

Probabilistic Aspects of Computer Science: Part 1

Rohit Chadha

LSV, ENS Cachan, CNRS and INRIA

E-mail address: `chadha@lsv.ens-cachan.fr`

September 18, 2011

Contents

Chapter 1. Preliminaries	3
1.1. Languages.	3
1.2. Basics of probability measure	3
Chapter 2. Markov Chains	5
2.1. Irreducible and aperiodic Markov chains	6
2.2. Stationary distribution	8
2.3. Existence of unique limiting stationary distributions	9
2.4. Computing Reachability Properties	12
Chapter 3. Probabilistic Automata	19
3.1. Definitions	19
3.2. Extremal Cut-Points	21
3.3. Non-extremal cut-points and regular languages	23
3.4. Undecidability of decision problems for threshold languages	25
Chapter 4. Basics of information theory	27
4.1. Log sum inequality	27
4.2. Basic Coding theory	28
4.3. Huffman Coding	32

CHAPTER 1

Preliminaries

We set some notations that we shall use throughout the notes.

- (1) The set of natural numbers shall be denoted by \mathbb{N} (we will include 0 in \mathbb{N}). The set of integers shall be denoted by \mathbb{Z} . The set of real numbers shall be denoted by \mathbb{R} .
- (2) We shall assume that the reader is familiar with some basic set theory. Given a universe \mathcal{U} , the empty set shall be denoted by \emptyset , the subset relation by \subseteq (with \subsetneq standing for strict inclusion), union, intersection, set difference and product by $\cup, \cap, \setminus, \times$ respectively. As usual a n -ary relation on sets A_1, \dots, A_n is a subset of $A_1 \times A_2 \dots A_n$ and a function $f : A \rightarrow B$ from *domain* A to *codomain* B is a relation on A and B which is total and functional.
- (3) As usual, the power set of a set A is the set of all subsets of A and shall be denoted as 2^A .
- (4) Given $r \in \mathbb{R}$, $|r|$ will denote the absolute value of r .

1.1. Languages.

Sequences/Strings/Words. Given a finite set S , $|S|$ will denote the number of elements of S . Given a finite sequence/string/word $\kappa = s_0, s_1, \dots, s_{n-1}$ over S , $|\kappa|$ will denote the length n of the sequence, and $\kappa[i]$ will denote the i th element s_i of the sequence. As usual S^* will denote the set of all finite sequences/strings/words over S and S^+ will denote the set of all finite non-empty sequences/strings/words over S . Given $\eta, \kappa \in S^*$, $\eta\kappa$ is the sequence obtained by concatenating the two sequences in order. Given $L_1, L_2 \subseteq \Sigma^*$, the set L_1L_2 is defined to be $\{\eta\kappa \mid \eta \in L_1 \text{ and } \kappa \in L_2\}$. Given natural numbers $i, j \leq |\kappa|$, $\kappa[i : j]$ is the finite sequence s_i, \dots, s_j , where $s_k = \kappa[k]$.

Languages. A language L of finite words over a finite alphabet Σ is a subset of Σ^* .

1.2. Basics of probability measure

Let Ω be any set and let $\mathcal{F} \subseteq 2^\Omega$. We shall say that \mathcal{F} is a *probability field* (or *algebra*) if $\emptyset, \Omega \in \mathcal{F}$ and \mathcal{F} is closed under basic set operations (intersection, union and set difference). Elements of \mathcal{F} are called *events*. A

map $\mu : \mathcal{F} \rightarrow [0, 1]$ is said to a *probability distribution* (or *probability measure*) if μ satisfies the following conditions.

- (1) $\mu(\emptyset) = 0$.
- (2) $\mu(\Omega) = 1$.
- (3) If $E_1, E_2 \subseteq \Omega$ are disjoint sets then $\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2)$.

The triple $(\Omega, \mathcal{F}, \mu)$ is said to be a *probability space*. Given an event $E \in \mathcal{F}$, $\mu(E)$ is said to be the *probability that E happens*.

EXERCISE 1.1. Show that if $(\Omega, \mathcal{F}, \mu)$ is a probability space and $A, B \subseteq \Omega$ then $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$ and $\mu(A \setminus B) = \mu(A) - \mu(A \cap B)$.

EXAMPLE 1.2. (Tossing a fair coin). When a coin is tossed, there are two outcomes, heads **H** and tails **T**. A fair coin toss is then modeled as a probability space $(\Omega, \mathcal{F}, \mu)$ with $\Omega = \{\mathbf{H}, \mathbf{T}\}$, $\mathcal{F} = \{\emptyset, \{\mathbf{H}\}, \{\mathbf{T}\}, \{\mathbf{H}, \mathbf{T}\}\}$, $\mu(\emptyset) = 0$, $\mu(\mathbf{H}) = \mu(\mathbf{T}) = \frac{1}{2}$ and $\mu(\{\mathbf{H}, \mathbf{T}\}) = 1$.

EXERCISE 1.3. Consider the case of tossing 4 fair coins. How can we model this situation as a probability space? What is the probability of getting at least 2 heads? What is the probability of getting exactly 2 tails?

One way of constructing a probability field is to consider let $\mathcal{F} = 2^\Omega$. In case Ω is finite, a probability measure on (Ω, \mathcal{F}) can be constructed by just defining it on single-element sets (*singletons*) of Ω as follows–

Definition: Given a finite set $\Omega = \{q_1, q_2, \dots, q_k\}$, a *distribution over Ω* is a $1 \times k$ row vector ρ such that for each $1 \leq i \leq k$, $\mu_i \geq 0$ and $\sum_{1 \leq i \leq k} \rho_i = 1$.

EXERCISE 1.4. Given a finite set $\Omega = \{q_1, q_2, \dots, q_k\}$ and a distribution ρ over Ω , let $\mu : 2^\Omega \rightarrow [0, 1]$ be defined such that

$$\mu(E) = \sum_{q \in E} \rho(q).$$

Show that μ is a probability measure on Ω .

CHAPTER 2

Markov Chains

We often need to study probabilistic systems which change with time. Markov chains are an important class of probabilistic systems which change with time. In order to define Markov chains, we need the following definition.

Definition: A *Markov chain* with state space Q is a sequence $\mu^{(0)}, \mu^{(1)}, \dots$ of distributions such that there is a $k \times k$ -matrix δ which satisfies the following

$$\forall n \in \mathbb{N}. \mu^{(n+1)} = \mu^{(n)}\delta.$$

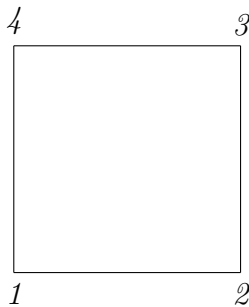
The vector $\mu^{(0)}$ is the initial distribution of Markov chain and δ the transition matrix. The quantity $\mu_i^{(n)}$ is said to be the probability of being in state q_i at time n .

EXERCISE 2.1. Show that if μ^0, μ^1, \dots is a Markov chain with transition matrix δ then for each $n \geq 0$,

- $\mu^{(n)} = \mu^{(0)}\delta^n$.
- $\sum_{1 \leq i \leq k} \mu_i^{(n)} = 1$.

Remark: A Markov chain is fully determined by its state space Q , initial distribution $\mu^{(0)}$ and the transition matrix δ . Therefore, in the rest of the chapter, we shall just denote a Markov chain as a triple $(Q, \mu^{(0)}, \delta)$. It shall also be useful to think of the transition matrix δ of M as a weighted directed graph as follows. The nodes of the graph are states q_i and there is an edge from node q_i to q_j with weight $p > 0$ iff $\delta_{i,j} = p$. This graph is called the *transition graph* of M .

EXAMPLE 2.2. Consider a person walking on the following square.



The person starts at intersection 1. Then the person tosses a fair coin and if the coin turns up heads then the person moves anti-clockwise to intersection 2, otherwise the person moves clockwise to intersection 4. At intersection 2, the person again tosses a coin and if the coin turns up heads then the person moves anti-clockwise to 3 otherwise the person moves clockwise to 1. At intersection 3, the person again tosses a coin and if the coin turns up heads then the person moves anti-clockwise to 4 otherwise the person moves clockwise to 2. At intersection 4, the person again tosses a coin and if the coin turns up heads then the person moves to anti-clockwise 1 otherwise the person moves clockwise to 3.

The position of the person can be thought of as a Markov chain $M_{sq} = (Q, \mu^0, \delta)$ with $Q = \{1, 2, 3, 4\}$, $\mu^{(0)} = [1, 0, 0, 0]$ and transition matrix

$$\delta = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

Note that $\mu^1 = [0, \frac{1}{2}, 0, \frac{1}{2}]$ which means that the probability of the person being at intersection 2 at time instant 1 is $\frac{1}{2}$; and so is the probability of the person being at intersection 4 at time instant 1. The distribution of the person's position at time instant 2 is given by $\mu^{(2)} = [\frac{1}{2}, 0, \frac{1}{2}, 0]$.

EXERCISE 2.3. Consider the Example 2.2.

- (1) What is the transition graph of the Markov chain M_{sq} .
- (2) Compute $\mu^{(n)}$ for all natural numbers n ?
- (3) What is $\lim_{n \rightarrow \infty} \mu^{(n)}$?

EXAMPLE 2.4. Consider the Example 2.2. Assume that instead of tossing one fair coin, a person tosses two fair coins. If the coin turns up heads then the person decides to stay in the position it was before; otherwise the person tosses another coin. If the second coin turns up heads, the person moves clockwise otherwise the person moves anti-clockwise. Let the resulting Markov chain be M_{sq2} . The person is again assumed to start in position 1.

- (1) What is the transition matrix of M_{sq2} ? Also draw the transition graph of M_{sq2} . What is the initial distribution of M_{sq2} ?
- (2) Show by induction that for all time instants $n > 0$, the distribution $\mu^{(n)}$ of M_{sq2} is $[\frac{1}{4} + \frac{1}{2^{n+1}}, \frac{1}{4}, \frac{1}{4} - \frac{1}{2^{n+1}}, \frac{1}{4}]$.
- (3) What is $\lim_{n \rightarrow \infty} \mu^{(n)}$?

2.1. Irreducible and aperiodic Markov chains

We identify some special kinds of Markov chains.

A Markov chain M is said to be irreducible if the transition graph of M is *strongly connected*, that is, for each pair of nodes n_1 and n_2 in the transition graph of M there is a directed path from n_1 to n_2 . Formally,

Definition: Let $Q = \{q_1, \dots, q_k\}$ be a finite set and let $M = (Q, \mu^{(0)}, \delta)$ be a Markov chain. M is said to be *irreducible* if for each $1 \leq i, j \leq k$, there is a $n_{i,j} > 0$ such that $(\delta^{n_{i,j}})_{i,j} > 0$. A Markov chain is said to be *reducible* if it is not irreducible.

EXERCISE 2.5. *Is the Markov chain M_{sq2} in Example 2.4 irreducible?*

Another property of Markov chains that we will be interested in is *aperiodicity*. Informally, a state q of a Markov chain is said to be aperiodic if the greatest common divisors of all the length of cycles which contain q is 1. Otherwise q is said to be periodic. A Markov chain is aperiodic iff all of its states are aperiodic, otherwise it is said to be periodic.

Definition: Let $Q = \{q_1, \dots, q_k\}$ be a finite set and let $M = (Q, \mu^{(0)}, \delta)$ be a Markov chain. For a state q_i , the period $d(q_i)$ is defined to be $\gcd(\{n \mid n > 0, (\delta^n)_{i,i} > 0\})$. The state q_i is said to be *aperiodic* if $d(q_i) = 1$; otherwise q_i is said to be *periodic*.

