Fano-Type Inequalities Based on General Conditional Information Measures over Countably Infinite Alphabets with List Decoding

Yuta Sakai, Student Member, IEEE.

Abstract

Fano’s inequality is one of the most elementary, ubiquitous and important tools in information theory. This study generalizes Fano’s inequality in the following four ways: (i) the alphabet $X$ of the random variable $X$ to be estimated is countably infinite; (ii) the probability distribution $P_X$ of $X$ is fixed to a given discrete probability distribution $Q$; (iii) the inequality is established for a general conditional information measure $h_\phi(X \mid Y)$; and (iv) the decoding rule is a list decoding scheme, in contrast to a unique decoding scheme. In other words, our main results concern tight upper bounds on $h_\phi(X \mid Y)$ subject to an admissible list decoding error probability and a fixed $X$-marginal $P_X = Q$. Since $h_\phi(X \mid Y)$ admits the conditional Shannon and Rényi’s entropies as special cases, our Fano-type inequalities subsume known generalizations of Fano’s inequality. Moreover, since $h_\phi(X \mid Y)$ is a general definition without explicit form of a function $\phi$, our Fano-type inequalities also provide some insights on how to measure conditional information. As an application of our Fano-type inequalities, we investigate various asymptotic estimates on the equivocations under the condition that the error probability vanishes. Such asymptotic estimates are important consequences of Fano’s inequality. Most interestingly, a consequence of our results is a novel characterization of the asymptotic equipartition property (AEP).

Index Terms

Fano’s inequality; general conditional information; countably infinite alphabet; list decoding; asymptotic equipartition property.

I. INTRODUCTION

Inequalities relating probabilities to various information measures are fundamental tools for proving various coding theorems in information theory. Fano’s inequality [19] is one such paradigmatic example of an information-theoretic inequality; it elucidates the interplay between the conditional Shannon entropy $H(X \mid Y)$ and the error probability $\mathbb{P}\{X \neq Y\}$. If we denote $h_2(u) := -u \log u - (1 - u) \log(1 - u)$ as the binary entropy function on $[0, 1]$ with $h_2(0) = h_2(1) = 0$, Fano’s inequality can be written as

$$
\max_{(X, Y) : \mathbb{P}\{X \neq Y\} \leq \varepsilon} H(X \mid Y) = h_2(\varepsilon) + \varepsilon \log(M - 1)
$$

(1)

for each $0 \leq \varepsilon \leq 1 - 1/M$, where the maximum is taken over the two $\{1, \ldots, M\}$-valued random variables (r.v.’s) $X$ and $Y$ satisfying $\mathbb{P}\{X \neq Y\} \leq \varepsilon$, and log stands for the natural logarithm throughout the paper.

An important consequence of Fano’s inequality (1) is that if the error probability vanishes (i.e., $\mathbb{P}\{X^n \neq Y^n\} = o(1)$), this implies that the normalized Shannon’s equivocation also vanishes (i.e., $H(X^n \mid Y^n) = o(n)$). Here, $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$ are $\{1, \ldots, M\}$-valued random vectors. This key in proving weak converse results in various communication models (cf. [9], [18], [60]). Moreover, Fano’s inequality (1) also shows that $\mathbb{P}\{X_n \neq Y_n\} = o(1)$ implies that the unnormalized equivocation $H(X_n \mid Y_n) = o(1)$ for $\{1, \ldots, M\}$-valued r.v.’s $X_n$ and $Y_n$, $n \geq 1$. This is useful to prove that some Shannon’s information measures with respect to the error metric $\mathbb{P}\{X \neq Y\}$ (cf. [11]) are continuous.

A. Beyond the Finiteness of Alphabet $X$

While Fano’s inequality (1) is elementary and useful, it is not obvious to generalize Fano’s inequality from $\{1, \ldots, M\}$-valued to $\{1, 2, \ldots\}$-valued r.v.’s $X$ and $Y$, i.e., these r.v.’s are defined on a countably infinite alphabet $X = \{1, 2, \ldots\}$. In fact, it is easy to see that (1) diverges as $M \to \infty$, provided that $\mathbb{P}\{X \neq Y\} > 0$. Moreover, if $X_n$ and $Y_n$ are $X$-valued for each $n \geq 1$, one can construct a counterexample so that $\mathbb{P}\{X_n \neq Y_n\} = o(1)$ but $H(X_n \mid Y_n) = \infty$ for every $n \geq 1$ (cf. [60], Example 2.49). This counterexample implies that when generalizing Fano’s inequality (1) from a finite to a countably infinite alphabet $X$, we need to impose additional conditions on the r.v.’s.

