DIMACS Technical Report 97-36
August 1997

FDOD Function and the Information Discrepancy Contained in Multiple Probability Distributions [1]

by

Weiwu Fang[2]
DIMACS Center, Rutgers University
IAM-CAS and APORC

[1]This work has been supported by DIMACS Center, the Chinese NSF under the grant 79370082, and American NSF under grant BIR 94-12594
[2]IAM-CAS: Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing
APORC: Asian-Pacific Operations Research Center

DIMACS is a cooperative project of Rutgers University, Princeton University, AT&T Bell Laboratories and Bellcore.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

ABSTRACT

The concept of Shonnon information has played a significant role in a variety of scientific and engineering areas. The question naturally arises: how can we measure information differences contained in two or more probability distributions ? The answer to this problem will be very interesting in both theory and practice. Some measures for the cases of two or three distributions have presented by the pioneers, but these measures have some disadvantages; moreover, there doesn't exist a measure for n distributions so far.

A FDOD function with many good properties has been introduced in the study of information discrepancy of judgments of multiple experts (FW 1994). In this paper, based on the ideas concerned with Shannon information and measures of difference, we propose an axiom set for measuring the information discrepancy contained in a group of distributions, and prove that the only function satisfying the axiom set is of the FDOD form. The final results and even the intermediate results indeed show the close connection of the FDOD function with Shannon information and the measures of difference in statistics.

Key words: Shannon information, difference, measure, multiple distributions

1 Introduction

The concept of Shonnon Entropy has played a significant role in a variety of scientific and engineering areas. Its original meaning is a measure of uncertainty information of a probability distribution, but it is also used to denote other concepts, such as disorder, equality (discrepancy), diversity and so on.

Let

$$\Gamma_t := \{(p_1, p_2, \cdots, p_t) \mid \sum_{k=1}^{t} p_k = 1 \text{ and } p_k \geq 0\}, \quad (t = 2, 3, \cdots).$$

Then Shonnon entropy can be written as

$$H(\mathbf{P}) = -\sum_{i=1}^{t} x_i \log x_i.$$

The question naturally arises : how can we measure information differences contained in two or more probability distributions ? Apparently, the answer to this problem will be very interesting in both theory and applications, and it will motivate applications of the entropy concept on a higher level. The direct use of Shannon entropy is not a good idea, e.g., the probability distributions $\mathbf{P_1}(p_1, p_2, p_3, \cdots, p_t)$ and $\mathbf{P_2}(p_2, p_3, \cdots, p_t, p_1)$ has the same value of Shannon entropy, but for most of situations they are different distributions and contain different information. Some measures for the cases of two or three distributions have presented by the pioneers, such as Kullback-Leiber entropy and crossing entropy (Kullback, 1959)

$$H(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{t} p_i \log \frac{p_i}{q_i} \quad \text{and} \quad H(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{t} p_i \log \frac{p_i}{q_i} + \sum_{i=1}^{t} q_i \log \frac{q_i}{p_i}, \mathbf{p}, \mathbf{q} \in \Gamma_t,$$

for mathematically measuring information discrepancy between two probability distributions. As to three distributions $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \Gamma_t$, Nath (1968), Renyi (1965), and Kannappan and Rathie (1972) have given their measures respectively as follows:

$$H(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \sum_{k=1}^{t} p_k \cdot \log \frac{q_k}{r_k}$$

(Nath),

$$H(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \frac{1}{\alpha - 1} \log \sum_{k=1}^{t} p_k q_k^{\alpha - 1} r_k^{1-\alpha}$$

and (Renyi)

$$H(\mathbf{p},\mathbf{q},\mathbf{r}) = (2^{\alpha-1}-1)^{-1}\left(\sum_{k=1}^{t} p_k q_k^{\alpha-1} r_k^{1-\alpha} - 1\right)$$

(Kannappan and Rathie).

Although they have different names, the most important fact is that all of them satisfy the identity property, i.e., if $\mathbf{p}=\mathbf{q}=\mathbf{r}$, then $H(\mathbf{p},\mathbf{q},\mathbf{r})=0$.

These measures have some disadvantages. For example, they are not symmetric, and their values could be infinite; it is hard to understand this result because the amount of Shannon information is, as is well known, finite. Let us see a specific example. Given s distributions $\mathbf{p_i} \in \Gamma_t$ $(i=1,\cdots,s)$, $s=t$,

$$\begin{array}{rcl} \mathbf{p_1} & = & (1,0,\cdots,0), \\ \mathbf{p_2} & = & (0,1,\cdots,0), \\ \vdots & & \vdots \\ \mathbf{p_s} & = & (0,0,\cdots,1), \end{array}$$

the outcome of each distribution has complete certainty and all outcomes are completely different from each other. From the meaningfulness of Shannon entropy, their discrepancy should be maximum. In this case, the methods simply can not be used to measure their discrepancies even for any two distributions.

Especially, how to measure information discrepancy contained in $4,5,6,\cdots\cdots$ distributions ? will it be more complicated ?

A bounded FDOD function (Function of Degree of Disagreement) with many good properties has been introduced in the study of information discrepancy of judgments of multiple experts (FW,1994). In the case of vector set, it is reduced as

$$B(\mathbf{U_1},\mathbf{U_2},\cdots,\mathbf{U_s}) = \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}\log\frac{u_{ik}}{\sum_{k=1}^{s} u_{ik}/s},$$

and if we use this function to measure the information discrepancy of the group of distributions mentioned above, the FDOD function has the maximum,

$$B(\mathbf{p_1},\mathbf{p_2},\cdots,\mathbf{p_s}) = s\cdot\log s$$

(see FW 1994, 1997). The result is a little surprising because the maximum of Shannon entropy is, as is well known, $\log s$ for $\mathbf{p_i} \in \mathbf{\Gamma_t}$. Does this result with other good properties imply some kind of closer relationship with Shannon entropy ?

All of these facts motivate us to investigate whether there is a reasonable set of axioms concerned with the basic axioms of Shannon entropy, so that a measure

of information discrepancy based on this axiom set is reasonable and unique for multiple distributions. It is our purpose for this paper. In fact, Kullback (1959), Renyi(1965), Nath(1968), and Kannappan and Rathie(1972) are pioneers in this area although the task has been not completed so far.

In Section 2 of this paper we propose some basic axioms based on the ideas concerned with Shannon axioms and measures of difference, then we find out the general function satisfying the axioms in Section 3. In Section 4, we have given the axiomatic characterization of the information discrepancy. It is easy to find that the intermediate results in Section 3 are surprisingly similar to that of Shannon entropy. To some extent it implies that the axioms proposed in this paper are indeed concerned with Shannon entropy. In Section 5, we discuss other properties of the FDOD function, and they also show many features concerned with the Shannon entropy and measures of difference in statistics. A simple comparison and some possible applications are described in the last two sections.

2 The basic axioms

In order to measure the information discrepancy contained in two or more probability distributions, the axiom set of measurement should be as close to Shannon entropy's axiom set as possible, and also should have the basic features of measures of difference.

Shonnon entropy can be obtained from different axiom sets (Aczél and Darózy, 1975, Forte and Daróczy, 1968). We think the main axioms are:

1. non-negative,
2. symmetric,
3. continuous, and
4. recursive (or additive).

From the axioms 2, 3, and 4, one can get the general solution satisfying the axioms: $H(\mathbf{P}) = -c \cdot \sum_{i=1}^{t} p_i \log p_i$. If one adds the normalized axiom, then it leads to the unique solution: $H(\mathbf{P}) = -\sum_{i=1}^{t} p_i \log p_i$.

We will use these axioms to develop a measure of information discrepancy contained in a group of probability distributions. Of course, we have to make some modifications because it is a different from Shannon entropy and more like a kind of mathematical concept. At first, for a measure of difference defined on multiple quantities, the non-negative, symmetric, and identical axioms are the most basic axioms, so we need add an identical axiom. Secondly, the recursive

axiom should be modified due to the fact that the recursion happens on a group of distributions and not on a group of numbers.

So we propose the corresponding basic axioms as the following:

1. non-negative,
2. symmetric,
3. continuous,
4. identical, and
5. recursive.

For getting a unique solution, we also can add a reasonably normalized axiom (see Section 4).

We need some symbols before we introduce these axioms in detail. Suppose set $I = \{1, 2, \cdots, t\}$ and $K = \{1, 2, \cdots, s\}$ ($s \geq 2$ and $t \geq 1$). Let $G(\mathcal{T}, \mathcal{S})$ or $U := \{U_k | k = 1, \cdots, s\}$ denote a group of probability distributions, where

$$U_k := (u_{1k}, u_{2k}, \cdots, u_{tk})^T \in \Gamma_t.$$

A measure of information discrepancy of a group of probability distributions is defined as a function from (t, s)-tuples u_{ik} of non-negative real numbers to a real number. We denote it by $B(U_1, U_2, \cdots, U_s)$ or $B(V_1, V_2, \cdots, V_t)$ or $B(\mathcal{T}, \mathcal{S})$,

$$B(\mathcal{T}, \mathcal{S}) = B(U_1, U_2, \cdots, U_s) = B(V_1, V_2, \cdots, V_t) = B \begin{pmatrix} u_{11}, & u_{12}, & \cdots, & u_{1s} \\ u_{21}, & u_{22}, & \cdots, & u_{2s} \\ \vdots & \vdots & \vdots & \vdots \\ u_{t1}, & u_{t2}, & \cdots, & u_{ts} \end{pmatrix}$$

where V_i denotes the i-th row elements of all U_k's, i.e.,

$$V_i := (u_{i1}, u_{i2}, \cdots, u_{is}), \quad i = 1, 2, \cdots, t.$$

We should also mention that hereafter, the capitalized V_i or U_k denotes some data group (we sometimes call them as "vectors" due to the fact that we define $\sum_i V_i$ as the normal vector addition), and non-capitalized u_{ik} or u_k^i means some real number.

Axiom 1.*(Non-negative)*

$$B(U_1, U_2, \cdots, U_s) \geq 0.$$

Axiom 2. *(Symmetric) For any permutation σ of K or I*

$$B(U_1, U_2, \cdots, U_s) = B(U_{\sigma(1)}, U_{\sigma(2)}, \cdots, U_{\sigma(s)}) \quad or$$

$$B(V_1, V_2, \cdots, V_t) = B(V_{\sigma(1)}, V_{\sigma(2)}, \cdots, V_{\sigma(t)}).$$

Axiom 3.*(Continuous)* $B(U_1, U_2, \cdots, U_s)$ *is a continuous function of* $U_1, U_2, \cdots, U_s$.

