Universal Compression of Piecewise i.i.d. Sources

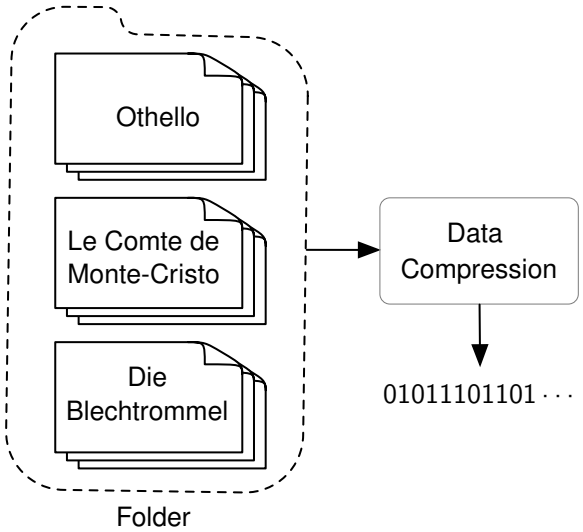
Badri N. Vellambi, Owen Cameron, **Marcus Hutter**
Australian National University

March 29, 2018

Data Compression Conference (DCC 2018)
Snowbird, UT

Introduction

- Consider the following applications:
  - jointly compressing a folder of images, audio and text files
  - compression of radio signals [speech + music]



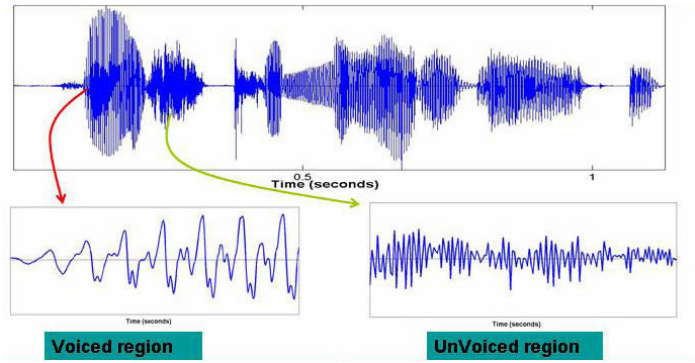


Figure: Frey Lab, Uni Toronto

- Data in the applications can be modelled by piecewise locally stationary sources
- Can we design **efficient universal** compression schemes for such sources?
  - Efficient: sub-linear redundancy = per-symbol redundancy vanishes asymptotically

Universal Data Compression: Relevant Results

- Memoryless Sources
  - KT Estimator: compression scheme universal over the class of i.i.d. sources [Krichevsky-Trofimov, 1981]
  - Lower bound on the expected per-symbol redundancy for universal codes over i.i.d. sources [Rissanen, 1984]
    - Establishes the optimality of the KT-estimator
- Variable-order Markov (Tree) Sources
  - Context-tree Weighting: optimal compression scheme universal over class of tree sources [Willems et al., 1995]
- Piecewise Stationary Sources
  - Lower bound on the redundancy of universal compression of piecewise i.i.d sources [Merhav, 1993]
  - Piecewise KT-estimators weighted based on all possible source transition patterns [Willems, 1995; Shamir-Merhav, 1999]
  - A weighted piecewise KT-estimators with CTW-style approach to partition weights [Veness et al. 2012]

Main Result of this paper: Efficient universal compression of piecewise i.i.d sources is achievable by appropriately modelling the source as a Markov process of sufficiently high order using the KT estimator.

The (Binary) Kirchevsky-Trofimov (KT) Estimator

- Let p_θ denote the Bernoulli distribution with parameter $\theta \in [0,1]$.
- The probability that the KT-estimator assigns to a sequence $x_{1:n} \in \{0,1\}^n$ with a zeros and $b = n - a$ is

$$P_{\mathrm{KT}}^n(x_{1:n}) = P_e(a,b) := \int_0^1 \underbrace{\frac{1}{\pi\sqrt{\theta(1-\theta)}}}_{Dir(\frac{1}{2})\text{prior}} \theta^a (1-\theta)^b \mathrm{d}\theta = \frac{\Gamma(a+\frac{1}{2})\Gamma(b+\frac{1}{2})}{\pi\Gamma(n+1)}$$

$$P_{\mathrm{KT}}^n(X_{n+1} = 0 | X_{1:n} = x_{1:n}) = \frac{a+\frac{1}{2}}{n+1} \quad \text{(Similar to Laplace estimator)}$$

