Hence, there exists a constant $C > 0$ such that

$$F_n \geq \frac{C}{(n+1)^{|\mathcal{X}|}} 2^{-n\, D(Q_X^{(n)} \,\|\, P_X)}, \quad \text{for all sufficiently large } n$$

which, together with the continuity of $D(Q_X \,\|\, P_X)$ with respect to Q_X, establishes (40) for case (B). □

ACKNOWLEDGMENT

The author is grateful to the Associate Editor and an anonymous reviewer for their helpful comments. In particular, the Associate Editor's comment on (n, k) linear codes was significant in revision of this correspondence. The author wishes to thank I. Nakano for discussions on the proof of Lemma 2.

REFERENCES

[1] C. H. Benett, G. Brassard, C. Crépeau, and U. M. Maurer, "Generalized privacy amplification," *IEEE Trans. Inf. Theory*, vol. 41, pp. 1915–1923, 1995.
[2] J. L. Carter and M. N. Wegman, "Universal classes of hash functions," *J. Comput. Syst. Sci.*, vol. 18, pp. 143–154, 1979.
[3] T. Cover, "A proof of data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inf. Theory*, vol. IT-21, pp. 226–228, 1975.
[4] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[5] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inf. Theory*, vol. IT-28, pp. 585–592, 1982.
[6] I. Csiszár and J. Körner, "Towards a general theory of source networks," *IEEE Trans. Inf. Theory*, vol. IT-26, pp. 155–165, 1980.
[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Source*. New York: Academic, 1981.
[8] H. Krawczyk, "LFSR-based hashing and authentication," in *Proc. Adv. Cryptol.—CRYPTO'95 (Lecture Notes in Computer Science)*. Berlin, Germany: Springer-Verlag, 1994, vol. 963, pp. 129–139.
[9] L. D. Davisson, "Comments on 'sequence time coding for data compression," *Proc. IEEE*, vol. 54, p. 2010, 1966.
[10] T. S. Han, *Information-Spectrum Methods in Information Theory*. New York: Springer-Verlag, 2003.
[11] H. Koga and H. Yamamoto, "Asymptotic properties on codeword lengths of an optimal FV code for general sources," *IEEE Trans. Inf. Theory*, vol. 51, pp. 1546–1555, 2005.
[12] K. Kurosawa and T. Yoshida, "Strongly universal hashing and identification codes via channels," *IEEE Trans. Inf. Theory*, vol. 45, pp. 2091–2095, 1999.
[13] T. J. Lynch, "Sequence time coding for data compression," *Proc. IEEE*, vol. 54, pp. 1490–1491, 1966.
[14] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[15] U. Maurer and S. Wolf, "Information-theoretic key agreement: From weak to strong secrecy for free," in *Adv. Cryptol.—EUROCRYPT'00 (Lecture Notes in Computer Science)*. Berlin, Germany: Springer-Verlag, 2000, vol. 1807, pp. 351–368.
[16] J. Muramatsu, "Source coding algorithms using the randomness of a past sequence," *IEICE Trans. Fund.*, vol. E88-A, pp. 1063–1083, 2005.
[17] D. R. Stinson, "Universal hashing and authentication codes," *Des., Codes Cryptogr.*, vol. 4, pp. 369–380, 1994.
[18] M. N. Wegman and J. L. Carter, "New hash functions and their use in authentication and set equality," *J. Comput. Syst. Sci.*, vol. 22, pp. 265–279, 1981.
[19] J. Ziv and A. Lempel, "Compression of individual sequence via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. IT-24, pp. 530–536, 1978.

On Defining Partition Entropy by Inequalities

Ping Luo, Guoxing Zhan, Qing He,
Zhongzhi Shi, *Senior Member, IEEE*, and Kevin Lü

***Abstract*—Partition entropy is the numerical metric of uncertainty within a partition of a finite set, while *conditional entropy* measures the degree of difficulty in predicting a decision partition when a condition partition is provided. Since two direct methods exist for defining conditional entropy based on its partition entropy, the inequality postulates of monotonicity, which conditional entropy satisfies, are actually additional constraints on its entropy. Thus, in this paper partition entropy is defined as a function of probability distribution, satisfying all the inequalities of not only partition entropy itself but also its conditional counterpart. These inequality postulates formalize the intuitive understandings of uncertainty contained in partitions of finite sets. We study the relationships between these inequalities, and reduce the redundancies among them. According to two different definitions of conditional entropy from its partition entropy, the convenient and unified checking conditions for any partition entropy are presented, respectively. These properties generalize and illuminate the common nature of all partition entropies.**

***Index Terms*—Conditional entropy, inequality, partition entropy, uncertainty.**

I. INTRODUCTION

Learning is an important cognitive process that allows the making of correct decisions and improves performance. From an information theory point of view, learning can be seen as a reduction of uncertainty and the amount by which the uncertainty is reduced can be an indicator of the speed of learning [1]. Thus, *partition entropy* [2], measuring uncertainty and impurity in a given *partition* of a finite set, is an important concept in cognitive and computer science.

Conditional Entropy [2], defined based on its partition entropy, is another significant concept. It describes the degree of difficulty in predicting a *decision partition* by a *condition partition*. It is also the measure of uncertainty left in a decision partition after a condition partition is provided. This concept is widely used in the field of Machine Learning, as heuristics to guide the greedy search for suboptimal solutions. For example, [3, Algorithm C4.5], which is a popular algorithm for building a decision tree, uses the Shannon conditional entropy as a metric to select the local "optimal" attribute to branch. In the algorithm for attribute reduction of information view [4], the Shannon conditional entropy is also selected as the measure of attribute importance for decision predicting. In these algorithms, the one with the minimal conditional entropy among all available options is chosen to continue the following steps. Thus, only the *relative* magnitude of entropies for

Manuscript received November 23, 2005; revised December 18, 2006. This work is supported by the National Science Foundation of China (No. 60435010, 90604017, 60675010), the 863 Project (No. 2006AA01Z128), National Basic Research Priorities Programme (No. 2003CB317004), and the Nature Science Foundation of Beijing (No. 4052025).

P. Luo is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China (e-mail: ping.luo@gmail.com). He is also with the Graduate University of the Chinese Academy of Sciences.

G. Zhan is with the Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, China. He is also with the Graduate University of the Chinese Academy of Sciences.

Q. He and Z. Shi are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China.

K. Lü is with the Brunel University, Uxbridge UB8 3PH, U.K.

Communicated by V. A. Vaishampayan, Associate Editor for At Large.
Digital Object Identifier 10.1109/TIT.2007.903124

these options is considered, rather than the *absolute* value. Therefore, in this paper, only the inequality postulates, which actually formalize the intuitive understanding of uncertainty, are used to define partition entropy, along with two inherent equality properties of *symmetry* and *expansibility* (defined below).

