Fifty-first Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 2 - 3, 2013

Results on the Optimal Memory-Assisted Universal Compression Performance for Mixture Sources

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

Abstract—**In this paper, we consider the compression of a sequence from a mixture of K parametric sources. Each parametric source is represented by a d-dimensional parameter vector that is drawn from Jeffreys' prior. The output of the mixture source is a sequence of length n whose parameter is chosen from one of the K source parameter vectors uniformly at random. We are interested in the scenario in which the encoder and the decoder have a common side information of T sequences generated independently by the mixture source (which we refer to as memory-assisted universal compression problem). We derive the minimum average redundancy of the memory-assisted universal compression of a new random sequence from the mixture source and prove that when $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ for some $\epsilon > 0$, the side information provided by the previous sequences results in significant improvement over the universal compression without side information that is a function of n, T, and d. On the other hand, as K grows, the impact of the side information becomes negligible. Specifically, when $K = \Omega\left(n^{\frac{d}{2}(1+\epsilon)}\right)$ for some $\epsilon > 0$, optimal memory-assisted universal compression almost surely offers negligible improvement over the universal compression without side information.**

Index Terms—**Universal Lossless Compression; Side Information; Redundancy-Capacity Theorem; Mixture Source.**

I. Introduction

The recent rapid growth in the network traffic has motivated new research directions that target to leverage the existing correlations in the sequences (network packets) in order to reduce the traffic. These solutions must be transparent to the user and the application and hence must perform on the network layer, where sequences in the network are present. The state-of-the-art solutions in network layer traffic reduction are based on deduplication and compression. Deduplication based solutions work well when there are exact retransmissions of duplicates of a data chunk in the network [1]–[3]. However, a great fraction of network traffic consists of data that are statistically correlated but not exact duplicates of each other, and hence, not exploited by deduplication mechanisms. On the other hand, although universal compression captures the statistical correlations between the data, compression based traffic reduction solutions usually need to observe a long sequence before they can effectively learn the existing patterns in the sequence for efficient compression. Therefore, the universal compression based solutions perform poorly on small sequences [4], [5] (which is the case for the network data packets), where sufficient training data is not available.

We investigate the problem of packet-level network traffic compression from an information-theoretic point of view. As shown in Fig. 1, we assume that each network packet is a sequence (sample) of length n from a mixture of K parametric sources with parameter vectors $\theta^{(1)}, \ldots, \theta^{(K)}$ such that $\theta^{(i)}$ is drawn independently from Jeffreys' prior. We assume that each output sequence from this mixture source is chosen from $\theta^{(S)}$, where the index S of the source is chosen uniformly at random from $\{1, \ldots K\}$. Please note that as M_1 and M_2 are two routers inside the Internet, they observe several sequences from different servers in the network, and hence, the mixture number K may be indeed very large. We consider the scenario where T sequences from the mixture source are memorized and shared as side information between the encoder (at the sender router M_1 in Fig. 1) and the decoder (at the receiver router M_2) as a result of the prior communication between the two intermediate nodes. We refer to this setup as memory-assisted universal compression problem, where we wish to derive the average redundancy of the *optimal* memory-assisted universal compression where optimality is defined in the sense of minimizing the average redundancy as a function of n, K, and T.

In [6]–[8], we derived the optimal memory-assisted compression performance for a single source, i.e., $K = 1$, and proved that significant improvement is obtained by applying the memory-assisted universal compression on small sequences with sufficient number of side information sequences. In [9], we extended the setup to fixed finite K which is known to the encoder and the decoder a priori. We further assumed that the indices $S(1), \ldots, S(T)$ of the sources that generated the T side information sequences are also known to both the encoder and the decoder. The latter makes sense as the source IP address is included in the header of the IP packets. Under these assumptions, we demonstrated that the memory-assisted compression using clustering of the side information by the index S of the sequence offers significant improvement over universal compression without side information. Inspired from this, in [10], we developed a clustering algorithm for memory-assisted

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1017234.

