DIMACS Technical Report 97-36
August 1997

The FDOD Function and 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 69772036, and partially supported by 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

ABSTRACT

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

The FDOD (Function of Degree of Disagreement) function has been introduced in the study of information discrepancies among some multiple information sources (FW 1994, 1997). It possesses some peculiar properties and may have applications in other areas. In this paper, based on the ideas concerned with Shannon entropy and measures of difference, we propose an condition set for measuring information discrepancies contained in a group of distributions, and prove that the only function satisfying the condition set is of the FDOD form. The final results and even intermediate results indeed show a close connection between the FDOD function and Shannon entropy as well as the measures of difference in statistics.

Key words: Shannon entropy, information discrepancy, measurement, multiple distributions

1 Introduction

The concept of Shannon information 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 $P := (p_1, p_2, \cdots, p_t)$ denote a distribution and

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

Shannon entropy can be written as

$$H(P) = -\sum_{i=1}^{t} p_i \log p_i, \qquad P \in \Gamma_t.$$

The question naturally arises : how can we measure information discrepancies contained in two or more distributions ? Apparently, explorations of this problem are of interest in both theory and applications. It is not appropriate to construct a measure by a direct use of Shannon entropy. For example, the Shannon entropy values of distributions $P_1(p_1, p_2, p_3, \cdots, p_t)$ and $P_2(p_2, p_3, \cdots, p_t, p_1)$ are the same, but for most situations these two distributions are different and with different information. Some measures for two or three distributions had been presented by some pioneers, such as Kullback-Leiber entropy and crossing entropy (Kullback, 1959),

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

for mathematically measuring information discrepancy between two probability distributions. As to the cases with three distributions $P, Q, R \in \Gamma_t$, Nath (1968), Renyi (1965), and Kannappan and Rathie (1972) had presented the following measures

$$H(P, Q, R) = \sum_{i=1}^{t} p_i \cdot \log \frac{q_i}{r_i} \qquad \text{(Nath)},$$

$$H(P, Q, R) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{t} p_i q_i^{\alpha-1} r_i^{1-\alpha} \qquad \text{(Renyi)},$$

and

$$H(P, Q, R) = (2^{\alpha-1} - 1)^{-1} \left(\sum_{i=1}^{t} p_i q_i^{\alpha-1} r_i^{1-\alpha} - 1 \right)$$

(Kannappan and Rathie)

respectively. Although the names of these measures are different (e.g., inaccuracy, error, and divergence etc.), all of them have the most important feature — identity property, i.e., if $P = Q = R$, then $H(P, Q, R) = 0$. These measures are quite useful, but there still exist some weaknesses of them. For example, most of them are not symmetric, they do not possess an appropriate recursiveness, can't be applied in singlur cases, their values could be infinite, and so on; it is hard to understand the infiniteness of information discrepancy because the amount of Shannon information is, as is well known, finite. Let us see a specific example. Given s distributions $P_i \in \Gamma_t$ $(i = 1, \cdots, s)$, $s = t$,

$$\begin{array}{rcl} P_1 & = & (1, 0, \cdots, 0), \\ P_2 & = & (0, 1, \cdots, 0), \\ \vdots & & \quad\vdots \\ P_s & = & (0, 0, \cdots, 1), \end{array}$$

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

Especially, there doesn't exist any measure of information discrepancies among n distributions.

The bounded FDOD function (Function of Degree of Disagreement) with many good properties has been introduced in the study of information discrepancy among multiple information sources (FW,1994). In the case of a group of distributions, it is reduced into

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

If we use this function to measure information discrepancies of the group of distributions mentioned above, the FDOD function has the maximum,

$$B(P_1, P_2, \cdots, P_s) = s \cdot \log s$$

(see FW 1994, 1997). The result is a little surprising because the maximum of Shannon entropy is, as is well known, $\log s$ for $P_i \in \Gamma_t$.

In contrast to other measures of entropy family, this function has some peculiar properties, such as monotonicity, invariance for averages, effectiveness for mutually singular distributions, and so on (see Table 1), so it may have applications in some areas, in which the traditional measures of information discrepancy become invalid, for example, in the study of multiple DNA comparison. In addition, the recursiveness is a very basic and important assumption of Shannon's theory, but all discrepancy measures in the entropy family do not possess a recursiveness with symmetric form. All of these facts motivate us to try to find an set of conditions concerned with the basic conditions of Shannon entropy and difference measures, so that the measure of information discrepancy based on this condition set is unique for multiple distributions.

In Section 2 we propose some basic conditions based on the ideas concerned with Shannon conditions and measures of difference, then we find the general solution satisfying the conditions in Section 3. In Section 4, we give the characterization of information discrepancy. It is easy to know that the intermediate results in Section 3 are surprisingly similar to that of Shannon entropy (Aczél, 1975). To some extent it implies that the conditions 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 conditions

A measure of information discrepancy should be concerned with the concepts of Shannon information and difference, so a condition set of measuring information discrepancies should be as close to Shannon's conditions as possible, and should also have the basic features of measures of difference.

