Fundamental Limits of Universal Lossless One-to-One Compression of Parametric Sources

Ahmad Beirami
Department of Electrical and Computer Engineering
Duke University, Durham, NC USA
Email: ahmad.beirami@duke.edu

Faramarz Fekri
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA USA
Email: fekri@ece.gatech.edu

***Abstract*—In this paper, the problem of universal lossless one-to-one compression (without prefix constraint) is studied. A converse bound is obtained on the average minimax (and maximin) redundancy that shows the redundancy is at least $(d-2)/2\log n+O(1)$ for the universal compression of a sequence of length n from a d-dimensional parametric source. Further, the type-size coding strategy is shown to be minimax optimal up to $o(\log n)$ for the class of memoryless sources, achieving the converse leading to characterization of the fundamental performance limit of universal compression for memoryless sources. Finally, through a numerical example, our results imply that the reduction on the codeword length due to relaxing the prefix constraint is negligible when compared to the cost of universality.**

I. Introduction

Staggering amount of data is churned daily around the world. This massive amount of data results in very high transmission costs, which accounts for a large fraction of the costs associated with dealing with this data. Ever since entropy rate was shown to be the lower bound on the average compression rate of any stationary source using prefix-free codes, many researchers have contributed toward the development of prefix-free codes with average codeword length approaching the entropy of the sequence. When the statistics of the information source are *known*, Huffman block coding achieves the entropy of a sequence with a negligible redundancy smaller than 1 bit on top of the entropy, which is due to the integer length constraint on the codewords [1]. In many applications, however, the sequence to be compressed does not follow a fixed known distribution requiring the compression to be universal [2]–[8].

Thus far, most of the literature on the universal compression has considered prefix-free codes (uniquely decodable codes) where the length function is required to satisfy Kraft's inequality. The prefix constraint (also known as unique decodability constraint) ensures that several blocks of data are encoded into a *stream* bit stream which can be uniquely decoded. For the fairly general class of parametric sources the average redundancy of prefix-free codes has been exactly characterized to be $d/2\log n+O(1)$ [9]–[13]. In [13], we further showed that the redundancy is a significant overhead on top of the entropy when the prefix-free universal compression of small sequences is concerned.

On the other hand, there are several applications that do not require the unique decodability of the concatenated blocks since the beginning and the end of each block is already determined via an external mechanism. For example, in the compression of network packets, the end of each IP packet is already determined by the header. In these cases, the unique decodability condition can be relaxed to the mapping (the code) to be injective so as to ensure that *one* block of length n can be uniquely decoded. These codes are known as one-to-one codes. While the average codeword length of prefix-free codes can never be smaller than the entropy, the average codeword length of one-to-one codes can go below the entropy (cf. [14]–[18] and the references therein).

The performance of one-to-one codes has been investigated for the case of known source parameter vectors, however, for the aforementioned reasons, our interest lies in the performance of *universal* one-to-one codes which is relatively unexplored. Universal one-to-one codes are only developed very recently for memoryless sources by Kosut and Sankar [19] as type-size codes, where it was shown that the average redundancy of type-size code scales as $\frac{d-2}{2}\log n+O(1)$. Although this result beats the fundamental limit of universal prefix-free codes, which is $d/2\log n+O(1)$, the redundancy is still bounded away from the entropy-rate rising to the question whether or not better universal one-to-one codes can be constructed. In [20], Kosut and Sankar proved that the type-size code is minimax optimal up to $o(\log n)$.

In this paper, we define the minimax and maximin games in the universal compression of parametric sources. We prove a converse bound on the average minimax redundancy for parametric sources, which gives back Kosut and Sankar's converse for memoryless sources as a special case. We also show that the reduction in the compression cost when prefix constraint is dropped is negligible when compared with the cost of universality in compression. Finally, we further show that the type-size code is minimax optimal up to $o(\log n)$ for the compression of memoryless sources.

