Available online at www.sciencedirect.com
ScienceDirect
Journal of Approximation Theory 158 (2009) 145–150
JOURNAL OF Approximation Theory
www.elsevier.com/locate/jat

On the entropy of a function

Rudolph A.H. Lorentz*

Fraunhofer Institute for Scientific Computations and Algorithms, Germany
University of Duisburg-Essen, Germany

Received 3 April 2008; accepted 4 July 2008
Available online 5 November 2008

Communicated by C.K. Chui and H.N. Mhaskar

To the memory of my father, both as a father and as a mathematical collaborator.

Abstract

A common statement made when discussing the efficiency of compression programs like JPEG is that the transformations used, the discrete cosine or wavelet transform, decorrelate the data. The standard measure used for the information content of the data is the probabilistic entropy. The data can, in this case, be considered as the sampled values of a function. However no sampling independent definition of the entropy of a function has been proposed. Such a definition is given and it is shown that the entropy so defined is the same as the entropy of the sampled data in the limit as the sample spacing goes to zero.

Keywords: Compression; Probabilistic entropy; Lossless

1. Introduction

This paper is concerned with the lossless compression of data consisting of the values of a real-valued function. An example is provided by a digitized photo. Such a photo can be compressed using the JPEG program, which in the older version involves a discrete cosine transform, while JPEG2000 is wavelet based. These transforms are said to decorrelate the data. Such a statement is based on the concept of entropy. The entropy of the transformed data is less than the entropy of the original data.

* Corresponding address: Texas A & M University at Qatar, Texas A & M Engineering Building, PO Box 23874, Education City, Doha, Qatar.
E-mail address: rudolph.lorentz@qatar.tamu.edu.

0021-9045/$ - see front matter
doi:10.1016/j.jat.2008.07.004

The concept of entropy used here is not the metric entropy discussed by my father on page 150 of [2] nor the Kolmogoroff complexity, but, as sons are wont to do, I consider the other type of entropy, the probabilistic entropy as introduced by Shannon [1], which is defined in the following section.

The entropy of a stream of data is a number H. There is a theorem (see e.g. [3] for an overview of compression) that there exist algorithms which compress the data stream into not more than H bits per datum on average and that if the data derive from a iid source, that one can do no better. Huffmann and arithmetic coding do this. Thus to quantify how well a data transformation decorrelates the data, one should compare the entropy before and after the transformation.

The standard way of measuring a function is to sample it, i.e., to record its values at a sequence of points. If the measured values are digitized (quantized) then one could also use the entropy of these values as the entropy of the function. This definition of the entropy of a function depends on the details of the sampling. We propose a definition of the entropy of a function which is sampling independent. It is shown that for two types of sampling, point sampling for a continuous function and mean sampling for a Lebesgue integrable function, the two concepts yield the same value as the sampling spacing tends to zero. Some simple examples are given.

2. Definitions

Definition 1 (*Entropy*). Let X be a discrete random variable which takes values in a finite alphabet A and let $p(x) > 0$ be the probability that $X = x$ for $x \in A$. Then the entropy $H(X)$ of X is

$$H(X) = -\sum_{x\in A} p(x)\log_2 p(x).$$

Note that the entropy of a data stream only depends on the probabilities of occurrence of the different letters of the alphabet and not on their meanings, which in our case is the interpretation as real numbers.

If we define $p\log_2 p$ to be zero when $p = 0$, then the entropy H is a continuous function of the probabilities in the extended range $p(x) \geq 0$.

For a given size alphabet, the entropy is the highest if the probabilities are all equal. Then $H(X) = \log_2 |A|$. The entropy is the lowest if one of the probabilities is 1. Then only one letter occurs and $H(X) = 0$.

Let f be a continuous real-valued function defined on an interval $I = [a, b]$. We want to define the entropy of f. There are two problems. The first is that there is a continuum of values. The second is that f is not a discrete random variable but a continuous random variable.

The approach usually taken to solve the first problem is to sample the function. Assume that f is a continuous real-valued function defined on $I = [a, b]$. Let a positive integer n be fixed, $h = (b - a)/n$ and $x_i = a + (\mathrm{i} + 1/2)h$ for $i = 0, \ldots, n - 1$. Then the sampled function is

$$S_n(f)(i) = f(x_i) \tag{1}$$

for $i = 0, \ldots, n - 1$. We call this point sampling, at the mid-points of the intervals with which the sample is associated.

