Universal Compression of Piecewise i.i.d. Sources

Badri N. Vellambi, Owen Cameron, and Marcus Hutter

Research School of Computer Science
Australian National University
Acton, ACT 2612, Australia
{badri.vellambi,u5163581,marcus.hutter}@anu.edu.au

Abstract

We study the problem of compressing piecewise i.i.d. sources, which models the practical application of jointly compressing multiple disparate data files. We establish that **universal compression** of piecewise i.i.d data is possible by modeling the data as a Markov process whose memory grows suitably with the size of the data using the Krichevsky-Trofimov (KT) estimator. The memory order is chosen large enough so that successful learning of the distribution of the each piece of the data from the corresponding contexts is possible for almost any realization of any piecewise i.i.d. data process. This is, *a priori*, a surprising result given that we are employing a *stationary* model to asymptotically optimally (model and) compress non-stationary data.

1 Introduction

Compression is an essential part of modern-day data storage and communication. It is not uncommon to imagine a situation where one jointly compresses unalike data. For example, consider the scenario illustrated in Fig. 1, where the a folder comprising of novels in different languages are compressed together. A typical data compressor will read the data from each of the files in a sequential manner and 'learn' the distribution corresponding to the source, and compress the data at hand. In this setting, it is conceivable that the statistics of the data in different files are quite dissimilar. A 'good' compressor for compressing such dissimilar data has to be: (a) **universal** in that it must learn (possibly) multiple distributions corresponding to each of the files as it sequentially reads them; and (b) **efficient** in that the redundancy of the compressor must be kept to a minimum.

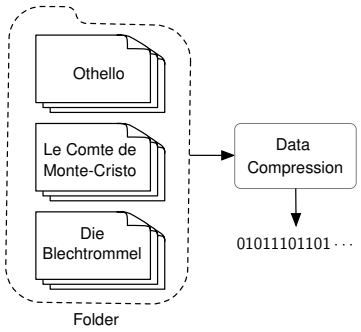

Figure 1: Compressing dissimilar files.

The notion of universal data compression was first introduced by Kolmogorov in [1]. In [2], Krichevsky and Trofimov introduced the KT estimator, and quantified the asymptotic redundancy incurred by imposing universality over the class of i.i.d. sources. Following this, Rissanen [3] presented a lower bound on the expected per-symbol redundancy, giving rise to a notion of 'optimality' for universal codes; this lower bound additionally establishes the optimality of the KT estimator in terms of the expected per-symbol redundancy.

A corresponding lower bound for variable-length piecewise i.i.d. sources (i.e., sources whose statistics change abruptly at unknown points in the data) and a strongly sequential distribution that achieves this bound was derived by Merhav [4]; this redundancy bound for the piecewise i.i.d. source contains an extra $\log n$ term per change in source parameters in addition to the terms specified by the Rissanen bound). Subsequently, Willems [5] gave efficient algorithms that achieve this bound in the limit, based on the KT distribution (both to model the boundaries between pieces and to predict for each piece). In [6], Shamir and Merhav extended [5] with even more efficient algorithms, but slightly poorer redundancy performance.

Another strand of literature, relates to universal compression for Markov processes and tree sources. Willems et al. [7] introduced the *context tree weighting (CTW)* algorithm that was shown to be optimal over the class of *time-homogeneous finite-depth tree sources*. These are a generalization of Markov sources, where the conditional distribution may depend (non-trivially) on variable-length contexts. Furthermore, the CTW distribution is efficiently computable and strongly sequential. While CTW is universal over finite-depth tree sources, and therefore finite-order Markov sources, it is unclear whether it performs well for non-stationary/time-inhomogeneous sources. An array of recent literature has recently investigated variants of CTW for wider classes of sources, using variants of KT which discount 'old' data [8], switch between a finite set of potential context trees [9] or model the boundaries between pieces [10].

In this work, we look at the problem of universal compression of piecewise i.i.d. sources. We establish the fact that the sequence of k-order KT distributions, where the order k grows logarithmically with the size n of each piece of the piecewise i.i.d. source is universal over the class of piecewise i.i.d. sources. A priori, this is surprising given that the k-KT distribution models 'time-homogenous/stationary' processes, whereas the piecewise i.i.d. source is not. However, by adapting the memory order k with n, we are able to achieve the goal of universality.

The remainder of this work is organized as follows. Section 2 presents the problem setup and the notation used. Section 3 presents the main result followed by pertinent remarks and extensions; auxiliary results required in Sec. 3 are relegated to Sec. 4.

2 Notation and Problem Setup

Random variables are denoted by upper case letters (A, B, etc), their realizations by lower case letters (a, b, etc), and their alphabets by calligraphic font letters ($\mathcal{A}$, $\mathcal{B}$, etc). A finite sequence of the first n random variables from a random process (also referred to as a source) $\{B_i\}_{i\in\mathbb{N}}$ is denoted by $B_{1:n}$, and its realization is given by $b_{1:n}$. Given $d \in \mathbb{N}$, Δ_d denotes the set of all d-dimensional categorical distributions, i.e., $\Delta_d = \{(r_1, \ldots, r_d) \in [0,1]^d : \sum_{\ell=1}^{d} r_d = 1\}$. We can now define the sources of interest.