Axiom 4.*(Identical)*

$B(U_1, U_2, \cdots, U_s) = 0$ *iff* $U_1 = U_2 = \cdots = U_s$.

The axioms 1, 2, 3 and 4 are very trivial.

As to recursiveness, for the case of two distributions Kannappan and Ng (1973, 1974) have presented the following recursive axiom (for the case of three distributions, this recursive axiom is substituted by several axioms which are very complex and cannot be described easily here (see the last section of Aczeĺ, 1975)) :

$$E_n(p_1, p_2, \quad \cdots, p_n; q_1, q_2, \cdots, q_n) = E_{n-1}(p_1 + p_2, \cdots, p_n; q_1 + q_2, \cdots, q_n) + \\ +(p_1 + p_2)E_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}; \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right).$$

There are some questions regarding this axiom. At first, why is it not symmetric ? Secondly, why do $\frac{p_1}{p_1+p_2}$ and $\frac{q_1}{q_1+q_2}$ play the same role in the measure ?

Let us discuss the second question in some detail. From the mathematical point of view, $\frac{p_1}{p_1+p_2}$ only means a ratio factor, and it implies a scale of the influence of p_1 in the merging operation. In the case dealing with a distribution, $\frac{p_1}{p_1+p_2}$ and $\frac{p_2}{p_1+p_2}$ can be used to scale the influences of p_1 and p_2 respectively. However, it will be seen from the following examples that they shouldn't directly be adopted in the case of a group of distributions. Suppose first that two merged terms of a distribution are 0.3 and 0.5, and those of another distribution are 0.0003 and 0.0005. After they are merged, the first distribution changes significantly and the second only a little, but $\frac{p_1}{p_1+p_2}$ and $\frac{q_1}{q_1+q_2}$ have the same value 3/8 and they will play the same role in the recursive axiom. As a figurative example, suppose we divide a cake into 5 pieces in proportion to 1:2:3:4:5, then continue to divide each piece into 5 smaller pieces. Can we think that any two smaller pieces have the same size or proportion relative to the whole cake ? Apparently, the answer is negative. So we think that this kind of ratio factors is not a good scaling in a recursive axiom for multiple distributions, and this is its essential difference in comparison to single distribution.

Thus, referring to the grouping axiom on Shannon entropy (Robert Ash, 1965

or 1990),

$$H(p_1,\cdots,p_t) = \quad H(p_1+\cdots+p_r, p_{r+1}+\cdots+p_t) + (p_1+\cdots+p_r)H\left(\frac{p_1}{\sum_{i=1}^{r} p_i},\cdots,\frac{p_r}{\sum_{i=1}^{r} p_i}\right)$$
$$+(p_{r+1}+\cdots+p_t)H\left(\frac{p_{r+1}}{\sum_{i=r+1}^{t} p_i},\cdots,\frac{p_t}{\sum_{i=r+1}^{t} p_i}\right), \qquad r=0,1,2,\cdots,t-1,$$

we propose a grouping recursive axiom of multiple distributions as the following.

Axiom 5. *(Recursive)*

$$B(U_1,\cdots,U_s) = B(V_1,\cdots,V_t) = B(V_1+\cdots+V_r,\quad V_{r+1}+\cdots+V_t) \ +$$
$$+\textstyle\sum_{k=1}^{s}(u_{1k}+\cdots+u_{rk})B\left(\frac{u_{1k}/J(\sum_{k=1}^{s} u_{1k})}{\sum_{i=1}^{r} u_{ik}/J(\sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik})},\cdots,\frac{u_{rk}/J(\sum_{k=1}^{s} u_{rk})}{\sum_{i=1}^{r} u_{ik}/J(\sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik})}\right)$$
$$+\textstyle\sum_{k=1}^{s}(u_{(r+1)k}+\cdots+u_{tk})B\left(\frac{u_{(r+1)k}/J(\sum_{k=1}^{s} u_{(r+1)k})}{\sum_{i=r+1}^{t} u_{ik}/J(\sum_{k=1}^{s}\sum_{i=r+1}^{t} u_{ik})},\cdots,\frac{u_{tk}/J(\sum_{k=1}^{s} u_{tk})}{\sum_{i=r+1}^{t} u_{ik}/J(\sum_{k=1}^{s}\sum_{i=r+1}^{t} u_{ik})}\right),$$

where $r = 0,1,2,\cdots,t-1$ *(*$r=0$ *means the second terms of the right side disappears),* $J(x)=x, x\in\mathbf{R}$ *or* J *is some function which can be considered to be reasonable for this axiom, and* $B(0,0,\cdots,0)=0$ *is defined when* $u_{1k}=u_{2k}=\cdots=u_{rk}=0$ *or* $u_{(r+1)k}=u_{(r+2)k}=\cdots=u_{tk}=0$.

The reason for Axiom 5 is as follows.

First of all, let us observe the grouping recursive axiom of Shannon entropy from the mathematical point of view. The meaningfulness of the first term of the right side of the axiom is obvious. As to the last two, we can regard them as additional terms used to make a correction, whose quantities depend on the values of each merged term p_i and its proportion to $\sum_{i=1}^{r} p_i$.

In the recursive axiom of a group of distributions, as in that of the Shannon entropy, the meaningfulness of the first term of the right side is trivial, and the influences of each merged u_{ik} and its proportion to summations are also considered in the additional terms.

Further we need to discuss the additional terms in more detail.

At first, in the recursive axiom of Shannon entropy, $\sum_{i=1}^{r} p_i$ denotes the merging of numbers $p_1, p_2, \cdots, p_r$, but for a group of distributions the corresponding merging should be a merging of vectors $V_1, V_2, \cdots, V_r$. There still exist s items after merging $V_1, V_2, \cdots, V_r$, so the summation over k from 1 to s is needed, i.e., $(p_1 + \cdots + p_r)H(\cdots)$ is substituted by $\sum_{k=1}^{s}(u_{1k} + \cdots + u_{rk})B(\cdots)$.

Secondly, $\frac{p_i}{\sum_{i=1}^{r} p_i}$ mathematically means just a ratio factor, and it measures the influence of a merged term p_i relative to $\sum_{i=1}^{r} p_i$. In the case of a group of distributions, there are s entries for each row i, and u_{ik} is just one of $u_{i1}, u_{i2}, \cdots, u_{is}$. A similar factor $u_{ik}/\sum_{i=1}^{r} u_{ik}$ is not connected to other distributions, and not a good scale as was mentioned above. Referring to the divided cake problem described before, we think that the influence of u_{ik} in a new ratio factor should be decreased with the increase of $\sum_{k=1}^{s} u_{ik}$, the latter is the simplest measurement concerned with other entries at the same row i. So we use $u_{ik}/J(\sum_{k=1}^{s} u_{ik})$ to correspond to p_i, where $J(x) = x, x \in \mathbf{R}$ or J is some function which can be considered to be reasonable ($J(x) = x$ can be regarded as reasonable from the analysis about the ratio factors, but we also assume that there might exist other functions which can be considered to be reasonable). Due to the same reasons, the $\sum_{i=1}^{r} u_{ik}$ should be changed as the term $\sum_{i=1}^{r} u_{ik}/J(\sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik})$. Finally, corresponding to $\frac{p_i}{\sum_{i=1}^{r} p_i}$ of Shannon's recursive axiom, the ratio factor is $\frac{u_{ik}/J(\sum_{k=1}^{s} u_{ik})}{\sum_{i=1}^{r} u_{ik}/J(\sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik})}$, and so on. The ratios $u_{ik}/\sum_{i=1}^{r} u_{ik}$ and $u_{ik}/\sum_{k=1}^{s} u_{ik}$ are combined in these factors.

In order to be consistent with the identical and continuous axiom, $B(0, 0, \cdots, 0) = 0$ is needed.

Corresponding to the Expansibility of Shannon Entropy, we propose

Axiom 6. *(Expansible)*

$$B(V_1, V_2, \cdots, V_t, \mathbf{0}) = B(V_1, V_2, \cdots, V_t)$$

where $\mathbf{0} = (0, 0, \cdots, 0)$.

Since $\mathbf{0} = (0, 0, \cdots, 0)$ means the identity of all distributions on the $t + 1$-th row and also means $\sum_{i=1}^{t} u_{ik} = \sum_{i=1}^{t+1} u_{ik}$ for any distribution, so its influence on information discrepancy can not be considered. In fact, this axiom serves to meet the mathematical strictness for some special cases, and the Shannon entropy also has this property.

3 The general function satisfying the basic axioms

We first need some lemmas. In the following sections, we assume $V_i \neq \mathbf{0}$ because we needn't consider the cases of $V_i = \mathbf{0}$ from Axiom 2 and Axiom 6.

Lemma 1. *(generalized grouping axiom)*

$$B(V_{11}, \cdots, V_{1m_1}; \cdots\cdots; V_{n1}, \cdots, V_{nm_n}) =$$
$$= B(V_1, \cdots, V_n) + \sum_{k=1}^{s} \sum_{i=1}^{n} u_k^i B\left(\frac{u_{1k}^i / J(\sum_{k=1}^{s} u_{1k}^i)}{u_k^i / J(\sum_{k=1}^{s} u_k^i)}, \cdots, \frac{u_{m_i k} / J(\sum_{k=1}^{s} u_{m_i k}^i)}{u_k^i / J(\sum_{k=1}^{s} u_k^i)} \right)$$

where $V_i = V_{i1} + V_{i2} + \cdots + V_{im_i}$ *and* $u_k^i = u_{1k}^i + \cdots + u_{m_i k}^i$ *, the superscript* i *denotes an entry* $u_{\cdot\cdot}^i \in$ *the* i*-th group.*

Proof. Using Axiom 5, we have

$$B(V_{11}, \cdots, V_{1m_1}; \cdots; V_{n1}, \cdots, V_{nm_n}) = B(V_{11}, \cdots, V_{1m_1}; \cdots; V_{(n-1)1}, \cdots, V_{(n-1)m_{(n-1)}}; V_n) +$$
$$+ \sum_{k=1}^{s} u_k^n B\left(\frac{u_{1k}^n / J(\sum_{k=1}^{s} u_{1k}^n)}{u_k^n / J(\sum_{k=1}^{s} u_k^n)}, \cdots, \frac{u_{m_n k} / J(\sum_{k=1}^{s} u_{m_n k}^n)}{u_k^n / J(\sum_{k=1}^{s} u_k^n)} \right).$$

Shifting V_n to the extreme left, we continue with the reduction, finally obtaining the desired result after n steps.

