

Shannon Entropy, Renyi Entropy, and Information

P.A. Bromiley, N.A. Thacker and E. Bouhova-Thacker

Last updated
26 / 7 / 2010

This document forms part of the **Statistics and Segmentation Series** (2008-001)
available from www.tina-vision.net.

- 2007-008 Tutorial: Defining Probability for Science.
- 2001-007 Performance Characterisation in Computer Vision:
The Role of Statistics in Testing and Design.
- 2002-007 The Effects of an Arcsin Square Root Transform on a Binomial Distributed Quantity.
- 2001-010 The Effects of a Square Root Transform on a Poisson Distributed Quantity.
- 2004-004 Shannon Entropy, Renyi Entropy, and Information.
- 2002-002 Validating MRI Field Homogeneity Correction Using Image Information Measures.
- 2004-001 Empirical Validation of Covariance Estimates for Mutual Information Coregistration.
- 2004-005 The Equal Variance Domain: Issues Surrounding the Use of Probability Densities in
Algorithm Design.
- 2009-008 Avoiding Zero and Infinity in Sample Based Algorithms.
- 2001-008 Derivation of the Renormalisation Formula for the Product of Uniform Probability
Distributions and Extension to Non-Integer Dimensionality.
- 2001-005 Model Selection and Convergence of the EM Algorithm.
- 2003-007 Noise Filtering and Testing for MR Using a Multi-Dimensional Partial Volume Model.
- 2002-004 A Novel Method for Non-Parametric Image Subtraction:
Identification of Enhancing Lesions in Multiple Sclerosis from MR Images.
- 2001-014 Bayesian and Non-Bayesian Probabilistic Models for Image Analysis.
- 1997-001 The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data.
- 1999-001 The Bhattacharyya Measure requires no Bias Correction.
- 1999-004 B-Fitting: An Estimation Technique With Automatic Parameter Selection.
- 2005-008 Tutorial: Beyond Likelihood.



Imaging Science and Biomedical Engineering,
School of Cancer and Imaging Sciences,
Stopford Building, The University of Manchester,
Oxford Road, Manchester M13 9PT, U.K.

Shannon Entropy, Renyi Entropy, and Information

P.A. Bromiley, N.A. Thacker and E. Bouhova-Thacker.
Imaging Science and Biomedical Engineering,
School of Cancer and Imaging Sciences,
Stopford Building, The University of Manchester,
Oxford Road, Manchester M13 9PT, U.K.
email: paul.bromiley@man.ac.uk

Abstract

This memo contains proofs that the Shannon entropy is the limiting case of both the Renyi entropy and the Tsallis entropy, or information. These results are also confirmed experimentally. We conclude with some general observations on the utility of entropy measures. A brief summary of the origins of the concept of physical entropy are provided in an appendix.

1 Introduction

The growth of telecommunications in the early twentieth century led several researchers to study the information content of signals. The seminal work of Shannon [12], based on papers by Nyquist [8, 9] and Hartley [6], rationalised these early efforts into a coherent mathematical theory of communication and initiated the area of research now known as information theory. Shannon states that a measure of the amount of information $H(p)$ contained in a series of events $p_1 \dots p_N$ should satisfy three requirements:

- H should be continuous in the p_i ;
- if all the p_i are equally probably, so $p_i = 1/N$, then H should be a monotonic increasing function of N ;
- H should be additive.

He then proved that the only H satisfying these three requirements is

$$H(P) = -K \sum_{i=1}^N p_i \ln p_i$$

where K is a positive constant. This quantity has since become known as the Shannon entropy. It has been used in a variety of applications: in particular, Shannon entropy is often stated to be the origin of the mutual information measure used in multi-modality medical image coregistration.

Extensions of Shannon's original work have resulted in many alternative measures of information or entropy. For instance, by relaxing the third of Shannon's requirements, that of additivity, Renyi [11] was able to extend Shannon entropy to a continuous family of entropy measures that obey

$$H_q(P) = \frac{1}{1-q} \ln \sum_{i=1}^N p_i^q$$

The Renyi entropy tends to Shannon entropy as $q \rightarrow 1$.

In addition, Kendall [10] defines the information content of a probability distribution in the discrete case as (see Section 5.1)

$$I_q(P) = \frac{1}{q-1} - \sum_{i=1}^N \frac{p_i^q}{q-1}$$

which again tends to the Shannon entropy as $q \rightarrow 1$.