978-1-4799-3410-2/13/$31.00 ©2013 IEEE

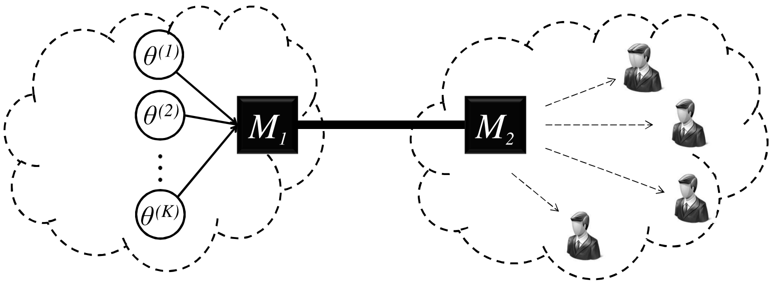

Fig. 1. The basic scenario of memory-assisted universal compression for a mixture source.

universal compression and demonstrated its effectiveness on real packets gathered from Internet. However, it remained an open problem what would be the performance of the *optimal* memory-assisted universal compression strategy in the sense that, given the side information, it achieves the minimum codeword length in the compression of a new sequence from the mixture source.

In this paper, we generalize the setting of [9] by letting K grow with n and drop the assumption that the indices $S(1), \ldots, S(T)$ are known. We further relax the assumption that K is known to both the encoder and the decoder. Please note that K can grow very large as a router in the Internet may communicate with several different servers. Further, as our experiments in [10] suggest, the mere clustering of packets based on the sender IP address is not a good job since there are several packets with high correlations sent from different IP addresses and several packets with the same sender IP address do not demonstrate much correlation to be leveraged. In this scenario, we formally characterize the average redundancy incurred in the optimal memory-assisted compression of a new sequence given that the encoder and the decoder have access to a shared memorized content of T sequences (each of length n from the mixture of K parametric sources) from the previous communication and compare with that of the universal compression without side information.

The rest of this paper is organized as follows. In Section II, we review the necessary background on universal compression. In Section III, we present the formal definition of the problem. In Section IV, we provide the main results and discuss their implications. In Section V, we provide the technical analysis of the results. Finally, Section VI concludes this paper.

II. Background on Universal Source Coding

In this section, we briefly review the necessary background on the universal compression of parametric sources. We let a parametric source be defined using a d-dimensional parameter vector $\theta = (\theta_1, ..., \theta_d) \in \Lambda$ that is a priori unknown, where d denotes the number of the source parameters and $\Lambda \subset \mathbb{R}^d$ is the space of d-dimensional parameter vectors of interest. Denote μ_θ as the parametric source (i.e., the probability measure defined by the parameter vector θ on sequences of length n).

Let $\mathcal{A}$ denote a finite alphabet. Let X^n denote a sample (random variable) from the probability measure μ_θ. We further denote $x^n = (x_1, ..., x_n) \in \mathcal{A}^n$ as a realization of the random variable X^n. Then, define $H_n(\theta) \triangleq H(X^n|\theta)$ as the source entropy given the parameter vector θ, i.e.,

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

Please note that throughout this paper $\log(\cdot)$ always denotes the logarithm in base 2 and expectations are taken over the random sequence X^n with respect to the probability measure μ_θ unless otherwise stated.

In this paper, we focus on the class of strictly lossless uniquely decodable fixed-to-variable codes defined as the following. The code $c_n : \mathcal{A}^n \to \{0,1\}^*$ is called strictly lossless (also called zero-error) on sequences of length n if 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 $l_n : \mathcal{A}^n \to \mathbb{R}$ denote the universal strictly lossless length function for the codeword $c_n(x^n)$ associated with the sequence x^n such that $l_n(\cdot)$ satisfies Kraft's inequality to ensure unique decodability. In this paper, we ignore the integer constraint on the length function, which results in a negligible $O(1)$ redundancy analyzed in [11], [12].

