VARIABLE-TO-FIXED LENGTH CODES FOR SOURCES WITH KNOWN AND UNKNOWN MEMORY

by

Serap Ayşe Savari

S.B., Electrical Engineering
Massachusetts Institute of Technology, 1990

S.M., Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 1991

S.M., Operations Research
Massachusetts Institute of Technology, 1991

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1996

Author .....
Department of Electrical Engineering and Computer Science
January 31, 1996

Certified by .....
Robert G. Gallager
Professor of Electrical Engineering
Thesis Supervisor

Accepted by .....
F. R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

VARIABLE-TO-FIXED LENGTH CODES FOR SOURCES WITH KNOWN AND UNKNOWN MEMORY

by

Serap Ayşe Savari

Submitted to the Department of Electrical Engineering and Computer Science
on March 26, 1996, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Lossless source coding aims at minimizing the expected number of code letters per source symbol used to encode the output of a discrete information source under the requirement that it is possible to correctly reconstruct any source string from its corresponding code string. A variable-to-fixed length encoding procedure is a mapping from a dictionary of variable length strings of source outputs to the set of codewords of a given length. For memoryless sources, the Tunstall procedure can be applied to construct optimal uniquely parsable dictionaries and the resulting codes are known to work especially well for sources with small entropies. For sources with memory, it seems plausible that variable-to-fixed length codes are well-suited to take advantage of any predictability in the source output. In this thesis, we consider a generalization of variable-to-fixed length codes for Markov sources, and analyze its performance as the dictionary size approaches infinity. We introduce the idea of plurally parsable dictionaries and use an example to illustrate that the optimal plurally parsable dictionary of a given size can outperform the Tunstall dictionary of the same size. We also investigate the asymptotic performance of the Lempel-Ziv incremental parsing rule and two of its variants. For each of the three algorithms, we demonstrate that the redundancy of encoding the first n letters of the source output is $\Theta\left(\frac{1}{\ln n}\right)$, and we upper bound the exact form of convergence. The Lempel-Ziv codes are universal variable-to-fixed length codes that are virtually standard in practical lossless data compression.

Thesis Supervisor: Robert G. Gallager
Title: Professor of Electrical Engineering

Acknowledgements

I would like to thank my mentor and thesis supervisor, Professor Robert G. Gallager, for his guidance and insights during the course of my graduate studies. His teachings inspired my interest in both source coding and discrete stochastic processes. Our interactions and collaborations have played an invaluable role in my professional growth.

I would like to thank my thesis committee members, Professors Peter Elias, Sanjoy Mitter, and John Tsitsiklis, for their interest in my thesis. Their comments about this work have been very useful.

I am grateful to AT&T Bell Laboratories for providing me with one of their Graduate Research Program for Women Fellowships. One of the windfalls of being a GRPW student was my frequent and helpful communication with my fellowship mentor, Dr. Ellen L. Hahne. I would like to thank Ellen for her generous sponsorship and friendship during my graduate studies. This work was also partially supported by Vinton Hayes Fellowships and by Army Research Office grant ARO DAAH04-95-1-0103.

I have been very fortunate in having Professor Dimitri P. Bertsekas as my faculty advisor and a mentor. I would like to thank Professor Bertsekas for his friendship, encouragement, and support during the last five years.

The Laboratory for Information and Decision Systems has been a very pleasant work environment. I have greatly enjoyed being part of the LIDS community, and I wish to express my appreciation to the many people who have been my friends and colleagues there over the years. I am particularly grateful to Li Shu and Dr. David Tse. My many stimulating conversations with each of them were influential in the writing of this thesis.

Finally, I would like to recognize my family. The continual love and nurturing of my parents, Mr. Aykut Savari and Mrs. Şirin Savari, has always been a vital source of sustenance for me. The affection and humor of my sister Phylis have enriched my life. I dedicate my thesis to the three of them with my love.

Contents

Chapter 1

Introduction

The construction of good variable-to-fixed length codes is an important problem in data compression; the performance of this family of codes is not well-understood for sources with memory. A variable-to-fixed length coder can be decomposed into a parser and a string encoder. The parser segments the source output into a concatenation of variable-length strings, each of which belongs to a dictionary with M entries. For example, suppose that we have a binary source and our dictionary consists of the strings {0, 100, 101, 11}. If the source sequence is 0 1 1 0 1 0 1 0 1 0 0 1 1 0 $\cdots$, the parser would divide it into the sequence of strings (0)(11)(0)(101)(0)(100)(11)(0) $\cdots$. The string encoder maps each dictionary entry into a fixed-length codeword.

1.1 Definitions and Background

A Markov source with finite alphabet $\{0, \ldots, K-1\}$ and set of states $\{0, \ldots, R-1\}$ is defined by specifying, for each state s and letter j,

1. the probability $p_{s,j}$ that the source emits j from state s
2. the next state $S[s,j]$ after j is emitted from state s.

In the underlying Markov chain, let $f_{s,r}$ denote the transition probability from state s to state r; then $f_{s,r} = \sum_{j:S[s,j]=r} p_{s,j}$. We assume the source has a single recurrent class of states; i.e., the underlying Markov chain has a single recurrent class of states or, equivalently, for each pair

of states s and r, there is a string σ for which $P(\sigma|s) > 0$ and $S[s,\sigma] = r$; i.e., σ has positive probability and drives the source to state r. Let π_s denote the steady state probability that the source is in state s and let $\mathcal{H}(s)$ represent the entropy, in natural units, of the next source symbol given that the source is in state s. If $\mathbf{F} = [f_{s,r}]$, $\pi = (\pi_0 \ \ldots \ \pi_{R-1})$, and $\mathbf{e}$ denotes the column vector of length R consisting of all ones, then π and $\mathcal{H}(s)$ are given by

$$\pi = \pi \cdot \mathbf{F} \tag{1.1}$$

$$\pi \cdot \mathbf{e} = 1 \tag{1.2}$$

$$\mathcal{H}(s) = -\sum_{j=0}^{K-1} p_{s,j} \ln p_{s,j}, \qquad s \in \{0 \ldots R-1\}. \tag{1.3}$$

The entropy of the source, $\mathcal{H}$, is then

$$\mathcal{H} = \sum_{s=0}^{R-1} \pi_s \mathcal{H}(s). \tag{1.4}$$

The rules defining the Markov source inductively specify the probability and the final state after an arbitrary source string σ is output from an initial state ψ_0; we let $P(\sigma|\psi_0)$ and $S[\psi_0,\sigma]$ denote this probability and state, respectively. The class of sources that can be modelled in this manner is fairly general and includes, for each $l \geq 1$, the family of sources for which each output, conditional on the l previous output symbols, is independent of all earlier symbols.

We also assume that the output of the source is encoded into a sequence of letters from a D-ary channel alphabet. More specifically, we are initially interested in one-to-one mappings of *strings* of source symbols, or dictionary entries, to the set of D-ary channel strings of length n; it is not necessary for any two distinct dictionary entries to consist of the same number of source symbols.

In [Sha48], Shannon established that the minimum number of code symbols per source symbol that can be achieved by any source coding technique is lower bounded by $\frac{\mathcal{H}}{\ln D}$. There are well-known families of coding procedures such as fixed-to-variable length codes, e.g., the Huffman code (see [Huf52]), and arithmetic codes (see [SG94]) for which the average number of code symbols per source symbol comes arbitrarily close to Shannon's entropy bound. Fixed-to-variable length codes encode each fixed length block of source symbols into a variable-length

string of code symbols, and the compression obtained by the Huffman algorithm approaches the entropy as the block size increases. Arithmetic coding does not follow the approach of mapping a given source string into a fixed code string. Instead, a sequence of source symbols is represented by a point on the unit interval and the corresponding code sequence is the base D representation of that point. For infinite length source sequences, ideal arithmetic codes achieve compression at the entropy bound. Then why should we be interested in variable-to-fixed length codes??

Intuitively, variable-to-fixed length codes seem likely to have a great potential for exploiting the statistical dependencies of the source output when dictionaries are chosen so that the long entries occur frequently and the short entries occur less often. For small to moderate dictionary sizes and sources with very predictable output, this flexibility in the choice of source dictionaries will often lead to better compression than that of the Huffman code designed for a comparably sized dictionary. For example (see [BCW90]), in the late eighteenth century, the British Admiralty used shutter telegraphs as an efficient way to send messages between London and the naval stations on the coast. This telegraph employed a series of cabins on hilltops; each cabin had six large shutters on its roof and a message was transmitted when a pattern on a set of shutters in London was set up. Each of the six shutters could be opened or closed, giving a total of sixty-four possible patterns. Since there were fewer than sixty-four characters to be encoded, the remaining patterns were used to represent common words and phrases such as "and," "the," "Portsmouth," "West," and "Sentence of court-martial to be put into execution." Note that the last of the these phrases has fifty-one characters and is represented by only six bits. For a fixed-length code to realize similar compression, the dictionary of source strings would probably have to be astronomical in size.

It is harder to compare variable-to-fixed length codes with arithmetic codes. In practice, arithmetic coding does not achieve the entropy bound because it requires computations that can be performed with only finite, rather than infinite, precision. Currently, the biggest shortcoming to arithmetic codes is that there is no method that is both theoretically and practically sound to extend their usefulness to sources with unknown memory. In contrast, the well-known Lempel-Ziv codes (see [ZL77] and [ZL78]) can be viewed as adaptive variable-to-fixed length codes. A better understanding of variable-to-fixed length codes for sources with memory may help

provide additional insight into the design of good universal codes.

Compared with fixed-length codes and arithmetic codes, there has been relatively little work on variable-to-fixed length codes. In the case of memoryless sources, runlength codes were the earliest variable-length codes to be designed and these codes have long been recognized to be particularly well-suited for binary sources with small entropies (see [SW49, p. 33] and [Eli55]). Tunstall [Tun67] considered the problem of generating an optimal variable-to-fixed length code for any discrete, memoryless source. In the next chapter, we will describe the Tunstall procedure. Khodak (see [Kho69] and [Kri94]) independently found the same algorithm. A summary of other work in this area can be found in [TW87].

A variable-to-fixed length encoding procedure for Markov sources was described and studied in [TW87]. The codes we will consider in Chapters 2 and 3 are different from those in [TW87] and our analysis is more complete than the ones there.

1.2 Goals of this Work

The body of this thesis consists of four parts. The next chapter

- introduces Tunstall codes and illustrates the difficulty in extending them to sources with memory,
- defines a broader class of codes, known as "generalized variable-to-fixed length codes" and the special case of "generalized Tunstall codes," and
- explains the role of renewal theory in understanding the asymptotic performance of generalized Tunstall codes and analyzes the exact form of the convergence of its redundancy from Shannon's entropy bound.

Chapter 3 contains a statement and proof of "the conservation of entropy" and uses it with the techniques of the previous chapter to develop better generalized variable-to-fixed length codes.

The fourth chapter describes LZ '78, i.e., the Lempel-Ziv incremental parsing rule of [ZL78], the Welch variation LZW, and a new modification to LZW. LZ '78 is often viewed as a universal version of the Tunstall code. For each of the three algorithms, we use renewal theory to upper

bound the number of phrases in the parsing of a string and the number of binary digits used to encode the string given the self-information of the string. Furthermore, we demonstrate that the redundancy of encoding the first n letters of the source output is $O\left(\frac{1}{\ln n}\right)$, and we upper bound the exact form of convergence.

In Chapters 2 and 3, we study dictionaries with the property that any source sequence has a unique prefix in the dictionary. Under this assumption, Tunstall codes are the optimal variable-to-fixed length codes for discrete, memoryless sources. In Chapter 5, we remove this constraint on the dictionaries and show an example where the Tunstall code is no longer optimal. We also analyze a code inspired by LZW.

Chapter 2

Generalized Tunstall Codes for Sources with Memory

2.1 The Tunstall Code and the Parsing Problem

We restrict our attention to dictionaries that are *uniquely parsable*; i.e., every source string, even those of zero probability, can be uniquely parsed into a concatenation of dictionary entries with a final string that is a non-null prefix of a dictionary entry. As an example, consider a ternary source.

- If the dictionary is {00, 1, 2}, it is not uniquely parsable because any source string beginning with the letters 0 1 cannot be parsed.
- The dictionary {00, 01, 02, 1, 2}, is uniquely parsable.
- However, adding the string 000 to the previous dictionary results in a new dictionary that is not uniquely parsable because the string 0 0 0 1 can either be segmented as (000)(1) or as (00)(01). We say that this dictionary is *plurally parsable.*

It is often convenient to picture the entries of a dictionary as the leaves of a rooted tree in which the root node corresponds to the null string, each edge is a source alphabet symbol, and each dictionary entry corresponds to the path from the root to a leaf. The tree corresponding to a uniquely parsable dictionary is complete in the sense that every intermediate node in the

tree has a full set of edges coming out of it; i.e., any single letter extension of a string that is a proper prefix of a dictionary entry is either a dictionary entry or a proper prefix of a dictionary entry.

If the dictionary contains M entries, unique parsability implies that $M = \alpha(K-1)+1$ for some integer α; here, α is the number of intermediate nodes in the dictionary tree, including the root. Efficiency also requires the number of strings to be as large as possible subject to $M \leq D^n$; i.e., M lies in the range $D^n - (K-2) \leq M \leq D^n$.

For the special case of a discrete, memoryless source, Tunstall found a very simple algorithm to construct a uniquely parsable dictionary that maximizes the expected number, $E[L]$, of source letters per dictionary string. To justify this criterion for building a dictionary, we note that as the length of the encoded source string increases, the number of code letters per source letter approaches $\frac{n}{E[L]}$ with probability 1. Since $E[L]$ is the expected length of the dictionary tree, it is straightforward to show by induction on the number of intermediate nodes that $E[L]$ is the sum of the probabilities associated with each intermediate node in the tree, including the root. Therefore, an optimal uniquely parsable dictionary will correspond to a set of intermediate nodes with maximal probabilities. Hence, the Tunstall algorithm given below finds the optimal uniquely parsable dictionary for a discrete, memoryless source:

1. Start with each source symbol as a dictionary entry.

2. If the total number of entries is less than $D^n - (K-2)$, then goto step 3, else stop.

3. Take the most probable entry σ and replace it with the K strings which are single letter extensions of σ. Do not alter the other entries. Goto step 2.

A discussion of Tunstall's algorithm and its performance appears in [JL75].

The preceding algorithm is known to maximize $E[L]$ only for the case of a memoryless source. The paramount obstacle in trying to generalize the algorithm to sources with memory lies in the insensitivity of the algorithm to the state probabilities at parsing points. More precisely, the probability of a dictionary entry, starting at a parsing point, depends on the state probabilities at parsing points, which in turn depend on the dictionary itself. In particular, as we will see in the following example, the parsing points are not guaranteed to leave the source in steady state. The expected length of a dictionary entry is generally a complex expression

because of the interdependence between the choice of dictionary and the state probabilities at parsing points.

Example 2.1: Consider a binary source where the state at any time is the last binary digit emitted. Suppose our dictionary is $\{0, 10, 11\}$ and define ρ_r as the steady-state probability of a parsing point after digit r, $r \in \{0, 1\}$. The following equations specify the interdependence between the probabilities of the dictionary entries and the state probabilities at parsing points.

$$\begin{aligned} \mathrm{P}(0) &= \mathrm{P}(0|0) \cdot \rho_0 + \mathrm{P}(0|1) \cdot \rho_1 && (2.1) \\ \mathrm{P}(10) &= \mathrm{P}(10|0) \cdot \rho_0 + \mathrm{P}(10|1) \cdot \rho_1 && (2.2) \\ \mathrm{P}(11) &= 1 - \mathrm{P}(0) - \mathrm{P}(10) && (2.3) \\ \rho_0 &= \mathrm{P}(0) + \mathrm{P}(10) && (2.4) \\ \rho_1 &= 1 - \rho_0. && (2.5) \end{aligned}$$

Let us consider a special case of this source in which the state-transition probability of a self-transition (both 0 to 0 and 1 to 1) is 0.99. In this case, the steady-state probability that the source will emit a zero or a one at any time is 0.5. What is the steady-state behavior of the parsing procedure on the source output? Intuitively, the dictionary entry 10 will rarely be used. Note that the source output will alternate between runs of zeroes and runs of ones. Since the distribution of the lengths of the runs of zeroes is the same as the distribution of the lengths of runs of ones, and since the string 11 is twice as long as the string 0, we would expect that the string 0 is used about twice as often as the string 11. Consequently, $\rho_0 \approx \mathrm{P}(0) \approx 2/3$ and $\rho_1 = \mathrm{P}(11) \approx 1/3$. To determine the exact solution, we can use (2.1) to (2.5) and find that

$$\mathrm{P}(0) = \frac{198}{298},\ \mathrm{P}(10) = \frac{1}{298},\ \mathrm{P}(11) = \frac{99}{298},\ \rho_0 = \frac{199}{298},\ \rho_1 = \frac{99}{298}.$$

For this example, the steady-state probability distribution differs from the probability distribution of states at parsing points. The steady-state expected length of a dictionary

entry is

$$E[L] = \mathrm{P}(0) + 2 \cdot \mathrm{P}(10) + 2 \cdot \mathrm{P}(11) = \frac{398}{298}.$$

For fixed-to-variable length codes for Markov sources, these parsing issues arise only for periodic sources. In this case, it is generally a good idea to choose the block length to be an integral multiple of the period in order to take advantage of the statistical dependencies of the source.

2.2 Generalized Tunstall Codes for Markov Sources

For general Markov sources, it appears to be difficult to find the optimal uniquely parsable dictionary of a given size. Therefore, to better understand encoding techniques for sources with memory, we will consider a family of codes that is more general than the class of variable-to-fixed length codes. We now assume that there is a uniquely parsable dictionary of size M associated with *each* state s; note, however, that not all M entries of a uniquely parsable dictionary necessarily have positive probability. Let $u^{(k)}$ be the subsequence of the source sequence starting with the k^{th} symbol emitted; $u^{(1)}$ is then the entire source output sequence. We apply the following encoding procedure to the source sequence:

1. Let $k = 1$, $i = 0$ and ψ_0 be the state of the source before any output is issued.

2. Look at the dictionary for state ψ_i and find the unique entry σ_i that is a prefix of $u^{(k)}$. σ_i has length $l(\sigma_i)$. The dictionary will specify the n-letter codeword to be emitted corresponding to state ψ_i and string σ_i.

3. $\psi_{i+1} = S[\psi_i, \sigma_i]$. $k \leftarrow k + l(\sigma_i)$. $i \leftarrow i + 1$. Goto step 2.

Note that any variable-to-fixed length code is a member of this larger class of codes with each state having the same dictionary.

To gain greater insights about the construction of good codes, we will take advantage of the tools of dynamic programming and renewal theory. We say that a collection of R uniquely parsable dictionaries of size M employed by the encoding procedure is a *policy* of size M. For any policy A, let $\mathcal{D}_s^A$ and $\mathcal{L}_s^A$ represent the dictionary corresponding to state s and the expected

number of source symbols in a dictionary entry when dictionary $\mathcal{D}_s^A$ is used, respectively. Then

$$\mathcal{L}_s^A = \sum_{\sigma \in \mathcal{D}_s^A} l(\sigma) P(\sigma|s). \tag{2.6}$$

For policy A, let $\mathbf{Q}^A = [q_{s,r}^A]$ be the transition probability matrix for the state of the source from one parsing point to the next. Then for $s, r \in \{0, \ldots, R-1\}$,

$$q_{s,r}^A = \sum_{\sigma \in \mathcal{D}_s^A : S[s,\sigma] = r} P(\sigma|s). \tag{2.7}$$

If the Markov chain corresponding to the transition probability matrix $\mathbf{Q}^A$ has a single recurrent class of states, then the steady state probability ρ_r^A of being in state r at a parsing point in the source sequence is given by the set of equations

$$\begin{cases} \rho_r^A = \sum_{s=0}^{R-1} \rho_s^A q_{s,r}^A, & r = 0, \ldots, R-1 \\ \sum_{r=0}^{R-1} \rho_r^A = 1. \end{cases} \tag{2.8}$$

If the Markov chain associated with $\mathbf{Q}^A$ has more than one recurrent class of states, then with probability 1, the chain will eventually enter and remain in one of these recurrent classes of states, say Γ. For any given recurrent class of states, say Γ, the steady state probability $\rho_r^A(\Gamma)$ of being in state r given that the chain is in the class of states Γ is specified by the following revision of (2.8).

$$\begin{cases} \rho_r^A(\Gamma) = 0, & r \notin \Gamma \\ \rho_r^A(\Gamma) = \sum_{s=0}^{R-1} \rho_s^A(\Gamma) q_{s,r}^A, & r \in \Gamma \\ \sum_{r=0}^{R-1} \rho_r^A(\Gamma) = 1. \end{cases} \tag{2.9}$$

Hence, if the Markov chain enters the single recurrent class Γ, the steady state expected length of a dictionary entry for policy A is

$$E[L^A(\Gamma)] = \sum_{r=0}^{R-1} \rho_r^A(\Gamma)\ \mathcal{L}_r^A \tag{2.10}$$

symbols, and as the length of the source string encoded tends to infinity, the strong law of large numbers (see [Chu60]) implies that the number of code letters per source symbol for policy A

approaches $\frac{n}{E[L^A(\Gamma)]}$ with probability 1. We define the *(best-case) code length* of policy A by

$$E[L^A] = \max_{\Gamma} E[L^A(\Gamma)]. \tag{2.11}$$

Since M, the dictionary size for each state, satisfies $D^n-(K-2) \leq M \leq D^n$, we will concentrate on evaluating $\frac{\ln M}{E[L^A]}$.

We will focus our attention upon the *generalized Tunstall policy*, denoted policy T. The generalized Tunstall policy is the policy that maximizes $\mathcal{L}_s$ for each $s \in \{0, \ldots, R-1\}$. It is straightforward to demonstrate that for policy T, the dictionary for state s can be constructed by using the Tunstall algorithm with the modification that the probability of an intermediate node corresponding to string σ is $P(\sigma|s)$. Policy T is not guaranteed to be optimal because the corresponding probability distribution of states at parsing points, and hence the relative frequency with which the various dictionaries are used, may not be ideal. However, we are going to show that as M increases, the generalized Tunstall policy becomes *asymptotically optimal* in the sense that for any M, the code length of policy T differs from the maximum achievable best-case code length by at most a constant that is independent of M. Hence, by selecting M sufficiently large, the number of code letters per source symbol for policy T, $\frac{\ln M}{E[L^T]}$ comes arbitrarily close to the minimum among policies of the same size.

We begin by stating more precisely the notion of a policy being close to optimal.

Definition 2.1 *In a class of policies $\mathcal{C}$, policy $B \in \mathcal{C}$ is said to be* ϵ-optimal *if $E[L^B] \geq E[L^A]-\epsilon$ for all policies $A \in \mathcal{C}$.*

We have the following result.

Lemma 2.1 *Let $\mathcal{C}$ be the class of policies of a given size. If the Markov chain corresponding to the transition probability matrix $\mathbf{Q}^T$ for the generalized Tunstall policy is recurrent and*

$$\epsilon = \max_{s \in \{0,\ldots,R-1\}} \mathcal{L}_s^T - E[L^T],$$

then policy T is ϵ-optimal for $\mathcal{C}$.

Proof: The definition of the Tunstall policy and (2.9) to (2.11) imply that for all $A \in \mathcal{C}$,

$$E[L^A] \leq \max_{s\in\{0,\ldots,R-1\}} \mathcal{L}_s^A \leq \max_{s\in\{0,\ldots,R-1\}} \mathcal{L}_s^T = \epsilon + E[L^T]. \qquad \square$$

Lemma 2.1 indicates that in order to show that the generalized Tunstall policy is asymptotically optimal, it is sufficient to establish that as M increases, the Markov chain associated with $\mathbf{Q}^T$ is recurrent and the expression $\max_{s\in\{0,\ldots,R-1\}} \mathcal{L}_s^T - E[L^T]$ is bounded by a constant which is independent of the dictionary size. Toward this end, we will use renewal theory to study $\mathbf{Q}^T$ and $\mathcal{L}^T$ when the dictionary size is very large.

To understand this problem, we investigate how the source generates self-information. We can model the generation of self-information as a *renewal process.* In a renewal process, renewals occur at randomly chosen epochs, and successive interrenewal periods, i.e., the intervals between renewals, are independent and identically distributed random variables. Renewal processes are conventionally used to describe processes that evolve in time; for our purposes, self-information plays the role of time. We choose the interrenewal periods to represent the self-information generated by the source between successive returns to some given state ψ; an epoch can be then be interpreted as the self-information of the source string upon an entrance of the source into state ψ. Because of the Markovian nature of the source, these interrenewal periods are independent random variables and all but the first interrenewal period are also identically distributed. To be more precise, if the source is initially in state ψ, the stochastic process defined above is a renewal process; otherwise, it is a *delayed* renewal process. For each state ψ and integer $k \geq 2$, we let $J_k^{(\psi)}$ symbolize the self-information, in natural units, generated by the source between the $k-1^{\text{st}}$ and the k^{th} occurrences of state ψ; $J_1^{(\psi)}$ denotes the self-information produced until the source reaches state ψ for the first time. Let $T_k^{(\psi)} = J_1^{(\psi)} + \ldots + J_k^{(\psi)}$, let $\{N^{(\psi,\psi_0)}(t); t \geq 0\}$ be the renewal or delayed renewal process defined, for each state ψ, and starting state ψ_0, by specifying the random variable $N^{(\psi,\psi_0)}(t)$ as the number of renewals until the self-information reaches t, i.e., the largest integer k for which $T_k^{(\psi)} \leq t < T_{k+1}^{(\psi)}$, and let $m^{(\psi,\psi_0)}(t) = E[N^{(\psi,\psi_0)}(t)]$.

