Forty-Eighth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 29 - October 1, 2010

On the Finite-Length Performance of Universal Coding for k-ary Memoryless Sources

Ahmad Beirami and Faramarz Fekri
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta GA 30332, USA
Email: {beirami, fekri}@ece.gatech.edu

***Abstract*— In this paper, we investigate the performance of universal coding schemes on *finite*-length memoryless sequences. Rissanen demonstrated that for the universal compression of k-ary memoryless sources, expected redundancy for regular codes is asymptotically lower bounded by $\frac{k-1}{2}\log n$ for almost all sources. Xie and Barron derived the minimax expected redundancy for k-ary memoryless sources, which characterizes the maximum redundancy over all possible source parameters. It does not provide much information about different source parameter values. This paper is a finite-length extension to Rissanen's result. Our treatment in this paper is probabilistic. In particular, we derive a lower bound on the probability measure of the sources that are not compressible with a redundancy smaller than a certain fraction of $\frac{k-1}{2}\log n$. In other words, we demonstrate a lower bound on the redundancy for a given percentile of sources. We demonstrate that as the length of the memoryless sequence decreases, the redundancy tends to become significant and comparable to the entropy of the sequence.**

I. Introduction

Recently, the amount of data that is being stored in storage systems has been increasing with a very high rate. Hence, the development of compression for storage systems has gained a lot of interest. In many cases, one complete database may be compressed to less than one tenth of its original size. The redundancy in the data may be leveraged to significantly reduce the cost of data maintenance as well as data transmission. However, most applications require that individual files be retrieved and updated separate from the rest of the database. Therefore, the compressed data for individual files need to be independently retrievable and updateable. On the other hand, the individual files are relatively very small. Moreover, the files may come from various natures, which calls for universal compression. It is well known that if a small file is compressed alone, existing universal compression techniques are unable to achieve good performance. Rissanen's result describes asymptotic achievable compression rates for universal compression of individual files [1], [2]. However, it does not provide much insight on the performance of universal coding for a small given file. The minimax redundancy [3] is concerned with the maximum redundancy over all parameters. However, it does not characterize the redundancy for other parameters.

Since Shannon's work on the analysis of communication systems [4], many researchers have contributed towards the development of source coding schemes with the compression rate as close as possible to the entropy rate of the information source. It is well known that using a prefix-free code, a stationary ergodic information source cannot be compressed at a rate lower than the entropy rate of the source [5]. Thus, the goal of source coding is to achieve a compression rate that approaches the entropy rate.

In this paper, we focus on the universal compression of memoryless sources with a finite-alphabet size. Let $\mathcal{S}$ denote a universal memoryless information source with k-ary alphabet $\alpha = \{\alpha_1, ..., \alpha_k\}$. Further, denote $\theta = (\theta_1, ..., \theta_k) \in Theta$ as the vector in the $(k-1)$-dimensional simplex of source parameters such that for a symbol X generated by $\mathcal{S}$, $\mathbf{P}[X = \alpha_i] = \theta_i$, for $1 \leq i \leq k$. Let $h(\theta)$ be the entropy rate of $\mathcal{S}$ given parameters θ, i.e., $h(\theta) = -\sum_j \theta_j \log \theta_j$. Finally, we use the notation $X^n = (X_1, ..., X_n)$ to present a vector of length n of iid random variables generated by $\mathcal{S}$.

Let $l(X^n) \in L$ denote the *regular* length function that describes the codeword associated with the sequence X^n. Denote $R_n(l, \theta)$ as the expected redundancy of the code with length function $l(.)$ and parameter vector θ on a sequence of length n, defined as the difference between the expected codeword length and the entropy of the sequence X^n:

$$R_n(l, \theta) = \mathbf{E}[l(X^n)] - nh(\theta). \tag{1}$$

For an asymptotically optimal code with length function $l(X^n)$, $\frac{R_n(l,\theta)}{n} \to 0$ as $n \to \infty$ for all θ.

Provided that the statistics of the information source are *known*, Huffman block coding achieves the entropy rate with a redundancy smaller than 1 bit per source symbol [5]. 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.