The rest of this paper is organized as follows. In Section II, we review the related background on universal prefix-free codes. In Section III, one-to-one codes are introduced. In Section IV, we provide our main results on universal one-to-one codes. In Section V, we demonstrate the significance of the results through a numerical example. Finally, Section VI

978-1-4799-5999-0/14/$31.00 ©2014 IEEE

concludes this paper.

II. Background on Universal Lossless Compression

In the following, we describe our source model together with necessary notations and related work. Let $\mathcal{A}$ denote a finite alphabet of size $|\mathcal{A}|$. Let the parametric source be defined using a d-dimensional parameter vector $\theta = (\theta_1, ..., \theta_d)$, where d denotes the number of the source parameters. Denote μ_θ as the probability measure defined by the parameter vector θ on sequences of length n. We also use the notation μ_θ to refer to the parametric source itself. We assume that the d parameters are unknown and lie in the d-dimensional space $\Lambda \subset \mathbb{R}^d$. Denote $\mathcal{P}_\Lambda^d$ as the *family* of parametric sources with d-dimensional unknown parameter vector θ such that $\theta \in \Lambda$. The family $\mathcal{P}_\Lambda^d$ contains all source models that have a *minimal* representation with a d-dimensional parameter vector θ. We use the notation $x^n = (x_1, ..., x_n) \in \mathcal{A}^n$ to represent a sequence of length n (which is assumed to be a realization of the random vector X^n that follows μ_θ unless otherwise stated).

In this paper, we consider the class of lossless fixed-to-variable codes. A lossless (also called zero-error) code is defined by an injective mapping $c_n : \mathcal{A}^n \to \{0,1\}^*$, i.e., there exists a reverse mapping $d_n : \{0,1\}^* \to \mathcal{A}^n$ such that $\forall x^n \in \mathcal{A}^n$, we have $d_n(c_n(x^n)) = x^n$. Further, let $l_n : \mathcal{A}^n \to \mathbb{Z}^+$ denote the universal lossless length function associated with the for the codeword $c_n(x^n)$ associated with the sequence x^n. The goal of compression is to minimize the expected codeword length, which is $\mathbf{E}l_n(x^n)$

Let $H_n(\theta)$ be the source entropy given parameter vector θ, i.e.,

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

In this paper $\log(\cdot)$ always denotes the logarithm in base 2.

If a code is prefix-free, then the corresponding length function must satisfy the well-known Kraft's inequality, i.e.,

$$\sum_{x^n \in \mathcal{A}^n} 2^{-l_n(x^n)} \leq 1. \quad (2)$$

Kraft's inequality ensures that when several blocks of length n are encoded using c_n there exists a uniquely decodable code such that all the blocks can be decoded. For the prefix-free codes, when the source parameter vector is known, the optimal code length is equal to the log-likelihood as given by

$$l_n(x^n) = \log\left(\frac{1}{\mu_\theta(x^n)}\right).^{2} \quad (3)$$

and the associated average codeword length is equal to the entropy, and hence, we conclude that the entropy is a lower limit on the average codeword length, i.e.,

$$\mathbf{E}l_n(X^n) \geq H_n(\theta). \quad (4)$$

[1]Throughout this paper all expectations are taken with respect to the distribution μ_θ induced by the true unknown parameter vector θ.

[2]Please note that the integer constraint on the codeword length is ignored, which results in a negligible redundancy upper bounded by 1 bit analyzed exactly in [1], [21].

On the other hand, when the parameter vector θ is unknown (and hence $\mu_\theta(x^n)$ is also unknown), the *universal* length function can no longer be a function of θ. Denote $R_n(l_n, \theta)$ as the expected redundancy of the code c_n with length function l_n on a sequence of length n for the parameter vector θ, defined as

$$R_n(l_n, \theta) = \mathbf{E}l_n(X^n) - H_n(\theta). \quad (5)$$

Note that the expected (average) redundancy is non-negative. Further, a code is called universal if its average codeword length normalized to the sequence length uniformly converges to the source entropy rate, i.e., $\lim_{n\to\infty} \frac{1}{n} R_n(l_n, \theta) = 0$ for all $\theta \in \Lambda$.

