A New Definition of the Entropy of General Probability Distributions Based on the Non-Standard Analysis

KAZUHIKO OZEKI

NHK Technical Research Laboratories, Tokyo, Japan

Based on the non-standard analysis by Robinson, a new definition of the entropy of general distributions is given as a straightforward extension of Shannon's discrete entropy. The newly defined entropy has natural properties such as positiveness and invariance under transformations unlike Shannon's continuous entropy. In the light of this new definition, the meaning of Shannon's continuous entropy will be more clearly seen.

1. INTRODUCTION

The entropy of a continuous distribution defined by Shannon (1948) is not a natural extension of the entropy of a discrete distribution despite their analogous forms. If one tries to define the entropy of a continuous distribution as the limit of a discrete case, a divergent term appears, which makes such a definition impossible. Therefore, the entropy of a continuous distribution has been defined by neglecting this divergent term. This rather artificial trick causes the entropy of a continuous distribution to have unnatural properties which the entropy of a discrete distribution does not have; it can be negative and is relative to the coordinate system. Another problem, which is even more serious, is that Shannon's definitions of the entropy do not lead to a natural definition of the entropy of a general mixed discrete and continuous distribution, because separate definitions are adopted for the discrete case and the continuous case. Since a probability distribution is not always purely discrete nor purely continuous, a more unified definition is desirable.

In spite of these problems, Shannon's definition of the continuous entropy is undeniably successful in practice, suggesting that its rationalization will be possible if a suitable piece of mathematics becomes available, just as a rationalization of Dirac's delta function was achieved through the notion of Schwartz' distribution (Schwartz (1951)).

It is quite natural that the entropy of a continuous distribution diverges when one tries to define it as the limit of the discrete case, because it is then
to represent the average amount of information we obtain when the value of a random variable is measurable with an infinitesimal resolution. Thus we see that the problem is not that it diverges, but that infinite numbers and infinitesimals do not have "citizenship" in the territory of the real numbers. Therefore, if we have a mathematical tool to handle infinite numbers and infinitesimals rightly, then we will be able to define the entropy of a continuous, and more general types of distributions in an intuitively natural way as a direct extension of the definition of the discrete entropy. The non-standard analysis by Robinson (1966), which is said to be a modern version of Leibnitz' infinitesimal calculus, is just the thing for this purpose.

In this paper, we shall give a new definition of the entropy of general distribution based on the non-standard analysis, and show that it has natural properties such as positiveness and invariance under transformations. The treatment is very straightforward and elementary. In the light of this new definition, the meaning of Shannon's continuous entropy will be more clearly seen.

2. Definition of the Entropy Based on the Non-Standard Analysis

We begin with the notational convention used in this paper. For the detail of non-standard analysis, the reader is referred to Keisler (1976). We denote by \( \mathbb{R} \) the field of real numbers, and by \( {}^*\mathbb{R} \) the field of hyperreal numbers. The natural extension of a real function \( f \) is denoted by \( {}^*f \). The natural extension of a set \( U \subseteq \mathbb{R} \) is denoted by \( {}^*U \). For \( \xi, \eta \in {}^*\mathbb{R} \), we write \( \xi \approx \eta \) if and only if they are infinitely close. The standard part of a finite hyperreal number \( \xi \) is denoted by \( \text{st}(\xi) \). An integral sign always means the Riemann integral.

Now let \( X \) be a real valued random variable defined on a probability space \((\Omega, \mathcal{A}, P)\), where \( \Omega \) is a set, \( \mathcal{A} \) is a \( \sigma \)-field on \( \Omega \) and \( P \) is a probability measure defined on \( \mathcal{A} \). We denote by \( F_X \) the distribution function of \( X \):

\[
F_X(x) = P\{X \leq x\}.
\]

Until Section 4, we assume for simplicity that the probability distribution is confined to a finite interval \([a_1, b_1]\): there exist \( a_1, b_1 \in \mathbb{R} \) such that \( F_X(a_1) = 0, F_X(b_1) = 1 \).