Shannon entropy can be obtained from different condition sets (Aczél and Darózy, 1975, Forte and Daróczy, 1968), but the basic conditions of Shannon information are:

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

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

Referring to these conditions, we will propose a condition set of measuring information discrepancy of multiple distributions. Of course, we have to make some modifications of conditions because the information discrepancy is a different from Shannon entropy and more like a mathematical concept. At first, for a measure of difference defined on multiple quantities, the non-negative, symmetric, and identical conditions are the most basic conditions, so it needs to add an identical condition in a new condition set. Secondly, the recursion assumption of Shannon entropy is very important and basic, we, certainly, should keep this assumption with symmetric form; however, it should be modified because new recursion processes will take place among a group of distributions but not a group of numbers.

So we propose the following conditions:

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

To get a unique solution, we can also add a reasonably normalized condition (see Section 4).

We first introduce some notations before describe these conditions. Given the index sets $I = \{1, 2, \cdots, t\}$ and $K = \{1, 2, \cdots, s\}$ ($s \geq 2$ and $t \geq 1$), let $G(\mathcal{T}, \mathcal{S}) := \{U_k | k = 1, \cdots, s\}$ denote a group of distributions, where

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

A measure of information discrepancy of a group of 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 $D(U_1, U_2, \cdots, U_s)$ or $D(V_1, V_2, \cdots, V_t)$ or $D(\mathcal{T}, \mathcal{S})$,

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

where V_i denotes the elements of all U_k's in the i-th row, 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 "vectors" due to the fact that $\sum_i V_i$ is defined as a normal vector addition), and the lower case u_{ik} or u_k^i means some real number.

Condition 1 *(Non-negative).*

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

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

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

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

Condition 3 *(Continuous). $D(U_1, U_2, \cdots, U_s)$ is a continuous function of $U_1, U_2, \cdots, U_s$.*
Condition 4*(Identical).*

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

The conditions 1, 2, 3 and 4 are trivial.

The recursiveness is a very basic and important assumption of Shannon entropy. But most of members in the entropy family do not have any kind of recursiveness. For the case of two distributions, Kannappan and Ng (1973, 1974) have presented the following recursion condition (for the case of three distributions, the recursion condition is substituted by several conditions which are very complex and cannot be described easily here (see the last section of Aczél, 1975)):

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

It is worth notice that this condition is not symmetric, and that $\frac{p_1}{p_1+p_2}$ and $\frac{q_1}{q_1+q_2}$ play the same role in the recursion formula. Let us discuss the latter observation in some detail. From the mathematical point of view, $\frac{p_1}{p_1+p_2}$ only means a ratio, and it implies a scale of the influence of p_1 in a 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 influences of p_1 and p_2 respectively. However, it will be seen from the following example that they are not appropriate in the case of a group of distributions. Suppose that two entries of a distribution P are 0.3 and 0.5, and the corresponding entries of another distribution Q are 0.0003 and 0.0005. After they are merged, the first distribution changes significantly and the second only a little; however, $\frac{p_1}{p_1+p_2}$ and $\frac{q_1}{q_1+q_2}$ equal the same value 3/8 and they will bring the same numerical influence in the recursive formula. So, we think it is not an appropriate scale in a recursion formula of multiple distributions, and this is an essential difference in comparison to the case of single distribution.

Referring to the grouping condition of Shannon entropy (Robert Ash, 1965 or

1990),

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

(where $r=0$ means that the second term of the right side disappears), we propose a grouping recursion condition of multiple distributions as the following.

Condition 5 *(Recursive).*

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

where $r=0,1,2,\cdots,t-1$, $J(x)=x, x\in\mathbf{R}$ *or* J *is some appropriate function, and* $D(0,0,\cdots,0)=0$ *when* $u_{1k}=u_{2k}=\cdots=u_{rk}=0$ *or* $u_{(r+1)k}=u_{(r+2)k}=\cdots=u_{tk}=0$.

The reason of Condition 5 is as follows.

First of all, let us observe the previous grouping condition of Shannon entropy from the mathematical point of view. The meaning of the first term of the right side of the formula is obvious. As to the last two, we can regard them as additional terms for making a correction, whose values depend on merged entries p_i and their proportions to $\sum_{i=1}^{r} p_i$.

In Condition 5, the meaning of the first term of the right side is similar to that in the previous grouping formula; the influences of merged u_{ik} and their proportions to summations are also reckoned in the additional terms.

We need to discuss the additional terms in more detail.

At first, in the previous recursion condition of Shannon entropy, $\sum_{i=1}^{r} p_i$ denotes the merging of numbers $p_1,p_2,\cdots,p_r$, but for the case of multiple distributions, a merging should be a merging of vectors $V_1, V_2, \cdots, V_r$. There still exist s entries after merging $V_1, V_2, \cdots, V_r$, and the influence of each entry should be also reckoned in a recursion formula. So, the summation over k from 1 to s is needed, i.e., $(p_1 + \cdots + p_r)H(\cdots)$ is substituted by $\sum_{k=1}^{s}(u_{1k} + \cdots + u_{rk})D(\cdots)$.

