DIFFERENTIAL ENTROPY

KENNETH HALPERN

April 14, 2019

CONTENTS

There are certain subtleties that arise when we consider the entropy of a continuous distribution. Although one can meaningfully define various entropy-like quantities, none possess all the qualities desired of a proper extension of the discrete entropy. Specifically, there is no existing definition which behaves like a measure of information or even a measure of information density. A direct continuum limit of the discrete entropy engenders divergences. The most commonly employed extension, "differential entropy", really is a discrete entropy, measuring uncertainty to within an implicit coarse-graining scale. The quantity typically used in physics calculations is just a sloppy form of differential entropy which ignores units. A meaningful metric for comparing the uncertainty of continuous distributions has yet to be devised and, in light of the problems discussed below, may not exist. We assume the reader is familiar with basic probability theory.

1. DISCRETE ENTROPY

Let us begin by reviewing the discrete version of entropy.

1.1. **Definition.** Recall that the entropy of a discrete random variable[1] with probability distribution $P = \{p_1 \dots p_m\}$ over m possible values is

$$\sigma[P] = -\sum_{1}^{m} p_i \log p_i \tag{1}$$

where m need not be finite. In information terms, the entropy is the average number of bits needed to optimally represent numbers drawn from the distribution.

Entropy is only defined to within a positive constant multiple, a freedom evident in both its axiomatic formulation and empirical origins. This corresponds to a choice of base for the logarithm[2]. We shall choose bits as our unit, and all logarithms are base 2.

1.2. **Interpretation.** The entropy often is interpreted in one of two equivalent ways:

(1) The average information we provide in making a choice of value for the associated random variable.
(2) Our uncertainty about the value of the random variable given only the probability distribution over it.

We refer to uncertainty and information interchangeably with the understanding that both are taken to have the same sign in this context[3].

1.3. **Uniform Distribution.** For a uniform distribution over m values, equation 1 becomes

$$\sigma[U_m] = \log m \tag{2}$$

This holds even in the limit $m \to \infty$, in which case the entropy is infinite as expected.

It is easy to show that for any non-uniform distribution P_m over m values,

$$\sigma[P_m] < \sigma[U_m]$$

1.4. **Full Knowledge.** The minimum entropy distribution is that corresponding to full knowledge: $m = 1$ and $P = \{1\}$. Denoting this C for certainty, it is clear that given any distribution $P \neq C$ (regardless of the number of values m):

$$\sigma[P] > \sigma[C]$$

.

[1] Although it is accurate to refer to the entropy of a random variable, this terminology can be slightly confusing. A probability distribution necessarily is associated with a random variable in order to have meaning. However, the entropy is purely a function of the probability distribution itself; the associated values of the random variable are immaterial (as long as they are distinct).

[2] Recall that $\log_b(x) = \frac{\ln a}{\ln b} \log_a(x)$.

[3] In other words, the information in question is what we must provide in making a choice, not that contained in the probability distribution. The latter increases as the entropy decreases and vice versa.

Note that the full knowledge distribution is identical to the 1-element uniform distribution U_1.

1.5. **Dimension.** Consider the entropy of a uniform distribution over points on an n-dimensional grid. Let us suppose there are m points per side. Then we have a uniform distribution over an m^n element space. The entropy is $\sigma(U_{n,m}) = n \log m$ and scales linearly with dimension.

1.6. **Axioms.** The discrete entropy can be derived from fairly weak axioms that embody our intuitive notion of the uncertainty represented by a probability distribution. In the following, we define a set of functions $\sigma_n(p_1 \dots p_n)$ and select the one with the appropriate number of arguments for a given distribution. Note that σ_n really is a function of $n-1$ independent arguments because $\sum p_i = 1$. We keep them separate but implicitly only require our conditions to be true over the constrained support, the set of values $\{p_1 \cdots p_n\}$ such that $\sum p_i = 1$ and $p_i \geq 0$.

There are several equivalent axiomatic formulations[4]. Let us first list the individual axioms that make an appearance. They all are true statements, but we may choose various minimal subsets of them as axioms.

We use P_n and Q_m for distributions of some random variables X and Y with n and m independent outcomes respectively. In the case where X and Y are not independent, we may write PQ_{nm} as the joint distribution of (X, Y) and P_n and Q_m as the marginal distributions: $P_j = \sum_{i=1}^{m} PQ_{ji}$ and $Q_i = \sum_{j=1}^{n} PQ_{ji}$.

