# Markov Chain Fundamentals (Contd.)

### Recap from last lecture:

- A random walk on a graph is a special type of Markov chain used for analyzing algorithms.

- An irreducible and aperiodic Markov chain does not have a unique stationary distribution ($\pi$) unless it is also ergodic.

- The stationary distribution is a fixed point which a Markov chain tends to approach irrespective of the starting point in the state space.

- Given an irreducible and ergodic Markov chain, we would like to know why(!) and how fast it approaches the stationary distribution.

- Given an exponential state space $\Omega$, we would like to design a Markov chain that samples from a probability distribution defined on the state space in $O(poly(\log|\Omega|))$ time.

## Reversible Markov Chains

These are Markov chains that satisfy the condition of *Detailed Balance*, i.e.

$$\pi(x) \cdot P(x,y) = \pi(y) \cdot P(y,x)$$

where $\pi$ is a stationary distribution corresponding to the transition matrix $P$. If the Markov chain is also ergodic, $\pi$ is *the unique* stationary distribution.

For finite state spaces, $\pi$ can be found from the condition of detailed balance above and the fact that $\pi$ is a discrete probability distribution ($\Sigma_x\ \pi(x) = 1$).

As a reverse implication, if we have a desirable stationary distribution $\pi_{des}$ and a set of transitions (a $Markov\ Kernel$ to be precise) defined on the state space, we can derive suitable values for the transition probabilities to achieve $\pi_{des}$.

**Note:** A Markov kernel is a transition matrix analogue for Markov chains in general (as opposed to those with a finite state space). It is essentially a succinct representation of transitions defined on a (usually exponential) state space.

**Example:** Given an undirected graph $G = (V, E)$, (say $C_4$), find a uniformly random indepedent set from $G$.

Here, the vertices in the state space are independent sets in $C_4$, and we connect any two sets $\sigma_i$ and $\sigma_j$, if one can be arrived from the other by adding or deleting exactly one vertex from $V$. To enforce aperiodicity on the resulting finite and irreducible Markov chain (hence making way for ergodicity and an easier analysis), we also add a self-loop to every $\sigma_i$.

Note that every $\sigma_i$ has $|\sigma_i|$ neighbours with one less vertex, where $|\sigma_i|$ is the size of the corresponding independent set. In this particular case, every $\sigma_{i\neq 0}$ has 3 neighbours (in the figure, solid bubbles indicate the vertices in the indepedent set).

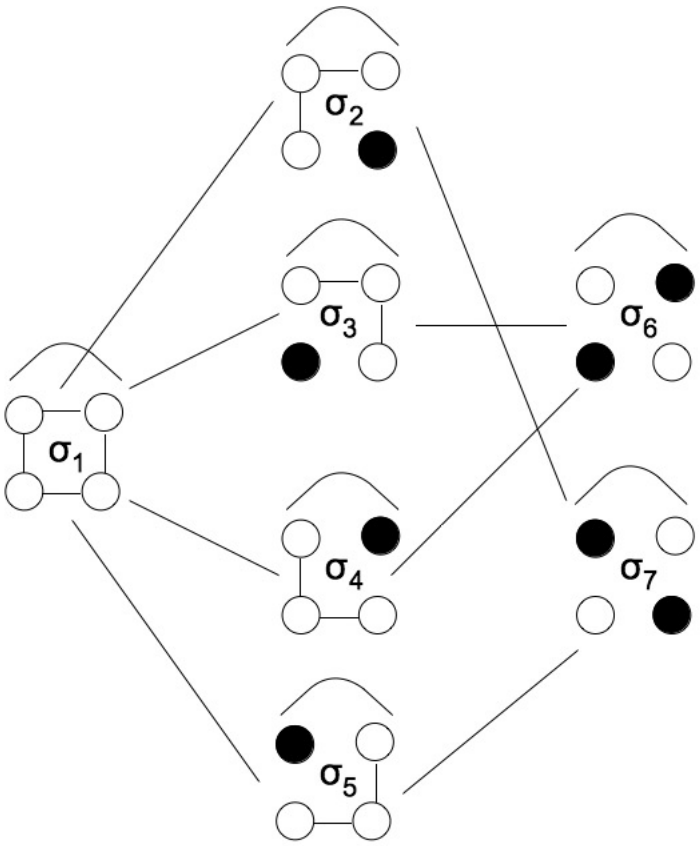


Figure 1: Markov kernel for sampling an independent set of $C_4$

Now let's try to configure the kernel for uniform sampling.

**Try 1:** Being in any state, choose one of the neighbours uniformly.

This gives us for example $P(\sigma_5, \sigma_1) = 1/3$ and $P(\sigma_1, \sigma_5) = 1/5$, and hence $\pi(1)/\pi(5) = 5/3$, which is a non-uniform distribution. Clearly this is not a good choice.

