Source Identification and Compression of Mixture Data from Finite Observations

Afshin Abdi* and Faramarz Fekri*

*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA

***Abstract*—We consider the problem of the identification of a mixture of ergodic stationary sources from a limited number of finite-length observations of a mixture. We propose an algorithm based on Bayesian Information Criterion and Expectation Maximization to identify the sources' models and estimate the mixture parameters. Based on this algorithm, the sources' distributions can be computed and used for nearly optimal memory-assisted coding of the sequences generated by the mixture. Further, we provide upper and lower bounds on the entropy of the mixture source and show that it converges to the upper bound as the length of the sequences increases and derive the convergence rate for the per-symbol entropy of the mixture of finite memory sources.**

I. Introduction

Consider the class of all ergodic stationary stochastic processes over a finite alphabet $\mathcal{A}$. Let $x_m^n = (x_m, \dots, x_n)$, $x_i \in \mathcal{A}$, be a sequence of length $n - m + 1$, generated by a stationary source. For simplicity, when $m = 1$, we may drop it in the notations. Also, we represent an arbitrary sequence by a boldface lower case letter $\mathbf{x}$ and its length by $l(\mathbf{x})$. For an ergodic stationary source S with the stochastic process $\{X_i\}_{-\infty}^{\infty}$, the probability distribution for an arbitrary sequence of length n is denoted by $p_s(x_1^n) := p(x_1^n|S) = \Pr\{X_1^n = x_1^n|S\}$.

A stationary source has a finite memory of length k if the outcome at any time depends only on at most the last k past samples, i.e. $\Pr\{X_0 = x_0|X_{-\infty}^{-1} = x_{-\infty}^{-1}\} = \Pr\{X_0 = x_0|X_{-k}^{-1} = x_{-k}^{-1}\}$. A finite memory source can be considered as a Markov chain process of order k and being described by a set of $|\mathcal{A}|^k\,(|\mathcal{A}| - 1)$ conditional probabilities. Context tree model [1] is the extension of the Markov processes in the sense that the conditional probabilities can be determined by strings of variable lengths (the contexts), and hence allows the description of the source with much less parameters. These sources are also called tree sources [2], [3].

Let $\mathbf{S} = (S_1, S_2, \dots, S_K)$ be a set of K stationary ergodic sources over alphabet $\mathcal{A}$ with $p_j(X^n) := p(X^n|S_j)$. Assume a prior distribution $\mathbf{w} = (w_1, \dots, w_K)$ on sources. At any time instant i, a source is chosen at random and generates a data sequence $\mathbf{x}_i$ from alphabet $\mathcal{A}$. Hence, the probability distribution for the mixture is given by

$$m_{\mathbf{\Theta}}(\mathbf{x}) := p(\mathbf{x}|\mathbf{\Theta}) = \sum_{j=1}^{K} w_j\, p(\mathbf{x}|S_j) \tag{1}$$

where $\mathbf{\Theta} = (\mathbf{S}, \mathbf{w})$ is the set of mixture's parameters. Our primary goal is to find the parameters of the mixture based on a finite number, N, of observed sequences $\mathbf{x}_1, \dots, \mathbf{x}_N$.

When the data is generated by a single source and the model is known (e.g. the order of a Markov chain or the context tree), the source identification problem is simplified to the estimation of the model parameters, often via the maximum likelihood estimator (MLE). For the mixture of sources, although MLE is applicable theoretically, due to its complexity and incompleteness of data[1], the expectation maximization (EM) based methods are usually preferred [4], [5]. On the other hand, when the source model is unknown a priori, considering a more general model and applying the MLE-based methods directly, would be problematic; First, as more complex models often give higher likelihoods, MLE tends to adapt them even if the source was truly from a simpler model. Second, there might not exist enough data samples to reliably estimate parameters of a complex model. So, we need an algorithm to find the simplest model that describes the data generation reliably and more accurately.

When the observations are generated by a single source, minimum description length principle (MDL) [6], [7] and Bayesian information criterion (BIC) have been successfully applied for model selection. For a Markov source, when there exist a priori known upper bound on its order, it is shown that BIC and MDL with KT [8] and Normalized Maximum Likelihood code-lengths are strongly consistent for order estimation (cf. [9]). without such a bound, the consistency of BIC order estimator is shown in [10]. For a general tree source, if it has finite memory, [11] showed the consistency of BIC estimator even if there is no restriction on the depth of the hypothetical context trees. For arbitrary ergodic stationary source, the consistency of BIC context tree estimators are shown in [12] and [13].

