Sample Complexity of Training Markov Chains

Roman Talyansky
Sample Complexity of Training Markov Chains

Research Thesis

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Roman Talyansky

Submitted to the Senate of the Technion — Israel Institute of Technology

Tamuz 5774 Haifa July 2014
The research thesis was done under the supervision of Prof. Alon Itai in the Computer Science Department.

The generous financial support of the Technion is gratefully acknowledged.
## Contents

Abstract 1

Abbreviations and Notations 3

1 Introduction 5

2 Related Work 10
   2.1 Probabilistic finite state automata 10
   2.2 Training Markov Chains 12
   2.3 Single-String Maximum Liklihood Model 13
   2.4 Hidden Markov Models 14
   2.5 Learning in the limit 14
   2.6 PAC-learning 14

3 Problem Statement 18

4 Mathematical and Information Theory Background 21
   4.1 Distinguishing Random Variables 21
   4.2 Matrix Perturbation 24
   4.3 Information Theory 26

5 Sample Complexity of Markov Chains 28
   5.1 Definitions 28
   5.2 Learnability vs. Distinguishability 29
## List of Figures

1.1 Sample Complexity ................................................................. 8

4.1 Perturbation of a MC .............................................................. 25

5.1 Edge learnability, assuming that $s$ is drawn from $M_1$ .................. 30

5.2 Node and edge stationary probabilities ........................................ 33

5.3 Strongly typical sequences ...................................................... 36

5.4 The part (5.4) of the upper bound (5.5) on the probability of non-strongly typical sequences for $\text{sign}=1$, $\delta = 0.1$ and $\theta = 0.01$ ..................... 39

5.5 The part (5.4) of the upper bound (5.5) on the probability of non-strongly typical sequences for $\text{sign}=-1$, $\delta = 0.1$ and $\theta = 0.01$ ..................... 40

5.6 The lower bound (5.7) on the probability of strongly typical sequences for $\theta = 0.01$ .............................................................. 41

5.7 The sample complexity bound on learning MC, whose transition matrix is given in Table 7.2 ......................................................... 47

5.8 The sample complexity bound on learning stationary edge probabilities of MC, whose transition matrix is given in Table 7.2 .................. 53

7.1 The practical Sample Complexity bound for $\delta = 0.1$ and $\theta = 1^{-4}$ . . . . . . . 72
Abstract

Markov chains have gained wide popularity in a diverse spectrum of applications, where there is a need to model time dependent processes. Markovian models appear extensively in thermodynamics and statistical mechanics, testing and telecommunication, finance and economics.

A Markov chain (MC) consists of the set of states, the initial state probabilities, the set of edges and the transition probabilities that define probabilities to traverse edges.

MCs are learned from a training data – sequences s of MC states. Training data are produced by a MC, called concept MC $M^{(c)}$ that cannot be observed directly.

Let dist be a distance between MCs and $\epsilon, \delta, \theta \in (0, 0.5)$. The goal of learning is to use s to construct a learned MC $M^{(l)}$ that is $\epsilon$-close to $M^{(c)}$ with confidence at least $(1 - \delta)$:

$$P_r(dist(M^{(c)}, M^{(l)}) < \epsilon) > 1 - \delta .$$ (1)

To develop sample complexity lower bounds we use methods of Information Theory: distinguishing between two random variables [5] and strongly typical sequences [5]. We also use matrix perturbation analysis [4].

We develop three bounds on the sample complexity of learning MCs. The first bound is expressed by the terms of the concept MC. This bound is theoretic, because in practice the concept MC is not known and thus the bound cannot be calculated. The second bound is expressed by the terms of the MC, that is learned from a sequence over the states of the concept MC and distributed according to the concept MC, but involves heavy computations. Finally the last bound is practical, since it is both expressed by the terms of the learned MC and may be calculated efficiently.
Abbreviations and Notations

\begin{itemize}
\item \( M \) — Markov Chain
\item \( M^{(c)} \) — concept Markov Chain that produces training data
\item \( M^{(\ell)} \) — learned Markov Chain that is learned from training data
\item \( M^{(h)}_1 \) — the perturbed MC, resulted from MC \( M^{(h)}_0 \), \( h \in \{c, \ell\} \)
\item \( V \) — a set of states
\item \( E \) — a set of edges
\item \( \iota \) — initial state probabilities
\item \( Q \) — a matrix of transition state probabilities
\item \( \sigma \) — denotes both stationary state and edge probabilities
\item \( \text{error}^g \) — the overall probability of error of the decision function \( g \)
\item \( A^c \) — the complement of the set \( A \)
\item \( B^{tr} \) — the transpose of the matrix \( B \)
\item \( \text{dist}(\cdot, \cdot) \) — a distance function
\item \( \text{dist}_\sigma \) — the distance over stationary edge probabilities
\item \( \text{dist}_Q \) — the distance over transition edge probabilities
\item \( \text{dist}_r \) — the distance over transition edge probabilities
\item \( SC_B(P_0, P_1, \delta) \) — the sample complexity for distinguishing between Bernoulli variables \( P_0 \) and \( P_1 \) with confidence \( (1 - \delta) \)
\item \( \overline{SC_B}(P_0, P_1, \delta) \) — the lower bound on the sample complexity for distinguishing between Bernoulli variables \( P_0 \) and \( P_1 \) with confidence \( (1 - \delta) \)
\item \( F \) — perturbation matrix
\item \( F_\epsilon(e_0, e_1) \) — the \( \epsilon \) perturbation matrix over edges \( e_0 \) and \( e_1 \)
\item \( \kappa(\cdot) \) — the condition number
\item \( H(\cdot) \) — entropy
\item \( I(\cdot; \cdot) \) — mutual information
\item \( M^{[n]} \) — the set of Markov Chains with \( n \) states
\item \( DF \) — a decision function to distinguish between Markov Chains
\end{itemize}
$s$ — a sequence of states $s_0,\ldots, s_{l-1}$

$start(s)$ — the first state in $s$

$end(s)$ — the last state in $s$

$P(\cdot|M,l)$ — the probability distribution over paths of MC $M$ of length $l$

$n(e|s)$ — the number of times edge $e$ appears in sequence $s$

$e$ — a sequence of edges $e_0,\ldots, e_{m-1}$

$n(e|s)$ — $(n(e_0|s), \ldots, n(e_{m-1}|s))$

$\tau(\theta,l)$ — the set of $\theta$-strongly typical sequences of length $l$ of MC

$\Xi_M$ — the lower bound on the probability of the set of $\theta$-strongly typical sequences of length $l$ in a MC $M$

$SC_M(\cdot,\cdot,\cdot,\cdot)$ — the lower bound on the sample complexity to distinguish between Markov Chains

$LA(s, n, \epsilon, \delta)$ — a learning function that learns Markov Chains with $n$ states from a state sequence $s$ with error $\epsilon$ and confidence $(1 - \delta)$

$LA_f(s, n, \epsilon, \delta)$ — an $f$-guarantee learning function

$S(\alpha, M)$ — the set of MCs $\alpha$-close to $M$

$K(e_0, e_1)$ — the sum of stationary edge probabilities of $e_0$ and $e_1$, maximized over $M_0$ and $M_1$

$K^{(h)}(e_0, e_1)$ — the sum of stationary edge probabilities of $e_0$ and $e_1$, maximized over $M_0$ and $M_1$ for concept MC ($h = c$) or learned MC ($h = \ell$)

$\Delta K$ — $|K^{(\ell)} - K^{(c)}|$

$\Psi(e_0, e_1)$ — the upper bound on $\Delta K$

$SCP_{mc}(M^{(\ell)}, \epsilon, \delta_B, e_0, e_1)$ — the practical lower bound on the sample complexity to learn MCs, by the terms of the learned MC $M^{(\ell)}$
Chapter 1

Introduction

Novelty in our work:

- Usually sample complexity is shown for learning an entire class of objects, e.g. all rectangles in a plane. In contrast., we show a way to calculate sample complexity for an individual MC.

- It is simpler to calculate sample complexity for a concept MC. However in practice concept MC is unknown and we have an access only to a training sequence, generated by the concept MC. Our bounds allow to calculate sample complexity of the concept MC using only the training sequence.

Markov chains (MC) have gained wide popularity in a diverse spectrum of applications. Markovian models appear extensively in thermodynamics ([11]) and statistical mechanics ([15]), testing [24] and telecommunication ([13]), finance and economics (Calvet [3]). Hidden Markov Models (hmm) and Finite State Machines are a generalization of Markov Chains and extensively used in information sciences. A wide overview of hmm and Finite State Machines may be found in [22] and [23]. In telephone networks hmms are used for error correction (Gupta and Mehra[19]), in Cryptanalysis (Karlof [9]) and Speech Recognition systems (Juang and Rabiner [16]), in speech synthesis (Zen et al [25]) and part-of-speech tagging and in many other domains.

We consider processes that are modeled by infinite sequences of states $s_0, s_1, \ldots$. Given a prefix $s_0, \ldots, s_{i-1}$, the distribution of the next state $s_i$ satisfies the Markovian property: it depends only on the previous state $s_{i-1}$, i.e.,

$$P(s_i|s_0, s_1, \ldots, s_{i-1}) = P(s_i|s_{i-1}) .$$

(1.1)
Note that for any state \( u \in V \), the probabilities of edges that emanate from \( u \) form a distribution, i.e.
\[
\sum_{v \in V} P(v|u) = 1.
\]

We follow the definitions of [8], [20, Chapters 3,4], and [17, Chapter 4]. A **Discrete Time Markov Chain** is a triple \( M = (V, \iota, Q) \), where \( V \) is the set of states; \( \iota \) is the initial state probabilities, for each \((u, v) \in E\), \( Q(u, v) \) is the transition probability and for each \( u \in V \), \( \{Q(u, v), v \in V\} \) is a distribution. We assume that all edges are permitted, i.e. \( E = V \times V \). Note that \( Q(u, v) = P(v|u) \). For simplicity we shorten Discrete Time Markov Chains to Markov Chains (MC).

A MC operates as follows: First an initial state is chosen at random according to \( \iota \). Assume that at time \( t \) the MC is at state \( u \). Then \( Q(u, v) \) is the probability that at time \( t + 1 \) the state is \( v \). Let \( s = s_0, s_1, \ldots, s_{l-1} \) be a sequence of states of length \( l \) over \( V \). A MC induces a distribution on sequences of states
\[
P(s_0, s_1, \ldots, s_{l-1}) = P(s_{l-1}|s_{l-2}, \ldots, s_0)P(s_{l-2}, \ldots, s_0) = \iota(s_0) \prod_{i=1}^{l-1} P(s_i|s_{i-1}, \ldots, s_0).
\]

Using the Markovian property (1.1)
\[
P(s_0, s_1, \ldots, s_{l-1}) = \iota(s_0) \prod_{i=1}^{l-1} P(s_i|s_{i-1}) = \iota(s_0) \prod_{i=1}^{l-1} Q(s_{i-1}, s_i).
\]

**Example: Brand Loyalty** Uslu and Cam [21] analyze brand loyalty using Markov chains. In that paper the authors conducted a survey of university students, in which students were asked the present brand of sport shoes they own and their preference for their next sport shoes purchase. The process of the change in brand preferences was modeled by a MC, where the states represent the present brand of sport shoes and the transitions between states represent the change in brand preference for their next sport shoes. The initial state probability describes the distribution of brands among the students when the survey was conducted. The output of this MC represents the buying preference for purchase of sport shoes.

In this example the underlying phenomenon is the process of change in brand preferences and the training data is collected using student surveys asking for the current brand and the brand preference for the next purchase.
A MC describes some underlying phenomenon and typically it is learned from training data that is produced by this phenomenon. In our work we assume that the phenomenon is governed by a concept MC and the training data consist of a sequence of states produced by the concept MC.

A learning function is a function, whose input consists of training data which is distributed according to the concept MC $M_0 = (V_0, i_0, Q_0)$, the number of states in $M_0$, and two constants – error $\epsilon$ and confidence $\delta$, $\epsilon, \delta \in (0, 0.5)$. The learning function produces a learned MC $M^{(l)} = (V_1, i_1, Q_1)$ over the same set of states such that with probability $1 - \delta$, $\max_{e \in V \times V} |Q_1(e) - Q_0(e)| < \epsilon$. In practice, the designer knows a priori the desired confidence and acceptable error and he strives to learn with the desired confidence a MC whose error does not exceed the acceptable error. Intuitively the larger the training set the lower the error. The minimal size of the training data, required to get the error below the acceptable error for the desired confidence is called sample complexity. If the size of the training data is below the sample complexity of the learning task, no learning strategy can obtain acceptable approximations to the concept MC. Instead we should invest our efforts in obtaining more training data or even abandon the effort to learn the MC, if gaining more training data is infeasible. In the absence of a tool for estimating the sample complexity, designers of MCs have only one way to validate that they have enough training data – they build a MC out of the training data and then estimate the accuracy of the resulting target MC. But even then overfitting may occur: the target MC fits the specific training data set but does not generalize to capture the underlying phenomenon and approach the concept MC. Here we study lower bounds on sample complexity of learning MCs. These bounds will help to estimate the minimum amount of training data required to learn a MC with the desired confidence and error. Many bounds in Machine Learning are expressed asymptotically as $O(\cdot)$, which hides very large constants, rendering them inapplicable in practice. In contrast, our bounds explicitly express the constants. We will show application of our bounds to the brand loyalty example.

**Example** To demonstrate our results, in Figure 1.1 we plot the sample complexity bound of learning a MC with 10 states, confidence 0.9, whose entries were randomly generated, according to the uniform distribution. This transition matrix is presented in Section 7.2, Table 7.2. The $x$ axis shows the tolerable learning error. From Figure 1.1 we can see that the sample complexity to learn this matrix starts to explode for values of tolerable error below 0.1, meaning that learning this MC with confidence below 0.1 and error below 0.1 requires an impractically large number of examples.
We now give an outline of the thesis. The next section presents related work and Section 3 gives a precise definition of the problem we investigate. The necessary back-
ground for our analysis, matrix theory and information theory, is presented in Section 4. Then, in Section 5, we start the discussion on distinguishing between two closely related MCs, since we use it as an underlying tool for our learnability bounds. This discussion is followed by our main theoretical results on learnability bounds that are still formulated by means of concept MC. Those bounds are interesting from a theoretical point of view, but as Lemma 5.4.8 and Theorem 5.4.10 reveal, they are formulated by terms of concept MCs that is unknown to the learning function and thus cannot be applied in practice. In Section 6 we formulate the learnability bounds for the concept MC by the terms of the learned MC. Those bounds, however, are still impractical, since they do not provide means for efficient bound calculation. To make our research applicable in practice, in Section 7 we further develop our theoretical results to bounds that may be efficiently calculated and depend only on the learned MC. We conclude by an example where we apply our practical bounds to a concrete example, introduced in [21].
Chapter 2

Related Work

We first note that developing lower bounds on the sample complexity is harder than developing upper bounds. To show an upper bound it is sufficient to show a learning strategy and then analyze its complexity. In contrast, in order to prove a lower bound, one should examine all possible learning strategies and show that for each strategy if the error is below the desired error with the desired confidence then the sample complexity must exceed the lower bound.

We next discuss a connection between the sample and the computation complexity bounds of learning models. Since a learning algorithm must invest at least one computational step to examine and process each sample point of the training set, the sample complexity of a learning algorithm is not larger than its computational complexity. Hence we may infer upper bounds on the sample complexity from upper bounds on the computational complexity. For this reason we discuss results on the computational complexity along with the results on the sample complexity.

We are now ready to introduce several models that extend Markov Chains. Then we discuss learnability results, known for those models.

2.1 Probabilistic finite state automata

A probabilistic finite state automaton (PFA) $(V,E,\iota,Q,\Sigma)$ extends Markov chains by adding the notion of a finite size output alphabet $\Sigma$. The transition probabilities $Q$ are
over edges and output symbols:

$$Q : V \times V \times \Sigma \rightarrow [0,1] .$$

Each mapping $Q(v,\cdot,\cdot)$ is required to be a distribution on the edges emanating from $v$. Thus

$$\sum_{v:(u,v) \in E,a \in \Sigma} Q(u,v,a) = 1 .$$

Note that several edges emanating from a node may be labeled with the same output symbol.

A large number of works study upper bounds on the sample complexity of PFAs, see Abe and Warmuth [1]. If a learning algorithm has information on the structure of a PFA, this information can help the algorithm to achieve better accuracy. The structure of a PFA is defined by the set of its states, the initial states of positive probability and the positive probability transitions. Abe and Warmuth formalize the notion of the information on a PFA structure in the form of a PFA constraint. A PFA constraint is a triple of sets: the set of states, the set of initial states and the set of transitions. A PFA $P$ satisfies a constraint $C$ if their sets of states are equal, the positive probability initial states of $P$ are included in the initial states of $C$ and the positive probability transitions of $P$ are included in the transitions of $C$.

Let $D$ and $Q$ be probability distributions over the same countable domain. To measure a distance among distributions Abe and Warmuth use the Kullback-Leibler divergence $d_{KL}(D,Q)$, also known as relative entropy – a well-known measure in Information Theory. The Kullback-Leibler divergence is defined later in the beginning of Section 4.

Let $\Sigma$ be a finite alphabet. An example of $\Sigma^n$ is an element of $\Sigma^n$. A sample $\Xi$ of $\Sigma^n$ is a finite sequence of examples of $\Sigma^n$, $\Xi = < w_1, \cdots, w_m >$, where $m$ is the sample size.

Below we give the Abe and Warmuth’s definition of a training algorithm. They assume that a randomized algorithm has an access to a fair coin and can flip it in a single time unit. Let $PA(C)$ denote the class of PFAs satisfying the constraint $C$.

**Definition 2.1.1 (Training PFA constraints).** A training algorithm takes as input a constraint $C$, a string length $n$, and a finite sample $\Xi$ of $\Sigma^n$ for some alphabet $\Sigma$, and outputs a probabilistic automaton which satisfies $C$. A (possible randomized) training algorithm $A$ trains a class of constraints $C$ with sample size $q(1/\epsilon, 1/\delta, n, t)$, if $A$, when given as input an arbitrary constraint $C \in C$ of size $t$, a string length $n$, and a finite sample $\Xi$ drawn independently at random from an arbitrary unknown distribution $D$ over $\Sigma^n$, is such that whenever
the sample size $m$ exceeds $q(1/\epsilon, 1/\delta, n, t)$, then provided that $\inf \{ d_{KL}(D, P) : P \in \mathcal{PA}(C) \}$ is finite, $A$’s output $Q$ satisfies with probability at least $1 - \delta$:

$$d_{KL}(D, Q) - \min \{ d_{KL}(D, P) : P \in \mathcal{PA}(C) \} \leq \epsilon ,$$

Here the probability is taken over the product distribution of $D$ producing the sample and the random coin flips of $A$, if $A$ is randomized.

- If there exists such a training algorithm, then we say that $C$ is trainable with sample complexity $q(1/\epsilon, 1/\delta, n, t)$.

- If there exists a training algorithm with a polynomial sample complexity which also runs in time polynomial in the total sample length, then we say that $C$ is polynomially trainable.

The main positive result of Abe and Warmuth states that PFAs that satisfy an arbitrary constraint of size $t$ are trainable with error $\epsilon$, confidence $\delta$ and sample complexity

$$O((n/\epsilon)^2 t \cdot \log^3 n t/\epsilon \cdot \log 1/\delta \cdot \log^2 \log 1/\delta) .$$

(2.1)

Let $C = (\Sigma_C, V_C, I_C, G_C)$ be a constraint. The constraint $C$ is a null constraint if it allows all the initial states and all transitions, i.e. $I_C = V_C$ and $G_C = V_C \times V_C \times \Sigma_C$.