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

$$R_n(l_n, \theta) = \mathbf{E} l_n(X^n) - H_n(\theta). \tag{2}$$

Note that the expected (average) redundancy is always non-negative. Further, a code is called universal if it uniformaly achieves the source entropy rate asymptotically, 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, i.e.,

$$\mathcal{I}(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n \log e} \mathbf{E}\left\{\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\left(\frac{1}{\mu_\theta(X^n)}\right)\right\}. \tag{3}$$

Fisher information matrix quantifies the amount of information, on the average, that each symbol in a sample sequence

x^n from the source conveys about the source parameters. Let Jeffreys' prior on the parameter vector θ be denoted by

$$p_J(\theta) \triangleq \frac{|\mathcal{I}(\theta)|^{\frac{1}{2}}}{\int |\mathcal{I}(\lambda)|^{\frac{1}{2}} d\lambda}. \tag{4}$$

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 [13]. Jeffreys' prior is particularly interesting because it also corresponds to the worst-case compression performance for the best compression scheme (called the capacity achieving prior). Define $\bar{R}_n$ as the minimum average redundancy when θ is chosen using Jeffreys' prior, i.e.,

$$\bar{R}_n = \min_{l_n} \int_{\theta \in \Lambda} R_n(l_n, \theta) p_J(\theta) d\theta. \tag{5}$$

It is evident that $\bar{R}_n$ is the average maximin redundancy since θ is chosen to follow Jeffreys' prior in our setup (i.e., the capacity achieving prior). Therefore, we have

$$\bar{R}_n = I(X^n; \theta) = H(X^n) - H_n(\theta), \tag{6}$$

Furthermore, the average maximin redundancy is equal to the average minimax redundancy for the case of parametric sources studied in this paper (cf. [5] and the references therein). The average minimax redundancy was characterized by Clarke and Barron for memoryless sources [13] and later generalized for Markov sources by Atteson [14] as given by

$$\bar{R}_n = \frac{d}{2} \log\left(\frac{n}{2\pi e}\right) + \log \int_{\theta \in \Lambda} |\mathcal{I}_n(\theta)|^{\frac{1}{2}} d\theta + O\left(\frac{1}{n}\right).^{1} \tag{7}$$

III. Problem Setup

In this section, we present the problem setup. We assume that each sequence of length n is generated using the parameter vector θ, which is supported on a countable support set Δ of $K \triangleq |\Delta|$ points on the space Λ of parameter vectors. Let Δ denote the set of all the parameter vectors in the support set, i.e., $\Delta \triangleq \{\theta^{(i)}\}_{i=1}^{K}$. Note that we let K deterministically scale with n. Denote $[K]$ as the set $\{1, ..., K\}$. We assume that $\forall i \in [K]$, we have $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_d^{(i)})$ is chosen at random according to the Jeffreys' prior on the d-dimensional parameter space Λ. Further, for the generation of each sequence, the generator source is selected uniformly at random from the mixture. In other words, $p(\theta|\Delta) = \frac{1}{K}\sum_{i=1}^{K} \delta(\theta - \theta^{(i)})$, where $\theta^{(i)}$ follows Jeffreys' prior on Λ and $\frac{1}{K}$ is the probability that the sequence is generated by source $\theta^{(i)}$ in the mixture. Please note that the random set Δ (which is unknown a priori) is randomly generated once according to Jeffreys' prior and is used thereafter for the generation of all packets at all times.

In this setup, as in Fig. 1, the source is a mixture of K parametric sources $\mu_{\theta^{(1)}}, \ldots, \mu_{\theta^{(K)}}$ with d-dimensional unknown parameter vectors $\theta^{(1)}, \ldots, \theta^{(K)}$. Let S be a random variable that determines the source index, and hence is uniform over $[K]$, i.e., $\mathbf{P}[S = i] = \frac{1}{K}$. Then, by definition, we have $\theta = \theta^{(S)}$ given Δ. Unlike Δ that is generated once, S is uniformly selected from $[K]$ every time a new sequence is generated.

We consider the following scenario. We assume that, in Fig. 1, both the encoder (at M_1) and the decoder (at M_2) have previously communicated T previous sequences (indexed by $[T]$) from the mixture of K parametric sources, where each of the sequences is independently generated according to the above procedure. Let $m \triangleq nT$ denote the total length of the previous T sequences from the mixture source.[2] Further, denote $\mathbf{y}^{n,T} = \{y^n(t)\}_{t=1}^{T}$ as the set of previous T sequences from shared between M_1 and M_2, where $y^n(t)$ is a sequence of length n generated from the source $\theta^{S(t)}$ and $S(t)$ follows a uniform distribution over $[K]$. In other words, $y^n(t) \sim \mu_{\theta^{(S(t))}}$. Further, denote $\mathbf{S}$ as the vector $\mathbf{S} = (S(1), ..., S(T))$, which contains the index of the source that generated the T previous side information sequences.