The other case that we consider is that when $f \in L_1([a, b])$. In this case, the i-th sample of f is

$$S_n(f)(i) = h^{-1}\int_{x_i-h/2}^{x_i+h/2} f(x)\mathrm{d}x \tag{2}$$

where h is defined as above and $i = 0, \ldots, n-1$. This we call mean sampling. Note that if f is measurable and bounded, then $f \in L_1([a, b])$.

Now we have a discrete variable which takes on a continuum of values. The next step is to quantize these values in order to obtain a finite number of values/letters. Given f as above let $q > 0$ be the quantum. The quantization of the real numbers which we take is

$$Q_q(f)(x) = (\mathrm{i} + 1/2)q \quad \text{if} \quad \mathrm{i}q \le x < (\mathrm{i}+1)q. \tag{3}$$

$Q_q(f)$ is a simple function which approximates f.

Generally speaking the formula for entropy is used even if it is not known whether the discrete random variable X satisfies the conditions of Definition 1. If one is given a string $x_1x_2 \ldots x_N$ of letters with x_i in a finite alphabet A, then one uses the relative frequencies of occurrences of the letters in place of the probabilities.

More precisely, let $c(x)$ denote the number of occurrences of the letter x in a string of length N. We define $p(x) = c(x)/N$ and use these as probabilities in the formula for the entropy. As an example, let us quantize the values of the samples of a bounded function $Q_q S_n(f)(i)$. Since f is bounded, only a finite number of values occur. Proceeding as above, we can calculate the "entropy" of the quantized sequence of samples.

Example 1. Let $f(x) = x$ on $[0, 1]$. Fix n and take $q = 1/n$. Then $S_n(f)(i) = (\mathrm{i} + 1/2)/n$ for $i = 0, \ldots, n-1$. The quantization has been chosen so that the values don't change $Q_q S_n(i) = (\mathrm{i} + 1/2)/n$. Each of the n values occurs exactly once

$$H(Q_q S_n(f)) = -\sum_{i=0}^{n-1} (1/n) \log_2(1/n) = \log_2 n.$$

So one of the simplest functions has the highest possible entropy. This also shows why probability based compression methods such as the Huffmann coding are unable to significantly compress files containing numbers. The explanation is that the letters/numbers are strongly correlated. The simplicity of the function is reflected in the correlation of the numbers, not in their probability distributions.

3. Entropy and sampling

In this section, we will assume for the sake of simplicity that $[a, b] = [0, 1]$.

Definition 2 (*Entropy of a Function*). Let f be a measurable essentially bounded real-valued function defined on $[0, 1]$ and let $q > 0$. Let $I_i = [\mathrm{i}q, (\mathrm{i}+1)q)$ and $B_i = f^{-1}(I_i)$. Then the entropy $H_q(f)$ of f at quantization level q is

$$H_q(f) = -\sum_i \mu(B_i) \log_2\{\mu(B_i)\}. \tag{4}$$

Here μ is the Lebesgue measure. Note that as f is measurable, so are the B_i (see e.g., [4] for details on the measure theory that we use here and in the proof of the following theorem). In addition,

$$[0, 1] = \bigcup_i B_i$$

as a finite disjoint union and so

$$\sum_i \mu(B_i) = 1$$

as a finite sum.

There is a somewhat similar concept, that of differential entropy, which is used for another purpose. Suppose that one has a real-valued bounded continuous random variable X with a continuous probability function $\rho(x)$. Let $h > 0$. Then for each i, there is an $x_i \in [(\mathrm{i}-1)h, \mathrm{i}h)$ with

$$\rho(x_i)h = \int_{(\mathrm{i}-1)h}^{\mathrm{i}h} \rho(x)\mathrm{d}x.$$

This converts the continuous random variable X into a discrete random variable X_h with probabilities $P(X_h = x_i) = \rho(x_i)h$. Only a finite number of these probabilities is non-zero since X is bounded. The entropy of X_h is

$$H(X_h) = -\sum_i \rho(x_i)h \log_2(\rho(x_i)h) = -\sum_i \rho(x_i) log_2(\rho(x_i))h - \log_2 h.$$

As $h \to 0$ the sum converges to the integral $-\int \rho(x) \log_2 \rho(x)\mathrm{d}x$ which is called the differential entropy of X. $\log_2 h$ converges to $-\infty$ unfortunately, but the integral expression is nevertheless used for other purposes.

The analogy to our case is when f is a continuous real-valued function $f : [0, 1] \to \mathbb{R}$ which induces the probability measure $\mu_f(A) = \mu(f^{-1}(A))$ for any $A \subset \mathbb{R}$. But this measure cannot in general be written as $\rho(x)\mathrm{d}\mu(x)$ since μ_f is not necessarily absolutely continuous with respect to μ, in particular, not when f is constant on a set of non-zero measure.

