IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024TAP0010

Publicized:2024/08/16

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER Special Section on Information Theory and Its Applications

A Variational Characterization of H-Mutual Information and its Application to Computing H-Capacity

Akira KAMATSUKA[†a)], ***Member***, **Koki KAZAMA**[†b)], ***Nonmember***, ***and*** **Takahiro YOSHIDA**[††c)], ***Senior Member***

SUMMARY H-mutual information (H-MI) is a wide class of information leakage measures, where $H = (\eta, F)$ is a pair of monotonically increasing function η and a concave function F, which is a generalization of Shannon entropy. H-MI is defined as the difference between the generalized entropy H and its conditional version, including Shannon mutual information (MI), Arimoto MI of order α, g-leakage, and expected value of sample information. This study presents a variational characterization of H-MI via statistical decision theory. Based on the characterization, we propose an alternating optimization algorithm for computing H-capacity.
key words: *H-mutual information, Arimoto–Blahut algorithm, statistical decision theory, value of information*

1. Introduction

Shannon mutual information (MI) $I(X;Y)$ [1] is a typical quantity that quantifies the amount of information a random variable Y contains about a random variable X. Several ways to generalize the Shannon MI are available in literature. A well-known generalization of Shannon MI is a class of α-mutual information (α-MI) $I_{\alpha}^{(\cdot)}(X;Y)$ [2], where $\alpha \in (0,1) \cup (1,\infty)$ is a tunable parameter. The α-MI class includes Sibson MI $I_{\alpha}^{\mathrm{S}}(X;Y)$ [3], Arimoto MI $I_{\alpha}^{\mathrm{A}}(X;Y)$ [4], and Csiszár MI $I_{\alpha}^{\mathrm{C}}(X;Y)$ [5]. These MIs share common properties such as non-negativity and data-processing inequality (DPI).

In problems on information security, Shannon MI can be interpreted as a measure of information leakage, i.e., a measure of how much information observed data Y leak about secret data X. Recently, various operationally meaningful leakage measures were proposed for privacy-guaranteed data-publishing problems. For example, Calmon and Fawaz introduced the *average cost gain* [6] and Issa *et al.* introduced the *maximal leakage*. Extending the maximal leakage, Liao *et al.* introduced *α-leakage* and *maximal α-leakage* [7]. Alvim *et al.* proposed *g-leakage* [8–10], a rich class of information leakage measures; g-leakage was extended to *maximal g-leakage* by Kurri *et al.* [11]. Note that these information leakage measures are based on the adversary's decision-making on X from the observed data Y and a gain (utility) or loss (cost) function.

Research on quantifying leaked information from the observed data Y based on a decision-making problem can be traced back to the 1960s. In a pioneering work by Raiffa and Schlaifer on quantifying the *value of information* (VoI) [12], the *expected value of sample information* (EVSI) was formulated in a statistical decision-theoretic framework. EVSI was defined as the largest increase in maximal Bayes expected gain (or the largest reduction of minimal Bayes risk) compared to those without using Y. Thus, information leakage measures in the information disclosure problem can be interpreted as variants of EVSI.

Recently, Américo *et al.* proposed a wide class of information leakage measures, referred to as H-mutual information (H-MI) $I_H(X;Y)$ [13, 14]. Here, $H = (\eta, F)$ is a pair of a continuous real-valued function $F\colon \Delta_{\mathcal{X}} \to \mathbb{R}$ and a continuous and strictly increasing function $\eta\colon F(\Delta_{\mathcal{X}}) \to \mathbb{R}$, where $\Delta_{\mathcal{X}}$ is a probability simplex on a finite set $\mathcal{X}$ and $F(\Delta_{\mathcal{X}})$ is the image of F. When η is an identity map and $F(p_X) := -\sum_x p_X(x) \log p_X(x)$, $H = (\eta, F)$ represents the Shannon entropy $S(X)$. Thus $H = (\eta, F)$ can be regarded as a generalized entropy. H-MI is defined as the difference between the generalized entropy $H = (\eta, F)$ and its conditional version $H(X|Y)$, which includes Shannon MI, Arimoto MI of order α, g-leakage, and EVSI. In [13, 14], Américo *et al.* provided the necessary and sufficient conditions (referred to as *core-concavity* (CCV) condition) for $I_H(X;Y)$ to satisfy non-negativity and DPI when the conditional entropy $H(X|Y)$ satisfies the *η-averaging* (EAVG) condition.

In this study, we present a variational characterization of H-MI that satisfies DPI via statistical decision theory. Our variational characterization transforms H-MI into the following optimization problem:

$$I_H(X;Y) = \max_{q_{X|Y}} \mathcal{F}_H(p_X, q_{X|Y}), \tag{1}$$

where $p_X \in \Delta_{\mathcal{X}}$ is a distribution on X and $q_{X|Y} = \{q_{X|Y}(\cdot \mid y)\}_{y \in \mathcal{Y}}$ is a set of conditional distributions of X, given $Y = y$. This variational characterization allows us to derive an alternating optimization algorithm (also known as Arimoto–Blahut algorithm [15], [16]) for computing H-capacity $C_H := \max_{p_X} I_H(X;Y)$, such as the channel capacity $C := \max_{p_X} I(X;Y)$ and Arimoto capacity $C_{\alpha}^{\mathrm{A}} := \max_{p_X} I_{\alpha}^{\mathrm{A}}(X;Y)$ [†] [4, 17], [18].

[†]The authors are with the Department of Information Science, Faculty of Engineering, Shonan Institute of Technology, Tsujido-Nishikaigan, Fujisawa, 251–8511, Japan.
[††]The author is with the Department of Business Administration, Nihon University, 5–2–1 Kinuta, Setagaya-ku, Tokyo 152–8570, Japan.
a) E-mail: kamatsuka@info.shonan-it.ac.jp
b) E-mail: kazama@info.shonan-it.ac.jp
c) E-mail: yoshida.takahiro@nihon-u.ac.jp

[†]It is worth mentioning that Liao *et al.* reported the operational

1.1 Main Contributions

The main contributions of this study are as follows:

- We provide a variational characterization of H-MI (Theorem 2) using the fact that every concave function F has a statistical decision-theoretic variational characterization [19, Section 3.5.4].
- On the basis of variational characterization, we build an alternating optimization algorithm for calculating H-capacity $C_H := \max_{p_X} I_H(X;Y) = \max_{p_X} \max_{q_{X|Y}} \mathcal{F}_H(p_X, q_{X|Y})$ (Algorithm 1) (see Section 4). Moreover, we show that the algorithms for computing Arimoto capacity C_α^{A} derived from our approach coincide with the previous algorithms reported in [17], [18].

1.2 Organization of the Paper

The remainder of this paper is organized as follows. We review the statistical decision theory and H-MI in Section 2. In Section 3, we present the variational characterization of H-MI. In Section 4, we derive an alternating optimization algorithm for computing H-capacity $C_H := \max_{p_X} I_H(X;Y)$ based on the characterization.

2. Preliminaries

2.1 Notations

Let X, Y be random variables on finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, drawn according to a joint distribution $p_{X,Y} = p_X p_{Y|X}$. Let p_Y be a marginal distribution of Y and $p_{X|Y}(\cdot|y) := \frac{p_X(\cdot)p_{Y|X}(y|\cdot)}{\sum_x p_X(x)p_{Y|X}(y|x)}$ be a posterior distribution on X given $Y = y$, respectively. The set of all distributions p_X is denoted as $\Delta_{\mathcal{X}}$. We often identify $\Delta_{\mathcal{X}}$ with $(m-1)$-dimensional probability simplex $\left\{(p_1, \ldots, p_m) \in [0,1]^m \,\middle|\, \sum_{i=1}^m p_i = 1\right\}$, where $m := |\mathcal{X}|$. Given a function $f\colon \mathcal{X} \to \mathbb{R}$, we use $\mathbb{E}_X[f(X)] := \sum_x f(x)p_X(x)$ and $\mathbb{E}_X[f(X)|Y=y] := \sum_x f(x)p_{X|Y}(x|y)$ to denote expectation on $f(X)$ and conditional expectation on $f(X)$ given $Y = y$, respectively. We also use $\mathbb{E}_X^{p_X}[f(X)]$ to emphasize that we are taking expectations p_X. We use $S(X)$, $S(X|Y)$, $I(X;Y) := S(X) - S(X|Y)^\dagger$, and $D(p||q)$ to denote Shannon entropy, conditional entropy, Shannon MI, and relative entropy, respectively. Let $\mathcal{A}$ be an action space (decision space) and $\delta\colon \mathcal{Y} \to \mathcal{A}$ be a decision rule for a decision maker (DM). Let $A := \delta(Y)$ be an action (decision) of the DM. We use $\ell(x,a)$ and $g(x,a)$ to denote the loss (cost) function and gain (utility) function of the DM, respectively. Throughout this paper, we use log to denote the natural logarithm and $\|p_X\|_p := (\sum_x p_X(x)^p)^{\frac{1}{p}}$ represents the p-norm of $p_X \in \Delta_{\mathcal{X}}$.

