UDC 519.2
DOI: https://doi.org/10.17721/1812-5409.2025/1.7

Hening HUANG, PhD, Researcher
ORCID ID: 0000-0003-1882-0042
e-mail: heninghuang1@gmail.com
Teledyne RD Instruments, San Diego, USA

THE THEORY OF INFORMITY: A NOVEL PROBABILITY FRAMEWORK

This paper proposes a novel probability framework, called the theory of informity. We define a mathematical quantity called "informity" to quantitatively measure the degree of informativeness of a probability distribution (or a probability system). We also define two other quantities: cross-informity and joint informity. We propose an informity metric that can be used as an alternative to entropy metric. The informities for twelve continuous distributions are given. Three examples are presented to demonstrate the practicability of the proposed informity metric.

Keywords: ***informity, information content, probability, probability distribution.***

AMS 2020 classification: 60E05, 94A17.

Introduction

Shannon (1948) introduced the concepts of "information content" and "information entropy". The Shannon information content (also called self-information, surprisal, or Shannon information) of an outcome x_i with probability $P(x_i)$ is defined as

$$I(x_i) := -\log_2 P(x_i). \tag{1}$$

For a discrete random variable *X* with the probability mass function (PMF) $P(x)$, the entropy (also called Shannon entropy or information entropy) of *X* is defined as

$$\mathrm{H}(X) := -\sum_{i=1}^{N} P(x_i)\log_2 P(x_i) = \mathrm{E}[-\log_2 P(X)], \tag{2}$$

where *N* is the number of outcomes of the probability system $\{X; P(x)\}$, and $\mathrm{E}[\cdot]$ means taking the "mathematical expectation".

For a continuous random variable *Y* with the probability density function (PDF) $p(y)$, the entropy of *Y* is defined as

$$\mathrm{H}(Y) := -\int p(y)\log_2 p(y)\,dy = \mathrm{E}[-\log_2 p(Y)]. \tag{3}$$

When discussing Shannon information theory, "information" is actually shorthand for "information content," which refers to the amount of information conveyed by the outcome x_i based on its probability $P(x_i)$. Since most people may not explicitly think of "information" as shorthand for "information content," we use "information content" to avoid confusion. Furthermore, the notion of "information content" does not involve the meaning of information.

Needless to say, Shannon information content and entropy are key concepts in Shannon information theory, which provides a framework for transmission, storage, compression, and processing of information in communications engineering. However, the use of Shannon information content and entropy in fields other than communications engineering may not be consistent with our common sense about information or informativeness. For example, entropy metric is used in machine learning for splitting decision tree nodes. However, the so-called "information gain" is actually the "Shannon information content loss" corresponding to the reduction of entropy. Petty (2018) discussed some of the shortcomings of Shannon entropy as a measure of information content in indirect measurements of continuous variables. They highlighted a number of significant limitations of Shannon entropy as a basis for quantifying the information content of remote sensing observations. Based on a historical review of the problems with conceptualization of information content, Schroeder (2004) demonstrated the need for an alternative to entropy for measuring information content more generally. He stated, "…there is nothing wrong with entropy as long as we analyze communication systems and are interested in the 'engineering problem'… The question is whether we can use entropy equally successfully for measuring information [content] in different contexts. Here I believe the answer is in the negative."

Shannon information theory links "Shannon information content" with "probability" as shown in Eq. (1). However, "probability" is a (normalized) *relative* quantity ranging from 0 to 1 and has no unit, while "Shannon information content" is an *absolute* quantity ranging from 0 to infinity and has units such as "bits". By definition, a relative quantity is expressed as a ratio or percentage compared to (or normalized to) a reference value, whereas an absolute quantity is a numerical value that does not rely on comparison with another value. The Appendix explains how the *relative* quantity "probability" is converted into the *absolute* quantity "Shannon information content". Since Shannon information content and entropy are absolute quantities, they can be encoded, transmitted, compressed, and stored in communications engineering. However, in other fields or in daily life, people are more familiar with and use the *relative* quantity "probability". For example, when talking about the possibility of rain tomorrow, people would say "the probability of rain tomorrow is 80 %", rather than "the chance of rain tomorrow is 0.32 bits".

The aim of this study is to develop a novel probability framework, called the theory of informity, that is more applicable to fields other than communications engineering. Unlike Shannon information theory, which deals with *absolute* information content and uses log-transformation of probabilities, the proposed informity theory treats probability as *relative* information content and does not use log-transformation. The greater the probability of an event, the greater the relative information content of the event (i. e. the event is more informative). Conversely, the smaller the probability of an event, the smaller the relative information content of the event (i. e. the event is less informative). We believe this view of probability is intuitive and consistent with our common sense or perceptual understanding about information or informativeness. The relative information content has the opposite meaning from Shannon information content, which is the absolute information content explained in the Appendix.