(1) POS: $\sigma_n \geq 0$. The sensible requirement that the information content be non-negative. We mention this even though it does not appear in the sets of axioms we consider.
(2) BOUND: The set of values of σ_n over all n is bounded on both sides. The information content can't be infinite and has a lower bound.
(3) SYM: Each σ_n is symmetric under all permutations of its arguments. Our uncertainty doesn't depend on the labels for our outcomes. We denote by m-SYM the weaker requirement that we have symmetry for all σ_i with $i \leq m$.
(4) CONT: Each σ_n must be continuous in all its arguments (subject to $\sum p_i = 1$). We denote by m-CONT the weaker requirement that the σ_i be continuous in their arguments for $i \leq m$.
(5) SUP: $\sigma_{n+1}[p_1 \dots p_i, 0, p_{i+1} \cdots p_n] = \sigma_n[p_1 \dots p_n]$ (for any placement including to the left of p_1). Only the support matters; we can't affect things by adding fictitious outcomes.
(6) ADD: $\sigma_{n \cdot m}[P_n \otimes Q_m] = \sigma_n[P_n] + \sigma_m[Q_m]$ for independent distributions. The information contained in two independent distributions is just additive. Stated differently, we may choose from distribution P and then distribution Q.
(7) JOINT: $\sigma_{n \cdot m}[PQ_{nm}] \leq \sigma_n[P_n] + \sigma_m[Q_m]$. There is more uncertainty in choosing from two independent distributions that from any joint distribution over the same set of outcome pairs.
(8) COND: $\sigma_{n \cdot m}[PQ_{nm}] = \sigma_n[P_n] + \sum_{i=1}^{n} p_i \sigma_m(QP_i)$ where QP_i denotes the conditional distribution $Q|p_i$ which is just the i^{th} row of QP. This just says that if we choose p_i from P and then conditionally choose from $Q|_{p_i}$ we have the same uncertainty as if we chose jointly. This axiom my also be written $\sigma[PQ] = \sigma[P] + \sigma[Q|P]$.
(9) SEP: $\sigma_n[p_1 \cdots p_n] = \sum_{i=1}^{n} \gamma(p_i)$ for some function γ. That is, the entropy is separable into a function of outcome.
(10) RED: This axiom has two equivalent common forms, so we list them both. If we needn't distinguish between two of the outcomes then the information needed to choose (or our uncertainty) decreases. The change in overall information is just that of choosing between the two outcomes times the probability that we must do so[5].
  (a) $\sigma_n[p_1 \cdots p_n] = \sigma_{n-1}[p_1 + p_2, p_3 \cdots p_n] + (p_1 + p_2)\sigma_2[\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}]$.

[4]The reader is referred to [5] for an excellent discussion of these.

[5]We must be careful to only use this in the presence of SYM or the position of the conjoined outcomes would be significant.

(b) Let us divide the outcomes into two disjoint sets with corresponding distributions Q and R (with $\sum q_i + \sum r_j = 1$). Then $\sigma[P] = \sigma[Q \cup R] = \sigma_2[(\sum q_i, \sum r_i)] + (\sum q_i)\sigma[\frac{Q}{\sum q_i}] + (\sum r_i)\sigma[\frac{R}{\sum r_i}]$ where $\frac{Q}{\sum r_i}$ just means the q_i normalized so $\sum q_i = 1$.

(11) MON: $\sigma_n[\frac{1}{n} \cdots \frac{1}{n}]$ increases monotonically with n.

(12) UNI: σ_n for any n is maximized by the uniform distribution $p_i = \frac{1}{n}$.

(13) NORM: $\sigma_2[(\frac{1}{2}, \frac{1}{2})] = 1$. This fixes the constant multiple, corresponding to a choice of base-2 for our logarithm.

The following is a list of some of the sets of axioms that uniquely lead to the formula for entropy within a positive constant multiple. The use of NORM fixes that constant by setting the logarithm base to 2.

(1) CONT, COND, MON: Shannon[4].
(2) CONT, UNI, COND, SUP: Khinchin[11].
(3) RED, 3-SYM, 2-CONT: Faddeev[6].
(4) RED, 3-SYM, BOUND: Diderrich[7].
(5) ADD, SEP (with γ continuous): Chaundry, McLeod[8].
(6) SYM, SUP, ADD, JOINT, 2-CONT: Aczel, Forte, Ng[9].

1.7. **Renyi and Tsallis Entropy.** There are various mathematical generalization of the entropy, the most common of which are the Renyi and Tsallis entropies. Both are sets of entropy functions parametrized by a number α. The ordinary discrete entropy is recovered in both cases as the limit $\alpha \to 1$.

1.7.1. *Renyi Entropy.* The Renyi entropy takes as an argument a "generalized probability distribution", where $p_i \geq 0$ but $\sum p_i \leq 1$ instead of equaling one. It is a family of functions of the form (where $\alpha \geq 0$ and $\alpha \neq 1$):

$$\sigma_\alpha[P] \equiv \frac{1}{1-\alpha} \log(\sum p_i^\alpha)$$

From an axiomatic standpoint, Renyi started by demonstrating that the simple generalization of entropy to generalized distributions given by $\sigma[P] = \frac{-\sum p_i \log p_i}{\sum p_i}$ (with P a generalized distribution) follows from five axioms:

(1) SYM
(2) 1-CONT. Note that only 2-CONT and above were meaningful for the ordinary entropy.
(3) $\sigma[(\frac{1}{2})] = 1$. This is the equivalent of NORM for ordinary entropy.
(4) ADD
(5) $\sigma[P \cup Q] = \frac{(\sum p_i)\sigma[P] + (\sum q_j)\sigma[Q]}{\sum p_i + \sum q_j}$ when $\sum p_i + \sum q_j \leq 1$ (and thus represents a valid generalized distribution). Though it is stated differently[6], this is the equivalent of the RED axiom for ordinary entropy.

Obviously if P is a probability distribution, then $\sigma[P]$ is the ordinary entropy.

