Entropy has many faces

J. Ziv
Faculty of Electrical Engineering
Technion—Israel Institute of Technology
Haifa, 32000, Israel

"In the beginning, there was Entropy..."

Let

$$X_1^\ell = X_1, X_2, \ldots, X_\ell\,; \qquad\qquad X_i \in \alpha$$
$$|\alpha| = A.$$

For stationary sources, the ℓ-th order per letter entropy (entropy- rate) is:

$$H(X_1^\ell) = -\frac{1}{\ell}\,E\,\text{log}\;P(X_1^\ell)$$

and

$$H = \lim_{\ell\to\infty}\;H(X_1^\ell)\,. \tag{1}$$

Quantities of the form H play a central role in information theory as measure of information, choice and uncertainty.

However, "The real justification of these definitions, however, will reside in their implications" (C.E.Shannon, 1948)

In this lecture we are going to discuss different implications, interpretations and applications of the the notion of entropy, as well as novel definitions of the notion of entropy which are derived from different physical models describing some real-world communication, classification and prediction problems,

A) Entropy and Typical Sequences

Let P be an ergodic process of entropy H. For $\epsilon > 0$ and $n \geq 1$ define

$$T_n(\epsilon) = [X_1^n : 2^{-n(H+\epsilon)} \leq P(X_1^n) \leq 2^{-n(H-\epsilon)}]$$

$For\ each\ \epsilon > 0,\ X_1^n \in T_n(\epsilon),\ eventually\ almost\ surely.$ (A.E.P.)

Thus, if the number of "Typical sequences" with a total measure $\alpha \in (0,1)$ is $N_n(\alpha, \epsilon)$, then:

$$H = \lim_{\epsilon \to \infty} \lim_{n \to \infty} N_n(\alpha, \epsilon)\,. \tag{2}$$

B) Entropy and Data Compression

C.E. Shannon was the first one to point out that for a given source, the entropy is the lowest average number of bits per input letter that a noiseless (i.e. error-free) encoder can achieve. Indeed, for a given source code, let $L(X_1^\ell)$ denote the length function of X_1^n defined to be the number of bits that represent X_1^n.

By (2), it follows that there exits a code or n-vectors such that,

$$H = \lim_{n\to\infty} \frac{1}{\ell} EL(X_1^n), . \tag{3}$$

Furthermore, for prefix-codes,

$$-\log P(X_1^\ell) \le L(X_1^\ell) \le -\log P(X_1^\ell) + 1$$

(Huffman code, Shannon code)

hence,

$$H(X_1^\ell) \le \frac{1}{\ell}\, EL(X_1^\ell) \le H(X_1^\ell) + \frac{1}{\ell}\,. \qquad (4)$$

C) Entropy and Individual Sequences

Almost surely, for any ergodic source with measure $P(.)$

$$H = \lim_{n\to\infty} \frac{1}{n}[-\log P(X_1^n)]. \qquad (5)$$

$-\log P(X_1^n)$ is the Self-Information of X_1^n and is a function of X_1^N via $P(.)$.

Q: ARE THERE OTHER FUNCTIONS $F(X_1^n)$ THAT EXPLICITLY DEPEND ON X_1^n (NOT VIA $P(.)$), WHICH MAY BE UNKNOWN TO THE SOURCE ENCODER), SUCH THAT:

$$H = \lim_{n\to\infty} \frac{1}{n}[-\log F(X_1^n)] \ ?$$

A: YES!!

We shall introduce a number of such functions, each associated with a different, unique set-up: either in connection with universal data-compression, entropy and recurrence time, universal classification or universal prediction.

C) Entropy of Individual Sequences

1) Minimal Description Length (MDL)

If the source is known to be a Markov source, but the order (and the probability measure) is not known, define $M(X_1^n, K)$ to be the binary description of the empirical K-th order Markov measure and let $L_M(X_1^n, K)$ be the binary encoding of X_1^N, relative to this empirical measure.

Let

$$\hat{L}(X_1^n) = \min_K \{M(X_1^K, K) + L_M(X_1^N, K)\}\,.$$

Then

$$\lim_{n\to\infty} \frac{1}{n}\hat{L}(X_1^n) = H(X)$$

almost surely

Another alternative for the MDL approach is the following:

$$\tilde{Q}(x_1^\ell) = \frac{1}{n-\ell+1} \sum_{i=1}^{n-\ell+1} 1\{X_i^{i+\ell-1} = x_1^\ell\}$$

This quantity is called the ℓ^{th}-order empirical probability measure. We also define $\tilde{H}(X_1^\ell)$ to be the entropy rate of the empirical probability $\tilde{Q}$.

Now, by the concavity of the entropy function and Jensen's inequality it follows that:

$$E\tilde{H}(X_1^\ell) = \frac{1}{\ell} E\{-\sum_{\alpha^\ell} \tilde{Q}(X_1^\ell) \log \tilde{Q}(X_1^\ell)\} \leq H(X_1^\ell)$$

Similarly to the Markov case the rest follows for ANY finite-alphabet stationary source.

What if nothing is known about the source measure?

2) Essential Topological Entropy

Consider an infinite sequence $\mathbf{x} = x_0, x_1, x_2 \cdots$.