Secondly, in the case of multiple distributions, there are s entries for each row i, and u_{ik} is just one of $u_{i1}, u_{i2}, \cdots, u_{is}$. As mentioned above, a simple ratio $u_{ik}/\sum_{i=1}^{r} u_{ik}$ is not an appropriate scale because it doesn't take account of the ratios between u_{ik} and $u_{i1}, u_{i2}, \cdots, u_{is}$ at the same row i. It is natural that the influence of a u_{ik} in a recursion formula should be decreased with increase of $\sum_{k=1}^{s} u_{ik}$, and the latter is the simplest measure concerned with all entries at the same row i. So, we use the ratio $u_{ik}/J(\sum_{k=1}^{s} u_{ik})$ to correspond to the original p_i, where $J(x) = x, x \in \mathbf{R}$ or J is some appropriate function. ($J(x) = x$ is appropriate from the analysis of ratios above, but we also assume that there might exist other functions which could be appropriate). Due to the same reasons, the ratio $(\sum_{i=1}^{r} u_{ik})/J(\sum_{k=1}^{s}\sum_{i=1}^{r} u_{ik})$ corresponds to the $\sum_{i=1}^{r} p_i$. Finally, corresponding to the ratio $\frac{p_i}{\sum_{i=1}^{r} p_i}$ in the previous grouping condition, the new ratio 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.

In order to be consistent with the identical and continuous condition, $D(0, 0, \cdots, 0) = 0$ is defined; moreover, corresponding to the Expansibility of Shannon entropy, we propose

Condition 6 *(Expansible).*

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

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 implies $\sum_{i=1}^{t} u_{ik} = \sum_{i=1}^{t+1} u_{ik}$ for any distribution, so the influence of $\mathbf{0} = (0, 0, \cdots, 0)$ is nothing to a measure of discrepancies. In fact, Shannon entropy also satisfies a similar property. And both of them serve to meet the mathematical strictness for some special cases.

3 The general function satisfying the basic conditions

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

Lemma 1 *(generalized grouping condition).*

$$D(V_{11},\cdots,V_{1m_1};\cdots\cdots;V_{n1},\cdots,V_{nm_n}) =$$
$$= D(V_1,\cdots,V_n)+\textstyle\sum_{k=1}^{s}\sum_{i=1}^{n} u_k^i D\left(\dfrac{u_{1k}^i/J(\sum\limits_{k=1}^{s}u_{1k}^i)}{u_k^i/J(\sum\limits_{k=1}^{s}u_k^i)},\cdots,\dfrac{u_{m_ik}/J(\sum\limits_{k=1}^{s}u_{m_ik}^i)}{u_k^i/J(\sum\limits_{k=1}^{s}u_k^i)}\right)$$

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

Proof. Using Condition 5, we have

$$D(V_{11},\ \cdots,V_{1m_1};\cdots;V_{n1},\cdots,V_{nm_n}) = D(V_{11},\cdots,V_{1m_1};\cdots;V_{(n-1)1},\cdots,V_{(n-1)m_{(n-1)}};V_n)+$$
$$+\textstyle\sum_{k=1}^{s} u_k^n D\left(\dfrac{u_{1k}^n/J(\sum\limits_{k=1}^{s}u_{1k}^n)}{u_k^n/J(\sum\limits_{k=1}^{s}u_k^n)},\cdots,\dfrac{u_{m_nk}/J(\sum\limits_{k=1}^{s}u_{m_nk}^n)}{u_k^n/J(\sum\limits_{k=1}^{s}u_k^n)}\right).$$

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

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

Proof. It is trivial from $U_1=U_2=\cdots=U_s=1$ and Condition 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 $D(U_1,U_2)=0$ from Condition 4. Using Condition 5, we have

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

From Lemma 2 and Condition 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, it is easy to get the following lemma:

Lemma 4 *(generalized grouping lemma).*

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

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 that the entry* $u_{..}$ *or* $u_{.} \in$ *the* i*-th group.*

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

Due to the symmetric condition, $f(t) = D(0,t,0,\cdots,0) = D(0,0,t,\cdots,0) = \cdots = D(0,0,\cdots,0,t)$.

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 condition is still valid for this case; one gets

$$\begin{aligned} &D(V_1,\cdots,V_t) = \\ &= D(V_1,\cdots,V_t) + \textstyle\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik} D\left(\dfrac{u_{ik}/\sum\limits_{k=1}^{s} u_{ik}}{u_{ik}/\sum\limits_{k=1}^{s} u_{ik}}\right) \\ &= D(V_1,\cdots,V_t) + \textstyle\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik} D(1) \\ &= D(V_1,\cdots,V_t) + s\cdot f(1). \end{aligned}$$

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 a *typical discrepancy group with parameter* a denoted by $G(\mathcal{S},\mathcal{S})^a$ if it is with 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* $D(U_1,U_2,\cdots,U_s)$ *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 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

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

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 all the same to $n = 1$.