In previous work, partition entropy is defined as the *Schur-concave* function [5], which satisfies the inequalities on the majorization lattice of probability distributions [6][7]. However, these inequalities concern only partition entropy itself and do not involve its conditional counterpart. Conditional entropy, as a function of a condition partition argument and a decision partition argument, inherently owns the following two inequalities, which coincide with the intuitive understandings of uncertainty in cognitive science as follows:

- if the decision partition is fixed, the finer a condition partition is, the more competent it is to predict the decision partition, and thus the less the conditional entropy is;
- if the condition partition is fixed, the finer a decision partition is, the more difficult it is to be predicted from the condition partition, and thus the greater the conditional entropy is.

These two postulates indicate that conditional entropy is monotonic in the condition partition argument and dually monotonic in the decision partition argument. These are the monotonicity properties, which conditional entropy holds inherently. Since direct method exists for defining conditional entropy based on its partition entropy, these two inequality postulates are actually additional constraints on partition entropy.

In this correspondence, we add the aforementioned postulates to a new definition of partition entropy, and reduce the redundancies in all the inequality postulates. According to two different definitions of conditional entropy based on its partition entropy, we present the checking conditions (sufficient and necessary or sufficient only) for any partition entropy, respectively. It should be noted that the partition entropy resulted from the axiomatization method [2] by the equalities is within the family of partition entropies defined in this paper. These results generalize and illuminate the common nature of all partition entropies.

II. Basic Notations and Notions

In the following, we adopt the notations in [2] and [6], and more information on partition entropy can be found there. The set of reals, the set of positive reals, the set of natural numbers, and the set of positive natural numbers are denoted by $\mathbb{R}, \mathbb{R}_{>0}, \mathbb{N}, \mathbb{N}_{>0}$, respectively. All other sets considered in the following discussion are nonempty and finite:

$\pi = \{A_1, \ldots, A_m\}$ is a partition of a set A, iff $\cup_{i=1}^{m} A_i = A$ and $A_i \cap A_j = \emptyset (i \neq j)$. A *block* of a partition refers to any element in a partition of a set A. Let $PART(A)$ be the set of partitions of set A. The class of all partitions of finite sets is denoted by $PART$. The one-block partition of A is denoted by ι_A. The partition $\{\{a\} \mid a \in A\}$ is denoted by ω_A. Thus, ι_A is the most coarse partition of A, while ω_A is the finest partition of A.

Let $\pi, \pi' \in PART(A)$, then $\pi \subseteq \pi'$ if every block of π is included in a block of π'. It is obvious that $\omega_A \subset \iota_A$.

If A,B are two disjoint sets, $\pi \in PART(A), \sigma \in PART(B)$, where $\pi = \{A_1, \ldots, A_m\}, \sigma = \{B_1, \ldots, B_n\}$, then the partition $(\pi + \sigma) \in PART(A \cup B)$ is given by

$$\pi + \sigma = \{A_1, \ldots, A_m, B_1, \ldots, B_n\}.$$

Let $\pi \in PART(A)$ and $C \subseteq A$. The "trace" of π on C is given by

$$\pi_C = \{A_i \cap C \mid A_i \in \pi \text{ such that } A_i \cap C \neq \emptyset\}.$$

It is clear that $\pi_C \in PART(C)$.

Let $\pi, \sigma \in PART(A)$ (two partitions defined on the same set A), where $\pi = \{A_1, \ldots, A_m\}, \sigma = \{B_1, \ldots, B_n\}$. The partition $\pi \wedge \sigma$ whose blocks consist of the nonempty intersections of the blocks of π and σ can be written as

$$\pi \wedge \sigma = \pi_{B_1} + \cdots + \pi_{B_n} = \sigma_{A_1} + \cdots + \sigma_{A_m}.$$

A. Partition Entropy

Partition entropy is a mapping

$$\mathcal{H} : PART \to \mathbb{R} \tag{1}$$

satisfying some additional conditions.

If $\pi = \{A_1, \ldots, A_n\}$ is a partition of a set A, then the probability distribution vector attached to π is $P(\pi) = (p_1, \ldots, p_n)$, where $p_i = \frac{|A_i|}{|A|}$ for $1 \leq i \leq n$. Thus, it is straightforward to consider the notion of partition entropy via the entropy of the corresponding probability distribution. We define the measure function of $\mathcal{H}$ as a mapping

$$\mathcal{M} : \Delta \to \mathbb{R}$$

such that $\mathcal{H}(\pi) = \mathcal{M}(P(\pi))$ for every $\pi \in PART$, where $\Delta = \{P(\pi) \mid \pi \in PART\}$.

The blocks in a partition π are unordered while the elements in $P(\pi)$ are ordered. Thus, the inherent postulate of $\mathcal{M}$ is that it is symmetric in the sense that

$$\mathcal{M}(P(\pi)) = \mathcal{M}(P'(\pi)) \tag{2}$$

where $P'(\pi)$ is any permutation of $P(\pi)$.

The other equality postulate of $\mathcal{M}$ is expansibility in the sense that for every $P \in \Delta_m$

$$\mathcal{M}(P) = \mathcal{M}(P') \tag{3}$$

where $P = (p_1, \ldots, p_m), P' = (p_1, \ldots, p_m, 0)$, and $\Delta_m = \{(p_1, \ldots, p_m) : 0 \leq p_i \leq 1 \text{ for } i = 1, \ldots, m, p_1 + \cdots + p_m = 1\}$.

Formulas (2) and (3) are the only two equalities the partition entropy in this paper must satisfy.

B. Entropically Comparable Relationship Between Partitions

For general $p, q \in \Delta_m$, it is hard to say precisely when the prediction under q is not easier than under p. However, there are special p and q for which this can be done. Namely, if $p, q \in \Delta_2$ and $p = (\alpha, 1-\alpha), q = (\beta, 1-\beta)$, then the prediction under p is not easier than that under q if and only if $(\alpha, 1-\alpha)$ is at least as close as $(\beta, 1-\beta)$ to the uniform distribution $(\frac{1}{2}, \frac{1}{2})$. To be "at least as close" naturally means

$$\left|\alpha - \frac{1}{2}\right| \leq \left|\beta - \frac{1}{2}\right|. \tag{4}$$

This observation can be applied to all vectors $p, q \in \Delta_m$ with only two different coordinates. p is said to be a *smoothing* of q, in symbol $p = Sm(q)$, if there exist $1 \leq j \neq k \leq m$ such that all coordinates of p and q coincide except the jth and kth, and these two coordinates satisfy (4) for

$$\alpha = \frac{p_j}{p_j + p_k} \quad \text{and} \quad \beta = \frac{q_j}{q_j + q_k}.$$

We extend the above smoothing relationship [5] between two probability distributions to the entropically comparable relationship between partitions. Let $\pi = \{A_1, \ldots, A_m\}, \sigma = \{B_1, \ldots, B_n\}$, and $p = P(\pi), q = P(\sigma)$. Without loss of generality, if $m < n$ we can add

$(n-m)$ 0's to the right side of p to make the dimensions of the two vectors equal while keeping the entropy unchanged. Then, the partial order $\preceq$ on $PART \times PART$ is defined as follows. For any $\pi, \sigma \in PART, \pi \preceq \sigma$ iff $p = Sm(q)$ or exists $\kappa \in PART(r = P(\kappa))$ s.t. $\kappa \preceq \sigma$ and $p = Sm(r)$. If $\pi \preceq \sigma$ or $\sigma \preceq \pi$, it is easy to tell which partition entropy is bigger between π and σ because smoothing a probability distribution means to increase its partition entropy. This time it can be said that π and σ are *entropically comparable* with each other.

