CHARACTERIZING THE REDUNDANCY OF UNIVERSAL SOURCE CODING FOR FINITE-LENGTH SEQUENCES

A Thesis
Presented to
The Academic Faculty

by

Ahmad Beirami

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
August 2011

CHARACTERIZING THE REDUNDANCY OF UNIVERSAL SOURCE CODING FOR FINITE-LENGTH SEQUENCES

Approved by:

Professor Faramarz Fekri, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor John Barry
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Steven W. McLaughlin
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Raghupathy Sivakumar
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 25 April 2011

To my parents.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my research advisor Prof. Faramarz Fekri for his continuous encouragement, advice, and guidance. Throughout the last two years, I have enjoyed his company as a supportive and insightful friend rather than a formal research advisor.

My sincere thanks go to my incredible friends Amir Hossein, Mehrsa, and Reza who have been there for me through the difficult times and have provided me with nonstop support and caring.

I thank my labmates Nima, Mohsen, Erman, and Arash for stimulating research discussions and the fun we have had in the lab during the past couple of years.

I am grateful to my friends Laleh, Sasan, Sara, Arash, Ehsan, Farshid, Zohreh, Amir Saeed, Navid, Josep, Massimiliano, and Yahya for making Atlanta a more fun place to live. I also would like to acknowledge my other friends who are spread all over the world: Meisam, Pooyan, Farid, Ali, Aliresa, Sadegh, Mohammadreza, Meysam, Hessam, Sina, Soroush, Iman, Mohammad, Arman, Sepideh, Saba, and Shiva for their support, encouragement, enthusiasm that has made my life much more enjoyable.

Last but not least, this thesis would not have been possible if it was not for the support and motivation of my family throughout my life: my dad Karim, my mom Farah, and my sister Elnaz. I also would like to thank my aunts and uncles Fariba, Saeed, Afsaneh, Mehdi, and Hossein for their support, inspirations, and care.

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

In this thesis, we first study what is the average redundancy resulting from the universal compression of a single finite-length sequence from an unknown source. In the universal compression of a source with d unknown parameters, Rissanen demonstrated that the expected redundancy for regular codes is asymptotically $\frac{d}{2}\log n + o(\log n)$ for almost all sources, where n is the sequence length. Clarke and Barron also derived the asymptotic average minimax redundancy for memoryless sources. The average minimax redundancy is concerned with the redundancy of the worst parameter vector for the best code. Thus, it does not provide much information about the effect of the different source parameter values. Our treatment in this thesis is probabilistic. In particular, we derive a lower bound on the probability measure of the event that a sequence of length n from an FSMX source chosen using Jeffreys' prior is compressed with a redundancy larger than a certain fraction of $\frac{d}{2}\log n$. Further, our results show that the average minimax redundancy provides good estimate for the average redundancy of most sources for large enough n and d. On the other hand, when the source parameter d is small the average minimax redundancy overestimates the average redundancy for small to moderate length sequences. Additionally, we precisely characterize the average minimax redundancy of universal coding when the coding scheme is restricted to be from the family of two–stage codes, where we show that the two–stage assumption incurs a negligible redundancy for small and moderate length n unless the number of source parameters is small. Our results, collectively, help to characterize the non-negligible redundancy resulting from the compression of small and moderate length sequences. Next, we apply these results to the compression of a small to moderate length sequence provided that the context present in a sequence

of length $\mathcal{M}$ from the same source is memorized. We quantify the achievable performance improvement in the universal compression of the small to moderate length sequence using context memorization.

CHAPTER I

INTRODUCTION

This work is broadly motivated by the universal compression of a finite-length sequence from a library of finite-length sequences with similar context in storage systems. The tremendous increase in the amount of storage data has raised a great deal of interest in the compression of storage data since the removal of redundancy can significantly reduce the cost of data maintenance as well as data transmission. In many cases, however, the data consists of several small files that need to be compressed and retrieved individually, i.e., a *finite*-length compression problem. Moreover, different data sets may be of various natures, hence little a priori assumptions may be made regarding the probability distribution of the data, i.e., universal compression [2, 9, 13, 14, 27, 39, 42]. When an entire database is concerned, in many cases, it can be compressed to less than one tenth of its original size. However, most applications require that individual files from the database be retrieved and updated separately from the rest of the database. On the other hand, the individual file sizes may be relatively small raising the question that how effectively redundancy is removed by the universal compression of files separately. Therefore, we have two main objectives in this thesis. First, we wish to investigate the compression performance of a single finite-length sequence. We will establish that the redundancy can be very significant when a single finite-length sequence is compressed alone. Second, we aim to determine the fundamental gains achieved in the compression of a single small to moderate length sequence provided that both the encoder and the decoder have access to a memorized context from the same source. This could be possibly used to

determine the performance improvement by considering the context when compressing a library of sequences from the same source. While previous work has developed efficient clustering of data for the best compression performance for sequences from similar context (cf. [19] and the references therein), our goal is to study memorization to improve compression of the finite-length sequences.

Since Shannon's seminal work on the analysis of communication systems [30], many researchers have contributed toward the development of source coding schemes with the code length as close as possible to the entropy of the sequence. It is well known that using a prefix-free code, the entropy of a sequence is the absolute lower bound on the expected codeword length of any stationary ergodic information source [7]. Provided that the statistics of the information source are *known*, Huffman block coding achieves the entropy of a sequence with a redundancy smaller than 1 bit per source symbol, where the redundancy term is due to the integer length constraint on the codewords [34,35]. However, the assumption of known source statistics fails to hold for many practical applications. We usually cannot assume a priori knowledge on the statistics of the source although we still wish to compress the *unknown* stationary ergodic source to its entropy rate. This is known as the universal compression problem. However, unfortunately, universality imposes an inevitable redundancy based on the richness of the class of the sources with respect to which the code is universal.

In [31], Shields showed that it is not possible to find a universal redundancy for the class of stationary ergodic sources by proving that there exists a stationary ergodic source whose redundancy rate dominates any given rate. Therefore, in this thesis, we focus our study on the fairly general class of FSMX (Finite State Machine X) sources [8, 21, 37]. The asymptotic average redundancy of FSMX sources has been investigated in the past [4, 24, 37, 38].

In the following, we describe our source model together with necessary notations and related work. Denote $\mathcal{A}$ as a finite alphabet. Let $\mathcal{A}^*$ be the set of all strings over

the alphabet $\mathcal{A}$, i.e.,

$$\mathcal{A}^* \triangleq \bigcup_{i=0}^{\infty} \mathcal{A}^i, \tag{1}$$

where $\mathcal{A}^i$ $(i > 1)$ denotes the set of the strings of length i. We use $\mathcal{A}^0$ to represent the empty string. Further, let the context tree T be defined as a finite subset of $\mathcal{A}^*$ such that for all $s \in T$, each postfix of s also belongs to T. Let ∂T be defined as

$$\partial T \triangleq \{xs : x \in \mathcal{A}, s \in T\} \cup \mathcal{A}^0 \setminus T. \tag{2}$$

Note that ∂T forms a complete postfix set, i.e., no element in ∂T is a postfix of another element. Assume that for all $s \in \mathcal{A}^*$, there exists a unique element in ∂T that is a postfix of the s. Let $\tau(s)$ be this unique postfix, which we refer to as the context or state. Note that τ defines a state transition function.

Let an FSMX source P be defined as an information source, in which for each context s the probability of the successive character is determined by a source parameter θ_s. Let $d = |\partial T|$ denote the number of the source parameters. Further, let $\theta = (\theta_1, ..., \theta_d)$ be the d-dimensional parameter vector associated with source P. We assume that the d parameters are unknown and lie in the space $\Theta^d \triangleq (0, 1)^d$. Denote $\mathcal{P}^d$ as the *family* of sources with d-dimensional unknown parameter vector $\theta \in \Theta^d$. We use the notation $x^n = (x_1, ..., x_n) \in \mathcal{A}^d$ to present a sequence of length n generated by the source P. Let μ_θ denote the probability measure defined by the parameter vector θ on sequences of length n. Let $H_n(\theta)$ be the source entropy given parameter vector θ, i.e.,

$$H_n(\theta) = \mathbf{E} \log \left(\frac{1}{\mu_\theta(X^n)} \right) = \sum_{x^n} \mu_\theta(x^n) \log \left(\frac{1}{\mu_\theta(x^n)} \right).^1 \tag{3}$$

In this thesis $\log(\cdot)$ always denotes the logarithm in base 2. Let $C_n : \mathcal{A}^n \to \{0, 1\}^*$ be an injective mapping from the set $\mathcal{A}^n$ of the sequences of length n over $\mathcal{A}$ to the set $\{0, 1\}^*$ of binary sequences. Further, denote $l(C_n, x^n) = l_n(x^n)$ as the *regular* length

[1] Throughout this thesis all expectations are taken with respect to the true unknown parameter vector θ.

function that describes the codeword length associated with the sequence x^n. Denote L_n as the set of all regular length functions on an input sequence of length n.

Let $r_n(l_n, \theta, x^n)$ denote the redundancy of the code with length function l_n and the parameter vector θ on the individual sequence x^n, defined as

$$r_n(l_n, \theta, x^n) = l_n(x^n) - \log\left(\frac{1}{\mu_\theta(x^n)}\right). \tag{4}$$

Note that the redundancy for an individual sequence x^n is not necessarily non-negative in general. Some past works [9, 32] have studied the worst-case minimax redundancy defined as

$$r_n = \min_{l_n \in L_n} \max_{\theta \in \Theta^d} \max_{x^n} \{r_n(l_n, \theta, x^n)\}. \tag{5}$$

The worst-case minimax redundancy characterizes the compression for the worst-case individual sequence [4, 9, 12, 15, 18, 28, 32, 36, 37, 39, 41]. It has been shown that the leading term in r_n is asymptotically $\frac{d}{2}\log n$. In particular, Szpankowski derived the asymptotic behavior of the worst-case minimax redundancy and precisely derived all the terms up to $O(n^{-3/2})$ [12]. The worst-case minimax redundancy, by definition, is a good metric whenever bad compression performance is not tolerated on any individual sequence. However, in the storage compression scenario, we are interested in reducing the average size of the stored data, and hence, the worst-case minimax redundancy would not be a good metric. We consider the average performance as opposed to the redundancy associated with the individual sequences.

Denote $R_n(l_n, \theta)$ as the expected redundancy of the code on a sequence of length n, defined as the difference between the expected codeword length and the entropy. That is

$$R_n(l_n, \theta) = \mathbf{E} r_n(l_n, \theta, X^n) = \mathbf{E} l_n(X^n) - H_n(\theta). \tag{6}$$

The expected redundancy is always non-negative. The code that asymptotically achieves the entropy rate with length function l_n would satisfy $\frac{1}{n}R_n(l_n, \theta) \to 0$ as $n \to \infty$ for all θ. Rissanen demonstrated that for the universal compression

of the family $\mathcal{P}^d$ of the FSMX sources with parameter vector θ, the redundancy of the codes with *regular* length functions l_n is asymptotically lower bounded by $R_n(l_n, \theta) \geq (1-\epsilon)\frac{d}{2}\log n$ [24, 25], for all $\epsilon > 0$ and almost all parameter vector θ. This asymptotic lower bound is tight since there exist coding schemes that achieve the bound asymptotically [24, 39]. This result was later extended in [14, 22] to more general classes of sources. Rissanen further proved that the redundancy for individual sequences is also asymptotically equal to the average minimax redundancy with high probability [28]. However, these results do not provide much insight on the performance of universal coding for a small to moderate length (size) sequence.

Let the maximum expected redundancy for a code with length function l_n be given as $R_n(l_n) = \max_{\theta\in\Theta^d} R_n(l_n, \theta)$, which may be minimized over all codes to achieve the average minimax expected redundancy [6, 10, 40]

$$R_n = \min_{l_n \in L_n} \max_{\theta\in\Theta^d} R_n(l_n, \theta). \tag{7}$$

The average minimax redundancy is concerned with the maximum redundancy over all parameter space, i.e., it describes the performance of the best code for the worst source parameter. Therefore, it does not characterize the average redundancy for all parameters, which determines the fundamental limits of compression in the above-mentioned setting. The leading term in the average minimax redundancy is asymptotically $\frac{d}{2}\log n$, similar to that of the redundancy for individual sequences and the worst-case minimax redundancy [12, 40].

In the first part of this thesis, we extend Rissanen's probabilistic treatment of redundancy to the universal compression in *finite*-length regime using the two–stage and conditional two–stage codes. In [5], we considered the redundancy of the universal compression in *finite*-length memoryless sources for the family of two–stage codes. Although the two–stage code assumption is restrictive and incurs an extra redundancy, the constraint could be relaxed by considering the conditional two–stage codes that are optimal in the sense that they achieve the average minimax redundancy.