Renyi then loosened the fifth axiom to allow more general averaging:

RED': There exists a function $g(x)$ such that $g^{-1}\big[\sigma[P \cup Q]\big] = \frac{(\sum p_i)g(\sigma[P]) + (\sum q_j)g(\sigma[Q])}{\sum p_i + \sum q_j}$ when $\sum p_i + \sum q_j \leq 1$ and where g is continuous and strictly monotonic.

[6]This statement is elegant, but requires the use of generalized distributions to be meaningful.

The function g represents a more general averaging[7]. Renyi's original function $\sigma[P]$ is obtained when g is linear in x, representing an arithmetic mean. There may be a variety of alternatives for g, but Renyi focused on the family $g_\alpha(x) = 2^{(\alpha-1)x}$ and demonstrated that the functions σ_α uniquely satisfy the associated axioms for all $\alpha \geq 0$ and $\alpha \neq 1$. As mentioned, we recover the ordinary entropy as $\sigma[P] = \lim_{\alpha\to 1} \sigma_\alpha[P]$ (where P now is a probability distribution).

1.7.2. *Tsallis Entropy.* The Tsallis entropy[8] is a family of the form

$$\sigma_\alpha[P] \equiv \frac{1}{\alpha - 1}(1 - \sum p_i^\alpha)$$

Starting with the Shannon-Khinchin axioms (CONT, UNI, COND, SUP), one modifies COND in an α dependent manner to:

COND': $\sigma_\alpha[P, Q] = \sigma_\alpha[P] + \sigma_\alpha[Q|P] + (1 - \alpha)\sigma_\alpha[P]\sigma_\alpha[Q|P]$.

This uniquely yields the Tsallis entropy with parameter α.

2. CONTINUOUS DISTRIBUTIONS

The continuous analogue of a discrete probability distribution is a probability density $p(x)$ defined on some domain[9]. For the moment let us take this domain to be the 1-dimensional real interval $S \equiv [0, L)$. The density can be thought of as the limit $m \to \infty$ of discrete m-element distributions[10]. For a fixed m, the "values" are the subintervals $S_i \equiv \left[\frac{(i-1)L}{m}, \frac{iL}{m}\right)$. Each p_i appearing in the distribution $P = \{p_1 \ldots p_m\}$ corresponds to the probability that a real number sampled from S falls within the corresponding S_i. Defining a function $p_m(x) \equiv p_{[\frac{xm}{L}]+1}$ (where we use $[]$ to denote a step function), and defining $dx_m \equiv \frac{L}{m}$ we see that $p_m(x)$ really is a probability per unit length and $p_m(x)dx_m$ is an actual probability. In the limit $m \to \infty$ we get a function $p(x)$ which is a density and whose corresponding probability on an infinitesimal interval dx is $p(x)dx$.

The extension to multiple dimensions is straightforward, with x denoting a point in the sample space and dx a volume element. For simplicity, we require this space to be well behaved in whichever ways we choose. For the most part, any distribution is assumed to have support on a well-defined volume of fixed dimension. Basically, we want to be able to take Riemann integrals and continuum limits without worrying about measure theory or topology. To keep things intuitive, it is best to think of the space as a subspace of some R^n with a cartesian metric and volume element. We assume that any domains are big enough to include the support for our distributions, and we will be sloppy about whether boundaries are open or closed.

---

[7] A more general mean of some set $\{x_1 \ldots x_n\}$ can be constructed from any well-behaved function $f(x)$ as $f^{-1}\left(\frac{\sum f(x_i)}{n}\right)$. To see this, consider that for $f(x) = x$ we get the arithmetic mean, for $f(x) = \ln x$ we get the geometric mean, and so on. It is this generalization of the arithmetic mean that Renyi sought to capture.

[8] This treatment is based on the paper by Abe [10].

[9] The more general mathematical concept is that of a probability measure, which takes less for granted about the underlying space. Probability densities suffice for the purpose of our discussion.

[10] It also may be defined on its own, in which case the discrete distributions approximate it at increasingly small granularities. The latter still provide a useful means of interpreting the density as well as a practical mechanism for computing functions of it (such as the entropy).

3. CONTINUUM LIMIT

Having reviewed discrete entropy and probability densities, let us turn to the task of defining a meaningful continuum version of the entropy. A number of issues arise, and they are not easily dealt with.

3.1. **Simple Attempts.**

Let us begin by taking the most direct route, and mechanically convert the sum to an integral. Then, $p_i \to p(x)dx$ and $\sum_i \to \int_D$ for some domain D. The result is $-\int p(x)dx \log(p(x)dx)$. This can be converted to $(-\log dx) - \int dx p(x) \log p(x)$. The notation $\log dx$ is mathematically meaningless, but we adopt it as a means of representing the divergence in the process of taking the continuum limit.

It is natural to hope we made a mistake in our analysis (we didn't) and things were meant to converge (they don't). Clearly, we need to take the limit more carefully to understand what really is happening. But before we do this, let's explore a simple alternative.