It is clear that the relationship $\preceq$ between partitions is reflexive, transitive and anti-symmetric.

C. Equivalents of Entropically Comparable Relationship

The *decreasing* rearrangement of $P(\pi)$ is denoted by $P_{\downarrow}(\pi) = (p_{[1]}, p_{[2]}, \ldots, p_{[n]})$, where $p_{[1]} \geq p_{[2]} \geq \cdots \geq p_{[n]}$. For entropically comparable relationships between partitions, the following conditions are equivalent [6]:

Let $\pi = \{A_1, \ldots, A_m\}, \sigma = \{B_1, \ldots, B_n\}$, and $p = P_{\downarrow}(\pi), q = P_{\downarrow}(\sigma)$, supposing $m \leq n$ and adding $(n-m)$ 0's to the right side of p as follows:

1) $\pi \preceq \sigma$;
2) $p = qA$ for some doubly stochastic matrix A (Matrix $A = (a_{jk})_{j,k=1}^{n}$ is said to be doubly stochastic if all its row and column vectors belong to Δ_n);
3) $\sum_{i=1}^{k} p_{[i]} \leq \sum_{i=1}^{k} q_{[i]}$ $(k = 1, \ldots, n-1)$, $\sum_{i=1}^{n} p_{[i]} = \sum_{i=1}^{n} q_{[i]}$.

Equivalent 2) shows that vector q post-multiplied by a doubly stochastic matrix is equivalent to smoothing q. Equivalent 3) provides a convenient method to check whether two partitions are entropically comparable.

D. Two Definitions of Conditional Entropy

Given a set A, conditional entropy is a mapping

$$\mathcal{C} : PART(A)^2 \to \mathbb{R}. \tag{5}$$

The first argument refers to a condition partition while the second one refers to a decision partition. If π, σ are two partitions of $A, \mathcal{C}(\pi, \sigma)$ measures the degree of difficulty in predicting σ by π. Based on an existing partition entropy, we give the definition of conditional entropy.

Definition II.1: Let $\pi, \sigma \in PART(A), \pi = \{A_1, \ldots, A_m\}, \sigma = \{B_1, \ldots, B_n\}$. A conditional entropy $\mathcal{C}^1$ is a function $\mathcal{C}$ in (5) such that

$$\mathcal{C}^1(\pi, \sigma) = \sum_{i=1}^{m} \frac{|A_i|}{|A|} \cdot \mathcal{H}(\sigma_{A_i})$$

where σ_{A_i} is the "trace" of σ on A_i.

Definition II.1 states that the conditional entropy $\mathcal{C}^1$ is the expected value of the entropies calculated according to the conditional distributions. Namely, $\mathcal{C}^1(\pi, \sigma) = E_{A_i}(\mathcal{H}(\sigma_{A_i})), A_i \in \pi$.

Definition II.2: Let $\pi, \sigma \in PART(A), \pi = \{A_1, \ldots, A_m\}, \sigma = \{B_1, \ldots, B_n\}$. A conditional entropy $\mathcal{C}^2$ is a function $\mathcal{C}$ in (5) such that

$$\mathcal{C}^2(\pi, \sigma) = \mathcal{H}(\pi \wedge \sigma) - \mathcal{H}(\pi).$$

Definition II.2 states that the conditional entropy $\mathcal{C}^2$ is the difference between two entropies: one is of the intersection of the condition and decision partition, while the other is of the condition partition only.

The equality $\mathcal{C}^1(\pi, \sigma) = \mathcal{C}^2(\pi, \sigma)$ yields the Shannon entropy. Thus, this famous axiomatization of the Shannon entropy shows the rationality of these two definitions.

III. Inequality Postulates of Partition Entropy

All the inequalities partition entropy and its corresponding conditional counterpart must satisfy are listed in this section.

Postulate III.1: For any $\pi \in PART(A)$

$$\mathcal{H}(\iota_A) \leq \mathcal{H}(\pi) \leq \mathcal{H}(\omega_A).$$

Postulate III.2: Let any $\pi, \pi' \in PART(A)$ and $\pi \subseteq \pi'$, then

$$\mathcal{H}(\pi') \leq \mathcal{H}(\pi).$$

Postulate III.3: Let any $\pi, \pi' \in PART(A)$ and $\pi \preceq \pi'$, then

$$\mathcal{H}(\pi') \leq \mathcal{H}(\pi).$$

When a function $\mathcal{H}$ defined by (1) satisfies Postulate III.3, its corresponding measure function $\mathcal{M}$ is Schur-concave. In the following, $\mathcal{H}$ is Schur-concave or concave if and only if its corresponding measure function $\mathcal{M}$ is Schur-concave or concave.

Postulate III.4: Let any $\pi, \pi', \sigma \in PART(A)$ and $\pi \subseteq \pi'$, then

$$\mathcal{C}(\pi, \sigma) \leq \mathcal{C}(\pi', \sigma).$$

Postulate III.5: Let any $\pi, \sigma, \sigma' \in PART(A)$ and $\sigma \subseteq \sigma'$, then

$$\mathcal{C}(\pi, \sigma') \leq \mathcal{C}(\pi, \sigma).$$

Postulate III.4 and III.5 state that conditional entropy $\mathcal{C}$ should be monotonic in the first argument and dually monotonic in the second argument. Postulate III.4 shows that finer condition partition has more ability for predicting, while Postulate III.5 shows that coarser decision partition relaxes the requirement of precision for classification and thus decreases the difficulty for predicting. They are two postulates conditional entropy holds inherently.

IV. Relationships Between Inequality Postulates of Partition Entropy

In this section we study the relationships among these inequality postulates, reduce the redundancies in them, and give a new definition of partition entropy.

Theorem IV.1. [5]: If a function $\mathcal{H}$ defined by (1) satisfies Postulate III.3, it satisfies Postulate III.1.