Suppose τ_s is the self-information of the last intermediate node chosen for the dictionary tree for state s. For any string σ and letter j, define the string $\sigma \circ j$ as the string formed by

appending j to the string σ. For any real number x, let $[x]_+$ denote the positive part of x; i.e., $[x]_+ = \max\{x, 0\}$. We have the following result:

Lemma 2.2 *For the Tunstall dictionary for state s, consider all pairs of states ψ and symbols j. For each string σ such that $S[s,\sigma] = \psi$ and the self-information of σ, $I(\sigma|s) = -\ln P(\sigma|s)$, is in the interval $([\tau_s + \ln p_{\psi,j}]_+, \tau_s)$, the string $\sigma \circ j$ is an entry of the dictionary; conversely, if $\sigma \circ j$ is in the dictionary, then $I(\sigma|s) \in [[\tau_s + \ln p_{\psi,j}]_+, \tau_s]$.*

The convention for the null string $\emptyset$ is that its self-information is zero and $S[s,\emptyset] = s$.

Proof: Suppose $I(\sigma|s) \in ([\tau_s + \ln p_{\psi,j}]_+, \tau_s)$. Then σ must correspond to an intermediate node in the dictionary tree; if this were not the case, then σ or some prefix of σ, say σ', would be an entry in the dictionary with $I(\sigma'|s) \leq I(\sigma|s) < \tau_s$, contradicting the construction of the dictionary. Since σ corresponds to an intermediate node, $\sigma \circ j$ is either a dictionary entry or a proper prefix of one. $\sigma \circ j$ is a dictionary entry because it is less probable than the string corresponding to the last intermediate node chosen; i.e., $I(\sigma \circ j|s) = I(\sigma|s) + I(j|S[s,\sigma]) = I(\sigma|s) - \ln p_{\psi,j}$, which is in the interval $(\tau_s, \tau_s - \ln p_{\psi,j})$. Conversely, every dictionary entry can be represented in the form $\sigma \circ j$ for some string σ and some symbol j with $I(\sigma \circ j|s) \geq \tau_s$. Hence, $I(\sigma|s) \geq [\tau_s + \ln p_{S[s,\sigma],j}]_+$. We also have that $I(\sigma|s) \leq \tau_s$ since σ is at least as probable as the last intermediate node selected. Consequently, $I(\sigma \circ j|s) \leq \tau_s - \ln p_{S[s,\sigma],j}$. □

To find $q_{s,r}^T$ and the relationship between τ_s and M, we use Lemma 2.2 to characterize the strings ϕ in the dictionary for state s that drive the source to state r. For each such string ϕ, there exists a string σ, a symbol j and a state ψ such that $\phi = \sigma \circ j$, $\psi = S[s,\sigma]$, $r = S[\psi, j]$ and $I(\sigma|s) \in [[\tau_s + \ln p_{\psi,j}]_+, \tau_s]$; here, $P(\phi|s) = P(\sigma|s) \cdot p_{\psi,j}$. Note that each such σ corresponds to a renewal in the process $\{N^{(\psi,s)}(t)\}$; the expected number of renewals in the interval $(t, t+dt]$ is

$$m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t) = \sum_{\sigma: I(\sigma|s) \in (t, t+dt],\ i = S[s,\sigma]} P(\sigma|s). \tag{2.12}$$

We will first assume that the self-information corresponding to the source symbols issued has a *non-arithmetic* distribution; i.e., there is no constant Λ such that $-\ln p_{\psi,j}$ is an integer multiple

of Λ for all pairs of states ψ and symbols j such that $p_{\psi,j} > 0$. Lemma 2.2 implies that

$$\sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \int_{([\tau_s+\ln p_{\psi,j}]_+)^+}^{\tau_s^-} p_{\psi,j}\, dm^{(\psi,s)}(x) \;\leq$$

$$q_{s,r}^T \;\leq\; \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \int_{([\tau_s+\ln p_{\psi,j}]_+)^-}^{\tau_s^+} p_{\psi,j}\, dm^{(\psi,s)}(x). \quad (2.13)$$

For each string σ with $I(\sigma|s) \in (t, t+dt]$, $e^{-t-dt} \leq P(\sigma|s) < e^{-t}$. Therefore, it follows from (2.12) that the number of strings σ with $I(\sigma|s) \in (t, t+dt]$ and $\psi = S[s,\sigma]$ is in the interval $(e^t[m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t)], e^{t+dt}[m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t)]]$. Hence,

$$\sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \int_{([\tau_s+\ln p_{\psi,j}]_+)^+}^{\tau_s^-} e^x\, dm^{(\psi,s)}(x) \;\leq\; M \;\leq\; \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \int_{([\tau_s+\ln p_{\psi,j}]_+)^-}^{\tau_s^+} e^x\, dm^{(\psi,s)}(x). \quad (2.14)$$

In Appendix A, we establish the following result.

Theorem 2.1 *If the self-information associated with the source symbols emitted has a non-arithmetic distribution, then for all sufficiently large dictionaries, the Markov chain corresponding to* $\mathbf{Q}^T$ *is recurrent. In particular, as the number of entries for the dictionary for each state increases,*

$$q_{s,r}^T \;\longrightarrow\; \rho_r^* \doteq \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \frac{-\pi_\psi p_{\psi,j} \ln p_{\psi,j}}{\mathcal{H}}, \qquad r,s \in \{0 \ldots R-1\} \quad (2.15)$$

$$\tau_s - \ln M \;\longrightarrow\; \ln\left(\frac{\mathcal{H}}{K-1}\right), \qquad s \in \{0 \ldots R-1\}. \quad (2.16)$$

From (2.15), it is clear that as $M \to \infty$, the steady state parsing probabilities ρ_r^T approach

$$\rho_r^T \;\longrightarrow\; \rho_r^*. \quad (2.17)$$

(2.16) implies the following fundamental relationship between the entropy of the source and the number of strings with a bounded self-information.

Corollary 2.1 *For a non-arithmetic Markov source,*

$$\lim_{\tau_s \to \infty} \mathcal{H} e^{-\tau_s} \cdot |\{\sigma : I(\sigma|s) < \tau_s\}| \;=\; \lim_{\tau_s \to \infty} \mathcal{H} e^{-\tau_s} \cdot |\{\sigma : I(\sigma|s) \leq \tau_s\}| \;=\; 1,\; s \in \{0, \ldots, R-1\},$$

where $|\chi|$ *denotes the cardinality of set* χ.

Proof: (2.16) is equivalent to

$$\lim_{\tau_s \to \infty} \mathcal{H} e^{-\tau_s} \cdot \frac{M}{K-1} \;=\; 1,\; s \in \{0, \ldots, R-1\}. \tag{2.18}$$

Let α be the number of intermediate nodes in the tree. Recall, then, that

$$\alpha \;=\; \frac{M-1}{K-1}.$$

Hence, (2.18) implies that

$$\lim_{\tau_s \to \infty} \mathcal{H} e^{-\tau_s} \alpha \;=\; 1,\; s \in \{0, \ldots, R-1\}. \tag{2.19}$$

By definition of τ_s and the construction of the Tunstall tree, we know that every string with self-information less than τ_s corresponds to an intermediate node in the tree. Furthermore, since intermediate nodes are added one at a time, the number of intermediate nodes with self-information τ_s is between 1 and $|\{\sigma : I(\sigma|s) = \tau_s\}|$. Thus,

$$|\{\sigma : I(\sigma|s) < \tau_s\}| + 1 \;\leq\; \alpha \;\leq\; |\{\sigma : I(\sigma|s) \leq \tau_s\}| \tag{2.20}$$

The result follows from (2.19) and (2.20). □

Next, we will briefly consider the situation in which the self-information corresponding to the source symbols issued has an *arithmetic* distribution with period Λ. Observe that step 3 of the Tunstall procedure does not specify how to choose the dictionary entry σ when the most probable entry is not unique. For non-arithmetic distributions, Theorem 2.1 holds for any order in which the most probable dictionary leaves are converted to intermediate nodes. Unfortunately, Theorem 2.1 is not true in general for arithmetic distributions. However, there are some analogous results that we present in Appendix B. For the remainder of the chapter,

we consider only non-arithmetic distributions.

We next investigate the asymptotic behavior of $\mathcal{L}_s^T$. Let H_s denote the entropy of the entries in the dictionary for state s. Just as we can define a renewal process where the inter-renewal variable is the self-information generated by the source, we can view the process by which the source generates self-information as a semi-Markov process with the properties that whenever the source enters state ψ:

1. The next state it will enter is state r with probability $f_{\psi,r}$.
2. Given that the next symbol to be emitted is letter j, the amount of self-information generated until the transition from ψ to $S[\psi, j]$ is $-\ln p_{\psi,j}$.

The self-information of a sample dictionary entry can be interpreted as the first transition after τ_s in the semi-Markov process. We'll use the following result from [Ros83, §4.8] to relate H_s and τ_s:

Lemma 2.3 *Let $Z(t)$ denote the state of the source at the last transition for which the self-information is at most t. Let $Y(t)$ represent the information growth from t until the next transition. If G_ψ represents the distribution of information growth that the semi-Markov process produces in state ψ before making a transition, then*

$$\lim_{t\to\infty} Prob\{Z(t) = \psi,\ Y(t) > x\} \ = \ \frac{\pi_\psi}{\mathcal{H}} \int_x^\infty [1 - G_\psi(y)]dy.$$

From Lemma 2.3, we can deduce the following theorem:

Theorem 2.2 *As the dictionary size increases,*

$$H_s - \tau_s \ \longrightarrow \ \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}}.$$

Proof: Since we view the self-information of a dictionary entry as the first transition epoch after τ_s in the semi-Markov process, we have that

$$H_s = \tau_s + E[Y(\tau_s)].$$

Using the preceding lemma, we have that

$$\begin{aligned}
\lim_{\tau_s \to \infty} E[Y(\tau_s)] &= \lim_{\tau_s \to \infty} \int_0^\infty \text{Prob}\{Y(\tau_s) > x\} dx \\
&= \sum_{\psi=0}^{R-1} \frac{\pi_\psi}{\mathcal{H}} \int_0^\infty \int_x^\infty [1 - G_\psi(y)]\, dy\, dx \\
&= \sum_{\psi=0}^{R-1} \frac{\pi_\psi}{\mathcal{H}} \int_0^\infty [1 - G_\psi(y)] \cdot y\, dy \\
&= \sum_{\psi=0}^{R-1} \frac{\pi_\psi}{\mathcal{H}} \left(\frac{y^2}{2} [1 - G_\psi(y)]|_0^\infty - \int_0^\infty \frac{y^2}{2}\, d[1 - G_\psi(y)] \right), \text{by integration by parts} \\
&= \sum_{\psi=0}^{R-1} \frac{\pi_\psi}{\mathcal{H}} \int_0^\infty \frac{y^2}{2}\, dG_\psi(y).
\end{aligned}$$

To complete the proof, we observe that $\int_0^\infty y^2\, dG_\psi(y)$ is the second moment of the time spent in state ψ during a visit to that state. □

An immediate consequence of Theorems 2.1 and 2.2 is

Corollary 2.2 *As the number of entries for the dictionary for each state increases,*

$$H_s - \ln M \longrightarrow \ln\left(\frac{\mathcal{H}}{K-1}\right) + \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}},\ s \in \{0, \ldots, R-1\}.$$

To evaluate the asymptotic behavior of $\mathcal{L}_s^T$, we will use some results from the theory of Markov reward processes. Suppose that for every symbol issued, the source collects a reward equal to the self-information of the symbol. Then from state s, the average reward for the next symbol to be emitted is $\mathcal{H}(s)$ and the steady-state average reward is $\mathcal{H}$. Since the source has a single recurrent class of states, it is known (see [Gal96, §4.5]) that there is a unique vector

$\mathcal{W} = (\mathcal{W}_0, \ldots, \mathcal{W}_{R-1})'$ such that $\mathcal{W}_0 = 0$ and

$$\sum_{r=0}^{R-1} f_{s,r}\mathcal{W}_r \;+\; \mathcal{H}(s) \;=\; \mathcal{H} \;+\; \mathcal{W}_s, \quad s \in \{0, \ldots, R-1\}. \tag{2.21}$$

$\mathcal{W}$ is called the relative reward vector and $\mathcal{W}_\psi$ is interpreted as the asymptotic relative gain in self-information of starting in state ψ relative to state 0. Note that $\mathcal{W}$ is defined in terms of source parameters and is independent of the choice of codes. In Appendix C, we establish the following result.

Lemma 2.4 *As the dictionary size increases, for each state s,*

$$H_s - \mathcal{L}_s^T \cdot \mathcal{H} \longrightarrow \mathcal{W}_s - \sum_{r=0}^{R-1} \rho_r^* \mathcal{W}_r. \tag{2.22}$$

Corollary 2.2 and Lemma 2.4 imply

Theorem 2.3 *As the dictionary size increases, for all states s,*

$$\ln M - \mathcal{L}_s \cdot \mathcal{H} \;\longrightarrow\; \ln\left(\frac{K-1}{\mathcal{H}}\right) - \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j}(-\ln p_{\psi,j})^2}{2\mathcal{H}} - \sum_{r=0}^{R-1} \rho_r^* \mathcal{W}_r + \mathcal{W}_s \tag{2.23}$$

and for all pairs of states r and s,

$$\mathcal{L}_s^T - \mathcal{L}_r^T \;\longrightarrow\; \frac{\mathcal{W}_r - \mathcal{W}_s}{\mathcal{H}}. \tag{2.24}$$

Hence, there is some constant γ such that

$$|\mathcal{L}_s^T - \mathcal{L}_r^T| \;\leq\; \gamma \tag{2.25}$$

for every dictionary size M and for each $r, s \in \{0, \ldots, R-1\}$.

Since $E[L^T] = \sum_{r=0}^{R-1} \rho_r^T \mathcal{L}_r^T$, we see from (2.17) that as $M \to \infty$,

$$E[L^T] - \sum_{\psi=0}^{R-1} \rho_\psi^* \mathcal{L}_\psi^T \;\longrightarrow\; 0. \tag{2.26}$$

Next, we will establish the asymptotic relationship between M and $E[L^T]$. Multiplying both sides of (2.22) by ρ_s^* and summing over s, and then using (2.26) and the fact (from Corollary 2.2) that H_s is independent of s in the limit as $M \to \infty$, we find that for each state s,

$$H_s - E[L^T] \cdot \mathcal{H} \longrightarrow 0. \tag{2.27}$$

From (2.24), (2.26) , and Lemma 2.1, we observe

Theorem 2.4 *In the limit as the dictionary size approaches infinity, the generalized Tunstall policy is* $\frac{1}{\mathcal{H}}(\sum_{r=0}^{R-1} \rho_r^* \mathcal{W}_r - \min_{s\in\{0,\ldots,R-1\}} \mathcal{W}_s)$*-optimal among policies of the same size.*

We have the following results.

Theorem 2.5 *As the number of entries for the dictionary for each state increases,*

$$\ln M - E[L^T] \cdot \mathcal{H} \longrightarrow \ln\left(\frac{K-1}{\mathcal{H}}\right) - \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j}(-\ln p_{\psi,j})^2}{2\mathcal{H}}, \tag{2.28}$$

$$\textit{and so} \qquad \frac{\ln M}{E[L^T]} - \mathcal{H} \longrightarrow 0, \tag{2.29}$$

$$\textit{and} \quad (\ln M) \cdot \left(\frac{\ln M}{E[L^T]} - \mathcal{H}\right) \longrightarrow \mathcal{H} \ln\left(\frac{K-1}{\mathcal{H}}\right) - \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j}(-\ln p_{\psi,j})^2}{2}. \tag{2.30}$$

Proof: The results of Corollary 2.2 and (2.27) imply (2.28). From (2.28), $\lim_{M\to\infty} E[L^T] = \infty$. Hence (2.29) follows from dividing both sides of (2.28) by $E[L^T]$. (2.28), (2.29) and the observation that

$$(\ln M) \cdot \left(\frac{\ln M}{E[L^T]} - \mathcal{H}\right) = \frac{\ln M}{E[L^T]} \cdot (\ln M - E[L^T] \cdot \mathcal{H})$$

imply (2.30). □

(2.30) shows not only that $\frac{\ln M}{E[L^T]} \to \mathcal{H}$ as $\Theta(\frac{1}{\ln M})$, but also gives the exact form of convergence. Even for the special case of a discrete, memoryless source, this result on the asymptotic performance of the Tunstall algorithm is new and comparable asymptotic results for the Huffman code or arithmetic codes do not currently exist. (2.30) demonstrates that variable-to-fixed length codes are particularly effective for very predictable sources, i.e., sources in which the entropy is very small; that result also provides an upper bound to the difference between the

compression obtained by the Tunstall code and the minimum number of code letters per source symbol among policies of the same size, in the limit as the dictionary size approaches infinity. To conclude this section, we will specify a different asymptotic upper bound. Let $E[L^*]$ symbolize the maximum achievable best-case code length among policies of the same size as the Tunstall policy under consideration. From Theorems 2.4 and 2.5, we have the following result.

Theorem 2.6 *As the dictionary size increases, we have that*

$$\limsup_{M\to\infty} (\ln M)\cdot\left(\frac{\ln M}{E[L^T]}-\frac{\ln M}{E[L^*]}\right) \le \mathcal{H}\left(\sum_{r=0}^{R-1}\rho_r^*\mathcal{W}_r - \min_{s\in\{0,\ldots,R-1\}}\mathcal{W}_s\right).$$

Proof: The definition of $E[L^*]$ and Shannon's entropy bound imply that

$$\mathcal{H} \le \frac{\ln M}{E[L^*]} \le \frac{\ln M}{E[L^T]},$$

and hence it follows from (2.29) that as the dictionary size increases

$$\frac{\ln M}{E[L^*]} \longrightarrow \mathcal{H}. \tag{2.31}$$

From Theorem 2.4 and the definition of ϵ-optimality, we have that

$$\limsup_{M\to\infty} E[L^*]-E[L^T] \le \frac{1}{\mathcal{H}}\Big(\sum_{r=0}^{R-1}\rho_r^*\mathcal{W}_r - \min_{s\in\{0,\ldots,R-1\}}\mathcal{W}_s\Big) \tag{2.32}$$

Note that

$$(\ln M)\cdot\left(\frac{\ln M}{E[L^T]}-\frac{\ln M}{E[L^*]}\right) = \frac{\ln M}{E[L^T]}\cdot\frac{\ln M}{E[L^*]}\cdot(E[L^*]-E[L^T]),$$

and so the theorem follows from (2.29), (2.31) and (2.32). □

2.3 Variable-to-Fixed Length Codes for Markov Sources

Now that we have acquired some insights about the performance of good policies, it is appropriate to re-examine the problem which originally interested us, namely, the design of good uniquely parsable dictionaries. We observed earlier that any variable-to-fixed length code can

be viewed as a policy with each state employing the same dictionary. Hence, Theorems 2.5 and 2.6 imply

Theorem 2.7 *As the size of the dictionary increases, the compression ratio, $\frac{\ln M}{E[L]}$, of any variable-to-fixed length code satisfies*

$$\liminf_{M\to\infty} (\ln M)\cdot\left(\frac{\ln M}{E[L]} - \mathcal{H}\right) \geq$$
$$\mathcal{H}\ln\left(\frac{K-1}{\mathcal{H}}\right) - \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1}\frac{\pi_\psi p_{\psi,j}(-\ln p_{\psi,j})^2}{2} + \mathcal{H}\left(\min_{s\in\{0,\ldots,R-1\}}\mathcal{W}_s - \sum_{r=0}^{R-1}\rho_r^*\mathcal{W}_r\right). \quad (2.33)$$

Comment: The right-hand side of (2.33) is positive for the sources we have tested empirically and it is positive for all binary sources with $\mathcal{H}(0) = \mathcal{H}(1) = \cdots = \mathcal{H}(R-1)$.

We are going to construct a variable-to-fixed length code with a compression ratio that is asymptotically close to optimal by merging a collection of Tunstall dictionaries, one for each state s. In particular, choose a set of state weights $\{\beta_0, \ldots, \beta_{R-1}\}$ with the property that $\beta_s \geq 0$ for all s and $\sum_{s=0}^{R-1}\beta_s = 1$. For each s, let

$$M_s = \beta_s \cdot M, \quad (2.34)$$

let $\mathcal{D}_s$ represent the largest Tunstall dictionary with at most M_s entries corresponding to state s, and let $\mathcal{D}'$ denote the union of the entries of $\mathcal{D}_s$ for each s. The dictionary $\mathcal{D}$ associated with our variable-to-fixed length code consists of the entries of $\mathcal{D}'$ that are not proper prefixes of other entries of $\mathcal{D}'$; i.e., the dictionary tree corresponding to $\mathcal{D}$ is the union of the dictionary trees associated with the dictionaries $\{\mathcal{D}_s\}$. $\mathcal{D}$ is uniquely parsable since all of the $\mathcal{D}_s$ are uniquely parsable. Furthermore, $\mathcal{D}$ has at most M entries. To illustrate the construction of the dictionary $\mathcal{D}$, we consider the following example.

Example 2.2: As in Example 1, suppose that we have a binary, Markov source where the state is given by the most recent binary digit emitted and the state-transition probability of a 0 to 0 or a 1 to 1 is 0.99. Suppose $M = 10$ and we set $\beta_0 = \beta_1 = 0.5$. Then $M_0 = M_1 = 5$,

$$\mathcal{D}_0 = \{0000,\ 0001,\ 001,\ 01,\ 1\},\ \mathcal{D}_1 = \{0,\ 10,\ 110,\ 1110,\ 1111\},$$

$$\mathcal{D}' = \{0,\ 0000,\ 0001,\ 001,\ 01,\ 1,\ 10,\ 110,\ 1110,\ 1111\},$$

and $$\mathcal{D} = \{0000,\ 0001,\ 001,\ 01,\ 10,\ 110,\ 1110,\ 1111\}.$$

Observe that $\mathcal{D}$ has eight entries.

As usual, let $\mathcal{L}_s$ denote the expected number of source symbols in an entry of $\mathcal{D}_s$ when the last parsing point left the source in state s. Recall that for all $s \in \{0, \ldots, R-1\}$,

$$\mathcal{L}_s = \sum_{\text{intermediate nodes } \sigma \text{ for } \mathcal{D}} P(\sigma|s) \tag{2.35}$$

$$\text{and so} \quad \mathcal{L}_s \geq \sum_{\text{intermediate nodes } \sigma \text{ for } \mathcal{D}_s} P(\sigma|s). \tag{2.36}$$

Since the right-hand side of (2.36) represents the expected length of a Tunstall dictionary with at most M_s entries corresponding to state s, (2.23) implies that for all $s \in \{0, \ldots, R-1\}$,

$$\limsup_{\tau_s \to \infty} \ln M_s - \mathcal{L}_s \cdot \mathcal{H} \leq \ln\left(\frac{K-1}{\mathcal{H}}\right) - \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}} - \sum_{r=0}^{R-1} \rho_r^* \mathcal{W}_r + \mathcal{W}_s. \tag{2.37}$$

From (2.34) and (2.37), we see that for all $s \in \{0, \ldots, R-1\}$,

$$\limsup_{\tau_s \to \infty} \ln M - \mathcal{L}_s \cdot \mathcal{H} \leq \ln\left(\frac{K-1}{\mathcal{H}}\right) - \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}} - \sum_{r=0}^{R-1} \rho_r^* \mathcal{W}_r + \mathcal{W}_s - \ln \beta_s. \tag{2.38}$$

If ρ_s is the steady-state probability of being in state s at a parsing point, then the expected length of a dictionary entry is $E[L] = \sum_{s=0}^{R-1} \rho_s \mathcal{L}_s$. In general, we cannot evaluate the set of steady-state probabilities. Therefore, we use the bound

$$E[L] \geq \min_{s \in \{0,\ldots,R-1\}} \mathcal{L}_s. \tag{2.39}$$

Hence, (2.38) and (2.39) imply that

$$\limsup_{\tau_s \to \infty} \ln M - E[L] \cdot \mathcal{H} \leq$$
$$\ln\left(\frac{K-1}{\mathcal{H}}\right) - \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}} - \sum_{r=0}^{R-1} \rho_r^* \mathcal{W}_r + \max_{s \in \{0,\ldots,R-1\}} (\mathcal{W}_s - \ln \beta_s). \tag{2.40}$$

The expression $\max_{s\in\{0,\ldots,R-1\}}(\mathcal{W}_s - \ln\beta_s)$ is minimized when the weights are chosen so that for all s, $\mathcal{W}_s - \ln\beta_s$ is a constant independent of s. This implies that

$$\beta_s \;=\; e^{\mathcal{W}_s}\cdot\left(\sum_{r=0}^{R-1} e^{\mathcal{W}_r}\right)^{-1}. \tag{2.41}$$

Combining (2.40) and (2.41), we have that

Theorem 2.8 *For the choice of weights given by (2.41), the asymptotic performance of the dictionary created by merging the R Tunstall dictionaries with sizes determined by (2.34) is*

$$\limsup_{M\to\infty}\,(\ln M - E[L]\cdot\mathcal{H}) \;\le\; \ln\left(\frac{K-1}{\mathcal{H}}\sum_{s=0}^{R-1} e^{\mathcal{W}_s}\right) - \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1}\frac{\pi_\psi p_{\psi,j}(-\ln p_{\psi,j})^2}{2\mathcal{H}} - \sum_{r=0}^{R-1}\rho_r^*\mathcal{W}_r$$

and so

$$\limsup_{M\to\infty}(\ln M)\cdot\left(\frac{\ln M}{E[L]} - \mathcal{H}\right) \le \mathcal{H}\ln\left(\frac{K-1}{\mathcal{H}}\sum_{s=0}^{R-1} e^{\mathcal{W}_s}\right) - \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1}\frac{\pi_\psi p_{\psi,j}(-\ln p_{\psi,j})^2}{2} - \mathcal{H}\sum_{r=0}^{R-1}\rho_r^*\mathcal{W}_r.$$

Theorems 2.7 and 2.8 establish that for large dictionary sizes, this dictionary achieves compression that is arbitrarily close to that of the optimal uniquely parsable dictionary.