We have not been able to find proofs for the assertions that these expressions regenerate Shannon entropy in the limit, and we therefore present such proofs here, and confirm the results experimentally on a sample of uniform probabilities. We conclude with some observations on the theoretical validity of entropy measures in general.

2 Shannon Entropy and Renyi Entropy

Given a sample of probabilities p_i

$$\sum_{i=1}^N p_i = 1$$

the Renyi entropy of the sample is given by

$$H_q(P) = \frac{1}{1-q} \ln \sum_{i=1}^N p_i^q$$

At $q = 1$ the value of this quantity is potentially undefined as it generates the form $0/0$. In order to find the limit of the Renyi entropy, we apply l'Hopital's Theorem

$$\lim_{q \rightarrow a} \frac{f(q)}{g(q)} = \lim_{q \rightarrow a} \frac{f'(q)}{g'(q)}$$

where in this case $a = 1$. We put

$$f(q) = \ln \sum_{i=1}^N p_i^q \quad g(q) = 1 - q$$

Then

$$\frac{d}{dq} g(q) = -1$$

and, applying the chain rule

$$\frac{d}{dq} f(q) = \frac{1}{\sum_{i=1}^N p_i^q} \sum_{i=1}^N \frac{d}{dq} p_i^q$$

The form a^x can be differentiated w.r.t. x by putting

$$\frac{d}{dx} a^x = \frac{d}{dx} e^{x \ln a} = e^{x \ln a} \frac{d}{dx} x \ln a = a^x \ln a$$

Therefore

$$\frac{d}{dq} f(q) = \frac{1}{\sum_{i=1}^N p_i^q} \sum_{i=1}^N p_i^q \ln p_i$$

Letting $q \rightarrow 1$, we have

$$\frac{d}{dq} f(q) = \frac{1}{\sum_{i=1}^N p_i} \sum_{i=1}^N p_i \ln p_i$$

Since the p_i sum to unity this gives

$$\lim_{q \rightarrow 1} \frac{1}{1-q} \ln \sum_{i=1}^N p_i^q = - \sum_{i=1}^N p_i \ln p_i$$

which is the Shannon entropy.

3 Shannon Entropy and Information

The information of a sample of probabilities p_i where

$$\sum_{i=1}^N p_i = 1$$

is given by

$$I_q(P) = - \sum_{i=1}^N \frac{p_i^q}{q-1} + \frac{1}{q-1} = \sum_{i=1}^N \frac{p_i - p_i^q}{q-1}$$

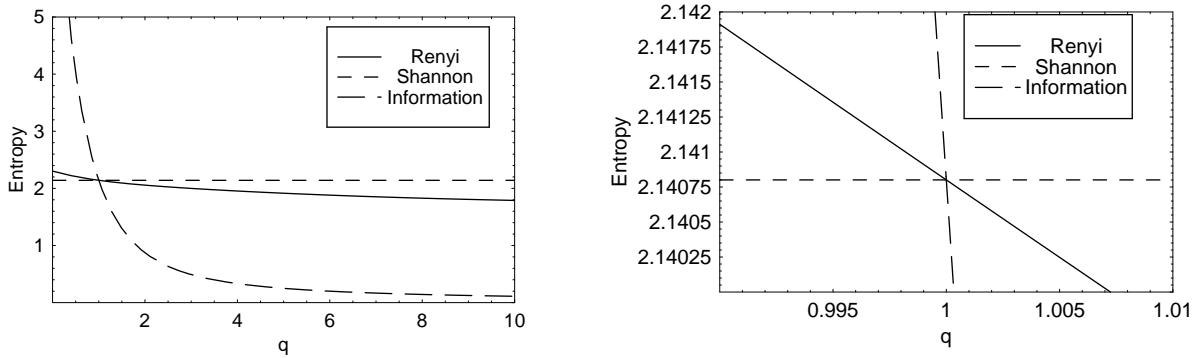


Figure 1: Various entropy measures for a sample of uniform probabilities with $N = 10$. The Renyi entropy and information converge to the Shannon entropy for $q \rightarrow 1$. The right-hand image is a magnified view of the intersection point in the left-hand image.