In the following sections, Section 2 defines a mathematical quantity called "informity" to quantitively measure the degree of informativeness of a probability distribution (or a probability system). Section 3 defines cross-informity and joint informity. Section 4 presents informity metric. Section 5 provides discussion. Section 6 gives three application examples. Section 7 provides conclusions.

1. The definition of informity

We consider the word "informity" to be a technical term meaning "the degree of informativeness". This is similar to the technical term "uncertainty" in measurement science, which means "degree of uncertainess". The mathematical definition of informity is different for discrete random variables and continuous random variables.

1.1. Discrete random variables

Definition 1. Let $\beta(X)$ denote the informity of a discrete random variable *X* with the PMF $P(x)$. $\beta(X)$ is given by the mathematical expectation of $P(X)$

$$\beta(X) := \mathrm{E}[P(X)] = \sum_{i=1}^{N}[P(x_i)]^2. \tag{4}$$

In the context of relative information content, $\beta(X)$ is the average amount of relative information contained in the probability system $\{X; P(x)\}$. Similar to how $\mathrm{H}(X)$ is called discrete entropy, $\beta(X)$ is called discrete informity. $\beta(X)$ quantitively measures the degree of informativeness (or certainty) of the probability system $\{X; P(x)\}$. It has the opposite meaning from the discrete entropy $\mathrm{H}(X)$. $\beta(X)$ ranges between 0 and 1 and is dimensionless.

Consider tossing a coin, which is a two-state probability system: $\{X; P(x)\} = \{\text{head}, \text{tail}; P(\text{head}), P(\text{tail})\}$, where $P(\text{tail}) = 1 - P(\text{head})$. The informity $\beta(X)$ of this system can be calculated as

$$\beta(X) = [P(\text{head})]^2 + [P(\text{tail})]^2. \tag{5}$$

Figure 1 shows the informity $\beta(X)$ as a function of $P(\text{head})$, compared with the corresponding entropy (in bits): $\mathrm{H}(X) = -P(\text{head})\log_2 P(\text{head}) - P(\text{tail})\log_2 P(\text{tail})$.

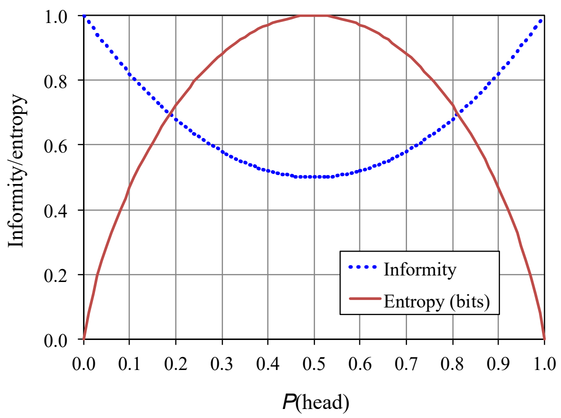

Fig. 1. Informity of the coin-tossing system as a function of the bias represented by the probability of heads, compared with the corresponding entropy

As can be seen in Fig. 1, when the coin is fair, i. e. $P(\text{head}) = P(\text{tail}) = 0.5$, the informity $\beta(X)$ is the lowest and the entropy $\mathrm{H}(X)$ is the largest, indicating that the system is least informative or has the greatest uncertainty. When a coin is highly biased towards the tail or head, the informity increases and the entropy decreases. In the most extreme case of $P(\text{tail}) = 1$ or $P(\text{head}) = 1$, the informity is the highest: $\beta(X) = 1$ and the entropy is the lowest: $\mathrm{H}(X) = 0$, indicating that the system is highest informative or has the greatest certainty.

1.2. Continuous random variables

Definition 2. Let $\beta(Y)$ denote the informity of a continuous random variable *Y* with the PDF $p(y)$. $\beta(Y)$ is given by the mathematical expectation of $p(Y)$

$$\beta(Y) := \mathrm{E}[p(Y)] = \int [p(y)]^2\, dy. \tag{6}$$

Similar to how $\mathrm{H}(Y)$ is called continuous entropy, $\beta(Y)$ is called continuous informity. $\beta(Y)$ quantitively measures the degree of informativeness (or certainty) of the probability system $\{Y; P(y)\}$. It has the opposite meaning from the continuous entropy $\mathrm{H}(Y)$. Table 1 shows the continuous informities of twelve probability distributions.

