus if there is an infinite sequence of integers satisfying $\frac{t_{\text{mix}}^{(n)}}{t_{\text{rel}}^{(n)}-1} \leq M_0$ for some $M_0 > 0$, then for every $0 < \varepsilon \leq \frac{1}{2}e^{-2M_0}$, we have

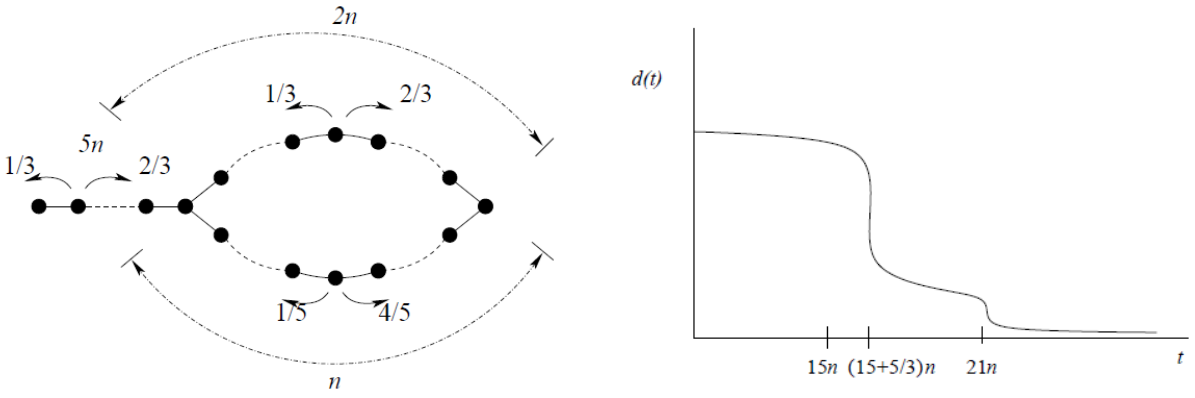
$$\liminf_{n\to\infty} \frac{t_{\text{mix}}(\varepsilon)}{t_{\text{mix}}(1-\varepsilon)} \geq \frac{1}{M_0} \log\left(\frac{1}{2\varepsilon}\right) \geq 2. \qquad \square$$

Remark 5.3. $t_{\text{rel}} - 1 = \frac{1-\gamma_*}{\gamma_*} = \lambda_* t_{\text{rel}}$, so as long as $\lambda_*^{(n)}$ is bounded away from 0, the necessary condition for cutoff in Corollary 5.1 can be phrased as $\gamma_*^{(n)} t_{\text{mix}}^{(n)} \to \infty$ or $t_{\text{rel}}^{(n)} = o\big(t_{\text{mix}}^{(n)}\big)$.

Example 5.1. For simple random walk on the n-cycle, we have $d_n(t) \leq \frac{3}{4}\exp\left(-\frac{\pi^2 t}{2n^2}\right)$, so $t_{\text{mix}}^{(n)} \leq \frac{2\log(3)}{\pi^2} n^2$. Since $\cos\left(\frac{2\pi}{n}\right) \geq 1 - \frac{2\pi^2}{n^2}$ is an eigenvalue, we have $t_{\text{rel}}^{(n)} \geq \frac{n^2}{2\pi^2}$, hence $\frac{t_{\text{mix}}^{(n)}}{t_{\text{rel}}^{(n)}} \leq 4\log(3)$ and we conclude that cutoff does not occur.

The condition given in Corollary 5.1 concerns *precutoff*, the requirement that $\frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)}$ is bounded above. However, this is not enough to ensure convergence to one, and there are families of chains for which $\sup_{0<\varepsilon<\frac{1}{2}} \limsup_{n\to\infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} < \infty$ but cutoff does not occur.

The graphic below illustrates one such example due to David Aldous.



There has been some nice work establishing that something like the condition in Corollary 5.1 is sufficient for cutoff with respect to L^p distances when $p > 1$, and some hitting time conditions implying cutoff in lazy reversible chains have been found as well, but the search for an elegant abstract characterization is still very much ongoing.