The rest of this thesis is organized as follows. In Chapter 2, after a review of the previous work, we formally state the problem of redundancy for finite-length universal compression of FSMX sources using the two–stage codes and the conditional two–stage codes. In Chapter 3, we present our main results on the average redundancy for universal compression of finite-length sequences. In Chapter 4, we tailor the main results to the class of finite-alphabet memoryless sources and restate the main results. In Chapter 5, we demonstrate the significance of our results through several examples using memoryless sources as well as finite alphabet finite memory Markov sources. In Chapter 6, we present an application of our results on determining the gain of context memorization in the universal compression of finite-length sequences. Finally, the conclusion is given in Chapter 7.

CHAPTER II

BACKGROUND REVIEW AND PROBLEM STATEMENT

In this section, after a brief review of the previous work, we state the finite-length redundancy problem. Let l_n^θ denote the (non-universal) length function induced by a parameter $\theta \in \Theta^d$. We require that the length function l_n^θ be *regular*, i.e.,

$$l_n^\theta(x^n) \geq \log\left(\frac{1}{\mu_\theta(x^n)}\right) \quad \forall x^n \in \mathcal{A}^n. \tag{8}$$

Note that the requirement (8) is not restrictive since all codes that we know are regular [7, 24].

Denote l_n as the regular length function on the input sequence of length n. Denote $R_n(l_n, \theta)$ as the expected redundancy of the universal compression of source $P \in \mathcal{P}^d$ using the length function l_n. Let $I_n(\theta)$ be the Fisher information matrix for parameter vector θ and a sequence of length n,

$$I_n(\theta) = \{I_n^{ij}(\theta)\} = \frac{1}{n \log e} \mathbf{E}\left\{\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\left(\frac{1}{\mu_\theta(X^n)}\right)\right\}. \tag{9}$$

Fisher information matrix quantifies the amount of information, on the average, that each symbol in a sample sequence of length n from the source conveys about the source parameters.

In this thesis, we assume that the following conditions hold:

1. $\lim_{n\to\infty} I_n(\theta)$ exists and the limit is denoted by $I(\theta)$.
2. All elements of the Fisher information matrix $I_n(\theta)$ are continuous in Θ^d.
3. $\int_{\Theta^d} |I(\theta)|^{\frac{1}{2}} d\theta < \infty$.
4. The family $\mathcal{P}^d$ has a minimal representation with the d-dimensional parameter vector θ.

Rissanen proved an asymptotic lower bound on the universal compression of an information sources with d parameters as [24, 25]:

Fact 1 *For all parameters θ, except in a set of asymptotically Lebesgue volume zero, we have*

$$\lim_{n\to\infty} \frac{R_n(l_n,\theta)}{\frac{d}{2}\log n} \geq 1-\epsilon, \quad \forall \epsilon > 0. \tag{10}$$

While Fact 1 describes the asymptotic fundamental limits of the universal compression of FSMX sources, it does not provide much insight for the case of *finite*-length n. Moreover, the result excludes an asymptotically volume zero set of parameter vectors θ that has non-zero volume for any finite n.

In [6], Clarke and Barron derived the expected minimax redundancy R_n for memoryless sources, later generalized in [1] by Atteson for Markov sources, as the following:

Fact 2 *The average minimax redundancy is asymptotically given by*

$$R_n = \frac{d}{2}\log\left(\frac{n}{2\pi}\right) + \log\int |I_n(\theta)|^{\frac{1}{2}} d\theta + O\left(\frac{1}{n}\right). \tag{11}$$

The average minimax redundancy characterizes the maximum redundancy over the space Θ^d of the parameter vectors. However, it does not say much about the rest of the space of the parameter vectors. The average minimax redundancy is obtained when the parameter vector θ follows Jeffreys' prior, which is [40]

$$p(\theta) = \frac{|I(\theta)|^{\frac{1}{2}}}{\int |I(\lambda)|^{\frac{1}{2}} d\lambda}. \tag{12}$$

Rissanen further proved that the redundancy for individual sequences defined in (4), except for a Lebesgue volume zero set of sequences, is asymptotically given as

$$r_n(l_n,\theta,x^n) = \frac{d}{2}\log\left(\frac{n}{2\pi}\right) + \log\int |I(\theta)|^{\frac{1}{2}} d\theta + O\left(\frac{1}{n}\right). \tag{13}$$

In other words, the redundancy for almost all sequences is asymptotically equal to the expected minimax redundancy, i.e., the redundancy is highly concentrated around

its mean. However, this result still does not characterize the redundancy for finite n. Next, we state the redundancy problem in the finite-length regime, where we will consider both two–stage and conditional two–stage codes.

2.1 Two–Stage Codes

In a two–stage code, the compression scheme attributes m bits to identify an estimate for the unknown source parameters. Then, in the second stage of the compression, it is assumed that the source with the estimated parameter has generated the sequence. In this case, there will be 2^m possible estimate points in the parameter space for the identification of the source. Let $\Phi^m = \{\phi_1, ..., \phi_{2^m}\}$ denote the set of all estimate points with an m-bit estimation budget. Note that for all i, we have $\phi_i \in \Theta^d$ [3,17,26].

Denote l_n^{2p} as the two–stage length function for the compression of sequences of length n. For each sequence x^n, there exists an estimate point in the set of the estimate points, i.e., $\gamma = \gamma(x^n, m) \in \Phi^m$, which is optimal in the sense that it minimizes the code length and the average redundancy. In other words, γ is the maximum likelihood estimation of the unknown parameter in the set of the estimate parameters. That is

$$\gamma = \arg \min_{\phi_i \in \Phi^m} \log \left(\frac{1}{\mu_{\phi_i}(x^n)} \right) = \arg \max_{\phi_i \in \Phi^m} \mu_{\phi_i}(x^n). \tag{14}$$

The two–stage universal length function for the sequence x^n is then given by

$$l_n^{2p}(x^n) = m + l_n^{\gamma}(x^n), \tag{15}$$

where l_n^{γ} denotes the regular length function induced by the parameter $\gamma \in \Phi^m$. Let L_n^{2p} be the set of all two–stage codes that could be described as in (15). Further denote $\mu_{\gamma}(x^n)$ as the probability measure induced by γ.

Increasing the bit budget m for the identification of the unknown source parameters results in an exponential growth in the number of estimate points, and hence, smaller $l_n^{\gamma}(x^n)$ on the average due to the more accurate estimation of the unknown source parameter vector. On the other hand, m directly appears as part of the

compression overhead in (15). Therefore, it is desirable to find the optimal m that minimizes the total expected codeword length, which is $\mathbf{E}l_n^{2p}(X^n) = m + \mathbf{E}l_n^{\gamma}(X^n)$.

Since we assumed the code is regular, we may use (8) to bound the average redundancy of two–stage codes

$$R_n(l_n^{2p}, \theta) \geq m + \mathbf{E}\log\left(\frac{1}{\mu_\gamma(X^n)}\right) - H_n(\theta). \tag{16}$$

Our goal in Sec. 3.1 is to better characterize the lower bound on the universal compression of two–stage codes in (16) in the small to moderate length regime.

Further, let R_n^{2p} denote the average minimax redundancy of the two–stage codes, i.e.,

$$R_n^{2p} = \min_{l_n^{2p} \in L_n^{2p}} \max_{\theta \in \Theta^d} R_n(l_n^{2p}, \theta). \tag{17}$$

In Sec. 3.3, we precisely derive R_n^{2p}.

2.2 Conditional Two–Stage Codes

In a two–stage code, we already have some knowledge about the sequence x^n through the optimally estimated parameter $\gamma(x^n)$ (maximal likelihood estimation) that can be leveraged for encoding x^n using the length function $l_n^{\gamma}(x^n)$. The two–stage length function in (15) defines an incomplete coding, which is not optimal in the sense that it does not achieve the optimal compression among all regular length functions. Further, it does not achieve the average minimax redundancy of the regular codes [5, 17]. Conditioned on $\gamma(x^n)$, the length of the codeword for x^n may be further decreased [26].

Let $S_m(\gamma)$ be the collection of all x^n for which the optimally estimated parameter is γ, i.e.,

$$S_m(\gamma) \triangleq \{x^n \in \mathcal{A}^n : \ \mu_\gamma(x^n) \geq \mu_{\phi_i}(x^n) \ \ \forall \phi_i \in \Phi^m\}. \tag{18}$$

Further, let $A_m(\gamma)$ denote the total probability measure of all sequences in the set $S_m(\gamma)$, i.e.,

$$A_m(\gamma) = \sum_{x^n \in S_m(\gamma)} \mu_\gamma(x^n). \tag{19}$$

Thus, the knowledge of $\gamma(x^n)$ in fact changes the probability distribution of the sequence. Denote $\mu_\gamma(x^n|x^n \in S_m(\gamma))$ as the conditional probability measure of x^n given γ, i.e., the probability distribution that is normalized to $A_m(\gamma)$. That is

$$\mu_\gamma(x^n|x^n \in S_m(\gamma)) = \frac{\mu_\gamma(x^n)}{A_m(\gamma)}. \tag{20}$$

Note that $\mu_\gamma(x^n|x^n \in S_m(\gamma)) \geq \mu_\gamma(x^n)$ due to the fact that $A_m(\gamma) \leq 1$. Let $l_n^\gamma(x^n|x^n \in S_m(\gamma))$ be the codeword length corresponding to the conditional probability distribution, which is decreased to $\mathbf{E}\log\left(\frac{A_m(\gamma(X^n))}{\mu_\gamma(X^n)}\right)$. Denote l_n^{c2p} as the conditional two–stage length function for the compression of sequences of length n using the normalized maximum likelihood, which is given by

$$l_n^{c2p} = m + l_n^\gamma(x^n|x^n \in S_m(\gamma)). \tag{21}$$

Therefore, the average redundancy of the conditional two–stage scheme is lower bounded as

$$R_n(l_n^{c2p}, \theta) \geq m + \mathbf{E}\log\left(\frac{A_m(\gamma(X^n))}{\mu_\gamma(X^n)}\right) - H_n(\theta). \tag{22}$$

Denote L_n^{c2p} as the set of the conditional two–stage codes that are described using (21). Let R_n^{c2p} denote the average minimax redundancy of the conditional two–stage codes, i.e.,

$$R_n^{c2p} = \min_{l_n^{c2p} \in L_n^{c2p}} \max_{\theta \in \Theta^d} R_n(l_n^{c2p}, \theta). \tag{23}$$

Rissanen demonstrated that this conditional version of two–stage codes is in fact optimal in the sense that it achieves the average minimax redundancy [28]. In other words, $R_n^{c2p} = R_n$, where R_n is the average minimax redundancy of the regular codes in (11). In Chapter 3.4, our goal is to investigate the performance of the conditional two–stage codes using (22). In particular, we derive a lower bound on the average redundancy for the compression of FSMX sources using regular conditional two–stage codes.

CHAPTER III

MAIN RESULTS ON THE REDUNDANCY

In Chapter 3.1, we present a lower bound on the redundancy for the universal compression of two–stage codes. In Chapter 3.2, we prove the main result. In Chapter 3.3, we precisely characterize the average minimax redundancy for two–stage codes. In Chapter 3.4, we drop the two–stage assumption on the length function and extend the main result on the average redundancy to the conditional two–stage sources.

3.1 Two–Stage Code Redundancy

In this section, we restrict the code to the set of two–stage length functions, i.e., $l_n^{2p} \in L_n^{2p}$. The two–stage assumption asymptotically incurs an extra redundancy in the compression. We derive a lower bound on the probability of the event that P is compressed with redundancy greater than the redundancy level R_0 for finite-length n. In other words, we find a lower bound on $\mathbf{P}[R_n(l_n^{2p}, \theta) > R_0]$. Let

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \tag{24}$$

denote Euler's gamma function.

Theorem 1 *Consider the universal compression of the family of FSMX sources $\mathcal{P}^d$ with the parameter vector θ that follows Jeffreys' prior. Let ϵ be a real number. Then,*

$$\mathbf{P}\left[\frac{R_n(l_n^{2p}, \theta)}{\frac{d}{2}\log n} \geq 1 - \epsilon\right] \geq 1 - \frac{C_d}{\int |I(\theta)|^{\frac{1}{2}} d\theta} \left(\frac{d}{en^\epsilon}\right)^{\frac{d}{2}}, \tag{25}$$

where C_d is the volume of the d-dimensional unit ball, which is

$$C_d = \frac{\Gamma\left(\frac{1}{2}\right)^d}{\Gamma\left(\frac{d}{2}+1\right)}. \tag{26}$$

As we shall see in the following, the proof of Theorem 1 is constructive, and hence, the lower bound is indeed tight and achievable.