Let $\mathcal{I}(\theta)$ be the Fisher information matrix associated with the parameter vector θ, i.e.,

$$\mathcal{I}(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n \log e} \mathbf{E}\left\{\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\left(\frac{1}{\mu_\theta(X^n)}\right)\right\}. \quad (6)$$

Please note that we assume that the source is ergodic such that the above limit converges. Let Jeffreys' prior on the parameter vector θ be denoted by

$$p_J(\theta) \triangleq \frac{|\mathcal{I}(\theta)|^{\frac{1}{2}}}{\int |\mathcal{I}(\lambda)|^{\frac{1}{2}} d\lambda}. \quad (7)$$

Jeffreys' prior is optimal in the sense that the average minimax redundancy is asymptotically achieved when the parameter vector θ is assumed to follow Jeffreys' prior [22]–[24]. Jeffreys' prior is particularly interesting because it also corresponds to the worst-case prior for the best compression scheme (called the capacity achieving prior), which is the average maximin redundancy [24]. Define $\bar{R}_n$ as the average minimax redundancy, i.e.,

$$\bar{R}_n = \min_{l_n} \sup_{\theta} R_n(l_n, \theta). \quad (8)$$

Gallager proved that the average minimax redundancy is equal to the average maximin redundancy [24], which is defined as

$$\underline{R}_n = \sup_{\omega} \min_{l_n} \int_{\theta\in\Lambda} R_n(l_n, \theta)\omega(\theta) d\theta. \quad (9)$$

The average maximin redundancy is associated with the best code under the worst prior on the space of parameter vectors (i.e., the capacity achieving Jeffreys' prior).

The average minimax (maximin) redundancy is well studied for parametric sources given by the following theorem.

Theorem 1 ([12], [22], [23]) *The average minimax redundancy is given by*

$$\bar{R}_n = \frac{d}{2}\log\left(\frac{n}{2\pi e}\right) + \log\int_{\theta\in\Lambda} |\mathcal{I}_n(\theta)|^{\frac{1}{2}} d\theta + O\left(\frac{1}{\sqrt{n}}\right).^{3} \quad (10)$$

Remark. According to Theorem 1, the average minimax redundancy scales as $\frac{d}{2}\log n + O(1)$. This redundancy may indeed be a significant overhead on top of the entropy for small sequences, as the constant term in (10) could be relatively large for small n (cf. [13]).

[3]$f(n) = O(g(n))$ if and only if $\limsup_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| < \infty$.

III. One-to-One Codes

Next, we introduce one-to-one codes. Let $l_n^\star(\cdot)$ denote a strictly lossless one-to-one length function. Further, denote $L_n^\star$ as the collection of all one-to-one codes (bijective mappings to binary sequences) on sequences of length n. The following result due to Alon and Orlitsky sets a lower limit on the one-to-one average codeword length.

Theorem 2 [14]: *Let the entropy of the random sequence X^n be equal to $H(X^n)$. Then,*

$$\mathbf{E}l_n^\star(X^n) \geq H(X^n) - \log(H(X^n)+1) - \log e. \tag{11}$$

Remark. Theorem 2 is indeed a significant result stating that the reduction in the average codeword length associated with a random sequence X^n is bounded from above by $\log(H(X^n)+1)+\log e$. Further, Alon and Orlitsky showed that if X^n follows the geometric distribution, the lower limit is attained.

When the source statistics are known, we can order all probabilities of the 2^n sequences in a decreasing fashion and then assign a codeword length $\lfloor \log j + 1 \rfloor$ to the j-th message sequence. It is straightforward to see that this coding strategy is the optimal one-to-one code but what is perhaps not straightforward is to analyze the average codeword length resulting from this coding strategy. In [15], Szpankowski derived the average codeword length of the non-universal one-to-one codes for binary memoryless sources, recently generalized by Kontoyiannis and Verdu [16] for finite-alphabet memoryless sources as the following.