From the upper bound (2.1) on sample complexity we can deduce an upper bound on the sample complexity of training PFAs with the null-constraint

$$O((n/\epsilon)^2 |\Sigma|^2 |\Sigma| \cdot \log^3 n |\Sigma|/|\Sigma|/\epsilon \cdot \log 1/\delta \cdot \log^2 \log 1/\delta) .$$

(2.2)

2.2 Training Markov Chains

We now use the upper bound (2.2) on the sample complexity to get an upper bound on the sample complexity of learning MCs. First we adapt Definition 2.1.1 of training PFAs to a similar definition of training MCs. To shorten the definition we just point out the differences to Definition 2.1.1. First in the definition of a MC constraint the transitions are a subset of $V_C \times V_C$. Instead of requiring similarity between the distributions over output sequences, we require similarity between distributions over sequences of states. Thus the finite sample $\Xi$ is of $V^n$. Obviously the output of the training algorithm is a MC.
Next we show a reduction of the MCs training problem to the problem of training PFAs. Let \( CM \) be a MC constraint with set of states \( V_M \), transition set \( G_M \) and \( \Xi \) be finite sample of \( V^n \). Our purpose is to show an algorithm for training a MC constraint \( C_M \), based on the input \( C_M \) and \( \Xi \). We construct for \( C_M \) a PFA constraint \( C_P \) with the extended set of states 
\[
V_P = V_M \cup \{v_0\},
\]
where \( v_0 \notin V_M \). The output alphabet of \( \Sigma_P \) is the set of the states of \( C_M \). For each transition \((u,v) \in G_M\) we define the transition \((u,v,v)\) in \( G_P \). We also add a transition \((v_i,u,u)\) for every \( u \in V_M \). Let \( P \) be a PFA that satisfies \( C_P \). Note that the traversal probabilities of \( P \) imply traversal probabilities of \( M \). Thus training the PFA constraint \( C_P \) with error \( \epsilon \), confidence \( \delta \) and the sample complexity \( q(1/\epsilon, 1/\delta, t, n) \) implies training the MC constraint \( C_M \) with the same error, confidence and sample complexity. Thus we can use the upper bound (2.2) on sample complexity of training PFAs to write the following upper bound on the sample complexity of training MCs

\[
O((n/\epsilon)^2 |V_M|^3 \cdot \log^3 n |V_M|/\epsilon \cdot \log 1/\delta \cdot \log^2 \log 1/\delta).
\]

### 2.3 Single-String Maximum Likelihood Model

Let \( P = (V, E, \iota, Q, \Sigma) \) be a PFA, \( w = w_1, w_2, \cdots, w_l \) be a word in \( \Sigma^l \). The generation probability \( Q(w) \), assigned to \( w \) by \( P \) is defined as follows

\[
Q(w) = \sum_{<v_0,v_1,\cdots,v_l> \in V^{l+1}} \prod_{j=0}^{l-1} Q(v_j, v_{j+1}, w_{j+1}).
\]

Abe and Warmuth [1] also defined a Single-String MLM (Maximum Likelihood Model) Problem for a class of constraints, where the input is a constraint \( C \) and a word \( w \) in \( \Sigma^l \) and the output is a PFA \((V, E, \iota, Q^*)\), where \( Q^* \) satisfies the constraint \( C \) and maximizes the generation probability of \( w \)

\[
Q^*(w) = \arg\max\{Q(w) : Q \text{ is a stochastic matrix satisfying the constraint } C\}.
\]

Abe and Warmuth show the equivalence between the learning PFA problem and approximating the single-string MLM problem [1, Theorem 4.1]. Finally they develop a hardness result [1, Theorem 5.1]: the single-string MLM problem for the class of 2-state null PFA constraints is not approximable in time polynomial in the alphabet size and the length of
the input word, unless $P = NP$. Obviously this does not imply an upper bound on the sample complexity of training PFAs.

A finite state automaton is deterministic (pdfa) if no edges emanating out of the same node are labeled by the same output symbols. Kearns and Rubinfeld [10] studied pdfas with output alphabet of size two. They show that under The Noisy Parity Conjecture \(^1\) pdfas cannot be learned in time polynomial in $1/\delta$ and $1/\epsilon$, where $\delta$ is the confidence and $\epsilon$ is the error.

### 2.4 Hidden Markov Models

The Hidden Markov Model (HMM) is another important extension of Markov chains. It also uses the notion of a finite size output alphabet $\Sigma$. However in addition to the transition probabilities between nodes, HMMs associate with each node a distribution $d_v$ over $\Sigma$. When an HMM arrives at node $v$, it produces a symbol from $\Sigma$, according to the distribution $d_v$. Vidal et al [23, Propositions 4 and 5] show that (non-deterministic) PFAs and HMMs generate the same distributions and in this sense they are equivalent. This means that expression (2.1) also upper bounds the sample complexity of learning HMMs.

### 2.5 Learning in the limit

Rudich [18] studied the learning in the limit model, where an HMM is learned from an output sequence of unbounded length that is generated by this HMM. He considers HMMs with only two outgoing edges from each node and did not study bounds on the sample complexity for learning such HMMs.

### 2.6 PAC-learning

Definition 2.1.1 defines training PFA constraints. This definition is an adaptation of the PAC-learning paradigm to PFAs (Section 4.1, A PAC learning for probabilistic automata, [7]). We now briefly summarize these results.

Let $X$ be the set of all possible instances over which the target functions may be defined. Let $C$ be a set of target concepts that a learner might be called upon to learn. \(^1\)The Noisy Parity Conjecture conjectures that learning the parity function is hard

---

\(^1\)The Noisy Parity Conjecture conjectures that learning the parity function is hard
Each concept \( c \in C \) corresponds to some subset of \( X \), or equivalently to some boolean-valued function \( c : X \to \{0, 1\} \). We assume instances are generated at random from \( X \) according to some stationary probability distribution \( \mathcal{D} \), i.e. \( \mathcal{D} \) does not change over time. Training examples are generated by drawing an instance \( x \) at random according to \( \mathcal{D} \), then presenting \( x \) along with its target value, \( c(x) \), to the learner.

The learner \( L \) considers some set \( M \) of possible hypotheses when attempting to learn the target concept. After observing a sequence of training examples of the target concept \( c \), \( L \) must output some hypothesis \( h \) from \( M \), which is its estimate of \( c \). We evaluate the success of \( L \) by the performance of \( h \) over new instances drawn randomly from \( X \) according to \( \mathcal{D} \), the same distribution, used to generate the training data.

Within this setting, we are interested in characterizing the performance of various learners \( L \) using various hypothesis spaces \( M \), when learning individual target concepts drawn from various classes \( C \).

**Error of a hypothesis**

Because we are interested in how closely the learner’s output hypothesis \( h \) approximates the actual target concept \( c \), let us begin by defining the true error of a hypothesis \( h \) with respect to target concept \( c \) and instance distribution \( \mathcal{D} \). Informally, the true error of \( h \) is just the error rate we expect while applying \( h \) to future instances drawn according to the probability distribution \( \mathcal{D} \).

**Definition 2.6.1** The true error (denoted \( \text{error}_{\mathcal{D}}(h) \)) of hypothesis \( h \) with respect to target concept \( c \) and distribution \( \mathcal{D} \) is the probability that \( h \) will misclassify an instance drawn at random according to \( \mathcal{D} \).

\[
\text{error}_{\mathcal{D}}(h) = \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)].
\]

Here the notation \( \Pr_{x \in \mathcal{D}} \) indicates that the probability is taken over the instance distribution \( \mathcal{D} \).

**PAC-Learnability**

Within the PAC setting the learner probably learns a hypothesis that is approximately correct – hence the term probably approximately correct learning, or PAC learning for short. Let \( \text{size}(c) \) be the encoding length of \( c \) in \( C \), assuming some representation for \( C \). Next we define PAC-learnability
Definition 2.6.2 Consider a concept class $C$ defined over a set of instances $X$ of length $n$ and a learner $L$ using a hypothesis space $M$. $C$ is PAC-learnable by $L$ using $M$ if for all $c \in C$, distributions $D$ over $X$, $\epsilon$ such that $0 < \epsilon < 1/2$, and $\delta$ such that $0 < \delta < 1/2$, learner $L$ will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\text{error}_D \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, and size($c$).

Sample complexity for infinite hypothesis spaces

In this section we consider a measure of the complexity of a hypothesis space $M$, called the Vapnik-Chervonenkis dimension of $M$ (Section 7.4.2, [14]). We will denote it VC-dimension of $VC(M)$, for short. As we shall see, we can state bounds on sample complexity that use $VC(M)$. These bounds allow us to characterize the sample complexity of many infinite hypothesis spaces, and can be shown to be fairly tight.

Using $VC(M)$ as a measure for the complexity of $M$, it is possible to derive an answer to the following question: “How many randomly drawn training examples suffice to probably approximately learn any target concept in $C$”. Equation (2.3) provides an upper bound on the number of training examples sufficient to probably approximately learn any target concept in $C$, for any desired $\epsilon$ and $\delta$.

$$m \geq \frac{1}{\epsilon} \left( 4 \log_2(2/\delta) + 8 VC(M) \log_2(13/\epsilon) \right). \quad (2.3)$$

It is also possible to obtain a lower bound, as summarized in the following theorem

**Theorem 2.6.3** Lower bound on sample complexity (Theorem 7.3, [14]). Consider any concept class $C$ such that $VC(C) \geq 2$, any learner $L$, and any $0 < \epsilon < 1/8$, and $0 < \delta < 1/100$. Then there exists a distribution $D$ and target concept in $C$ such that if $L$ observes fewer examples than

$$\max \left[ \frac{1}{\epsilon} \log(1/\delta), \frac{VC(C) - 1}{32\epsilon} \right]$$

then with probability at least $\delta$, $L$ outputs a hypothesis $h$, having $\text{error}_D(h) > \epsilon$.

This theorem states that if the number of training examples is too small, then no learner can PAC-learn every target concept in any nontrivial $C$. Thus, this theorem provides a lower bound on the number of training examples necessary for successful learning, complementing the earlier upper bound that gives a sufficient number.

Our results provide lower bounds on sample complexity. However in contrast to Theorem 2.6.3, we consider Markov Chains, whose learning model does not suit the PAC-
learning settings. Indeed in PAC learning a concept \( c \) or hypothesis \( h \) are dichotomies that label each element \( x \) of the sample space \( X \) to either 0 or 1. Also in PAC-learning training examples are pairs of an element \( x \in X \), chosen according to an unknown distribution \( D \) over the sample space, and its label \( c(x) \). In Markov Chains a concept and a hypothesis are individual Markov Chains; training examples are state sequences that are generated by the concept MC. Thus we cannot apply bounds on sample complexity of PAC-learning to the problem of learning MCs.
Chapter 3

Problem Statement

A MC $M = (V, \iota, Q)$ defines a graph $G = (V, E)$, where $E$ is the set of pairs of states $(u, v)$ such that $Q(u, v) > 0$. When using graph terminology (nodes, edges, paths) we refer to the underlying graph.

Let $\rho_t$ denote the vector of state probabilities at time $t$. After $M$ traverses a single edge according to $Q$, the state probability vector is

$$\rho_{t+1} = \rho_t^e Q .$$

If the limit $\lim_{t \to \infty} \rho_t$ exists, the limit is called the stationary state probabilities. Throughout this work we assume that MCs are ergodic: there is a directed path between any two states and no integer greater than 1 divides the lengths of all the directed cycles [2, Proposition 3.20]. For ergodic MCs the limit exists, this limit is unique and does not depend on the initial state probabilities $\iota$.

**Definition 3.0.4** Let $M$ be an ergodic MC. The stationary node probability vector is defined as

$$\sigma_{\text{state}} = \lim_{t \to \infty} \rho_t .$$

Let $e = (u, v) \in E$. The stationary edge probability of $e = (u, v) \in E$ is defined as

$$\sigma_{\text{edge}}(e) = \sigma_{\text{state}}[u]Q(e) .$$ (3.1)

Note that the stationary edge probability $\sigma_{\text{edge}}$ is a matrix of the same size as $Q$. 

18
We will abuse notation and use $\sigma$ for both $\sigma^{\text{state}}$ and $\sigma^{\text{edge}}$.

Most of the research on MC focuses only on their stationary behavior and since the stationary behavior does not depend on the initial state probabilities (Theorem 3.17 in [2]), the initial state probabilities do not form part of the definition of Markov Chains. Since our aim is to estimate the lower bound on the sample complexity, and the stationary behavior of a MC becomes apparent only for sufficiently long sequences, we cannot neglect the initial state probabilities. However, as the discussion following Lemma 5.4.8 will reveal, our bounds do not depend on the initial state probabilities. They were included only to show that the bounds do not depend on them.

Let $s = v_0, v_1, \ldots, v_l$ be a path of $M$, $\text{start}(s)$ the first node of $s$. Then $P(\cdot|M, l)$, the probability distribution over paths $s$ of length $l$, satisfies

$$P^{[l]}(s|M) = \nu(\text{start}(s)) \prod_{i=1}^{l} Q[v_{i-1}, v_i] .$$

Adopting the notations of [5], let $X_1, X_2, \ldots, X_l$ be a sequence of $l$ i.i.d. random variables and let $P^{[i]}_i, i \in \{0, 1\}$ be two probability distributions over this sequence.

In Section 5.4 we shall reduce the problem of distinguishing between two MCs to that of distinguishing between distributions over sequences of i.i.d. random variables. For this reason we first define the problem of distinguishing between two distributions over sequences of random variables and then extend it to the problem of distinguishing between two MCs.

**Problem 1** We consider two hypotheses:

- $Hyp_0: P = P^{[0]}_0$.
- $Hyp_1: P = P^{[1]}_1$.

Let $\Omega^{(l)}$ be the set of all sequences of length $l$ over a random variable $X$. Let $A^c = \Omega^{(l)} - A$ denote the complement of the set $A$. Consider the general decision function $g(s)$, where $g(s) = 0$ implies that $Hyp_0$ is accepted and $g(s) = 1$ implies that $Hyp_1$ is accepted. Since the function takes only two values, the test may be defined by specifying the set $A$ over which $g(s)$ is 1. We define two probabilities of error:

$$\alpha^{[l]} = P(g(s) = 1|Hyp_0) = P^{[0]}_0(A^c) \quad (3.2)$$
and 
\[ \beta[l] = P(g(s) = 0|\text{Hyp}_1) = P[l](A) . \]

We follow a Bayesian approach, in which we assign prior probabilities to the two hypotheses. In this case we wish to minimize the overall probability of error given by the weighted sum of the individual probabilities of error.

For the two hypotheses, let \( P = P_0 \) with prior probability \( \text{prior}_0 \) and \( P = P_1 \) with prior probability \( \text{prior}_1 \). The overall probability of error of the decision function \( g \) is

\[
error_g[l] = \text{prior}_0 \alpha[l] + \text{prior}_1 \beta[l] 
\] (3.3)

This probability of error is called the Bayesian probability of error.

Now we extend Problem 1 and the definition of error probability (3.3) to MCs. Let \( M_0 \) and \( M_1 \) be two MCs over the same set of nodes. Instead of considering sequences of random variables of length \( l \), we will consider paths of length \( l \), generated according to either \( M_0 \) or \( M_1 \). The distinguishability problem for MCs is defined as follows

**Problem 2** We consider two hypotheses:

- \( \text{Hyp}_0 : M = M_0 \).
- \( \text{Hyp}_1 : M = M_1 \).

The probability of error (3.3) is also extended by considering paths, generated by either \( M_0 \) or \( M_1 \) instead of sequences over random variables, distributed according to either distribution \( P_0 \) or \( P_1 \).

The distinguishability problem is only a milestone to a more important practical problem – the problem of learning MCs from training data. In the settings of this problem there is a concept MC \( M^{(c)} \). One cannot observe \( M^{(c)} \), however he can observe a state sequence \( s \), distributed according to \( M^{(c)} \). Let \( dist \) be a distance between MCs and \( \epsilon, \delta \in (0, 0.5) \). The goal of learning is to use \( s \) to construct a learned MC \( M^{(l)} \) that will be \( \epsilon \)-close to \( M^{(c)} \) with confidence at least (1 − \( \delta \)).

**Problem 2** Construct \( M^{(l)} \) such that

\[
P(dist(M^{(c)}, M^{(l)}) < \epsilon) > 1 - \delta .
\]
4.1 Distinguishing Random Variables

We resume the discussion of the problem of distinguishing between two i.i.d.-s. Let $P_0$ and $P_1$ be two distributions over the alphabet $\Sigma$. The Kullback-Leibler divergence between $P_0$ and $P_1$ is defined as follows:

$$D(P_0||P_1) = \sum_{a \in \Sigma} P_0(a) \log \frac{P_0(a)}{P_1(a)}.$$ 

Let $|\Sigma|$ denote the size of the alphabet and $g$ be a decision function. Also let us recall that each decision function $g$ is associated with a set $A \in \Omega^{(l)}$ such that for each $s \in \Omega^{(l)}$: $g(s) = 1$ iff $s \in A$. Let

$$D^* = \lim_{l \to \infty} \min_{g} \frac{1}{l} \log error_g^{[l]}(A).$$

Let $ex$ be a function that takes as input a pair of probability distributions, returns a positive real number such that the following inequalities hold:

$$2^{-l \cdot ex(P_0,P_1)} \leq error_g^{[l]} \leq (l + 1)^{|\Sigma|} 2^{-l \cdot ex(P_0,P_1)}.$$ 

Such a function $ex$ is an exponent in the Bayesian probability of error. The following theorem provides the best achievable exponent [5].
Theorem 4.1.1  The best achievable exponent in the Bayesian probability of error is
\[ D(P_{\lambda^*} || P_0) = D(P_{\lambda^*} || P_1) \]

with
\[ P_{\lambda}(x) = \frac{P_0^\lambda(x) P_1^{1-\lambda}(x)}{\sum_{a \in \Sigma} P_0^\lambda(a) P_1^{1-\lambda}(a)} \]

and \( \lambda^* \) is the value of \( \lambda \) such that
\[ D(P_{\lambda^*} || P_0) = D(P_{\lambda^*} || P_1) \equiv C(P_0, P_1). \]

\( C(P_0, P_1) \) is the highest achievable exponent for the probability of error and is called the Chernoff information. An equivalent definition of the Chernoff information is
\[ C(P_0, P_1) = - \min_{0 \leq \lambda \leq 1} \log \left( \sum_{a \in \Sigma} P_0^\lambda(a) P_1^{1-\lambda}(a) \right). \]

Cover and Thomas [5] showed that
\[ 2^{-lC(P_0, P_1)} \leq \alpha[l], \beta[l] \leq (l + 1)^{|\Sigma|} 2^{-lC(P_0, P_1)}, \]

where \( \alpha[l] \) and \( \beta[l] \) are defined in (3.2). Hence,
\[ 2^{-lC(P_0, P_1)} \leq \text{error}^{[l]}_g \leq 2(l + 1)^{|\Sigma|} 2^{-lC(P_0, P_1)}. \quad (4.1) \]

When \( P_0 \) and \( P_1 \) are Bernoulli variables such that \( P_0(0) = p \) and \( P_1(0) = p + \epsilon \), the Chernoff information \( C(P_0, P_1) \) is attained at
\[ \lambda^* = \ln \left\{ \left[ -\ln \left( \frac{1-p}{1-p-\epsilon} \left( \frac{p}{p+\epsilon} \right)^{-1} \right) \right]^{-1} \frac{1-p-\epsilon}{p+\epsilon} \right\} \left( \ln \left( \frac{p}{1-p} \frac{1-p-\epsilon}{p+\epsilon} \right) \right). \]

By (4.1) the decision error is lower bounded by \( \text{error}^{[l]}_g \geq 2^{-lC(P_0, P_1)} \). For a given desired decision error \( \delta \in (0, 1) \), the decision error \( \text{error}^{[l]}_g > \delta \) when
\[ \text{error}^{[l]}_g \geq 2^{-lC(P_0, P_1)} > \delta \]
or

\[
l < \frac{\log_2 \frac{1}{\delta}}{C(P_0, P_1)} = LB_B(P_0, P_1, \delta),
\]

where \( LB_B(P_0, P_1, \delta) \) denotes the upper bound.