3.2 Proof of Theorem 1

To prove Theorem 1, first, we rewrite (16) as

$$R_n(l_n^{2p}, \theta) \geq m + \mathbf{E} \log \left(\frac{\mu_\theta(X^n)}{\mu_\gamma(X^n)} \right). \tag{27}$$

In order to bound the average redundancy in (27), we use the following lemma

Lemma 1

$$\mathbf{E} \log \left(\frac{\mu_\theta(X^n)}{\mu_\gamma(X^n)} \right) = D_n(\mu_\theta || \mu_\beta) + O\left(2^{-D_n(\mu_\theta || \mu_\beta)}\right), \tag{28}$$

where β is defined as

$$\beta \triangleq \arg \min_{\phi_i \in \Phi^m} D_n(\mu_\theta || \mu_{\phi_i}), \tag{29}$$

and

$$D_n(\mu_\theta || \mu_\beta) = \mathbf{E} \log \left(\frac{\mu_\theta(x^n)}{\mu_\beta(x^n)} \right). \tag{30}$$

Proof: See Appendix A. ■

Note that $D_n(\mu_\theta || \mu_\beta)$ is the non-negative Kullback–Leibler divergence between the probability measures μ_θ and μ_β. We use a probabilistic treatment in order to bound $D_n(\mu_\theta || \mu_\beta)$ for a certain fraction of the source parameters. We assume that the parameter vector θ follows Jeffreys' prior. This distribution is particularly interesting since it results in uniform convergence of redundancy over the space of the parameter vectors and hence the achievement of the average minimax expected redundancy [20, 40].

In order to bound the average redundancy $R_n(l_n^{2p}, \theta)$, in the following, we find an upper bound on the Lebesgue measure of the volume defined by $\frac{1}{n} D_n(\mu_\theta || \mu_\beta) < \delta$ in the d-dimensional space of θ. Since $\beta \in \Phi^m$, the total probability measure of

the volume defined by $\min_{\phi_i \in \Phi^m} \frac{1}{n} D_n(\theta||\phi_i) < \delta$ may be upper bounded as well. This represents the total measure of the sources that have a small redundancy, i.e., it provides us with a lower bound on the probability measure of the sources with $R_n(l_n^{2p}, \theta) \geq \delta$.

Lemma 2 *Assume that the parameter vector θ follows Jeffreys' prior. Then,*

$$\mathbf{P}\left[\frac{1}{n} D_n(\mu_\theta||\mu_\beta) < \delta\right] \leq \frac{C_d}{\int |I(\theta)|^{\frac{1}{2}} d\theta} \left(\frac{2\delta}{\log e}\right)^{\frac{d}{2}}. \tag{31}$$

Further, we have

$$\mathbf{P}\left[\min_i \frac{1}{n} D_n(\mu_\theta||\mu_{\phi_i}) < \delta\right] \leq 2^m \frac{C_d}{\int |I(\theta)|^{\frac{1}{2}} d\theta} \left(\frac{2\delta}{\log e}\right)^{\frac{d}{2}}. \tag{32}$$

Proof: See Appendix B. ■

Lemma 2 states that the probability of the event that $\frac{1}{n} D_n(\mu_\theta||\mu_\beta) < \delta$ does not depend on β when θ follows Jeffreys' prior. Further, the probability of the event $\mathbf{P}[\min_i \frac{1}{n} D_n(\mu_\theta||\mu_{\phi_i}) < \delta]$ is only a function of m. In fact, it is independent of the choice of the points in Φ^m in the space of θ, as long as the points are chosen far apart so that the probability measures of the sources that are covered by each point do not overlap. We are now equipped to prove the main result given in Theorem 1.

Proof of Theorem 1: Note that for all values of m, using Lemma 1, we can rewrite (27) as:

$$R_n(l_n^{2p}, \theta) \geq \min_{\phi_i \in \Phi^m} \{m + D_n(\mu_\theta||\mu_{\phi_i})\} + O\left(2^{-D_n(\mu_\theta||\mu_\beta)}\right). \tag{33}$$

As we shall see, $D_n(\mu_\theta||\mu_\beta) = O(\log n)$, and hence, the error term is $O\left(\frac{1}{n}\right)$, which is negligible compared to the main term. Therefore,

$$\mathbf{P}\left[\frac{R_n(l_n^{2p}, \theta)}{\frac{d}{2}\log n} \leq 1 - \epsilon\right] \tag{34}$$

$$\leq \mathbf{P}\left[\min_i \{m + D_n(\mu_\theta||\mu_{\phi_i})\} \leq (1-\epsilon)\frac{d}{2}\log n\right] \tag{35}$$

$$= \mathbf{P}\left[\min_i \frac{1}{n} D_n(\mu_\theta || \mu_{\phi_i}) \leq (1-\epsilon)\frac{d}{2n}\log n - \frac{m}{n}\right], \tag{36}$$

$$\leq \left\{ 2^m \frac{C_d}{\int |I(\theta)|^{\frac{1}{2}} d\theta} \left(\frac{2\delta(m)}{\log e}\right)^{\frac{d}{2}} \right\}. \tag{37}$$

The last inequality is obtained using Lemma 2. Here, $\delta(m)$ is given by

$$\delta(m) = (1-\epsilon)\frac{d}{2n}\log n - \frac{m}{n}. \tag{38}$$

The inequality in (37) holds for all values of m. We can minimize the right hand side to find an upper bound that is independent of the value of m:

$$\mathbf{P}\left[\frac{R_n(l_{2p},\theta)}{\frac{d}{2}\log n} \leq 1-\epsilon\right]$$

$$\leq \min_m \left\{ 2^m \frac{C_d}{\int |I(\theta)|^{\frac{1}{2}} d\theta} \left(\frac{2\delta(m)}{\log e}\right)^{\frac{d}{2}} \right\}. \tag{39}$$

Carrying out the minimization in (37) leads to the optimal value of m, denoted by m_{op}:

$$m_{op} = (1-\epsilon)\frac{d}{2}\log\left(\frac{n}{e}\right). \tag{40}$$

Using m_{op} in (39), the desired result in Theorem 1 is obtained. ■

3.3 Average Minimax Redundancy of Two–Stage Codes

In this section, we characterize the average minimax redundancy when the coding scheme is restricted to be from the family of two–stage codes. The following Theorem is the main result in this section.

Theorem 2 *In the universal compression of the family of FSMX sources $\mathcal{P}^d$, the average minimax redundancy of two–stage codes is obtained by*

$$R_n^{2p} = R_n + g(d) + O\left(\frac{1}{n}\right). \tag{41}$$

Here, R_n is the average minimax redundancy defined in (11) and $g(d)$ is the two–stage penalty term given by

$$g(d) = \log\Gamma\left(\frac{d}{2}+1\right) - \frac{d}{2}\log\left(\frac{d}{2e}\right). \tag{42}$$

Proof: Let $F(n,d,\theta,\epsilon) \triangleq \frac{C_d}{\int |I(\theta)|^{\frac{1}{2}} d\theta} \left(\frac{d}{en^{\epsilon}}\right)^{\frac{d}{2}}$. Denote $R_\epsilon \triangleq (1-\epsilon)\frac{d}{2}\log n$ as a redundancy level. Then, according to Theorem 1, for any ϵ such that $1 - F(n,d,\theta,\epsilon) > 0$, R_ϵ is a lower bound on the maximum redundancy. This is due to the fact that $P[R_n(l_n,\theta) > R_\epsilon] > 0$, i.e., there exists at least one parameter θ such that $R_n(l_n,\theta) > R_\epsilon$. Moreover, note that the average minimax redundancy is achieved when the parameters follow Jeffreys' prior [6,20]. Therefore, the maximum redundancy in our case is the average minimax redundancy and we have $R_n^{2p} > R_\epsilon$. Note that as described in Sec. 3.2, the lower bound in Theorem 1 is tight and achievable. If we minimize ϵ (maximize R_ϵ) with the constraint that $F(n,d,\theta,\epsilon) < 1$, we get the tightest lower bound on the average minimax redundancy as

$$R_n^{2p} = \frac{d}{2}\log n - \log C_d + \log \int |I(\theta)|^{\frac{1}{2}} d\theta - \frac{d}{2}\log\left(\frac{d}{e}\right), \tag{43}$$

Theorem 2 is inferred if C_d is substituted from (26) in 43. ■

3.4 Conditional Two–Stage Code Redundancy

Thus far, we established a lower bound on the average redundancy for the universal compression of the family of FSMX sources when the coding scheme is restricted to the two–stage codes. Now, we relax this constraint and obtain the lower bound on the average redundancy of universal compression for conditional two–stage coding.

Theorem 3 *Assume that the parameter vector θ follows Jeffreys' prior in the universal compression of the family of FSMX sources $\mathcal{P}^d$. Let ϵ be a real number. Then,*

$$\mathbf{P}\left[\frac{R_n(l_n^{c2p},\theta)}{\frac{d}{2}\log n} \geq 1-\epsilon\right] \geq 1 - \frac{1}{\int |I(\theta)|^{\frac{1}{2}} d\theta}\left(\frac{2\pi}{n^{\epsilon}}\right)^{\frac{d}{2}}. \tag{44}$$

Note that it is straightforward to deduce Fact 1 for the case of conditional two–stage codes from Theorem 3 for $\epsilon > 0$ by letting $n \to \infty$. The key in the proof of Theorem 3

is the following lemma that upper bounds the saving achieved by using the conditional two–stage codes.

Lemma 3 *The penalty term in the redundancy of the two–stage coding is upper bounded as*

$$R_n(l_n^{2p},\theta) - R_n(l_n^{c2p},\theta) \leq g(d) + O\left(\frac{1}{n}\right). \tag{45}$$

Proof: See Appendix C ■

We may now prove Theorem 3.

Proof of Theorem 3: First note that according to Lemma 3, we have

$$R_n(l_n^{2p},\theta) \leq R_n(l_n,\theta) + g(d) + O\left(\frac{1}{n}\right). \tag{46}$$

The $O\left(\frac{1}{n}\right)$ term is much smaller than the main term $R_n(l_n,\theta)$, which is $O(\log n)$ and could be ignored even in the small to moderate n regime. Thus,

$$\begin{aligned}
&\mathbf{P}\left[\frac{R_n(l_n^{c2p},\theta)+g(d)}{\frac{d}{2}\log n} \geq 1-\hat{\epsilon}\right] \\
&\quad\geq \mathbf{P}\left[\frac{R_n(l_n^{2p},\theta)}{\frac{d}{2}\log n} \geq 1-\hat{\epsilon}\right] \\
&\quad\geq 1 - \frac{C_d}{\int |I(\theta)|^{\frac{1}{2}} d\theta}\left(\frac{d}{en^{\hat{\epsilon}}}\right)^{\frac{d}{2}},
\end{aligned} \tag{47}$$

where the second inequality is due to Theorem 1, for any $\hat{\epsilon}$. Now, the desired result is then achieved if we set ϵ such that

$$(1-\epsilon)\frac{d}{2}\log n = (1-\hat{\epsilon})\frac{d}{2}\log n - g(d).$$

■

CHAPTER IV

MEMORYLESS SOURCES

In this section, we tailor the results for the class of memoryless sources due to the importance of memoryless information sources. Further, we highlight some of the intermediate results. The proofs for the material in this section may be found in [5]. Let $\mathcal{M}_0$ denote the family of memoryless sources. Denote the alphabet as $\mathcal{A}$ and let $k = |\mathcal{A}|$ be the alphabet size. Let $\theta = (\theta_1, ..., \theta_k)$ be the parameter vector, where $\theta_j = \mathbf{P}[X = \alpha_j]$ and $\sum_j \theta_j = 1$. Note that the parameters live in a $(k-1)$-dimensional simplex, i.e., $d = k - 1$. Let r_i count the appearance of symbol α_i in sequence x^n. Let f_i denote the empirical mass function for the symbol α_i, i.e., $f_i = r_i/n$. Then, the probability measure μ_θ over a memoryless source with parameter vector θ is

$$\mu_\theta(x^n) = \mathbf{P}[X^n = x^n|\theta] = \prod_{i=1}^{k} \theta_i^{r_i}. \tag{48}$$

Let Φ^m denote the set of 2^m estimate points. Further, let $\gamma = (\gamma_1, ..., \gamma_k) \in \Phi^m$ denote the optimal estimated point for the sequence x^n. Then, the probability measure defined by γ is

$$\mu_\gamma(x^n) = \prod_{i=1}^{k} {\gamma_i}^{r_i}. \tag{49}$$

In the following, we state the main results on the compression of finite-alphabet memoryless sources:

Corollary 1 *Consider the universal compression of the family of memoryless sources $\mathcal{M}_0$. Assume that the parameter vector θ follows Jeffreys' prior. Let ϵ be a real number. Then,*

$$\mathbf{P}\left[\frac{R_n(l_n^{2p}, \theta)}{\frac{k-1}{2}\log n} \geq 1 - \epsilon\right] \geq 1 - \left(\frac{k-1}{en^\epsilon}\right)^{\frac{k-1}{2}} B_k, \tag{50}$$

where

$$B_k = \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)}\sqrt{\frac{1}{\pi}}. \tag{51}$$