M is said to be *aperiodic* if all states of M are aperiodic; otherwise M is said to be *periodic*.

EXERCISE 2.6. *Is the Markov chain M_{sq2} in Example 2.4 aperiodic?*

We have the following result.

THEOREM 2.7. *Let $Q = \{q_1, \dots, q_k\}$ be a finite set and let $M = (Q, \mu^{(0)}, \delta)$ be a Markov chain. If M is irreducible and aperiodic, then there is a N such that*

$$\forall n \geq N, \forall 1 \leq i, j \leq k. (\delta^n)_{i,j} > 0.$$

EXERCISE 2.8. *We sketch the proof of Theorem 2.7. We use the following fact from elementary number theory.*

Fact: *A set $S = \{n_1, n_2, \dots\}$ of natural numbers is said to be closed under addition if $\ell + k \in S$ whenever ℓ and k are in S . If $S = \{n_1, n_2, \dots\}$ is closed under addition and $\text{g.c.d.}(n_1, n_2, \dots) = 1$, then there is a M such that for all $n \geq M$, $n \in S$.*

- (1) *Show using the above fact that if δ is the transition matrix of an aperiodic Markov chain on $Q = \{q_1, \dots, q_k\}$, then there is a natural number M such that for all $n \geq M$ and for all $1 \leq i \leq k$ $(\delta^n)_{i,i} > 0$.*
- (2) *Using Part 1 of the exercise, prove Theorem 2.7.*

EXERCISE 2.9. *Let $Q = \{q_1, \dots, q_k\}$ be a finite set and $M = (Q, \mu^{(0)}, \delta)$ be an irreducible Markov chain. Show that if there is an i such that $\delta_{i,i} > 0$ then M is aperiodic also.*

2.2. Stationary distribution

Consider the example of the person walking randomly on a square in Example 2.2 modeled as the Markov chain M_{sq} . As we saw in Exercise 2.3 the distribution $\mu^{(n)}$ keeps on fluctuating and the limit $\lim_{n \rightarrow \infty} \mu^{(n)}$ does not exist. In contrast, the modification of the walk on square M_{sq2} we encountered in Example 2.4 does have the nice property that as $n \rightarrow \infty$ the distributions $\mu^{(n)}$ tend to the distribution $\mu_{\text{stat}} = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$. It is easy to show that μ_{stat} enjoys two more properties.

EXERCISE 2.10. Consider the Markov chain $M_{sq2} = (Q, \mu^{(0)}, \delta)$ in Exercise 2.4. Show that

- (1) $\mu_{\text{stat}} = \mu_{\text{stat}}\delta$.
- (2) If μ is any distribution on Q such that $\mu = \mu\delta$ then $\mu = \mu_{\text{stat}}$.

In fact, we will show that there for any aperiodic and irreducible Markov chain $M = (Q, \mu^{(0)}, \delta)$, there is a *unique* distribution μ such that $\mu = \mu\delta$. Furthermore, we show that it is the case $\lim_{n \rightarrow \infty} \mu^{(n)} = \mu$. Note that aperiodicity and irreducibility are properties of δ , and so even if the Markov chain were to start in a different distribution, say $\mu'^{(0)}$, we would still have $\lim_{n \rightarrow \infty} \mu'^{(n)} = \mu$.

Before, we prove this result, we state one useful definition.

Definition: Let $M = (Q, \mu^{(0)}, \delta)$ be a Markov chain. A distribution μ_{stat} on Q is said to be a stationary distribution for M if $\mu_{\text{stat}} = \mu_{\text{stat}}\delta$.

EXERCISE 2.11. Consider the Markov chain M_{sq} in Example 2.2. Does M_{sq} have a stationary distribution? If yes, how many? Can you list all of them?

It turns out that for Markov chains on finite states always have stationary distributions (they may not be unique). If the distributions of a Markov chain converge to a distribution then it is easy to see that the limit is a stationary distribution. However, Markov chains may not converge (even when Markov chains have unique stationary distribution).

EXERCISE 2.12. Show that if a Markov Chain has two stationary distributions, then it has an infinite number of stationary distributions.

EXERCISE 2.13. Given a Markov chain $M = \{Q, \mu^{(0)}, \delta\}$ on the state space $Q = \{q_1, q_2, \dots, q_k\}$, a distribution μ on Q is said to be reversible if for each $1 \leq i, j \leq k$ $\mu_i \delta_{i,j} = \mu_j \delta_{j,i}$. A Markov chain M is said to be reversible if M has a reversible distribution.

Show that if μ is a reversible distribution for the Markov chain M , then μ is a stationary distribution.

EXERCISE 2.14. If $\mu^{(n)}$ are the distributions at time n of a Markov chain M such that $\lim_{n \rightarrow \infty} \mu^{(n)}$ exists then show that $\lim_{n \rightarrow \infty} \mu^{(n)}$ is a stationary distribution of the Markov chain M .

2.3. Existence of unique limiting stationary distributions

We will show a remarkable result: for every aperiodic and irreducible Markov chain, the distribution at time n of the Markov chain always converges as time goes to infinity. Furthermore the limit is independent of the initial distribution. Also, note that Exercise 2.14 implies that the limit is a stationary distribution of the Markov chain.

We start by defining the distance between two distributions.

Definition: Given a finite set Q and two distributions $\mu, \nu : Q \rightarrow [0, 1]$, the *distance* between μ and ν is denoted by $d(\mu, \nu)$ and given by the formula

$$d(\mu, \nu) = \sum_{q \in Q} \frac{|\mu(q) - \nu(q)|}{2}.$$

We have the following properties of the distance.

EXERCISE 2.15. *Let Q be a set of k elements. For every distribution μ, ν, ρ on Q we have the following.*

- (1) $d(\nu, \mu) = d(\mu, \nu) \geq 0$ and $(\mu, \nu) = 0$ iff $\mu = \nu$.
- (2) $d(\mu, \nu) \leq d(\mu, \rho) + d(\rho, \nu)$.
- (3) $d(\mu, \nu) \leq 1$.
- (4) *If a $k \times k$ matrix δ is a transition matrix, i.e., for each $1 \leq i \leq k$, $\sum_{j=1}^k \delta(q_i, q_j) = 1$ then*

$$d(\mu\delta, \nu\delta) \leq d(\mu, \nu).$$

Let $Q = \{q_1, \dots, q_k\}$ be a finite set and let $M = (Q, \mu^{(0)}, \delta)$ be a Markov chain. If M is irreducible and aperiodic, then recall (see Theorem 2.7) that there is a N such that

$$\forall 1 \leq i, j \leq k. (\delta^n)_{i,j} > 0.$$

LEMMA 2.16. *Given an irreducible and aperiodic Markov Chain $(Q, \mu^{(0)}, \delta)$, let Q have k elements and let $N > 0$ be such that $\delta_{i,j}^N > 0$ for each $1 \leq i, j \leq k$. There exists $0 \leq \alpha < 1$ such that for every pair of distributions μ, ν on Q ,*

$$d(\mu\delta^N, \nu\delta^N) < \alpha d(\mu, \nu).$$

PROOF. Let $\varepsilon = \min_{i,j} \delta_{i,j}$. Note that $0 < \varepsilon \leq \frac{1}{k}$. We have the following computation.

$$\begin{aligned}
2d(\mu\delta^N, \nu\delta^N) &= \sum_{j=1}^k |\mu\delta^N(j) - \nu\delta^N(j)| \\
&= \sum_{j=1}^k \left| \sum_{i=1}^k (\mu(i)\delta_{i,j}^N - \nu(i)\delta_{i,j}^N) \right| \\
&= \sum_{j=1}^k \left| \sum_{i=1}^k (\mu(i)(\delta_{i,j}^N - \varepsilon + \varepsilon) - \nu(i)(\delta_{i,j}^N - \varepsilon + \varepsilon)) \right| \\
&= \sum_{j=1}^k \left| \sum_{i=1}^k (\mu(i)(\delta_{i,j}^N - \varepsilon) - \nu(i)(\delta_{i,j}^N - \varepsilon)) + \right. \\
&\quad \left. \varepsilon(\sum_{i=1}^k \mu(i) - \sum_{i=1}^k \nu(i)) \right| \\
&= \sum_{j=1}^k \left| \sum_{i=1}^k (\mu(i)(\delta_{i,j}^N - \varepsilon) - \nu(i)(\delta_{i,j}^N - \varepsilon)) + \varepsilon(1 - 1) \right| \\
&\leq \sum_{j=1}^k \sum_{i=1}^k |\mu(i)(\delta_{i,j}^N - \varepsilon) - \nu(i)(\delta_{i,j}^N - \varepsilon)| \\
&\leq \sum_{j=1}^k \sum_{i=1}^k |\mu(i) - \nu(i)|(\delta_{i,j}^N - \varepsilon) \\
&\leq \sum_{i=1}^k \sum_{j=1}^k |\mu(i) - \nu(i)|(\delta_{i,j}^N - \varepsilon) \\
&\leq \sum_{i=1}^k (|\mu(i) - \nu(i)| \sum_{j=1}^k (\delta_{i,j}^N - \varepsilon)) \\
&\leq \sum_{i=1}^k (|\mu(i) - \nu(i)| (1 - k\varepsilon)) \\
&\leq 2(1 - k\varepsilon)d(\mu, \nu).
\end{aligned}$$

The Lemma follows by letting $\alpha = 1 - k\varepsilon$. □

THEOREM 2.17. *Given an irreducible and aperiodic Markov Chain (Q, μ, δ) , the limit*

$$\lim_{n \rightarrow \infty} \mu\delta^n$$

exists. The following properties are true of the limit $\mu_{\text{stat}} = \lim_{n \rightarrow \infty} \mu\delta^n$.

- (1) μ_{stat} is a stationary distribution of the Markov chain.
- (2) If ν is also a stationary distribution of the Markov chain then $\mu_{\text{stat}} = \nu$.
- (3) For any other distribution μ' , we have $\lim_{n \rightarrow \infty} \mu'\delta^n = \mu_{\text{stat}}$.