Definition 1. *Given a finite set $\mathcal{A} = \{\mathsf{a}_1, \ldots, \mathsf{a}_d\}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \Delta_d$, let $p_{\boldsymbol{\theta}}$ to denote probability measure of the i.i.d. random process with each random variable distributed according to $\boldsymbol{\theta}$, i.e., for each $n \in \mathbb{N}$ and $x_{1:n} \in \mathcal{A}^n$,*

$$p_{\boldsymbol{\theta}}(x_{1:n}) := \theta_1^{t_1} \cdot \theta_2^{t_2} \cdots \theta_d^{t_d}, \tag{1}$$

where for $i = 1, \ldots, d$, t_i is the number of occurrences of a_i in $x_{1:n}$. As an abuse of notation, we also let $p_{\boldsymbol{\theta}}$ operate on $[0,\infty)^d$ by defining

$$p_{\boldsymbol{\theta}}(t_1, \ldots, t_{|\mathcal{A}|}) = \theta_1^{t_1} \cdot \theta_2^{t_2} \cdots \theta_d^{t_d}, \quad (t_1, \ldots, t_d) \in [0,\infty)^d. \tag{2}$$

Note that by the above notation, if $(t_1, \ldots, t_d) \in (\mathbb{N} \cup \{0\})^d$, then for some $x_{1:n} \in \mathcal{A}^n$ containing t_i occurrences of a_i for each $i = 1, \ldots, d$, $p_{\boldsymbol{\theta}}(t_1, \ldots, t_d) = p_{\boldsymbol{\theta}}(x_{1:n})$. ∎

Definition 2. *Given $m \in \mathbb{N}$, we let an m-piece i.i.d. source $\{p_{\boldsymbol{\theta}_i} : \boldsymbol{\theta}_i \in \Delta_d\}_{i=1}^m$ to be a collection of m independent i.i.d. processes $(\{X_{1,i}\}_{i\in\mathbb{N}}, \ldots, \{X_{m,i}\}_{i\in\mathbb{N}})$, where for each $j = 1, \ldots, m$, the i.i.d. process $\{X_{j,i}\}_{i\in\mathbb{N}}$ is distributed according to $p_{\boldsymbol{\theta}_j}$.* ∎

We now present a brief introduction to the Krichevsky-Trofimov Estimator.

2.1 The Krichevsky-Trofimov Estimator

To define the KT estimator, we require the following notation for counting the number of occurrences of substrings in the source data. Throughout this part, we assume that $\mathcal{A} = \{\mathsf{a}_1, \ldots, \mathsf{a}_d\}$ is the (finite) alphabet corresponding to the source of interest.

Definition 3. *Let $k, n \in \mathbb{N}$ with $k \leqslant n$ be given. For each $\boldsymbol{s} \in \mathcal{A}^k$ and $x_{1:n} \in \mathcal{A}^n$, let 'count $c_{\boldsymbol{s}}(x_{1:n})$ of $\boldsymbol{s}$ in $x_{1:n}$' denote the number of times the string $\boldsymbol{s}$ occurs as a (contiguous) substring of $x_{1:n}$, i.e., $c_{\boldsymbol{s}}(x_{1:n}) := \sum_{i=1}^{n-k+1} \mathbb{1}_{\boldsymbol{s}}(x_{i:i+k-1})$.*

Also, for each $\boldsymbol{s} \in \mathcal{A}^k$ and $x_{1:n} \in \mathcal{A}^n$, let $\mathsf{C}_{\boldsymbol{s}}(x_{1:n}) := (c_{\boldsymbol{s}\mathsf{a}_1}(x_{1:n}), \ldots, c_{\boldsymbol{s}\mathsf{a}_d}(x_{1:n}))$. For each $\boldsymbol{s}$, $\mathsf{C}_{\boldsymbol{s}}(x_{1:n})$ denotes the vector of counts of $\boldsymbol{s}\mathsf{a}$ in $x_{1:n}$ for each $\mathsf{a} \in \mathcal{A}$. ∎

The **KT distribution** for $\mathcal{A}$ is the $\mathcal{A}$-valued stochastic process distribution $\{P_{KT}^n\}_{n\in\mathbb{N}}$, where for each $n \in \mathbb{N}$, $P_{KT}^n : \mathcal{A}^n \to [0,1]$ is defined by

$$P_{KT}^n(x_{1:n}) := \Gamma(\tfrac{d}{2}) \frac{\prod_{\sigma\in\mathcal{A}} \Gamma(c_\sigma(x_{1:n}) + 1/2)}{\Gamma(1/2)^d \cdot \Gamma(n + d/2)}, \tag{3}$$

where Γ is the Gamma function. Note that the above estimator naturally yields a sequential estimator for the next symbol given by

$$P_{KT}^{n+1}(X_{n+1} = \mathsf{a} | X_{1:n} = x_{1:n}) := \frac{P_{KT}^{n+1}(X_{1:n+1} = x_{1:n}\mathsf{a})}{P_{KT}^n(X_{1:n} = x_{1:n})} = \frac{c_{\mathsf{a}}(x_{1:n}) + 1/2}{n + d/2}, \tag{4}$$

which is a regularized frequency estimator (also known as the add-$\frac{1}{2}$ estimator. As in Definition 1, we let P_{KT}^n operate over $\{(n_1, \ldots, n_d) \in \mathbb{N}^d : n_1 + \cdots + n_d = n\}$ by

$$P_{KT}^n((n_1, \ldots, n_d)) := \Gamma(d/2) \frac{\prod_{i=1}^d \Gamma(n_i + 1/2)}{\Gamma(1/2)^d \cdot \Gamma(n + d/2)}. \tag{5}$$

Note that $P_{KT}^n((n_1, \ldots, n_d)) = P_{KT}^n(x_{1:n})$ for a sequence $x_{1:n}$ that contains n_i occurrences of a_i for each $i = 1, \ldots, d$.

Given $k \in \mathbb{N}$, the **k-KT distribution** for $\mathcal{A}$ is the stochastic process distribution $\{P_{k\text{-}KT}^n\}_{n\in\mathbb{N}}$, where for each $n \in \mathbb{N}$, $P_{k\text{-}KT}^n : \mathcal{A}^n \to [0,1]$ defined by

$$P_{k\text{-}KT}^n(x_{1:n}) := \begin{cases} d^{-n} & n \leqslant k \\ d^{-k} \prod_{\boldsymbol{s}\in\mathcal{A}^k} P_{KT}^n(\mathsf{C}_{\boldsymbol{s}}(x_{1:n})) & k < n \end{cases}. \tag{6}$$

We end this section with an upper bound for the redundancy of the KT estimator [2].