Theorem 3 [15], [16]: *In the non-universal one-to-one compression of finite-alphabet memoryless sources, the average codeword length is given by*

$$\mathbf{E}l_n^\star(X^n) = H_n(\theta) - \frac{1}{2}\log n + O(1). \tag{12}$$

Szpankowski refers to the second-order term the *anti-redundancy* [15], which is the average codeword length reduction below the entropy. Therefore, the anti-redundancy in the non-universal one-to-one compression of finite-alphabet memoryless sources is $\frac{1}{2}\log n + O(1)$ when Kraft's inequality is relaxed.

IV. Universal One-to-One Codes

Thus far, it was shown that the optimal average codeword length is below the entropy for non-universal one-to-one codes. On the other hand, several challenges arise when universal one-to-one codes are concerned. First, the optimal codeword length assignment is no longer obvious. Further, the performance analysis of a given codeword length assignment is not straightforward.

Let $R_n^\star(l_n^\star, \theta)$ denote the average redundancy of the one-to-one code, which is defined in the usual way as

$$R_n^\star(l_n^\star, \theta) \triangleq \mathbf{E}l_n^\star(X^n) - H_n(\theta). \tag{13}$$

Further, define the one-to-one average maximin redundancy $\underline{R}_n^\star$ as

$$\underline{R}_n^\star = \sup_p \min_{l_n^\star \in L_n^\star} \int_{\theta\in\Lambda} R_n^\star(l_n^\star, \theta) p(\theta) d\theta, \tag{14}$$

where the supremum is taken over all distributions over the space Λ. Let the one-to-one average minimax redundancy $\bar{R}_n^\star$ be defined as

$$\bar{R}_n^\star = \min_{l_n^\star \in L_n^\star} \sup_{\theta\in\Lambda} R_n^\star(l_n^\star, \theta). \tag{15}$$

Theorem 4 *The one-to-one average minimax redundancy is no smaller than the one-to-one average maximin redundancy. That is*

$$\bar{R}_n^\star \geq \underline{R}_n^\star. \tag{16}$$

Sketch of the proof: The proof is straightforward by following the lines of proof of the same result for prefix-free codes in [24]. ■

Remark. According to Theorem 4, the average minimax redundancy is always at least equal to the average maximin redundancy. Please note that for the case of prefix-free codes it can be shown that they are equivalent [24], while the equivalence would not readily extend to the one-to-one codes.

To the best of our knowledge, the only existing work on universal one-to-one codes is by Kosut and Sankar [19], who proposed a so-called *type-size* coding scheme based on the type of the sequences [25]. The type of sequence x^n is given by

$$t_{x^n}(a) = \frac{|i : x_i = a|}{n} \text{ for } a \in \mathcal{A}. \tag{17}$$

For a type t, let the type class T_t be defined as

$$T_t = \{x^n \in \mathcal{A}^n : t_{x^n} = t\}. \tag{18}$$

Therefore, $|T_t|$ denotes the size of the type class of the type t, i.e., the total number of sequences with type t. Here, we will present a slightly modified version of the type-size code for the purpose of clarity of discussion, which has essentially the same performance. The type-size code essentially sorts the sequences based on the size of the corresponding type classes in a descending order. Therefore, the sequence x^n may appear before y^n only if $|T_{t_{x^n}}| < |T_{t_{y^n}}|$. Then, the rest is performed by assigning a codeword length $\lfloor \log j + 1 \rfloor$ to the j-th message sequence. Let l_n^{tsc} denote the length function associated with the type-size code. The performance of the type-size code was analyzed in [19].