In this paper, we extend the above results to the mixture of ergodic stationary sources. In section II, we show that under some conditions, our proposed algorithm can determine the source models and the mixture distribution as number of observations increases. Next, in section III, the compression and entropy of mixture source is addresses. Finally, the simulation results are provided, followed by conclusion.

This work is partially founded by Samsung Inc. through the GRO Award.

[1] i.e. it is unknown that each data sequence is generated by which source.

A. Notations

Let $\mathbb{E}_p(.)$ denotes the expectation taken with respect to the distribution given by p. For two stochastic processes given by p and q, the Kullback-Leibler divergence (KLD) is defined as $D(p\|q) = \lim_{n\to\infty} \frac{1}{n} D^{(n)}(p\|q)$ where $D^{(n)}(p\|q)$ is the KLD for sequences of length n, given by

$$D^{(n)}(p\|q) = \sum_{x^n \in \mathcal{A}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)}$$

where $\log$ represents the base-2 logarithm.

II. Mixture Characterization

Consider the mixture source $\mathbf{\Theta} = (\mathbf{S}, \mathbf{w})$ as defined earlier. Suppose that we have observed N data samples, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, where $\mathbf{x}_i$ is a sequence of length $l(\mathbf{x}_i)$ generated by an (unknown) source, independently from other samples, i.e.

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mathbf{\Theta}) = \prod_{i=1}^{N} m_{\mathbf{\Theta}}(\mathbf{x}_i) \tag{2}$$

Our goal is to use these N observation sequences and find the underlying mixture distribution; i.e. finding number of sources K, the prior distribution of sources $\mathbf{w}$, and each source's statistics $p(.|S_j)$. Note that there is no general straight-forward method to find the number of sources in the mixture. It is usually done by running the ”algorithm” for different values of K and comparing the ”results” to determine the optimum value. Hence, in the following, we assume that K is fixed and the optimization is done with respect to $\mathbf{\Theta} = (\mathbf{w}, \mathbf{S})$.

Assume that y_i is the index of the source that generated the sequence $\mathbf{x}_i$. Therefore, for $1 \leq k \leq K$,

$$p(y_i = k|\mathbf{\Theta}) = w_k\,, \tag{3a}$$

$$p(\mathbf{x}_i|y_i = k, \mathbf{\Theta}) = p(\mathbf{x}_i|S_k)\,, \tag{3b}$$

$$p(y_i = k|\mathbf{x}_i, \mathbf{\Theta}) = w_k\, p(\mathbf{x}_i|S_k)/m_{\mathbf{\Theta}}(\mathbf{x}_i) \tag{3c}$$

If the class of the sources' models were known, the EM algorithm would have been a method of choice to find the statistics by iteratively maximizing the log-likelihood of data

$$\max_{\mathbf{\Theta}} \mathcal{L}(\mathbf{w}, \mathbf{S}) = \max_{\mathbf{\Theta}} \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} w_k\, p(\mathbf{x}_i|S_k) \right)$$

Specifically, if the estimated parameters are $\mathbf{\Theta}^j = (\mathbf{w}^j, \mathbf{S}^j)$ at the j^{th} iteration of the algorithm, the updated parameters at the $(j+1)^{\text{th}}$ iteration are given by

$$w_k^{j+1} = \frac{1}{N} \sum_{i=1}^{N} p\big(y_i = k|\mathbf{x}_i, \mathbf{\Theta}^j\big)$$

and the parameters of the k^{th} source are found by maximizing $\sum_{i=1}^{N} p\big(y_i = k|\mathbf{x}_i, \mathbf{\Theta}^j\big) \log p(\mathbf{x}_i|S)$ over S.

However, when the exact source model is unknown, the maximum likelihood approach tends to consider the most complex model as it gives the highest likelihood. This over-estimation of the sources is undesirable when the length of sequences are not enough to reliably estimate model's parameters, we are interested in finding the exact source model or, as in compression applications, when it is required to send source parameters as well as data. Therefore, we need a reliable approach to find the ”simplest” model that best describes the mixture source.