Lemma 2. $B(1, 1, \cdots, 1) = 0$.

Proof. It is trivial from $U_1 = U_2 = \cdots = U_s = 1$ and Axiom 4.

Lemma 3. *The only form of* $J(x)$ *is* $c \cdot x$*, where* c *is an arbitrary non-zero real number.*

Proof. Given a group $G(2, 2)$ in which $U_1 = U_2 = (x, 1-x)^T$ and $x \in]0, 1[$, it is easy to know $B(U_1, U_2) = 0$ from Axiom 4. Using Axiom 5, we have

$$B(U_1, U_2) = B(1, 1) + 2 \cdot B(\frac{x / J(2 * x)}{1 / J(2)}, \frac{(1-x) / J(2-2x)}{1 / J(2)}).$$

From Lemma 2 and Axiom 4, one gets

$$\frac{x}{J(2 * x)} = \frac{(1-x)}{J(2-2x)} \quad \text{i.e.,} \quad \frac{x}{1-x} = \frac{J(2 * x)}{J(2 * (1-x))}.$$

Because x is an arbitrary real number in the interval $]0,1[$, the unique and elementary solution of the last equation is $J(x) = c \cdot x$, where c is an arbitrary non-zero real number.

According to Lemma 1 and Lemma 3, we have the following lemma:

Lemma 4. *(generalized grouping lemma)*

$$B(V_{11},\cdots,V_{1m_1};\cdots;V_{n1},\cdots,V_{nm_n}) =$$
$$= B(V_1,\cdots,V_n) + \sum_{k=1}^{s}\sum_{i=1}^{n} u_k^i B\left(\frac{u_{1k}^i/\sum_{k=1}^{s} u_{1k}^i}{u_k^i/\sum_{k=1}^{s} u_k^i},\cdots,\frac{u_{m_ik}/\sum_{k=1}^{s} u_{m_ik}^i}{u_k^i/\sum_{k=1}^{s} u_k^i}\right)$$

where $u_k^i = u_{1k}^i + \cdots + u_{m_ik}^i$, $V_i = V_{i1} + V_{i2} + \cdots + V_{im_i}$, *and the superscript* i *denotes an entry* $u_{\cdot\cdot}^i \in$ *the* i*-th group.*

Definition 1. $f(t) := B(\overbrace{t,0,0,\cdots,0}^{t})$ *(*t *is an integer).*

Due to the symmetric axiom, $f(t) = B(0,t,0,\cdots,0) = B(0,0,t,\cdots,0) = \cdots = B(0,0,\cdots,0,t)$.

Lemma 5.

$$f(1) = 0.$$

Proof. For a $G(\mathcal{T},\mathcal{S})$, we can divide it into t subgroups, i.e., each subgroup contains only a row. The generalized grouping axiom is still valid for this case, and one gets

$$B(V_1,\cdots,V_t) =$$
$$= B(V_1,\cdots,V_t) + \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik} B\left(\frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_{ik}/\sum_{k=1}^{s} u_{ik}}\right)$$
$$= B(V_1,\cdots,V_t) + \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik} B(1)$$
$$= B(V_1,\cdots,V_t) + s \cdot B(1).$$

It means that $f(1)$ should equal zero due to $s \geq 2$.

Definition 2. A group $G(\mathcal{S},\mathcal{S})$ of distributions is called as a typical discrepancy

group with the parameter a and denoted by $G(\mathcal{S},\mathcal{S})^a$ if it has the form

$$\begin{pmatrix} a,0, & \cdots\cdots, & 0 \\ 0,a, & \cdots\cdots, & 0 \\ \vdots & \ddots & \vdots \\ 0,0, & \cdots\cdots, & a \end{pmatrix}.$$

Lemma 6. *The $B(U_1,U_2,\cdots,U_s)$ of a typical discrepancy group $G(\mathcal{S},\mathcal{S})^1$ is $s\cdot f(s)$.*

Proof. We divide $G(\mathcal{S},\mathcal{S})^1$ into only one group, i.e., this group contains s rows. In this case, $\sum_{i=1}^{t} u_{ik}=1$, $\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}=s$, and $\sum_{k=1}^{s} u_{ik}=1$. From Lemma 2 and Lemma 4, we can obtain

$$\begin{aligned} B(V_1,V_2,\cdots,V_s) &= B(1,1,\cdots,1)+s\cdot B(s,0,0,\cdots,0) \\ &= 0+s\cdot f(s). \end{aligned}$$

Lemma 7. *$f(t)$ is a completely additive number theoretical function, i.e.,*

$$f(n\cdot m)=f(n)+f(m) \quad \textit{holds for} \quad m,n=1,2,\cdots.$$

Proof. If $m=1$, then from Lemma 5

$$f(n)=f(n\cdot 1)=f(n)+0=f(n)+f(1),$$

i.e.,$f(n\cdot m)=f(n)+f(m)$ holds for $m=1$. It is the same for $n=1$.

Suppose $m\geq 2$ and $n\geq 2$. Given a typical discrepancy $G(\mathcal{S},\mathcal{S})^1$ and $s=t=n\cdot m$, we divide it into n subgroups from the top row to the bottom row and each subgroup i contains m rows.

Then according to Lemma 4, we have

$$\begin{aligned} &B(V_1,V_2,\cdots,V_s)=B(V_{11},\cdots,V_{1m};\cdots\cdots;V_{n1},\cdots,V_{nm})= \\ &=B(V_{11}+\cdots+V_{1m},\cdots\cdots,V_{n1}+\cdots+V_{nm})+ \\ &+\sum_{k=1}^{s}\sum_{i=1}^{n} u_k^i B\left(\frac{u_{1k}^i/\sum_{k=1}^{s} u_{1k}^i}{u_k^i/\sum_{k=1}^{s} u_k^i},\cdots,\frac{u_{mk}^i/\sum_{k=1}^{s} u_{mk}^i}{u_k^i/\sum_{k=1}^{s} u_k^i}\right) \end{aligned}$$

where $u_k^i=u_{1k}^i+\cdots+u_{mk}^i$. In the last equation, there is at most one non-zero term among $u_{1k}^i,u_{2k}^i,\cdots,u_{mk}^i$, so u_k^i equals 1 or 0; further, in $u_1^i,u_2^i,\cdots,u_s^i$ there

exists only m terms which equal 1, so $\sum_{k=1}^{s} u_k^i = m$; in addition, $\sum_{k=1}^{s} u_{jk}^i = 1$ for any row j. Thus the last equation becomes

$$\begin{aligned} &B(V_1, V_2, \cdots, V_s) = \\ &= B(V_{11} + \cdots + V_{1m}, \cdots, V_{n1} + \cdots + V_{nm}) + n \cdot m \cdot f(m) \end{aligned} \tag{1}$$

according to Definition 1 and $0 \cdot B(0, 0, \cdots, 0) = 0$.

We continue to merge the rows in the first term of the right side of Equation (1) and get

$$\begin{aligned} &B(V_1, V_2, \cdots, V_s) = \\ &= B(1, 1, \cdots, 1) + \sum_{k=1}^{s} u_k B\left(\frac{u_k^1 / \sum_{k=1}^{s} u_k^1}{u_k / \sum_{k=1}^{s} u_k}, \cdots, \frac{u_k^n / \sum_{k=1}^{s} u_k^n}{u_k / \sum_{k=1}^{s} u_k} \right) + n \cdot m \cdot f(m) \end{aligned} \tag{2}$$

where $u_k = \sum_{i=1}^{n} u_k^i$. In (2), $u_k = \sum_{i=1}^{n} u_k^i = 1$ because there is only one non-zero term in each column; further, $\sum_{k=1}^{s} u_k = s = m \cdot n$ and $\sum_{k=1}^{s} u_k^i = m$. Thus (2) is reduced as

$$B(V_1, V_2, \cdots, V_s) = 0 + s \cdot f(n) + s \cdot f(m). \tag{3}$$

We also can directly obtain

$$B(V_1, V_2, \cdots, V_s) = s \cdot f(s)$$

from Lemma 6. From (3) and the last equation, one gets

$$s \cdot f(s) = s \cdot f(n) + s \cdot f(m), \quad \text{i.e.,} \quad f(m \cdot n) = f(n) + f(m). \tag{5}$$

Lemma 8. (Aczel,1976) *Let ϕ be a completely additive number theoretical function, If, and only if, we have*

$$n \cdot \phi(n) \leq (n+1) \cdot \phi(n+1) \quad \textit{for} \quad n = 1, 2, \cdots,$$

will there exist a constant $c \geq 0$ such that

$$\phi(n) = c \cdot \log n \qquad (n = 1, 2, \cdots).$$

Lemma 9. $f(t) = c \cdot \log t \quad t = 1, 2, \cdots,$ *where c is a constant ($c \geq 0$).*

Proof. If $n = 1$, $1 \cdot f(1) \leq (1+1) \cdot f(1+1) = 2 \cdot B(2,0)$ from Axiom 1 and Lemma 5.

Suppose $s \geq 2$, and given a typical discrepancy $G(s+1, s+1)^1$, we first merge the first s rows into one row and get

$$\begin{aligned} & B(U_1, U_2, \cdots, U_{s+1}) \\ &= B\begin{pmatrix} 1,1,\cdots,1,0 \\ 0,0,\cdots,0,1 \end{pmatrix} + s \cdot f(s) \\ &= s \cdot B(\tfrac{s+1}{s}, 0) + B(0, s+1) + s \cdot f(s). \end{aligned}$$

Directly using Lemma 6, one gets $B(U_1, U_2, \cdots, U_{s+1}) = (s+1) \cdot f(s+1)$ and then has

$$(s+1) \cdot f(s+1) = s \cdot B(\frac{s+1}{s}, 0) + B(0, s+1) + s \cdot f(s).$$

It means

$$(s+1) \cdot f(s+1) \geq s \cdot f(s)$$

because the first two terms of the right side of the last equation are greater or equal to zero from Axiom 1. Finally, the conclusion is obtained immediately from Lemma 8 and 7.

Lemma 10.

$$B(r, 0, \cdots, 0) = c \cdot \log r \qquad \textit{holds, where } r \textit{ is a real number and } \; r \geq 1.$$

Proof.