Similar to the above difficulty of generalizing Fano’s inequality, it is not straightforward to generalize some information-theoretic tools for systems defined on a finite alphabet to one defined on a countably infinite alphabet. In fact, Ho–Yeung
[30] showed that Shannon’s information measures are not continuous with respect to the $\chi^2$-divergence, the relative entropy, and the variational distance. In addition, while weak typicality [9, Chapter 3] (or entropy-typical sequences [11, Problem 2.5]) is a convenient tool in proving achievability theorems for sources and channels with defined on countably infinite (or even uncountable) alphabets, strong typicality [11] is only amenable in situations with finite alphabets. To ameliorate this issue, Ho–Yeung [31] proposed a notion known as unified typicality that ensures that the desirable properties of weak and strong typicality are retained when one is working with countably infinite alphabets.

Particularly, Ho–Verdú [28] generalized Fano’s inequality (1) from a finite to a countably infinite alphabet $X$ by adding another condition on the r.v. $X$: the induced probability distribution $P_X$ of $X$ is fixed to a discrete probability distribution $Q$, i.e., $P_X = Q$. Let us call this inequality the Ho–Verdú–Fano inequality. Since the Ho–Verdú–Fano inequality [28] is given by the maximum of $H(X \mid Y)$ subject to the constraints $\mathbb{P}(X \neq Y) = \varepsilon$ and $P_X = Q$, it can be regarded as a rate-distortion function, i.e., the minimum of the mutual information $\mathbb{I}(X \mid Y)$ subject to the distortion constraint $\mathbb{E}(d(X,Y)) \leq \varepsilon$, where $d(x,y) = 1$ if $x = y$ and $d(x,y) = 0$ otherwise. Moreover, using the Ho–Verdú–Fano inequality, the authors in [28, Section V] provided some sufficient conditions on a source $(X_n)_{n=1}^{\infty}$ in which vanishing error probability implies vanishing normalized/unnormalized equivocations.

B. Other Previous Works of Generalizing Fano’s inequality

In addition to the Ho–Verdú–Fano inequality [28], many researchers have considered various directions to generalize Fano’s inequality. An interesting study involves reversing the usual Fano inequality. In this regard, lower bounds on $H(X \mid Y)$ subject to $\mathbb{P}(X \neq Y) = \varepsilon$ were independently established by Kovalevsky [35] and Tebбе–Dwyer [53] (see also Feder–Merhav’s study [21]).

Fano’s inequality with list decoding was initiated by Ahlswede–Gács–Körner [2]. They [2] proved the strong converse property (in Wolfowitz’s sense [59]) of degraded broadcast channels under the maximum error probability criterion by combining Fano’s inequality with list decoding and the blowing-up technique [11, Chapter 5] (see also [41, Section 3.6.2]). Extending Ahlswede–Gács–Körner’s proof technique [2] together with the wringing technique, Dueck [12] proved the strong converse property of multiple-access channels under the average error probability criterion. Since these proofs rely on a combinatorial lemma (cf. [11, Lemma 5.1]), they work only when the channel output alphabet is finite; but see recent work by Fong–Tan [16], [17] in which such techniques have been extended to Gaussian channels.