The objective is to analyze the average redundancy in the compression of a new sequence x^n that is independently generated by the same mixture source with source index Z (which is also uniformly chosen over $[K]$). We investigate the fundamental limits of the memory-assisted universal compression when both the encoder and the decoder have access to the side information sequence $\mathbf{y}^{n,T}$ and compare with that of the universal compression without side information of the previous sequences. It is straightforward to verify that this problem is equivalent to the analysis of $H(X^n|\mathbf{Y}^{n,T})$ and $H(X^n)$ for different values of the sequence length n, memory (side information) size $m = nT$, and number of sources in the mixture K. In other words, we are seeking the minimum number of bits that is required for representing a random sequence X^n when $\mathbf{Y}^{n,T}$ is present (at both the encoder and the decoder) or not.

IV. Main Results

Before we state the main results of this paper, we need to introduce a fundamental quantity that will be used in the derivations.

Definition. Let $H_n(\Delta, Z) \triangleq H(X^n|\Delta, Z)$ be defined as the entropy of a random sequence X^n from the mixture source given that the source parameters are known (Δ is known) and the index of the source that has generated the sequence (i.e., Z) is also known. In other words, the parameter vector $\theta^{(Z)}$ associated with sequence X^n is known. Then, it is simply shown in this case that

$$H_n(\Delta, Z) = \frac{1}{K}\sum_{i=1}^{K} H_n(\theta^{(i)}), \tag{8}$$

where $H_n(\theta^{(i)})$ is the entropy of source $\mu_{\theta^{(i)}}$ given $\theta^{(i)}$ defined in (1). Please note that $H_n(\Delta, Z)$ is *not* the achievable

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

[2] For simplicity of the discussion we consider the length of all sequences to be equal to n. However, most of the results are readily extendible to the case where the sequences are not necessarily equal in length.

performance of compression. It is merely introduced so as to make the presentation convenient.

In the sequel, we first state the compression performance in the case of known source parameters. Then, we derive the impact of universality (i.e., the source parameter being unknown) on the universal compression without side information as well as the memory-assisted universal compression. The sketches of the proofs are deferred to Section V.

A. Known Source Parameters

First, we derive the entropy of the mixture source (which sets the asymptotic fundamental lower limit on the codeword length) for the known source parameters case, i.e., Δ is known. Define $H_n(\Delta) \triangleq H(X^n|\Delta)$.

Theorem 1. **(a)** *If $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ for some $\epsilon > 0$, then*

$$H_n(\Delta) = H_n(\Delta, Z) + \log K + O\left(\frac{1}{n}\right) \ a.s.^{3,4}$$

(b) *If $K = \Omega\left(n^{\frac{d}{2}(1+\epsilon)}\right)$ for some $\epsilon > 0$,*[5] *then*

$$H_n(\Delta) = H_n(\Delta, Z) + \bar{R}_n + O\left(\frac{1}{n}\right) \ a.s.,$$

where $\bar{R}_n$ is given by (7).