We initially review statistical decision theory [20] and H-MI [13, 14].

2.2 Statistical Decision Theory and Scoring Rules

In this subsection, we review statistical decision theory. In particular, we review a problem of deciding the optimal probability mass function (pmf) considering a loss or a gain function (referred to as a *scoring rule*), which is historically known as a *probability forecasting* problem.

Suppose that a DM makes action $A \in \mathcal{A}$ from observed data $Y \in \mathcal{Y}$ using a decision rule $\delta\colon \mathcal{Y} \to \mathcal{A}$. We assume that the DM uses the decision rule δ^* that minimizes Bayes risk $r(\delta) := \mathbb{E}_{X,Y}[\ell(X, \delta(Y))]$ (or maximizes Bayes expected gain $G(\delta) := \mathbb{E}_{X,Y}[g(X, \delta(Y))]$). Figure 1 shows the system model for this problem.

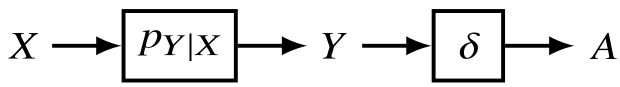

Fig. 1 System model of the statistical decision theory

Proposition 1 ([20, Result 1], [21, Thm 2.7]): The minimal Bayes risk is given by

$$\min_\delta r(\delta) = r(\delta^*) \tag{2}$$

$$= \mathbb{E}_Y\left[\min_{a\in\mathcal{A}} \mathbb{E}_X[\ell(X,a) \mid Y]\right] \tag{3}$$

$$= \sum_y p_Y(y)\left[\min_{a\in\mathcal{A}} \sum_x \ell(x,a)p_{X|Y}(x \mid y)\right], \tag{4}$$

with the optimal decision rule $\delta^*\colon \mathcal{Y} \to \mathcal{A}$ given by

$$\delta^*(y) := \operatorname*{argmin}_{a\in\mathcal{A}} \mathbb{E}_X[\ell(X,a) \mid Y = y]. \tag{5}$$

Similarly, the maximal Bayes expected gain and the optimal decision rule $\delta^*\colon \mathcal{Y} \to \mathcal{A}$ are given by

$$\max_\delta G(\delta) = G(\delta^*) \tag{6}$$

$$= \mathbb{E}_Y\left[\max_{a\in\mathcal{A}} \mathbb{E}_X[g(X,a) \mid Y]\right], \tag{7}$$

$$\delta^*(y) := \operatorname*{argmax}_{a\in\mathcal{A}} \mathbb{E}_X[g(X,a) \mid Y = y]. \tag{8}$$

Remark 1: Let $\ell(x,a)$ be a loss function. Let us define a gain function $g(x,a) := c\ell(x,a) + d$, where $c < 0$ and d are constants. One can easily see that if δ^* minimize Bayes risk $r(\delta) := \mathbb{E}_{X,Y}[\ell(X,\delta(Y))]$ then the rule δ^* also maximizes the Bayes expected gain $G(\delta) := \mathbb{E}_{X,Y}[g(X,\delta(Y))]$. The reverse is also true.

meaning of Arimoto capacity and Sibson capacity in the privacy-guaranteed data-publishing problems [7, Thm 2]; these capacities are essentially equivalent to the maximal α-leakage.

†Note that, throughout this paper, the notations $H(X)$ and $H(X|Y)$ are used to denote generalized forms of entropy and conditional entropy introduced in Definitions 2 and 4.

Example 1: Let $\hat{X}$ be an estimator of X. Suppose that a DM conducts a point estimation on X, i.e., $A = \hat{X} \in \mathcal{X}$ considering 0-1 loss $\ell_{0\text{-}1}(x, \hat{x}) = \mathbb{1}_{\{x=\hat{x}\}}$, where $\mathbb{1}_{\{\cdot\}}$ is an indicator function. Then the minimal Bayes risk and the optimal decision rule δ^* are given as follows:

$$\min_{\delta} r(\delta) = 1 - \mathbb{E}_Y\left[\max_x p_{X|Y}(x \mid Y)\right], \tag{9}$$

$$\delta^*(y) = \operatorname*{argmax}_x p_{X|Y}(x \mid y). \quad \text{(MAP estimation)} \tag{10}$$

Example 2: Suppose that a DM decides the optimal pmf $q \in \mathcal{A} = \Delta_{\mathcal{X}}$ considering log-score $g_{\log}(x, q) := \log q(x)$ [22]. Then, the maximal Bayes expected gain and the optimal decision rule are given as

$$\min_{\delta} r(\delta) = S(X \mid Y), \tag{11}$$

$$\delta^*(y) = p_{X|Y}(\cdot \mid y), \tag{12}$$

where $S(X|Y) = -\sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log p_{X|Y}(x|y)$ is the conditional entropy.

Remark 2: Historically, the problem of deciding the optimal pmf $q \in \Delta_{\mathcal{X}}$ considering a loss $\ell(x, q)$ or a gain $g(x, q)$ is called a *probability forecasting* problem [23], [24]. In the problem, the loss or gain function is called the *scoring rule*.

Remark 3: Note that finding the optimal decision rule $\delta\colon \mathcal{Y} \to \Delta_{\mathcal{X}}$ that minimizes $r(\delta)$ (*resp.* maximizes $G(\delta)$) is equivalent to finding the optimal set of conditional distributions $q_{X|Y} = \{q_{X|Y}(\cdot \mid y)\}_{y\in\mathcal{Y}}$ that minimizes $r(q_{X|Y}) := \mathbb{E}_{X,Y}\left[\ell(X, q_{X|Y}(X \mid Y))\right]$ (*resp.* maximizes $G(q_{X|Y}) := \mathbb{E}_{X,Y}\left[g(X, q_{X|Y}(X \mid Y))\right]$). Thus we call $r(q_{X|Y})$ (*resp.* $G(q_{X|Y})$) as Bayes risk (*resp.* Bayes expected gain) for $q_{X|Y}$ and denote the optimal set of conditional distribution as $q^*_{X|Y}$.

Example 3: Besides the log-score $g_{\log}(x, q)$ in Example 2, there exist other scoring rules that give the same optimal set of conditional distribution $q^*_{X|Y}$. Some examples are shown below:

- $g_{\text{PS}}(x, q) := \frac{1}{\alpha-1}\left(\frac{q(x)}{\|q\|_\alpha}\right)^{\alpha-1}$ (the *pseudo-spherical score* [25])
- $g_{\text{Power}}(x, q) := \frac{\alpha}{\alpha-1} \cdot q(x)^{\alpha-1} - \|q\|_\alpha^\alpha$ (the *power score* [26] (also known as *Tsallis score* [24]))

†

Note that the log-score $g_{\log}(x, q)$, pseudo-spherical score $g_{\text{PS}}(x, q)$, and power score $g_{\text{Power}}(x, q)$ are all *proper scoring rules* (PSR) defined as follows.

Definition 1: The scoring rule $g(x, q)$ is *proper* if for all $q \in \Delta_{\mathcal{X}}$,

$$\mathbb{E}_X^{p_X}\left[g(X, p_X)\right] \geq \mathbb{E}_X^{p_X}\left[g(X, q)\right]. \tag{13}$$

If the equality holds if and only if $q = p_X$, then the scoring rule $g(x, q)$ is called *strictly proper*††.

Example 4: Recently, Liao *et al.* proposed *α-loss* $\ell_\alpha(x, q) := \frac{\alpha}{\alpha-1}\left(1 - q(x)^{\frac{\alpha-1}{\alpha}}\right)$ [7, Def 3] in the privacy-guaranteed data-publishing context. In [7, Lemma 1], they proved that

$$\operatorname*{argmin}_q \mathbb{E}_X^{p_X}\left[\ell_\alpha(X, q)\right] = p_{X_\alpha}, \tag{14}$$

where p_{X_α} is the *α-tilted distribution of p_X* (also known as *scaled distribution [2] and escort distribution [27]*) defined as follows:

$$p_{X_\alpha}(x) := \frac{p_X(x)^\alpha}{\sum_x p_X(x)^\alpha}. \tag{15}$$

Thus, α-loss $\ell_\alpha(x, q)$ can be regard as a scoring rule that is *not* proper.