The main theorem which justifies the above definition is:

Theorem 1. *Using the terminology of Definition 2, let f be continuous for point sampling and f be measurable and essentially bounded for mean sampling. The sampling spacing is $1/n$. Let $S_n(f)$ be the corresponding sampling of Eq. (1) respectively (2). Fix $q > 0$ and let $Q_q S_n$ be the quantization of the samples with resolution q as given in Eq. (3). Let $c_n(i)$ be the number of occurrences of the value $(\mathrm{i} + 1/2)q$ in $Q_q S_n$ and $p_n(i)$ be the relative probability of the occurrence of the value $(\mathrm{i} + 1/2)q$*

$$p_n(i) = \frac{c_n(i)}{\sum_j c_n(j)} = \frac{c_n(i)}{n}.$$

Then

$$\lim_{n\to\infty} -\sum_i p_n(i) \log_2 p_n(i) = H_q(f). \tag{5}$$

In other words, the entropy of the quantized samples of a function converges to the entropy of the quantized function as the sampling spacing goes to zero.

Proof. Let $\epsilon > 0$ be given. Let A_i be the union of those sampling intervals $[r_{i,j}(1/n), (r_{i,j} + 1)(1/n))$ for which $\mathrm{i}q \le f((r_{i,j} + 1/2)(1/n)) < (\mathrm{i} + 1)q$ for point sampling and

$$\mathrm{i}q \le \int_{r_{i,j}(1/n)}^{(r_{i,j}+1)(1/n)} f(x)\mathrm{d}x < (\mathrm{i} + 1)q$$

for mean sampling. Then $\bigcup_i A_i = [0, 1]$, $c_n(i) = \mu(A_i)/(1/n)$ and $p_n(i) = \mu(A_i)$.

Using the terminology of the definition of the entropy of f, $B_i = \{x \mid \mathrm{i}q \le f(x) < (\mathrm{i}+1)q\}$ and we have $\bigcup_i B_i = [0,1]$. Let M be the number of sets B_i with positive measure. Let $\delta > 0$ be such that

$$\left| \sum_i \mu(B_i) \log_2 \mu(B_i) - \sum_i p_i \log_2 p_i \right| < \epsilon$$

whenever $\max_i |\mu(B_i) - p_i| < \delta$. We take the sum only over the M non-zero B_i.

For each B_i, there is a closed set C_i contained in B_i such that $\mu(B_i \setminus C_i) \le \delta/(20M)$. As a closed subset of $[0,1]$, C_i is a countable disjoint union of closed intervals $J_{i,j}$. Choose m_i so that

$$\mu(C_i) - \sum_{j=1}^{m_i} \mu(J_{i,j}) \le \delta/(20M).$$

Replace each $J_{i,j}$ by half-open intervals (closed on the left and open on the right) $\tilde{J}_{i,j}$ whose end points are multiples of the sample spacing $1/n$, such that $\tilde{J}_{i,j} \subseteq J_{i,j}$, and which are the largest intervals with this property. Now choose n so large that

$$\sum_{j=1}^{m_i} \mu(J_{i,j}) - \sum_{i=1}^{m_i} \mu(\tilde{J}_{i,j}) \le \delta/(20M).$$

As a consequence,

$$\mu(C_i) - \sum_{j=1}^{m_i} \mu(\tilde{J}_{i,j}) \le \delta/(10M).$$

Note that in each interval $\tilde{J}_{i,j}$, we have $\mathrm{i}q \le f < (\mathrm{i}+1)q$. Thus the sampled value for each sampling subinterval is $(\mathrm{i}+1/2)q$ and $\bigcup_{j=1}^{m_i} \tilde{J}_{i,j} \subseteq A_i$ for both types of sampling.

Since $B_i \supseteq C_i \supseteq \bigcup_{j=1}^{m_i} \tilde{J}_{i,j}$,

$$\mu(B_i) - \mu\left(\bigcup_{j=1}^{m_i} \tilde{J}_{i,j}\right) \le \delta/(10M).$$

Summing over i

$$1 - \mu\left(\bigcup_i \bigcup_{j=1}^{m_i} \tilde{J}_{i,j}\right) \le \delta/10.$$

Now $\bigcup_i A_i = [0,1]$, so

$$\mu\left(\bigcup_i A_i\right) - \mu\left(\bigcup_i \bigcup_{j=1}^{m_i} \tilde{J}_{i,j}\right) = \sum_i \left(\mu(A_i) - \mu\left(\bigcup_{j=1}^{m_i} \tilde{J}_{i,j}\right)\right) \le \delta/10.$$

