Information theory based on nonadditive information content

Takuya Yamano
Department of Applied Physics, Faculty of Science, Tokyo Institute of Technology, Oh-okayama, Meguro-ku, Tokyo, 152-8551, Japan
(Received 6 October 2000; published 23 March 2001)

We generalize Shannon's information theory in a nonadditive way by focusing on the source coding theorem. The nonadditive information content we adopted is consistent with the concept of the form invariance structure of the nonextensive entropy. Some general properties of the nonadditive information entropy are studied, in addition, the relation between the nonadditivity q and the codeword length is pointed out.

DOI: 10.1103/PhysRevE.63.046105 PACS number(s): 02.50.−r, 05.20.−y, 89.70.+c

I. INTRODUCTION

The intuitive notion of what a quantitative expression for information should be has been addressed in the development of transmission of information, which led to the information theory (IT). The IT today is considered to be the most fundamental field that connects other various fields such as physics (thermodynamics), electrical engineering (communication theory), mathematics (probability theory and statistics), computer science (Kolmogorov complexity) and so on [1]. Accordingly, the selection of the information measure becomes an influential matter. The introduction of logarithmic form of information measure dates back to Hartley [2]. He defined the practical measure of information as the logarithm of the number of possible symbol sequences. After that, Shannon established the logarithmic based IT from the reasons: (a) practical usefulness, (b) closeness to our intuitive feeling, and (c) easiness of mathematical manipulation [3,4].

On the other hand, however, nonlogarithmic form of (or nonextensive) entropy is currently considered as a useful measure in describing thermostatistical properties of a certain class of physical systems, which entail long-range interactions, long-time memories, and (multi) fractal structures. The form of the nonextensive entropy proposed by Tsallis [5] has been intensively applied to many such systems [6]. The reason why the formalism violating the extensivity of the statistical-mechanical entropy seems to be essential for convincing description of these systems is not sufficiently revealed in the present status. Nevertheless the successful application to some physical systems seems to lead us to investigate into the possibility of the nonadditive IT since Shannon's information entropy has the same form as the logarithmic statistical-mechanical entropy.

It is desirable to employ the nonadditive information content that the associated IT contains Shannon's IT in a special case. The concept of the form invariance to the structure of nonextensive entropies was considered to provide a guiding principle for a clear basis for generalizations of logarithmic entropy [7]. This structure seems to give a hint at the selection of the nonadditive information content. The form invariant structure requires normalization of the original Tsallis entropy by $\Sigma_i p_i^q$, where p_i is a probability of event i and q is a real number residing in the interval (0,1) from the preservation of concavity of the entropy. In addition, the Kullback-Leibler (KL) relative entropy, which measures distance between two probability distributions, is also modified [7].

This paper explores consequences of adopting nonadditive information content in the sense that the associated average information, i.e., entropy takes a form of the modified Tsallis entropy. The use of modified form of Tsallis's entropy is in conformity with the appropriate definitions of expectation value (the normalized q-expectation value [8]) of the nonadditive information content. Since the information theoretical entropy is defined as the *average* of information content, it is desirable to unify the meaning of the *average* as the normalized q-expectation value throughout the nonadditive IT. Moreover we shall later see how Shannon's additive IT is extended to the nonadditive one by addressing the source coding theorem, which is one of the most fundamental theorems in IT.

The organization of this paper is as follows. In Sec. II, we present the mathematical preliminaries of the nonadditive entropy and the generalized KL entropy. Section III addresses an optimal code word within the framework of nonadditive context. We shall attempt to give a possible meaning of nonadditive index q in terms of codeword length there. Section IV deals with the extension of Fano's inequality, which gives upper bound on the conditional entropy with an error probability in a channel. We devote the last section to concluding remarks.