**Definition 4.1.2** Let \( \delta \in (0, 0.5) \). Two random variables are distinguishable with \( l \) examples and error \( \delta \) if there exists a decision function \( g \) such that when given \( \delta \) and \( l \) examples drawn according to \( X \), \( g \) returns a hypothesis \( \text{Hyp}_0 \) or \( \text{Hyp}_1 \) with error bounded by \( \text{error}^{[l]} < \delta \).

Now we can define the true sample complexity of distinguishing random variables.

**Definition 4.1.3** Let \( \delta \in (0, 0.5) \), \( P_0 \) and \( P_1 \) be two random variables. The sample complexity to distinguish Bernoulli random variables is a function \( SC_B(P_0, P_1, \delta) \) over the same domain as \( LB_B \), such that

- For any \( l < SC_B(P_0, P_1, \delta) \), \( P_0 \) and \( P_1 \) are not distinguishable with \( l \) examples and confidence \( 1 - \delta \);
- For any \( l \geq SC_B(P_0, P_1, \delta) \), \( P_0 \) and \( P_1 \) are distinguishable with \( l \) examples and confidence \( 1 - \delta \).

The following lemma summarizes our findings:

**Lemma 4.1.4** For \( l \) that satisfies (4.2), two random variables with probabilities \( P_0 \) and \( P_1 = P_0 + \epsilon \), \( \epsilon \in [0; 1-P_0] \), are not distinguishable with \( l \) examples and error \( \delta \).

We can draw the following corollary from Lemma 4.1.4

**Corollary 4.1.5** Let \( \delta \in (0, 0.5) \), \( P_0 \) and \( P_1 \) be two random variables. Then

\[
LB_B(P_0, P_1, \delta) \leq SC_B(P_0, P_1, \delta).
\]

(4.3)

Table 4.1 provides numerical data on the highest achievable exponent of the probability of error for two Bernoulli variables with probabilities \( P_0 \) and \( P_1 = P_0 + \epsilon \), \( \epsilon \in [0; 1-P_0] \).
Table 4.1: The error exponent for Bernoulli variables $P_0$ and $P_1 = P_0 + \epsilon$

<table>
<thead>
<tr>
<th>$P_0$</th>
<th>$\epsilon$</th>
<th>$\delta$</th>
<th>$C(P_0, P_1)$</th>
<th>$l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.01</td>
<td>0.01</td>
<td>1.330972-04</td>
<td>49917</td>
</tr>
<tr>
<td>0.2</td>
<td>0.01</td>
<td>0.01</td>
<td>7.671175e-05</td>
<td>86608</td>
</tr>
<tr>
<td>0.3</td>
<td>0.01</td>
<td>0.01</td>
<td>5.897408e-05</td>
<td>112657</td>
</tr>
<tr>
<td>0.4</td>
<td>0.01</td>
<td>0.01</td>
<td>5.187552e-05</td>
<td>128073</td>
</tr>
<tr>
<td>0.5</td>
<td>0.01</td>
<td>0.01</td>
<td>5.00075e-05</td>
<td>132857</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.01</td>
<td>0.01012452</td>
<td>656</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.01</td>
<td>0.006745688</td>
<td>984</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.01</td>
<td>0.005531722</td>
<td>1201</td>
</tr>
<tr>
<td>0.4</td>
<td>0.1</td>
<td>0.01</td>
<td>0.00507677</td>
<td>1308</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>0.01</td>
<td>0.00507677</td>
<td>1308</td>
</tr>
</tbody>
</table>

4.2 Matrix Perturbation

We now introduce the notion of a matrix perturbation – a small change of the matrix entries – and the effect of such a perturbation on the matrix stationary node probabilities and its other characteristics.

**Definition 4.2.1** Let $M_0 = (V, \iota, Q, \Gamma, W)$ be a MC and $F$ be a $|V| \times |V|$ matrix. Matrix $F$ is an $\epsilon$-perturbation for $M$ if

(i) $\max_{i,j} |F_{i,j}| \leq \epsilon$ ;

(ii) $Q + F$ is a transition matrix, i.e. $\forall u,v \in V, 0 \leq (Q + F)_{u,v} \leq 1$ and $\forall u \in V, \sum_{v \in V} (Q + F)_{u,v} = 1$ .

We will abuse notation and denote the perturbed matrix by $M_1 = M_0 + F$. Figure 4.1 illustrates this definition with

$$F = \begin{bmatrix} 0 & 0.01 & -0.01 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Let $\sigma_i = \sigma(M_i)$ be the stationary node distribution of $M_i$. Table 4.2 shows the influence of the perturbation on the stationary node probabilities of the MC.

24
Next we discuss ways to bound the influence of a transition matrix perturbation on the stationary node probabilities of MC. Cho and Meyer [4] studied the influence of transition matrix perturbation on its stationary node probabilities in the following form

$$||\sigma_1 - \sigma_0|| \leq \kappa(Q_0)||F||.$$ (4.4)

The scalar $\kappa(Q_0)$ is called the condition number. A concrete choice of a specific norm in (4.4) is not significant, since all norms are equivalent up to a multiplicative constant [12, page 24, Example 5.1.3] and this constant may be incorporated into $\kappa(Q_0)$. We now define a specific condition number $\kappa(Q_0)$.

Let $M$ be a MC with transition matrix $Q$ and $A = I - Q$, where $I$ is the identity matrix. Let $\sigma$ be the vector of stationary state probabilities. The fundamental matrix is defined by

$$Z = (A + 1\sigma^T) ;$$

and the condition number:

$$\kappa(Q) = ||Z||_{\infty} .$$ (4.5)

<table>
<thead>
<tr>
<th>Node</th>
<th>$\sigma_0(V)$</th>
<th>$\sigma_1(V)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$u_2$</td>
<td>0.25</td>
<td>0.255</td>
</tr>
<tr>
<td>$u_3$</td>
<td>0.25</td>
<td>0.245</td>
</tr>
</tbody>
</table>

Figure 4.1: Perturbation of a MC

Table 4.2: Perturbation of a MC
Cho et al ([4]) show

$$||\sigma_0 - \sigma_1||_\infty \leq \kappa(Q)||F||_\infty.$$  \hfill (4.6)

The condition number of the MC $M_0$ in Figure 4.1 is 1.1481.

Cho et al ([4]) discuss alternative condition numbers. We chose (4.5) because it is relatively easy to compute.

**Definition 4.2.2** Matrices, whose condition numbers are bounded by a small constant, are called well-conditioned.

### 4.3 Information Theory

For a random variable $X$ let $\text{Dom}(X)$ denote its domain. For random variables $X$ and $Y$, their values $x \in \text{Dom}(X)$ and $y \in \text{Dom}(Y)$, let $p(x, y)$ be the joint probability mass function and $p(x)$, $p(y)$ the marginal probability mass functions. The *entropy* $H(X)$ of a discrete random variable $X$ determines the amount of information that $X$ has and it is defined as

$$H(X) = - \sum_{x \in \text{Dom}(X)} p(x) \log_2 p(x).$$

If $(X, Y)$ is distributed according to $p(x, y)$ then the conditional entropy $H(Y|X)$ determines the amount of information that $Y$ has when $X$ is known and it is defined as

$$H(Y|X) = \sum_{x \in \text{Dom}(X)} p(x) H(Y|X = x).$$

By the chain rule

$$H(X, Y) = H(X) + H(Y|X).$$

*The mutual information* $I(X; Y)$ between two random variables $X$ and $Y$ is defined as:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$  \hfill (4.7)

It can be shown that

$$I(X; Y) = H(X) - H(X|Y).$$

Next we define a Markov chain of random variables that will be used in the definition of *Data-Processing inequality* below.
**Definition 4.3.1** Random variables $X,Y,Z$ are said to form a Markov chain (denoted by $X \rightarrow Y \rightarrow Z$) if the conditional distribution of $Z$ depends only on $Y$ and is conditionally independent of $X$. Specifically

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

Theorem 4.3.2 demonstrates that no processing of $Y$, deterministic or random, can increase the information that $Y$ contains about $X$.

**Theorem 4.3.2** *(Data-processing inequality, [5, Theorem 2.8.1])* If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$. The equality $I(X; Y) = I(X; Z)$ holds if and only if $I(X; Y|Z) = 0$, i.e. $X \rightarrow Z \rightarrow Y$.

**Corollary 4.3.3** If $Z = g(Y)$, we have $I(X; Y) \geq I(X; g(Y))$.

Suppose we have a family of probability mass functions $\{f_\Theta(x)\}$, indexed by $\Theta$, $X$ is a sample from a distribution in this family and $T(X)$ be any statistic (function of the sample). Then $\Theta \rightarrow X \rightarrow T(X)$ and by the data processing inequality $I(\Theta; T(X)) \leq I(\Theta; X)$. Next we define sufficient statistic [5]:

**Definition 4.3.4** A function $T(X)$ is said to be a sufficient statistic relative to the family $\{f_\Theta(x)\}$ if $X$ is independent of $\Theta$ given $T(X)$, i.e. $\Theta \rightarrow T(X) \rightarrow X$ forms a Markov chain.

Intuitively a sufficient statistic $T(X)$ for $\Theta$ contains all the information in $X$ about $\Theta$. 


Chapter 5

Sample Complexity of Markov Chains

We now develop sample complexity bounds on the problem of distinguishability between MCs (Problem 2, 3). We also define the problem of learning MCs and develop the appropriate sample complexity bounds.

5.1 Definitions

Let us start with defining distinguishability and learnability of MCs for an arbitrary distance between MCs.

Consider paths of length $l$ over states $V$ of a MC. Instead of defining a subset of paths of positive probability, we will work with the entire set of all paths of length $l$ $V^l$, i.e., the cross product $V \times V \times \cdots \times V$ $l$ times. This set contains all positive probability paths as well as paths of probability 0.

**Definition 5.1.1** Let $M_0$ and $M_1$ be two MCs. A decision function is a function $DF : V^l \rightarrow \{0, 1\}$. We say that $DF$ decides that $s$ was produced by $M_0$ if and only if $DF(s) = 0$.

Let us recall the definition of the error probability (3.3). A decision function $DF$ is a function, whose input is a sequence $s$ of length $l$ and whose output is in $\{0, 1\}$. For $i = 0, 1$ let $Hyp_i$ be the hypothesis that $s$ was produced by $M_i$. Its error probability is defined as
follows

\[ \text{error}^{(l)}_{DF}(M_0, M_1) = \text{prior}_0 \cdot P(DF(s) = 1|Hyp_0) + \text{prior}_1 \cdot P(DF(s) = 0|Hyp_1). \]

Next we define distinguishability between MCs.

**Definition 5.1.2** Let \(\epsilon, \delta \in (0, 0.5), l \in \mathbb{N}^+\), \(\text{dist}\) be a distance between MCs and let \(M_0, M_1\) be two MCs, such that \(\text{dist}(M_0, M_1) \geq \epsilon\).

The MCs \(M_0\) and \(M_1\) are \((l, \epsilon, \delta)_{\text{dist}}\)-distinguishable if there exists a decision function \(DF\) such that \(\text{error}^{(l)}_{DF}(M_0, M_1) < \delta\).

Let \(\mathcal{M}^{[n]}\) be the set of MCs with \(n\) states.

\[ \mathcal{M}^{[n]} = \{ M : M \text{ is a MC with } n \text{ states} \}. \]

Next we introduce the notion of a learning function and learnability of MCs.

**Definition 5.1.3** Let \(l, n \in \mathbb{N}^+, \epsilon, \delta \in (0, 0.5)\), \(V\) be a set of \(n\) states and \(\text{dist}\) be a distance between MCs. The class \(\mathcal{M}^{[n]}\) is \((l, n, \epsilon, \delta)_{\text{dist}}\)-learnable if there exists a function \(LA : V^{l+1} \times \mathbb{N}^+ \times (0, 0.5) \times (0, 0.5) \to \mathcal{M}^{[n]}\), such that for any MC \(M_0 \in \mathcal{M}^{[n]}\), given a path \(s\) of length \(l\), chosen according to \(M_0, \epsilon, \delta \in (0, 0.5)\), LA returns a MC \(M_1\) such that \(P(\text{dist}(M_0, M_1) < \epsilon) > 1 - \delta\).

In this case we say that \(M_1 = LA(s, n, \epsilon, \delta)\) is learned based on a path \(s\), and LA is a learning function.

### 5.2 Learnability vs. Distinguishability

In this section we develop lower sample complexity bounds for an arbitrary distance between MCs. First we show that learning MCs is at least as hard as distinguishing between them.

**Theorem 5.2.1** Let \(\text{dist}\) be a pseudo-distance between MCs \(^1\) Also let \(\epsilon, \delta \in (0, 0.5), n \in \mathbb{N}, M_0, M_1 \in \mathcal{M}^{[n]}\), such that \(\text{dist}(M_0, M_1) \geq \epsilon\) and \(l \in \mathbb{N}\) be such that \(M_0\) and \(M_1\) are not \((l, \epsilon, \delta)_{\text{dist}}\)-distinguishable. Then the class \(\mathcal{M}^{[n]}\) is not \((l, n, \epsilon/2, \delta)_{\text{dist}}\)-learnable.

\(^1\)A pseudo-distance satisfies the axioms of a distance, except that \(\text{dist}(M, M') = 0\) does not entail \(M = M'\).
Proof We show a reduction from the problem of MC learnability to the problem of MC distinguishability: we will assume that the class $M[n]$ is $(l, n, \epsilon/2, \delta)_{dist}$-learnable and show that in this case $M_0$ and $M_1$ are $(l, \epsilon, \delta)_{dist}$-distinguishable.

Assume that $M[n]$ is $(l, n, \epsilon/2, \delta)_{dist}$-learnable. Then there exists a learning function $LA$ that realizes $(l, n, \epsilon/2, \delta)_{dist}$-learnability, meaning that it receives a sequence $s$ of length $l$, distributed according to a MC $M \in M[n]$, error $\epsilon/2$ and confidence $\delta$ and with confidence $\delta$ returns a MC $\tilde{M}$ such that $dist(M, \tilde{M}) < \epsilon/2$.

Next we describe a decision function $DF$ that uses $LA$ to achieve $(l, \epsilon, \delta)_{dist}$-distinguishability. The input of the function is a sequence $s$ of length $l$, drawn according to either $M_0$ or $M_1$. The function applies $LA$ with input $(s, n, \epsilon/2, \delta)$, resulting in MC $M_2$. If $dist(M_2, M_0) < \epsilon/2$, the function returns 0. Otherwise it returns 1. Now we analyze the error probability

Figure 5.1: Edge learnability, assuming that $s$ is drawn from $M_1$.

of the function above. Let $j \in \{0, 1\}$ be the index of the MC, from which $s$ is drawn (see Figure 4.1). From the definition of $LA$, with confidence $(1 - \delta)$ $dist(M_2, M_j) < \epsilon/2$. Since from the assumptions of this lemma $dist(M_0, M_1) \geq \epsilon$, and $dist$ obeys triangular inequality, it follows from the triangle inequality that $dist(M_2, M_{1-j}) > \epsilon/2$. Thus, the decision of the procedure above is correct with confidence $(1 - \delta)$. 

Theorem 5.2.1 provides the sample complexity bound on learning MCs in the class $M[n]$, given non-distinguishability of two MCs in this class. Next we discuss how Theorem 5.2.1 may be applied to compute the sample complexity bound of a given MC. Let us recall Definition 4.2.1 of a matrix perturbation. First we define a specific perturbation, that effects only two edges of the MC. 

30
Definition 5.2.2 Let \( M_0 = (V, \iota, Q, \Gamma, W) \) be a MC, \( \epsilon \in (0, 0.5) \), \( e_0, e_1 \in E \) be two edges that emanate from the same state, such that \( Q(e_0) + \epsilon \leq 1, Q(e_1) - \epsilon \geq 0 \). The \( F_\epsilon(e_0, e_1) \)-perturbation is a \(|V| \times |V|\) matrix that has \( \epsilon \) in entry \( e_0 \), \( -\epsilon \) in entry \( e_1 \) and 0 in all the other entries.

Corollary 5.2.3 Given a MC \( M_0 \in \mathcal{M}[n] \), choose a pair of edges \( e_0, e_1 \in E \) that emanate from the same node, \( \epsilon, \delta \in (0, 0.5) \). Let \( M_1 = M_0 + F_\epsilon(e_0, e_1), l \in \mathbb{N} \) be such that \( M_0 \) and \( M_1 \) are not \((l, \epsilon, \delta)_{dist}\)-distinguishable. Then \( M_0 \) is not \((l, n, \epsilon/2, \delta)_{dist}\)-learnable.

5.3 Distances Examples

The previous section discussed arbitrary pseudo-distances. We now introduce two specific distances, and compare them in the context of distinguishability and learnability of MCs.

Definition 5.3.1 Let \( M_0 \) and \( M_1 \) be two MCs with the same number of states. Then the transition edge distance between \( M_0 \) and \( M_1 \) is defined as
\[
\text{dist}_Q(M_0, M_1) = \max_{e \in E} |Q_1[e] - Q_0[e]|
\]
and the stationary edge pseudo-distance is defined as
\[
\text{dist}_\sigma(M_0, M_1) = \max_{e \in E} |\sigma_1[e] - \sigma_0[e]|.
\]

When learning MCs, one may want to learn either transition or stationary edge probabilities. As the example of brand loyalty in the Introduction shows, learning stationary edge probabilities is usually more interesting from a practical point of view.

Now, when we have two approaches to learning MCs, we clarify what approach is simpler in terms of sample complexity. Lemma 5.3.2 shows that learning stationary probabilities implies learning transition probabilities.

Lemma 5.3.2 Let \( M_0 \) and \( M_1 \) be two ergodic MCs that share the set of the states and the set of edges of positive stationary probabilities. Let \( \epsilon \in (0, \frac{1}{2}) \) and let us assume that for each edge \( e \in V^2 \) with positive stationary probability, i.e. \( \sigma_0(e) > 0 \), the following condition holds
\[
\frac{\sigma_0(e)}{\sigma_1(e)} \in (1 - \epsilon, 1 + \epsilon).
\]
Then for each $e \in E$ with positive stationary probability, i.e. $\sigma_0(e) > 0$, the following conditions hold

(i) $\frac{\sigma_0[u]}{\sigma_1[u]} \in (1 - \epsilon, 1 + \epsilon)$;

(ii) $\frac{Q_0(e)}{Q_1(e)} \in ([1 - \epsilon]^2, [1 - \epsilon]^{-2})$.

**Proof** First we note that since we assume that MCs are ergodic, for all $u \in V$, $\sigma_i[u] > 0, i = 0, 1$. Hence for any edge $e$, $\sigma_i[e] > 0$ iff $Q_i[e] > 0, i = 0, 1$. Recall that for any vertex $v$, $\sigma_i[u,v] = \sigma_i[u]Q_i[u,v]$. Note that the denominator of (5.1) $\sigma_1[e]$ must be positive, implying $Q_1[u,v] > 0$. Equation (5.1) implies $\sigma_0[u,v] > 0$, consequently $Q_0[u,v] > 0$.