Suppose $m \geq 2$ and $n \geq 2$. Given a typical discrepancy group $G(\mathcal{S}, \mathcal{S})^1$ where $s = t = n \cdot m$, 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

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

where $u_k^i = u_{1k}^i + \cdots + u_{mk}^i$. In the last equation, there is at most one non-zero entry 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

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

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

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

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

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 entry 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 to

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

We can also obtain

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

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

$$s \cdot f(s) = s \cdot f(n) + s \cdot f(m), \quad \text{i.e.,} \quad f(m \cdot n) = f(n) + f(m). \tag{4}$$

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

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

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 $t = 1$, $1 \cdot f(1) \leq (1+1) \cdot f(1+1) = 2 \cdot D(2, 0)$ from Condition 1 and Lemma 5.

Suppose $t \geq 2$. Given a typical discrepancy group $G(t+1, t+1)$[1], we first merge the first t rows into one and get

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

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

$$(t+1) \cdot f(t+1) = t \cdot D(\frac{t+1}{t}, 0) + D(0, t+1) + t \cdot f(t).$$

It means

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

because the first two terms of the right side of the last equation are greater or equal to zero from Condition 1. Finally, the conclusion is obtained from Lemma 8 and 7 immediately.
Lemma 10.

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

Proof.

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

Assume p is a rational number, $p = \frac{m}{s}$ and $0 < m < s$, where m and s are integers. Given a typical discrepancy group $G(\mathcal{S},\mathcal{S})^1$, we get

$$D(\mathcal{S},\mathcal{S})^1 = s \cdot D(\frac{1/1}{1/s}, 0, \cdots, 0) = s \cdot D(s, 0, \cdots, 0) = s \cdot f(s) \tag{5}$$

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

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

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

$$\begin{aligned} D(\mathcal{S},\mathcal{S})^1 \; & = D(1,\cdots,1) + m \cdot D(\tfrac{1/m}{1/s}, 0) + (s-m) D(\tfrac{1/(s-m)}{1/s}, 0) + \\ & + m \cdot D(m,0,0,\cdots,0) + (s-m) D(s-m,0,0,\cdots,0) \\ & = 0 + m \cdot D(\tfrac{s}{m}, 0) + (s-m) D(\tfrac{s}{(s-m)}, 0) + m f(m) + (s-m) f(s-m). \end{aligned} \tag{6}$$

From (5), (6) and Lemma 9, one gets

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

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

$$\begin{aligned} p \cdot D(\tfrac{1}{p}, 0) + (1-p) D(\tfrac{1}{1-p}, 0) \; & = c \cdot (\log s - \tfrac{m}{s} \log m - \tfrac{(s-m)}{s} \log(s-m)) \\ & = c \cdot (\log \tfrac{s}{(s-m)} + \tfrac{m}{s} \log \tfrac{(s-m)}{s}) \\ & = c \cdot (\log \tfrac{1}{1-p} + p \log \tfrac{1-p}{p}) \\ & = c \cdot (-\log(1-p) + p \log(1-p) - p \log p) \\ & = c \cdot (p \log \tfrac{1}{p} + (1-p) \log \tfrac{1}{(1-p)}). \end{aligned} \tag{7}$$

Because p is an arbitrary rational number less than 1 and greater than zero, the elementary and unique solution of (7) is

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

This result can be extended to all irrational $r > 1$ by the continuity conditions immediately.

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

$$D(r,0) = c\cdot\log r \quad \text{for} \quad r \geq 1. \tag{8}$$

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

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

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

merging the bottom $s-1$ rows of $D(\mathcal{S}, 2*(\mathcal{S}-1))$ by Lemma 4 and (8), one gets

$$\begin{aligned} &D(\mathcal{S}, 2*(\mathcal{S}-1)) = D\begin{pmatrix} \overbrace{p\ ,p\ ,\cdots\ ,p\ ,}^{s-1} & \overbrace{0\ ,0\ ,\cdots\ ,0}^{s-1} \\ 0\ ,0\ ,\cdots\ ,0\ , & q\ ,q\ ,\cdots\ ,q \end{pmatrix} + q(s-1)D(s-1,0,0,\cdots,0) \\ &= D(p\ ,\cdots\ ,p\ ,q\ ,\cdots\ ,q\ ) + p(s-1)D(\tfrac{1}{p},0) + q(s-1)D(\tfrac{1}{q},0) + cq(s-1)\log(s-1) \\ &= D(p\ ,\cdots\ ,p\ ,q\ ,\cdots\ ,q\ ) + cp(s-1)\log(\tfrac{1}{p}) + cq(s-1)\log(\tfrac{1}{q}) + cq(s-1)\log(s-1). \end{aligned} \tag{9}$$

We can also merge all s rows of $D(\mathcal{S}, 2*(\mathcal{S}-1))$ by Lemma 4 directly and get

$$\begin{aligned} D(\mathcal{S}, 2*(\mathcal{S}-1)) \quad &= D(p\ ,\cdots\ ,p\ ,q\ ,\cdots\ ,q\ ) + p(s-1)D(\overbrace{\frac{1}{p},0,\cdots,0}^{s}) + \\ &+q(s-1)D(\overbrace{\frac{s-1}{q},0,\cdots,0}^{s}). \end{aligned} \tag{10}$$

Combining (9) and (10), we obtain

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

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

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

Similarly, we can extend it to the case of $r = 1$ due to the continuous condition.
Lemma 11.