Chapter 3

Variable-to-Fixed Length Codes and the Conservation of Entropy

In the preceding chapter, a dictionary was called *uniquely parsable* if every source sequence has exactly one prefix in the dictionary. Under this definition of unique parsability, it is possible for the dictionary to include many strings with zero probability. Since it may be undesirable to construct dictionaries with numerous entries that will never be used, we will weaken the definition of unique parsability and consider a broader class of codes. A dictionary of source strings is now said to be uniquely parsable if every source string with *positive* probability can be uniquely parsed into a concatenation of dictionary entries with a final string that is a non-null prefix of a dictionary entry.

We follow the notation developed in Chapter 2. The parser maintains the procedure for segmenting a source output string described in the last chapter; i.e., it determines the source state s after each parsing point and subsequently uses $\mathcal{D}_s$ to find the next parsed string. However, we no longer require that two dictionaries $\mathcal{D}_r$ and $\mathcal{D}_s$ have the same number of entries. Instead, we assume that each dictionary $\mathcal{D}_s$ has at least $M - K + 2$ and at most M entries.

To make our new notion of unique parsability concrete, we assume that for every state ψ, there is a set of symbols $\tilde{K}(\psi)$ that correspond to the single letter extensions out of state ψ in any dictionary constructed. More precisely, if σ is a proper prefix of an entry of $\mathcal{D}_s$, then the single letter extensions of σ that are either proper prefices of entries or entries of $\mathcal{D}_s$ are

$\{\sigma \circ j : j \in \tilde{K}(S[s,\sigma])\}$. A common choice for $\tilde{K}(\psi)$ is either

- $\tilde{K}(\psi) = \{0, \ldots, K-1\}$ for all ψ or
- $\tilde{K}(\psi) = \{j : p_{\psi,j} > 0\}$ for all ψ.

We let $K_\psi = |\tilde{K}(\psi)|$.

We would like to find a good way to design the dictionaries $\mathcal{D}_s$. As we saw in the last chapter, the expected length $\mathcal{L}_s$ of a dictionary entry for $\mathcal{D}_s$ is given by

$$\mathcal{L}_s = \sum_{\text{intermediate nodes } \sigma \text{ for } \mathcal{D}_s} P(\sigma|s).$$

In Chapter 2, $\mathcal{D}_s$ was chosen to maximize $\mathcal{L}_s$ for each state s. The resulting generalized Tunstall code maximizes the expected number of source symbols per parse for each state, but is not guaranteed to be optimal because it does not necessarily lead to good parsing probabilities.

3.1 The Conservation of Entropy

We will see in the next section that the relationship between the code length and the average self-information between consecutive parses provides insight into finding codes that have a smaller compression ratio than the generalized Tunstall code.

For any choice of dictionaries $\mathcal{D}_s$, let $\mathbf{Q}$ denote the transition probability matrix for the state of the source from one parsing point to the subsequent one. If $\mathbf{Q}$ does not consist of a single recurrent class of states, then the chain will ultimately enter and stay in one of the recurrent classes of states, say Γ.

For a collection of dictionaries resulting in a transition matrix $\mathbf{Q}$ with a single recurrent class of states, let H represent the steady-state average self-information between successive parsing points. For other sets of dictionaries, H is a function of the recurrent class of states entered. We have the following result.

Theorem 3.1 *Let* $\mathbf{Q}$ *symbolize the transition probability matrix associated with the source states at parsing points.*

- *If the corresponding Markov chain has a single recurrent class of states, then*

$$H = \mathcal{H} \cdot E[L].$$

- *Otherwise, if the Markov chain enters the class of states* Γ*, then*

$$H(\Gamma) = \mathcal{H} \cdot E[L(\Gamma)].$$

Proof: Let $\mathcal{I}(k)$ denote the cumulative self-information generated by the source after the k^{th} symbol has been emitted. Let l_ψ symbolize the number of symbols in the ψ^{th} string parsed. Then

$$\frac{\mathcal{I}(\sum_{\psi=1}^{k} l_\psi)}{k} = \frac{\mathcal{I}(\sum_{\psi=1}^{k} l_\psi)}{\sum_{\psi=1}^{k} l_\psi} \cdot \frac{\sum_{\psi=1}^{k} l_\psi}{k}.$$

To complete the proof, we note that the strong law of large numbers (see [Chu60]) implies that with probability 1,

$$\lim_{k\to\infty} \frac{\mathcal{I}(\sum_{\psi=1}^{k} l_\psi)}{k} = H \text{ (or } H(\Gamma)),$$

$$\lim_{k\to\infty} \frac{\mathcal{I}(\sum_{\psi=1}^{k} l_\psi)}{\sum_{\psi=1}^{k} l_\psi} = \mathcal{H},$$

and

$$\lim_{k\to\infty} \frac{\sum_{\psi=1}^{k} l_\psi}{k} = E[L] \text{ (or } E[L(\Gamma)]). \qquad \square$$

We will conclude this section by making a few comments about the theorem and its proof. For memoryless sources, this "conservation of entropy" theorem was established for codes with one uniquely parsable dictionary in [JS72]. The theorem is very general and can be applied to *any* deterministic parsing rule in which the number of dictionaries is finite and the expected length of a parsed phrase is finite. For example, we don't need to assume that the dictionaries are uniquely parsable or have the same number of entries. Hence, the result will also apply for the fixed data base version (see [Wyn93]) of the 1977 Lempel-Ziv algorithm, which is often abbreviated LZ '77 (see [ZL77]). For the sliding window implementation of LZ '77 (see [WZ94]), let ω be the size of the window. If we interpret a window as a dictionary, there are at most K^ω dictionaries and therefore, by redefining the set of states in the right way, the theorem holds

for this parsing procedure as well.

The theorem can be applied to other types of parsing rules. In Lemma A.1 of Appendix A, we established that $E[J_2^{(\psi)}]$, the average self-information generated between successive entrances by the source to state ψ, is equal to $\frac{\mathcal{H}}{\pi_\psi}$. We can provide an alternate proof using the theorem and Kac's lemma (see [Wyn93]). Let L_ψ denote the expected number of symbols generated between consecutive entrances by the source to state ψ. The theorem indicates that $E[J_2^{(\psi)}] = \mathcal{H} \cdot L_\psi$ and Kac's lemma asserts that $L_\psi = \frac{1}{\pi_\psi}$. In fact, for Markov sources, Kac's lemma itself follows from a simple modification to Theorem 3.1. Consider an additive Markov process where the "weight" of any string increases by one every time the source enters state ψ and remains the same whenever the source enters any other state. Then the analogue to Theorem 3.1 would state that the average increase in weight between successive entrances to state ψ, which is one, is equal to the time-average steady-state drift of this process, which is π_ψ, times the average number of symbols between consecutive entrances to state ψ.

In Appendix D, we state and prove a more comprehensive relationship between the self-information and the length of a parsed string. We use this relationship to provide an alternate proof of Theorem 3.1.

3.2 Greedy Variable-to-Fixed Length Codes

Let H_s symbolize the entropy of the entries in $\mathcal{D}_s$. We have the following relationship between H_s and the set of proper prefixes in the dictionary for state s.

Theorem 3.2 *For any uniquely parsable dictionary $\mathcal{D}_s$,*

$$H_s = \sum_{\text{intermediate nodes } \sigma \text{ for } \mathcal{D}_s} P(\sigma|s)\mathcal{H}(S[s,\sigma]),\ 0 \leq s \leq R-1.$$

Theorem 3.2 follows from the "leaf entropy" theorem of [Mas83]; alternatively, it can be demonstrated by means of induction on the number of intermediate nodes in the dictionary tree.

The expressions for $\mathcal{L}_s$ and H_s suggest that we may wish to consider dictionaries that maximize

$$W(s) = \sum_{\text{intermediate nodes } \sigma \text{ for } \mathcal{D}_s} P(\sigma|s) w_{S[s,\sigma]}$$

for some choice of $\mathbf{w} = \{w_0, \ldots, w_{R-1}\}$. w_ψ is called the *weight* of state ψ and $P(\sigma|s)w_{S[s,\sigma]}$ is the *state s reward* of string σ.

A desirable feature of the generalized Tunstall code is that it can be constructed in a *greedy* manner; i.e., for each state s and string σ, the state s reward of σ is at most the state s reward of any proper prefix of σ, so the nodes with the largest state s reward can be selected one by one starting with the null string. A necessary and sufficient condition for a greedy construction is that the weight vector $\mathbf{w}$ satisfies

$$P(\sigma \circ j|s) w_{S[s,\sigma \circ j]} \leq P(\sigma|s) w_{S[s,\sigma]}, \text{ for all states } s, \text{ strings } \sigma, \text{ and source symbols } j.$$

Equivalently, the weight vector must be in the set of greedy vectors

$$\mathcal{G} = \{\mathbf{w} = (w_0, \ldots, w_{R-1}) : \mathbf{w} > \mathbf{0},\ p_{r,j} w_{S[r,j]} \leq w_r,\ \forall\ r,\ j\}.$$

Given that $\mathbf{w} \in \mathcal{G}$, we can use the following greedy procedure to find a dictionary $\mathcal{D}_s$ that maximizes $W(s)$ if $K_s = K_r$ for all states s and r, i.e., if the number of intermediate nodes in the dictionary tree is predetermined, and otherwise approximately maximizes $W(s)$ when the dictionary size is large:

1. Start with each source symbol as a dictionary entry.

2. Find the entry σ that has maximum state s weight. If the number of entries in the dictionary is at most $M - K_{S[s,\sigma]} + 1$, then goto step 3, else stop.

3. Replace σ with the $K_{S[s,\sigma]}$ strings which are single letter extensions of σ. Do not alter the other entries. Goto step 2.

Let policy W denote the resulting set of dictionaries. Note that choosing $K_s = K$ for all s and $\mathbf{w} = (1, 1, \ldots, 1)$ results in the generalized Tunstall policy.

In general, the asymptotic analysis of policy W is very similar to that of the generalized Tunstall policy in Chapter 2 and we will closely follow the steps of that analysis. Let $\mathbf{Q}^W = [q_{s,r}^W]$ denote the transition probability matrix for the state of the source from one parsing point to the next for policy W. Eventually, the Markov chain corresponding to the sequence of states at parsing points will eventually enter and remain in a recurrent class of states, say Γ. Let $\rho_r^W(\Gamma)$ represent the steady-state probability of being in state r at a parsing point given the recursive class of states Γ. Then Theorem 3.1 implies that the steady state expected length of a dictionary entry for policy W is

$$E[L^W(\Gamma)] \;=\; \frac{H^W(\Gamma)}{\mathcal{H}} \;=\; \frac{1}{\mathcal{H}} \sum_{r=0}^{R-1} \rho_r^W(\Gamma)\; H_r^W. \tag{3.1}$$

For the special case when $K_s = K_r$ for all $s, r \in \{0, \ldots, R-1\}$, we will find the asymptotically best greedy policy; in the special case when $\mathbf{w} = \{\mathcal{H}(0), \ldots, \mathcal{H}(R-1)\} \in \mathcal{G}$, we are going to show that as M increases, the corresponding policy, call it policy $*$, is not only the asymptotically best greedy code, but it also becomes *asymptotically optimal* in the sense that its code length differs from the maximum achievable best-case code length by at most a constant that approaches zero.

As in Chapter 2, let $\{N^{(\psi,\psi_0)}(t); t \geq 0\}$ be the renewal or delayed renewal process defined, for each state ψ, and starting state ψ_0, by the number of renewals, i.e., entrances of the source into state ψ, until the self-information generated by the source reaches t, and let $m^{(\psi,\psi_0)}(t) = E[N^{(\psi,\psi_0)}(t)]$.

For any greedy policy, we define the *pseudo-self-information* $\tilde{\tau}(s,\sigma)$ of string σ from starting state s by

$$\tilde{\tau}(s,\sigma) \;=\; -\ln[P(\sigma|s) w_{S[s,\sigma]}] \;=\; I(\sigma|s) - \ln w_{S[s,\sigma]} \tag{3.2}$$

and let $\tilde{\tau}_s$ represent the pseudo-self-information of the last intermediate node chosen for the dictionary tree for state s. We have the following result:

Lemma 3.1 *For each string σ such that $S[s,\sigma] = \psi$ and $\tilde{\tau}(s,\sigma)$ is in the interval $([\tilde{\tau}_s + \ln p_{\psi,j} + \ln w_{S[\psi,j]} - \ln w_\psi]_+, \tilde{\tau}_s)$, the string $\sigma \circ j$ is an entry of dictionary $\mathcal{D}_s$; conversely, if $\sigma \circ j$ is in dictionary $\mathcal{D}_s$, then $\tilde{\tau}(s,\sigma) \in [[\tilde{\tau}_s + \ln p_{\psi,j} + \ln w_{S[\psi,j]} - \ln w_\psi]_+, \tilde{\tau}_s]$.*

Proof: Note that for any state s, string σ, and symbol j,

$$\begin{aligned}\tilde{\tau}(s,\sigma\circ j) &= I(\sigma\circ j|s) - \ln w_{S[s,\sigma\circ j]} \\ &= I(\sigma|s) - \ln p_{\psi,j} - \ln w_{S[\psi,j]} \\ &= \tilde{\tau}(s,\sigma) - \ln p_{\psi,j} - \ln w_{S[\psi,j]} + \ln w_\psi. \end{aligned} \tag{3.3}$$

Given this relationship between $\tilde{\tau}(s,\sigma\circ j)$ and $\tilde{\tau}(s,\sigma)$ and our procedure for constructing $\mathcal{D}_s$, the proof of the lemma is analogous to the proof of Lemma 2.2 found in Chapter 2. □

We next use the preceding result to specify the entries ϕ in $\mathcal{D}_s$ that drive the source to state r. For each such string ϕ, there exists a string σ, a symbol j and a state ψ such that $\phi = \sigma\circ j$, $\psi = S[s,\sigma]$, $r = S[\psi,j]$ and $\tilde{\tau}(s,\sigma) \in [\tilde{\tau}_s + \ln p_{\psi,j} + \ln w_r - \ln w_\psi, \tilde{\tau}_s]$. Consequently, $I(\sigma|s) \in [\tilde{\tau}_s + \ln p_{\psi,j} + \ln w_r, \tilde{\tau}_s + \ln w_\psi]$. Each such σ corresponds to a renewal in the process $\{N^{(\psi,s)}(t)\}$; the expected number of renewals in the interval $(t, t+dt]$ is given by

$$m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t) = \sum_{\sigma: I(\sigma|s)\in(t,t+dt],\ \psi=S[s,\sigma]} P(\sigma|s). \tag{3.4}$$

As we observed in the last chapter, if $I(\sigma|s) \in (t, t+dt]$, then $e^{-t-dt} \le P(\sigma|s) < e^{-t}$. Hence, (3.4) implies that the number of strings σ with $I(\sigma|s) \in (t, t+dt]$ and $\psi = S[s,\sigma]$ is in the interval $(e^t[m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t)], e^{t+dt}[m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t)]]$. For the remainder of this section, we assume that the pseudo-self-information corresponding to the symbols emitted by the source is *non-arithmetic*. We have that

$$\begin{aligned}\sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \int_{([\tilde{\tau}_s+\ln p_{\psi,j}+\ln w_{S[\psi,j]}]_+)^+}^{(\tilde{\tau}_s+\ln w_\psi)^-} e^x\, dm^{(\psi,s)}(x) \ \le \\ M \ \le \ \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \int_{([\tilde{\tau}_s+\ln p_{\psi,j}+\ln w_{S[\psi,j]}]_+)^-}^{(\tilde{\tau}_s+\ln w_\psi)^+} e^x\, dm^{(\psi,s)}(x) \end{aligned} \tag{3.5}$$

In Appendix E, we prove the following theorem.

Theorem 3.3 *If the set* $\{\ln\left(\frac{w_\psi}{p_{\psi,j}w_{S[\psi,j]}}\right) : p_{\psi,j} > 0\}$ *has a non-arithmetic distribution, then as the number of entries for the dictionary for each state increases,*

$$\ln M - \tilde{\tau}_s \longrightarrow \ln\left(\sum_{\psi=0}^{R-1} \frac{\pi_\psi(K_\psi - 1)w_\psi}{\mathcal{H}}\right), \qquad s \in \{0 \ldots R-1\}.$$

We next probe into the relationship between H_s^W and M as the dictionary size approaches infinity. Following the analysis we carried out in Chapter 2, we now regard the generation of pseudo-self-information as a semi-Markov process. Whenever the source enters state ψ,

1. the process will next enter state r with probability $f_{\psi,r}$;
2. letting j be the next source output symbol, the amount of pseudo-self-information generated until the transition from ψ to $S[\psi,j]$ is $-\ln p_{\psi,j} - \ln w_{S[\psi,j]} + \ln w_\psi$.

We can construe the pseudo-self-information of a sample dictionary entry as the first transition after $\tilde{\tau}_s$ in the semi-Markov process. To relate H_s^W and $\tilde{\tau}_s$, we'll apply the following result from [Ros83, §4.8]:

Lemma 3.2 *Let* $S^W(t)$ *denote the state of the source at the first transition for which the pseudo-self-information is at least* t. *Let* $Y(t)$ *represent the pseudo-self-information growth from* t *until the next transition. If* $G_{\psi,r}$ *represents the distribution of pseudo-self-information growth that the semi-Markov process produces in state* ψ *before making a transition given that the transition is into state* r, *then for all* s,

$$\lim_{t\to\infty} Prob\{Y(t) > x,\ S^W(t) = r \mid \psi_0 = s\} = \sum_{\psi=0}^{R-1} \frac{f_{\psi,r}\int_x^\infty [1 - G_{\psi,r}(y)]dy}{J_2^{(\psi)}}.$$

Using this lemma and Lemma A.1, we can prove the following theorem:

Theorem 3.4 *As the dictionary size approaches infinity,*

$$q_{s,r}^W \longrightarrow \rho_r^W \doteq \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j}}{\mathcal{H}} \ln\left(\frac{w_\psi}{p_{\psi,j}w_r}\right) \tag{3.6}$$

$$H_s^W - \tilde{\tau}_s \longrightarrow \sum_{\psi=0}^{R-1} \frac{\pi_\psi \mathcal{H}(\psi) \ln w_\psi}{\mathcal{H}} + \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}}. \tag{3.7}$$

Proof: We begin by establishing (3.6). In Lemma A.1 and earlier in this chapter, we showed that $J_2^{(\psi)} = \frac{\mathcal{H}}{\pi_\psi}$. Therefore, Lemma 3.2 implies that

$$\lim_{t\to\infty} \text{Prob}\{Y(t) > x,\ S^W(t) = r \mid \psi_0 = s\} = \sum_{\psi=0}^{R-1} \frac{\pi_\psi f_{\psi,r}}{\mathcal{H}} \int_x^\infty [1 - G_{\psi,r}(y)]dy. \tag{3.8}$$

Since $q_{s,r}^W = \text{Prob}\{S^W(t) = r \mid \psi_0 = s\}$, we see from (3.8) that

$$q_{s,r}^W \longrightarrow \sum_{\psi=0}^{R-1} \frac{\pi_\psi f_{\psi,r}}{\mathcal{H}} \int_0^\infty [1 - G_{\psi,r}(y)]dy. \tag{3.9}$$

Note that $\int_0^\infty [1 - G_{\psi,r}(y)]dy$ is the mean pseudo-self-information growth that is generated by a source between a transition from state ψ to state r. Hence,

$$\int_0^\infty [1 - G_{\psi,r}(y)]dy = \frac{1}{f_{\psi,r}} \sum_{j:S[\psi,j]=r} p_{\psi,j} \ln \left(\frac{w_\psi}{p_{\psi,j} w_r} \right). \tag{3.10}$$

(3.6) follows from (3.9) and (3.10).

To demonstrate (3.7), we observe that the pseudo-self-information of a dictionary entry is the first transition after $\tilde{\tau}_s$ in the semi-Markov process. Thus, the relationship between the self-information and pseudo-self-information of a string given in (3.2) implies that

$$H_s^W = \sum_{r=0}^{R-1} q_{s,r}^W \left(\tilde{\tau}_s + E[Y(\tilde{\tau}_s) \mid \psi_0 = s,\ S^W(\tilde{\tau}_s) = r] + \ln w_r \right),$$

and using (3.6), we see that

$$H_s^W - \tilde{\tau}_s \longrightarrow \sum_{r=0}^{R-1} \rho_r^W \left(\ln w_r + \lim_{\tilde{\tau}_s \to \infty} E[Y(\tilde{\tau}_s) \mid \psi_0 = s,\ S^W(\tilde{\tau}_s) = r] \right). \tag{3.11}$$

Combining (3.6) and (3.8), we find that

$$\begin{aligned}
\lim_{\tilde{\tau}_s\to\infty} E[Y(\tilde{\tau}_s) \mid \psi_0 = s,\ S^W(\tilde{\tau}_s) = r] &= \int_0^\infty dx \left(\frac{1}{\rho_r^W} \sum_{\psi=0}^{R-1} \frac{\pi_\psi f_{\psi,r}}{\mathcal{H}} \int_x^\infty [1 - G_{\psi,r}(y)]\, dy \right) \\
&= \frac{1}{\rho_r^W} \sum_{\psi=0}^{R-1} \frac{\pi_\psi f_{\psi,r}}{\mathcal{H}} \int_0^\infty [1 - G_{\psi,r}(y)] \cdot y\, dy \\
&= \frac{1}{\rho_r^W} \sum_{\psi=0}^{R-1} \frac{\pi_\psi f_{\psi,r}}{\mathcal{H}} \int_0^\infty \frac{y^2}{2}\, dG_{\psi,r}(y), \text{integrating by parts} \\
&= \frac{1}{\rho_r^W} \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j}}{2\mathcal{H}} \left(\ln \left(\frac{w_\psi}{p_{\psi,j} w_r} \right) \right)^2 \qquad (3.12)
\end{aligned}$$

(3.11) and (3.12) imply

$$\begin{aligned}
H_s^W - \tilde{\tau}_s \longrightarrow & \sum_{r=0}^{R-1} \left(\rho_r^W \ln w_r + \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j}}{2\mathcal{H}} \left(\ln \left(\frac{w_\psi}{p_{\psi,j} w_r} \right) \right)^2 \right) \\
= & \sum_{r=0}^{R-1} \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j}}{2\mathcal{H}} \left(2(\ln w_r) \cdot \ln \left(\frac{w_\psi}{p_{\psi,j} w_r} \right) + \left(\ln \left(\frac{w_\psi}{p_{\psi,j} w_r} \right) \right)^2 \right), \text{ by (3.6)} \\
= & \sum_{r=0}^{R-1} \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j}}{2\mathcal{H}} \left(-2(\ln w_\psi)(\ln p_{\psi,j}) + (\ln p_{\psi,j})^2 + (\ln w_\psi)^2 - (\ln w_r)^2 \right) \\
= & \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j}}{2\mathcal{H}} \left(-2(\ln w_\psi)(\ln p_{\psi,j}) + (\ln p_{\psi,j})^2 \right) \\
& + \sum_{r=0}^{R-1} \sum_{\psi=0}^{R-1} \frac{\pi_\psi f_{\psi,r}}{2\mathcal{H}} \left((\ln w_\psi)^2 - (\ln w_r)^2 \right) \\
= & \sum_{\psi=0}^{R-1} \frac{\pi_\psi \mathcal{H}(\psi) \ln w_\psi}{\mathcal{H}} + \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (\ln p_{\psi,j})^2}{2\mathcal{H}} \\
& + \sum_{\psi=0}^{R-1} \frac{\pi_\psi (\ln w_\psi)^2}{2\mathcal{H}} \sum_{r=0}^{R-1} f_{\psi,r} - \sum_{r=0}^{R-1} \frac{\pi_r (\ln w_r)^2}{2\mathcal{H}} \\
= & \sum_{\psi=0}^{R-1} \frac{\pi_\psi \mathcal{H}(\psi) \ln w_\psi}{\mathcal{H}} + \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}}.
\end{aligned}$$

For any greedy policy W with weight vector $\mathbf{w} = (w_0, \ldots, w_{R-1})$, let

$$\mu(W) \doteq \ln\left(\sum_{\psi=0}^{R-1} \frac{\pi_\psi(K_\psi - 1)w_\psi}{\mathcal{H}}\right) - \sum_{\psi=0}^{R-1} \frac{\pi_\psi \mathcal{H}(\psi) \ln w_\psi}{\mathcal{H}} - \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j}(-\ln p_{\psi,j})^2}{2\mathcal{H}}.$$

An immediate result from Theorems 3.3 and 3.4 is

Theorem 3.5 *As the size of the dictionary for each state increases,*

$$\ln M - H_s^W \longrightarrow \mu(W), \qquad s \in \{0, \ldots, R-1\}. \tag{3.13}$$

Consequently, the conservation of entropy implies that

$$\begin{aligned} \ln M - L^W \cdot \mathcal{H} &\longrightarrow \mu(W), \\ \textit{and so} \qquad (\ln M) \cdot \left(\frac{\ln M}{L^W} - \mathcal{H}\right) &\longrightarrow \mathcal{H} \cdot \mu(W). \end{aligned} \tag{3.14}$$

Theorem 3.5 indicates that the asymptotically optimal greedy policy can be found by minimizing $\mu(W)$ subject to the functional constraints that the corresponding weight vector $\mathbf{w}$ is in the set of greedy vectors $\mathcal{G}$. Note that $\mu(W)$ does not change if we multiply the weight vector by any positive constant. This is what we would expect because we chose our dictionaries $\mathcal{D}_s$ to maximize $W(s)$. Since the solution to the optimization problem is unique up to a scale factor, let us add the set constraint that

$$\sum_{\psi=0}^{R-1} \frac{\pi_\psi(K_\psi - 1)w_\psi}{\mathcal{H}} = 1.$$

We observe that any state ψ for which K_ψ is one does not cause any difficulties because unique parsability implies that $\mathcal{H}(\psi)$ is zero and hence the choice of weight w_ψ is irrelevant in the expression $\mu(W)$. The resulting convex programming problem can be solved by standard techniques (see [Lue84]). In the special case where

$$\mathbf{w}^* = \left(\frac{\mathcal{H}(0)}{K_0 - 1}, \cdots, \frac{\mathcal{H}(R-1)}{K_{R-1} - 1}\right) \tag{3.15}$$

is in the set of greedy vectors, it is straightforward to demonstrate that the corresponding policy is the unique optimal greedy policy with

$$\mu(*) = \sum_{\psi=0}^{R-1} \frac{\pi_\psi \mathcal{H}(\psi)}{\mathcal{H}} \ln\left(\frac{K_\psi - 1}{\mathcal{H}(\psi)}\right) - \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j}(-\ln p_{\psi,j})^2}{2\mathcal{H}}.$$

For the special case in which $K_r = K_s$ for all states r and s, the number of intermediate nodes in each dictionary tree is preestablished. If $(\mathcal{H}(0), \ldots, \mathcal{H}(R-1)) \in \mathcal{G}$, then (3.15) indicates that maximizing H_s for each s leads to the asymptotically optimal greedy code. Furthermore, (3.13) implies that as the dictionary size increases, $H_s - H_r$ approaches zero for all states r and s, and thus, $\max_s H_s - H$ approaches zero. From the conservation of entropy, it follows that this code is asymptotically optimal in the sense that no other choice of dictionaries will result in a smaller first-order term in the rate of convergence of the code redundancy to zero.