Remark. Theorem 1 determines the minimum codeword length when the parameter vectors are known. Please note that $H_n(\Delta)$ serves as a trivial lower bound on the codeword length for the case of universal compression (unknown parameter vectors) as well. According to Part (a), for sufficiently small K, the codeword length in the optimal compression converges to the entropy of the mixture (i.e., $H(X^n|Z,\Delta)$) plus the average price $\log K$ required to determine the respective source parameter in the encoder (i.e., $H(Z|\Delta)$). Therefore, the optimal coding strategy when the source parameters are known (almost surely) would be to encode the source index Z and then use the optimal code (e.g., Huffman code) associated with parameter $\theta^{(Z)}$ for sequences of length n to encode the sequence x^n. In fact, if $\lim_{n\to\infty} \frac{\log K}{\frac{d}{2}\log n} < 1$, then the cost of encoding the parameter is asymptotically smaller than the universal coding of the parameter and hence it is still beneficial to encode the parameter using an average of $\log K$ bits. Further, if $K = 1$, then $\Delta = \theta^{(1)}$ and $Z = 1$ would be deterministic. Hence, $H_n(\Delta) = H_n(\Delta, Z) = H_n(\theta^{(1)})$, which was introduced in (1) as the average compression limit for the single known parameter case. The interesting phenomenon here is that in Part (b) when all the (known) parameter vectors are chosen from Jeffreys' prior, the entropy converges to the mixture entropy plus an extra term $\bar{R}_n$, which is exactly the average minimax redundancy in the *universal* compression of parametric sources with d *unknown* parameters given in (7). At first, it may seem odd that the codeword length in the case that the source parameters are *known* incurs a term that is associated with the universal compression of a source with *unknown* parameters. However, in this case the cost of encoding the source index surpasses the cost of universal encoding of the source parameter. Hence, it no longer makes sense to encode the parameter for the compression of the sequence x^n. More rigorously speaking, as will be shown in Section V-A, the probability distribution of x^n given Δ would converge to the probability distribution of x^n when the source has one *unknown* parameter vector that follows Jeffreys' prior. This in turn results in the $\bar{R}_n$ term in the compression performance.

B. Unknown Source Parameters

In order to see the impact of the universality on the compression performance, i.e., to investigate the impact of Δ being unknown, we will analyze the redundancy for the following two schemes.

Definition. We refer to Ucomp as the universal compression without side information. We further refer to $R_{\text{Ucomp}}(n, K)$ as the average redundancy of the universal compression of a sequence of length n (in our problem setup described in Section III). In other words,

$$R_{\text{Ucomp}}(n, K) \triangleq H(X^n) - H_n(\Delta). \tag{9}$$

Theorem 2. *In the case of* Ucomp*,*
(a) *if $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ for some $\epsilon > 0$, then*

$$R_{\text{Ucomp}}(n, K) = \bar{R}_n - \log K + O\left(\frac{1}{n}\right) a.s.$$

(b) *If $K = \Omega\left(n^{\frac{d}{2}(1+\epsilon)}\right)$ for some $\epsilon > 0$, then*

$$R_{\text{Ucomp}}(n, K) = O\left(\frac{1}{n}\right) a.s.$$

Remark. According to Theorem 2, in the universal compression of a sequence of length n from the mixture source for sufficiently small K, the main term of the redundancy is $\bar{R}_n - \log K$ which can be significantly large. Again, if $K = 1$, then $R_{\text{Ucomp}}(n, 1) = \bar{R}_n$, which is exactly the average minimax redundancy in the case of one unknown source parameter described in (7). On the other hand, for large K, we almost surely expect no extra redundancy associated with universality. This is not surprising as even in the *known* sources case, the performance converges to that of the *unknown* source parameters that follow Jeffreys' prior. Therefore, there is no extra penalty when the source parameters are indeed unknown.

Theorem 2 also suggests that independent of K and $\log K$, the price of universality is given by $\bar{R}_n$ (which is defined as the price of universal compression of a sequence of length n

[3] An event A happens a.s. (almost surely) if and only if $\mathbb{P}[A] = 1$.

[4] Please note that the sample space is the set of all source parameter vectors $\Delta = \{\theta^{(i)}\}_{i=1}^{K}$ such that $\theta^{(i)}$ is drawn independently from Jeffreys' prior.

[5] $f(n) = \Omega(g(n))$ if and only if $g(n) = O(f(n))$.

from a source with unknown parameter that follows Jeffreys' prior) on top of $H_n(\Delta, Z)$ (i.e., the entropy when Δ and Z are known).

