

**Stochastic Modeling for Efficient Computation
of Information Theoretic Quantities**

by

Adi Schreibman

Submitted to the Institute of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

THE HEBREW UNIVERSITY OF JERUSALEM
JERUSALEM, ISRAEL.

March 2000

© The Hebrew University of Jerusalem
Jerusalem, Israel. 2000

Signature of Author
Institute of Computer Science
23-2-2000

Certified by
Naftali Tishby
Professor, Hebrew University of Jerusalem
Thesis Supervisor

Certified by
Ran El-Yaniv
Doctor, Technion
Thesis Supervisor

Accepted by
Arthur C. Clarke
Chairperson, Department Committee on Graduate Students

**Stochastic Modeling for Efficient Computation
of Information Theoretic Quantities**

by
Adi Schreiber

Submitted to the Institute of Computer Science
on 23-2-2000, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Replace with the thesis abstract

Thesis Supervisor: Naftali Tishby
Title: Professor, Hebrew University of Jerusalem

Thesis Supervisor: Ran El-Yaniv
Title: Doctor, Technion

Contents

1	Introduction	5
2	Background	9
2.1	Stochastic Processes and Information Sources	9
2.1.1	Markov Chains and Markov Processes	10
2.2	Representation of a Markov Source, Tree Models	13
2.2.1	Types and Frequency functions with Respect to a Markov Model	16
2.2.2	The Type of a Corpus with Respect to a Markov Model	17
2.3	Basic Information Theoretical Quantities	18
2.3.1	Entropy and Relative Entropy	19
2.3.2	Entropy of a Process, Entropy Rate	21
2.4	The Method of Types, Definitions and Basic Results	22
2.5	Probability Estimation via the Context Tree Weighting Method	23
3	The Mutual Source and the J Function	27
3.1	Definitions	28
3.2	The Empirical Mutual Source and the Statistic J	30
3.3	The J Function for Multiple Sources	31
3.4	Properties of J	32
4	The J Function and the Two Samples Problem	37
4.1	Hypotheses Testing	37
4.1.1	Definitions	37

4.1.2	Testing Simple Hypotheses	38
4.1.3	Testing Composite Hypotheses	40
4.2	The Two Samples Problem	42
4.2.1	Introduction	42
4.2.2	Optimal Tests for the Two Samples Problem	44
5	Monotone Decrease of the Function J	47
5.1	Introduction	47
5.2	An Illustrative Example	48
5.3	Proofs of main results	49
5.4	The Two Samples Problem Revisited	54
5.4.1	Optimal Test	54
6	Conclusions and Future Research	56
6.1	The Generalization Power of Compression Algorithms	56
6.1.1	The Two Samples Problem, a Different Angle	56
6.1.2	Model Selection and the Two Samples Problem	58
6.2	Generalization to Transducers	60
6.2.1	Definitions	61
6.2.2	Implications	62

Chapter 1

Introduction

Information Theory addresses some fundamental questions about systems that store or communicate data. Among the most fundamental quantities of the theory are entropy and mutual information introduced by Shannon, and relative entropy introduced by Kullback and Leibler, at the end of the first half of the last century. In this work we focus our attention on a rather new quantity, which we refer to as J ([7]). It has meaning in coding theory, is related to other fundamental quantities and it appears in statistical analyses, such as in the solution of the two samples problem for finite Markov sources ([13]). Hence efficient computation of the J function appears to be important for many applications. For an intuitive definition of the J function, consider the following random source of binary bits: at every step it flips a coin with probability λ for head. If a head was drawn then a bit is drawn from some source P , otherwise a bit is drawn from another source Q . Then $J(P, Q)$ is the minimal cost (in bits) induced by assuming the bits came from a single source. The λ -mutual source of P and Q is a source that achieves this minimum.

From the practitioners point of view, this work provides an efficient way for computing the J function when the statistical model is known. This involves the computation of entropy, for which many procedures can be used, and in particular compression algorithms. From a more theoretical point of view, we also discuss the problems of computing the J function when the statistical model is not known. We show that knowledge of the statistical model is crucial for meaningful definition of the J function, and as a result for a meaningful solution of the two samples problem. In broad terms, the two samples problem is to decide whether two samples

are drawn from the same distribution or not. We will show how the J function serves as a statistic in the solution of this problem for Markov sources.

Throughout the work we use the idea of a representation of a Markov source (e.g models) and in particular finite order tree models ([32]). The use of models enriches the theoretical results as well as reflects on the translation of the results to applicative methods.

In more details, this thesis is primarily concerned with the following subjects:

- A rigorous definition of the J function and an organization of its properties. In particular, we investigate carefully its relationship with a closely related function, denoted JS ([18]), and remove some confusion that exists between the two. We contribute new properties that we have discovered for the J function, and we illuminate its importance in various applications.
- Efficient computation of the J function. We present two observations, which facilitate an efficient computation of J . The first observation has to do with computing properties of the mutual source. We show that in order to compute the entropy of the mutual source there is no need to generate a sample from it. In fact the computation of the J function for two samples \mathbf{x} and \mathbf{y} reduces to the following intuitive procedure: try to compress \mathbf{x} and \mathbf{y} together, compare the result to the result of compressing each separately. The second observation has to do with the model that is used to represent the sources. We show that any tree model can be used for computing the J function, assuming it describes the sources, and not only full trees as used so far.
- A discussion on the “two samples problem” for Markov sources. The J function appears in a test for the “two samples problem” suggested by Gutman [13]. This is an important aspect of the definition of the J function, and the discussion on the J function reflects on the understanding of the solution to the two samples problem. For example understanding the behavior of the J function when the complexity of the model increases helps to understand the difficulty of solving the two samples problem when the model is unknown. This is why we devote much attention to the “two samples problem” and build the discussion from the basics of hypothesis testing.
- Maximum Likelihood versus Bayesian estimation. In [13] Gutman show that a test based

on the statistic J is optimal for the two samples problem under the Neyman and Pearson asymptotic optimality criterion. The J statistic is based on probability estimation via the maximum likelihood estimator. We show that a family of related statistics can be defined and is optimal under the same Neyman and Pearson criterion. This family differs from J in using different probability estimators. We focus our attention on the Bayesian probability estimator, which is related to the context tree weighting (CTW) algorithm.

- Monotonicity of J . We show that J is monotonically decreasing with increasing complexity of the model being used. This observation that we make and prove is relevant for the use of the J function as a statistic for the two samples problem when the model is unknown. The difficulty is that for given two samples, one can select a sufficiently complex model and describe the two samples accurately enough using this model.
- Conditional sources. We discuss shortly the generalization of the J function to transducers, or as we refer to it, to conditional sources. Possible implications of such a generalization are discussed as well.

The rest of this thesis is organized as follows. Chapter 2 covers thoroughly but briefly all the definitions and results needed in the rest of the thesis. It includes the definition of stochastic processes, information sources, a type, entropy, relative entropy, entropy rate, and it gives necessary results from the theory of Markov chains and the theory of types. Lastly it specifies the basic definitions and results for the context tree weighting algorithm. In Chapter 3 we focus on the J function. We give the definition of the mutual source and the J function, and define the mutual source and the statistic J for two samples. We briefly discuss the generalization of the above definitions for multiple sources and finally give some properties of the J function and the mutual source. In Chapter 4 we discuss the J function as related to the two samples problem for finite Markov sources. We begin with an introduction to hypotheses testing and the two samples problem. Chapter 5 gives the proofs of monotonicity of the J function and discusses its relation to the solution of the two samples problem when the model is unknown. Finally Chapter 6 suggests some conclusions and ideas for future work. Mainly it discusses generalization to transducers, and the following question of selecting a test for the two samples problem: is the J function based on the maximum likelihood is best or is the known intuition

that the Bayesian estimation generalize better works also for this case.

Chapter 2

Background

This chapter is aimed to give the reader all necessary background in order to read the rest of the thesis. Section 2.1 gives the definitions of a stochastic process and an information source and states a basic theorem from the theory of Markov chains. Section 2.2 discusses the representation of a Markov source and defines the type of a sequence. Section 2.3 introduces basic information theoretic quantities and some of their properties. Section 2.4 gives some basic definitions and results from the method of types. Finally Section 2.5 introduces definitions relating to the CTW algorithm (Bayesian estimation).

2.1 Stochastic Processes and Information Sources

We wish to give a probability measure to describe an experiment which takes place in stages. This leads to the definition of a stochastic process. A *stochastic process* [5, p.60] over a finite alphabet \mathcal{X} is an indexed sequence of random variables $\{X_i\}_{i=1}^{\infty}$. The process is characterized by a family of joint probability mass functions $\Pr\{(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)\}$, $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ for all $n = 1, 2, \dots$. The family of joint probability mass functions can be thought of as the generator of the stochastic process and is sometimes treated independently, this leads to the definition of an information source. An *information source* over a finite alphabet \mathcal{X} [31] is defined by a family of joint probability mass functions $P^n(\mathbf{x})$ for all $n = 0, 1, \dots$. Each distribution P^n is defined on the set \mathcal{X}^n of all strings \mathbf{x} of length n (In general we will denote a string by \mathbf{x} , x_1^n or explicitly x_1, \dots, x_n). This family must satisfy the

marginality condition

$$P^n(\mathbf{x}) = \sum_{a \in \mathcal{X}} P^{n+1}(\mathbf{x} \circ a),$$

for every n . Here $\mathbf{x} \circ a = x_1, x_2, \dots, x_n, a$ is the concatenation of \mathbf{x} and a and $P_0(\emptyset) = 1$ where \emptyset is the empty string. Every stochastic process defines a unique information source and vice versa, thus the use of both terminologies is somewhat confusing and as a result there are different definitions for properties of stochastic processes. To demonstrate the use of both terminologies, we now give two equivalent definitions of stationarity and proof their equivalence. A stochastic process is said to be **stationary** if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index, i.e.,

$$\Pr\{(X_1, X_2, \dots, X_n) = \mathbf{x}\} = \Pr\{(X_{l+1}, X_{l+2}, \dots, X_{l+n}) = \mathbf{x}\} \quad (2.1)$$

for every l and for all $\mathbf{x} \in \mathcal{X}^n$. An information source is said to be stationary if the condition

$$P^n(\mathbf{x}) = \sum_{a \in \mathcal{X}} P^{n+1}(a\mathbf{x}) \quad (2.2)$$

holds for all n . Formally the two definitions are equivalent but they demonstrate the different views raised by the definitions of a stochastic process and an information source.

2.1.1 Markov Chains and Markov Processes

The theory of Markov chains [9, 12] considers the simplest generalization of a sequence of independent random variables. This generalization is formed by permitting any random variable X_{n+1} in the sequence to depend on the variable directly preceding it X_n , but given the preceding variable, X_{n+1} is conditionally independent of all other variables in the sequence.

Definition 2.1.1 ([5]) *A Markov process or a Markov chain*¹ *is a stochastic process over*

¹The distinction between Markov chains and Markov processes varies between different books. For example [12] defines a Markov chain as a time invariant Markov Process, while [9] distinguishes between the two definitions only in that Markov processes are described in terms of random variable, and Markov chains in terms of trials. We do not distinguish between the two, as in [5].

a finite alphabet $\mathcal{X} = \{a_i\}_{i=1}^m$, such that $P^n(x_1^n)$ is given by

$$P^n(x_1^n) = P^1(x_1) \prod_{i=2}^n P^i(x_i|x_{i-1}),$$

for all $n > 1$ and $x_1^n \in \mathcal{X}^n$.

Definition 2.1.2 ([5, 9]) A Markov chain is said to be **time invariant** or **with constant transition probabilities** if the conditional probability $P^i(x_i|x_{i-1})$ does not depend on i .

For a Markov process the letters of \mathcal{X} are also referred to as **states**. A time invariant Markov chain is characterized by its initial probability vector $\pi = (P^1(a_1), \dots, P^1(a_m))$ and its **transition matrix** $T = (p_{ij}) = (P^2(x_i|x_j))$, and is a stationary process if

$$\pi_k = \sum_j \pi_j p_{jk}. \quad (2.3)$$

A probability distribution π satisfying Eq. (2.3) is called stationary. In the sequel we assume that any Markov chain is time invariant unless otherwise is stated. Next we give some basic definitions and results for Markov chains. We denote by $p_{jk}^{(n)}$ the probability that the state at time n is a_k given that the initial state is a_j ; that is, $p_{jk}^{(n)}$ is defined recursively by,

$$p_{jk}^{(1)} = p_{jk}, \quad (2.4)$$

$$p_{jk}^{(n+1)} = \sum_i p_{ji} p_{ik}^{(n)}. \quad (2.5)$$

Definition 2.1.3 ([9]) A set C of states is called **closed** if $p_{jk}^{(n)} = 0$ for all $a_j \in C$, $a_k \in \mathcal{X} \setminus C$ and $n > 0$. A Markov chain is called **irreducible** if there exists no closed set other than \mathcal{X} .

Definition 2.1.4 ([9]) A state a_i is said to have **period** $t > 1$ if $p_{ii}^{(n)} = 0$ whenever n is not divisible by t and t is the smallest such integer with this property. In this case the state is called **periodic** with period t . A state a_i is called **aperiodic** if there is no $t > 1$ such that a_i has period t . A Markov chain is called **aperiodic** if all its states are aperiodic.

Denote by f_{ij} the probability of visiting state a_j after leaving state a_i (Note that this is not the transition probability), and by H_{ij} the expected number of steps from state i to state j .

Definition 2.1.5 ([9]) A state a_i is called **persistent** if $f_{ii} = 1$. A state a_i is called a **null state** if $H_{ii} = \infty$. Persistent states which are neither periodic nor null states will be called **ergodic**.

Theorem 2.1.1 ([9]) If a Markov chain is irreducible, finite with n states, and aperiodic, then

1. All its states are ergodic.
2. There exists a unique stationary distribution π , and for all i such that $1 \leq i \leq n$, $\pi_i > 0$.
3. For all i such that $1 \leq i \leq n$, $f_{ii} = 1$ and $H_{ii} = \frac{1}{\pi_i}$,
4. Let $N(i, t)$ be the number of visits the Markov chain makes to state i in t steps. Then $\frac{N(i, t)}{t}$ converges to π_i with probability 1 as t goes to infinity.

Definition 2.1.6 ([9]) A Markov chain is said to be **ergodic** if it is irreducible and aperiodic.

Higher order processes Sometimes we are interested in processes, where the distribution of every random variable depends not only on the variable directly preceding it, but rather on the first D variables preceding it.

Definition 2.1.7 ([31, 9]) A stochastic process over a finite alphabet \mathcal{X} , is said to have the **Markov property** if there exists a natural positive number D such that $P^n(x_1^n)$ is given by

$$P^n(x_1^n) = P^D(x_1, \dots, x_D) \prod_{i=D+1}^n P^i(x_i | x_{i-1}, \dots, x_{i-D}), \quad (2.6)$$

for all $n > 1$ and $x_1^n \in \mathcal{X}^n$. The minimal number D such that Eq. (2.6) holds for, is called the **order of the process**.

A process with the Markov property and order D will be termed, a **D -order Markov process**². The class of such processes are termed **finite order Markov processes**. It is not difficult to see that a D -order Markov process over the alphabet \mathcal{X} defines a unique Markov-chain over the alphabet \mathcal{X}^D . Thus all the definitions and results for Markov chains, can be translated to the case of finite order Markov processes. For example, for a D -order Markov process, we refer to the elements of \mathcal{X}^D as the states of the process.

²Processes with the Markov property and order 1, we will term Markov chains.

2.2 Representation of a Markov Source, Tree Models

We usually describe a Markov process by specifying the probability of the next character, given every possible history string. For example we say: “consider the Markov process over $\{0, 1\}$ that after a string that ends with 1 gives probability $\frac{1}{4}$ to see 1, and gives probability $\frac{1}{2}$ to see 1 otherwise”³. Another example might be: “consider the Markov process over $\{0, 1\}$ that gives probability $\frac{1}{4}$ to see 1 if the history string is a code of a program that stops on its own input and gives probability $\frac{1}{2}$ to see 1 otherwise”. Both examples are of well defined Markov chains, but the first one can easily be simulated by a computer, while the second one not. When considering applications we must concentrate on processes that have a “simple representation”. This section discusses the concept of a representation and introduces tree models representation.

Assume we receive the sequence $x = 0100010100011110$ drawn from some stationary and ergodic source P , and we are asked to estimate the probability $P(x)$ assigned to this sequence by P . If we knew that P is a 1-order Markov source then we could partition x into two subsequences of symbols $x_1 = 0001110$ and $x_2 = 10011001$ each drawn from a 0-order source (i.e. sequences of Bernoulli trials). x_1 represents the symbols that were drawn when P was in state 1 and x_2 represents the symbols that were drawn when P was in state 0. The first symbol does not appear in x_1 or x_2 since the sequence does not reveal the state of P prior to the generation of x . We can then use the maximum likelihood estimator or any other estimator to estimate the probability assigned to x_1 and x_2 by P . This procedure is very simple and can be computed using one pass over the sequence x . Assume next that our knowledge was that P is a 4-order Markov source. We can follow the lines of the above procedure and partition x into $2^4 = 16$ subsequences x_1, \dots, x_{16} leaving the 4 first sequences out. The problem is that some of the sequences are very short and our estimation error will be very big. We would like to be able to combine some of the shorter sequences, e.g. x_1, \dots, x_4 into single sequences yielding longer sequences. These combined sequences are no longer necessarily drawn from a 0-order source, so we expect our estimation of their probability (by assuming they are i.i.d) to bear an error but we hope that it will be smaller than the one for estimating the shorter sequences⁴. We would like that our

³Note that this defines only the transition probabilities, but assuming that the process is stationary and ergodic the initial distribution is also fixed.

⁴In fact this is an example of a mutual source and the error we will bear is equal to the J function between

algorithm will still work efficiently if we do these kind of operations. An example of another important property is that the algorithm be online, this is crucial for compression algorithms. Not every representation will allow for efficient implementation of the above procedures. For example if we have decided for 010001 that it is not the binary code of a program that stops on it's own input it usually gives us no knowledge on 0100010 or 01000101.