Let \( \Delta x \) be a positive real number. We divide the interval \([a_1, b_1]\) as

\[
x_i = a_1 + i \cdot \Delta x \quad (i = 0, \ldots, n(\Delta x))
\]

where \( n(\Delta x) \) is the least integer \( n \) such that

\[
a_1 + n \cdot \Delta x \geq b_1.
\]
Then the quantity
\[ H(X, \Delta x) = -\sum_{i=1}^{n(\Delta x)} u_i \log_2(u_i), \]
where
\[ u_i = F_x(x_i) - F_x(x_{i-1}) \]
can be interpreted as the entropy of \( X \) measured with the resolution \( \Delta x \). The mapping

\[ H(X, \cdot): \Delta x \mapsto H(X, \Delta x) \]
is defined on \((0, \infty)\). Therefore, its natural extension, denoted by \( *H(X, \cdot) \), is defined on \(*(0, \infty)\) (Keisler (1976, Sect. 1D, Theorem 1)).

**Definition 1.** Let \( \delta x \) be positive infinitesimal: \( \delta x > 0 \), \( \delta x \approx 0 \). The entropy of \( X \) measured with the resolution \( \delta x \) is defined by \( *H(X, \delta x) \).

**Example 1.** We shall calculate the newly defined entropy for some common cases. By the Lebesgue decomposition theorem (Loève (1963, Sect. 11)), the probability distribution function \( F_x \) can be uniquely decomposed into three components as

\[ F_x = F_x^{(ac)} + F_x^{(d)} + F_x^{(s)}, \]
where

(i) \( F_x^{(ac)} \) is absolutely continuous;

(ii) \( F_x^{(d)} \) is a jump function having the following form:

\[ F_x^{(d)}(x) = \sum_k p_k h(x - t_k), \]

the sum being finite or countable, where

\[ h(x) = 0; \quad \text{for} \; x < 0, \]
\[ = 1; \quad \text{for} \; x \geq 0, \]

and \( p_k > 0 \); the discontinuities \( t_k \) are identical with those of \( F_x \);

(iii) \( F_x^{(s)} \) is a singular function.

Some of the three components may be missing.

**Case 1.** \( F_x^{(ac)} \) and \( F_x^{(s)} \) are missing, and \( F_x^{(d)} \) has only a finite number of discontinuous points.
In this case \( F_X \) can be written as
\[
F_X(x) = \sum_{k=1}^{N} p_k h(x - t_k), \quad a_1 < t_1 < t_2 < \cdots < t_N \leq b_1.
\]

Let
\[
m = \min_{k=1, \ldots, N-1} (t_{k+1} - t_k).
\]

Then it is easy to see that for any \( \Delta x \) satisfying
\[
0 < \Delta x < m,
\]
the following equation holds:
\[
H(X, \Delta x) = - \sum_{k=1}^{N} p_k \log_2(p_k).
\]

In other words, every real solution of (1) is a real solution of (2). Then by Theorem 4 in Keisler (1976, Sect. 1C), every hyperreal solution of (1) is a hyperreal solution of (2). Thus we have for any positive infinitesimal \( \delta x \)
\[
\ast H(X, \delta x) = - \sum_{k=1}^{N} p_k \log_2(p_k).
\]

This shows that in this case the entropy defined by Definition 1 coincides with Shannon’s discrete entropy.

\textbf{Case 2.} \( F_X^{(d)} \) and \( F_X^{(s)} \) are missing, and the probability density function \( P_X \) of \( F_X^{(ac)} \) is continuous.

By the mean value theorem, there exists \( x'_{i-1} \in [x_i, x_{i-1}] \) such that
\[
F_X(x_i) - F_X(x_{i-1}) = \int_{x_{i-1}}^{x_i} P_X(x) \, dx = P_X(x'_{i-1})(x_i - x_{i-1}),
\]
so that
\[
H(X, \Delta x) = \Phi(\Delta x) - \log_2(\Delta x),
\]
where
\[
\Phi(\Delta x) = - \sum_{i=1}^{\pi(\Delta x)} (P_X(x'_{i-1}) \log_2(P_X(x'_{i-1}))) \Delta x.
\]
Hence, for any positive infinitesimal $\delta x$

$$^*H(X, \delta x) = ^*\Phi(\delta x) - ^*\log_2(\delta x). \quad (3)$$

It should be noted that for any positive infinitesimal $\delta x$

$$\text{st}(^*\Phi(\delta x)) = - \int_{a_1}^{b_1} P_X(x) \log_2 P_X(x) \, dx.$$  

This shows that the standard part of the first term of $^*H(X, \delta x)$ coincides with Shannon’s continuous entropy. The second term $-^*\log_2(\delta x)$ is a positive infinite hyperreal number, and is independent of the distribution function.