Theorem 1. *Consider the class of distributions* $\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Delta_d\}$. *Given* $x_{1:n} \in \mathcal{A}^n$, *let* $p_{\theta_{ML}(x_{1:n})}$ *denote the distribution defined by*

$$\boldsymbol{\theta}_{ML}(x_{1:n}) := \arg\max_{\boldsymbol{\theta}\in\Delta_d} p_{\boldsymbol{\theta}}(x_{1:n}). \tag{7}$$

Then the coding redundancy $\rho_{p_{\boldsymbol{\theta}},P^n_{KT}}$ *(see Def. 4 below) with respect to the maximum likelihood i.i.d. distribution* $p_{\boldsymbol{\theta}_{ML}}$ *and the KT estimator* $P^n_{KT}(x_{1:n})$ *satisfies*

$$\sup_{\boldsymbol{\theta}\in\Delta_d} \rho_{p_{\boldsymbol{\theta}},P^n_{KT}}(x_{1:n}) \leqslant \rho_{p_{\boldsymbol{\theta}_{ML}(x_{1:n})},P^n_{KT}}(x_{1:n}) \leqslant \tfrac{(d-1)}{2}\log n + \log d. \tag{8}$$

2.2 Problem Statement

We begin with the specific notion of universality that we employ.

Definition 4. *Given a random process* $\{X_{\theta,i}\}_{i\in\mathbb{N}}$ *distributed according to* p_θ *from a class* $\mathfrak{C} = \{p_\theta : \theta \in \Theta\}$ *of* $\mathcal{A}$*-valued distributions, a sequence* $\{Q^n\}_{n\in\mathbb{N}}$, *where* Q^n *is a distribution over* $\mathcal{A}^n$, *is said to be* ***universal almost surely for*** $\mathfrak{C}$ *if for each* $\theta \in \Theta$,

$$\mathbb{P}\left[\lim_{n\to\infty} n^{-1}\rho_{p_\theta,Q^n}(X_{\theta,1:n}) = 0\right] = 1, \tag{9}$$

where the redundancy $\rho_{p_\theta,Q^n}(x_{1:n}) := \log \frac{p_\theta(x_{1:n})}{Q^n(x_{1:n})}$.

Our formal problem statement is as follows. Suppose that an m-piece i.i.d. source $(\{X_{1,i}\}_{i\in\mathbb{N}}, \ldots, \{X_{m,i}\}_{i\in\mathbb{N}})$ distributed according to $\{p_{\boldsymbol{\theta}_i} : \boldsymbol{\theta}_i \in \Delta_d\}_{i=1}^m$ is given. We are interested in modeling the data $\boldsymbol{X} := (X_{1,1:n}, X_{2,1:n}, \ldots, X_{m,1:n})$ by a KT distribution of appropriate order k_n that grows with n in a way that the coding redundancy $\rho_{\{p_{\boldsymbol{\theta}_i}\}_{i=1}^m,k_n\text{-}KT}$ grows sub-linearly in n, and the sequence of appropriately growing k_n-KT distributions is **universal almost surely** for $\left(\Delta_d\right)^m$. At a glance, this seems counterintuitive since the k-KT distribution is 'stationary,' unlike the piecewise i.i.d. source. However, by adapting the memory order k with n, we show that we can achieve both universality as well as optimal compression (i.e., sub-linear redundancy).

3 Main Result

Theorem 2 (Main theorem)**.** *Let* $m \in \mathbb{N}$, *and* $\mathcal{A} = \{\mathsf{a}_1, \ldots, \mathsf{a}_d\}$ *be a finite alphabet, and let* $\mathfrak{C} := (\Delta_d)^m$ *be the class of* m*-piece i.i.d. processes. For* $k_n \in o(\log n) \cap \omega(1)$, *the sequence of* k_n*-KT distributions* $\left\{P^{mn}_{k_n\text{-}KT}\right\}_{n\in\mathbb{N}}$ *for* $\mathcal{A}$ *is* ***universal almost surely*** *for the class* $\mathfrak{C}$ *for compressing equal number of symbols of each piece. In other words, for an* m*-piece i.i.d. source* $(\{X_{1,i}\}_{i\in\mathbb{N}}, \ldots, \{X_{m,i}\}_{i\in\mathbb{N}})$ *distributed according to* $\{p_{\boldsymbol{\theta}_i} : \boldsymbol{\theta}_i \in \Delta_d\}_{i=1}^m$

$$\mathbb{P}\left[\lim_{n\to\infty} n^{-1}\rho_{\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m,k_n\text{-}KT}(X_{1,1:n}, X_{2,1:n}, \ldots, X_{m,1:n}) = 0\right] = 1. \tag{10}$$

Prior to presenting the formal proof of the main result, let us intuitively argue why an appropriate k_n-KT distribution might be universal for m-piece i.i.d. sources.

On the one hand, by modeling the data using a k_n-KT distribution, where k_n is sufficiently large, we are guaranteed by the law of large numbers that the empirical frequencies of different symbols (of the alphabet) in the context of length k_n at any position is 'close' to the distribution of the piece corresponding to that position. Thus, learning form longer contexts enables one to: (a) cluster the majority of contexts into m sets such that each context cluster is likely to be seen overwhelmingly in only one piece as illustrated in the bell portion of each distribution in Fig. 2; and (b) minimize the probability of observing contexts that are seen with similar frequency in more than one piece; these contexts correspond to overlapping tail portions of the distributions in the figure. Hence, it is preferable that k_n grows with n. On the other hand, by modelling the source as a k_n^{th}-order Markov process, the number of model parameters that must be learned from the data is $d^{k_n}(d-1)$ (i.e., one for each element of $\mathcal{A}^{k_n+1}$). Hence, k_n ought to be small enough to observe the contexts sufficiently often. That is, $\frac{n}{d^{k_n}(d-1)}$ must also grow with n. The proof formalizes this argument.