Table 1 summarizes examples of scoring rules described above, their optimal values, and the optimal set of conditional distributions $q^*_{X|Y}$.

2.3 H-Mutual information (H-MI) [13, 14]

In this subsection, we review H-MI and show that H-MI includes well-known information leakage measures.

Definition 2 ([13, Def. 11]): Let p_X be a pmf of X, $F\colon \Delta_{\mathcal{X}} \to \mathbb{R}$ and $\eta\colon F(\Delta_{\mathcal{X}}) \to \mathbb{R}$ be continuous functions, and η be strictly increasing. Given $H = (\eta, F)$, the *unconditional form of entropy* is defined as follows:

$$H(X) := \eta(F(p_X)). \tag{16}$$

Definition 3 (CCV [13, Def. 12]): $H = (\eta, F)$ is *core-concave* (CCV) if F is concave. We say that $H(X)$ is core-concave entropy if $H = (\eta, F)$ is CCV.

Definition 4 (EAVG [13, Def. 13]): ††† Given a joint distribution $p_{X,Y} = p_X p_{Y|X}$ and $H = (\eta, F)$, a functional $H(p_X, p_{Y|X})$ satisfies *η-averaging* (EAVG) if it is represented as follows:

$$H(p_X, p_{Y|X}) = \eta\left(\mathbb{E}_Y^{p_Y}\left[F(p_{X|Y}(\cdot \mid Y))\right]\right) \tag{17}$$

$$= \eta\left(\sum_y p_Y(y) F(p_{X|Y}(\cdot \mid y))\right), \tag{18}$$

where $p_{X|Y}(x|y) := \frac{p_X(x)p_{Y|X}(y|x)}{\sum_x p_X(x)p_{Y|X}(y|x)}$ is the posterior distribution of X given $Y = y$ and $p_Y(y) := \sum_x p_X(x)p_{Y|X}(y|x)$ is the marginal distribution of Y. We say that $H(p_X, p_{Y|X})$ is conditional entropy of $H = (\eta, F)$ and it is denoted by $H(X|Y)$.

†The pseudo-spherical score and the power score are originally defined for $\alpha > 1$. We multiply the original definitions by $\frac{1}{\alpha-1}$ so that we can define them for $\alpha \in (0, 1) \cup (1, \infty)$.

††Similarly, we can define a (strictly) proper loss $\ell(x, q)$.

†††We slightly modified the definition of EAVG.

Table 1 Typical scoring rules for deciding $q \in \Delta_{\mathcal{X}}$ and the optimal decision rules

$\ell(x,q)$, $g(x,q)$	$\mathrm{argmin}_q \mathbb{E}_X[\ell(X,q)]$ $= \mathrm{argmax}_q \mathbb{E}_X[g(X,q)]$	$\min_q \mathbb{E}_X[\ell(X,q)]$, $\max_q \mathbb{E}_X[g(X,q)]$	$\mathrm{argmin}_{q_{X\|Y}} \mathbb{E}_{X,Y}\left[\ell(X,q_{X\|Y}(\cdot\|Y))\right]$ $= \mathrm{argmax}_{q_{X\|Y}} \mathbb{E}_{X,Y}\left[g(X,q_{X\|Y}(\cdot\|Y))\right]$	$\min_{q_{X\|Y}} \mathbb{E}_{X,Y}\left[\ell(X,q_{X\|Y}(\cdot\|Y))\right]$, $\max_{q_{X\|Y}} \mathbb{E}_{X,Y}\left[g(X,q_{X\|Y}(\cdot\|Y))\right]$
$-\log q(x)$ (log-loss), $\log q(x)$ (log-score [22])	p_X	$S(X)$, $-S(X)$	$p_{X\|Y}(\cdot\|y), y \in \mathcal{Y}$	$S(X\|Y)$, $-S(X\|Y)$
$\frac{1}{\alpha-1}\left(1-\left(\frac{q(x)}{\\|q\\|_\alpha}\right)^{\alpha-1}\right)$, $\frac{1}{\alpha-1}\cdot\left(\frac{q(x)}{\\|q\\|_\alpha}\right)^{\alpha-1}$ (pseudo-spherical score [25])	p_X	$\frac{1}{\alpha-1}(1-\\|p_X\\|_\alpha)$ (Harvda–Tsallis entropy), $\frac{1}{\alpha-1}\cdot\\|p_X\\|_\alpha$	$p_{X\|Y}(\cdot\|y), y \in \mathcal{Y}$	$\frac{1}{\alpha-1}\left(1-\mathbb{E}_Y\left[\\|p_{X\|Y}(\cdot\|Y)\\|_\alpha\right]\right)$, $\frac{1}{\alpha-1}\cdot\mathbb{E}_Y\left[\\|p_{X\|Y}(\cdot\|Y)\\|_\alpha\right]$
$\frac{\alpha}{\alpha-1}\left(1-q(x)^{\alpha-1}\right)+\\|q\\|_\alpha^\alpha$, $\frac{\alpha}{\alpha-1}\cdot q(x)^{\alpha-1}-\\|q\\|_\alpha^\alpha$ (power score [26], Tsallis score [24])	p_X	$\frac{1}{\alpha-1}(1-\\|p_X\\|_\alpha^\alpha)$, $\frac{1}{\alpha-1}\cdot\\|p_X\\|_\alpha^\alpha$	$p_{X\|Y}(\cdot\|y), y \in \mathcal{Y}$	$\frac{1}{\alpha-1}\left(1-\mathbb{E}_Y\left[\\|p_{X\|Y}(\cdot\|Y)\\|_\alpha^\alpha\right]\right)$, $\frac{1}{\alpha-1}\cdot\mathbb{E}_Y\left[\\|p_{X\|Y}(\cdot\|Y)\\|_\alpha^\alpha\right]$
$\frac{\alpha}{\alpha-1}\left(1-q(x)^{\frac{\alpha-1}{\alpha}}\right)$ (α-loss [7]), $\frac{\alpha}{\alpha-1}\cdot q(x)^{\frac{\alpha-1}{\alpha}}$ (α-score)	p_{X_α}	$\frac{\alpha}{\alpha-1}(1-\\|p_X\\|_\alpha)$, $\frac{\alpha}{\alpha-1}\cdot\\|p_X\\|_\alpha$	$p_{X_\alpha\|Y}(\cdot\|y), y \in \mathcal{Y}$	$\frac{\alpha}{\alpha-1}\left(1-\mathbb{E}_Y\left[\\|p_{X\|Y}(\cdot\mid Y)\\|_\alpha\right]\right)$, $\frac{\alpha}{\alpha-1}\cdot\mathbb{E}_Y\left[\\|p_{X\|Y}(\cdot\|Y)\\|_\alpha\right]$

Theorem 1 ([14, Thm. 2] and [13, Thm. 4]): Given $H = (\eta, F)$, H-MI is defined as

$$I_H(X;Y) := H(X) - H(X \mid Y), \tag{19}$$

where $H(X|Y)$ satisfies EAVG. Then, the following are equivalent†:

(CCV) $H = (\eta, F)$ is core-concave.
(Non-negativity) $I_H(X;Y) \geq 0$.
(DPI) If $X - Y - Z$ forms a Markov chain, then

$$I_H(X;Z) \leq I_H(X;Y). \tag{20}$$

Table 2 lists examples of H-MI, $H = (\eta, F)$, and $H(X|Y)$ described below that satisfy the conditions in Theorem 1 (For more examples, see [13,14], [28, Table I]).

Example 5: Let $\alpha \in (0,1) \cup (1,\infty)$. Shannon MI $I(X;Y) := S(X) - S(X|Y)$ and Arimoto MI $I_\alpha^{\mathrm{A}}(X;Y) := H_\alpha(X) - H_\alpha^{\mathrm{A}}(X \mid Y)$ are examples of H-MI, where

$$H_\alpha(X) := \frac{\alpha}{1-\alpha}\log\|p_X\|_\alpha = \frac{1}{1-\alpha}\log\|p_X\|_\alpha^\alpha \tag{21}$$

$$= -\log\|p_X\|_\alpha^{\frac{\alpha}{\alpha-1}}, \tag{22}$$

$$H_\alpha^{\mathrm{A}}(X \mid Y) := \frac{\alpha}{1-\alpha}\log\sum_y p_Y(y)\sum_x \left\|p_{X|Y}(\cdot \mid y)\right\|_\alpha \tag{23}$$

are the Rènyi entropy of order α and the Arimoto conditional entropy of order α [4], respectively.