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

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,

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

From Condition 2, it is easy to know

$$D_1(\mathcal{R} * \mathcal{S}, \mathcal{S}) = D_2(\mathcal{R} * \mathcal{S}, \mathcal{S}). \tag{11}$$

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

$$
\begin{aligned}
D_1(\mathcal{R}*\mathcal{S},\mathcal{S}) &= D\begin{pmatrix} u_{11},u_{12},\cdots,u_{1s} \\ u_{21},u_{22},\cdots,u_{2s} \\ \vdots \quad \ddots \quad \vdots \\ u_{r1},u_{r2},\cdots,u_{rs} \end{pmatrix} + u_{11}D(\frac{u_{11}/u_{11}}{u_{11}/\sum_{k=1}^{s}u_{1k}},0,\cdots,0)+ \\
&+u_{12}D(0,\frac{\sum_{k=1}^{s}u_{1k}}{u_{12}},0,\cdots,0)+\cdots+u_{rs}D(0,0,\cdots,0,\frac{\sum_{k=1}^{s}u_{rk}}{u_{rk}}) \\
&= D(u_1,u_2,\cdots,u_s)+\sum_{k=1}^{s}u_kD\left(\frac{u_{1k}/\sum\limits_{k=1}^{s}u_{1k}}{u_k/\sum\limits_{k=1}^{s}u_k},\cdots,\frac{u_{rk}/\sum\limits_{k=1}^{s}u_{rk}}{u_k/\sum\limits_{k=1}^{s}u_k}\right) \\
&+c\cdot\sum_{k=1}^{s}\sum_{i=1}^{r}u_{ik}\cdot\log\frac{\sum_{k=1}^{s}u_{ik}}{u_{ik}}
\end{aligned}
\tag{12}
$$

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

$$
\begin{aligned}
D_2(\mathcal{R}*\mathcal{S},\mathcal{S}) &= D\begin{pmatrix} u_1,\ 0,\cdots,\ 0 \\ 0,u_2,\cdots,\ 0 \\ \vdots \quad \ddots \quad \vdots \\ 0,\ 0,\ \cdots,\ u_s \end{pmatrix} + s\cdot D(1,1,\cdots,1)+r(s-1)\cdot 0\cdot D(0,0,\cdots,0) \\
&= D(u_1,u_2,\cdots,u_s)+\sum_{k=1}^{s}u_kD(\frac{\sum_{k=1}^{s}u_k}{u_k},0,\cdots,0) \\
&= D(u_1,u_2,\cdots,u_s)+c\cdot\sum_{k=1}^{s}u_k\cdot\log\frac{\sum_{k=1}^{s}u_k}{u_k}
\end{aligned}
\tag{13}
$$

according to Lemma 4 and Lemma 10.
Combining (11), (12) and (13), we have

$$
\begin{aligned}
&\sum_{k=1}^{s}u_kD\left(\frac{u_{1k}/\sum\limits_{k=1}^{s}u_{1k}}{u_k/\sum\limits_{k=1}^{s}u_k},\cdots,\frac{u_{rk}/\sum\limits_{k=1}^{s}u_{rk}}{u_k/\sum\limits_{k=1}^{s}u_k}\right) \\
&= c\cdot\sum_{k=1}^{s}u_k\cdot\log\frac{\sum_{k=1}^{s}u_k}{u_k}-c\cdot\sum_{k=1}^{s}\sum_{i=1}^{r}u_{ik}\cdot\log\frac{\sum_{k=1}^{s}u_{ik}}{u_{ik}} \\
&= c\cdot\sum_{k=1}^{s}\left[(\sum_{i=1}^{r}u_{ik})\cdot\log\frac{\sum_{k=1}^{s}u_k}{u_k}-\sum_{i=1}^{r}u_{ik}\cdot\log\frac{\sum_{k=1}^{s}u_{ik}}{u_{ik}}\right] \\
&= c\cdot\sum_{k=1}^{s}\sum_{i=1}^{r}u_{ik}\log\frac{u_{ik}/\sum_{k=1}^{s}u_{ik}}{u_k/\sum_{k=1}^{s}u_k}.
\end{aligned}
\tag{14}
$$

Definition 3. *A function $B(U_1,U_2,\cdots,U_s):\Gamma_t\to R$ is called a FDOD function* if

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

where $K_i = \sum_{k=1}^{s} u_{ik}/s$ and $0 \cdot \log 0 = 0$ are defined (cf.Corollary 1 in FW 1994).
Theorem 1. *If $D(U_1, U_2, \cdots, U_s) : \Gamma_t \to R$ is non-negative, symmetric, continuous, identical, recursive and expansible, then $D(U_1, U_2, \cdots, U_s)$ is the FDOD function with a non-negative constant, i.e.,*

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

for all

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

Proof.