Note that $B_k \approx \sqrt{\frac{2}{k\pi}}$ for $k \gg 2$. Although Corollary 1 can be directly obtained from Theorem 1, some of the intermediate results are insightful and could be simplified in this case. In the following, we highlight the important steps in the proof, where we introduce simple notation for some of the previously defined quantities. In the case of the memoryless sources, the entropy could be rewritten as

$$H_n(\theta) = nH(\theta) = n\mathbf{E}\sum_{i=1}^{k} f_i \log\left(\frac{1}{\theta_i}\right). \tag{52}$$

We can further simplify (16) to

$$R_n(l_n^{2p}, \theta) \geq m + n\mathbf{E}\sum_{i=1}^{k} f_i \log\left(\frac{\theta_i}{\gamma_i(X^n)}\right). \tag{53}$$

In order to bound the average redundancy in (53), in the following, we find a lower bound on $\mathbf{E}\sum_{i=1}^{k} f_i \log\frac{\theta_i}{\gamma_i(X^n)}$. Note that this term implicitly depends on m since γ is a function of m. Let β be defined as

$$\beta = \arg\min_{\phi_i} D(\theta||\phi_i). \tag{54}$$

Then, we have the following.

$$\mathbf{E}\sum_{i=1}^{k} f_i \log\left(\frac{\theta_i}{\gamma_i(X^n)}\right) \geq D(\theta||\beta) + O\left(\frac{1}{n^2}\right), \tag{55}$$

where

$$D(x||y) = \sum_{j=1}^{k} x_j \log\left(\frac{x_j}{y_j}\right). \tag{56}$$

This is directly resulted from Lemma 1. In the case of memoryless sources, Jeffreys' prior for the parameter vector θ is given by

$$p(\theta) = \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k}\prod_{j=1}^{k}\frac{1}{\sqrt{\theta_j}}, \tag{57}$$

where $\Gamma(\cdot)$ is Euler's gamma function defined in (24). This is in fact the $(\frac{1}{2},...,\frac{1}{2})$ Dirichlet distribution.

Further, The square root of the determinant of the Fisher information matrix may be analytically integrated to be

$$\int_{\Theta^d} |I(\theta)|^{\frac{1}{2}} d\theta = \frac{\Gamma\left(\frac{1}{2}\right)^k}{\Gamma\left(\frac{k}{2}\right)}. \tag{58}$$

This analytical integration enables us to further simplify the main results in the following.

In order to bound the average redundancy $R_n(l_n, \theta)$, we would need an upper bound on the Lebesgue measure of the volume defined by $D(\theta||\gamma) < \delta$ in the $(k-1)$-dimensional simplex of θ. According to Lemma 2, we have

$$\mathbf{P}[D(\theta||\gamma) < \delta] \leq \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} \sqrt{\frac{1}{\pi}} \left(\frac{2\delta}{\log e}\right)^{\frac{k-1}{2}}. \tag{59}$$

Further,

$$\mathbf{P}\left[\min_{\phi_i \in \Phi^m} D(\theta||\phi_i) < \delta\right] \leq 2^m \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} \sqrt{\frac{1}{\pi}} \left(\frac{2\delta}{\log e}\right)^{\frac{k-1}{2}}. \tag{60}$$

The rest of the proof could be carried out following the lines of the proof of Theorem 1.

In [6], Clarke and Barron demonstrated that the expected minimax redundancy R_n for the memoryless sources is asymptotically given by

$$R_n = \frac{k-1}{2} \log\left(\frac{n}{2\pi}\right) + \log\left(\frac{\Gamma\left(\frac{1}{2}\right)^k}{\Gamma\left(\frac{k}{2}\right)}\right) + O\left(\frac{1}{n}\right). \tag{61}$$

Using Theorem 2, the average minimax redundancy for the two–stage codes for the case of the memoryless sources could be rewritten as follows.

Corollary 2 *In the universal compression of the family of memoryless sources $\mathcal{M}_0$, the average minimax redundancy of two–stage codes is obtained by*

$$R_n^{2p} = R_n + \log\Gamma\left(\frac{k+1}{2}\right) - \frac{k-1}{2} \log\left(\frac{k-1}{2e}\right), \tag{62}$$

where R_n is the average minimax redundancy for memoryless sources defined in (61).

Theorem 2 gives the extra redundancy due to the two–stage coding of the memoryless sources.

In the following, we present Theorem 3 for the special case of the memoryless sources.

Corollary 3 *Assume that the parameter vector θ follows Jeffreys' prior in the universal compression of the family of memoryless sources $\mathcal{M}_0$. Let ϵ be a real number. Then,*

$$\mathbf{P}\left[\frac{R_n(l_n,\theta)}{\frac{k-1}{2}\log n} \geq 1-\epsilon\right] \geq 1 - \frac{\Gamma\left(\frac{k}{2}\right)}{\sqrt{\pi}}\left(\frac{2}{n^{\epsilon}}\right)^{\frac{k-1}{2}}. \tag{63}$$

CHAPTER V

ELABORATION ON THE RESULTS

In this section, we elaborate on the significance of our results. In Chapter 5.1, we demonstrate that the average minimax redundancy underestimates the performance of source coding in the small to moderate length n for sources with small d. In Chapter 5.2, we compare the performance of two–stage codes with conditional two–stage codes. We show that the penalty term of two–stage coding is negligible for sources with large d as well as for the sequences of long n. In Chapter 5.3, we demonstrate that as the number of source parameters grow, the minimax redundancy well estimates the performance of the source coding. In Chapter 5.4, we show that the redundancy is significant in the compression of small to medium length sequences with large number of parameters.

5.1 Redundancy in Finite-Length Sequences with Small d

In Figures 1 and 2, the x-axis denotes a fraction P_0 and the y-axis represents a redundancy level R_0. The solid curves demonstrate the derived lower bound on the average redundancy of the conditional two–stage codes R_0 as a function of the fraction P_0 of the sources with redundancy larger than R_0, i.e., $\mathbf{P}[R_n(l_n^{c2p}, \theta) \geq R_0] \geq P_0$. In other words, at least a fraction P_0 of the sources that are chosen from Jeffreys' prior have an expected redundancy that is greater than R_0.

First, we consider a ternary memoryless information source denoted by $\mathcal{M}_0^3$. Let k be the alphabet size, where $k = 3$. This source may be parameterized using two parameters, i.e., $d = 2$. The unknown parameter vector is chosen from Jeffreys' prior in all of the examples. In Fig. 1, our results are compared to the average minimax redundancy, i.e., R_n from (11). Since the conditional two–stage codes achieve the

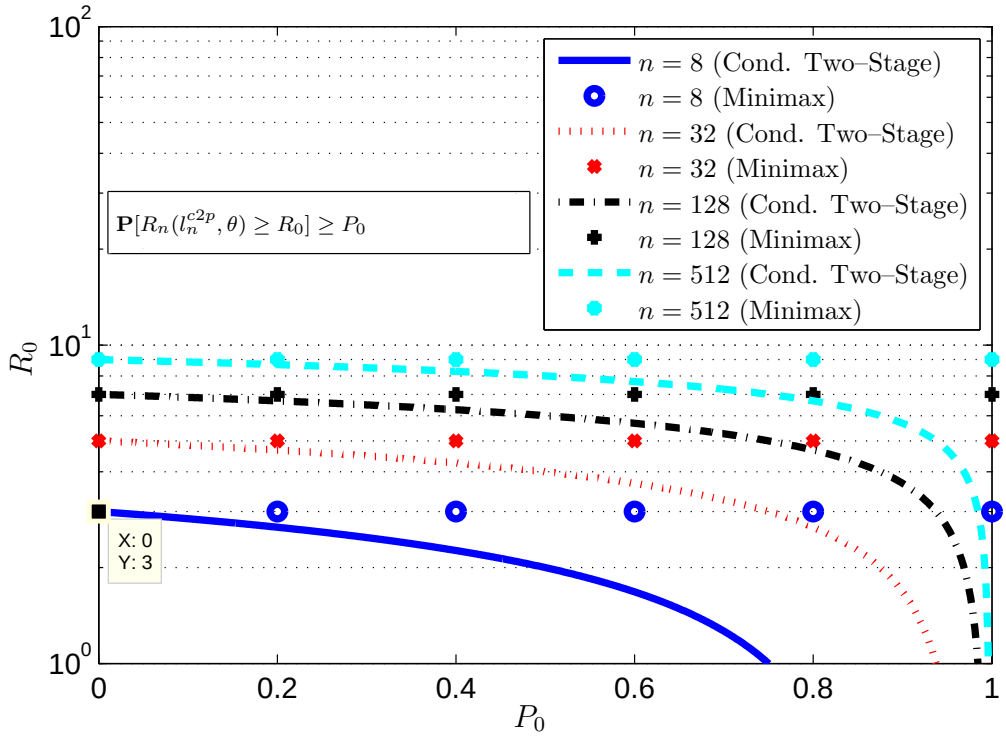

Figure 1: Average redundancy of the conditional two–stage codes (*Cond. Two–Stage*) and the average minimax redundancy (*Minimax*) as a function of the fraction of sources P_0 with $R_n(l_n^{c2p}, \theta) > R_0$. Memoryless source $\mathcal{M}_0^3$ with $k = 3$ and $d = 2$.

minimax redundancy of the regular codes, R_n is in fact the average minimax redundancy for the conditional two–stage codes (R_n^{c2p}) as well. The results are presented in bits. As shown in Fig. 1, at least 40% of ternary memoryless sequences of length $n = 32$ ($n = 128$) may not be compressed beyond a redundancy of 4.26 (6.26) bits. Also, at least 60% of ternary memoryless sequences of length $n = 32$ ($n = 128$) may not be compressed beyond a redundancy of 3.67 (5.68) bits. Note that as $n \to \infty$, the average redundancy approaches the average minimax redundancy for most sources.

Further, let $\mathcal{M}_1^2$ denote a binary first-order Markov source ($d = 2$). We present the finite-length compression results in Fig. 2 for different values of sequence length n. The values of n are chosen such that they are almost log(3) times the values of n for the ternary memoryless source in the first example. This choice has been made to equate the amount of information in the two sequences from $\mathcal{M}_0^3$ and $\mathcal{M}_1^2$

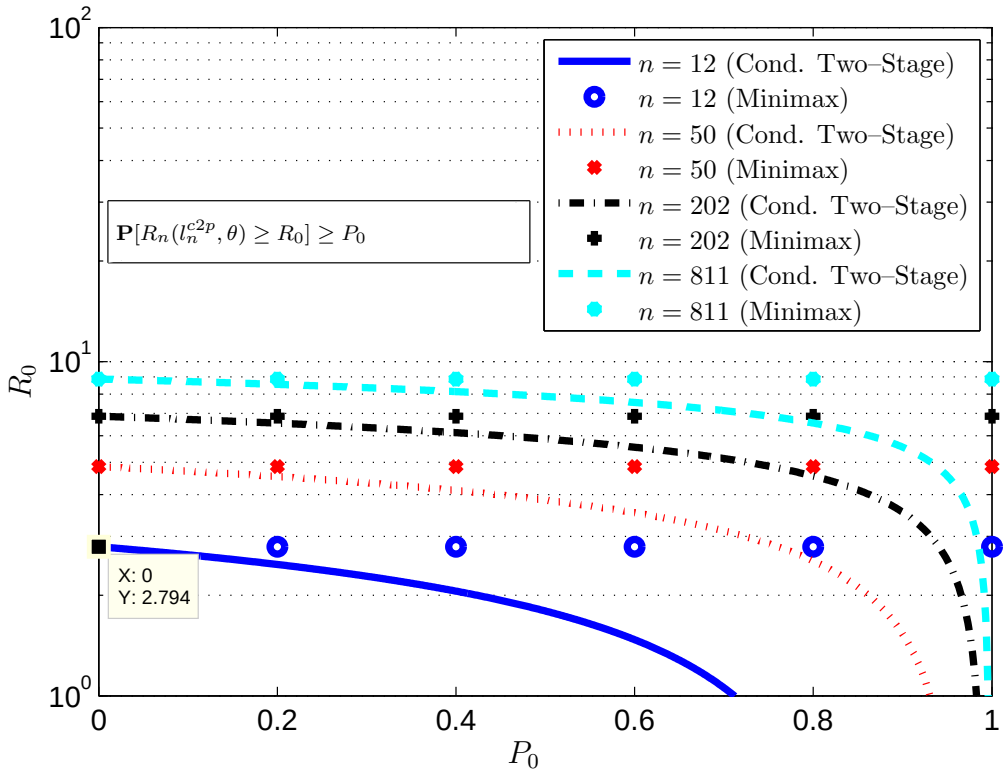

Figure 2: Average redundancy of the conditional two–stage codes (*Cond. Two–Stage*) and the average minimax redundancy (*Minimax*) as a function of the fraction of sources P_0 with $R_n(l_n^{c2p}, \theta) > R_0$. First-order Markov source $\mathcal{M}_1^2$ with $k = 2$ and $d = 2$.

allowing a fair comparison. For example, a sequence of length $n = 8$ from source $\mathcal{M}_0^3$, consisted of 8 ternary symbols, is equivalent to $8\log(3)$ bits of information that is almost equivalent to 12 bits in $\mathcal{M}_1^2$.