- It was shown in [Krichevsky-Trofimov, 1981] that

$$P_{\mathrm{KT}}^n(x_{1:n}) \geq \frac{1}{2\sqrt{n}} \max_{\theta\in[0,1]} p_\theta(x_{1:n})$$

- The KT distribution is **universal** in the class of i.i.d. sources; the redundancy ρ is upper bounded by

$$\rho_{p_\theta, P_{\mathrm{KT}}^n}(x_{1:n}) := \log_2 \frac{p_\theta(\cdot)}{P_{\mathrm{KT}}^n(\cdot)} \leq \frac{1}{2}\log_2 n + 1,$$

i.e., the additional cost of modelling n realizations of an i.i.d. Bernoulli source by the KT distribution grows at most logarithmically in n independent of the realization and the Bernoulli source parameter.

The (Binary) kKT Distribution

- The KT-distribution can be extended to model a k^{th}-order Markov source for $k \geq 0$.
- Given $x_{1:n} \in \{0,1\}^n$, let for context $\boldsymbol{s} \in \{0,1\}^k$, let $a_{\boldsymbol{s}}$ and $b_{\boldsymbol{s}}$ denote the number of occurrences of $\boldsymbol{s}0$ and $\boldsymbol{s}1$, respectively.
- The kKT distribution over $\{0,1\}^n$ is given by

$$P^n_{k\text{KT}}(x_{1:n}) := \begin{cases} 2^{-n} & n \leq k \\ 2^{-k} \prod_{\boldsymbol{s} \in \{0,1\}^k} P_{\text{KT}}(a_{\boldsymbol{s}}, b_{\boldsymbol{s}}) & n > k \end{cases}$$

- $P^n_{0KT} = P^n_{\text{KT}}$.
- The kKT distribution is **universal** in the class of k^{th}-order Markov processes; the redundancy $\rho_{p_{\{\theta_{\boldsymbol{s}}\}}, P^n_{k\text{KT}}}(\cdot)$ of modelling a k^{th}-order Markov process $p_{\{\theta_{\boldsymbol{s}}\}}$ with true parameters $\{\theta_{\boldsymbol{s}} : \boldsymbol{s} \in \{0,1\}^k\}$ with the kKT distribution can be upper bounded by

$$\rho_{p_{\{\theta_{\boldsymbol{s}}\}}, P^n_{k\text{KT}}}(x_{1:n}) \leq 2^{k-1} \log_2 n + 1,$$

  i.e., the additional cost of modelling n realizations of a k^{th}-order Markov source by the kKT distribution grows no more than logarithmically in n independent of the realization and the source parameters.

Universal Compression of Piecewise i.i.d. Sources

- Existing universal compression approaches [Willems, 1995; Shamir-Merhav, 1999; Veness et al. 2012]:
  - Given $x_{1:n}$, model data as a weighted average of piecewise i.i.d sources over all possible partitions of $\{1, \ldots, n\}$.

$$Q^n(x_{1:n}) = \sum_{T \in \mathcal{T}_n} w(T) Q^n(x_{1:n}|T),$$

where

$$\begin{aligned}
\mathcal{T}_n &= \text{set of all partitions of } \{1, \ldots, n\} \\
&= \{(t_1, \ldots, t_k) : 1 \leq t_1 < t_2 < \cdots < t_k \leq n \text{ for some } k \in \mathbb{N}\} \\
T &= \text{partition indicating where the source parameters change} \\
Q^n(x_{1:n}|T) &= \text{product of KT-estimators for each partition segment} \\
w(\cdot) &= \text{Probability distribution over } \mathcal{T}_n.
\end{aligned}$$

- Schemes differ in the weights for each partition, and the way partitions are combined.
- Runtime: n^3 [Willems, 1995]; n^2 [Shamir-Merhav, 1999]; $n \log n$ [Veness et al. 2012].

