On the CTW-based Entropy Estimator

Ronit Bustin

General Motors Advanced Technical Center - Israel

May, 2021

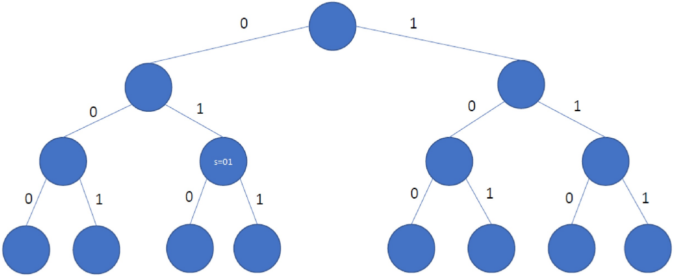
Problem Setting

- The problem: estimating the entropy of a given sequence of discrete observations
- Rises in many different fields and there are numerous different applications, specifically in neuroscience (entropy is a measure of information transmitted between neurons).
- Gao et al. 2008 conducted an extensive comparison between some of the most popular and effective entropy estimation methods used in practice.
- They have shown that the context tree weighted (CTW) method repeatedly and consistently provides the most accurate result.

What is the CTW estimator?

- Willems, Shtarkov and Tjalkens, 1995, conceived the CTW algorithm as a *universal lossless compression* algorithm.
- The algorithm is built on the idea of the *context tree*, suggested by Rissanen and Langdon in 1981.
- A context tree is a data structure that elegantly facilitates the estimation of the probability of a given sequence by using enumerations.
- Each node in the tree is equivalent to a context

Context Tree



- Each node in the tree is equivalent to a context, e.g. the context 01.
- The enumeration in that node is for the subsequent symbol.

Context Tree

Consider the following sequence, where we already marked the context 01:

10111011100100001110101010101100010101001010101001010111010101010

We now extract the subsequence of subsequent symbols:

10111011100100001110101010101100010101001010101001010111010101010

In this example the resulting subsequence is:

10111011100100001110101010101100010101001010101001010111010101010

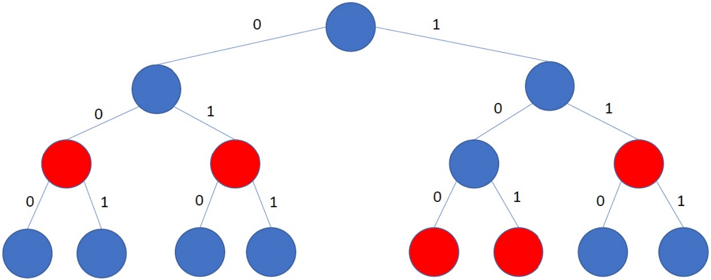
...

1101000010000000000010000

The enumeration in this node facilitates the estimation of the probability of this subsequence assuming independence of its symbols.

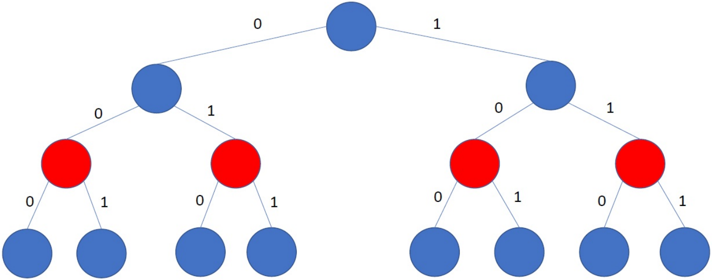
Different models within the tree

- Assuming we limit ourselves to $D = 3$, meaning the maximum context is of length three (and correspondingly, the context tree is of depth three).
- The context tree contains multiple possible models - C_D, denotes the *model class*.



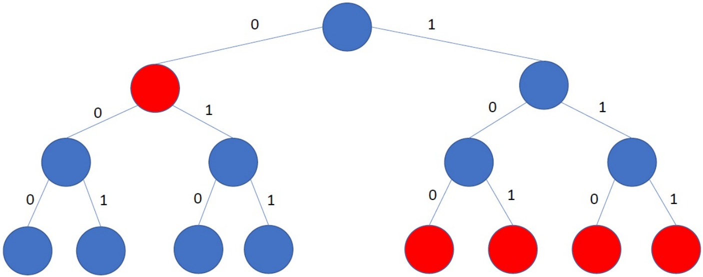
Different models within the tree

- Assuming we limit ourselves to $D = 3$, meaning the maximum context is of length three (and correspondingly, the context tree is of depth three).
- The context tree contains multiple possible models - C_D, denotes the *model class*.



Different models within the tree

- Assuming we limit ourselves to $D = 3$, meaning the maximum context is of length three (and correspondingly, the context tree is of depth three).
- The context tree contains multiple possible models - C_D, denotes the *model class*.



The novelty in CTW

- The usage of the Krichevski-Trofimov (KT) estimator for the probability of the independent subsequence in each node (context). The advantages of this are:
  - The estimator of the probability can be computed sequentially
  - Has a lower bound that allows bounding uniformly the parameter redundancy in the CTW compression
- More importantly - the weighing of the models. The weighted probability in each node is defined as

$$P_w^s = \begin{cases} \frac{1}{2} P_e(a_s, b_s) + \frac{1}{2} P_w^{0s} P_w^{1s}, & \text{for } 0 \leq \ell(\boldsymbol{s}) < D \\ P_e(a_{\boldsymbol{s}}, b_{\boldsymbol{s}}), & \text{for } \ell(\boldsymbol{s}) = D \end{cases}$$

The novelty in CTW - cont'd

This weighing results with the following interesting result:

Lemma 2 [Willems, Shtarkov and Tjalkens, 1995]

The weighted probability P_w^s of a node $s \in \mathcal{T}_D$ with $\ell(\boldsymbol{s}) = d$ for $d \in [0, D]$ satisfies

$$P_w^s = \sum_{U \in \mathcal{C}_{D-d}} 2^{-\Gamma_{D-d}(U)} \prod_{u \in U} P_e(a_{us}, b_{us})$$

with

$$\sum_{U \in \mathcal{C}_{D-d}} 2^{-\Gamma_{D-d}(U)} = 1.$$

where $\Gamma_{D-d}(U)$ is the cost of a model U with respect to the model class $\mathcal{C}_{D-d}$. If one uses the natural code to represent the model it is the number of bits required to represent the model.