At first we prove $B(r, 0) = c \cdot \log r$.

We assume p is a rational number, and $p = \frac{m}{s}$ and $m < s$. Given a typical discrepancy group $G(\mathcal{S}, \mathcal{S})^1$, we get

$$B(\mathcal{S}, \mathcal{S})^1 = s \cdot B(\frac{1/1}{1/s}, 0, \cdots, 0) = s \cdot B(s, 0, \cdots, 0) = s \cdot f(s) \tag{6}$$

by merging all rows. We also can merge the top m rows and the bottom $s - m$ rows respectively, it leads to

$$\begin{aligned} B(\mathcal{S}, \mathcal{S})^1 = \; & B\begin{pmatrix} \overbrace{1,1,\cdots,1,}^{m} & 0,\ 0,\cdots,0 \\ 0,\ 0,\cdots,0, & \underbrace{1,1,\cdots,1}_{s-m} \end{pmatrix} + m \cdot B(\tfrac{1/1}{1/m}, 0, \cdots, 0) + \\ & + (s-m) \cdot B(\tfrac{1/1}{1/(s-m)}, 0, \cdots, 0). \end{aligned}$$

Continuing to merge the rows in the first term of the right side of the last equation, one gets

$$\begin{aligned} B(\mathcal{S},\mathcal{S})^1 &= B(1,\cdots,1) + m\cdot B(\tfrac{1/m}{1/s},0) + (s-m)B(\tfrac{1/(s-m)}{1/s},0)+ \\ &+m\cdot B(m,0,0,\cdots,0) + (s-m)B(s-m,0,0,\cdots,0) \\ &= 0 + m\cdot B(\tfrac{s}{m},0) + (s-m)B(\tfrac{s}{(s-m)},0) + mf(m) + (s-m)f(s-m). \end{aligned} \tag{7}$$

From (6), (7) and Lemma 9, one gets

$$\frac{m}{s}B(\frac{s}{m},0) + \frac{(s-m)}{s}B(\frac{s}{(s-m)},0) = c\cdot(\log s - \frac{m}{s}f(m) - \frac{(s-m)}{s}f(s-m)).$$

Noticing $p = m/s$, from the last equation we have

$$\begin{aligned} p\cdot B(\tfrac{1}{p},0) + (1-p)B(\tfrac{1}{1-p},0) &= c\cdot(\log s - \tfrac{m}{s}\log m - \tfrac{(s-m)}{s}\log(s-m)) \\ &= c\cdot(\log\tfrac{s}{(s-m)} + \tfrac{m}{s}\log\tfrac{(s-m)}{s}) \\ &= c\cdot(\log\tfrac{1}{1-p} + p\log\tfrac{1-p}{p}) \\ &= c\cdot(-\log(1-p) + p\log(1-p) - p\log p) \\ &= c\cdot(p\log\tfrac{1}{p} + (1-p)\log\tfrac{1}{(1-p)}). \end{aligned} \tag{8}$$

Because p is an arbitrary rational number less than 1 and greater than zero, the elementary and unique solution of (8) is

$$B(q,0) = c\cdot\log q \quad (q \text{ is any rational number and } q > 1).$$

By the continuity axioms, this result immediately extends to all irrational $r > 1$.

Since $\lim_{r\to 1+} B(r,0) = \lim_{r\to 1+} c\cdot\log r = 0$, and $c\cdot\log r = 0$ for $r = 1$, from the continuous axiom we have

$$B(r,0) = c\cdot\log r. \quad \text{for} \quad r \geq 1. \tag{9}$$

Now we prove $B(\overbrace{r,0,\cdots,0}^{s}) = c\cdot\log r$ holds for any s and $r \geq 1$.

Given the following $G(\mathcal{S}, 2*(\mathcal{S}-1))$ in which p is a real number less than 1 and greater than zero,

$$G(\mathcal{S}, 2*(\mathcal{S}-1)) = \begin{pmatrix} \overbrace{p\ ,p\ ,\cdots\ ,p\ ,}^{s-1} & \overbrace{0\ ,0\ ,\cdots\ ,0}^{s-1} \\ 0\ ,0\ ,\cdots\ ,0\ , & q\ ,0\ ,\cdots\ ,0 \\ 0\ ,0\ ,\cdots\ ,0\ , & 0\ ,q\ ,\cdots\ ,0 \\ \vdots & \vdots \quad \ddots \quad \vdots \\ 0\ ,0\ ,\cdots\ ,0\ , & 0\ ,0\ ,\cdots\ ,q \end{pmatrix} \qquad \text{where} \quad q = 1-p.$$

Merging the bottom $s-1$ rows by Lemma 4 and (9), one gets

$$\begin{aligned} B(\mathcal{S}, 2*(\mathcal{S}-1)) &= B\begin{pmatrix} \overbrace{p\ ,p\ ,\cdots\ ,p\ ,}^{s-1} & \overbrace{0\ ,0\ ,\cdots\ ,0}^{s-1} \\ 0\ ,0\ ,\cdots\ ,0\ , & q\ ,q\ ,\cdots\ ,q \end{pmatrix} + q(s-1)B(s-1,0,0,\cdots,0) \\ &= B(p\ ,\cdots\ ,p\ ,q\ ,\cdots\ ,q\ ) + p(s-1)B(\tfrac{1}{p},0) + q(s-1)B(\tfrac{1}{q},0) + cq(s-1)\log(s-1) \\ &= B(p\ ,\cdots\ ,p\ ,q\ ,\cdots\ ,q\ ) + cp(s-1)\log(\tfrac{1}{p}) + cq(s-1)\log(\tfrac{1}{q}) + cq(s-1)\log(s-1). \end{aligned} \tag{10}$$

We also can directly merge all s rows by Lemma 4 and get

$$\begin{aligned} B(\mathcal{S}, 2*(\mathcal{S}-1)) \ &= B(p\ ,\cdots\ ,p\ ,q\ ,\cdots\ ,q\ ) + p(s-1)B(\overbrace{\frac{1}{p},0,\cdots,0}^{s}) + \\ &\quad + q(s-1)B(\overbrace{\frac{s-1}{q},0,\cdots,0}^{s}). \end{aligned} \tag{11}$$

Combining (10) and (11), we obtain

$$\begin{aligned} pB(\overbrace{\frac{1}{p},0,\cdots,0}^{s}) + (1-p)B(\overbrace{\frac{s-1}{1-p},0,\cdots,0}^{s}) \ &= cp\log(\tfrac{1}{p}) + cq\log(\tfrac{1}{1-p}) + cqt\log(s-1) \\ &= cp\log(\tfrac{1}{p}) + c(1-p)\log(\tfrac{s-1}{1-p}). \end{aligned}$$

As p is an arbitrary real number ($0 < p < 1$) and s is an arbitrary integer, we have the following elementary solution

$$B(\overbrace{r,0,\cdots,0}^{s}) = c\cdot\log r \qquad (r > 1).$$

Similarly, we can extend it to the case of $r = 1$ due to the continuous axiom.

Lemma 11.

$$\sum_{k=1}^{s} u_k B\left(\frac{u_{1k}/\sum_{k=1}^{s} u_{1k}}{u_k/\sum_{k=1}^{s} u_k}, \cdots, \frac{u_{rk}/\sum_{k=1}^{s} u_{rk}}{u_k/\sum_{k=1}^{s} u_k}\right) = c \cdot \sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik} \log \frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k/\sum_{k=1}^{s} u_k}$$

where $u_k = \sum_{i=1}^{r} u_{ik}$.

Proof.

We construct two groups $G_1(\mathcal{R} * \mathcal{S}, \mathcal{S})$ and $G_2(\mathcal{R} * \mathcal{S}, \mathcal{S})$ as the following,

$$G_1(\mathcal{R}*\mathcal{S}, \mathcal{S}) = \begin{pmatrix} u_{11}, & 0 & , & 0 & , \cdots & , & 0 \\ 0 & , u_{12}, & 0 & , \cdots & , & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & , & 0 & , & 0 & , \cdots & , u_{1s} \\ u_{21}, & 0 & , & 0 & , \cdots & , & 0 \\ 0 & , u_{22}, & 0 & , \cdots & , & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & , & 0 & , & 0 & , \cdots & , u_{2s} \\ & \vdots & & \vdots & & \vdots & \\ & \vdots & & \vdots & & \vdots & \\ u_{r1}, & 0 & , & 0 & , \cdots & , & 0 \\ 0 & , u_{r2}, & 0 & , \cdots & , & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & , & 0 & , & 0 & , \cdots & , u_{rs} \end{pmatrix} \text{ and } G_2(\mathcal{R}*\mathcal{S}, \mathcal{S}) = \begin{pmatrix} u_{11}, & 0 & , & 0 & , \cdots & , & 0 \\ u_{21}, & 0 & , & 0 & , \cdots & , & 0 \\ \vdots & & \vdots & & \vdots & & \\ u_{r1}, & 0 & , & 0 & , \cdots & , & 0 \\ 0 & , u_{12}, & 0 & , \cdots & , & 0 \\ 0 & , u_{22}, & 0 & , \cdots & , & 0 \\ \vdots & & \vdots & & \vdots & & \\ 0 & , u_{r2}, & 0 & , \cdots & , & 0 \\ & \vdots & & \vdots & & \vdots & \\ & \vdots & & \vdots & & \vdots & \\ 0 & , & 0 & , & 0 & , \cdots & , u_{1s} \\ 0 & , & 0 & , & 0 & , \cdots & , u_{2s} \\ \vdots & & \vdots & & \vdots & & \\ 0 & , & 0 & , & 0 & , \cdots & , u_{rs} \end{pmatrix}.$$

From Axiom 2, it is easy to know

$$B_1(\mathcal{R} * \mathcal{S}, \mathcal{S}) = B_2(\mathcal{R} * \mathcal{S}, \mathcal{S}). \tag{12}$$