PROOF. Let $N > 0$ and $0 \leq \alpha < 1$ be such that $d(\nu\delta^N, \rho\delta^N) < \alpha d(\nu, \rho)$ for every pair of distributions ν and ρ on Q . Note that it suffices to show that

$$\lim_{i,j \rightarrow \infty} d(\mu\delta^i, \mu\delta^j) = 0.$$

Without loss of generality, let

$$\begin{aligned}
i &= m_i N + r_i & \text{where } m_i \geq 0 \text{ and } 0 \leq r_i \leq N - 1 \\
j &= m_j N + r_j & \text{where } m_j \geq 0 \text{ and } 0 \leq r_j \leq N - 1.
\end{aligned}$$

Furthermore, w.l.o.g., let $j \geq i$, i.e. $m_j \geq m_i$.

$$\begin{aligned}
d(\mu\delta^i, \nu\delta^j) &= d((\mu\delta^{r_i})\delta^{m_i N}, (\mu\delta^{r_j})\delta^{m_j N}) \\
&\leq d((\mu\delta^{r_i})\delta^{m_i N}, (\mu\delta^{r_j})\delta^{(m_i+1)N}) + \\
&\quad d((\mu\delta^{r_i})\delta^{(m_i+1)N}, (\mu\delta^{r_j})\delta^{(m_i+2)N}) + \\
&\quad \cdots + \\
&\quad d((\mu\delta^{r_i})\delta^{(m_j-1)N}, (\mu\delta^{r_j})\delta^{(m_j)N}) \\
&\leq \alpha^{m_i} d(\mu\delta^{r_i}, (\mu\delta^{r_j})\delta^N) + \\
&\quad \alpha^{m_i+1} d(\mu\delta^{r_i}, (\mu\delta^{r_j})\delta^N) + \\
&\quad \cdots + \\
&\quad \alpha^{m_j-1} d(\mu\delta^{r_i}, (\mu\delta^{r_j})\delta^N)
\end{aligned}$$

Let β be the $\max_{0 \leq m, n \leq N-1} d(\mu\delta^m, \mu\delta^n \delta^N)$. We get therefore

$$\begin{aligned}
d(\mu\delta^i, \mu\delta^j) &\leq \alpha^{m_i} \beta + \alpha^{m_i+1} \beta + \cdots + \alpha^{m_j-1} \beta \\
&\leq \alpha^{m_i} \beta + \alpha^{m_i+1} \beta + \cdots + \alpha^{m_j-1} \beta + \alpha^{m_j} \beta + \alpha^{m_j+1} \beta + \cdots \\
&\leq \beta \frac{\alpha^{m_i}}{1-\alpha}
\end{aligned}$$

Note that if i goes to infinity, m_i also goes to infinity. Since $0 \leq \alpha < 1$, α^{m_i} goes to 0. Therefore

$$\lim_{i, j \rightarrow \infty} d(\mu\delta^i, \mu\delta^j) = 0.$$

Thanks the limit $\mu_{\text{stat}} = \lim_{n \rightarrow \infty} \mu\delta^n$ exists. The desired properties of μ_{stat} can be proved as follows.

- (1) Follows from Exercise 2.14,
- (2) If ν is also a stationary distribution then we $\nu\delta = \nu$. We also have that $d(\mu_{\text{stat}}\delta^N, \nu\delta^N) \leq \alpha d(\mu_{\text{stat}}, \nu)$. As ν and μ_{stat} are stationary distributions, $\nu\delta^N = \nu$ and $\mu_{\text{stat}}\delta^N = \mu_{\text{stat}}$. Therefore $d(\mu_{\text{stat}}, \nu) < \alpha d(\mu_{\text{stat}}, \nu)$. Since $\alpha < 1$, the only way for that to happen is to have $d(\mu_{\text{stat}}, \nu) = 0$.
- (3) Note that δ is irreducible and aperiodic. Hence, $\lim_{n \rightarrow \infty} \mu\delta^n$ exists and is a stationary distribution. Thanks to property 2, we have that this limit must be μ_{stat} . \square

EXAMPLE 2.18. *We give a very simplified version of the PageRank algorithm employed by search algorithms (for example, by Google). The algorithm assumes that the internet consists of some webpages which have hyperlinks to other webpages. The person browsing the internet decides (probabilistically) whether to click on one of these links or visit a new page by entering it in the address bar.*

Assuming that there are N webpages in the world, the PageRank algorithm creates a Markov chain M with N states follows. Let the webpages be named p_1, p_2, \dots, p_N . Then the set $\{p_1, p_2, \dots, p_N\}$ are the states of M . Now the transition function δ of M is defined as follows. If the webpage p_i has links to every other page then $\delta_{i,j} = \frac{1}{N-1}$. If p_i has links to $N' < N-1$ webpages then $\delta_{i,j} = \frac{0.85}{N'}$ if p_i has a link to page p_j and $\delta_{i,j} = \frac{0.15}{N-N'+1}$ if p_i does not have a link to page p_j . The initial distribution of M assigns probability $\frac{1}{N}$ to each

of the states. The Markov chain is easily seen to be aperiodic and irreducible. Therefore, it must have a unique stationary distribution to which the Markov chain converges to irrespective of the starting distribution. The PageRank algorithm computes this distribution and the pages with higher probability in the distribution get ranked higher in the search by the PageRank algorithm.

2.4. Computing Reachability Properties

We have seen examples of Markov chains that do not converge to a stationary distribution. Sometimes, even if Markov chains do not converge to a stationary distribution, there might be states q of the Markov chain such that the probability of being in state q at time n converges as n goes to ∞ . One special case is if q is an absorbing state, namely a state from where you cannot escape to another state.

Definition: Given $Q = \{q_1, q_2, \dots, q_k\}$ and a Markov chain $\mathcal{M} = (Q, \mu^{(0)}, \delta)$, a state q_{i_0} is said to *absorbing* if $\delta_{i_0, i_0} = 1$.

PROPOSITION 2.19. *If the state q_{i_0} of a Markov chain $\mathcal{M} = (Q, \mu^{(0)}, \delta)$ on $Q = \{q_1, q_2, \dots, q_k\}$ is absorbing then for each $1 \leq j \leq k$, the sequence $\{(\delta^n)_{j, i_0} \mid n \in \mathbb{N}\}$ is a non-decreasing sequence. Therefore, $\lim_{n \rightarrow \infty} \delta_{j, i_0}^n$ exists. Furthermore, $\lim_{n \rightarrow \infty} \mu_{i_0}^{(n)}$ also exists.*

PROOF. Note that

$$\begin{aligned} (\delta^{n+1})_{j, i_0} &= \sum_{1 \leq \ell \leq k} (\delta^n)_{j, \ell} \delta_{\ell, i_0} \\ &= (\delta^n)_{j, i_0} \delta_{i_0, i_0} + \sum_{1 \leq \ell \leq k, \ell \neq i_0} (\delta^n)_{j, \ell} \delta_{\ell, i_0} \\ &\geq (\delta^n)_{j, i_0}. \end{aligned}$$

The proposition now follows from observing $\mu_{i_0}^{(n)} = \sum_{1 \leq \ell \leq k} (\mu_{\ell}^{(0)}) (\delta^n)_{\ell, i_0}$. \square

We will be interested in calculating the limit $\lim_{n \rightarrow \infty} \mu_{i_0}^{(n)}$ for a Markov chain if i_0 is an absorbing state. This limit is called the probability of reaching i_0 .

Definition: Given $Q = \{q_1, q_2, \dots, q_k\}$, a Markov chain $\mathcal{M} = (Q, \mu^{(0)}, \delta)$, an absorbing state q_{i_0} , and $1 \leq j \leq k$, the quantity $\lim_{n \rightarrow \infty} (\delta^n)_{j, i_0}$ is said to be the *probability* of reaching i_0 having started in j , and the quantity $\lim_{n \rightarrow \infty} (\mu^{(n)})_{i_0}$ is said to be the probability of reaching i_0 .

In order to compute the probability of reaching a state, we need the following definition.

Definition: Given a Markov chain $\mathcal{M} = (Q, \mu^{(0)}, \delta)$ on $Q = \{q_1, q_2, \dots, q_k\}$ and $1 \leq i, j \leq k$, we say that $i \mapsto j$ if there is a path from i to j in the transition graph of δ . If there is no path from i to j , we say that $i \not\mapsto j$.

If q_{i_0} is an absorbing state of \mathcal{M} , we say that $Bad_{i_0} = \{j \mid j \not\mapsto i_0\}$, $Good_{i_0} = \{j \mid j \mapsto i_0\}$ and $VeryGood_{i_0} = \{j \mid \forall \ell \in Bad_{i_0}, j \not\mapsto \ell\}$.

Intuitively, bad states are states from which one cannot reach the absorbing state, and good states are all those states that are not bad. Very good states are those good states from which one cannot reach bad states.

EXERCISE 2.20. *If the state q_{i_0} of a Markov chain $\mathcal{M} = (Q, \mu^{(0)}, \delta)$ on $Q = \{q_1, q_2, \dots, q_k\}$ is absorbing, then for each $j \in \text{Bad}_{i_0}$, $\lim_{n \rightarrow \infty} (\delta^n)_{j, i_0} = 0$.*

The following lemma states that from a good state, the probability of reaching a state that is neither a bad state nor the absorbing state goes to 0 as n goes to ∞ .

LEMMA 2.21. *If the state q_{i_0} of a Markov chain $\mathcal{M} = (Q, \mu^{(0)}, \delta)$ on $Q = \{q_1, q_2, \dots, q_k\}$ is absorbing, then for each $j \in \text{Good}_{i_0}$,*

$$\lim_{n \rightarrow \infty} \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^n)_{j, m} = 0.$$

Thus, for each $m \in \text{Good}_{i_0} \setminus \{i_0\}$, $\lim_{n \rightarrow \infty} (\delta^n)_{j, m} = 0$.

PROOF. Given $j \in \text{Good}_{i_0}$, fix a path p in the transition graph from q_j to q_{i_0} in the transition graph of δ (one always exists thanks to definition of Good_{i_0}). Let the length of this path be n_j and the weight be w_j (weight of a path is the product of the probabilities in the path). Now let $N = \max\{n_j \mid j \in \text{Good}_{i_0}\}$ and $w = \min\{w_j \mid j \in \text{Good}_{i_0}\}$. Clearly $w > 0$. It is easy to see that the following hold true.

(1) For each $\ell \in \text{Good}_{i_0}$, $(\delta^N)_{\ell, i_0} \geq w$. Therefore,

$$\sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^N)_{\ell, m} \leq 1 - w.$$

(2) For each $m \in \text{Good}_{i_0}$ and $\ell \in \text{Bad}_{i_0}$, $\delta_{\ell, m}^N = 0$.

We are now ready to prove the lemma. Consider the sum $\sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^{N+1})_{j, m}$.

We have

$$\begin{aligned} \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^{N+1})_{j, m} &= \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} \left(\sum_{1 \leq \ell \leq k} \delta_{j, \ell} (\delta^N)_{\ell, m} \right) \\ &= \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} \left(\sum_{\ell \in \text{Bad}_{i_0}} \delta_{j, \ell} (\delta^N)_{\ell, m} + \sum_{\ell \in \text{Good}_{i_0}} \delta_{j, \ell} (\delta^N)_{\ell, m} \right) \\ &= \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} \left(0 + \delta_{j, i_0} (\delta^N)_{i_0, m} + \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j, \ell} (\delta^N)_{\ell, m} \right) \\ &= \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} \left(0 + 0 + \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j, \ell} (\delta^N)_{\ell, m} \right) \\ &= \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j, \ell} (\delta^N)_{\ell, m} \\ &= \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j, \ell} (\delta^N)_{\ell, m} \\ &= \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j, \ell} \left(\sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^N)_{\ell, m} \right) \\ &\leq \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j, \ell} (1 - w) \\ &\leq (1 - w) \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j, \ell}. \end{aligned}$$

Similarly, we can show that

$$\sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^{2N+1})_{j,m} \leq (1-w) \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^{N+1})_{j,m} \leq (1-w)^2 \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j,\ell}.$$

In general, we will have by induction,

$$\sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^{rN+1})_{j,m} \leq (1-w)^r \sum_{\ell \in \text{Good}_{i_0} \setminus \{i_0\}} \delta_{j,\ell}.$$

Therefore,

$$\lim_{r \rightarrow \infty} \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^r N + 1)_{j,m} = 0.$$

Simialrly, we can show that for all $0 \leq s \leq N-1$,

$$\lim_{r \rightarrow \infty} \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^r N + s)_{j,m} = 0.$$

Hence,

$$\lim_{n \rightarrow \infty} \sum_{m \in \text{Good}_{i_0} \setminus \{i_0\}} (\delta^n)_{j,m} = 0.$$

□

EXERCISE 2.22. If the state q_{i_0} of a Markov chain $\mathcal{M} = (Q, \mu^{(0)}, \delta)$ on $Q = \{q_1, q_2, \dots, q_k\}$ is absorbing, then for each $j \in \text{VeryGood}_{i_0}$, $\lim_{n \rightarrow \infty} (\delta^n)_{j,i_0} = 1$.