The transition from discrete to continuous informity can be justified as follows. The continuous distribution is approximated by discrete quantization (divided into intervals of size Δy) to form a discrete probability system: $\{Y_\Delta; p(y)\Delta y\}$. The discrete informity of Y_Δ is given by,

$$\beta(Y_\Delta) = \sum_{i=1}^{N}[p(y_i)\Delta y]^2 = \Delta y \sum_{i=1}^{N}[p(y)]^2 \Delta y. \tag{7}$$

Normalizing $\beta(Y_\Delta)$ by Δy and then taking the limit of the normalized $\beta(Y_\Delta)$ yields

$$\lim_{\Delta y \to 0} \frac{\beta(Y_\Delta)}{\Delta y} = \lim_{\Delta y \to 0} \sum_{i=1}^{N}[p(y)]^2 \Delta y = \int [p(y)]^2\, dy = \beta(Y)\,. \tag{8}$$

This procedure shows that while continuous informity is analogous to discrete informity, it is not exactly a limit in the strict sense, but rather a renormalized version that accounts for the continuous nature of the underlying probability space. Note that this process is essentially similar to the transition from discrete to continuous (differential) entropy.

Table 1

Continuous informities of twelve probability distributions

Distribution	Probability density function $p(y)$	Informity $\beta(Y)$	Support
Uniform	$\frac{1}{2a}$	$\frac{1}{2a}$	$[-a,a]$
Triangular	$\begin{cases}\frac{(y+a)}{a^2}, & -a\le y\le 0\\ \frac{(a-y)}{a^2}, & 0\le y\le a\end{cases}$	$\frac{2}{3a}$	$[-a,a]$
Quadratic	$\frac{3}{4a}\left[1-\left(\frac{1}{a}y\right)^2\right]$	$\frac{3}{5a}$	$[-a,a]$
Raised cosine	$\frac{1}{2a}\left[1+\cos\left(\frac{\pi}{a}y\right)\right]$	$\frac{3}{4a}$	$[-a,a]$
Half-cosine	$\frac{\pi}{4a}\cos\left(\frac{\pi}{2a}y\right)$	$\frac{\pi^2}{16a}$	$[-a,a]$
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$	$\frac{1}{2\sigma\sqrt{\pi}}$	$(-\infty,\infty)$
Exponential	$\lambda\exp(-\lambda y)$	$\frac{\lambda}{2}$	$[0,\infty)$
Cauchy	$\frac{1}{\pi\gamma[1+\frac{1}{\gamma^2}(y-y_0)^2]}$	$\frac{1}{2\pi\gamma}$	$(-\infty,\infty)$
Rayleigh	$\frac{y}{\sigma^2}\exp\left(-\frac{y^2}{2\sigma^2}\right)$	$\frac{\sqrt{\pi}}{4\sigma}$	$[0,\infty)$
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}\exp(-\beta y)$	$\frac{\beta}{[\Gamma(\alpha)]^2}2^{-(2\alpha-1)}\Gamma(2\alpha-1)$	$[0,\infty)$
Laplace	$\frac{1}{2b}\exp\left(-\frac{\lvert y-\mu\rvert}{b}\right)$	$\frac{1}{4b}$	$(-\infty,\infty)$
Weibull	$\frac{k}{\lambda^k}y^{k-1}\exp(-\frac{y^k}{\lambda^k})$	$\frac{k}{\lambda}2^{-\frac{2k-1}{k}}\Gamma\left(\frac{2k-1}{k}\right)$	$[0,\infty)$

Just like the continuous entropy $\mathrm{H}(Y)$ does not have all the properties of the discrete entropy $\mathrm{H}(X)$, the continuous informity $\beta(Y)$ does not have all the properties of the discrete informity $\beta(X)$. Unlike $\beta(X)$, which is dimensionless, $\beta(Y)$ is not dimensionless. $\beta(Y)$ has dimensions of $[y]^{-1}$. If the quantity Y has physical units of length, the inverse of $\beta(Y)$ also has units of length. Sussmann (1997) introduced a quantity called "uncertainty length" in quantum mechanics.

Definition 3. The uncertainty length is given by the reciprocal of the continuous informity $\beta(Y)$

$$\delta(Y) := \frac{1}{\beta(Y)} = \frac{1}{\int[p(y)]^2dy}. \quad (9)$$

For the normal distribution $N(\mu,\sigma^2)$, $\delta(Y) = 2\sigma\sqrt{\pi} = 3.545\sigma$. Therefore, the uncertainty length $\delta(Y)$ is proportional to the standard deviation σ.