**Try 2** (*Metropolis Method*): Choose $(v, b) \in_u Vx\{0, 1\}$. If $b = 0$, try to remove vertex $v$ from the indepedent set. If $b = 1$, try to add vertex $v$ to the indepedent set.

Let's consider $\sigma_4$. There is a $1/2n = 1/8$ chance that we move to $\sigma_1$ ($v$ = vertex in the independent set, $b = 0$), $1/8$ chance that we move to $\sigma_6$ ($v$ = vertex diagonally opposite to the one in the independent set, $b = 1$), and a $3/4$ chance that we stay put (owing to either, a violation of the independent set property or the action suggested by the method being inapplicable). Likewise, the chance that we move from $\sigma_1$ (or $\sigma_6$) to $\sigma_4$ is also $1/8$. Other cases can be worked out in a similar

manner. We will see that the resulting distribution is uniform, i.e, $\pi(i)/\pi(i \neq j) = 1$ (whenever $(\sigma_i, \sigma_j)$ is an edge), even though it may be approached very slowly (why?).

Incidentally, an identical distribution is achieved by applying *Glauber Dynamics* (or *Heat Bath*), but that does not hold true in general.

## Total Variation Distance

In order to be able to calculate the rate of convergence to the stationary distribution, given an initial distribution, we must know "how far" the current distribution is from the stationary distribution at any point of time; this is measured by the *Total Variation Distance*.

For any set $E \in \Omega$, let's define

$$P(E) = \Sigma_{x \in E} P(x)$$

Then given two distributions $P$ and $Q$ on $\Omega$, the total variation distance between them is given by

$$||P - Q||_{TV} = Max_{E \subseteq \Omega} |P(E) - Q(E)|$$

**Lemma:** Given two distributions $P$ and $Q$ on $\Omega$,

$$||P - Q||_{TV} = \tfrac{1}{2} \Sigma_{x \in \Omega} |P(x) - Q(x)| = \tfrac{1}{2} ||P - Q||_1$$

**Proof:** Let $A = \{x \in \Omega : P(x) \geq Q(x)\}$ and $B$ be any subset of $\Omega$. Then

$$P(B) - Q(B) \leq P(A \cap B) - Q(A \cap B) \leq P(A) - Q(A)$$

The first inequality is true because any $x \in A \cap \bar{B}$ satisfies $P(x) - Q(x) \leq 0$, so the difference in probability cannot decrease when such elements are eliminated.. For the second inequality too, note that including more elements of $B$ cannot decrease the difference in probability.

By exact parallel reasoning,

$$Q(B) - P(B) \leq Q(\bar{A}) - P(\bar{A}) = P(A) - Q(A)$$

Thus for any set $B \subseteq \Omega$,

$$|P(B) - Q(B)| \leq Q(\bar{A}) - P(\bar{A}) = P(A) - Q(A)$$

Thus, by definition,

$$\begin{aligned} ||P - Q||_{TV} &= Max_{B \subseteq \Omega} |P(B) - Q(B)| = \tfrac{1}{2}[(Q(\bar{A}) - P(\bar{A})) + (P(A) - Q(A))] \\ &= \tfrac{1}{2}(\Sigma_{x \in \bar{A}} |Q(x) - P(x)| + \Sigma_{x \in A} |P(x) - Q(x)|) \\ &= \tfrac{1}{2} \Sigma_{x \in \Omega} |P(x) - Q(x)| \end{aligned}$$

□

Figure 2 presents an intuitive picture of the lemma. Since both P and Q are probability distributions, the area under each individual curve is 1. So, the sum of areas marked '+' equals the sum of areas marked '-' $= \frac{1}{2}$ (total area enclosed between the curves) $= TV$.

