ISSN 2186-6570

An Introductory Review on Some Information-Theoretic Measures for Discrimination Between Probability Distributions

Mitsuru Hamada

Quantum Information Science Research Center
Quantum ICT Research Institute
Tamagawa University
6-1-1 Tamagawa-gakuen, Machida, Tokyo 194-8610, Japan

Tamagawa University Quantum ICT Research Institute Bulletin, Vol.11, No.1, 23-25, 2021

©Tamagawa University Quantum ICT Research Institute 2021
All rights reserved. No part of this publication may be reproduced in any form or by any means electrically, mechanically, by photocopying or otherwise, without prior permission of the copy right owner.

Tamagawa University Quantum ICT Research Institute Bulletin
Vol.11 No.1 : 23–25 (2021)

An Introductory Review on Some Information-Theoretic Measures for Discrimination Between Probability Distributions

Mitsuru Hamada
Quantum Information Science Research Center
Quantum ICT Research Institute
Tamagawa University
6-1-1 Tamagawa-gakuen, Machida, Tokyo 194-8610, Japan

***Abstract*—Chernoff's approach to statistical hypothesis testing has influenced information theory considerably. This review article focuses on Chernoff's quantity proposed in 1952 in his influential work. Its quantitative relations to a quantity better-known in information theory, i.e., the Kullback-Leibler information, are presented in terms of inequalities and equalities.**

I. Introduction

Most of fundamental coding theorems in classical information theory, which has been initiated by Shannon, can be written in terms of Shannon entropy, $H(P)$, and Kullback-Leibler information, $D(P||Q)$. The definitions of H and D are given in what follows; they are functions of probability distributions, denoted by P and Q here. For example, in the source coding theorem, $H(P)$ appears as the limit of data compression rate. Regarding channel coding, the mutual information $I(P,W)$, which is a function of a probability distribution P and a channel W, appears in the universal channel coding theorem of Csiszar and Koerner [1], [2]. One should note that $I(P,W)$ can also be written with D. This theorem shows that the decoding error of some sequence of channel codes goes to zero exponentially in the code-length, and it also gives a bound on the best exponent in a large-deviation-theoretic manner.

Such a large-deviation-theoretic bound was obtained earlier in [3], [4], [5], [6], [7], [8]. In particular, Gallager's work [5], [8] is famous for a simple proof of the attainability of the bound. There, the channel coding theorem is stated in a manner different from that of Csiszar and Koerner. While Csiszar and Koerner have proved the result in such a way that D and I appear directly, Gallager's approach resembles Chernoff's [9]. They exploit moment generating functions or their logarithms, called cumulant generating functions.

The attainable error exponent, which may be interpreted as speed of convergence, obtained in these two manners look different, but they equal each other quantitatively [1]. The aim of this article is not to discuss such a specialized matter in details, but to draw the reader's attention to some relationships among fundamental quantities that have appeared in the results described above (or, at least, in their underlying ideas). The importance of these quantities has already been illustrated by the above examples. These seem to be useful for diverse applications.

In this memorandum, we explain how Chernoff's quantity, specifically, the cumulant generating function suggested in Chernoff's work [9], is related to $D(P||Q)$. We will focus on showing quantitative relations (in terms of inequalities or equalities) among D and Chernoff's quantity, and only give brief references to specific pieces of the literature for the reader who is interested in how these quantities emerge in existing theories.

II. Preliminaries

Throughout, we use natural logarithms, i.e., logarithms are to base e. We only treat probability distributions on some finite set. We use the following notation on the method of types [1], [10], [11]. We denote by $\mathcal{P}(\mathcal{X})$ the set of all probability distributions on a set $\mathcal{X}$ [11]. We denote the type of $x \in \mathcal{X}^n$ by P_x. This means that the number of appearances of $u \in \mathcal{X}$ in $x \in \mathcal{X}^n$ is $n\mathsf{P}_x(u)$.

We follow the convention to denote by P_X the probability distribution of a random variable X. The expectation operation with respect to a random variable X taking values in a set $\mathcal{X}$ is denoted by E_X:

$$\mathsf{E}_X F(X) = \sum_{u \in \mathcal{X}} P_X(u) F(u)$$

where F is a real-valued function on $\mathcal{X}$. For a probability distribution $P \in \mathcal{P}(\mathcal{X})$, P^n denotes the product of n copies of P defined by $P^n(x_1, \ldots, x_n) = P(x_1) \cdots P(x_n)$.