2. The definitions of cross-informity and joint informity

Definition 4. The cross-informity of two discrete random variables X_1 and X_2 with PMFs $P_1(x)$ and $P_2(x)$ is given by

$$\beta(X_1\cap X_2) := \sum_{i=1}^{N} P_1(x_i)P_2(x_i) = \mathrm{E}_{P_1}[P_2(x)] = \mathrm{E}_{P_2}[P_1(x)], \quad (10)$$

where $\mathrm{E}_{P_1}[P_2(x)]$ is the expected value of $P_2(x)$ with respect to $P_1(x)$, and $\mathrm{E}_{P_2}[P_1(x)]$ is the expected value of $P_1(x)$ with respect to $P_2(x)$. Since $\beta(X_1\cap X_2) = \beta(X_2\cap X_1)$, the cross-informity is symmetric.

Definition 5. The cross-informity of two continuous random variables Y_1 and Y_2 with PDFs $p_1(y)$ and $p_2(y)$ is given by

$$\beta(Y_1\cap Y_2) := \int p_1(y)p_2(y)dy = \mathrm{E}_{p_1}[p_2(y)] = \mathrm{E}_{p_2}[p_1(y)], \quad (11)$$

where $\mathrm{E}_{p_1}[p_2(y)]$ is the expected value of $p_2(y)$ with respect to $p_1(y)$, and $\mathrm{E}_{p_2}[p_1(y)]$ is the expected value of $p_1(y)$ with respect to $p_2(y)$. The symmetry also holds, i.e. $\beta(Y_1\cap Y_2) = \beta(Y_2\cap Y_1)$.

Definition 6. Like $\beta(Y)$, $\beta(Y_1\cap Y_2)$ has dimensions of $[y]^{-1}$. Thus, similar to the definition of uncertainty length, the discrepancy length is given by the reciprocal of the cross-informity $\beta(Y_1\cap Y_2)$

$$\delta(Y_1\cap Y_2) := \frac{1}{\beta(Y_1\cap Y_2)} = \frac{1}{\int p_1(y)p_2(y)dy}. \quad (12)$$

If the distributions of Y_1 and Y_2 are identical, the discrepancy length $\delta(Y_1\cap Y_2)$ reduces to the uncertainty length $\delta(Y)$.

Definition 7. The joint informity of two discrete random variables X_1 and X_2 is given by

$$\beta(X_1,X_2) := \sum_{i=1}^{N}\sum_{j=1}^{N}\left[P(x_{1,i},x_{2,j})\right]^2, \quad (13)$$

where $P(x_{1,i},x_{2,j})$ is the joint PMF of X_1 and X_2. If X_1 and X_2 are independent,

$$\beta(X_1,X_2) = \sum_{i=1}^{N}[P_1(x_i)]^2\sum_{j=1}^{N}\left[P_2(x_j)\right]^2 = \beta(X_1)\beta(X_2). \quad (14)$$

Definition 8. The joint informity of two continuous random variables Y_1 and Y_2 is given by

$$\beta(Y_1, Y_2) := \iint [p(y_1, y_2)]^2 \, dy_1 dy_2, \tag{15}$$

where $p(y_1, y_2)$ is the joint PDF of Y_1 and Y_2. If Y_1 and Y_2 are independent,

$$\beta(Y_1, Y_2) = \iint [p_1(y_1)]^2 [p_2(y_2)]^2 \, dy_1 dy_2 = \beta(Y_1)\beta(Y_2). \tag{16}$$

The joint informity $\beta(Y_1, Y_2)$ has dimensions of $[y]^{-2}$. As a two-dimensional analogy to the one-dimensional "uncertainty length", Sussmann (1997) introduced a quantity called "uncertainty area" for the Wigner quasi-distribution in quantum mechanics.

The justification for the transition from discrete to continuous cross-informity or joint informity is similar to the procedure of the transition from discrete to continuous informity described in Section 2, so it is not repeated here.

Definition 9. The uncertainty area is given by the reciprocal of the joint informity $\beta(Y_1, Y_2)$

$$\delta(Y_1, Y_2) := \frac{1}{\beta(Y_1, Y_2)} = \frac{1}{\iint [p(y_1, y_2)]^2 dy_1 dy_2}. \tag{17}$$

If Y_1 and Y_2 are independent, the uncertainty area is the product of the two uncertainty lengths

$$\delta(Y_1, Y_2) = \frac{1}{\beta(Y_1)\beta(Y_2)} = \delta(Y_1)\delta(Y_2). \tag{18}$$

3. Informity metric

Informity can be used as a metric to compare statistical distributions (models) or splits (in machine learning). Informity metric provides an alternative to entropy metric or information criteria based on maximum likelihood, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC).