It is tempting to use the obvious carryover from the discrete case:

$$\sigma_d[P] = -\int dx \big[p(x) \log p(x)\big] \tag{3}$$

This also corresponds to the non-divergent term in our first attempt, lending it credibility. In fact, there is good reason to think that the divergence may be a harmless artifact. We could argue that if we simply wish to compare the uncertainty or information content of distributions, we need only concern ourselves with differences in entropy. Clearly $(-\log dx)$ vanishes when we subtract two entropies, doesn't it?

This turns out not to be the case. Before we proceed to a detailed discussion, let us try a simple calculation. Consider a uniform distribution with support on volume V. Our candidate for entropy is $\sigma_d = \log V$. That this is not unitless furnishes our first clue something is wrong. The significance of units is not confined to physical applications. They provide a convenient means of accounting, and warn us of a dependence on both dimension and an arbitrary choice of domain. Recall that the entropy should be independent of the labels we assign to the associated variable as long as they are distinct. The continuous version of this is more complicated, and we will discuss it later. However, one aspect of it is that the choice of scale should not matter. If we use $[0, 2]$ instead of $[0, 1]$ as our domain and halve the probability density, the answer should not change. But σ_d does. This will be clarified further shortly.

As another example, suppose $L = 1$ in our units. Then $\log L^2 = \log L^3 = 0$. The entropy of uniform distributions on the unit square and unit cube are equal. But from the discrete case, we expect entropy to scale linearly with dimension. From a common sense standpoint, there has to be more uncertainty in a cube than in a square the size of one of its sides. The issues with units and dimensional scaling are two symptoms of the same problem.

3.2. **Continuum limit.**

To properly analyze the continuum limit of the discrete entropy, let us consider a step along the way. We suppose that our domain of integration is $[0, L]$ in one-dimension and we have broken it into m intervals. Each interval has length $\frac{L}{m}$ and the probability associated with it is sampled from the function $p(x)$ somehow. For well-behaved functions the exact method doesn't matter, but let's suppose the sample points are $x_j = \frac{j \cdot L}{m}$. Then, $p_j = p(x_j) \cdot \frac{L}{m}$ for $j = 1 \ldots m$. The entropy at this stage is

$$\sigma_m[P] = -\sum_{j=1}^{m} p_j \log p_j$$

which yields

$$\sigma_m[P] = -\sum_{j=1}^{m} p(x_j)\frac{L}{m} \log \big[p(x_j)\frac{L}{m}\big]$$

and reduces to

$$\sigma_m[P] = -\log \frac{L}{m} - \sum_{j=1}^{m} p(x_j)\frac{L}{m} \log p(x_j) \tag{4}$$

In the limit $m \to \infty$, the second term becomes $-\int_0^L dx \big[p(x) \log p(x)\big]$. When we wrote $(-\log dx)$ earlier, what we really meant was $\lim_{m\to\infty}(-\log \frac{L}{m})$. We denote the continuum limit of the σ_m as σ_c (not to be confused with our notation $\sigma[C]$ for the full-knowledge entropy).

$$\sigma_c[P] = -\log dx - \int dx \big[p(x) \log p(x)\big] \tag{5}$$

Notationally we have separated the terms, but we may not always be justified in taking their limits separately.

Consider two cases: the uniform and full-knowledge distributions.

3.2.1. *Uniform Distribution.* In the uniform case, there is no mystery: $p(x) = \frac{1}{L}$ and we have

$$\sigma_c[U_L] = -\log dx + \log L$$

Note that the argument of $\log$ really is unitless ($\frac{L}{dx}$), but we separated the terms.

3.2.2. *Full Knowledge.* The full-knowledge[11] case requires a bit more care. How do we define $p(x)$ when only one point matters? Notationally, we may use the Dirac delta function $p(x) = \delta(x)$, where we have chosen the relevant point to be $x = 0$.

A cursory glance shows that this looks problematic. $\int f(x)\delta(x)dx = f(0)$, so the entropy should be

$$\sigma_c[C] = -\log dx - \log \delta(0)$$

We now have two terms that are, strictly speaking, mathematically meaningless. $\delta(0)$ is not a well-defined expression. The δ function is a distribution, the dual of a function, and can only meaningfully

[11] This sometimes is referred to as the "degenerate" distribution. It corresponds to certainty of a particular value.

be used in an integrand. This is one of the cases where we cannot separate the terms in σ_c before we take the limit. The discrete version of $p(x) = \delta(x)$ at stage m in our limit is $\frac{m}{L}\delta_{j,1}$ where δ_{ij} is the Kronecker delta (1 for $i = j$ and 0 otherwise). Then, $p_i = \frac{L}{m}\frac{m}{L}\delta_{j,1} = \delta_{j,1}$. In other words, we approximate $p(x)$ with a rectangle of height $\frac{m}{L}$ and width $\frac{L}{m}$.

We then get

$$\sigma_c[C] = -\lim_{m\to\infty}\sum_{j=1}^{m}\delta_{j,1}\log\delta_{j,1}$$

The summand is well-defined and always 0. Thus $\sigma_c[C] = 0$. This is consistent with the corresponding discrete entropy. It should make no difference whether we call something discrete or continuous. Certainty is certainty.

It is easy to verify that this result does not change if we increase the dimension.