As shown in Example 5, the Rènyi entropy $H_\alpha(X)$ can be represented in at least three different ways. The corresponding $H = (\eta, F)$ for these expressions are shown in Table 2. Thus, we can define novel MIs as follows:

Definition 5 (Hayashi MI, Fehr–Berens MI): Hayashi MI of order $\alpha \in (0,1) \cup (1,\infty)$ and Fehr–Berens MI of order $\alpha > 1$ are defined as follows:

$$I_\alpha^{\mathrm{H}}(X;Y) := H_\alpha(X) - H_\alpha^{\mathrm{H}}(X \mid Y), \tag{24}$$

$$I_\alpha^{\mathrm{FB}}(X;Y) := H_\alpha(X) - H_\alpha^{\mathrm{FB}}(X \mid Y), \tag{25}$$

where

$$H_\alpha^{\mathrm{H}}(X;Y) := \frac{1}{1-\alpha}\log\sum_y p_Y(y)\sum_x \left\|p_{X|Y}(\cdot|y)\right\|_\alpha^\alpha, \tag{26}$$

$$H_\alpha^{\mathrm{FB}}(X;Y) := -\log\sum_y p_Y(y)\left\|p_{X|Y}(\cdot|y)\right\|_\alpha^{\frac{\alpha}{\alpha-1}} \tag{27}$$

are the Hayashi conditional entropy of order α [29, Section II.A] and the Fehr–Berens conditional entropy of order α [30, Section III.E, 5)], respectively.

Since $H_\alpha^{\mathrm{A}}(X|Y) \geq H_\alpha^{\mathrm{H}}(X|Y)$ [31, Prop 1], it follows that Hayashi MI is greater than or equal to Arimoto MI.

Proposition 2: Let $\alpha \in (0,1) \cup (1,\infty)$.

$$I_\alpha^{\mathrm{A}}(X;Y) \leq I_\alpha^{\mathrm{H}}(X;Y). \tag{28}$$

The amount of information that the observed data Y contain about X can also be quantified using the framework of a decision-making problem. In the 1960s, the EVSI was proposed by Raiffa and Schaifer [12]. Recently, equivalents or variants of the EVSI have been proposed in the context of privacy-guaranteed data-publishing problems. For example, Calmon and Fawaz proposed average (cost) gain [6] and Alvim *et al.* proposed g-leakage [8–10].

Definition 6: Let $g(x,a)$ be a gain function. The EVSI [12], also known as *average gain* [6] and *additive g-leakage* [8–10], is defined as the largest increase in the maximal Bayes expected gain compared to those without using Y, i.e.,

$$\mathrm{EVSI}^g(X;Y) := \max_\delta G(\delta) - \max_a \mathbb{E}_X[g(X,a)] \tag{29}$$

$$= -\max_a \mathbb{E}_X[g(X,a)] - \mathbb{E}_Y\left[-\max_a \mathbb{E}_X[g(X,a) \mid Y]\right], \tag{30}$$

where the equality in (30) follows from Proposition 1. The EVSI can also be defined using a loss function $\ell(x,a)$ as the

†Note that the original statement of the theorem is stated in terms of conditional entropy $H(X|Y)$ instead of H-MI $I_H(X;Y)$.

Table 2 Examples of H-mutual information (H-MI)

Name of H-MI	$H(X)$	$\eta(t)$	$F(p_X)$	$H(X\|Y)$
Shannon MI $I(X;Y)$ [1]	$-\sum_x p_X(x)\log p_X(x)$	t	$-\sum_x p_X(x)\log p_X(x)$	$-\sum_y p_Y(y)\sum_x p_{X\|Y}(x\|y)\log p_{X\|Y}(x\|y)$
Arimoto MI $I_\alpha^{\mathrm{A}}(X;Y)$ [4]	$\frac{\alpha}{1-\alpha}\log\\|p_X\\|_\alpha$	$\begin{cases}\frac{\alpha}{1-\alpha}\log t, & 0<\alpha<1,\\ \frac{\alpha}{1-\alpha}\log(-t), & \alpha>1\end{cases}$	$\begin{cases}\\|p_X\\|_\alpha, & 0<\alpha<1,\\ -\\|p_X\\|_\alpha, & \alpha>1\end{cases}$	$\frac{\alpha}{1-\alpha}\log\sum_y p_Y(y)\sum_x \\|p_{X\|Y}(\cdot\|y)\\|_\alpha$
Hayashi MI $I_\alpha^{\mathrm{H}}(X;Y)$	$\frac{1}{1-\alpha}\log\\|p_X\\|_\alpha^\alpha$	$\begin{cases}\frac{1}{1-\alpha}\log t, & 0<\alpha<1,\\ \frac{1}{1-\alpha}\log(-t), & \alpha>1\end{cases}$	$\begin{cases}\\|p_X\\|_\alpha^\alpha, & 0<\alpha<1,\\ -\\|p_X\\|_\alpha^\alpha, & \alpha>1\end{cases}$	$\frac{1}{1-\alpha}\log\sum_y p_Y(y)\sum_x \\|p_{X\|Y}(\cdot\|y)\\|_\alpha^\alpha$
Fehr–Berens MI $I_\alpha^{\mathrm{FB}}(X;Y)$, $\alpha>1$	$-\log\\|p_X\\|_\alpha^{\frac{\alpha}{\alpha-1}}$	$-\log(-t)$	$-\\|p_X\\|_\alpha^{\frac{\alpha}{\alpha-1}}$	$-\log\sum_y p_Y(y)\\|p_{X\|Y}(\cdot\|y)\\|_\alpha^{\frac{\alpha}{\alpha-1}}$
$\mathrm{EVSI}^{(\cdot)}(X;Y)$ [12], [6], [8–10]	$\min_q \mathbb{E}_X[\ell(X,q)]$, $-\max_q \mathbb{E}_X[g(X,q)]$	t	$\min_q \mathbb{E}_X[\ell(X,q)]$, $-\max_q \mathbb{E}_X[g(X,q)]$	$\sum_y p_Y(y)\min_q \mathbb{E}_X[\ell(X,q)\mid Y=y]$, $-\sum_y p_Y(y)\max_q \mathbb{E}_X[g(X,q)\mid Y=y]$

largest reduction of the minimal Bayes risk compared with those without using Y, i.e.,

$$\mathrm{EVSI}^{\ell}(X;Y) := \min_a \mathbb{E}_X[\ell(X,a)] - \min_\delta r(\delta) \tag{31}$$

$$= \max_a \mathbb{E}_X[\ell(X,a)] - \mathbb{E}_Y\left[\min_a \mathbb{E}_X[\ell(X,a)\mid Y]\right]. \tag{32}$$

Example 6: Suppose that a DM decides a pmf $q \in \Delta_{\mathcal{X}}$ considering log-loss $\ell_{\log}(x,q) := -\log q(x)$ or log-score $g_{\log}(x,q) := \log q(x)$. From Example 2, we obtain

$$\mathrm{EVSI}^{\ell_{\log}}(X;Y) = \mathrm{EVSI}^{g_{\log}}(X;Y) = I(X;Y). \tag{33}$$

Instead of examining the differences between $G(\delta)$ and $\mathbb{E}_X[g(X,a)]$, one can quantify information leakage by examining their ratio. Alvim *et al.* proposed *multiplicative* g-leakage [8–10] as follows:

Definition 7 (multiplicative g-leakage): † Let $g(x,a)$ be a non-negative or non-positive gain function and $c(g)$ be a function of g such that its sign is equal to $\mathrm{sign}(g)$††. Then the *multiplicative g-leakage* is defined as the largest multiplicative increase of the maximal Bayes expected gain compared to those of without Y, i.e.,

$$\mathrm{MEVSI}^{g}(X;Y) := c(g)\log\frac{\max_\delta G(\delta)}{\max_a \mathbb{E}_X[g(X,a)]} \tag{34}$$

$$= c(g)\log\frac{\mathbb{E}_Y[\max_a \mathbb{E}_X[g(X,a)\mid Y]]}{\max_a \mathbb{E}_X[g(X,a)]}. \tag{35}$$

Similarly, we can define $\mathrm{MEVSI}^{\ell}(X;Y)$ using a loss function $\ell(x,a)$.