3.3 Future Work

There are a number of issues that need to be investigated in order to finish the work in this chapter. A very obvious question is whether or not Theorems 3.3-3.5 hold when the weight vector $\mathbf{w}$ is not in the set $\mathcal{G}$ of greedy vectors. In particular, when $K_r = K_s$ for all states r and s, will we *always* get the asymptotically optimal code by maximizing H_s for each s? From the conservation of entropy, it is clear that this choice of dictionaries will do at least as well asymptotically as the best greedy policy.

For thoroughness, it is necessary to at least sketch out the results for the case where the pseudo-self-information of the symbols emitted has an *arithmetic* distribution. The techniques to find and prove these results will parallel those in Appendix B.

Chapter 4

Notes on the Lempel-Ziv incremental parsing rule

4.1 Background

An interesting and challenging avenue of research is trying to better understand the Lempel-Ziv incremental parsing rule (see [ZL78]), which can be viewed as an adaptive version of the Tunstall algorithm. Let us continue to assume a Markov source with finite alphabet $\{0, \ldots, K-1\}$ and set of states $\{0, \ldots, R-1\}$ and maintain the notation used to analyze the generalized Tunstall code. The Lempel-Ziv incremental parsing rule starts off with a dictionary consisting of the K source symbols. At any parsing point, the next parsed phrase σ is the unique dictionary entry which is a prefix of the unparsed source output. Once this phrase has been selected, the dictionary is enlarged by replacing σ with its K single letter extensions. As an example, suppose that we have a ternary source, and the source output is the string 0 0 0 0 2 $\cdots$.

- Initially, the dictionary is {0, 1, 2}.
- The first parsed string is 0, and the dictionary is updated to {00, 01, 02, 1, 2}. At this point, the unparsed source output is 0 0 0 2 $\cdots$.
- The second parsed string is 00, and the revised dictionary is {000, 001, 002, 01, 02, 1, 2}. Now, the unparsed source sequence is 0 2 $\cdots$.

- The third parsed string is 02, resulting in the dictionary {000, 001, 002, 01, 020, 021, 022, 1, 2}.

The encoding procedure associated with the Lempel-Ziv incremental parsing rule is often called LZ '78. Observe that LZ '78 is a universal version of the Tunstall algorithm in which the coder's guess for the most probable dictionary entry at any time is the next parsed phrase.

Practical implementations of the Lempel-Ziv incremental parsing rule often differ somewhat from the original LZ '78 algorithm. We will focus on the LZW algorithm introduced by Welch in [Wel84]. Initially, the dictionary entries are the K source symbols. At any parsing point, the next parsed phrase is the longest dictionary entry which is a prefix of the unparsed source output. Thus far, the parsing rule is identical to the one used by LZ '78. The difference is in the way the dictionaries are updated in the two procedures. In the Lempel-Ziv incremental parsing rule, the last parsed phrase is replaced by its K single letter extensions, and hence, the resulting dictionary is always uniquely parsable. For LZW, the dictionary is enlarged by adding the last parsed phrase concatenated with the first symbol of the unparsed source output and thus the LZW dictionary generally is plurally parsable. According to [Wel84], LZW achieves very similar compression to LZ '78, but is easier to implement. Miller and Wegman discussed a "character extension improvement" algorithm in [MW85] that is identical to LZW. They claimed that the algorithm empirically achieves better compression than LZ '78 on English text, especially for small dictionary sizes. Miller and Wegman attributed the empirical success of LZW to the addition of at most one new dictionary string per parsed string versus the net gain of $K-1$ dictionary strings per parsed string created by LZ '78; i.e., each parsed string is represented by approximately $\log_D(K-1)$ fewer code symbols. Note that any string can appear as a parsed phrase at most once for the original Lempel-Ziv incremental parsing rule, while it can occur as a parsed phrase up to K times for LZW. Let us continue the previous example by examining how the LZW parser would segment the source output sequence 0 0 0 0 2 $\cdots$.

- Initially, the dictionary is {0, 1, 2}.
- The first parsed string is 0, and the remaining source output is 0 0 0 2 $\cdots$. Hence, the dictionary is enlarged to {0, 00, 1, 2}.

- The next parsed string is 00, and the unparsed source sequence is now 0 2 $\cdots$. The dictionary is expanded to $\{0,\ 00,\ 000,\ 1,\ 2\}$.

- The third parsed string is 0, and the rest of the source output is 2 $\cdots$. The new dictionary is $\{0,\ 00,\ 02,\ 000,\ 1,\ 2\}$ and the fourth parsed phrase is 2.

For LZ '78, it is clear that the decoder can use the sequence of code symbols to simulate the evolution of the parser's dictionary and subsequently reconstruct the source output; it is less obvious that the LZW decoder has this property. The LZW decoder can easily determine the first source output symbol u_1. The new dictionary entry is of the form $u_1 \circ j$ for some source symbol j. To find j, the decoder looks at the code letters corresponding to the second phrase. If these code letters indicate that the second parsed phrase is u_1 or $u_1 \circ j$, then j and u_1 are the same symbol. Otherwise, the second parsed phrase is some u_2 which is distinct from u_1, and therefore, j is the same as u_2. This argument can be extended to show that it is possible to accurately decode any source string from its corresponding string of code letters.

There are many small modifications that can be made to LZ '78 or LZW in order to create new encoding rules. For example, in [Gal95], Gallager proposed a variant of LZW. Suppose that a string σ has occurred $K-2$ times as a parsed string for LZW. Then it has two single letter extensions, say $\sigma \circ j_1$ and $\sigma \circ j_2$, which are not dictionary entries. Without loss of generality, assume that σ is next used as a parsed string when $\sigma \circ j_1$ is a prefix of the unparsed source output starting from a parsing point. Then $\sigma \circ j_1$ will be the new dictionary entry and σ will be used as a parsed string for the K^{th} time if and only if there is a parsing point at which $\sigma \circ j_2$ is a prefix of the unparsed source output. In Gallager's encoding rule, when a string is used as a parsed string for the $K-1^{\text{st}}$ time, the dictionary is updated by replacing the string with its two single letter extensions which are not already in the dictionary. Note that the size of the dictionary for Gallager's code grows by one each time a string is parsed, and a string can be used as a parsed phrase up to $K-1$ times. For $K=2$, Gallager's encoding rule is the same as LZ '78. Let us continue our example and see how the Gallager parser would segment the source output sequence 0 0 0 0 2 $\cdots$.

- Initially, the dictionary is $\{0,\ 1,\ 2\}$.

- The first parsed string is 0, and the remaining source output is 0 0 0 2 $\cdots$. The new dictionary is {0, 00, 1, 2}.

- The second parsed string is 00, and the dictionary is enlarged to {0, 00, 000, 1, 2}. Now, the unparsed source sequence is 0 2 $\cdots$.

- The next parsed string is 0 and the remainder of the source output is 2 $\cdots$. Since this is the second time that 0 is a parsed string, it will be removed from the dictionary and the strings 01 and 02 will be added. The new dictionary is {00, 01, 02, 000, 1, 2} and the fourth parsed phrase is 2.

4.2 New Redundancy Bound

Let u_1^n symbolize the string $u_1, \ldots, u_n$. Let $\mathcal{L}^{LZ}(u_1^n)$, $\mathcal{L}^{W}(u_1^n)$, and $\mathcal{L}^{G}(u_1^n)$ denote the length of the encoding of the string u_1^n in bits for LZ '78, LZW, and Gallager's code, respectively. The redundancies $\mathcal{R}^{LZ}$, $\mathcal{R}^{W}$, and $\mathcal{R}^{G}$ of the codes in bits are

$$\mathcal{R}^{LZ} = E_{u_1^n}\left(\frac{1}{n}\mathcal{L}^{LZ}(u_1^n)\right) - \mathcal{H}\log_2 e \tag{4.1}$$

$$\mathcal{R}^{W} = E_{u_1^n}\left(\frac{1}{n}\mathcal{L}^{W}(u_1^n)\right) - \mathcal{H}\log_2 e \tag{4.2}$$

and

$$\mathcal{R}^{G} = E_{u_1^n}\left(\frac{1}{n}\mathcal{L}^{G}(u_1^n)\right) - \mathcal{H}\log_2 e, \tag{4.3}$$

where the expectations are taken over all n-tuples. In [PWZ92], it is established that for a binary source,

$$\mathcal{R}^{LZ} \leq \frac{\ln\ln n}{\ln n} + o\left(\frac{\ln\ln n}{\ln n}\right).$$

In [LS95], it is claimed that for a binary, memoryless source, there exists a constant $\mathcal{C}$ which is a function of source parameters that satisfies

$$\mathcal{R}^{LZ} = \frac{\mathcal{C}}{\ln n} + O\left(\frac{\ln\ln n}{(\ln n)^2}\right)$$

and that an extension of this result exists for Markov sources. Our approach to analyzing the redundancy of LZ '78 is new. Furthermore, we provide the first asymptotic bound on the

redundancy of LZW and Gallager's code.

In evaluating $\mathcal{L}^{LZ}(u_1^n)$, $\mathcal{L}^{W}(u_1^n)$, and $\mathcal{L}^{G}(u_1^n)$, we will use the following result.

Lemma 4.1 *For any integer $k \geq 2$, and real number $x \geq 0$,*

$$\begin{aligned}\sum_{i=1}^{k}\lceil x+\log_2 i\rceil &\leq k\left(\lfloor\log_2(2^x k)\rfloor\right)+k-2\cdot 2^{-x+\lfloor\log_2(2^x k)\rfloor}+O(\ln k)\\ &\leq k\log_2 k+kx+k\log_2\left(\frac{\log_2 e}{e}\right)+O(\ln k).\end{aligned}$$

Proof: Let $\epsilon = x - \lfloor x\rfloor$. Then $0 \leq \epsilon < 1$ and

$$\sum_{i=1}^{k}\lceil x+\log_2 i\rceil = k\lfloor x\rfloor+\sum_{i=1}^{k}\lceil \epsilon+\log_2 i\rceil,$$

so it is sufficient to prove the lemma assuming $0 \leq x < 1$. Let $z = \lfloor\log_2(2^x k)\rfloor$ and suppose that l is the largest integer for which $z = \lceil\log_2(2^x l)\rceil$; i.e., $2^x l \leq 2^z < 2^x(l+1)$, and so $l \leq 2^{z-x} < l+1$. We have that

$$\begin{aligned}\sum_{i=1}^{k}\lceil x+\log_2 i\rceil &= \sum_{i=1}^{k}\lceil\log_2(2^x i)\rceil\\ &= \sum_{j=1}^{z} j\cdot|\text{ integers } i: 2^{j-1}<2^x i\leq 2^j|+(k-l)(z+1)\\ &= \sum_{j=1}^{z} j\cdot|\text{ integers } i: 2^{j-x-1}<i\leq 2^{j-x}|+(k-l)(z+1)\\ &\leq \sum_{j=1}^{z} j\cdot 2^{j-x-1}+(k-2^{z-x}+1)(z+1)\\ &= 2^{z-x}(z-1)+2^{-x}+(k-2^{z-x})(z+1)+z+1\\ &= kz+k-2\cdot 2^{z-x}+O(z),\end{aligned}\tag{4.4}$$

which is equivalent to the first inequality. Let $y = 2^{z-x}$. Since $y = 2^{\lfloor\log_2(2^x k)\rfloor-x}$, $\frac{k}{2} < y \leq k$. The expression $k\log_2 y - 2y$ is maximized at $y = \frac{k\log_2 e}{2}$ and for this value of y, $z = \log_2 k + x + \log_2(\log_2 e) - 1$. Substituting this value of z into (4.4), we obtain the second inequality. □

Let c^{LZ}, c^{W} and c^{G} represent the number of complete phrases obtained by parsing u_1^n according to LZ '78, LZW, and Gallager's code respectively. For LZ '78, the dictionary starts

with K entries and has a net gain of $K-1$ symbols per parse; hence, the size of the dictionary used to select the ψ^{th} parsed string is $\psi(K-1)+1$. Since $\lceil \log_2 M \rceil$ bits are used to encode any entry of a dictionary of size M for each parsing rule, $\mathcal{L}^{LZ}(u_1^n)$ satisfies

$$\mathcal{L}^{LZ}(u_1^n) \;<\; \sum_{j=1}^{c^{LZ}+1} \lceil \log_2(j(K-1)+1) \rceil \;<\; \sum_{j=1}^{c^{LZ}+1} \lceil \log_2(K-1) + \log_2(j+1) \rceil.$$

Therefore, it follows from Lemma 4.1 that

$$\mathcal{L}^{LZ}(u_1^n) \;\leq\; c^{LZ} \log_2 c^{LZ} + c^{LZ} \log_2 \left(\frac{(K-1)\log_2 e}{e} \right) + O\left(\ln c^{LZ} \right). \tag{4.5}$$

For LZW and Gallager's code, the number of possibilities for the ψ^{th} parsed string is $\psi + K - 1$ and thus Lemma 4.1 implies that

$$\mathcal{L}^{W}(u_1^n) \;\leq\; c^{W} \log_2 c^{W} + c^{W} \log_2 \left(\frac{\log_2 e}{e} \right) + O\left(\ln c^{W} \right), \tag{4.6}$$

and

$$\mathcal{L}^{G}(u_1^n) \;\leq\; c^{G} \log_2 c^{G} + c^{G} \log_2 \left(\frac{\log_2 e}{e} \right) + O\left(\ln c^{G} \right). \tag{4.7}$$

From (4.5)-(4.7), we see that upper bounds on c^{LZ}, c^{W}, and c^{G} lead to upper bounds on $\mathcal{L}^{LZ}$, $\mathcal{L}^{W}$, and $\mathcal{L}^{G}$, respectively. We have the following results.

Theorem 4.1 *Assume that the source has positive entropy. Let $I = I(u_1^n|s_0) < \infty$. For the three encoding rules we are studying, we have the following asymptotic relationships between the number of phrases associated with the parsing of u_1^n and the self-information of u_1^n.*

$$c^{LZ} \cdot \frac{\ln I}{I} \;\leq\; 1 + o(1) \tag{4.8}$$

$$(c^{LZ} \log_2 c^{LZ} - I \log_2 e) \cdot \frac{\ln I}{I} \;\leq\; \log_2 \left(\frac{Re}{\mathcal{H}} \right) + o(1) \tag{4.9}$$

$$c^{W} \cdot \frac{\ln I}{I} \;\leq\; 1 + o(1) \tag{4.10}$$

$$(c^{W} \log_2 c^{W} - I \log_2 e) \cdot \frac{\ln I}{I} \;\leq\; \log_2 \left(\frac{RKe}{\mathcal{H}} \right) + o(1) \tag{4.11}$$

$$c^{G} \cdot \frac{\ln I}{I} \;\leq\; 1 + o(1) \tag{4.12}$$

$$(c^{G} \log_2 c^{G} - I \log_2 e) \cdot \frac{\ln I}{I} \;\leq\; \log_2 \left(\frac{R(K-1)e}{\mathcal{H}} \right) + o(1). \tag{4.13}$$

Hence,

$$\ln n \cdot \left(\frac{\mathcal{L}^{LZ}(u_1^n) - I \log_2 e}{n} \right) \leq \frac{I}{n} \log_2 \left(\frac{R(K-1)e}{\mathcal{H}} \right) + o(1) \tag{4.14}$$

$$\ln n \cdot \left(\frac{\mathcal{L}^{W}(u_1^n) - I \log_2 e}{n} \right) \leq \frac{I}{n} \log_2 \left(\frac{RKe}{\mathcal{H}} \right) + o(1) \tag{4.15}$$

$$\ln n \cdot \left(\frac{\mathcal{L}^{G}(u_1^n) - I \log_2 e}{n} \right) \leq \frac{I}{n} \log_2 \left(\frac{R(K-1)e}{\mathcal{H}} \right) + o(1). \tag{4.16}$$

Proof: We introduce the following notation for the Lempel-Ziv incremental parsing rule. Let

- σ_i denote the i^{th} phrase of the source output
- ψ_i denote the source state just before phrase σ_i
- $\Omega = \{\tau : I(\sigma|s) = \tau \text{ for some string } \sigma \text{ and state } s\}$
- $c(\tau) = |\{\sigma_i : 1 \leq i \leq c^{LZ},\ I(\sigma_i|\psi_i) = \tau\}|$
- $\gamma_s(\tau) = |\{\sigma :\ I(\sigma|s) = \tau\}|$
- $\gamma(\tau) = \sum_{s=0}^{R-1} |\{\sigma :\ I(\sigma|s) \leq \tau\}|$
- $\tilde{I} = \sum_{i=1}^{c^{LZ}} I(\sigma_i|\psi_i)$; $\tilde{I} \leq I$ because of the final partial phrase.

Note that

$$c^{LZ} = \sum_{\tau \in \Omega} c(\tau) \tag{4.17}$$

and

$$\tilde{I} = \sum_{\tau \in \Omega} \tau \cdot c(\tau). \tag{4.18}$$

To upper bound c^{LZ}, we maximize $\sum_{\tau \in \Omega} c(\tau)$ subject to the constraints $\sum_{\tau \in \Omega} \tau \cdot c(\tau) \leq I$ and $0 \leq c(\tau) \leq \sum_{s=0}^{R-1} \gamma_s(\tau)$ for all $\tau \in \Omega$. We will show that the number of phrases is maximized by selecting as many phrases with small self-information as possible. In particular, we are going to pick a "threshhold" self-information $\overline{\tau}$ and consider the set $\mathcal{S}$ of strings with self-information

upper-bounded by $\overline{\tau}$. Our choice of $\overline{\tau}$ is determined by the criterion that the cumulative self-information of the strings in $\mathcal{S}$ is approximately I. We will upper bound c^{LZ} by the size of $\mathcal{S}$. More precisely, we have the following result.

Lemma 4.2 *An upper bound on c^{LZ} is given by*

$$c^{LZ} \leq \gamma(\overline{\tau}) \tag{4.19}$$

where $\overline{\tau}$ is chosen so that

$$\sum_{s=0}^{R-1} \sum_{\tau \in \Omega: \tau < \overline{\tau}} \tau \cdot \gamma_s(\tau) < I \leq \sum_{s=0}^{R-1} \sum_{\tau \in \Omega: \tau \leq \overline{\tau}} \tau \cdot \gamma_s(\tau). \tag{4.20}$$

Proof of Lemma 4.2: To arrive at a contradiction, suppose that (4.19) is false. Then $c^{LZ} \geq \gamma(\overline{\tau}) + 1$. The encoding rule ensures that every complete phrase in the parsing is distinct. Let h_i denote the self-information of phrase i; i.e., $h_i = I(\sigma_i|\psi_i)$. Without loss of generality, reorder the self-informations so that $h_1 \leq h_2 \leq \cdots \leq h_{c^{LZ}}$. Now consider the $\gamma(\overline{\tau})$ distinct pairs (s, λ) with $I(\lambda|s) \leq \overline{\tau}$ and order the pairs so that $I(\lambda_1|s_1) \leq I(\lambda_2|s_2) \leq \cdots \leq I(\lambda_{\gamma(\overline{\tau})}|s_{\gamma(\overline{\tau})})$. Observe that for each $i \in \{1, 2, \ldots, \gamma(\overline{\tau})\}$, $h_i \geq I(\lambda_i|s_i)$ and for each $i > \gamma(\overline{\tau})$, $h_i \geq \overline{\tau}$. Hence,

$$\begin{aligned} I \geq \tilde{I} = \sum_{i=1}^{c^{LZ}} h_i \geq \sum_{i=1}^{\gamma(\overline{\tau})+1} h_i &> \sum_{i=1}^{\gamma(\overline{\tau})} I(\lambda_i|s_i) + \overline{\tau} \\ &= \sum_{s=0}^{R-1} \sum_{\tau \in \Omega: \tau \leq \overline{\tau}} \tau \cdot \gamma_s(\tau) + \overline{\tau} \\ &\geq I + \overline{\tau}, \end{aligned}$$

which is a contradiction. □

As in the last two chapters, let $\{N^{(\psi,s)}(t); t \geq 0\}$ be the renewal or delayed renewal process defined, for each state ψ, and starting state s, by the number of renewals, i.e., entrances of the source into state ψ, until the self-information generated by the source reaches t, and let $m^{(\psi,s)}(t) = E[N^{(\psi,s)}(t)]$. Given the starting state s and the source output $v_1, v_2, \ldots$, each prefix v_1^i with $S[s, v_1^i] = \psi$ corresponds to a renewal in the corresponding sample function; the

expected number of renewals in the interval $(t, t+dt]$ is given by

$$m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t) = \sum_{\sigma: I(\sigma|s) \in (t,t+dt],\, \psi = S[s,\sigma]} P(\sigma|s). \tag{4.21}$$

As we have remarked in earlier chapters, if $I(\sigma|s) \in (t, t+dt]$, then $e^{-t-dt} \leq P(\sigma|s) < e^{-t}$. Hence, (4.21) implies that the number of strings σ with $I(\sigma|s) \in (t, t+dt]$ and $\psi = S[s,\sigma]$ is in the interval $(e^t[m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t)], e^{t+dt}[m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t)]]$. For the remainder of this chapter, we assume that the self-information corresponding to the symbols emitted by the source is *non-arithmetic*. The case where the self-information of source symbols is arithmetic can be handled using the ideas in this chapter and those in Appendix B. We have that

$$\gamma(\overline{\tau}) = \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\overline{\tau}^+} e^x \, dm^{(\psi,s)}(x) \tag{4.22}$$

and

$$\sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\overline{\tau}^-} x e^x \, dm^{(\psi,s)}(x) < I \leq \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\overline{\tau}^+} x e^x \, dm^{(\psi,s)}(x). \tag{4.23}$$

Hence, by multiplying both sides of (4.22) and (4.23) by $e^{-\overline{\tau}}$, we find that

$$\gamma(\overline{\tau}) \cdot e^{-\overline{\tau}} = \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\overline{\tau}^+} e^{x-\overline{\tau}} \, dm^{(\psi,s)}(x) \tag{4.24}$$

and

$$\begin{aligned} \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\overline{\tau}^-} x e^{x-\overline{\tau}} \, dm^{(\psi,s)}(x) &< I \cdot e^{-\overline{\tau}} \leq \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\overline{\tau}^+} x e^{x-\overline{\tau}} \, dm^{(\psi,s)}(x) \\ &= \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\overline{\tau}^+} (x-\overline{\tau}) e^{x-\overline{\tau}} \, dm^{(\psi,s)}(x) + \overline{\tau}\gamma(\overline{\tau}) \cdot e^{-\overline{\tau}}. \end{aligned} \tag{4.25}$$