Theorem IV.2. [5]: If a function $\mathcal{H}$ defined by (1) satisfies Postulate III.3, it satisfies Postulate III.2.

Proof: $\pi \subseteq \pi'$ implies $\pi \preceq \pi'$. Then it follows the conclusion. □

Theorem IV.3: If a function $\mathcal{H}$ defined by (1) satisfies Postulate III.3, its conditional counterpart $\mathcal{C}^1$ satisfies Postulate III.5.

Proof: $\sigma \subseteq \sigma'$ implies $\sigma_{A_i} \subseteq \sigma'_{A_i}$ for every $A_i \in \pi$. Thus, $\sigma_{A_i} \preceq \sigma'_{A_i}$ for every $A_i \in \pi$. From Postulate III.3, $\mathcal{H}(\sigma'_{A_i}) \leq \mathcal{H}(\sigma_{A_i})$. It follows the conclusion immediately. □

Theorem IV.4: If a function $\mathcal{H}$ defined by (1) satisfies Postulate III.3, its conditional counterpart $\mathcal{C}^2$ satisfies Postulate III.5.

Proof: $\sigma \subseteq \sigma'$ implies $(\pi \wedge \sigma) \subseteq (\pi \wedge \sigma')$. Then it follows that $\mathcal{H}(\pi \wedge \sigma) - \mathcal{H}(\pi \wedge \sigma') \geq 0$. Thus
$$\mathcal{C}^2(\pi, \sigma) - \mathcal{C}^2(\pi, \sigma') = \mathcal{H}(\pi \wedge \sigma) - \mathcal{H}(\pi \wedge \sigma') \geq 0. \qquad \square$$

From the above theorems the definition of partition entropy is given as follows.

Definition IV.1: When a function defined by (1) satisfies Postulate III.3 and Postulate III.4, and its corresponding measure function $\mathcal{M}$ is symmetric and expansible, it is a partition entropy.

V. Checking Conditions for Partition Entropy

A. When Conditional Entropy Defined as $\mathcal{C}^1$

Next we give a sufficient and necessary checking condition for any partition entropy when conditional entropy is defined as $\mathcal{C}^1$. We first give the definition of concavity for functions of n-dimensional inputs as follows.

Definition V.1: Suppose $X \subset \mathbb{R}^n$ is a convex set. $f : X \to \mathbb{R}$ is concave if for any $x, y \in X$, we have, for all $\lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y).$$

A direct property of concave functions is: if $f : X \to \mathbb{R}$ is concave in a convex set $X \subset \mathbb{R}^n$, for any $x_1, \ldots, x_m \in X, \lambda_1, \ldots, \lambda_m \in (0,1)$ and $\sum_{i=1}^m \lambda_i = 1$, we have

$$f\left(\sum_{i=1}^m \lambda_i x_i\right) \geq \sum_{i=1}^m \lambda_i f(x_i).$$

Lemma V.1. [6]: If a function $\mathcal{H}$ defined by (1) is concave and its corresponding measure function is symmetric, it is Schur-concave.

Lemma V.2: When conditional entropy is defined as $\mathcal{C}^1$, if and only if its corresponding $\mathcal{H}$ is concave, it satisfies Postulate III.4.

Proof $\Rightarrow$: Let $\pi = \{A_{11}, \ldots, A_{1u_1}, \ldots, A_{m1}, \ldots, A_{mu_m}\}$, $\pi' = \{A_1, \ldots, A_m\}$, $A_k = \cup_{i=1}^{u_k} A_{ki}$ and $\sigma = \{B_1, \ldots, B_n\}$. And let $\frac{|A_{ki} \cap B_j|}{|A_{ki}|} = a_{ij}^k$ and $\lambda_i^k = \frac{|A_{ki}|}{|A_k|}$, thus $\frac{|A_k \cap B_j|}{|A_k|} = \sum_{i=1}^{u_k} \lambda_i^k a_{ij}^k$ for $k = 1, \ldots, m$ and $j = 1, \ldots, n$. Because $\mathcal{H}$ is concave and $\mathcal{M}$ is the measure function of $\mathcal{H}$, it follows:

$$\mathcal{M}\left(\sum_{i=1}^{u_k} \lambda_i^k a_{i1}^k, \ldots, \sum_{i=1}^{u_k} \lambda_i^k a_{in}^k\right) \geq \sum_{i=1}^{u_k} \lambda_i^k \mathcal{M}\left(a_{i1}^k, \ldots, a_{in}^k\right).$$

Because $\mathcal{M}(\sum_{i=1}^{u_k} \lambda_i^k a_{i1}^k, \ldots, \sum_{i=1}^{u_k} \lambda_i^k a_{in}^k) = \mathcal{H}(\sigma_{A_k})$ and $\mathcal{M}(a_{i1}^k, \ldots, a_{in}^k) = \mathcal{H}(\sigma_{A_{ki}})$, it follows:

$$\mathcal{H}(\sigma_{A_k}) \geq \sum_{i=1}^{u_k} \lambda_i^k \mathcal{H}(\sigma_{A_{ki}}),$$

$$\frac{|A_k|}{|A|} \mathcal{H}(\sigma_{A_k}) \geq \sum_{i=1}^{u_k} \frac{|A_{ki}|}{|A|} \mathcal{H}(\sigma_{A_{ki}}), \qquad \text{for } k = 1, \ldots, m.$$

Then

$$\sum_{k=1}^m \frac{|A_k|}{|A|} \mathcal{H}(\sigma_{A_k}) \geq \sum_{k=1}^m \sum_{i=1}^{u_k} \frac{|A_{ki}|}{|A|} \mathcal{H}(\sigma_{A_{ki}}).$$

It follows that

$$\mathcal{C}^1(\pi, \sigma) \leq \mathcal{C}^1(\pi', \sigma).$$

$\Leftarrow$.

Let $\pi = \{A_{11}, \ldots, A_{1u_1}, A_2, \ldots, A_m\}$, $\pi' = \{A_1, A_2, \ldots, A_m\}$, $A_1 = \cup_{i=1}^{u_1} A_{1i}$, and $\sigma = \{B_1, \ldots, B_n\}$. And let $\frac{|A_{1i} \cap B_j|}{|A_{1i}|} = a_{ij}^1$ and $\lambda_i^1 = \frac{|A_{1i}|}{|A_1|}$, thus $\frac{|A_1 \cap B_j|}{|A_1|} = \sum_{i=1}^{u_1} \lambda_i^1 a_{ij}^1$. Because $\mathcal{C}^1(\pi, \sigma) \leq \mathcal{C}^1(\pi', \sigma)$, it follows that

$$\mathcal{H}(\sigma_{A_1}) \geq \sum_{i=1}^{u_1} \lambda_i^1 \mathcal{H}(\sigma_{A_{1i}})$$

which means that $\mathcal{H}$ is concave. $\square$

Theorem V.1: When conditional entropy is defined as $\mathcal{C}^1$, if and only if its corresponding $\mathcal{H}$ is concave and the measure function $\mathcal{M}$ of $\mathcal{H}$ is symmetric and expansible, $\mathcal{H}$ is a partition entropy.