Since for $i = 0, 1$: $(Q_i(u,v))_v$ are distributions, $\sum_v Q_i(u,v) = 1$. Thus $\max \frac{Q_0(u,v)}{Q_1(u,v)} \geq 1$ and $\min \frac{Q_0(u,v)}{Q_1(u,v)} \leq 1$. Let the minimum and maximum value of the fraction be obtained at $v^-$ and $v^+$:

$$
\sigma_0[u] = \sigma_0(u, v^-)/Q_0(u, v^-) = \sigma_0(u, v^-) \frac{Q_0(u, v^-)}{Q_1(u, v^-)}^{-1} \geq \frac{\sigma_0(u, v^-)}{\sigma_1(u, v^-)} > 1 - \epsilon
$$

$$
\sigma_0[u] = \sigma_0(u, v^+)/Q_0(u, v^+) = \sigma_0(u, v^+) \frac{Q_0(u, v^+)}{Q_1(u, v^+)}^{-1} < \frac{\sigma_0(u, v^+)}{\sigma_1(u, v^+)} < 1 + \epsilon
$$

To prove (ii):

$$
\frac{Q_0(e)}{Q_1(e)} = \frac{Q_0(e)\sigma_0[u]}{Q_1(e)\sigma_1[u]} \frac{\sigma_0[u]}{\sigma_1[u]}^{-1} = \frac{\sigma_0(e)}{\sigma_1(e)} \frac{(\sigma_0[u])^{-1}}{(\sigma_1[u])^{-1}}
$$

From (5.1) and (i)

$$
\frac{Q_0(e)}{Q_1(e)} = \frac{\sigma_0(e)}{\sigma_1(e)} \frac{\sigma_0[u]}{\sigma_1[u]}^{-1} < \frac{1 + \epsilon}{1 - \epsilon} < (1 - \epsilon)^{-2}.
$$

Lemma 5.3.2 shows that learning stationary probabilities implies learning transition edge probabilities. This means that learning transition probabilities is at least as hard as learning stationary probabilities. For this reason learning transition probabilities is sufficient. However in certain cases learning transition probabilities may not be necessary, as in the example, drawn in Figure 5.2.

32
Figure 5.2 illustrates a MC and Table 5.1 shows its stationary node probabilities for \( n = 5 \). The structure of this MC is a chain, where for each node, except for the last one, traversing the next edge returns the MC to \( u_1 \) with probability 0.5. Thus, if the chain was infinitely long, this property would hold for every node. For an infinite chain, the stationary node probability of \( u_i \) is \( \sigma[u_i] = 0.5^i \). Hence edges at the end of this chain are hardly visited.

![Figure 5.2: Node and edge stationary probabilities](image)

<table>
<thead>
<tr>
<th>Node ( u_i )</th>
<th>( u_1 )</th>
<th>( u_2 )</th>
<th>( u_3 )</th>
<th>( u_4 )</th>
<th>( u_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stationary node probabilities</td>
<td>0.5161</td>
<td>0.2581</td>
<td>0.1290</td>
<td>0.0645</td>
<td>0.0323</td>
</tr>
</tbody>
</table>

First we discuss the difficulty of learning the transition probabilities of this MC. Note that to estimate \( Q(u_{n-1}, u_n) \) we need to visit node \( u_{n-1} \) more than once. However since the stationary probability of node \( u_{n-1} \) is \( 2^{-(n-1)} \), an edge sequence of length \( 2^{n-1} \), chosen according to this MC, visits this node on average only once. Hence, it is impossible to estimate its transition probabilities, and thus it is easy to show a lower bound on learning the transition edge probabilities of this MC.

Assume we need to learn stationary edge probabilities of the edges in the end of this chain up to the distance \( \epsilon \). Note that the stationary edge probabilities of these edges are exponentially close to 0. Thus for nodes \( u_i, i > \log \frac{1}{\epsilon} \), to reach the above goal it is sufficient to assign any value in \([0,1]\) as the learned value of the stationary edge probabilities of these edges. This means that to learn stationary edge probabilities at the end of this chain, there is no need for a long training sequence, in turn meaning that learning stationary edge probabilities of these edges requires less training data than learning their transition edge probabilities. A similar phenomenon occurs with PAC-learning – an event of low probability does not require many training samples, since the training error, associated
with this event, is weighted by the probability of this event.

5.4 Transition Probabilities

Distinguishing between MCs

In this section we develop the main results for distinguishability of MCs. Let us recall Definition 4.2.1 of an \( \epsilon \)-perturbation of a MC. Let \( M_0 \) be a MC, \( F \) be an \( \epsilon \)-perturbation of MC and \( M_1 \) the perturbated matrix \( M_1 = M_0 + F \). We shall identify \( F \) with the edges corresponding to its non-zero entries, i.e., \( f = (u, v) \in F \) iff \( F_{u,v} \neq 0 \). Let \( m = |F| \) and let \( f = (f_0, \ldots, f_{m-1}) \) be a fixed ordering of the edges of \( F \). Thus \( F = F(e) \). Let \( n(f_j|s) \) be the number of times \( f_j \) appears in \( s \) and \( n(f|s) = (n(f_0|s), \ldots, n(f_{m-1}|s)) \).

Let \( s \) be a path of length \( l \), chosen according to \( M_i \). Then following (3.3), the probability of choosing \( s \) is defined as follows:

\[
P^{[l]}(s) = P^{[l]}(s|M_0)\text{prior}_0 + P^{[l]}(s|M_1)\text{prior}_1 .
\]

Let \( \text{Ind} \) be a random variable that returns the index \( i \in \{0, 1\} \) of the hypothesis \( M_i \) with prior probability \( \text{prior}_i \), as defined in (3.3). Let \( k \in \mathbb{N}^m \) and \( Z(s) \) be any random variable that is defined on \( s \). We will show that the information in \( Z(s) \) about \( \text{Ind} \) is incorporated in \( n(f|s) \) and thus cannot improve the accuracy of an optimal decision function.

**Lemma 5.4.1** Let \( M_1 = M_0 + F(e) \). Then \( n(f|s) \) is a sufficient statistic for \( \text{Ind} \), i.e. \( \text{Ind} \rightarrow n(f|s) \rightarrow s \) forms a Markov chain.

**Proof** Consider the probability space of paths of length \( l \). Let \( s \) be such a path, chosen according to either \( M_0 \) or \( M_1 \). To show that \( \text{Ind} \rightarrow n(f|s) \rightarrow s \), we need to show that if \( n(f|s) \) is kown, then \( s \) is independent of \( \text{Ind} \). For a sequence of \( m \) integers \( k = (k_0, k_1, \ldots, k_{m-1}) \in \mathbb{N}^m \), consider the conditional probabilities \( P(s|n(f|s) = k, M_i) \), i.e., the probability that \( M_i \) produced the path \( s \) subject to the condition that the path traversed \( f_j \) \( k_j \) times, \( 0 \leq j \leq m - 1 \). Let

\[
P(k|M_i) = P((k_0, \ldots, k_{m-1})|M_i) = \sum_{t \in V^{(l+1)}: n(f|t) = n(f|s)} P(t|M_i) .
\]
Then

\[ P(s|M_i) = \sum_{k \in \mathbb{N}_m} P(s|n(f|s) = k)P(k|M_i). \]

Since for \( k \neq n(f|s) \), \( P(s|n(f|s) = k, M_i) = 0 \), it follows that

\[ P(s|M_i) = P(s|n(f|s) = k, M_i)P(n(f|s) = k|M_i). \]

From the expression above

\[ P(s|n(f|s) = k, M_i) = \frac{P(s|M_i)}{P(n(f|s) = k|M_i)} \]

\[ = \frac{\ell(start(s)) \prod_{f \in F} Q_1(f)^{n(f|s)} \prod_{e \in F} Q_1(e)^{n(e|s)}}{\sum_{t \in V^{(l+1)}:n(f|t) = n(f|s)} \ell(start(t)) \prod_{f \in F} Q_1(f)^{n(f|t)} \prod_{e \in F} Q_1(e)^{n(e|t)}} \]

\[ = \frac{\ell(start(s)) \prod_{e \in F} Q_0(e)^{n(e|s)}}{\sum_{t \in V^{(l+1)}:n(f|t) = n(f|s)} \ell(start(t)) \prod_{e \in F} Q_0(e)^{n(e|t)}} \]

Since for \( e \notin F \), \( Q_0(e) = Q_1(e) \), we have

\[ P(s|n(f|s) = k, M_i)) = \frac{\ell(start(s)) \prod_{e \in F} Q_0(e)^{n(e|s)}}{\sum_{t \in V^{(l+1)}:n(f|t) = n(f|s)} \ell(start(t)) \prod_{e \in F} Q_0(e)^{n(e|t)}} = P(s|n(f|s) = k, M_0). \]

We conclude that when \( n(f|s) \) is known, \( s \) is independent of \( Ind \), which implies \( Ind \rightarrow n(f|s) \rightarrow s \).

**Corollary 5.4.2** \( I(Ind; n(f|s)) \geq I(Ind; Z(s)) \), where \( I \) is the mutual information, as defined by (4.7).

One of the central tools in our analysis is **strong typicality**. Next we introduce this tool in the form of strongly typical edge sequences.

**Definition 5.4.3** Let \( \theta \in [0, 1] \). A sequence \( s \) of length \( l \) in a MC \( M_i \) is \( \theta \)-strongly typical if for every edge \( e \in E \), the average number of times \( e \) is traversed in \( s \) is \( \theta \)-close to \( \sigma(e) \):

\[ |n(e|s)/l - \sigma(e)| < \theta. \]

(5.2)
Let $\tau_i(\theta, l)$ denote the set of $\theta$-strongly typical sequences of length $l$ in $M_i$. To shorten the notation we will write $\tau_i$ when $\theta$ and $l$ is understood from the context.

Let $M$ be a MC and let us examine the set of sequences that $M$ produces. The Asymptotic Equipartition Property of MCs (Theorem A.1.1) states that the probability of the set of strongly typical sequences $\tau(\theta, l)$ of $M$ approaches 1 as $l$ goes to infinity

$$\lim_{l \to \infty} \sum_{s \in \tau(\theta, l)} P(s) = 1$$

and all the sequences in $\tau(\theta, l)$ have similar probability. The latter property follows from (5.2). Figure 5.3 demonstrates this property. Equation (5.3) implies that for sufficiently long sequences produced by a MC, we can assume that we observe only strongly typical sequences.

Next we develop a computationally efficient lower bound on the probability of strongly typical sequences.

Let $M$ be a MC. Recall that $\iota$ denotes a column vector of initial probabilities of a MC. Let $e, e' \in E$. For any $\alpha > 0$ let the matrix $A^{(\alpha, e)}$ be

$$A^{(\alpha, e)}(e') = \begin{cases} Q(e') & e' \neq e \\ Q(e') \exp(\alpha) & e' = e \end{cases}$$

Note that for $\alpha > 0$, $A^{(\alpha, e)}$ is not a stochastic matrix, since row $e$ does not sum up to 1, due to the multiplicative factor $\exp(\alpha)$ in the entry $e$. Recall that $1 = (1, ..., 1)^{tr}$. Let
\( \alpha > 0, \text{sign} \in \{-1,1\} \) and let \( \pi_M(\tau(\theta,l)(\theta),\alpha,\text{sign},e) \) be the scalar
\[
\pi_M(\theta,l,\alpha,\text{sign},e) = \exp(-\text{sign} \cdot \alpha l(\sigma(e) + \text{sign} \cdot \theta))1^{tr}(A^{(\text{sign} \cdot \alpha,e)})1 .
\] (5.4)

Let \( S^c \) denote the complement of a set \( S \). The following lemma is proved in Appendix A.2.

**Lemma 5.4.4** Let \( M \) be an MC and \( s = (e_1,e_2,\cdots,e_l) \) be a path, chosen according to \( M \) and \( e \in E \). Then
\[
P\left( \left| \frac{1}{l} p(e|s) - \sigma(e) \right| > \theta \right) \leq \min_{\alpha>0} \pi_M(\theta,l,\alpha,1,e) + \min_{\alpha>0} \pi_M(\theta,l,\alpha,-1,e) .
\] (5.5)

Note that (5.5) states that the condition (5.2) of the strongly typical sequences does not hold for a specific edge \( e \). A state sequence is not strongly typical if at least one of its edges violates the condition (5.2). We formalize this reasoning in the following corollary.

**Corollary 5.4.5** Let
\[
\pi_M(\theta,l) = 1 - \sum_{e \in V \times V} \left( \min_{\alpha>0} \pi_M(\theta,l,\alpha,1,e) + \min_{\alpha>0} \pi_M(\theta,l,\alpha,-1,e) \right) .
\] (5.6)

The set of \( \theta \)-strongly typical sequences of length \( l \) satisfies
\[
P\left( (\tau(\theta,l))^c \right) \leq \sum_{e \in V \times V} \min_{\alpha>0} \pi_M(\theta,l,\alpha,1,e) + \min_{\alpha>0} \pi_M(\theta,l,\alpha,-1,e)
\]
or alternatively
\[
P\left( \tau(\theta,l) \right) \geq \pi_M(\tau(\theta,l)) .
\] (5.7)

Note that Corollary 5.4.5 provides the means to calculate the lower bound on the probability of strongly typical sequences. However there is no guarantee that this bound converges to 1 as the sequence length approaches infinity. For the sake of completeness in Appendix A.1 we provide another lower bound on the probability of strongly typical sequences that does converge to 1 as the length of strongly typical sequences approaches infinity. We still prefer to use the bound (5.7), since in practice it converges to 1 and, as we note in the end of this appendix, it has exponential convergence rate, while the convergence rate of the bound from Appendix A.1 is linear in the length of strongly typical sequences.

37
In Section 7.2 we apply our bounds to a MC, whose transition matrix is shown in Table 7.2. Figures 5.4 and 5.5 show the values of the lower bound of Lemma 5.4.4 for edge $(3, 2)$ of this MC. Figure 5.4 shows expression (5.4) with $\text{sign} = 1$ as a function of $\alpha$ for sequences of lengths 1,500, 2,500 and 3,500. Figure 5.5 shows this expression with $\text{sign} = -1$ as a function of $\alpha$ for sequences of lengths 500, 1,000 and 2,000. Figure 5.6 shows the probability of strongly typical sequences of that MC as a function of sequence length.
Figure 5.4: The part (5.4) of the upper bound (5.5) on the probability of non-strongly typical sequences for sign=1, $\delta = 0.1$ and $\theta = 0.01$
The upper bound
The part of the upper bound on the probability of non-strongly typical sequences for sign=1

Figure 5.5: The part (5.4) of the upper bound (5.5) on the probability of non-strongly typical sequences for sign=-1, δ = 0.1 and θ = 0.01
Let us recall Problem 2, where a sequence $s$ of edges of length $l$ is chosen either...
according to the MC $M_0$ with prior probability $\text{prior}_0$ or according to the MC $M_1$ with prior probability $\text{prior}_1$. Thus the probability of this sequence is

$$P(s) = P(s|M_0)\text{prior}_0 + P(s|M_1)\text{prior}_1.$$  

Furthermore, the probability of the set of sequences $\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)$ is

$$P(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)) = \begin{cases} P(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)|M_0)\text{prior}_0 + \text{prior}_0, \\ P(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)|M_1)\text{prior}_1. \end{cases}$$  

From Lemma 5.4.4 we get the following corollary:

**Corollary 5.4.6** Let $M_0$ and $M_1 = M_0 + F$ be two MCs. Then

$$P(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)) \geq \max_{i=0,1} \{\pi_{M_i} (\theta, l) \cdot \text{prior}_i\},$$  

where $\pi_{M_i}$ is defined in (5.6).

Above we used a perturbation $F$ with arbitrary set of non-zero entries. Now we define a specific perturbation.

**Definition 5.4.7** Let $M$ be a MC, $\epsilon \in (0, 0.5)$ and $e_0, e_1 \in V^2$ be two edges such that they share the source state and $Q(e_0) + \epsilon \leq 1, Q(e_1) - \epsilon \geq 1$. An $F_\epsilon(e_0, e_1)$ perturbation is defined by means of a perturbation matrix $F_\epsilon(e_0, e_1)$, with 0 in all entries, except for the entries for $e_0$ and $e_1$, in which $F_\epsilon(e_0, e_1)[e_0] = \epsilon$ and $F_\epsilon(e_0, e_1)[e_1] = -\epsilon$.

Let

$$K(e_0, e_1) = \max_{i=0,1} \{\sigma_i(e_0) + \sigma_i(e_1)\}.$$  

From the definition of strongly typical sequences, the expression $(K(e_0, e_1) + 2\theta)l$ is an upper bound on the number of times a strongly typical sequence $s \in \tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)$ passes through either $e_0$ or $e_1$.

Next we provide sufficient conditions on MC non-distinguishability. We construct a reduction from the problem of distinguishing between MCs to the problem of distinguishing between random variables. To construct the reduction we choose a sequence $s$, distributed according to either a MC $M_0$ or $M_1 = M_0 + F_\epsilon(e_0, e_1)$, and observe the events when $s$ traverses either $e_0$ or $e_1$. These events will be =a random variable with two possible
outcomes: pass over $e_0$ or pass over $e_1$. Then we use Lemma 4.1.4 to derive the bound on the length of any sequence $s$ needed to distinguish between the two MCs.

The following lemma provides conditions on $l$-edge non-distinguishability.

**Lemma 5.4.8** Let $\epsilon, \delta_B \in (0,0.5)$, $M_0$ be a MC and $M_1 = M_0 + F_\epsilon(e_0, e_1)$, such that $\epsilon \leq \min \{1 - Q_0(e_0), Q_0(e_1)\}$. Let

$$P_{B,i} = \frac{Q_i(e_0)}{Q_i(e_0) + Q_i(e_1)}, \quad (5.11)$$

and define

$$SC_M(M_0, \epsilon, \delta_B, e_0, e_1) = \frac{SC_B(P_{B,0}, P_{B,1}, \delta_B)}{K(e_0, e_1) + 2\theta}. \quad (5.12)$$

Also let

$$l \leq SC_M(M_0, \epsilon, \delta_B, e_0, e_1), \quad (5.13)$$

and

$$\delta = \delta_B \cdot P(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)). \quad (5.14)$$

Then $M_0$ and $M_1$ are not $(l, \delta, \epsilon)$-distinguishable.

**Proof** For an event $A$, let $\text{error}^{|l|}_{DF}(M_0, M_1)|_A$ be the probability of the decision error for the distance $\text{dist}$ given the event $A$.

The error of a decision function can be decomposed as follows:

$$\text{error}^{|l|}_{DF}(M_0, M_1) = \text{error}^{|l|}_{DF}(M_0, M_1)|_{\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)} P(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)) + \text{error}^{|l|}_{DF}(M_0, M_1)|_{(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1))}^C P((\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1))^C).$$

Since we are interested in a lower bound on the decision error:

$$\text{error}^{|l|}_{DF}(M_0, M_1) \geq \text{error}^{|l|}_{DF}(M_0, M_1)|_{\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)} P(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)) \quad (5.15)$$

We get the lower bound on $P(\tau(\theta, l|0) \cup \tau(\theta, l|1))$ from Corollary 5.4.6.

**Lower bound on** $\text{error}^{|l|}_{DF}(M_0, M_1)|_{s \in \tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)}$: 

43
When \( s \) traverses node \( u = \text{start}(e_0) \) in \( M_i \) \((i = 0, 1)\), it leaves \( u \) by one of its outgoing edges. Let us restrict our attention to the events in which \( s \) leaves \( u \) by either \( e_0 = (u, v_0) \) or \( e_1 = (u, v_1) \). These events form a sequence \( S = (X_1, ..., X_m) \) of i.i.d. Bernoulli random variables \( X_j \in \{0, 1\} \), where \( m \leq l \). The outcome of \( X_j \) is 0 if at the \( j^{th} \) time that \( s \) leaves \( u \) by either \( e_0 \) or \( e_1 \), it leaves by \( e_0 \) and 1 if it leaves by \( e_1 \).

