A note on the Shannon entropy of short sequences

H. M. de Oliveira and R. Ospina

Statistics Department, Federal University of Pernambuco (UFPE), Recife-PE, Brazil.

Introduction

Entropy is one of the most fundamental concepts of Science.

Scope of Information Theory (IT): it is characterized in the process of coding discrete memoryless sources (DMS).

It even seems a little odd that the entropy being associated with a statistic mean, but no parameters have been associated with the variance of the information produced by the symbols of the source.

entropy $H(U)$ = first moment of a random variable that measures the information about a symbol emitted by the source;

information fluctuation $F(U)$ = square root of the second central moment of the same variable.

Fluctuation of Information

DMS source U is a random variable over $\mathbb{A} := \{a_k\}_{k=1}^K$ with probability of occurrence at l-time instant $\{p_l(a_k)\}_{k=1}^K$, where $p_l(a_k) = P_l(U = a_k)$ for $l = 1, 2, \cdots, L$

stationary model: $p_l(a_k) = p(a_k) = P(U = a_k)$.

for each symbol a_k

$$\mathcal{I}(a_k) = -\log_2(p(a_k)),\ a_k \in \mathbb{A}. \tag{1}$$

Let us denote $p(a_k) = p_k$.

Shannon entropy of the DMS:

$H(U) := -\sum_{k=1}^K p_k \cdot \log_2 p_k$.

Here, we propose to use the *information fluctuation*:

$F^2(U) := \sum_{k=1}^K p_k \cdot \log_2^2 p_k - H^2(U)$.

Therefore, the fluctuation can be computed by ($F^2 \geq 0$ by Jensen's inequality):

$$F(U) = \sqrt{\sum_{k=1}^{K}(p_k - p_k^2)\log_2^2 p_k - \sum_{i\neq j} p_i p_j \log_2 p_i \log_2 p_j}. \tag{2}$$

Definition 2.1 (degenerated sources). *Let U^* and U^{**} be two kinds of degenerated DMS, with distributions:*
type I source U^***:** $\{p_k\}_1^K$ *where* $\exists\, k^* | p_{k^*} = 1$, *and* $p_k = 0$ $(\forall\, k \neq k^*)$.
type II source U^{**}**:** $\{p_k\}_1^K$, *where* $p_k = \frac{1}{J}$, $\forall\, k \in I \subset \{1, 2, \cdots, K\}$ *and* $J = ||I||$. *Here, the* $||\cdot||$ *indicates the cardinality of set.* □

Proposition 2.1. *The fluctuation of information is null if and only if the source is degenerated, i.e.,* $F(U) = 0 \Leftrightarrow U^*$ *or* U^{**}.

Although U^*-sources can be included as a particular case of U^{**}-sources, we define them as different types of sources.

Consider now a binary DMS source where the output is either a 0 with probability p or 1 with a probability $q := 1 - p$.:

$$H_2(p) := -p \log_2 p - q \log_2 q, \tag{3}$$

$$F_2(p) := \sqrt{pq \left(\log_2 p - \log_2 q\right)^2}, \tag{4}$$

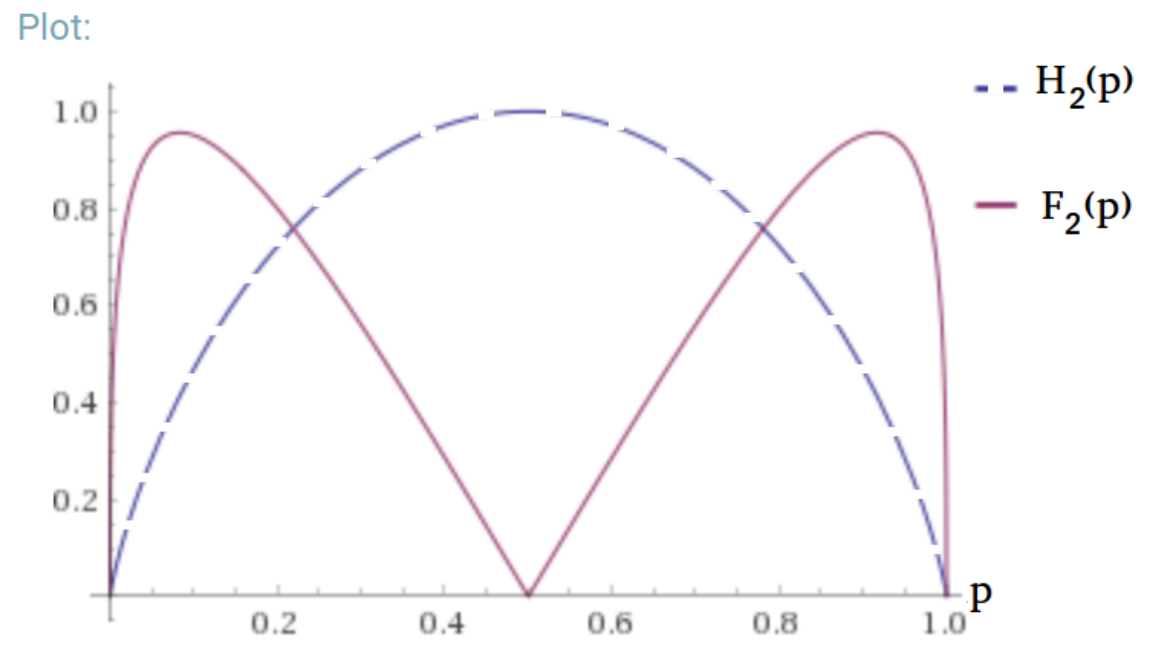

Figure 1: Binary fluctuation information $F_2(p)$ expressed in Shannons.

There are five attributions to p worth to mention:
$p \in \{0, 0.0832217\cdots, 0.5, 0.9167783\cdots, 1\}$.

Three limit cases have no variance (binary degenerated sources):
$p = 0$ and $p = 1$, and $p = 0.5$.

With equiprobable sequences and maximum entropy, has no variance in the amount of information: whatever the length of the sequence of symbols of the source, the average information per source letter is *exactly* equal to the entropy.

The points of maximum variability on the information content seem to be less known in the literature.

We start by investigating

$$\frac{\mathrm{d}F_2(p)}{\mathrm{d}p} = \frac{\log(\frac{1}{p}-1).\left\{(1-2p)\tanh^{-1}(1-2p)-1\right\}}{\sqrt{-p(1-p)\log^2(\frac{1}{p}-1)}}. \tag{5}$$

Three points have infinite derivative values. The two critical points occurs when $\tanh\left(\frac{1}{1-2p}\right) = 1-2p$, and the numerical solution is: $p^* \approx \frac{1}{2}[1 \pm 0.833557]$. The *saltus* at $p = 0.5$ is circa 5.77078.

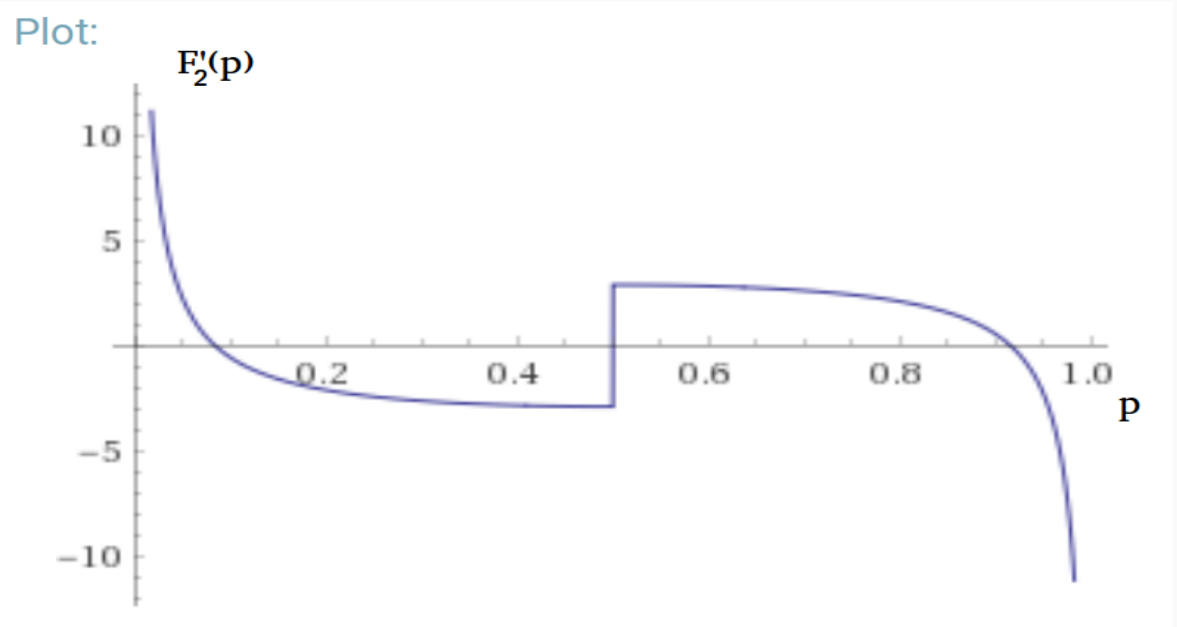

Figure 2: Derivative of the binary fluctuation $F_2(p)$.