The maximum average redundancy for a length function of a code with length function l is given as $R_n(l) = \max_{\theta\in\Theta} R_n(l, \theta)$, which is lower bounded by the minimax average redundancy $R_n = \min_{l\in\mathcal{L}} \max_{\theta\in\Theta} R_n(l, \theta)$ [3]. The leading term of the average minimax redundancy is asymptotically $\frac{k-1}{2}\log n$. According to Rissanen's results, for the universal compression of memoryless sources with

uniformly distributed parameter vector θ, the redundancy of *regular* codes is asymptotically lower bounded by $R_n(l,\theta) \geq (1-\delta)\frac{k-1}{2}\log n$ [2], for all $\epsilon > 0$ and almost all sources. These results were later extended in [6], [7] to more general classes of sources. The asymptotic lower bound is tight since there exist coding schemes that achieve the bound asymptotically [2], [8]. We will formally state Rissanen's result in Sec. II.

In this paper, we study the redundancy for the universal compression in *finite*-length regime. As the first step, we consider universal coding for k-ary memoryless sources. The work can be viewed as the extension of Rissanen's probabilistic treatment to the finite-length sequences. The rest of this paper is organized as follows. In Section II, we formally state the finite-length universal compression problem followed by our main result. In Section III, we present the proof of the main result. In Section IV, we demonstrate the significance of the main result through two main examples. Finally, the conclusion is given in Section V.

II. Problem Statement and Main Results

In this section, we formally state the finite-length redundancy problem and present our main result and its implications. We focus on two-part codes, where the compression scheme utilizes m bits for the identification of an estimate for the unknown source parameters. The estimate parameter is then used for the compression of the sequence. It has been demonstrated that the two-part assumption brings about a small $O(1)$ redundancy term [9]. This corresponds to 2^m possible estimate points in the parameter space for the unknown parameter. In other words, the compression scheme chooses the best among 2^m estimate points in the parameter space as it compresses the input sequence. Let $\Phi = \{\phi_1, ..., \phi_{2^m}\}$ denote the set of all estimate points. Note that each ϕ_i is a point in the $(k-1)$-dimensional simplex of θ. For each sequence X^n, there is an estimate point $\beta = \beta(X^n) \in \Phi$ that minimizes the redundancy.

Let r_i denote the number of times symbol α_i appears in the sequence X^n. Let f_i denote the relative frequency of symbol α_i, i.e., $f_i = r_i/n$. Let μ_θ denote the probability measure defined over a memoryless source with parameter vector θ as

$$\mu_\theta(X^n) = \mathbf{P}[X^n|\theta] = \theta_1^{r_1}...\theta_k^{r_k}. \tag{2}$$

We require the length function $l_\theta(X^n)$ be *regular*, i.e.,

$$l_\theta(X^n) \geq \log\left(\frac{1}{\mu_\theta(X^n)}\right) \quad \forall X^n, \ ^1 \tag{3}$$

where $\mu_\theta(X^n)$ is the memoryless probability measure defined by the parameter vector θ. Note that the requirement (3) is not restrictive since all codes that we know are regular [2]. Let $l_\beta(X^n)$ denote the length function induced by $\beta \in \Phi$. Further denote $\mu_\beta(X^n)$ as the probability measure induced by β:

$$\mu_\beta(X^n) = {\beta_1}^{r_1}...{\beta_k}^{r_k}. \tag{4}$$

[1] In this paper $\log(x)$ always denotes the logarithm of x in base 2.

Increasing m results in an exponential growth in the number of estimate points and more accurate estimate for the unknown source parameter vector hence better compression. On the other hand, increasing m directly increases the compression overhead. Therefore, it is desirable to find the best m that minimizes the total codeword length as

$$l(X^n) = \min_m \{m + l_\beta(X^n)\}. \tag{5}$$

Since we assumed the code is regular, we may use (3) to bound the redundancy rate

$$R_n(\theta) \geq \min_m \left\{ m + \mathbf{E}\log\frac{1}{\mu_\beta(X^n)} \right\} - nh(\theta). \tag{6}$$