Homework 1

(1) Consider the following process: Two animals are mated and among their direct descendants two individuals of opposite sex are selected at random. These individuals are mated and the process continues. Suppose that each individual can be of one of three genotypes, AA, Aa, aa, and suppose that the type of offspring is determined by selecting a letter from each parent. With these rules, the pair of genotypes in the n^{th} generation is a Markov chain with six states,

$$(AA, AA), (AA, Aa), (AA, aa), (Aa, Aa), (Aa, aa), (aa, aa).$$

(a) Compute its transition matrix.

(b) Compute $\mathbb{P}\big(X_6 = (Aa, Aa) \big| X_0 = (AA, Aa)\big)$. (Feel free to get help from a computer.)

(2) Show that if the transition matrix P is symmetric (so $P(x, y) = P(y, x)$ for all $x, y \in \mathcal{S}$), then the uniform distribution on $\mathcal{S}$ is stationary for P.

(3) Prove that simple random walk on a finite, connected, and undirected graph is aperiodic if and only if the graph is not bipartite.
($G = (V, E)$ is bipartite if $V = V_1 \bigsqcup V_2$ with $E \subseteq \big\{\{x, y\} : x \in V_1, y \in V_2\big\}$. Equivalently, G is bipartite if it contains no cycles of odd length.)

(4) Suppose P is the transition matrix of an irreducible Markov chain X_n having stationary distribution π. Define the *time reversal* of X_n to be the chain $\widehat{X}_n$ with transition matrix

$$\widehat{P}(x, y) = P(y, x)\frac{\pi(y)}{\pi(x)}.$$

Show that π is stationary for $\widehat{X}_n$, and for any $x_0, ..., x_t \in \mathcal{S}$, we have

$$\mathbb{P}_\pi\big(X_0 = x_0, ..., X_t = x_t\big) = \mathbb{P}_\pi\big(\widehat{X}_0 = x_t, ..., \widehat{X}_t = x_0\big).$$

(5) Let P be a Markov chain with state space $\mathcal{S}$. We say that $A \subseteq \mathcal{S}$ is *closed* if $x \in A$ and $P(x, y) > 0$ implies $y \in A$. A is *irreducible* if $x, y \in A$ implies $P^n(x, y) > 0$ for some n.
Give an example of a Markov chain and sets $B, C \subseteq \mathcal{S}$ such that B is closed but not irreducible and C is irreducible but not closed.

(6) Give an example of a Markov chain X_n on a countable state space S and a function g on S such that $g(X_n)$ is not a Markov chain. Can you give any conditions on X_n and g which ensure that $g(X_n)$ is a Markov chain?

(7) Let P be the transition matrix for simple random walk on the n-cycle ($X_k = X_{k-1} + \xi_k \pmod n$ where $\xi_1, \xi_2, ...$ are i.i.d. with $P(\xi_1 = 1) = P(\xi_1 = -1) = \frac{1}{2}$) for n odd.

Find the smallest value of r so that $P^s(x, y) > 0$ for all $s \geq r$ and all $x, y \in \mathbb{Z}/n\mathbb{Z}$.

Homework 2

(1) Compute the expected number of moves it takes a knight to return to its initial position if it starts on the corner of a chessboard, assuming that there are no other pieces on the board and that each time it chooses a move at random from its legal moves.

The following table gives the number of legal knight moves from each square.

2	3	4	4	4	4	3	2
3	4	6	6	6	6	4	3
4	6	8	8	8	8	6	4
4	6	8	8	8	8	6	4
4	6	8	8	8	8	6	4
4	6	8	8	8	8	6	4
3	4	6	6	6	6	4	3
2	3	4	4	4	4	3	2

(2) Suppose that q is a transition function on $\mathcal{S}$ with $q(x,y) = q(y,x)$ for all $x, y \in \mathcal{S}$. Let π be a probability on $\mathcal{S}$ satisfying $\pi(x) > 0$ for all $x \in \mathcal{S}$. Define a new Markov chain as follows:

When $X_n = x$, sample y from $q(x,\cdot)$. If $\pi(y) \geq \pi(x)$, set $X_{n+1} = y$. Otherwise, flip a coin with heads probability $\pi(y)/\pi(x)$, set $X_{n+1} = y$ if the coin lands heads, and set $X_{n+1} = x$ if the coin lands tails.