Proposition 1. *Suppose that among a set of candidate distributions, we want to select one that best fits the given data or information. Let $p_k(y)$ denote the PDF of the k-th candidate distribution and $\beta_k(Y)$ denote the corresponding informity. Then, the best distribution is the one with the maximum informity. That is,*

$$best\ distribution = \arg\max_k\ [\beta_k(Y)], \qquad k = 1, 2, 3 \ldots M, \tag{19}$$

where M is the number of candidate distributions.

We designate Proposition 1 as "the maximum informity criterion". The maximum informity criterion can be seen as the counterpart of the minimum entropy criterion proposed by Huang (2024), and in fact it can be derived from the minimum entropy criterion because informity is inversely proportional to entropy. In principle, the informity and entropy metrics are compliable. However, the informity metric does not necessarily give exactly the same results as the entropy metric. This is because the continuous informity is computed in the original probability space, while the continuous entropy is computed in the log-transformed probability space.

Definition 10. Suppose a set is split into several disjoint subsets. The sum of the elements in the subsets is equal to the elements in the original set. The informity gain $\Delta\beta$ is the difference in informity before and after the split. That is,

$$\Delta\beta := \beta_{\text{after}}(X) - \beta_{\text{before}}(X), \tag{20}$$

where $\beta_{\text{before}}(X)$ is the informity before the split and $\beta_{\text{after}}(X)$ is the informity after the split. $\beta_{\text{after}}(X)$ is calculated as the weighted average of the informities of the subsets.

Proposition 2. *When training decision trees in machine learning, consider M candidate splits for a mixed class (the original set). Let $\Delta\beta_k$ denote the informity gain of the k-th candidate split. The best split among the M candidate splits is the one with the maximum informity gain. That is,*

$$best\ split = \arg\max_k\ [\Delta\beta_k], \qquad k = 1, 2, 3 \ldots M. \tag{21}$$

We designate Proposition 2 as "the maximum informity gain criterion". The maximum informity gain criterion can be seen as the counterpart of the "information gain" criterion based on the maximum entropy reduction (e.g. Zhou, 2019a).

4. Discussion

The continuous informity has the opposite meaning from the continuous entropy. They provide insights into a probability system from different perspectives.

Figure 2 shows the continuous informity and continuous entropy of the normal distribution as a function of the standard deviation σ (scale).

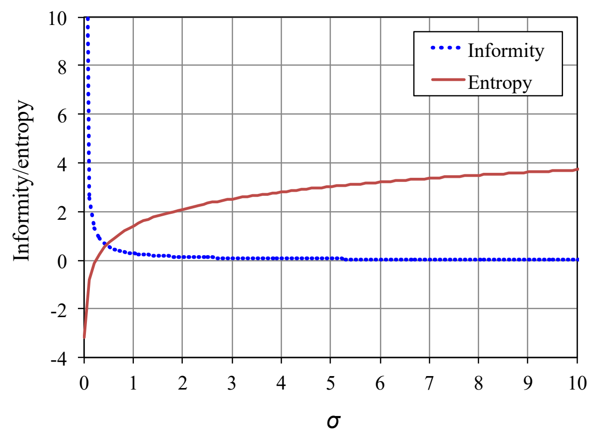

Fig. 2. Continuous informity and continuous entropy of normal distribution as a function of the standard deviation σ (scale)

It can be seen from Fig. 2 that, as the scale parameter *σ* increases, the continuous informity decreases and the continuous entropy increases. When the continuous informity is very large, the continuous entropy becomes negative.

Remark 1. When the population parameters (e.g. location and scale parameters) are known, the true value of the continuous informity can be obtained. When the population parameters are unknown and estimated from a sample, the continuous informity can be calculated based on the estimated parameter values and becomes a statistic. Therefore, using the proposed informity theory in practice may involve point estimates of population parameters.

Consider a sample (*n* observations) drawn from the normal distribution of *Y* with mean μ and standard deviation *σ*. When μ and *σ* are known, the true value of the informity $\beta(Y)$ is $1/(2\sigma\sqrt{\pi})$. When μ and *σ* are unknown, the sample mean $\bar{y}$ is an unbiased estimate of μ and $s/c_{4,n}$ is an unbiased estimate of *σ*, where *s* is the sample standard deviation and $c_{4,n}$ is the bias-correction factor; $c_{4,n}=\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$; Γ(.) stands for Gamma function (Wadsworth, 1989). Accordingly, the estimated informity is