We consider the context tree $\mathcal{T}$ of a stationary source, S. The tree $\mathcal{T}$ consists of all sequences $\mathbf{t}$ such that none of them is a suffix of another sequence in $\mathcal{T}$. Additionally, for all sequences $x_{-\infty}^{0}$, there exists a *unique* $\mathbf{t} \in \mathcal{T}$ such that $p_s(X_0 = x_0|X_{-\infty}^{-1} = x_{-\infty}^{-1}) = p_s(X_0 = x_0|X_{-l(\mathbf{t})}^{-1} = \mathbf{t})$. The restriction of tree to depth D, denoted by $\mathcal{T}|_D$, is defined as

$$\begin{aligned}\mathcal{T}|_D =& \{\mathbf{t} \in \mathcal{T},\ l(\mathbf{t}) \leq D\} \cup \\ & \{\mathbf{t},\ l(\mathbf{t}) = D,\ \mathbf{t} \text{ is a suffix of some } \mathbf{t}' \in \mathcal{T}\}\end{aligned}$$

We would like to find a consistent estimation of the tree $\mathcal{T}$ and its parameters $\boldsymbol{\theta} = \{\theta(\mathbf{t}, a) : \forall \mathbf{t} \in \mathcal{T}, \forall a \in \mathcal{A}\}$ where $\theta(\mathbf{t}, a) := p_s(X_0 = a|X_{-l(\mathbf{t})}^{-1} = \mathbf{t})$.

Denote the numbers of occurrences of sequence $\mathbf{t}$ followed by letter a in the observation $\mathbf{x} = x_1^n$ by

$$n_{\mathbf{x}}(\mathbf{t}, a) = \left| \left\{ i : l(\mathbf{t}) < i \leq l(\mathbf{x}),\ x_{i-l(\mathbf{t})}^{i-1} = \mathbf{t}, x_i = a \right\} \right|$$

Let $n_{\mathbf{x}}(\mathbf{t}) = \sum_{a \in \mathcal{A}} n_{\mathbf{x}}(\mathbf{t}, a)$ and $n_{\mathbf{x}} = \sum_{t \in \mathcal{T}} n_{\mathbf{x}}(\mathbf{t})$. Hence

$$p(x_1^n|S) \propto \prod_{\mathbf{t} \in \mathcal{T}} \prod_{a \in \mathcal{A}} \big(\theta(\mathbf{t}, a)\big)^{n_{\mathbf{x}}(\mathbf{t}, a)} \tag{4}$$

The maximum likelihood of sequence x^n with respect to tree $\mathcal{T}$ is defined by maximizing the right hand side of (4):

$$ML_{\mathcal{T}}(x^n) = \prod_{\mathbf{t} \in \mathcal{T}, a \in \mathcal{A}} \left(\frac{n_{\mathbf{x}}(\mathbf{t}, a)}{n_{\mathbf{x}}(\mathbf{t})} \right)^{n_{\mathbf{x}}(\mathbf{t}, a)} \tag{5}$$

with the convention that $\left(\frac{0}{0}\right)^0 = 1$.

Next, we review the results for the context tree estimation for a single source (mainly from [12] and [13]) and then we present our algorithm for the parameter estimation of a mixture of sources.

A. Single Source

In [12] and [13], the problem of estimating context tree for ergodic stationary sources has been investigated. For a tree, $\mathcal{T}$, the Bayesian Information Criterion (BIC) is defined as

$$BIC_{\mathcal{T}}(x^n) = -\log ML_{\mathcal{T}}(x^n) + \frac{(|\mathcal{A}| - 1)\,|\mathcal{T}|}{2} \log n \tag{6}$$

and the BIC estimator of the context tree is given by

$$\widehat{\mathcal{T}}_{BIC}(x^n) = \operatorname*{argmin}_{\mathcal{T}} BIC_{\mathcal{T}}(x^n) \tag{7}$$

Theorem 1 (2.11 [13])**.** *For any stationary ergodic source with context tree $\mathcal{T}_0$, for any constant integer D, $\widehat{\mathcal{T}}_{BIC}(x^n)|_D \to \mathcal{T}_0|_D$ almost surely as $n \to \infty$.*

Further, the maximum likelihood estimates $\hat{\theta}(\mathbf{t}, a) = \frac{n_{\mathbf{x}}(\mathbf{t}, a)}{n_{\mathbf{x}}(\mathbf{t})}$ converges to the source parameters $p_s(a|\mathbf{t})$.

