Chapter 12

Markov Chains & PageRank

Let us try to predict the weather! How long until it is rainy the next time? What about the weather in ten days? What is the local “climate”, i.e., the “average” weather?

![Weather Graph](image)

Figure 12.1: According to a self-proclaimed weather expert, the above graph models the weather in Zürich. On any given day, the weather is either sunny, cloudy, or rainy. The probability to have a cloudy day after a sunny day is $\frac{1}{2}$. In the context of Markov chains the nodes, in this case sunny, rainy, and cloudy, are called the states of the Markov chain.

Remarks:
- Figure 12.1 above is an example of a Markov chain—see the next section for a formal definition.
- If the weather is currently sunny, the predictions for the next few days according to the model from Figure 12.1 are:

<table>
<thead>
<tr>
<th>Day</th>
<th>sunny</th>
<th>cloudy</th>
<th>rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.611</td>
<td>0.222</td>
<td>0.167</td>
</tr>
<tr>
<td>3</td>
<td>0.574</td>
<td>0.259</td>
<td>0.167</td>
</tr>
<tr>
<td>4</td>
<td>0.568</td>
<td>0.247</td>
<td>0.185</td>
</tr>
</tbody>
</table>

108 CHAPTER 12. MARKOV CHAINS & PAGERANK

12.1 Markov Chains

Markov chains are a tool for studying stochastic processes that evolve over time.

**Definition 12.2** (Markov Chain). Let $S$ be a finite or countably infinite set of states. A (discrete time) Markov chain is a sequence of random variables $X_0, X_1, X_2, \ldots \in S$ that satisfies the Markov property (see below).

**Definition 12.3** (Markov Property). A sequence $\{X_t\}$ of random variables has the Markov property if for all $t$, the probability distribution for $X_{t+1}$ depends only on $X_t$, but not on $X_0, X_1, \ldots, X_{t-1}$. More formally, for all $t \in \mathbb{N}_{\geq 0}$ and $s_0, \ldots, s_{t+1} \in S$ it holds that $\Pr[X_{t+1} = s_{t+1} | X_0 = s_0, X_1 = s_1, \ldots, X_t = s_t] = \Pr[X_{t+1} = s_{t+1} | X_t = s_t]$.

Remarks:
- A sequence of random variables is also called a discrete time stochastic process. Processes that satisfy the Markov property are also called memoryless.
- The probability distribution of $X_0$ does not depend on a previous state (since there is none). It is called the initial distribution, and we denote it by the vector $q_0 = (q_{0,s})_{s \in S}$ with the entries $\Pr[X_0 = s]$ for every state $s \in S$. If the first day is sunny, the initial distribution is $q_0 = (1, 0, 0)$.

**Definition 12.4** (Time Homogeneous Markov Chain). A Markov chain is time homogeneous if $\Pr[X_{t+1} = s_{t+1} | X_t = s_t]$ is independent of $t$, and in that case $p_{i,j} = \Pr[X_{t+1} = j | X_t = i]$ is well defined.

Remarks:
- We will only consider time homogeneous Markov chains.
- Markov chains are often modeled using directed graphs, as in Figure 12.1. The states are represented as nodes, and an edge from state $i$ to state $j$ is weighted with probability $p_{i,j}$.
- Just like directed graphs, Markov chains can be written in matrix form (using the adjacency matrix). In this context, the matrix is called the transition matrix, and we denote it by $P$. For the example from Figure 12.1, the transition matrix is:

<table>
<thead>
<tr>
<th>From</th>
<th>sunny</th>
<th>cloudy</th>
<th>rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>2/3</td>
<td>1/3</td>
<td>0</td>
</tr>
<tr>
<td>cloudy</td>
<td>1/2</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>rainy</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

- Let $q_0 = (q_{0,s})_{s \in S}$ be the probability distribution on $S$ for time $t$, i.e., $q_{0,s} = \Pr[X_0 = s]$. The probability to be in state $j$ at time $t + 1$ is $q_{t+1,j} = \sum_{s \in S} q_{t,s} p_{s,j}$. This can be written as the vector-matrix-multiplication $q_{t+1} = q_t \cdot P$. 

\[ \]
• The state distribution at time $t$ is $q_t = q_0 \cdot P^t$. We denote by $p^{(t)}_{i,j}$ the entry at position $i, j$ in $P^t$, i.e., the probability of reaching $j$ from $i$ in $t$ steps.