According to Lemma 5.4.1, \( n(e_0, e_1|s) \) is a sufficient statistics for \( \text{Ind} \). Since \( n(e_0, e_1|s) \) can be calculated from \( S \), \( S \) is also a sufficient statistics for \( \text{Ind} \). Thus to distinguish between \( M_0 \) and \( M_1 \), it is sufficient to analyze \( S \). Since we assume that \( s \in \tau(\theta, l|M_0) \cup \tau(\theta, l|M_1) \), \(|S| \leq l(K(e_0, e_1) + 2\theta)\). Thus, if we choose \( l \) such that

\[
l(K(e_0, e_1) + 2\theta) < LB_B(P_{B,0}, P_{B,1}, \delta_B)
\]

or

\[
l < \frac{LB_B(P_{B,0}, P_{B,1}, \delta_B)}{(K(e_0, e_1) + 2\theta)},
\]

the following inequalities hold

\[
|S| \leq l(K(e_0, e_1) + 2\theta) < LB_B(P_{B,0}, P_{B,1}, \delta_B).
\]

and, according to Lemma 4.1.4, \( P_{B,0} \) and \( P_{B,1} \), and thus \( M_0 \) and \( M_1 \) are not distinguishable with error \( \delta_B \). Consequently when (5.16) holds, the error of any decision function satisfies

\[
\text{error}^{[\text{DF}]}(M_0, M_1)|_{\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)} > \delta_B.
\]

According to (5.15) the decision error is at least

\[
\text{error}^{[\text{DF}]}(M_0, M_1) > P(\tau(\theta, l|M_0) \cup \tau(\theta, l|M_1)) \cdot \delta_B.
\]

Note that since we use bound (5.9), the statement of Lemma 5.4.8 does not depend on the initial node probabilities of the MC \( M_0 \), meaning that our bounds do not depend on them. For the sake of completeness in Appendix B we also show that given any initial node probabilities, the node probabilities of any MC converge to the stationary node probabilities exponentially fast.
Distinguishability of MCs: An Example

Table 5.2 shows the bound of Lemma 5.4.8 on the error of distinguishing between the MC of Figure 5.2 from its \((u_4, u_1), (u_4, u_5)\)-0.01-perturbation. It shows a tradeoff between the values of \(\theta\) that controls the tightness of strongly typical sequences and the bound on their length and finally the error of distinguishing between MCs. For smaller values of \(\theta\) we get error bounds for longer strongly typical sequences at the expense of getting worse bounds on the error to distinguish MCs. For example for \(\theta = 0.5\), the error bound is for length 3 and the error is 0.077, while for \(\theta = 0.1\), the length is 12 and the error bound drops to 0.012.

Table 5.2: Error to distinguish MC in Figure 5.2 from its perturbation

<table>
<thead>
<tr>
<th>(\theta)</th>
<th>(l)</th>
<th>(\delta)</th>
<th>(\delta')</th>
<th>(\pi(\tau(\theta, l_0)))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>3.12</td>
<td>0.1</td>
<td>0.077</td>
<td>0.766</td>
</tr>
<tr>
<td>0.4</td>
<td>3.84</td>
<td>0.1</td>
<td>0.059</td>
<td>0.588</td>
</tr>
<tr>
<td>0.3</td>
<td>5</td>
<td>0.1</td>
<td>0.049</td>
<td>0.486</td>
</tr>
<tr>
<td>0.2</td>
<td>7.15</td>
<td>0.1</td>
<td>0.035</td>
<td>0.351</td>
</tr>
<tr>
<td>0.1</td>
<td>12.56</td>
<td>0.1</td>
<td>0.012</td>
<td>0.121</td>
</tr>
<tr>
<td>0.05</td>
<td>20.19</td>
<td>0.1</td>
<td>0.0032</td>
<td>0.0032</td>
</tr>
</tbody>
</table>

Learning MCs

Note that from the assumptions of Lemma 5.4.8 \(dist_Q(M_0, M_1) \geq \epsilon\). Thus we can apply Theorem 5.2.1 to get the following corollary

**Corollary 5.4.9** Let \(l \in \mathbb{N}^+, \epsilon, \delta_B \in (0, 0.5)\). If there exists \(M_0 \in \mathcal{M}^{[n]}\) and \(e_0, e_1 \in E\) such that (5.13) holds, then for \(\delta\), given by (5.14), the class \(\mathcal{M}^{[n]}\) is not \((l, n, \epsilon/2, \delta)_{dist_Q}\)-learnable.

Again, since the bound of Lemma 5.4.8 does not depend on the initial node probabilities of the MC, the learnability bounds also do not depend on those probabilities.

Above we proved the distinguishability bound (Lemma 5.4.8) and the learnability bound (Corollary 5.4.9) for a specific pair of edges \(e_0\) and \(e_1\). We can improve the result considering all pairs of edges that emanate from the same node. The following theorem generalizes Corollary 5.4.9:
**Theorem 5.4.10** Let $\epsilon, \delta'$ and $M_0$ be as in Lemma 5.4.8 and let

\[
l = \max \{ SCM(M_0, \epsilon, \delta, (u, v_0), (u, v_1)) : u, v_0, v_1 \in V, v_0 \neq v_1, Q_0(u, v_0) \leq 1 - \epsilon, Q_0(u, v_1) \geq \epsilon \}.
\]

Then the class $M^{[n]}$ is not $(l, n, \epsilon/2, \delta)_{distQ}$-learnable.

Figure 5.7 exemplifies the bound of Theorem 5.4.10 for MC, whose transition probability matrix is given in Table 7.2.
Figure 5.7: The sample complexity bound on learning MC, whose transition matrix is given in Table 7.2
5.5 Stationary Edge Probabilities

Theorem 5.4.10 showed a lower complexity bound on learning transition edge probabilities of MCs. Next we extend this learnability bound to learning stationary edge probabilities. We start with Lemma 5.5.1 that provides us with the means to develop a lower bound on the sample complexity of learning stationary edge probabilities of MCs.

Lemma 5.5.1 Let $M_0$ be a MC, $e_0 = (u,v_0), e_1 = (u,v_1) \in E$ be two edges of $M_0$, $F_\epsilon(e_0,e_1)$ be an $\epsilon$-perturbation matrix and $M_1 = M_0 + F_\epsilon(e_0,e_1)$ the perturbated matrix. Also let

$$
\epsilon_\sigma = \epsilon_\sigma(M_0, e_0, e_1, \epsilon) = \epsilon \min_{i=0,1} \{ \sigma_0[u] - 2\kappa(Q_0) \left(Q_0(e_i) + (-1)^i\epsilon\right) \} .
$$

Then

$$
|\sigma_1(e_i) - \sigma_0(e_i)| \geq \epsilon_\sigma, \ i = 0, 1 .
$$

If

$$
\sigma_0[u] > \max_{i=0,1} 2\kappa(Q_0)(Q_0(e_i) + (-1)^i\epsilon) ,
$$

then

$$
\epsilon_\sigma > 0 .
$$

Proof Note that (5.19) follows from (5.17) and (5.18).

We prove this lemma separately for $e_0$ and $e_1$.

Case 1, $e_0$: By definition

$$
\sigma_1(e_0) - \sigma_0(e_0) = \sigma_1[u]Q_1(e_0) - \sigma_0[u]Q_0(e_0) .
$$

Since $M_1 = M_0 + F_\epsilon(e_0,e_1)$

$$
\sigma_1(e_0) - \sigma_0(e_0) = \sigma_1[u](Q_0(e_0) + \epsilon) - \sigma_0[u]Q_0(e_0) .
$$

From (4.6) and $||F_\epsilon(e_0,e_1)|| = 2\epsilon$ it follows that

$$
\sigma_0[u] - 2\epsilon\kappa(Q_0) \leq \sigma_1[u] \leq \sigma_0[u] + 2\epsilon\kappa(Q_0) .
$$
We use this expression to eliminate $\sigma_1[u]$ from (5.20)

$$
\sigma_1(e_0) - \sigma_0(e_0) \geq (\sigma_0[u] - 2\epsilon\kappa(Q_0))(Q_0(e_0) + \epsilon) - \sigma_0[u]Q_0(e_0)
$$

$$
= -2\epsilon\kappa(Q_0)Q_0(e_0) + \epsilon\sigma_0[u] - 2\epsilon^2\kappa(Q_0)
$$

$$
= \epsilon\sigma_0[u] - 2\epsilon\kappa(Q_0)(Q_0(e_0) + \epsilon)
$$

$$
\geq \epsilon_\sigma .
$$

This completes the proof for $e_0$.

Case 2, $e_1$: By definition

$$
\sigma_1(e_1) - \sigma_0(e_1) = \sigma_1[u]Q_1(e_1) - \sigma_0[u]Q_0(e_1) .
$$

Since $M_1 = M_0 + F_\epsilon(e_0,e_1)$

$$
\sigma_1(e_1) - \sigma_0(e_1) = \sigma_1[u](Q_0(e_1) - \epsilon) - \sigma_0[u]Q_0(e_1) .
$$

Using the right hand side of inequality (5.21),

$$
\sigma_1[e_1] - \sigma_0[e_1] = \sigma_1[e_1]Q_1(e_1) - \sigma_0[e_1]Q_0(e_1)
$$

$$
\leq (\sigma_0[u] + 2\epsilon\kappa(Q_0))(Q_0(e_1) - \epsilon) - \sigma_0[u]Q_0(e_1)
$$

$$
= 2\epsilon\kappa(Q_0)Q_0(e_1) - \epsilon\sigma_0[u] - 2\epsilon^2\kappa(Q_0)
$$

$$
= \epsilon [-\sigma_0[u] + 2\kappa(Q_0)Q_0(e_1) - 2\epsilon\kappa(Q_0)]
$$

$$
= \epsilon [-\sigma_0[u] + 2\kappa(Q_0)(Q_0(e_1) - \epsilon)]
$$

Multiplying by $-1$:

$$
\sigma_0[e_1] - \sigma_1[e_1] \geq \epsilon [\sigma_0[u] - 2\kappa(Q_0)(Q_0(e_1) - \epsilon)] \geq \epsilon_\sigma .
$$

By (5.19) $\epsilon_\sigma > 0$. Hence,

$$
|\sigma_1[e_1] - \sigma_0[e_1]| \geq \epsilon_\sigma .
$$

Next we develop one of our main theoretic results: a learnability bound for MCs. We will assume that two MCs $M_0$ and $M_1$ are not $(l, \delta, \epsilon)_\text{dist}_Q$-distinguishable, i.e. when distance $\text{dist}_Q$ is over transition edge probabilities. Then from $M_0$ we will derive a similar
non-distinguishability result for distance $\text{dist}_\sigma$ over stationary edge probabilities. Finally we will use the last result to derive non-learnability result for distance $\text{dist}_\sigma$ over stationary edge probabilities.

**Theorem 5.5.2** Let $l \in \mathbb{N}^+, \epsilon, \delta_B \in (0, 0.5)$. If there exists $M_0 \in \mathcal{M}^{[n]}$ and $e_0, e_1 \in E$ such that

$$l \leq SC_M(M_0, \epsilon, \delta_B, e_0, e_1)$$

and

$$\sigma_0[u] > \max_{i=0,1} 2\kappa(Q_0)(Q_0(e_i) + (-1)^i\epsilon),$$

then for $\delta$, given by (5.14), and $\epsilon_\sigma$, given by (5.17), the class $\mathcal{M}^{[n]}$ is not $(l, n, \epsilon_\sigma/2, \delta)_{\text{dist}_\sigma}$-learnable.

**Proof** Let us recall that expression (5.17) is used to calculate $\epsilon_\sigma$ from $\epsilon$ and $M_0$. From Lemma 5.4.8, $M_0$ and $M_1$ are not $(l, \delta, \epsilon)_{\text{dist}_Q}$-distinguishable, when the distance between $M_0$ and $M_1$ is over transition edge probabilities. Now we show that this implies that they are not $(l, \delta, \epsilon_\sigma)_{\text{dist}_\sigma}$-distinguishable. To show this we assume by contradiction that $M_0$ and $M_1$ are $(l, \delta, \epsilon_\sigma)_{\text{dist}_\sigma}$-distinguishable. By Definition 5.1.2 it means that there exists a decision function $DF$ that, given $M_0, M_1, \epsilon_\sigma$ and $\delta$, distinguishes between $M_0$ and $M_1$, when the distance between them is over stationary edge probabilities. Since $\epsilon_\sigma$ is a function of $\epsilon$ and $M_0$, $DF$ may be used to distinguish between these two MCs, when the distance between them is over transition edge probabilities: given $M_0, M_1, \epsilon$ and $\delta$, calculate $\epsilon_\sigma$ and apply $DF$ to $M_0, M_1, \epsilon_\sigma$, and $\delta$. This means that $M_0$ and $M_1$ are $(l, \delta, \epsilon_\sigma)_{\text{dist}_Q}$-distinguishable, contradicting Lemma 5.4.8.

Now, after we showed that $M_0$ and $M_1$ are not $(l, \delta, \epsilon_\sigma)_{\text{dist}_\sigma}$-distinguishable, when the distance between $M_0$ and $M_1$ is over stationary edge probabilities, we can assign $(l, \delta, \epsilon_\sigma)_{\text{dist}_\sigma}$-non-distinguishability, $\epsilon_\sigma$ and $\text{dist}_\sigma$ into Theorem 5.2.1 to complete the proof.

Theorem 5.5.2 is proved for the edge pair $e_0$ and $e_1$. Note that going over all edge pairs, we can maximize either for $l$ or for $\epsilon_\sigma$. Theorem 5.5.3 maximizes for $l$ over all edge pairs.

50
Theorem 5.5.3 Let $\epsilon, \delta_B \in (0, 0.5), M \in M^{[n]}$ and

$$l = \max \{ SCM(M, \epsilon, \delta, (u, v_0), (u, v_1)) :$$

$$u, v_0, v_1 \in V, Q[u, v_0] \leq 1 - \epsilon, Q[u, v_1] \geq \epsilon,$$

$$\sigma[u] > 2\kappa(Q) \max \{Q[u, v_0] + \epsilon, Q[u, v_1] - \epsilon\} \}.$$  \hspace{1cm} (5.22)

Let $u, v_0, v_1$ be the states on which the maximum of (5.22) is obtained, $e_i = (u, v_i)$ and $\epsilon_\sigma = \epsilon_\sigma(M_0, e_0, e_1, \epsilon)$. Then the class $M^{[n]}$ is not $(l, n, \epsilon_\sigma/2, \delta)_{\text{dist}_\sigma}$-learnable.

Theorem 5.5.3 maximizes the results of Theorem 5.5.2 for $\epsilon_\sigma$ over all edge pairs.

Theorem 5.5.4 Let $\epsilon, \delta_B \in (0, 0.5), M \in M^{[n]}$ and

$$\epsilon_{\sigma, \max} = \max \{ \epsilon_\sigma(M_0, (u, v_0), (u, v_1), \epsilon) :$$

$$u, v_0, v_1 \in V, Q[u, v_0] \leq 1 - \epsilon, Q[u, v_1] \geq \epsilon,$$

$$\sigma[u] > 2\kappa(Q) \max \{Q[u, v_0] + \epsilon, Q[u, v_1] - \epsilon\} \}.$$ \hspace{1cm} (5.23)

Let $u, v_0, v_1$ be the states on which the maximum of (5.23) is obtained, $e_i = (u, v_i)$ and $l = SCM(M, \epsilon, \delta, e_0, e_1)$. Then the class $M^{[n]}$ is not $(l, n, \epsilon_{\sigma, \max}/2, \delta)_{\text{dist}_\sigma}$-learnable.

Figure 5.8 shows the application of Theorem 5.5.3 (circle-marked lines) and Theorem 5.5.3 (star-marked lines) to the MC, whose transition matrix is given in Table 7.2. The upper plot shows the upper bound on the sample complexity to learn the MC as a function of the perturbation of the transition edge probabilities $\epsilon$. The lower plot shows the lower bound on the perturbation of the stationary edge probabilities. The circle-marked line shows the bound when the sample complexity $l$ is maximized. The star-marked line shows the bound when the stationary edge probability is maximized.

One can see that while the difference between the sample complexity bounds is not very significant, the difference between stationary edge probabilities is significant. Thus it is preferable to use the bound when the stationary edge probability is maximized.

Bounds (5.22) and (5.23) are maximized over MC edges. For this reason for varying values of $\epsilon$ the maximum can be achieved at different edges. These bounds find the maximum over a set of edges, however, when $\epsilon$ increases, the set of relevant edges decreases. Thus when an edge leaves the set of relevant edges, a discontinuity may occur. As Figure 5.8 reveals, there are two discontinuity points: one around x coordinate 2.7 and the other around 2.9. Examining the edges on which the maximum occurs, we find that at
the first point the edges change from \( e_0 = (4, 10), e_1 = (4, 9) \) to \( e_0 = (9, 2), e_1 = (9, 10) \). Between these points there is no change in the maximization edges. Finally at the second point there is a change from edges \( e_0 = (9, 2), e_1 = (9, 10) \) to \( e_0 = (9, 2), e_1 = (9, 3) \). Thus the maximization in (5.22) over MC edges explains discontinuities in the sample complexity bounds.
Figure 5.8: The sample complexity bound on learning stationary edge probabilities of MC, whose transition matrix is given in Table 7.2
Chapter 6

Concept MC vs. learned MC

The bounds, developed in Theorem 5.2.1 and Corollary 5.2.3, were in terms of a concept MC and a sequence, produced by it. However, when learning, this MC is not known, hence the learner cannot know whether he has received a sufficient number of examples. Our aim is to find bounds in terms of the learned MC. Thus, whenever the learner postulates a MC, he can assess whether he has received a sufficient number of examples.

To differentiate between the concept and learned target MCs we use the superscript \((c)\) for a concept MC and \((\ell)\) for a learned MC.

Let \(\alpha : E \rightarrow [0, 1]\), \(M\) be a MC and let \(S(\alpha, M)\) denote the set of all the MCs over the same set of states:

\[
S(\alpha, M) = \{ M' : |\sigma_M(e) - \sigma_{M'}(e)| \leq \alpha(e), \forall e \in E \}.
\]

Note that \(S(\alpha, M)\) is an ellipsoid\(^1\) in \(\mathbb{R}^{|E|}\), where the size of the \(e\)-axis is \(\sigma(e)\). Also note that the ellipsoid is defined using the stationary edge probabilities. We prove our results only for the \(\text{dist}_\sigma\) distance. A similar result may be proved for transition edge probabilities. However for the sake of brevity we omit this proof.

While developing the practical bounds we assume that there exists a learning function with guarantees on the learning error. Definition 6.0.5 formulates our assumption. It further develops the notion of a learning function (Definition 5.1.3).

\textbf{Definition 6.0.5} Let \(f : E \times \mathbb{N}^+ \times (0, 0.5) \rightarrow [0, 1]\), \(M^{(c)}\) be a concept MC with \(n\) states, \(l_{LA} \in \mathbb{N}^+\), \(\delta_{LA} \in (0, 0.5)\). An \(f\)-guarantee learning function \(LA_f\) is a learning function

\(^1\)In \(L_1\).
such that, when given the number of states $n$ in $M^{(c)}$ and a sequence $s$ of length $l_{LA}$, distributed according to $M^{(c)}$, $LAF$ returns a MC $M^{(ℓ)}$ with $n$ states such that with probability at least $1 - δ_{LA}$, $M^{(ℓ)} \in S(f(\cdot, l_{LA}, δ_{LA}), M^{(c)})$.

We now develop a learnability bound for the concept MC by terms of the learned MC. Within this bound an $f$-guarantee learning function is applied to a training sequence $s$, resulting in a learned MC that is $f$-close to the concept MC with confidence $(1 - δ_{LA})$, i.e. with this confidence, the concept MC is located in a sphere around the learned MC. Minimizing the lower bound on the sample complexity of learning MCs over all MCs in this sphere, will provide us with the lower bound for the concept MC.

**Theorem 6.0.6** Let $M^{(c)} \in \mathcal{M}^{[n]}$ be a concept MC, $s$ be a sequence of length $l_{LA}$, distributed according to $M^{(c)}$, $LAf$ be an $f$-guarantee learning function, $δ_{LA} \in (0, 0.5)$ and $M^{(ℓ)} = LAf(s, n, δ_{LA})$. Let $e_0, e_1 \in E$ be two edges that emanate from the same node, $ε_{rv}, δ_{rv} \in (0, 0.5)$ and $SC_M$ be as defined in (5.12). Let

$$S = S(f(\cdot, l_{LA}, δ_{LA}), M^{(ℓ)}) ,$$

$$l = \min\{SC_M(M, ε_{rv}, δ_B, e_0, e_1) :$$

$$M \in S, σ[u] > 2κ(Q) \max\{Q[e_0] + ε_{rv}Q[e_1] - ε_{rv}\}\} ,$$

$$δ = (1 - δ_{LA}) · δ_B · P(τ(θ, l|M_0) \cup τ(θ, l|M_1)|M^{(c)}) .$$

Let $M$ be a MC on which minimum of (6.1) is obtained and $ε_σ = ε_σ(M, e_0, e_1, ε_{rv})$. Then the class $\mathcal{M}^{[n]}$ is not $(l, n, ε_σ/2, δ)_{dist_{σ}}$-learnable.