Merging each s rows of $G_1(\mathcal{R} * \mathcal{S}, \mathcal{S})$ from top to bottom, one gets

$$\begin{aligned} B_1(\mathcal{R} * \mathcal{S}, \mathcal{S}) &= B\begin{pmatrix} u_{11}, u_{12}, \cdots, u_{1s} \\ u_{21}, u_{22}, \cdots, u_{2s} \\ \vdots \quad \ddots \quad \vdots \\ u_{r1}, u_{r2}, \cdots, u_{rs} \end{pmatrix} + u_{11}B(\tfrac{u_{11}/u_{11}}{u_{11}/\sum_{k=1}^{s} u_{1k}}, 0, \cdots, 0) + \\ &+ u_{12}B(0, \tfrac{\sum_{k=1}^{s} u_{1k}}{u_{12}}, 0, \cdots, 0) + \cdots + u_{rs}B(0, 0, \cdots, 0, \tfrac{\sum_{k=1}^{s} u_{rk}}{u_{rk}}) \\ &= B(u_1, u_2, \cdots, u_s) + \sum_{k=1}^{s} u_k B\begin{pmatrix} u_{1k}/\sum_{k=1}^{s} u_{1k} & \cdots & u_{rk}/\sum_{k=1}^{s} u_{rk} \\ \hline u_k/\sum_{k=1}^{s} u_k & & u_k/\sum_{k=1}^{s} u_k \end{pmatrix} \\ &+ c \cdot \sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik} \cdot \log \tfrac{\sum_{k=1}^{s} u_{ik}}{u_{ik}} \end{aligned} \tag{13}$$

from Lemma 4 and Lemma 10. It is also easy to obtain

$$\begin{aligned} B_2(\mathcal{R} * \mathcal{S}, \mathcal{S}) &= B\begin{pmatrix} u_1, & 0, \cdots, & 0 \\ 0, & u_2, \cdots, & 0 \\ \vdots & \ddots & \vdots \\ 0, & 0, \cdots, & u_s \end{pmatrix} + s \cdot B(1, 1, \cdots, 1) + r(s-1) \cdot 0 \cdot B(0, 0, \cdots, 0) \\ &= B(u_1, u_2, \cdots, u_s) + \sum_{k=1}^{s} u_k B(\tfrac{\sum_{k=1}^{s} u_k}{u_k}, 0, \cdots, 0) \\ &= B(u_1, u_2, \cdots, u_s) + c \cdot \sum_{k=1}^{s} u_k \cdot \log \tfrac{\sum_{k=1}^{s} u_k}{u_k} \end{aligned} \tag{14}$$

according to Lemma 4 and Lemma 10.
Combining (12), (13) and (14), we have

$$\begin{aligned} &\sum_{k=1}^{s} u_k B\begin{pmatrix} u_{1k}/\sum_{k=1}^{s} u_{1k} & \cdots & u_{rk}/\sum_{k=1}^{s} u_{rk} \\ \hline u_k/\sum_{k=1}^{s} u_k & & u_k/\sum_{k=1}^{s} u_k \end{pmatrix} \\ &= c \cdot \sum_{k=1}^{s} u_k \cdot \log \tfrac{\sum_{k=1}^{s} u_k}{u_k} - c \cdot \sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik} \cdot \log \tfrac{\sum_{k=1}^{s} u_{ik}}{u_{ik}} \\ &= c \cdot \sum_{k=1}^{s} \left[(\sum_{i=1}^{r} u_{ik}) \cdot \log \tfrac{\sum_{k=1}^{s} u_k}{u_k} - \sum_{i=1}^{r} u_{ik} \cdot \log \tfrac{\sum_{k=1}^{s} u_{ik}}{u_{ik}} \right] \\ &= c \cdot \sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik} \log \tfrac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k/\sum_{k=1}^{s} u_k}. \end{aligned} \tag{15}$$

Definition 3. *The function $F(U_1, U_2, \cdots, U_s) : \Gamma_t^s \to R$ is called as a FDOD function if*

$$F(U_1, U_2, \cdots, U_s) = \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik} \log(u_{ik}/K_i)$$

where $K_i = \sum_{k=1}^{s} u_{ik}/s$, *and* $0 \cdot \log 0 = 0$ *is defined (cf.Corollary 1 in FW 1994).*

Theorem 1. *If* $B(U_1, U_2, \cdots, U_s) : \Gamma_t^s \to R$ *is non-negative, symmetric, continuous, identical, recursive and expansible, then* $B(U_1, U_2, \cdots, U_s)$ *is the FDOD function with a non-negative constant, i.e.,*

$$B(U_1, U_2, \cdots, U_s) = c \cdot F(U_1, U_2, \cdots, U_s),$$

for all

$$U_1, U_2, \cdots, U_s \in \Gamma_t \quad and \quad s \geq 2.$$

Proof.

Suppose $U_1, U_2, \cdots, U_s \in \Gamma_t$, and $B(U_1, U_2, \cdots, U_s)$ satisfies all conditions mentioned above. Using Lemma 11 and noticing $u_k = \sum_{i=1}^{t} u_{ik} = 1$ in this case, we merge all rows and obtain

$$\begin{aligned}
B(U_1, U_2, \cdots, U_s) &= B\begin{pmatrix} u_{11}, u_{12}, \cdots, u_{1s} \\ u_{21}, u_{22}, \cdots, u_{2s} \\ \vdots \quad \ddots \quad \vdots \\ u_{t1}, u_{t2}, \cdots, u_{ts} \end{pmatrix} = \\
&= \sum_{k=1}^{s} u_k B\left(\frac{u_{1k}/\sum_{k=1}^{s} u_{1k}}{u_k/\sum_{k=1} u_k}, \cdots, \frac{u_{tk}/\sum_{k=1}^{s} u_{tk}}{u_k/\sum_{k=1} u_k} \right) + B(1, 1, \cdots, 1) \\
&= c \cdot \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log \frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k/\sum_{k=1}^{s} u_k} + 0 \\
&= c \cdot \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log(u_{ik}/K_i) = c \cdot F(U_1, U_2, \cdots, U_s)
\end{aligned} \tag{16}$$

where $K_i = \sum_{k=1}^{s} u_{ik}/s$.

4 The normalized axiom and the characterization of the information discrepancy

In order to obtain a unique solution, a normalized axiom, $I(1/2, 1/2) = \log 2$, was introduced in the study of Shannon entropy. We can also add a reasonably normalized axiom for getting a unique solution. In fact, the normalized axiom is given just for determining a scale. In the case of a group of distributions,

$$B\begin{pmatrix} 1/2, & 1/2 \\ 1/2, & 1/2 \end{pmatrix} = 0$$

from the identical axiom, so we can't directly use the original normalized axiom and it should be modified. We have known that the value of Shannon entropy of $p_3 = (1/2, 1/2) \in \Gamma_2$ is $\max_{p\in\Gamma_2} H(p) = \log 2$, and that $H(p_1) = H(p_2) = 0$ for $p_1 = (1, 0)$ and $p_2 = (0, 1)$; further, $p_1 = (1, 0)$ means a complete certainty, $p_2 = (0, 1)$ represents an oppositely complete certainty, and $p_3 = (1/2, 1/2)$ implies the complete uncertainty. Thus a reasonable assumption is: the information discrepancy between p_1 and p_2 should be two times the maximum $\log 2$. So we propose the normalized axiom as follows.

Axiom 7. *(normalized).*

$$B\begin{pmatrix} 1, & 0 \\ 0, & 1 \end{pmatrix} = 2 \cdot \log 2. \tag{17}$$

Lemma 12. *(FW,1994).*

If $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = a > 0, \quad x_i > 0, \ \textit{and} \ \ y_i > 0, \ \textit{then} \ \ \sum_{i=1}^{n} x_i \log(x_i/y_i) \geq 0$

with equality iff $x_i = y_i$ *for all* i.

If define $0 \cdot \log 0 = 0$ *and* $0 \cdot \log(0/0) = 0$, *Then the condition of this lemma can be weakened to as* $\sum_{i=1}^{n} x_i = a$, $\sum_{i=1}^{n} y_i = a$ *where* $x_i \geq 0$ *and* $y_i > 0$, *or* $x_i = 0$ *and* $y_i = 0$.

Theorem 2. *If, and only if,* $B(U_1, U_2, \cdots, U_s) : \Gamma_t^s \to R$ *is non-negative, symmetric, continuous, identical, recursive, expansible and normalized,* $B(U_1, U_2, \cdots, U_s)$ *is the FDOD function, i.e.,*

$$B(U_1, U_2, \cdots, U_s) = F(U_1, U_2, \cdots, U_s),$$

for all

$$U_1, U_2, \cdots, U_s \in \Gamma_t \quad \textit{and} \quad s \geq 2.$$

Proof.

The "if"part follows from the previous results. From (15) of Theorem 1 and Axiom 6, it is easy to get $c = 1$ and

$$B(U_1, U_2, \cdots, U_s) = \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log(u_{ik}/K_i) = F(U_1, U_2, \cdots, U_s). \tag{18}$$

As to the "only if" part, It is trivial for the symmetric, normalized, expansible and continuous axioms.

Since $\sum_{i=1}^{t} K_i = \sum_{i=1}^{t}\sum_{k=1}^{s}\frac{u_{ik}}{s} = \frac{1}{s}\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}$=1, $\sum_{i=1}^{t}\sum_{k=1}^{s} K_i = s$, and $\sum_{k=1}^{s} K_i = \sum_{k=1}^{s} u_{ik}$, we have

$$\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}\log(u_{ik}/K_i) \geq 0, \quad \sum_{i=1}^{t} u_{ik}\log(u_{ik}/K_i) \geq 0, \quad \text{and} \sum_{k=1}^{s} u_{ik}\log(u_{ik}/K_i) \geq 0$$

from Lemma 12, i.e., the FDOD function satisfies the non-negative axiom.

Now we prove that FDOD function also satisfies the recursive axiom (let $J(x) = c \cdot x$).