II. NONADDITIVE ENTROPY AND THE GENERALIZED KL ENTROPY

A. Properties of nonadditive entropy of information

For a discrete set of states with probability mass function $p(x)$, where x belongs to alphabet $\mathcal{H}$, we consider the following nonadditive information content $I_q(p)$,

$$I_q(p) \equiv -\ln_q p(x), \tag{1}$$

where $\ln_q x$ is a q-logarithm function defined as $\ln_q x = (x^{1-q}-1)/(1-q)$. In the limit $q \to 1$, $\ln_q x$ recovers the standard logarithm $\ln x$. In Shannon's additive IT, an information content is expressed as $-\ln p(x)$ in NAT unit, which is monotonically decreasing function with respect to $p(x)$. This behavior matches our intuition in that we get more information in the case the least-probable event occurs and less information in the case the event with high probability occurs. It is worth noting that this property is qualitatively valid for nonadditive information content for all q except the fact that there exists upper limit $1/(1-q)$ at $p(x)=0$. Therefore the Shannon reason (b) we referred to in Sec. I is considered to be no crucial element for determining the logarithmic form. Moreover, it is easy to see that the Renyi information of order q [9], $H_q^R=\ln \Sigma_{x\in\mathcal{H}} p^q(x)/(1-q)$, which is an additive information measure, can be written with this nonadditive information content as

$$H_q^R=\frac{\ln \sum_{x\in\mathcal{H}} \{1-(1-q)I_q[p(x)]\}^{q/(1-q)}}{1-q}. \tag{2}$$

The entropy $H_q(X)$ of a discrete random variable X is defined as an average of the information content, where the *average* means the normalized q-expectation value [8],

$$H_q(X)=\frac{-\sum_{x\in\mathcal{H}} p^q(x)\ln_q p(x)}{\sum_{x\in\mathcal{H}} p^q(x)}=\frac{1-\sum_{x\in\mathcal{H}} p^q(x)}{(q-1)\sum_{x\in\mathcal{H}} p^q(x)}, \tag{3}$$

where we have used the normalization of probability $\Sigma_{x\in\mathcal{H}} p(x)=1$. In a similar way, we define the nonadditive conditional information content and the joint one as follows:

$$I_q(y|x)=\frac{p^{1-q}(y|x)-1}{q-1}, \tag{4}$$

$$I_q(x,y)=\frac{p^{1-q}(x,y)-1}{q-1}, \tag{5}$$

where y belongs to a different alphabet $\mathcal{Y}$. Corresponding entropy conditioned by x and the joint entropy of X and Y becomes

$$H_q(Y|x)=\frac{\sum_{y\in\mathcal{Y}} p^q(y|x)I_q(y|x)}{\sum_{y\in\mathcal{Y}} p^q(y|x)}=\frac{1-\sum_{y\in\mathcal{Y}} p^q(y|x)}{(q-1)\sum_{y\in\mathcal{Y}} p^q(y|x)} \tag{6}$$

and

$$H_q(X,Y)=\frac{\sum_{x\in\mathcal{H},y\in\mathcal{Y}} p^q(x,y)I_q(x,y)}{\sum_{x\in\mathcal{H},y\in\mathcal{Y}} p^q(x,y)}=\frac{1-\sum_{x\in\mathcal{H},y\in\mathcal{Y}} p^q(x,y)}{(q-1)\sum_{x\in\mathcal{H},y\in\mathcal{Y}} p^q(x,y)}, \tag{7}$$

respectively. Then we have the following theorem.

Theorem 1: The joint entropy satisfies

$$H_q(X,Y)=H_q(X)+H_q(Y|X)+(q-1)H_q(X)H_q(Y|X). \tag{8}$$

Proof: From Eq. (7) we can rewrite $H_q(X,Y)$ with the relation $p(x,y)=p(x)p(y|x)$ as

$$H_q(X,Y)=\frac{1}{q-1}\left[\frac{1}{\sum_{x\in\mathcal{H}} p^q(x)\sum_{y\in\mathcal{Y}} p^q(y|x)}-1\right]. \tag{9}$$

Since Eq. (6) gives $\Sigma_{y\in\mathcal{Y}} p^q(y|x)=[1+(q-1)H_q(Y|x)]^{-1}$, we get

$$H_q(X,Y)=\frac{1}{q-1}\left[\frac{1}{\sum_{x\in\mathcal{H}} p^q(x)/[1+(q-1)H_q(Y|x)]}-1\right]. \tag{10}$$

Here, we introduce the following definition [10]

$$\left\langle \frac{1}{1+(q-1)H_q(Y|x)}\right\rangle_q^{(X)}\equiv\frac{1}{1+(q-1)H_q(Y|X)}, \tag{11}$$

where the bracket denotes the normalized q-expectation value with respect to $p(x)$. Then we have from Eq. (10),

$$H_q(X,Y)=\frac{1}{q-1}\left[\frac{1+(q-1)H_q(Y|X)}{\sum_{x\in\mathcal{H}} p^q(x)}-1\right]. \tag{12}$$