• If we start a Markov chain in a single state (with probability 1), we do a random walk.

**Definition 12.5 (Random Walk).** Let $G = (V,E)$ be a directed graph, and let $\omega : E \rightarrow [0,1]$ be a weight function so that $\sum_{v \in G} \omega(u,v) = 1$ for all nodes $u$. Let $u \in V$ be the starting node. A weighted random walk on $G$ starting at $u$ is the following discrete Markov chain in discrete time. Beginning with $X_0 = u$, in every step $t$, the node $X_{t+1}$ is chosen according to the weights $\omega(X_t,v)$, where $v$ are the neighbors of $X_t$. If $G$ is undirected and unweighted, then $X_{t+1}$ is chosen uniformly at random among $X_t$’s neighbors and the random walk is called simple.

Remarks:

• In Section 12.5 we will study some simple random walks.

• Now that we understand the definitions, we want to do some calculations. E.g., if it is sunny today, how long will it stay sunny?

### 12.2 Hitting Time & Arrival Probability

**Definition 12.6 (Sojourn Time).** The sojourn time $T_i$ of state $i$ is the time the process stays in state $i$.

Remarks:

• It holds that $\Pr[T_i = k] = p^{k-1}_{i,i} \cdot (1 - p_{i,i})$, i.e., $T_i$ is geometrically distributed. For example $\mathbb{E}[T_{\text{sojourn}}] = 3$.

• The sojourn time $T_i$ does not depend on the time the process has spent in state $i$ already (memoryless property). The geometric distribution is the only discrete distribution that is memoryless.

• If it is currently sunny, how long does it take until we see the first rainy day?

**Definition 12.7 (Hitting Time & Arrival Probability).** Let $i$ and $j$ be two states. The hitting time $T_{i,j}$ is the random variable counting the number of steps until visiting $j$ the first time when starting from state $i$, i.e., the value of $T_{i,j}$ is the smallest integer $t \geq 1$ for which $X_t = j$ under the condition that $X_0 = i$. The expected hitting time from $i$ to $j$ is the expected value $\mathbb{E}[T_{i,j}]$.

The arrival probability from $i$ to $j$ is the probability $f_{i,j} = \Pr[T_{i,j} < \infty]$.

**Remarks:**

• The time $c_{i,j} = h_{i,j} + f_{i,j}$ is referred to as the commute time between $i$ and $j$.

• The following lemma states that the expected hitting time can be computed by solving a system of linear equations.

**Lemma 12.8.** If $h_{i,j}$ exists for all $i, j \in S$, then the expected hitting times are

$$h_{i,j} = 1 + \sum_{k \neq j} p_{i,k} h_{k,j}.$$

Proof. Plugging in the definition of $h_{i,j}$ and applying the law of total probability we get that

$$h_{i,j} = \mathbb{E}[T_{i,j}] = \sum_{k \leq S} \mathbb{E}[T_{i,j} \mid X_1 = k] \cdot p_{i,k}.$$

Taking the $j^{th}$ term out, we obtain

$$h_{i,j} = \mathbb{E}[T_{i,j} \mid X_1 = j] \cdot p_{i,j} + \sum_{k \neq j} \mathbb{E}[T_{i,j} \mid X_1 = k] \cdot p_{i,k} = 1 \cdot p_{i,j} + \sum_{k \neq j} (1 + \mathbb{E}[T_{k,j}]) \cdot p_{i,k}.$$

Since $p_{i,j}$ together with all the values $p_{i,k}$ sum up to 1, we can simplify to

$$h_{i,j} = 1 + \sum_{k \neq j} \mathbb{E}[T_{k,j}] \cdot p_{i,k} = 1 + \sum_{k \neq j} p_{i,k} h_{k,j}.$$

**Remarks:**

• On a sunny day it takes in expectation 8 days until it starts raining.