We also calculate the intersection points of the two curves in Fig. 1, where $H_2(p) = F_2(p)$. We agreed to use this range of p values to characterize binary DMS with low entropy variability.
Using the coefficient of variation, $CV := 100\frac{F_2(p)}{H_2(p)}$, we plot Fig. 3.

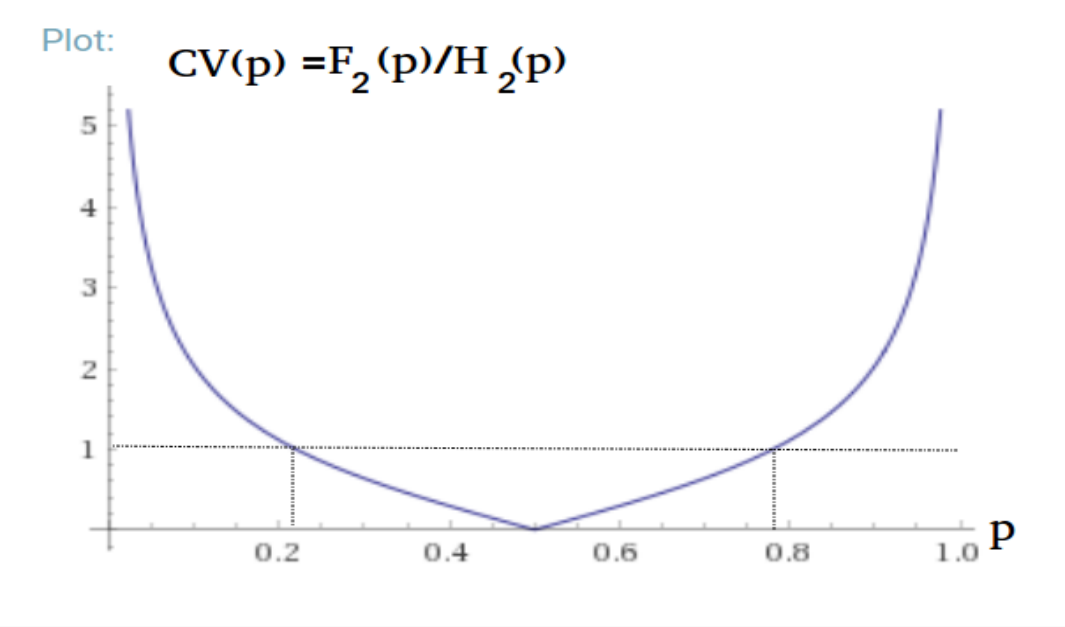

Figure 3: Coefficient of variation of entropy: $CV := 100\frac{F_2(p)}{H_2(p)}$.

Definition 2.2. *A binary memoryless source is called a low entropy variability DMS if and only if* $0.21907592 \leq p \leq 0.78092407$. □

Entropy of Finite Sequences: Statistical Evaluation

Let $\{x_l\}$ be a sequence of observed counts of symbols in the alphabet in a DMS sample of size L and $\widehat{p}_k = m_k/L$ as being the sample relative frequency of the kth symbol a_k.

Nonparametric estimation of $H(U)$ (plug-in estimator)

$$\widehat{H}(U) := -\sum_{k=1}^{K} \widehat{p}_k \cdot \log_2 \widehat{p}_k$$

The plugin estimator of information fluctuation the statistic

$$\widehat{F}^2(U) := \sum_{k=1}^{K} \widehat{p}_k \cdot \log_2^2 \widehat{p}_k - \widehat{H}^2(U). \tag{6}$$

The central limit theorem can be used to establish that $\widehat{H}(U)$ converges to the normal distribution

$$\widehat{H}(U) \sim \mathcal{N}(H(U), \frac{F^2(U)}{L}). \tag{7}$$

Using a similar heuristic, is possible to note that $\widehat{F^2}(U)$ is in the fact an asymptotic estimator of he information fluctuation, i.e. $\mathbb{V}ar(\widehat{H}(U))$ is the variance of entropy.