Putting $\Sigma_{x\in\mathcal{H}} p^q(x)=[1+(q-1)H_q(X)]^{-1}$ into this yields the theorem.□

This theorem has a remarkable similarity to the relation with which the Jackson basic number in q-deformation theory satisfies, which was pointed out in Ref. [7]. That is, $[X]_q\equiv(q^X-1)/(q-1)$ is the Jackson basic number of a quantity X. Then, for the sum of two quantities X and Y, the associated basic number $[X+Y]_q$ is shown to become $[X]_q+[Y]_q+(q-1)[X]_q[Y]_q$. Obviously this theorem recovers the ordinary relation $H(X,Y)=H(X)+H(Y|X)$ in the limit $q\to1$. In this modified Tsallis formalism, q appears as $q-1$ instead of as $1-q$ [11,12]. When X and Y are independent events from each other, Eq. (8) gives the pseudoadditivity relation [8]. However, it is converse to the case of the original Tsallis one in that $q>1$ yields superadditivity and $q<1$ subadditivity. It is worth mentioning that the concept of nonextensive conditional entropy in the framework of the original Tsallis entropy was first introduced for discussing quantum entanglement in Ref. [11]. From this theorem, we immediately have the following corollary concerning the equivocation.

Corollary:

$$H_q(Y|X)=\frac{H_q(Y,Z|X)-H_q(Z|Y,X)+(q-1)\{H_q(X)H_q(Y,Z|X)-H_q(X,Y)H_q(Z|Y,X)\}}{1+(q-1)H_q(X)} \tag{13}$$

Proof: In Eq. (8), when we see Y as Y,Z, we have

$$H_q(X,Y,Z)=H_q(X)+H_q(Y,Z|X)+(q-1)H_q(X)H_q(Y,Z|X), \tag{14}$$

on the other hand, when we regard X as Y,X and Y as Z, we get

$$H_q(X,Y,Z)=H_q(X,Y)+H_q(Y,Z|X)+(q-1)H_q(X,Y)H_q(Z|Y,X). \tag{15}$$

Subtracting both sides of the above two equations and arranging with respect to $H_q(Y|X)$ with Eq. (8), we obtain the corollary. □

Remarks: In the additive limit$(q\to 1)$, we recover the relation $H(Y|X)=H(Y,Z|X)-H(Z|Y,X)$.

Moreover, with the help of Eq. (13), we have the following theorem.

Theorem 2: Hierarchical structure of entropy H_q—The joint entropy of n random variables $X_1,X_2,\ldots,X_n$ satisfies

$$H_q(X_1,X_2,\ldots,X_n)=\sum_{i=1}^{n}[1+(q-1)H_q(X_{i-1},\ldots,X_1)]\times H_q(X_i|X_{i-1},\ldots,X_1). \tag{16}$$

Proof: From Eq. (8), $H_q(X_1,X_2)=H_q(X_1)+[1+(q-1)H_q(X_1)]H_q(X_2|X_1)$ holds. Next, from Eq. (14), we have

$$H_q(X_1,X_2,X_3)=H_q(X_1)+H_q(X_2,X_3|X_1)+(q-1)H_q(X_1)H_q(X_2,X_3|X_1). \tag{17}$$

Since $H_q(X_2,X_3|X_1)$ is written using Eq. (13) as

$$H_q(X_2,X_3|X_1)=H_q(X_2|X_1)+\frac{1+(q-1)H_q(X_1,X_2)}{1+(q-1)H_q(X_1)}H_q(X_3|X_2,X_1), \tag{18}$$

Eq. (17) can be rewritten as

$$H_q(X_1,X_2,X_3)=H_q(X_1)+[1+(q-1)H_q(X_1)]H_q(X_2|X_1)+[1+(q-1)H_q(X_1,X_2)]H_q(X_3|X_2,X_1). \tag{19}$$

Similarly, repeating application of the corollary gives the theorem. □

Remark: In the additive limit, we get $H(X_1,X_2,\ldots,X_n)=\sum_{i=1}^{n}H(X_i|X_{i-1},\ldots,X_1)$ which states that the entropy of n variables is constituted by the sum of the conditional entropies (*chain rule*).