The classical Kullback-Leibler information (also called relative entropy) [12], [13] is denoted by D and Shannon entropy by H. Specifically, for probability distributions P and Q on a finite set $\mathcal{X}$, we define $D(P||Q)$ by

$$D(P||Q) = \sum_{u \in \mathcal{X}} P(u) \log_{\mathrm{e}} \frac{P(u)}{Q(u)}$$

and $H(Q)$ by $H(Q) = -\sum_{u \in \mathcal{X}} Q(u) \log_{\mathrm{e}} Q(u)$. Here, the conventions $0 \log 0 = 0$, $0 \log(0/0) = 0$, and $p \log(p/0) = +\infty$ for $p > 0$ should be understood.

A short straightforward calculation gives

Fact 1: For $Q \in \mathcal{P}(\mathcal{X})$ and $x \in \mathcal{X}^n$,

$$Q^n(x) = \exp\{-n[H(\mathsf{P}_x) + D(\mathsf{P}_x||Q)]\}.$$

At this stage, $H(P)$ and $D(P||Q)$ have already appeared. We will move on to relate $D(P||Q)$ with some known quantity.

III. Chernoff's Quantity

Now fix a set $\mathcal{X} = \{a_1, \ldots, a_m\}$, where $m \geq 2$, and consider a pair of probability distributions P and $Q \in \mathcal{P}(\mathcal{X})$. For simplicity, write $p_i = P(a_i)$ and $q_i = Q(a_i)$ for $i = 1, \ldots, m$. To avoid cumbersomeness in presentation, we assume $p_i > 0$ and $q_i > 0$ for $i = 1, \ldots, m$ in what follows.

Some quantity, which may be viewed as a measure of closeness of P and Q, is known [9]:

$$C_\alpha(P, Q) = -\log \sum_{i=1}^{m} p_i^\alpha q_i^{1-\alpha}, \tag{1}$$

where $\alpha \in \mathbb{R}$.

Fact 2:

$$\left[\frac{dC_\alpha(P,Q)}{d\alpha}\right]_{\alpha=0} = D(Q||P)$$

and

$$\left[\frac{dC_\alpha(P,Q)}{d\alpha}\right]_{\alpha=1} = -D(P||Q).$$

Note that $C_\alpha(P, Q) = 0$ for $\alpha = 0, 1$, and $C_\alpha(P, P) = 0$.

Fact 3: $C_\alpha(P, Q)$ is a concave function of α. If $P \neq Q$, then $C_\alpha(P, Q)$ is a strictly concave.

We can also define

$$C_\alpha^*(P, Q) = \frac{1}{1-\alpha} C_\alpha(P, Q) \tag{2}$$

for $\alpha \neq 1$.

Fact 4:

$$\lim_{\alpha \to 1} C_\alpha^*(P, Q) = D(P||Q).$$

Fact 5: Assume $P \neq Q$. Then, $C_\alpha^*(P, Q)$ is a monotonically increasing function of α.

Hence, for $\alpha \leq 0$,

$$C_\alpha^*(P, Q) \leq 0, \tag{3}$$

for $0 \leq \alpha < 1$,

$$0 \leq C_\alpha^*(P, Q) \leq D(P||Q), \tag{4}$$

and for $\alpha > 1$,

$$C_\alpha^*(P, Q) \geq D(P||Q). \tag{5}$$

These facts presented in this section are taken from [14, Chapter 2, Exercises], so that proofs may be left to the reader as suggested exercises. However, the next section is helpful to understand these facts easily based on general backgrounds.

IV. Moment Generating Function

In this section, we show that Chernoff's quantity is obtained by applying some general notion to a special case.

For a random variable Z, which takes values in some finite subset $\mathcal{Z}$ of $\mathbb{R}$, we can define a function

$$M(t) = \mathsf{E}_Z\{\exp[tZ]\}, \quad t \in \mathbb{R}, \tag{6}$$

which is known as the moment generating function of Z. The quantity

$$\log M(t) \tag{7}$$

is called the cumulant generating function of Z.