A Markov source can be represented by a function $s : \mathcal{X}^* \rightarrow S$ from the space of all finite strings to a finite space S that defines all possible states of the process. Every state in S corresponds to a transition distribution over \mathcal{X} . The structure of s is the more complicated part of a representation of a source, and different families were considered, e.g. FSM and FSMX sources defined by Rissanen ([23]). In this section we will describe tree models, also defined by Rissanen ([31]). It can be shown that tree models representation allows for efficient algorithms for the operations stated above. Variants of these algorithms are used for example in many model based compression algorithms ([32, 2, 4]). They are referred to as context, context tree or a suffix tree algorithms (see also [23, 22]).⁵

Definition 2.2.1 *A set $B \subset \mathcal{X}^n$ is called **proper** if every semi infinite string \dots, x_{n-1}, x_n has a suffix in B . B is called **complete** if no element of B is a suffix of another element of B .*

Definition 2.2.2 ([32]) *A **tree model** is a set $S \subset \mathcal{X}^n$ that is proper and complete.*

A tree model is termed this way since every tree model S can be represented by a tree \mathcal{T} labeled by strings over \mathcal{X} as follows (see Fig. 2-1). The root of \mathcal{T} is labeled by the empty string, and every other node v is labeled by some string s . If s is in S then v is a leaf of the tree, else it has $|\mathcal{X}|$ sons labeled by $s \circ a$ for every $a \in \mathcal{X}$ (where $s \circ a$ is the concatenation of s and a). A tree represented in this way is sometimes called a **suffix tree**, and S is also called a **suffix set**.

Given a model S , there are many sources that can be represented by it, one for every assignment of a distribution vector to it's states. Many times it is convenient to refer to all the sources that a model can describe. This is the motivation for the definition of a model class.

the distributions that we combine (we will elaborate on this later).

⁵Note that we use the model representations for theoretical purposes and will not discuss these practical issues. But it is relevant when coming to translate the results we state in this thesis into algorithms that the representations we use allow for efficient computation.

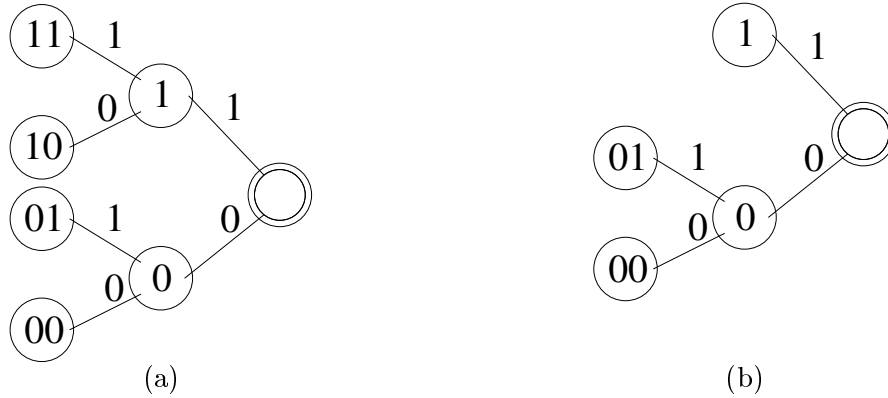


Figure 2-1: (a) A full binary suffix tree of depth 2. Corresponds to the suffix set $\{01, 00, 11, 10\}$
 (b) A subtree of the tree in (a). Corresponds to the suffix set $\{1, 01, 00\}$

Definition 2.2.3 Let S be a tree model, then the class of all stationary and ergodic sources that can be represented by S is called the **model class** of S , and is denoted by $\mathbf{P}(S)$.

By saying that a source P can be represented by S we mean that, if D is the maximum length of a string in S , then for every string v over \mathcal{X} that is longer than D , and every $a \in \mathcal{X}$, it holds that

$$P(a|v) = P(a|S(v)),$$

where $S(v)$ is the suffix of v in S .

Remark 2.2.1 The fact we have specified that all sources in $\mathbf{P}(S)$ are stationary and ergodic entails that any information source in $\mathbf{P}(S)$ is defined by a single conditional distribution rather than an infinite family of distributions. To explain this let D denote the longest suffix in S and let $P \in \mathbf{P}(S)$. To see that P is defined by a finite number of distributions is easy, all distributions P^n such that n exceeds D are redundant. To see that $P^D(X_n|X_{n-1}, \dots, X_{n-D})$ is enough, we use an example. Assume $D = 1$, then by Theorem 2.1.1 we know there is a unique stationary distribution π such that

$$\pi(i) = \sum_j \pi(j)P(i|j) \quad i, j \in \{0, 1\}. \quad (2.7)$$

To see that $P(j|i)$ defines $P^2(i, j)$, it remains to show that the unique stationary distribution π

is determined by $P(j|i)$. But this can easily be done by solving the $|\mathcal{X}| - 1$ by $|\mathcal{X}| - 1$ equations.

To understand the intuition for the definition of a tree source, consider a D -order Markov source P over \mathcal{X} , and assume that for some $s \in \mathcal{X}^{D-1}$ and all $b \in \mathcal{X}^D$

$$P(a|s \circ b) = P(a|s),$$

for all $a \in \mathcal{X}$. Thus using all the strings $s \circ b$ when describing the model is redundant, and we can describe the model by a smaller set of strings that still has a simple representation. We will not go into the details of what makes some representations simpler to use than others, and just note that there can be many cases of redundant parameters, but it is not always easy in practice to lump them all together.

2.2.1 Types and Frequency functions with Respect to a Markov Model

We now present the basic definition of empirical Markov distribution (type) of a sequence. Then we will extend the definition for several sequences. These definitions demonstrate the use of models in theoretical discussions; when defining the empirical distribution of a sequence we must assume a model, the more descriptive the model is, the more descriptive our theoretical results are. See for example Eq. (2.34)-(2.37).

Let \mathbf{x} be a sequence of n symbols from an alphabet \mathcal{X} and let $S \subset \mathcal{X}^*$ be a model, where \mathcal{X}^* is the set of finite strings of symbols in \mathcal{X} . The **type** $P_{\mathbf{x}}^S$ of \mathbf{x} with respect to the model S is the empirical joint probability distribution, defined as follows

$$P_{\mathbf{x}}^S(s, a) = \frac{N_{\mathbf{x}}(s, a)}{n} \quad \text{for all } s \in S, a \in \mathcal{X},$$

where $N_{\mathbf{x}}(s, a)$ is the number of times the letter a had occurred in the state s in \mathbf{x} . For example if S is a tree model then $N_{\mathbf{x}}(s, a)$ is the number of times the substring $s \circ a$ occurred in \mathbf{x} . In order to ensure that $\sum_{s \in S, a \in \mathcal{X}} P_x^S(s, a) = 1$ we use a convention where \mathbf{x} is cyclic; that is, the first symbol in the sequence \mathbf{x} is considered to be the symbol which follows the last symbol in \mathbf{x} . In the sequel we omit the superscript S from $P_{\mathbf{x}}^S$ whenever it is clear from the context. We

denote the marginal probability of a state as

$$P_{\mathbf{x}}^S(s) = \sum_{a \in \mathcal{X}} P_{\mathbf{x}}^S(s, a) = \frac{N_{\mathbf{x}}(s)}{n} \quad \text{for all } s \in S,$$

and the conditional probability of a symbol given a state as

$$P_{\mathbf{x}}^S(a|s) = \frac{P_{\mathbf{x}}^S(s, a)}{P_{\mathbf{x}}^S(s)} = \frac{N_{\mathbf{x}}(s, a)}{N_{\mathbf{x}}(s)} \quad \text{for all } s \in S, a \in \mathcal{X}.$$

Lemma 2.2.1 (Maximum Likelihood) *For all $\mathbf{x} \in \mathcal{X}^n$*

$$P_{\mathbf{x}}^S = \arg \max_{P \in \mathbf{P}(S)} P(\mathbf{x}). \quad (2.8)$$

Proof. By simple derivation. See also Theorem 2.2.2. ■

2.2.2 The Type of a Corpus with Respect to a Markov Model

A set $\{\mathbf{x}_i\}_{i=1}^m$ of m sequences of arbitrary lengths $\mathbf{x}_i = x_1^i \dots x_{n_i}^i$ over the alphabet \mathcal{X} is called a **corpus**. The **type of the corpus** $\{\mathbf{x}_i\}_{i=1}^m$ with respect to a Markov model S , denoted $P_{\{\mathbf{x}_i\}}^S$, is its empirical conditional probability distribution defined as follows

$$P_{\{\mathbf{x}_i\}}^S(s, a) = \frac{\sum_{i=1}^m N_{\mathbf{x}_i}(s, a)}{\sum_{i=1}^m n_i} \quad \text{for all } s \in S, a \in \mathcal{X}.$$

Remark 2.2.2 *Let $\bar{\mathbf{x}} = \mathbf{x}^1 \dots \mathbf{x}^m$ be the concatenation of all the sequences in the corpus. Note that $P_{\{\mathbf{x}_i\}}^S$ is different from $P_{\bar{\mathbf{x}}}^S$ only in that \mathbf{x}_i and \mathbf{x}_j are treated as independent samples. This means that the type of the corpus does not account for the occurrences of letters of \mathbf{x}_i after suffixes in \mathbf{x}_{i-1} , while the type of the concatenation does.*

Lemma 2.2.2 (Maximum Likelihood) *For all $\mathbf{x}_i \in \mathcal{X}^{n_i}$ $i = 1..m$*

$$P_{\{\mathbf{x}_i\}}^S = \arg \max_{P \in \mathbf{P}(S)} \prod_{i=1}^m P(\mathbf{x}_i). \quad (2.9)$$

Proof. Assume $P \in \mathbf{P}(\emptyset)$, the general case is easily deduced. For simplicity denote $\mathcal{X} = \{a_i\}_{i=1}^k$, $p_i = P(a_i)$ $i = 1..k$ and $n_{i,j} = N_{\mathbf{x}_j}(a_i)$ $i = 1..k, j = 1..m$. We can express Eq. (2.9) by,

$$\arg \max_{P \in \mathbf{P}(\emptyset)} \prod_{j=1}^m P(\mathbf{x}_j) = \arg \min_{P \in \mathbf{P}(\emptyset)} - \prod_{j=1}^m P(\mathbf{x}_j) = \arg \min_{P \in \mathbf{P}(\emptyset)} - \sum_{j=1}^m \log P(\mathbf{x}_j) = \quad (2.10)$$

$$\arg \min_{P \in \mathbf{P}(\emptyset)} - \sum_{i=1}^k \sum_{j=1}^m \log P_i^{n_{i,j}} = \arg \min_{P \in \mathbf{P}(\emptyset)} - \sum_{i=1}^k \sum_{j=1}^m n_{i,j} \log P_i. \quad (2.11)$$

We can write the constraint minimization using Lagrange multipliers as the minimization of

$$\mathcal{J} = - \sum_{i=1}^k \sum_{j=1}^m n_{i,j} \log P_i + c \left(\sum p_i \right). \quad (2.12)$$

Differentiating with respect to p_i , we obtain

$$\frac{\partial \mathcal{J}}{\partial p_i} = - \frac{\sum_{j=1}^m n_{i,j}}{p_i} + c, \quad (2.13)$$

setting the derivative to 0, we obtain

$$p_i = \frac{\sum_{j=1}^m n_{i,j}}{c}, \quad (2.14)$$

substituting in the constraints to find c , we find $c = \sum_{i=1}^k \sum_{j=1}^m n_{i,j}$ and hence

$$p_i = \frac{\sum_{j=1}^m n_{i,j}}{\sum_{i=1}^k \sum_{j=1}^m n_{i,j}} = P_{\{\mathbf{x}_i\}}^S(a_i). \quad (2.15)$$

■

2.3 Basic Information Theoretical Quantities

In this section we give some of the basic definitions and properties of information theoretic quantities. All random variables and distributions are assumed to be discrete. For more information see [5]. We use the notation \times for Cartesian product, e.g. $\{a, b\} \times \{c, d\} = \{(a, c)(b, c)(a, d)(b, d)\}$. And we also use the notation \mathcal{X}^n for the n times Cartesian product of

\mathcal{X} with itself, e.g. $(\{a, b\})^2 = \{(a, a)(b, a)(a, b)(b, b)\}$.

2.3.1 Entropy and Relative Entropy

Definition 2.3.1 *The entropy $H(X)$ of a random variable X is defined as*

$$H(X) = - \sum_{a \in \mathcal{X}} P(a) \log P(a), \quad (2.16)$$

where X is distributed according to P . If $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ then $H(X)$ is also called the **joint entropy** of X_1, \dots, X_n and is denoted $H(X_1, \dots, X_n)$, where X_i is the projection of X onto \mathcal{X}_i . The **conditional entropy** of X_n given X_{n-1}, \dots, X_1 is defined by

$$H(X_n | X_{n-1}, \dots, X_1) = - \sum_{s \in \mathcal{X}^{n-1}} P(s) \sum_{a \in \mathcal{X}} P(a|s) \log P(a|s). \quad (2.17)$$

We will identify random variables with their distributions and thus denote the entropy by $H(P)$ and refer to it as the entropy of the distribution. Note that we will also denote conditional entropy by $H(P)$ and will state explicitly that it is conditional entropy only when it is not clear from context.

Definition 2.3.2 *The relative entropy $D_{kl}(P||Q)$ of the distribution P with respect to the distribution Q is defined by*

$$D_{kl}(P||Q) = \sum_{a \in \mathcal{X}} P(a) \log \frac{P(a)}{Q(a)}.$$

Here again, if $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ then $D_{kl}(P||Q)$ is also called the **joint relative entropy** of $P(X_1, \dots, X_n)$ and $Q(X_1, \dots, X_n)$. The **conditional relative entropy** of $P(X_n | X_{n-1}, \dots, X_1)$ with respect to $Q(X_n | X_{n-1}, \dots, X_1)$ is defined by

$$D_{kl}(P||Q) = \sum_{s \in \mathcal{X}^{n-1}} P(s) \sum_{a \in \mathcal{X}} P(a|s) \log \frac{P(a|s)}{Q(a|s)}. \quad (2.18)$$

Again note that we use the same notation for conditional and unconditional relative entropy. Following are some useful properties:

Lemma 2.3.1 (Chain rules [5])

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1), \quad (2.19)$$

$$D_{kl}(P(X_1, X_2)||Q(X_1, X_2)) = D_{kl}(P(X_1)||Q(X_1)) + D_{kl}(P(X_2|X_1)||Q(X_2|X_1)) \quad (2.20)$$

The next lemma which proof is omitted states that when S is a tree model, then the conditional entropy of a source P and the conditional relative entropy of P with respect to a source Q can be written as the conditional entropy (relative entropy) with respect to each state in S , averaged over the probabilities of the states.

Lemma 2.3.2 *Let P and Q be two distributions over \mathcal{X}^n , and assume the dependencies induced by them are described by the tree model S then*

$$D_{kl}(P||Q) = \sum_{s \in S} P(s) \sum_{a \in \mathcal{X}} P(a|s) \log \frac{P(a|s)}{Q(a|s)}, \quad (2.21)$$

$$H(P) = - \sum_{s \in S} P(s) \sum_{a \in \mathcal{X}} P(a|s) \log P(a|s), \quad (2.22)$$

where both the entropy and the relative entropy are conditional. In the sequel we will mention this only when it is not clear from context. The next two lemmas are useful properties.

Lemma 2.3.3 *Let $P \in \mathbf{P}(S)$, $\mathbf{x} \in \mathcal{X}^n$ and let $P_{\mathbf{x}}$ be the type of \mathbf{x} with respect to S , then*

$$-\frac{1}{n} \log(P_{\mathbf{x}}(\mathbf{x})) = H(P_{\mathbf{x}}), \quad (2.23)$$

$$-\frac{1}{n} \log(P(\mathbf{x})) = D_{kl}(P_{\mathbf{x}}||P) + H(P_{\mathbf{x}}). \quad (2.24)$$

Proof. Eq. (2.23) follows easily from Eq. (2.24). To see Eq. (2.24),

$$-\frac{1}{n} \log(P(\mathbf{x})) = -\frac{1}{n} \log \left(\prod_{s \in S} \prod_{a \in \mathcal{X}} P(a|s)^{N_{\mathbf{x}}(s,a)} \right) \quad (2.25)$$

$$= -\frac{1}{n} \sum_{s \in S} \sum_{a \in \mathcal{X}} \log \left(P(a|s)^{N_{\mathbf{x}}(s,a)} \right) \quad (2.26)$$

$$= -\frac{1}{n} \sum_{s \in S} \sum_{a \in \mathcal{X}} N_{\mathbf{x}}(s, a) \log (P(a|s)) \quad (2.27)$$

$$= - \sum_{s \in S} \frac{N_{\mathbf{x}}(s)}{n} \sum_{a \in \mathcal{X}} \frac{N_{\mathbf{x}}(s, a)}{N_{\mathbf{x}}(s)} \log(P(a|s)) \quad (2.28)$$

$$= - \sum_{s \in S} P_{\mathbf{x}}(s) \sum_{a \in \mathcal{X}} P_{\mathbf{x}}(a|s) \log(P(a|s)) \quad (2.29)$$

$$= \sum_{s \in S} P_{\mathbf{x}}(s) \sum_{a \in \mathcal{X}} P_{\mathbf{x}}(a|s) \log \left(\frac{P_{\mathbf{x}}(a|s)}{P(a|s)} \right) - \sum_{s \in S} P_{\mathbf{x}}(s) \sum_{a \in \mathcal{X}} P_{\mathbf{x}}(a|s) \log(P_{\mathbf{x}}(a|s)) \quad (2.30)$$

$$= D_{kl}(P_{\mathbf{x}}||P) + H(P_{\mathbf{x}}). \quad (2.31)$$

■

Note the importance of the cyclic definition of $N_{\mathbf{x}}$ to the neatness of the identity.

Lemma 2.3.4 (Convexity of relative entropy [5, p.30]) *For all probability distributions P_1, P_2, Q_1, Q_2 and $\lambda \in [0, 1]$,*

$$D_{kl}(\lambda P_1 + (1 - \lambda)P_2 || \lambda Q_1 + (1 - \lambda)Q_2) \leq \lambda D_{kl}(P_1 || Q_1) + (1 - \lambda)D_{kl}(P_2 || Q_2).$$

2.3.2 Entropy of a Process, Entropy Rate

So far we have defined entropy and relative entropy for random variables, or equivalently for distributions. In this section we will define the entropy and relative entropy of stochastic processes and information sources. We will specify a Lemma based on Remark 2.2.1, saying that for finite order stationary and ergodic Markov processes these definitions reduces to the equivalent definitions for distributions.