From this relation Eq. (16), we need all joint entropy below the level of n random variables to acquire the joint entropy $H_q(X_1,\ldots,X_n)$; this situation is similar to the Bogoliubov-Born-Green-Kirkwood-Yvon hierarchy in the N-body distribution function. Let us next define the mutual information $\mathcal{I}_q(Y;X)$, which quantifies the amount of information that can be gained from one event X about another event Y,

$$\mathcal{I}_q(Y;X)\equiv H_q(Y)-H_q(Y|X)=\frac{H_q(X)+H_q(Y)-H_q(X,Y)+(q-1)H_q(X)H_q(Y)}{1+(q-1)H_q(X)}. \tag{20}$$

Therefore $\mathcal{I}_q(Y;X)$ expresses the reduction in the uncertainty of Y due to the acquisition of knowledge of X. Here we postulate that the mutual information in nonadditive case is non-negative. The non-negativity may be considered to be a requirement rather than the one to be proved in order to be consistent with the usual additive mutual information. $\mathcal{I}_q(Y;X)$ also converges to the usual mutual information $\mathcal{I}(Y;X)=H(Y)-H(Y|X)=H(X)+H(Y)-H(X,Y)$ in the additive case $(q\to 1)$. We note that the mutual information of a random variable with itself is the entropy itself $\mathcal{I}_q(X;X)=H_q(X)$. When X and Y are independent variables, we have $\mathcal{I}_q(Y;X)=0$ [13]. Then, we have the following theorem.

Theorem 3: Independence bound on entropy H_q

$$H_q(X_1,X_2,\ldots,X_n)\leqslant\sum_{i=1}^{n}[1+(q-1)H_q(X_{i-1},\ldots,X_1)]H_q(X_i) \tag{21}$$

with equality if and only if each X_i is independent.

Proof: From the assumption of $\mathcal{I}_q(X;Y)\geqslant 0$ introduced above, we have

$$\sum_{i=1}^{n}H_q(X_i|X_{i-1},\ldots,X_1)\leqslant\sum_{i=1}^{n}H_q(X_i) \tag{22}$$

with equality, if and only if, each X_i is independent of $X_{i-1},\ldots,X_1$. Then the theorem holds from the previous theorem Eq. (16). □

B. The generalized KL entropy

The KL entropy or the relative entropy is a measure of the distance between two probability distributions $p_i(x)$ and $p_i'(x)$. Here, we define it as the normalized q-expectation value of the change of the nonadditive information content $\Delta I_q \equiv I_q[p'(x)] - I_q[p(x)]$ [14],

$$D_q[p(x)\|p'(x)] \equiv \frac{\sum_{x\in\mathcal{H}} p^q(x)\Delta I_q}{\sum_{x\in\mathcal{H}} p^q(x)} = \frac{\sum_{x\in\mathcal{H}} p^q(x)[\ln_q p(x) - \ln_q p'(x)]}{\sum_{x\in\mathcal{H}} p^q(x)}. \tag{23}$$

We note that the above generalized KL entropy satisfies the form invariant structure, which has been introduced in Ref. [7,15]. We review the positivity of the generalized KL entropy in the case of $q>0$, which can be considered to be a necessary property to develop the IT.

Theorem 4: Information inequality—For $q>0$, we have

$$D_q(p(x)\|p'(x)) \geqslant 0 \tag{24}$$

with equality, if and only if, $p(x)=p'(x)$ for all $x\in\mathcal{H}$.

Proof: The outline of the proof is the same as the one in Refs. [16,17] except for the factor $\Sigma_{x\in\mathcal{H}} p^q(x)$. From the definition Eq. (23),

$$\begin{aligned} D_q[p(x)\|p'(x)] &= \frac{1}{1-q}\sum_{x\in\mathcal{H}} p(x) \times \left\{1-\left(\frac{p'(x)}{p(x)}\right)^{1-q}\right\} \Big/ \sum_{x\in\mathcal{H}} p^q(x) \qquad (25)\\ &\geqslant \frac{1}{1-q}\left\{1-\left(\sum_{x\in\mathcal{H}} p(x)\frac{p'(x)}{p(x)}\right)^{1-q}\right\} \Big/ \sum_{x\in\mathcal{H}} p^q(x)\\ &=0, \qquad (26)\end{aligned}$$

where Jensen's inequality for the convex function has been used $\Sigma_x p(x) f(x) \geqslant f[\Sigma_x p(x) x]$ with $f(x)=-\ln_q(x)$, $f''(x)>0$. We have equality in the second line, if and only if, $p'(x)/p(x)=1$ for all x, accordingly $D_q[p(x)\|p'(x)]=0$. □