Put $q - 1 = a$, so that as $q \rightarrow 1$ $a \rightarrow 0$, and $p_i = 1 - x_i$. Then

$$I_a(X) = \sum_{i=1}^N \frac{(1 - x_i) - (1 - x_i)^{a+1}}{a}$$

Taking out one power of p_i immediately gives

$$I_a(X) = \sum_{i=1}^N \frac{(1 - x_i)(1 - (1 - x_i)^a)}{a}$$

The binomial expansion

$$(1 + x)^n = 1 + nx + n(n-1)\frac{x^2}{2!} + n(n-1)(n-2)\frac{x^3}{3!} \dots$$

can be applied to the first term of this equation to give

$$\frac{(1 - x_i)^a - 1}{a} = -x_i + (a-1)\frac{x_i^2}{2!} - (a-1)(a-2)\frac{x_i^3}{3!} \dots$$

In the limit of $a \rightarrow 0$ this becomes

$$= -x_i - \frac{x_i^2}{2} - \frac{x_i^3}{3} \dots$$

Which is the well known series expansion for the natural logarithm

$$\ln(1 - x) = -x - \frac{x^2}{2} - \frac{x^3}{3} \dots$$

Therefore,

$$\lim_{a \rightarrow 0} \frac{(1 - x_1) - (1 - x_i)^{a+1}}{a} = -(1 - x_i) \ln(1 - x_i)$$

and

$$I_1(P) = - \sum_{i=1}^N p_i \ln p_i$$

which is the Shannon entropy.

4 Experimental Testing

The above results were confirmed by plotting the Shannon entropy, Renyi entropy, and information against q for a sample of uniform probabilities. Ten random samples from a uniform distribution were generated and normalised such that they summed to unity. Then the Shannon and Renyi entropies and information were plotted against q . The results are shown in Fig. 1. As expected, the three measures converge as $q \rightarrow 1$. The behaviour around this point is well behaved.

5 Conclusions

This memo has demonstrated that, in the limit of $q \rightarrow 1$, both the Renyi entropy $H_q(p)$ and the information $I_q(p)$ tend to the Shannon entropy. Also, the Renyi entropy is a monotonic function of the information. However, as Kendall states [10] these measures are scale-dependent when applied to continuous distributions, and so their absolute values are meaningless. Therefore, they can generally only be used in comparative or differential processes. The monotonic relationship therefore implies that Renyi entropy and information can be used interchangeably in any practical applications.

Although these entropy measure form a self-consistent family of functions, their scale-dependence limits their utility as they cannot then be considered as well-formed statistics. For instance, concepts of Shannon entropy can be used to derive the mutual information measure commonly used in information-theoretic multi-modality medical image coregistration [15]. However, recent work [13, 4, 1, 2, 3] has shown that mutual information is in fact a biased maximum likelihood technique, and in the original application of Shannon entropy, calculating the information content of signals composed from a discrete alphabet of independent symbols, Shannon entropy is identical to the likelihood function. Therefore, although the Renyi entropy could be used to derive a continuous family of mutual information measures that could be applied, for instance, to coregistration, the statistical validity of such techniques would be questionable.

5.1 A Note on “Information”

In this document we refer to the expression for the information content of a probability distribution given in Section 3.40 of Kendall’s Advanced Theory of Statistics, Vol. 1 [10]

$$I_q(P) = \frac{1}{q-1} - \sum_{i=1}^N \frac{p_i^q}{q-1}$$

Kendall does not provide a reference for the origins of this quantity: However, it is identical to the structural α -entropy proposed by Havrda and Charvat [7]. More recently, Tsallis has proposed the use of the same quantity as a physical entropy measure [14], although this has provoked considerable controversy e.g. [5] (see also the letters in response to this article: Science, 298 p. 1171-1172, 2002; Science, 300 p. 249-251, 2003). The quantity is now commonly called the Tsallis entropy. The authors thank Frank Neilsen for pointing this out.

References

- [1] P A Bromiley, M Pokric, and N A Thacker. Computing covariances for mutual information coregistration. In *Proceedings MIUA 2004*, 2004.
- [2] P A Bromiley, M Pokric, and N A Thacker. Empirical evaluation of covariance matrices for mutual information coregistration. In *Proceedings MICCAI 2004*, 2004.
- [3] P A Bromiley, M. Pokric, and N A Thacker. Tina memo 2004-001: Empirical evaluation of covariance matrices for mutual information coregistration. Technical report, Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester, 2004.
- [4] P A Bromiley and N A Thacker. Tina memo 2003-002: Computing covariances for mutual information coregistration 2. Technical report, Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester, 2003.
- [5] A Cho. A fresh take on disorder, or disorderly science? *Science*, 297:1268–1269, 2002.
- [6] R V L Hartley. Transmission of information. *Bell Systems Technical Journal*, page 535, July 1928.
- [7] J Havrda and F Charvat. Quantification method of classification processes: concept of structural α -entropy. *Kybernetika*, 3:30–35, 1967.
- [8] H Nyquist. Certain factors affecting telegraph speed. *Bell Systems Technical Journal*, page 324, April 1924.
- [9] H Nyquist. Certain topics in telegraph transmission theory. *A.I.E.E. Trans.*, page 617, April 1928.
- [10] K Ord and S Arnold. *Kendall’s Advanced Theory of Statistics: Distribution Theory*. Arnold, 1998.