The k block universe of $\mathbf{x}$ is the set $\mathcal{U}_K(\mathbf{x})$ of all vectors a_1^K that appears in $\mathbf{x}$

$\mathcal{U}_K(\mathbf{x}) = \{a_1^K;\ x_i^{i+K-1} = a_1^K, \text{ for some } i > 1\}$.

The *topological entropy* of $\mathbf{x}$ is defined by:

$$h(\mathbf{x}) = \lim_{K \to \infty} \frac{1}{K} \log |\mathcal{U}_K(\mathbf{x})|$$

Clearly by parsing $\mathbf{x}$ into consecutive K-blocks, one does not need more than $h(\mathbf{x})$ bits per letter to faithfully represent $\mathbf{x}$ (for large K).

However, one can do better!
Strings with too small frequency of occurrence can be eliminated by making small limiting density of changes. For example:

Let

$$\overline{d}(\mathbf{x}, y) = \limsup_{n\to\infty} d_n(x_1^n, y_1^n)$$

Hamming distance between x_1^n and y_1^n

The *essential topological entropy* is defined by

$$H(\mathbf{x}) = \lim_{\epsilon\to 0} \inf_{\overline{d}(\mathbf{x},\mathbf{y})<\epsilon} h(y)$$

("A.E.P." for individual sequence).

Theorem: *If P is an ergodic process with entropy H then*

$$H(\mathbf{X}) = H, \text{ almost surely}$$

Thus $H(\mathbf{x})$ is a "proper" self information.

Coding: Find the smallest $\mathcal{U}_K(\mathbf{x})$ such that $\overline{d}(x, y) \le \epsilon$, for small enough ϵ and big enough K.

Do we have to work so hard?

Not really!

Express an infinite sequence $\mathbf{x}$ (via parsing) as a concatenation

$$\mathbf{x} = w(1)\, w(2)\, w(3)\ \cdots$$

of distinct variable-length blocks called *phrases*.

The next phase is
the shortest new word

0 0 0 1 0 0 1 0 0 1 1 1 1

0, 0 0, 1, 0 0 1, 0 0 1 1, 1 1

One generates a new distinct phrase only when he has to! (LZ parsing)

Let $C(x_1^n)$ be the number of complete phrases in x_1^n.

Theorem:

If P is an ergodic process with entropy H, then

$$\frac{1}{n} C(X_1^n) \log C(X_1^n) \to H$$

almost surely.

Coding: Pointer information (about $\log C(x_1^n)$ bits) + the new letter at the end of the phrase. (LZ algorithm)

Test for Randomness

A source is $truely$ random if H=$\log A$ (The maximal entropy for a source with A letters).

An individual sequence is $truely$ random if:

$$\frac{1}{n} C(X_1^n) \log C(X_1^n) \to \log A$$

A RANDOM SEQUENCE IS AN

$INCOMPRESSIBLE$ ONE!

D) Entropy and Pattern Matching

For x_{-n+1}^{n}, define $L(x_{-n+1}^{n})$ to be the largest integer $L < n$ such that

$$x_1^L = \underline{x}_{-M}^{-M+L-1}$$

for some $M \in [1, n-L+1]$

1 0 0 0 0 1 1 1 , 0 1 1 0

L = 3

M = 4

Theorem:

1. $\dfrac{\log n}{L(X_{-n+1}^{0}, X_1^n)} \to H$ in probability

2. $\dfrac{\log n + 0(1)}{E\,L(X_{-n+1}^{0}, X_1^n)} = H$

   a $n \to \infty$ (subject to some technical conditions).

   This leads to a good estimate of H, namely

$$\frac{\log n}{\frac{1}{K}\sum_{j=1}^{K} L(X_j^{j+n-1}, X_{j+n}^{j+2n-1})}$$

E) Characterization of a Probability Law

Suppose that we observe $X_1, X_2, \cdots, X_{N_0}$ with unknown ℓ-th order probability distribution $P_\ell(z), z \in \mathcal{A}^\ell$. How large an N_0 do we need so that we can determine $P_\ell(\cdot)$?

We formulate the problem as follows. Let $\mathcal{M}_\ell$ be the set of probability distributions on $\mathcal{A}^\ell$. We seek a function $f_c : \mathcal{A}^\ell \times \mathcal{M}_\ell \to \{0, 1\}$, such that with high probability $f_c(\mathbf{X}_1^{N_0}, P_t) = 0$, and $f_c(\mathbf{X}_1^{N_0}, Q_\ell) = 1$, whenever the Kullback-Leibler divergence

$$D_\ell(Q_\ell \| P_\ell) \stackrel{\triangle}{=} \frac{1}{\ell} \sum_{z \in \mathcal{A}^\ell} Q_\ell(z) \log \frac{Q_\ell(z)}{P_\ell(z)} \geq \Delta\,.$$

Thus $f_c(\cdot)$ is a *classifier* that tells us if $P \equiv Q$ or $D_\ell(Q \| P) \geq \Delta$. Using pattern matching theorems we show that (*roughly*) such a classifier exists if and only if $N_0 \geq 2^{H_\ell \ell}$.

More topics

Entropy and universal prediction
Entropy and gambling
Entropy and guessing
etc.