III. SOURCE CODING THEOREM

Having presented some properties of the nonadditive entropy and the generalized KL entropy as a preliminary, we are now in a status to address our main results that Shannon's source coding theorem can be extended to the nonadditive case. Let us consider encoding the sequence of source letters generated from an information source to the sequence of binary codewords as an example. If a code is allocated for four source letters X_1, X_2, X_3, X_4 as 0,10,110,111, respectively, the source sequence $X_2X_4X_3X_2$ is coded into 1011111010. On the other hand, if another code assigns them as 0,1,00,11, then the codeword becomes 111001. The difference between the two codes is striking in the coding. In the former codeword, the first two letters 10 corresponds to X_2 and not the beginning of any other codeword, then we can observe X_2. Next there are no codewords corresponding to 1 and 11 but 111 is, and is decoded into X_4. Then the next 110 is decoded into X_3, leaving 10, which is decoded into X_2. Therefore we can uniquely decode the codeword. In the latter case, however, we have possibilities to interpret the first three letters 111 as $X_2X_2X_2$, X_2X_4, or as X_4X_2. Namely, this code cannot be uniquely decoded into the source letter that gave rise to it. Accordingly, we need to deal with so-called the *prefix code* or the *instantaneous code* such as the former case. The prefix code is a code that no codeword is a prefix of any other codeword (prefix condition code) [1,18]. We recall that any code that satisfies the prefix condition over the alphabet size D($D=2$ is a binary case) must satisfy the Kraft inequality [1,18],

$$\sum_i^M D^{-l_i} \leqslant 1, \tag{27}$$

where l_i is a code length of ith codeword $(i=1,\ldots M)$. Moreover if a code is uniquely decodable, the Kraft inequality holds for it [1,18]. We usually hope to encode the sequence of source letters to the sequence of codewords as short as possible, that is, our problem is finding a prefix condition code with the minimum average length of a set of codewords $\{l_i\}$. The optimal code is given by minimizing the following functional constrained by the Kraft inequality,

$$J = \frac{\sum_i p_i^q l_i}{\sum_i p_i^q} + \lambda\left(\sum_i D^{-l_i}\right), \tag{28}$$

where p_i is the probability of realization of the word length l_i and λ is a Lagrange multiplier. We have assumed equality in the Kraft inequality and have neglected the integer constraint on l_i. Differentiating with respect to l_i and setting the derivative to 0, yields

$$D^{-l_i} = \frac{p_i^q}{\left(\sum_i p_i^q\right)\lambda \log D}. \tag{29}$$

Here, it is worth noting that from the Kraft inequality the Lagrange multiplier λ is related as $\lambda \geqslant (\log D)^{-1}$. Furthermore when the equality holds, the fraction $D^{-l_i^*}$, which is given by the optimal codeword length l_i^*, is expressed as

$$D^{-l_i^*}=\frac{p_i^q}{\sum_i p_i^q}. \tag{30}$$

Therefore l_i^* can be written as $\log_D(\Sigma_i p_i^q)-q\log_D p_i$ and in the additive limit, we obtain $l_i^*=-\log_D p_i$. However, we cannot always determine the optimal codeword length like this since the l_i's must be integers. We have the following theorem.

Theorem 5: The average codeword length $\langle L\rangle_q$ of any prefix code for a random variable X satisfies

$$\langle L\rangle_q \geqslant H_q(X) \tag{31}$$

with equality, if and only if, $p_i=[1-(1-q)l_i]^{1/(1-q)}$.