$$\hat{\beta}(Y)=\frac{c_{4,n}}{2s\sqrt{\pi}}. \quad (22)$$

Remark 2. In this paper, the sum of the squares of PMF $\sum_{i=1}^{N}P(x_i)^2$ is defined as the discrete informity and the integral of the square of PDF $\int[p(y)]^2\,dy$ is defined as the continuous informity. The author found through online search that these two quantities or their related quantities have different names and uses in some literature. The sum of the squares of PMF is called "repeat rate" or the Simpson index, which is a measure of concentration with a long history (Rousseau, 2018). The quantity (1- repeat rate) is called Gini impurity, which measures the disorder or impurity of a dataset. It is also known as logical entropy, which is interpreted as "the probability that a distinction or dit (elements in different blocks) is obtained in two independent draws from the underlying set" (Ellerman, 2013; 2022). On the other hand, Lad et al. (2015) defined the *negative half* of the integral of the square of PDF as differential *extropy*, i. e. $J(Y)=-\frac{1}{2}\int[p(y)]^2\,dy$. However, this definition of differential *extropy* is inconsistent with their definition of discrete *extropy*: $J(X)=-\sum_{i=1}^{N}[1-P(x_i)]\log_2[1-P(x_i)]$. Note that if $P(x_i)$ is replaced by the density value *p*(*y*), the differential *extropy* would be undefined when $p(y)>1$. Lad et al. (2015) acknowledged this problem and used an approximation: $J(X)\approx 1-\frac{1}{2}\sum_{i=1}^{N}P(x_i)^2$. However, we argue that it is questionable to deduce $J(Y)=-\frac{1}{2}\int[p(y)]^2\,dy$ from the approximation $J(X)\approx 1-\frac{1}{2}\sum_{i=1}^{N}P(x_i)^2$. Moreover, puzzlingly, $J(Y)$ is always negative. This casts doubt about the true meaning of $J(Y)$.

Although these two quantities are not new, their specific perspective as the average amount of information contained in a probability system is novel and unique. Therefore, the idea of informity is new and the proposed informity theory is a novel probability framework.

5. Application examples

5.1. Selecting distribution of calibration error

Consider the calibration error of a measuring instrument. The calibration error is confined within the interval $[-a,+a]$, but has a central tendency. In measurement uncertainty analysis, we want to select the best distribution among a set of candidate distributions for the calibration error. Huang (2024) considered five candidate distributions: uniform, triangular, quadratic, raised cosine, and half-cosine. He determined that the raised cosine distribution is the best distribution according to the minimum entropy criterion.

We evaluate these five candidate distributions using the informity metric. Similar to defining the "unit entropy" by setting *a* = 1 in Huang (2024), we define "unity informity" by setting $a=1$ in the informity expressions of these five distributions shown in Tab. 1. The results for unit informity are shown in Tab. 2, together with the unit entropy.

Table 2

Unit informities and unit entropies of five candidate distributions

Distribution	Unit informity	Unit entropy (nats)
Uniform	0.5	0.693
Triangular	0.667	0.5
Quadratic	0.6	0.567
Raised cosine	**0.75**	**0.386**
Half-cosine	0.617	0.548

As can be seen from Tab. 2, the raised cosine distribution has the highest unit informity and the lowest unit entropy among the five candidate distributions. Therefore, it is the best distribution according to the maximum informity criterion and the minimum entropy criterion. This result is consistent with Castrup's (2004) argument based on physical reality considerations.

Note from Tab. 2 that, as expected, the uniform distribution has the lowest unit informity and the highest unit entropy. It is the least informative distribution among the five candidate distributions from both the informity and entropy perspectives.

5.2. Evaluating four distributions given a sample drawn from a Weibull distribution

Xie (2011) evaluated four candidate distributions: exponential, Weibull, Gamma, and normal to fit a sample (*n*=200) randomly drawn from a Weibull distribution with shape 1.8 and scale 1, using Akaike information criterion (AIC) and Pearson's Q-statistic. He estimated the parameters using the routine maximum likelihood estimation procedure, which gives: exponential distribution: rate = 1.167, Weibull distribution: shape = 1.662 and scale = 0.9588, Gamma distribution: shape = 2.255 and scale = 0.3081, and normal distribution: mean = 0.8570 and standard deviation = 0.5277. Huang (2024) considered this example using the entropy metric. We consider this example using the informity metric. Table 3 shows the results obtained using the four different metrics.

As can be seen from Tab. 3, the Weibull distribution is the best distribution according to the AIC (Xie, 2011) or the minimum entropy criterion (Huang, 2024). The Gamma distribution is the best distribution according to the Q-statistic (Xie, 2011) or the maximum informity criterion (this study). Xie (2011) pointed out that the Gamma distribution was also significantly supported by the data, as evidenced by its low ΔAIC value 4.7. Xie (2011) also pointed out, "AIC selects a "good" model which is not necessarily the true model". His argument should also apply to the other three metrics. Therefore, we can say that the four metrics give consistent results for this example.