Proof: By Lemma V.1, Lemma V.2 and the definition of partition entropy, it holds directly. $\square$

Corollary V.1: Let $f : [0,1] \to \mathbb{R}, \mathcal{M}(p_i, \ldots, p_m) = \sum_{i=1}^m f(p_i)$ be the measure function of a function $\mathcal{H}$ defined by (1) and $f(0) = 0$. When conditional entropy is defined as $\mathcal{C}^1$, if f is concave in $[0,1]$, $\mathcal{H}$ is a partition entropy.

Proof: It is clear that $\mathcal{M}$ is symmetric and expansible. If f is concave in $[0,1]$, $\mathcal{M}$ is concave in $[0,1]^m$. By Theorem V.1, $\mathcal{H}$ is a partition entropy. $\square$

B. When Conditional Entropy Defined as $\mathcal{C}^2$

Next we give a sufficient checking condition for any partition entropy when conditional entropy is defined as $\mathcal{C}^2$. For convenience and clarity, we first give the following definition.

Definition V.2: A function $f : [0,1] \longrightarrow \mathbb{R}$ is called additivity-concave if for any $n \in \mathbb{N}_{>0}$ the following inequality is satisfied:

$$f(s) + f(t) - f(s+t) \geq \sum_{i=1}^n [f(a_i s) + f(b_i t) - f(a_i s + b_i t)] \quad (6)$$

where $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i = 1, 0 \leq s \leq 1, 0 \leq t \leq 1, 0 \leq s+t \leq 1$, and $a_i \geq 0, b_i \geq 0$ for $i = 1, \ldots, n$.

Lemma V.3: Let $\pi = \{A_1, A_2, \ldots, A_m\}$. When conditional entropy is defined as $\mathcal{C}^2$ and $f(0) = 0$, if and only f is additivity-concave, $\mathcal{H}(\pi) = \sum_{i=1}^m f(\frac{|A_i|}{A})$ satisfies Postulate III.4.

Proof: First, because $\mathcal{H}(\pi) = \sum_{i=1}^m f(\frac{|A_i|}{A})$, $\mathcal{H}(\pi)$ is symmetric. Let $\pi' = \{A_1 \cup A_2, \ldots, A_m\}, A = \cup_{i=1}^m A_i$, $\sigma = \{B_1, B_2, \ldots, B_n\}, s = \frac{|A_1|}{|A|}, t = \frac{|A_2|}{|A|}, a_i = \frac{|A_1 \cap B_i|}{|A_1|}$, and $b_i = \frac{|A_2 \cap B_i|}{|A_2|}$ for $i = 1, \ldots, n$

$$\mathcal{C}^2(\pi', \sigma) - \mathcal{C}^2(\pi, \sigma) = f(s) + f(t) - f(s+t) - \sum_{i=1}^n [f(a_i s) + f(b_i t) - f(a_i s + b_i t)]$$

$\Rightarrow$:

if f is additivity-concave, then $\mathcal{C}^2(\pi', \sigma) - \mathcal{C}^2(\pi, \sigma) \geq 0$, which means that the combination of any two blocks in the condition partition will increase the conditional entropy. Thus, it satisfies Postulate III.4.

$\Leftarrow$: trivial. $\square$

Theorem V.2: Suppose a function $f : [0,1] \longrightarrow \mathbb{R}$ is continuous and concave, $f(0) = 0$, the second derivative f'' exists in $(0,1)$ and is continuous in $(0,1)$. And f satisfies the following inequality:

$$[f''(u) + f''(v)] \cdot [f''(x) + f''(y)] \leq [f''(u) + f''(v) + f''(x) + f''(y)] \cdot [f''(u+x) + f''(v+y)]$$

whenever $u, v, x, y \in (0,1), u+x<1, v+y<1, u+v<1, x+y<1$. Let $\pi = \{A_1, A_2, \ldots, A_m\}$. Then $\mathcal{H}(\pi) = \sum_{i=1}^m f(\frac{|A_i|}{A})$ is a partition entropy when its conditional counterpart is defined as $\mathcal{C}^2$.

Proof: Directly by Lemma V.3 and Theorem I.1 in the Appendix. □

Corollary V.2: Suppose a function $f : [0,1] \longrightarrow \mathbb{R}, f(0) = 0$, f is continuous on $[0,1]$, the third derivative f''' exists in $(0,1)$, $f''(x) \leq 0$ and $f'''(x) \leq 0$ for any $x \in (0,1)$. Let $\pi = \{A_1, A_2, \ldots, A_m\}$. Then $\mathcal{H}(\pi) = \sum_{i=1}^{m} f(\frac{|A_i|}{A})$ is a partition entropy when its conditional counterpart is defined as $\mathcal{C}^2$.

Proof: Directly by Lemma V.5 and Corollary I.1 in the Appendix. □

VI. Examples of Partition Entropy

This section gives some examples of partition entropy with constraints from the two definitions of conditional entropy $\mathcal{C}^1$ and $\mathcal{C}^2$, respectively. All these examples are under the following assumption: let $\pi = \{A_1, \ldots, A_n\}$ be a partition of a set A, the probability distribution vector attached to π be $P(\pi) = (p_1, \ldots, p_n)$, where $p_i = \frac{|A_i|}{|A|}$ for $1 \leq i \leq n$.

A. Examples With the Constraints From the Conditional Entropy $\mathcal{C}^1$

The examples in this subsection are partition entropies when their conditional counterparts are defined as $\mathcal{C}^1$.

Example 1 (The Shannon Entropy): $\mathcal{H}(\pi) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$ is a partition entropy.

Example 2 ([8]): $\mathcal{H}(\pi) = \sum_{i=1}^{n} p_i e^{1-p_i}$ is a partition entropy. A generalized form of this partition entropy is $\mathcal{H}(\pi) = \sum_{i=1}^{n} p_i(e^{a-p_i} + b)$, where $a, b \in \mathbb{R}$.

Example 3: $\mathcal{H}(\pi) = \sum_{i=1}^{n} p_i(1 - p_i)$ is a partition entropy. Its conditional counterpart, defined as $\mathcal{C}^1$, is referred to as the Gini Index in [9]. A generalized form of this partition entropy is $\mathcal{H}(\pi) = \sum_{i=1}^{n} a_1 p_i^3 + a_2 p_i^2 + a_3 p_i$, where $a_1, a_2, a_3 \in \mathbb{R}$, $(a_1 \geq 0 \wedge 3a_1 + a_2 \leq 0)$ or $(a_1 \leq 0 \wedge a_2 \leq 0)$.