3.3. **Dimension.** We now have a big problem. The full-knowledge entropy is 0, while the continuous entropy of a uniform distribution is infinite. The divergence does not cancel out when we compare entropies. However, this is only one instance of a broader issue regarding dimension.

It sometimes is useful to think of the discrete case as a 0-dimensional version of the continuous case. We can fix[12] $dx = 1$, convert the integral to a sum, and change the probability densities to discrete probabilities.

By this token, when we compare the entropies of the uniform and full-knowledge distributions we really are comparing 1-dimensional and 0-dimensional distributions.

Let us consider a probability distribution in n dimensions. It is not hard to generalize equation 5:

$$\sigma_c[P] = -\log d^n x - \int_\Omega d^n x\big[p(x_1\cdots x_n)\log p(x_1\cdots x_n)\big]$$

where Ω is the n-dimensional support.

For a uniform distribution this becomes[13]:

$$\sigma_c[U_\Omega] = -n\log dx + \log\Omega$$

Considering the case $\Omega = [0, L]^n$, we have

$$\sigma_c[U_{L,n}] = -n\log dx + n\log L$$

The problem is clear: the divergence differs with dimension. This also is why our problem with units arose when comparing σ_d of a uniform distribution over an area with that over a volume. The two could not be compared because the resulting equation was not unitless:

[12]We reason that $dx = dl^n$ for an n-dimensional volume element, with dl a 1-dimensional length, so $dx = dl^0 = 1$ for any finite dl and hence for the limit.

[13]For a uniform distribution we may use a single limit dx because there is uniform convergence in all variables. Hence the $n\log dx$.

$$\sigma_d[U_{[0,L]^3}] - \sigma_d[U_{[0,L]^2}] = \log L$$

This is not a problem with σ_c, which accurately tracks units:

$$\sigma_c[U_{[0,L]^3}] - \sigma_c[U_{[0,L]^2}] = -n \log dx + \log L$$

However, we then have a divergence because we cannot compare σ_c across dimensions.

So what do we do? We cheat. But first let's understand the intuitive meaning of the divergence in σ_c.

3.4. **Meaning of the Divergence.** To understand the divergence, we interrupt the process of taking the limit in the calculation of σ_c. This time, let's separate the first term[14] instead of using it to form dx.

$$\sigma_m[P] = \log m - \sum_{j=1}^{m} p(x_j) \frac{L}{m} \log(p(x_j)L)$$

If we stop part-way, there is a clear interpretation: the divergent term $\log m$ can be taken to represent the number of bits used to approximate a continuous point in our support $[0, L]$. That is the coarse-graining resolution we have chosen. What if the dimension is higher? As we saw, this becomes $n \log m$. The interpretation endures: it is the information needed to approximate n real numbers[15].

With this interpretation, the source of the divergence is clear. We require infinite information to specify a single point in a continuum. We recall from cardinal arithmetic that $2^{\aleph_0} = \aleph_1$. There are an uncountable number of countably long binary sequences. Conversely, we can represent any real with a countable number of bits. What about an integer? There are $\aleph_0$ finite binary sequences. Any integer can be described by a finite sequence of bits[16]. The description of an integer is finite while that of almost any real is infinite[17].

On this note, we should point out that when we speak of requiring m bits to approximate a real, it does not matter *which* reals we are approximating. There are 2^m of them at that stage, and the m bits label them.

In summary, the divergence is common to every distribution of a given dimension rather than every distribution overall.

3.5. **Differential Entropy.** Returning to the question of how to define a version of entropy for continuous distribution, we clearly have been thwarted so far. So what do people do? We cheat. We use a quantity called "differential entropy". This is nothing more than σ_d, our failed attempt to convert discrete entropy to an integral and ignore the divergence.

[14]Now the purpose of $\log L$ is clear: it defines the units. Note that it still depends on dimension. For n-dimensions it would be $-n \log L$ representing (log of) the unit of n-dimensional volume.

[15]This is true even if we don't take a uniform limit, and use separate $m_1 \cdots m_n$ instead. Each is the resolution in its respective direction.

[16]Though we require sequences of arbitrary finite length to describe all integers, a given integer contains finite information.

[17]We say "almost" because from an information theory standpoint, the specification of the *value* of certain reals is simple, requiring finite information. We know this must be the case because the integers are real. There also are reals with simple algorithmic descriptions, such as π. Obviously there cannot be more than a denumerable number of reals with such finite descriptions.

TABLE 1. Behavior of various entropies under doubling of volume (n dimensions, stage m of limit)

Scenario			Entropy		
Domain	Density p	Volume unit	σ_d	$\tilde{\sigma}_d$	σ_c
V	$\frac{1}{V}$	L_0^n	$\log V$	$\log \frac{V}{L_0^n}$	$n \log m$
$2V$	$\frac{1}{2V}$	L_0^n	$\log(2V)$	$\log \frac{2V}{L_0^n}$	$n \log m$
V	$\frac{1}{V}$	$2(L_0^n)$	$\log V$	$\log \frac{V}{2L_0^n}$	$n \log m$
$2V$	$\frac{1}{2V}$	$2(L_0^n)$	$\log(2V)$	$\log \frac{V}{L_0^n}$	$n \log m$

Recall that

$$\sigma_d[P] = -\int dx \big[p(x) \log p(x)\big]$$

This is what appears in physics textbooks and countless other places. However, it isn't what it seems. It really is a discrete entropy in disguise.