Note that although the above theorem was stated for a single sequence whose length, n, increases, it still holds if we consider independent data samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of lengths more than D and let the number of observations, N, goes to infinity. For all $\mathbf{t} \in \mathcal{T}$ and $a \in \mathcal{A}$, let $\overline{n}(t,a) = \sum_{i=1}^{N} n_{\mathbf{x}_i}(t,a)$ and define $\overline{n}(t)$ and $\overline{n}$, similarly. Therefore,

$$ML_{\mathcal{T}}(\mathbf{x}_1,\dots,\mathbf{x}_N) = \prod_{\mathbf{t}\in\mathcal{T},a\in\mathcal{A}} \left(\frac{\overline{n}(\mathbf{t},a)}{\overline{n}(\mathbf{t})}\right)^{\overline{n}(\mathbf{t},a)}$$

$$\begin{aligned} BIC_{\mathcal{T}}(\mathbf{x}_1,\dots,\mathbf{x}_N) =& -\log ML_{\mathcal{T}}(\mathbf{x}_1,\dots,\mathbf{x}_N) \\ &+ \frac{(|\mathcal{A}|-1)\,|\mathcal{T}|}{2}\,\log\overline{n} \end{aligned} \tag{8}$$

and the BIC estimator is given by

$$\widehat{\mathcal{T}}_{BIC}(\mathbf{x}_1,\dots,\mathbf{x}_N) = \underset{\mathcal{T}}{\operatorname{argmin}}\, BIC_{\mathcal{T}}(\mathbf{x}_1,\dots,\mathbf{x}_N) \tag{9}$$

Corollary 2. *For any stationary ergodic source with context tree $\mathcal{T}_0$ and for a constant integer D, $\widehat{\mathcal{T}}_{BIC}(\mathbf{x}_1,\dots,\mathbf{x}_N)|_D = \mathcal{T}_0|_D$ eventually almost surely as $N \to \infty$, provided that $l(\mathbf{x}_i) > D$ infinitely often.*

Therefore, instead of a long sequence, we can use multiple (independently generated) relatively short sequences to estimate source's parameters.

B. Mixture of Sources

Assume that there exist K unknown sources which generated data samples $\mathbf{x}_1,\dots,\mathbf{x}_N$, with $l(\mathbf{x}_i) \le l_{max}$, for some constant l_{max}. When considering maximum-likelihood estimation of a source S_k, each sequence $\mathbf{x}_i$ affects the estimation proportional to $p(y_i = k|\mathbf{x}_i,\mathbf{\Theta})$. Therefore, it seems intuitive to use $p(y_i = k|\mathbf{x}_i,\mathbf{\Theta})\, n_{\mathbf{x}_i}(t,a)$ instead of $n_{\mathbf{x}_i}(t,a)$ for estimation of each source. However, note that at this point $p(y_i = k|\mathbf{x}_i,\mathbf{\Theta})$ cannot be computed as it requires knowing source statistics.

Lemma 3. *Assume that data samples $\mathbf{x}_i$, $1 \le i \le N$ are generated by a mixture of ergodic and stationary sources with parameters $\mathbf{\Theta} = (\mathbf{S},\mathbf{w})$. If $w_1 > 0$, then*

$$\frac{\sum_{i=1}^{N} P(y_i = 1|\mathbf{x}_i)\, n_{\mathbf{x}_i}(\mathbf{t},a)}{\sum_{i=1}^{N} P(y_i = 1|\mathbf{x}_i)\, n_{\mathbf{x}_i}(\mathbf{t})} \to p_1(a|\mathbf{t}) \tag{10}$$

almost surely as $N \to \infty$, provided that $P(y_i = 1|\mathbf{x}_i) \neq 0$ and $l(\mathbf{x}_i) > l(\mathbf{t}a)$ infinitely often.

Proof. See appendix A. ■

For an arbitrary ergodic stationary source S_k (with true context tree $\mathcal{T}_k$) and a hypothetical tree $\mathcal{T}$, for all $\mathbf{t} \in \mathcal{T}$ and $a \in \mathcal{A}$, let $\overline{n}_k(\mathbf{t},a) = \sum_{i=1}^{N} P(S_k|\mathbf{x}_i)\, n_i(\mathbf{t},a)$. Define $\overline{n}_k(\mathbf{t})$ and $\overline{n}_k$ similarly as before. The maximum likelihood and BIC for the kth source with respect to tree $\mathcal{T}$ are computed as

$$\hat{\theta}_k(\mathbf{t},a) = \frac{\overline{n}_k(\mathbf{t},a)}{\overline{n}_k(\mathbf{t})} \tag{11}$$

$$ML_{\mathcal{T}}(\mathbf{x_1},\dots,\mathbf{x}_N;k) = \prod_{\mathbf{t}\in\mathcal{T},a\in\mathcal{A}} \left(\hat{\theta}_k(\mathbf{t},a)\right)^{\overline{n}_k(\mathbf{t},a)} \tag{12}$$