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

$$\begin{aligned} D(U_1, U_2, \cdots, U_s) &= D \begin{pmatrix} u_{11}, u_{12}, \cdots, u_{1s} \\ u_{21}, u_{22}, \cdots, u_{2s} \\ \vdots \quad \ddots \quad \vdots \\ u_{t1}, u_{t2}, \cdots, u_{ts} \end{pmatrix} = \\ &= \textstyle\sum_{k=1}^{s} u_k D \left(\dfrac{u_{1k}/\sum\limits_{k=1}^{s} u_{1k}}{u_k/\sum\limits_{k=1}^{s} u_k}, \cdots, \dfrac{u_{tk}/\sum\limits_{k=1}^{s} u_{tk}}{u_k/\sum\limits_{k=1}^{s} u_k} \right) + D(1, 1, \cdots, 1) \\ &= c \cdot \textstyle\sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log \frac{u_{ik}/\sum_{k=1}^{s} u_{ik}}{u_k/\sum_{k=1}^{s} u_k} + 0 \\ &= c \cdot \textstyle\sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log(u_{ik}/K_i) = c \cdot B(U_1, U_2, \cdots, U_s) \end{aligned} \tag{15}$$

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

4 The normalized condition and the characterization of information discrepancy

In order to obtain a unique solution, a normalized condition, $I(1/2, 1/2) = \log 2$, was introduced in the study of Shannon entropy. In the case of a group of distributions, we can also add an appropriate normalized condition to get a unique solution.

Because

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

from the identical condition, so we can't use the normalized condition of Shannon entropy directly 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 just the maximum $\log 2$ of $H(p)$ $p \in \Gamma_2$, 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 condition as follows.

Condition 7 *(normalized).*

$$D\begin{pmatrix} 1, & 0 \\ 0, & 1 \end{pmatrix} = 2 \cdot \log 2. \tag{16}$$

Lemma 12 *(FW,1994).*

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

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* $\sum_{i=1}^{n} x_i = a$*,* $\sum_{i=1}^{n} y_i = a$ *where* $x_i \geq 0$ *and* $y_i > 0$*, or* $x_i = 0$ *and* $y_i = 0$*.*

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

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

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 Theorem 1 and Condition 7, it is easy to get $c = 1$ and

$$D(U_1, U_2, \cdots, U_s) = \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log(u_{ik}/K_i) = B(U_1, U_2, \cdots, U_s). \tag{17}$$

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

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 condition.

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

Let $u_k^1 = u_{1k} + u_{2k} + \cdots + u_{rk} = \sum_{i=1}^{r} u_{ik}$ and $u_k^2 = u_{(r+1)k} + u_{(r+2)k} + \cdots + u_{tk} = \sum_{i=r+1}^{t} u_{ik}$, then $(u_k^1 + u_k^2) = 1$. Directly using (17), 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

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

according to Lemma 11. Thus, these formulae lead to

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

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

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

This means $B(V_1,V_2,\cdots,V_t)$ satisfies the recursive condition.

5 The other properties of FDOD function

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

First of all, the FDOD function satisfies 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

$$J_t(V_1, V_2) =$$

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

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*

$$B(V_1, V_2, \cdots, V_t) = \sum_{i=1}^{t} G(V_i) \tag{18}$$

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

(18) 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.

Let

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

denote the information discrepancy of a distribution U_k relative to the remainder. It is easy to obtain the two following 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.

Properties concerned with 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 $u_{1k}, u_{2k}, \cdots, u_{tk}(k \in K)$.

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 *a group of distributions 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*

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, when $s > t$, 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 or equal to 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 results:

Invariance for its average (FW,1994).

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

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

Similarity to some measures of difference in statistics.

$B(U_1, U_2, \cdots, U_s)$ is given by

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

where $x_a = - \sum_{i=1}^{t} u_{ia} \log u_{ia}$ is the Shannon entropy of the average of all distributions and $x_k = - \sum_{i=1}^{t} u_{ik} \log u_{ik}$ is the Shannon entropy of a distribution. Combining the last formula with the non-negativity, identity monotonicity, and invariance for averages of the FDOD function, we feel that this function is similar to some difference formulae in statistics. For example, given a statistical sample, $\{y_i\}$ is a set of n observations, $\overline{y}$ is the mean of these observations, and $(y_i - \overline{y})$ is the deviation of a observation from the mean of all observations. Thus, $(x_a - x_k)$ is somewhat similar to the deviation $(y_i - \overline{y})$, $\sum_{k=1}^{s}(x_a - x_k)$ 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 averages (fortunately, the sum of these "deviations" $(x_a - x_k)$ is not always equal to zero, so we need not work with the sum of their squares).