Han–Verdú [24] generalized Fano’s inequality on a countably infinite alphabet $X$ by investigating lower bounds on the mutual information $I(X \wedge Y) \geq d(\mathbb{P}(X \neq Y), \mathbb{P}(\tilde{X} \neq Y))$ via the data-processing inequality without additional conditions on the r.v.’s $X$ and $Y$, where $d(a \parallel b) = a \log(a/b) + (1 - a) \log((1 - a)/(1 - b))$ stands for the binary relative entropy, and $\tilde{X}$ and $Y$ are mutually independent r.v.’s having marginals of $X$ and $Y$ respectively. Instead of the Han–Verdú–Fano inequality [24], using the Donsker–Varadhan lemma [41, Equation (3.4.67)], Liu–Verdú [33] investigated the second-order asymptotics of the mutual information via a novel technique known as the pumping-up technique. It is worth mentioning that while the blowing-up technique [11, Chapter 5] works well only on finite output alphabets, the pumping-up technique works well not only on finite output alphabets but also on infinite output alphabets.

Generalizations of Fano’s inequality from the conditional Shannon entropy $H(X \mid Y)$ to Arimoto’s conditional Rényi entropy $H_\alpha(X \mid Y)$ were recently and independently investigated by Sakai–Iwata [45] and Sason–Verdú [47]. Sakai–Iwata [45] provided sharp upper/lower bounds on $H_\alpha(X \mid Y)$ for fixed $H_\alpha^*(X \mid Y)$ with two distinct orders $\alpha \neq \beta$. Since $H_\beta(X \mid Y)$ is a strictly monotone function of the minimum average probability of error if $\beta = \infty$, Sakai–Iwata’s results can be directly reduced to both the forward and reverse Fano inequalities on $H_\alpha(X \mid Y)$ (cf. [45, Section V in the arXiv paper]). Sason–Verdú [47] also gave generalizations of the forward and reverse Fano’s inequalities on $H_\alpha(X \mid Y)$ together with applications to M-ary Bayesian hypothesis testing. Moreover, in the forward Fano inequality on $H_\alpha(X \mid Y)$, they [47, Theorem 8] generalized the decoding rules from unique decoding to list decoding. The reverse Fano inequalities [45], [47] do not require the finiteness of an alphabet $X$; on the other hand, the forward Fano inequality [45], [47] work only on a finite alphabet $X$.

C. Contributions of This Study

The contributions of our study are as follows:

1) For the purpose of stating our generalized Fano-type inequalities, we introduce a general conditional information $h_\phi(X \mid Y)$. With appropriate choices of $\phi$, $h_\phi(X \mid Y)$ can be specialized to Shannon’s and Rényi’s information measures, and some other conditional quantities. Our main results are Fano-type inequalities relating $h_\phi(X \mid Y)$ to the error probability. Moreover, we also show that $h_\phi(X \mid Y)$ admits the following desirable properties: the data-processing inequality and the property that conditioning reduces information. In this regard, $h_\phi(X \mid Y)$ is similar to the f-divergence [3], [10], which is a generalized divergence measure between probability measures that possesses some desirable properties. Naturally, our Fano-type inequalities on $h_\phi(X \mid Y)$ also reduces to the Ho–Verdú–Fano inequality [28] and Fano’s inequality on Rényi’s information measures [45], [47].

2) We investigate Fano-type inequalities on $h_\phi(X \mid Y)$ subject to small probability of list decoding error. Namely, our Fano-type inequalities can be applied to list decoding schemes for a source on a countably infinite alphabet $X$. In addition, our Fano-type inequality can also be specialized to the Sason–Verdú–Fano inequality [47, Theorem 8].
3) We prove our Fano-type inequalities via a novel application of majorization theory [38]. To simplify our analysis, we shall employ the infinite-dimensional Birkhoff’s theorem [43]. This approach results in a novel characterization of Fano’s inequality on a countably infinite alphabet $X$ via majorization theory. Moreover, when the side-information $Y$ is defined on a finite alphabet $\mathcal{Y}$, we refine our Fano-type inequality via an application of the finite-dimensional Birkhoff’s theorem [5].

4) Via our Fano-type inequalities, we investigate sufficient conditions on a general source $X = \{X_n\}_{n=1}^\infty$ in which vanishing error probability implies vanishing equivocation. We show that the asymptotic equipartition property (AEP) as defined in Verdú–Han [58] is indeed such a sufficient condition. This is, to the best of the author’s knowledge, a novel connection between the AEP and Fano’s inequality. Moreover, we extend Ho–Verdú’s sufficient conditions [28, Section V] and Sason–Verdú’s sufficient conditions [47, Theorem 4] on $X = \{X_n\}_{n=1}^\infty$ to Rényi’s equivocation measures with list decoding schemes. Lastly, the symbol-wise error criterion is also considered.