Definition. We refer to UcompOM as the optimal memory-assisted universal compression strategy in the sense that it achieves the minimum average redundancy given the side information. We further refer to $R_{\text{UcompOM}}(n, m, K)$ as the average redundancy of the optimal memory-assisted universal compression (with total available memory size m) in our problem setup described in Section III, where $T = \frac{m}{n}$ sequences (samples) from the mixture source are observed and are available to both the encoder and the decoder as side information. In other words,

$$R_{\text{UcompOM}}(n, m, K) \triangleq H(X^n | \mathbf{Y}^{n,T}) - H_n(\Delta). \tag{10}$$

Theorem 3. *In the case of* UcompOM*,*
(a) *if* $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ *for some* $\epsilon > 0$*, then*

$$R_{\text{UcompOM}}(n, m, K) = \hat{R} + O\left(\frac{1}{\sqrt{T}} + \frac{1}{n}\right) a.s.,$$

where

$$\hat{R} \triangleq \frac{d}{2} \log\left(1 + \frac{n}{\frac{1}{K}m}\right). \tag{11}$$

(b) *If* $K = \Omega\left(n^{\frac{d}{2}(1+\epsilon)}\right)$ *for some* $\epsilon > 0$*, then*

$$R_{\text{UcompOM}}(n, m, K) = O\left(\frac{1}{n}\right) a.s.$$

Remark. Theorem 3 characterizes the redundancy of the optimal memory-assisted universal compression scheme, which uses a memory of size $m = nT$ (T sequences of size n) in the compression of a new sequence of length n. It is expected that memorization decreases the redundancy. As suggested by Part (a) of the theorem, when $\log K$ or roughly K is sufficiently small the redundancy of the UcompOM decreases. As an important special case if $K = 1$, then $R_{\text{UcompOM}}(n, m, 1) = \frac{d}{2} \log\left(1 + \frac{n}{m}\right) + O\left(\frac{1}{n} + \frac{1}{\sqrt{T}}\right)$, which gives back Theorem 2 of [8] about the average minimax redundancy of the single source. Further, it is deduced from Theorem 3 that $\lim_{T\to\infty} R_{\text{UcompOM}}(n, m, K) = O\left(\frac{1}{n}\right)$ (regardless of K), i.e., the price of universality would be negligible given that sufficiently large memory (side information) is available. Thus, the benefits of optimal memory-assisted universal compression would be substantial when $\log K$ is sufficiently small. On the other hand, when $\log K$ grows very large, there is no benefit obtained from the memory-assisted universal compression and the performance improvement becomes negligible. This is due to the fact that in light of Theorem 2(b) the compression performance for the known source parameters case is already that of the universal compression.

Let $B(n, m, K) \triangleq R_{\text{Ucomp}}(n, K) - R_{\text{UcompOM}}(n, m, K)$ denote the performance improvement of UcompOM over Ucomp. The next corollary which is a direct consequence of Theorems 2 and 3 characterizes $B(n, m, K)$.

Corollary 4. (a) *If* $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ *for some* $\epsilon > 0$*, then we have*

$$\begin{aligned} B(n, m, K) &= \bar{R}_n - \log K - \frac{1}{K} \sum_{i=1}^{K} \frac{d}{2} \log\left(1 + \frac{nK}{m}\right) \\ &+ O\left(\frac{1}{\sqrt{T}} + \frac{1}{n}\right) a.s. \end{aligned}$$

(b) *If* $K = \Omega\left(n^{\frac{d}{2}(1+\epsilon)}\right)$ *for some* $\epsilon > 0$*, then*

$$B(n, m, K) = O\left(\frac{1}{n}\right) a.s.$$

Remark. In light of Corollary 4, it is relatively straightforward to see that if T is sufficiently large, we have $B(n, m, K) = \Theta(\log n)$.[6] In particular, this implies that the optimal memory-assisted universal compression partially compensates the main extra term in universal compression on top of $H_n(\Delta, Z)$ (which is $\frac{d}{2} \log n$) for sufficiently small K. Further, if $T \to \infty$ and K is constant, then $B(n, m, K) = \frac{d}{2} \log n + O(1)$, which completely cancels out the main extra redundancy term. In this case, UcompOM achieves $H_n(\Delta, Z)$ with a constant negligible extra term.

V. Technical Analysis

In this section we provide the sketches of the proofs of Theorems 1, 2, and 3.