• Lemma 12.9 for the arrival probabilities can be established similarly to Lemma 12.8.

**Lemma 12.9.** For all $i, j \in S$, the arrival probability is

$$f_{i,j} = p_{i,j} + \sum_{k \neq j} p_{i,k} f_{k,j}.$$

### 12.3 Stationary Distribution & Ergodicity

What is the “climate” in Zürich? Often one is particularly interested in the long term behavior of Markov chains and random walks. The mathematical notion that captures a Markov chain’s long term behavior is the stationary distribution, which we will introduce and study in the following.
Remarks:
• The entries in \( P^t \) contain the probability of entering a certain weather condition (state). What happens for large values of \( t \)? The matrix seems to converge.

\[
P^3 = \begin{pmatrix} 0.574 & 0.259 & 0.167 \\ 0.556 & 0.222 & 0.222 \\ 0.537 & 0.259 & 0.204 \end{pmatrix}, \quad P^{10} = \begin{pmatrix} 0.563 & 0.250 & 0.187 \\ 0.562 & 0.250 & 0.187 \\ 0.562 & 0.250 & 0.188 \end{pmatrix}
\]

• No matter what the initial weather \( q_0 \) is, the product \( q_0 \cdot P^t \) seems to approach \( \bar{q} \approx (0.563, 0.250, 0.188) \) as \( t \) grows. Moreover, if we multiply the vector \( \bar{q} \) with \( P \) we almost get \( \bar{q} \) again. In other words, \( \bar{q} \) is almost an eigenvector of \( P \) with eigenvalue 1.

Definition 12.10 (Stationary Distribution). A distribution \( \pi \) over the states is called stationary distribution of the Markov chain with transition matrix \( P \) if \( \pi = \pi \cdot P \).

Remarks:
• Our weather Markov chain converges towards \( \pi = (9/16, 4/16, 3/16) \), which is an eigenvector of \( P \) with eigenvalue 1. We conclude that in the long run, 9 out of 16 days are sunny in Zürich. The weather model appears to be not as accurate as the weather expert led us to believe.

• Consider the sequence \( q_0 = q_{16} \cdot P \), where \( q_0 \) is the initial distribution. In general, this sequence does not necessarily converge as \( t \) grows. However, if it does converge to some distribution \( \pi \), then it must hold that \( \pi = \pi \cdot P \).

Lemma 12.11. The transition matrix of every Markov chain has a left eigenvector with eigenvalue 1.

Proof. Let \( P \) be the transition matrix of a Markov chain, and denote by \( e = (1, \ldots, 1)^\top \) the all-ones vector. Because in \( P \) the entries in each row sum up to 1 (\( P \) is row stochastic), it holds that \( Pe = e \). Denoting by \( I \) the identity matrix, it follows that \((P - I)e = 0\). In other words, \( e \) is an eigenvector with eigenvalue 0 for \((P - I)\), which implies that \((P - I)^\dagger\) is singular, i.e., not invertible. Thus, also \((P - I)^\dagger P = P^\dagger - \pi P \) is singular, and it follows that there is a vector \( \pi \neq 0 \) so that \( 0 = (P - I)^\dagger \pi = P^\dagger \pi - \pi \pi \). Transposing and rearranging we obtain that \( \pi^\top P = \pi^\top \), as desired.

Remarks:
• Using Brouwer’s fixed point theorem one can show that there is also a left eigenvector \( \pi \) that corresponds to a probability distribution.

• The stationary distribution is not necessarily unique, see Figure 12.12.

The issue is that some states are not reachable from all other states.
by switching the summation on the right hand side. Since \( \pi \) is an eigenvector with eigenvalue 1, it holds that \( \sum_{i} \pi_{i} h_{i,j} = \pi_{j} \), and thus the equation becomes

\[
\pi_{j} h_{j,j} + \sum_{i \neq j} \pi_{i} h_{i,j} = 1 + \sum_{k \neq j} h_{k,j} \pi_{k}.
\]

Noting that all \( h_{i,j} \geq 1 \), we conclude that \( \pi_{j} = 1 / h_{j,j} \), as desired. In the remaining case where \( \sum_{i} \pi_{i} = 0 \), the equation turns into

\[
\pi_{j} h_{j,j} + \sum_{i \neq j} \pi_{i} h_{i,j} = \sum_{k \neq j} h_{k,j} \pi_{k},
\]

yielding that \( \pi_{j} = 0 \) for all \( j \). This contradicts that \( \pi \) is an eigenvector. \( \square \)

Remarks:

• Irreducible Markov chains with an infinite number of states do not necessarily have a stationary distribution.