Let $u_k^1 = u_{1k} + u_{2k} + \cdots + u_{rk} = \sum_{i=1}^{r} u_{ik}$ and $u_k^2 = u_{(r+1)k} + u_{(r+2)k} + \cdots + u_{tk} = \sum_{i=r+1}^{t} u_{ik}$, and $(u_k^1 + u_k^2) = 1$. Directly using (18), we have

$$B(V_1 + V_2 + \cdots + V_r, V_{r+1} + V_{r+2} + \cdots + V_t) = B\begin{pmatrix} u_1^1, u_2^1, \cdots, u_s^1 \\ u_1^2, u_2^2, \cdots, u_s^2 \end{pmatrix}$$
$$= \sum_{k=1}^{s} u_k^1 \log\frac{u_k^1/\sum_{k=1}^{s} u_k^1}{1/s} + \sum_{k=1}^{s} u_k^2 \log\frac{u_k^2/\sum_{k=1}^{s} u_k^2}{1/s}$$

and

$$B(V_1, V_2, \cdots, V_t) = \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}\log(u_{ik}/K_i)$$
$$= \sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik}\log(u_{ik}/K_i) + \sum_{k=1}^{s}\sum_{i=r+1}^{t} u_{ik}\log(u_{ik}/K_i);$$

further, we get

$$\sum_{k=1}^{s} u_k^1 \quad B\left(\frac{u_{1k}/\sum_{k=1}^{s} u_{1k}}{u_k^1/\sum_{k=1}^{s} u_k^1}, \cdots, \frac{u_{rk}/\sum_{k=1}^{s} u_{rk}}{u_k^1/\sum_{k=1}^{s} u_k^1}\right) +$$
$$+ \sum_{k=1}^{s} u_k^2 B\left(\frac{u_{(r+1)k}/\sum_{k=1}^{s} u_{(r+1)k}}{u_k^2/\sum_{k=1}^{s} u_k^2}, \cdots, \frac{u_{tk}/\sum_{k=1}^{s} u_{tk}}{u_k^2/\sum_{k=1}^{s} u_k^2}\right)$$
$$= \sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik}\log\frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k^1/\sum_{k=1}^{s} u_k^1} + \sum_{k=1}^{s}\sum_{i=r+1}^{t} u_{ik}\log\frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k^2/\sum_{k=1}^{s} u_k^2}$$

according to Lemma 11. Thus, these formulae lead to

$$B(V_1+\cdots+V_r, V_{r+1}+\cdots+V_t)+\sum_{k=1}^{s} u_k^1 B\left(\frac{u_{1k}/\sum_{k=1}^{s} u_{1k}}{u_k^1/\sum_{k=1}^{s} u_k^1},\cdots,\frac{u_{rk}/\sum_{k=1}^{s} u_{rk}}{u_k^1/\sum_{k=1}^{s} u_k^1}\right)+$$

$$+\sum_{k=1}^{s} u_k^2 B\left(\frac{u_{(r+1)k}/\sum_{k=1}^{s} u_{(r+1)k}}{u_k^2/\sum_{k=1}^{s} u_k^2},\cdots,\frac{u_{tk}/\sum_{k=1}^{s} u_{tk}}{u_k^2/\sum_{k=1}^{s} u_k^2}\right)$$

$$\begin{aligned}
&=\sum_{k=1}^{s} u_k^1 \log\frac{u_k^1/\sum_{k=1}^{s} u_k^1}{1/s}+\sum_{k=1}^{s} u_k^2 \log\frac{u_k^2/\sum_{k=1}^{s} u_k^2}{1/s}\\
&+\sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik}\log\frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k^1/\sum_{k=1}^{s} u_k^1}+\sum_{k=1}^{s}\sum_{i=r+1}^{t} u_{ik}\log\frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k^2/\sum_{k=1}^{s} u_k^2}\\
&=\sum_{k=1}^{s}(\sum_{i=1}^{r} u_{ik})\log\frac{u_k^1/\sum_{k=1}^{s} u_k^1}{1/s}+\sum_{k=1}^{s}(\sum_{i=r+1}^{t} u_{ik})\log\frac{u_k^2/\sum_{k=1}^{s} u_k^2}{1/s}\\
&+\sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik}\log\frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k^1/\sum_{k=1}^{s} u_k^1}+\sum_{k=1}^{s}\sum_{i=r+1}^{t} u_{ik}\log\frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k^2/\sum_{k=1}^{s} u_k^2}\\
&=\sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik}\log(u_{ik}/K_i)+\sum_{k=1}^{s}\sum_{i=r+1}^{t} u_{ik}\log(u_{ik}/K_i)\\
&=\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}\log(u_{ik}/K_i)=F(V_1,V_2,\cdots,V_t).
\end{aligned}$$

That means $F(V_1, V_2, \cdots, V_t)$ satisfies the recursive axiom.

5 The other properties of FDOD function

There are other algebraic, analytic, and statistical properties on FDOD function, which also show some connections of the function with Shannon entropy and difference in statistics.

First of all, the FDOD function satisfies the two generalized properties as follows.

Branching property. *There exists a sequence of functions*

$$J_t : \{(V_1, V_2) : V_1, V_2 \in \Gamma_t\} \to R(t = 3, 4, \cdots)$$

such that

$$B_t(V_1, V_2, \cdots, V_t) - B_{t-1}(V_1 + V_2, V_3, \cdots, V_t) = J_t(V_1, V_2)$$

for all $U_1, U_2, \cdots, U_s \in \Gamma_t \quad t = 3, 4, \cdots.$

In fact we can choose

$$J_t(V_1, V_2) =$$
$$= \sum_{k=1}^{s} u_k B\left(\frac{u_{1k}/\sum_{k=1}^{s} u_{1k}}{(u_k/\sum_{k=1}^{s} u_k} , \frac{u_{2k}/\sum_{k=1}^{s} u_{rk}}{u_k/\sum_{k=1}^{s} u_k} \right) = \sum_{k=1}^{s} \sum_{i=1}^{2} u_{ik} \log \frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k/\sum_{k=1}^{s} u_k}$$

and $J_t(\mathbf{0},\mathbf{0}) = 0$, for all $t \geq 3$, where $u_{1k} + u_{2k} = u_k$.

Sum property. *There exists a function G such that*

$$B(V_1, V_2, \cdots, V_t) = \sum_{i=1}^{t} G(V_i).$$

for all $U_1, U_2, \cdots, U_s \in \Gamma_t \quad t = 2, 3, \cdots$.

It is also obvious if we let $G(V_i) = \sum_{k=1}^{s} u_{ik} \log \frac{u_{ik}}{K_i}$ in the FDOD function.

If

$$B_k(U_1, U_2, \cdots, U_s) = \sum_{i=1}^{t} u_{ik} \log \frac{u_{ik}}{K_i} \tag{19}$$

is defined as an information discrepancy of a distribution U_k relative to the remainder, then it is easy to obtain the two following properties from other properties.

Linear homogeneity property (FW, 1994).

$$B(rU_1, rU_2, \cdots, rU_s) = rB(U_1, U_2, \cdots, U_s) \quad \textit{and}$$
$$B_k(rU_1, rU_2, \cdots, rU_s) = rB_k(U_1, U_2, \cdots, U_s)$$

where r is a positive real number.

Symmetric property relative to a distribution.

$B_k(U_1, U_2, \cdots, U_s)$ is a function with non-negative, symmetric, continuous, identical, bounded, expansible, and linearly homogeneous properties.

Convexity property (FW, 1996).

$B(\mathcal{T}, \mathcal{S})$ is a convex function with respect to any U_k, i.e., to $u_{1k}, u_{2k}, \cdots, u_{tk}, (k = 1, \cdots s)$.

Given t and s, we denote s by $s = m \cdot t + n$ ($m = [s/t]$ is the greatest integer not greater than s/t).

Definition 4. A $G(\mathcal{T}, \mathcal{S})$ is called as a distribution group with symmetric structure if $s \equiv 0 \pmod{t}$.

Bounded property (FW, 1994). *Given s and t, an upper bound of $B(\mathcal{T},\mathcal{S})$ is $s \cdot \log t$.*

From Theorem 5 of FW,1997, we have a reduced result:

Maximum property. *For any s and t, the maximum of $\{B(\mathcal{T},\mathcal{S})\}$ is indeed*

1. $s \cdot \log\ t$ if $s \geq t$ and $\{G(\mathcal{T},\mathcal{S})\}$ is the set of distribution groups with symmetry structure; or

2. $s \cdot \log s$ if $s \leq t$.

In addition, we have already proved that there always exists a sort of typical groups of distributions, whose FDOD values are not less than $0.918 \cdot s \cdot \log t$ and close to $s \cdot \log t$ (FW,1997).

Monotonicity (FW,1994).

$$B(U_1, U_2, \cdots, U_s) \leq B(U_1, \cdots, U_s, U_{s+1}).$$

This property implies that the FDOD value of a new group of distributions will be higher than that of the original group if the new group includes all original members. This property implies the accumulation of discrepancy.

Suppose $U_1, U_2, \cdots, U_s \in \Gamma_t$ and U_a is the average distribution of $U_1, U_2, \cdots, U_s \in \Gamma_t$, i.e., $U_a = (u_{1a}, u_{2a}, \cdots, u_{ta})^T$ and $u_{ia} = \sum_{k=1}^{s} \frac{u_{ik}}{s}$, then we have the following result:

Invariance for its average (FW,1994).

$$B(U_1, U_2, \cdots, U_s) = B(U_1, \cdots, U_s, U_a) \quad \textit{and}$$

$$BK_k(U_1, U_2, \cdots, U_s) = BK_k(U_1, \cdots, U_s, U_a).$$

Similarity to some measures of difference in statistics.

We can write $B(U_1, U_2, \cdots, U_s)$ as

$$\begin{aligned}
&B(U_1, U_2, \cdots, U_s) = \textstyle\sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log u_{ik} - \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log(\frac{\sum_{k=1}^{s} u_{ik}}{s}) \\
&= \textstyle\sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log u_{ik} - s \sum_{i=1}^{t} u_{ia} \log u_{ia} \\
&= \textstyle\sum_{k=1}^{s} \left((-\sum_{i=1}^{t} u_{ia} \log u_{ia}) - (-\sum_{i=1}^{t} u_{ik} \log u_{ik}) \right) \\
&= \textstyle\sum_{k=1}^{s} (x_a - x_k)
\end{aligned}$$

where $x_a = -\sum_{i=1}^{t} u_{ia} \log u_{ia}$ is the Shannon information of the average of all distributions and $x_k = -\sum_{i=1}^{t} u_{ik} \log u_{ik}$ is the Shannon information of an individual

distribution. Combining the last formula with the non-negativity, identity monotonicity, and invariance for its average, we think FDOD function is somewhat similar to some difference formulae in statistics. For example, given a statistical sample, $\{y_i\}$ is a set of n observations and $\overline{y}$ is its mean of the observations, then $(x_a - x_k)$ is somewhat similar to the deviation $(y_i - \overline{y})$ of a observation from the mean of the observations, and $\sum_{k=1}^{s}(x_a - x_k)$ is similar to $\sum_{i=1}^{n}(y_i - \overline{y})^2$ of n deviations due to the fact that both of them are non-negative, identical, monotonic, and invariant for its average (fortunately, the sum of our "deviations" is not always equal to zero, so it is not necessary to work with the sum of their squares).