$$\begin{aligned} BIC_{\mathcal{T}}(\mathbf{x}_1,\dots,\mathbf{x}_N;k) =& -\log ML_{\mathcal{T}}(\mathbf{x}_1,\dots,\mathbf{x}_N;k) \\ &+ \frac{(|\mathcal{A}|-1)\,|\mathcal{T}|}{2}\,\log\overline{n}_k \end{aligned} \tag{13}$$

and the BIC tree estimator for source S_k is given by

$$\widehat{\mathcal{T}}_k(\mathbf{x}_1,\dots,\mathbf{x}_N) = \underset{\mathcal{T}}{\operatorname{argmin}}\, BIC_{\mathcal{T}}(\mathbf{x}_1,\dots,\mathbf{x}_N;k) \tag{14}$$

Using lemma 3 with lemmas 3.1 and 3.2 from [12], it can be shown that

Theorem 4. *For a constant D, assume that $l(\mathbf{x}_i) > D$. Then $\widehat{\mathcal{T}}_k(\mathbf{x}_1,\dots,\mathbf{x}_N)|_D = \mathcal{T}_k|_D$ almost surely as $N \to \infty$. Moreover, the maximum likelihood estimates of the parameters converge to the source parameters.*

Using the above theorem to find the sources' parameters requires prior knowledge of $p(y_i|\mathbf{x}_i)$ which is not available. As such, we propose using the EM algorithm to estimate $p(y_i|\mathbf{x}_i)$ and refine the estimation of w_k, $\mathcal{T}_k$ and $\boldsymbol{\theta}_k$, for $1 \le k \le K$, at each iteration of the algorithm.

Assume that w_k^l and $S_k^l = (\mathcal{T}_k^l, \boldsymbol{\theta}_k^l)$ are the estimated parameters of the kth source at the lth iteration of the algorithm, giving the estimated mixture distribution $m_{\mathbf{\Theta}}^l(\mathbf{x})$. The estimations for the next iteration are given by

$$\widehat{P}^l(y_i = k|\mathbf{x}_i) = w_k^l \frac{P(\mathbf{x}_i|S_k^l)}{m_{\mathbf{\Theta}}^l(\mathbf{x}_i)} \tag{15}$$

We use $\widehat{P}^l(y_i = k|\mathbf{x}_i)$ to estimate $\overline{n}_k(\mathbf{t},a)$, $\overline{n}_k(\mathbf{t})$ and $\overline{n}_k$, $\forall\mathbf{t} \in \mathcal{T}$, $a \in \mathcal{A}$ and $1 \le k \le K$. Finally,

$$\hat{\theta}_k^{l+1}(\mathbf{t},a) = \frac{\overline{n}_k(\mathbf{t},a)}{\overline{n}_k(\mathbf{t})} \tag{16a}$$

$$\mathcal{T}_k^{l+1} = \underset{\mathcal{T}}{\operatorname{argmin}}\, BIC_{\mathcal{T}}(\mathbf{x}_1,\dots,\mathbf{x}_N;k) \tag{16b}$$

and the weights of sources in the mixture are updated as

$$w_k^{l+1} = \frac{1}{N}\sum_{i=1}^{N} \widehat{P}^l(y_i = k|\mathbf{x}_i) \tag{16c}$$

III. Entropy of Mixture Source

For a mixture of K stationary ergodic sources with parameters $\mathbf{\Theta} = (\mathbf{S},\mathbf{w})$, by definition, the entropy of sequences of length n is

$$H(X^n|\mathbf{S},\mathbf{w}) = -\sum_{x^n\in\mathcal{A}^n} m_{\mathbf{\Theta}}(x^n)\,\log m_{\mathbf{\Theta}}(x^n) \tag{17}$$

Using mutual information inequalities, the following upper and lower bounds can be easily obtained for all choices of $\mathbf{w}$ and $\mathbf{S}$:

$$\sum_{j=1}^{K} w_j H(X^n|S_j) \le H(X^n|\mathbf{S},\mathbf{w}) \le \sum_{j=1}^{K} w_j H(X^n|S_j) + H(\mathbf{w})$$

Note that the lower bound can be interpreted as when both encoder and decoder know the index of the source that generated x^n and use the optimum code designed for that specific source to compress x^n. Similarly, the upper bound can be seen as if the encoder knows the ”active” source, encode it with entropy $H(\mathbf{w})$ and send it to the decoder along with that source's optimum code for x^n.