Example 4 ([2], [10]): When $\beta > 1, \mathcal{H}(\pi) = k(1 - \sum_{i=1}^{n} p_i^{\beta})$ is a partition entropy, where $k \in \mathbb{R}_{>0}$.

Example 5 ([2], [10]): When $0 < \beta < 1, \mathcal{H}(\pi) = k(\sum_{i=1}^{n} p_i^{\beta} - 1)$ is a partition entropy, where $k \in \mathbb{R}_{>0}$.

Example 6: $\mathcal{H}(\pi) = 1 - \max_{i=1}^{n} p_i$ is a partition entropy. Its conditional counterpart is referred to as the Goodman-Kruskal coefficient in [11][12].

By Corollary V.1, Examples 1 through 5 are proved to be partition entropies. Because $\max_{i=1}^{n} p_i$ is convex in $[0,1]^n$ for any $n \in \mathbb{N}_{>0}$, Example 6 is easily proved to be a partition entropy by Theorem V.1.

B. Examples With the Constraints From the Conditional Entropy $\mathcal{C}^2$

The examples in this subsection are partition entropies when their conditional counterparts are defined as $\mathcal{C}^2$.

Example 1: The Shannon entropy is also a partition entropy when its conditional counterpart is defined as $\mathcal{C}^2$.

Example 2: $\mathcal{H}(\pi) = \sum_{i=1}^{n} p_i(1 - p_i)$ is a partition entropy. Its conditional counterpart defined as $\mathcal{C}^2$ is presented in [13], [14]. A generalized form of this partition entropy is $\mathcal{H}(\pi) = \sum_{i=1}^{n} a_1 p_i^3 + a_2 p_i^2 + a_3 p_i$, where $a_1, a_2, a_3 \in \mathbb{R}, a_1 \leq 0 \wedge a_2 \leq 0$.

Example 3: When $\beta > 2, \mathcal{H}(\pi) = k(1 - \sum_{i=1}^{n} p_i^{\beta})$ is a partition entropy, where $k \in \mathbb{R}_{>0}$.

The Shannon entropy satisfy the sufficient condition in Theorem V.2, thus it is a partition entropy when the conditional entropy is defined as $\mathcal{C}^2$. The results in Examples 2 and 3 can be easily proved by Corollary V.2.

VII. Conclusion

In this correspondence, the monotonicity properties of conditional entropy are formalized in Postulate III.4 (monotonicity in condition partition argument of $\mathcal{C}$) and III.5 (dual monotonicity in decision partition argument of $\mathcal{C}$). We add these properties of conditional entropy to the definition of partition entropy, and reduce the redundancies among all the inequality postulates. This new definition of partition entropy is more strict than the previous one [5], which is Schur-concave only. The main theoretical contributions of this paper are Theorems IV.3 and IV.4, Lemma V.2, Theorems V.1 and V.2. Theorems IV.3 and IV.4 show that the dual monotonicity in the decision partition argument of conditional entropy (defined as both $\mathcal{C}^1$ and $\mathcal{C}^2$) is a property of a Schur-concave function. Lemma V.2 demonstrates that the monotonicity in the condition partition argument of the conditional entropy $\mathcal{C}^1$ is actually equivalent to the concavity of its partition entropy. Theorem V.1 gives a sufficient and necessary condition for any partition entropy when conditional entropy is defined as $\mathcal{C}^1$, while the condition in Theorem V.2 are sufficient, but not necessary for any partition entropy when conditional entropy is defined as $\mathcal{C}^2$. These results present the mathematical insights into monotonicity properties of conditional entropy, provide the convenient and unified checking methods for any partition entropy. It should be noted that it is still an open problem to find the sufficient and necessary condition for any partition entropy when conditional entropy is defined as $\mathcal{C}^2$, which is actually the sufficient and necessary condition for additivity-concave functions.

The theorems in this paper focus on partitions of finite sets, which can be naturally defined by grouping objects with common values in certain attributes and are widely used in Machine Learning. They illuminate a family of partition entropies, which can be used as heuristics in the algorithms of Machine Learning. The existence of various types of entropies suggests that different entropies should be used to produce rather distinct patterns for classification and clustering.

Appendix I

Lemma I.1: A function $f : [0,1] \longrightarrow \mathbb{R}$ is additivity-concave if and only if the inequality (6) holds when $n = 2$.

Proof:

$\Rightarrow$: Obvious.

$\Leftarrow$: Suppose the inequality (6) holds when $n = 2$. We prove by induction on n that (6) holds for any positive integer n. We assume that the inequality (6) is satisfied when $n = k - 1$, then when $n = k$

$$\begin{aligned} f(s) + f(t) - f(s+t) &\geq f\left(\left(\sum_{i=1}^{k-1} a_i\right)s\right) + f\left(\left(\sum_{i=1}^{k-1} b_i\right)t\right) \\ &\quad - f\left(\left(\sum_{i=1}^{k-1} a_i\right)s + \left(\sum_{i=1}^{k-1} b_i\right)t\right) \\ &\quad + f(a_k s) + f(b_k t) - f(a_k s + b_k t) \quad (7) \\ &\geq \sum_{i=1}^{k-1}[f(a_i s) + f(b_i t) - f(a_i s + b_i t)] \\ &\quad + f(a_k s) + f(b_k t) - f(a_k s + b_k t) \quad (8) \\ &= \sum_{i=1}^{k}[f(a_i s) + f(b_i t) - f(a_i s + b_i t)] \end{aligned}$$

where (7) follows from the condition that the inequality holds when $n = 2$ and (8) follows from the inductive assumption. □

Lemma I.1 suggests that A function $f : [0,1] \longrightarrow \mathbb{R}$ is additivity-concave if and only if the following inequality holds:

$$\begin{aligned} &f(s)+f(t)-f(s+t) \\ &\quad \geq f(as)+f(bt)-f(as+bt) \\ &\qquad + f((1-a)s)+f((1-b)t)-f((1-a)s+(1-b)t)) \end{aligned} \tag{9}$$

where $0 \leq a \leq 1, 0 \leq b \leq 1, 0 \leq s \leq 1, 0 \leq t \leq 1, s+t \leq 1$.

Lemma I.2: Suppose a function $f : [0,1] \longrightarrow \mathbb{R}$ is continuous and concave, $f(0) \leq 0$, the second derivative f'' exists in $(0,1)$ (thus $f'' \leq 0$ in $(0,1)$) and is continuous in $(0,1)$. If f satisfies the following inequality:

$$\begin{aligned} &[f''(u)+f''(v)] \cdot [f''(x)+f''(y)] \\ &\quad \leq [f''(u)+f''(v)+f''(x)+f''(y)] \\ &\qquad \cdot [f''(u+x)+f''(v+y)] \end{aligned} \tag{10}$$

whenever $u, v, x, y \in (0,1), u+x<1, v+y<1, u+v<1, x+y<1$. Then f satisfies (9).