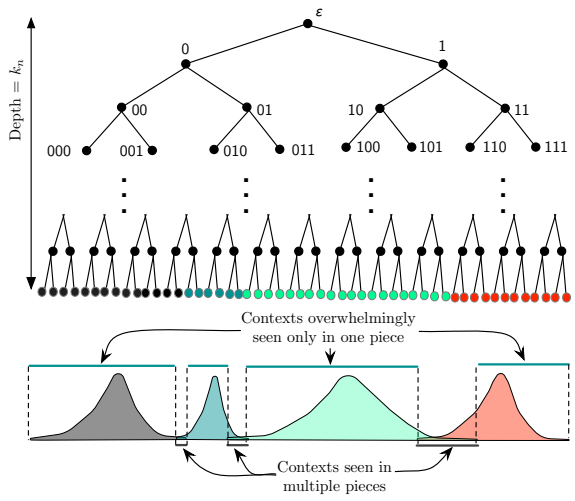

Figure 2: Intuition behind Theorem 2.

Proof. *We will start by establishing that the sequence of distributions* $\left\{P^{mn}_{k_n\text{-}KT}\right\}_{n\in\mathbb{N}}$ *offers vanishing per-symbol redundancy for most source realizations. To do that, let us start with the* ***typical*** *sequences* $\boldsymbol{A}(n,k_n,\alpha_n) = \times_{\ell=1}^{m} A_\ell(n,k_n,\alpha_n)$*, where*

$$A_\ell(n,k_n,\alpha_n) := \left\{x_{1:n} : \left|c_{\boldsymbol{s}}(x_{\ell,1:n}) - \mathbb{E}[c_{\boldsymbol{s}}(X_{\ell,1:n})]\right| \leqslant \alpha_n \mathbb{E}[c_{\boldsymbol{s}}(X_{\ell,1:n})] \text{ for all } \boldsymbol{s}\in\mathcal{A}^{k_n+1}\right\},$$

and parameters $\alpha_n > 0$*, and* k_n *will be chosen later. From Lemma 3, it follows that*

$$\Pr\left[(X_{1,1:n},\ldots,X_{m,1:n}) \notin \boldsymbol{A}(n,k_n,\alpha_n)\right] \leqslant 2\sum_{\ell=1}^{m}\sum_{\boldsymbol{s}:p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s})>0} e^{\frac{-(\alpha_n\mathbb{E}[c_{\boldsymbol{s}}(X_{\ell,1:n})])^2}{2k_n^2(n-k_n+1)}}. \tag{11}$$

Now, pick $\boldsymbol{x} := (x_{1,1:n},\ldots,x_{m,1:n}) \in \boldsymbol{A}(n,k_n,\alpha_n)$*. As a shorthand, let* $\boldsymbol{x}_\ell = x_{\ell,1:n}$*,* $\ell = 1,\ldots,m$*. Then, the redundancy* $\rho_{\{p_{\boldsymbol{\theta}_i}\}_{i=1}^m,k_n\text{-}KT}(\boldsymbol{x})$ *is given by*

$$\rho_{\{p_{\boldsymbol{\theta}_i}\}_{i=1}^m,k_n\text{-}KT}(\boldsymbol{x}) = \log\frac{\prod_{\ell=1}^m p_{\boldsymbol{\theta}_\ell}(\boldsymbol{x}_\ell)}{d^{-k_n}\prod_{\boldsymbol{s}\in\mathcal{A}^{k_n}} P^{mn}_{KT}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}))} \tag{12}$$

$$= \underbrace{\sum_{\boldsymbol{s}\in\mathcal{A}^{k_n}} \log\frac{\prod_{\ell=1}^m p_{\boldsymbol{\theta}_\ell}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}{P^{mn}_{KT}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}))}}_{=:\mathsf{T}_1} + \underbrace{\log\left(d^{k_n}\prod_{\ell=1}^m p_{\boldsymbol{\theta}_\ell}(x_{\ell,1:k_n})\right)}_{=:\mathsf{T}_2} \tag{13}$$

Since term $\mathsf{T}_2 \leqslant k_n \log d$*, the per-symbol redundancy can be made to vanish with* n*, if*

$$\boxed{\lim_{n\to\infty} n^{-1}k_n = 0 \Leftrightarrow k_n = o(n)} \tag{14}$$

Now, let us define a Bayesian mixture of the m i.i.d. distributions as follows.

$$\boldsymbol{\lambda}_{\boldsymbol{s}} := \begin{cases} \frac{\sum_{\ell=1}^{m} p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s})\boldsymbol{\theta}_\ell}{\sum_{\ell=1}^{m} p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s})} & \min\{p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s}) : \ell = 1, \dots, m\} > 0 \\ d^{-1} & \textit{otherwise} \end{cases} . \tag{15}$$

Now, we can introduce the above mixture distribution into T_1 as follows.

$$\mathsf{T}_1 = \underbrace{\sum_{\boldsymbol{s} \in \mathcal{A}^{k_n}} \log \frac{\prod_{\ell=1}^{m} p_{\boldsymbol{\theta}_\ell}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}{\prod_{\ell=1}^{m} p_{\boldsymbol{\lambda}_s}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}}_{=:\mathsf{T}_{1,1}} + \underbrace{\sum_{\boldsymbol{s} \in \mathcal{A}^{k_n}} \log \frac{\prod_{\ell=1}^{m} p_{\boldsymbol{\lambda}_s}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}{P_{KT}^{mn}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}))}}_{=:\mathsf{T}_{1,2}} . \tag{16}$$