Blackwell's Theorem (see Theorem A.1), and Lemma A.1 imply that for all s, as $\overline{\tau}$ increases,

$$\int_0^{\overline{\tau}^+} e^{x-\overline{\tau}} \, dm^{(\psi,s)}(x) \longrightarrow \frac{\pi_\psi}{\mathcal{H}} \tag{4.26}$$

$$\int_0^{\overline{\tau}^-} (x-\overline{\tau}) e^{x-\overline{\tau}} \, dm^{(\psi,s)}(x) \longrightarrow -\frac{\pi_\psi}{\mathcal{H}} \tag{4.27}$$

and

$$\int_0^{\overline{\tau}^+} (x-\overline{\tau}) e^{x-\overline{\tau}} \, dm^{(\psi,s)}(x) \longrightarrow -\frac{\pi_\psi}{\mathcal{H}}. \tag{4.28}$$

It follows from (4.24) to (4.28) that as $\overline{\tau}$ increases,

$$\gamma(\overline{\tau}) \cdot e^{-\overline{\tau}} = \frac{R}{\mathcal{H}} + o(1) \tag{4.29}$$

and

$$(\overline{\tau}\gamma(\overline{\tau}) - I) \cdot e^{-\overline{\tau}} = \frac{R}{\mathcal{H}} + o(1). \tag{4.30}$$

Taking the logarithm of both sides of (4.29), we find that as $\overline{\tau}$ increases,

$$\log_2 \gamma(\overline{\tau}) - \overline{\tau} \log_2 e = \log_2 \left(\frac{R}{\mathcal{H}}\right) + o(1). \tag{4.31}$$

Multiplying (4.29) by (4.31), we see that

$$(\gamma(\overline{\tau}) \log_2 \gamma(\overline{\tau}) - \overline{\tau}\gamma(\overline{\tau}) \log_2 e) \cdot e^{-\overline{\tau}} = \frac{R}{\mathcal{H}} \log_2 \left(\frac{R}{\mathcal{H}}\right) + o(1). \tag{4.32}$$

Multiplying both sides of (4.30) by $\log_2 e$ and adding the resulting expression to (4.32), we find that as $\overline{\tau}$ increases

$$(\gamma(\overline{\tau}) \log_2 \gamma(\overline{\tau}) - I \log_2 e) \cdot e^{-\overline{\tau}} = \frac{R}{\mathcal{H}} \log_2 \left(\frac{Re}{\mathcal{H}}\right) + o(1). \tag{4.33}$$

Next, we would like to determine the asymptotic relationship between $e^{-\overline{\tau}}$ and I. Substituting (4.29) into (4.30), we see that

$$\overline{\tau}\frac{R}{\mathcal{H}} + o(\overline{\tau}) - I \cdot e^{-\overline{\tau}} = 0. \tag{4.34}$$

Define δ to satisfy

$$\overline{\tau} = \ln \left(\frac{\frac{\mathcal{H}I}{R}}{\ln\left(\frac{\mathcal{H}I}{R}\right)} (1 + \delta) \right). \tag{4.35}$$

Substituting (4.35) into (4.34), we find that

$$\frac{R}{\mathcal{H}} \ln \left(\frac{\mathcal{H}I}{R}\right) + \frac{R}{\mathcal{H}} \ln(1+\delta) - \frac{R}{\mathcal{H}} \ln \ln \left(\frac{\mathcal{H}I}{R}\right) + o\left(\ln\left(\frac{\mathcal{H}I}{R}(1+\delta)\right)\right) - \frac{R}{\mathcal{H}(1+\delta)} \ln \left(\frac{\mathcal{H}I}{R}\right) = 0,$$

and dividing both sides of this equation by $\frac{R}{\mathcal{H}} \ln\left(\frac{\mathcal{H}I}{R}\right)$, we see that

$$1 + \frac{\ln(1+\delta)}{\ln\left(\frac{\mathcal{H}I}{R}\right)} - o(1) + o\left(1 + \frac{\ln(1+\delta)}{\ln\left(\frac{\mathcal{H}I}{R}\right)}\right) - \frac{1}{1+\delta} \;=\; 0.$$

Thus,

$$\delta \;=\; o(1),$$

and hence, as $\overline{\tau}$ and I increase,

$$e^{-\overline{\tau}} \;=\; \left(\frac{R}{\mathcal{H}I} \ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1 + o(1)). \tag{4.36}$$

Substituting (4.36) into (4.29) and (4.33), we see that as $\overline{\tau}$ and I increase,

$$\gamma(\overline{\tau}) \cdot \left(\frac{R}{\mathcal{H}I} \ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1 + o(1)) \;=\; \frac{R}{\mathcal{H}} + o(1), \tag{4.37}$$

and

$$(\gamma(\overline{\tau}) \log_2 \gamma(\overline{\tau}) - I \log_2 e) \cdot \left(\frac{R}{\mathcal{H}I} \ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1 + o(1)) \;=\; \frac{R}{\mathcal{H}} \log_2\left(\frac{Re}{\mathcal{H}}\right) + o(1). \tag{4.38}$$

By Lemma 4.2, $c^{LZ} \leq \gamma(\overline{\tau})$. Hence, by (4.37) and (4.38), as I increases,

$$c^{LZ} \cdot \left(\frac{R}{\mathcal{H}I} \ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1 + o(1)) \;\leq\; \frac{R}{\mathcal{H}} + o(1), \tag{4.39}$$

and

$$(c^{LZ} \log_2 c^{LZ} - I \log_2 e) \cdot \left(\frac{R}{\mathcal{H}I} \ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1 + o(1)) \;\leq\; \frac{R}{\mathcal{H}} \log_2\left(\frac{Re}{\mathcal{H}}\right) + o(1). \tag{4.40}$$

(4.39) and (4.40) are equivalent to (4.8) and (4.9), respectively. From (4.5), (4.8), and (4.9), we have

$$(\mathcal{L}^{LZ}(u_1^N) - I \log_2 e) \cdot \frac{\ln I}{I} \;\leq\; \log_2\left(\frac{R(K-1)e}{\mathcal{H}}\right) + o(1). \tag{4.41}$$

Since the source has positive entropy, $I = \Theta(n)$, and therefore, (4.14) is equivalent to (4.41).

We use the same ideas to prove the rest of Theorem 4.1. For LZW and Gallager's code, the lemma corresponding to Lemma 4.2 is

Lemma 4.3

$$c^W \;\leq\; K\gamma(\tau^W)$$

and

$$c^G \leq (K-1)\gamma(\tau^G),$$

where τ^W *and* τ^G *are chosen so that*

$$K\sum_{s=0}^{R-1}\sum_{\tau\in\Omega:\tau<\tau^W}\tau\cdot\gamma_s(\tau) < I \leq K\sum_{s=0}^{R-1}\sum_{\tau\in\Omega:\tau\leq\tau^W}\tau\cdot\gamma_s(\tau)$$

$$(K-1)\sum_{s=0}^{R-1}\sum_{\tau\in\Omega:\tau<\tau^G}\tau\cdot\gamma_s(\tau) < I \leq (K-1)\sum_{s=0}^{R-1}\sum_{\tau\in\Omega:\tau\leq\tau^G}\tau\cdot\gamma_s(\tau)$$

Proof of Lemma 4.3: The difference between Lemma 4.2 and Lemma 4.3 lies in the number of times a given string can appear as a parsed phrase for each encoding rule. In LZ'78, a string can occur at most once. Any string can occur up to K times as a parsed phrase for LZW and up to $K-1$ times as a parsed phrase for Gallager's code. □

For LZW, the counterparts to (4.29) and (4.30) are that as τ^W increases,

$$\gamma(\tau^W)\cdot e^{-\tau^W} = \frac{R}{\mathcal{H}} + o(1) \tag{4.42}$$

and

$$(K\tau^W\gamma(\tau^W) - I)\cdot e^{-\tau^W} = \frac{RK}{\mathcal{H}} + o(1) \tag{4.43}$$

and for Gallager's code, as τ^G increases,

$$\gamma(\tau^G)\cdot e^{-\tau^G} = \frac{R}{\mathcal{H}} + o(1) \tag{4.44}$$

and

$$((K-1)\tau^G\gamma(\tau^G) - I)\cdot e^{-\tau^G} = \frac{R(K-1)}{\mathcal{H}} + o(1). \tag{4.45}$$

With these modifications, the proofs of (4.10)-(4.13), (4.15) and (4.16) are identical to the proofs of (4.8), (4.9) and (4.14). □

An immediate consequence of Theorem 4.1 is

Corollary 4.1 *Assume that the source has positive entropy. Let* h_{max} *denote the maximum self-information generated by the source upon emitting a symbol. Note that* h_{max} *is finite. Then for all source output strings* u_1^n *with non-zero probability,*

$$\frac{\mathcal{L}^{LZ}(u_1^n) - I\log_2 e}{n} \leq \frac{h_{max}}{\ln n}\left(\log_2\left(\frac{R(K-1)e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right)$$

$$\begin{aligned}
\frac{\mathcal{L}^W(u_1^n) - I\log_2 e}{n} &\leq \frac{h_{max}}{\ln n}\left(\log_2\left(\frac{RKe}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
\frac{\mathcal{L}^G(u_1^n) - I\log_2 e}{n} &\leq \frac{h_{max}}{\ln n}\left(\log_2\left(\frac{R(K-1)e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right)
\end{aligned}$$

Proof: We have that $I \leq n \cdot h_{\max}$. This fact and (4.14)-(4.16) imply the result. □

For a very probable collection of strings, we can further tighten the bound presented in Corollary 4.1. In particular, we have the following result.

Corollary 4.2 *Assume that the source has finite entropy. With probability* $1 - O\left(\frac{1}{\sqrt{n}}\right)$,

$$\begin{aligned}
\frac{\mathcal{L}^{LZ}(u_1^n) - I\log_2 e}{n} &\leq \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
\frac{\mathcal{L}^W(u_1^n) - I\log_2 e}{n} &\leq \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{RK\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
\frac{\mathcal{L}^G(u_1^n) - I\log_2 e}{n} &\leq \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right).
\end{aligned}$$

Proof: Since $\lim_{n\to\infty} E_{u_1^n}\left(\frac{I(u_1^n|s_0)}{n}\right) = \mathcal{H}$, and the variance and third moment of the self-information of every source symbol emitted is finite, it follows from [GK68] that

$$\mathcal{H} - o\left(\frac{1}{\ln n}\right) \leq \frac{I}{n} \leq \mathcal{H} + o\left(\frac{1}{\ln n}\right) \quad \text{with probability } 1 - O\left(\tfrac{1}{\sqrt{n}}\right). \tag{4.46}$$

The corollary follows from (4.46) and (4.14)-(4.16). □

It is also easy to derive upper bounds on the redundancy of the codes using Theorem 1 and (4.46). We have the following result.

Theorem 4.2 *Assume the source has positive entropy. Then*

$$\begin{aligned}
\mathcal{R}^{LZ} &\leq \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
\mathcal{R}^{W} &\leq \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{RK\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
\textit{and } \mathcal{R}^{G} &\leq \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right).
\end{aligned}$$

Proof: Taking the expected value of both sides of (4.14), (4.15) and (4.16) with respect to u_1^n, we see that

$$\ln n \cdot E_{u_1^n}\left(\frac{\mathcal{L}^{LZ}(u_1^n) - I(u_1^n|s_0)\log_2 e}{n}\right) \leq E_{u_1^n}\left(\frac{I(u_1^n|s_0)}{n}\right) \cdot \log_2\left(\frac{R(K-1)e}{\mathcal{H}}\right) + o(1) \quad (4.47)$$

$$\ln n \cdot E_{u_1^n}\left(\frac{\mathcal{L}^{W}(u_1^n) - I(u_1^n|s_0)\log_2 e}{n}\right) \leq E_{u_1^n}\left(\frac{I(u_1^n|s_0)}{n}\right) \cdot \log_2\left(\frac{RKe}{\mathcal{H}}\right) + o(1) \quad (4.48)$$

$$\ln n \cdot E_{u_1^n}\left(\frac{\mathcal{L}^{G}(u_1^n) - I(u_1^n|s_0)\log_2 e}{n}\right) \leq E_{u_1^n}\left(\frac{I(u_1^n|s_0)}{n}\right) \cdot \log_2\left(\frac{R(K-1)e}{\mathcal{H}}\right) + o(1) \quad (4.49)$$

We have that for all string u_1^n,

$$0 \leq \frac{I(u_1^n|s_0)}{n} \leq h_{\max}. \quad (4.50)$$

(4.46) and (4.50) imply that

$$\mathcal{H} - o\left(\frac{1}{\ln n}\right) \leq E\left(\frac{I(u_1^n|s_0)}{n}\right) \leq \mathcal{H} + o\left(\frac{1}{\ln n}\right). \quad (4.51)$$

(4.47)-(4.49) and (4.51) imply the theorem. □

It is interesting to compare the redundancy of the code associated with the Lempel-Ziv incremental parsing rule with that of a different type of Lempel-Ziv code. LZ '77, i.e., the encoding rule developed by Ziv and Lempel in 1977 (see [ZL77]), uses a greedy parsing scheme. Essentially, if the first n symbols u_1^n of the source output have been parsed, the next parsed string is the longest prefix σ of the unparsed source output that is of the form u_m^{m+l-1} for some $m \leq n$. The string is encoded by using $\log_2 n$ bits to represent m and $O(\log l)$ bits to represent l. For a fixed database implementation of LZ '77, Wyner demonstrated in [Wyn93] that as n increases, the length l of the next parsed phrase is asymptotically normally distributed with mean $\frac{\ln n}{\mathcal{H}} + O(1)$ and variance $O(\ln n)$. This suggests that the redundancy of encoding the first n letters of the source output using LZ '77 is $\Theta\left(\frac{\ln\ln n}{\ln n}\right)$. Since LZ '78 and its variants can also be viewed as greedy procedures, it was conjectured that these schemes would also have redundancy $\Theta\left(\frac{\ln\ln n}{\ln n}\right)$. In LZ '78 and the related codes, we upper bounded the redundancies of the algorithms by minimizing the self-information per parsed phrase. The goal of LZ '77 is to

maximize the self-information per phrase, but the resulting decrease in the number of phrases is more than offset by the size of the "dictionary" corresponding to the parsed string u_1^n, and this is why LZ '78 and its variants asymptotically outperform LZ '77. However, recent trends in practical lossless data compression suggest that LZ '77 and its variants perform as well as, if not better than, LZ '78 and its variants. There are a few explanations for this. For small to moderate values of n, the difference between $\Theta\left(\frac{\ln\ln n}{\ln n}\right)$ and $\Theta\left(\frac{1}{\ln n}\right)$ is not significant and an understanding of the lower order terms is very important in order to make a fair comparison. The other limitations to our analysis are the assumptions that the source statistics do not change over time and that the dictionary of source strings can grow arbitrarily large. If either of these assumptions are violated, then the situation may be very different.

4.3 Bound on Pointwise Code Length

The results of the last section and earlier analyses on the compression achieved by the Lempel-Ziv codes assume a model for the source and bound the average number of code symbols per source symbol used by the code for a random output from this source. In practical situations, we would like to be able to bound the number of code symbols per source symbol needed for the encoding of a particular string u_1^n. The appropriate way to modify the analysis carried out in the last section is to select a source model and then choose the parameters of this model to maximize the likelihood that u_1^n is emitted. More precisely, suppose the source letters come from a finite alphabet $\{0, 1, \ldots, K-1\}$. We will apply a model for the source in which there are a set of states $\{0, 1, \ldots, R-1\}$ with an initial state s_0 and $S[s, j]$ defines the next state if symbol j is emitted from state s. In order to complete the definition of the model, we need to specify $\theta_{s,j}$, the probability that the source emits symbol j from state s. Let $\hat{p}_{s,j}$ and $\hat{\pi}_s$ be the empirical probability that j is emitted from state s and the empirical probability that the source is in state s, respectively. That is,

$$\hat{p}_{s,j} = \frac{\text{number of times } j \text{ is emitted from state } s \text{ in } u_1^n}{\text{number of times the source is in state } s \text{ in } u_1^n}$$

and

$$\hat{\pi}_s = \frac{\text{number of times the source is in state } s \text{ in } u_1^n}{n}$$

The empirical entropy $\hat{\mathcal{H}}_n$ for this model is given by

$$\hat{\mathcal{H}}_n \;=\; -\sum_{s=0}^{R-1}\sum_{j=0}^{K-1} \hat{\pi}_s \hat{p}_{s,j} \ln \hat{p}_{s,j},$$

and the self-information $I(u_1^n|s_0)$ of u_1^n assuming the model is

$$I(u_1^n|s_0) \;=\; n\sum_{s=0}^{R-1}\sum_{j=0}^{K-1} \hat{\pi}_s \hat{p}_{s,j} \ln\left(\frac{1}{\theta_{s,j}}\right).$$

Lemma 4.4 *The choice of the probabilities $\theta_{s,j}$ to minimize the self-information $I(u_1^n|s_0)$ of u_1^n is $\theta_{s,j} = \hat{p}_{s,j}$ for all states s and symbols j.*

Proof: The problem of minimizing $I(u_1^n|s_0)$ is equivalent to selecting the $\theta_{s,j}$ to minimize the *empirical divergence* $\hat{D}_n$ defined by

$$\hat{D}_n \;\doteq\; -\sum_{s=0}^{R-1}\sum_{j=0}^{K-1} \hat{\pi}_s \hat{p}_{s,j} \ln\left(\frac{\theta_{s,j}}{\hat{p}_{s,j}}\right).$$

Using the inequality that for all $x > 0$, $\ln x \leq x - 1$ with equality if and only if $x = 1$, we have that

$$\begin{aligned}\hat{D}_n \;&\geq\; -\sum_{s=0}^{R-1}\sum_{j=0}^{K-1} \hat{\pi}_s \hat{p}_{s,j}\left(\frac{\theta_{s,j}}{\hat{p}_{s,j}} - 1\right)\\ &=\; 0\end{aligned}$$

and $\hat{D}_n = 0$ if and only if $\theta_{s,j} = p_{s,j}$ for all states s and symbols j. □

Let $\hat{I}(u_1^n|s_0)$ be the self-information of the string assuming the empirical model; i.e., when $\theta_{s,j} = \hat{p}_{s,j}$ for all states s and symbols j. Note that Lemma 4.4 implies that for this model of the source, we have that

$$\hat{I}(u_1^n|s_0) \;=\; n\hat{\mathcal{H}}_n. \tag{4.52}$$

Suppose that $\hat{\mathcal{H}}_n$ is positive. Then the analysis on individual sequences carried out in the last section applies for u_1^n assuming the empirical model of the source. The counterparts to (4.14)-(4.16) are

Theorem 4.3 *If $\hat{\mathcal{H}}_n$ is positive, then*

$$\begin{aligned}
\frac{1}{n}\mathcal{L}^{LZ}(u_1^n) - \hat{\mathcal{H}}_n \log_2 e &\leq \frac{\hat{\mathcal{H}}_n}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\hat{\mathcal{H}}_n}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
\frac{1}{n}\mathcal{L}^{W}(u_1^n) - \hat{\mathcal{H}}_n \log_2 e &\leq \frac{\hat{\mathcal{H}}_n}{\ln n}\left(\log_2\left(\frac{RK\log_2 e}{\hat{\mathcal{H}}_n}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
\frac{1}{n}\mathcal{L}^{G}(u_1^n) - \hat{\mathcal{H}}_n \log_2 e &\leq \frac{\hat{\mathcal{H}}_n}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\hat{\mathcal{H}}_n}\right)\right) + o\left(\frac{1}{\ln n}\right).
\end{aligned}$$

Now suppose that $\hat{\mathcal{H}}_n$ is zero. We can select another model for the source which results in a small, but positive empirical divergence $\hat{\mathcal{D}}_n$. For this source model, $\hat{I}(u_1^n|s_0) = n\hat{\mathcal{D}}_n$, and the entropy of the source is $\hat{\mathcal{D}}_n$. Hence, Theorem 4.3 will hold with $\hat{\mathcal{H}}_n$ replaced by $\hat{\mathcal{D}}_n$. Since $\hat{\mathcal{D}}_n$ can be chosen arbitrarily close to zero, we have

Theorem 4.4 *If $\hat{\mathcal{H}}_n$ is zero, then*

$$\begin{aligned}
\frac{1}{n}\mathcal{L}^{LZ}(u_1^n) &\leq o\left(\frac{1}{\ln n}\right) \\
\frac{1}{n}\mathcal{L}^{W}(u_1^n) &\leq o\left(\frac{1}{\ln n}\right) \\
\frac{1}{n}\mathcal{L}^{G}(u_1^n) &\leq o\left(\frac{1}{\ln n}\right).
\end{aligned}$$

Sometimes Theorem 4.4 provides useful information and other times it does not. As an example of the former situation, suppose we assume a binary, memoryless source and the source output u_1^n is the all-zero string. In this case, it is straightforward to demonstrate that the three encoding rules will all parse the string as 0 00 000 $\cdots$, and so the number of parsed phrases is $\sqrt{2n} + O(1)$ and the number of binary digits used to encode the string by all three algorithms satisfies

$$\frac{\mathcal{L}(u_1^n)}{n} \leq \frac{\log_2 n}{\sqrt{2n}} + O\left(\frac{1}{\sqrt{n}}\right) = o\left(\frac{1}{\ln n}\right).$$

Theorems 4.3 and 4.4 give little or no information about the number of binary digits used in the encoding of u_1^n when the number of states, R, in the source model is too large in terms of n. For example, if $R = n$, the model can be chosen so that u_1^n is emitted with probability

1. In this case, $\hat{\mathcal{H}}_n = 0$, but nothing can be said about $\mathcal{L}(u_1^n)$ for any of the three encoding techniques. In general, the results in the previous section were asymptotic results; i.e., they hold as $I(u_1^n|s_0)$ tends to infinity. When $I(u_1^n|s_0)$ is small, the lower order terms are significant in our bounds on $\mathcal{L}(u_1^n)$ for all three algorithms.

A standard source model assumes that each output depends statistically only on the l previous output symbols. In this case, the first l symbols of u_1^n form the "initial state" and the empirical model is based on the $n-l$ $(l+1)$-tuples $u_i^{i+l}, 1 \leq i \leq n-l$. We let $\hat{\mathcal{H}}_{n,l}$ denote the empirical entropy of the resulting distribution. Then

$$\hat{I}(u_1^n|u_1^l) = (n-l)\hat{\mathcal{H}}_{n,l}.$$

Let R_l be the number of distinct source states that have occurred in u_1^n. Then $R_l \leq \min\{n-l, K^l\}$. We have

Theorem 4.5 *For large n and $l = O(1)$, we have the results of Theorems 4.3 and 4.4, with $\hat{\mathcal{H}}_n$ replaced by $\hat{\mathcal{H}}_{n,l}$ and R replaced by R_l.*

Note that as l increases, R_l increases. Hence, there is a point beyond which increasing the complexity of the source model provides no further insight about the number of code letters per source symbol needed to encode the source.

Chapter 5

Variable-to-Fixed Length Codes and Plurally Parsable Dictionaries

In Chapters 2 and 3, we investigated dictionaries with the property that every source string with positive probability can be uniquely parsed into a concatenation of dictionary entries with a final string that is a non-null prefix of a dictionary entry. In practice, many variable-to-fixed length codes use dictionaries that are *plurally parsable*, i.e., each source sequence can be segmented into a concatenation of dictionary entries in at least one way, and there exist source sequences that can be parsed into a concatenation of dictionary entries in two or more ways. At a parsing point, the most common rule for designating the next parsed phrase from a plurally parsable dictionary is to select the longest dictionary entry which is a prefix of the unparsed source output. This is the parsing rule that we'll assume throughout this chapter, unless we indicate otherwise. An example of a code employing a plurally parsable dictionary and this parsing rule is the British Admiralty's shutter telegraph (see Section 1.1).

In this chapter, we will restrict our attention to discrete, memoryless sources. To our knowledge, the design of good plurally parsable dictionaries has not been addressed, even for this simple class of sources. Since the Tunstall procedure is known to specify the optimal uniquely parsable dictionary of a given size and since the results in Chapter 2 and earlier papers in the literature have provided some insights into the performance of this algorithm, it is important to determine whether or not it is possible to construct a plurally parsable dictionary

which yields a significantly larger average length of a parsed string than that of the Tunstall dictionary of the same size.