We are ready to show the main theorem of this section which allows us to compute the probability of reaching an absorbing state.

THEOREM 2.23. Let q_{i_0} be an absorbing state of a Markov chain $\mathcal{M} = (Q, \mu^{(0)}, \delta)$ on the set of states $Q = \{q_1, q_2, \dots, q_k\}$. Let $\{x_j, 1 \leq j \leq k\}$ be a set of k -variables. Then, $\{\lim_{n \rightarrow \infty} \delta_{j,i_0}\}$ is the unique set of solutions of the following system, S , of k equations.

- $x_j = 0$ if $j \in \text{Bad}_{i_0}$.
- $x_j = 1$ if $j \in \text{VeryGood}_{i_0}$.
- $x_j = \sum_{\ell \in \text{Good}_{i_0} \setminus \text{VeryGood}_{i_0}} \delta_{j,\ell} x_\ell + \sum_{\ell \in \text{VeryGood}_{i_0}} \delta_{j,\ell}$ if $j \in \text{Good}_{i_0} \setminus \text{VeryGood}_{i_0}$.

PROOF. We will show that $\{\lim_{n \rightarrow \infty} \delta_{j,i_0}\}$ satisfies the set of equations. Note that for $j \in \text{Bad}_{i_0}$ and $j \in \text{VeryGood}_{i_0}$, the result follows from Exercise 2.20 and Exercise 2.22. Now, consider the case $j \in \text{Good}_{i_0} \setminus \text{VeryGood}_{i_0}$. Observe that we have

$$\begin{aligned} (\delta^{n+1})_{j,i_0} &= \sum_{1 \leq \ell \leq k} \delta_{j,\ell} (\delta^n)_{\ell,i_0} \\ &= \sum_{\ell \in \text{Good}_{i_0}} \delta_{j,\ell} (\delta^n)_{\ell,i_0} + \sum_{\ell \in \text{Bad}_{i_0}} \delta_{j,\ell} (\delta^n)_{\ell,i_0} \\ &= \sum_{\ell \in \text{Good}_{i_0}} \delta_{j,\ell} (\delta^n)_{\ell,i_0} + 0. \end{aligned}$$

The result now follows from taking the limit $n \rightarrow \infty$ on both sides. We now show the uniqueness of the solution.

We only need to show that $\{\lim_{n \rightarrow \infty} (\delta^n)_{j,i_0} \mid j \in \text{Good}_{i_0} \setminus \text{VeryGood}_{i_0}\}$ is the unique solution of the set of equations:

$$x_j = \sum_{\ell \in \text{Good}_{i_0} \setminus \text{VeryGood}_{i_0}} \delta_{j,\ell} x_\ell + \sum_{\ell \in \text{VeryGood}_{i_0}} \delta_{j,\ell}.$$

Let r be the number of elements of $\text{Good}_{i_0} \setminus \text{VeryGood}_{i_0}$. If $r = 0$, then we have nothing to prove. Without loss of generality, we can assume that the set $\text{Good}_{i_0} \setminus \text{VeryGood}_{i_0}$ is $\{q_1, q_2, \dots, q_r\}$.

Now consider the matrix $(r+1) \times (r+1)$ matrix:

$$\delta' = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \dots & \delta_{1,r} & \sum_{\ell \in \text{VeryGood}_{i_0}} \delta_{1,\ell} \\ \delta_{2,1} & \delta_{2,2} & \dots & \delta_{2,r} & \sum_{\ell \in \text{VeryGood}_{i_0}} \delta_{2,\ell} \\ \vdots & & & & \\ \delta_{r,1} & \delta_{r,2} & \dots & \delta_{r,r} & \sum_{\ell \in \text{VeryGood}_{i_0}} \delta_{r,\ell} \\ 0 & 0 & & 0 & 1 \end{pmatrix}.$$

Note that for each $1 \leq j \leq r$, $\sum_{1 \leq \ell \leq r+1} ((\delta')^n)_{j,\ell}$ is the sum of all weights of paths of length n that start in j and end in a good state. Therefore, $1 - \sum_{1 \leq \ell \leq r+1} ((\delta')^n)_{j,\ell}$ is the sum of all weights of paths of length n that start in ℓ and end in a bad state. Now, there must exist an N such for each $1 \leq j \leq r$, $1 - \sum_{1 \leq \ell \leq r+1} ((\delta')^N)_{j,\ell} > 0$ (Why?) Fix one such N and let $w = \min\{1 - \sum_{1 \leq \ell \leq r+1} ((\delta')^N)_{j,\ell} \mid 1 \leq j \leq r\}$.

Now if $\{y_j \mid 1 \leq j \leq r\}$ and $\{z_i \mid 1 \leq i \leq r+1\}$ are solutions to the system of equations S , then consider the $(r+1)$ -column vectors,

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \\ 1 \end{pmatrix} \quad \text{and} \quad \vec{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_r \\ 1 \end{pmatrix}.$$

Now clearly, we have that $\delta' \vec{y} = \vec{y}$ and $\delta' \vec{z} = \vec{z}$. This implies that $(\delta')^N \vec{y} = \vec{y}$ and $(\delta')^N \vec{z} = \vec{z}$.

Theorefore, $\vec{y} - \vec{z} = (\delta')^N (\vec{y} - \vec{z})$. Hence, for each $1 \leq j \leq r$, we have

$$y_j - z_j = \sum_{1 \leq \ell \leq r} ((\delta')^N)_{j,\ell} (y_\ell - z_\ell).$$

Taking absolute values, we get

$$\begin{aligned}
|y_j - z_j| &= \left| \sum_{1 \leq \ell \leq r} ((\delta')^N)_{j,\ell} (y_\ell - z_\ell) \right| \\
&\leq \sum_{1 \leq \ell \leq r} |((\delta')^N)_{j,\ell} (y_\ell - z_\ell)| \\
&\leq (\max_{\{1 \leq \ell \leq r\}} |y_\ell - z_\ell|) \sum_{1 \leq \ell \leq r} ((\delta')^N)_{j,\ell} \\
&\leq (1 - w) (\max_{\{1 \leq \ell \leq r\}} |y_\ell - z_\ell|).
\end{aligned}$$

Since the inequality holds for each j , we have that

$$\max_{\{1 \leq \ell \leq r\}} |y_\ell - z_\ell| \leq (1 - w) (\max_{\{1 \leq \ell \leq r\}} |y_\ell - z_\ell|).$$

Since $w > 0$, the only way for the equality to hold is to require that $\max_{\{1 \leq \ell \leq r\}} |y_\ell - z_\ell| = 0$, from which the uniqueness follows. \square

EXAMPLE 2.24. *Alice has the option of either joining swimming lessons or taking judo lessons. She prefers judo slightly, but she hears that the swimming teacher is an Olympic champion. Unable to make up her mind, Alice decides to toss coins. She tosses two fair coins. If at most one coin turns up heads, she joins judo lessons. If both coins turn up heads, she joins swimming lessons, and if neither is the case she starts over again.*

This can be modeled as a Markov chain M with three states $\{q_0, q_1, q_2\}$ with the following transition matrix.

$$\begin{pmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here q_1 is the state that models that Alice decides to do judo and q_2 is the state that models that Alice joins swimming lessons. Note that q_1 and q_2 are both absorbing states. M is aperiodic but not irreducible.

Now, what is the probability that Alice chooses judo lessons. In order to compute this, we first compute $\lim_{n \rightarrow \infty} \delta_{i1}^n$. Note that $\text{Good}_1 = \{q_0, q_1\}$, $\text{Bad}_1 = \{q_2\}$ and $\text{VeryGood}_1 = \{q_1\}$.

Therefore $\lim_{n \rightarrow \infty} \delta_{11}^n = 1$ and $\lim_{n \rightarrow \infty} \delta_{21}^n = 0$. And $\lim_{n \rightarrow \infty} \delta_{01}^n = 1$ is the solution to the equation

$$x = \frac{1}{4}x + \frac{1}{2}.$$

That is $\lim_{n \rightarrow \infty} \delta_{01} = \frac{2}{3}$.

If we make the reasonable assumption that the initial distribution is $[1, 0, 0]$, we can see that the probability of Alice taking judo lessons is $1 \times \frac{2}{3} + 0 \times 1 + 0 \times 0 = \frac{2}{3}$.

EXERCISE 2.25. *We now discuss a synchronous consensus protocol. Assume that there are N processes trying to agree upon a leader. Let the processes be p_1, p_2, \dots, p_N . The protocol proceeds in rounds. In each round, each process p_i tosses a N -faced fair dice and if the result is j , then p_i declares j to*

be its choice for the leader. If all processes agree on the choice then the agreed choice is declared to be the leader. If they disagree, then start again by tossing the coins. Let $X^{(n)}$ be the random variable that denotes the leader at beginning of round n (if there is no leader then $X^{(n)}$ takes the value 0.) Note that $X^{(0)}, X^{(1)}, \dots$ is a Markov chain. Let us call this Markov chain SM .

- (1) Give the transition matrix for the Markov chain SM .
- (2) If δ is the transition matrix of SM , does $\lim_{n \rightarrow \infty} \delta^n$ exists? If yes, please compute it. Does $\lim_{n \rightarrow \infty} X^{(n)}$ exists? If yes, please compute it.
- (3) What are stationary distributions of SM ?
- (4) What is the probability that a leader is elected?

CHAPTER 3

Probabilistic Automata

3.1. Defintions

Informally, a probabilistic finite automaton (PFA) is like a finite-state deterministic automaton except that the transition function from a state on a given input is described as a probability distribution which determines the probability of the next state.

Definition: A *probabilistic finite automaton* (PFA) over a finite alphabet Σ is a tuple $\mathcal{A} = (Q, q_s, Q_f, \delta)$ where Q is a finite set of *states*, $q_s \in Q$ is the *initial state*, $Q_f \subseteq Q$ is the set of *accepting/final states*, and $\delta : Q \times \Sigma \rightarrow \text{Dist}(Q)$ is the *transition relation*.

Intuitively, the PFA \mathcal{A} starts in the initial state q_s and if after reading $a_0, a_1 \dots, a_i$ results in state q , then it moves to state q' with probability $\delta(q, a_{i+1})(q')$ on symbol a_{i+1} .

Definition: Given a PFA \mathcal{A} on Σ and a word $u \in \Sigma^*$, the *probability space generated by \mathcal{A} and u* is the probability space $(\Omega_{\mathcal{A},u}, \mathcal{F}_{\mathcal{A},u}, \mu_{\mathcal{A},u})$ where $\Omega_{\mathcal{A},u}, \Sigma_{\mathcal{A},u}$ and $\mu_{\mathcal{A},u}$ are defined as follows.