Proof: First, we prove a weaker result : under the hypotheses of the Lemma I.2, if f satisfies the following inequality instead of (10) (notice the small difference between $<$ and $\leq$!)

$$\begin{aligned} &[f''(u)+f''(v)] \cdot [f''(x)+f''(y)] \\ &\quad < [f''(u)+f''(v)+f''(x)+f''(y)] \\ &\qquad \cdot [f''(u+x)+f''(v+y)] \end{aligned} \tag{11}$$

then f satisfies (9).

For fixed $s,t \in [0,1]$, we define a function on $[0,1] \times [0,1]$

$$\begin{aligned} H(a,b) = f(as)+f(bt)-f(as+bt)+f((1-a)s) \\ + f((1-b)t-f((1-a)s+(1-b)t)) \end{aligned} \tag{12}$$

where $a, b \in [0,1]$.

Since $s+t \leq 1, s=1$ if $t=0$ and $s=0$ if $t=1$. It is obvious that the inequality (9) holds if $s=1$ and $t=0$ or $t=1$ and $s=0$. In the following, we assume that $s,t \in (0,1)$.

When $a=1, H(1,b)=f(s)+f(bt)-f(s+bt)+f(0)$. $\frac{\partial H(1,b)}{\partial b} = tf'(bt)-tf'(s+bt)$. Because f is concave, $f'' \leq 0$, which means f' is a decreasing function on $(0,1)$. It follows that $\frac{\partial H(1,b)}{\partial b} \geq 0$. Thus, $H(1,b)$ is increasing as a function of b, and the following inequality holds:

$$\begin{aligned} H(1,b) \leq H(1,1) &= f(s)+f(t)-f(s+t)+f(0) \\ &\leq f(s)+f(t)-f(s+t). \end{aligned} \tag{13}$$

When $a=0, H(0,b)=f(s)+f((1-b)t)-f(s+(1-b)t)+f(0)$. Using the similar method, we can prove that $H(0,b)$ is decreasing as a function of b, and the following inequality holds:

$$\begin{aligned} H(0,b) \leq H(0,0) &= f(s)+f(t)-f(s+t)+f(0) \\ &\leq f(s)+f(t)-f(s+t). \end{aligned} \tag{14}$$

Using similar methods, we can also prove that

$$H(a,1) \leq f(s)+f(t)-f(s+t) \tag{15}$$

$$H(a,0) \leq f(s)+f(t)-f(s+t). \tag{16}$$

From (13), (14), (15) and (16), we see that $H(a,b) \leq f(s)+f(t)-f(s+t)$ holds when (a,b) lies in the boundary of $[0,1] \times [0,1]$. Since H is continuous in $[0,1] \times [0,1]$, it reaches its maximum at the boundary of $[0,1] \times [0,1]$ if H can not reach its maximum in the interior of $[0,1] \times [0,1]$. In that case (9) holds. In the following we shall prove that if f satisfies the inequality (11), then H can not reach its maximum in $(0,1) \times (0,1)$.

If f satisfies the inequality (11), then

$$\begin{aligned} &[f''(as)+f''((1-a)s)] \cdot [f''(bt)+f''((1-b)t)] \\ &\quad < [f''(as)+f''((1-a)s)+f''(bt)+f''((1-b)t] \\ &\qquad \cdot [f''(as+bt)+f''((1-a)s+(1-b)t)]. \end{aligned} \tag{17}$$

With direct calculations, it follows from the inequality (17) that

$$\det \begin{bmatrix} \frac{\partial^2 H}{\partial a^2} & \frac{\partial^2 H}{\partial a \partial b} \\ \frac{\partial^2 H}{\partial a \partial b} & \frac{\partial^2 H}{\partial b^2} \end{bmatrix} < 0. \tag{18}$$

Since the determinant of the Hessian matrix

$$\begin{bmatrix} \frac{\partial^2 H}{\partial a^2} & \frac{\partial^2 H}{\partial a \partial b} \\ \frac{\partial^2 H}{\partial a \partial b} & \frac{\partial^2 H}{\partial b^2} \end{bmatrix}$$

of H is negative in $(0,1) \times (0,1)$, this matrix must be indefinite, and thus H can not reach its maximum in $(0,1) \times (0,1)$. Thus we have proved the weaker result we claim at the beginning of the proof.

Next, we will prove the lemma under the condition (10). Now f satisfies (10). We take any continuous function $g : [0,1] \longrightarrow \mathbb{R}$ with the following properties: $g(0) \leq 0, g''(x) < 0, g'''(x) \leq 0$ for any $x \in (0,1)$. (Such a 'g' exists, for instance, we may set $g(x) = -x^2$.) Then g satisfies (11). We shall show that $f+g$ satisfies (11). For fixed u, v, x, y, define

$$\begin{aligned} D(f) = &[f''(u)+f''(v)+f''(x)+f''(y)] \\ &\cdot [f''(u+x)+f''(v+y)] - [f''(u)+f''(v)] \\ &\cdot [f''(x)+f''(y)]. \end{aligned} \tag{19}$$

Similarly we define $D(g), D(f+g)$. It is obvious that $D(f) \geq 0, D(g) > 0$.

$$\begin{aligned} D(f+g) = &D(f)+D(g)+[f''(u)+f''(v)+f''(x)+f''(y)] \\ &\cdot [g''(u+x)+g''(v+y)] - [f''(u)+f''(v)] \\ &\cdot [g''(x)+g''(y)] \\ &+ [g''(u)+g''(v)+g''(x)+g''(y)] \\ &\cdot [f''(u+x)+f''(v+y)] - [g''(u)+g''(v)] \\ &\cdot [f''(x)+f''(y)] \\ \geq &D(f)+D(g)+[f''(x)+f''(y)] \\ &\cdot [g''(u+x)+g''(v+y)] \\ &+ [g''(u)+g''(v)+g''(x)+g''(y)] \\ &\cdot [f''(u+x)+f''(v+y)] \\ &- [g''(u)+g''(v)] \cdot [f''(x)+f''(y)] \\ \geq &D(f)+D(g)+[f''(x)+f''(y)] \\ &\cdot [g''(u+x)+g''(v+y)] \\ &- [g''(u)+g''(v)] \cdot [f''(x)+f''(y)] \\ \geq &D(f)+D(g) > 0. \end{aligned} \tag{20}$$

So $f+g$ satisfies (11), and thus satisfies (9). Since for any positive integer m, the function $\frac{g}{m}$ satisfies the same conditions as g does, (9) also holds for $f+\frac{g}{m}$. Taking the limit

$$\lim_{m \to \infty} f + \frac{g}{m} = f$$

we see that (9) holds for f. Now the result is established. □

Theorem I.1: If a function $f : [0,1] \longrightarrow \mathbb{R}$ satisfies all the hypotheses in Lemma I.2, f is additivity-concave.