Theorem 5 [19]: *In the universal one-to-one compression of the class of memoryless sources over a finite alphabet $\mathcal{A}$, for any $\epsilon > 0$, we have*

$$R_n^\star(l_n^{tsc}, \theta) \leq (1+\epsilon)\frac{|\mathcal{A}|-3}{2}\log n + O(1). \tag{19}$$

Remark. According to Theorem 5, the one-to-one average redundancy for memoryless sources of alphabet size $|\mathcal{A}|$ is

asymptotically bounded from above by $\frac{|\mathcal{A}|-3}{2}\log n+o(\log n)$, which is smaller than $\frac{|\mathcal{A}|-1}{2}\log n+O(1)$ attributed to prefix-free universal codes. However, it remains open to see whether this bound can be further improved and to asses how significant the improvement is.

Next, our main results on the universal one-to-one compression performance are presented.

Theorem 6 *Assume that the unknown parameter vector θ follows Jeffreys' prior $p_J(\cdot)$ given in* (7)*, where θ lies in the $|\mathcal{A}|-1$ dimensional simplex of memoryless parameter vectors. Then, type-size code is asymptotically optimal for the universal one-to-one compression of the family of finite-alphabet memoryless sources. That is*

$$l_n^{tsc}=\arg\min_{l_n^\star\in L_n^\star}\int_{\theta\in\Lambda}R_n^\star(l_n^\star,\theta)p_J(\theta)d\theta. \tag{20}$$

Sketch of the proof: We have

$$\mathbf{P}[X^n=x^n]=\int_{\theta\in\Lambda}\mu_\theta(x^n)p_J(\theta)d\theta. \tag{21}$$

On the other hand, since Jeffreys' prior is asymptotically capacity achieving, i.e., maximizes $I(\theta;X^n)$ [12], [24], it asymptotically results in equiprobable types. In other words,

$$\mathbf{P}[t_{X^n}=t]\simeq\frac{1}{\binom{n+|\mathcal{A}|-1}{n}}.^4 \tag{22}$$

where $\binom{n+|\mathcal{A}|-1}{n}$ denotes the total number of type classes, which is a constant with respect to x^n. Hence,

$$\mathbf{P}[X^n=x^n]\simeq\frac{1}{\binom{n+|\mathcal{A}|-1}{n}}\frac{1}{|T_{t_{x^n}}|}. \tag{23}$$

Therefore, by definition, the type-size coding orders the sequences in a descending fashion based on their probabilities, which completes the proof. ∎

Remark. According to Theorem 6, the type-size coding is optimal for the universal one-to-one compression of finite-alphabet memoryless sources when the unknown parameter vector follows Jeffreys' prior. Furthermore, it is easily seen that the average redundancy of type-size coding serves as a lower limit on the average maximin redundancy. However, our goal is to deduce a meaningful converse for d-dimensional parametric sources, which is carried out in the next theorem.

Theorem 7 *The one-to-one average maximin redundancy for the family $\mathcal{P}_\Lambda^d$ of d-dimensional parametric sources is bounded from below by*

$$\underline{R}_n^\star\geq\frac{d-2}{2}\log\frac{n}{2\pi e}-\log 2\pi e^2+\int_{\theta\in\Lambda}|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta+O\left(\frac{1}{\sqrt{n}}\right). \tag{24}$$

Sketch of the proof: We have

$$H(X^n)=H(X^n|\theta)+I(X^n;\theta) \tag{25}$$

[4] $f(n)\simeq g(n)$ if and only if $\lim_{n\to\infty}\frac{f(n)}{g(n)}=1$.