Then, given P and $Q \in \mathcal{P}(\mathcal{X})$, as in the previous section, for a random variable X taking values in $\mathcal{X}$ and satisfying $P_X = Q$, the moment generating function of

$$Z = \log \frac{P(X)}{Q(X)} \tag{8}$$

is calculated as

$$M(t) = \sum_{i=1}^{m} q_i \exp\left[t \log \frac{p_i}{q_i}\right] = \sum_{i=1}^{m} p_i^t q_i^{1-t}.$$

Thus,

$$C_\alpha(P, Q) = -\log M(\alpha)$$

for a moment generating function $M(\alpha)$ of Z in (8), where $P_X = Q$.[1] In other words, the cumulant generating function of Z in (8) is $-C_\alpha(P, Q)$. Note also that

$$C_\alpha^*(P, Q) = \frac{-C_\alpha(P, Q) - [-C_1(P, Q)]}{\alpha - 1}, \tag{9}$$

so that $C_\alpha^*(P, Q)$ is the slope of the line passing through the points $(1, 0) = (1, -C_1(P, Q))$ and $(\alpha, -C_\alpha(P, Q))$ in the Cartesian plane.

Now one can see that each of those facts listed in the previous section either trivially follows or easily follows from the general properties of cumulant (or moment) generating functions, in particular, that the cumulant generating functions are convex. See a short text [15, Chapter 1, Section 9, pp. 145–152], which treats cumulant generating functions and Chernoff's result with an application to hypothesis testing.

The reader may wish to know how these quantities $D(P||Q)$ and $C_t(P, Q)$ appear in theories on statistical hypothesis testing or information. See the aforementioned text [15, Chapter 1, Section 9] for a readable exposition. See also the original paper [9], or [10, pp. 43–44], [16]. In information theorists' texts [17, Chapter 12], [18, Chapter 11], one may find a theorem (attributed to Chernoff) comparable to coding theorems. This is based on a result called Stein's lemma [17, Theorem 12.8.1] or the Chernoff-Stein Lemma [18, Theorem 11.8.3]. Kullback and Leibler [12] called $D(P||Q)$ the mean information for discrimination between the two hypotheses with P and Q per observation from P.

[1] In fact, Chernoff's suggestion [9] is to use $\inf_{\alpha \in (0,1)} M(\alpha)$ or $-\log \inf_{\alpha \in (0,1)} M(\alpha)$.

As for channel coding, the approach taken by Gallager [4], [5], [6], [7] to derive exponential bounds on the best decoding error probabilities of channel codes resembles Chernoff's. In particular, he began with arguments in the case of two code words in [8]. This is understood to be an argument on hypothesis testing.

V. Concluding Remarks

This review article has introduced a classical quantity of Chernoff's and described its relation to the Kullback-Leibler information (mean information for discrimination between two hypotheses). These quantities have appeared in treatments on statistical hypothesis testing. Issues on hypothesis testing have already been extended to quantum settings. Similarly to the classical case, hypothesis testing in quantum settings [19], [20, Chapter 8], [21] would be fundamental.

References

[1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. NY: Academic, 1981.

[2] ——, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Information Theory*, vol. IT-27, no. 1, pp. 5–12, Jan. 1981.

[3] R. M. Fano, *Transmission of Information, A Statistical Theory of Communications*. NY: Wiley, 1961.

[4] R. G. Gallager, *Low-density parity-check codes*. Cambridge, MA: MIT Press, 1963.

[5] ——, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Information Theory*, vol. IT-11, no. 1, pp. 3–18, Jan. 1965.

[6] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I," *Information and Control*, vol. 10, pp. 65–103, January 1967.

[7] ——, "Lower bounds to error probability for coding on discrete memoryless channels. II," *Information and Control*, vol. 10, pp. 522–552, May 1967.

[8] R. G. Gallager, *Information Theory and Reliable Communication*. NY: John Wiley & Sons, 1968.

[9] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

[10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. NY: Cambridge Univ. Press, 2011.

[11] I. Csiszár, "The method of types," *IEEE Trans. Information Theory*, vol. IT-44, no. 6, pp. 2505–2523, Oct. 1998.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[13] S. Kullback, *Information Theory and Statistics*. NY: Dover, 1968.

[14] T. S. Han and K. Kobayashi, *Joho to Fugoka no Suri*. Tokyo: Baifukan, 1999, in Japanese.

[15] P. Billingsley, *Probability and Measure*, 3rd ed. NY: Wiley, 1995.

[16] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Berlin: Springer, 1998.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. NY: Wiley, 1991.

[18] ——, *Elements of Information Theory*, 2nd ed. NY: Wiley, 2006.

[19] C. W. Helstrom, *Quantum Detection and Estimation Theory*. NY: Academic Press, 1976.

[20] D. Petz, *Quantum Information Theory and Quantum Statistics*. Berlin: Springer, 2008.

[21] T. Ogawa and H. Nagaoka, "Strong converse and Stein's lemma in quantum hypothesis testing," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2428–2433, 2000.