Using (4), we get

$$R_n(\theta) \geq \min_m \left\{ m + n\mathbf{E}\sum_{i=1}^{k} f_i \log\frac{1}{\beta_i} \right\} - nh(\theta). \tag{7}$$

Our goal is to better characterize the lower bound in (7). In [2], Rissanen proved an asymptotic lower bound on the universal compression of parametric sources that can be represented with k parameters. The following asymptotic lower bound on the redundancy rate of universal coding of k-ary information sources is a direct consequence of Rissanen's result:

Theorem 1 *Let $\mathcal{S}$ denote a k-ary memoryless information source with parameter vector θ. Let X^n denote a sequence of length n produced by $\mathcal{S}$. Let $l(X^n)$ denote any regular length function for universal compression. Then, for all parameters θ, except in a set of asymptotically Lebesgue volume zero, we have*

$$\lim_{n\to\infty} \frac{R_n(\theta)}{\frac{k-1}{2}\log n} \geq 1-\epsilon, \quad \forall \epsilon > 0, \tag{8}$$

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

In [3], Xie and Barron derive the expected minimax redundancy R_n for memoryless sources, where they demonstrate that

$$R_n = \frac{k-1}{2}\log\left(\frac{n}{2\pi}\right) + \log\left(\frac{\Gamma(1/2)^k}{\Gamma(k/2)}\right) + o(1), \tag{9}$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the gamma function. The minimax redundancy characterizes the maximum redundancy on the space of all possible parameter vectors but does not imply much about the rest of the space of the parameter vectors.

In this paper, we derive a lower bound on the probability that a source with parameter vector θ is compressed with redundancy rate $R_n(\theta) > R_0$ for any $R_0 > 0$. In other words, we find a lower bound on $\mathbf{P}[R_n(\theta) > R_0]$. Using this

result, we demonstrate the fundamental limits of the universal compression for finite-length n. The following is our main result:

Theorem 2 *Let $\mathcal{S}$ denote a k-ary memoryless information source with parameter vector θ that follows the Dirichlet distribution. Let X^n denote a sequence of length n produced by $\mathcal{S}$. Let $l(X^n)$ denote any regular length function. Let ϵ be a real number such that $0 < \epsilon < 1$. Then, the probability that $R_n(\theta)$ is greater than $(1-\epsilon)$ times the asymptotic bound is lower bounded as follows:*

$$\mathbf{P}\left[\frac{R_n(\theta)}{\frac{k-1}{2}\log n} \geq 1-\epsilon\right] \geq 1 - B_k\left(\frac{k-1}{e n^{\epsilon}}\right)^{\frac{k-1}{2}}, \quad (10)$$

where $B_k = \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)}\sqrt{\frac{1}{\pi}}$. Note that $B_k \approx \sqrt{\frac{2}{k\pi}}$ for $k \gg 2$.

Proof: See Section III. ∎

Note that it is straightforward to deduce Theorem 1 from Theorem 2 by letting $n \to \infty$. Note that for any ϵ such that $1 - B_k\left(\frac{k-1}{e n^{\epsilon}}\right)^{\frac{k-1}{2}} > 0$, $R_\epsilon = (1-\epsilon)\frac{k-1}{2}\log n$ is a lower bound on the minimax redundancy since there exists a source that has a redundancy of at least R_ϵ. We will later demonstrate that this lower bound is actually tight, and gives back the minimax redundancy.

III. Proof of the Main Result

In this section, we present the proof of Theorem 2. We break down the main proof to a few intermediate lemmas that constitute the main proof. First, we simplify (7) as

Lemma 1 *The redundancy rate $R_n(\theta)$ is lower bounded by*

$$R_n(\theta) \geq \min_m \left\{ m + n\mathbf{E}\sum_{i=1}^{k} f_i \log\frac{\theta_i}{\beta_i(X^n)} \right\}.$$

Proof: This is straightforward since $h(\theta) = \mathbf{E}\sum_{i=1}^{k} f_i \log\frac{1}{\theta_i}$. ∎

In order to bound the redundancy in Lemma 1, in the following, we bound $\mathbf{E}\sum_{i=1}^{k} f_i \log\frac{\theta_i}{\beta_i(X^n)}$. Note that this term implicitly depends on m since β is a function of m.