Definition 2.3.3 ([5]) *The **entropy rate** of a stochastic process $\{X_i\}_{i=1}^{\infty}$ is defined by*

$$H(\mathbf{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n), \quad (2.32)$$

or alternatively by

$$H'(\mathbf{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_n | X_{n-1}, \dots, X_1), \quad (2.33)$$

when the limits exists.

Lemma 2.3.5 ([5, p.64]) *For a stationary process $\{X_i\}_{i=1}^{\infty}$, $H(\mathbf{X})$ and $H'(\mathbf{X})$ are defined and equal.*

The next lemma characterizes the entropy rate of a stationary and ergodic Markov process as the entropy of the distribution that defines the process.

Lemma 2.3.6 *Let $P \in \mathbf{P}(S)$, and let $\{X_i\}_{i=1}^\infty$ be the process defined by P . Then the entropy rate of the process equals the conditional entropy of P .*

Proof. The proof is based on the claim of Remark 2.2.1. We omit the full details here. ■

For two Information sources it is possible to define their relative entropy rate in a similar form as in Eq. (2.33) (see [37]). Again for $P, Q \in \mathbf{P}(S)$ this rate will equal the conditional relative entropy of P and Q .

2.4 The Method of Types, Definitions and Basic Results

We give in this section a short overview of the basic definitions and results from the method of types that is used in this thesis. For a thorough exposition see [5, ch.12], [6]. As usual, in this section S denotes a tree model, P denotes a distribution in $\mathbf{P}(S)$ and we consider sequences in \mathcal{X}^n . Recall the definition of a type from Section 2.2.1.

Definition 2.4.1 ([5]) *Denote by \mathcal{P}_n^S the set of all possible empirical distributions (types) of a sample of length n with respect to the model S , that is*

$$\mathcal{P}_n^S = \{P_{\mathbf{x}}^S : \mathbf{x} \in \mathcal{X}^n\}.$$

Definition 2.4.2 ([5]) *Let $P \in \mathcal{P}_n^S$. The set of all sequences of length n with type P is called the **type class** of P , and is denoted by $T(P)$, that is*

$$T(P) = \{\mathbf{x} \in \mathcal{X}^n : P_{\mathbf{x}}^S = P\}.$$

We now state some basic results of the method of types:

$$|\mathcal{P}_n^S| \leq (n+1)^{|S||\mathcal{X}|}, \quad (2.34)$$

$$P(\mathbf{x}) = 2^{-n(D_{kl}(P_{\mathbf{x}}||P)+H(P_{\mathbf{x}}))}, \quad (2.35)$$

$$\frac{1}{(n+1)^{|S||\mathcal{X}|}} 2^{-nH(P_{\mathbf{x}})} \leq |T(P_{\mathbf{x}})| \leq 2^{-nH(P_{\mathbf{x}})}, \quad (2.36)$$

$$\frac{1}{(n+1)^{|S||\mathcal{X}|}} 2^{-nD_{kl}(P_{\mathbf{x}}||P)} \leq P(T(P_{\mathbf{x}})) \leq 2^{-nD_{kl}(P_{\mathbf{x}}||P)}. \quad (2.37)$$

Where $P_{\mathbf{x}}$ is the type of \mathbf{x} with respect to the model S . Note that $|S||\mathcal{X}|$ is the number of parameters of the distribution P .

Definition 2.4.3 A sequence \mathbf{x} is called ϵ -**typical** with respect to the distribution P if $D_{kl}(P_{\mathbf{x}}||P) \leq \epsilon$. The set of all sequences of length n that are ϵ -typical with respect to the distribution P is called the ϵ -**typical set** for P , and is denoted by

$$T_P^\epsilon = \{\mathbf{x} \in \mathcal{X}^n : D_{kl}(P_{\mathbf{x}}||P) \leq \epsilon\}. \quad (2.38)$$

The next theorem relates to large deviation theory. It bounds the probability of the non-typical set for all n .

Theorem 2.4.1 (Sanov's theorem [5, p.292]) Let E be a set of probability distributions and let x_1^n be a sample drawn from $P \in \mathbf{P}(S)$. Denote by $P(E \cap \mathcal{P}_n^S)$ the probability of all sequences \mathbf{x} such that their type is in E , that is $P(E \cap \mathcal{P}_n^S) = P\left(\bigcup_{Q \in E \cap \mathcal{P}_n^S} T(Q)\right)$, then

$$P(E \cap \mathcal{P}_n^S) \leq (n+1)^{|S||\mathcal{X}|} 2^{-nD_{kl}(P^*||P)}, \quad (2.39)$$

where $P^* = \arg \min_{\hat{P} \in E} D_{kl}(\hat{P}||P)$.

Corollary 2.4.1

$$P(\overline{T_P^\epsilon}) \leq (n+1)^{|S||\mathcal{X}|} 2^{-n\epsilon}, \quad (2.40)$$

where $\overline{T_P^\epsilon} = \mathcal{X}^n - T_P^\epsilon$ is the complement of T_P^ϵ .

Proof. Let E denote all distributions that are not ϵ -close to P in relative entropy. Then $\overline{T_P^\epsilon} = E \cap \mathcal{P}_n^S$ and $\min_{\hat{P} \in E} D_{kl}(\hat{P}||P) \geq \epsilon$ by definition. ■

2.5 Probability Estimation via the Context Tree Weighting Method

Probability estimation refers to the problem of estimating the probability of a sample drawn from an unknown information source. To solve this problem we must assume something about

the family of possible sources. We describe here a method for estimating the probability of a sample drawn from an unknown probability in $\mathbf{P}(S)$ for some unknown model S . This method is termed the Context Tree Weighting method and is introduced in [32]. The main idea behind this method is that to solve the probability estimation problem it is not mandatory to estimate the model or the parameters of the unknown source. Instead it is possible to use a Bayesian framework and average over all possible models according to some prior distribution. The novelty in [32] is that they showed an algorithm for the computation of this average for a family of priors. We give here the main definitions and results from [32]. For elaborations see [10, 33, 34]

Definition 2.5.1 ([32]) *The **cost** of a model S with respect to the order D is defined by*

$$\Gamma_D(S) = |S| - 1 + |\{s : s \in S, |s| \neq D\}|.$$

It is not hard to show [32] that $\Gamma_D(S)$ is the number of bits needed to distinguish the model S from all other binary tree models of order smaller than or equal to D .

For the rest of this section, let $|\mathcal{X}| = k$ and $\mathcal{X} = \{a_1, \dots, a_k\}$. Next we define a probability distribution over \mathcal{X}^n by averaging over all distributions in $\mathbf{P}(\emptyset)$. The distributions $P \in \mathbf{P}(\emptyset)$ are identified by their parameter set $\Theta = (\theta_1, \dots, \theta_k) = (P(a_1), \dots, P(a_k))$.

Definition 2.5.2 ([32]) *The **Krichevski-Trofimov ([17]) estimated probability** for a sequence $\mathbf{x} \in \mathcal{X}^n$ is defined as*

$$P_e(\mathbf{x}) \triangleq \int P(\Theta)P(\mathbf{x}|\Theta)d\Theta,$$

where $P(\mathbf{x}|\Theta) = \prod_i \theta_i^{N_{\mathbf{x}}(a_i)}$, and

$$P(\Theta) = \frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})^k} \prod_{i=1}^k \theta_i^{\frac{1}{2}},$$

is the **beta distribution** ([14]) with parameters $(\frac{1}{2}, \frac{1}{2}, \dots, \frac{1}{2})$. Used as a prior distribution it is also termed the $(\frac{1}{2}, \frac{1}{2}, \dots, \frac{1}{2})$ -**Dirichlet prior**. It is possible to show that this prior is a **conjugate** or **natural prior** for this case, i.e. the posterior have the same form as the prior.

Some properties of the KT-probability estimator are listed in the following lemmas. Lemma 2.5.1 states that the KT-estimator can be computed sequentially. Lemma 2.5.2 bounds the estimation error induced by using the KT-probability estimator.

Lemma 2.5.1 (Sequential computation [32]) $P_e(\emptyset) = 1$, and for all $\mathbf{x} \in \mathcal{X}^n$ and $a \in \mathcal{X}$

$$P_e(\mathbf{x} \circ a) = \frac{N_{\mathbf{x}}(a) + \frac{1}{2}}{n + \frac{k}{2}} \cdot P_e(\mathbf{x}).$$

Lemma 2.5.2 ([32]) For $\mathbf{x} \in \mathcal{X}^n$

$$-\log P_e(\mathbf{x}) \leq -\log P_{\mathbf{x}}(\mathbf{x}) + \frac{k}{2} \log n + O(1).$$

For a given model S we can estimate the probability of a sequence by partitioning it to $|S|$ conditionally independent subsequences (Note that some of the subsequences have zero length). We will denote this probability estimator by,

$$P_e^S(\mathbf{x}) \triangleq \prod_{s \in S} P_e(\mathbf{x}_s),$$

where \mathbf{x}_s consists of all the symbols that follows s in \mathbf{x} . As in the definition of a type of a sequence we consider \mathbf{x} to be cyclic. The next lemma gives a bound on the error estimation for P_e^S . The function γ is used to specify the error in order to separate the case where the sample is longer than the size of the model from the case where it is not.

Definition 2.5.3 ([32])

$$\gamma(z) \triangleq \begin{cases} z & \text{for } 0 \leq z \leq 1 \\ \frac{k}{2} \log z + O(1) & \text{for } z > 1 \end{cases} \quad (2.41)$$

Lemma 2.5.3 ([32]) For $\mathbf{x} \in \mathcal{X}^n$

$$-\log P_e^S(\mathbf{x}) \leq -\log P_{\mathbf{x}}(\mathbf{x}) + |S| \gamma\left(\frac{n}{|S|}\right).$$

When the model is not known, it is possible to estimate the probability of a sequence by weighting over $P_e^S(\mathbf{x})$ with respect to all possible models. The weighted probability of a sequence $\mathbf{x} \in \mathcal{X}^n$ is defined by

$$P_w(\mathbf{x}) = \sum_{S \in \mathcal{C}_D} 2^{-\Gamma_D(S)} P_e^S(\mathbf{x}), \quad (2.42)$$

where \mathcal{C}_D is the class of all tree models with finite order D . The next lemma gives a bound on the error estimation for P_w .

Lemma 2.5.4 ([32]) For $\mathbf{x} \in \mathcal{X}^n$

$$-\log P_w(\mathbf{x}) \leq \min_{S \in \mathcal{C}_D} \left\{ \min_{P \in \mathbf{P}(S)} \left\{ -\log P(\mathbf{x}) + \Gamma_D(S) + |S| \gamma \left(\frac{n}{|S|} \right) \right\} \right\}. \quad (2.43)$$

The CTW algorithm is related to the Minimum Description Length (MDL) principle for model selection (see [28, 1, 29, 30]). Given a sequence, the MDL principle suggests to select the model that minimizes the code length needed to describe the model plus the code length needed to describe the sequence relative to the model. Eq. (2.43) suggests that $P_w(\mathbf{x})$ is exactly the minimum description length. This is reasonable since the CTW algorithm does not use any model for compression. Thus the model suggested by the MDL principle is given by

$$S^{MDL} = \arg \min_{S \in \mathcal{C}_D} (-\log P_e^S(\mathbf{x}) + \Gamma_D(S)). \quad (2.44)$$

Note that the above is true if the class of models to choose from is the class of all tree models of order D .

Chapter 3

The Mutual Source and the J Function

The quantities defined in Section 2.3, arise in many results in statistics and coding theory. For example, the entropy of a random source P , is a lower bound on the average description length of any code assuming the symbols are drawn according to P , while the relative entropy $D_{kl}(P||Q)$ is the cost induced by assuming the source is Q . The function J defined in this section, also has coding theoretical motivation and arises in many places, e.g. an optimal test for the two samples problem described in Section 4.2. We give in this section the definition of the J function and the related definition of a mutual source. We also give a basic identity describing the mutual source in terms of the related sources. In Section 3.2 we give the definition of the empirical mutual source and the statistic J of two samples. In Section 3.3 we shortly discuss the generalization of the these quantities to multiple sources/samples. Section 3.4 gives some properties of the J function, among others, relating it to the basic quantities of information theory.

3.1 Definitions

Definition 3.1.1 (λ -Mutual Source [7]) Let S be a tree model, and $P_1, P_2 \in \mathbf{P}(S)$. The λ -Mutual Source of P_1 and P_2 with respect to the model S is defined by

$$P_\lambda^S = \arg \min_{P \in \mathbf{P}(S)} \lambda D_{kl}(P_1||P) + (1 - \lambda) D_{kl}(P_2||P),$$

where $0 \leq \lambda \leq 1$.

Definition 3.1.2 (J Function [7]) Let $0 \leq \lambda \leq 1$, and S be a tree model. Then the function $J_\lambda^S : \mathbf{P}(S) \times \mathbf{P}(S) \rightarrow \mathbb{R}^+$ is defined as follows, for any two sources $P_1, P_2 \in \mathbf{P}(S)$

$$\begin{aligned} J_\lambda^S(P_1, P_2) &= \min_{P \in \mathbf{P}(S)} \lambda D_{kl}(P_1||P) + (1 - \lambda) D_{kl}(P_2||P) \\ &= \lambda D_{kl}(P_1||P_\lambda^S) + (1 - \lambda) D_{kl}(P_2||P_\lambda^S), \end{aligned}$$

where P_λ^S is the λ -mutual source of P_1 and P_2 with respect to the model S .

To understand the coding theory intuition for the J function, consider the following random source of binary bits: at every step it flips a coin with probability λ for head. If a head was drawn then a bit is drawn from some source $P_1 \in \mathbf{P}(S)$, otherwise a bit is drawn from another source $P_2 \in \mathbf{P}(S)$. Then $J_\lambda^S(P_1, P_2)$ is the minimal cost (in bits) induced by assuming the bits came from a single source in $\mathbf{P}(S)$. The λ -mutual source of P_1 and P_2 with respect to the model S is a source that achieves this minimum. The next two lemmas gives an explicit expression of the mutual source.

Lemma 3.1.1 For any $P_1, P_2 \in \mathbf{P}(\emptyset)$ and $\alpha, \beta \in \mathbb{R}^+$, denote

$$P_{\alpha, \beta} = \arg \min_{P \in \mathbf{P}(\emptyset)} \alpha D_{kl}(P_1||P) + \beta D_{kl}(P_2||P), \quad (3.1)$$

then

$$\forall a \in \mathcal{X} \quad P_{\alpha, \beta}(a) = \frac{\alpha}{\alpha + \beta} P_1(a) + \left(1 - \frac{\alpha}{\alpha + \beta}\right) P_2(a). \quad (3.2)$$

Proof. For notational convenience denote $p_i = P_{\alpha,\beta}(a_i)$ where $\mathcal{X} = \{a_i\}$. We can express Eq. (3.1) as constrained minimization using Lagrange multipliers, as the minimization of

$$\mathcal{J} = \alpha D_{kl}(P_1||P) + \beta D_{kl}(P_2||P) + c \left(\sum p_i \right) \quad (3.3)$$

$$= \alpha \sum_i P_1(a_i) \log \frac{P_1(a_i)}{p_i} + \beta \sum_i P_2(a_i) \log \frac{P_2(a_i)}{p_i} + c \left(\sum p_i \right). \quad (3.4)$$

Differentiating with respect to p_i , we obtain

$$\frac{\partial \mathcal{J}}{\partial p_i} = -\alpha \frac{P_1(a)}{p_i} - \beta \frac{P_2(a)}{p_i} + c, \quad (3.5)$$

equating the derivative to 0, and solving for p_i we obtain,

$$p_i = \frac{\alpha P_1(a) + \beta P_2(a)}{c}, \quad (3.6)$$

substituting in the constraints to find c , we find $c = \alpha + \beta$ and hence

$$p_i = \frac{\alpha P_1(a) + \beta P_2(a)}{\alpha + \beta}. \quad (3.7)$$

■

Corollary 3.1.1 ([7]) *The λ -mutual source P_λ^\emptyset of two memoryless sources P_1 and P_2 satisfies*

$$\forall a \in \mathcal{X} \quad P_\lambda^\emptyset(a) = \lambda P_1(a) + (1 - \lambda) P_2(a). \quad (3.8)$$

The next Lemma was shown in [7]. We give here a simpler proof.

Lemma 3.1.2 *Let S be a tree model. The λ -mutual source P_λ^S of two sources $P_1, P_2 \in \mathbf{P}(S)$ satisfies, $\forall a \in \mathcal{X}, s \in S$*

$$P_\lambda^S(a|s) = \frac{\lambda P_1(s)}{\lambda P_1(s) + (1 - \lambda) P_2(s)} P_1(a|s) + \frac{(1 - \lambda) P_2(s)}{\lambda P_1(s) + (1 - \lambda) P_2(s)} P_2(a|s). \quad (3.9)$$

Proof. For all $P \in \mathbf{P}(S)$ and $\alpha, \beta \in \mathbb{R}^+$

$$\begin{aligned} \alpha D_{kl}(P_1||P) + \beta D_{kl}(P_2||P) &= \sum_{s \in S} \alpha P_1(s) \sum_{a \in \mathcal{X}} P_1(a|s) \log \frac{P_1(a|s)}{P(a|s)} + \sum_{s \in S} \beta P_2(s) \sum_{a \in \mathcal{X}} P_2(a|s) \log \frac{P_2(a|s)}{P(a|s)} \\ &= \sum_{s \in S} \left[\alpha P_1(s) \sum_{a \in \mathcal{X}} P_1(a|s) \log \frac{P_1(a|s)}{P(a|s)} + \beta P_2(s) \sum_{a \in \mathcal{X}} P_2(a|s) \log \frac{P_2(a|s)}{P(a|s)} \right] \end{aligned}$$

by Lemma 3.1.1 P_λ^S minimizes

$$\alpha P_1(s) \sum_{a \in \Sigma} P_1(a|s) \log \frac{P_1(a|s)}{P(a|s)} + \beta P_2(s) \sum_{a \in \Sigma} P_2(a|s) \log \frac{P_2(a|s)}{P(a|s)},$$

for all $s \in S$, which completes the proof. ■

Remark 3.1.1 Let S be a tree model and $P_1, P_2 \in \mathbf{P}(S)$. Denote by M_λ the joint probability distribution over $S \times \mathcal{X}$ defined by

$$M_\lambda(s, a) = \lambda P_1(s, a) + (1 - \lambda) P_2(s, a).$$

Then $\forall a \in \mathcal{X}, s \in S$

$$M_\lambda(a|s) = P_\lambda^S(a|s),$$

where P_λ^S is the λ -mutual source of P_1 and P_2 with respect to the model S .