Now, the counts of each string $\boldsymbol{s} \in \mathcal{A}^{k_n}$ in each piece $\boldsymbol{x}_\ell$ and $\boldsymbol{x}$ are related by

$$\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}) = \sum_{\ell=1}^{m} \mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell) + \sum_{\ell=2}^{m} \mathsf{C}_{\boldsymbol{s}}(x_{\ell-1,n-k_n+1:n}, x_{\ell,1:k_n}). \tag{17}$$

Thus, it follows that

$$\prod_{\boldsymbol{s}} \frac{P_{KT}^{mn}(\sum_{\ell=1}^{m} \mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}{P_{KT}^{mn}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}))} = \prod_{\boldsymbol{s}} \left(\frac{\prod_{i=1}^{d} \Gamma(\sum_{\ell=1}^{m} c_{\boldsymbol{s}\mathsf{a}_i}(\boldsymbol{x}_\ell) + \frac{1}{2})}{\Gamma(\sum_{\ell=1}^{m} \|\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell)\|_1 + \frac{d}{2})} \cdot \frac{\Gamma(\|\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x})\|_1 + \frac{d}{2})}{\prod_{i=1}^{d} \Gamma(c_{\boldsymbol{s}\mathsf{a}_i}(\boldsymbol{x}) + \frac{1}{2})} \right) \tag{18}$$

$$\overset{(17)}{\leqslant} \prod_{\boldsymbol{s}} (\|\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x})\|_1 + d/2)^{\|\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x})\|_1 - \sum_{\ell=1}^{m} \|\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell)\|_1} \leqslant (mn + d/2)^{mk_n} . \tag{19}$$

Now, using (8) and (19), we can bound $\mathsf{T}_{1,2}$ as follows.

$$\mathsf{T}_{1,2} = \sum_{\boldsymbol{s} \in \mathcal{A}^{k_n}} \log \frac{\prod_{\ell=1}^{m} p_{\boldsymbol{\lambda}_s}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}{P_{KT}^{mn}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}))} \tag{20}$$

$$\overset{(a)}{=} \sum_{\boldsymbol{s} \in \mathcal{A}^{k_n}} \log \frac{p_{\boldsymbol{\lambda}_s}(\sum_{\ell=1}^{m} \mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}{P_{KT}^{mn}(\sum_{\ell=1}^{m} \mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))} + \log \left(\prod_{\boldsymbol{s}} \frac{P_{KT}^{mn}(\sum_{\ell=1}^{m} \mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}{P_{KT}^{mn}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}))} \right) \tag{21}$$

$$\overset{(8),(19)}{\leqslant} d^{k_n} \left(\tfrac{(d-1)}{2} \log(mn) + \log d \right) + mk_n \log (mn + d/2) . \tag{22}$$

As before, the above terms grow sub-linearly in n provided

$$\boxed{d^{k_n} \left(\tfrac{(d-1)}{2} \log(mn) + \log d \right) + mk_n \log (mn + d/2) = o(n)} \tag{23}$$

We are now left to bound $\mathsf{T}_{1,1}$. To do so, we proceed as follows.

$$\mathsf{T}_{1,1} = \sum_{\boldsymbol{s} \in \mathcal{A}^{k_n}} \sum_{\ell=1}^{m} \log \frac{p_{\boldsymbol{\theta}_\ell}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))}{p_{\boldsymbol{\lambda}_s}(\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell))} \leqslant \sum_{\boldsymbol{s} \in \mathcal{A}^{k_n}} \sum_{\ell=1}^{m} \log \frac{p_{\boldsymbol{\theta}_\ell}(\mathbb{E}[\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell)](1-\alpha_n))}{p_{\boldsymbol{\lambda}_s}(\mathbb{E}[\mathsf{C}_{\boldsymbol{s}}(\boldsymbol{x}_\ell)](1+\alpha_n))} \tag{24}$$

$$= \sum_{\boldsymbol{s} \in \mathcal{A}^{k_n}} \sum_{\ell=1}^{m} \sum_{i=1}^{d} \left(\log \theta_{\ell,i}^{(n-k_n)p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s})\theta_{\ell,i}(1-\alpha_n)} - \log \lambda_{\boldsymbol{s},i}^{(n-k_n)p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s})\theta_{\ell,i}(1+\alpha_n)} \right) \tag{25}$$

$$= (n-k_n)\Bigg[(1+\alpha_n)\sum_{\boldsymbol{s}\in\mathcal{A}^k}\sum_{\ell=1}^{m} p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s}) D_{KL}\left(\boldsymbol{\theta}_\ell||\boldsymbol{\lambda}_{\boldsymbol{s}}\right) + 2\alpha_n \sum_{\ell=1}^{m} H(\boldsymbol{\theta}_\ell)\Bigg], \tag{26}$$

where D_{KL} refers to the Kullback-Leibler divergence. Upon using Lemma 1 to bound the KL divergence term, we see that the expression in (26) *grows sub-linearly in n, if*

$$\boxed{\alpha_n = o(1) \text{ and } e^{-k_n\epsilon^2\delta} = o(1).} \tag{27}$$