**Case 3.** $F_X^{(s)}$ is missing, and $F_X^{(d)}$ has only a finite number of discontinuous points $t_1 < t_2 < \cdots < t_N$:

$$F_X^{(d)}(x) = \sum_{i=1}^{N} p_i h(x - t_i).$$

Furthermore the density function $P_X$ of $F_X^{(ac)}$ is continuous.

A simple calculation similar to the preceding two cases leads to

$$^*H(X, \delta x) \approx - \int_{a_1}^{b_1} P_X(x) \log_2(P_X(x)) \, dx$$

$$- (F_X^{(ac)}(b_1) - F_X^{(ac)}(a_1)) \log_2(\delta x)$$

$$- \sum_{j=1}^{N} p_j \log_2(p_j).$$

Now we consider the 2-dimensional case and define the joint entropy, the conditional entropy and the average mutual information in a non-standard way.

Let $Y$ be another random variable defined on the same probability space as $X$. We denote the joint probability distribution function of $(X, Y)$ by $F_{(X,Y)}$, and the marginal distribution function of $Y$ by $F_Y$. We also assume that the distribution function of $Y$ is confined to a finite interval $[a_2, b_2]$;

$$F_Y(a_2) = 0, \quad F_Y(b_2) = 1,$$

so that

$$F_{(X,Y)}(a_1, a_2) = 0, \quad F_{(X,Y)}(b_1, b_2) = 1.$$
Let $Ay > 0$ be a real number, and divide the interval $[a_2, b_2]$ as
\[ y_j = a_2 + j \cdot Ay \quad (j = 0, \ldots, n(Ay)), \]
where $n(Ay)$ is the least integer $n$ satisfying
\[ a_2 + n \cdot Ay \geq b_2. \]

Define $r_{ij}$, $v_j$ and $s_{ij}$ by
\[ r_{ij} = F(x, y)(x_i, y_j) - F(x, y)(x_{i-1}, y_j) \]
\[ \quad - F(x, y)(x_i, y_{j-1}) + F(x, y)(x_{i-1}, y_{j-1}), \]
\[ v_j = \sum_{k=1}^{n(Ax)} r_{kj}, \]
and
\[ s_{ij} = \frac{r_{ij}}{v_j}; \quad v_j \neq 0, \]
\[ = 0 \quad v_j = 0. \]

Then the joint entropy of $X$ and $Y$, the conditional entropy of $X$ under $Y$, the average mutual information between $X$ and $Y$, all measured with the resolution $(Ax, Ay)$, can be defined by
\[ H(X, Ax; Y, Ay) = - \sum_{i=1}^{n(Ax)} \sum_{j=1}^{n(Ay)} r_{ij} \log_2(r_{ij}), \]
\[ H(X, Ax \mid Y, Ay) = - \sum_{i=1}^{n(Ax)} \sum_{j=1}^{n(Ay)} r_{ij} \log_2(s_{ij}), \]
and
\[ I(X, Ax; Y, Ay) = H(X, Ax) - H(X, Ax \mid Y, Ay), \]
respectively. These are functions of $(Ax, Ay)$ defined on $(0, \infty) \times (0, \infty)$. Therefore, their natural extensions $*H(X, \cdot ; Y, \cdot)$, $*H(X, \cdot \mid Y, \cdot)$, $*I(X, \cdot ; Y, \cdot)$ are defined on $*(0, \infty) \times *(0, \infty)$.