• Depending on the choice of the initial distribution, even an irreducible Markov chain does not necessarily converge towards its stationary distribution, see Figure 12.16.

![Figure 12.16](image-url)  
Figure 12.16: This Markov chain is irreducible, and has the unique stationary distribution \( \pi = (0.5, 0.5) \). In this particular chain, each state can only be reached every other step, or in other words, both states have period 2. Therefore, the initial distribution is attained in every second step, and only \( q_{0} = \pi \) “converges” towards the stationary distribution.

Definition 12.17 (Aperiodic Markov Chains). The period of a state \( j \in S \) is the largest \( \xi \in \mathbb{N} \) such that

\[
\{ n \in \mathbb{N} | p_{j}^{(n)} > 0 \} \subseteq \{ i : \xi | i \in \mathbb{N} \}
\]

A state with period \( \xi = 1 \) is called aperiodic, and the Markov chain is aperiodic if all its states are.

Remarks:

• One can show that if the Markov chain is irreducible, then all states have the same period.

• A state \( j \) with \( p_{j,j} > 0 \) is trivially aperiodic.

• If \( p_{j,j} = 0 \), then one can check whether state \( j \) is aperiodic by testing, as illustrated in Figure 12.18, if the following holds: Does \( j \) lie on two directed cycles of lengths \( k \) and \( l \) (counting the edges in the chain) so that \( k \) and \( l \) are relatively prime, i.e., have a greatest common divisor of 1? Or, using the \( k \)th and \( l \)th powers of \( P \), are there relatively prime \( k \) and \( l \) such that both \( p_{j,j}^{(k)} \) and \( p_{j,j}^{(l)} > 0 \)?

Theorem 12.20. If a Markov chain is ergodic it holds that

\[
\lim_{t \to \infty} q_{t} = \pi,
\]

where \( \pi \) is the unique stationary distribution of the chain.

Remarks:

• The theorem holds regardless of the initial distribution.

• The stationary distribution of ergodic Markov chains can thus be approximated efficiently, namely by successively multiplying a vector with a matrix instead of computing the powers of a matrix.

12.4 PageRank Algorithm

Google’s PageRank algorithm is based on a Markov chain obtained from a variant of a random walk.

Remarks:

• Google provides search results that match the user’s search terms. Under the hood Google maintains a ranking among websites to make sure “better” or “more important” websites appear early in the search results. Instead of solving the whole problem at once, this ranking is first established globally (independent of the search terms), and only later websites matching the search query are sorted according to some rank. In this section we focus on the ranking part.

• The first step to ranking websites is to crawl the web graph, i.e., a directed graph in which the nodes are websites, and an edge \( (u, v) \) indicates that website \( u \) contains a hyperlink to website \( v \).
12.4. PAGERANK ALGORITHM

Figure 12.21: An example of a web graph with 5 websites. Website x does not link to any other website, i.e., x is a sink.

- A naive approach is to rank the sites by the number of incoming hyperlinks. In the example from Figure 12.21 this yields the same rank for websites w and x. One could, however, argue that the link from w to x means that x is more important than w.
- Google’s idea is to model a random surfer who follows hyperlinks in the web graph, i.e., performs a simple random walk. After sufficiently many steps, the websites can be ranked by how many times they were visited. The intuition is that websites are visited more often if they are linked to by many other sites, which should be a good measure of how important a website is.
- Since the walk is directed, the random surfer can get stuck in sinks (nodes with no outgoing edges). To fix this issue, a random website is chosen for the next step whenever the random surfer reaches a sink.
- Let us denote the random surfer matrix describing this simple random walk by W.
- Simulating the simple random walk described by W to find a stationary distribution is not feasible: There are over 1 billion websites—meaning that a lot of steps have to be simulated to get a good estimation of the stationary distribution. Using our knowledge about Markov chains we can simulate many random walks at once by repeatedly multiplying some initial distribution q0 with W.
- There is no guarantee that this process converges to a stationary distribution. We know that this can be fixed by making the Markov chain ergodic.
- One way to make a Markov chain ergodic is to insert an edge between every two nodes.