Proof: Directly by Lemma I.1 and Lemma I.2. □

Corollary I.1: Let $f : [0,1] \longrightarrow \mathbb{R}$ be a continuous function, $f(0) \leq 0$, the third derivative f''' exists in $(0,1)$, $f''(x) \leq 0$ and $f'''(x) \leq 0$ for any $x \in (0,1)$. Then f is additivity-concave.

Proof: If $f''' \leq 0$, then f'' is a decreasing function. Combining $f'' \leq 0$, the inequality (10) is easily verified. By Lemma I.2, f is additivity-concave. □

ACKNOWLEDGMENT

The authors wish to acknowledge the careful and judicious review of the anonymous reviewers and the editor, which helped them to greatly improve the presentation and substance of this correspondence. Ping Luo would like to acknowledge the judicious observations of Dr. Shiwei Ye from the School of Information Science and Engineering, Graduate School of the Chinese Academy of Sciences. Kevin Lü would like to express his appreciation to the Wang Kuan Cheng Science Foundation, Chinese Academy of Sciences, for the funding that enabled him to conduct this research.

REFERENCES

[1] R. V. Belavkin and F. E. Ritter, "The use of entropy for analysis and control of cognitive models," in *Proc. Fifth Int. Conf. Cognitive Modeling*, 2003, pp. 21–26.

[2] D. A. Simovici and S. Jaroszewicz, "An axiomatization of partition entropy," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 2138–2142, Jul. 2002.

[3] J. R. Quinlan, *C4.5: Programs for Machine Learning*. New York: Morgan Kaufmann, 1993.

[4] G. Wang, "Rough reduction in algebra view and information view," *Int. J. Intell. Syst.*, vol. 18, no. 6, pp. 679–688, 2003.

[5] D. Morales, L. Pardo, and I. Vajda, "Uncertainty of discrete stochastic systems: General theory and statistical inference," *IEEE Trans. Syst., Man, Cybern.*, vol. 26, no. 6, pp. 681–697, 1996.

[6] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*. New York: Academic, 1979.

[7] F. Cicalese and U. Vaccaro, "Supermodularity and subadditivity properties of the entropy on the majorization lattice," *IEEE Trans. Inf. Theory*, vol. 48, no. 4, pp. 933–938, Apr. 2002.

[8] N. R. Pal and S. K. Pal, "Entropy: A new definition and its applications," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 5, pp. 1260–1270, 1991.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.

[10] D. A. Simovici and S. Jaroszewicz, "Generalized entropy and decision trees," *EGC-Journes Francophones d'Extraction et de Gestion des Connaissances*, 2003.

[11] D. A. Simovici and S. Jaroszewicz, "A metric approach to building decision trees based on goodman-kruskal association index," in *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining*, 2004, pp. 181–190.

[12] S. Jaroszewicz, D. A. Simovici, W. Kuo, and L. Ohno-Machado, "The goodman-kruskal coefficient and its applications in the genetic diagnosis of cancer," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1095–1102, Jul. 2004.

[13] J. Liang, K. S. Chin, C. Dang, and R. C. M. Yam, "A new method for measuring uncertainty and fuzziness in rough set theory," *Int. J. General Syst.*, vol. 31, no. 4, pp. 331–342, Jul. 2002.

[14] P. Luo, Q. He, and Z. Shi, "Theoretical study on a new information entropy and its use in attribute reduction," in *Proc. IEEE Int. Conf. Cognitive Information*, 2005.

Application of Tauberian Theorem to the Exponential Decay of the Tail Probability of a Random Variable

Kenji Nakagawa, *Member, IEEE*

***Abstract*—In this correspondence, we give a sufficient condition for the exponential decay of the tail probability of a nonnegative random variable. We consider the Laplace–Stieltjes transform of the probability distribution function of the random variable. We present a theorem, according to which if the abscissa of convergence of the LS transform is negative finite and the real point on the axis of convergence is a pole of the LS transform, then the tail probability decays exponentially. For the proof of the theorem, we extend and apply so-called a finite form of Ikehara's complex Tauberian theorem by Graham–Vaaler.**

***Index Terms*—Complex Tauberian theorem, exponential decay, Graham–Vaaler's finite form, Laplace transform, tail probability of random variable.**

I. INTRODUCTION

The purpose of this correspondence is to give a sufficient condition for the exponential decay of the tail probability of a nonnegative random variable. For a nonnegative random variable X, $P(X < x)$ is called the *tail probability* of X. The tail probability *decays exponentially* if the limit

$$\lim_{x \to \infty} \frac{1}{x} \log P(X > x) \tag{1}$$

exists and is a negative finite value.

For the random variable X, the probability distribution function of X is denoted by $F(x) = P(X \leq x)$ and the Laplace–Stieltjes transform of $F(x)$ is denoted by $\varphi(s) = \int_0^\infty e^{-sx} dF(x)$. We will give a sufficient condition for the exponential decay of the tail probability $P(X > x)$ based on analytic properties of $\varphi(s)$.

In [11], we obtained a result that the exponential decay of the tail probability $P(X > x)$ is determined by the singularities of $\varphi(s)$ on its axis of convergence. In this correspondence, we investigate the case where $\varphi(s)$ has a pole at the real point of the axis of convergence, and reveal the relation between analytic properties of $\varphi(s)$ and the exponential decay of $P(X > x)$.

The results obtained in this correspondence will be applied to queueing analysis. In general, there are two main performance measures of queueing analysis, one is the number of customers Q in the system and the other is the sojourn time W in the system. Q is a discrete random variable and W is a continuous one. It is important to evaluate the tail probabilities $P(Q > q)$ and $P(W > w)$ for designing the buffer size or link capacity in communication networks. Even in the case that the probability distribution functions $P(Q \leq q)$ or $P(W \leq w)$ cannot be calculated explicitly, their generating functions $Q(z) = \sum_{q=0}^{\infty} P(Q = q) z^q$ or $W(s) = \int_0^\infty e^{-sw} dP(W \leq w)$ can be obtained explicitly in many queues. Particularly, in *M*/*G*/1 queue,

Manuscript received January 20, 2002; revised November 18, 2002. The material in this correspondence was presented at International Symposium on Information Theory and Its Applications (ISITA2006), Seoul, Korea, October 2006.

The author is with the Department of Electrical Engineering, Nagaoka University of Technology, Nagaoka, Niigata 940-2188, Japan (e-mail: nakagawa@nagaokaut.ac.jp).

Communicated by X. Wang, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2007.903114