**Definition 2.** Let $\delta x$ and $\delta y$ be positive infinitesimal. The joint entropy of $X$ and $Y$, the conditional entropy of $X$ under $Y$, the average mutual information between $X$ and $Y$, all measured with the resolution $(\delta x, \delta y)$, are defined by $*H(X, \delta x; Y, \delta y)$, $*H(X, \delta x \mid Y, \delta y)$ and $*I(X, \delta x; Y, \delta y)$, respectively.
EXAMPLE 2. Assume that the probability distribution function $F_{(x,y)}$ is absolutely continuous, and let its density function $P_{(x,y)}$ be continuous. Then a simple computation yields

$$\begin{align*}
*H(X, \delta x; Y, \delta y) &
\approx - \int_{a_1}^{b_1} \int_{a_2}^{b_2} P_{(x,y)}(x,y) \log_2(P_{(x,y)}(x,y)) \, dx \, dy \\
&\quad - \log_2(\delta x) - \log_2(\delta y), \\
*H(X, \delta x \mid Y, \delta y) &
\approx - \int_{a_1}^{b_1} \int_{a_2}^{b_2} P_{(x,y)}(x,y) \log_2(P(x,y)) \, dx \, dy - \log_2(\delta x), \\
*I(X, \delta x; Y, \delta y) &
\approx \int_{a_1}^{b_1} \int_{a_2}^{b_2} P_{(x,y)}(x,y) \log_2 \left( \frac{P_{(x,y)}(x,y)}{P_x(x)P_y(y)} \right) \, dx \, dy.
\end{align*}$$

From this example we see that Shannon's joint entropy of $X$ and $Y$, and the conditional entropy of $X$ under $Y$ are equal to the standard parts of

$$*H(X, \delta x; Y, \delta y) + \log_2(\delta x) + \log_2(\delta y),$$

and

$$*H(X, \delta x \mid Y, \delta y) + \log_2(\delta x),$$

respectively. Shannon's mutual information between $X$ and $Y$ is the standard part of $*I(X, \delta x; Y, \delta y)$ itself. The average mutual information $*I(X, \delta x; Y, \delta y)$ differs from $*H(X, \delta x; Y, \delta y)$ and $*H(X, \delta x \mid Y, \delta y)$ in that it is "almost" independent of the infinitesimal resolution $(\delta x, \delta y)$: if $\delta x, \delta y, \delta x'$ and $\delta y'$ are positive infinitesimals, then

$$*I(X, \delta x; Y, \delta y) \approx *I(X, \delta x'; Y, \delta y').$$

This fact is closely related to the fact that Shannon's mutual information is independent of the coordinate system.
3. Properties of the Entropy and Its Relatives

In this section we shall show that the newly defined entropy and its relatives have natural properties.

**Proposition 1.** For every positive infinitesimal $\delta x$,

$$\star H(X, \delta x) \geq 0.$$  

**Proof.** Every real solution $\Delta x$ of

$$\Delta x > 0$$  

is a real solution of

$$H(X, \Delta x) \geq 0,$$  

because $H(X, \Delta x)$ is Shannon’s discrete entropy. Therefore, by Theorem 4 in Keisler (1976, Sect. 1C), every hyperreal solution of (4) is a hyperreal solution of (5). Q.E.D.

**Proposition 2.** If the distribution function of $X$ is of the type in Case 2 in Section 2, then $\star H(X, \delta x)$ is monotone in the following sense: if $\delta x$ and $\delta x'$ are positive infinitesimal and if there exists $c \in R$, $c > 1$ such that $\delta x' \geq c \cdot \delta x$, then

$$\star H(X, \delta x) > \star H(X, \delta x').$$

**Proof.** By (3),

$$\star H(X, \delta x) - \star H(X, \delta x') = (\Phi(\delta x) - \Phi(\delta x')) + \star \log_2(\delta x'/\delta x)$$

$$\geq (\Phi(\delta x) - \Phi(\delta x')) + \log_2(c).$$

The first term on the right-hand side is infinitely close to $0$ and the second term is a positive real number. Hence the right-hand side is positive. Q.E.D.

The next proposition asserts that the familiar formulas hold for the newly defined entropy and its relatives.