Example 7: Suppose that a DM decides a pmf $q \in \Delta_{\mathcal{X}}$ considering pseudo-spherical score $g_{\mathrm{PS}}(x,q) := \frac{1}{\alpha-1}\cdot\left(\frac{q(x)}{\|q\|_\alpha}\right)^{\alpha-1}$ or $g_\alpha(x,q) := \frac{\alpha}{\alpha-1}\cdot q(x)^{\frac{\alpha-1}{\alpha}}$ (referred to as α-score). Define $c(g_{\mathrm{PS}}) = c(g_\alpha) := \frac{\alpha}{\alpha-1}$. From Table 1, we obtain

$$\mathrm{MEVSI}^{g_{\mathrm{PS}}}(X;Y) = \mathrm{MEVSI}^{g_\alpha}(X;Y)$$
$$= I_\alpha^{\mathrm{A}}(X;Y). \quad \text{(Arimoto MI)} \tag{36}$$

Example 8: Suppose that a DM decides a pmf $q \in \Delta_{\mathcal{X}}$ considering a power score $g_{\mathrm{Power}}(x,q) := \frac{\alpha}{\alpha-1}\cdot q(x)^{\alpha-1} - \|q\|_\alpha^\alpha$. Define $c(g_{\mathrm{Power}}) := \frac{1}{\alpha-1}$. From Table 1, we obtain

$$\mathrm{MEVSI}^{g_{\mathrm{Power}}}(X;Y) = I_\alpha^{\mathrm{H}}(X;Y). \quad \text{(Hayashi MI)} \tag{37}$$

Note that we can easily show that $F(p_X) := -\mathbb{E}_X^{p_X}[g(X,a)]$ and $F(p_X) := \mathbb{E}_X^{p_X}[\ell(X,a)]$ are concave with respect to p_X and that $H(X|Y) := \mathbb{E}_Y[-\max_a \mathbb{E}_X[g(X,a)\mid Y]]$ and $H(X|Y) := \mathbb{E}_Y[\min_a \mathbb{E}_X[\ell(X,a)\mid Y]]$ satisfy the EAVG condition given in Definition 4 (see also [14, Sec V.F]). Thus, we obtain the following result.

Proposition 3 ([14, Sec V.F]): $\mathrm{EVSI}^{(\cdot)}(X;Y)$ and $\mathrm{MEVSI}^{(\cdot)}$ are members of H-MI.

Conversely, can we represent H-MI $I_H(X;Y)$ by a decision-theoretic quantity? In the next section, we will show that this is possible. Furthermore, we derive a variational characterization of H-MI using this representation.

3. Variational Characterization of H-MI

In this section, we provide a variational characterization of H-MI $I_H(X;Y)$ using the fact that every continuous concave function F has a statistical decision-theoretic variational characterization [19, Section 3.5.4].

Grünwald and Dawid showed that every concave function $F\colon \Delta_{\mathcal{X}} \to \mathbb{R}$ has the following variational characterization.

Proposition 4 ([19, Section 3.5.4]): Let $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ and $F\colon \Delta_{\mathcal{X}} \to \mathbb{R}$ be a continuous concave functions. Suppose that a DM decide a pmf $q \in \Delta_{\mathcal{X}} \subseteq [0,1]^m$ considering the following proper loss function $\ell_F(x,q)$ defined as

$$\ell_F(x,q) := F(q) + z^\top(\mathbb{1}^x - q), \tag{38}$$

where

- $\mathbb{1}^x$ is the m-dimensional vector having $\mathbb{1}_j^x = 1$ if $j = x$,

†We slightly modified the definition of the multiplicative g-leakage so that we can define it using non-positive gain function $g(x,a)$ by multiplying $c(g)$.

††$\mathrm{sign}(g) := 1$, if $g(x,a) \geq 0, \forall(x,a)$, -1; otherwise.

0; otherwise,

- $z \in \partial F(q) \subseteq \mathbb{R}^m$ is a subgradient in subdifferential of $F(q)$[†].

Then, the following holds:

$$F(p_X) = \min_q \mathbb{E}_X^{p_X}\left[\ell_F(X, q)\right], \tag{39}$$

where the minimum is achieved at $q = p_X$.

Example 9: Some examples of the proper loss function $\ell_F(x, q)$ in Proposition 4 are listed below:

- If $F(p_X) = -\sum_x p_X(x)\log p_X(x)$, then $\ell_F(x,q) = \ell_{\log}(x,q) = -g_{\log}(x,q) = -\log q(x)$.
- If $F(p_X) = \|p_X\|_\alpha, 0 < \alpha < 1$, then $\ell_F(x,q) = \left(\frac{q(x)}{\|q\|_\alpha}\right)^{\alpha-1} = (\alpha-1)g_{\mathrm{PS}}(x,q)$. If $F(p_X) = -\|p_X\|_\alpha, \alpha > 1$, then $\ell_F(x,q) = (1-\alpha)g_{\mathrm{PS}}(x,q)$.
- If $F(p_X) = \|p_X\|_\alpha^\alpha, 0 < \alpha < 1$, then $\ell_F(x,q) = \alpha q(x)^{\alpha-1} - (\alpha-1)\|q\|_\alpha^\alpha = (\alpha-1)g_{\mathrm{Power}}(x,q)$. If $F(p_X) = -\|p_X\|_\alpha^\alpha, \alpha > 1$, then $\ell_F(x,q) = (1-\alpha)g_{\mathrm{Power}}(x,q)$.
- If $F(p_X) = -\|p_X\|_\alpha^{\frac{\alpha}{\alpha-1}}, \alpha > 1$, then $\ell_F(x,q) = \|q\|_\alpha^{\alpha-1} - \frac{\alpha}{\alpha-1}(\|q\|_\alpha^\alpha - q(x)^{\alpha-1})$.

Using Proposition 4, we obtain the following variational characterization of H-MI.

Theorem 2 (Variational characterization of H-MI): Suppose that $H = (\eta, F)$ satisfies the CCV condition and $H(X \mid Y)$ satisfies the EAVG condition, respectively. Then, there exists a functional $\mathcal{F}_H(p_X, q_{X|Y})$ such that

$$I_H(X;Y) = \max_{q_{X|Y}} \mathcal{F}_H(p_X, q_{X|Y}). \tag{40}$$

Proof. From Proposition 4, there exists a proper loss function $\ell_F(x,q)$ such that $F(p_X) = \min_q \mathbb{E}_X^{p_X}\left[\ell_F(X,q)\right]$. Since $H(X|Y)$ satisfies EAVG, it can be written as

$$H(X \mid Y) = \eta\left(\mathbb{E}_Y\left[F(p_{X|Y}(\cdot \mid Y))\right]\right) \tag{41}$$

$$= \eta\left(\mathbb{E}_Y\left[\min_q \mathbb{E}_X^{p_{X|Y}(\cdot|Y)}\left[\ell_F(X,q)\right]\right]\right) \tag{42}$$

$$= \eta\left(\mathbb{E}_Y\left[\min_q \mathbb{E}_X\left[\ell_F(X,q) \mid Y\right]\right]\right) \tag{43}$$

$$\overset{(a)}{=} \eta\left(\min_{q_{X|Y}} \mathbb{E}_{X,Y}\left[\ell_F(X, q_{X|Y}(X \mid Y))\right]\right) \tag{44}$$

$$\overset{(b)}{=} \min_{q_{X|Y}} \eta\left(\mathbb{E}_{X,Y}\left[\ell_F(X, q_{X|Y}(X \mid Y))\right]\right), \tag{45}$$

where

- (a) follows from Proposition 1 and Remark 3,
- (b) follows from the assumption that η is strictly increasing.