The transition function for $\{X_k\}_{k=0}^{\infty}$ is thus

$$p(x,y) = \begin{cases} r(x,y)q(x,y), & y \neq x \\ 1 - \sum_{z \neq x} r(x,z)q(x,z), & y = x \end{cases}$$

where $r(x,y) = \frac{\pi(y)}{\pi(x)} \wedge 1$. Show that π is stationary for p.

(The same basic trick works if the *proposal chain* q is not symmetric. In this case, one must accept the move from x to y with probability

$$\widetilde{r}(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \wedge 1.$$

This is called the *Metropolis-Hastings algorithm* and is useful for generating approximate samples from a distribution π that is only known up to a normalizing constant.)

(3) Suppose that p is a Markov chain with countable state space $\mathcal{S}$. We say $\mu : \mathcal{S} \to [0,\infty)$ is *stationary* for p if $\mu \not\equiv 0$ and $\mu(y) = \sum_x \mu(x)p(x,y)$ for all y.

(a) Show that $\mu p^n = \mu$ for all $n \geq 1$.

(b) Prove that if p is irreducible and μ is stationary, then $\mu(y) > 0$ for all $y \in \mathcal{S}$.

(4) Show that the expectation of an $\mathbb{N}_0$-valued random variable N can be computed as

$$\mathbb{E}[N] = \sum_{k=1}^{\infty} \mathbb{P}(N \geq k).$$

(5) Suppose that $\{X_n\}_{n=0}^{\infty}$ is a Markov chain with countable state space $\mathcal{S}$ and transition function p. Define $T_y^0 = 0$ and $T_y^k = \inf\{t > T_y^{k-1} : X_t = y\}$ for $k \in \mathbb{N}$. Let $\rho_{xy} = \mathbb{P}_x(T_y^1 < \infty)$ denote the probability that the chain started at x visits y in finitely many steps. We say y is *recurrent* if $\rho_{yy} = 1$ and *transient* otherwise.

You may take it as given that $\mathbb{P}_x(T_y^k < \infty) = \rho_{xy}\rho_{yy}^{k-1}$. (This is a consequence of the strong Markov property.)

(a) Suppose that y is transient and let $N(y) = \sum_{k=1}^{\infty} 1\{X_k = y\}$ denote the number of visits to y at positive times. Show that $\mathbb{E}_x[N(y)] = \frac{\rho_{xy}}{1-\rho_{yy}}$.
(Hint: $N(y) \geq k$ if and only if $T_y^k < \infty$.)

(b) Show that y is recurrent if and only if $\mathbb{E}_y[N(y)] = \infty$.

(c) Show that if x is recurrent and $\rho_{xy} > 0$, then $\rho_{yx} = 1$.

(d) Show that if x is recurrent and $\rho_{xy} > 0$, then y is recurrent.

(6) Suppose that p is the transition function for a Markov chain with state space $\mathcal{S}$. Prove that if π is a stationary distribution for p and $\pi(y) > 0$, then y is recurrent.
(Hint: Compute $\sum_x \pi(x)\mathbb{E}_x[N(y)]$.)

(7) Compute the stationary distribution for the Markov chain on $\mathbb{N}$ defined by $p(x, (x-1) \vee 1) = q$ and $p(x, x+1) = 1 - q$, $q > 1/2$.

Homework 3

(1) Given a probability μ on $\mathcal{S}$ and a function $f : \mathcal{S} \to \mathbb{R}$, define $\mu(f) := \sum_{x \in \mathcal{S}} \mu(x) f(x)$, the expectation of f with respect to μ. Also, write $\|f\|_\infty := \sup_{x \in \mathcal{S}} |f(x)|$.
Prove that $\|\mu - \nu\|_{TV} = \frac{1}{2} \max_{\|f\|_\infty \leq 1} \big|\mu(f) - \nu(f)\big|$.