- $\Omega_{\mathcal{A},u} = \{w \in Q^{n+1} \mid |u| = n \ \& \ w[0] = q_s\}$.
- $\mathcal{F} = 2^{\Omega_{\mathcal{A},u}}$.
- If $u = a_0 a_1 \dots a_{n-1}$, then $\mu_{\mathcal{A},u}$ is the unique probability measure such that for any $q_0 \dots q_n \in Q^{n+1}$,

$$\mu(\{q_0 q_1 \dots q_n\}) = \delta(q_0, a_0)(q_1) \delta(q_1, a_1)(q_2) \dots \delta(q_{n-1}, a_{n-1})(q_n).$$

EXAMPLE 3.1. Consider the automaton \mathcal{A} drawn in Figure 1. Note that here q is the initial state and also the final state. Now consider the word $U = aba$. The probability measure $\mu_{\mathcal{A},u}$ is such that –

$$\begin{aligned} \mu_{\mathcal{A},u}(qqqq) &= 0 & \mu_{\mathcal{A},u}(qqqs) &= 0 \\ \mu_{\mathcal{A},u}(qqsq) &= \frac{2}{9} & \mu_{\mathcal{A},u}(qqss) &= \frac{1}{9} \\ \mu_{\mathcal{A},u}(qssq) &= 0 & \mu_{\mathcal{A},u}(qqsq) &= 0 \\ \mu_{\mathcal{A},u}(qsqq) &= \frac{2}{9} & \mu_{\mathcal{A},u}(qsqs) &= \frac{4}{9}. \end{aligned}$$

Definition: Given a word $u \in \Sigma^*$ and a PFA $\mathcal{A} = (Q, q_s, Q_f, \delta)$ on Σ , let $|u| = n$ and $(\Omega_u, \Sigma_u, \mu_u)$ be the probability distribution generated by \mathcal{A} and u . Let *Acc* be the event $\{w \in \Omega_u \mid w[n] \in Q_f\}$. The *probability of \mathcal{A} accepting*

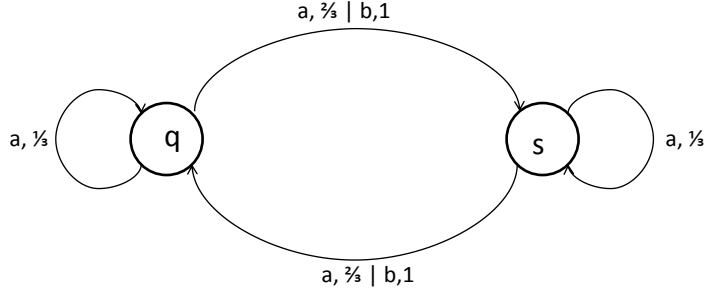


FIGURE 1. Automaton \mathcal{A}

u is $\mu_u(\text{Acc})$. We shall denote this probability by $\mu_{\mathcal{A},u}^{\text{acc}}$. The probability of rejecting u , denoted by $\mu_{\mathcal{A},u}^{\text{rej}}$, is $1 - \mu_{\mathcal{A},u}^{\text{acc}}$.

EXAMPLE 3.2. Let \mathcal{A} and u be as in Example 3.1. We see that $\mu_{\mathcal{A},u}^{\text{acc}} = \frac{4}{9}$.

Notation: The transition function δ of PFA \mathcal{A} on input a can be seen as a square matrix δ_a of order $|Q|$ with the rows labeled by “current” state, columns labeled by “next state” and the entry $\delta_a(q, q')$ equal to $\delta(q, a)(q')$. Given a word $u = a_0 a_1 \dots a_n \in \Sigma^+$, δ_u is the matrix product $\delta_{a_0} \delta_{a_1} \dots \delta_{a_n}$. For an empty word $\varepsilon \in \Sigma^*$ we take δ_ε to be the identity matrix. Finally for any $Q_0 \subseteq Q$, we say that $\delta_u(q, Q_0) = \sum_{q' \in Q_0} \delta_u(q, q')$.

Given a state $q \in Q$ and a word $u \in \Sigma^+$, $\text{post}(q, u) = \{q' \mid \delta_u(q, q') > 0\}$.

EXAMPLE 3.3. Let \mathcal{A} and u be as in Example 3.1. Note that

$$\delta_a = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix} \quad \delta_b = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

A simple matrix calculation gives us that

$$\delta_{aba} = \begin{pmatrix} \frac{4}{9} & \frac{5}{9} \\ \frac{5}{9} & \frac{4}{9} \end{pmatrix}.$$

EXERCISE 3.4. Given a PFA $\mathcal{A} = (Q, q_s, Q_f, \delta)$ on Σ and a word $u \in \Sigma^*$, show that

$$\mu_{\mathcal{A}, u}^{acc} = \delta_u(q_s, Q_f).$$

3.1.1. Languages. Unlike the definition of language accepted by a finite automaton, a PFA \mathcal{A} does not recognize a unique language. Instead, the notion of a language accepted by a PFA requires the notion of a *threshold* probability, also called a *cut-point*.

Definition: Given a *cut-point* $x \in [0, 1]$ and a PFA \mathcal{A} on alphabet Σ , we can define two *languages*:

- $\mathcal{L}_{>x}(\mathcal{A}) = \{\alpha \in \Sigma^\omega \mid \mu_{\mathcal{A}, \alpha}^{acc} > x\}$, and
- $\mathcal{L}_{\geq x}(\mathcal{A}) = \{\alpha \in \Sigma^\omega \mid \mu_{\mathcal{A}, \alpha}^{acc} \geq x\}$.

The language $\mathcal{L}_{>x}(\mathcal{A})$ is said to be a *strict cut-point language* while the language $\mathcal{L}_{\geq x}(\mathcal{A})$ is said to be a *non-strict cut-point language*.

One can of course also define $\mathcal{L}_{=x}(\mathcal{A})$. However, we will usually not concern ourselves with $\mathcal{L}_{=x}(\mathcal{A})$ except for the special case when $x = 1$ in which case $\mathcal{L}_{=x}(\mathcal{A})$ coincides with $\mathcal{L}_{\geq x}(\mathcal{A})$. We can also define $\mathcal{L}_{<x}(\mathcal{A})$ and $\mathcal{L}_{\leq x}(\mathcal{A})$, but we do not explicitly study them. This is because the following result holds.

EXERCISE 3.5. For any PFA \mathcal{A} and a *cut-point* $x \in [0, 1]$, there is a PFA \mathcal{B} and a $y \in [0, 1]$ such that $\mathcal{L}_{<x}(\mathcal{A}) = \mathcal{L}_{>y}(\mathcal{B})$ and $\mathcal{L}_{\leq x}(\mathcal{A}) = \mathcal{L}_{\geq y}(\mathcal{B})$.

The thresholds 0 and 1 are said to be *extremal cut-points* while $x \in (0, 1)$ is said to be a *non-extremal cut-point*. The following is an easy consequence of the definitions.

PROPOSITION 3.6. For any PFA \mathcal{A} , we have $\mathcal{L}_{\geq 0}(\mathcal{A}) = \Sigma^*$ and $\mathcal{L}_{>1}(\mathcal{A}) = \emptyset$.

Therefore, for extremal cut-points, the only interesting languages are $\mathcal{L}_{>0}(\mathcal{A})$ and $\mathcal{L}_{=1}(\mathcal{A})$. This is what we shall discuss next.

3.2. Extremal Cut-Points

We start with an example.

EXAMPLE 3.7. Let \mathcal{A} and u be as in Example 3.1. Observe that $\mathcal{L}_{>0}(\mathcal{A}) = \{a, b\}^* \setminus (b^2)^*b$.

The main result of this section is that the set of languages recognized with PFA with extremal cut-point coincides with the set of regular languages.

PROPOSITION 3.8. Given an alphabet Σ and a PFA \mathcal{A} on Σ , the language $\mathcal{L}_{>0}(\mathcal{A})$ is a regular language. For any regular language $L \subseteq \Sigma^*$, there is a PFA \mathcal{A} such that $L = \mathcal{L}_{>0}(\mathcal{A})$.

PROOF. Given a PFA $\mathcal{A} = (Q, q_s, Q_f, \delta)$ on Σ consider the *non-deterministic finite automaton* $\mathcal{B} = (Q, q_s, Q_f, \Delta)$ where $\Delta \subseteq Q \times \Sigma \times Q$ is defined as follows–

$$\forall q, q' \in Q, a \in \Sigma. ((q, a, q') \in \Delta \Leftrightarrow \delta(q, a, q') > 0).$$

The reader is invited to prove that $\mathcal{L}_{>0}(\mathcal{A}) = \mathcal{L}(\mathcal{B})$.

The other direction is left as exercise to the reader. \square

EXERCISE 3.9. *Given an alphabet Σ and a regular language $L \subseteq \Sigma^*$, there is a PFA \mathcal{A} such that $L = \mathcal{L}_{>0}(\mathcal{A})$.*

Please recall that checking emptiness of a finite automaton is decidable in polynomial time while checking universality is decidable in **PSPACE**. The proof of Proposition 3.8 then yields the following.

COROLLARY 3.10. *Given an alphabet Σ and a PFA \mathcal{A} on Σ , the problem of checking emptiness of $\mathcal{L}_{>0}(\mathcal{A})$ is decidable in polynomial time, while the problem of checking universality of $\mathcal{L}_{>0}(\mathcal{A})$ is decidable in **PSPACE**.*

The lower bound for checking universality of $\mathcal{L}_{>0}(\mathcal{A})$ also turns out to be **PSPACE**.

EXERCISE 3.11. *Given an alphabet Σ and a PFA \mathcal{A} on Σ , the problem of checking universality of $\mathcal{L}_{>0}(\mathcal{A})$ is **PSPACE**-complete.*

The corresponding results for the cut-point 1 is left to the reader.

EXERCISE 3.12. *Given an alphabet Σ and a PFA \mathcal{A} on Σ ,*

- (1) $\mathcal{L}_{=1}(\mathcal{A})$ is regular. For any regular language $L \subseteq \Sigma^*$, there is a PFA \mathcal{A} such that $L = \mathcal{L}_{=1}(\mathcal{A})$.
- (2) The problem of checking emptiness of $\mathcal{L}_{=1}(\mathcal{A})$ is **PSPACE**-complete.
- (3) The problem of checking universality of $\mathcal{L}_{=1}(\mathcal{A})$ is decidable in polynomial time.

We conclude this section by observing that probabilistic automata, like non-deterministic automata can be exponentially more succinct than the deterministic finite automata.

EXERCISE 3.13. *Let $\Sigma = \{a, b\}$ and n be a non-zero natural number. Consider the language $L_n = \Sigma^* \setminus (\Sigma^* a \Sigma^n a \Sigma^*)$. Now, L_n is regular and it can be shown that any deterministic finite automaton that recognizes L_n has $O(2^n)$ states. Construct a PFA \mathcal{A} such that the following hold–*

- \mathcal{A} has $O(n)$ -states.
- Every word in L_n is accepted by \mathcal{A} with probability 1 and every word not in L_n is accepted by \mathcal{A} with probability $\leq 1 - 1/2^n$.

3.3. Non-extremal cut-points and regular languages

We start by observing that we only need to consider the case when cut-point is $\frac{1}{2}$ in light of the following observation.