To find the asymptotic behavior of the entropy of the mixture source, we need the following lemma which is an extension of Thm. 2 in [14] to the more general class of fading memory sources, $\mathcal{F}$, [2] (which includes finite memory sources as well).

Lemma 5. *Assume that K stationary sources from $\mathcal{F}$ are given. If there exists $\lambda > 0$ such that $D(p_1\|p_i) > \lambda$ for all $i \neq 1$, then, $\forall K$ and all prior distributions $\mathbf{w}$ with $w_1 > 0$,*

$$\lim_{n\to\infty} D^{(n)}(p_1\|m_{\mathbf{\Theta}}) = -\log w_1 \tag{18}$$

Proof. See appendix B. ■

The following theorem is a direct result of the above lemma.

Theorem 6. *Assume that stationary ergodic sources in $\mathcal{F}$ are separated from each other, i.e. $\exists\lambda > 0$ such that for all $i \neq j$, $D(p_i\|p_j) > \lambda$. Then, for any K and distribution $\mathbf{w} > 0$,*

$$\lim_{n\to\infty} \Big(H(X^n|\mathbf{S},\mathbf{w}) - \sum_{j=1}^{K} w_j\, H(X^n|S_j) \Big) = H(\mathbf{w}) \tag{19}$$

That is, the entropy of mixture asymptotically converges to the upper bound. Therefore, to achieve optimum compression *asymptotically*, it is sufficient to find the index of the source at the encoder, compress the data with that statistics and send both the source index and data to the decoder.

Although the per-symbol entropy $\frac{1}{n}H(X^n|\mathbf{S},\mathbf{w})$ converges to $\sum_{j=1}^{K} w_j\, H(X|S_j)$, but the convergence rate depends on the individual rates at which $\frac{1}{n}H(X^n|S_j)$ converges to $H(X|S_j)$. For finite memory sources, the convergence rate is $O(1/n)$, and hence, the per symbol mixture entropy is

$$\frac{1}{n}H(X^n|\mathbf{S},\mathbf{w}) = \sum_{j=1}^{K} w_j\, H(X|S_j) + O\big(\frac{1}{n}\big) \tag{20}$$

A. Memory-Assisted Compression of Mixture Source

In universal compression of data from single sources, it is known that the redundancy of compressing a sequence of length n from a Markov source is $\frac{d}{2}\log(n) + O(1)$ where d is the number of free (unknown) parameters. It is shown that using a memory of length m reduces the redundancy to $\frac{d}{2}\log(1+\frac{n}{m}) + o(1)$ [17] and for a mixture of sources, almost similar results are provided in [18].

One way of achieving the above performance is to estimate parameters of the mixture. Although, clustering is asymptotically optimum, it is far from optimum for short data sequences and using the estimated mixture distribution as the statistical model is closer to the entropy, i.e. if $\widehat{\mathbf{\Theta}}$ is the estimated parameter, the code-length would be approximately $m_{\widehat{\mathbf{\Theta}}}(\mathbf{x})$ instead of the optimal $m_{\mathbf{\Theta}}(\mathbf{x})$. Note that the estimated parameters can be accurate enough even for short blocks of data if the number of training data samples (N) is large enough. The redundancy of resulting memory-assisted universal compression equals $D^{(n)}(m_{\mathbf{\Theta}}\|m_{\widehat{\mathbf{\Theta}}})$. As a result of Lemma 5, it can be shown that if the estimated sources are close enough to the true ones, the main term of redundancy would be $D(\mathbf{w}\|\widehat{\mathbf{w}})$ for large n.

IV. Simulation Results

To verify our algorithm, we created four random sources over an alphabet of size 4; an i.i.d., a Markov(1), a tree source over $\mathcal{T} = \{00,\ldots,30,1,02,\ldots,32,03,\ldots,33\}$ and a Markov(2). The entropy and prior distribution of each source is given in table I. The asymptotic per symbol entropy of the mixture is 1.71. We generated $N = 10000$ sequences of length $n = 100$. Note that the number of free parameters of a Markov(2) process, 48, is comparable to the length of each sequence and hence, algorithms based on single sequence statistics for source estimation or classification usually fail. To initialize the algorithm, we simply used K random tree sources and uniform initial $\mathbf{w}_0$. The average per symbol code-length (i.e. normalized log-likelihood of the whole sequences) is considered as the cost function. Results of simulations for different values of K are given in table II. We noticed that the average code length generally decreases by increasing number of hypothetical sources, K, but for $K \geq 4$ the improvement is negligible. Therefore, by comparing the cost functions, we conclude that $K = 4$ is the optimum choice for the number of sources in the mixture. The sources' models, found through the proposed algorithm, are given in table III, which agrees with the original sources. Note that the case $K = 1$ approximately equals the universal compression using a single model. Hence, using our proposed algorithm for identification and estimation of the mixture model, the compression ratio is improved and becomes within 1.2% of the asymptotic per-symbol entropy of the mixture. We expect to see higher compression gains when the alphabet size increases, the length of each sequence decreases or data is generated from a more complex models.