First, let us modify it to explicitly incorporate the units and dimension. These are important, but tend to be glossed over. Though we won't solve our deeper problems by including them, we at least can clarify what we are dealing with. Let us choose some value L_0 as our unit[18]. We define the "Differential Entropy" as:

$$\tilde{\sigma}_d[P] = -\int_V d^n x \big[p(x) \log(p(x) L_0^n)\big] \tag{6}$$

Clearly, σ_d scales linearly with dimension, changes additively as V scales (for fixed dimension), and is independent of units. It generally is finite and comparable across dimensions[19].

Aside from some minor cosmetics, all we really did in choosing σ_d over σ_c was ignore the divergent terms. But as we saw, those divergent terms represent the information needed to locate a vertex on a grid of m^n points (that is, m points along each axis). Comparing more closely, we see that at stage m,

$$\sigma_c[P] - \tilde{\sigma}_d[P] = n \log m - n \log \frac{L}{L_0}$$

The two are the same when $m = \frac{L}{L_0}$. This reveals how we should interpret the differential entropy. It is no more than discrete entropy $\sigma_m[P]$ at stage m of the limit and with $m = \frac{L}{L_0}$. As such, it represents the information we must provide to get to within volume L_0^n (i.e. our uncertainty to within a volume L_0^n). We simply have discretized to resolution L_0 in each direction.

Note that it is important not to confuse L (or dx) with our choice of units. Both $\tilde{\sigma}_d$ and σ_c are independent of the units, but σ_d is not. Table 1 illustrates the various behaviors of the entropies under a change of scale for a uniform distribution.

[18]Because we cannot compare entropies across dimensions we could pick a different unit for each dimension. However there is no reason not to use L_0^n, and we do so.

[19]That is, it is a finite number. Whether it is useful to compare it across dimensions is another question.

In light of this, it is no surprise that the differential entropy $\tilde{\sigma}_d$ behaves well and allows comparison of distributions of differing dimension. The latter is necessary for it to be of use in fields such as statistical mechanics. There is nothing wrong with this, as long as we do not fool ourselves into believing that it is anything more than a discrete entropy which conveniently identifies the coarse graining scale with our choice of units.

There is another aspect of discretization that we should mention. Recall that in the definition of discrete entropy we explained that the values assigned to the random variable are immaterial as long as they are distinct. The analogous statement for a continuous distribution is not obvious. What does it mean for the values to be "distinct"? It is not enough for them to be unequal as reals (or n-tuples of reals). The limit process is countable and requires some sort of countable covering. If we apply a coarse graining, then we may speak of disjoint volumes above the discretization scale as distinct – or say that two values are distinct to within that scale. While the volume-preserving transformations we shall discuss shortly provide a useful analogy to permutations of discrete random variable values, the coarse graining scale is essential for this analogy to have meaning.

4. Other Approaches

Returning to the question of whether we can define a meaningful true differential entropy, we may try a number of other approaches, but they all have problems.

4.1. **Entropy Densities.**

4.1.1. *Simple Entropy Density.* We could try to use some sort of entropy density. Just as the probability density is the useful concept for continuous distributions, an entropy density could correspondingly prove meaningful. However, the obvious definition doesn't work. The entropy scales linearly with dimension, not with volume. For example, a uniform distribution over $2V$ does not double the entropy of one over V. As discussed, it has only 1-bit more of entropy. The entropy doubles if we go to V^2. So any such quantity would have to be logarithmic and couldn't easily correspond to a density.

4.1.2. *Entropic Dimensional Density.* As an alternative, we could attempt to define a new quantity whose logarithmic base varies with dimension as 2^n. In the above example, if $V = L^n$ then $\log_{2^n} V = \log_2 L$ while $\log_{2^{2n}} V^2 = \log_2 L$ also. This could be a useful measure of some sort of entropic dimensional density, but is not directly applicable as an entropy density.

4.2. **Relative Entropy and Mutual Information.**

In information theory there are a number of comparative measures, mutual information and relative entropy being the most prominent.

4.2.1. *Mutual Information.* Mutual Information describes the interdependence of two random variables. This information is contained in the joint distribution to the extent it deviates from a simple product distribution. But this is of no use to us. We simply wish to compare the individual information contents of two separate distributions. The only way to decouple the random variables (and hence the distributions) is if they are independent.

4.2.2. *Relative Entropy.* The Relative Entropy[20] is a more plausible candidate. In some sense, it is the additional information required if we assume the wrong distribution[21]. We can think of it as an asymmetric distance between distributions over the same random variable. It is defined in the discrete case as

$$D(P||Q) = \sum_i p_i \log \frac{p_i}{q_i}$$

We could symmetrize it by using $D(P||Q) + D(Q||P)$, which looks like a metric though it isn't quite one.