Note that ϵ, δ are constants that depend only on $\{\boldsymbol{\theta}_\ell : \ell = 1, \ldots, m\}$. Now, suppose that $\{k_n\}_{n\in\mathbb{N}}$ is chosen so that:

$$\lim_{n\to\infty} k_n = \infty \text{ and } k_n < \beta \log n \quad \text{for} \quad 0 < \beta < \tfrac{-1}{2\log\delta}, \tag{28}$$

and α_n is chosen to be a suitable function that is $o(1)$ (say, $\alpha_n = 1/\log\log n$). Then, the three boxed constraints given in (14), (23), *and* (27) *are satisfied. Further, for this choice of parameters, it can be shown that for **any** $\boldsymbol{x} \in \boldsymbol{A}(n, k_n, \alpha_n)$, the redundancy $\rho_{\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m, k_n\text{-}KT}(\boldsymbol{x}) < n^{1-\eta}$ for sufficiently large n and for a suitable $\eta > 0$, and that*

$$\lim_{n\to\infty} \sum_{\ell=1}^{m} \sum_{\boldsymbol{s}: p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s})>0} e^{\frac{-(\alpha_n \mathbb{E}[c_{\boldsymbol{s}}(X_{\ell,1:n})])^2}{2k_n^2(n-k_n+1)}} = 0. \tag{29}$$

Thus, we are guaranteed that

$$\lim_{n\to\infty} \Pr\left[n^{-1}\rho_{\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m, k_n\text{-}KT}(X_{1,1:n}, \ldots, X_{m,1:n}) < n^{-\eta}\right] = 1. \tag{30}$$

In addition to (29), *for this selection of parameters, we can also establish that*

$$\sum_{n=1}^{\infty}\sum_{\ell=1}^{m} \sum_{\boldsymbol{s}: p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s})>0} e^{\frac{-(\alpha_n \mathbb{E}[c_{\boldsymbol{s}}(X_{\ell,1:n})])^2}{2k_n^2(n-k_n+1)}} < \infty. \tag{31}$$

Then, by the Borel-Cantelli lemma [11], we are assured that

$$\Pr\left[\lim_{n\to\infty} n^{-1}\rho_{\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m, k_n\text{-}KT}(X_{1,1:n}, \ldots, X_{m,1:n}) = 0\right] = 1, \tag{32}$$

which is precisely the universality result we aimed to establish. ■

We conclude this section with a few remarks and observations.

- The main result above can be straightforwardly extended to unequal piece-length setting.
- The proof is critically hinged on the fact that the typical set for the m-piece i.i.d. source covers most of the probability measure, and for each typical sequence, the redundancy offered by the k-KT distribution grows only sub-linearly. The proof does not provide an upper bound for the redundancy of the KT distribution for non-typical source realizations; however, it seems plausible that the universality holds in the stronger worst-case sense, i.e., for **all** finite-length sequences over $\mathcal{A}$, as opposed to only the asymptotic setting discussed above; however, the worst-case result remains open.

- The sequence of KT distributions are not necessarily **strongly sequential**, i.e., the following equality need not hold.

$$P_{KT}^{mn}(x_{1:n},\ldots,x_{m,1:n}) = \sum_{x_{1:m,n+1}} P_{KT}^{m(n+1)}(x_{1,1:n+1},\ldots,x_{m,n+1}) \tag{33}$$

The potential inconsistency arises because of the fact that the 'model' parameter k_n increases with the size n of the data. However, the infinite-depth context-tree weighting algorithm [7] allows us to devise a sequence of distributions that is strongly sequential. The universality of the CTW distributions follows from the fact that the CTW is a mixture of tree-source models of which one corresponds to the complete tree source $T_{k_n}^*$ of depth k_n; the distribution corresponding to this complete tree $T_{k_n}^*$ is precisely the k_n-KT distribution $P_{k_n\text{-}KT}^n$. That is,

$$P_{CTW}(x_{1:n}) = \sum_T \frac{P_T(x_{1:n})}{2^{\Gamma(T)}} \geqslant \frac{P_{k_n\text{-}KT}^n(x_{1:n})}{2^{\Gamma(T_{k_n}^*)}} = \frac{P_{k_n\text{-}KT}^n(x_{1:n})}{2^{2^{k_n+1}-1}}, \tag{34}$$

where $\Gamma(T)$ is the weight of the CTW mixture associated with the finite context tree T. Consequently, for k_n satisfying (28),

$$\frac{1}{n}\log\frac{P_{k_n\text{-}KT}^n(x_{1:n})}{P_{CTW}(x_{1:n})} \leqslant \frac{2^{k_n+1}-1}{n} \overset{n\to\infty}{\longrightarrow} 0. \tag{35}$$

- Although we have established that the k-KT distribution is universal for the class of piecewise i.i.d processes, the proof does not establish if the compression is uniform, i.e., the rate of convergence of the per-letter redundancy to zero could potentially vary with the actual distribution of various pieces of the data. For example, the required data size to achieve a certain per-symbol redundancy when the distributions of different pieces are 'close' might be larger than that required to achieve the same per-symbol redundancy when the distributions of pieces are 'very different.'

4 Auxiliary Results

Lemma 1. *Let $\mathcal{A} = \{\mathsf{a}_1,\ldots,\mathsf{a}_d\}$, $\hat{m}\in\mathbb{N}$, and $\eta_1\ldots,\eta_{\hat{m}}\in\mathbb{N}$. Let for $\ell=1,\ldots,\hat{m}$, and $k\in\mathbb{N}$, $p_{k,\boldsymbol{\theta}_\ell}$ denote the multinomial distribution with parameters n and $\boldsymbol{\theta}_\ell\in\mathbb{R}^{+d}$. Suppose that the $\hat{m}$ distributions are distinct, i.e., $\boldsymbol{\theta}_\ell=\boldsymbol{\theta}_{\ell'}$ iff $\ell=\ell'$. Let for $\boldsymbol{s}\in\mathcal{A}^k$,*

$$\boldsymbol{\lambda}_{\boldsymbol{s}} := \sum_{\ell=1}^{\hat{m}} \frac{\eta_\ell\, p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s})}{\sum_{j=1}^{\hat{m}}\eta_j p_{k,\boldsymbol{\theta}_j}(\boldsymbol{s})}\boldsymbol{\theta}_\ell, \tag{36}$$