TABLE I
Sources' models

Source	1	2	3	4
Model	iid	Markov(1)	Tree $\mathcal{T}$	Markov(2)
Entropy	1.4218	1.7053	1.8423	1.8236
Weight	0.2	0.3	0.1	0.4

TABLE II
Average code-length for different values of K

K	1	2	3	4	5
cost	1.9017	1.8207	1.802	1.7346	1.7345

TABLE III
Sources' models found using the proposed algorithm

Source	1	2	3	4
Model	iid	Markov(1)	Tree $\mathcal{T}$	Markov(2)
Weight	0.198	0.301	0.107	0.394

V. Conclusion

In this paper, we proposed an iterative algorithm to find the characteristics of a mixture of ergodic stationary sources based on a finite number of observations. We showed that under some conditions, as the number of observations increases, this algorithm converges almost surely to the true statistics of the

[2] For the exact definition, see [15]

mixture. Next, we considered the problem of data compression for a mixture source and derived some upper and lower bounds on the optimum coding rates and showed that as the length of data sequences increases, the entropy of the mixture converges to $\sum_{j=1}^{K} w_j\, H(X^n|S_j) + H(\mathbf{w})$, justifying the asymptotic optimality of clustering for compression. We can use the proposed algorithm to identify mixture distribution and use it for memory-assisted compression, which can result in significant gain over traditional universal compression algorithms.

Appendix A
Proof of lemma 3

First, we assume that all sequences have the same length. Let $n = l(\mathbf{x}_i) - l(\mathbf{t}a)$. (The generalization to arbitrary length sequences with $l(\mathbf{x}_i) \leq l_{max}$ for some constant l_{max} is straightforward) Hence, we can assume that $\mathbf{x}_1, \dots, \mathbf{x}_N$ are i.i.d. samples from a source with distribution $m(\mathbf{x})$. Clearly,

$$\mathbb{E}_m\left[n_{\mathbf{x}}(\mathbf{t},a)\, P(S_1|\mathbf{x})\right] = P(S_1)\, \mathbb{E}_{S_1}\left[n_{\mathbf{x}}(\mathbf{t},a)\right]$$

By strong law of large number, $\frac{1}{N}\sum_{i=1}^{N} P(S_1|\mathbf{x}_i)\, n_i(\mathbf{t},a)$ converges almost surely to $nP(S_1)\, p_1(\mathbf{t},a)$. Similarly, as $N \to \infty$, $\frac{1}{N}\sum_{i=1}^{N} P(S_1|\mathbf{x}_i) n_i(\mathbf{t})$ converges almost surely to $nP(S_1)p_1(\mathbf{t})$. Therefore,

$$\frac{\sum_{i=1}^{N} P(S_1|\mathbf{x}_i)\, n_i(\mathbf{t},a)}{\sum_{i=1}^{N} P(S_1|\mathbf{x}_i)\, n_i(\mathbf{t})} \xrightarrow{a.\,s.} p_1(a|\mathbf{t}) \;\text{ as } N \to \infty$$

Appendix B
Proof of lemma 5

Obviously, $-\log w_1 \geq \log \frac{p_1(x^n)}{m_w(x^n)}$. Therefore,

$$D^{(n)}(p_1\|m_{\mathbf{\Theta}}) = \mathbb{E}_{p_1} \log \frac{p_1(X^n)}{m_{\mathbf{\Theta}}(X^n)} \leq -\log w_1 \tag{21}$$

To prove the lemma, we need the following result from [15], for the class of fading memory sources:

Lemma 7. *$\forall P_1, P_2 \in \mathcal{F}$, if $D(P_1\|P_2) > \lambda$, for some constant $\lambda > 0$, then*

$$\lim_{n\to\infty} P_1(x^n : \frac{1}{n} \log \frac{p_1(x^n)}{p_2(x^n)} > \lambda) = 1$$