**Proposition 3.** (i) $\star H(X, \delta x; Y, \delta y) = \star H(Y, \delta y) + \star H(X, \delta x | Y, \delta y)$,

(ii) $\star H(X, \delta x) + \star H(Y, \delta y) \geq \star H(X, \delta x; Y, \delta y)$,

(iii) $\star H(X, \delta x) \geq \star H(X, \delta x | Y, \delta y)$.

**Proof.** We shall prove (i) only, because the other formulas can be proved similarly.
Every real solution \((\Delta x, \Delta y)\) of
\[
\Delta x > 0, \quad \Delta y > 0,
\]
is a real solution of
\[
H(X, \Delta x; Y, \Delta y) = H(Y, \Delta y) + H(X, \Delta x \mid Y, \Delta y).
\]
Therefore, by Theorem 4 in Keisler (1976, Sect. 1C), every hyperreal solution of (6) is a hyperreal solution of (7). Q.E.D.

Before discussing the invariance of \(*H(X, \delta x)\) under transformations, we introduce the notion of the entropy measured with a distributed resolution.

Let \(\varphi: R \to R\) be a continuous and strictly increasing function, and let
\[
a_1 = \varphi^{-1}(a_1), \quad b_1 = \varphi^{-1}(b_1),
\]
where \(a_1\) and \(b_1\) are the values from Section 2. We divide \([a_1', b_1']\) as
\[
x_i = a_1' + i \cdot \Delta x \quad (i = 0, 1, \ldots, n(\Delta x)),
\]
where \(n(\Delta x)\) is the least integer \(n\) satisfying
\[
a_1' + n \cdot \Delta x \geq b_1'.
\]

Now consider
\[
- \sum_{i=1}^{n(\Delta x)} (F_x(y_i) - F_x(y_{i-1})) \log_2(F_x(y_i) - F_x(y_{i-1})),
\]
where
\[
y_i = \varphi(x_i).
\]
This is a function of \(\Delta x\), which we denote by
\[
H(X, \varphi' \cdot): \Delta x \mapsto H(X, \varphi' \Delta x).
\]
Based on the natural extension \(*H(X, \varphi' \cdot)\) of this function, we give the following definition.

**Definition 3.** Let \(\varphi: R \to R\) be a continuous and strictly increasing function, and let \(\delta x\) be a positive infinitesimal. We define the entropy of \(X\) measured with the distributed resolution \(\varphi' \delta x\) by \(*H(X, \varphi' \delta x)\).

The entropy \(*H(X, \delta x)\) is invariant under transformations in the following sense.
PROPOSITION 4. Let \( \phi: \mathbb{R} \to \mathbb{R} \) be a continuous and strictly increasing function. Then for any positive infinitesimal \( \delta x \),

\[
*H(X, \delta x) = *H(\phi(X), \phi' \delta x).
\]

Proof. Noting that

\[
F_{\phi(x)}(x) = F_x(\phi^{-1}(x)),
\]

where \( F_{\phi(x)} \) is the distribution function of \( \phi(X) \), we have

\[
H(\phi(X), \phi' \Delta x)
\]

\[
= - \sum_{i=1}^{n(\Delta x)} (F_{\phi(x)}(y_i) - F_{\phi(x)}(y_{i-1})) 
\times \log_2(F_{\phi(x)}(y_i) - F_{\phi(x)}(y_{i-1}))
\]

\[
= - \sum_{i=1}^{n(\Delta x)} (F_{x}(x_i) - F_{x}(x_{i-1})) \log_2(F_{x}(x_i) - F_{x}(x_{i-1}))
\]

\[
= H(X, \Delta x).
\]

Since this holds for any \( \Delta x > 0 \),

\[
*H(\phi(X), \phi' \delta x) = *H(X, \delta x)
\]

holds for any positive hyperreal \( \delta x \).

Q.E.D.

If \( \phi(x) = k \cdot x (k \in \mathbb{R}, k > 0) \), then it is easy to show that

\[
*H(X, \phi' \delta x) = *H(X, k \cdot \delta x).
\]

A simple example of a transformation is a change of the unit of measurement. Suppose that \( X \) is a measurement of length and is measured in meter. If the unit of measurement is changed to centimeter, \( X \) is transformed to \( 100X \). From Proposition 4 and (8), we see that

\[
*H(100X, 100\delta x) = *H(X, \delta x),
\]

that is, the entropy of \( X \) meter measured with the resolution of \( \delta x \) meter is equal to the entropy of \( 100X \) centimeter measured with the resolution of \( 100\delta x \) centimeter. This might seem trivial. Shannon's continuous entropy, however, is unable to express this natural situation.
4. Case of Unbounded Random Variable

In this section, we shall extend the definition of entropy to the case in which the random variable is not bounded.