Definition 12.22 (Google Matrix). Let W be a random surfer matrix, and let $\alpha \in (0, 1)$ be a constant. Denote further by R the matrix in which all entries are $1/n$. The following matrix M is called the Google Matrix:

$$M = \alpha \cdot W + (1 - \alpha) \cdot R.$$  

Remarks:
- The intuition behind R is that in every step, with probability $1 - \alpha$, the random surfer “gets bored” by the current website and surfs to a new random site.
- While the R-component in M ensures that the Markov chain converges, it also changes the stationary distribution. To ensure the impact is not too large, $\alpha$ should be chosen close to 1. A typical value for $\alpha$ is 0.85.
- The rate at which the process converges depends on the magnitude of M’s second largest eigenvalue. One can show that for M the second largest eigenvalue is at most $\alpha$, and that the error decreases by a factor of $\alpha$ in each step.
- In the example from Figure 12.21, the page ranks are

<table>
<thead>
<tr>
<th>Website</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>0.384615</td>
</tr>
<tr>
<td>w</td>
<td>0.230769</td>
</tr>
<tr>
<td>u</td>
<td>0.153846</td>
</tr>
<tr>
<td>v</td>
<td>0.153846</td>
</tr>
<tr>
<td>y</td>
<td>0.0769231</td>
</tr>
</tbody>
</table>

- This initial version of the PageRank algorithm worked well at the time it was invented. However, it can be (and has been) fooled. Consider the following example.

Figure 12.23: Website u wants to improve its PageRank, which is $\approx 0.23$ in the initial setting on the left. First, all outgoing links to websites that do not link back are removed. The PageRank improves to $\approx 0.27$. In a Sybil attack (right) the owner of u creates fake websites $u'$ and $u''$ whose purpose is to exchange links with u. Moreover, the new websites increase the probability to visit u after a sink. Now, website u is the highest ranked site in the network with a rank of $\approx 0.41$.

- Attacks where a single party pretends to be more than one individual are called Sybil Attacks.
- It is unknown how exactly Google ranks websites today, and specifically how the engineers at Google mitigate the effects of attacks.
• A different kind of attack on Google is Google bombing. This attack relies on the fact that the search terms for which a website is considered relevant also take the anchor text of hyperlinks to website into account. If, for instance, many websites link to http://www.ethz.ch using the anchor text “Smartest People Alive”, then a search query for smart people might end up presenting ETH’s website.

12.5 Simple Random Walks

In this section, all random walks are considered to be simple. This means that the edges are undirected, and the node for the next step is chosen uniformly at random among the current node’s neighbors.

Lemma 12.24. Let \( G \) be a graph with \( m \) edges. The stationary distribution \( \pi \) of any simple random walk on \( G \) is

\[
\pi_u = \frac{\delta(u)}{2m}.
\]

Proof. The Markov chain underlying the random walk is irreducible, with Lemma 12.15 it has a unique stationary distribution. We first verify that \( \pi \) from above satisfies the equation \( \pi = \pi \cdot P \) from Definition 12.10:

\[
\pi_u = \sum_{v \in N(u)} \pi_v \cdot p_{v,u} = \sum_{v \in N(u)} \frac{\delta(v)}{2m} \cdot \frac{1}{\delta(v)} = \frac{\delta(u)}{2m},
\]

for some arbitrary node \( u \in V \). Since \( \sum_{v \in V} \delta(v) = 2m \), \( \pi \) is a distribution. This proves that \( \pi \) is the unique stationary distribution. \( \square \)

Remarks:

• It follows from Lemma 12.15 that for a simple random walk, \( h_{u,u} \) is \( 2m/\delta(u) \).