Define $B_i^{(n)} = \{x^n : \frac{1}{n} \log \frac{p_1(x^n)}{p_i(x^n)} > \lambda\}$, $B^{(n)} = \bigcap_{i=1}^{K} B_i^{(n)}$ and $C^{(n)} = (B^{(n)})^c$. From lemma 7, as n goes to infinity, $P_1(B_i^{(n)}) \to 1$ and hence $P_1(B^{(n)}) \to 1$.

$$\mathbb{E}_{p_1} \log \frac{p_1(X^n)}{m_{\mathbf{\Theta}}(X^n)} = \sum_{x^n \in B^{(n)}} p_1(x^n) \log \frac{p_1(x^n)}{m_{\mathbf{\Theta}}(x^n)} \tag{22}$$

$$+ \sum_{x^n \in C^{(n)}} p_1(x^n) \log \frac{p_1(x^n)}{m_{\mathbf{\Theta}}(x^n)} \tag{23}$$

For all $x^n \in B^{(n)}$:

$$\frac{m_{\mathbf{\Theta}}(x^n)}{p_1(x^n)} \leq w_1 + \sum_{i \neq 1} w_i\, 2^{-\lambda\, n}$$

$$\Rightarrow (22) \geq -\log(w_1 + 2^{-\lambda\, n})\, P_1(B^{(n)}) \tag{24}$$

On the other hand, since $-\log(.)$ is a convex function,

$$(23) \geq P_1(C^{(n)})\, \log P_1(C^{(n)}) \tag{25}$$

Therefore, the following lower bound is obtained for $D^{(n)}(p_1\|m_{\mathbf{\Theta}})$:

$$-\log(w_1 + 2^{-\lambda\, n})\, P_1(B^{(n)}) + P_1(C^{(n)})\, \log P_1(C^{(n)})$$

As n goes to infinity, the first term approaches $-\log(w_1)$ and the second term goes to zero. Therefore, combining with eqn. (21), we get $\lim_{n\to\infty}, D^{(n)}(p_1\|m_{\mathbf{w}}) = -\log w_1$.

Corollary 8. *Similar arguments show that $\frac{m_{\mathbf{\Theta}}(x^n)}{p_1(x^n)}$ converges to w_1 almost surely (with measure p_1) as $n \to \infty$.*

References

[1] J. Rissanen, "A universal data compression system," *Information Theory, IEEE Transactions on*, vol. 29, no. 5, pp. 656–664, Sep 1983.

[2] M. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite memory sources," *Information Theory, IEEE Transactions on*, vol. 38, no. 3, pp. 1002–1014, May 1992.

[3] F. M. J. 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.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[5] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," International Computer Science Institute, Tech. Rep. TR-97-021, 1998.

[6] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2743–2760, Oct 1998.

[7] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length." *JASA. Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

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

[9] L. Finesso, C.-C. Liu, and P. Narayan, "The optimal error exponent for markov order estimation," *Information Theory, IEEE Transactions on*, vol. 42, no. 5, pp. 1488–1497, Sep 1996.

[10] I. Csiszar and P. Shields, "The consistency of the bic markov order estimator," in *Information Theory, 2000. Proceedings. IEEE International Symposium on*, 2000, pp. 26–.

[11] A. Garivier, "Consistency of the unlimited BIC context tree estimator," *Information Theory, IEEE Transactions on*, vol. 52, no. 10, pp. 4630–4635, Oct 2006.

[12] I. Csiszar and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *Information Theory, IEEE Transactions on*, vol. 52, no. 3, pp. 1007–1016, March 2006.

[13] Z. Talata and T. Duncan, "BIC context tree estimation for stationary ergodic processes," *Information Theory, IEEE Transactions on*, vol. 57, no. 6, pp. 3877–3886, June 2011.

[14] B. S. Clarke and A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *J. Statistical Planning and Inference*, vol. 41, pp. 37–60, 1994.

[15] J. Ziv, "On classification with empirically observed statistics and universal data compression," *Information Theory, IEEE Transactions on*, vol. 34, no. 2, pp. 278–286, Mar 1988.

[16] Z. Rached, F. Alajaji, and L. Campbell, "The kullback-leibler divergence rate between markov sources," *Information Theory, IEEE Transactions on*, vol. 50, no. 5, pp. 917–921, May 2004.

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

[18] A. Beirami, M. Sardari, and F. Fekri, "Results on the optimal memory-assisted universal compression performance for mixture sources," in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, Oct 2013, pp. 890–895.