5.1 Analysis of a Small Plurally Parsable Dictionary

To better grasp the possible advantages of a plurally parsable dictionary, let's consider a binary source with $p_0 \geq p_1$ and find the optimal dictionary of size three. We know that the Tunstall dictionary for this source is $\{00,\ 01,\ 1\}$ and the expected number of source letters per dictionary string is $1 + p_0$. For sources with very small p_1, the code is inefficient because the strings 01 and 1 are rarely used. Let 0^l denote the string of l zeroes. Any plurally parsable dictionary that has a larger expected length of a parsed string than the Tunstall dictionary must be of the form $\{0,\ 0^l,\ 1\}$ for some integer $l \geq 2$. To analyze this code, we study the steady-state behavior of a Markov chain with set of states $\{0^l,\ 1,\ 0_1,\ \ldots,\ 0_{l-1}\}$; here, 0^l and 1 are the states of the Markov chain when the next parsed source string is 0^l and 1, respectively, and for $1 \leq i \leq l-1$, 0_i is the state of the chain when the unparsed source output has exactly i zeroes before the next one. Let π_φ represent the steady-state probability that the Markov chain is in state φ. Then we have

$$\begin{aligned}
\pi_{0^l} &= {p_0}^l \cdot \pi_{0^l} + {p_0}^l \cdot \pi_1 \\
\pi_{0_{l-1}} &= {p_0}^{l-1} p_1 \cdot \pi_{0^l} + {p_0}^{l-1} p_1 \cdot \pi_1 \\
\pi_{0_i} &= {p_0}^i p_1 \cdot \pi_{0^l} + {p_0}^i p_1 \cdot \pi_1 + \pi_{0_{i+1}},\ 1 \leq i \leq l-2 \\
\pi_1 &= p_1 \cdot \pi_{0^l} + p_1 \cdot \pi_1 + \pi_{0_1}.
\end{aligned}$$

It is straightforward to verify that

$$\begin{aligned}
\pi_{0_i} &= \left(\frac{1}{{p_0}^{l-i} - 1}\right) \cdot \pi_{0^l},\ 1 \leq i \leq l-1 \\
\pi_1 &= \left(\frac{1}{{p_0}^l - 1}\right) \cdot \pi_{0^l}.
\end{aligned}$$

Since $\pi_{0^l} + \pi_1 + \sum_{i=1}^{l-1} \pi_{0_i} = 1$, we have that

$$\pi_{0^l} = \frac{p_1}{\left(\frac{1}{p_0}\right)^l - 1 - p_1(l-1)} \tag{5.1}$$

$$\pi_0 \doteq \sum_{i=1}^{l-1} \pi_{0_i} = \frac{\left(\frac{1}{p_0}\right)^{l-1} - 1 - p_1(l-1)}{\left(\frac{1}{p_0}\right)^l - 1 - p_1(l-1)} \tag{5.2}$$

$$\pi_1 = \frac{p_1\left(\left(\frac{1}{p_0}\right)^l - 1\right)}{\left(\frac{1}{p_0}\right)^l - 1 - p_1(l-1)} \tag{5.3}$$

and hence,

$$E[L] = \pi_0 + \pi_1 + l \cdot \pi_{0^l} = \frac{\left(\frac{1}{p_0}\right)^l - 1}{\left(\frac{1}{p_0}\right)^l - 1 - p_1(l-1)} = \left(1 - \frac{p_1(l-1)}{\left(\frac{1}{p_0}\right)^l - 1}\right)^{-1}. \tag{5.4}$$

To maximize the expected length of a parsed string, we would like to find the integer $l \geq 1$ that maximizes (5.4). This is the smallest integer $l \geq 2$ for which

$$\frac{l-1}{\left(\frac{1}{p_0}\right)^l - 1} \geq \frac{l}{\left(\frac{1}{p_0}\right)^{l+1} - 1},$$

or equivalently,

$$p_0{}^{l+1} + p_1 l - 1 = p_1\left(l - \sum_{i=0}^{l} p_0{}^i\right) \geq 0. \tag{5.5}$$

For example,

- $l = 2$ is optimal when

$$0.5 \leq p_0 \leq p(2) \doteq \frac{-1+\sqrt{5}}{2} \approx 0.618033988. \tag{5.6}$$

- $l = 3$ is optimal when

$$p(2) \leq p_0 \leq p(3) \doteq \frac{1}{3}\left(\sqrt[3]{\frac{3\sqrt{417}+61}{2}} - \sqrt[3]{\frac{3\sqrt{417}-61}{2}} - 1\right) \approx 0.810535713. \tag{5.7}$$

- $l = 4$ is optimal when

$$p(3) \leq p_0 \leq p(4) \approx 0.888179661. \tag{5.8}$$

The exact value of $p(4)$ is defined by

$$\begin{aligned} p(4) &= \frac{-\alpha + \sqrt{\alpha^2 - 4\beta}}{2} \\ \text{where} \qquad \gamma &= \frac{2}{3} + \sqrt[3]{\frac{\sqrt{27057}}{18} - \frac{155}{54}} - \sqrt[3]{\frac{\sqrt{27057}}{18} + \frac{155}{54}} \\ \alpha &= \frac{1 + \sqrt{1 - 4\gamma}}{2} \\ \beta &= \frac{1 - \alpha + \alpha^2 - \alpha^3}{1 - 2\alpha} \end{aligned}$$

From (5.4), it follows that the dictionary $\{0,\ 0^l,\ 1\}$ has a larger average length of a parsed string than the Tunstall dictionary $\{00,\ 01,\ 1\}$ when

$$\left(1 - \frac{p_1(l-1)}{\left(\frac{1}{p_0}\right)^l - 1}\right)^{-1} = \left(1 - \frac{{p_0}^l(l-1)}{\sum_{i=0}^{l-1} {p_0}^i}\right)^{-1} > 1 + p_0;$$

this reduces to the condition

$$\sum_{i=0}^{l-2} {p_0}^i - (l-2){p_0}^{l-1} - (l-1){p_0}^l < 0. \tag{5.9}$$

From (5.9), we see that the dictionary $\{0,\ 00,\ 1\}$ always has a smaller expected length of a parsed string than the Tunstall dictionary. When $l = 3$, condition (5.9) is satisfied when

$$p_0 > \sqrt[3]{\frac{11}{54} + \frac{\sqrt{177}}{72}} + \sqrt[3]{\frac{11}{54} - \frac{\sqrt{177}}{72}} - \frac{1}{6} \approx 0.829483541. \tag{5.10}$$

The dictionary $\{0,\ 0000,\ 1\}$ is better than $\{00,\ 01,\ 1\}$ when

$$p_0 > p^* \doteq \sqrt[3]{\frac{245}{1458} + \frac{\sqrt{741}}{162}} + \sqrt[3]{\frac{245}{1458} - \frac{\sqrt{741}}{162}} + \frac{1}{9} \approx 0.824122621. \tag{5.11}$$

Conditions (5.5) to (5.11) imply that the optimal plurally parsable dictionary has a larger average length of a parsed string than the Tunstall dictionary if and only if $p_0 > p^*$.

To understand the performance of the optimal plurally parsable dictionary as p_0 approaches one, we will maximize (5.4) and ignore the integer constraint. Let $c = \ln\left(\frac{1}{p_0}\right)$. We have that

$$\frac{d}{dl}\left(\frac{l-1}{\left(\frac{1}{p_0}\right)^l - 1}\right) = -\frac{e^{cl}\left(cl - c - 1 + e^{-cl}\right)}{\left(e^{cl}-1\right)^2}, \tag{5.12}$$

and hence, we would like to find the value of l, say l^*, for which the right-hand side of (5.12) is equal to zero; i.e.,

$$cl^* - c - 1 + e^{-cl^*} = 0. \tag{5.13}$$

For small values of $|x|$, $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + o(x^3)$. If cl^* is small, then

$$e^{-cl^*} = 1 - cl^* + \frac{(cl^*)^2}{2} - \frac{(cl^*)^3}{6} + o\left((cl^*)^3\right) \tag{5.14}$$

and

$$p_1 = 1 - e^{-c} = c - \frac{c^2}{2} + o\left(c^2\right). \tag{5.15}$$

Substituting (5.14) into (5.13), we see that cl^* satisfies

$$-c + \frac{(cl^*)^2}{2} - \frac{(cl^*)^3}{6} + o\left((cl^*)^3\right) = 0. \tag{5.16}$$

There is some $\delta > 0$ for which

$$cl^* = \sqrt{2c(1+\delta)} \tag{5.17}$$

and it follows from (5.16) that

$$\delta = \frac{\sqrt{2c}}{3} + o\left(\sqrt{c}\right). \tag{5.18}$$

For small values of $|x|$, $\sqrt{1+x} = 1 + \frac{x}{2} + o(x)$. Hence, (5.16) to (5.18) imply that

$$l^* = \sqrt{\frac{2}{c}\left(1 + \frac{\sqrt{2c}}{3} + o\left(\sqrt{c}\right)\right)} = \sqrt{\frac{2}{\ln\left(\frac{1}{p_0}\right)}} + \frac{1}{3} + \frac{o\left(\sqrt{\ln\left(\frac{1}{p_0}\right)}\right)}{\sqrt{\ln\left(\frac{1}{p_0}\right)}}. \tag{5.19}$$

For $l = l^*$,

$$
\begin{aligned}
E[L] &= \left(1 - \frac{p_1(l-1)}{\left(\frac{1}{p_0}\right)^l - 1}\right)^{-1} \\
&= \left(1 - \frac{p_1(1 - cl^* + c)}{c}\right)^{-1}, \qquad \text{by (5.13)} \\
&= \left(1 - \frac{1}{c} \cdot \left(c - \frac{c^2}{2} + o\left(c^2\right)\right) \cdot \left(1 - \sqrt{2c} + \frac{2c}{3} + o(c)\right)\right)^{-1}, \qquad \text{by (5.15) and (5.19)} \\
&= \left(\sqrt{2c}\left(1 - \frac{\sqrt{2c}}{12} + o\left(\sqrt{c}\right)\right)\right)^{-1} \\
&= \sqrt{\frac{1}{2\ln\left(\frac{1}{p_0}\right)}} + \frac{1}{12} + \frac{o\left(\sqrt{\ln\left(\frac{1}{p_0}\right)}\right)}{\sqrt{\ln\left(\frac{1}{p_0}\right)}}, \qquad \text{since } (1+x)^{-1} = 1 - x + o(x) \\
\pi_{0^l} &= \frac{p_1 \cdot E[L]}{\left(\frac{1}{p_0}\right)^l - 1} \\
&= \frac{E[L] - 1}{l^* - 1}, \qquad \text{by (5.4)} \\
&= \frac{1}{2} - \frac{7}{24}\sqrt{2\ln\left(\frac{1}{p_0}\right)} + o\left(\sqrt{\ln\left(\frac{1}{p_0}\right)}\right) \\
\pi_0 &= 1 - e^{cl^*} \cdot \pi_{0^l} \\
&= 1 - \frac{1}{1 - cl^* + c} \cdot \left(\frac{1}{2} - \frac{7}{24}\sqrt{2c} + o\left(\sqrt{c}\right)\right), \qquad \text{by (5.13)} \\
&= 1 - \left(1 + \sqrt{2c} + o\left(\sqrt{c}\right)\right) \cdot \left(\frac{1}{2} - \frac{7}{24}\sqrt{2c} + o\left(\sqrt{c}\right)\right) \\
&= \frac{1}{2} - \frac{5}{24}\sqrt{2\ln\left(\frac{1}{p_0}\right)} + o\left(\sqrt{\ln\left(\frac{1}{p_0}\right)}\right) \\
\pi_1 &= p_1 \cdot E[L] = \frac{1}{2}\sqrt{2\ln\left(\frac{1}{p_0}\right)} + o\left(\sqrt{\ln\left(\frac{1}{p_0}\right)}\right).
\end{aligned}
$$

Note that as p_0 approaches one, $E[L]$ grows without bound, in contrast to the expected length of the Tunstall dictionary. Furthermore, the dictionary entries are much closer to the ideal of being equiprobable than the entries of the Tunstall dictionary. This example suggests that a plurally parsable dictionary can be significantly better than the Tunstall dictionary of the same size when the dictionary is small and the source is very predictable.

5.2 Variation on a theme by Welch and Gallager

For the remainder of the chapter, we will restrict our attention to sources in which the self-information corresponding to the source symbols emitted has a *non-arithmetic* distribution. We saw in Chapter 2 that the average number of code symbols per source symbol of the Tunstall code will approach the entropy bound as the dictionary size increases. An important question about plurally parsable dictionaries is whether or not they can outperform the Tunstall code asymptotically. In this section, we will consider a code with a plurally parsable dictionary that seems likely to have a smaller redundancy than the Tunstall code. Our analysis will indicate that for large dictionary sizes, this code is inferior to the Tunstall code.

One of our objectives in studying variable-to-fixed length codes with known memory is to gain insight into the design of good universal codes. In this subsection, we will consider a variable-to-fixed length code that is motivated by Gallager's variant of LWZ, which we described in Section 4.1.

The Gallager code suggests that it would be interesting to consider a variable-to-fixed length code in which the dictionary is the set of strings formed by removing the last symbol from every entry of a Tunstall dictionary. More precisely, fix some self-information τ which is larger than the self-information of the least probable symbol. Suppose the Tunstall dictionary consists of all strings σ such that the self-information of σ is greater than τ and the self-informations of all proper prefixes of σ is less than or equal to τ. Then for our code, the corresponding dictionary contains strings with self-information upper-bounded by τ that have at least one single letter extension with self-information greater than τ; we will call this code the "Welch-Gallager variation" and abbreviate it as WGV.

For the rest of this chapter, we will model the generation of self-information by the source as a renewal process, where the interrenewal periods represent the self-information generated by the source upon emitting a symbol. A renewal epoch can then be viewed as the self-information of the corresponding source string. Note that if we use WGV and look at the generation of self-information starting from a parsing point, the resulting process is a *delayed* renewal process $\{\tilde{N}(t); t \geq 0\}$ because the probability distribution of the first symbol will generally differ from the probability distribution of subsequent symbols.

Let p_{min} denote the probability of the least probable source symbol. We have the following

result:

Lemma 5.1 *For each string σ with self-information $I(\sigma)$ in the interval $(\tau + \ln p_{min}, \tau)$, the string σ is an entry of the WGV dictionary; conversely, if σ is in the WGV dictionary, then $I(\sigma) \in [\tau + \ln p_{min}, \tau]$.*

Proof: We saw in Lemma 2.2 that if σ satisfies $I(\sigma) \in (\tau + \ln p_j, \tau)$, then $\sigma \circ j$ is in the Tunstall dictionary, and if $\sigma \circ j$ is in the Tunstall dictionary, then $I(\sigma) \in [\tau + \ln p_j, \tau]$. Note that the union of the intervals $(\tau + \ln p_j, \tau)$ is the interval $(\tau + \ln p_{min}, \tau)$. □

We introduce the following notation. Let

- M^{WGV} denote the size of the dictionary corresponding to the Welch-Gallager variation,
- M_j^{WGV} represent the number of WGV dictionary entries beginning with j, and
- $m(t) = E[N(t)]$.

To find the relationship between M_j^{WGV} and τ, we can use Lemma 5.1. Let σ be a dictionary entry, let j be the first symbol in σ, and define $\tilde{\sigma}$ by $\sigma = j \circ \tilde{\sigma}$. Then σ corresponds to a renewal in the delayed renewal process $\{\tilde{N}(t); t \geq 0\}$. After the symbol j is emitted, successive renewals of $\{\tilde{N}(t); t \geq 0\}$ correspond to renewals in a non-delayed renewal process. Note that $\tilde{\sigma}$ corresponds to a renewal in $\{N(t)\}$ renewal process $\{N(t); t \geq 0\}$, $\tilde{\sigma}$ corresponds to a renewal in this process. Note that $I(\tilde{\sigma}) = I(\sigma) + \ln p_j$. For the ordinary renewal process $\{N(t)\}$, the expected number of renewals in the interval $(t, t + dt]$ is

$$m(t + dt) - m(t) = \sum_{\tilde{\sigma}: I(\tilde{\sigma}) \in (t, t+dt]} P(\tilde{\sigma}). \tag{5.20}$$

For each string $\tilde{\sigma}$ with $I(\tilde{\sigma}) \in (t, t + dt]$, $e^{-t-dt} \leq P(\tilde{\sigma}) < e^{-t}$. Therefore, it follows from (5.20) that the number of strings $\tilde{\sigma}$ with $I(\tilde{\sigma}) \in (t, t + dt]$ is in the interval $(e^t[m(t + dt) - m(t)]$, $e^{t+dt}[m(t + dt) - m(t)]]$. Hence,

$$\int_{(\tau + \ln p_j + \ln p_{min})^+}^{(\tau + \ln p_j)^-} e^{x-\tau}\, dm(x) \;\leq\; M_j^{WGV} e^{-\tau} \;\leq\; \int_{(\tau + \ln p_j + \ln p_{min})^-}^{(\tau + \ln p_j)^+} e^{x-\tau}\, dm(x). \tag{5.21}$$

(5.21), Blackwell's Theorem (see Theorem A.1), and Lemma A.1 imply that

$$M_j^{WGV} e^{-\tau} \longrightarrow p_j \cdot \frac{1 - p_{min}}{\mathcal{H}}. \tag{5.22}$$

Since $M^{WGV} = \sum_{j=0}^{K-1} M_j^{WGV}$, by summing both sides of (5.22) over j and then taking the logarithm, we find that

Theorem 5.1 *As the dictionary size increases,*

$$\ln M^{WGV} - \tau \longrightarrow \ln\left(\frac{1 - p_{min}}{\mathcal{H}}\right).$$

Let H^{WGV} symbolize the expected self-information of a parsed string for the Welch-Gallager variation. We would like to understand the asymptotic relationship between H^{WGV} and τ. Let $Z(t)$ represent the information growth back from t to the most recent renewal epoch in the process $\{\tilde{N}(t)\}$; in the renewal theory literature, $Z(t)$ is called the *age* of the renewal process at epoch t. From [Gal96, §3.5 and §3.7], we have the following result.

Lemma 5.2 *As t increases,*

$$E[Z(t)] \longrightarrow \frac{\sum_{j=0}^{K-1} p_j(-\ln p_j)^2}{2\mathcal{H}}.$$

As a consequence, we find that

Corollary 5.1 *As the dictionary size increases,*

$$\tau - H^{WGV} \longrightarrow \frac{\sum_{j=0}^{K-1} p_j(-\ln p_j)^2}{2\mathcal{H}}.$$

Proof: The self-information of a sample dictionary entry can be interpreted as the last renewal epoch before τ in the process $\{\tilde{N}(t)\}$. Hence, $\tau - H^{WGV} = E[Z(\tau)]$, and the result follows from Lemma 5.2. □

Let $E[L^{WGV}]$ denote the expected length of a parsed string for the Welch-Gallager variation. The conservation of entropy (Theorem 3.1), Theorem 5.1, and Corollary 5.1 imply

Theorem 5.2 *As the dictionary size increases,*

$$\ln M^{WGV} - \mathcal{H} \cdot E[L^{WGV}] \longrightarrow C^{WGV} \doteq \ln\left(\frac{1-p_{min}}{\mathcal{H}}\right) + \sum_{j=0}^{K-1} \frac{p_j(-\ln p_j)^2}{2\mathcal{H}}$$

Therefore, $$\frac{\ln M^{WGV}}{E[L^{WGV}]} - \mathcal{H} \longrightarrow 0,$$

and so $$\ln M^{WGV} \cdot \left(\frac{\ln M^{WGV}}{E[L^{WGV}]} - \mathcal{H}\right) \longrightarrow \mathcal{H} \cdot C^{WGV}.$$

Recall from Theorem 2.5 that for the Tunstall code,

$$(\ln M^T) \cdot \left(\frac{\ln M^T}{E[L^T]} - \mathcal{H}\right) \longrightarrow \mathcal{H} \cdot C^T$$

where

$$C^T \doteq \ln\left(\frac{K-1}{\mathcal{H}}\right) - \sum_{j=0}^{K-1} \frac{p_j(-\ln p_j)^2}{2\mathcal{H}}.$$

In Appendix F, we establish the following result.

Lemma 5.3 *For all memoryless sources,* $C^{WGV} \geq C^T$.

Lemma 5.3 implies

Theorem 5.3 *For sufficiently large dictionary sizes, the expected length of a dictionary entry for the Tunstall code is larger than the average length of the Welch-Gallager variation dictionary with the same number of entries.*

As we mentioned in the last chapter, Miller and Wegman claimed that LZW empirically outperformed LZ '78 on English text, and our asymptotic bounds on the redundancy of LZ '78, LZW, and Gallager's code did not indicate that LZ '78 is a better code than Gallager's code when the length of the encoded source string approaches infinity. Therefore, Theorem 5.2 is a rather surprising result. How can we explain it? Miller and Wegman suggested that LZW is experimentally better than LZ '78 on English text because the dictionary size is smaller

and hence, fewer bits are required to encode each parsed phrase. We see from our analysis that the construction of a WGV dictionary from a Tunstall dictionary results in a decrease in the dictionary size by a factor asymptotically approaching $\frac{K-1}{1-p_{min}}$, but any improvement in compression achieved by reducing the dictionary size is more than offset by the loss of self-information of the last letter of a Tunstall dictionary entry. Note, however, that the dictionaries for the universal codes LZ '78 and LZW tend to differ from their counterparts for sources with memory near the leaves; hence, the loss of self-information of the last symbol in the universal coding setting may be less significant than the loss incurred by converting a Tunstall dictionary into a Welch-Gallager variation dictionary. Finally, Theorem 5.2 is an asymptotic result and Miller and Wegman pointed out that any advantage of using LZW over LZ '78 diminishes as the length of the encoded source string increases.

Chapter 6

Conclusions and Future Work

We have seen that techniques from discrete stochastic processes, particularly renewal theory, are powerful tools in the study of variable-to-fixed length codes. We first employed renewal theory in Chapter 2 to study the asymptotic behavior of the Tunstall code and our generalization of Tunstall codes to Markov sources. In Chapter 3, we stated and proved the "conservation of entropy" theorem which followed immediately from the strong law of large numbers. Using the conservation of entropy and renewal theory, we obtained better variable-to-fixed length codes for Markov sources. In the next chapter, we applied renewal theory to our investigation of the redundancy of the Lempel-Ziv incremental parsing rule and two of its variants. Our fifth chapter introduced plurally parsable dictionaries. We used a Markov chain to determine the optimal plurally parsable dictionary of size three for a binary, memoryless source and we employed renewal theory and the conservation of entropy to analyze a variable-to-fixed length code motivated by Gallager's variant of LZW.

A number of open issues remain. In Chapter 3, we found that if the weight vector $(\mathcal{H}(0), \cdots, \mathcal{H}(\mathcal{R}-1))$ is in the set of greedy vectors, then the asymptotically optimal generalized variable-to-fixed length code is the one that maximizes the entropies of the entries of the dictionaries. We conjecture that this result is always true.

The claims in [LS95] suggest that the constant factor in our bounds of the redundancy of LZ '78 and its variants can be reduced. It would be interesting to find a simple way to do so.

Finally, very little is known about plurally parsable dictionaries and their potential to outperform the Tunstall code. We hope to pursue this line of research.

Appendix A

If $g^+_{\psi,j}(x) = \begin{cases} p_{\psi,j}, & x \in [0, -\ln p_{\psi,j}] \\ 0, & x \notin [0, -\ln p_{\psi,j}] \end{cases}$, $g^-_{\psi,j}(x) = \begin{cases} p_{\psi,j}, & x \in (0, -\ln p_{\psi,j}) \\ 0, & x \notin (0, -\ln p_{\psi,j}) \end{cases}$,

$h^+_{\psi,j}(x) = \begin{cases} e^{-x}, & x \in [0, -\ln p_{\psi,j}] \\ 0, & x \notin [0, -\ln p_{\psi,j}] \end{cases}$, and $h^-_{\psi,j}(x) = \begin{cases} e^{-x}, & x \in (0, -\ln p_{\psi,j}) \\ 0, & x \notin (0, -\ln p_{\psi,j}) \end{cases}$, then (2.13) and (2.14) can be rewritten as

$$\sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \int_0^{\tau_s} g^-_{\psi,j}(\tau_s - x)\, dm^{(\psi,s)}(x) \le q^T_{s,r} \le \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \int_0^{\tau_s} g^+_{\psi,j}(\tau_s - x)\, dm^{(\psi,s)}(x) \tag{A.1}$$

$$\sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \int_0^{\tau_s} h^-_{\psi,j}(\tau_s - x)\, dm^{(\psi,s)}(x) \le Me^{-\tau_s} \le \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \int_0^{\tau_s} h^+_{\psi,j}(\tau_s - x)\, dm^{(\psi,s)}(x). \tag{A.2}$$

In order to prove Theorem 2.1, we employ the following well-known renewal theorem (see [Ros83, §3.4 and §3.5]).

Theorem A.1 *If $J^{(\psi)}_k, k > 1$ has a non-arithmetic distribution and if $h(t)$ is directly Riemann integrable, then*

$$\lim_{t\to\infty} \int_0^t h(t-x) dm^{(\psi,\psi_0)}(x) = \frac{1}{E[J^{(\psi)}_2]} \int_0^\infty h(t)dt.$$

We have the following relationship among $E[J^{(\psi)}_2]$, $\mathcal{H}$, and π_ψ.

Lemma A.1 $E[J^{(\psi)}_2] = \frac{\mathcal{H}}{\pi_\psi}$, $\psi \in \{0 \ldots R-1\}$.

Proof of Lemma A.1: If we view the emission of self-information as the semi-Markov process

defined immediately before Lemma 2.3, then $\mathcal{H}(\psi)$ is the mean information growth produced by this semi-Markov process from its entrance into state ψ until it makes a transition. $E[J_2^{(\psi)}]$ can now be interpreted as the information growth between successive transitions into state ψ. To complete the proof, we note that the long run proportion of self-information generated while the process is in state ψ can be shown (see [Ros83, §4.8]) to be equal to both $\frac{\mathcal{H}(\psi)}{E[J_2^{(\psi)}]}$ and $\frac{\pi_\psi \mathcal{H}(\psi)}{\sum_{r=0}^{R-1} \pi_r \mathcal{H}(r)} = \frac{\pi_\psi \mathcal{H}(\psi)}{\mathcal{H}}$. □

Proof of Theorem 2.1: As M increases, τ_s approaches infinity for each state s. Therefore, (2.15) follows directly from (A.1), Theorem A.1 and Lemma A.1. Equation (A.2), Theorem A.1 and Lemma A.1 imply that for all s,

$$\begin{aligned} Me^{-\tau_s} &\longrightarrow \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi}{\mathcal{H}} (1 - p_{\psi,j}) \\ &= \frac{1}{\mathcal{H}} \left(\sum_{\psi=0}^{R-1} \pi_\psi K - \sum_{j=0}^{K-1} \sum_{\psi=0}^{R-1} \pi_\psi p_{\psi,j} \right) \\ &= \frac{K-1}{\mathcal{H}}. \end{aligned}$$

□

Appendix B

Throughout this appendix, we assume the self-information associated with the source symbols emitted has an arithmetic distribution with period Λ. Furthermore, we presume that for all pairs of states ψ and symbols j, $p_{\psi,j} < 1$. If necessary, it is possible to change the alphabet and set of states in order to satisfy this assumption.