A. Sketch of the Proof of Theorem 1

It is straightforward to show that

$$H(X^n | \Delta) = H(X^n | \Delta, Z) + I(X^n; Z | \Delta) \tag{12}$$

Further, $I(X^n; Z|\Delta) = H(Z|\Delta) - H(Z|X^n, \Delta)$. Clearly, $H(Z|\Delta) = \log K$ by definition. For sufficiently small K, we have the following lemma.

Lemma 1. *If* $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ *for some* $\epsilon > 0$*, then* $H(Z|X^n, \Delta) = O\left(\frac{1}{n}\right)$*.*

Therefore, $I(X^n; Z|\Delta) = \log K + O\left(\frac{1}{n}\right)$ which completes the proof of Part (a).

Lemma 1 states that when the number of the source parameters is sufficiently small, the Maximum-Likelihood (ML) estimate of the parameter obtained from sequence X^n with high probability will determine the true unknown source parameter. It is proved by showing that the ML estimate is with high probability closest to the true unknown source parameter vector with high probability.

The following lemma is the key to the proof of Part (b) for sufficiently large K.

[6] $f(n) = \Theta(g(n))$ if and only if $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

Lemma 2. *If $K = \Omega\left(n^{\frac{d}{2}(1+\epsilon)}\right)$ for some $\epsilon > 0$, then*

$$\frac{1}{K}\sum_{i=1}^{K}\mu_{\theta^{(i)}}(x^n) \sim \int_{\theta\in\Lambda}\mu_\theta(x^n)p_J(\theta)d\theta \ a.s.^7$$

Lemma 2 states that the probability measure induced on a random sequence X^n for sufficiently large K converges to that of the universal measure induced by one unknown parameter that follows Jeffreys' prior, completing the proof of Part (b). Lemma 2 is proved by showing that when K is sufficiently large there are almost surely $\omega(1)$ source parameter vectors in the vicinity of the ML estimate.

B. Sketch of the Proof of Theorem 2

It is straightforward that

$$H(X^n) = H(X^n|\Delta, Z) + I(X^n; \Delta, Z) \tag{13}$$

Lemma 3. *We have $I(X^n; \Delta, Z) = I(X^n; \theta^{(Z)}|Z)$.*

According to Lemma 3, all the information that X^n carries about the set Δ of the unknown parameter vectors and index Z is contained in $I(X^n; \theta^{(Z)}|Z)$. Since each of the unknown parameter vectors follow Jeffreys' prior, we have $I(X^n; \theta^{(Z)}|Z = z)$ is equal to the average minimax redundancy [5]. Thus,

$$I(X^n; \theta^{(Z)}|Z) = \frac{1}{K}\sum_{i=1}^{K} I(X^n; \theta^{(Z)}|Z = z) = \bar{R}_n, \tag{14}$$

which completes the proof if combined with Theorem 1.

C. Sketch of the Proof of Theorem 3

In the case of UcompOM, we have

$$H(X^n|\mathbf{Y}^{n,T}) = H(X^n|\mathbf{Y}^{n,T}, \mathbf{S}, Z) + I(\mathbf{S}, Z; X^n|\mathbf{Y}^{n,T}). \tag{15}$$

On the other hand, we also have

$$H(X^n|\mathbf{Y}^{n,T}\!,\mathbf{S}, Z) = H_n(\Delta, Z) + I(X^n; \theta^{(Z)}|\mathbf{Y}^{n,T}\!,\mathbf{S}, Z). \tag{16}$$

We need the following lemmas to complete the proof of Part (a).

Lemma 4. *If $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ for some $\epsilon > 0$, then*

$$I(X^n; \theta^{(Z)}|\mathbf{Y}^{n,T}, \mathbf{S}, Z) = \hat{R} + O(T^{-\frac{1}{2}}),$$

where $\hat{R}$ is defined in (11).

The proof of Lemma 4 is carried out by rewriting the LHS as $I(X^n, \mathbf{Y}^{n,T}; \theta^{(Z)}|\mathbf{S}, Z) - I(\mathbf{Y}^{n,T}; \theta^{(Z)}|\mathbf{S}, Z)$. Then, the first term is shown to converge to $\hat{R}$ whilst the second term is asymptotically vanishing.