Thus, FDOD function is not only concerned with the Shannon information entropy, but also with the difference concepts in statistics, so we can think it to be a measure of information discrepancy contained in multiple probability distributions.

6 A simple comparison with other measures

We give a simple comparison shown in the following table.

Basic properties	Shannon entropy	K-L measure	FDOD function
Basic meaning of concept	Uncertainty, disorder, ... on a distribution	discrepancy of two distributions	discrepancy of a group of distributions
Data amount (number of distributions)	1	2	$s \geq 2$
Non-negativity	Yes	Yes	Yes
Identity		Yes	Yes
Symmetry	Yes	No	Yes
Boundedness	Yes	No	Yes
Uniform continuity	Yes	No	Yes
Upper bound	$\ln t$	∞	$s \cdot \ln t$
Convexity or concavity w.r.t $\mathbf{x} \in \mathbf{\Gamma_t}$	concavity	convexity	convexity
Monotonicity		No	Yes
Invariance for averages		No	Yes
Grouping Recursiveness	Yes	No	Yes
Smallness for small probabilities	Yes	No	Yes
Expansibility	Yes	Yes	Yes
Linear homogeneity		Yes	Yes
Sum	Yes	Yes	Yes
Limitation for data except for $\sum_{i=1}^{t} x_i = 1$	No	Yes (if $q_i = 0$)	No

7 Some possible applications

A measure of information discrepancy of multiple distributions may have applications in many areas, we describe a few possible applications here.

1. Principle on average effect and its application

Principle on average effect. *If U_a is the average distribution of $U_1, U_2, \cdots, U_s \in$*

Γ_t, *i..e.,* $U_a = (u_{1a}, u_{2a}, \cdots, u_{ta})^T$ *and* $u_{ia} = \sum_{k=1}^{s} \frac{u_{ik}}{s}$, *then*

$$-\frac{1}{s}\sum_{k=1}^{s}(\sum_{i=1}^{t} u_{ik}\log u_{ik}) \leq -\sum_{i=1}^{t} u_{ia}\log u_{ia}$$

and the equality holds iff $U_1 = U_2 = \cdots = U_s$, *i.e., the Shannon entropy of the average distribution is greater than or equal to the average of the sum of Shannon entropies of all distributions.*

Proof. From the non-negative property of FDOD function and $K_i = \sum_{k=1}^{s} u_{ik}/s$, we have

$$\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}\log\frac{u_{ik}}{K_i} \geq 0, \qquad \text{and}$$

$$\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}\log u_{ik} \geq \sum_{i=1}^{t}\sum_{k=1}^{s} u_{ik}\log(\frac{\sum_{k=1}^{s} u_{ik}}{s}),$$

$$\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}\log u_{ik} \geq s\sum_{i=1}^{t}(\frac{\sum_{k=1}^{s} u_{ik}}{s})\log(\frac{\sum_{k=1}^{s} u_{ik}}{s}), \qquad \text{and}$$

$$-\frac{1}{s}\sum_{k=1}^{s}(\sum_{i=1}^{t} u_{ik}\log u_{ik}) \leq -\sum_{i=1}^{t} u_{ia}\log u_{ia}.$$

From Lemma 11, one can get that the equality holds iff $\sum_{k=1}^{s} u_{ik}/s = u_{ik}$ holds for any i and k, it also means that the equality holds iff $U_1 = U_2 = \cdots = U_s$.

This principle may be used to explain some phenomena in different areas.

For example, if we measure the equality of wealth by Shannon entropy, then a government can decrease the difference measured between poverty and wealth by averaging statistical data of local states.

Another example may be interesting for both statistical mechanics and information theory. As some authors indicated, " The basic premises in statistical mechanics raise many questions of a fundamental nature at the conceptual level, such as irreversibility and its relation to mechanics, ergodicity, and so on" (Eu, 1992). In addition, since the 1960's an interesting argument has been : is there some kind of linkage between the information entropy and the thermodynamic entropy or not ? It is still a challenging problem in many respects. So many of the confusions in statistical mechanics are sometimes caused by the lack of proper mathematical tools, which can be used to handle a system with n particles. We feel that Shannon entropy, average principle and other results on the FDOD function together might be able to lead to a new approach to investigate the concept

of system entropy and irreversible processes; moreover, it would connect the information theory with the classical theory of thermodynamic equilibrium smoothly. In this paper, we present some ideas for further investigations.

As an example, an entropy model of a simple discrete system with irreversible adiabatic changes may be as follows.

This model is not constructed the ensemble way but in a more direct fashion. Given a system at equilibrium or non-equilibrium state, s denotes the number of particles of the system, and t is the number of possible energy levels, $\epsilon_1, \epsilon_2, \cdots, \epsilon_t$. Suppose u_{ik} is the probability that the particle k in the system has energy ϵ_i, then the entropy of a particle k is $-c_0 \sum_{i=1}^{t} u_{ik} \log u_{ik}$, and the entropy of the system can be naturally defined as

$$GS := -c_0 \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log u_{ik} \quad \text{and}$$

$$\sum_{i=1}^{t} u_{ik} = 1, \ k = 1, \cdots, s, \tag{20}$$

where c_0 is some physical constant.

In this basic model, only "entropy" and "energy" are involved considering the first statement of the Second Law enunciated by Clausius (1879) (The energy of the world is constant. The entropy of the world tends to a maximum). For this model, we use a simple and natural postulate as similarly described by Hecht (1990): *all particles in a system will have equal a priori probability over each possible energy level only when the system is at a thermodynamic equilibrium state;* however, we do not require the critical assumption that" There exists a one to one correspondence between states of thermodynamic equilibrium and states of maximum entropy" (Mackey, 1991).

At first we find out its maximum in a linear constraint and prove this value is exactly equal to that of the classical equilibrium theory in a linear constraint.

If we know the expected energy of all particles in the system is E, then the constraint is

$$\sum_{k=1}^{s} \sum_{i=1}^{t} \epsilon_i u_{ik} = E. \tag{21}$$

The Lagrangian is

$$L \equiv -c_0 \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log u_{ik} - \sum_{k=1}^{s} (\lambda_k - c_0)(\sum_{i=1}^{t} u_{ik} - 1) - \mu_0 \sum_{k=1}^{s} \sum_{i=1}^{t} (\epsilon_i u_{ik} - E).$$

Setting the derivatives of the Lagrangian with respect to u_{ik} equal to zero, one gets

$$-c_0 \log u^*_{ik} - \lambda_k - \mu_0 \epsilon_i = 0, \quad i = 1, \cdots, t, \quad k = 1, \cdots, s$$

or

$$u^*_{ik} = e^{-\frac{1}{c_0}(\lambda_k + \mu_0 \epsilon_i)}.$$

From (20) and the last equality, one gets

$$\sum_{i=1}^{t} u_{ik} = \sum_{i=1}^{t} e^{-\frac{1}{c_0}(\lambda_k + \mu_0 \epsilon_i)} = e^{-\frac{1}{c_0}\lambda_k} \sum_{i=1}^{t} e^{-\frac{\mu_0}{c_0}\epsilon_i} = 1,$$

i.e.,

$$e^{-\frac{1}{c_0}\lambda_k} = \frac{1}{\sum_{i=1}^{t} e^{-\frac{\mu_0}{c_0}\epsilon_i}} = 1, \quad \text{and } u^*_{ik} = \frac{e^{-\frac{\mu_0}{c_0}\epsilon_i}}{\sum_{i=1}^{t} e^{-\frac{\mu_0}{c_0}\epsilon_i}}.$$

Let $\mu = \mu_0 / c_0$, we have

$$u^*_{ik} = \frac{e^{-\mu\epsilon_i}}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}}, \quad k = 1, 2, \cdots, s. \tag{22}$$

It is a unique stationary solution, and its physical meaning is that all particles have the same probability over the energy level ϵ_i. This fact implies that the system is at an equilibrium state from the postulate; moreover, the entropy GS of the system at this state is just maximum according to the principle on average effect. It also means that a process from a non-equilibrium state to a equilibrium will lead to an increase of the entropy GS in the fixed circumstances and E.

The last equation is just the Maxwell-Boltzmann distribution in traditional equilibrium statistical mechanics. We will show that the entropy GS of the system will exactly equal the entropy of traditional equilibrium statistical mechanics by choosing proper physical parameters only.

From the book of E. Schrödinger, a Nobel laureate of Physics, (1989,1952, chapter 2 and 3), we have known the following results of traditional equilibrium statistical mechanics:

$$\frac{E}{s} = \frac{\sum_{i=1}^{t} \epsilon_i e^{-\mu\epsilon_i}}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} \tag{23}$$

where $\mu = \frac{1}{kT}$ and T is the temperature of an equilibrium system, and

$$S(\text{entropy}) = k \log \sum_{i=1}^{t} e^{-\epsilon_i / kT} + \frac{1}{T} \frac{\sum_{i=1}^{t} \epsilon_i e^{-\epsilon_i / kT}}{\sum_{i=1}^{t} e^{-\epsilon_i / kT}}. \tag{24}$$

Substituting (22) into (21), (21) becomes (23), i.e., we can take $\mu = \frac{1}{kT}$ in (22); further, substituting (22) into GS and taking $c_0 = k/s$, we have

$$\begin{aligned} GS &= -c_0 \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik}^* \log u_{ik}^* = -c_0 s \sum_{i=1}^{t} \frac{e^{-\mu\epsilon_i}}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} \log \frac{e^{-\mu\epsilon_i}}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} \\ &= -c_0 s \Big[\frac{1}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} [\sum_{i=1}^{t} e^{-\mu\epsilon_i} (-\mu\epsilon_i - \log \sum_{i=1}^{t} \epsilon^{-\mu\epsilon_i})] \Big] \\ &= c_0 s \Big[\frac{1}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} [\mu \sum_{i=1}^{t} \epsilon_i e^{-\mu\epsilon_i} + (\sum_{i=1}^{t} e^{-\mu\epsilon_i})(\log \sum_{i=1}^{t} \epsilon^{-\mu\epsilon_i})] \Big] \\ &= c_0 s \Big[\mu \frac{\sum_{i=1}^{t} \epsilon_i e^{-\mu\epsilon_i}}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} + \log \sum_{i=1}^{t} e^{-\mu\epsilon_i} \Big] \\ &= k \log \sum_{i=1}^{t} e^{-\epsilon_i/kT} + \frac{1}{T} \frac{\sum_{i=1}^{t} \epsilon_i e^{-\epsilon_i/kT}}{\sum_{i=1}^{t} e^{-\epsilon_i/kT}}. \end{aligned} \tag{25}$$

The last result (25) exactly equals (24). Thus GS can be rewritten as

$$GS = -\frac{k}{s} \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log u_{ik} \tag{26}$$

and (26) can be reduced as $GS = -k \sum_{i=1}^{t} u_i \log u_i$ when a system is at equilibrium state.