Table 3

Evaluation of four candidate distributions using four different metrics

Distribution	**AIC** (Xie, 2011)	**Q-statistic** (Xie, 2011)	**Entropy (nats)** (Huang, 2024)	**Informity** (this study)
Exponential	340.3	48.1	0.846	0.5835
Weibull	**276.6**	7.07	**0.680**	0.5835
Gamma	281.3	**6.34**	0.694	**0.6010**
Normal	315.9	27.6	0.780	0.5346

5.3. Informity gain when splitting a mixed class

Zhou (2019a) considered an example of splitting a mixed class (a set) of 5 blues and 5 greens. He split the class into two branches (two subsets): the left branch had 4 blues and the right branch had 1 blue and 5 greens. He calculated that the entropy before the split was 1 and after the split was 0.39. The information gain (IG), defined as the difference in entropy before and after the split, is 0.61 bits.

We consider this example using the informity gain. Before the split, the informity of the mixed class $\beta_{\text{before}}(X) = 0.5$. After the split, $\beta_{\text{after}}(X) = 0.8333$. Then, the informity gain $\Delta\beta = 0.3333$.

Zhou (2019b) also considered this example using Gini impurity metric. Gini impurity (denoted by *G*) can be written as

$$G(X) := \sum_{i=1}^{N} P(x_i)[1 - P(x_i)] = 1 - \sum_{i=1}^{N} P(x_i)^2. \tag{23}$$

Gini gain ΔG is defined as

$$\Delta G := G_{\text{before}}(X) - G_{\text{after}}(X). \tag{24}$$

Note that Gini impurity *G* is related to the discrete informity as $G(X) = 1 - \beta(X)$. Therefore, the Gini gain ΔG is the same as the informity gain $\Delta\beta$. For this example, $\Delta G = \Delta\beta = 0.3333$.

Discussion and conclusions

The proposed informity theory (a novel probability framework) is established in the original probability space; it does not involve the log-transformation of probabilities like Shannon information theory and is more applicable to fields other than communications engineering. The mathematical quantity "informity" provides a quantitative measure of the degree of informativeness of a probability system. The idea of informity is intuitive and consistent with our common sense or perceptual understanding about information or informativeness. Informity is an important property of a probability distribution. It is the counterpart of entropy and provides insights into probability distributions from a different perspective than entropy. We have derived the informities for twelve continuous distributions. Informities for many more distributions need to be derived. Furthermore, just as the entropy for a distribution is often included in Wikipedia, the informity for the distribution may also be included in Wikipedia.

The proposed informity metric can be used as an alternative to entropy metric. For a given set of data or information, the proposed maximum informity criterion can be used to select the best distribution among a set of candidate distributions. When training decision trees in machine learning, the proposed maximum informity gain criterion can be used to select the best split among a set of candidate splits. The three examples presented demonstrate the practicability of the proposed informity metric.

The author envisions that the proposed informity theory will have broad application prospects. Currently, the author has used the proposed informity theory to develop a new index for measuring the tail-heaviness of a probability distribution (called the tail-heaviness index), two new indices for measuring the difference between two probability distributions (one called the "distribution similarity index (DSI)" and the other called the "distribution discrepancy index (DDI)"), and a new index for measuring the peakedness of a probability distribution (called the peakedness index). The manuscripts describing these developments have been posted as preprints on ResearchGate. Further research is needed to extend and enhance the proposed informity theory and its applications.

Appendix

Converting the *relative* quantity "probability" into the *absolute* quantity "Shannon information content"

Consider a discrete random variable *X* with the PMF $P(x)$ that has *N* possible outcomes. The total amount of information contained in the probability system $\{X; P(x)\}$ is normalized to 1, i. e. $\sum_{i=1}^{N} P(x_i) = 1$. We view probability $P(x_i)$ as the *relative* information content of the outcome x_i. Let n_i denote the reciprocal of $P(x_i)$, i. e. $n_i = \frac{1}{P(x_i)}$. This means that the outcome x_i has one chance of occurring out of n_i possible outcomes. In Shannon information theory, n_i is "the number of messages in the set" and "… this number or any monotonic function of this number can be regarded as a measure of the information [content] produced when one message is chosen from the set, all choices being equally likely (Shannon, 1948)." Therefore, either n_i or any monotonic function of n_i can be regarded as the *absolute* amount of information conveyed (or held) by the outcome x_i, i. e. the *absolute* information content of the outcome x_i.