Lemma 2 *Assume that $\exists \beta \in \Phi$ such that $1 \leq \forall i \leq 2^m$; $D(\theta||\beta) \leq D(\theta||\phi_i)$. Then, we have*

$$\mathbf{E}\sum_{i=1}^{k} f_i \log\frac{\theta_i}{\beta_i(X^n)} \geq D(\theta||\beta), \quad (11)$$

where

$$D(x||y) = \sum_{j=1}^{k} x_j \log\left(\frac{x_j}{y_j}\right). \quad (12)$$

Proof: See Appendix I ∎

Note that $D(\theta||\beta)$ is the non-negative Kullback–Leibler divergence between the probability measures defined by θ and β. We use a probabilistic treatment in order to bound $D(\theta||\beta)$ for a certain percentage of source parameters. We assume that the parameter vector θ follows the Dirichlet distribution. The Dirichlet prior distribution is particularly interesting since it results in uniform redundancy over the parameter vector space, which results in the achievement of the minimax expected redundancy [3], [10]. The Dirichlet probability distribution for the parameter vector θ is given by:

$$p(\theta) = \frac{\Gamma(k/2)}{\Gamma(1/2)^k}\prod_{j=1}^{k}\frac{1}{\sqrt{\theta_j}}, \quad (13)$$

where $\Gamma(.)$ is the gamma function defined in Theorem 2.

In order to bound the redundancy rate $R_n(\theta)$, in the following, we find an upper bound on the Lebegue measure of the volume defined by $D(\theta||\beta) < \delta$ in the $(k-1)$-dimensional simplex of θ. Since $\beta \in \Phi$, the total measure of the volume defined by $\exists\phi_i \in \Phi;\ D(\theta||\phi_i) < \delta$ may be upper bounded as well. This provides with a lower bound on the measure of the sources with $R_n(\theta) \geq \delta$.

Lemma 3 *Let $\beta = (\beta_1, ..., \beta_k)$ and $\theta = (\theta_1, ..., \theta_k)$ be a fixed and a variable point in the $(k-1)$-dimensional simplex of θ, respectively. If θ follows the Dirichlet distribution, the probability $\mathbf{P}[D(\theta||\beta) < \delta]$ is upper bounded by*

$$\mathbf{P}[D(\theta||\beta) < \delta] \leq \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)}\sqrt{\frac{1}{\pi}}\left(\frac{2\delta}{\log e}\right)^{\frac{k-1}{2}}. \quad (14)$$

Moreover, $\mathbf{P}[\exists\phi_i \in \Phi;\ D(\theta||\phi_i) < \delta]$ is upper bounded by

$$\mathbf{P}[\exists\phi_i \in \Phi;\ D(\theta||\phi_i) < \delta] \leq 2^m \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)}\sqrt{\frac{1}{\pi}}\left(\frac{2\delta}{\log e}\right)^{\frac{k-1}{2}}. \quad (15)$$

Proof: See Appendix II. ∎

Lemma 3 states that the probability measure that is covered $\mathbf{P}[D(\theta||\beta) < \delta]$ does not depend on β when θ follows the Dirichlet prior. In other words, the choice of the set of the parameter points Φ does not affect the performance of the compression.

We are now equipped to prove the main result given in Theorem 2.

Proof: **[Proof of Theorem 2]** Lemma 3 bounds the probability measure of the parameter vectors that may be compressed with a small redundancy. The condition in Lemma 2 is satisfied if $m < \frac{k-1}{2}\log n$ for points close to some $\phi \in \Phi$. This is because $2^m = o(n^{\frac{k-1}{2}})$ and thus the distance between the points is $\omega(1/\sqrt{n})$. Using Lemmas 1 and 2,

$$R_n(\theta) \geq \min_m \min_i \left\{ m + nD(\theta||\phi_i) \right\}. \quad (16)$$

Therefore,

$$\mathbf{P}\left[\frac{R_n(\theta)}{\frac{k-1}{2}\log n} \leq 1-\epsilon\right] \quad (17)$$