3.2 The Empirical Mutual Source and the Statistic J

Let S be a tree model, and let x_1^n and y_1^m be two sequences of symbols in \mathcal{X} . Denote the sum of lengths of \mathbf{x} and \mathbf{y} by $l = n + m$. We define the **statistic** $J(\mathbf{x}, \mathbf{y})$ of the two sequences by, $J(\mathbf{x}, \mathbf{y}) = J_{\frac{n}{l}}^S(P_{\mathbf{x}}, P_{\mathbf{y}})$. The λ -mutual source of $P_{\mathbf{x}}$ and $P_{\mathbf{y}}$ for $\lambda = \frac{n}{l}$ is termed the **empirical mutual source** of the samples \mathbf{x} and \mathbf{y} . The following lemma states that the empirical mutual source of the two samples \mathbf{x} and \mathbf{y} is the type of the corpus consisting of the two samples (see Section 2.2.2).

Lemma 3.2.1 Let x_1^n and y_1^m be two sequences of symbols in \mathcal{X} . Denote the empirical mutual source of \mathbf{x} and \mathbf{y} by $M_{x,y}$, then $M_{\mathbf{x},\mathbf{y}} = P_{\mathbf{x},\mathbf{y}}$.

Proof. By Lemma 3.1.2 we know that for all $a \in \mathcal{X}$ and $s \in S$,

$$\begin{aligned} M_{\mathbf{x},\mathbf{y}}(a|s) &= \arg \min_{M \in \mathbf{P}(S)} |\mathbf{x}| D_{kl}(P_{\mathbf{x}}||M) + |\mathbf{y}| D_{kl}(P_{\mathbf{y}}||M) \\ &= \frac{|\mathbf{x}|P_{\mathbf{x}}(s)}{|\mathbf{x}|P_{\mathbf{x}}(s) + |\mathbf{y}|P_{\mathbf{y}}(s)} P_{\mathbf{x}}(a|s) + \frac{|\mathbf{y}|P_{\mathbf{y}}(s)}{|\mathbf{x}|P_{\mathbf{x}}(s) + |\mathbf{y}|P_{\mathbf{y}}(s)} P_{\mathbf{y}}(a|s), \end{aligned}$$

thus,

$$\begin{aligned} M_{\mathbf{x},\mathbf{y}}(a|s) &= \frac{|\mathbf{x}|P_{\mathbf{x}}(s)}{|\mathbf{x}|P_{\mathbf{x}}(s) + |\mathbf{y}|P_{\mathbf{y}}(s)} P_{\mathbf{x}}(a|s) + \frac{|\mathbf{y}|P_{\mathbf{y}}(s)}{|\mathbf{x}|P_{\mathbf{x}}(s) + |\mathbf{y}|P_{\mathbf{y}}(s)} P_{\mathbf{y}}(a|s) \\ &= \frac{|\mathbf{x}| \frac{N_{\mathbf{x}}(s)}{|\mathbf{x}|}}{|\mathbf{x}| \frac{N_{\mathbf{x}}(s)}{|\mathbf{x}|} + |\mathbf{y}| \frac{N_{\mathbf{y}}(s)}{|\mathbf{y}|}} \frac{N_{\mathbf{x}}(a|s)}{N_{\mathbf{x}}(s)} + \frac{|\mathbf{y}| \frac{N_{\mathbf{y}}(s)}{|\mathbf{y}|}}{|\mathbf{x}| \frac{N_{\mathbf{x}}(s)}{|\mathbf{x}|} + |\mathbf{y}| \frac{N_{\mathbf{y}}(s)}{|\mathbf{y}|}} \frac{N_{\mathbf{y}}(a|s)}{N_{\mathbf{y}}(s)} \\ &= \frac{N_{\mathbf{x}}(a|s)}{N_{\mathbf{x}}(s) + N_{\mathbf{y}}(s)} + \frac{N_{\mathbf{y}}(a|s)}{N_{\mathbf{x}}(s) + N_{\mathbf{y}}(s)} \\ &= \frac{N_{\mathbf{x}}(a|s) + N_{\mathbf{y}}(a|s)}{N_{\mathbf{x}}(s) + N_{\mathbf{y}}(s)} \\ &= P_{\mathbf{x},\mathbf{y}}(a|s). \end{aligned}$$

■

We conclude that

$$J(\mathbf{x}, \mathbf{y}) = J_{\frac{n}{l}}(P_{\mathbf{x}}, P_{\mathbf{y}}) = \frac{n}{l} D_{kl}(P_{\mathbf{x}}||P_{\mathbf{x},\mathbf{y}}) + (1 - \frac{n}{l}) D_{kl}(P_{\mathbf{y}}||P_{\mathbf{x},\mathbf{y}}).$$

3.3 The J Function for Multiple Sources

The J function can be defined for any number of sources in the same way as it was defined for two sources. Most of the properties of the J function for two sources are easily generalized for multiple sources. For simplicity we will describe most of the results in this thesis only for two sources, but in some cases the general definition is needed, hence we give it here.

Definition 3.3.1 Let $P_1, P_2, \dots, P_t \in \mathbf{P}(S)$ and $0 \leq \lambda_1, \lambda_2, \dots, \lambda_t \leq 1$, $\sum_{i=1}^t \lambda_i = 1$, then

$$J_{\{\lambda_i\}_{i=1}^t}^S(\{P_i\}_{i=1}^t) \triangleq \min_{P \in \mathbf{P}(S)} \left\{ \sum_{i=1}^t \lambda_i D_{kl}(P_i||P) \right\} \quad (3.10)$$

We call the source that achieves the minimum in Eq. (3.10) the **mutual source** of $\{P_i\}$ with respect to the prior $\{\lambda_i\}$, and the model S .

3.4 Properties of J

We state in this section some properties of the J function defined in the sections above. For additional properties see [18, 7].

Lemma 3.4.1 ([7]) *Let $P_1, P_2 \in \mathbf{P}(S)$ and $\lambda \in [0, 1]$, then*

$$J_\lambda^S(P_1, P_2) = H(P_\lambda^S) - (\lambda H(P_1) + (1 - \lambda)H(P_2)).$$

Corollary 3.4.1 *Let $P_1, P_2 \in \mathbf{P}(\emptyset)$ and $\lambda \in [0, 1]$, then*

$$J_\lambda(P_1, P_2) = JS_\lambda(P_1, P_2).$$

Where $JS_\lambda(P_1, P_2)$ is the Jensen-Shannon divergence Measure defined in [18].

Lemma 3.4.2 *Let $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{y} \in \mathcal{X}^m$, $l = n + m$ and denote $\lambda = \frac{n}{l}$, then*

$$J_\lambda^S(\mathbf{x}, \mathbf{y}) = \frac{1}{l} \log \frac{\sup_{P_1, P_2 \in \mathbf{P}(S)} P_1(\mathbf{x})P_2(\mathbf{y})}{\sup_{P \in \mathbf{P}(S)} P(\mathbf{x})P(\mathbf{y})}$$

Proof. Straightforward using Lemma 3.2.1, Lemma 3.4.1, Eq. (2.23) and Lemma 2.2.2. ■

Lemma 3.4.3 *Let $P \in \mathbf{P}(S)$, then for any $s \in S$ the conditional distribution $P(X|s)$ is the mutual source of the distributions $\{P(X|sa)\}_{a \in \mathcal{X}}$ with respect to the parameters $\{P(a|s)\}_{a \in \mathcal{X}}$, and the empty model.*

Proof. For any $s \in S$ and $b \in \mathcal{X}$

$$P(b|s) = \sum_{a \in \mathcal{X}} P(a, b|s) = \sum_{a \in \mathcal{X}} P(a|s)P(b|s, a). \quad (3.11)$$

The lemma then follows using the generalization of Lemma 3.1.1 to arbitrary number of sources ■

Example 3.4.1 Let X_1 and X_2 be two random variables over the same alphabet \mathcal{X} , with joint probability distribution $P(X_1 = a, X_2 = b)$ such that the two marginals are equal. Then

$$P(x) = \sum_y P(x, y) = \sum_y P(y)P(x|y) = \sum_x P(x)P(x|y), \quad (3.12)$$

where $P(x)$ means $P(X_1 = a)$. Note that if we think of a Markov process of order one, then it is defined by the joint distributions $P(X_i, X_{i+1})$, and the stationarity condition implies that the marginals are equal.

Definition 3.4.1 Let X_1 and X_2 be two random variables with joint probability $P(X_1, X_2)$, the **mutual information** $I(X_1, X_2)$ of X_1 and X_2 is defined by

$$I(X_1, X_2) \triangleq \sum_{a \in \mathcal{X}_1} \sum_{b \in \mathcal{X}_2} P(a, b) \log \frac{P(a, b)}{P(a)P(b)},$$

where $P(X_1)$ and $P(X_2)$ are the marginals of P .

Lemma 3.4.4 Let X_1, X_2 be two random variables distributed by $P(X_1, X_2)$, such that the marginals $P(X_1)$ and $P(X_2)$ are equal, then

$$I(X_1, X_2) = J_{\{P(a_i)\}_{a_i \in \mathcal{X}}}^{\emptyset}(\{P(X_2|a_i)\}_{a_i \in \mathcal{X}}).$$

Proof.

$$I(X_1, X_2) = \sum_{a,b} P(a, b) \log \frac{P(a, b)}{P(a)P(b)} \quad (3.13)$$

$$= \sum_{a,b} P(a, b) \log \frac{P(b|a)}{P(b)} \quad (3.14)$$

$$= - \sum_{a,b} P(a, b) \log P(b) + \sum_{a,b} P(a, b) \log P(b|a) \quad (3.15)$$

$$= - \sum_a P(a) \log P(a) - \left(- \sum_a P(a) \sum_b P(b|a) \log P(b|a) \right) \quad (3.16)$$

$$= H(P(X_1)) - \left(\sum_a P(a) H(P(X_2|a)) \right) \quad (3.17)$$

$$= J_{\{P(a_i)\}_{a_i \in \mathcal{X}}}^{\emptyset}(\{P(X_2|a_i)\}_{a_i \in \mathcal{X}}), \quad (3.18)$$

where Eq. (3.18) is due to the generalization of Lemma 3.4.3 to arbitrary number of sources.

■

Lemma 3.4.5 *Let $P_1, P_2, P_3 \in \mathbf{P}(S)$ and $\lambda, \eta \in [0, 1]$, then*

$$J_{\{\lambda\eta, \lambda(1-\eta), 1-\lambda\}}^S(P_1, P_2, P_3) = \lambda J_\eta^S(P_1, P_2) + J_\lambda^S(P_\eta^S, P_3),$$

where P_η^S is the η -mutual source of P_1 and P_2 .

Proof.

$$J_{\{\lambda\eta, \lambda(1-\eta), 1-\lambda\}}^S(P_1, P_2, P_3) \tag{3.19}$$

$$= H(\lambda\eta P_1 + \lambda(1-\eta)P_2 + (1-\lambda)P_3) - \lambda\eta H(P_1) - \lambda(1-\eta)H(P_2) \tag{3.20}$$

$$- (1-\lambda)H(P_3) \tag{3.21}$$

$$= H(\lambda[\eta P_1 + (1-\eta)P_2] + (1-\lambda)P_3) - \lambda[\eta H(P_1) + (1-\eta)H(P_2)] \tag{3.22}$$

$$- (1-\lambda)H(P_3) \tag{3.23}$$

$$= H(\lambda P_\eta^S + (1-\lambda)P_3) - \lambda H(P_\eta^S) - (1-\lambda)H(P_3) \tag{3.24}$$

$$+ \lambda[H(P_\eta^S) - \eta H(P_1) - (1-\eta)H(P_2)] \tag{3.25}$$

$$= \lambda J_\eta^S(P_1, P_2) + J_\lambda^S(P_\eta^S, P_3). \tag{3.26}$$

■

Lemma 3.4.6 *Let P_1, P_2 be two distributions over \mathcal{X} , then $J_\lambda(P_1, P_2) = 0$ if and only if $\lambda = 0$ or $\lambda = 1$ or $P_1 = P_2$.*

Proof. The lemma directly follows from the “information inequality” Theorem (see [5, p.26]), stating that $D_{kl}(P_1||P_2) \geq 0$ with equality if and only if $P_1(a) = P_2(a)$ for all $a \in \mathcal{X}$. ■

Lemma 3.4.7 ([7]) *$J(P, Q)$ is convex in the pair (P, Q) .*

Proof. Follows directly from Lemma 2.3.4. ■

The following lemmas give bounds on the J function.

Lemma 3.4.8 ([7]) Let $\mathcal{X} = \{a_i\}_{i=1}^k$. For any two distributions $P, Q \in \mathbf{P}(\emptyset)$ over \mathcal{X} ,

$$0 \leq J_\lambda(P, Q) \leq h(\lambda),$$

where $h(\lambda)$ is the binary entropy defined by

$$h(\lambda) = -\lambda \log \lambda - (1 - \lambda) \log(1 - \lambda).$$

Proof. The first inequality follows directly from the “information inequality” theorem (see above). To show the second inequality we use the convention $0 \log 0 = 0$,

$$\begin{aligned} D_{kl}(P \parallel \lambda P + (1 - \lambda) Q) &= \sum_{i=1}^k P(a_i) \log \frac{P(a_i)}{\lambda P(a_i) + (1 - \lambda) Q(a_i)} \\ &\leq \sum_{i: P(a_i) \neq 0} P(a_i) \log \frac{P(a_i)}{\lambda P(a_i)} \\ &= -\log \lambda. \end{aligned}$$

Similarly $D_{kl}(Q \parallel \lambda P + (1 - \lambda) Q) \leq -\log(1 - \lambda)$. hence

$$\begin{aligned} J_\lambda(P, Q) &= \lambda D_{kl}(P \parallel \lambda P + (1 - \lambda) Q) + (1 - \lambda) D_{kl}(Q \parallel \lambda P + (1 - \lambda) Q) \\ &\leq -\lambda \log \lambda - (1 - \lambda) \log(1 - \lambda) \\ &= h(\lambda). \end{aligned}$$

Note that equality holds if and only if $P(a_i) > 0$ implies $Q(a_i) = 0$ and $Q(a_i) > 0$ implies $P(a_i) = 0$. ■

Lemma 3.4.9 ([7, 18]) Let $\mathcal{X} = \{a_i\}_{i=1}^k$. For any two distributions $P, Q \in \mathbf{P}(\emptyset)$ over \mathcal{X} , define $p = (P(a_1), \dots, P(a_k))$ and $q = (Q(a_1), \dots, Q(a_k))$, then

$$J_\lambda(P, Q) \leq \frac{1}{2} h(\lambda) \cdot \|p - q\|_1 \quad \lambda \in [0, 1]$$

Proof. Define v to be the minimum components of p and q , i.e. $v_i = \min(p_i, q_i)$ and define v^p and v^q to be the residual vector between p and q respectively, i.e.

$$\begin{aligned} v^p &= p_i - v_i \\ v^q &= q_i - v_i. \end{aligned}$$

Since $\sum_i v_i^p = \sum_i v_i^q = \frac{1}{2} \|p - q\|_1$ (see [5, p. 299]) then normalizing the vectors v^p , v^q and v results in

$$\begin{aligned} \widehat{v}^p &= \frac{v^p}{\frac{1}{2} \|p - q\|_1} \\ \widehat{v}^q &= \frac{v^q}{\frac{1}{2} \|p - q\|_1} \\ \widehat{v} &= \frac{v}{1 - \frac{1}{2} \|p - q\|_1}. \end{aligned}$$

It then follows that the pair (p, q) may be expressed as the following convex combination

$$(p, q) = \left(1 - \frac{1}{2} \|p - q\|_1\right) (\widehat{v}, \widehat{v}) + \frac{1}{2} \|p - q\|_1 (\widehat{v}^p, \widehat{v}^q),$$

hence, by the convexity of J_λ we may write

$$J_\lambda(p, q) \leq \left(1 - \frac{1}{2} \|p - q\|_1\right) J_\lambda(\widehat{v}, \widehat{v}) + \frac{1}{2} \|p - q\|_1 J_\lambda(\widehat{v}^p, \widehat{v}^q) \quad (3.27)$$

$$\leq \frac{1}{2} \|p - q\|_1 J_\lambda(\widehat{v}^p, \widehat{v}^q) \quad (3.28)$$

$$\leq \frac{1}{2} \|p - q\|_1 h(\lambda), \quad (3.29)$$

where Eq. (3.27) is by the convexity of J_λ , Eq. (3.28) is by Lemma 3.4.6 and Eq. (3.29) is by Lemma 3.4.8 ■

Chapter 4

The J Function and the Two Samples Problem

4.1 Hypotheses Testing

4.1.1 Definitions

Consider a parameter θ whose value is unknown but must lie in a given parameter space Θ . Let Θ_1, Θ_2 be a partition of Θ . Let H_0 denote the hypothesis that $\theta \in \Theta_1$ and H_1 denote the hypothesis that $\theta \in \Theta_2$. A problem of deciding which of H_0 or H_1 is true is called a ***hypotheses testing problem***. The decision is usually made after seeing a random sample x_1, \dots, x_n from a distribution that is dependent on the unknown parameter θ . A procedure for deciding whether to accept H_0 or H_1 is called a ***test***. A test can be specified by partitioning the sample space \mathcal{X}^n into two subsets Λ_1 and Λ_2 . When $x_1^n \in \Lambda_1$ the hypothesis H_0 is accepted, otherwise it is rejected and the hypothesis H_1 is accepted. Traditionally the two hypotheses are not treated equally. The hypotheses H_0 is called ***the null hypothesis*** and it is given higher “priority” than H_1 which is called ***the alternative hypothesis***. Thus the subset Λ_2 of samples that lead to the rejection of the null hypothesis is also treated differently and is called the ***critical region*** of the test. In the sequel we will sometimes identify a test with its critical region, i.e. instead of writing “let $\delta = (\Lambda_1, \Lambda_2)$ be a test ...”, we will write “let Λ be a test ...”, where Λ is the critical region (equivalent to Λ_2 in the above formulation). Since $\Lambda_1 = \mathcal{X}^n / \Lambda_2$, Λ_2

completely characterizes the test.