The *absolute* amount of information contained in the probability system $\{X; P(x)\}$ can be calculated as

$$\sum_{i=1}^{N} P(x_i) n_i = \sum_{i=1}^{N} \frac{1}{n_i} n_i = N. \tag{25}$$

At first glance, this equation seems trivial. However, if we take the logarithm of n_i (note: logarithm is a monotonic function), Eq. (25) becomes the expression for entropy

$$\mathrm{H}(X) = \sum_{i=1}^{N} P(x_i) \log_2(n_i) = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i). \tag{26}$$

It is important to note that the logarithm of a quantity *only* changes the scale of the quantity; it *does not* change the nature or innate character of the quantity. Therefore, both Shannon information content $\log_2(n_i) = -\log_2 P(x_i)$ and entropy $\mathrm{H}(X)$ are the *absolute* quantities.

However, for a continuous random variable *Y* with the PDF $p(y)$, we cannot define the absolute information content of a *y* value. But we can define the absolute information content of a probability interval as the reciprocal of the coverage probability of the probability interval. For example, for a coverage probability of 0.9, the absolute amount of information contained in the probability interval is 1/0.9=11.11, or $-\log_2 0.9 = 0.152$ bits.

References

Castrup, H. (2004). Selecting and applying error distributions in uncertainty analysis. *Measurement Science Conference*. Anaheim. http://www.isgmax.com/Articles_Papers/Selecting%20and%20Applying%20Error%20Distributions.pdf

Ellerman, D. (2013). An Introduction to Logical Entropy and its Relation to Shannon Entropy. *International Journal of Semantic Computing*, 7(2), 121–145. https://doi.org/10.1142/S1793351X13400059

Ellerman, D. (2022). Introduction to logical entropy and its relationship to Shannon entropy. *4open*, 5, 1–33. https://doi.org/10.1051/fopen/2021004

Huang, H. (2024). A minimum entropy criterion for distribution selection for measurement uncertainty analysis. *Measurement Science and Technology,* 35, 035014. https://iopscience.iop.org/article/10.1088/1361-6501/ad1476

Lad, F., Sanfilippo, G. & Agrò, G. (2015). Extropy: Complementary Dual of Entropy. *Statistical Science,* 30(1), 40–58. https://doi.org/10.1214/14-STS430

Petty, G. W. (2018). On some shortcomings of Shannon entropy as a measure of information content in indirect measurements of continuous variables. *Journal of Atmospheric and Oceanic Technology,* 35(5), 1011–1021. https://doi.org/10.1175/JTECH-D-17-0056.1

Rousseau, R. (2018). The repeat rate: from Hirschman to Stirling. *Scientometrics,* 116, 645–653. https://doi.org/10.1007/s11192-018-2724-8

Shannon, C. E. (1948). A mathematical theory of communications. *The Bell System Technical Journal,* 27(3), 379–423.

Schroeder, M. J. (2004). An alternative to entropy in the measurement of information. *Entropy,* 6, 388–412.

Sussmann, G. (1997). Uncertainty relation: from inequality to equality. *Zeitschrift fur Naturforschung A,* 52, 1–2. https://doi.org/10.1515/zna-1997-1-214

Wadsworth, Jr. H. M. (1989). Summarization and interpretation of data. In Harrison M Wadsworth Jr. (Ed.), *Handbook of Statistical Methods for Engineers and Scientists* (2.1–2.21). McGRAW-HILL In.

Xie, G. (2011). *Further developments of two point process models for fine-scale time series.* Massey University, New Zealand.

Zhou, V. (2019a). *A simple explanation of information gain and entropy.* https://victorzhou.com/blog/information-gain

Zhou, V. (2019b). *A simple explanation of Gini impurity.* https://victorzhou.com/blog/gini-impurity

Отримано редакцією журналу / Received: 19.01.25
Прорецензовано / Revised: 13.03.25
Схвалено до друку / Accepted: 14.03.25

Хенінг ХУАНГ, д-р філософії, наук. співроб.
ORCID ID: 0000-0003-1882-0042
e-mail: heninghuang1@gmail.com
Теледін РД Інструментс, Сан-Дієго, США

ТЕОРІЯ ІНФОРМАТИВНОСТІ: НОВИЙ ПІДХІД ДО ТЕОРІЇ ЙМОВІРНОСТЕЙ

Запропоновано новий підхід до теорії ймовірностей, що отримав назву теорія інформативності. Для кількісного вимірювання рівня інформативності ймовірнісного розподілу (або системи ймовірностей) визначено математичну величину під назвою "інформативність". Також встановлено дві інші величини: перехресну інформативність і спільну інформативність. Запропоновано метрику інформативності, яку можна використовувати як альтернативу метриці ентропії. Наведено значення інформативності для дванадцяти неперервних розподілів. Подано три приклади, що демонструють практичність запропонованої метрики інформативності.

Ключові слова: ***інформативність, інформаційний контент, ймовірність, розподіл ймовірностей.***

Автор заявляє про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.