From universal compression to entropy estimation

The output of CTW compression algorithm is an estimated probability for the given sequence, weighing all possible models. How do we extract an estimation for the entropy?

[Shannon-McMillan-Breiman] result

For any stationary and ergodic process $\{X_i\}_{-\infty}^{\infty}$ with entropy rate H,

$$-\frac{1}{T}\log P(x_1^T) \to H, \text{ with probability } 1 \text{ as } T \to \infty$$

where $P(x_1^T)$ is the probability of the sequence x_1^T.

So what is the problem?

- The probability outputted from the CTW compression algorithm is a weighted combination over all possible models:

$$P_c(x_1^T) = \sum_{U \in C_D} 2^{-\Gamma(U)} \prod_{u \in U} P_e^u.$$

  where $\prod_{u \in U} P_e^u$ is the estimated probability of he sequence, given a specific model U.
- There is no gurantee of convergence according to the SMB result.
- Can we provide more insight to what it is that we are estimating? Can we bound the over-estimation?

Diving into the details of the CTW we provide insights to the performance of the CTW entropy estimator.

Before stating our contribution...

- We define $\boldsymbol{U}$ to be a random variable distributed according to $2^{-\Gamma(U)}$ for all $U \in \mathcal{C}_{D-d}$
- $\hat{H}_{\boldsymbol{s}}^{CTW} \equiv -\frac{1}{T_{\boldsymbol{s}}} \log P_w(\boldsymbol{s})$ denotes the CTW based entropy estimator at node $\boldsymbol{s}$

Definition

$\hat{H}(\boldsymbol{x_s}|\boldsymbol{U} = U)$ is the Shannon-McMillan-Breiman estimator of the entropy assuming the model describing the subsequence $\boldsymbol{x_s}$ (the subsequence constructed from the symbols in $\boldsymbol{x}$ appearing after the context $\boldsymbol{s}$) is U, meaning:

$$\hat{H}(\boldsymbol{x_s}|\boldsymbol{U} = U) = -\frac{1}{T_{\boldsymbol{s}}} \log P_a^U.$$

where P_a^U is the actual probability of the subsequence (from node $\boldsymbol{s}$) assuming U is the model.
Thus, $\hat{H}(\boldsymbol{x_s}|\boldsymbol{U})$ is the conditional Shannon-McMillan-Breiman estimator of the entropy.

Our contribution

Theorem 1

Consider a context tree $\mathcal{T}$. At each node $\boldsymbol{s}$ with $\ell(\boldsymbol{s}) = d$ we have the following property:

$$\hat{H}_{\boldsymbol{s}}^{CTW} \leq \hat{H}\left(\boldsymbol{x}_{\boldsymbol{s}} | \boldsymbol{U}\right) + \frac{1}{T_{\boldsymbol{s}}} \sum_{U \in \mathcal{C}_{D-d}} 2^{-\Gamma(U)} |U| \gamma\left(\frac{T_{\boldsymbol{s}}}{|U|}\right)$$

where $T_{\boldsymbol{s}}$ is the length of the subsequence that corresponds to node $\boldsymbol{s}$ (meaning the subsequence of symbols appearing after the context $\boldsymbol{s}$ in the full sequence).

$$\gamma(z) = \begin{cases} z, & \text{for } 0 \leq z < 1 \\ \frac{1}{2}\log z + 1, & \text{for } z \geq 1 \end{cases}$$

Our contribution - cont'd

Theorem 2

Consider a context tree $\mathcal{T}$. At each node $\boldsymbol{s}$ with $\ell(\boldsymbol{s}) = d$ we have the following property:

$$\begin{aligned}\hat{H}_{\boldsymbol{s}}^{CTW} &\leq 2^{-\Gamma(U^\star)} \hat{H}\left(\boldsymbol{x}_{\boldsymbol{s}} | \boldsymbol{U} = U^\star\right) + \frac{1}{T_{\boldsymbol{s}}} 2^{-\Gamma(U^\star)} |U^\star| \gamma\left(\frac{T_{\boldsymbol{s}}}{|U^\star|}\right) \\ &\leq \hat{H}\left(\boldsymbol{x}_{\boldsymbol{s}} | \boldsymbol{U} = U^\star\right) + \frac{1}{T_{\boldsymbol{s}}} 2^{-\Gamma(U^\star)} |U^\star| \gamma\left(\frac{T_{\boldsymbol{s}}}{|U^\star|}\right)\end{aligned}$$

where $T_{\boldsymbol{s}}$ is the length of the subsequence that corresponds to node $\boldsymbol{s}$ (meaning the subsequence of symbols appearing after the context $\boldsymbol{s}$ in the full sequence). $U^\star$ is any specific model in $\mathcal{C}_{D-d}$.

About the proof

- The understanding of the CTW algorithm
- Jensen's inequality
- The bounds used in the performance analysis of the CTW: parameter redundancy and model redundancy.

Thank You!

email: ronit.bustin@gm.com