**Proof** The definition of an $f$-guarantee learning function means that with confidence $(1 - δ_{LA})$, $M^{(c)}_0 \in S(f(\cdot, δ_{LA}, l_{LA}), M^{(ℓ)})$. Since in (6.1) we minimize $ε_σ$ over this sphere, with confidence $(1 - δ_{LA})$

$$ε \leq ε_σ(M^{(c)}, e_0, e_1, ε_{rv}) .$$

Since in (6.1) we minimize $SC_M$ over the same sphere, with confidence $(1 - δ_{LA})$

$$\min_{M \in S} SC_M(M, ε_{rv}, δ_B, e_0, e_1) \leq SC_M(M^{(c)}, ε_{rv}, δ_B, e_0, e_1) .$$

Now the theorem statement follows from Theorem 5.5.2.

Theorem 6.0.7 shows a sample complexity bound if only the edges $e_0, e_1 \in E$ were perturbed. We can get a tighter bound if we apply this bound to all pairs of edges.
in \( E \) that emanate from the same node and choose a pair that maximizes the bound. Obviously improving the tightness is at the expense of increasing the computations. The next theorem summarizes this bound.

**Theorem 6.0.7** Let \( M^{(c)} \in \mathcal{M}^{[n]} \) be a concept MC; \( s \) be a sequence of length \( l_{LA} \), distributed according to \( M^{(c)} \), LA\( f \) be an \( f \)-guarantee learning function, \( \delta_{LA} \in (0, 0.5) \) and \( M^{(\ell)} = LA_f(s, n, \delta_{LA}) \). Let \( \epsilon_{rv}, \delta_{rv} \in (0, 0.5) \) and \( SC_M \) be as defined in (5.12). Let

\[
S = S(f(\cdot, l_{LA}, \delta_{LA}), M^{(\ell)}) ,
\]

\[
l = \max \{ \min \{ SC_M(M, \epsilon_{rv}, \delta_B, e_0, e_1) : \}
\]

\[
M \in S, \sigma[u] > 2\kappa(Q) \max \{ Q[e_0] + \epsilon, Q[e_1] - \epsilon \} : \]

\[
e_0 = (v, u_0), e_1 = (u, v_1) \}
\]

\[
\delta = (1 - \delta_{LA}) \cdot \delta_B \cdot P \left( \tau(\theta, l|M_0) \cup \tau(\theta, l|M_1) | M^{(c)} \right).
\]

Let \( e_0, e_1 \) be two edges, on which the maximum in (6.2) is obtained and \( M \) be a MC on which minimum of (6.2) is obtained and \( \epsilon_\sigma = \epsilon_\sigma(M, e_0, e_1, \epsilon_{rv}) \). Then the class \( \mathcal{M}^{[n]} \) is not \((l, n, \epsilon_\sigma/2, \delta)_{dist_\sigma}\)-learnable.
Chapter 7

Efficient Calculation

While Theorem 6.0.7 enables us to express a lower bound on the sample complexity of learning a concept MC by terms of the learned MC, it involves minimizing the lower bound (5.12) over an ellipse of MCs around the learned MC. Since the lower bound (5.12), as a function of $M_0$, is not guaranteed to be convex, this theorem does not provide a means to calculate a lower bound on the sample complexity of learning MCs. In this section we develop a weaker version of our bounds, that allows us to efficiently calculate lower bounds on the sample complexity of learning MCs.

7.1 Derivation

Let us recall the definition (5.10) of $K$

$$K(e_0, e_1) = \max_{i=0,1} \{ \sigma_i(e_0) + \sigma_i(e_1) \}.$$  \hspace{1cm} (7.1)

We first provide an outline of this section. Let $e_0 = (u,v_0), e_1 = (u,v_1)$ be two edges that
emanate from the same node. Our main building block is the calculation of the bound $SC_M(M(c), \epsilon, \delta_B, e_0, e_1)$ that, according to (5.12), is given by

$$SC_M(M(c), \epsilon, \delta_B, e_0, e_1) = \frac{LB_B(P_{B,0}, P_{B,1}, \delta_B)}{K^{(c)}(e_0, e_1) + 2\theta},$$

where $P_{B,0}, P_{B,1}$ and $K^{(c)}(e_0, e_1)$ are calculated for $M^{(c)}$. From the proof of Lemma 5.4.8, the nominator of (7.2) is a lower bound on the sample complexity needed to distinguish between two Bernoulli variables $P_{B,0}, P_{B,1}$ with confidence $(1 - \delta_B)$. We will replace this lower bound by another lower bound on the sample complexity to distinguish the same Bernoulli variables, but the new lower bound will depend only on $M^{(\ell)}$. Since we use an $(f, \text{dist})$-guarantee learning function $LA_f$, we know that $M^{(c)} \in S(f(\cdot, \delta_{LA}, l_{LA}), M^{(\ell)})$. For this reason we will follow the logic of Theorem 6.0.7 and provide an upper bound on the denominator of (7.2), over all MCs in $S(f(\delta_{LA}, l_{LA}), M^{(\ell)})$, which includes $M^{(c)}$ with probability $(1 - \delta_{LA})$. This bound will depend only on the learned MC.

We extend the notation $M^{(h)}, h = c, \ell$ that we used in Section 6 and in (7.2) to $M^{(h)}_i, i = 0, 1, h = c, \ell$: to differentiate between the concept (learned) MC and its $F_\epsilon(e_0, e_1)$-perturbation, we will use the subscripts 0 and 1. For example $M^{(c)}_0$ and $M^{(c)}_1 = M^{(c)}_0 + F_\epsilon(e_0, e_1)$ will mean the concept MC and an $F_\epsilon(e_0, e_1)$-perturbation of the concept MC, while $M^{(\ell)}_0$ and $M^{(\ell)}_1 = M^{(\ell)}_0 + F_\epsilon(e_0, e_1)$ will mean the target learned MC and an $F_\epsilon(e_0, e_1)$-perturbation of $M^{(\ell)}_0$.

We will bound the change of $K$, when the concept MC is replaced by the learned MC. Before developing this bounding, we prove the following auxiliary lemma.

**Lemma 7.1.1**

$$\left| K^{(h)}_0 - K^{(h)}_1 \right| \leq \frac{2\epsilon K Q^{(h)}_0}{\sigma^{(h)}_0[u]} \sum_{i \in \{0,1\}} \sigma^{(h)}_0(e_i).$$
Proof

\[ |K_0^{(h)} - K_1^{(h)}| = \left| \sum_{i \in \{0,1\}} \sigma_0^{(h)}(e_i) - \sum_{i \in \{0,1\}} \sigma_1^{(h)}(e_i) \right| \]

(7.3)

\[ = \left| \sum_{i \in \{0,1\}} \sigma_0^{(h)}(u)Q_0^{(h)}(e_i) - \sum_{i \in \{0,1\}} \sigma_1^{(h)}(u)Q_1^{(h)}(e_i) \right| \]

\[ = \left| \sum_{i \in \{0,1\}} \sigma_0^{(h)}(u)Q_0^{(h)}(e_i) - \sigma_1^{(h)}(u) \left( Q_0^{(h)}(e_0) + \epsilon + Q_1^{(h)}(e_0) - \epsilon \right) \right| \]

\[ = \left| \sigma_0^{(h)}(u) - \sigma_1^{(h)}(u) \right| \sum_{i \in \{0,1\}} Q_i^{(h)}(e_i) . \]

By definition \( Q_1^{(h)} = Q_0^{(h)} + F(c(e_0, e_1), h = c, \ell) \). From (4.6) and \( ||F(e_0, e_1)||_{\infty} = 2\epsilon \) it follows that

\[ \left| \sigma_0^{(h)}(u) - \sigma_1^{(h)}(u) \right| \leq \kappa \left( Q_0^{(h)} \right) 2\epsilon . \]

Substituting this expression into (7.3)

\[ |K_0^{(h)} - K_1^{(h)}| \leq \kappa \left( Q_0^{(h)} \right) 2\epsilon \sum_{i \in \{0,1\}} Q_i^{(h)}(e_i) . \]

(7.4)

Let us recall that \( \sigma_0^{(h)}(e_i) = \sigma_0^{(h)}(u)Q_0^{(h)}(e_i) \), \( i = 0, 1 \). Using this expression and (7.4), we get

\[ |K_0^{(h)} - K_1^{(h)}| \leq \frac{2\epsilon \kappa \left( Q_0^{(h)} \right)}{\sigma_0^{(h)}(u)} \sum_{i \in \{0,1\}} \sigma_0^{(h)}(e_i) . \]

(7.5)

The notion of an \( f \)-guarantee learning function allows us to bound the difference between stationary edge probabilities of the concept and learned MCs. Next we use this notion to bound the difference between stationary state probabilities of the concept and learned MCs.
Lemma 7.1.2 Let $M_0^{(c)}$ be a concept MC with $n$ states, $s$ be a path of length $l_{LA}$, distributed according to $M_0^{(c)}$, $\epsilon_{LA}, \delta_{LA} \in (0, 0.5)$, $LA_f$ be an $f$-guarantee learning function, 

$$M_0^{(\ell)} = LA_f(s, n, \epsilon_{LA}, \delta_{LA})$$

and $u \in V$. Then with confidence at least $(1 - \delta_{LA})$

$$\left| \sigma_0^{(c)}[u] - \sigma_0^{(\ell)}[u] \right| \leq \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) .$$

Proof Let $M$ be a MC and $z \in V$. From ([2, page 88]) the stationary node probability of a node $z$ is equal to the sum of the stationary edge probabilities of the incoming edges to $z$

$$\sigma[z] = \sum_{w \in V} \sigma(w, z) .$$

Using this expression

$$\left| \sigma_0^{(c)}[u] - \sigma_0^{(\ell)}[u] \right| = \sum_{v \in V} \left| \sigma_0^{(c)}(v, u) - \sum_{v \in V} \sigma_0^{(\ell)}(v, u) \right| \leq \sum_{v \in V} \left| \sigma_0^{(c)}(v, u) - \sigma_0^{(\ell)}(v, u) \right| .$$

Since $M_0^{(\ell)} = LA_f(s, n, \delta_{LA})$, with confidence at least $(1 - \delta_{LA})$, $M_0^{(c)} \in S(f(\cdot, \delta_{LA}, l_{LA}), M_0^{(\ell)})$. Using this information, we get that with confidence at least $(1 - \delta_{LA})$

$$\left| \sigma_0^{(c)}[u] - \sigma_0^{(\ell)}[u] \right| \leq \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) .$$

We now bound the change of $K$, when the concept MC is replaced by the learned MC. Let us denote this change

$$\Delta K = \left| K^{(\ell)} - K^{(c)} \right| .$$

(7.6)

Lemma 7.1.3 provides an upper bound on $\Delta K$. 60
Lemma 7.1.3 Assume the settings of Theorem 6.0.7. Also let

\[ \Psi_{0,0} = \sum_{i \in \{0, 1\}} f(e_i, \delta_{LA}, l_{LA}) , \]

\[ \Psi_{0,1} = \Psi_{0,0} + \frac{2 \epsilon \kappa \left( Q_0^{(c)} \right)}{\sigma_0^{(t)}[u]} - \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) \sum_{i \in \{0, 1\}} \left( \sigma_0^{(t)}(e_i) + f(e_i, \delta_{LA}, l_{LA}) \right) , \]

\[ \Psi_{1,0} = \Psi_{0,0} + \frac{2 \epsilon \kappa \left( Q_0^{(t)} \right)}{\sigma_0^{(t)}[u]} \sum_{i \in \{0, 1\}} \sigma_0^{(t)}(e_i) , \]

\[ \Psi_{1,1} = \Psi_{0,1} + \frac{2 \epsilon \kappa \left( Q_0^{(t)} \right)}{\sigma_0^{(t)}[u]} \sum_{i \in \{0, 1\}} \sigma_0^{(t)}(e_i) . \]

Then with confidence \((1 - \delta_{LA})\)

\[ \Delta K \leq \Psi = \max \{ \Psi_{0,0}, \Psi_{0,1}, \Psi_{1,0}, \Psi_{1,1} \} . \] (7.7)

**Proof** From (7.1) and (7.6)

\[ \Delta K \leq \max \left\{ \left| K_0^{(t)} - K_0^{(c)} \right|, \left| K_0^{(t)} - K_1^{(c)} \right|, \left| K_1^{(t)} - K_0^{(c)} \right|, \left| K_1^{(t)} - K_1^{(c)} \right| \right\} . \] (7.8)

Now we develop upper bounds for the four constituents on the right hand side of (7.8).

**Case 1, \( \left| K_0^{(t)} - K_0^{(c)} \right| \):**

\[ \left| K_0^{(t)} - K_0^{(c)} \right| = \left| \sum_{i \in \{0, 1\}} \sigma_0^{(t)}(e_i) - \sum_{i \in \{0, 1\}} \sigma_0^{(c)}(e_i) \right| . \]

Since \( M_0^{(t)} = LA_f(s, n, \delta_{LA}) \), with confidence at least \((1 - \delta_{LA})\), \( M_0^{(c)} \in S(f(\cdot, \delta_{LA}, l_{LA}), M_0^{(t)}) \) and thus with confidence at least \((1 - \delta_{LA})\),

\[ \left| K_0^{(t)} - K_0^{(c)} \right| \leq \sum_{i \in \{0, 1\}} \left| \sigma_0^{(t)}(e_i) - \sigma_0^{(c)}(e_i) \right| \leq \sum_{i \in \{0, 1\}} f(e_i, \delta_{LA}, l_{LA}) = \Psi_{0,0} . \] (7.9)
Case 2, $|K_0^{(f)} - K_1^{(c)}|$:

First note that

$$|K_0^{(f)} - K_1^{(c)}| = |K_0^{(f)} + K_0^{(c)} - K_0^{(c)} - K_1^{(c)}| \leq |K_0^{(f)} - K_0^{(c)}| + |K_0^{(c)} - K_1^{(c)}|.$$  \hspace{1cm} (7.10)

We will separately bound the two factors on the right-hand side of the above expression. From Case 1 with confidence at least $(1 - \delta_{LA})$,

$$|K_0^{(f)} - K_0^{(c)}| \leq \Psi_{0,0}.$$  \hspace{1cm} (7.11)

From Lemma 7.1.1

$$|K_0^{(c)} - K_1^{(c)}| \leq \frac{2\epsilon \kappa \left(Q_0^{(c)}\right)}{\sigma_0^{(c)}[u]} \sum_{i \in \{0,1\}} \sigma_0^{(c)}(e_i).$$  \hspace{1cm} (7.12)

Since $M_0^{(f)} = LA_f(s, n, \delta_{LA})$, with confidence at least $(1 - \delta_{LA})$, $M_0^{(c)} \in S(f(\cdot, \delta_{LA}, l_{LA}), M_0^{(f)})$. Using this information in (7.12), we get that with confidence $(1 - \delta_{LA})$

$$|K_0^{(c)} - K_1^{(c)}| \leq \frac{2\epsilon \kappa \left(Q_0^{(c)}\right)}{\sigma_0^{(c)}[u]} \sum_{i \in \{0,1\}} \left(\sigma_0^{(f)}(e_i) + f(e_i, \delta_{LA}, l_{LA})\right).$$  \hspace{1cm} (7.13)

Finally substituting the above expression and (7.11) into (7.10) and using Lemma 7.1.2 we get that with confidence at least $\delta_{LA}$

$$|K_0^{(f)} - K_1^{(c)}| \leq \Psi_{0,0} + \frac{2\epsilon \kappa \left(Q_0^{(c)}\right)}{\sigma_0^{(c)}[u]} \sum_{(v,u) \in V} f((v,u), \delta_{LA}, l_{LA}) \sum_{i \in \{0,1\}} \left(\sigma_0^{(f)}(e_i) + f(e_i, \delta_{LA}, l_{LA})\right) = \Psi_{0,1}.$$  \hspace{1cm} (7.14)

Case 3, $|K_1^{(f)} - K_0^{(c)}|$:

Note that

$$|K_1^{(f)} - K_0^{(c)}| = |K_1^{(f)} + K_0^{(f)} - K_0^{(f)} - K_0^{(c)}| \leq |K_0^{(f)} - K_0^{(c)}| + |K_0^{(f)} - K_1^{(c)}|.$$  \hspace{1cm} (7.14)

We will separately bound each of the two factors on the right-hand side of the above
expression. From Lemma 7.1.1
\[ |K_1^{(l)} - K_0^{(l)}| \leq \frac{2\epsilon \kappa \left( Q_0^{(l)} \right)}{\sigma_0^{(l)}[u]} \sum_{i \in \{0,1\}} \sigma_0^{(l)}(e_i). \] (7.15)

Substituting the above expression and (7.9) into (7.14), we get that with confidence \((1 - \delta_{LA})\)
\[ |K_1^{(l)} - K_1^{(c)}| \leq \Psi_{0,0} + \frac{2\epsilon \kappa \left( Q_0^{(l)} \right)}{\sigma_0^{(l)}[u]} \sum_{i \in \{0,1\}} \sigma_0^{(l)}(e_i) = \Psi_{1,0}. \]

Case 4, \( |K_1^{(l)} - K_1^{(c)}| \).

Note that
\[ |K_1^{(l)} - K_1^{(c)}| = |K_1^{(l)} + K_0^{(l)} - K_0^{(c)} + K_0^{(c)} - K_1^{(c)}| \]
\[ \leq |K_0^{(l)} - K_0^{(c)}| + |K_1^{(l)} - K_0^{(l)}| + |K_0^{(c)} - K_1^{(c)}|. \]

Substituting (7.9), (7.13) and (7.15) into the above expression, with confidence \((1 - \delta_{LA})\) we get
\[ |K_1^{(l)} - K_1^{(c)}| \leq \Psi_{0,1} + \frac{2\epsilon \kappa \left( Q_0^{(l)} \right)}{\sigma_0^{(l)}[u]} \sum_{i \in \{0,1\}} \sigma_0^{(l)}(e_i). \]

Now note that we bounded each one of the factors on the right hand side of (7.8) under the condition that \( M_0^{(c)} \in S(f(\cdot, \delta_{LA}, l_{LA}), M_0^{(l)}) \), which happens with probability at least \((1 - \delta_{LA})\), since \(M_0^{(l)} = LA_f(s_n, \delta_{LA})\). This implies that each one of the bounds on the factors of the right hand side of (7.8) is conditioned by the same single event \( M_0^{(c)} \in S(f(\cdot, \delta_{LA}, l_{LA}), M_0^{(l)}) \). This implies that (7.7) holds with probability of this event, i.e. probability at least \((1 - \delta_{LA})\).