4.1.2 Testing Simple Hypotheses

If Θ_i ($i = 1, 2$) contain only a single value then H_i is called a *simple hypothesis*, otherwise it is called a *composite hypothesis*. In the case where both hypotheses are simple, we wish to build a test for deciding between the following two hypotheses:

$$\begin{aligned} H_0 & : \theta = \theta_0 \\ H_1 & : \theta = \theta_1. \end{aligned} \tag{4.1}$$

Given a test $\delta = (\Lambda_1, \Lambda_2)$ for H_0 and H_1 , there are two kinds of possible errors. The first is rejecting H_0 while $\theta = \theta_0$, and is termed *error of the first kind*. The second is accepting H_0 while $\theta = \theta_1$, and it is termed *error of the second kind*. Consider an indexed family of joint probability functions (or probability density functions) $\{f_\theta(x_1^n)\}_{\theta \in \Theta}$ (In this case $\Theta = \{\theta_0, \theta_1\}$). The probability of the two kinds of errors are $f_{\theta_0}(\Lambda_2)$ and $f_{\theta_1}(\Lambda_1)$ respectively. These probabilities are usually denoted by $\alpha(\delta)$ and $\beta(\delta)$.

We wish to use tests that are optimal under a certain optimality criterion. One possible criterion is to say that a test δ^* is optimal if for every other test δ , $\alpha(\delta^*) \leq \alpha(\delta)$ and $\beta(\delta^*) \leq \beta(\delta)$. Except for trivial cases there are no such optimal tests. Another approach is to minimize a linear combination of the two kinds of error. That is, for every pair of positive numbers a and b we wish to find a test $\delta_{(a,b)}^*$ that satisfies, for every other test δ

$$a\alpha(\delta_{(a,b)}^*) + b\beta(\delta_{(a,b)}^*) \leq a\alpha(\delta) + b\beta(\delta). \tag{4.2}$$

The following critical region defines a test that satisfies this criterion (see [14, p.443]):

$$\Lambda_{(a,b,n)}^* = \{x_1^n \in \mathcal{X}^n : bf_{\theta_1}(x_1^n) \geq af_{\theta_0}(x_1^n)\}. \tag{4.3}$$

This criterion allows one to specify the ratio of importance of the two hypotheses. Another well known criterion is the one defined by Neyman and Pearson : Among all tests δ satisfying $\alpha(\delta) \leq \alpha_0$ for a fixed α_0 , select one that minimizes the probability of error of the second kind.

This criterion only considers tests with error of the first kind bounded by a constant α_0 . An upper bound α_0 specified in this way is called the **level of significance** of the test. This approach emphasizes the importance of H_0 over H_1 . The possibility that an error of the first kind exceeds a given constant is not considered. A well known example is the one of a test made to decide whether a patient has a certain fatal disease, the null hypothesis is that he has the disease and the alternative is that he does not. Then the error of the first kind (false alarm) is considered less important than the second type of error (not discovering the disease).

It was shown by Neyman and Pearson that the **likelihood ratio test (LRT)** is optimal under this criterion (For every choice of α_0). The critical region of the LRT is defined by

$$\Lambda_n^{LRT} = \{x_1^n \in \mathcal{X}^n : \frac{f_{\theta_1}(x_1^n)}{f_{\theta_0}(x_1^n)} \geq c_n\}, \quad (4.4)$$

where c_n is defined so $f_{\theta_0}(\Lambda_n^{LRT}) = \alpha_0$. The ratio $\frac{f_{\theta_1}(x_1^n)}{f_{\theta_0}(x_1^n)}$ is called the **likelihood ratio** of the sample x_1^n , hence the name of the test. Next we give an example for constructing an optimal test for deciding between two normal distributions.

Example 4.1.1 ([14]) Let $\Theta = \{a, b\}$, $a, b \in \mathbb{R}$ and $f_\theta(x_i) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp(-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2)$. Given a sample x_1, \dots, x_n we wish to decide whether it was drawn from a normal distribution with mean a and variance 1, or from a normal distribution with mean b and variance 1, i.e. we wish to test between the following hypotheses:

$$\begin{aligned} H_0 & : \theta = a \\ H_1 & : \theta = b \end{aligned} \quad (4.5)$$

The likelihood ratio for this case can be written in the form

$$\frac{f_b(\mathbf{x})}{f_a(\mathbf{x})} = \exp(n(\bar{x}_n - \frac{1}{2})) \quad (4.6)$$

Thus the LRT can be equivalently written as follows

$$\Lambda_n^{LRT} = \{x_1^n \in \mathcal{X}^n : \bar{x}_n \geq \bar{c}_n\} \quad (4.7)$$

Where \bar{c}_n is chosen to achieve $f_a(\Lambda_n^{LRT}) = \alpha_0$. This test is optimal in the Neyman-Pearson sense.

4.1.3 Testing Composite Hypotheses

Suppose next that H_1 is a composite hypothesis, i.e. Θ_2 has more than a single element. That is, we wish to find a test for deciding between the hypotheses,

$$\begin{aligned} H_0 & : \theta \in \Theta_1 \\ H_1 & : \theta \in \Theta_2. \end{aligned} \tag{4.8}$$

H_0 might be simple or composite. The Neyman-Pearson criterion and the likelihood ratio test are no longer applicable, since the second distribution is unknown. It is only known to belong to a certain family of distributions. In this section we will consider extensions of the Neyman-Pearson criterion to the case of composite hypotheses. When H_1 is composite then we can no longer consider the error of the first kind as before, instead we consider an upper bound over all the possible errors of the first kind (related to the different parameters in Θ_1), more formally: the *size* of a test Λ is defined as $\sup_{\theta \in \Theta_1} \{f_\theta(\Lambda)\}$. Then one straightforward extension of the Neyman-Pearson criterion is to choose among all tests with size bounded by a given level of significance, a test that minimizes the second kind of error uniformly for all $\theta \in \Theta_2$. More formally : The tests considered are the ones with critical regions Λ that satisfies

$$\sup_{\theta \in \Theta_1} \{f_\theta(\Lambda)\} \leq \alpha_0. \tag{4.9}$$

Among all tests satisfying (4.9), select a test Λ^* that also satisfies: for every other test Λ satisfying (4.9) and every $\theta \in \Theta_2$

$$f_\theta(\Lambda) \leq f_\theta(\Lambda^*).$$

A test satisfying this criterion is called a ***uniformly most powerful test (UMP)***. Next we give an example of such a test.

Example 4.1.2 ([16]) Let $\Theta = [a, b] \subset \mathbb{R}$ and $f_\theta(x_i) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp(-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2)$. Continuing Example 4.1.1, given a sample x_1, \dots, x_n we wish to decide whether it came from a normal

distribution with variance 1 and mean μ , or with variance 1 and mean greater than μ . i.e. we wish to test between the following hypotheses:

$$H_0 : \theta = \mu \tag{4.10}$$

$$H_1 : \theta > \mu.$$

Consider the LRT defined in (4.7), it does not depend on the parameter b of the alternative hypotheses. That means it is optimal in the Neyman-Pearson sense for all such b . It follows then that it is a uniformly most powerful test.

As Example 4.1.2 shows, the requirements of the criterion given above are a bit strong; when H_0 in (4.8) is simple, then a UMPT is required for all $\theta \in \Theta_2$ to be optimal for that θ and yet independent of it. Another approach asks only for asymptotic optimality [20, 19, 36, 13, 35]. The search for a test Λ is replaced by a search for a sequence of tests $\{\Lambda_n\}_{n \geq 1}$. The demand for error bounded by a constant is replaced by a demand that the size dominating the error exponent as n grows is bounded by a constant. More formally, the Neyman-Pearson asymptotic optimality criterion is defined as:

Definition 4.1.1 (*Asymptotic Optimality Criterion*) Among all sequences of tests $\{\Lambda_n\}_{n \geq 1}$ that do not depend on the unknown parameters and satisfy

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \sup_{\theta \in \Theta_1} f_{\theta}(\Lambda_n) < -\lambda, \tag{4.11}$$

for a given $\lambda > 0$, select a sequence that maximizes the second kind of error exponent, $-\limsup_{n \rightarrow \infty} \frac{1}{n} \log f_{\theta}(\bar{\Lambda}_n)$, uniformly for all $\theta \in \Theta_2$.

The Asymptotic Optimality (AO) criterion is similar to the UMP in that it requests for a uniform behavior for all parameters in Θ_2 . It relaxes though the demand for an optimal behavior for all finite n , and in exchange asks for asymptotic optimality. Next we give an example of a test that is optimal under the OA criterion.

Example 4.1.3 In our discussions up to this point we did not restrict the nature of the family of distributions that our sample is drawn from (excluding examples). In this example and also

later in this section we restrict the discussion to the family of stationary and ergodic Markov sources with a specified tree model, over some finite alphabet \mathcal{X} . Let S denote such a tree model. Assume we see a sample x_1, \dots, x_n from a distribution in $\mathbf{P}(S)$, and we wish to decide whether \mathbf{x} was drawn from a distribution $P_1 \in \mathbf{P}(S)$ or from a distribution in a set $\mathbf{P}_2 \subset \mathbf{P}(S)$ not containing P_1 . If we denote by P the distribution that \mathbf{x} was drawn from, then we wish to test between the following hypotheses:

$$\begin{aligned} H_0 & : P = P_1 \\ H_1 & : P \in \mathbf{P}_2. \end{aligned} \tag{4.12}$$

To solve this problem, define a sequence of critical regions $\{\Lambda_n^*\}$ as follows:

$$\Lambda_n^* = \{\mathbf{x} \in \mathcal{X}^n \mid D_{kl}(P_{\mathbf{x}} \parallel P_1) > \lambda\}, \tag{4.13}$$

then this sequence of tests is asymptotically optimal by the AO criterion. This test was originally suggested by Hoeffding [15]. A straightforward extension is presented in [35]. The analysis is based on results from large deviation theory, e.g. Theorem 2.4.1. The methods used in [35] are very similar to the methods used in the proof of Theorem 4.2.1.

Example 4.1.4 Continuing Example 4.1.3, another well known test for the hypotheses in (4.12) is the **generalized likelihood ratio test (GLRT)** defined as follows

$$\Lambda_n^{GLRT} = \{\mathbf{x} \in \mathcal{X}^n \mid \frac{1}{n} \log \sup_{P_2 \in \mathbf{P}_2} P_2(\mathbf{x}) - \frac{1}{n} \log P_1(\mathbf{x}) > \lambda\} \tag{4.14}$$

In [35], conditions for asymptotic optimality of the GLRT are studied and discussed.

4.2 The Two Samples Problem

4.2.1 Introduction

The main reason we were able to construct a test for the hypotheses in (4.1) and (4.8), is that a sample of length n “separates” the two hypotheses. For example in (4.12), every distribution in $\mathbf{P}(S)$ has an ϵ -typical (Definition 2.4.3) set of sequences. And for small enough ϵ those sets are

disjoint for every two distinct distributions (The size of ϵ depends on the two distributions). This is why it was crucial in (4.12) that \mathbf{P}_2 will not contain P_1 . In this section we are interested in a case where we don't have such knowledge, i.e. we cannot specify the hypotheses in the simple form (4.1), nor in the composite form (4.8). As in previous cases denote by Θ a parameter space. Ω will denote a sample space (previously we assumed that $\Omega = \mathcal{X}^n$). Denote by $\{f_\theta(\mathbf{x})\}_{\theta \in \Theta}$ an indexed family of distributions over Ω . We observe a sample $\mathbf{x} \in \Omega$ drawn from a distribution f_θ and we wish to decide whether it was drawn from f_{θ_1} or f_{θ_2} for fixed unknown θ_1 and θ_2 . Instead of assuming an ability to associate θ_1 and θ_2 both with two distinct subsets of Θ , we assume that we can sample from f_{θ_i} and that this sampling is independent from the observed sample \mathbf{x} . Denote by \mathbf{y} a sample drawn from f_{θ_1} , then the problem of deciding whether \mathbf{x} was drawn from f_{θ_1} or not reduces to the problem of deciding whether \mathbf{x} and \mathbf{y} are realizations of the same source or not. This problem is known as the “general two samples problem”, defined more formally next.

Definition 4.2.1 (The General Two Samples Problem) *Let \mathbf{x} and \mathbf{y} be two samples independently drawn according to two unknown distributions in a predefined family. The **general two samples problem** is the testing of the following two hypotheses,*

$$\begin{aligned} H_0 & : \text{The samples are drawn from the same distribution} & (4.15) \\ H_1 & : \text{The samples are drawn from two different distributions} \end{aligned}$$

In the rest of this section we address the problem of finding an optimal test for these hypotheses, when the family of distributions is the class of stationary and ergodic Markov sources with a specified tree model. In parallel to Example 4.1.3 let S denote a tree model, and let $\mathbf{P}(S)$ denote all stationary and ergodic sources with model S . Note that the sample space Ω for this case is not \mathcal{X}^n but rather $\mathcal{X}^n \times \mathcal{X}^m$ where n is the length of \mathbf{x} and m is the length of \mathbf{y} . The specification of a test changes accordingly.

In [13] it was shown that the GLRT is an asymptotically optimal test for this problem, when $\mathbf{P}(S)$ was assumed to be the set of all Markov sources with order bounded by a fixed integer D , i.e. S represents the full tree of depth D . We will give a proof in Theorem 4.2.1 similar to the one in [13], but for general finite order tree sources. In addition we prove that

a modified version of the GLRT based on the CTW compression algorithm (see Section 2.5) is also asymptotically optimal. In Section 5.4 we address the question of finding an optimal test for the two samples problem without assuming a bound on the order.

4.2.2 Optimal Tests for the Two Samples Problem

A test for the two samples problem is a partition of the Cartesian product $\Omega = \mathcal{X}^m \times \mathcal{X}^n$, of all pairs of m -tuples and n -tuples of elements in \mathcal{X} . We will denote by l the sum of lengths of the two samples, i.e. $l = n + m$. The GLRT for this case takes the following form

$$\Lambda_l^{GLRT} = \left\{ \mathbf{x}, \mathbf{y} \in \Omega : \frac{1}{l} \log \frac{\sup_{P_1, P_2 \in \mathbf{P}(S)} P_1(\mathbf{x})P_2(\mathbf{y})}{\sup_{P \in \mathbf{P}(S)} P(\mathbf{x})P(\mathbf{y})} > \lambda \right\}.$$

By Lemma 3.2.1 and the statistic tested by the GLRT, for the model S is $J^S(\mathbf{x}, \mathbf{y})$, i.e.

$$\Lambda_l^{GLRT} = \{ \mathbf{x}, \mathbf{y} \in \Omega : J^S(\mathbf{x}, \mathbf{y}) > \lambda \}.$$

Theorem 4.2.1 (Gutman89 [13]) *If $n = \Theta(m)$, then the GLRT is asymptotically optimal in the Neyman-Pearson sense (see Definition 4.1.1).*

Proof. First we will show that

$$\limsup_{l \rightarrow \infty} \frac{1}{l} \log \sup_{P \in \mathbf{P}(S)} P(\Lambda_l^{GLRT}) \leq -\lambda. \quad (4.16)$$

For convenience of notations, in the rest of this proof we will denote Λ_l^{GLRT} by Λ^* and $\mathbf{P}(S)$ by \mathbf{P} . Assuming that the types of the samples are sufficient statistics (this is not essential, see [13, lemma 1]) then

$$\log \sup_{P \in \mathbf{P}} P(\Lambda^*) = \log \sup_{P \in \mathbf{P}} \sum_{(\mathbf{x}, \mathbf{y}) \in \Lambda^*} P(\mathbf{x})P(\mathbf{y}) \quad (4.17)$$

$$= \log \sup_{P \in \mathbf{P}} \sum_{T_{\mathbf{x}}, T_{\mathbf{y}} \subset \Lambda^*} P(T_{\mathbf{x}})P(T_{\mathbf{y}}) \quad (4.18)$$

$$\leq \log \sup_{P \in \mathbf{P}} \sum 2^{-nD(P_{\mathbf{x}}||P) - mD(P_{\mathbf{y}}||P)} \quad (4.19)$$

$$\leq \log \sup_{P \in \mathbf{P}} |\mathcal{P}_n^S| \sup_{(\mathbf{x}, \mathbf{y}) \subset \Lambda^*} 2^{-nD(P_{\mathbf{x}}||P) - mD(P_{\mathbf{y}}||P)} \quad (4.20)$$

$$\leq \sup_{(\mathbf{x}, \mathbf{y}) \subset \Lambda^*} \left\{ - \inf_{P \in \mathbf{P}} \{nD(P_{\mathbf{x}}||P) + mD(P_{\mathbf{y}}||P)\} + O(\log n + \log m) \right\} \quad (4.21)$$

$$= \sup_{(\mathbf{x}, \mathbf{y}) \subset \Lambda^*} \{-J^S(\mathbf{x}, \mathbf{y}) \cdot l\} + O(\log n + \log m) \quad (4.22)$$

$$< -\lambda \cdot l + O(\log n + \log m), \quad (4.23)$$

where Eq. (4.19) is due to Eq. (2.37), Eq. (4.20) is by the definition of $|\mathcal{P}_n^S|$, Eq. (4.21) is due to Eq. (2.34), Eq. (4.22) is by definition of $J^S(\mathbf{x}, \mathbf{y})$ and Eq. (4.23) is by the definition of the GLRT. Thus Eq. (4.16) holds. For the second condition of asymptotic optimality, denote by Λ an arbitrary test that satisfies (4.16), then for small enough ϵ and large enough l

$$\frac{1}{l} \log \sup_{P \in \mathbf{P}} P(\Lambda) < -\lambda - \epsilon,$$

and for every $\mathbf{x}, \mathbf{y} \in \Lambda$

$$\frac{1}{l} \log \sup_{P \in \mathbf{P}} P(\Lambda) \geq \frac{1}{l} \log \sup_{P \in \mathbf{P}} P(T_{\mathbf{x}}, T_{\mathbf{y}}) \quad (4.24)$$

$$\geq - \inf_{P \in \mathbf{P}} \left(\frac{n}{l} D(P_{\mathbf{x}}||P) + \frac{m}{l} D(P_{\mathbf{y}}||P) \right) - o(l) \quad (4.25)$$

$$= -J^S(\mathbf{x}, \mathbf{y}) - o(l), \quad (4.26)$$

where Eq. (4.25) is due to Eq. (2.37). Thus for every $\mathbf{x}, \mathbf{y} \in \Lambda$

$$J^S(\mathbf{x}, \mathbf{y}) + o(l) > \lambda + \epsilon,$$

and for sufficiently large l , $J^S(\mathbf{x}, \mathbf{y}) > \lambda$. It follows that $\Lambda_l \subset \Lambda_l^*$ for all sufficiently large l , which concludes the proof. ■

Corollary 4.2.1 *Let $\tilde{J}^S(\mathbf{x}, \mathbf{y})$ be a function from Ω to \mathbb{R} with the following property: for all n, m and $(\mathbf{x}, \mathbf{y}) \in \Omega$*

$$\left| J^S(\mathbf{x}, \mathbf{y}) - \tilde{J}^S(\mathbf{x}, \mathbf{y}) \right| \leq o(n + m).$$

Then the test

$$\tilde{\Lambda}_l^{GLRT} = \{\mathbf{x}, \mathbf{y} : \tilde{J}^S(\mathbf{x}, \mathbf{y}) > \lambda\},$$

based on \tilde{J} is asymptotically optimal in the Neyman-Pearson sense.