(2) Let μ and ν be two probabilities on a countable set $\mathcal{S}$. A fair coin is flipped and if it lands heads then a point is sampled from μ, while if it lands tails a point is sampled from ν. You are told which point was sampled and must decide whether the coin landed heads or tails. Show that your chance of being correct is $\frac{1}{2} + \frac{1}{2} \|\mu - \nu\|_{TV}$.

(3) Let U be the uniform distribution on S_n.

(a) Compute $\|P - U\|_{TV}$ where P is uniform over all permutations satisfying $\pi(1) = 1$.
(In terms of cards, P corresponds to "card 1 is on top an the rest are random.")

(b) Compute $\|Q - U\|_{TV}$ where Q is uniform over $(1, 2, \ldots, n), (2, \ldots, n, 1), \ldots, (n, 1, \ldots, n-1)$.
(Under Q, any particular card is equally likely to be in each position.)

(4) Let μ and ν be probabilities on $\mathbb{Z}$. We define their product as $(\mu \times \nu)(x, y) = \mu(x)\nu(y)$ and their convolution as $(\mu * \nu)(x) = \sum_y \mu(x - y)\nu(y)$.

(a) Show that $\|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\|_{TV} \leq \|\mu_1 - \nu_1\|_{TV} + \|\mu_2 - \nu_2\|_{TV}$.

(b) Show that $\|\mu_1 * \mu_2 - \nu_1 * \nu_2\|_{TV} \leq \|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\|_{TV}$.

(5) Compute $d(t)$ and $t_{\text{mix}}(\varepsilon)$ for the chain with transition matrix $P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$.

(6) Establish the identity $(1 + x)^n \leq e^{nx}$ for $x \geq -1$.

(7) The random-to-top shuffle proceeds at each time step by choosing a card uniformly at random and placing it on the top of the deck. We can model this as the left-invariant random walk on S_n driven by the probability $\mu(\sigma) = \frac{1}{n} 1\{\sigma = (k \cdots 1) \text{ for some } k \in [n]\}$, so the stationary distribution is given by $U(\sigma) = \frac{1}{n!}$ and the distribution after t-steps is $P^t(id, \cdot) = \mu^{*t}(\cdot)$, the t-fold convolution of μ with itself. (The distance to stationarity is the same for all initial states, so we may as well start with the ordering in which card k is in position k for $k = 1, \ldots, n$.)

Use a coupling argument to show that $d(t) = \|P^t(id, \cdot) - U\|_{TV}$ satisfies

$$d\big(n \log(n) + cn\big) \leq e^{-c}.$$

Homework 4

(1) Compute all irreducible representations of $G = \mathbb{Z}/n\mathbb{Z}$. Write the Fourier transform and inversion formula for $f : G \to \mathbb{C}$.

(2) Recall that the trace of a matrix $M \in \mathbb{C}^{d\times d}$ is defined as $\mathrm{Tr}(M) = \sum_{k=1}^{d} M_{kk}$.

(a) Suppose that $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times m}$. Prove that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

(b) Show that the trace is basis independent. In other words, if $T : V \to V$ is a linear transformation, B is the matrix for T with respect to a basis $\mathcal{B}$, and C is the matrix for T with respect to a basis $\mathcal{C}$, then $\mathrm{Tr}(B) = \mathrm{Tr}(C)$.

(c) Suppose that $A \in \mathbb{C}^{n\times n}$ has eigenvalues (counting multiplicity) $\lambda_1, \ldots, \lambda_n$. Show that $\mathrm{Tr}(A) = \sum_{k=1}^{n} \lambda_k$.

(3) Given representations $\rho : G \to GL(V)$ and $\eta : G \to GL(W)$, show that the characters of their direct sum and tensor product satisfy

$$\chi_{\rho\oplus\eta}(s) = \chi_\rho(s) + \chi_\eta(s)$$

and

$$\chi_{\rho\otimes\eta}(s) = \chi_\rho(s)\chi_\eta(s).$$

(4) Consider the following lazy random walk on S_3 generated by transpositions: Let $\varepsilon \in (0, 1)$ and define $P(id) = \varepsilon$, $P(12) = P(13) = P(23) = (1 - \varepsilon)/3$.

(a) Compute $\widehat{P}(\rho)$ for each irreducible representation of S_3 (trivial, sign, and standard).