- [11] A Renyi. On measures of entropy and information. In *Proc. Fourth Berkeley Symp. Math. Stat. Prob., 1960*, volume 1, page 547, Berkeley, 1961. University of California Press.
- [12] C E Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 and 623–656, Jul and Oct 1948.
- [13] N A Thacker and P A Bromiley. Tina memo 2001-013: Computing covariances for mutual information coregistration. Technical report, Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester, 2001.
- [14] C Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.
- [15] P Viola and W M Wells. Alignment by maximisation of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

Appendix: Entropy in Physics

The term ‘entropy’ is often used to add credibility to a particular approach to understanding a problem. If we are to use a physics terms in this way perhaps we should take some time to explain what it is. Boltzmann was able to show that by quantising the states of a gas equally as a function of momentum (mv) he could regenerate the mathematical term of entropy, which had already been found to be necessary in classical thermodynamics. He and others argued for an interpretation of this construct in terms of a measure of system complexity; the possible number of indistinguishable re-arrangements of a system. This is not the only interpretation which could have been made at that time.

Temperature can be interpreted as the momentum tied up in a system which is associated with random motion. When considering a macroscopic volume of gas, we can therefore interpret temperature in terms of the distribution of velocities. In particular, we can define a mean velocity, this is the average motion of the volume, which is available to do useful work using mechanical devices. We can also define the variance of the velocity distribution around this mean. As the components of velocity for each molecule associated with this variance are incoherent, this energy is not available for useful work using simple kinematics (see Szilard below). It is only natural that any “energy sum” level description of this macroscopic system will need to consider the possible interchange of both forms of energy. Thus, if we think about the distribution of velocities for a body of gas we can say that the shape of the distribution in velocity space is characteristic of temperature, but the absolute location of the centroid of the distribution has no effect on this concept.

In order to make this quantitative we need some way of converting a known distribution into temperature. We could do this by defining a mean and then estimating a variance, but if you try this you will find that this is quite cumbersome. Alternatively, we can apply the entropy calculation $-\sum p_i \log p_i$. The entropy formula is well known as a way to measure the “peakiness” of a distribution. For example if our distribution of velocities is a simple Gaussian ($\exp - (x - x_m)^2 / 2\sigma^2$), then the application of the entropy formula gives a value of $(\sqrt{\pi}\sigma)$, ie: it is proportional to the standard deviation of the velocity distribution.

The above result raises an interesting issue which isn’t likely to be found in popular texts. In the absence of a first principles proof for the concept of entropy we have a problem. Assuming that all distributions of physical velocity converge to some well defined form, which is always of the same characteristic shape under conditions of thermodynamic equilibrium, then entropy is potentially nothing more than one of an infinite number of ways to obtain an estimate which correlates with the quantity of kinematic randomness in a system. Its main appeal being mathematical simplicity. Proof of fundamental validity would require us to deliberately construct specific velocity distributions and then confirm the physical predictions based upon $-\sum_i p_i \log p_i$ as opposed to any other estimator of this quantity. However, although entropy forms part of a theory of energy conservation, it is quite clear that for a fixed value of entropy we can adjust the distribution of incoherent velocities of particles (simply by allowing particle interactions) in a way which violates the principle of conservation of energy. Therefore the restrictions upon the physical circumstances in which entropy can be defined are generally extended to include ideal gasses as well as thermal equilibrium. Combining this with the idea that the concept can only be computed for un-physically simple systems, the kind of experiment we need to conduct is logically prohibited by the accepted definition of the quantity we wish to test. We therefore need another way of assessing Boltzmann’s idea.

Szilard’s single particle thought experiment (in which he shows that energy can be continually extracted from a single particle in a thermal bath) can also be interpreted in the light of this result. Entropy is an effective parameter which can only be defined for multi-particle systems, and it should not be considered valid for single particle systems where the associated momentum is simply recoverable. In conclusion, Szilard’s thought experiment is the philosophical equivalent of the sound of one hand clapping. However, contrary to this, many physicists interpret his analysis as a fundamental justification for the concept of entropy as information. Stating that it is the state of knowledge of the system which needs to be correctly included in order to prevent useful work from being done by a thermal bath. This brings us to our last test of Boltzmann’s entropy concept, which is based upon consideration of the computational form itself.