Assuming that θ follows Jeffreys' prior, we can get

$$H(X^n)=\bar{H}_n+\bar{R}_n, \tag{26}$$

where $\bar{R}_n$ is the average minimax redundancy for prefix-free codes given in (10) and $\bar{H}_n$ is given by

$$\bar{H}_n=\int_{\theta\in\Lambda}H_n(\theta)p_J(\theta)d\theta. \tag{27}$$

Hence, we can now use Theorem 2 to provide a lower bound on $\mathbf{E}l_n^\star(X^n)$. The proof is completed by seeing that $\log\bar{H}_n\leq\log n$ and noting that the average redundancy for the case where θ follows Jeffreys' prior provides a lower limit on the average maximin redundancy. ∎

Remark. Theorem 7 basically states that the one-to-one average maximin redundancy is bounded from below by $\underline{R}_n^\star\geq\frac{d-2}{2}\log n+O(1)$. By using Theorem 4, we can deduce that the bound also holds for the average minimax redundancy, i.e., $\bar{R}_n^\star\geq\frac{d-2}{2}\log n+O(1)$.

Finally, let us consider the performance of universal one-to-one codes for the case of memoryless sources. We have the following.

Corollary 8 *In the universal compression of memoryless sources with finite alphabet $\mathcal{A}$, for any $\epsilon>0$, we have*

$$\bar{R}_n^\star\geq\underline{R}_n^\star\geq\frac{|\mathcal{A}|-3}{2}\log n+O(1), \tag{28}$$

$$\underline{R}_n^\star\leq\bar{R}_n^\star\leq(1+\epsilon)\frac{|\mathcal{A}|-3}{2}\log n+O(1). \tag{29}$$

Remark. According to Corollary 8, for memoryless sources the average minimax and average maximin redundancy scale as $\frac{|\mathcal{A}|-3}{2}\log n+O(1)$ Theorem 8 implies that type-size coding is minimax (and maximin) optimal up to $o(\log n)$ for the universal one-to-one compression of memoryless sources. Please note that the lower bound on the average minimax redundancy was already known as it readily results from [20]. Our result further implies that the same bound also holds for the average maximin redundancy.

V. Numerical Example

It is desirable to see how much reduction is offered by universal one-to-one compression compared with the prefix-free universal compression. We compare the performance of universal one-to-one codes with that of the universal prefix-free codes through a numerical example from [13]. We consider a first-order Markov source with alphabet size $|\mathcal{A}|=256$, where the number of source parameters is $d=256\times255=62580$. Please note that our results did not provide an actual code for the compression of a parametric source. Hence, we compare the converse bound of Theorem 7 on the average minimax (maximin) redundancy of universal one-to-one codes with the performance of the minimax optimal universal prefix-free code.

Fig. 1 compares the minimum average number of bits per symbol required to compress the class of the first-order Markov sources normalized to the entropy of the sequence

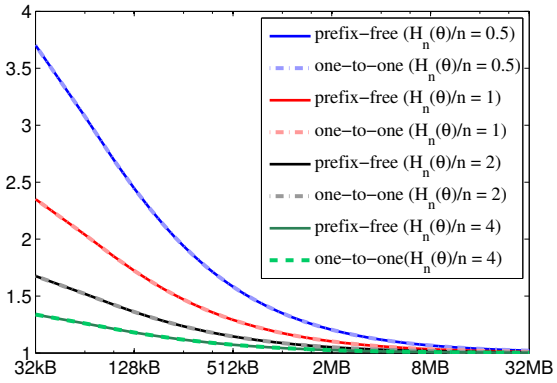

Fig. 1. The Lower bound on the average minimax redundancy as a function of sequence length n for prefix-free and one-to-one codes for different values of entropy rate $H_n(\theta)/n$.

for different values of entropy rates in bits per source symbol (per byte). As can be seen, relaxing the prefix constraint at its best does not offer meaningful performance improvement on the compression performance as the curves for the prefix-free codes and one-to-one codes almost coincide. This leads to the conclusion that the universal one-to-one codes are not of much practical interest.