D. Paper Organization

The rest of this paper is organized as follows: We establish our Fano-type inequalities in Section II. Section III investigates several asymptotic estimates on equivocations of a general source under the vanishing error probability. We conclude this study in Section IV with some remarks. The proofs of the results are given in the appendices of this paper.

II. FANO-TYPE INEQUALITY

This section establishes our Fano-type inequalities based on a general conditional information measure $h_\phi(X \mid Y)$ under a fidelity criterion that is related to list decoding schemes. The quantity $h_\phi(X \mid Y)$ is defined in Section II-A; the notion of list decoding schemes is introduced in Section II-B; our Fano-type inequalities are established in Section II-C; and we reduce our Fano-type inequalities from $h_\phi(X \mid Y)$ to known Shannon’s and Rényi’s information measures in Section II-D.

A. General Conditional Information Measures

Let $X = \{1, 2, \ldots\}$ be a countably infinite alphabet. In this paper, a discrete probability distribution on $X$ is called an $X$-marginal. Given an $X$-marginal $P$, a decreasing rearrangement of $P$ is denoted by $P^↓$, i.e., it fulfills

$$P^↓(1) \geq P^↓(2) \geq P^↓(3) \geq P^↓(4) \geq P^↓(5) \geq \cdots .$$

The following definition gives us the notion of majorization for $X$-marginals.

**Definition 1** (Majorization [38]). An $X$-marginal $P$ is said to be majorized by another $X$-marginal $Q$ if

$$\sum_{i=1}^k P^↓(i) \leq \sum_{i=1}^k Q^↓(i)$$

for every $k \geq 1$. If this relation holds we also say that $Q$ majorizes $P$. This relation is denoted by $P < Q$ or $Q > P$.

The following definitions\(^1\) are important postulates of a function $\phi : \mathcal{P}(X) \to [0, \infty]$ playing the role of an information measure of an $X$-marginal, where $\mathcal{P}(X)$ is the set of $X$-marginals.

**Definition 2.** A function $\phi : \mathcal{P}(X) \to [0, \infty]$ is said to be symmetric if it is invariant for any permutation of probability masses, i.e., $\phi(P) = \phi(P^↓)$ for every $P \in \mathcal{P}(X)$.

**Definition 3.** A function $\phi : \mathcal{P}(X) \to [0, \infty]$ is said to be lower semicontinuous if for any $P \in \mathcal{P}(X)$, it holds that $\liminf_n \phi(P_n) \geq \phi(P)$ for every pointwise convergent sequence $P_n \to P$.

**Definition 4.** A function $\phi : \mathcal{P}(X) \to [0, \infty]$ is said to be convex if $\phi(R) \leq \lambda \phi(P) + (1 - \lambda)\phi(Q)$ with $R = \lambda P + (1 - \lambda)Q$ for every $P, Q \in \mathcal{P}(X)$ and $0 \leq \lambda \leq 1$.

**Definition 5.** A function $\phi : \mathcal{P}(X) \to [0, \infty]$ is said to be quasiconvex if the sublevel set $\{P \in \mathcal{P}(X) \mid \phi(P) \leq c\}$ is convex for every $P \in \mathcal{P}(X)$ and $c \in [0, \infty)$.

**Definition 6.** A function $\phi : \mathcal{P}(X) \to [0, \infty]$ is said to be Schur-convex if $P < Q$ implies $\phi(P) \leq \phi(Q)$.

We now introduce a generalized conditional information measure of an $X$-valued r.v. $X$ given a $\mathcal{Y}$-valued r.v. $Y$, where $\mathcal{Y}$ is an arbitrary alphabet\(^2\). Denote by $\mathbb{E}[-]$ the expectation operator. Given a symmetric, concave, and lower semicontinuous function $\phi : \mathcal{P}(X) \to [0, \infty]$, the generalized conditional information measure is defined by

$$h_\phi(X \mid Y) := \mathbb{E}[\phi(P_{X|Y})],$$

1Definitions 4–6, each term or its suffix convex is replaced by concave if $-\phi$ fulfills the condition.