Therefore, we obtain the following variational characterization of H-MI:

$$I_H(X;Y) := \eta(F(p_X)) - \eta\left(\mathbb{E}_Y\left[F(p_{X|Y}(X \mid Y))\right]\right) \tag{46}$$

$$= \eta(F(p_X)) - \min_{q_{X|Y}} \eta\left(\mathbb{E}_{X,Y}\left[\ell_F(X, q_{X|Y}(X \mid Y))\right]\right) \tag{47}$$

$$= \max_{q_{X|Y}} \underbrace{\left(\eta(F(p_X)) - \eta\left(\mathbb{E}_{X,Y}\left[\ell_F(X, q_{X|Y}(X \mid Y))\right]\right)\right)}_{=:\mathcal{F}_H(p_X, q_{X|Y})}. \tag{48}$$

□

Example 10: From Theorem 2 and Example 9 we obtain the variational characterization for specific H-MIs as follows:

$$I(X;Y) = \max_{q_{X|Y}} \mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[\log \frac{q_{X|Y}(X \mid Y)}{p_X(X)}\right], \tag{49}$$

$$I_\alpha^{\mathrm{A}}(X;Y) = \max_{q_{X|Y}} \frac{\alpha}{\alpha-1} \log \frac{\mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[\left(\frac{q_{X|Y}(X|Y)}{\|q_{X|Y}(\cdot|Y)\|_\alpha}\right)^{\alpha-1}\right]}{\|p_X\|_\alpha}, \tag{50}$$

$$I_\alpha^{\mathrm{H}}(X;Y) = \max_{q_{X|Y}} \frac{1}{\alpha-1} \times \log \frac{\mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[\alpha q_{X|Y}(X \mid Y)^{\alpha-1} - (\alpha-1)\left\|q_{X|Y}(\cdot \mid Y)\right\|_\alpha^\alpha\right]}{\|p_X\|_\alpha^\alpha}, \tag{51}$$

$$I_\alpha^{\mathrm{FB}}(X;Y) = \max_{q_{X|Y}} \log \frac{\mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[\ell^{\mathrm{FB}}(X, q_{X|Y}(X \mid Y))\right]}{\|p_X\|_\alpha^{\frac{\alpha}{\alpha-1}}}, \tag{52}$$

where $\ell^{\mathrm{FB}}(x,q) := \|q\|_\alpha^{\alpha-1} - \frac{\alpha}{\alpha-1}(\|q\|_\alpha^\alpha - q(x)^{\alpha-1})$.

Remark 4: From Example 7, we obtain another variational characterization with $\ell_F(x,q) = -g_\alpha(x,q)$ that is *not* proper as follows:

$$I_\alpha^{\mathrm{A}}(X;Y) = \max_{q_{X|Y}} \frac{\alpha}{\alpha-1} \log \frac{\mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[q_{X|Y}(X \mid Y)^{\frac{\alpha-1}{\alpha}}\right]}{\|p_X\|_\alpha}. \tag{53}$$

4. Application: Deriving Algorithm For Computing H-Capacity

In information theory, the notion of capacity often characterizes the theoretical limits of performance in the problem. For example, channel capacity $C := \max_{p_X} I(X;Y)$ characterizes supremum of achievable rate in channel coding [1]. Recently, Liao *et al.* reported the operational

[†]Note that if F is differentiable, then the subdifferential $\partial F(q)$ is singleton, i.e., $\partial F(q) = \{\nabla F(q)\}$, where $\nabla F(q)$ is the gradient of $F(q)$.

meaning of Arimoto capacity $C_\alpha^{\mathrm{A}} := \max_{p_X} I_\alpha^{\mathrm{A}}(X;Y)$ in the privacy-guaranteed data-publishing problems [7, Thm 2]. The Arimoto–Blahut algorithm (ABA), which is a well-known alternating optimization algorithm for computing capacity C, proposed by Arimoto [15] and Blahut [16]. Extending his results, Arimoto derived an ABA for computing Arimoto capacity C_α^{A} in [17]. Recently, we derived another ABA for computing C_α^{A} using a variational characterization of $I_\alpha^{\mathrm{A}}(X;Y)$ different from Arimoto's method [18]. These algorithms are based on a double maximization problem using the variational characterization of MIs. In this section, we derive an alternating optimization algorithm for computing H-capacity $C_H := \max_{p_X} I_H(X;Y)$ based on the variational characterization of H-MI and ABA. Moreover, we show that the algorithms for computing Arimoto capacity C_α^{A} from our approach coincide with the previous algorithms [17], [18].

From Theorem 2, H-capacity $C_H := \max_{p_X} I_H(X;Y)$ can be represented as a double maximization problem as follows:

$$C_H = \max_{p_X} \max_{q_{X|Y}} \mathcal{F}_H(p_X, q_{X|Y}), \tag{54}$$

where

$$\begin{aligned}&\mathcal{F}_H(p_X, q_{X|Y}) \\ &:= \left(\eta(F(p_X)) - \eta\left(\mathbb{E}_{X,Y}\left[\ell_F(X, q_{X|Y}(X \mid Y))\right]\right)\right).\end{aligned} \tag{55}$$

Based on the representation in (54), we can derive an alternating optimization algorithm for computing C_H as described in Algorithm 1, where $p_X^{(0)}$ is an initial distribution of the algorithm..

Algorithm 1 Arimoto–Blahut algorithm for computing C_H

Input:
  $p_X^{(0)}, p_{Y|X}, \epsilon \in (0,1)$
Output:
  approximation of C_H
1: **Initialization:**
  $q_{X|Y}^{(0)} \leftarrow \mathrm{argmax}_{q_{X|Y}} \mathcal{F}_H(p_X^{(0)}, q_{X|Y})$
  $F^{(0,0)} \leftarrow \mathcal{F}_H(p_X^{(0)}, q_{X|Y}^{(0)})$
  $k \leftarrow 0$
2: **repeat**
3:   $k \leftarrow k+1$
4:   $p_X^{(k)} \leftarrow \mathrm{argmax}_{p_X} \mathcal{F}_H(p_X, q_{X|Y}^{(k-1)})$
5:   $q_{X|Y}^{(k)} \leftarrow \mathrm{argmax}_{q_{X|Y}} \mathcal{F}_H(p_X^{(k)}, q_{X|Y})$
6:   $F^{(k,k)} \leftarrow \mathcal{F}_H(p_X^{(k)}, q_{X|Y}^{(k)})$
7: **until** $\left|F^{(k,k)} - F^{(k-1,k-1)}\right| < \epsilon$
8: **return** $F^{(k,k)}$

From Propositions 1 and 4, the optimum $q_{X|Y}^* = \mathrm{argmax}_{q_{X|Y}} \mathcal{F}_H(p_X, q_{X|Y})$ for a fixed p_X is obtained as follows.

Proposition 5: For a fixed p_X, $\mathcal{F}_H(p_X, q_{X|Y})$ is maximized by

$$q_{X|Y}^*(x \mid y) = p_{X|Y}(x \mid y) = \frac{p_X(x) p_{Y|X}(y \mid x)}{\sum_x p_X(x) p_{Y|X}(y \mid x)}. \tag{56}$$

Proof. It can be easily checked that finding the optimum $q_{X|Y}^* = \mathrm{argmax}_{q_{X|Y}} \mathcal{F}_H(p_X, q_{X|Y})$ for fixed p_X is equivalent to finding the optimum $q_{X|Y}^* = \mathrm{argmin}_{q_{X|Y}} \mathbb{E}_{X,Y}\left[\ell_F(X, q_{X|Y}(X \mid Y))\right]$. From Proposition 1, the problem of finding $q_{X|Y} = \{q_{X|Y}(\cdot|y)\}_{y\in\mathcal{Y}}$ that minimizes $\mathbb{E}_{X,Y}\left[\ell_F(X, q_{X|Y}(X \mid Y))\right]$ becomes equivalent to the problem of finding the optimal conditional distribution $q_{X|Y}(\cdot|y)$ for each $y \in \mathcal{Y}$ that minimizes $\mathbb{E}_X\left[\ell(X, q_{X|Y}(\cdot|y)) \mid Y = y\right] = \mathbb{E}_X^{p_{X|Y}(\cdot|y)}\left[\ell(X, q_{X|Y}(\cdot|y))\right]$. Since $\ell_F(x,q)$ defined in (38) is proper, the optimum is obtained as $q_{X|Y}^*(\cdot|y) = p_{X|Y}(\cdot|y), y \in \mathcal{Y}$. □

Remark 5: On the other hand, whether the optimum $p_X^* = \mathrm{argmax}_{p_X} \mathcal{F}_H(p_X, q_{X|Y})$ for a fixed $q_{X|Y}$ can be obtained explicitly depends on $H = (\eta, F)$. For example, Arimoto [15] and Blahut [16] derived the explicit formula for p_X^*, where $\mathcal{F}(p_X, q_{X|Y}) := \mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[\log \frac{q_{X|Y}(X|Y)}{p_X(X)}\right]$ is defined in (49). Table 3 lists the explicit updating formulae for computing channel capacity C. However, when computing Hayashi capacity $C_\alpha^{\mathrm{H}} := \max_{p_X} I_\alpha^{\mathrm{H}}(X;Y)$ and Fehr–Berens capacity $C_\alpha^{\mathrm{FB}} := \max_{p_X} I_\alpha^{\mathrm{FB}}(X;Y)$, it seems that there is no explicit updating formula for p_X^* for a fixed $q_{X|Y}$. Therefore, one must find it numerically.