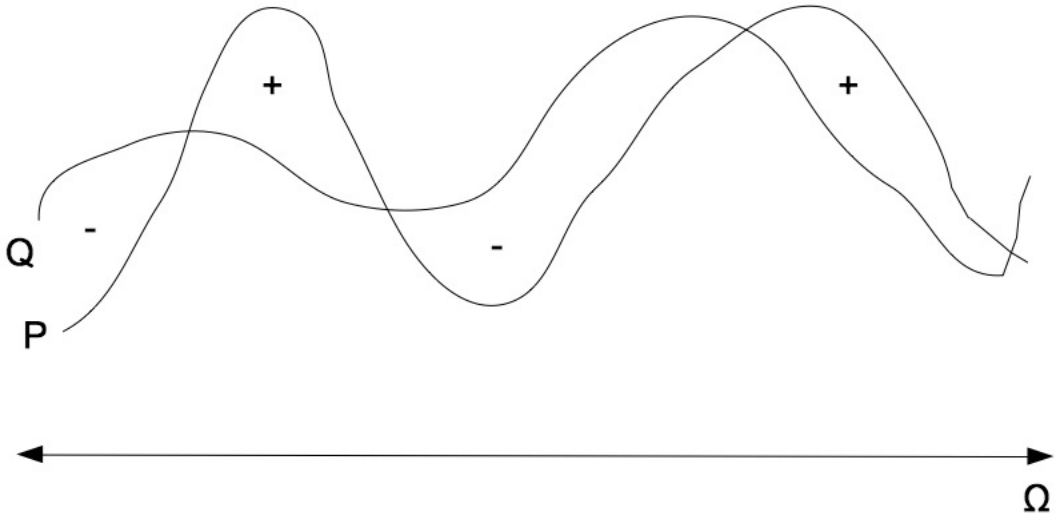


Figure 2: Total Variation Distance between P and Q

## Spectral Gap and Mixing Time

The *Mixing Time* $\tau_\varepsilon$ of a Markov chain with a stationary probability distribution $\pi$ (or $P_{eq}$) is given by

$$\tau_\varepsilon = Min_{\forall P_0}\{t : ||\pi_t - \pi||_{TV} < \varepsilon\}$$

**Note:** Some texts also define mixing time as $\tau = \tau_{1/4}$.

From elementary linear algebra, we know that any irreducible and ergodic Markov chain with a transition matrix $P$ will have $\pi$ as an eigenvector with eigenvalue 1 since

$$\pi \cdot P = 1 \cdot \pi$$

Also, the $Perron - Frobenius$ theorem states that any real $n$ x $n$ matrix $A$ with strictly positive entries ($a_{ij} > 0, 1 \leq i, j \leq n$) has a positive, real eigenvalue $\lambda_0$ (called the Perron-Frobenius eigenvalue) such that for every other eigenvalue, $|\lambda_i| < \lambda_0$ for all $i \neq 0$. This eigenvalue is characterized by a unique associated eigenvector $v_0$ with strictly positive components.

Thus, in the context of irreducible and ergodic Markov chains, $\lambda_0 = 1$ and $v_0 = \pi$.

If $|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_{N-1}|$ be the remaining eigenvalues of $P$, where $N = |\Omega|$, $\delta = 1 - |\lambda_1|$ is referred to as the *Spectral Gap*, the difference between the two dominating eigenvalues.

The following inequality (which we will prove in next lecture) relates the mixing time of a Markov chain to its spectral gap.

$$\tau_\varepsilon \leq \frac{ln(N\varepsilon^{-1})}{\delta}$$

Clearly, smaller the spectral gap (closer the two dominating eigenvalues), larger the mixing time.

**Example:** Given the two-state Markov chain in the figure below and an initial probability distribution of $(1, 0)$, calculate the total variation distance from $\pi$ at time t.

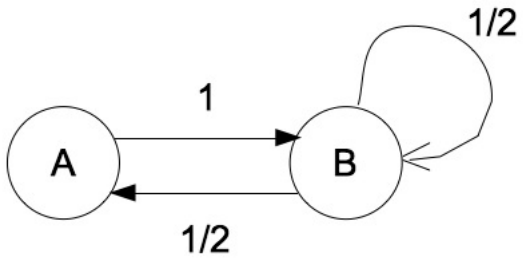


Figure 3: Example

Here $P = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}$, and $\pi_0 = [1\ 0]$. Using $\pi_t = \pi_0 \cdot P^t$, we get: $\pi_1 = [0\ 1], \pi_2 = [1/2\ 1/2], \pi_3 = [1/4\ 3/4]$, and so on.

Also, $\lambda_0 = 1$, and the other eigenvalue $(= -1/2)$ can be obtained from the trace of $P$ ( = sum of eigenvalues). The corresponding eigenvectors are $v_0 = [1/3\ 2/3]$ and $v_1 = [1\ -1]$.

Now since every vector can be written as a linear combination of the eigenvectors, $\pi_0 = 1 \cdot v_0 + \frac{2}{3}v_1$ (note here that the multiplicity of $v_0 (= \pi)$ is 1 as expected), and $\pi_t = \pi_0 \cdot P^t = (v_0 + \frac{2}{3}v_1) \cdot P^t = v_0 + \frac{2}{3}\lambda_1^t \cdot v_1$.

So, $||\pi_t - \pi||_{TV} = \frac{1}{2}||\frac{2}{3}\lambda_1^t \cdot v_1||$.

Note that any starting probability distribution $\pi_0$ can be represented as a linear combination of the eigenvectors including $v_0$ (stationary distibution) and hence its evolution $\pi_t$ with $t$ as shown below,

$$\pi_0 = v_0 + \Sigma_{i \neq 0} a_i v_i; \pi_t = v_0 + \Sigma_{i \neq 0} a_i \lambda_i^t v_i$$

and since $|\lambda_{i \neq 0}| < 1$, the $v_{i \neq 0}$ components diminish as $t$ grows larger, making $\pi_t \to v_0$.