Figure 2 shows that the average minimax redundancy of two–stage codes for the case of $n = 12$ is given as $R_{12} \approx 2.794$ bits. Comparing Fig. 1 with Fig. 2, we conclude that the average redundancy of universal compression for a binary first-order Markov source is very similar to that of the ternary memoryless source, suggesting that d is the most important parameter in determining the redundancy of finite-length sources. This subtle difference becomes even more negligible as $n \to \infty$ since the dominating factor of redundancy for both cases approaches to $\frac{d}{2}\log n$.

As demonstrated in Figs. 1 and 2, there is a significant gap between the known

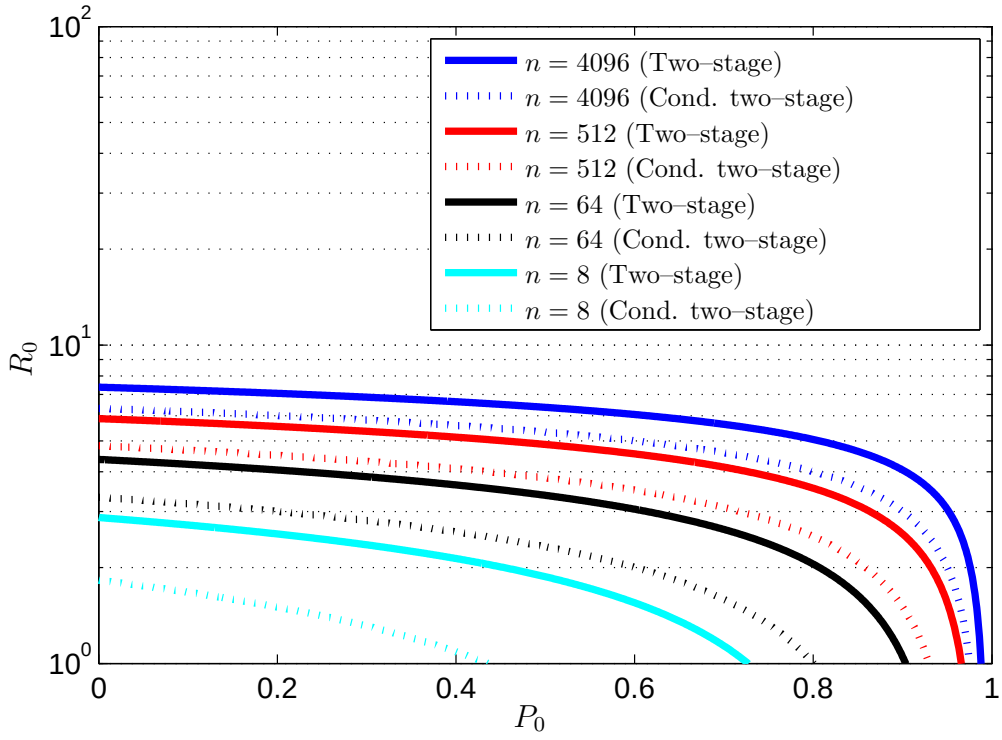

Figure 3: Average redundancy of the two–stage codes (solid) vs average redundancy of the conditional two—stage codes (dotted) as a function of the fraction of sources P_0. Memoryless source $\mathcal{M}_0^2$ with $k = 2$ and $d = 1$.

result by the average minimax redundancy and the finite-length results obtained in this thesis when a high fraction P_0 of the sources is concerned. Hence, for many sources, the average minimax redundancy overestimates the average redundancy in universal source coding of finite-length sequences where the number of the parameters is small. In other words, the compression performance of a high fraction of finite-length sources would be better than that of the average minimax redundancy estimate.

5.2 Two–Stage Codes Vs Conditional Two–Stage Codes

We now compare the finite-length performance of the two–stage codes with the conditional two–stage codes on the class of binary memoryless source $\mathcal{M}_0^2$ with $k = 2$ ($d = 1$). The results are presented in Figure 3. The solid line (dotted line) demonstrates the the lower bound for the two–stage codes (conditional two–stage codes). As

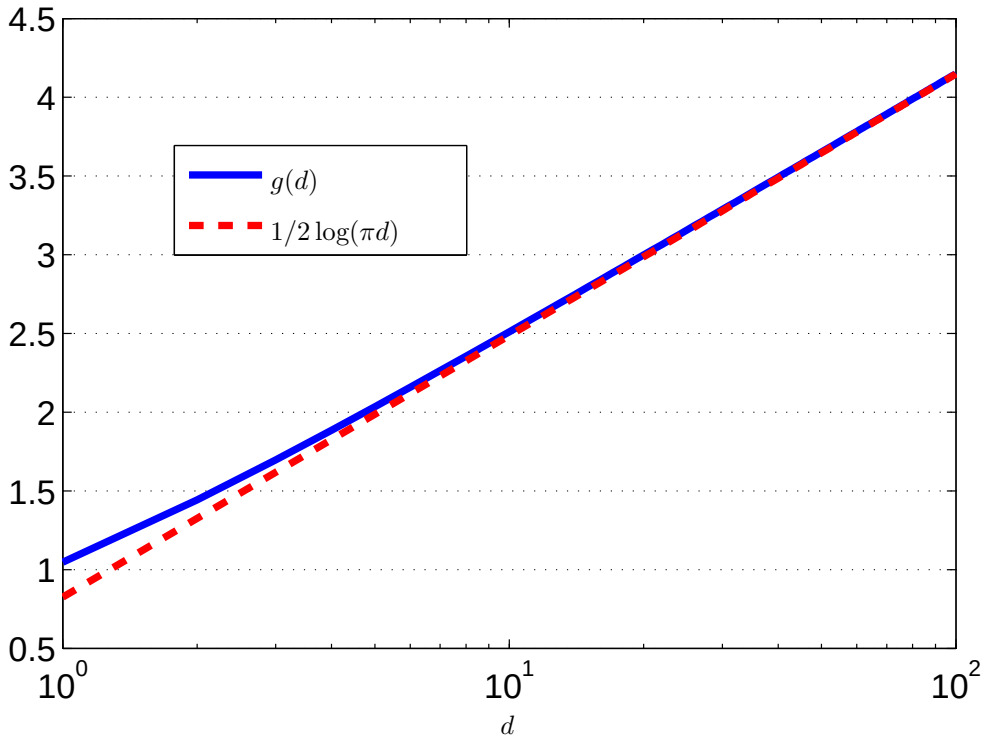

Figure 4: The extra redundancy incurred due to the two–stage assumption on the code as a function of d.

can be seen, the gap between the achievable compression using two–stage codes and that of the conditional two–stage codes constitutes a significant fraction of the average redundancy for small n. For a Bernoulli source, the average minimax redundancy of the two–stage code is given in (62) as

$$R_n^{2p} = R_n + \frac{1}{2}\log\left(\frac{\pi e}{2}\right) \approx R_n + 1.048. \tag{64}$$

The average minimax redundancy of two–stage codes for the case of $n = 8$ is given as $R_8^{2p} \approx 2.86$ bits while that of the conditional two–stage codes (i.e., all regular codes) is $R_8 \approx 1.82$. Thus, the two–stage codes incur an extra compression overhead of more than 50% for $n = 8$.

In Theorem 2, we derived that the extra redundancy $g(d)$ incurred by the two–stage assumption. We further use Stirling's approximation for sources with large number of parameters in order to show the asymptotic behavior of $g(d)$ as $d \to \infty$.

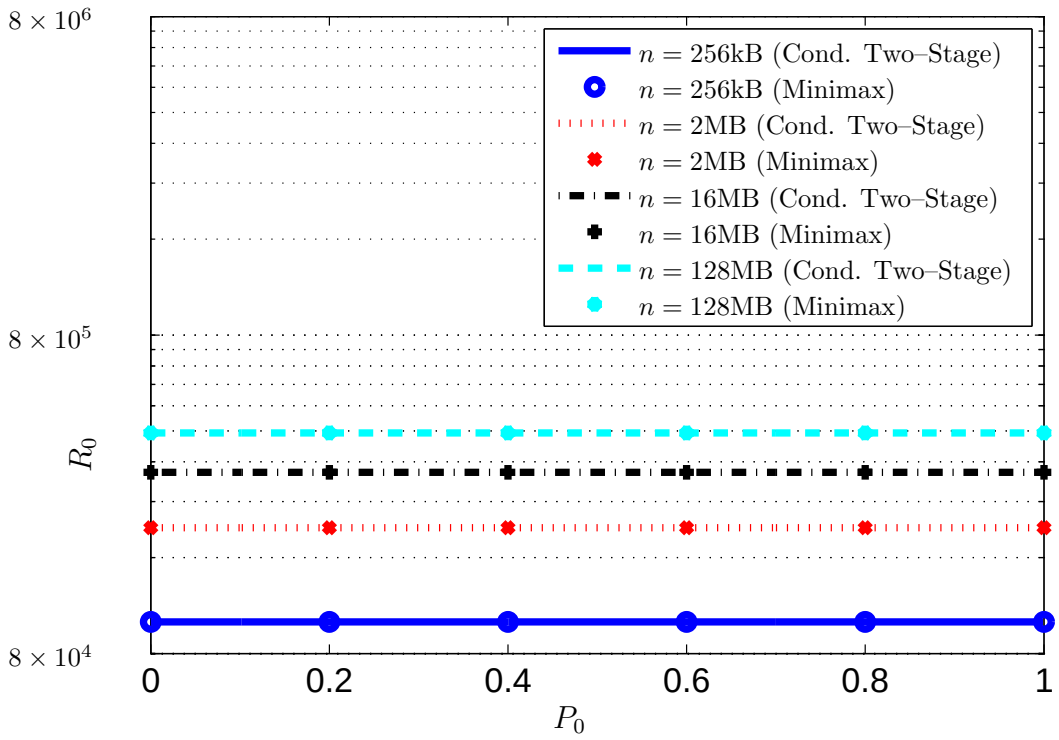

Figure 5: Average redundancy of the conditional two–stage codes (*Cond. Two–Stage*) and the average minimax redundancy (*Minimax*) as a function of the fraction of sources P_0 with $R_n(l_n^{c2p}, \theta) > R_0$. First-order Markov source with $k = 256$ and $d = 65280$. The sequence length n is measured in bytes (B).

That is, asymptotically, we have

$$g(d) = \frac{1}{2} \log (\pi d) + o(1). \tag{65}$$

Note that $o(1)$ denotes a function of d and not n here. As demonstrated in Figure 4, $g(d)$ is increasing logarithmically with d as $d \to \infty$. Finally, we must note that the main term of redundancy in R_n is $\frac{d}{2} \log n$, which is linear in d, but the penalty term $g(d)$ is logarithmic in d. Hence, the effect of the two–stage assumption becomes negligible for the families of sources with larger d.

5.3 Redundancy in Finite-Length Sequences with Large d

The results of this thesis can be used to quantify the significance of redundancy in finite-length compression. We consider a first-order Markov source with alphabet size

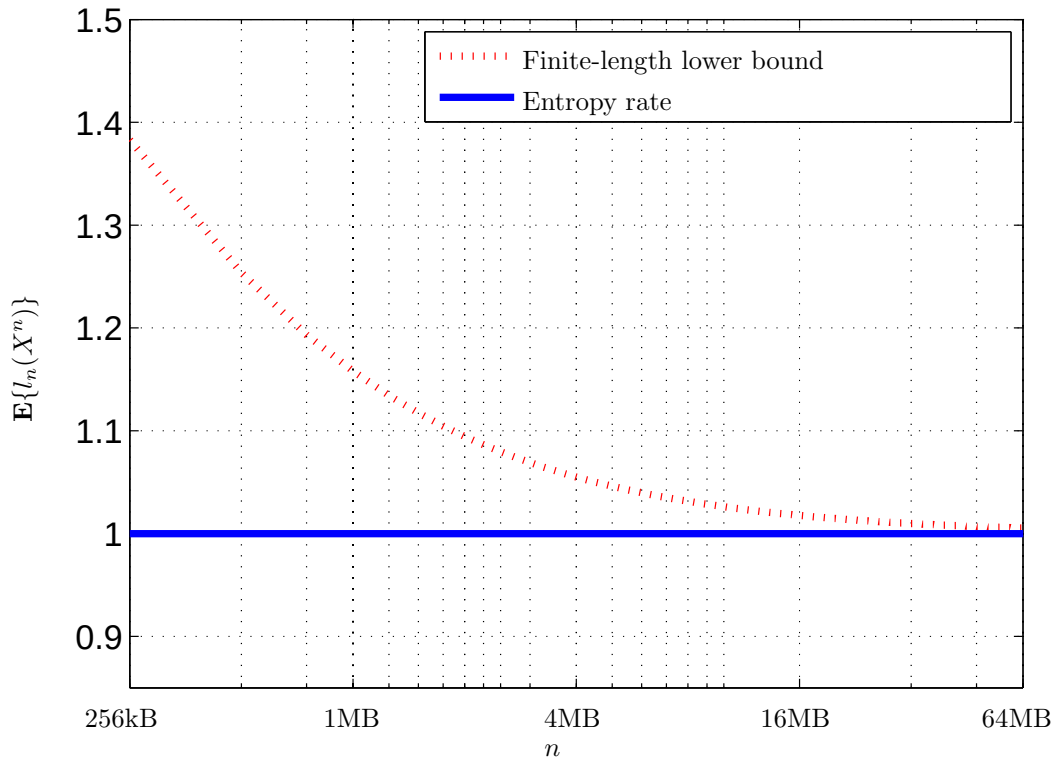

Figure 6: The Lower bound on compression for at least 95% of the sources as a function of sequence length n.