(b) Use Fourier inversion to find $P^k(id, \sigma)$ for $\sigma \in S_n$.

(5) Show that the convolution of class functions is a class function.

(6) The uniform distribution on a finite group G is given by $U(g) = \frac{1}{|G|}$ for all $g \in G$.

(a) Show that $\widehat{U}(\rho)$ is 1 if ρ is the trivial representation and the zero matrix if ρ is any other irreducible representation.

(b) Let μ be a probability on G and define $\check{\mu}(s) = \mu(s^{-1})$. Show that $\mu * \check{\mu} = U$ if and only if $\mu = U$.

(7) Define a Markov chain on $Irr(G)$, the irreducible representations of a finite group G, as follows: Fix at the outset a (not necessarily irreducible) representation η whose character is real-valued, and transition from λ to ρ with probability

$$P_\eta(\lambda, \rho) = \frac{d_\rho m_\rho(\lambda \otimes \eta)}{d_\lambda d_\eta}$$

where $m_\rho(\lambda \otimes \eta) = (\chi_\rho\,,\, \chi_{\lambda\otimes\eta})$ is the multiplicity of ρ in the direct sum decomposition of $\lambda \otimes \eta$.

Write $\pi(\rho) = \frac{d_\rho^2}{|G|}$ for the *Plancherel measure* on $Irr(G)$.

(a) Show that P_η defines a Markov chain on $Irr(G)$, π defines a probability on $Irr(G)$, and P_η is reversible with respect to π.

(b) Argue by induction that the k-step transitions are given by

$$P_\eta^k(\lambda, \rho) = \frac{d_\rho m_\rho(\lambda \otimes \eta^k)}{d_\lambda d_\eta^k}$$

where η^k is the k-fold tensor product of η with itself.

(Hint: The second orthogonality relation will be useful here.)

(c) Let $s_0, s_1, \ldots, s_{m-1}$ be a full set of conjugacy class representatives with $s_0 = 1$. Define

$$\beta_i = \frac{\chi_\eta(s_i)}{d_\eta}, \quad \psi_i(\rho) = \frac{|\mathbf{cl}(s_i)|^{\frac{1}{2}}}{d_\rho}\chi_\rho(s_i).$$

Show that ψ_i is an eigenfunction of P_η with eigenvalue β_i for $i = 0, 1, \ldots, m-1$.

Homework 5

(1) Let G be a group of order n and fix $\varepsilon \in [0,1]$. Define a probability μ on G by

$$\mu(id) = 1-\varepsilon \quad \text{and} \quad \mu(s) = \frac{\varepsilon}{n-1},\ s \in G \setminus \{id\}.$$

(a) Writing $P^k(t) = \mu^{*k}(t)$, show that

$$\begin{aligned} P^k(id) &= \frac{1}{n} + \frac{n-1}{n}\left(1 - \frac{\varepsilon n}{n-1}\right)^k \\ P^k(s) &= \frac{1}{n} - \frac{1}{n}\left(1 - \frac{\varepsilon n}{n-1}\right)^k,\ s \neq id \end{aligned} . \tag{5.2}$$

Conclude that $\left\|P^k - U\right\|_{TV} = \dfrac{n-1}{n}\left(1 - \dfrac{\varepsilon n}{n-1}\right)^k$.

(b) Show that $\displaystyle\sum_{\rho \in Irr^*(G)} d_\rho \mathrm{Tr}\Big(\widehat{\mu}(\rho)^k \big(\widehat{\mu}(\rho)^k\big)^\dagger\Big) = (n-1)\left(1 - \frac{\varepsilon n}{n-1}\right)^{2k}$.

(2) Show that for any $m, k \in \mathbb{N}$,

$$\sum_{j=1}^{2m} \cos^{2k}\left(\frac{2\pi j}{2m+1}\right) = 2\sum_{j=1}^{m} \cos^{2k}\left(\frac{2\pi j}{2m+1}\right) = 2\sum_{j=1}^{m} \cos^{2k}\left(\frac{\pi j}{2m+1}\right).$$

(3) Consider the random walk on $G = \mathbb{Z}/n\mathbb{Z}$ driven by $\mu(\pm 1) = \frac{1}{2}$ and started at 0. This is an irreducible Markov chain with finite state space so all states will eventually be visited with full probability. Prove that the point which is last visited is uniformly distributed over $G \setminus \{0\}$.