Next, we consider driving the algorithms for computing the Arimoto capacity C_α^{A}. Based on the variational characterizations (53) and (50), we define functionals $\mathcal{F}_\alpha{}^{\mathrm{A1}}(p_X, q_{X|Y})$ and $\mathcal{F}_\alpha{}^{\mathrm{A2}}(p_X, q_{X|Y})$ as follows:

$$\mathcal{F}_\alpha{}^{\mathrm{A1}}(p_X, q_{X|Y}) := \frac{\alpha}{\alpha-1} \log \frac{\mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[q_{X|Y}(X \mid Y)^{\frac{\alpha-1}{\alpha}}\right]}{\|p_X\|_\alpha}, \tag{57}$$

$$\mathcal{F}_\alpha{}^{\mathrm{A2}}(p_X, q_{X|Y}) := \frac{\alpha}{\alpha-1} \log \frac{\mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[\left(\frac{q_{X|Y}(X|Y)}{\|q_{X|Y}(\cdot|Y)\|_\alpha}\right)^{\alpha-1}\right]}{\|p_X\|_\alpha}. \tag{58}$$

Simple calculations yield the following result.

Proposition 6:

$$\begin{aligned}&\mathcal{F}_\alpha{}^{\mathrm{A1}}(p_X, q_{X|Y}) \\ &= \frac{\alpha}{\alpha-1} \log \sum_{x,y} p_{X_\alpha}(x)^{\frac{1}{\alpha}} p_{Y|X}(y|x) q_{X|Y}(x|y)^{\frac{\alpha-1}{\alpha}},\end{aligned} \tag{59}$$

$$\begin{aligned}&\mathcal{F}_\alpha{}^{\mathrm{A2}}(p_X, q_{X|Y}) \\ &= \frac{\alpha}{\alpha-1} \log \sum_{x,y} p_{X_\alpha}(x)^{\frac{1}{\alpha}} p_{Y|X}(y|x) q_{X_\alpha|Y}(x|y)^{\frac{\alpha-1}{\alpha}},\end{aligned} \tag{60}$$

where p_{X_α} is the α-tilted distribution of p_X defined in (15)

Table 3 Formulae for updating $p_X^{(k)}$ and $q_{X|Y}^{(k)}$ in the Arimoto–Blahut Algorithm for calculating H-capacity C_H (cited from [18, Table I])

Name	$\mathcal{F}_H(p_X, q_{X\|Y})$	$p_X^{(k)}$	$q_{X\|Y}^{(k)}$
ABA for computing C [15], [16]	$\mathbb{E}_{X,Y}^{p_X p_{Y\|X}}\left[\log \frac{q_{X\|Y}(X\|Y)}{p_X(X)}\right]$	$\frac{\prod_y q_{X\|Y}^{(k-1)}(x\|y)^{p_{Y\|X}(y\|x)}}{\sum_x \prod_y q_{X\|Y}^{(k-1)}(x\|y)^{p_{Y\|X}(y\|x)}}$	$\frac{p_X^{(k)}(x)p_{Y\|X}(y\|x)}{\sum_x p_X^{(k)}(x)p_{Y\|X}(y\|x)}$
ABA for computing C_α^{A} [17]	$\frac{\alpha}{\alpha-1}\log\sum_{x,y} p_{X_\alpha}(x)^{\frac{1}{\alpha}} p_{Y\|X}(y\|x) q_{X\|Y}(x\|y)^{\frac{\alpha-1}{\alpha}}$	$\frac{\left(\sum_y p_{Y\|X}(y\|x) q_{X\|Y}^{(k-1)}(x\|y)^{\frac{\alpha-1}{\alpha}}\right)^{\frac{1}{\alpha-1}}}{\sum_x\left(\sum_y p_{Y\|X}(y\|x) q_{X\|Y}^{(k-1)}(x\|y)^{\frac{\alpha-1}{\alpha}}\right)^{\frac{1}{\alpha-1}}}$	$\frac{p_X^{(k)}(x)^\alpha p_{Y\|X}(y\|x)^\alpha}{\sum_x p_X^{(k)}(x)^\alpha p_{Y\|X}(y\|x)^\alpha}$
ABA for computing C_α^{A} [18]	$\frac{\alpha}{\alpha-1}\log\sum_{x,y} p_{X_\alpha}(x)^{\frac{1}{\alpha}} p_{Y\|X}(y\|x) q_{X_\alpha\|Y}(x\|y)^{\frac{\alpha-1}{\alpha}}$	$\frac{\left(\sum_y p_{Y\|X}(y\|x) q_{X_\alpha\|Y}^{(k-1)}(x\|y)^{\frac{\alpha-1}{\alpha}}\right)^{\frac{1}{\alpha-1}}}{\sum_x\left(\sum_y p_{Y\|X}(y\|x) q_{X_\alpha\|Y}^{(k-1)}(x\|y)^{\frac{\alpha-1}{\alpha}}\right)^{\frac{1}{\alpha-1}}}$	$\frac{p_X^{(k)}(x)p_{Y\|X}(y\|x)}{\sum_x p_X^{(k)}(x)p_{Y\|X}(y\|x)}$

and $q_{X_\alpha|Y} = \{q_{X_\alpha|Y}(\cdot|y)\}_{y\in\mathcal{Y}}$ is a set of α-tilted distribution of $q_{X|Y}(\cdot|y)$ defined as $q_{X_\alpha|Y}(x|y) := \frac{q_{X|Y}(x|y)^\alpha}{\sum_x q_{X|Y}(x|y)^\alpha}$.

The variational characterization $I_\alpha^{\mathrm{A}}(X;Y) = \max_{q_{X|Y}} \mathcal{F}_\alpha{}^{\mathrm{A1}}(p_X, q_{X|Y})$ is equivalent to that presented in [17, Eq. (7.103)] by Arimoto (see also [18, Prop 4 and Remark 4]). On the other hand, the variational characterization $I_\alpha^{\mathrm{A}}(X;Y) = \max_{q_{X|Y}} \mathcal{F}_\alpha{}^{\mathrm{A2}}(p_X, q_{X|Y})$ is equivalent to that presented in [18, Thm 1]. Therefore, Algorithm 1 applied for computing the Arimoto capacity C_α^{A} is equivalent to those previously presented in [17], [18]. Table 3 lists the explicit updating formulae for computing Arimoto capacity C_α^{A} of each algorithm.

Finally, we discuss the global convergence property of Algorithm 1. In general, there is no guarantee that Algorithm 1 exhibits global convergence property, and whether it does or not depends on the given $H = (\eta, F)$. However, the following sufficient condition on $H = (\eta, F)$ for the global convergence can be immediately obtained from [32, Thm 10.5].

Proposition 7: Let $\{p_X^{(k)}\}_{k=0}^\infty$ and $\{q_{X|Y}^{(k)}\}_{k=0}^\infty$ be sequences of distributions obtained from Algorithm 1. If $(p_X, q_{X|Y}) \mapsto \mathcal{F}_H(p_X, q_{X|Y})$ is jointly concave, then

$$\lim_{k\to\infty} \mathcal{F}_H(p_X^{(k)}, q_{X|Y}^{(k)}) = C_H. \qquad (61)$$

Remark 6: $\mathcal{F}(p_X, q_{X|Y}) := \mathbb{E}_{X,Y}^{p_X p_{Y|X}}\left[\log \frac{q_{X|Y}(X|Y)}{p_X(X)}\right]$ is a typical example that satisfies this condition (see [32, Section 10.3.2]). Note that even if $H = (\eta, F)$ does not satisfy this sufficient condition, it may be possible to show the global convergence property of Algorithm 1. For example, Kamatsuka *et al*. [18, Cor 2] proved that

$$\lim_{k\to\infty} \mathcal{F}_\alpha^{\mathrm{A1}}(p_X^{(k)}, q_{X|Y}^{(k)}) = \lim_{k\to\infty} \mathcal{F}_\alpha^{\mathrm{A2}}(p_X^{(k)}, q_{X|Y}^{(k)}) = C_\alpha^{\mathrm{A}} \qquad (62)$$

by showing the equivalence of the proposed algorithm with the alternating optimization algorithm for which global convergence is guaranteed by Arimoto [33, Thm 3].