$k = 256$. We intentionally picked this alphabet size as it is a common practice to use the byte as a source symbol. This source may be represented using $d = 256 \times 255 = 62580$ parameters. In Figure 5, the achievable redundancy is demonstrated for four different values of n. Here, again the redundancy is measured in bits. The curves are almost flat when d and n are very large validating our results that the average minimax redundancy provides a good estimate on the achievable compression for most sources. The sequence length in this example is presented in bytes (B). We observe that for $n = 256kB$, we have $R_n(l_n, \theta) \geq 100,000$ bits for most sources. Further, the extra redundancy due to the two–stage coding $g(d) \approx 8.8$ bits, which is a negligible fraction of the redundancy of $100,000$ bits.

5.4 Significance of Redundancy in Finite-Length Compression

Figure 6 demonstrates the average number of bits per symbol required to compress the class of the first-order Markov sources whose entropy rates are 1 bit per source symbol (per byte). We have chosen this value since many practical sources have an entropy rate that is smaller than 1 bit per source symbol. The dashed curve demonstrates the lower bound on the achievable compression for at least 95% of the sources, i.e., at least 95% of the sources from this class may not be compressed with a redundancy smaller than the dashed curve. The solid curve denotes the entropy rate of the source. As can be seen, the compression overhead is 38%, 16%, 5.5%, 1.7%, and 0.5% for sequences of lengths 256kB, 1MB, 4MB, 16MB, and 64MB, respectively. Hence, we conclude that redundancy may be significant for the compression of small sequences of length up to 1MB. On the other hand, redundancy is negligible for sequences of length 64MB and higher.

CHAPTER VI

APPLICATION: CHARACTERIZING CONTEXT MEMORIZATION GAIN

In this section, we use the results we developed thus far to characterize the achievable improvement due to the context memorization when compressing a small or moderate size sequence from the same source. In Chapter 6.1, we present the problem setup for the context memorization gain followed by the review of the existing results that we need in the derivation of our main results. In Chapter 6.2, we present our main results on the gain. In Chapter 6.3, we demonstrate the significance of our results.

6.1 Problem Setup and Background

Thus far, we established a lower bound on the average redundancy in the universal compression of a single finite-length sequence. We learned that, on the average, significant redundancy is present when a single finite-length sequence is compressed. In this section, we consider a new scenario in which the source encoder wishes to compress a single finite-length sequence. However, we assume that the encoder (and hence the decoder) has already visited, i.e., compressed, another sequence from the same context, i.e., generated by the same information source. We refer to this new scenario as compression with context memory. Further, we assume that the problem is a universal compression problem and the context is unknown a priori, i.e., the parameter vector of the information source in unknown. While relevant, this problem of compressing a sequence with context memorization is different from those addressed by the distributed source coding techniques that are concerned with multiple sources sending correlated information to the same destination [16, 23, 29, 33]. Instead, we

with to establish the fundamental benefits of context memorization in the universal compression of a single small or moderate length sequence from the same context (originating from the same source).

Let x^n be a sequence of length n from the FSMX source P that is to be compressed. Further, assume that $y^{\mathcal{M}}$ is a sequence of length $\mathcal{M}$ from the same source P. Note that we assume that the sequences x^n and $y^{\mathcal{M}}$ are samples from the statistically independent random sequences X^n and $Y^{\mathcal{M}}$ generated by P. By context memorization, we mean that both the encoder and the decoder have already visited the sequence $y^{\mathcal{M}}$.

Our goal is to investigate the fundamental benefits of context memorization in the universal compression of the finite-length sequence x^n. In particular, we show that although X^n and $Y^{\mathcal{M}}$ are *independent* given that the source model is *known*, context memorization can indeed result in the reduction of the expected length of the codeword associated with x^n, i.e., better compression of an individual sequence of length n on the average. This seemingly counter intuitive phenomenon is due to the fact that the source model is *unknown* a priori and the compression is *universal*. Therefore, the sequence $y^{\mathcal{M}}$ indeed contains useful information about the unknown source parameters, that can be used in the compression of x^n.

As an example, consider the extreme case when $\mathcal{M} \to \infty$. In this case, the sequence $y^{\mathcal{M}}$ exactly determines the unknown source parameters. The fact that both encoder and the decoder have access to $y^{\mathcal{M}}$ simply means that both the encoder and the decoder exactly know the source parameters, and hence, the problem reduces to coding of x^n with known source parameters. Consequently, the sequence x^n can be efficiently compressed with a redundancy smaller than 1 bit, e.g., using a Huffman code, as opposed to the redundancy of about $\frac{d}{2}\log n$ of the universal compression of a source with d unknown parameters.

In order to demonstrate the benefits of context memorization, we will compare

two schemes:

- Comp (Universal compression of an individual sequence with no context memorization), which applies a sole universal compression on a sequence without using context memorization.

- CompCM (Universal compression of an individual sequences with context memorization), which assumes that the encoder and decoder have access to a memorized context and utilizes the context when compressing x^n.

In what follows, we obtain the context memorization gain for the conditional two–stage codes. Therefore, the performance of Comp may be characterized by the expected redundancy of the conditional two–stage coding, $R_n(l_n^{c2p}, \theta)$. For the ease of notation, we drop the notation $c2p$ although all the results in this section are derived for conditional two–stage codes. In the case of CompCM, let $l_{n|\mathcal{M}}$ be the regular length function with the context memorization, where the encoder and the decoder have access to a memorized context $y^{\mathcal{M}}$. Further, denote $R_n(l_{n|\mathcal{M}}, \theta)$ as the expected redundancy of encoding a sequence of length n form the source with parameter vector θ using the length function $l_{n|\mathcal{M}}$, i.e., with context memorization. The average redundancy of CompCM is given by

$$R_n(l_{n|\mathcal{M}}, \theta) = \mathbf{E} l_{n|\mathcal{M}}(X^n) - H_n(\theta). \tag{66}$$

Let $Q(l_n, l_{n|\mathcal{M}}, \theta)$ be defined as the ratio of the expected codeword length of Comp to that of CompCM as

$$Q(l_n, l_{n|\mathcal{M}}, \theta) \triangleq \frac{\mathbf{E} l_n(X^n)}{\mathbf{E} l_{n|\mathcal{M}}(X^n)} = \frac{H_n(\theta) + R_n(l_n, \theta)}{H_n(\theta) + R_n(l_{n|\mathcal{M}}, \theta)}. \tag{67}$$

Let δ be a real number such that $0 < \delta < 1$. We denote $g(n, \mathcal{M}, \delta)$ as the fundamental gain of the context memorization on the family of FSMX sources $\mathcal{P}^d$ on a sequence of length n using a context sequence of length $\mathcal{M}$ for a fraction $1 - \delta$ of the sources,

which is defined as follows:

$$g(n, \mathcal{M}, \delta) = \sup_{z \in \mathbb{R}} \left\{ z : \ \mathbf{P}\left[Q(l_n, l_{n|\mathcal{M}}, \theta) \geq z \right] \geq 1 - \delta \right\}. \tag{68}$$

In other words, the fundamental gain of memorization is at least $g(n, \mathcal{M}, \delta)$ for a fraction $1 - \delta$ of the sources in the family. The following is a trivial lower bound on the context memorization gain.

Fact 3 *The fundamental gain of context memorization is:* $g(n, \mathcal{M}, \delta) \geq 1$.

Proof: Note that there exists $l_{n|\mathcal{M}}$ such that for all l_n, we have $R_n(l_{n|\mathcal{M}}, \theta) \leq R_n(l_n, \theta)$. The rational is that CompCM works no worse than Comp. Thus, we have $Q(l_n, l_{n|\mathcal{M}}, \theta) \geq 1$ and the claim follows. ■

Fact 3 simply states that the context memorization does not degrade the performance of the universal compression. To obtain a lower bound on the memorization gain $g(n, \mathcal{M}, \delta)$, we need a lower bound on $R_n(l_n, \theta)$. Further, we require a useful upper bound on $R_n(l_{n|\mathcal{M}}, \theta)$. In Theorem 3, we established the desired lower bound on $R_n(l_n, \theta)$, i.e., the redundancy of Comp. In order to obtain an upper bound on the redundancy of CompCM, we use the known results on Context Tree Weighting (CTW) [39]. Note that we do not claim that the CTW with memorization is the optimal scheme. In fact, the sequential compression of the symbols results in the performance loss. However, the CTW compression using context memorization provides us with a useful upper bound on the average redundancy with context memorization $R_n(l_{n|\mathcal{M}}, \theta)$. Willems et. al. proved that CTW is an optimal compression scheme in the sense that it achieves Rissanen's lower bound in Fact 1. Let λ_n denote the CTW universal length function. Drmota et. al. derived the average redundancy of the CTW [11]:

Fact 4 *The average redundancy of the Context Tree Weighting (CTW) algorithm is*

given by

$$R_n(\lambda_n, \theta) = \frac{d}{2} \log \left(\frac{2n}{\pi e} \right) + 2d + f_0(\theta) - E_n + O\left(\frac{1}{\sqrt{n}} \right), \tag{69}$$

where $f_0(\theta)$ is only a function of the parameter vector θ and $E_n \approx \frac{d}{2}$ is the erratic part of the redundancy.

Note that Fact 4 is a strong result in the sense that it precisely characterizes the average redundancy for all parameters θ and all sequences for all ranges of n. Denote $\lambda_{n|\mathcal{M}}$ as the CTW universal length function using memorization of the context of length $\mathcal{M}$. By memorization of the context, we mean that a sequence of length $\mathcal{M}$ from the source has already been encoded using the CTW. Then, the resulting context tree is used for the compression of the sequence of length n. Let $R_n(\lambda_{n|\mathcal{M}}, \theta)$ denote the expected redundancy of encoding the source with parameter θ using the CTW length function $\lambda_{n|\mathcal{M}}$. The following characterizes the redundancy for the compression of the length n sequence.

Fact 5 *The average redundancy of CompCM (the compression with context memorization) is upper bounded as*

$$R_n(l_{n|\mathcal{M}}, \theta) \leq \hat{R}(n, \mathcal{M}), \tag{70}$$

where

$$\hat{R}(n, \mathcal{M}) \triangleq \frac{d}{2} \log \left(1 + \frac{n}{\mathcal{M}} \right) + 2. \tag{71}$$

Proof: Fact 4 describes the average redundancy of the CTW. As can be seen in (69), only the first term and the erratic redundancy are significant functions of n. In other words, the rest are upper bounded by a constant overhead for any sequence length. Therefore, when the context is already memorized, there is no extra overhead cost except for the penalty of the integer-length codeword requirement. We further assume that the difference between the erratic redundancy terms at lengths $\mathcal{M}$ and

$n + \mathcal{M}$ is negligible. Since the CTW algorithm is sequential, $R_n(\lambda_{n|\mathcal{M}}, \theta)$ is obtained as

$$R_n(\lambda_{n|\mathcal{M}}, \theta) = R_{\mathcal{M}+n}(\lambda_{\mathcal{M}+n}, \theta) - R_{\mathcal{M}}(\lambda_{\mathcal{M}}, \theta) + C_1, \tag{72}$$

where C_1 arises due to the integer-length codeword requirement and $C_1 < 2$. Then, the result in Fact 5 is obtained by noting that

$$R_n(l_{n|\mathcal{M}}, \theta) \leq R_n(\lambda_{n|\mathcal{M}}, \theta).$$

■

Fact 5 sets an upper bound on the average redundancy of compression using context memorization.