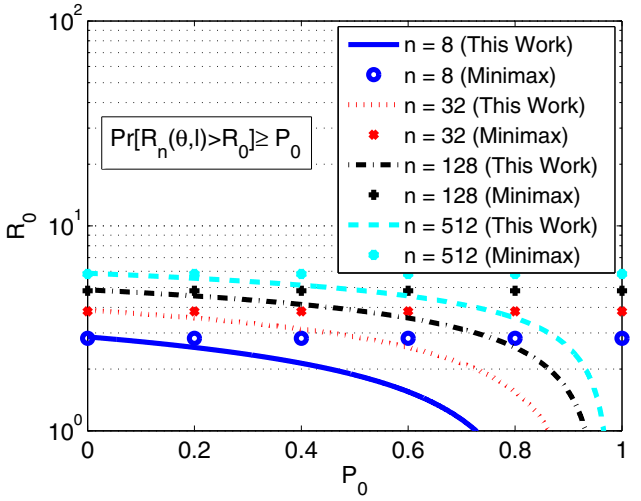

Fig. 1. Redundancy level R_0 as a function of percentile of sources P_0 with $R_n(l,\theta) > R_0$. Alphabet size: $k = 2$.

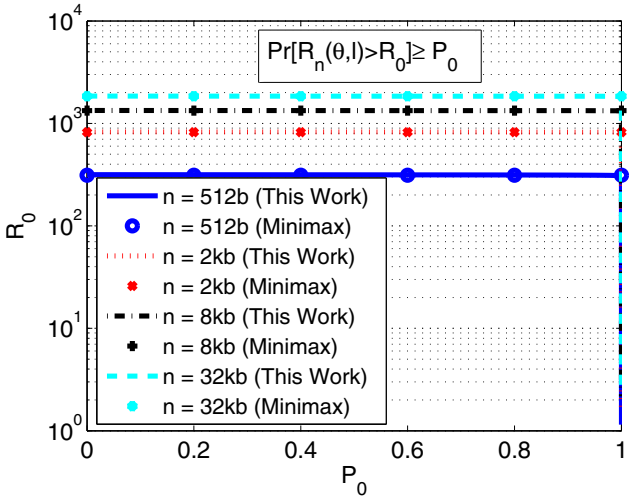

Fig. 2. Redundancy level R_0 as a function of percentile of the sources P_0 with $R_n(l,\theta) < R_0$. Alphabet size: $k = 256$.

$$\leq \mathbf{P}\left[\min_m \min_i \{m + nD(\theta||\phi_i)\} \leq (1-\epsilon)\frac{k-1}{2}\log n\right] \quad (18)$$

$$= \min_m \min_i \mathbf{P}\left[D(\theta||\phi_i) \leq (1-\epsilon)\frac{k-1}{2n}\log n - \frac{m}{n}\right] \quad (19)$$

$$\leq \min_m \left\{2^m \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)}\sqrt{\frac{k}{\pi}}\left(\frac{2\delta(m)}{\log e}\right)^{\frac{k-1}{2}}\right\}. \quad (20)$$

The last inequality is obtained using Lemma 3. Here, $\delta(m)$ is given by

$$\delta(m) = (1-\epsilon)\frac{k-1}{2n}\log n - \frac{m}{n}. \quad (21)$$

Carrying out the minimization in (20) leads to the desired result in Theorem 2. ■

IV. Elaboration of the Results and Examples

In this section, we illustrate the results through two examples. In Figures 1 and 2, the x-axis is a percentile P_0 and the y-axis represents a redundancy level R_0. The solid curves demonstrate the derived lower bound on the redundancy as a function of the percentile of the sources that have redundancy beyond the specified level based on Theorem 2, i.e., we have $\mathbf{P}[R_n(\theta) \geq R_0] \geq P_0$. In other words, at least a fraction P_0 of the sources that are chosen from the Dirichlet prior have an expected redundancy that is greater than R_0.