The most influential contribution to entropy in the last century was made by Shannon. Though, as he was considering the properties of the mathematical expression and not entropy as a physical concept, it has been suggested that he should not have been calling it ‘entropy’. He found non-physical systems for which the idea of system complexity was completely appropriate. However, the formulation of entropy measures has specific problems when applied to continuous variables. Shannon even warns us about this in the appendix of his original paper. Non-invariance under non-linear transformation means that if Boltzmann had chosen to quantise over kinetic energy ($mv^2/2$), which might have been considered an equally reasonable alternative, he would not have generated the desired result. When looking for a logical argument to explain why momentum is the correct choice for quantisation (such as the obvious one of physical dimensionality), we can find arguments that it isn’t. Although

momentum is a conserved quantity in classical physics, we now know that momentum is not an absolute quantity but varies due to the effects of relativity. It therefore cannot be used as a fundamental quantity from which to define a quantisation of states without violating Einstein's equivalence principle (ie: the states cannot be meaningful as they change under Lorentz boost). We can try to avoid this by defining the expression using integrals, rather than a sum. This does not solve the problem, as it implicitly defines the continuous variable as a specific choice of metric (ie: a continuous form of quantisation).

So what about other theoretical interpretations, such as maximum entropy? Here the concept can be applied in order to solve the most likely distribution of state variables (f) given the known physics (generally expected probability distributions p) of the system. However, as the maximum of $\sum_i p_i \log(f_i)$ subject to the constraint that $\sum_i f_i = \text{constant}$ is obtained when $p_i = f_i$ this can simply be interpreted as a consistency argument in frequentist probability, ie: in equilibrium the proportion of particles found in any state will be proportional to the theoretically predicted probabilities (this IS thermal equilibrium). Again, as was the case with the use of entropy for assessing random velocities, there are a multitude of formula which have the characteristic of being a maxima when $p_i = f_i$. And once again, it is the properties of the solution for (untestable) conditions away from equilibrium which would need to be tested in order to show that the entropy calculation is in any way fundamental.

We can conclude that physical entropy is a term in an effective (but not fundamental), physics model which has the property of correlating with the required summary variable under very restricted circumstances (thermal equilibrium in ideal gasses). This equilibrium process is present so that observed data densities (obtained for finite quantities of data) are meaningful estimates of the underlying probability used in the theory. Despite the time which has elapsed since Boltzmann, thermodynamics has been taught on degree courses with this (modest) perspective for decades. Raising the status of the theory to anything more fundamental, involving information or claiming that the construct can be applied to finite samples, is a relatively recent (bold) perspective. Presumably, those who take this view have good reason to do so, but what we can see is that the concept of entropy and arguments for its applicability and validity are non-trivial, even in physics. Once we move away from physics, we cannot expect the concept to be meaningful for any finite sampled distribution we care to characterise. The role of quantisation and use of finite samples as a substitute for probabilities seem to be key here.

Personal Comment from NAT

Good physicists are very wary of the perils of mixing effective theories in order to make predictions. In the absence of first principle derivations it is possible they make contradictory assumptions. This is one of the reasons why theoretical physics predictions are of no real value unless they are testable by experiment (e.g.: not inside a black hole, or an unobservable parallel universe), this isn't just a utility judgement. Experiment is often the only way of confirming the theoretical approach, and mathematical analysis of valid systems of equations also often generate un-physical solutions along with the physical ones¹. This requires physicists to be very careful regarding what they choose to combine and conclude. Those wishing to use analogous concepts as components in a theory of data analysis should be equally wary. We cannot simply pick up formulae from a physics book and start using them as part of a data analysis because they are convenient, or even because others tell you "they work" when applied to specific data sets. It's all too easy to generate absurd (self contradictory) ideas which defy basic logic, i.e. entertaining pseudoscience which is the analysts equivalent of time travel. In general the valid use of any mathematical theory requires justification. For example, if you think you are computing something equivalent to an entropy, what is the equivalent of 'thermal equilibrium' for your system? When we see people using these and other expressions outside the context of physics we should not be impressed and assume that all is well, instead we should ask ; Why? And keep asking until we get a good answer.

¹My favourite example here is inferring the radius of a circle from a measured area. Square-roots of course have two solutions, but that doesn't mean we must conclude the existence of another world in which circles have negative radii.