Proof: From Eq. (23) the generalized KL entropy between two distributions p and r is written as

$$D_q(p\|r)=\frac{\sum_i p_i^q(\ln_q p_i-\ln_q r_i)}{\sum_i p_i^q}=\frac{1-\sum_i p_i^q r_i^{1-q}}{(1-q)\sum_i p_i^q}$$
$$=\frac{1-\sum_i p_i^q}{(1-q)\sum_i p_i^q}-\frac{\sum_i p_i^q(r_i^{1-q}-1)}{(1-q)\sum_i p_i^q}. \tag{32}$$

If we take the information content associated with probability r as the ith codeword length l_i, i.e., $-(r_i^{1-q}-1)/(1-q)=l_i$, then the average codeword length can be written from Eq. (32) as

$$\langle L\rangle_q=H_q(X)+D_q(p\|r). \tag{33}$$

Since $D_q(p\|r)\geqslant 0$ for $q>0$ from Theorem 4, we have the theorem. The equality holds if and only if, $p_i=r_i$. □

We note that the relation $-(r_i^{1-q}-1)/(1-q)=l_i$ means that the codeword length l_i equals the information content different from p, $l_i=I_q(r)$. When the equality is realized in the above theorem, we can derive an interesting interpretation on the nonadditivity parameter q. The condition of the equality states that the probability is expressed as the Tsallis canonical ensemblelike factor in an i-wise manner. Then each l_i has the limit in length corresponding to q such as $l_i^{max}=1/(1-q)$.

Since $\log_D(\Sigma_i p_i^q)-q\log_D p_i$ obtained by the optimization problem is not always equal to an integer, we impose the integer condition on the codewords $\{l_i\}$ by rounding it up as $l_i=\lceil\log_D(\Sigma_i p_i^q)-q\log_D p_i\rceil$, where $\lceil x\rceil$ denotes the smallest integer $\geqslant x$ [1]. Moreover the relation $\lceil\log_D(\Sigma_i p_i^q)-q\log_D p_i\rceil \geqslant \log_D(\Sigma_i p_i^q)-q\log_D p_i$ leads to

$$\sum_i D^{-\lceil\log_D(\Sigma_i p_i^q)-q\log_D p_i\rceil}\leqslant\sum_i D^{-(\log_D(\Sigma_i p_i^q)-q\log_D p_i)}$$
$$=\sum_i \frac{p_i^q}{\sum_i p_i^q}=1. \tag{34}$$

Hence $\{l_i\}$ satisfies the Kraft inequality. Moreover, we have the following theorem.

Theorem 6: The average codeword length assigned by $l_i=\lceil\log_D(\Sigma_i p_i^q)-q\log_D p_i\rceil$ satisfies

$$H_q(p)+D_q(p\|r)\leqslant\langle L\rangle_q<H_q(p)+D_q(p\|r)+1. \tag{35}$$

Proof: The integer codeword length satisfies

$$\log_D\left(\sum_i p_i^q\right)-q\log_D p_i\leqslant l_i<\log_D\left(\sum_i p_i^q\right)-q\log_D p_i+1. \tag{36}$$

Multiplying by $p_i^q/\Sigma_i p_i^q$ and summing over i with Eq. (33) yields the theorem. □

This means that the distribution, different from the optimal one, provokes a correction of $D_q(p\|r)$ in the average codeword length as it does in the case of additive one.

We have discussed the properties of the nonadditive entropy in the case of one letter so far. Next, let us consider the situation when we transmit a sequence of n letters from the source in such a way that each letter is to be generated independently as identically distributed random variables according to $p(x)$. Then the average codeword length per letter $\langle L_n\rangle_q=\langle l(X_1,\ldots,X_n)\rangle_q/n$ is bounded as we derived in the preceding theorem,

$$H_q(X_1,\ldots,X_n)\leqslant\langle l(X_1,\ldots,X_n)\rangle_q<H_q(X_1,\ldots,X_n)+1. \tag{37}$$

Since we are now considering independently, identically distributed random variables $X_1,\ldots,X_n$, we obtain

$$\frac{\sum_{i=1}^{n}[1+(q-1)H_q(X_{i-1},\ldots,X_1)]H_q(X_i)}{n}$$
$$\leqslant\langle L_n\rangle_q<\frac{\sum_{i=1}^{n}[1+(q-1)H_q(X_{i-1},\ldots,X_1)]H_q(X_i)}{n}$$
$$+\frac{1}{n}, \tag{38}$$

where we have used Theorem 3. This relation can be considered to be the generalized source coding theorem for the finite number of letters. We note that we obtain $H(X)\leqslant\langle L_n\rangle<H(X)+1/n$ in the additive limit since $H(X_1,\ldots,X_n)=\Sigma_i H(X_i)=nH(X)$ holds.