Thus, it seems that the FDOD function is not only connected with the Shannon information, but also with some differences in statistics, so we can consider it as a measure of information discrepancy contained in multiple 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
Absolute continuity	Yes	No	Yes
Maximum	$\ln t$	∞	$s \cdot \ln s \ \ (s \cdot \ln t)$
Convexity or concavity w.r.t $\mathbf{x} \in \mathbf{\Gamma}$	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
Sense for mutually singular distributions	No	Yes (if $q_i = 0$)	No

7 Some possible applications

A measure of information discrepancy of multiple distributions can be applied in many areas, we describe some possible applications here. At first, we need a principle on average.

The following principle can be obtained from the concavity of Shannon entropy directly; however, to show some kind of relationship between Shannon entropy and the FDOD function, we prefer using the non-negative property of the FDOD function to prove it.

Principle on average effect.

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

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

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

Proof. From

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

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

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

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

According to 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 can 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.

1. On an entropy model in statistical mechanics

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). A lot of confusions in statistical mechanics may be concerned with 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 about the FDOD function together may be able to lead to a new approach to investigate the concept of system entropy and irreversible processes; moreover, the connection between the new way and the classical theory of thermodynamic equilibrium will be smooth. 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 is as follows. (This model is constructed not by the traditional 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, $\epsilon_1, \epsilon_2, \cdots, \epsilon_t$ at some time. Suppose u_{ik} is the probability that the particle k in the system has energy ϵ_i, then the state entropy of a particle k is defined as $-c_0 \sum_{i=1}^{t} u_{ik} \log u_{ik}$, and the entropy of the system is 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.

For this model, we remain a simple and natural postulate, which was similarly described by Hecht (1990):

Assumption A. 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 try to find its maximum in a linear constraint and prove that this value is exactly equal to that of the classical equilibrium theory.

Suppose that the expected energy of all particles in the system is E, then the constraint can be written by

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

The Lagrangian is

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

Letting the derivative of the Lagrangian with respect to u_{ik} equal 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, we get

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

i.e.,

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

Letting $\mu = \mu_0/c_0$, 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 Assumption A; moreover, the entropy GS of the system at this state is just maximum according to the principle on average effect; (22) 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 $\mu = \frac{1}{k_0 T}$ (T is the temperature of an equilibrium system and k_0 is the Boltzmann constant) and

$$S(\text{entropy}) = k_0 \log \sum_{i=1}^{t} e^{-\epsilon_i/k_0 T} + \frac{1}{T} \frac{\sum_{i=1}^{t} \epsilon_i e^{-\epsilon_i/k_0 T}}{\sum_{i=1}^{t} e^{-\epsilon_i/k_0 T}}. \tag{24}$$

Substitute (22) into (21), then (21) becomes (23); thus, we can take $\mu = \frac{1}{k_0 T}$ in (22); further, substituting (22) into GS and taking $c_0 = k_0/s$, we have

$$\begin{aligned} GS &= -c_0 \textstyle\sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik}^* \log u_{ik}^* = -c_0 s \sum_{i=1}^{t} \frac{e^{-\mu\epsilon_i}}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} \log \frac{e^{-\mu\epsilon_i}}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} \\ &= -c_0 s \Big[\tfrac{1}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} [\textstyle\sum_{i=1}^{t} e^{-\mu\epsilon_i} (-\mu\epsilon_i - \log \sum_{i=1}^{t} \epsilon^{-\mu\epsilon_i})] \Big] \\ &= c_0 s \Big[\tfrac{1}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} [\mu \textstyle\sum_{i=1}^{t} \epsilon_i e^{-\mu\epsilon_i} + (\sum_{i=1}^{t} e^{-\mu\epsilon_i})(\log \sum_{i=1}^{t} \epsilon^{-\mu\epsilon_i})] \Big] \\ &= c_0 s \Big[\mu \tfrac{\sum_{i=1}^{t} \epsilon_i e^{-\mu\epsilon_i}}{\sum_{i=1}^{t} e^{-\mu\epsilon_i}} + \log \textstyle\sum_{i=1}^{t} e^{-\mu\epsilon_i} \Big] \\ &= k_0 \log \textstyle\sum_{i=1}^{t} e^{-\epsilon_i/k_0 T} + \frac{1}{T} \frac{\sum_{i=1}^{t} \epsilon_i e^{-\epsilon_i/k_0 T}}{\sum_{i=1}^{t} e^{-\epsilon_i/k_0 T}}. \end{aligned} \tag{25}$$

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

$$GS = -\frac{k_0}{s} \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log u_{ik} \tag{26}$$

where k_0 is Boltzmann constant and u_{ik} is the probability that the particle k in the system has energy ϵ_i; moreover, (26) can be reduced to $GS = -k_0 \sum_{i=1}^{t} u_i \log u_i$ when a system is at equilibrium state.