(Hint: Condition on which neighbor was first visited and apply a gambler's ruin argument.)

(4) Recall that for a partition $\lambda \vdash n$, we can define a representation $(\rho_\lambda, M^\lambda)$ of S_n by $\rho_\lambda(\pi)\mathbf{e}_{\{t\}} = \mathbf{e}_{\{\pi t\}}$ where $\{\mathbf{e}_{\{t\}} : \{t\} \text{ a tabloid of shape } \lambda\}$ is a basis for M^λ.

For $\sigma \in S_n$, $S \subseteq S_n$, $s \subseteq [n]$, define $\sigma S = \{\sigma\pi : \pi \in S\}$, $S\sigma = \{\pi\sigma : \pi \in S\}$, and $\sigma s = \{\sigma(i) : i \in s\}$.

(a) Given a tableau t of shape λ, let c_j denote the set of entries in the j^{th} column of t and define the column stabilizer of t by $C_t = \{\pi \in S_n : \pi c_j = c_j \text{ for all } j\}$. Show that for any $\sigma \in S_n$, $C_{\sigma t} = \sigma C_t \sigma^{-1}$.

(b) Using the result from part (a), show that the polytabloid $\mathbf{f}_t = \sum_{\pi \in C_t} \mathrm{sgn}(\pi)\mathbf{e}_{\{\pi t\}} \in M^\lambda$ satisfies $\rho_\lambda(\sigma)\mathbf{f}_t = \mathbf{f}_{\sigma t}$.

(5) Define $m = \lceil \frac{4}{5} n \rceil$. Prove that there is a constant $B > 0$ such that

$$\sum_{j=n-m+1}^{n-1} \binom{n}{j} \frac{n!}{(n-j)!} \left(1 - \frac{j}{n}\right)^{n \log(n)} \leq B$$

for all n.

(Hint: Show that the summands are decreasing for $n \geq 2$ so that the sum can be bounded by the first term times the number of terms. Using the Stirling approximation $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$, prove that this quantity goes to 0 as $n \to \infty$ and therefore is uniformly bounded.)

(6) For the abelian sandpile model on $G = (V, E)$, let $\mathcal{S}$ be the commutative monoid of stable configurations under the operation $\oplus$ of addition followed by stabilization. Recall that $\mathcal{I} \subseteq \mathcal{S}$ is an *ideal* if $\mathcal{I} \oplus \mathcal{S} \subseteq \mathcal{I}$, and $\mathcal{I}$ is *minimal* if $\mathcal{I} \subseteq \mathcal{J}$ for every ideal $\mathcal{J}$ of $\mathcal{S}$.

(a) Writing η_* for the saturated configuration, $\eta_*(v) = \deg(v) - 1$, show that the set of recurrent states $\mathcal{G} = \eta_* \oplus \mathcal{S}$ is the minimal ideal of $\mathcal{S}$.

(b) Prove that any finite commutative semigroup has a minimal ideal which forms an abelian group under the inherited operation.

(7) The Dihedral group of order 6, which we will denote by D_6, has presentation $\langle r, s \,|\, r^3 = 1, s^2 = 1, (rs)^2 = 1 \rangle$. Label the group elements $g_1 = id$, $g_2 = r$, $g_3 = r^2$, $g_4 = s$, $g_5 = rs$, $g_6 = rs^2$ and consider a general probability measure μ on D_6 where $\mu(g_k) = p_k$. (The conjugacy classes of D_6 are $\{g_1\}$, $\{g_2, g_3\}$, and $\{g_4, g_5, g_6\}$, so μ is a class function precisely when $p_2 = p_3$ and $p_4 = p_5 = p_6$.)

(a) Write down the transition matrix for the random walk (D_6, μ).

(b) Compute the unitary irreps of D_6.

(c) Use part (b) to 'block diagonalize' the transition matrix as in Section 4.3 of the class notes.