6.2 Main Results on the Context Memorization Gain

We are now equipped to establish a lower bound on the fundamental gains obtained when context memorization is leveraged to achieve a better compression rate in CompCM. The next theorem characterizes the fundamental gains:

Theorem 4 *Assume that the parameter vector θ follows Jeffreys' prior in the universal compression of the family of FSMX sources $\mathcal{P}^d$. Then,*

$$g(n, \mathcal{M}, \delta) \geq 1 + \frac{R_n + \log(\delta) - \hat{R}(n, \mathcal{M})}{H_n(\theta) + \hat{R}(n, \mathcal{M})}, \tag{73}$$

where R_n is the average minimax redundancy defined in (11).

Proof: First note that

$$\begin{aligned} Q(l_n, l_{n|\mathcal{M}}, \theta) &= \frac{\mathbf{E}l_n(X^n)}{\mathbf{E}l_{n|\mathcal{M}}(X^n)} \geq \frac{\mathbf{E}l_n(X^n)}{\mathbf{E}\lambda_{n|\mathcal{M}}(X^n)} && (74)\\ &= \frac{H_n(\theta) + R_n(l_n, \theta)}{H_n(\theta) + R_n(\lambda_{n|\mathcal{M}}, \theta)} && (75)\\ &\geq \frac{H_n(\theta) + R_n(l_n, \theta)}{H_n(\theta) + \hat{R}(n, \mathcal{M})} && (76)\\ &\triangleq \hat{Q}(l_n, \mathcal{M}, \delta), && (77) \end{aligned}$$

where the inequality in (76) is due to Fact 5. Further, let $\hat{g}(n, m, \delta)$ be defined as

$$\hat{g}(n, \mathcal{M}, \delta) \triangleq \sup_{z \in \mathbb{R}} \left\{ z : \ \mathbf{P}\left[\hat{Q}(l_n, \mathcal{M}, \delta) \geq z \right] \geq 1 - \delta \right\}. \tag{78}$$

Equation (77) implies

$$\mathbf{P}\left[Q(l_n, l_{n|\mathcal{M}}, \theta) \geq z \right] \geq \mathbf{P}\left[\hat{Q}(l_n, \mathcal{M}, \delta) \geq z \right]. \tag{79}$$

Thus,

$$g(n, \mathcal{M}, \delta) \geq \hat{g}(n, \mathcal{M}, \delta). \tag{80}$$

We can now apply Theorem 3 to $R_n(l_n, \theta)$ in (76) with the proper choice of ϵ in order to obtain a lower bound on $\hat{g}(n, \mathcal{M}, \delta)$, which will complete the proof. ■

Let $g(n, \infty, \delta)$ be defined as the achievable gain of context memorization where there is no constraint on the size of the context, i.e, the encoder and the decoder have access to an infinite length sequence from the source P.

$$g(n, \infty, \delta) \triangleq \lim_{\mathcal{M} \to \infty} g(n, \mathcal{M}, \delta) \tag{81}$$

The following Corollary quantifies the achievable context memorization gain when there is no restriction on the memory size, which is obtained by taking $\lim_{\mathcal{M} \to \infty} \hat{R}(n, \mathcal{M})$.

Corollary 4 *Assume that the parameter vector θ follows Jeffreys' prior in the universal compression of the family of FSMX sources $\mathcal{P}^d$. Then,*

$$g(n, \infty, \delta) \geq 1 + \frac{R_n + \log(\delta) - 2}{H_n(\theta) + 2}. \tag{82}$$

Next, we consider the case where the sequence length n grows to infinity. Intuitively, we would expect that the context memorization gain become negligible for the compression of long sequences, i.e., large n. Let $g(\infty, \mathcal{M}, \delta)$ be the context memorization gain for the universal compression of an infinite length sequence, defined as

$$g(\infty, \mathcal{M}, \delta) \triangleq \lim_{n \to \infty} g(n, \mathcal{M}, \delta). \tag{83}$$

In the following, we claim that the context memorization does not provide any benefit in this case

Fact 6 *The context memorization gain for an infinite-length sequence is given as* $g(\infty, \mathcal{M}, \delta) = 1$.

Proof: As $n \to \infty$, an asymptotically optimal universal compression scheme (one that asymptotically achieves the entropy of the sequences) would have a redundancy that is $o(n)$. Therefore,

$$\lim_{n\to\infty} \frac{\mathbf{E}l_n(X^n)}{H_n(\theta)} = 1. \tag{84}$$

Since entropy is the absolute lower bound on the achievable expected codeword length and $\mathbf{E}l_{n|\mathcal{M}}(X^n) \leq \mathbf{E}l_n(X^n)$, we have

$$\lim_{n\to\infty} \frac{\mathbf{E}l_{n|\mathcal{M}}(X^n)}{H_n(\theta)} = 1, \tag{85}$$

which proves the claim. ■

Finally, we note that these results are also applicable to the practical scenario, where we have a library of sequences that consists of several *independent* sequences from the same source P that need to be compressed and stored individually. Therefore, the sequences may not be concatenated and stored. Our results in the first part of the thesis quantify the performance limits of the compression, when each file is compressed and stored without regard to the rest of the sequences in the library. On the other hand, since the sequences share the same context, we may consider the concatenation of all of the sequences in the library as a memorized context $y^{\mathcal{M}}$ for the compression of the sequence x^n. Roughly speaking, this approximation is valid when the length of the individual sequences in the library are fairly large compared to the depth of the context tree so that they capture the statistics of the source. Therefore, the results in the second part of the thesis will quantify the benefits of the context memorization in the compression of x^n.

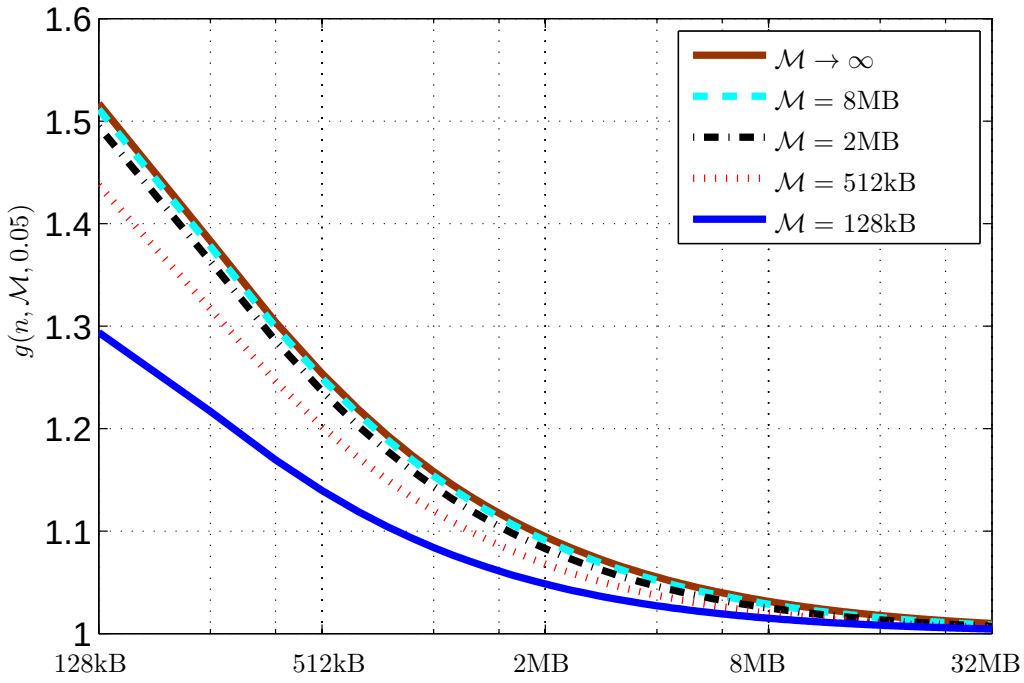

Figure 7: Lower bound on the context memorization gain, $g(n, \mathcal{M}, 0.05)$, as a function of sequence length n for various $\mathcal{M}$.

6.3 Significance of the Results

In this section, we demonstrate the significance of the context memorization gain through an example. We again consider a first-order Markov source with alphabet size $k = 256$. In Figure 7, the lower bound on the context memorization gain is demonstrated as a function of the sequence length n for different values of $\mathcal{M}$. As can be seen, significant improvement in the compression may be achieved using context memorization. For example, the lower bound on $g(1\text{MB}, \mathcal{M}, 0.05)$ is equal to 1.084, 1.120, 1.144, and 1.154, when the context parameter $\mathcal{M}$ is 128kB, 512kB, 2MB, and 8MB, respectively. Further, $g(1\text{MB}, \infty, 0.05) = 1.158$. As demonstrated in Figure 7, the gain of a context of length 8MB is very close to $g(n, \infty, \delta)$, and hence, increasing the context memory beyond 8MB does not result in the significant increase

of the memorization gain. We further observe that as $n \to \infty$, the context memorization gain becomes negligible. For example, at $n = 32\text{MB}$ and $\mathcal{M} \to \infty$, we have $g(32\text{MB}, \infty, 0.05) \approx 1.01$, which is a negligible improvement. On the other hand, at $n = 128\text{kB}$ and $\mathcal{M} = 8\text{MB}$, we have $g(32\text{MB}, 8\text{MB}, 0.05) \approx 1.51$, i.e., more than 50% improvement is achieved in the compression performance with memorization of a context of length 8MB.

CHAPTER VII

CONCLUSION

In this thesis, we investigated the average redundancy rate of universal coding schemes on FSMX sources in the *finite*-length regime. We derived a lower bound on the probability of the event that an information source chosen using Jeffreys' prior from the family of FSMX information sources is not compressible beyond any certain redundancy level. This work may be viewed as the finite-length extension of the previous asymptotic results. We demonstrated that the average minimax redundancy underestimates the performance of source coding in the small to moderate length sequences for sources with small number of parameters. We also compared the performance of two–stage codes with conditional two–stage codes, where we showed that the penalty term of two–stage coding is negligible for sources with large d as well as for the sequences of sufficient lengths. Further, we demonstrated that as the number of source parameters grow very large, the minimax redundancy provides accurate estimate for the performance of the source coding. We also showed that the redundancy is significant in the compression of small to medium length sequences with large number of source parameters. Finally, we concluded that the context memorization can significantly improve the performance of source coding when the sequence length is small or moderate and the number of source parameters are not too small.

APPENDIX A

PROOF OF LEMMA 1

We have

$$\mathbf{E}\log\left(\frac{\mu_\theta(X^n)}{\mu_\gamma(X^n)}\right) = \sum_{x^n}\mu_\theta(x^n)\log\left(\frac{\mu_\theta(x^n)}{\mu_\gamma(x^n)}\right)$$

$$= D_n(\mu_\theta||\mu_\beta) + \sum_{x^n}\mu_\theta(x^n)\log\left(\frac{\mu_\beta(x^n)}{\mu_\gamma(x^n)}\right)$$

$$= D_n(\mu_\theta||\mu_\beta) + \sum_{x^n}\min_{\phi_i\in\Phi^m}\mu_\theta(x^n)\log\left(\frac{\mu_\beta(x^n)}{\mu_{\phi_i}(x^n)}\right) \tag{86}$$

$$\geq D_n(\mu_\theta||\mu_\beta) - f_n(\Phi^m,\theta), \tag{87}$$

where $f_n(\Phi^m,\theta)$ is given by

$$f_n(\Phi^m,\theta) = \sum_{x^n:\gamma\neq\beta}\mu_\theta(x^n)\log\left(\frac{1}{\mu_\beta(x^n)}\right). \tag{88}$$

Note that the sum is taken over all sequences x^n whose best estimate γ is not equal to β. Our goal is to show find an upper bound on $f_n(\Phi^m,\theta)$.