Proof. Similar to the proof of Theorem 4.2.1. ■

Corollary 4.2.2 *The test*

$$\Lambda_l^e = \{\mathbf{x}, \mathbf{y} : \frac{1}{l} \log \frac{P_e^S(\mathbf{x})P_e^S(\mathbf{y})}{P_e^S(\mathbf{x}, \mathbf{y})} > \lambda\},$$

based on probability estimation by using the KT-estimator is asymptotically optimal in the Neyman-Pearson sense.

Proof. Follows from Corollary 4.2.1 and Lemma 2.5.3 ■

Remark 4.2.1 *Note the large gap left by the asymptotic optimality criterion. It does not distinguish between a test based on estimating the probability via empirical distribution and a test based on estimating the probability via a Bayesian estimate which generalizes much better. In Section 6.1.1 we give an example where a test based on the Bayesian estimation is superior to the test based on the empirical distribution.*

Chapter 5

Monotone Decrease of the Function

J

5.1 Introduction

Assume two samples x_1, \dots, x_n and y_1, \dots, y_m drawn from two unknown information sources. We wish to answer the question "have the two samples emerged from the same source?". We have shown an asymptotically optimal test for this question under the constraints that all sources are members of the family of stationary and ergodic Markov sources with a given model S . In order to analyze the behavior of a test for the case where S is unknown, we analyze in this section the behavior of the statistic $J^S(\mathbf{x}, \mathbf{y})$ for different models. In particular we show that $J^S(\mathbf{x}, \mathbf{y})$ decrease monotonically when exceeding the true model. In Section 5.2 we give an example that clearly exhibits this behavior. In Section 5.3 we give proofs of the main results. We start with some necessary definitions.

Definition 5.1.1 Let S_1 and S_2 be two tree models. we say that S_1 **extends** model S_2 and denote $S_1 \prec S_2$ if every $s \in S_2$ is a suffix of some $\hat{s} \in S_1$.

Definition 5.1.2 A **structure** is a sequence of models such that $S_1 \prec S_2 \prec \dots \prec S_i \prec \dots$

An example for a sequence of models is $S_i = \mathcal{X}^i$. The corresponding classes of sources form the sequence $\mathbf{P}(S_1) \subset \mathbf{P}(S_2) \subset \dots \subset \mathbf{P}(S_d) \subset \dots$

5.2 An Illustrative Example

Assume a binary alphabet $\mathcal{X} = \{0, 1\}$. Let P_1 and P_2 be two Bernoulli sources such that $P_1(1) = 1$ and $P_2(1) = 0$. Assume two samples $\mathbf{x} = 11, \dots, 1$ and $\mathbf{y} = 00, \dots, 0$ of lengths n and m . We wish to determine whether they came from a single mutual source. If we assume that $S = \emptyset$ (i.e. $\mathbf{P}(S)$ is the set of all zero-order Markov sources) then it is easy to see that the most likely mutual source is the binary zero order Markov source with parameter $\Pr(1) = \frac{1}{2}$ and that $J^S(\mathbf{x}, \mathbf{y}) > 0$, which means that there exists an optimal sequence of critical regions $\{\Lambda_n\}$ for this case, such that $\mathbf{y} \in \Lambda_n$ for all large n . Intuitively, this means that the two sources are more likely not to have come from the same source. This conclusion appears to be natural. But if we now assume that $S = \{0, 1\}$, i.e. $\mathbf{P}(S)$ is the set of Markov sources of order bounded by $D = 1$, then we get a different conclusion. The most likely mutual source is given by the initial and transition probabilities as given below,

$$\begin{aligned} P(x = 1) &= \frac{1}{2}, P(x = 0) = 1 - P(x = 1) = \frac{1}{2} & (5.1) \\ P(x_2 = 1|x_1 = 1) &= 1, P(x_2 = 0|x_1 = 1) = 1 - P(x_2 = 1|x_1 = 1) = 0 \\ P(x_2 = 1|x_1 = 0) &= 0, P(x_2 = 0|x_1 = 0) = 1 - P(x_2 = 1|x_1 = 0) = 1, \end{aligned}$$

and $J^S(\mathbf{x}, \mathbf{y}) = 0$, which means that for all optimal sequences of critical regions $\{\Lambda_n\}$ for this case, $\mathbf{y} \in \bar{\Lambda}_n$ for all large n . Note that for a source that is just slightly different from the source (5.1) in that there are no zero probabilities but rather very small ones, the behavior is not much different although the source is ergodic.

In the next section we will show that the behavior in the general case is much like as described here only much less extreme. The main intuition is as follows, assume two samples \mathbf{x} and \mathbf{y} . It is possible to choose a sufficiently large order source flexible enough to describe both samples (this will be stated more formally later). Since taking long enough strings any two distinct sources act like the two sources shown in Section 5.2. Each source gives close to one probability to the set of strings in its type (with respect to some model), and close to zero probability to the rest of the strings. The mutual source will have a type which is a unification of the two types of the two sources. In the next section, more formal proofs to this intuition are given.

5.3 Proofs of main results

Through out this section we assume a structure $S_1 \prec S_1 \prec \dots \prec S_i \prec \dots$ of models, where $S_i = \mathcal{X}^{i-1}$, and let $P_1, P_2 \in \mathbf{P}(S_D)$. In this section we show a monotonic property of $J^{S_i}(P_1, P_2)$. The structure of the section is as follows: Lemma 5.3.2 shows that $J^{S_i}(P_1, P_2)$ decreases monotonically for all $i > D$. Lemma 5.3.3 shows that this decrease is monotone in a strong sense up to the point where $J^{S_i}(P_1, P_2)$ is zero. Finally Lemma 5.3.4 shows that in the limit $J^{S_i}(P_1, P_2)$ is always zero. We start with a useful property.

Lemma 5.3.1 *Let $P \in \mathbf{P}(S_D)$, then for all $i \in \mathbb{N}$*

$$H^{S_i}(P) - H^{S_{i+1}}(P) = \sum_{\bar{s} \in S_i} P(\bar{s}) J(\{P(X|\bar{s}a)\}_{a \in \mathcal{X}}). \quad (5.2)$$

Proof.

$$\begin{aligned} H^{S_i}(P) - H^{S_{i+1}}(P) &= \sum_{\bar{s} \in S_i} P(\bar{s}) \sum_{a \in \mathcal{X}} P(a|\bar{s}) \log P(a|\bar{s}) - \sum_{s \in S_{i+1}} P(s) \sum_{a \in \mathcal{X}} P(a|s) \log P(a|s) \\ &= \sum_{\bar{s} \in S_i} P(\bar{s}) \sum_{a \in \mathcal{X}} P(a|\bar{s}) \log P(a|\bar{s}) - \sum_{a \in \mathcal{X}} \sum_{\bar{s} \in S_i} P(\bar{s}) P(a|\bar{s}) \sum_{a \in \mathcal{X}} P(a|s) \log P(a|s) \\ &= \sum_{\bar{s} \in S_i} P(\bar{s}) \sum_{a \in \mathcal{X}} P(a|\bar{s}) \log P(a|\bar{s}) - \sum_{s \in S_i} P(\bar{s}) \sum_{a \in \mathcal{X}} P(a|\bar{s}) \sum_{a \in \mathcal{X}} P(a|s) \log P(a|s) \\ &= \sum_{\bar{s} \in S_i} P(\bar{s}) H(P(X|\bar{s})) - \sum_{\bar{s} \in S_i} P(\bar{s}) \sum_{a \in \mathcal{X}} P(a|\bar{s}) H(P(X|\bar{s}a)) \\ &= \sum_{\bar{s} \in S_i} P(\bar{s}) J(\{P(X|\bar{s}a)\}_{a \in \mathcal{X}}) \end{aligned}$$

■

Corollary 5.3.1 *For every Markov source P the following inequality holds*

$$H^{S_i}(P) - H^{S_{i+1}}(P) \geq 0,$$

and equality hold if and only if S_i extends the model of P .

Lemma 5.3.2 *Let $P_1, P_2 \in \mathbf{P}(S_D)$, then for all $i \geq D$*

$$J_\lambda^{S_i}(P_1, P_2) \geq J_\lambda^{S_{i+j}}(P_1, P_2), \forall j > 0. \quad (5.3)$$

Proof. It is sufficient to show that Eq. (5.3) holds for $j = 1$. We know that

$$J_\lambda^{S_i}(P_1, P_2) - J_\lambda^{S_{i+1}}(P_1, P_2) = H^{S_i}(P_\lambda^{S_i}) - H^{S_{i+1}}(P_\lambda^{S_{i+1}}) - \quad (5.4)$$

$$[H^{S_i}(P_1) - H^{S_{i+1}}(P_1) + H^{S_i}(P_2) - H^{S_{i+1}}(P_2)]. \quad (5.5)$$

By Corollary 5.3.1 the term (5.5) equals zero for $i \geq d$ and thus

$$J_\lambda^{S_i}(P_1, P_2) - J_\lambda^{S_{i+1}}(P_1, P_2) = H^{S_i}(P_\lambda^{S_i}) - H^{S_{i+1}}(P_\lambda^{S_{i+1}}) \quad (5.6)$$

$$= H^{S_i}(P_\lambda^{S_{i+1}}) - H^{S_{i+1}}(P_\lambda^{S_{i+1}}) \geq 0, \quad (5.7)$$

and equality holds if and only if $P_\lambda^{S_i} = P_\lambda^{S_{i+1}}$, in Lemma 5.3.3 we show necessary and sufficient conditions for this. Note that Eq. (5.7) is due to Corollary 5.3.1 ■

Lemma 5.3.3 *Let $P_1, P_2 \in \mathbf{P}(S_d)$ and let $i \geq d$ then*

a. *If $P_\lambda^{S_i} = P_\lambda^{S_{i+1}}$ then $J_\lambda^{S_i}(P_1, P_2) = 0$.*

b. *If $P_\lambda^{S_i} \neq P_\lambda^{S_{i+1}}$ then $P_1 \neq P_2$ and $J_\lambda^{S_i}(P_1, P_2) \neq 0$.*

Proof. For any $a, b \in \mathcal{X}$ and $\bar{s} \in S_i$, denote by s the suffix $\bar{s} \circ b$ in S_{i+1} , then

$$P_\lambda^{S_{i+1}}(a|s) = \frac{\lambda P_1(s)}{\lambda P_1(s) + (1-\lambda) P_2(s)} P_1(a|s) + \frac{(1-\lambda) P_2(s)}{\lambda P_1(s) + (1-\lambda) P_2(s)} P_2(a|s) \quad (5.8)$$

$$= \frac{\lambda P_1(s)}{\lambda P_1(s) + (1-\lambda) P_2(s)} P_1(a|\bar{s}) + \frac{(1-\lambda) P_2(s)}{\lambda P_1(s) + (1-\lambda) P_2(s)} P_2(a|\bar{s}) \quad (5.9)$$

$$= \frac{\lambda P_1(\bar{s}) P_1(b|\bar{s})}{\lambda P_1(\bar{s}) P_1(b|\bar{s}) + (1-\lambda) P_2(\bar{s}) P_2(b|\bar{s})} P_1(a|\bar{s}) + \quad (5.10)$$

$$\frac{(1-\lambda) P_2(\bar{s}) P_2(b|\bar{s})}{\lambda P_1(\bar{s}) P_1(b|\bar{s}) + (1-\lambda) P_2(\bar{s}) P_2(b|\bar{s})} P_2(a|\bar{s}) \quad (5.11)$$

$$= \frac{\lambda P_1(\bar{s})}{\lambda P_1(\bar{s}) + (1-\lambda) P_2(\bar{s}) \frac{P_2(b|\bar{s})}{P_1(b|\bar{s})}} P_1(a|\bar{s}) + \quad (5.12)$$

$$\frac{(1-\lambda) P_2(\bar{s})}{\lambda P_1(\bar{s}) \frac{P_1(b|\bar{s})}{P_2(b|\bar{s})} + (1-\lambda) P_2(\bar{s})} P_2(a|\bar{s}),$$

where in (5.12) we assume that $P_1(b|\bar{s})$ and $P_2(b|\bar{s})$ are both strictly positive. On the other hand we know that

$$P_\lambda^{S_i}(a|\bar{s}) = \frac{\lambda P_1(\bar{s})}{\lambda P_1(\bar{s}) + (1-\lambda) P_2(\bar{s})} P_1(a|\bar{s}) + \frac{(1-\lambda) P_2(\bar{s})}{\lambda P_1(\bar{s}) + (1-\lambda) P_2(\bar{s})} P_2(a|\bar{s}). \quad (5.13)$$

It follows that $P_\lambda^{S_{i+1}}(a|s) = P_\lambda^{S_i}(a|\bar{s})$ if and only if one of the following conditions is true:

1. $P_1(\bar{s}) = 0$ or $P_2(\bar{s}) = 0$.
2. $P_1(a|\bar{s}) = P_2(a|\bar{s})$.
3. $P_1(a|\bar{s}) \neq P_2(a|\bar{s})$ but $P_1(b|\bar{s}) = P_2(b|\bar{s})$.

In addition, note that $J_\lambda^S(P_1, P_2)$ can be broken just like the entropy and D_{kl} to the following convex sum,

$$\sum_{s \in S} P_\lambda^S(s) J_{\lambda_s}^\emptyset(P_1(X|s), P_2(X|s)), \quad (5.14)$$

where $\lambda_s = \frac{\lambda P_1(s)}{\lambda P_1(s) + (1-\lambda) P_2(s)}$. Next we prove (a) and (b). (a) If $P_\lambda^{S_i} = P_\lambda^{S_{i+1}}$ then for all $a, b \in \mathcal{X}$ and $\bar{s} \in S_i$, one of the conditions (1)-(3) are true. Then in particular they are true for all $a = b$. In this case condition (3) is never true and we get that for any \bar{s} , either condition (1) holds which means that $\lambda_{\bar{s}} \in \{0, 1\}$, or condition (2) holds for all a which means $P_1(X|\bar{s}) = P_2(X|\bar{s})$. In either case, using Lemma 3.4.6 the element related to \bar{s} in Eq. (5.14) is 0, and thus the sum is zero. (b) If on the other hand $P_\lambda^{S_i} \neq P_\lambda^{S_{i+1}}$, then there exist $a, b \in \mathcal{X}$ and $\bar{s} \in S_i$ such that all conditions (1)-(3) does not hold. Condition (2) does not hold if for some $c \in \mathcal{X}$ $P_1(c|\bar{s}) \neq P_2(c|\bar{s})$ (c being either a or b), and thus $P_1(X|\bar{s}) \neq P_2(X|\bar{s})$. Condition (1) not holding means that $\lambda_{\bar{s}} \in (0, 1)$. Using Eq. (5.14) again and Lemma 3.4.6 it implies that $J_\lambda^{S_i}(P_1, P_2) \neq 0$. ■

Lemma 5.3.4 $\inf_{i>0} \left\{ J_\lambda^{S_i}(P_1, P_2) \right\} = 0$.

Proof. Recall Eq. (5.14). The idea of the proof is as follows: For all suffixes s that are not typical with respect to P_1 nor with respect to P_2 their total probability is negligible, as the length of the suffixes grows. For all suffixes s that are typical with respect to P_1 and not typical with respect to P_2 or vice versa, λ_s is very small. Finally, for distinct P_1 and P_2 there exists an ϵ such that their sets of typical sets are “far enough” (i.e. the sets are separable).

- Let $\epsilon \geq 0$ be a real number such that $T_{P_1}^\epsilon \subset \overline{T_{P_2}^{2\epsilon}}$ and $T_{P_2}^\epsilon \subset \overline{T_{P_1}^{2\epsilon}}$, i.e. all ϵ typical sequences with respect to P_1 are 2ϵ not typical sequences with respect to P_2 and vice versa. Lemma 5.3.5 shows that such a number always exists.
- We break (5.14) into two terms as follows,

$$\begin{aligned} & \sum_{s \in S} P_\lambda^S(s) J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) = \\ & \sum_{s \in T_{P_1}^\epsilon \cap T_{P_2}^\epsilon} P_\lambda^S(s) J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) + \sum_{s \in S, s \notin T_{P_1}^\epsilon \cap T_{P_2}^\epsilon} P_\lambda^S(s) J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)). \end{aligned}$$

Next we bound each of the terms.