First, we consider sources that are *acyclic* in the sense that there is no integer D greater than one for which the self-information generated by the source between successive occurrences of state ψ is an integer multiple of $D\Lambda$ for all ψ. We use Blackwell's theorem (see [Ros83, §3.4 and §3.5]):

Theorem B.1 *If $J_k^{(\psi)}, k \geq 1$ has an arithmetic distribution with period Λ, then*

$$\lim_{m\to\infty} E[\textit{number of renewals at } \ m\Lambda] \ = \ \frac{\Lambda}{E[J_2^{(\psi)}]}.$$

Since $p_{\psi,j} < 1$ for all pairs of states ψ and symbols j, at time $m\Lambda$, there will either be no renewal or one rewenal. From Lemma A.1, $E[J_2^{(\psi)}] = \frac{\mathcal{H}}{\pi_\psi}$ (see [Gal96, §5.6]). Hence, it follows from Theorem B.1 and Lemma A.1 that as m increases, the probability of a renewal at $m\Lambda$ (for the process associated with returns to state ψ) is $\frac{\pi_\psi \Lambda}{\mathcal{H}}$. For all states ψ and starting states s,

$$\begin{aligned}
\frac{\pi_\psi \Lambda}{\mathcal{H}} &= \lim_{m\to\infty} \text{Prob}\{\sigma : I(\sigma|s) = m\Lambda \text{ and } S[s,\sigma] = i\} \\
&= \lim_{m\to\infty} \sum_{\sigma: I(\sigma|s)=m\Lambda, S[s,\sigma]=i} P(\sigma|s) \\
&= \lim_{m\to\infty} e^{-m\Lambda} \left| \{\sigma : I(\sigma|s) = m\Lambda \text{ and } S[s,\sigma] = i\} \right| . \qquad \text{(B.1)}
\end{aligned}$$

Consequently, we have the following result.

Lemma B.1 *Starting from any state s, as m increases, the number of strings with self-information $m\Lambda$ satisfies*

$$\lim_{m\to\infty} e^{-m\Lambda} \, |\{\sigma : I(\sigma|s) = m\Lambda\}| = \frac{\Lambda}{\mathcal{H}}.$$

For our dictionary tree, let X denote the number of intermediate nodes with self-information τ_s. Note that X changes every time an intermediate node is added to the dictionary tree. From Lemma B.1, we see that as the dictionary size increases,

$$\lim_{M\to\infty} \mathcal{H} e^{-\tau_s} X \in (0, \Lambda]. \tag{B.2}$$

We have the following counterpart to Theorem 2.1 and Corollaries 2.1 and 2.2 for acyclic, arithmetic sources.

Theorem B.2 *Assume the source is arithmetic with period Λ and acyclic. There is a code for which (2.15) remains valid. If X represents the number of intermediate nodes with self-information τ_s, then for this code, the analogues to (2.16) and Corollaries 2.1 and 2.2 are that for all s,*

$$\lim_{M\to\infty} \tau_s - \ln M + \ln\left(\frac{\Lambda}{e^\Lambda - 1} + \mathcal{H} e^{-\tau_s} X\right) = \ln\left(\frac{\mathcal{H}}{K-1}\right) \tag{B.3}$$

$$\lim_{\tau_s\to\infty} \mathcal{H} e^{-\tau_s} \cdot |\{\sigma : I(\sigma|s) < \tau_s\}| = \frac{\Lambda}{e^\Lambda - 1} \tag{B.4}$$

$$\lim_{\tau_s\to\infty} \mathcal{H} e^{-\tau_s} \cdot |\{\sigma : I(\sigma|s) \le \tau_s\}| = \frac{\Lambda e^\Lambda}{e^\Lambda - 1} \tag{B.5}$$

$$\textit{and} \quad \lim_{M\to\infty} H_s - \ln M + \ln\left(\frac{\Lambda}{e^\Lambda - 1} + \mathcal{H} e^{-\tau_s} X\right) + \frac{\Lambda}{2} - \mathcal{H} e^{-\tau_s} X =$$
$$\ln\left(\frac{\mathcal{H}}{K-1}\right) + \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}}, \tag{B.6}$$

respectively.

Proof: Assume throughout that we are considering the Tunstall dictionary for state s. We

will first establish (B.3) to (B.6). Let w_ψ represent the set of intermediate nodes with self-information τ_s that leave the source in state ψ; i.e.,

$$w_\psi \;=\; \{\sigma :\; I(\sigma|s) = \tau_s,\; S[s,\sigma] = i,\; \sigma \text{ is an intermediate node}\}.$$

Then X, the number of intermediate nodes with self-information τ_s, satisfies

$$X \;=\; \sum_{\psi=0}^{R-1} |\, w_\psi \,| \,. \tag{B.7}$$

Define $\Omega_{m,i}$ as the set of strings with self-information $\tau_s - m\Lambda$ that drive the source to state ψ; i.e.,

$$\Omega_{m,i} \;=\; \{\sigma :\; I(\sigma|s) = \tau_s - m\Lambda,\; S[s,\sigma] = i\}. \tag{B.8}$$

Since all strings with self-information less than τ_s are intermediate nodes in the dictionary tree and there are X intermediate nodes with self-information τ_s, the total number of intermediate nodes in the tree is

$$\frac{M-1}{K-1} \;=\; X \;+\; \sum_{\psi=0}^{R-1} \sum_{m=1}^{\frac{\tau_s}{\Lambda}} |\, \Omega_{m,i} \,| \,. \tag{B.9}$$

From (B.1), for any fixed $m \geq 0$,

$$\lim_{\tau_s \to \infty} |\, \Omega_{m,i} \,|\, e^{-(\tau_s - m\Lambda)} \;=\; \frac{\pi_\psi \Lambda}{\mathcal{H}}, \tag{B.10}$$

and so

$$\lim_{\tau_s \to \infty} \sum_{m=1}^{\frac{\tau_s}{\Lambda}} |\, \Omega_{m,i} \,|\, e^{-\tau_s} \;=\; \frac{\pi_\psi \Lambda}{\mathcal{H}} \sum_{m=1}^{\infty} e^{-m\Lambda} \;=\; \frac{\pi_\psi \Lambda}{\mathcal{H}(e^\Lambda - 1)}, \tag{B.11}$$

proving (B.4). (B.5) follows from (B.4) and Lemma B.1. (B.9) and (B.11) imply that

$$\lim_{M \to \infty} e^{-\tau_s} \left(\frac{M}{K-1} - X \right) \;=\; \frac{\Lambda}{\mathcal{H}(e^\Lambda - 1)},$$

which is equivalent to (B.3).

Next we demonstrate (B.6). We have that

$$H_s \;=\; \sum_{\text{leaves } \sigma \circ j} P(\sigma \circ j|s) \cdot [-\ln P(\sigma \circ j|s)].$$

Using Lemma 2.2 to characterize the leaves $\phi = \sigma \circ j$ in the Tunstall tree, we find that

$$\begin{aligned}
H_s &= -\sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1}\sum_{m=1}^{-\left(\frac{\ln p_{\psi,j}+\Lambda}{\Lambda}\right)}\sum_{\sigma\in\Omega_{m,i}} P(\sigma\circ j|s)\ln P(\sigma\circ j|s) \\
&\quad -\sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1}\sum_{\sigma\in w_\psi} P(\sigma\circ j|s)\ln P(\sigma\circ j|s) \;-\; \sum_{\psi=0}^{R-1}\sum_{\phi\in\Omega_{0,i},\ \phi \text{ is a leaf}} P(\phi|s)\ln P(\phi|s) \\
&= \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1}\sum_{m=1}^{-\left(\frac{\ln p_{\psi,j}+\Lambda}{\Lambda}\right)} |\,\Omega_{m,i}\,|\, p_{\psi,j}e^{-(\tau_s-m\Lambda)}\cdot(\tau_s - m\Lambda - \ln p_{\psi,j}) \\
&\quad +\sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} |\,w_\psi\,|\, p_{\psi,j}e^{-\tau_s}\cdot(\tau_s - \ln p_{\psi,j}) \;+\; \sum_{\psi=0}^{R-1}(|\,\Omega_{0,i}\,| - |\,w_\psi\,|)e^{-\tau_s}\cdot\tau_s \\
&= \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1}\sum_{m=1}^{-\left(\frac{\ln p_{\psi,j}+\Lambda}{\Lambda}\right)} |\,\Omega_{m,i}\,|\, p_{\psi,j}e^{-(\tau_s-m\Lambda)}\cdot(\tau_s - m\Lambda - \ln p_{\psi,j}) \\
&\quad +\sum_{\psi=0}^{R-1} |\,w_\psi\,|\, e^{-\tau_s}\cdot\mathcal{H}(\psi) \;+\; \sum_{\psi=0}^{R-1} |\,\Omega_{0,i}\,|\, e^{-\tau_s}\cdot\tau_s. \qquad \text{(B.12)}
\end{aligned}$$

For each state ψ, let $\pi_\psi(s,\tau_s)$ represent the fraction of strings with self-information τ_s that drive the source to state ψ; i.e.,

$$\pi_\psi(s,\tau_s) \;=\; \frac{|\,\{\sigma : I(\sigma|s) = \tau_s \text{ and } S[s,\sigma] = i\}\,|}{|\,\{\sigma : I(\sigma|s) = \tau_s\}\,|}. \qquad \text{(B.13)}$$

It is possible to select the sets w_ψ so that for all states ψ,

$$\pi_\psi(s,\tau_s)\cdot X \;-\; 1 \;\le\; |\,w_\psi\,| \;\le\; \pi_\psi(s,\tau_s)\cdot X \;+\; 1. \qquad \text{(B.14)}$$

From (B.1) and Lemma B.1, we have that

$$\lim_{M\to\infty}\pi_\psi(s,\tau_s) \;=\; \pi_\psi. \qquad \text{(B.15)}$$

It follows from (B.10), (B.14), and (B.15) that

$$
\begin{aligned}
H_s \;-\; & \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \sum_{m=1}^{-\left(\frac{\ln p_{\psi,j}+\Lambda}{\Lambda}\right)} |\Omega_{m,i}|\, p_{\psi,j} e^{-(\tau_s - m\Lambda)} \cdot (\tau_s - m\Lambda - \ln p_{\psi,j}) \\
& - \sum_{\psi=0}^{R-1} |w_\psi|\, e^{-\tau_s} \cdot \mathcal{H}(\psi) - \sum_{\psi=0}^{R-1} |\Omega_{0,i}|\, e^{-\tau_s} \cdot \tau_s \\
\longrightarrow \quad H_s \;-\; & \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \sum_{m=1}^{-\left(\frac{\ln p_{\psi,j}+\Lambda}{\Lambda}\right)} \frac{\pi_\psi \Lambda}{\mathcal{H}} \cdot p_{\psi,j}(\tau_s - m\Lambda - \ln p_{\psi,j}) \\
& - \sum_{\psi=0}^{R-1} \pi_\psi X e^{-\tau_s} \cdot \mathcal{H}(\psi) \;-\; \sum_{\psi=0}^{R-1} \frac{\pi_\psi \Lambda}{\mathcal{H}} \cdot \tau_s \\
= \quad H_s \;-\; & \tau_s \;-\; \sum_{\psi=0}^{R-1}\sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j}}{2\mathcal{H}} (-\ln p_{\psi,j})^2 \;+\; \frac{\Lambda}{2} \;-\; \mathcal{H} X e^{-\tau_s}. \qquad \text{(B.16)}
\end{aligned}
$$

Combining (B.3), (B.12), and (B.16) gives (B.6).

Finally, we consider the asymptotic behavior of $q_{s,r}^T$. By definition,

$$
q_{s,r}^T \;=\; \sum_{\text{leaves } \phi: S[s,\phi]=r} P(\phi|s). \qquad \text{(B.17)}
$$

As in our analysis of H_s, we will break up the right-hand side of (B.17) into three terms. The first accounts for those leaves $\phi = \sigma \circ j$ for which $I(\sigma|s) < \tau_s$ and $I(\phi|s) > \tau_s$, and the second and third terms incorporate the effect of the leaves having a proper prefix with self-information τ_s and the leaves with self-information τ_s, respectively. We have that

$$
\begin{aligned}
q_{s,r}^T \;=\; & \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \sum_{m=1}^{-\left(\frac{\ln p_{\psi,j}+\Lambda}{\Lambda}\right)} \sum_{\sigma \in \Omega_{m,i}} P(\sigma \circ j|s) \\
& + \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \sum_{\sigma \in w_\psi} P(\sigma \circ j|s) \;+\; \sum_{\phi \in \Omega_{0,r},\ \phi \text{ is a leaf}} P(\phi|s) \\
= \; & \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \sum_{m=1}^{-\left(\frac{\ln p_{\psi,j}+\Lambda}{\Lambda}\right)} |\Omega_{m,i}|\, p_{\psi,j} e^{-(\tau_s - m\Lambda)}
\end{aligned}
$$

$$+ \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} |w_\psi| \, p_{\psi,j} e^{-\tau_s} + (|\Omega_{0,r}| - |w_r|) e^{-\tau_s} \tag{B.18}$$

$$\longrightarrow \quad - \frac{\pi_\psi p_{\psi,j} \ln p_{\psi,j}}{\mathcal{H}} + [\sum_{\psi=0}^{R-1} |w_\psi| \, q_{\psi,r} - |w_r|] e^{-\tau_s}, \tag{B.19}$$

using (B.11).

From (B.14) and (B.15), we know that for all pairs of states ψ and r, as the dictionary size increases,

$$\limsup_{M\to\infty} | \, |w_\psi| \, q_{\psi,r} - \pi_\psi q_{\psi,r} X| \; \leq \; q_{\psi,r},$$

and hence,

$$\limsup_{M\to\infty} | \sum_{\psi=0}^{R-1} |w_\psi| \, q_{\psi,r} - |w_r| \, | \; = \; \limsup_{M\to\infty} | \left(\sum_{\psi=0}^{R-1} [|w_\psi| \, q_{\psi,r} - \pi_\psi q_{\psi,r} X|] \right) + (\pi_r X - |w_r|)| \; \leq \; 2. \tag{B.20}$$

(2.15) follows from (B.19) and (B.20). □

Lemma 2.4, Theorem 2.3, and (2.27) remain valid. Hence, combining these results with (B.6) yields

Theorem B.3 *Assume the source is arithmetic with period Λ and acyclic. As the number of entries for the dictionary for each state increases,*

$$\lim_{M\to\infty} \ln M \; - \; E[L^T] \cdot \mathcal{H} \; - \; \ln \left(\frac{\Lambda}{e^\Lambda - 1} \; + \; \mathcal{H} e^{-\tau_s} X \right) \; - \; \frac{\Lambda}{2} \; + \; \mathcal{H} e^{-\tau_s} X$$
$$= \; \ln \left(\frac{K-1}{\mathcal{H}} \right) \; - \; \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2\mathcal{H}}$$

and so $$\lim_{M\to\infty} (\ln M) \cdot \left(\frac{\ln M}{E[L^T]} - \mathcal{H} \right) - \mathcal{H} \ln \left(\frac{\Lambda}{e^\Lambda - 1} + \mathcal{H} e^{-\tau_s} X \right) - \frac{\Lambda \mathcal{H}}{2} + \mathcal{H}^2 e^{-\tau_s} X$$
$$= \; \mathcal{H} \ln \left(\frac{K-1}{\mathcal{H}} \right) - \sum_{\psi=0}^{R-1} \sum_{j=0}^{K-1} \frac{\pi_\psi p_{\psi,j} (-\ln p_{\psi,j})^2}{2}.$$

Since Theorem 2.4 still holds, it is possible to combine Theorems 2.4 and B.3 to find the

counterparts to the results in Section 2.3.

We next consider the situation in which the set of source states is *cyclic*. Let D be the maximum integer for which the self-information generated by the source between consecutive occurrences of any given state is an integer multiple of $D\Lambda$. As an example, we consider

Example B.1: Suppose the source has states $\{0, 1, 2\}$ and always emits a symbol corresponding to the next state. Let

$$\mathbf{F} = \begin{pmatrix} 0.25 & 0.25 & 0.5 \\ 0.25 & 0.25 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}. \tag{B.21}$$

It is straightforward to demonstrate that this source is arithmetic with period $\ln 2$ and the self-information generated between consecutive occurrences of any given state is an integer multiple of $2 \ln 2$.

The set of states can be partitioned into $1 < d \leq D$ categories with the property that state ψ is in category c, denoted $\mathcal{C}(c)$, where $c \in \{0, \ldots, D-1\}$, if and only if the (possibly delayed) renewal process for state ψ (from starting state s) has renewals at epochs of the form $mD\Lambda + c\Lambda$. The counterpart to Theorem B.1 is

Theorem B.4 *If state ψ is in category c and $J_k^{(\psi)}, k > 1$ has an arithmetic distribution with period $D\Lambda$, then for $c' \in \{0, \ldots, D-1\}$,*

$$\lim_{m\to\infty} E[\textit{number of renewals at } \ mD\Lambda + c\Lambda] = \begin{cases} \frac{D\Lambda}{E[J_2^{(\psi)}]}, & c' = c \\ 0, & c' \neq c \end{cases}. \tag{B.22}$$

Since Lemma A.1 continues to be valid, the analogue to (B.1) is

$$\begin{aligned} \frac{\pi_\psi D\Lambda}{\mathcal{H}} &= \lim_{m\to\infty} \text{Prob}\{\sigma : I(\sigma|s) = mD\Lambda + c\Lambda \text{ and } S[s,\sigma] = i \in \mathcal{C}(c)\} \\ &= \lim_{m\to\infty} \sum_{\sigma: I(\sigma|s)=mD\Lambda+c\Lambda,\ S[s,\sigma]=i\in\mathcal{C}(c)} P(\sigma|s) \\ &= \lim_{m\to\infty} e^{-mD\Lambda-c\Lambda} \,|\{\sigma : I(\sigma|s) = mD\Lambda + c\Lambda \text{ and } S[s,\sigma] = i \in \mathcal{C}(c)\}|. \end{aligned} \tag{B.23}$$

For a given category c and a given τ_s, define $\gamma(\tau_s, c)$ as the smallest positive integer such

that $\frac{\tau_s}{\Lambda} - \gamma(\tau_s, c) \equiv c \pmod{D}$; i.e.,

$$\gamma(\tau_s, c) = D - [c - \frac{\tau_s}{\Lambda} \pmod{D}]. \tag{B.24}$$

For any statement A, let 1_A denote the function that is 1 if A is true and 0 if A is false. We have the following analogue to Theorem 2.1 and Corollary 2.1 for cyclic sources.

Theorem B.5 *Assume the source is arithmetic with period Λ and cyclic. Let $D > 1$ be the maximum integer for which the self-information generated by the source between consecutive occurrences of any given state is an integer multiple of $D\Lambda$. If X represents the number of intermediate nodes with self-information τ_s then there is a code such that for all states s,*

$$\lim_{M\to\infty} q_{s,r}^T + \sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j} \ln p_{\psi,j}}{\mathcal{H}} - \frac{\sum_{i\in\mathcal{C}(\frac{\tau_s}{\Lambda} \pmod D)} \pi_\psi f_{\psi,r} e^{-\tau_s} X}{\sum_{\psi\in\mathcal{C}(\frac{\tau_s}{\Lambda} \pmod D)} \pi_\psi}$$
$$- 1_{r\in\mathcal{C}(\frac{\tau_s}{\Lambda} \pmod D)} \cdot \left(\frac{\pi_r D\Lambda}{\mathcal{H}} - \frac{\pi_r e^{-\tau_s} X}{\sum_{\psi\in\mathcal{C}(\frac{\tau_s}{\Lambda} \pmod D)} \pi_\psi} \right)$$
$$= \sum_{c=0}^{D-1} \sum_{i\in\mathcal{C}(c)} \sum_{j:S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j}}{\mathcal{H}} \Lambda \left(D - 1 - \gamma(\tau_s, c) - \left(-\frac{\ln p_{\psi,j}}{\Lambda} - 1 - \gamma(\tau_s, c) \pmod D \right) \right) \tag{B.25}$$

$$\lim_{M\to\infty} \tau_s - \ln M + \ln \left(\frac{D\Lambda}{e^{D\Lambda} - 1} \sum_{c=0}^{D-1} \sum_{i\in\mathcal{C}(c)} \pi_\psi e^{\Lambda(c - \frac{\tau_s}{\Lambda} \pmod D)} + \mathcal{H} e^{-\tau_s} X \right) = \ln \left(\frac{\mathcal{H}}{K-1} \right) \tag{B.26}$$

$$\lim_{\tau_s\to\infty} \mathcal{H} e^{-\tau_s} \cdot |\{\sigma : I(\sigma|s) < \tau_s\}| = \sum_{i=0}^{R-1} \frac{\pi_\psi D\Lambda e^{-\gamma(\tau_s,c)\Lambda}}{1 - e^{-D\Lambda}} \tag{B.27}$$

$$\lim_{\tau_s\to\infty} \mathcal{H} e^{-\tau_s} \cdot |\{\sigma : I(\sigma|s) \le \tau_s\}| = \sum_{i\in\mathcal{C}(\frac{\tau_s}{\Lambda} \pmod D)} \frac{\pi_\psi D\Lambda}{1 - e^{-D\Lambda}}$$
$$+ \sum_{i\notin\mathcal{C}(\frac{\tau_s}{\Lambda} \pmod D)} \frac{\pi_\psi D\Lambda e^{-\gamma(\tau_s,c)\Lambda}}{1 - e^{-D\Lambda}}. \tag{B.28}$$

Proof: To prove (B.25), we note that (B.18) remains valid. By (B.22) to (B.24), we have that

$$\sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} \sum_{m=1}^{-\left(\frac{\ln p_{\psi,j} + \Lambda}{\Lambda}\right)} |\Omega_{m,i}|\, p_{\psi,j} e^{-(\tau_s - m\Lambda)}$$

$$= \sum_{c=0}^{D-1} \sum_{i \in \mathcal{C}(c)} \sum_{j:S[\psi,j]=r} \sum_{m=1}^{\lfloor \frac{-\ln p_{\psi,j} - (1+\gamma(\tau_s,c))\Lambda}{D\Lambda} \rfloor} |\Omega_{\gamma(\tau_s,c)\Lambda+mD\Lambda,i}|\, p_{\psi,j} e^{-(\tau_s - \gamma(\tau_s,c)\Lambda - mD\Lambda)}$$

$$\longrightarrow \sum_{c=0}^{D-1} \sum_{i \in \mathcal{C}(c)} \sum_{j:S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j} D\Lambda}{\mathcal{H}} \cdot \lfloor \frac{-\ln p_{\psi,j}(D - 1 - \gamma(\tau_s,c))\Lambda}{D\Lambda} \rfloor. \tag{B.29}$$

Using the observation that for integers A and B,

$$B\lfloor \frac{A}{B} \rfloor \;=\; A \;-\; [A \pmod B],$$

we can use (B.29) to obtain the first two expressions on the right-hand side of (B.25).

To finish establishing (B.25), we note that (B.22) implies

$$\sum_{\psi=0}^{R-1} \sum_{j:S[\psi,j]=r} |w_\psi|\, p_{\psi,j} e^{-\tau_s} \;+\; (|\Omega_{0,r}| - |w_r|) e^{-\tau_s}$$

$$= \sum_{i \in \mathcal{C}\left(\frac{\tau_s}{\Lambda} \pmod D\right)} q_{\psi,r} |w_\psi|\, e^{-\tau_s} \;+\; 1_{r \in \mathcal{C}\left(\frac{\tau_s}{\Lambda} \pmod D\right)} \cdot (|\Omega_{0,r}| - |w_r|) e^{-\tau_s}. \tag{B.30}$$

(B.13) and (B.14) still hold. The counterpart to (B.15) is that as the dictionary size increases,

$$\pi_\psi(s,\tau_s) \longrightarrow \begin{cases} \frac{\pi_\psi}{\sum_{\psi \in \mathcal{C}\left(\frac{\tau_s}{\Lambda} \pmod D\right)} \pi_\psi}, & i \in \mathcal{C}\left(\frac{\tau_s}{\Lambda} \pmod D\right) \\ 0, & i \notin \mathcal{C}\left(\frac{\tau_s}{\Lambda} \pmod D\right). \end{cases} \tag{B.31}$$

(B.23), (B.30), and (B.31) imply the last two terms on the right-hand side of (B.25).