Note that Lemma 5.5.1 provides us with means to estimate the difference between stationary edge probabilities of the concept MC and its \( F_\epsilon(e_0, e_1) \)-perturbation. Since in practice the concept MC is not known, we need means to estimate this difference based on the learned MC. The next lemma provides us with such means.
Lemma 7.1.4 Assume the settings of Theorem 6.0.7 with \( M(c) = M_0(c) \), \( M(\ell) = M_0(\ell) \), 
\( e_0 = (u, v_0) \), \( e_1 = (u, v_1) \), except for the definition of \( \epsilon \). Let

\[
\epsilon = \epsilon_{rv} \min_{i=0,1} \left\{ \sigma_0^{(\ell)}[u] - \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) \right. \]

\[
- 2\kappa \left( Q_0^{(c)} \right) \left( \frac{\sigma_0^{(\ell)}(e_i) + f(e_i, \delta_{LA}, l_{LA})}{\sigma_0^{(\ell)}[u] - \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA})} + (-1)^i \epsilon_{rv} \right) \right\} .
\]

If

\[
\sigma_0^{(\ell)}[u] > \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) ,
\]

then with confidence at least \( 1 - \delta_{LA} \)

\[
|\sigma_0^{(c)}(e_i) - \sigma_0^{(c)}(e_i)| \geq \epsilon, \ i = 0, 1 .
\]

If

\[
\sigma_0^{(\ell)}[u] > \max_{i=0,1} \left\{ \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) + 2\kappa \left( Q_0^{(c)} \right) \left( \frac{\sigma_0^{(\ell)}(e_i) + f(e_i, \delta_{LA}, l_{LA})}{\sigma_0^{(\ell)}[u] - \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA})} + (-1)^i \epsilon_{rv} \right) \right\} ,
\]

then

\[
\epsilon > 0 .
\]

**Proof** Note that (7.19) follows from (7.16) and (7.18).

From the proof of Lemma 5.5.1

\[
|\sigma_1^{(c)}(e_i) - \sigma_0^{(c)}(e_i)| > \epsilon_{rv} \min_{i=0,1} \left\{ \sigma_0^{(c)}[u] - 2\kappa \left( Q_0^{(c)} \right) \left( Q_0^{(c)}(e_i) + (-1)^i \epsilon_{rv} \right) \right\} .
\]

We will prove the lemma separately for \( i = 0 \) and \( i = 1 \). We start with \( i = 0 \).

Case 1 \((i = 0)\): Let us examine the argument of min of (7.20). Since \( \sigma_0^{(c)}(e_0) = \sigma_0^{(c)}[u]Q_0^{(c)}(e_0) \), and from (7.17),

\[
\sigma_0^{(c)}[u] - 2\kappa \left( Q_0^{(c)} \right) \left( Q_0^{(c)}(e_0) + \epsilon_{rv} \right) = \sigma_0^{(c)}[u] - 2\kappa \left( Q_0^{(c)} \right) \left( \frac{\sigma_0^{(c)}(e_0)}{\sigma_0^{(c)}[u]} + \epsilon_{rv} \right) .
\]
Since we assume the settings of Theorem 6.0.7, where \( M_0^{(\ell)} = LA_f(s,n,\delta_{LA}) \), with confidence at least \((1 - \delta_{LA})\), \( M_0^{(c)} \in S(f(\cdot, \delta_{LA}, l_{LA}), M_0^{(\ell)}) \). This implies that with confidence at least \((1 - \delta_{LA})\)

\[
\sigma_0^{(c)}(e_0) - f(e_0, \delta_{LA}, l_{LA}) \leq \sigma_0^{(c)}(e_0) \leq \sigma_0^{(\ell)}(e_0) + f(e_0, \delta_{LA}, l_{LA}).
\]

These inequalities allow us to eliminate \( \sigma_0^{(c)}(e_0) \) from (7.21)

\[
\sigma_0^{(c)}[u] - 2\kappa \left( Q_0^{(c)}(e_0) + \epsilon_{rv} \right) \geq \frac{\sigma_0^{(\ell)}(e_0) + f(e_0, \delta_{LA}, l_{LA})}{\sigma_0^{(c)}[u]} + \epsilon_{rv}.
\]

From Lemma 7.1.2, with confidence at least \(1 - \delta_{LA}\)

\[
\sigma_0^{(\ell)}[u] - \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) \leq \sigma_0^{(c)}[u] \leq \sigma_0^{(\ell)}[u] + \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}).
\]

These inequalities allow us to eliminate \( \sigma_0^{(c)}[u] \) from (7.22)

\[
\sigma_0^{(c)}[u] - 2\kappa \left( Q_0^{(c)}(e_0) + \epsilon_{rv} \right) \geq \frac{\sigma_0^{(\ell)}(e_0) + f(e_0, \delta_{LA}, l_{LA})}{\sum_{v \in V} f((v, u), \delta_{LA}, l_{LA})} + \epsilon_{rv}.
\]

Note that because in (7.23) we divide by \( \sigma_0^{(\ell)}[u] - \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) \), we should make sure that this quantity is not zero. Thus we require

\[
\sigma_0^{(\ell)}[u] > \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}).
\]

Note that this requirement is a part of our lemma statement.

Case 2 \( (i = 1) \): From (7.20), we get the same lower bound on \(|\sigma_1(e_1) - \sigma_0(e_1)|\) as (7.23), where the sign of \(\epsilon_{rv}\) is “−1”.

Recall definition (5.12) of \( SCM \). Now we prove a lower bound on the sample com-
plexity by terms of the learned MC.

**Theorem 7.1.5** Assume the settings of Theorem 6.0.7 with \( M^c = M_0^c \), \( M^l = M_0^l \), except for the definition of \( l \). Let \( u = \text{start}(e_0) \). Also let

\[
\begin{align*}
P_{\text{min}} &= \frac{\sigma_0^l(e_0) - f(e_0, \delta_{LA}, l_{LA})}{\sum_{i=0,1}(\sigma_0^l(e_i) + f(e_i, \delta_{LA}, l_{LA}))}, \\
P_{\text{max}} &= \frac{\sigma_0^l(e_0) + f(e_0, \delta_{LA}, l_{LA}) + \epsilon \left( \sigma_0^l[u] + \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) \right)}{\sum_{i=0,1}(\sigma_0^l(e_i) - f(e_i, \delta_{LA}, l_{LA}))}.
\end{align*}
\]

Let

\[
l \leq SCP_{\text{mc}}(M_0^l, \epsilon, \delta_B, e_0, e_1) \equiv \frac{LB_B(P_{\text{min}}, P_{\text{max}}, \delta_B)}{K^l(e_0, e_1) + \Psi + 2\delta}, \tag{7.24}
\]

and \( \epsilon \) be as in Lemma 7.1.4. Then the class \( M^{[n]} \) is not \((l, n, \epsilon/2, \delta)_{\text{dist}, \epsilon}\)-learnable.

**Proof** Let \( P_{B,i} \) be as defined in (5.11) for \( M_0 = M_0^c \):

\[
P_{B,i} = \frac{Q_{i}^c(e_0)}{Q_{i}^c(e_0) + Q_{i}^c(e_1)},
\]

Let us recall that \( M_1^c \) results from \( M_0^c \) by adding \( \epsilon \) to \( Q_0^c(e_0) \) and substructing \( \epsilon \) from \( Q_0^c(e_1) \):

\[
\begin{align*}
P_{B,0} &= \frac{Q_0^c(e_0)}{Q_0^c(e_0) + Q_0^c(e_1)}, \tag{7.25} \\
P_{B,1} &= \frac{Q_0^c(e_0) + \epsilon}{Q_0^c(e_0) + \epsilon + Q_0^c(e_1)} - \epsilon = \frac{Q_0^c(e_0) + \epsilon}{Q_0^c(e_0) + Q_0^c(e_1)}.
\end{align*}
\]

Since \( \sigma_0^c(e_0) = \sigma_0^c(u)Q_0^c(e_0) \), multiplying the nominator and the denominator of (7.25) by \( \sigma_0^c(u) \) we get

\[
\begin{align*}
P_{B,0} &= \frac{\sigma_0^c(e_0)}{\sigma_0^c(e_0) + \sigma_0^c(e_1)}, \tag{7.26} \\
P_{B,1} &= \frac{\sigma_0^c(e_0) + \epsilon \sigma_0^c(u)}{\sigma_0^c(e_0) + \sigma_0^c(e_1)}.
\end{align*}
\]
By Lemma 7.1.2
\[ P_{B,1} \leq \frac{\sigma_0^c(e_0) + \epsilon \left( \sigma_0^c(u) + \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) \right)}{\sigma_0^c(e_0) + \sigma_0^c(e_1)} . \] (7.27)

First we show that in (7.2) we can replace the nominator \( LB(P_{B,0}, P_{B,1}, \delta_B) \) by \( LB(P_{min}, P_{max}, \delta_B) \). By adding \((-\sigma_0^c(e_0) + \sigma_0^c(e_0))\) to the nominator of (7.26) and adding \((-\sigma_0^c(e_0) + \sigma_0^c(e_0) - \sigma_0^c(e_1) + \sigma_0^c(e_1))\) to its denominator, we get
\[ P_{B,0} = \frac{\sigma_0^c(e_0) - \sigma_0^c(e_1) + \sigma_0^c(e_0)}{\sigma_0^c(e_0) - \sigma_0^c(e_0) + \sigma_0^c(e_0) - \sigma_0^c(e_1) + \sigma_0^c(e_1)} . \] (7.28)

Since we assume the settings of Theorem 6.0.7, where \( M^c = LAf(s, n, \delta_{LA}) \), it follows that with confidence \((1 - \delta_{LA})\),
\[ -f(e_i, \delta_{LA}, l_{LA}) \leq \sigma_0^c(e_i) - \sigma_0^c(e_i) \leq f(e_i, \delta_{LA}, l_{LA}), \quad i = 0, 1 . \] (7.29)

Substituting \(-f(e_i, \delta_{LA}, l_{LA})\) \(\leq \sigma_0^c(e_i) - \sigma_0^c(e_i)\) into the nominator of (7.28) and \(\sum_{i=0,1} \sigma_0^c(e_i) - \sigma_0^c(e_i) \leq \sum_{i=0,1} f(e_i, \delta_{LA}, l_{LA})\) into its denominator, we get
\[ P_{min} \leq P_{B,0} . \] (7.30)

Next we add \((-\sigma_0^c(e_0) + \sigma_0^c(e_0))\) to the nominator of (7.27) and \((-\sigma_0^c(e_0) + \sigma_0^c(e_0) - \sigma_0^c(e_1) + \sigma_0^c(e_1))\) to its denominator, we get
\[ P_{B,1} \leq \frac{\sigma_0^c(e_0) - \sigma_0^c(e_0) + \sigma_0^c(e_0) + \epsilon \left( \sigma_0^c(u) + \sum_{v \in V} f((v, u), \delta_{LA}, l_{LA}) \right)}{\sigma_0^c(e_0) - \sigma_0^c(e_0) + \sigma_0^c(e_0) - \sigma_0^c(e_1) + \sigma_0^c(e_1)} . \] (7.31)

By (7.29) with confidence \((1 - \delta_{LA})\) we can substitute \(\sigma_0^c(e_0) - \sigma_0^c(e_0) \leq f(e_i, \delta_{LA}, l_{LA})\) into the nominator of (7.31) and \(-\sum_{i=0,1} f(e_i, \delta_{LA}, l_{LA}) \leq \sum_{i=0,1} \sigma_0^c(e_i) - \sigma_0^c(e_i)\) into its denominator:
\[ P_{B,1} \leq P_{max} . \] (7.32)

Let us recall Definition 4.1.3 of the sample complexity to distinguish random variables and inequality (4.3). As the probability \(P_1\) gets further away from \(P_0\), it becomes easier to distinguish between those two variables and thus a shorter sequence \(s\) is sufficient to
distinguish between them. Thus $SC_B(P_0 - b, P_1 + a, \delta_B)$ monotonically decreases with $a, b > 0$ for $0 \leq P_0 - b$ and $P_1 + a \leq 1$. Using this conclusion and (4.3)

$$LB_B(P_0 - b, P_1 + a, \delta_B) \leq SC_B(P_0 - b, P_1 + a, \delta_B) \leq SC_B(P_0, P_1, \delta_B).$$

Assigning $P_0 \leftarrow P_{B,0}$, $b \leftarrow P_{B,0} - P_{\min}$, $P_1 \leftarrow P_{B,1}$ and $a \leftarrow P_{\max} P_{B,1}$ and using (7.30) and (7.32)

$$LB_B(P_{\min}, P_{\max}, \delta_B) \leq SC_B(P_{B,0}, P_{B,1}, \delta_B).$$

**Conclusion:** In expression (7.2) we can replace the nominator by $LB_B(P_{\min}, P_{\max}, \delta_B)$.

Let us recall that according to Lemma 7.1.3 we can replace the denominator of (7.2) by $K^{(f)}(e_0, e_1) + \Psi + 2\theta$. Replacing both the nominator and the denominator results in the right hand side of expression (7.24). Thus, when inequality (7.24) holds, according to Lemma 5.4.8, $M^{(c)}_0$ and $M^{(c)}_1$ are not $(l, \epsilon, \delta)$-distinguishable. Finally, according to Theorem 5.5.2, the class $M^{[n]}$ is not $(l, n, \epsilon/2, \delta)_{dist}^{\sigma}$-learnable.

Maximizing over all pairs of edges that emanate from the same node, we get the following result:

**Theorem 7.1.6** Assume the settings of Theorem 7.1.5, except for the definition of $l$, and let

$$l \leq \max_{e_0, e_1 \in E: \text{start}(e_0) = \text{start}(e_1), \epsilon \leq \min(1 - Q_0(e_0), Q_0(e_1))} \frac{LB_{mc}(M^{(f)}_0, \epsilon, \delta_B, e_0, e_1)}{\epsilon_{e_0, e_1}}$$

Then the class $M^{[n]}$ is not $(l, n, \epsilon/2, \delta)_{dist}^{\sigma}$-learnable.

### 7.2 Example

In this section we discuss application of Theorem 7.1.6 to the real problem that was studied in [21] and was briefly discussed in the Introduction. That work analyzes brand loyalty using Markov chains. The authors chose 11 sport shoes brands and conducted a survey on university students, in which students were asked the present brand of sport shoes they own and the brand preference they have for their next sport shoes purchase (among the 11 brand selected). 531 students participated in this survey. The results of the survey were represented as the transition matrix $T$ of a markov chain, shown in Table 7.2. Then the authors found the stationary node probabilities of matrix $T$, which show the strength
of each brand over a long period of time, i.e. the buying preferences of customers in the long run.

Table 7.1: Brand loyalty transition matrix

<table>
<thead>
<tr>
<th></th>
<th>Adidas</th>
<th>Nike</th>
<th>Reebok</th>
<th>Puma</th>
<th>Slazenger</th>
<th>Kappa</th>
<th>Diadora</th>
<th>Ellese</th>
<th>Le coq</th>
<th>Six</th>
<th>Converse</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adidas</td>
<td>0.042</td>
<td>0.563</td>
<td>0.204</td>
<td>0.024</td>
<td>0.060</td>
<td>0.006</td>
<td>0.012</td>
<td>0.006</td>
<td>0.012</td>
<td>0.006</td>
<td>0.018</td>
<td>0.047</td>
</tr>
<tr>
<td>Nike</td>
<td>0.472</td>
<td>0.016</td>
<td>0.276</td>
<td>0.057</td>
<td>0.041</td>
<td>0.008</td>
<td>0.033</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.073</td>
</tr>
<tr>
<td>Reebok</td>
<td>0.360</td>
<td>0.488</td>
<td>0.012</td>
<td>0.023</td>
<td>0.047</td>
<td>0.000</td>
<td>0.012</td>
<td>0.000</td>
<td>0.012</td>
<td>0.000</td>
<td>0.000</td>
<td>0.034</td>
</tr>
<tr>
<td>Puma</td>
<td>0.438</td>
<td>0.250</td>
<td>0.000</td>
<td>0.123</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.063</td>
<td>0.063</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Slazenger</td>
<td>0.395</td>
<td>0.289</td>
<td>0.184</td>
<td>0.000</td>
<td>0.053</td>
<td>0.000</td>
<td>0.026</td>
<td>0.053</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.080</td>
<td>0.500</td>
<td>0.025</td>
<td>0.000</td>
<td>0.250</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Diadora</td>
<td>0.111</td>
<td>0.333</td>
<td>0.000</td>
<td>0.111</td>
<td>0.111</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.111</td>
<td>0.112</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Ellese</td>
<td>0.000</td>
<td>0.600</td>
<td>0.000</td>
<td>0.000</td>
<td>0.200</td>
<td>0.200</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Le coq sportif</td>
<td>0.000</td>
<td>0.571</td>
<td>0.286</td>
<td>0.143</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Six</td>
<td>0.167</td>
<td>0.000</td>
<td>0.167</td>
<td>0.167</td>
<td>0.166</td>
<td>0.000</td>
<td>0.000</td>
<td>0.166</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Converse</td>
<td>0.400</td>
<td>0.200</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Other</td>
<td>0.309</td>
<td>0.400</td>
<td>0.055</td>
<td>0.000</td>
<td>0.036</td>
<td>0.000</td>
<td>0.036</td>
<td>0.180</td>
<td>0.000</td>
<td>0.000</td>
<td>0.018</td>
<td>0.128</td>
</tr>
</tbody>
</table>

We assume that there is a concept $M_0^{(c)}$ that produced the training data - student surveys. The matrix above is the transition matrix of the $M_0^{(f)}$ which is learned from the training data.

To apply the bounds of Theorem 7.1.5 and Theorem 7.1.6, we need to characterize the $(f, \text{dist}_{\sigma})$-learning function $LA_{f,\text{dist}_{\sigma}}$ that was used to construct the transition matrix, presented in Table 7.2. Let $M$ be a MC, $s$ be a path in $MC$ and $e \in E$. Let us recall that $n(e|s)$ denotes the number of times $s$ traverses $e$. We extend this notation to node traversals. For $v \in V$, let $n(v|s)$ denote the number of times $s$ traverses state $v$. Now we can characterize $LA_{f,\text{dist}_{\sigma}}$. The transition edge probabilities of the learned MC $M^{(f)}$ were estimated as follows

$$\forall e = (u,v) \in E, \; Q^{(f)}(e) \leftarrow n(e|s)/n(u|s) .$$

Edge probabilities of this MC were estimated as follows

$$\forall e = (u,v) \in E, \; P^{(f)}(e) \leftarrow n(e|s)/|s| . \quad (7.33)$$

From [2, page 88] edge probability distribution $P$ is stationary if and only if for each state $u \in V$ the sum of the probabilities of the incoming edges is equal to the sum of the probabilities of the outgoing edges:

$$P \text{ is stationary edge distribution iff } \forall u \in V, \; \sum_{v \in V} P(v,u) = \sum_{v \in V} P(u,v) . \quad (7.34)$$

69
We can use (7.34) to prove the following necessary and sufficient condition on edge probability (7.33)

**Lemma 7.2.1** Edge probability distribution, defined in (7.33), is stationary iff $s$ starts and ends at the same state.

**Proof** From (7.33), we can rephrase (7.34) for $P(\ell)$ as follows

$P(\ell)$ is stationary edge distribution iff

$$
\forall u \in V, \sum_{v \in V} n((v, u)|s) = \sum_{v \in V} n((u, v)|s) ,
$$

i.e. $s$ enters and leaves each state the same number of times.

Let us examine now training sequence $s$. If it starts and ends at different states, the number of times $s$ enters the first state of $s$ is less by 1 than the number of times $s$ leaves this state. From (7.2), $P(\ell)$ is not stationary edge probability distribution.

If $s$ starts and ends at the same state, $s$ enters and leaves each state the same number of times and, thus, by (7.2), (7.33) defines stationary edge probabilities. 

Now we characterize the $(f, dist_\sigma)$-learning function $LA_{f, dist_\sigma}$.

**Lemma 7.2.2** Let $M^{(c)}$ be a concept MC, $s$ a path, distributed according to $M^{(c)}$, and $M^{(\ell)}$ be the MC, learned from $s$ using (7.33). Let $\theta \in (0, 0.5)\), $\tau(\theta, |s|)$ be the set of $\theta$-strongly typical sequences of $M^{(c)}$ of length $|s|$, $e' \in E$, $\pi_{M^{(c)}}$ be as defined in (5.6) and let

$$
\delta_{LA} \leftarrow 1 - \max_{e'} \pi_{M^{(c)}}(\tau(\theta, |s|), e'),
$$

Then with confidence at least $1 - \delta_{LA}$,

$$
|\sigma^{(c)}(e) - \sigma^{(\ell)}(e)| < \theta + |P^{(\ell)}(e) - \sigma^{(\ell)}(e)|, \ \forall e \in E . \tag{7.35}
$$

**Proof** First note that

$$
|\sigma^{(c)}(e) - \sigma^{(\ell)}(e)| = |\sigma^{(c)}(e) - P^{(\ell)}(e) + P^{(\ell)}(e) - \sigma^{(\ell)}(e)|
$$

$$
\leq |\sigma^{(c)}(e) - P^{(\ell)}(e)| + |P^{(\ell)}(e) - \sigma^{(\ell)}(e)| .
$$
According to Lemma 5.4.4, with confidence at least $1 - \delta_{LA}$, $s \in \tau(\theta, |s|)$. Thus by Definition 5.4.3, with confidence at least $1 - \delta_{LA}$,

$$|n(e|s)/|s| - \sigma^{(c)}(e)| < \theta, \forall e \in E.$$ 

Using (7.33), we get from the last expression

$$|P^{(L)}(e) - \sigma^{(c)}(e)| < \theta, \forall e \in E$$

with confidence at least $1 - \delta_{LA}$. 