A. Minimax Redundancy of Two-Part Codes: Lower Bound

Note that the solid curves are a lower bound on the minimax redundancy for all values of $P_0 > 0$. This is due to the fact that a non-zero percentage of the sources may not be compressed beyond the level R_0 hence the worst case is included in this region. Therefore, we may use our result to derive a lower bound on the minimax redundancy of two-part codes when $P_0 \to 0$ as a side product:

$$R_{n,2p} \geq \frac{k-1}{2}\log n - \log B_k - \frac{k-1}{2}\log(k-1) + \frac{k-1}{2}\log e, \quad (22)$$

where B_k is defined in Theorem 2. This may be rewritten as follows:

$$R_{n,2p} \geq R_n + \log\Gamma\left(\frac{k+1}{2}\right) - \frac{k-1}{2}\log\left(\frac{k-1}{2e}\right). \quad (23)$$

Note that it is straightforward to demonstrate that the lower bound $R_{n,2p} \to R_n$ as $k \to \infty$, i.e., the effect of two-part assumption will be negligible.

B. Example 1: $k = 2$

First, for a Bernoulli information source, i.e., $k = 2$, we demonstrate the fundamental limits for the compression of finite-length sequences. The plots are given for various n. The markers in this Figure labeled as *Minimax* demonstrate the minimax redundancy for a two-part code. For a Bernoulli source, the minimax redundancy of the two-part code is given by [9]:

$$R_{n,2p} = R_n + \frac{1}{2}\log\left(\frac{\pi e}{2}\right) \approx R_n + 1.048, \quad (24)$$

where R_n is the minimax redundancy and is defined in (9).

As shown in Figure 1, at least 40% of Bernoulli sequences of length $n = 32$ ($n = 128$) may not be compressed beyond a redundancy of 3.5 (4.5) bits per symbol. Also, at least 60% of Bernoulli sequences of length $n = 32$ ($n = 128$) may not be compressed beyond a redundancy of 2.5 (3.5) bits per symbol.

Note that as $n \to \infty$, the redundancy approaches the minimax redundancy for more sources, and asymptotically redundancy approaches the minimax redundancy for all sources as given in Theorem 1. Also, using (22) we get the following lower bound on the minimax redundancy of two-part codes for $k = 2$:

$$R_{n,2p} \geq \frac{1}{2}\log n + \frac{1}{2}\log\left(\frac{\pi e}{2}\right), \quad (25)$$

which is indeed equal to the minimax redundancy.

C. Example 2: $k = 256$

Next, we consider $k = 256$, which is a common practice to use the byte as a symbol. In Figure 2, the achievable redundancy is demonstrated for four different values of n. Here, the redundancy is measured in bits per source symbol (byte). We observe that for $n = 512b$, we have $R_n(l, \theta) \geq 300$ bits/symbol for almost all sources. This is considered significant specially for sources with small entropy rate. For example, if the entropy rate is around 1 bit/symbol, there will be almost always more than 50% redundancy in the compression of a sequence of length $n = 512b$. We also observe that the two-part assumption does not incur further redundancy for large k.

V. Conclusion

In this paper, we investigated the redundancy rate of universal coding schemes on memoryless input sequences in the *finite*-length regime. We derived a lower bound on the probability measure of information sources that may not be compressed beyond any certain redundancy level. Our result may be viewed as the finite-length extension of the previous asymptotic results. Our result may be used to evaluate the performance of universal source coding on finite-length sequences. We observe that redundancy may be very significant in the compression of finite-length low-entropy information sources.

Appendix I
Proof of Lemma 2

Proof:

$$\mathbf{E}\sum_{i=1}^{k} f_i \log\left(\frac{\theta_i}{\beta_i(X^n)}\right) \tag{26}$$

$$= \sum_{\substack{r_1,\ldots,r_k \\ \sum_j r_j = n}} \frac{n!}{r_1!\ldots r_k!}\theta_1^{r_1}\ldots\theta_k^{r_k}\sum_i \frac{r_i}{n}\log\left(\frac{\theta_i}{\beta_i(X^n)}\right), \tag{27}$$

$$= \sum_i \sum_{\substack{s_1,\ldots,s_k \\ \sum_j s_j = n-1}} \frac{(n-1)!}{s_1!\ldots s_k!}\theta_1^{s_1}\ldots\theta_k^{s_k}\theta_i \log\left(\frac{\theta_i}{\beta_i(X^n)}\right), \tag{28}$$

where $s_i = r_i - 1$ and $s_j = r_j,\ j \neq i$. Note that this could now be viewed as taking expectations over a different set of variables. Thus, (28) could be rewritten as

$$\mathbf{E}\sum_{i=1}^{k} \theta_i \log\left(\frac{\theta_i}{\beta_i(X^n)}\right). \tag{29}$$