To demonstrate (B.26), we observe that (B.9) still holds and that if ψ is in category c, the counterpart to (B.11) is

$$\lim_{\tau_s \to \infty} \sum_{m=1}^{\frac{\tau_s}{\Lambda}} |\Omega_{m,i}|\, e^{-\tau_s} \;=\; \frac{\pi_\psi D\Lambda}{\mathcal{H}} \sum_{m=0}^{\infty} e^{-(\gamma(\tau_s,c) + Dm\Lambda)} \;=\; \frac{\pi_\psi D\Lambda e^{-\gamma(\tau_s,c)\Lambda}}{\mathcal{H}(1 - e^{-D\Lambda})}, \tag{B.32}$$

proving (B.27). (B.27) and (B.23) imply (B.28). (B.26) follows from (B.9), (B.32), and (B.24).□

It is possible to find the analogues to (B.6), Theorem B.3, and Section 2.3, but the resulting equations are complicated and do not provide insights; we omit them here.

Appendix C

Proof of Lemma 2.4: Let us now view the process by which the source generates self-information as an additive Markov process (see [NN87]). Suppose the state is initially ψ_0 and u_k is the k^{th} letter emitted from the source. Let ψ_k and $\mathcal{I}_k$ denote the state of the source and the cumulative self-information generated by the source after the k^{th} symbol has been issued, respectively. Then ψ_k is determined by the recursive rule

$$\psi_k \;=\; S[\psi_{k-1}, u_k] \tag{C.1}$$

and $\mathcal{I}_k$ can be written

$$\mathcal{I}_k \;=\; \sum_{i=1}^{k} I(u_\psi|\psi_{i-1}). \tag{C.2}$$

We wish to relate $\mathcal{I}_k$ to H_s and $\mathcal{L}_s^T$ for all $s \in \{0, \ldots, R-1\}$. Define $\mathcal{J}_k$ by

$$\mathcal{J}_k \;\equiv\; \mathcal{I}_k - k \cdot \mathcal{H} + \mathcal{W}_{\psi_k} - \mathcal{W}_{\psi_0}, \tag{C.3}$$

where $\mathcal{W}$ is defined in (2.21). We can rewrite (C.3) as

$$\mathcal{J}_k \;=\; \mathcal{J}_{k-1} + I(u_k|\psi_{k-1}) - \mathcal{H} + \mathcal{W}_{\psi_k} - \mathcal{W}_{\psi_{k-1}}. \tag{C.4}$$

Note that

$$\begin{aligned} E[I(u_k|\psi_{k-1}) - \mathcal{H} + \mathcal{W}_{\psi_k} - \mathcal{W}_{\psi_{k-1}} \mid \psi_{k-1} = s] \\ = \;\; \mathcal{H}(s) - \mathcal{H} - \mathcal{W}_s + E[\mathcal{W}_{\psi_k} \mid \psi_{k-1} = s] \end{aligned}$$

$$
\begin{aligned}
&= \mathcal{H}(s) - \mathcal{H} - \mathcal{W}_s + \sum_{r=0}^{R-1} f_{s,r}\mathcal{W}_r \\
&= 0, \quad \text{by (2.21).}
\end{aligned} \tag{C.5}
$$

Therefore, the sequence $\{\mathcal{J}_k; k \geq 1\}$ is a martingale with $E[\mathcal{J}_1] = 0$; furthermore $\{\mathcal{J}_k; k \geq 1\}$ is a martingale relative to the joint process $\{\mathcal{J}_k, \psi_k; k \geq 1\}$ (see [Gal96, §7]). In our additive Markov process with $\psi_0 = s$, the length of a sample entry ϕ of the dictionary for state s is given by the stopping rule $L_s = \min\{k \geq 1 : \mathcal{I}_k \geq \tau_s$ **and** ϕ is a leaf$\}$. The optional stopping theorem (see [Fel71] and [Gal96, §7]) implies that for all $s \in \{0, \ldots, R-1\}$,

$$
E[\mathcal{I}_{L_s} - L_s \cdot \mathcal{H} + \mathcal{W}_{\psi_{L_s}} - \mathcal{W}_s \mid \psi_0 = s] = 0, \tag{C.6}
$$

and since $H_s = E[\mathcal{I}_{L_s}]$ and $\mathcal{L}_s^T = E[L_s]$,

$$
\begin{aligned}
H_s &= \mathcal{L}_s^T \cdot \mathcal{H} + \mathcal{W}_s - E[\mathcal{W}_{\psi_{L_s}} \mid \psi_0 = s] \\
&= \mathcal{L}_s^T \cdot \mathcal{H} + \mathcal{W}_s - \sum_{r=0}^{R-1} q_{s,r}^T \mathcal{W}_r.
\end{aligned} \tag{C.7}
$$

Let $\delta_s = \mathcal{W}_s - \sum_{r=0}^{R-1} q_{s,r}^T \mathcal{W}_r$. As the dictionary size increases, we have from Theorem 2.1 that $q_{s,r}^T \longrightarrow \rho_r^*$ for all states s and r and hence,

$$
\delta_s \longrightarrow \mathcal{W}_s - \sum_{r=0}^{R-1} \rho_r^* \mathcal{W}_r. \tag{C.8}
$$

Appendix D

As in the previous appendix, we view the process by which the source generates self-information as an additive Markov process and we continue to use the notation defined in Chapters 1 and 2 and Appendix C. Now let

$$g_{s,r}(x) \;=\; E\left(e^{xI(u_k)} \mid \psi_{k-1} = s,\ \psi_k = r\right)$$

and let $\Upsilon(x) = [f_{s,r}g_{s,r}(x)]$. Since $\Upsilon(x)$ is a non-negative, irreducible matrix, the Perron-Frobenius theorem (see [Gal96, §4.4]) demonstrates that $\Upsilon(x)$ has a positive real eigenvalue $\hbar(x)$ with an associated positive right eigenvalue $\nu(x)$ that is unique up to a scale factor.

The process $\{\mathcal{M}_k(x);\ k \geq 1\}$ defined by

$$\mathcal{M}_k(x) \;=\; \frac{e^{xI(u_k)}\ \nu_{\psi_k}(x)}{[\hbar(x)]^k\ \nu_{\psi_0}(x)}$$

is shown in [Gal96, §7.6, Example 8]) to be a product type Martingale for all real numbers x. Since the entries of a dictionary are finite in length and since parsing can be viewed as a stopping rule, it follows from Theorem 6, equation (97), and exercise 7.25 of [Gal96, §7]) that for all dictionaries $\mathcal{D}_s$,

$$E_{\sigma\in\mathcal{D}_s}\left(\frac{e^{xI(\sigma|s)}\ \nu_{S[s,\sigma]}(x)}{[\hbar(x)]^{l(\sigma)}\ \nu_s(x)}\right) \;=\; 1, \tag{D.1}$$

where $l(\sigma)$ is the length of the string σ. We have the following result.

Theorem D.1 *Let* $\mathbf{Q} = [q_{s,r}]$ *represent the transition probability matrix for the state of the source from one parsing point to the next.*

- *If the Markov chain corresponding to* $\mathbf{Q}$ *has a single recurrent class of states, let* ρ_r

denote the steady-state probability of being in state r at a parsing point. Then for all real numbers x,

$$\sum_{r=0}^{R-1} \nu_r(x) \left(\rho_r - \sum_{s=0}^{R-1} \rho_s q_{s,r} \cdot E_{\sigma \in \mathcal{D}_s} \left(\frac{e^{xI(\sigma|s)}}{[\hbar(x)]^{l(\sigma)}} \mid S[s,\sigma] = r \right) \right) = 0. \tag{D.2}$$

- *Otherwise, the Markov chain will eventually enter and remain in one of the recurrent classes of states, say Γ. Let $\rho_r(\Gamma)$ symbolize the steady-state probability of being in state r at a parsing point given that the chain is in the class of states Γ. Then for all x,*

$$\sum_{r=0}^{R-1} \nu_r(x) \left(\rho_r(\Gamma) - \sum_{s=0}^{R-1} \rho_s(\Gamma) q_{s,r} \cdot E_{\sigma \in \mathcal{D}_s} \left(\frac{e^{xI(\sigma|s)}}{[\hbar(x)]^{l(\sigma)}} \mid S[s,\sigma] = r \right) \right) = 0. \tag{D.3}$$

Proof: By (D.1), for all states s and real numbers x, we have that

$$\sum_{r=0}^{R-1} q_{s,r} \nu_r(x) \cdot E_{\sigma \in \mathcal{D}_s} \left(\frac{e^{xI(\sigma|s)}}{[\hbar(x)]^{l(\sigma)}} \mid S[s,\sigma] = r \right) = \nu_s(x). \tag{D.4}$$

The theorem follows by multiplying both sides of (D.4) by ρ_s or $\rho_s(\Gamma)$ and summing over s.□

Comment: The observations at the end of Section 3.1 on the generality of Theorem 3.1 are valid for Theorem D.1 as well.

We now demonstrate that Theorem D.1 implies the conservation of entropy.

Alternate proof of Theorem 3.1: By taking the derivative of both sides of (D.2) or (D.3) with respect to x and setting the resulting equations equal to zero, we obtain the conservation of entropy. To see this, we will have to calculate $\nu_r(0)$, $\hbar(0)$, and $\frac{d\hbar(x)}{dx}|_{x=0}$. Note that $\Upsilon(0) = \mathbf{F}$ and $\frac{d\Upsilon(x)}{dx}|_{x=0} = [f_{s,r} E(I(u_k) \mid \psi_{k-1} = s,\ \psi_k = r)]$. Since

$$\Upsilon(x)\nu(x) = \hbar(x)\nu(x), \tag{D.5}$$

for all real numbers x, we have that $\hbar(0) = 1$, and we can choose $\nu(0) = \mathbf{e}$, the vector with every component equal to one.

Taking the derivative of both sides of (D.5), we find that

$$\begin{aligned} \Upsilon(x)\cdot\frac{d\nu(x)}{dx}+\frac{d\Upsilon(x)}{dx}\nu(x) &= \hbar(x)\frac{d\nu(x)}{dx}+\frac{d\hbar(x)}{dx}\nu(x) \\ \text{and so } \mathbf{F}\cdot\frac{d\nu(x)}{dx}|_{x=0}+(\mathcal{H}(0)\ldots\mathcal{H}(R-1))' &= \frac{d\nu(x)}{dx}|_{x=0}+\frac{d\hbar(x)}{dx}|_{x=0}\cdot\mathbf{e} \end{aligned} \tag{D.6}$$

According to (2.21), $\frac{d\hbar(x)}{dx}|_{x=0}=\mathcal{H}$. Taking the derivative of both sides of (D.2) with respect to x, we see that

$$\begin{aligned} &\sum_{r=0}^{R-1}\frac{d\nu_r(x)}{dx}\left(\rho_r-\sum_{s=0}^{R-1}\rho_s q_{s,r}\cdot E_{\sigma\in\mathcal{D}_s}\left(\frac{e^{xI(\sigma|s)}}{[\hbar(x)]^{l(\sigma)}}\mid S[s,\sigma]=r\right)\right) \\ &\quad-\sum_{r=0}^{R-1}\nu_r(x)\left(\sum_{s=0}^{R-1}\rho_s q_{s,r}\cdot E_{\sigma\in\mathcal{D}_s}\left(\frac{e^{xI(\sigma|s)}\left(I(\sigma|s)\cdot\hbar(x)-l(\sigma)\cdot\frac{d\hbar(x)}{dx}\right)}{[\hbar(x)]^{l(\sigma)+1}}\mid S[s,\sigma]=r\right)\right) = 0. \end{aligned} \tag{D.7}$$

Hence, by setting x to zero in (D.7), we have that

$$\sum_{r=0}^{R-1}\frac{d\nu_r(x)}{dx}|_{x=0}\left(\rho_r-\sum_{s=0}^{R-1}\rho_s q_{s,r}\right)-\sum_{s=0}^{R-1}\rho_s\, E_{\sigma\in\mathcal{D}_s}(I(\sigma|s)-\mathcal{H}\cdot l(\sigma))=0. \tag{D.8}$$

Since ρ_r is the steady-state probability of being in state r at a parsing point, we have that $\rho_r=\sum_{s=0}^{R-1}\rho_s q_{s,r}$ for all r, and thus

$$\sum_{s=0}^{R-1}\rho_s\, E_{\sigma\in\mathcal{D}_s}(I(\sigma|s)-\mathcal{H}\cdot l(\sigma))=0. \tag{D.9}$$

By using the identical argument starting from (D.3) instead of (D.2), we find that for any recurrent class of states Γ,

$$\sum_{s=0}^{R-1}\rho_s(\Gamma)\cdot E_{\sigma\in\mathcal{D}_s}(I(\sigma|s)-\mathcal{H}\cdot l(\sigma))=0. \tag{D.10}$$

(D.9) and (D.10) imply the conservation of entropy. □

In principle, Theorem D.1 can be used to determine the relationship between higher moments of the self-information and length of a parsed string.

Appendix E

Proof of Theorem 3.3: Let $\chi_\psi = \tilde{\tau}_s + \ln w_\psi$,

$$h^+_{\psi,j}(x) = \begin{cases} e^{-x}, & x \in [0, \ln\left(\frac{w_\psi}{p_{\psi,j} w_{S[\psi,j]}}\right)] \\ 0, & x \notin [0, \ln\left(\frac{w_\psi}{p_{\psi,j} w_{S[\psi,j]}}\right)] \end{cases}, \text{ and } h^-_{\psi,j}(x) = \begin{cases} e^{-x}, & x \in (0, \ln\left(\frac{w_\psi}{p_{\psi,j} w_{S[\psi,j]}}\right)) \\ 0, & x \notin (0, \ln\left(\frac{w_\psi}{p_{\psi,j} w_{S[\psi,j]}}\right)) \end{cases}.$$

Then (3.5) can be rewritten as

$$\sum_{\psi=0}^{R-1} \sum_{j \in \tilde{K}(\psi)} w_\psi \int_0^{\chi_\psi} h^-_{\psi,j}(\chi_\psi - x)\, dm^{(\psi,s)}(x) \le M e^{-\tilde{\tau}_s} \le \sum_{\psi=0}^{R-1} \sum_{j \in \tilde{K}(\psi)} w_\psi \int_0^{\chi_\psi} h^+_{\psi,j}(\chi_\psi - x)\, dm^{(\psi,s)}(x).$$

Using Blackwell's theorem, we have

$$\begin{aligned} M e^{-\tilde{\tau}_s} &\longrightarrow \sum_{\psi=0}^{R-1} \sum_{j \in \tilde{K}(\psi)} w_\psi \cdot \frac{\pi_\psi}{\mathcal{H}} \left(1 - \frac{p_{\psi,j} w_{S[\psi,j]}}{w_\psi}\right) \\ &= \sum_{\psi=0}^{R-1} \frac{\pi_\psi w_\psi}{\mathcal{H}} \cdot K_\psi - \sum_{\psi=0}^{R-1} \sum_{r=0}^{R-1} \sum_{j: S[\psi,j]=r} \frac{\pi_\psi p_{\psi,j} w_r}{\mathcal{H}} \\ &= \sum_{\psi=0}^{R-1} \frac{\pi_\psi K_\psi w_\psi}{\mathcal{H}} - \sum_{r=0}^{R-1} \sum_{\psi=0}^{R-1} \frac{\pi_\psi f_{\psi,r} w_r}{\mathcal{H}} \\ &= \sum_{\psi=0}^{R-1} \frac{\pi_\psi (K_\psi - 1) w_\psi}{\mathcal{H}}. \end{aligned}$$

□

Appendix F

Proof of Lemma 5.3: Suppose $p_0 \leq p_j$ for all j. If we let $\mathcal{H}$ and each of the probabilities be variables, then we would like to show that the minimum of

$$\ln\left(\frac{1-p_0}{K-1}\right) + \sum_{j=0}^{K-1} \frac{p_j(\ln p_j)^2}{\mathcal{H}} \tag{F.1}$$

subject to the constraints

$$\sum_{j=0}^{K-1} p_j - 1 = 0 \tag{F.2}$$

$$\sum_{j=0}^{K-1} p_j \ln p_j + \mathcal{H} = 0 \tag{F.3}$$

$$p_0 - p_j \leq 0 \tag{F.4}$$

is non-negative for all choices of $\mathbf{p} = (p_0, p_1, \ldots, p_{K-1})$. We will first fix p_0 and minimize $\sum_{j=0}^{K-1} \frac{p_j(\ln p_j)^2}{\mathcal{H}}$ over the constraints (F.2)-(F.4). If $p_0 = \frac{1}{K}$, then (F.2) and (F.4) imply that $p_j = \frac{1}{K}$ for all j. So assume that $p_0 < \frac{1}{K}$. Let us introduce the Lagrange multipliers λ_1, λ_2, and $\overline{\mu} = (\mu_1, \ldots, \mu_{K-1})$ (see [Lue84, §10.8]) for constraints (F.2), (F.3), and (F.4), respectively. Define

$$f(\mathbf{p}, \mathcal{H}) \doteq \sum_{j=0}^{K-1} \frac{p_j(\ln p_j)^2}{\mathcal{H}} + \lambda_1 \left(\sum_{j=0}^{K-1} p_j - 1\right) + \lambda_2 \left(\sum_{j=0}^{K-1} p_j \ln p_j + \mathcal{H}\right) + \sum_{j=1}^{K-1} \mu_j(p_0 - p_j). \tag{F.5}$$

At a relative minimum $(\mathcal{H}^*, p_1^*, \ldots, p_{K-1}^*)$, we have that the Lagrange multipliers can be chosen

so that

$$\overline{\mu} \geq \mathbf{0} \tag{F.6}$$

$$\nabla f(\mathbf{p}^*, \mathcal{H}^*) = 0 \tag{F.7}$$

and

$$\mu_j(p_0 - p_j) = 0 \text{ for all } j. \tag{F.8}$$

We have that $\frac{\partial}{\partial \mathcal{H}} f(\mathbf{p}, \mathcal{H}) = \lambda_2 - \sum_{j=0}^{K-1} \frac{p_j (\ln p_j)^2}{\mathcal{H}^2} = 0$. Hence,

$$\lambda_2 = \sum_{j=0}^{K-1} \frac{p_j (\ln p_j)^2}{\mathcal{H}^2}. \tag{F.9}$$

Furthermore, for $j \geq 1$,

$$\frac{\partial}{\partial p_j} f(\mathbf{p}, \mathcal{H}) = \frac{1}{\mathcal{H}} \left((\ln p_j + 1)^2 - 1 \right) + \lambda_1 + \lambda_2 (\ln p_j + 1) - \mu_j. \tag{F.10}$$

Suppose $p_j > p_0$ for all $j \geq 1$. By (F.8), $\mu_j = 0$ for all j and hence, (F.10) implies that $p_1 = p_2 = \cdots = p_{K-1} = \frac{1-p_0}{K-1}$. This is a potential candidate for the solution. We will demonstrate that it's the only candidate. To arrive at a contradiction, suppose there is a different relative minimum. Without loss of generality, assume the other minimum point satisfies $p_0 \leq p_1 \leq \cdots \leq p_{K-1}$, and for $1 < i \leq K-1$, suppose that $p_0 = p_1 = \cdots = p_{i-1} < p_i$. By (F.8), $\mu_j = 0$ for all $j \geq i$. Therefore, it follows from (F.10) that

$$p_i = \cdots = p_{K-1} = \frac{1 - i p_0}{K - i} > p_0. \tag{F.11}$$

Hence,

$$\begin{aligned}
\mathcal{H} &= -\sum_{j=0}^{K-1} p_j \ln p_j = -i p_0 \ln p_0 - (1 - i p_0) \ln \left(\frac{1 - i p_0}{K - i} \right) \\
\lambda_2 &= \frac{i p_0 (\ln p_0)^2 + (1 - i p_0) \ln \left(\frac{1 - i p_0}{K - i} \right)^2}{\mathcal{H}^2} > 0 \\
\lambda_1 &= \frac{1}{\mathcal{H}} \left(1 - (\ln p_i + 1)^2 \right) - \lambda_2 (\ln p_i + 1) \\
&= \frac{1}{\mathcal{H}} \left(1 - \left(\ln \left(\frac{1 - i p_0}{K - i} \right) + 1 \right)^2 \right) - \lambda_2 \left(\ln \left(\frac{1 - i p_0}{K - i} \right) + 1 \right)
\end{aligned}$$

$$\begin{aligned}
\mu_1 = \cdots = \mu_{i-1} &= \frac{1}{\mathcal{H}}\left((\ln p_0 + 1)^2 - 1\right) + \lambda_1 + \lambda_2(\ln p_0 + 1) \\
&= \frac{1}{\mathcal{H}}\left((\ln p_0 + 1)^2 - \left(\ln\left(\frac{1 - ip_0}{K - i}\right) + 1\right)^2\right) + \lambda_2\left(\ln p_0 - \ln\left(\frac{1 - ip_0}{K - i}\right)\right) \\
&< 0, \text{ because of (F.11).}
\end{aligned}$$

The last inequality violates the non-negativity constraint on $\mu_1, \cdots, \mu_{i-1}$. Hence, for fixed p_0, $p_i = \cdots = p_{K-1} = \frac{1-ip_0}{K-i} > p_0$ minimizes (F.1). Then by minimizing

$$\ln\left(\frac{1 - p_0}{K - 1}\right) - \frac{p_0(\ln p_0)^2 + (1 - p_0)\left(\ln\left(\frac{1-p_0}{K-1}\right)\right)^2}{p_0 \ln p_0 + (1 - p_0)\ln\left(\frac{1-p_0}{K-1}\right)}$$

over $0 \leq p_0 \leq \frac{1}{K}$, we find that the minimum occurs at $p_0 = \frac{1}{K}$, and the value of (F.1) at $p_0 = \cdots = p_{K-1} = \frac{1}{K}$ is zero.

Bibliography

[BCW90] T. C. Bell, J. G. Cleary, I. H. Witten, *Text Compression,* Prentice-Hall, New Jersey, 1990.

[Chu60] K. L. Chung, *Markov Chains with Stationary Transition Probabilites*, Springer-Verlag, Berlin 1960.

[Eli55] P. Elias, "Predictive Coding," *I.R.E. Trans. Inform. Theory* IT-1, 16-33, March 1955.

[Fel71] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 2, Wiley, New York 1966 (2nd ed., 1971).

[Gal95] R. G. Gallager, Personal Communication.

[Gal96] R. G. Gallager, *Discrete Stochastic Processes*, Kluwer, Boston 1996.

[GK68] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables,* Addison-Wesley Pub. Co., Cambridge, MA, 1968.

[Huf52] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. I.R.E.* 40, 1098-1101, 1952.

[JL75] F. Jelinek and G. Longo, "Algorithms for source coding" in *Coding and Complexity*, G. Longo, Ed. Springer Verlag, Wien, 1975.

[JS72] F. Jelinek and K. S. Schneider, "On variable-length-to-block coding," *I.E.E.E. Trans. Inform. Theory* IT-18, 765-774, 1972.

[Kho69] G. L. Khodak, "Delay-redundancy relation of VB-encoding," All-union Conference on Theoretical Cybernetics. Novobirsk, 1969. (Russian)

[Kri94] R. Krichevsky, *Universal Compression and Retrieval,* Kluwer, Dordrecht 1994.

[LS95] G. Louchard and W. Szpankowski, "On the average redundancy rate of the Lempel-Ziv code," Preprint.

[Lue84] D. G. Luenberger, *Linear and Nonlinear Programming,* Addison-Wesley, Reading 1984.

[Mas83] J. L. Massey, "The entropy of a rooted tree with probabilities," IEEE International Symposium on Information Theory, 1983.

[MW85] V. S. Miller and M. N. Wegman, "Variations on a theme by Ziv and Lempel" in *Combinatorial Algorithms on Words*, NATO ASI Series, Vol. F12, A. Apostolico and Z. Galil, Ed. Springer Verlag, Berlin, 131-140, 1985.

[NN87] P. Ney and E. Nummelin, "Markov additive processes. I. Eigenvalues properties and limit theorems," *Annals of Prob.* 15, 561-592, 1987.

[PWZ92] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm," *I.E.E.E. Trans. Inform. Theory*, IT-38, 66-72, 1992.

[Ros83] S. M. Ross, *Stochastic Processes,* Wiley, New York 1983.

[SG94] S. A. Savari and R. G. Gallager, "Arithmetic coding for finite-state noiseless channels," *I.E.E.E. Trans. Inform. Theory* IT-40, 100-107, 1994.

[Sha48] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.* 27, 379-423, 623-656, 1948.

[SW49] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication,* University of Illinois Press, Urbana, 1949.

[Tun67] B. P. Tunstall, "Synthesis of noiseless compression codes," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, 1967.

[TW87] T. J. Tjalkens and F. M. J. Willems, "Variable to fixed-length codes for Markov sources," *I.E.E.E. Trans. Inform. Theory* IT-33, 246-257, 1987.

[Wel84] T. A. Welch, "A technique for high-performance data compression," *I.E.E.E. Computer* 17:6, 8-19, 1984.

[Wyn93] A. J. Wyner, *String Matching Theorems and Applications to Data Compression and Statistics*, Ph.D. thesis, Stanford University, 1993.

[WZ94] A. D. Wyner and J. Ziv, "The sliding-window Lempel-Ziv algorithm is asymptotically optimal" in *Communications and Cryptography, Two Sides of One Tapestry*, R. E. Blahut, D. J. Costello, Jr., U. Maurer, T. Mittelholzer, ed. Kluwer, Boston 1996.

[ZL77] J. Ziv and A. Lempel, "A universal algorithm for data compression," *I.E.E.E. Trans. Inform. Theory* IT-23, 337-343, 1977.

[ZL78] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *I.E.E.E. Trans. Inform. Theory* IT-24, 530-536, 1978.