In the continuum limit, the problematic volume elements disappear within the log because of the ratio. In fact, if we think of P and Q as probability measures with appropriate continuity, we have[22]

$$D(P||Q) = -\int \big(\log \frac{dQ}{dP}\big) dP$$

Note that this depends on a specific assignment of the distributions to an underlying variable. We cannot reorder the elements of P or Q arbitrarily because their alignment with one another matters. Nor is this resolved by the invariance under volume-preserving transformations discussed in section 5. The relative entropy (symmetric or asymmetric) inextricably depends on the positioning of the two distributions. Even in the discrete case, if we permute Q but leave P the same we get a different answer. Such a measure cannot be suitable for our purposes.

4.2.3. *Attempt at Order-Independent Relative Entropy.* We could attempt to remove the dependence on positioning by summing or integrating over all relative orderings with (an assumed) equal weight.

We can keep the order of P fixed and vary Q. Considering the discrete asymmetric case, we have (where all expectations are over P):

$$D'(P||Q) = -<\ln p> - \frac{1}{n!} \sum_{q \in Perm(Q)} <\ln q>$$

The second term can be rewritten $-\frac{1}{n!} \sum_i \sum_{q \in Perm(Q)} \ln q_i$.

By symmetry, the latter sum must be constant across all i. The first sum then has no effect ($\sum p_i = 1$). So the second term is simply a constant (for given n), and D' gives us nothing new.

4.3. **Renyi and Tsallis Entropies.** We could attempt to use the Renyi or Tsallis entropies to construct a meaningful generalization and then take the limit as $\alpha \to 1$.

The direct extension of these to the continuous case clearly involves replacing $\sum p_i^\alpha$ above with $\int [p(x)dx]^\alpha$, a notationally sloppy way of including the infinitesimal scaling. This doesn't give us anything directly. However, the interesting thing about σ_1 is that there are two limits: $\alpha \to 1$ and

[20] Also called the Kullback-Leibler divergence.

[21] In fact, the mutual information is the relative entropy between a joint distribution and the product of marginal distributions – that is, the extra information needed beyond that of the joint distribution if we mistakenly assume that the variables are independent. This reinforces our view of the mutual information as inappropriate for our purposes.

[22] The use of dP notationally in place of $p(x)dx$ make sense if we regard P as the cumulative probability distribution $\int_{-\infty}^{x} p(x)dx$ and $p(x)dx = \frac{dP}{dx}dx = dP$. This is standard practice.

$dx \to 0$. Let us assume that they may be swapped (this requires careful justification, but we gloss over that for now). In the case of the Renyi entropy, L'Hospital's rule gives us a limit which is $-\int p(x)dx \log[p(x)dx]$ as before. The same holds true for the Tsallis Entropy. Therefore, these generalizations do not yield a fruitful extension.

4.4. **Axiomatic Extensions.** Since the discrete entropy proves a problematic starting point, perhaps we could directly construct a functional $\sigma[P] : \Omega \to \mathbb{R}$, where P has support over an arbitrary subset of Ω, to allow for comparison of distributions of differing dimension. Although we won't attempt such a formulation here, let us list the continuous counterparts of our discrete axioms. We assume Ω has a measure on it.

(1) POS: $\sigma[P] \geq 0$. No change is needed.
(2) BOUND: $\sigma[P]$ is bounded on both sides. No change is needed.
(3) SYM: This is difficult to find an analogue for. We could consider arbitrary volume preserving transformations. They are discussed in the next section.
(4) CONT: Uniform continuity is the natural extension.
(5) SUP: We could say that $\sigma[P'] = \sigma[P]$ if P and P' differ only in their kernels (i.e. are the same for all nonzero probabilities).
(6) ADD: $\sigma[P \otimes Q] = \sigma[P] + \sigma[Q]$ for independent distributions. No change is needed.
(7) JOINT: $\sigma[PQ] \leq \sigma[P] + \sigma[Q]$. No change is needed.
(8) COND: $\sigma[PQ] = \sigma[P] + < \sigma[Q|p] >_P$.
(9) SEP: $\sigma[P] = \int_X dx g(p(x))$ for some function γ.
(10) RED: For any disjoint subsets $\Omega_1 \cup \Omega_2 = \Omega$, define $P_1 = \int_{\Omega_1} p(x)dx$ and $P_2 = \int_{\Omega_2} p(x)dx$ as the respective probabilities over the two regions ($P_1 + P_2 = 1$). Then we require that $\sigma[P] = P_1 \cdot \sigma[\frac{P(\Omega_1)}{P_1}] + P_2 \cdot \sigma[\frac{P(\Omega_2)}{P_2}] + \sigma[(P_1, P_2)]$, where $\frac{P(\Omega_1)}{P_1}$ is the normalized partial probability distribution.
(11) MON: $\sigma[U(V)]$ increases monotonically with V.
(12) UNI: $\sigma[P]$ is maximized over all P with support of measure ω by $U(\omega)$, the uniform distribution.

5. Coordinate Transformations

One important characteristic of a probability distribution is the manner in which it transforms under a change of parametrization (or coordinate transformation). Naturally, the same is true of any function of such a distribution. In order to be meaningful, there should be no dependence on our arbitrary choice of parametrization. However, we expect that the structure of the underlying space may play a role[23]. In some sense, a change of parametrization is the continuous version of a permutation of the values for a discrete random variable[24].