Let $a, b \in R, a < b$, and let $\Delta x > 0, \Delta x \in R$. We divide the interval $[a, b]$ just as in Section 2:

$$x_i = a + i \cdot \Delta x \quad (i = 0, ..., n(\Delta x)).$$

Define

$$H(X, \Delta x, a, b) = - \sum_{i=1}^{n(\Delta x)} (F_X(x_i) - F_X(x_{i-1})) \log_2(F_X(x_i) - F_X(x_{i-1})).$$

The real function $H(X, \cdot, \cdot, \cdot)$ is defined on the domain

$$D = \{(\Delta x, a, b) | \Delta x, a, b \in R, \Delta x > 0, a < b\},$$

and its natural extension $^*H(X, \cdot, \cdot, \cdot)$ is defined on

$$^*D = \{(\varepsilon, \xi, \eta) | \varepsilon, \xi, \eta \in ^*R, \varepsilon > 0, \xi < \eta\}.$$

Now let $\delta x, A, B$ be positive infinitesimal, negative infinite, positive infinite, respectively. We define the entropy of $X$ measured with the resolution $\delta x$ within the range $[A, B]$ by $^*H(X, \delta x, A, B)$.

In most of the interesting cases, which negative infinite $A$ and positive infinite $B$ are chosen is of little importance. For example, suppose that the density function $P_X$ of the absolutely continuous component of $F_X$ is continuous and Riemann integrable in the wide sense and that $F_X^{(d)}$ has only a finite number of discontinuous points $t_k, k = 1, ..., N$:

$$F_X^{(d)} = \sum_{k=1}^{N} p_k h(x - t_k).$$

Furthermore, suppose that the singular component $F_X^{(s)}$ is missing. Then it is easy to show that

$$^*H(X, \delta x, A, B) \approx - \int_{-\infty}^{\infty} P_X(x) \log_2 P_X(x) \, dx - \sum_{k=1}^{N} p_k \log_2 p_k - (F_X^{(ac)}(+\infty) - F_X^{(ac)}(-\infty)) \log_2(\delta x),$$

where

$$F_X^{(ac)}(+\infty) = \lim_{x \to +\infty} F_X^{(ac)}(x),$$
and
\[ F_X^{(ac)}(-\infty) = \lim_{x \to -\infty} F_X^{(ac)}(x). \]

Other quantities such as the joint entropy, the conditional entropy and the average mutual information can be similarly extended to the case of unbounded random variables.

5. Conclusion

Based on the non-standard analysis, the entropy and its relatives of general probability distributions were defined as a direct extension of the definition of Shannon's discrete entropy. The newly defined entropy is relative to the resolution, and it is the author's opinion that it should be so. All the natural properties of the discrete entropy and its relatives are inherited by the newly defined ones.

Interpreted in the new point of view set out here, Shannon's information theory in the continuous case can be regarded as one in which a certain infinitesimal resolution is chosen and fixed, and infinite terms such as \(-\delta \log_2(\delta x)\), which is constant as long as the resolution \(\delta x\) is kept fixed, are neglected.

There seems to be two possible attitudes in defining the entropy of general distributions rigorously; one allows only a real, non-infinitesimal resolution or one allows an infinitesimal resolution. The former may be physically more reasonable, because we know that no measurement can be without limit in precision. However, the problem is, as Amari (1970, p. 150) pointed out, what the value of the resolution should be in constructing a general theory of entropy. This problem motivates the latter attitude: allowance of an infinitesimal resolution as an idealization. Shannon's definition of continuous entropy can be regarded as an attempt along this line. This paper has shown that the non-standard analysis is very useful to refine Shannon's idea in a rigorous and natural manner.

Received: April 16, 1980; revised: July 23, 1980

References