Note that since $\sigma^{(L)}(e)$ is efficiently computable, the bound 7.35 is also efficiently computable.

Let us recall that Theorem 7.1.5 provides us with the means to efficiently calculate the lower sample complexity bound on learning concept MCs by the terms of learned MCs. This theorem requires a characterization of an $(f, \text{dist}_\sigma)$-learning function. In Lemma 7.2.2 we developed the characterization of an $(f, \text{dist}_\sigma)$-learning function and this characterization depends on the estimation of edge probabilities $P^{(L)}(e)$ of the learned MC. Unfortunately the data of the brand loyalty example, described in the beginning of this section, lacks the details, required for such an estimation. For this reason we apply our bounds to a MC, whose entries are randomly generated.

In Table 7.2 we show the transition matrix of a MC, whose probabilities were randomly generated, according to the uniform distribution. We use this MC to examplify bounds, stated in Theorem 5.5.3 and Theorem 5.5.4. However when we used this MC to examplify the practical bounds and calculated the bound (7.16) of Lemma 7.1.4, we did not find an edge pair for which the right hand side of the bound is positive.

Since Table 7.2 is not suitable for demonstrating practical bounds, for this purpose we use MC, whose transition matrix is given in Table 7.3.

Figure 7.2 presents the result of applying Lemma 7.1.4 with $\delta_B = 0.1$ and $\theta = 10^{-4}$ to the transition matrix in Table 7.3.
Table 7.2: Transition matrix of the example MC.

<table>
<thead>
<tr>
<th></th>
<th>v1</th>
<th>v2</th>
<th>v3</th>
<th>v4</th>
<th>v5</th>
<th>v6</th>
<th>v7</th>
<th>v8</th>
<th>v9</th>
<th>v10</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>0.0974</td>
<td>0.0912</td>
<td>0.0677</td>
<td>0.1166</td>
<td>0.0582</td>
<td>0.1376</td>
<td>0.1323</td>
<td>0.0861</td>
<td>0.0894</td>
<td>0.0975</td>
</tr>
<tr>
<td>v2</td>
<td>0.1168</td>
<td>0.1399</td>
<td>0.0969</td>
<td>0.0203</td>
<td>0.0324</td>
<td>0.0853</td>
<td>0.0929</td>
<td>0.2051</td>
<td>0.1911</td>
<td>0.0194</td>
</tr>
<tr>
<td>v3</td>
<td>0.1585</td>
<td>0.0991</td>
<td>0.0666</td>
<td>0.0957</td>
<td>0.1068</td>
<td>0.1245</td>
<td>0.1617</td>
<td>0.0597</td>
<td>0.1071</td>
<td>0.0203</td>
</tr>
<tr>
<td>v4</td>
<td>0.0575</td>
<td>0.1568</td>
<td>0.1660</td>
<td>0.1153</td>
<td>0.0570</td>
<td>0.1531</td>
<td>0.0851</td>
<td>0.1460</td>
<td>0.0336</td>
<td>0.0296</td>
</tr>
<tr>
<td>v5</td>
<td>0.0649</td>
<td>0.1666</td>
<td>0.1281</td>
<td>0.1757</td>
<td>0.0991</td>
<td>0.0902</td>
<td>0.1569</td>
<td>0.0895</td>
<td>0.0408</td>
<td>0.1384</td>
</tr>
<tr>
<td>v6</td>
<td>0.0226</td>
<td>0.1883</td>
<td>0.1463</td>
<td>0.0919</td>
<td>0.1431</td>
<td>0.0037</td>
<td>0.0722</td>
<td>0.1540</td>
<td>0.0771</td>
<td>0.0938</td>
</tr>
<tr>
<td>v7</td>
<td>0.1686</td>
<td>0.0392</td>
<td>0.1674</td>
<td>0.0706</td>
<td>0.0436</td>
<td>0.0594</td>
<td>0.1450</td>
<td>0.1379</td>
<td>0.1343</td>
<td>0.0340</td>
</tr>
<tr>
<td>v8</td>
<td>0.1173</td>
<td>0.0192</td>
<td>0.1767</td>
<td>0.1220</td>
<td>0.0804</td>
<td>0.0771</td>
<td>0.1372</td>
<td>0.0304</td>
<td>0.1560</td>
<td>0.0899</td>
</tr>
<tr>
<td>v9</td>
<td>0.1030</td>
<td>0.0236</td>
<td>0.0412</td>
<td>0.1592</td>
<td>0.1477</td>
<td>0.0580</td>
<td>0.0810</td>
<td>0.1851</td>
<td>0.1696</td>
<td>0.0317</td>
</tr>
<tr>
<td>v10</td>
<td>0.1827</td>
<td>0.0182</td>
<td>0.0397</td>
<td>0.1487</td>
<td>0.1027</td>
<td>0.0563</td>
<td>0.0618</td>
<td>0.2830</td>
<td>0.0911</td>
<td>0.0157</td>
</tr>
</tbody>
</table>

Table 7.3: Transition matrix to exemplify practical of the example MC.

<table>
<thead>
<tr>
<th></th>
<th>v1</th>
<th>v2</th>
<th>v3</th>
<th>v4</th>
<th>v5</th>
<th>v6</th>
<th>v7</th>
<th>v8</th>
<th>v9</th>
<th>v10</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>0.0974</td>
<td>0.0912</td>
<td>0.0677</td>
<td>0.1166</td>
<td>0.0582</td>
<td>0.1376</td>
<td>0.1323</td>
<td>0.0860</td>
<td>0.0894</td>
<td>0.0975</td>
</tr>
<tr>
<td>v2</td>
<td>0.1168</td>
<td>0.1399</td>
<td>0.0969</td>
<td>0.0203</td>
<td>0.0324</td>
<td>0.0853</td>
<td>0.0929</td>
<td>0.2051</td>
<td>0.1911</td>
<td>0.0194</td>
</tr>
<tr>
<td>v3</td>
<td>0.1585</td>
<td>0.0991</td>
<td>0.0666</td>
<td>0.0957</td>
<td>0.1068</td>
<td>0.1245</td>
<td>0.1617</td>
<td>0.0597</td>
<td>0.1071</td>
<td>0.0203</td>
</tr>
<tr>
<td>v4</td>
<td>0.0575</td>
<td>0.1568</td>
<td>0.1659</td>
<td>0.1153</td>
<td>0.0570</td>
<td>0.1531</td>
<td>0.0851</td>
<td>0.1459</td>
<td>0.0335</td>
<td>0.0296</td>
</tr>
<tr>
<td>v5</td>
<td>0.0649</td>
<td>0.1666</td>
<td>0.1280</td>
<td>0.1756</td>
<td>0.0991</td>
<td>0.0902</td>
<td>0.1569</td>
<td>0.0895</td>
<td>0.0408</td>
<td>0.1384</td>
</tr>
<tr>
<td>v6</td>
<td>0.0244</td>
<td>0.2031</td>
<td>0.1578</td>
<td>0.0991</td>
<td>0.1543</td>
<td>0.0040</td>
<td>0.0811</td>
<td>0.1703</td>
<td>0.0046</td>
<td>0.1012</td>
</tr>
<tr>
<td>v7</td>
<td>0.1685</td>
<td>0.0392</td>
<td>0.1674</td>
<td>0.0706</td>
<td>0.0436</td>
<td>0.0594</td>
<td>0.1450</td>
<td>0.1378</td>
<td>0.1342</td>
<td>0.0340</td>
</tr>
<tr>
<td>v8</td>
<td>0.1173</td>
<td>0.0192</td>
<td>0.1767</td>
<td>0.1220</td>
<td>0.0803</td>
<td>0.0771</td>
<td>0.1372</td>
<td>0.0304</td>
<td>0.1560</td>
<td>0.0899</td>
</tr>
<tr>
<td>v9</td>
<td>0.1030</td>
<td>0.0236</td>
<td>0.0412</td>
<td>0.1592</td>
<td>0.1477</td>
<td>0.0579</td>
<td>0.0810</td>
<td>0.1851</td>
<td>0.1696</td>
<td>0.0317</td>
</tr>
<tr>
<td>v10</td>
<td>0.1827</td>
<td>0.0182</td>
<td>0.0397</td>
<td>0.1487</td>
<td>0.1027</td>
<td>0.0562</td>
<td>0.0617</td>
<td>0.2829</td>
<td>0.0911</td>
<td>0.0156</td>
</tr>
</tbody>
</table>

The sample complexity to learn stationary edge probabilities.

maximize lower bound
maximize edge perturbation

The perturbation of stationary edge probabilities.

maximize lower bound
maximize edge perturbation

Figure 7.1: The practical Sample Complexity bound for $\delta = 0.1$ and $\theta = 1 - 4$. 

72
Appendix A

The Probability of Strongly Typical Sequences

In this appendix we provide known results for convergence of the probability of strongly typical sequences, as their length approaches infinity. In Section A.1 we show that the probability of strongly typical sequences approaches 1, as the length of the sequences approaches infinity. However this convergence is linear in the length of the sequences. In Section A.2 we show an alternative lower bound on the probability of strongly typical sequences. When this bound converges, it converges to 1 exponentially fast. In contrast to Section A.1, the bound of Section A.2 is not guaranteed to converge.

A.1 Convergence to 1

For brevity we state Theorem A.1.1 without a proof. This theorem establishes the upper bound on the probability of sequences in a MC that are not strongly typical.

**Theorem A.1.1** ([2, Problem 3.38]) (Asymptotic Equipartition Property) Let \( M = (V, \sigma, Q) \) be an irreducible MC such that the algebraic multiplicity of its eigenvalues is 1. Let \( \mu \) be the largest absolute value of any eigenvalue of \( Q \) other than 1. Then

\[
\sigma ([\tau(\theta, l)]^c) \leq \frac{|V|^2}{\theta^2 l} \left(1 + 2|V| \frac{\mu}{1 - \mu}\right). \tag{A.1}
\]
Note that since the length $l$ of the strongly typical sequences in (A.1) appears only in the denominator,
\[
\lim_{l \to \infty} \sigma(\tau(\theta, l)) \to 1.
\]
Also note also that the convergence rate is linear in $l$.

### A.2 Exponential Rate Convergence

In Section A.1 we showed that the probability of strongly typical sequences approaches 1 linearly as the sequence length approaches infinity. In this section we follow [6] and present the lower bound (Theorem 3.1.2) on the probability of strongly typical sequences. When the bound converges, it approaches 1 exponentially fast. The drawback of this bound is that we cannot show that the bound always converges.

Let $M$ be a MC. Recall definitions (5.4) and (5.6) of $A(\alpha, e)$ and $\pi_M(\theta, l)$. Now we prove Lemma 5.4.4.

**Proof** Let $M$ be an MC and $s = (e_1, e_2, \cdots, e_l)$ be a path, chosen according to $M$ and $e \in E$. Note that $\frac{1}{l} n(e|s)$ is the empirical probability of edge $e$ on the path $s$. The probability that $s$ is not $\theta$-strongly typical is

\[
P\left(\left| \frac{n(e|s)}{l} - \sigma(e) \right| > \theta \right) = P\left(\frac{n(e|s)}{l} > \sigma(e) + \theta \right) + P\left(\frac{n(e|s)}{l} < \sigma(e) - \theta \right).
\]

We will bound each additive factor in the expression above. For any $\alpha > 0$

\[
Pr\left(\frac{n(e|s)}{l} > \sigma(e) + \theta \right) = Pr\left(\exp\left(\alpha n(e|s)\right) > \exp\left(\alpha l(\sigma(e) + \theta)\right)\right).
\]

Let $E$ denote expectation. Using Markov’s inequality

\[
Pr\left(\frac{n(e|s)}{l} > \sigma(e) + \theta \right) \leq \exp(-\alpha l(\sigma(e) + \theta))E\left(\exp\left(\alpha n(e|s)\right)\right)
\]

(A.2)

Let $\chi$ be the indicator function over edges, i.e. $\chi(e, e')$ is equal 1 if $e = e'$ and 0 otherwise. Note that for a state sequence $s = (e_1, e_2, \cdots, e_l)$

\[
n(e|s) = \sum_{t=1}^{l} \chi(e_t, e_t).
\]

74
Substituting (A.3) into (A.2), we get

\[
Pr \left( \frac{n(e|s)}{l} > \sigma(e) + \theta \right) \leq \exp(-\alpha l(\sigma(e) + \theta)) \sum_s \lambda_{\text{start}(s)} \prod_{i=1}^l [Q(e_i) \exp(\alpha \chi(e_i, e))] \\
= \exp(-\alpha l(\sigma(e) + \theta)) l^{tr} \left( A^{(\alpha,e)} \right)^l 1 \leq \exp(-\alpha l(\sigma(e) + \theta)) l^{tr} \left( A^{(\alpha,e)} \right)^l 1 .
\]

(A.4)

Similarly for any \( \alpha > 0 \)

\[
Pr \left( \frac{n(e|s)}{l} < \sigma(e) - \theta \right) \leq \exp(\alpha l(\sigma(e) - \theta)) l^{tr} \left( A^{(-\alpha,e)} \right)^l 1 .
\]

(A.5)

Now note that expressions (A.4) and (A.5) may be minimized over \( \alpha > 0 \).

Note that this estimation on the probability of strongly typical sequences of a MC does not depend on the initial node probabilities of that MC.

According to the proof of Theorem B.1.1 ([2, Theorem 3.17]), the \( l \)-th power of a primitive matrix \( A \) can be decomposed as follows

\[
A^l = \lambda^l xy^T + E^{(l)} ,
\]

where \( E^{(l)} \) is a matrix of the same order of \( A \), whose entries satisfy

\[
\left| E^{(l)}_{u,v} \right| = O \left( l^{(h-1)} \mu^l \right)
\]

for every \( u \) and \( v \).

We can use (A.6) to study exponential behavior of (A.4) and (A.5) using the largest eigenvalue \( \lambda(A^{(\alpha,e)}) \) of \( A^{(\alpha,e)} \) and \( \lambda(A^{(-\beta,e)}) \) of \( A^{(-\beta,e)} \). Alternatively we can directly compute the values of expressions (A.4) and (A.5), minimizing it over the possible values of \( \alpha \) and \( \beta \).
Appendix B

Stationary State Probability

B.1 Convergence Rate to Stationary Probability

At the end of Section 5.4 we pointed out that our bounds do not depend on the initial state probabilities of a MC. For the sake of completeness we also show that given any initial state probabilities, the state probabilities of any MC converge to the stationary state probabilities exponentially fast.

**Theorem B.1.1** ([2, Theorem 3.17]) Let $A$ be a primitive matrix and $x$ and $y^t$ be strictly positive right and left eigenvectors of $A$, associated with eigenvalue $\lambda = \lambda(A)$, normalized so that $y^t x = 1$. Let also $A = P \Lambda P^{-1}$ be the Jordan canonical form of $A$, $\mu$ be the largest absolute value of any eigenvalue of $A$ other than $\lambda$. Also let $h$ be the algebraic multiplicity of any eigenvalue of $A$ whose absolute value equals $\mu$. Then there exists a matrix $M$ of the same size as $A$ such that for any $i, j$ $|M_{i,j}| < 1$ and

$$A^t = \lambda^t y^t x + (\mu/\lambda)^{t^{h-1}} P \Lambda P^{-1}, \quad (B.1)$$

The proof of the theorem directly follows from the details of the proof of Theorem 3.17 in [2]. We also state the following feature of matrix eigenvalues.

**Lemma B.1.2** ([2, Proposition 3.16]) Let $A$ be a primitive matrix with $\lambda(A) = \lambda$. Then $|\mu| < \lambda$ for every eigenvalue $\mu \neq \lambda$ of $A$.

The convergence result follows from the following lemma that is based on [2, Proposition 3.20].
Lemma B.1.3 Let $Q$ be a stochastic $|V| \times |V|$ matrix, $\sigma^{tr}$ the stationary state probability and let $\iota^{tr} = (\iota_u)_{u \in V}$ be such that $\forall u \in V \ i[u] \geq 0$ and $\sum_{u \in V} \iota_u = 1$. Also let $\mu$, $h$, $P$ and $M$ be as in Theorem B.1.1. Then,

$$\iota^{tr} Q^l = \sigma^{tr} + \left(\frac{\mu}{\lambda}\right)^{l-1} iPM^1.$$

(B.2)

Proof Since $Q$ is stochastic we have $Q1 = 1$, where $1$ is the all-one column vector; that is, $1$ is a right eigenvector associated with the Perron eigenvalue $\lambda(Q) = 1$. Hence by Theorem B.1.1 we have

$$\iota^{tr} Q^l = \sigma^{tr} + \mu^l i^{h-1} iPM^1,$$

where for any $i, j$, $|M_{i,j}| < 1$. 

Note that $\iota^{tr} Q^l$ is the state probability after MC starts, according to initial state probabilities $\iota$, and traversed $l$ edges. Since our study is focused on ergodic MCs and ergodic matrices are primitive, from Lemma B.1.2 $|\mu| < 1$, which by (B.2) implies that state probability of any MC converges to the stationary state probability exponentially fast.

77
Bibliography


In the learning setting, the problem of distinguishing between two Markov chains with a given distance \( l = l_{\text{distinguish}}(M, M', \varepsilon, \delta) \) for every \( \varepsilon, \delta \) in the class \( M' \) is characterized by

\[
l = l_{\text{learn}}(M, M', \varepsilon, \delta) \geq \min_{v_0, v_1} Q(u, v_0) = Q(u, v_1) + \varepsilon
\]

or

\[
l = l_{\text{learn}}(M, M', \varepsilon, \delta) \geq \min_{v_0, v_1} Q(u, v_0) = Q(u, v_1) - \varepsilon
\]

This allows us to raise the following problem: for a fixed Markov chain with \( \varepsilon, \delta \) such that \( l \geq M \) and a sequence of \( Q(u, v_0) = Q(u, v_1) \), we can distinguish the two chains with a distance \( \varepsilon \).
The research on Markov models and learning theories, and in particular, has shown that the learning of a Markov model is a topic of interest in the field of computer science. The learning process is defined by the model $M = (V, i, Q)$, where $V$ represents the vocabulary, $i$ represents the initial state, and $Q$ represents the transition probabilities.

Moreover, the learning process is characterized by the transition probabilities $q$, which determine the behavior of the model. The learning process aims to find the model $M$ that best matches the data.

The learning process is often evaluated using metrics such as the error rate or the likelihood of the model. In this context, the study of learning Markov models is crucial for understanding the behavior of complex systems, especially in the field of natural language processing.

In conclusion, the learning of Markov models is a fundamental topic in computer science, and its study is essential for the development of more efficient and accurate models in various applications.
המחקר נערך במנהלת פרופ' אלון אייתי בפקולטה להנדסת מחשבים.

אני מודד לטכניקת (ולקן שינת מנהל או בבלית) על התמיכות הכספית והדינמית בשיתופי פעולה.
סיבוכיות מדגם ללימודי שרשראות מרקוב

חיבור על מחקר

לשם مليוי חלקי של הדרישת לכתבל התואר
דוקטור לפילוסופיה

רומן טלינסקי

הㄥשلطנט ה ltd חכימי — מוכן טכנולוגי ייזראל
ורן הרש"ץ
חתפתיי
יול 2014
סיבוכיות מדגמי למידת שרשרת מרקוב

רומן טליאנסקי