Our Approach

- ***m*-piece source:** a tuple of m independent i.i.d. sources $(\{X_{1,i}\}_{i\in\mathbb{N}}, \ldots, \{X_{m,i}\}_{i\in\mathbb{N}})$, where for each $j = 1, \ldots, m$, the i.i.d. source $\{X_{j,i}\}_{i\in\mathbb{N}}$ is distributed according to some distribution p_{θ_j} over a finite alphabet $\mathcal{A}$.
- **Universality**: Given:
  (1) a class $\mathfrak{C} = \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in [0,1]^{|\mathcal{A}|-1}\}$ of distributions over $\mathcal{A}$; and
  (2) a source (random process) $\{X_{\theta,i}\}_{i\in\mathbb{N}}$ distributed according to $p_{\boldsymbol{\theta}} \in \mathfrak{C}$,
  Q^n distributed over $\mathcal{A}^n$ is **universal almost surely for $\mathfrak{C}$** if for each $\boldsymbol{\theta} \in [0,1]^{|\mathcal{A}|-1}$,

$$\frac{\rho_{p_{\boldsymbol{\theta}},Q^n}(X_{\boldsymbol{\theta},1:n})}{n} \xrightarrow{n\to\infty} 0 \qquad \text{with probability 1.}$$

  where the redundancy $\rho_{p_{\boldsymbol{\theta}},Q^n}(x_{1:n}) := \log \frac{p_{\boldsymbol{\theta}}(x_{1:n})}{Q^n(x_{1:n})}$.
  - Universality is **not** worst-case, i.e., ~~$\max_{\theta\in\Theta} \rho_{p_{\boldsymbol{\theta}},Q^n}(x_{1:n}) \leq (\ldots)$~~.
  - Universality is **asymptotic** and **almost surely**.

The Main Result

Theorem

Let $\{k_n\}_{n\in\mathbb{N}}$ *be such that* $\lim_n k_n = \infty$ *and* $\lim_n \frac{k_n}{\log n} = 0$. *The* k_n*KT distribution* $P^{mn}_{k_nKT}$ *is* **universal almost surely** *for the class of all m-piece i.i.d. sources when compressing equal number of symbols of each piece, i.e.,*

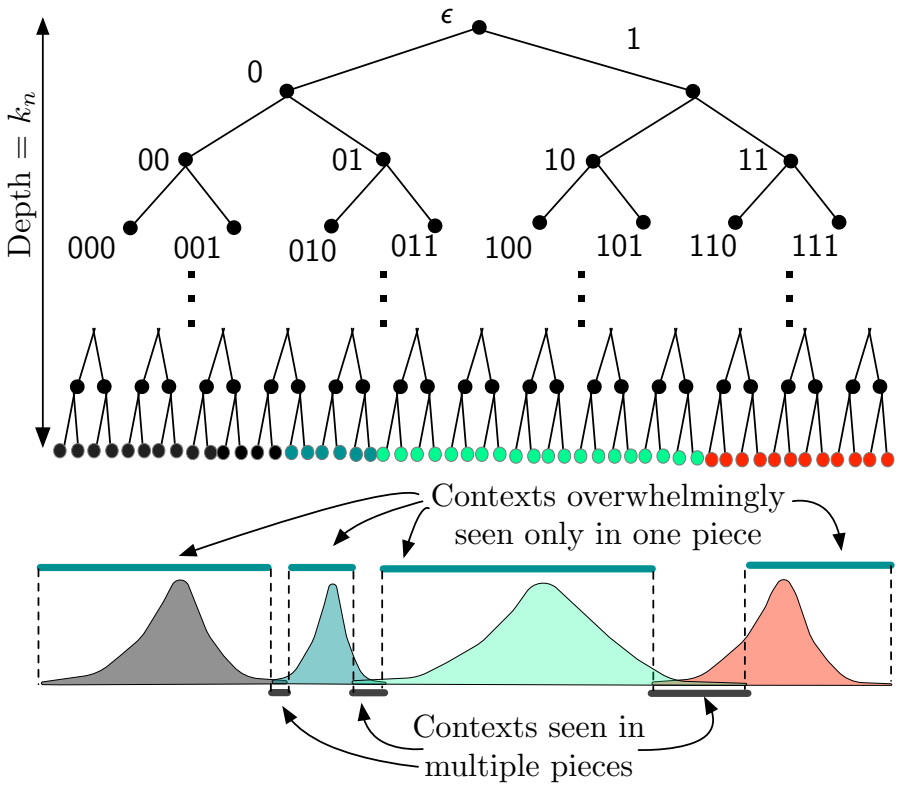
$$\frac{\rho_{\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m, P^{mn}_{k_nKT}}(X_{1,1:n}, X_{2,1:n}, \ldots, X_{m,1:n})}{n} \xrightarrow{n\to\infty} 0 \qquad \textit{with probability 1.}$$