EXERCISE 3.14. *Given an alphabet Σ , a PFA \mathcal{A} on Σ , and two rational numbers $x, y \in (0, 1)$ there is a PFA \mathcal{B} such that $\mathcal{L}_{>x}(\mathcal{A}) = \mathcal{L}_{>y}(\mathcal{B})$ and $\mathcal{L}_{\geq x}(\mathcal{A}) = \mathcal{L}_{\geq y}(\mathcal{B})$.*

EXERCISE 3.15. *Let $\Sigma = \{\mathbf{0}, \mathbf{1}\}$. Consider the PFA $\mathcal{A} = (Q, q_s, Q_f, \delta)$ where the set $Q = (q_s, q_a, q_r)$, $Q_f = \{q_a\}$ and δ is defined as follows.*

- $\delta(q_s, \mathbf{0}, q_s) = \delta(q_s, \mathbf{0}, q_r) = \frac{1}{2}$.
- $\delta(q_s, \mathbf{1}, q_s) = \delta(q_s, \mathbf{1}, q_a) = \frac{1}{2}$.
- $\delta(q_a, \mathbf{0}, q_a) = \delta(q_a, \mathbf{1}, q_a) = 1$.
- $\delta(q_r, \mathbf{0}, q_r) = \delta(q_r, \mathbf{1}, q_r) = 1$.

(1) *Show that for any word $u = a_1 \cdots a_n \in \Sigma^*$. the probability of \mathcal{A} accepting u is*

$$\sum_{i=1}^n \frac{\text{num}(a_i)}{2^i}$$

where $\text{num}(a_i)$ is the number 0 if a_i is $\mathbf{0}$ and is number 1 one when a_i is 1.

(2) *Compute $\mathcal{L}_{>\frac{1}{2}}(\mathcal{A})$. Is it regular?*

We have seen that with extremal cut-points PFAs are as expressive as finite automata. This is no longer true for non-extremal cut-points. We shall demonstrate that PFAs can recognize non-regular languages with non-extremal cut-points. Unlike, the case of non-extremal cut-points, one can recognize non-regular languages with PFAs.

THEOREM 3.16. *There is a PFA \mathcal{A} such that $\mathcal{L}_{>\frac{1}{2}}(\mathcal{A})$ is not regular.*

PROOF. Let \mathcal{A} be the automata in Exercise 3.15. We show that the $\mathcal{L}_{>\frac{1}{\sqrt{2}}}(\mathcal{A})$ is not regular. We can then apply the construction from Exercise 3.18 to get an automata \mathcal{B} such that $\mathcal{L}_{>\frac{1}{2}}(\mathcal{B})$ is not rational. We have that for $u = a_1 \cdots a_n \in \Sigma^*$. the probability of \mathcal{A} accepting u is

$$\sum_{i=1}^n \frac{\text{num}(a_i)}{2^i}$$

where $\text{num}(a_i)$ is the number 0 if a_i is $\mathbf{0}$ and is number 1 one when a_i is 1. Thus, $\mathcal{L}_{>\frac{1}{\sqrt{2}}}(\mathcal{A})$ is the set of all words whose binary expansion is > 0 .

Let $L = \mathcal{L}_{>\frac{1}{\sqrt{2}}}(\mathcal{A})$. We show that L is not regular by the showing that the Myhill-Nerode index \equiv_L of L is infinite. Let $0.b_1b_2\cdots$ be the binary expansion of $\frac{1}{\sqrt{2}}$. Let

$$IsOne = \{i \in \mathbb{N} \mid b_i = 1\}.$$

We make the following observations.

- (1) *IsOne* is infinite.
- (2) If $i, j \in \text{IsOne}$ are such that $i \neq j$ then there exists $k > 0$ such that $b_{i+k} \neq b_{j+k}$ (see exercise 3.17).

Let $\bar{0}$ be $\mathbf{0}$ and $\bar{1}$ be $\mathbf{1}$. Now for $i \in \mathbb{N}$ let u_i be the word $\bar{b}_1 \cdots \bar{b}_i$. The theorem follows from the following claim.

Claim: For $i, j \in \text{IsOne}$ such that $i < j$, we have that $u_i \not\equiv_L u_j$.

PROOF OF THE CLAIM: Let k be the smallest natural number such that $b_{i+k} \neq b_{j+k}$. We thus have that for

$$1 \leq \ell < k, \quad b_{i+\ell} = b_{j+\ell}.$$

Assume first that $b_{i+k} = 0$ and that $b_{j+k} = 1$ (the other case is symmetric). Let $w = \bar{b}_{i+1} \cdots \bar{b}_{i+k-1} \mathbf{1}$. Now observe the following.

- (1) $0.b_1 \cdots b_i b_{i+1} \cdots b_{i+k-1} \mathbf{1} = 0.b_1 \cdots b_i b_{i+1} \cdots b_{i+k-1} 011111 \cdots > \frac{1}{\sqrt{2}}$.
- (2) $0.b_1 \cdots b_j b_{j+1} \cdots b_{j+k-1} \mathbf{1} = 0.b_1 \cdots b_j b_{j+1} \cdots b_{j+k-1} \mathbf{1} < \frac{1}{\sqrt{2}}$.

Thus $u_i w \in L$ but $u_j w \notin L$. Therefore $u_i \not\equiv_L u_j$. (END-PROOF OF THE CLAIM) \square

Remark: We have shown a non-regular language that is recognized by a PFA with a strict cut-point. One may ask if a non-regular language can be recognized by a PFA with non-strict cut-point. This is indeed true, and the construction of Theorem 3.16 works for this case as well.

EXERCISE 3.17. Show that if $0.b_1 b_2 \cdots$ is the binary expansion of $\frac{1}{\sqrt{2}}$, then for each $i, j \in \mathbb{N}$ with $i \neq j$ there exists a $k \geq 0$ such that $b_{i+k} \neq b_{j+k}$. (Hint: use the irrationality of $\frac{1}{\sqrt{2}}$).

EXERCISE 3.18. Given any PFA \mathcal{A} , show that there is another automata \mathcal{B} such that:

- (1) $\mathcal{L}_{>\frac{1}{2}}(\mathcal{B}) = \mathcal{L}_{>\frac{1}{\sqrt{2}}}(\mathcal{A})$.
- (2) In general for any $x \in [0, 1]$, $\mathcal{L}_{>x^2}(\mathcal{B}) = \mathcal{L}_{>x}(\mathcal{A})$.

3.3.1. Isolated cut-points. We have seen that PFAs with non-extremal cut-points can define non-regular languages. There is however, an important class of PFAs for which non-extremal cut-points yield only regular languages.

Definition: Given a PFA $\mathcal{A} = (Q, q_s, Q_f, \delta)$ on alphabet Σ , we say that \mathcal{A} is (x, ε) -robust if for each $u \in \Sigma^*$,

$$|\mu_{\mathcal{A}, u}^{acc} - x| > \varepsilon.$$

x is said to be *isolated cut-point* of \mathcal{A} if there an $\varepsilon > 0$ such that \mathcal{A} is (x, ε) -robust.

For any (x, ε) -robust \mathcal{A} , note that $\mathcal{L}_{>=x}(\mathcal{A}) = \mathcal{L}_{>x}(\mathcal{A})$. We shall prove the fact shortly that $\mathcal{L}_{>x}(\mathcal{A})$ is regular for an (x, ε) -robust PFA \mathcal{A} , but we first state a result that is needed in the proof.

EXERCISE 3.19. *Given $n \in \mathbb{N}, n > 0$, let $\mathcal{P}_n \subseteq \mathbb{R}^n$ be the set of vectors defined as $\{(\xi_1, \dots, \xi_n) \mid 0 \leq \xi_j \leq 1, \sum_{j=1}^n \xi_j = 1\}$. Given $\varepsilon \in \mathbb{R}, \varepsilon > 0$, let $\mathcal{U} \subset \mathcal{P}_n$ be a set such that for any $(\xi_1, \dots, \xi_n), (\xi'_1, \dots, \xi'_n) \in \mathcal{U}$, $\sum_j |\xi_j - \xi'_j| > \varepsilon$. Then \mathcal{U} must be finite.*

THEOREM 3.20. *For any (x, ε) -robust PFA $\mathcal{A} = (Q, q_s, Q_f, \delta)$ on alphabet Σ , the language $\mathcal{L}_{>x}(\mathcal{A})$ is a regular language.*

PROOF. We shall use the Myhill-Nerode theorem to prove the result. That is we show that the Myhill-Nerode index of the language $\mathcal{L}_{>x}(\mathcal{A})$ is finite if PFA \mathcal{A} is x -robust. Let $L = \mathcal{L}_{>x}(\mathcal{A})$ and consider the Myhill-Nerode equivalence relation \equiv_L .

Fix $u, v \in \Sigma^*$ such that $u \not\equiv_L v$. This implies that there is a word $w \in \Sigma^*$ such that either $uw \in L$ and $vw \notin L$ or $uw \notin L$ and $vw \in L$. Consider first the case such that $uw \in L$ and $vw \notin L$. Since $uw \in L$ we get that $\delta uw(q_s, Q_f) > x + \varepsilon$ and $\delta vw(q_s, Q_f) < x - \varepsilon$. Therefore, $\delta uw(q_s, Q_f) - \delta vw(q_s, Q_f) > 2\varepsilon$. Now

$$\begin{aligned} & \delta uw(q_s, Q_f) - \delta vw(q_s, Q_f) > 2\varepsilon \\ \Rightarrow & |\delta uw(q_s, Q_f) - \delta vw(q_s, Q_f)| > 2\varepsilon \\ \Rightarrow & |(\sum_{q \in Q} \delta_u(q_s, q) \delta_w(q, Q_f)) - (\sum_{q \in Q} \delta_v(q_s, q) \delta_w(q, Q_f))| > 2\varepsilon \\ \Rightarrow & \sum_{q \in Q} |\delta_w(q, Q_f) (\delta_u(q_s, q) - \delta_v(q_s, q))| > 2\varepsilon \\ \Rightarrow & \sum_{q \in Q} |\delta_w(q, Q_f)| |\delta_u(q_s, q) - \delta_v(q_s, q)| > 2\varepsilon \\ \Rightarrow & \sum_{q \in Q} |\delta_u(q_s, q) - \delta_v(q_s, q)| > 2\varepsilon. \end{aligned}$$

Similarly if $uw \notin L$ and $vw \in L$, we can show that $\sum_{q \in Q} |\delta_v(q_s, q) - \delta_u(q_s, q)| > 2\varepsilon$. Therefore if $u \not\equiv_L v$, then $\sum_{q \in Q} |\delta_v(q_s, q) - \delta_u(q_s, q)| > 2\varepsilon$. In light of Exercise 3.19, L has a finite Myhill-Nerode index. \square

EXERCISE 3.21. *Given a PFA \mathcal{A} on Σ , $x \in [0, 1]$ and an $\varepsilon > 0$ such that $|\mu_{\mathcal{A}, u}^{acc} - x| > \varepsilon$ for all $u \in \Sigma^*$. \mathcal{A} is a regular language thanks to Theorem 3.20. Give an upper bound on the number of states of the finite deterministic automaton recognizing the PFA \mathcal{A} .*

EXERCISE 3.22. *Consider the automaton \mathcal{A} in Exercise 3.13. Is $\frac{1}{2}$ an isolated cut-point of \mathcal{A} ?*

EXERCISE 3.23. *Consider the automaton \mathcal{A} in Exercise 3.15. Is $\frac{1}{2}$ an isolated cut-point of \mathcal{A} ?*

3.4. Undecidability of decision problems for threshold languages

We had seen earlier that checking emptiness (as well as universality) of the language $\mathcal{L}_{>0}(\mathcal{A})$ (as well as $\mathcal{L}_{=1}(\mathcal{A})$) for a PFA \mathcal{A} is decidable. In contrast the corresponding decision problems for $\mathcal{L}_{\geq \frac{1}{2}}(\mathcal{A})$ (as well as $\mathcal{L}_{> \frac{1}{2}}(\mathcal{A})$) turn

out to be undecidable. Due to lack of time, we will not go over the proof in the class. Therefore, you will not be required to know this proof for either your homework or the examinations, but knowledge of the result itself may be called upon in the homework or the examinations.