- For the first term we use the inequality,

$$\sum_{s \in T_{P_1}^\epsilon \cap T_{P_2}^\epsilon} P_\lambda^S(s) J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) \leq \max_{s \in T_{P_1}^\epsilon \cap T_{P_2}^\epsilon} \left\{ J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) \right\}, \quad (5.15)$$

and give a bound for $\max_{s \in T_{P_1}^\epsilon \cap T_{P_2}^\epsilon} \left\{ J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) \right\}$. For any $s \in T_{P_2}^\epsilon$, denote $|s| = d$, then

$$\begin{aligned} \lambda_s &= \frac{\lambda P_1(s)}{\lambda P_1(s) + (1-\lambda) P_2(s)} \leq \frac{\lambda P_1(s)}{(1-\lambda) P_2(s)} \\ &= \exp \left\{ -d \left(H(P_s) + D_{kl}(P_s||P_1) - \frac{\log \lambda}{d} \right) \right\} \\ &\quad \cdot \exp \left\{ +d \left(H(P_s) + D_{kl}(P_s||P_2) - \frac{\log(1-\lambda)}{d} \right) \right\} \\ &= \exp \left\{ -d \left(D_{kl}(P_s||P_1) - D_{kl}(P_s||P_2) - \frac{O(1)}{d} \right) \right\} \\ &\leq \exp \left\{ -d \left(\epsilon - \frac{O(1)}{d} \right) \right\} \text{ taking large enough } d \\ &\leq \exp \left\{ -d \left(\frac{\epsilon}{2} \right) \right\}. \end{aligned}$$

By Lemma 3.4.8, $J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s))$ goes to 0 as λ_s goes to 0. Thus for every $\gamma \geq 0$ and large enough d ,

$$\max_{s \in T_{P_2}^\epsilon} \left\{ J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) \right\} \leq \gamma,$$

where $s \in \mathcal{X}^d$. An equivalent derivation can be done for any $s \in T_{P_1}^\epsilon$, hence achieving a bound for (5.15).

- For the second term we use the following inequality,

$$\begin{aligned} & \sum_{s \in S, s \notin T_{P_1}^\epsilon \cap T_{P_2}^\epsilon} P_\lambda^S(s) J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) \leq \\ & \max_s \left\{ J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) \right\} \sum_{s \in S, s \notin T_{P_1}^\epsilon \cap T_{P_2}^\epsilon} P_{\lambda_s}^S(s). \end{aligned} \quad (5.16)$$

By Lemma 3.4.8,

$$\max_s \left\{ J_{\lambda_s}^\theta(P_1(X|s), P_2(X|s)) \right\} \leq 1.$$

Next we give a bound for $\sum_{s \in S, s \notin T_{P_1}^\epsilon \cap T_{P_2}^\epsilon} P_{\lambda_s}^S(s)$. Denote by \mathcal{N} the set $\overline{T_{P_1}^\epsilon \cap T_{P_2}^\epsilon}$, then it follows from Theorem 2.4.1 that,

$$\begin{aligned} P_1(\mathcal{N}) & \leq \exp \left\{ -d \left(\epsilon - O \left(\frac{\log d}{d} \right) \right) \right\}, \\ P_2(\mathcal{N}) & \leq \exp \left\{ -d \left(\epsilon - O \left(\frac{\log d}{d} \right) \right) \right\}, \end{aligned}$$

thus,

$$P_\lambda^S(\mathcal{N}) \leq \exp \left\{ -d \left(\epsilon - O \left(\frac{\log d}{d} \right) \right) \right\},$$

where $S = \mathcal{X}^d$. So for any $\gamma \geq 0$ and large enough d ,

$$P_\lambda^S(\mathcal{N}) \leq \gamma.$$

■

Lemma 5.3.5 *Let $P_1, P_2 \in \mathbf{P}(S)$ for some model S , be two distinct sources, then there exists a real number $\epsilon > 0$ such that $T_{P_1}^\epsilon \subset \overline{T_{P_2}^{2\epsilon}}$ and $T_{P_2}^\epsilon \subset \overline{T_{P_1}^{2\epsilon}}$.*

Proof. For all $\alpha > 0$ denote

$$P_\alpha^* \triangleq \arg \min_{P \in \mathbf{P}(S), D_{kl}(P||P_1) \leq \alpha} D_{kl}(P||P_2).$$

Let $\alpha_1 \in (0, D_{kl}(P_2||P_1))$, then $P_{\alpha_1}^* \neq P_2$ and $D_{kl}(P_{\alpha_1}^*||P_2) \triangleq \delta > 0$. Obviously for all $0 \leq \alpha_2 \leq \alpha_1$, $D_{kl}(P_{\alpha_2}^*||P_2) \geq \delta$. Now for $\epsilon \triangleq \min\{\frac{\delta}{2}, \alpha_1\}$ we have that $T_{P_1}^\epsilon \subset \overline{T_{P_2}^{2\epsilon}}$, since for all $P \in \mathbf{P}(S)$ such that $D_{kl}(P||P_1) \leq \epsilon$ we have that

$$\begin{aligned} D_{kl}(P||P_2) &\geq D_{kl}(P_\epsilon^*||P_2) \\ &\geq \delta \geq 2\epsilon. \end{aligned}$$

The statement of the lemma than follows. ■

5.4 The Two Samples Problem Revisited

In Section 4.2.2 we gave a proof that the GLRT is an asymptotically optimal test for the two samples problem when assuming the two sources P_1 and P_2 are in a known family $\mathbf{P}(S)$. In particular we assumed that we know a bound on the order of the unknown sources. This test was shown to be based on the statistic $J^S(\mathbf{x}, \mathbf{y})$, which is strongly dependent on the model S as Section 5.3 shows. What if the model S is not known, i.e. there is no known bound on the order of the two sources?. We address this question in the rest of the section. We show that the same line of proof as in Section 4.2.2, can be used to show that the GLRT is an asymptotically optimal for this case as well. Only the test is based on the minimum of all statistics $J^S(\mathbf{x}, \mathbf{y})$ over all models S that can describe the samples. Using the results of Section 5.3 we then argue why this test is trivial in some sense.

5.4.1 Optimal Test

Consider sources $P_1, P_2 \in \mathbf{P}(S)$, and samples \mathbf{x} and \mathbf{y} as in the last sections. In this section we assume that S is unknown. Denote by Λ_n^* the following critical region

$$\Lambda_n^* = \{\mathbf{x}, \mathbf{y} : \inf_{i>0} \frac{1}{l} \log \frac{\sup_{P_1, P_2 \in \mathbf{P}(S_i)} P_1(\mathbf{x})P_2(\mathbf{y})}{\sup_{P \in \mathbf{P}(S_i)} P(\mathbf{x}, \mathbf{y})} > \lambda\}$$

Theorem 5.4.1 *The sequence of critical regions $\{\Lambda_n^*\}_{n \geq 1}$ defined above, is asymptotically optimal.*

Proof. First we will show that

$$\limsup_{l \rightarrow \infty} \frac{1}{l} \log \sup_{P \in \mathbf{P}} P(\Lambda_n^*) \leq -\lambda \quad (5.17)$$

this can be shown by

$$\begin{aligned} \log \sup_{P \in \mathbf{P}} P(\Lambda^*) &= \log \sup_{P \in \mathbf{P}} \sum_{(\mathbf{x}, \mathbf{y}) \in \Lambda^*} P(\mathbf{x})P(\mathbf{y}) \\ &= \log \sup_{P \in \mathbf{P}} \sum_{T_{\mathbf{x}}, T_{\mathbf{y}} \subset \Lambda^*} P(T_{\mathbf{x}})P(T_{\mathbf{y}}) \\ &\leq \log \sup_{P \in \mathbf{P}} \sum 2^{-nD(P_{\mathbf{x}}||P) - mD(P_{\mathbf{y}}||P)} \\ &\leq \sup_{(\mathbf{x}, \mathbf{y}) \subset \Lambda^*} \left\{ - \inf_{P \in \mathbf{P}} \{nD(P_{\mathbf{x}}||P) + mD(P_{\mathbf{y}}||P)\} + O(\log n + \log m) \right\} \\ &\leq \sup_{(\mathbf{x}, \mathbf{y}) \subset \Lambda^*} \left\{ - \inf_{i > 0} \{J^{S_i}(\mathbf{x}, \mathbf{y}) \cdot l\} + O(\log n + \log m) \right\} \\ &< -\lambda \cdot l + O(\log n + \log m) \end{aligned}$$

Thus (5.17) holds. For the second condition of asymptotic optimality, denote by Λ an arbitrary test that satisfies (5.17), then for large enough l and any $(\mathbf{x}, \mathbf{y}) \in \Lambda$

$$\begin{aligned} -\lambda + \epsilon_l &> \frac{1}{l} \log \sup_{P \in \mathbf{P}} P(\Lambda) \geq \frac{1}{l} \log \sup_{P \in \mathbf{P}} P(T_{\mathbf{x}}, T_{\mathbf{y}}) \\ &\geq - \inf_{P \in \mathbf{P}} \left(\frac{n}{l} D(P_{\mathbf{x}}||P) + \frac{m}{l} D(P_{\mathbf{y}}||P) \right) + \epsilon_l \\ &\geq \inf_{i > 0} \{-J^S(\mathbf{x}, \mathbf{y})\} + \epsilon_l \end{aligned}$$

Where $\epsilon_l = O(\log(n) + \log(m))$. Thus $\Lambda_l \subset \Lambda_l^*$ for large enough l . which concludes the proof. ■

Chapter 6

Conclusions and Future Research

6.1 The Generalization Power of Compression Algorithms

6.1.1 The Two Samples Problem, a Different Angle

Many times, when the two samples problem comes up in practice, it has the following form; We are able to sample two information sources (e.g neurons) ¹, or we have two finite samples of this sources, and we wish to determine whether this two sources differ or not. Theorem 4.2.1 suggests that we test whether $J^S(\mathbf{x}, \mathbf{y}) > \lambda$, where $J(\mathbf{x}, \mathbf{y})^S$ is defined in Section 3 and λ is a positive real that determines the rate of decrease of the error of the first kind. But the following questions/problems arise:

- How to choose λ ?
- When λ is fixed, some sources will not be distinguished even for very long samples (i.e. different sources P_1, P_2 with $J^S(P_1, P_2) < \lambda$).
- How to choose a model S ?

But as part of the proof of Theorem 4.2.1 we proved the following lemma.

¹More accurately we assume that it is possible to model the unknown source as a discrete information source. Usually we also assume that this source is stationary and ergodic.

Lemma 6.1.1 *Let $P \in \mathbf{P}(S)$ then*

$$P \{ \mathbf{x}, \mathbf{y} : J^S(\mathbf{x}, \mathbf{y}) > \epsilon \} \leq 2^{l(\epsilon - \frac{n}{l}O(\frac{\log n}{n}) - \frac{m}{l}O(\frac{\log m}{m}))}. \quad (6.1)$$

Thus if we have two samples with lengths n and m the test $J^S(\mathbf{x}, \mathbf{y}) > \epsilon$ ensures confidence of at least $1 - 2^{l(\epsilon - \frac{n}{l}O(\frac{\log n}{n}) - \frac{m}{l}O(\frac{\log m}{m}))}$. Note that according to this analyses, we must have $\epsilon > \frac{n}{l}O(\frac{\log n}{n}) - \frac{m}{l}O(\frac{\log m}{m})$. Thus if we wish to solve the two samples problem in such a way that when there is enough statistics we include as many sources as possible then we should test whether $J^S(\mathbf{x}, \mathbf{y}) > \epsilon_{n,m}$ where $\epsilon_{n,m} > \frac{n}{l}O(\frac{\log n}{n}) - \frac{m}{l}O(\frac{\log m}{m})$. Next we show that testing whether $\tilde{J}^S(\mathbf{x}, \mathbf{y}) > 0$ is such a test, where $\tilde{J}^S(\mathbf{x}, \mathbf{y})$ defined in Lemma 4.2.2 is a test based on the Bayesian estimate of probability. To see that consider the following relation,

$$\tilde{J}^S(\mathbf{x}, \mathbf{y}) \quad (6.2)$$

$$= -\frac{1}{l} \log P_e^S(\mathbf{x}, \mathbf{y}) + \frac{n}{l} \frac{1}{n} \log P_e^S(\mathbf{x}) + \frac{m}{l} \frac{1}{m} \log P_e^S(\mathbf{x}) \quad (6.3)$$

$$\geq -\frac{1}{l} \log P_{\mathbf{x}, \mathbf{y}}^S(\mathbf{x}, \mathbf{y}) + \frac{n}{l} \frac{1}{n} \log P_{\mathbf{x}}^S(\mathbf{x}) + \frac{m}{l} \frac{1}{m} \log P_{\mathbf{y}}^S(\mathbf{x}) - \frac{n}{l} O(\frac{\log n}{n}) - \frac{m}{l} O(\frac{\log m}{m}) \quad (6.4)$$

$$= J^S(\mathbf{x}, \mathbf{y}) - \frac{n}{l} O(\frac{\log n}{n}) - \frac{m}{l} O(\frac{\log m}{m}) \quad (6.5)$$

where (6.4) is due to Lemma 2.5.3 and Lemma 2.2.2. Thus $\tilde{J}^S(\mathbf{x}, \mathbf{y}) > 0$ implies $J^S(\mathbf{x}, \mathbf{y}) > \frac{n}{l} O(\frac{\log n}{n}) - \frac{m}{l} O(\frac{\log m}{m})$. This gives another example to the generalization power of estimating probability via the Bayesian estimate (e.g. CTW).

The constants in the analyses just given are hidden behind the O notation. This is due to that the constants in Sanov's theorem does not match the ones in the analyses of the redundancy of the CTW, although intuitively they should. This relation, we think can be an interesting topic for further research.

extending to the case of unknown model

Consider the third question; many times in practice when analyzing a group of sequences we don't have full knowledge of the family of possible sources they might have been drawn from. But as shown in this thesis answering the two samples problem is strongly based on the assumed model of the sources, thus what model should we choose ?. Three possible solutions are:

- Assume the order of the sources is bounded by some integer D and take as model the full tree of depth D . For example it is a known rule of thumb that in English text the full tree of depth 5 is a “good enough” model. Two problems with this approach are; first that the assumption might be wrong. And second is that the full tree of depth D might extend the real model, but there will be many intermediate models. Thus the J statistic according to this model might be very small, which will demand much longer samples to “solve the two samples problem”.
- Use some model selection principle to decide on the model (e.g. the MDL).
- Use the model that maximize the statistic J , i.e.

$$S_m = \arg \max_S \{J^S(\mathbf{x}, \mathbf{y})\},$$

where \mathbf{x} and \mathbf{y} are finite sequences. Denote by S the real model for the sources that \mathbf{x} and \mathbf{y} are drawn from then it follows from Lemma 5.3.2 that S extends S_m , thus S_m is well defined. The advantage here is that J is maximal, the problem is that our confidence in the frequency count reduces.

We think that the relation between these methods could also be an interesting topic for further research.

6.1.2 Model Selection and the Two Samples Problem

The *model selection problem* has as its input a sample \mathbf{x} from some unknown distribution in $\mathbf{P}(S)$, for an unknown model S . Its desired output is S , with or without the actual parameters of the source, depending on the application. Next a general interpretation of the model selection problem² is described through the two samples problem. In particular a close relation is shown to the MDL principle (see Section 2.5).

Assume we wish to determine whether a Markov source P is of order 0, by examining a growing sample \mathbf{x} drawn from it. It is known that P has 0-order if and only if all infinite sub-samples in \mathbf{x} has the same statistics, i.e. all sub-samples in \mathbf{x} came from the same source

²For the specific model selection problem described here.

([9]). Knowing P is a finite order Markov source it is enough to examine all subsamples that depend on a possible state, e.g. all samples following some fixed letter in \mathcal{X} . It is claimed here that the MDL principle can be interpreted as executing exactly this test by breaking it to sub-tests and considering finite statistics.

Example 6.1.1 Consider an ergodic Markov source P of order 1, over the alphabet $\mathcal{X} = 0, 1$. Then P can be expressed by the zero-order model if and only if $P(X|1)$ is the same as $P(X|0)$, which is equivalent to $J_\lambda(P(X|1), P(X|0)) = 0$ for every positive λ , in particular for $\lambda = P(1)$.

Examine the identity given by lemma 5.3.1

$$H^{S_i}(P) - H^{S_{i+1}}(P) = \sum_{s \in S_i} P(s) J(\{P(X|s \cdot a)\}_{a \in \mathcal{X}}), \quad (6.6)$$

where P is a Markov source. Using Lemma 3.4.6 it follows that the following criterion gives the “right model” for P , for large enough n :

$$\arg \min_{S \in \bigcup_{D=1}^{\infty} \mathcal{C}_D} \left\{ H^S(P) + \frac{1}{n} \Gamma(S) \right\}, \quad (6.7)$$

where $\Gamma(S)$ is a simple generalization of the function $\Gamma_D(S)$ defined in Section 2.5 in which all leaves have cost one bit as well as the internal nodes³. Note that in this case we can simply find the first point where extending the model does not contribute anything. Using 6.6 it is easy to see that,

$$H^{S_{i+1}}(P) = H^{S_0}(P) - \sum_{s \in \bigcup_{D=1}^i \mathcal{C}_D} P(s) J(\{P(X|s \cdot a)\}_{a \in \mathcal{X}}), \quad (6.8)$$

and thus 6.7 can be rewritten,

$$\arg \min_i \left\{ H^{S_0}(P) - \sum_{s \in \bigcup_{D=1}^i \mathcal{C}_D} P(s) J(\{P(X|s \cdot a)\}_{a \in \mathcal{X}}) + \frac{1}{n} \Gamma(S) \right\}. \quad (6.9)$$

Note that the actual value of $J(\{P(X|s \cdot a)\}_{a \in \mathcal{X}})$ does not really matter for the result of 6.9, it only matters whether it is greater than zero (strictly positive) or not. Thus taking an

³Intuitively it gives the size of a coding of S without the assumption that the order is bounded by D .

indicating variable that is 0 if $J(\{P(X|s \cdot a)\}_{a \in \mathcal{X}}) = 0$ and some $\alpha > 0$ otherwise instead of $J(\{P(X|s \cdot a)\}_{a \in \mathcal{X}})$ in 6.9 will not change the results of taking the minimum.