IV. GENERALIZED FANO'S INEQUALITY

Fano's inequality is an essential ingredient to prove the converse to the channel coding theorem, which states that the probability of error that arises over a channel is bounded away from zero when the transmission rate exceeds the channel capacity [19]. In the estimation of an original message generated from the information source, the original variable X may be estimated as X' on the side of a destination. Therefore, we introduce the probability of error $P_e = Pr\{X' \neq X\}$ due to the noise of the channel through which the signal is transmitted. With an error random variable E defined as

$$E = \begin{cases} 1 & \text{if } X' \neq X \\ 0 & \text{if } X' = X, \end{cases} \tag{39}$$

we have the following theorem, which is considered to be the generalized (nonadditive version) Fano's inequality.

Theorem 7: The generalized Fano's inequality

$$H_q(X|Y) \leqslant H_q(P_e) + \frac{1+(q-1)H_q(E,Y)}{1+(q-1)H_q(Y)} \times \frac{P_e^q}{P_e^q + (1-P_e)^q} \frac{1-(|\mathcal{H}|-1)^{1-q}}{(q-1)(|\mathcal{H}|-1)^{1-q}}, \tag{40}$$

where $|\mathcal{H}|$ denotes the size of the alphabet of the information source.

Proof: The proof can be done along the line of the usual Shannon's additive case (e.g., Ref. [1]). Using the corollary Eq. (13), we have two different expressions for $H_q(E,X|Y)$,

$$H_q(E,X|Y) = H_q(X|Y) + \frac{1+(q-1)H_q(X,Y)}{1+(q-1)H_q(Y)} H_q(E|X,Y) \tag{41}$$

and

$$H_q(E,X|Y) = H_q(E|Y) + \frac{1+(q-1)H_q(E,Y)}{1+(q-1)H_q(Y)} H_q(X|E,Y), \tag{42}$$

where we have used the corollary by regarding $H_q(E,X|Y)$ as $H_q(X,E|Y)$ in the second expression Eq. (42). We are now considering the following situation. That is, we wish to know the genuine X; however, we only observe a random variable Y, which can be related to the X by the nonadditive conditional information content $I_q(y|x)$. Hence we calculate X', an estimate of X, as a function of Y such as $g(Y)$ [1]. Then we see $H_q(E|X,Y)$ becomes 0 since E is a function of X and Y by the definition Eq. (39). Therefore the first expression of $H_q(E,X|Y)$ reduces to $H_q(X|Y)$. On the other hand, from the non-negativity property of $\mathcal{I}(E;Y)$ we assumed, and from the relation $H_q(E) = H_q(P_e)$, we can evaluate $H_q(E|Y)$ as $H_q(E|Y) \leqslant H_q(E) = H_q(P_e)$. Moreover, $H_q(X|E,Y)$ can be written as

$$H_q(X|E,Y) = \frac{\sum_E (Pr\{E\})^q H_q(X|Y,E)}{\sum_E (Pr\{E\})^q} = \frac{(1-P_e)^q H_q(X|Y,0) + P_e^q H_q(X|Y,1)}{P_e^q + (1-P_e)^q}. \tag{43}$$

For $E=0$, $g(Y)$ gives X resulting in $H_q(X|Y,0)=0$ and for $E=1$, we have upper bound on $H_q(X|Y,1)$ by the maximum entropy comprised of the remaining outcomes $|\mathcal{H}|-1$,

$$H_q(X|Y,1) \leqslant \frac{1-(|\mathcal{H}|-1)^{1-q}}{(1-q)(|\mathcal{H}|-1)^{1-q}}. \tag{44}$$

Then it follows from Eq. (43) that

$$[P_e^q + (1-P_e)^q] H_q(X|E,Y) \leqslant P_e^q \frac{1-(|\mathcal{H}|-1)^{1-q}}{(1-q)(|\mathcal{H}|-1)^{1-q}}. \tag{45}$$

Combining the above results with Eq. (42), we have the theorem. □

Remark: In the additive limit, we obtain the usual Fano's inequality $H(X|Y) \leqslant H(P_e) + P_e \ln(|\mathcal{H}|-1)$ in NAT unit.