This may be lower bounded using the fact $\forall \phi_i \in \Phi$, $D(\theta||\phi_i) \geq D(\theta||\beta)$, which proves the claim. ∎

Appendix II
Proof of Lemma 3

Proof: Let $f(\theta) = D(\theta||\beta) = \sum_{i=1}^{k} \theta_i \log\left(\frac{\theta_i}{\beta_i}\right)$. Then, using Taylor expansion, we get

$$D(\theta||\beta) \geq L(\theta, \beta) + O(\theta_i - \beta_i)^3, \tag{30}$$

where

$$L(\theta, \beta) = \frac{\log e}{2}\sum_{i=1}^{k}\frac{1}{\beta_i}(\theta_i - \beta_i)^2.$$

Note that $L(\theta, \beta) \leq \delta$, where $\delta > 0$, defines an ellipsoid on the $(k-1)$-dimensional simplex of θ. It is straightforward to demonstrate that the volume of the ellipsoid is given by

$$V_k(\beta) = C_{k-1}\left(\frac{2\delta}{\log e}\right)^{\frac{k-1}{2}}\prod_{i=1}^{k}\sqrt{\beta_i}, \tag{31}$$

where $C_{k-1} = \frac{\Gamma(1/2)^{k-1}}{\Gamma((k+1)/2)}$ is the volume of the $(k-1)$-dimensional unit ball. Moreover, since θ follows the Dirichlet distribution, the probability measure of the covered ellipsoid is given by

$$\mathbf{P}[L(\theta, \beta) \leq \delta] = V_k(\beta)\left(\frac{\Gamma(k/2)}{\Gamma(1/2)^k}\prod_{j=1}^{k}\frac{1}{\sqrt{\beta_i}}\right)$$

$$= \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)}\sqrt{\frac{1}{\pi}}\left(\frac{2\delta}{\log e}\right)^{\frac{k-1}{2}}. \tag{32}$$

Since $D(\theta||\beta) \approx L(\theta, \beta)$, the volume defined by $D(\theta||\beta) < \delta$ is equal to the volume $L(\theta, \beta) < \delta$, which completes the proof of the first claim. Although the volume of the ellipsoid depends on the point β in the parameter space, the volume of the ellipsoid defined by $D(\theta||\beta) < \delta$ does not depend on β.

For the second claim, there are 2^m choices for ϕ_i. $\forall i$, $D(\theta||\phi_i)$ defines a measure that has the same upper bound. Thus, the second claim follows directly from the first claim using the union bound. ∎

References

[1] J. Rissanen, "Universal coding, information, prediction, and estimation," *Information Theory, IEEE Transactions on*, vol. 30, no. 4, pp. 629–636, Jul 1984.

[2] ——, "Complexity of strings in the class of Markov sources," *Information Theory, IEEE Transactions on*, vol. 32, no. 4, pp. 526–532, Jul 1986.

[3] Q. Xie and A. Barron, "Minimax redundancy for the class of memoryless sources," *Information Theory, IEEE Transactions on*, vol. 43, no. 2, pp. 646 –657, mar 1997.

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

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

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

[7] M. Feder and N. Merhav, "Hierarchical universal coding," *Information Theory, IEEE Transactions on*, vol. 42, no. 5, pp. 1354 –1364, sep 1996.

[8] F. Willems, Y. Shtarkov, and T. Tjalkens, "The context-tree weighting method: basic properties," *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 653–664, May 1995.

[9] P. D. Grunwald, *The minimum description length principle*. The MIT Press, 2007.

[10] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *Information Theory, IEEE Transactions on*, vol. 27, no. 2, pp. 199–207, March 1981.