On the other hand, if the source entropy rate is 1 bit per byte ($H_n(\theta)/n = 1$), the compression rate on sequences of length 32kB (for both prefix-free and one-to-one codes) is around 2.25 times the entropy-rate, which results in more than 100% overhead on top of the entropy-rate for both prefix-free and one-to-one universal codes. Hence, we conclude that average redundancy poses significant overhead in the universal compression of finite-length low-entropy sequences, such as the Internet traffic, which cannot be compensated by dropping the prefix constraint. Hence, the side information provided in memory-assisted universal compression to overcome the redundancy is essential even if the prefix constraint is dropped (cf. [26], [27] for details about universal memory-assisted compression.)

VI. Conclusion

In this paper, the fundamental limits of universal one-to-one codes (without prefix constraint) were investigated. It was proved that the type-size code proposed earlier in the literature is optimal up to $o(\log n)$ for universal one-to-one compression of memoryless sources. Further, a lower bound on the average minimax redundancy of universal one-to-one codes was derived. Finally, it was also demonstrated that the reduction on the average redundancy by relaxing the prefix constraint is negligible compared with the cost of universality in universal compression of low-entropy small sequences (e.g., network packets).

Acknowledgements

This research was carried out when A. Beirami was affiliated with Georgia Institute of Technology. The work of A. Beirami and F. Fekri was supported in part by the National Science Foundation under grant No. CNS-1017234.

References

[1] W. Szpankowski, "Asymptotic average redundancy of Huffman (and other) block codes ," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2434–2443, Nov. 2000.
[2] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 783 – 795, Nov. 1973.
[3] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, May 1977.
[4] F. Willems, Y. Shtarkov, and T. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.
[5] J. Rissanen and J. Langdon, G., "Universal modeling and coding," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 12 – 23, Jan. 1981.
[6] D. Baron and Y. Bresler, "An O(N) semipredictive universal encoder via the BWT," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 928–937, May 2004.
[7] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1034–1054, Jul. 1991.
[8] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. 27, no. 2, pp. 199–207, 1981.
[9] M. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 643 –652, May 1995.
[10] J. Rissanen, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 526–532, Jul. 1986.
[11] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743 –2760, Oct. 1998.
[12] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 714 –722, May 1995.
[13] A. Beirami and F. Fekri, "Results on the redundancy of universal compression for finite-length sequences," in *2011 IEEE International Symposium on Information Theory (ISIT '11)*, Jul. 2011, pp. 1604–1608.
[14] N. Alon and A. Orlitsky, "A lower bound on the expected length of one-to-one codes," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1670–1672, Sept. 1994.
[15] W. Szpankowski, "A one-to-one code and its anti-redundancy," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4762–4766, Oct. 2008.
[16] I. Kontoyiannis and S. Verdu, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.
[17] S. Leung-Yan-Cheong and T. Cover, "Some equivalences between shannon entropy and kolmogorov complexity," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 331–338, May 1978.
[18] W. Szpankowski and S. Verdu, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4017–4025, Jul. 2011.
[19] O. Kosut and L. Sankar, "Universal fixed-to-variable source coding in the finite blocklength regime," in *2013 IEEE International Symposium on Information Theory Proceedings (ISIT '13)*, Jul. 2013, pp. 649–653.
[20] ——, "New results on third-order coding rate for universal fixed-to-variable source coding: Converse and prefix codes," in *2014 International Symposium on Information Theory (ISIT 2014)*, Jul. 2014.
[21] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2686–2707, Nov. 2004.
[22] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453 –471, May 1990.
[23] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2104 –2109, Sept. 1999.
[24] R. G. Gallager, "Source coding with side information and universal coding," *unpublished*.
[25] I. Csiszár, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.
[26] A. Beirami, M. Sardari, and F. Fekri, "Results on the fundamental gain of memory-assisted universal source coding," in *2012 IEEE International Symposium on Information Theory (ISIT '12)*, Jul. 2012, pp. 1087–1091.
[27] ——, "Results on the optimal memory-assisted universal compression performance for mixture sources," in *51st Annual Allerton Conference*, Oct. 2013, pp. 890–895.