Secondly, using the FDOD function instead of the conditional entropy, we discuss a specific evolution of the system entropy GS from a non-equilibrium to equilibrium state. Some further postulates are needed for a dynamical system as usual. Because the probability distribution of all particles over the possible energy levels will tend to the same when t(time) $\to \infty$ (i.e., when the system tends to the equilibrium state), we assume:

Assumption A: For fixed circumstances and energy E, the total discrepancy of entropies among all particles will decrease with an increase time, i.e., if times $t_2 \geq t_1$ and the FDOD function is used to measure the discrepancy among all particles, then $F(t_2) \leq F(t_1)$.

Since the probability distributions of all particles will tend to the same, the value of the probability of a particle over an energy level can be considered as a random shift around a medium, so in the case of a great number of particles, we can expect:

Assumption B: For fixed circumstances and energy E, the average probability of all particles over an energy level is invariant in the process due to the Law of Large Numbers.

Assumption B is somewhat similar to the density property of canonical ensembles obtained from the Liouville equation: the entropy of the density will be

constant for all time and will have the value determined by the initial density with which the system is prepared(Mackey, 1990).

Thus, from Assumption A, (18), and (26), we have

$$\begin{aligned} F(t_2) \quad & -F(t_1) \\ & = \tfrac{k}{s}(\textstyle\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}^2 \log \frac{u_{ik}^2}{K_i^2} - \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}^1 \log \frac{u_{ik}^1}{K_i^1}) \\ & = \tfrac{k}{s}(\textstyle\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}^2 \log u_{ik}^2 - \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}^1 \log u_{ik}^1 - \sum_{i=1}^{t}(\sum_{k=1}^{s} u_{ik}^2)(\log K_i^2) + \\ & \quad \textstyle\sum_{i=1}^{t}(\sum_{k=1}^{s} u_{ik}^1)(\log K_i^1)) \\ & = GS^1 - GS^2 - \tfrac{k}{s}(\textstyle\sum_{i=1}^{t}(\sum_{k=1}^{s} u_{ik}^2)(\log K_i^2) - \sum_{i=1}^{t}(\sum_{k=1}^{s} u_{ik}^1)(\log K_i^1)) \quad \leq 0 \end{aligned}$$

In the last formula, $\sum_{k=1}^{s} u_{ik}^1/s = \sum_{k=1}^{s} u_{ik}^2/s$ and $K_i^1 = K_i^2$ from Assumption B, so one can get

$$GS^2 \geq GS^1,$$

and this means the increase of the entropy GS with a time increase.

In a word, this example shows us a new possibility of investigating the concept of system entropy and irreversible process. The physical meaningfulness of this model is quite clear and natural, it has the weaker premises, and its connection with the classical equilibrium theory is smooth and consistent.

Of course, a lot of work should be done before it becomes a real frame that would be of interest for both statistical mechanics and information theory.

2. On weighting problems

We explain this by a problem of weighting decision criteria.

Nutt reported a field study which compared direct and indirect criteria weighting methods [Nutt, 1980]. The average criteria weights assigned by the DM's for each method are shown in Table 2.

Criteria	Rank weight	Point assignment	ARS	Log-odds	Regression
Quality of life	0.37	0.23	0.24	0.30	0.30
Disability relieved	0.20	0.19	0.20	0.18	0.22
Years life gained	0.10	0.20	0.19	0.17	0.24
Target group size	0.21	0.22	0.23	0.23	0.14
Cost per case	0.12	0.16	0.13	0.13	0.10

Table 2. The criteria weight set by each method

To determine if these weights were significantly different, the Pearson product moment correlations were calculated as shown in Table 3.

	Regression	Rank weight	Point assignment	ARS	Log-odds
Regression		0.59	0.63	0.64	0.65
Rank weight	0.59		0.73	0.74	0.93
Point assignment	0.63	0.73		0.97	0.92
ARS	0.64	0.74	0.97		0.89
Log-odds	0.65	0.93	0.92	0.89	

Table 3. Pearson correlation matrix

According to Table 3, an expert could get an intuitive impression. Nutt's intuition was the following: "The rank weight method produced criteria weighting that were somewhat dissimilar from ARS and point assignment. In contrast the log-odds method produced weights that were quite similar to the ARS, point assignment, and rank weight methods, ... The observed differences between regression and the other methods could be attributed to ..... . Also, the rank weight method ..... also produced different weighting from two of the remaining three methods".

We can use the FDOD function to determine if these weights were significantly different.

At first, using the formula (19) and the Table 2, we can get the information discrepancy as follows:

$$\begin{aligned} B_{rank-weight} &= 3.23 \times 10^{-2}, \\ B_{point-assignment} &= 1.17 \times 10^{-2}, \\ B_{ARS} &= 0.588 \times 10^{-2}, \\ B_{log-odds} &= 0.267 \times 10^{-2}, \\ \text{and} B_{Regression} &= 2.57 \times 10^{-2} \end{aligned}$$

These results illustrate that the discrepancy of the log-odds method relative to the reminder is the smallest one, but that of the rank-weight method is bigger and it is also true for regression method. All are consistent with Nutt's intuition and it seems to be simpler and more clear.

Secondly, if we use the FDOD function to make paired comparisons for this example, then most of results are consistent with the Pearson correlation matrix. For the observed differences, the results of FDOD function seem to be more reasonable according to an intuitive analysis; further, if we use symmetric cross-entropy to make paired comparisons, all results for this example are consistent with that of the paired comparisons of FDOD function .

Thirdly, we also can obtain an overall evaluation about the original data by the procedure described in FW (1994).

3. On similarity of multiple sequences in molecular biology

There is a basic need in molecular biology to have a measure of similarity between multiple sequences. Some algorithms, depended on a lot of paired comparison of two sequences, have been proposed(Chen, Wong and Chiu, 1992). Referring to the good properties of the FDOD function, we are developing a algorithms to measure some information in molecular biology. It is involved in integrated and complex comparisons, so we will discuss it in another paper.

8 Conclusion

In this paper, based on the concept of Shannon information and measure of difference, we have proposed the axiom set for measuring the information discrepancy contained in multiple distributions, and proved that the FDOD function is the unique solution satisfying the axiom set. The final results indeed show the close connection of the FDOD function with Shannon information and the measures of difference, and they may be able to be applied in various areas.

Acknowledgments

This work has been supported by DIMACS Center, Chinese NSF under Grant 79370082, and American NSF under Grant BIR 94-12594 . The author gratefully thanks Prof. F. Roberts, Prof. H.K. Kim. Prof. F. Roush, Prof. G. Torrance, Prof. G. Steiner, and Prof. T. Daniel for their help.

References

[1] J.Aczél and Z.Daróczy, On Measures of Information and Their Characterization, Academic Press, New York, London, 1975.

[2] J.Aczél, Functional Equations in Several Variables : with Applications, Cambridge University Press, Cambridge, New York, 1989.

[3] Robert Ash, Information theory, Interscience Pub. New York, 1965, 1990.

[4] Chan, S.C., Wong, A.K. & Chiu, D.K., A Survey of Multiple Sequence Comparison Methods. Bulu. Math. Biol. 54, 1992, 563-598.

[5] Byung Chan Eu, Kinetic Theory and Irreversible Thermodynamics, John Wiley and Sons, New york, 1992.

[6] J. R. Dorfman, Research in Statistical Physics (Abstract), NSF Award Abstract - no. 9600428, 1996.

[7] Weiwu Fang, The Disagreement Degree of Multi-person Judgments in Additive Structure, Mathematical Social Sciences, vol. 28, no. 2 (1994) 85-111.

[8] Weiwu Fang, Some Properties of FDOD Function and Kullback Measure, CSIAM'96 VII:9, 1996.

[9] Weiwu Fang, On a Global Optimization Problem in the Study of Information Discrepancy, DIMACS Tec.Report 97-04 (to appear in J. of Global Optimization Vol.11, No.1, 1-22, 1997).

[10] A.Feinstein, Foundations of Information Theory, McGraw Hill, New York - Toronto, 1958

[11] B.Forte and Z. Daroczy, A Characterization of Shannon's Entropy. Boll. Un. Mat. Ital. (4) 1 (1968), 631-635.

[12] C.E.Hecht, Statistical Thermodynamics and Kinetic Theory, W.H.Freeman and Company, New York (1990).

[13] Pl. Kannappan and P.N. Rathie, An Application of a Functional Equation to Information Theory. Ann. Polon. Math. 26, 95-101, 1972.

[14] Pl. Kannappan and C.T. Ng, Measurable Solutions of Functional Equations Related to Information Theory. Proc. Math. Soc. 38,303-310, 1973.

[15] J.N.Kapur and H.K.Kesavan, Entropy Optimization Principles with Applications. Academic Press, San Diego, 1992.

[16] D.H.Krantz,R.D.Luce, P.Suppes, and A. Tversky, Foundations of Measurement, Vol.1, Academic Press, New York, 1971.

[17] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959

[18] P.M.Lee, On the Axioms of Information Theory. Ann. Math. Statist. 35 (1964), 414-441

[19] M.C. Mackey, Time's Arrow: The origins of Thermodynamic Behavior, Springer-Verlag, New york, 1991.

[20] P. Nath, On the Measures of Errors in Information, J. Math.Sci. 3, 1-16, 1968

[21] A.Renyi, On the Foundations of Information Theory. Rev. Inst. Internat. Statist. 33 (1965), 1-14.

[22] F.Roberts and R.D.Luce, Axiomatic Thermodynamics and Extensive Measurement, Syntheses 18 (1968) 311-326.

[23] Erwin Schrödinger, Statistical Thermodynamics, Dover Pub., Inc., New York, (1989,1952).

[24] C.E. Shannon, A Mathematical Theory of Communication, Bell System Tech. J. 27 (1948), 379-423, 623-656