Consequently the quantity

$$\frac{(L-1)\widehat{F^2}(U)}{F^2(U)} \sim \chi^2(L-1), \tag{8}$$

where $\chi^2(L-1)$ is the Chi-square distribution with $L-1$ degrees of freedom.

Set a significance level α, one can calculate a (right) confidence interval setting

$$\widehat{H}(U) + z_\alpha \frac{F(U)}{\sqrt{L}}, \tag{9}$$

where $z_\alpha = \phi^{-1}(\alpha)$ is the α-quantile of the Normal distribution, and $\phi(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\zeta^2/2} d\zeta$.

For short sequences is more appropriate to use the robust t-Student statistics:

$$H_{practical} := \widehat{H}(U) + t_{(\alpha;L-1)} \frac{\widehat{F}(U)}{\sqrt{L}}$$

We can thus expect to code the output with $H_{practical}$ bits per source symbol, instead of using the fundamental limit $H(U)$, which can be asymptotically achieved when the length of the source sequences grows indefinitely.

We propose to assess the efficiency of a source coding run for a particular given sequence (especially in the cases of short sequences and high entropy variability DMS), at a significance level α, according to

$$\eta_\alpha := \frac{\widehat{H}(U) + t_{(\alpha, L-1)}\widehat{F}(U)/\sqrt{L}}{\bar{L}}. \tag{10}$$

Atypical Sequences

Let us now shift the focus to long sequences. All sequences produced from the source that result in a sample entropy within the confidence interval

$$\left[\widehat{H}(U) - z_{\alpha/2}\frac{\widehat{F}(U)}{\sqrt{L}}, \widehat{H}(U) + z_{\alpha/2}\frac{\widehat{F}(U)}{\sqrt{L}}\right] \tag{11}$$

may are considered to be typical.

Now, set an arbitrary value $\epsilon > 0$. Any sequence with length L symbols for which the sample entropy is outside the range $[H - \epsilon, H + \epsilon]$ is called an ϵ-atypical sequence. What is the level α of significance at which the confidence interval (Eq. (11)) for the entropy coincides with this interval? The relationship

$$\phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \frac{\epsilon\sqrt{L}}{F(U)} \tag{12}$$

should be imposed.

Here, the significance level is interpreted as the probability of occurrence of an atypical sequence, i.e., $P(\text{atypical sequence}) = \alpha$. Therefore,

$$\lim_{L\to\infty} P(\text{atypical sequence}) = \lim_{L\to\infty} 2\left[1-\phi\left(\frac{\epsilon\sqrt{L}}{F(U)}\right)\right] = 0. \tag{13}$$

Another key concept of IT is the **asymptotic equipartition property (AEP)**. The extension U^L of the source U consider L-grams as the new symbols, it has an alphabet $\{\underline{u}_i\}_{i=1}^{K^L}$ with probabilities $\{P(\underline{u}_i)\}_1^{K^L}$. More than $P\left[\lim_{L\to\infty}\widehat{H}(U) = H(U)\right] = 1$, it is shown that ϵ-typical sequences hold

$$2^{-L(H(U)+\epsilon)} \leq P(\underline{u}_i) \leq 2^{-L(H(U)-\epsilon)}. \tag{14}$$

Therefore, $P(\underline{u}_i) \approx 2^{-L\cdot H(U)}$ (constant) for $i \in \mathbb{T} \subset \{1, 2, \cdots, K^L\}$, the set of typical sequences, with $\| \mathbb{T} \| = 2^{L\cdot H(U)}$. By definition of degenerated sources, we see that large extensions of a DMS becomes type II asymptotically degenerated. Its entropy becomes exactly $L \cdot H(U)$, with essentially no information fluctuation.

Conclusions

The idea of introducing a new parameter for a DMS, namely the information fluctuation, seems to be as basic as the concept of entropy itself.

In addition to proposing the calculation of a more “realistic” estimate for entropy, now depending on the length of the sequence by the source, a naive and didactic interpretation on “typical sequences” is presented.

For short sequences is not the entropy of the source that should be used to calculate the efficiency of source coders, but rather in terms of the sample entropy.

This work has been supported by Statistics Department UFPE, Brazil.