Lemma 5. *If $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ for some $\epsilon > 0$, then*

$$I(\mathbf{S}, Z; X^n|\mathbf{Y}^{n,T}) = \log K + O\left(\frac{1}{n} + \frac{1}{T}\right).$$

The proof of Lemma 5 is carried out by rewriting the LHS as $H(Z|\mathbf{Y}^{n,T}, \mathbf{S}) + H(\mathbf{S}|\mathbf{Y}^{n,T}) - H(\mathbf{S}, Z|\mathbf{Y}^{n,T}, X^n)$ and demonstrating that the last two terms asymptotically vanish. Part (a) is proved by combining Lemmas 4 and 5. For Part (b), when $K = \Omega\left(n^{\frac{d}{2}(1+\epsilon)}\right)$ for some $\epsilon > 0$, we have $R_{\text{Ucomp}}(n, K) = O\left(\frac{1}{n}\right) a.s.$ On the other hand, $R_{\text{UcompOM}}(n) \leq R_{\text{Ucomp}}(n, K)$, which completes the proof.

VI. Conclusion

In this paper, we derived the fundamental limits of optimal memory-assisted universal compression for a mixture of K parametric sources. Our results demonstrated that when $K = O\left(n^{\frac{d}{2}(1-\epsilon)}\right)$ for some $\epsilon > 0$, there is significant improvement offered by the memory-assisted universal compression. On the other hand, as K grows the benefits of memory-assisted universal compression vanish to the extent that when $K = \Omega\left(n^{\frac{d}{2}(1+\epsilon)}\right)$ for some $\epsilon > 0$ the performance of optimal memory-assisted universal compression almost surely becomes that of the universal compression without side information.

References

[1] Z. Zhuang, C.-L. Tsao, and R. Sivakumar, "Curing the amnesia: Network memory for the internet, Tech Report," 2009. [Online]. Available: http://www.ece.gatech.edu/research/GNAN/archive/tr-nm.pdf

[2] S. Sanadhya, R. Sivakumar, K.-H. Kim, P. Congdon, S. Lakshmanan, and J. P. Singh, "Asymmetric caching: improved network deduplication for mobile devices," in *Proc. of Mobicom '12*, 2012, pp. 161–172.

[3] M. Sardari, A. Beirami, and F. Fekri, "Memory-assisted universal compression of network flows," in *IEEE INFOCOM 2012*, March 2012, pp. 91–99.

[4] A. Beirami and F. Fekri, "Results on the redundancy of universal compression for finite-length sequences," in *2011 IEEE International Symp. on Info. Theory (ISIT 2011)*, July 2011, pp. 1604–1608.

[5] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Info. Theory*, vol. 41, no. 3, pp. 714 –722, May 1995.

[6] A. Beirami and F. Fekri, "Memory-assisted universal source coding," in *2012 Data Compression Conference (DCC '2012)*, April 2012, p. 392.

[7] M. Sardari, A. Beirami, and F. Fekri, "On the network-wide gain of memory-assisted source coding," in *2011 IEEE Information Theory Workshop (ITW' 2011)*, October 2011, pp. 476–480.

[8] A. Beirami and F. Fekri, "On lossless universal compression of distributed identical sources," in *2012 IEEE International Symp. on Info. Theory (ISIT 2012)*, July 2012, pp. 561–565.

[9] A. Beirami, M. Sardari, and F. Fekri, "Results on the fundamental gain of memory-assisted universal source coding," in *2012 IEEE International Symp. on Info. Theory (ISIT 2012)*, July 2012, pp. 1087–1091.

[10] M. Sardari, A. Beirami, J. Zou, and F. Fekri, "Content-aware network data compression using joint memorization and clustering," in *2013 IEEE Conference on Computer Networks (INFOCOM 2013)*, April 2013.

[11] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Info. Theory*, vol. 50, no. 11, pp. 2686 – 2707, November 2004.

[12] W. Szpankowski, "Asymptotic average redundancy of Huffman (and other) block codes ," *IEEE Trans. Info. Theory*, vol. 46, no. 7, pp. 2434–2443, November 2000.

[13] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Info. Theory*, vol. 36, no. 3, pp. 453 –471, May 1990.

[14] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Info. Theory*, vol. 45, no. 6, pp. 2104 –2109, September 1999.

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