Secondly, the entropy GS of the system will increase with increase of time when this system is at an irreversible process. Let

$$B(\tau) = \frac{k_0}{s} \sum_{k=1}^{s} \sum_{i=1}^{t} u_{ik} \log \frac{u_{ik}}{K_i} \tag{27}$$

denote the discrepancy of distributions of all particles at some time τ.

Using the $B(\tau)$ 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 required for a dynamical system as usual. Since probability distributions of all particles over the possible energy levels will tend to the same with τ (time) $\to \infty$ (i.e., the system tends to an equilibrium state with τ (time) $\to \infty$), we assume

Assumption B. For a fixed circumstance and energy E, the discrepancy of probability distributions among all particles will decrease with increase of time, i.e., if $\tau_2 \geq \tau_1$ and

$B(\tau)$ is adopted as a measure of discrepancy among distributions all particles at the time τ, then $B(\tau_2) \leq B(\tau_1)$.

Due to that there is a great number of particles, the probability value of a particle over an energy level at some time τ can be considered as a random shift around the average of probability values of all particles over the same energy level at the same time τ; moreover, the probability distributions of all particles will tend to be same. Thus, we present

Assumption C. For a fixed circumstances and energy E, the average of probability values of all particles over an energy level will be constant for all time.

Assumption C is somewhat similar to the density assumption of the canonical ensemble way 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); however, Assumption C is more natural.

Thus, from Assumption B, (17), and (26), we have

$$\begin{aligned} B(\tau_2) \quad & -B(\tau_1) \\ & = \tfrac{k_0}{s}\left(\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}^2 \log \tfrac{u_{ik}^2}{K_i^2} - \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}^1 \log \tfrac{u_{ik}^1}{K_i^1}\right) \\ & = \tfrac{k_0}{s}\left(\sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}^2 \log u_{ik}^2 - \sum_{k=1}^{s}\sum_{i=1}^{t} u_{ik}^1 \log u_{ik}^1 - \sum_{i=1}^{t}\left(\sum_{k=1}^{s} u_{ik}^2\right)(\log K_i^2) + \right. \\ & \quad \left. + \sum_{i=1}^{t}\left(\sum_{k=1}^{s} u_{ik}^1\right)(\log K_i^1)\right) \\ & = GS^1 - GS^2 - \tfrac{k_0}{s}\left(\sum_{i=1}^{t}\left(\sum_{k=1}^{s} u_{ik}^2\right)(\log K_i^2) - \sum_{i=1}^{t}\left(\sum_{k=1}^{s} u_{ik}^1\right)(\log K_i^1)\right) \quad \leq 0 \end{aligned} \tag{28}$$

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

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

and this means that the entropy GS of the system will increase with increase of time in an irreversible process.

In a word, this example shows a new possibility of investigating the concept of system entropy and irreversible process. The physical meaning of this model is clear and natural, and the connection between it and the classical equilibrium theory is smooth.

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

2. On weighting problems

This sort of applications is illustrated by an analysis of weighting decision criteria.

Nutt reported a field study in which direct and indirect criteria weighting methods were compared [Nutt, 1980]. The average criteria weights obtained by 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 also 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 the following:

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

These results illustrate that the discrepancy of the log-odds method relative to the reminder is the smallest one; the discrepancy of the rank-weight method is bigger, and this is also true for the 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 figure analysis; further, if we use the 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 can also obtain an overall evaluation about the original data by the procedure described in FW (1994).

3. On similarity of multiple information sequences

In molecular biology and communication, there is a basic need to compare multiple sequence information. Some algorithms, depended on a lot of paired comparison of two sequences, have been proposed(Chen, Wong and Chiu, 1992). Using the properties of the FDOD function, we have developed a new algorithms to compare multiple DNA sequences. The results are very good. It is involved in complex comparisons, so we will discuss it in another paper.

8 Conclusion

In this paper, based on the concepts of Shannon information and measure of difference, we have proposed the condition set for measuring the information discrepancy contained in multiple distributions, and proved that the FDOD function is the unique solution satisfying the condition set. The final and intermediate results indeed show the closer connection between the FDOD function and Shannon information as well as the measures of difference, and it may be able to be applied in various areas.

Acknowledgments

This work has been supported by DIMACS Center, Chinese NSF under Grant 69772036, 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] S.C. Chan, A.K. Wong & , D.K. Chiu, A Survey of Multiple Sequence Comparison Methods. Bulu. Math. Biol. 54, 563-598,1992.

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

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

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

[8] Weiwu Fang, On a Global Optimization Problem in the Study of Information Discrepancy, J. of Global Optimization Vol.11, 387-408, 1997.

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

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

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

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

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

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

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

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

[17] P.M. Lee, On the Conditions of Information Theory. Ann. Math. Statist. 35, 414-441,1994.

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

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

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

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

[22] Erwin Schrödinger, Statistical Thermodynamics, Dover Pub., Inc., New York, 1952 or 1989.

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