V. CONCLUDING REMARKS

We have attempted to extend Shannon's additive IT to the nonadditive case by using the nonadditive information content Eq. (1). In developing the nonadditive IT, this postulate of the nonadditive information content seems to plausible selection in terms of the unification of the meaning of *average* throughout the entire theory. As a consequence, the information entropy becomes the modified type of Tsallis's nonextensive entropy. We have shown that the properties of the nonadditive information entropy, conditional entropy, and the joint entropy in the form of theorem, which are necessary elements to develop IT. These results recover the usual Shannon ones in the additive limit ($q \to 1$). Moreover, the source coding theorem can be generalized to the nonadditive case. As we have seen in Theorem 5, the nonadditivity of the information content can be regarded as it determines the longest codeword we can transmit to the channel. The philosophy of the present attempt can be positioned as a reverse of Jaynes's pioneering work [20,21]. Jaynes has brought a concept of IT to statistical mechanics in the form of maximizing entropy of a system (Jaynes's maximum entropy principle). The information theoretical approach to statistical mechanics is now considered to be very robust in discussing some areas of physics. In turn, we have approached IT in a nonadditive way. We believe that the present consideration based on the nonadditive information content may trigger some practical future applications in such an area of the information processing.

ACKNOWLEDGMENTS

The author would like to thank Dr. S. Abe for useful comments at the Yukawa Institute for Theoretical Physics, Kyoto, Japan and for suggesting him the nonadditive conditional entropy form Eq. (11).

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[2] R. V. L. Hartley, Bell Syst. Tech. J. **7**, 535 (1928).
[3] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); *ibid.* **27**, 623 (1948).
[4] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois, Urbana, 1963).
[5] C. Tsallis, J. Stat. Phys. **52**, 479 (1988); E. M. F. Curado and C. Tsallis, J. Phys. A **24**, L69 (1991); Corrigenda **24**, 3187 (1991); **25**, 1019 (1992).
[6] http://tsallis.cat.cbpf.br/biblio.htm for an updated bibliography.
[7] A. K. Rajagopal and S. Abe, Phys. Rev. Lett. **83**, 1711 (1999).
[8] C. Tsallis, R. S. Mendes, and A. R. Plastino, Physica A **261**, 534 (1998).
[9] J. Balatoni and A. Renyi, Pub. Math. Int. Hungarian Acad Sci. **1**, 9 (1956); A. Renyi, *Probability Theory* (North-Holland, Amsterdam, 1970).
[10] The definition Eq. (11) corresponds to taking the nonadditive conditional entropy as the form $H_q(Y|X)=1/(q-1)[(\langle 1+(q-1)H_q(Y|x)\rangle_q^{(X)})^{-1}-1]$.
[11] S. Abe and A. K. Rajagopal, Physica A **289**, 157 (2001).
[12] S. Abe, Phys. Lett. A **271**, 74 (2000).
[13] From the definition, we can easily confirm $H_q(Y|X)=H_q(Y)$, if and only if, X and Y are independent variables. That is, $H_q(Y|x)=H_q(Y)$. Moreover, since $\langle [1+(q-1)H_q(Y)]^{-1}\rangle_q^{(X)}=\langle \Sigma_{y\in\mathcal{Y}} p^q(y)\rangle_q^{(X)}=\Sigma_{y\in\mathcal{Y}} p^q(y)$, we obtain $H_q(Y|X)=(1/\Sigma_{y\in\mathcal{Y}} p^q(y)-1)/(q-1)=H_q(Y)$.
[14] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951); S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).
[15] We also note that our definition of $\ln_q x$ corresponds to $-\ln_q x^{-1}$ in Ref. [7].
[16] C. Tsallis, Phys. Rev. E **58**, 1442 (1998).
[17] L. Borland, A. R. Plastino, and C. Tsallis, J. Math. Phys. **39**, 6490 (1998); Errata: **40**, 2196 (1999).
[18] R. G. Gallager, *Information Theory and Reliable Communication* (Wiley, New York, 1968).
[19] We will report the extension of the theorem elsewhere.
[20] E. T. Jaynes, Phys. Rev. **106**, 620 (1957); **108**, 171 (1957).
[21] E. T. Jaynes, in *Papers on Probability, Statistics and Statistical Physics*, edited by R. D. Rosenkrantz (D. Reidel, Boston, 1983); E. T. Jaynes, in *Statistical Physics*, edited by W. K. Ford (Benjamin, New York, 1963).