if $p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s})>0$ for some $1\leqslant\ell\leqslant\hat{m}$, and $\lambda_{\boldsymbol{s}}=[1/d,\ldots,1/d]$ otherwise. Let μ be the smallest positive element of $\{|\theta_{i,j}-\theta_{i',j}|:i,i'\in\{1,\ldots,\hat{m}\},j\in\{1,\ldots,d\}\}$, and let δ be the smallest positive element of $\{\theta_{i,j}:i\in\{1,\ldots,\hat{m}\},j\in\{1,\ldots,d\}\}$. Let $\mu/2>\epsilon>0$, and $m:=\sum_{\ell=1}^{\hat{m}}\eta_\ell$. Then,

$$\Delta := \sum_{\boldsymbol{s}\in\mathcal{A}^k}\sum_{\ell=1}^{\hat{m}} \eta_\ell p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s}) D_{KL}\left(\boldsymbol{\theta}_\ell||\boldsymbol{\lambda}_{\boldsymbol{s}}\right) \leqslant mh(2mde^{-k\epsilon^2\delta}) + 4dm^2e^{-k\epsilon^2\delta}\log m, \tag{37}$$

Proof. *Let* $\mathsf{S} := \{\boldsymbol{s} \in \mathcal{A}^k : p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s}) > 0 \text{ for some } 1 \leqslant \ell \leqslant \hat{m}\}$. *Define a jointly correlated random variables* M *and* $X_{1:k}$ *over* $\{1, \ldots, \hat{m}\}$ *and* $\mathcal{A}^k$, *respectively, by*

$$p_{M,X_{1:k}}(\ell, \boldsymbol{x}) = \frac{\eta_\ell}{m} p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{x}), \quad \ell \in \{1, \ldots, \hat{m}\}, \boldsymbol{x} \in \mathcal{A}^k. \tag{38}$$

Thus, $X_{1:k}$ *is distributed according to* $\sum_{j=1}^{\hat{m}} \frac{\eta_j}{m} p_{k,\boldsymbol{\theta}_j}$. *Then,*

$$\begin{aligned}
\Delta &:= \sum_{\boldsymbol{s} \in \mathcal{A}^k} \sum_{\ell=1}^{\hat{m}} \eta_\ell \, p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s}) D_{KL}\left(\boldsymbol{\theta}_\ell || \boldsymbol{\lambda}_{\boldsymbol{s}}\right) = \sum_{\boldsymbol{s} \in \mathsf{S}} \sum_{\ell=1}^{\hat{m}} \eta_\ell \, p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s}) D_{KL}\left(\boldsymbol{\theta}_\ell || \boldsymbol{\lambda}_{\boldsymbol{s}}\right) && (39)\\
&\leqslant \sum_{\boldsymbol{s} \in \mathsf{S}} \sum_{\ell=1}^{\hat{m}} \eta_\ell \, p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s}) \left(\sum_{i=1}^{d} \theta_{\ell,i} \log \frac{\sum_{j=1}^{m} \eta_j \, p_{k,\boldsymbol{\theta}_j}(\boldsymbol{s})}{\eta_\ell \, p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s})} \right) && (40)\\
&= \sum_{\boldsymbol{s} \in \mathsf{S}} \sum_{\ell=1}^{\hat{m}} \eta_\ell \, p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s}) \log \frac{\sum_{j=1}^{m} \eta_j \, p_{k,\boldsymbol{\theta}_j}(\boldsymbol{s})}{\eta_\ell \, p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s})} && (41)\\
&= m \sum_{\boldsymbol{s} \in \mathsf{S}} \sum_{\ell=1}^{\hat{m}} \frac{\eta_\ell \, p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s})}{m} \log \frac{\sum_{j=1}^{\hat{m}} \frac{\eta_j}{m} p_{k,\boldsymbol{\theta}_j}(\boldsymbol{s})}{\frac{\eta_\ell}{m} p_{k,\boldsymbol{\theta}_\ell}(\boldsymbol{s})} && (42)\\
&= m \left[H(M) + H(X_{1:k}|M) - H(X_{1:k}) \right] = m H(M|X_{1:k}). && (43)
\end{aligned}$$

Let for $\ell = 1, \ldots, \hat{m}$, $\mathsf{W}_{k,\epsilon}[\boldsymbol{\theta}_\ell]$ *denote the set of strongly typical sequences defined by*

$$\mathsf{W}_{k,\epsilon}[\boldsymbol{\theta}_\ell] := \left\{ x_{1:k} \in \mathcal{A}^k : \left| k^{-1} c_{\mathsf{a}_i}(x_{1:n}) - \theta_{\ell,i} \right| \leqslant \epsilon \theta_{\ell,i} \text{ for all } i = 1, \ldots, d \right\}. \tag{44}$$

It can be easily verified that the sets $\{\mathsf{W}_{k,\epsilon}[\boldsymbol{\theta}_\ell]\}_{\ell=1}^{\hat{m}}$ *are disjoint. Now, let us define* $Z := \sum_{\ell=1}^{m} \ell \cdot [\![X_{1:k} \in \mathsf{W}_{k,\epsilon}[\boldsymbol{\theta}_\ell]]\!]$. *Then, by [12, (1.9)], we have* $1 \leqslant \ell \leqslant \hat{m}$,

$$\Pr\left[Z = \ell \mid M = \ell \right] = \Pr\left[X_{1:k} \in \mathsf{W}_{k,\epsilon}[\boldsymbol{\theta}_\ell] \mid M = \ell \right] \geqslant 1 - 2de^{-k\epsilon^2\delta}. \tag{45}$$