As the length n of the sequence increases, the source parameters concentrate around the mean and the tail of the distribution becomes exponentially small. Let $\hat{\theta}(X^n)$ be an estimator of θ. Let $\Sigma_\theta(\hat{\theta}(X^n))$ be the covariance matrix for the estimator $\hat{\theta}(X^n)$. According to the Cramer-Rao bound, we have

$$\Sigma_\theta(\hat{\theta}(X^n)) \succeq nI_n(\theta), \tag{89}$$

i.e., for all $\theta\in\Theta^d$,

$$\frac{1}{n}\theta^T\Sigma_\theta(\hat{\theta}(X^n))\theta \geq \theta^T I_n(\theta)\theta. \tag{90}$$

Thus, assuming that the central limit theorem conditions hold, the probability distributions μ_θ and μ_β are well approximated with normal distributions. Let $\mathcal{G}_\theta(\cdot)$ be

defined as

$$\mathcal{G}_\theta(\hat{\theta}) = \frac{n^{\frac{1}{2}}|I_n(\theta)|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{n}{2}(\hat{\theta}-\theta)^T I_n(\theta)(\hat{\theta}-\theta)\right). \tag{91}$$

Thus, assuming that $\hat{\theta}$ is the maximum likelihood estimate for the unknown parameter θ given the sequence x^n, we have

$$\begin{aligned} f_n(\Phi^m, \theta) &\approx \int_{\mathcal{G}_\gamma(\hat{\theta}) > \mathcal{G}_\beta(\hat{\theta})} \mathcal{G}_\theta(\hat{\theta}) \log\left(\frac{1}{\mathcal{G}_\beta(\hat{\theta})}\right) \\ &\leq \int_{D_n(\mu_\theta||\mu_{\hat{\theta}}) \geq D_n(\mu_\theta||\mu_\beta)} \mathcal{G}_\theta(\hat{\theta}) \log\left(\frac{1}{\mathcal{G}_\beta(\hat{\theta})}\right) \end{aligned} \tag{92}$$

Therefore, by working out details we can show that

$$f_n(\Phi^m, \theta) = O\left(\exp\left(-\frac{n}{2}(\beta-\theta)^T I_n(\theta)(\beta-\theta)\right)\right) \tag{93}$$

Note that as discussed in the proof of Lemma 2, we have

$$\frac{n}{2}(\beta-\theta)^T I_n(\theta)(\beta-\theta) \approx \frac{1}{\log e} D_n(\mu_\theta||\mu_\beta), \tag{94}$$

and hence,

$$f_n(\Phi^m, \theta) = O\left(2^{-D_n(\mu_\theta||\mu_\beta)}\right), \tag{95}$$

which completes the proof. ■

APPENDIX B

PROOF OF LEMMA 2

Let $f(\theta) = D_n(\mu_\theta||\mu_\beta)$. We may use Taylor series to characterize $f(\theta)$ for θ close to β.

$$\frac{1}{n}D_n(\mu_\theta||\mu_\beta) \approx \mathcal{E}_\beta(\theta) + O(||\theta - \beta||^3), \tag{96}$$

where

$$\mathcal{E}_\beta(\theta) = \frac{\log e}{2}(\theta - \beta)^T I_n(\beta)(\theta - \beta) \tag{97}$$

Since we have

$$\lim_{n\to\infty} \frac{1}{n}D_n(\mu_\theta||\mu_\beta) = 0, \tag{98}$$

the error term of $O(||\theta - \beta||^3)$ is negligible. Further, note that $\mathcal{E}_\beta(\theta) \leq \delta$, where $\delta > 0$, defines an ellipsoid on the d-dimensional space of θ. It is straightforward to demonstrate that the volume of the ellipsoid is given by

$$V_d(\beta, \delta) = \frac{C_d}{|I(\beta)|^{\frac{1}{2}}} \left(\frac{2\delta}{\log e}\right)^{\frac{d}{2}}, \tag{99}$$

where C_d is the volume of the d-dimensional unit ball. Moreover, since θ follows Jeffreys' prior, the probability measure covered by the ellipsoid is given by

$$\begin{aligned}\mathbf{P}\left[\mathcal{E}_\beta(\theta) \leq \delta\right] &= V_d(\beta, \delta)\left(\frac{|I(\beta)|^{\frac{1}{2}}}{\int |I(\theta)|^{\frac{1}{2}} d\theta}\right) \\ &= \frac{C_d}{\int |I(\theta)|^{\frac{1}{2}} d\theta}\left(\frac{2\delta}{\log e}\right)^{\frac{d}{2}}.\end{aligned} \tag{100}$$

Thus, the volume defined by $\frac{1}{n}D_n(\mu_\theta||\mu_\beta) < \delta$ is almost equal to the volume $\mathcal{E}_\beta(\theta) < \delta$, which completes the proof of the first claim. Although the volume of the ellipsoid depends on the point β in the parameter space, the probability measure of the ellipsoid does not depend on β.

For the second claim, let the event $\mathcal{V}_i$ be the defined as

$$\mathcal{V}_i = \left\{ \omega \in \Omega : \ \frac{1}{n} D_n(\mu_\theta || \mu_{\phi_i}) < \delta \right\}. \tag{101}$$

Note that there are 2^m choices for ϕ_i. For all $1 < i < 2^m$, in the first claim, we found an upper bound on the probability of the event $\mathcal{V}_i$. Thus, using the union bound, we can upper bound the probability of $\bigcup_{i=1}^{2^m} \mathcal{V}_i$. Define the following event.

$$\mathcal{W} = \left\{ \omega \in \Omega : \ \min_i \frac{1}{n} D_n(\mu_\theta || \mu_{\phi_i}) < \delta \right\}. \tag{102}$$

The second claim is obtained by noting that

$$\mathcal{W} = \bigcup_{i=1}^{2^m} \mathcal{V}_i. \tag{103}$$

■

APPENDIX C

PROOF OF LEMMA 3

First, note that

$$R_n(l_n^{2p}, \theta) - R_n(l_n^{c2p}, \theta) = \mathbf{E} \log \left(\frac{1}{A_m(\gamma(X^n))} \right). \tag{104}$$

According to (40), m increases as ϵ decreases until ϵ is minimized and the average minimax redundancy is achieved as in (43). Let $|S_m(\gamma)|$ be the number of the sequences whose optimally estimated point (maximum likelihood estimation) is γ. Increasing m results in the increase of the number of the estimate points. Thus, $|S_m(\gamma)|$ decreases with m on the average and so does $A_m(\gamma)$. Therefore, we would conclude that $\mathbf{E} \log \left(\frac{1}{A_m(\gamma(X^n))} \right)$ is an increasing function of m. As discussed earlier, we optimized m in order to find the best lower bound on the redundancy in Theorem 1. As can be seen in (40), the optimal value of m is decreasing with ϵ. Thus, in order to maximize $\mathbf{E} \log \left(\frac{1}{A_m(\gamma(X^n))} \right)$, we would need to minimize ϵ. As discussed in the proof of Theorem 2, by minimizing ϵ, we obtain the average minimax redundancy. Therefore, we have

$$\mathbf{E} \log \left(\frac{1}{A_m(\gamma(X^n))} \right) \leq R_n^{2p} - R_n^{c2p}, \tag{105}$$

Note that the conditional two–stage codes achieve the average minimax redundancy of the regular codes, i.e., $R_n^{c2p} = R_n$. Thus,

$$R_n^{2p} - R_n^{c2p} = g(d) + O\left(\frac{1}{n} \right). \tag{106}$$

■

REFERENCES

[1] Atteson, K., "The asymptotic redundancy of bayes rules for markov chains," *IEEE Transactions on Information Theory*, vol. 45, pp. 2104 –2109, Sep 1999.

[2] Baron, D. and Bresler, Y., "An O(N) semipredictive universal encoder via the BWT," *IEEE Transactions on Information Theory*, vol. 50, pp. 928–937, May 2004.

[3] Baron, D., Bresler, Y., and Mihcak, M. K., "Two-part codes with low worst-case redundancies for distributed compression of bernoulli sequences," in *37th Annual Conference on Information Sciences and Systems (CISS '03)*, Mar 2003.

[4] Barron, A., Rissanen, J., and Yu, B., "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, pp. 2743 –2760, Oct 1998.

[5] Beirami, A. and Fekri, F., "On the finite-length performance of universal coding for k-ary memoryless sources," in *48th Annual Allerton Conference on Communication, Control, and Computing*, 2010.

[6] Clarke, B. and Barron, A., "Information-theoretic asymptotics of Bayes methods," *IEEE Transactions on Information Theory*, vol. 36, pp. 453 –471, May 1990.

[7] Cover, T. M. and Thomas, J. A., *Elements of Information Theory*. John Wiley and sons, 2006.

[8] Csiszar, I. and Talata, Z., "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, pp. 1007 –1016, Mar 2006.

[9] Davisson, L., "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. 19, pp. 783 – 795, Nov 1973.

[10] Davisson, L. and Leon-Garcia, A., "A source matching approach to finding minimax codes," *IEEE Transactions on Information Theory*, vol. 26, pp. 166 – 174, Mar 1980.

[11] Drmota, M., Hwang, H.-K., and Szpankowski, W., "Precise average redundancy of an idealized arithmetic coding," in *Data Compression Conference (DCC '02)*, pp. 222 – 231, 2002.

[12] Drmota, M. and Szpankowski, W., "Precise minimax redundancy and regret," *IEEE Transactions on Information Theory*, vol. 50, pp. 2686 – 2707, Nov 2004.

[13] Effros, M., Visweswariah, K., Kulkarni, S., and Verdu, S., "Universal lossless source coding with the Burrows Wheeler transform ," *IEEE Transactions on Information Theory*, vol. 48, pp. 1061–1081, May 2002.

[14] Feder, M. and Merhav, N., "Hierarchical universal coding," *IEEE Transactions on Information Theory*, vol. 42, pp. 1354 –1364, Sep 1996.

[15] Feder, M., Merhav, N., and Gutman, M., "Universal prediction of individual sequences," *IEEE Transactions on Information Theory*, vol. 38, pp. 1258 –1270, Jul 1992.

[16] Garcia-Frias, J. and Zhao, Y., "Compression of correlated binary sources using turbo codes," *IEEE Communications Letters*, vol. 5, pp. 417–419, Oct 2002.

[17] Grunwald, P. D., *The Minimum Description Length Principle*. The MIT Press, 2007.

[18] Jacquet, P. and Szpankowski, W., "Markov types and minimax redundancy for markov sources," *IEEE Transactions on Information Theory*, vol. 50, pp. 1393 – 1402, Jul 2004.

[19] Kontkanen, P., Myllymak, P., Buntine, W., Rissanen, J., and Tirri, H., "An MDL framework for data clustering," tech report, Helsinki Institute for Information Technology HIIT, 2004.

[20] Krichevsky, R. E. and Trofimov, V. K., "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, pp. 199–207, Mar 1981.

[21] Martin, A., Seroussi, G., and Weinberger, M., "Linear time universal coding and time reversal of tree sources via fsm closure," *IEEE Transactions on Information Theory*, vol. 50, pp. 1442 – 1468, Jul 2004.

[22] Merhav, N. and Feder, M., "The minimax redundancy is a lower bound for most sources," in *Data Compression Conference (DCC'94)*, pp. 52 –61, 29-31 1994.

[23] Pradhan, S. S. and Ramchandran, K., "Distributed source coding using syndromes (DISCUS): design and construction," in *Data Compression Conference (DCC '99)*, pp. 158–167, Mar 1999.

[24] Rissanen, J., "Complexity of strings in the class of Markov sources," *IEEE Transactions on Information Theory*, vol. 32, pp. 526–532, Jul 1986.

[25] Rissanen, J., "Stochastic complexity and modeling," *Annals of Statistics*, vol. 14, no. 3, pp. 1080–1100, 1986.

[26] Rissanen, J., "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Transactions on Information Theory*, vol. 47, pp. 1712 –1717, Jul 2001.

[27] Rissanen, J. and Langdon, G., J., "Universal modeling and coding," *IEEE Transactions on Information Theory*, vol. 27, pp. 12 – 23, Jan 1981.

[28] Rissanen, J., "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, pp. 40 –47, Jan 1996.

[29] Sartipi, M. and Fekri, F., "Lossy distributed source coding using LDPC codes," *IEEE Communications Letters*, vol. 13, Feb 2009.

[30] Shannon, C. E., "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, Jul, Oct 1948.

[31] Shields, P. C., "Universal redundancy rates do not exist," *IEEE Transactions on Information Theory*, vol. 39, pp. 520–524, Mar 1993.

[32] Shtarkov, Y., "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 175–186, 1987.

[33] Slepian, D. and Wolf, J. K., "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, pp. 471–480, Jul 1973.

[34] Szpankowski, W., "Asymptotic average redundancy of Huffman (and other) block codes ," *IEEE Transactions on Information Theory*, vol. 46, pp. 2434–2443, Nov 2000.

[35] Szpankowski, W., " Average Redundancy for Known Sources: Ubiquitous Trees in Source Coding ," *DMTCS Proceedings*, vol. 0, no. 1, 2008.

[36] Weinberger, M., Merhav, N., and Feder, M., "Optimal sequential probability assignment for individual sequences," *IEEE Transactions on Information Theory*, vol. 40, pp. 384–396, Mar 1994.

[37] Weinberger, M., Rissanen, J., and Feder, M., "A universal finite memory source," *IEEE Transactions on Information Theory*, vol. 41, pp. 643 –652, May 1995.

[38] Willems, F., "The context-tree weighting method: extensions," *IEEE Transactions on Information Theory*, vol. 44, pp. 792–798, Mar 1998.

[39] Willems, F., Shtarkov, Y., and Tjalkens, T., "The context-tree weighting method: basic properties," *IEEE Transactions on Information Theory*, vol. 41, pp. 653–664, May 1995.

[40] Xie, Q. and Barron, A., "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory*, vol. 43, pp. 646 –657, Mar 1997.

[41] Xie, Q. and Barron, A., "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, pp. 431 –445, Mar 2000.

[42] Ziv, J. and Lempel, A., "A universal algorithm for sequential data compression," *IEEE Transactions on Information Theory*, vol. 23, pp. 337–343, May 1977.