THEOREM 3.24. *Given a finite alphabet Σ and a PFA \mathcal{A} on Σ , the problem of checking emptiness of $\mathcal{L}_{>\frac{1}{2}}(\mathcal{A})$ is undecidable. More precisely, the set $\{(\Sigma, \mathcal{A}) \mid \mathcal{A} \text{ is a PFA on } \Sigma \text{ \& } \mathcal{L}_{\geq\frac{1}{2}}(\mathcal{A}) = \emptyset\}$ is **co-R.E.**-hard.*

CHAPTER 4

Basics of information theory

We shall briefly consider the topic of information theory. The central concept of information theory is *entropy*, which is a measurement of uncertainty of a random variable.

Definition: Let $\mathbf{X} : \omega \rightarrow Q$ be a Q -valued random variable, where Q is a finite set. The *entropy* of \mathbf{X} , denoted $H(\mathbf{X})$, is defined to be

$$H(\mathbf{X}) = - \sum_{q \in Q} \Pr(\mathbf{X} = q)(\log_2 \Pr(\mathbf{X} = q)).$$

Remark: We adopt the convention that $0 \log_2 0 = 0$. Note that $H(\mathbf{X})$ is the expected value of the random variable $\log_2(\Pr(\mathbf{X}))$. The unit of measurement of entropy is bits. It can be thought of as the average number of bits required to describe the random variable \mathbf{X} . One can also talk about entropy of a Q -valued random variable where Q is a countable set. In that case, we use the same expression as above.

It is easy to see that the following is true.

PROPOSITION 4.1. $H(\mathbf{X}) \geq 0$.

EXAMPLE 4.2. Let \mathbf{X} be a random variable which takes values in the set $\{\mathbf{0}, \mathbf{1}\}$. Let $\Pr(\mathbf{X} = \mathbf{0}) = p$. Therefore, $H(\mathbf{X}) = -p \log_2(p) + (1-p) \log_2(1-p)$. Now, $H(\mathbf{X})$ takes its maximum value at $p = \frac{1}{2}$ where it is exactly $\frac{1}{2}$, and takes the value 0 when $p = 0$ or $p = 1$.

EXAMPLE 4.3. Let \mathbf{X} be a random variable which takes values in the set $\{a, b, c, d\}$. Let $\Pr(\mathbf{X} = a) = \frac{1}{8}, \Pr(\mathbf{X} = b) = \frac{1}{4}, \Pr(\mathbf{X} = c) = \frac{1}{8}, \Pr(\mathbf{X} = d) = \frac{1}{2}$. Therefore, $H(\mathbf{X}) = \frac{7}{4}$.

4.1. Log sum inequality

Whenever we have quantitative measures, we often would like to compare them. For example, in security applications, we may want to compare how much information is leaked by running a program. One of the tools that we have to compare entropy is log sum inequality stated below without proof.

THEOREM 4.4. Let a_1, a_2, \dots, a_n and b_1, b_2, \dots, b_n be non-negative real numbers. We have that

$$\sum_{1 \leq i \leq n} a_i \log_2 \frac{a_i}{b_i} \geq \left(\sum_{1 \leq i \leq n} a_i \right) \log_2 \frac{\left(\sum_{1 \leq i \leq n} a_i \right)}{\left(\sum_{1 \leq i \leq n} b_i \right)}$$

with equality iff $\frac{a_i}{b_i} = \text{const}$.

Remark: We adopt the convention that $0 \log_2 \frac{0}{0} = 0$ and that $p \log_2 \frac{p}{0} = \infty$ if $p \neq 0$.

One of the consequences of the log sum inequality is the following.

THEOREM 4.5. Let $|Q| = n$. Show that for any random variable $\mathbf{X} : \Omega \rightarrow Q$,

$$H(\mathbf{X}) \leq \log_2(n)$$

with equality iff $\Pr(\mathbf{X} = q) = \frac{1}{n}$ for each $q \in Q$.

PROOF. We have

$$\begin{aligned} H(\mathbf{X}) &= - \sum_{q \in Q} \Pr(\mathbf{X} = q) (\log_2 \Pr(\mathbf{X} = q)) \\ &= - \sum_{q \in Q} \Pr(\mathbf{X} = q) \left(\log_2 \frac{\Pr(\mathbf{X} = q)}{1} \right) \\ &\leq - \left(\sum_{q \in Q} \Pr(\mathbf{X} = q) \right) \left(\log_2 \frac{\sum_{q \in Q} \Pr(\mathbf{X} = q)}{n} \right) \quad (\text{by log-sum inequality}) \\ &\leq - \log_2 \frac{1}{n} \quad \text{as } \sum_{q \in Q} \Pr(\mathbf{X} = q) = 1 \\ &\leq \log_2(n). \end{aligned}$$

Note the equality happens only when

$$\frac{\Pr(\mathbf{X} = q)}{1} = \frac{\Pr(\mathbf{X} = q')}{1} \text{ for all } q, q' \in Q.$$

Which implies that equality only happens when $\Pr(\mathbf{X} = q) = \frac{1}{n}$ for each $q \in Q$. \square

4.2. Basic Coding theory

One of the things that computer scientists are interested in is transmission of data. Data is usually transmitted in bits (or as 0s and 1s).

Definition: Let Q be a finite set. A *code* c is a map $c : Q \rightarrow \{0, 1\}^+$.

Remark: Strictly speaking the map $c : Q \rightarrow \{0, 1\}^+$ is said to be a *binary code*. We will drop the adjective binary.

A code is only useful for transmission as long as the receiver can distinguish different symbols being transmitted.

Definition: A code $c : Q \rightarrow \{0, 1\}^+$ is said to be *non-singular* if $c(q) \neq c(q')$ whenever $q \neq q'$.

Usually, a stream of symbols is transmitted and not just a single symbol. The receiver should be able to interpret the stream of symbols uniquely. One way of achieving it is to explicitly use separators between different symbols when transmitting. But this is expensive. Instead *uniquely decodable codes* are used.

Definition: A code $c : Q \rightarrow \{0, 1\}^+$ is said to be *uniquely decodable* if whenever $c(q_1)c(q_2)\dots c(q_n) = c(q'_1)c(q'_2)\dots c(q'_n)$, then $n' = n$ and for each $1 \leq i \leq n$, $q_i = q'_i$.

One way to ensure uniquely decodable codes is to have *prefix codes*.

Definition: A code $c : Q \rightarrow \{0, 1\}^+$ is said to be a *prefix code* if for each $q \neq q'$, $c(q)$ is **NOT** a prefix of $c(q')$.

EXERCISE 4.6. *Show that any prefix code is uniquely decodable.*

EXERCISE 4.7. *Give an example of a non-singular code that is not uniquely decodable. Give an example of an uniquely decodable code that is not a prefix code.*

THEOREM 4.8. *Let Q be a finite set let $c : Q \rightarrow \{0, 1\}^+$ be a prefix coding. Let $\ell : Q \rightarrow \mathbb{N}$ be the function that maps q to the length of the code $c(q)$. We have that*

$$\sum_{q \in Q} \frac{1}{2^{\ell(q)}} \leq 1.$$

PROOF. Let $\ell_{\max} = \max_{q \in Q} \ell(q)$. Consider the full binary tree \mathbb{T} of height ℓ_{\max} and label the edges of the tree \mathbb{T} as follows. If the edge connects a left child to its parent, label it 0 otherwise label it 1. Now for each $q \in Q$, let node N_q be the node which is connected to the root of \mathbb{T} by the path labeled as $c(q)$.

Consider the descendants of N_q at depth ℓ_{\max} . Now, there are $2^{\ell_{\max} - \ell(q)}$ descendants of N_q at the depth ℓ_{\max} . Thanks to c being a prefix code, we have that for $q' \neq q$, $N_{q'}$ is not a descendant of N_q . Therefore,

$$\sum_{q \in Q} 2^{\ell_{\max} - \ell(q)} \leq \text{number of nodes at depth } \ell_{\max}.$$

However, the number nodes at depth $\ell_{\max} = 2^{\ell_{\max}}$. The result now follows. \square

We also have the following converse.

THEOREM 4.9. *Without loss of generality, assume $\ell_1 \leq \ell_2 \leq \dots \leq \ell_n$. Let $Q = \{q_1, \dots, q_n\}$ and let $\ell_1, \ell_2, \dots, \ell_n$ be natural numbers such that*

$$\sum_{1 \leq i \leq n} \frac{1}{2^{\ell_i}} \leq 1.$$

Show that there exists a prefix code $c : Q \rightarrow \{0, 1\}^+$ such that length of $c(q_i)$ is ℓ_i .

PROOF. Let $\ell_{\max} = \max_{1 \leq n} \ell_i$. Consider the full binary tree \mathbb{T} of height ℓ_{\max} . In this binary tree, label the left edge as 0 and label the right edge as 1. We will choose nodes $n_{q_i}, 1 \leq i \leq n$ and construct codes $c(q_i), 1 \leq i \leq n$ inductively such that the following hold.

- (1) $c(q_i)$ is the label of the path from the root to n_{q_i} .
- (2) For each $j \neq k$, $c(q_k)$ is not a prefix of $c(q_j)$.
- (3) The number of leaf nodes of \mathbb{T} which have one of the n_{q_1}, \dots, n_{q_i} as an ancestor is

$$2^{\ell_{\max} - \ell_1} + \dots + 2^{\ell_{\max} - \ell_i}.$$

Base Case: We fix a node n_{q_1} whose distance from the root is ℓ_1 . We let $c(q_1)$ be the label of the path from the root to n_{q_1} . Now the number of leaf nodes of \mathbb{T} which have n_{q_1} as an ancestor is exactly $2^{\ell_{\max} - \ell_1}$.

Inductive step: Suppose, for some $1 \leq i < n$, we have picked nodes n_{q_1}, \dots, n_{q_i} and codes $c(q_1), \dots, c(q_i)$. Now, by inductive hypothesis, the number of leaf nodes of \mathbb{T} which have one of the n_{q_1}, \dots, n_{q_i} as an ancestor is

$$2^{\ell_{\max} - \ell_1} + \dots + 2^{\ell_{\max} - \ell_i}.$$

Thus the number of leaf nodes that do not have n_{q_1}, \dots, n_{q_i} as an ancestor is

$$\begin{aligned} 2^{\ell_{\max}} - (2^{\ell_{\max} - \ell_1} + \dots + 2^{\ell_{\max} - \ell_i}) &= 2^{\ell_{\max}} \left(1 - \left(\frac{1}{2^{\ell_1}} + \dots + \frac{1}{2^{\ell_i}} \right) \right) \\ &\geq 2^{\ell_{\max}} \left(\frac{1}{2^{\ell_{i+1}}} + \dots + \frac{1}{2^{\ell_n}} \right) \\ &> 0. \end{aligned}$$

Thus, there is a leaf node n' such that the path from the leaf to n' does not contain any node in n_{q_1}, \dots, n_{q_i} . Let $n_{q_{i+1}}$ be the node on this path such that the distance from the root to this node is ℓ_{i+1} . Let $c(q_{i+1})$ be the label of the path from the root to $n_{q_{i+1}}$. $n_{q_{i+1}}$ and $c(q_{i+1})$ is easily seen to satisfy the inductive hypothesis, *i.e.*.