Using Bayes' theorem and (38), *it can be shown that for each* $\ell = 1, \ldots, m$,

$$\Pr[M = \ell | Z = \ell] \geqslant \frac{(1 - 2de^{-k\epsilon^2\delta}) \Pr[M = \ell]}{\Pr[M = \ell] + \Pr[M \neq \ell] \cdot (2de^{-k\epsilon^2\delta})} > 1 - 2dme^{-k\epsilon^2\delta}. \tag{46}$$

Using the above bounds with [13, Thm. 1], we can show that

$$H(M|Z = \ell) \leqslant \begin{cases} h(2mde^{-k\epsilon^2\delta}) + 2mde^{-k\epsilon^2\delta} \log(m-1) & \ell \neq 0 \\ \log m & \ell = 0 \end{cases}. \tag{47}$$

Note that by definition, Z *is a function of* $X_{1:k}$ *and hence,* Z *and* M *are conditionally independent given* $X_{1:k}$. *Then, by the data processing inequality, it follows that*

$$H(M|X_{1:k}) \leqslant H(M|Z) \leqslant h(2mde^{-k\epsilon^2\delta}) + 4mde^{-k\epsilon^2\delta} \log m. \tag{48}$$

Lastly, combining the above with (43) *completes the claim.* ∎

Lemma 2. *Suppose $X_1, \dots, X_n$ are i.i.d. taking values in $\mathcal{A}$ according to the distribution $p_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Delta_{|\mathcal{A}|}$. Let $k \in \mathbb{N}$ and $\boldsymbol{s} \in \mathcal{A}^k$ such that $p_{\boldsymbol{\theta}}(\boldsymbol{s}) > 0$. Then the following concentration bound holds for the count of $\boldsymbol{s}$ in $X_{1:n}$.*

$$\mathbb{P}\left[\left|c_{\boldsymbol{s}}(X_{1:n}) - \mathbb{E}[c_{\boldsymbol{s}}(X_{1:n})]\right| > \epsilon\right] \leqslant 2e^{\frac{-\epsilon^2}{2k^2(n-k+1)}}. \tag{49}$$

Proof. *The claim immediately follows from McDiarmid's inequality [14].* ∎

Lemma 3. *Let $\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m$ be m distributions over $\mathcal{A} = \{1, \dots, d\}$ such that $p_{\boldsymbol{\theta}_\ell}(i) = \theta_{\ell,i}$ for $1 \leqslant l \leqslant m$ and $1 \leqslant i \leqslant d$. Let $\{(X_{\ell,1}, \dots, X_{\ell,n})\}_{\ell=1}^m$ be mn independent random variables with $X_{\ell,i} \sim p_{\boldsymbol{\theta}_\ell}$ for $1 \leqslant i \leqslant n$ and $1 \leqslant \ell \leqslant m$. Lastly, for $1 \leqslant \ell \leqslant m$, let*

$$A_\ell(n,k,\alpha) := \left\{x_{1:n} : \left|c_{\boldsymbol{s}}(x_{\ell,1:n}) - \mathbb{E}[c_{\boldsymbol{s}}(X_{\ell,1:n})]\right| \leqslant \alpha \mathbb{E}[c_{\boldsymbol{s}}(X_{\ell,1:n})] \text{ for all } \boldsymbol{s} \in \mathcal{A}^{k+1}\right\}.$$

Then,

$$\Pr\left[(X_{1,1:n}, \dots, X_{m,1:n}) \in \bigtimes_{l=1}^m A_l(n,k,\alpha)\right] \geqslant 1 - 2\sum_{\ell=1}^m \sum_{\boldsymbol{s}: p_{\boldsymbol{\theta}_\ell}(\boldsymbol{s})>0} e^{\frac{-\alpha^2(\mathbb{E}[c_{\boldsymbol{s}}(X_{\ell,1:n})])^2}{2k^2(n-k+1)}}. \tag{50}$$

Proof. *The claim follows directly from Lemma 2, and the union bound.* ∎

References

[1] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *International Journal of Computer Mathematics*, vol. 2, no. 1-4, pp. 157–168, 1968.

[2] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding." *IEEE Trans. Information Theory*, vol. 27, no. 2, pp. 199–206, 1981.

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

[4] N. Merhav, "On the minimum description length principle for sources with piecewise constant parameters," *IEEE Trans. Inf. Theory*, vol. 39, pp. 1962–1967, Nov. 1993.

[5] F. M. J. Willems, "Coding for a binary independent piecewise-identically-distributed source," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2210–2217, 1996.

[6] G. I. Shamir and N. Merhav, "Low complexity sequential lossless coding for piecewise stationary memoryless sources," *IEEE Trans. Inf. Theory*, 1999.

[7] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, pp. 653–664, 1995.

[8] A. O'Neill, M. Hutter, W. Shao, and P. Sunehag, "Adaptive context tree weighting," *CoRR*, vol. abs/1201.2056, 2012. [Online]. Available: http://arxiv.org/abs/1201.2056

[9] J. Veness, K. S. Ng, M. Hutter, and M. H. Bowling, "Context tree switching," *CoRR*, vol. abs/1111.3182, 2011. [Online]. Available: http://arxiv.org/abs/1111.3182

[10] J. Veness, M. White, M. Bowling, and A. György, "Partition tree weighting," in *2013 Data Compression Conference, Snowbird, UT, USA, March*, 2013, pp. 321–330.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications.* Wiley, January 1968, vol. 1.

[12] G. Kramer, "Topics in multi-user information theory," *Found. Trends Commun. Inf. Theory*, vol. 4, no. 4-5, pp. 265–444, 2007.

[13] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 259–266, Jan 1994.

[14] C. McDiarmid, *On the method of bounded differences*, ser. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989, pp. 148–188.