Each of the summands is positive, so for each i,

$$\mu(A_i) - \mu\left(\bigcup_{j=1}^{m_i} \tilde{J}_{i,j}\right) \le \delta/10.$$

Because $p_n(i) = \mu(A_i)$, we want to show that $|\mu(B_i) - \mu(A_i)| \le \delta$ for all i. But

$$|\mu(B_i) - \mu(A_i)| \le \left|\mu(B_i) - \mu\left(\bigcup_{j=1}^{m_i} \tilde{J}_{i,j}\right)\right| + \left|\mu(A_i) - \mu\left(\bigcup_{j=1}^{m_i} \tilde{J}_{i,j}\right)\right|$$
$$\le \delta/(10M) + \delta/10 < \delta. \quad \square$$

As can be seen from the proof, the theorem would hold for other types of sampling and other types of quantization.

4. Examples

Example 2. This is a continuation of Example 1 with $f(x) = x$ on $[0, 1]$. Take $q = 2^{-\ell}$. Then $Q_q(f)(x) = \mathrm{i} + 1/2$ for $\mathrm{i}2^{-\ell} \le x < (\mathrm{i}+1)2^{-\ell}$, $i = 0, \ldots, n-1$. Thus $\mu\{x \mid Q_q f(x) = (\mathrm{i}+1/2)q\} = 2^{-\ell}$ and $H_q(f) = \ell = -\log_2 q$.

Let us sample with $n = 2^k$. Then the sampling points are $x_i = (\mathrm{i}+1/2)2^{-k}$ with sampled values $S_n(f)(i) = (\mathrm{i}+1/2)2^{-k}$. If $k \le \ell$, that is if the quantization is finer than the sampling, then the quantization does not change the sampled values. There are 2^k different values. $H(Q_q S_n(f)) = k = \log_2 n$.

If $k > \ell$, each of the 2^ℓ quantization intervals contains $2^{k-\ell}$ samples. Thus $p_i = 2^{k-\ell}/2^k = 2^{-\ell}$ and $H(Q_q S_n(f)) = \ell = H_q(f)$ which is in accordance with the theorem.

Example 3. A continuation of the previous case in which differences are taken. As before, the sampled values $S_n(f)(i) = (\mathrm{i}+1/2)2^{-k}$. Instead of this sequence, we take $d(0) = (1/2)2^{-k}$ and $d(i) := S_n(f)(i) - S_n(f)(\mathrm{i}-1) = 2^{-k}$ for $i = 1, \ldots, n-1$. Then $c_n((1/2)2^{-k}) = 1$ while $c_n(2^{-k}) = n-1$. The entropy of the transformed sequence is

$$H(D) = -\frac{1}{n}\log_2\frac{1}{n} - \frac{n-1}{n}\log_2\frac{n-1}{n} \to 0$$

as $n \to \infty$. As the original sequence can be reconstructed from the first value and the differences, the transformation has produced a lossless compression with a high compression factor. The original signal had an entropy of $\log_2 n$ while the compressed signal has a entropy of $\mathrm{o}(n)$.

Example 4. Let $f(x) = \sqrt{x}$ on $[0, 1]$ and take $q = 2^{-\ell}$. Then $Q_q f(x) = (\mathrm{i}+1/2)q$ for $\mathrm{i}2^{-\ell} \le \sqrt{x} < (\mathrm{i}+1)2^{-\ell}$ or $\mathrm{i}^2 2^{-2\ell} \le x < (\mathrm{i}+1)^2 2^{-2\ell}$. So $\mu\{x \mid Q_q f(x) = (\mathrm{i}+1/2)q\} = (2\mathrm{i}+1)2^{-2\ell}$ and

$$H_q(f) = -\sum_{i=0}^{2^\ell-1} (2\mathrm{i}+1)2^{-2\ell}\log_2(2\mathrm{i}+1)\log_2((2\mathrm{i}+1)2^{-2\ell})$$
$$= 2\ell - 2^{-2\ell}\sum_{i=0}^{2^\ell-1}(2\mathrm{i}+1)\log_2(2\mathrm{i}+1).$$

References

[1] C. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423; 623–656.
[2] G.G. Lorentz, Approximation of Functions, Holt, Reinhart and Winston, New York, 1966.
[3] K. Sayood, Introduction to Data Compression, Elsevier, Amsterdam, 2006.
[4] H.L. Royden, Real Analysis, Macmillan, New York, 1963.