- (1) $c(q_{i+1})$ is the label of the path from the root to $n_{q_{i+1}}$.
- (2) For each $j, k \leq i, j \neq k$, $c(q_k)$ is not a prefix of $c(q_j)$.
- (3) The number of leaf nodes of \mathbb{T} which have one of the $n_{q_1}, \dots, n_{q_{i+1}}$ as an ancestor is

$$2^{\ell_{\max} - \ell_1} + \dots + 2^{\ell_{\max} - \ell_{i+1}}.$$

□

4.2.1. Coding and entropy. Usually, the messages being transmitted are modeled as being generated by a source which generates the symbols randomly.

Definition: A random variable $\mathbf{X} : \Omega \rightarrow Q$ is said to be a *random source* for Q . The expected length of a code $c : Q \rightarrow \{0, 1\}^+$ is said to be the sum $\sum_{q \in Q} \Pr(X = q)\ell(q)$.

Remark: The weak law of large numbers would imply that if a sequence of n symbols is generated independently according to the random variable \mathbf{X} , then the length of the binary code being transmitted will be close to n times the expected length of c as n goes to infinity.

We have that entropy is a lower bound on the expected length of a prefix code.

THEOREM 4.10. *Let Q be a finite set and let $\mathbf{X} : \Omega \rightarrow Q$ be a random source. Let $c : Q \rightarrow \{0, 1\}^+$ be a prefix code. Show that the expected length of the code $c \geq H(\mathbf{X})$.*

PROOF. Let $\ell : Q \rightarrow \mathbb{N}$ be the function that maps q to the length of the code $c(q)$. Now, the expected length of the code is $\sum_{q \in Q} \Pr(\mathbf{X} = q)\ell(q)$. Now,

$$\begin{aligned} (\sum_{q \in Q} \Pr(X = q)\ell(q)) - H(\mathbf{X}) &= \sum_{q \in Q} \Pr(X = q)(\log_2 2^{\ell(q)} + \log_2 \Pr(X = q)) \\ &= \sum_{q \in Q} \Pr(X = q)(-\log_2 2^{-\ell(q)} + \log_2 \Pr(X = q)) \\ &= \sum_{q \in Q} \Pr(X = q) \log_2 \frac{\Pr(X=q)}{2^{-\ell(q)}} \\ &\geq (\sum_{q \in Q} \Pr(X = q)) \log_2 \frac{\sum_{q \in Q} \Pr(X=q)}{\sum_{q \in Q} 2^{-\ell(q)}} \\ &\geq 1 \cdot \log_2 \frac{1}{\sum_{q \in Q} 2^{-\ell(q)}} \end{aligned}$$

But, thanks to Theorem 4.8, $\sum_{q \in Q} 2^{-\ell(q)} \leq 1$. Therefore,

$$\left(\sum_{q \in Q} \Pr(X = q)\ell(q)\right) - H(\mathbf{X}) \geq 0$$

as required. □

THEOREM 4.11. *Let Q be a finite set and let $\mathbf{X} : \Omega \rightarrow Q$ be a random source. Show that there is a prefix code $c : Q \rightarrow \{0, 1\}^+$ of expected length $\leq H(\mathbf{X}) + 1$.*

PROOF. Let for $q \in Q_i$ let ℓ_q be the unique ℓ such that

$$\frac{1}{2^\ell} \leq \Pr(X = q) \leq \frac{1}{2^{\ell-1}}.$$

By definition, we have

$$\sum_{q \in Q} \frac{1}{2^{\ell_q}} \leq \sum_{q \in Q} \Pr(X = q) = 1.$$

Therefore, by Theorem 4.9, we have that there is a prefix code $c : Q \rightarrow \Sigma^*$ such that length of $c(q)$ is ℓ_q . Now, we also have that

$$\Pr(X = q) \leq \frac{1}{2^{\ell_q - 1}}$$

which implies that

$$\log \Pr(X = q) \leq -\ell_q + 1$$

and hence

$$\log \Pr(X = q) \geq \ell_q - 1.$$

Thus, we get that

$$H(\mathbf{X}) \geq \sum_{q \in Q} \Pr(X = q)(\ell_q - 1) \geq \left(\sum_{q \in Q} \Pr(X = q)\ell_q \right) - 1.$$

The result follows from observing that $\sum_{q \in Q} \Pr(X = q)\ell_q$ is the expected length of the code c . \square

4.3. Huffman Coding

We shall now show how to construct an optimal prefix coding, namely, a prefix code that minimizes the expected length. The coding scheme that we shall give is called Huffman coding, after David A. Huffman, who designed the coding scheme.

Suppose we want to construct an optimal code for $\mathbf{X} : \Omega \rightarrow Q$ where Q is a finite set. The construction of the code uses a “greedy algorithm.” At each step of the construction, a forest \mathcal{F} of node-labeled binary trees, is maintained. If Q , the set of input symbols, has n nodes, then initially the forest \mathcal{F} consists of n trees all of which have exactly one node. There is one tree corresponding to each input symbol q whose root is labeled by $\Pr(X = q)$. At each step of the construction, the algorithm picks two trees \mathbb{T}_1 and \mathbb{T}_2 whose roots are labeled by the smallest two numbers. It then creates a new tree \mathbb{T}_{new} whose root has \mathbb{T}_1 and \mathbb{T}_2 as its children and is labeled by the sum of the labels of the roots of \mathbb{T}_1 and \mathbb{T}_2 . The new forest at the end of this step is $(\mathcal{F} - \{\mathbb{T}_1, \mathbb{T}_2\}) \cup \mathbb{T}_{\text{new}}$. The construction terminates when there is only one tree left in the forest \mathcal{F} . Let $\mathbb{T}_{\text{final}}$ be the only tree left after the construction. For each node in $\mathbb{T}_{\text{final}}$, label the left edge as 0 and the right edge as 1. Then, the code $c(q)$ is defined to be the label of the path from the root of $\mathbb{T}_{\text{final}}$ to the node corresponding to q .

Remark: Note that the choice of labeling the left edge as 0 and the right edge as 1 is completely arbitrary. So, in principle, there are several Huffman codes.

EXERCISE 4.12. Assuming that comparison and addition of real numbers take constant time, what is running time of computing Huffman code?

We shall now show that Huffman code is indeed optimal.

Definition: Let Q be a finite set containing at least 2 elements generated by a random source $\mathbf{X} : \Omega \rightarrow Q$. A prefix code $c : Q \rightarrow \{0, 1\}^+$ is said to be *optimal* if the expected length of c is less than or equal to the expected length of any prefix code $c' : Q \rightarrow \{0, 1\}^+$.

We need the following lemma.

LEMMA 4.13. *Let Q be a finite set containing at least 2 elements generated by a random source $\mathbf{X} : \Omega \rightarrow Q$. Then there is an optimal prefix code $c : Q \rightarrow \{0, 1\}^+$ and a function $\ell : Q \rightarrow \mathbb{N}$ is the function that maps q to the length of the code $c(q)$ then the following must happen.*

- For each $q_1, q_2 \in Q$, $\Pr(\mathbf{X} = q_1) < \Pr(\mathbf{X} = q_2)$ then $\ell(q_2) \leq \ell(q_1)$.
- Let $\ell_{\max} = \max_{q \in Q}(\ell(q))$. Then the set $\ell^{-1}(\ell_{\max})$ consists of at least two elements.
- Fix two elements q_1, q_2 of Q such that for each $q \in Q$, $\Pr(\mathbf{X} = q) \geq \max\{\Pr(\mathbf{X} = q_1), \Pr(\mathbf{X} = q_2)\}$. (In other words q_1, q_2 are the elements which are generated with the least probability). Then $c(q_1)$ and $c(q_2)$ only differ in the last bit.

We shall leave the proof of the above lemma as an exercise. Actually, we did prove it in class, but it might be useful to try to prove it again yourself.

EXERCISE 4.14. *Prove lemma 4.13.*

We have that Huffman coding is optimal.

THEOREM 4.15. *Let Q be a finite set containing at least 2 elements generated by a random source $\mathbf{X} : \Omega \rightarrow Q$. Let $c_h : Q \rightarrow \{0, 1\}^+$ be a Huffman code. c_h is an optimal prefix code.*

PROOF. We will show that the result holds by induction on $|Q|$, the number of elements of Q . Clearly, the result holds for $|Q| = 2$. Assume that the result holds for any Q_0 and any random source $\mathbf{X}' : \Omega' \rightarrow Q_0$ such that $|Q_0| = m$ where $m \geq 2$.

Let $|Q| = m + 1$ and $\mathbf{X} : \Omega \rightarrow Q$ be a random source. Let c_h be a Huffman code with expected length ℓ_h . Also, let $\mathbb{T}_{\text{final}}$ be the tree obtained in the final step of the construction of the Huffman code c_h .

Let q_1, q_2 be the two elements of Q , such that the following two conditions hold—

- For every $q \in Q$, $\Pr(\mathbf{X} = q) \geq \max\{\Pr(q_1), \Pr(q_2)\}$.
- $c_h(q_1)$ and $c_h(q_2)$ differ in the last bit. Let $b_h^{q_1, q_2}$ be the bitstring obtained from $c_h(q_1)$ by deleting the last bit.

Also let c_{opt} be an optimal code for Q such that $c_{\text{opt}}(q_1)$ and $c_{\text{opt}}(q_2)$ differ in the last bit. Let $b_{\text{opt}}^{q_1, q_2}$ be the bitstring obtained from $c_{\text{opt}}(q_1)$ by deleting the last bit. Let the expected length of c_{opt} be ℓ_{opt} . We have by definition,

$$\ell_{\text{opt}} \leq \ell_h.$$

Pick a new symbol q_{new} not in Q and let $Q' = (Q \setminus \{q_1, q_2\}) \cup \{q_{\text{new}}\}$ and let $\mathbf{X}' : \Omega \rightarrow Q'$ be the random variables that such that $\mathbf{X}'(\omega) = \mathbf{X}(\omega)$ if $\mathbf{X}(\omega) \notin$

$\{q_1, q_2\}$ and $\mathbf{X}'(\omega) = q_{\text{new}}$ otherwise. Consider the codes $c_1, c_2 : Q' \rightarrow \{0, 1\}^+$ defined as follows

- If $q \in Q \setminus \{q_1, q_2\}$, then $c_1(q) = c_h(q)$ and $c_2(q) = c_{\text{opt}}(q)$.
- $c_1(q_{\text{new}}) = b_h^{q_1, q_2}$ and $c_2(q_{\text{new}}) = b_{\text{opt}}^{q_1, q_2}$.

We make the following observations.

- (1) c_1 and c_2 are prefix codes for $\mathbf{X} : \Omega \rightarrow Q'$.
- (2) Let ℓ_1 and ℓ_2 be the expected lengths of c_1 and c_2 . Then

$$\ell_1 = \ell_h - (\Pr(X = q_1) + \Pr(X = q_2))$$

and

$$\ell_2 = \ell_{\text{opt}} - (\Pr(X = q_1) + \Pr(X = q_2)).$$

- (3) c_1 is a Huffman code for $\mathbf{X} : \Omega \rightarrow Q'$. Therefore, by induction hypothesis,

$$\ell_1 \leq \ell_2.$$

This implies that

$$\ell_h \leq \ell_{\text{opt}}.$$

Since, we already had $\ell_{\text{opt}} \leq \ell_h$, we get that $\ell_h = \ell_{\text{opt}}$. □