• The cover time \( \text{cov}(v) \) is the expected number of steps until all nodes in \( G \) were visited at least once, starting at \( v \).

• One could use the following Markov chain to compute the cover time of a simple random walk on the graph \( G = (V,E) \). The set of states is \( \{(v,I) \mid v \in V \text{ and } I \in 2^V \cup \{\emptyset\} \} \), where \( v \) denotes the current state, \( I \) denotes the visited states, and \( \emptyset \) is an additional sink state. The probabilities \( p_{(v,I), (w,J)} \) are 0 if either \( I = V \) or \( w \notin E \), and \( 1/\delta(v) \) else if \( (v,w) \in E \). Additionally, each state \( (v,I) \) has an edge to the sink state \( (\emptyset, I) \) with \( p_{(v,I), (\emptyset, I)} = 1 \). Then, the cover time is \( \text{cov}(v) = h_{(v,I), (\emptyset, I)} \).

Lemma 12.25. Let \( G = (V,E) \) be a graph with \( n \) nodes and \( m \) edges. It holds that \( \text{cov}(s) < 4m(n-1) \) for any starting node \( s \in V \).

Proof. Let \( \{u,v\} \in E \) be an edge. It holds that

\[
\frac{2m}{\delta(u)} = h_{u,u} = \frac{1}{\delta(u)} \sum_{s \in N(u)} (h_{s,u} + 1),
\]

and thus it must be true that \( h_{s,u} < 2m \). Next, observe that it is possible to traverse all nodes in \( G \) by using no more than \( 2n - 2 \) edges, e.g., by traversing a spanning tree rooted at \( s \). Since \( h_{s,u} < 2m \) holds for every edge \( \{u,v\} \) used in the traversal, it follows that \( \text{cov}(s) < (2n - 2) \cdot 2m = 4m(n-1) \), as desired. \( \square \)

Remarks:

• Consider the resistor network obtained from \( G \) by replacing every edge with a \( 1 \Omega \) resistor. Let \( u \) and \( v \) be two nodes in the resistor network.

It can be shown that the commute time \( c_{u,v} = 2m \cdot R(u,v) \), where \( R(u,v) \) denotes the effective resistance between \( u \) and \( v \).

• Foster’s Theorem states that for every connected graph \( G = (V,E) \) with \( n \) nodes,

\[
\sum_{(u,v) \in E} R(u,v) = n - 1,
\]

i.e., that adding/removing an edge in \( G \) reduces/increases the effective resistance, respectively.

Chapter Notes

Historic background on the development of Markov chains can be found in [1]. The short version is that in a 1902 paper [9], the theologian Pavel Alekseevich Nekrasov, in his effort to establish a new philosophical basis, (false) postulated that independence of events is necessary for the law of large numbers. Markov, being an atheist and considering Nekrasov’s reasoning an “abuse of mathematics”, set out to prove him wrong.

In 1906, Markov published his first findings on chains of pairwise dependent random variables [7]. This work already includes a variant of Theorem 12.20, thus disproving Nekrasov’s claim. Markov also studied the notion of irreducibility [8], proving that for irreducible Markov chains 1 is a single eigenvalue and thus disproving Nekrasov’s claim. Markov also studied the notion of irreducibility [8], proving that for irreducible Markov chains 1 is a single eigenvalue and the largest by magnitude. Today, Markov’s ideas are widely applied in, e.g., physics, chemistry, and economics.

Markov chains are the basis for queueing theory, an important transport layer concept. Another application in computer science is the PageRank algorithm [10]. The bound on the Google matrix’ second eigenvalue is from [5]. Sybil attacks were originally studied in the context of peer to peer systems [3], and PageRank’s sensitivity to such attacks was investigated in [2].

The connection from random walks to resistor networks is investigated in depth in [4]. By associating a word with each state, random walks can be used to generate random text [11]. More than 120 “scientific” papers were generated using such methods [6] and later withdrawn by the publishers.

This chapter was written in collaboration with Jochen Seidel.

Bibliography