- Informally: efficient universal compression of piecewise i.i.d sources is achievable by modelling them as a Markov process of sufficiently high order using the KT estimator.
- Counter-intuitive: non-stationary class of piecewise i.i.d. sources are universally compressed by a sequence of stationary/time-homogenous kKT distributions.

Advantages of Our Approach

- $\{k_n\}$ must grow unbounded, but can grow at any sub-logarithmic rate.
- No need for a Bayesian mixture over a class of distributions (as is done in previous approaches in the literature).
- Modelling the data as a Markov source with sufficient memory enables us to:
  - discriminate between different sources without needing to weight different sources;
  - naturally remembers and exploits statistics of past pieces
- No need to identify the locations of source transitions explicitly.
- Runtime reduced to $\omega(n)$ (by appropriately choosing k_n).

Intuition Why the Result Holds



- $k_n \in \omega(1)$ ensures that the contexts get longer as n grows
  - **atypical** realizations occur with vanishingly small probability (as k_n grows)
- $k_n \in o(\log n)$ ensures:
  - In **typical** realizations of the source, each context is seen with the frequency prescribed by the distribution of **one** of the m-pieces.
- Sufficiently long contexts likely in $\geq$ 2 pieces are progressively rare (as k_n grows)

- The redundancy bound for typical realizations follows standard bounds for KT and kKT estimators yielding an asymptotic almost surely result.

A Few Remarks

Theorem

Let $\{k_n\}_{n\in\mathbb{N}}$ *be such that* $\lim_n k_n = \infty$ *and* $\lim_n \frac{k_n}{\log n} = 0$. *The* k_n*KT distribution* $P^{mn}_{k_n\mathrm{KT}}$ *is* **universal almost surely** *for the class of all (equal length) m-piece i.i.d. sources.*

- The assumption that each piece has the same length can be dropped (as long as each piece grows linearly in n).
- The sequences of distributions $\{P^{mn}_{k\mathrm{KT}}\}$ are not necessarily **strongly sequential**, i.e.,

$$P^{mn}_{k_n\mathrm{KT}}(x_{1,1:n},\ldots,x_{m,1:n}) \neq \sum_{x_{1:m,n+1}} P^{m(n+1)}_{k_{n+1}\mathrm{KT}}(x_{1,1:n+1},\ldots,x_{m,n+1})$$

- The sequence of CTW distributions P_{CTW,k_n} of depth k_n are at least as good, and typically better, for universal compression of m-piece sources, since

$$\rho_{\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m,P_{\mathrm{CTW},k_n}} = \rho_{\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m,P^{mn}_{k_n\mathrm{KT}}} + \log\frac{P^{mn}_{k_n\mathrm{KT}}(\cdot)}{P_{\mathrm{CTW},k_n}} \leq \rho_{\{p_{\boldsymbol{\theta}_\ell}\}_{\ell=1}^m,P^{mn}_{k_n\mathrm{KT}}} + \underbrace{2^{k_n}}_{\in o(n)}$$

Future Directions

- The uniformity/non-uniformity of the convergence of the redundancy to zero.
- Extension to the setting when pieces which do not grow linearly in n.
- A worst-case upper bound for redundancy (finite n setting).
- Extensions to piecewise Markov and variable-order Markov sources.