5.1. **Transformation of Probability Measure.** For simplicity, let us restrict our domain Ω to points with nonzero probability density. Then the latter behaves like the reciprocal of a volume element[25]. This makes sense, since we require $\int_\Omega d^n x \big[p(x_1 \cdots x_n)\big] = 1$. From here on, we treat x and y as vectors[26] and write dx and dy for the volume elements.

[23]There may be a particular set of variables which lend themselves to the definition of our probability distribution in any given application. However, once defined the probability distribution should not be bound to them. If it were, they would constitute additional structure on top of that of a probability measure.

[24]Such as appears, for instance, in the SYM discrete axiom.

[25]In fact the probability density and volume element both are mathematical "measures", with the additional requirement that the former must sum to 1. This constraint forces them to act like reciprocals as far as their behavior under coordinate changes (or scaling under changes of unit).

[26]Which is possible in almost any situation where it makes sense to define a density and coordinate transformation.

If we change the parametrization of P from x to y, denoting by $y(x)$ and $x(y)$ the transformation and its inverse, then the volume element at any point changes to $dy = dx \cdot |detJ|$ while the probability density changes to $q(y) = \frac{p(x(y))}{|\det J|}$ where $J_{ij} = \frac{\partial y_i}{\partial x_j}$. Any expectation value then transforms as

$$\int_{\Omega_x} f(x)p(x)dx = \int_{\Omega_y} f(x(y))p(x(y))|\det J|\frac{dy}{|\det J|}$$

Obviously, the determinants cancel out and we get

$$\int_{\Omega_x} f(x)p(x)dx = \int_{\Omega_y} f(x(y))p(x(y))dy$$

This just is $< f >$ and means that the expectation value of any quantity is invariant under a well-behaved change of variable.

5.2. **Transformation of Differential Entropy.** We can write the differential entropy (with sloppy units) as

$$\sigma_d[P] = -<\log p>$$

The quantity $(-<\log p>)$ is an expectation value and, as discussed, is invariant under coordinate transformations.

However, $\sigma_d[P]$ isn't necessarily invariant; its definition changes as well. When we change variables, the probability density changes from $p(x)$ to $q(y) = \frac{p(x(y))}{|\det J|}$. The differential entropy then is $(-<\log q>)$, rather than $(-<\log p>)$ Each of the expectation values is invariant, but the change in underlying distribution from p to q in the definition of entropy is not.

$$\sigma_d[Q] =<\log(|\det J|)> - <\log p>=<\log(|\det J|)> +\sigma_d[P]$$

The two only are equal if $<\log(|\det J|)>= 0$. This obviously is the case[27] when $|\det J| = 1$. The case $|\det J| = 1$ corresponds to a volume-preserving coordinate transform. The sign of $\det J$ must be constant to preserve continuity[28]. Obviously, under a volume-preserving transformation both p and dx are invariant.

Note that σ_c is invariant under all coordinate transforms because $p(x)d(x)$ is invariant. The change in dx under the logarithm eliminates the $<\log(|\det J|)>$ term that causes problems for σ_d. However, the unit term $N \log L_0$ doesn't change, so $\tilde{\sigma_d}$ behaves like σ_d. This means that the difference between σ_c and σ_d arises from the $\lim_{m\to\infty}(\log m)$ term. This is to be expected because the rate at which $m \to \infty$ changes when we change variables. It is the salient piece of $\log dx$.

In summary, $< f >$ and σ_c are invariant under all well-behaved coordinate transforms while σ_d and $\tilde{\sigma}_d$ are invariant under volume-preserving coordinate transforms.

[27]Though $|\det J|$ is non-negative, $\log(|\det J|)$ need not be, and we conceivably could come up with special cases where $<\log(|\det J|)>= 0$ while $|\det J| \neq 1$, but we won't discuss those here.

[28]Which is part of what we meant by "well-behaved" earlier.

REFERENCES

[1] T. Cover and J. Thomas, "Elements of Information Theory," Wiley, 1991
[2] C. Tsallis, "Possible Generalization of Boltzmann-Gibbs Statistics," J. Stat. Phys. 52, 1988 p. 479-487
[3] A. Renyi, "On Measures of Information and Entropy," Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960 p. 547-561
[4] C.E. Shannon, "A Mathematical Theory of Communication," Bell Syst. Tech. J., 27 p. 379-423, p. 623-656
[5] I. Csiszar, "Axiomatic Characterizations of Information Measures," *Entropy* 2008, 10 p. 261-273
[6] D.K. Faddeev, "On the concept of entropy of a finite probability scheme," Uspehi Mat. Nauk 1956, 11 p. 227-231
[7] G. Diderrich, "The role of boundedness in characterizing Shannon entropy," Information and Control 1975, 29 p. 149-161
[8] T.W. Chaundry, J.B. McLeod, "On a functional equation," Edinburgh Mat. Notes 1960, 43 p. 7-8
[9] J. Aczel, B. Forte, C.T. Ng, "Why Shannon and Hartley entropies are "natural"," Adv. Appl. Probab. 1974, 6 p. 131-146
[10] S. Abe, "Axioms and uniqueness theorem for Tsallis entropy," Phys. Letters A 2000, 271 p. 74-79
[11] A.I. Khinchin, "Mathematical Foundations of Information Theory," Dover, 1957