Not knowing the real J in every node, we can use an estimate. This brings us back to the two samples problem and the problem of selecting the appropriate statistic for its testing. Next we argue that for this purpose again the statistic based on the Bayesian estimation superiors the one based on the ML. Examine the following criterion based on the ML,

$$\arg \min_{i \geq 0} \left\{ H^{S_0}(P_{\mathbf{x}}) + \sum_{s \in \bigcup_{D=1}^i \mathcal{C}_{\mathcal{D}}} P_{\mathbf{x}}(s) J(\{P_{\mathbf{x}}(X|s \cdot a)\}_{a \in \mathcal{X}}) + \frac{1}{n} \Gamma(S) \right\}, \quad (6.10)$$

this is not a good criterion since $J(\{P_{\mathbf{x}}(X|s \cdot a)\}_{a \in \mathcal{X}}) \geq 0$ and equality holds if and only if all the distributions $P_{\mathbf{x}}(X|s \cdot a)$ are identical. We are not likely to get such samples. Note that 6.10 is equivalent to

$$\arg \min_{S \in \bigcup_{D=1}^{\infty} \mathcal{C}_{\mathcal{D}}} \left\{ H^S(P_{\mathbf{x}}) + \frac{1}{n} \Gamma(S) \right\}, \quad (6.11)$$

which is known by its ... prone to choose over fitting models. The problem is we can not be sure that $P_{\mathbf{x}} = P$. But we can be relatively sure that $D_{kl}(P_{\mathbf{x}}||P) \leq O(\frac{\log n}{n})$. This brings us back to the discussion in (6.1.1) and to the suggestion of the following criterion,

$$\arg \min_{i \geq 0} \left\{ -\log P_e^S(\mathbf{x}) + \sum_{s \in \bigcup_{D=1}^i \mathcal{C}_{\mathcal{D}}} P_{\mathbf{x}}(s) \tilde{J}(\{P_{\mathbf{x}}(X|s \cdot a)\}_{a \in \mathcal{X}}) + \frac{1}{n} \Gamma(S) \right\}, \quad (6.12)$$

where \tilde{J} is the statistic based on the Bayesian estimation (CTW). Which is equivalent to

$$\arg \min_{S \in \bigcup_{D=1}^{\infty} \mathcal{C}_{\mathcal{D}}} \left\{ -\log P_w(\mathbf{x}) + \frac{1}{n} \Gamma(S) \right\}, \quad (6.13)$$

that was shown in Section 2.5 to be exactly the model chosen by the MDL principle.

6.2 Generalization to Transducers

The theory developed in this thesis so far was concerned mainly with Markov processes. That is processes, where the next symbol's distribution depends on the last D symbols. Many natural processes can be looked at as Markov processes, e.g. text, protein sequences, output of a neuron,

growth of the stock market etc. But this misses many other processes that can not be looked at as a single sequence, but rather as two depending sequences of input and output. e.g. the reaction of a neuron to input, the growth of the stock market with reaction to changes in the dollar value, a text as input and some property like whether the letter is a vowel or not as output etc. In this section we briefly discuss the generalization of the ideas presented in this thesis to the case of a processes with input and output. This approach, of generalizing known models (e.g. finite state automata) to fit to the case of an input output process is very common, see [3, 26, 8, 24, 25]. The term *transducer* is commonly used for finite state automata with input and output.

6.2.1 Definitions

Joint Information Sources and Conditional Information sources

A *conditional information source* over a finite alphabet \mathcal{Y} with respect to the process $\{X_i\}$ is defined by a family of conditional probability mass functions $Q^n(y|x_1^n)$ for every $n \geq 0$. This family must satisfy the marginality condition,

$$Q^n(y|\mathbf{x}) = \sum_{x \in \mathcal{X}} P(x|\mathbf{x})Q^{n+1}(y|\mathbf{x}x)$$

for all n . A conditional information source can be thought of as generating a sequence $\{Y_i\}$ of letters in \mathcal{Y} , such that the two sequences $\{X_i\}$ and $\{Y_i\}$ are aligned. A *conditional stochastic process* $\{Y_i|X_i\}_1^\infty$ is the process of generating $\{Y_i\}$ by a conditional information source with respect to $\{X_i\}$. Another way of seeing this process is as a stochastic input-output process, i.e. at the n th step x_1, \dots, x_n are entered as input to a box yielding as output the letter a_j with probability $Q(Y_n = a_j|x_1, \dots, x_n)$, $j = 1 \dots |\mathcal{Y}|$. A conditional stochastic process is called *stationary* if

$$Q(Y = y|X_1 = x_1, \dots, X_n = x_n) = Q(Y = y|X_{l+1} = x_1, \dots, X_{l+n} = x_n)$$

for every time shift l , $y \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}^n$. A conditional stochastic process will be called Markov of order D if

$$Q(Y = y|X_1 = x_1, \dots, X_n = x_n) = Q(Y = y|X_{n-D+1} = x_{n-D+1}, \dots, X_n = x_n)$$

For every $y \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}^n$. As in the case of information sources, it is necessary for applicative purposes to assume that

$$Q(y|\mathbf{x}) = Q(y|f(\mathbf{x}))$$

for every $y \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}^n$, where $f : \mathcal{X}^n \rightarrow S$ is a function from \mathcal{X}^n to a finite state space S . This set has similar functionality as in the case of information sources. The definitions from the Section 2.2 of a model and a tree model still stand. As one can see a stochastic process is a special case of a conditional stochastic process where the input and output sequences are the same, only the input is shifted one time unit to the left. Thus the generalization of most of the results shown in this thesis to the case of conditional sources is straightforward, as shown by the above definitions. For example in [26] a generalization of the CTW algorithm for this case is given, and the equivalent of all the results stated in Section 2.5 are deduced.

A **joint stochastic process** $\{(X_i, Y_i)\}_1^\infty$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$, is a stochastic process over the alphabet $\mathcal{X} \times \mathcal{Y}$. As we have seen in the last section this process is characterized by the family of joint distributions $P^n((x_1, y_1) \dots (x_n, y_n))$ for each $n = 1, 2, \dots$. A **joint information source** over the finite alphabets \mathcal{X} and \mathcal{Y} is an information source over the alphabet $\mathcal{X} \times \mathcal{Y}$.

6.2.2 Implications

As we have seen the ideas presented in this thesis can be naturally extended to the case of conditional sources, but why is this interesting?, we give here two reasons why we think the idea of conditional sources can lead to new perspectives. First, it gives us a tool to answer new kind of questions, e.g. does English and French differ in our ability to guess whether the next letter is 'a' or not, or equivalently does two families of proteins differ in our ability to guess whether the next amino acid is hydrophobic or not. We give next the general formalization of such questions.

Let \mathcal{X} be a finite alphabet, and let $\{X_i\}_{i=1}^\infty$ be a stochastic process over \mathcal{X} . Let \mathcal{X}_1 and \mathcal{X}_2 be two finite sets such that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, i.e. \mathcal{X} is the set of all pairs of elements of \mathcal{X}_1 and \mathcal{X}_2 . For example if $\mathcal{X} = \{1, 2, 3, 4\}$ we can take $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$ and write $1 = (0, 0), 2 = (1, 0), 3 = (0, 1), 4 = (1, 1)$. Now the process $\{X_i\}_{i=1}^\infty$ defines the joint process $\{(X_{1,i}, X_{2,i})\}$ where $X_{1,i} \in \mathcal{X}_1$ and $X_{2,i} \in \mathcal{X}_2$, and also the conditional processes $\{(X_{2,i}|X_{1,i})\}, \{(X_{1,i}|X_{2,i})\}, \{(X_{2,i}|X_i)\}$ etc. Thus, take two information sources $P_1, P_2 \in \mathbf{P}(S)$ over \mathcal{X} , then the general form of the questions above is asking whether $P_1(X_{1,i}|X_n, \dots, X_1)$ and $P_2(X_{2,i}|X_n, \dots, X_1)$ are realizations of the same conditional source.

The second purpose, the concept of conditional sources might be useful for, is decomposing processes over large alphabets into sub processes with smaller output alphabets. We elaborate on this idea next. Assume that the information source P defining the process $\{X_i\}_{i=1}^\infty$ is in $\mathbf{P}(S)$, and consider the following decomposition of the probability given by P ,

$$\begin{aligned} P(x_n|x_{n-1}, \dots, x_1) &= P((x_{1,n}, x_{2,n})|(x_{1,n-1}, x_{2,n-1}) \dots (x_{1,1}, x_{2,1})) \\ &= P_1(x_{1,n}|(x_{1,n-1}, x_{2,n-1}) \dots (x_{1,1}, x_{2,1})) \cdot \\ &\quad P_2(x_{2,n}|x_{1,n}, (x_{1,n-1}, x_{2,n-1}) \dots (x_{1,1}, x_{2,1})). \end{aligned}$$

This gives a decomposition of the information source to two conditional information sources. The first is the conditional information source defining the process $\{X_{1,i}|X_i\}$, while the second is the information source defining the process $\{X_{2,i}|\tilde{X}_i\}$ where $\tilde{X}_i \in \mathcal{X} \cup \mathcal{X}_1$. The output alphabets of these conditional sources are \mathcal{X}_1 and \mathcal{X}_2 and they can be as small as $\sqrt{|\mathcal{X}|}$, although this is not the only quality the decomposing alphabets should be chosen by. It is common knowledge that the CTW compression algorithm does much better when decomposed in this way⁴. To try and understand the reason for the improvement lets look at the CTW bounds (assuming they represent the real redundancy) for both the case of the decomposed process and of the joint one. For the joint process we have,

$$-\log P_w(\mathbf{x}) \leq \min_{S \in \mathcal{C}_D} \left(\min_{P \in \mathbf{P}(S)} -\log P(\mathbf{x}) + \Gamma_D(S) + |S| \left(\frac{|\mathcal{X}|}{2} \log \left(\frac{n}{|S|} \right) + O(1) \right) \right), \quad (6.14)$$

⁴Note that this is not the only way of improving the performance of the CTW algorithm on sequences over a large alphabet, for example a weighting scheme is also examined, see [27, 11]

assuming $n > |S|$. And for the two decomposed processes we have,

$$-\log P_w(\mathbf{x}_1|\mathbf{x}) \leq \min_{S \in \mathcal{C}_D} \left(\min_{P \in \mathbf{P}(S)} -\log P(\mathbf{x}_1|\mathbf{x}) + \Gamma_D(S) + |S| \left(\frac{|\mathcal{X}_1|}{2} \log \left(\frac{n}{|S|} \right) + O(1) \right) \right) \quad (6.15)$$

$$-\log P_w(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}) \leq \min_{S \in \mathcal{C}_D} \left(\min_{P \in \mathbf{P}(S)} -\log P(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}) + \Gamma_D(S) + |S| \left(\frac{|\mathcal{X}_2|}{2} \log \left(\frac{n}{|S|} \right) + O(1) \right) \right) \quad (6.16)$$

where $P_w(\mathbf{x}_1|\mathbf{x})$ and $P(\mathbf{x}_1|\mathbf{x})$ are not defined here, to understand the point it is enough to know that $-\log P_w(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}) + \log P(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x})$ intuitively represents the redundancy in coding \mathbf{x}_1 while modeling it via \mathbf{x} ⁵. The dominating term in the redundancy is $\frac{|S||\mathcal{X}|}{2} \log n$, that depends on the size of the alphabet and the size of the model. Denote by S the model achieving the minimum at 6.14, by S_1 the model achieving the minimum at 6.15 and by S_2 the model achieving the minimum at 6.16. The asymptotic gain in redundancy (or loss) is proportional to

$$|S||\mathcal{X}| - (|S_1||\mathcal{X}_1| + |S_2||\mathcal{X}_2|). \quad (6.17)$$

The relation between \mathcal{X}_1 , \mathcal{X}_2 and \mathcal{X} is known. What about the relation between S_1 , S_2 and S ? We now give an example and answer this question for that example. This will demonstrate that at least for some cases there can be gain in decomposing a process with large alphabet in the manner we have just shown.

Example 6.2.1 *Let $\mathcal{X} = \{1, 2, 3, 4\}$ and let $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$. Denote by P an information source over \mathcal{X} , and by $P_1(x_{1,n}|x_{n-1}, \dots, x_1)$ and $P_2(x_{2,n}|x_{1,n}, x_{n-1}, \dots, x_1)$ two conditional sources that make a decomposition of P . Then if the “marginal” processes $\{x_{1,i}\}$ and $\{x_{2,i}\}$ are independent we would get that the models S_1 and S_2 in 6.17 are at least as simple as S , and thus $6.17 \geq |S|(\mathcal{X} - \mathcal{X}_1 - \mathcal{X}_2)$.*

Note that this also gives a naive algorithm to discover relationships among groups of symbols of an alphabet as well as a door to enter prior knowledge. Finally we mention that the next step would be to further simplify the process by shrinking the input alphabet as well. In broad terms, consider the processes $\{x_{1,i}\}$ and $\{x_{2,i}\}$, if they are independent then we can break the process to two independent sub processes over smaller alphabets. If on the other hand they

⁵The redundancy is due to coding with finite statistics.

have large mutual information then we can throw (theoretically) all the duplication in $\{X_i\}$ and hence getting again a process over a smaller alphabet. A work in this direction can be found in [21], this also relates closely to [7, 10].

Bibliography

- [1] Jorma Rissanen Andrew Barron and Bin Yu. The minimum description length principle in coding and modeling. *IEEE Trans. on Information Theory*, 44(6):2743–2760, October 1998.
- [2] Cleary J.G. Bell T.C. and Witten I.H. *Probability and Statistics*. Prentice Hall, NJ, 1990.
- [3] Y. Bengio and P. Frasconi. An input/output hmm architecture. *Advances in Neural Information Processing Systems*, 1994.
- [4] John G. Cleary and W.J.Teahan. Unbounded length contexts for ppm. *The Computer Journal*, 36(5), 1993.
- [5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
- [6] Imre Csiszar. The method of types. *IEEE Trans. Inform. Theory*, 44(6):2505–2523, October 1998.
- [7] R. El-Yaniv, S. Fine, and N. Tishby. Agnostic classification of markovian sequences. *Advances in Neural Information Processing Systems*, 10:465–471, 1997.
- [8] R. Emmanuel and Y. Schabes. Deterministic part-of-speech tagging with finite-state transducers, 1995.
- [9] William Feller. *An Introduction to Probability Theory and Its Applications*. Volume I, second edition, John Wiley & Sons, inc, New York, 1961.

- [10] Y.M. Shtarkov F.M.J. Willems and Tj.J. Tjalkens. Context weighting for general finite-context sources. *IEEE Transactions on Information Theory*, pages 1514–1520, September 1996.
- [11] N. Friedman and Y. Singer. Efficient bayesian parameter estimation in large discrete domains, 1999.
- [12] Jhon G.Kemeny and J.Laurie Snell. *Finite Markov Chains*. Springer-Verlag, New York, 1976.
- [13] M. Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. on Information Theory*, 35(2):401–408, 1989.
- [14] Morris H.Degroot. *Probability and Statistics*. Addison-Wesley Pub Co, January 1986.
- [15] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *The Annal of Mathematical Statistics*, 36:369–401, 1965.
- [16] Steven H. Kim. *Statistics And Decisions, An Introduction to Foundations*. Van Nostrand Reinhold, New York, 1992.
- [17] R.E. Krichevsky and V.K. Trofimov. The performance of universal encoding. *IEEE Trans. Inform. Theory*, IT-27:199–207, March 1981.
- [18] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37, No. 1:145–151, 1991.
- [19] N. Merhav. The estimation of the model order in exponential families. *IEEE Trans. Inform. Theory*, 35:1109–1114, 1989.
- [20] M. Gutman N. Merhav and J. Ziv. On the estimation of the order of a markov chain and universal data compression. *IEEE Trans. Inform. Theory*, 35:1014–1019, 1989.
- [21] Naftali Tishby Noam Slonim. Agglomerative information bottleneck.
- [22] J. Rissanen. A universal data compression system. *IEEE, Trans. Inform. Theory*, IT-29(5):656–664, 1983.

- [23] J. Rissanen. Complexity of strings in the class of markov sources. *IEEE, Trans. Inform. Theory*, IT-32(4):526–532, July 1986.
- [24] E. Roche. Parsing with finite state transducers, 1997.
- [25] D. Searls and K. Murphy. Automata-theoretic models of mutation and alignment, 1995.
- [26] Y. Singer. Adaptive mixture of probabilistic transducers. *Neural Computation*, 9(8), 1997.
- [27] T. Tjalkens, Y. Shtarkov, and F. Willems. Sequential weighting algorithms for multi-alphabet sources, 1993.
- [28] P.M.B. Vitanyi and M. Li. Ideal mdl and its relation to bayesianism. *Proc. ISIS:Information, Statistics and Induction in Science World Scientific, Singapore*, pages 282–291, 1996.
- [29] Paul A.J. Volf and Frans M.J. Willems. Context maximizing: Finding mdl decision trees. 1994.
- [30] Paul A.J. Volf and Frans M.J. Willems. A study of the context tree maximizing method. 1995.
- [31] M. J. Weinberger, J. Rissanen, and M. Feder. A universal finite memory source. *IEEE Trans. on Information Theory*, 41(3):643–652, 1995.
- [32] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method: Basic properties. *IEEE Trans. on Information Theory*, 41(3):653–664, 1995.
- [33] F.M.J. Willems. The context-tree weighting method: Extentions. *IEEE Transactions on Information Theory*, pages 792–798, March 1998.
- [34] F.M.J. Willems and Tj.J. Tjalkens. Complexity reduction of the context-tree weighting method. In *Benelux, Veldhoven, The Netherlands*, May 1997.
- [35] O. Zeitouni, J. Ziv, and N. Merhav. When is the generalized likelihood ratio test optimal? *IEEE Trans. on Information Theory*, 38(5):1597–1602, 1992.

- [36] J. Ziv. On classification with empirically observed statistics and universal data compression. *IEEE Transactions on Information Theory*, 34:278–286, 1988.
- [37] J. Ziv and N. Merhav. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, 39(4):1270–1279, 1993.