5. Conclusion

In this study, we derived a variational characterization of H-MI $I_H(X;Y)$. On the basis of the characterization, we derived an alternating optimization algorithm for H-capacity $C_H := \max_{p_X} I_H(X;Y)$. We also showed that the algorithms applied for computing Arimoto capacity C_α^{A} coincide with the previously reported algorithms [17], [18]. In a future study, we will derive algorithms for the calculating Hayashi capacity $C_\alpha^{\mathrm{H}} := \max_{p_X} I_\alpha^{\mathrm{H}}(X;Y)$ and Fehr–Berens capacity $C_\alpha^{\mathrm{FB}} := \max_{p_X} I_\alpha^{\mathrm{FB}}(X;Y)$.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K16886.

References

[1] C.E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, vol.27, pp.379–423, 1948.

[2] S. Verdú, "α-mutual information," 2015 Information Theory and Applications Workshop (ITA), pp.1–6, 2015.

[3] R. Sibson, "Information radius," Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, vol.14, pp.149–160, 1969.

[4] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," 2nd Colloquium, Keszthely, Hungary, 1975, ed. I. Csiszar and P. Elias, Amsterdam, Netherlands: North Holland, pp.41–52, Colloquia Mathematica Societatis Jano's Bolyai, 1977.

[5] I. Csiszár, "Generalized cutoff rates and renyi's information measures," IEEE Transactions on Information Theory, vol.41, no.1, pp.26–34, 1995.

[6] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp.1401–1408, Oct 2012.

[7] J. Liao, O. Kosut, L. Sankar, and F. du Pin Calmon, "Tunable measures for information leakage and applications to privacy-utility tradeoffs," IEEE Transactions on Information Theory, vol.65, no.12, pp.8043–8066, 2019.

[8] M.S. Alvim, K. Chatzikokolakis, A. Mciver, C. Morgan, C. Palamidessi, and G. Smith, "Additive and multiplicative notions of leakage, and their capacities," 2014 IEEE 27th Computer Security Foundations Symposium, pp.308–322, 2014.

[9] M.S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith, "An axiomatization of information flow measures," Theoretical Computer Science, vol.777, pp.32–54, 2019. In memory of Maurice Nivat, a founding father of Theoretical Computer Science - Part I.

[10] M.S. Alvim, K. Chatzikokolakis, C. Palamidessi, and G. Smith, "Measuring information leakage using generalized gain functions," 2012 IEEE 25th Computer Security Foundations Symposium, pp.265–279, 2012.

[11] G.R. Kurri, L. Sankar, and O. Kosut, "An operational approach to information leakage via generalized gain functions," IEEE Transactions on Information Theory, pp.1–1, 2023.

[12] H. Raiffa and R. Schlaifer, Applied Statistical Decision Theory, Harvard Business School Publications, Division of Research, Graduate School of Business Adminitration, Harvard University, 1961.

[13] A. Américo and P. Malacaria, "Concavity, core-concavity, quasiconcavity: A generalizing framework for entropy measures," 2021 IEEE 34th Computer Security Foundations Symposium (CSF), pp.1–14, 2021.

[14] A. Américo, M. Khouzani, and P. Malacaria, "Conditional entropy and data processing: An axiomatic approach based on core-concavity," IEEE Transactions on Information Theory, vol.66, no.9, pp.5537–5547, 2020.

[15] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," IEEE Transactions on Information Theory, vol.18, no.1, pp.14–20, 1972.

[16] R. Blahut, "Computation of channel capacity and rate-distortion functions," IEEE Transactions on Information Theory, vol.18, no.4, pp.460–473, 1972.

[17] A. Suguru, Information Theory, Kyoritsu Suugaku Kouza (in Japanese), no.22, KYORITSU SHUPPAN, 1976.

[18] A. Kamatsuka, Y. Ishikawa, K. Kazama, and T. Yoshida, "New algorithms for computing sibson capacity and arimoto capacity," 2024. `https://arxiv.org/abs/2401.14241`.

[19] P.D. Grünwald and A.P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory," Ann. Statist., vol.32, no.4, pp.1367–1433, 08 2004.

[20] J. Berger, Statistical decision theory and Bayesian analysis, 2nd ed., Springer series in statistics, Springer, New York, NY, 1985.

[21] J.K. Ghosh, An introduction to Bayesian analysis : theory and methods / Jayanta K. Ghosh, Mohan Delampady, Tapas Samanta., Springer texts in statistics, Springer, New York.

[22] I.J. Good, "Rational decisions," Journal of the Royal Statistical Society: Series B (Methodological), vol.14, no.1, pp.107–114, 1952.

[23] T. Gneiting and A.E. Raftery, "Strictly proper scoring rules, prediction, and estimation," Journal of the American Statistical Association, vol.102, no.477, pp.359–378, 2007.

[24] A.P. Dawid and M. Musio, "Theory and applications of proper scoring rules," METRON, vol.72, no.2, pp.169–183, 2014.

[25] I.J. Good, Comments on "Measuring Information and Uncertainty" by R. J. Buehler, Rinehart and Winston, Toronto: Holt, 1971.

[26] R. Selten, "Axiomatic characterization of the quadratic scoring rule," Experimental Economics, vol.1, no.1, pp.43–61, 1998.

[27] S.i. Amari, Information Geometry and Its Applications, 1st ed., Springer Publishing Company, Incorporated, 2016.

[28] A. Américo, M. Khouzani, and P. Malacaria, "Channel-supermodular entropies: Order theory and an application to query anonymization," Entropy, vol.24, no.1, 2022.

[29] M. Hayashi, "Exponential decreasing rate of leaked information in universal random privacy amplification," IEEE Transactions on Information Theory, vol.57, no.6, pp.3989–4001, 2011.

[30] S. Fehr and S. Berens, "On the conditional rényi entropy," IEEE Transactions on Information Theory, vol.60, no.11, pp.6801–6810, 2014.

[31] M. Iwamoto and J. Shikata, "Information theoretic security for encryption based on conditional rényi entropies," Information Theoretic Security, ed. C. Padró, Cham, pp.103–121, Springer International Publishing, 2014.

[32] R.W. Yeung, A First Course in Information Theory (Information Technology: Transmission, Processing and Storage), Springer-Verlag, Berlin, Heidelberg, 2006.

[33] S. Arimoto, "Computation of random coding exponent functions," IEEE Transactions on Information Theory, vol.22, no.6, pp.665–671, 1976.

Akira KAMATSUKA received B.E., M.E. and Ph.D. degrees in the Department of Pure and Applied Mathematics from Waseda University, Tokyo, Japan, in 2012, 2014 and 2018, respectively. He is currently an associate professor in the Department of Information Science, Faculty of Engineering, Shonan Institute of Technology. His research interests are in information theory and statistical decision theory.

Koki Kazama received his B.E. degree in the Department of Applied Mathematics from Waseda University, Tokyo, Japan, in 2014. He received M.E. and Ph.D. degrees in the Department of Pure and Applied Mathematics from Waseda University, Tokyo, Japan, in 2016 and 2022, respectively. He is currently an assistant professor in the Department of Information Science, Faculty of Engineering, Shonan Institute of Technology. His research interests include information theory and information security.

Takahiro Yoshida received the B.E. and M.E. degrees from Musashi Institute of Technology in 1996 and 1998, respectively. He received Doctor of Engineering degree in Waseda University in 2010. From 2000 to 2003, he was a research associate in the School of Science and Engineering, Waseda University, Tokyo, Japan. From 2007 to 2010, he was a researcher in the Research and Development Initiative, Chuo University, Tokyo, Japan. From 2010 to 2013, he was a research associate in the College of Science and Engineering, Aoyama Gakuin University, Kanagawa, Japan. From 2013 to 2016, he was a lecturer, and from 2016 to 2019 he was an associate professor in the Faculty of Commerce, Yokohama College of Commerce, Kanagawa, Japan. Since 2020, he has been an associate professor in Nihon University College of Commerce, Tokyo, Japan. His research interests include information theory and its applications.