2Unless stated otherwise, we automatically assume that $\mathcal{Y}$ is standard Borel.
where $P_{X|Y}$ stands for the regular conditional probability distribution of $X$ given $\sigma(Y)$ (cf. [13, Section 5.1.3]), i.e., $P_{X|Y}(x)$ is a version of the conditional probability $\mathbb{P}\{X = x \mid \sigma(Y)\}$ for each $x \in X$.

The following lemma is one of the reasons why we assume that $\phi$ is symmetric, concave, and lower semicontinuous.

**Proposition 1.** Every symmetric and quasiconvex function $\phi : \mathcal{P}(X) \to [0, \infty]$ is Schur-convex.

**Proof of Proposition 1:** See Appendix A-A.

To apply the Schur-convavity property in the sequel, Proposition 1 suggests assuming that $\phi$ is symmetric and quasiconcave. In addition, to adopt Jensen’s inequality on the function $\phi$, it suffices to assume that $\phi$ is concave and lower semicontinuous, because the domain $\mathcal{P}(X)$ forms a closed convex bounded set in the variational distance topology (cf. [50, Proposition A-2]). Motivated by these properties, we impose these three postulates on $\phi$ in this study.

**B. Minimum Average Probability of list decoding Error**

Consider a certain communication model for which a $\mathcal{Y}$-valued r.v. $Y$ plays a role of the side information of an $\mathcal{X}$-valued r.v. $X$. A list decoding scheme with an (allowable) list size $1 \leq L < \infty$ is a decoding scheme producing $L$ or fewer candidates (i.e., a list of size at most $L$) for realizations of $X$ when we observe a realization of $Y$. The minimum average error probability under list decoding is defined by

$$P_e^{(L)}(X \mid Y) := \min_{f : \mathcal{Y} \to \mathcal{X}^{(L)}} \mathbb{P}\{X \neq f(Y)\},$$

where the minimization is taken over all set-valued functions $f : \mathcal{Y} \to \mathcal{X}^{(L)}$ with the decoding range

$$\mathcal{X}^{(L)} := \bigcup_{i=1}^{L} \{X\}_{l=1}^{|X|},$$

$$\{X\}_{l=1}^{|X|} := \{D \subset X \mid |D| = l\},$$

and $|\cdot|$ stands for the cardinality of a set. If $L = 1$, then (5) coincides with the average error probability of the maximum a posteriori (MAP) decoding scheme. For short, we write $P_e(X \mid Y) := P_e^{(1)}(X \mid Y)$. Obviously, the inequality $\mathbb{P}\{X \neq f(Y)\} \leq \varepsilon$ implies $P_e^{(L)}(X \mid Y) \leq \varepsilon$ for any list decoder $f : \mathcal{Y} \to \mathcal{X}^{(L)}$ and any tolerated probability of error $\varepsilon \geq 0$. Therefore, it suffices to consider the constraint $P_e^{(L)}(X \mid Y) \leq \varepsilon$ rather than $\mathbb{P}\{X \neq f(Y)\} \leq \varepsilon$ in our subsequent analyses.

The following proposition shows that $P_e^{(L)}(X \mid Y)$ can be calculated similarly to that for MAP decoding.

**Proposition 2.** It holds that

$$P_e^{(L)}(X \mid Y) = 1 - \mathbb{E}\left[\sum_{x=1}^{L} P_{X|Y}(x)\right].$$

**Proof of Proposition 2:** See Appendix A-B.

Define the counting function of a set $A$ by

$$\#(A) := \begin{cases} |A| & \text{if } A \text{ is finite}, \\ \infty & \text{if } A \text{ is infinite}. \end{cases}$$

(9)

The following proposition provides fundamental upper and lower bounds on $P_e^{(L)}(X \mid Y)$.

**Proposition 3.** If $P_X = Q$, then

$$1 - \sum_{x=1}^{\#(\mathcal{Y})} Q^L(x) \leq P_e^{(L)}(X \mid Y) \leq 1 - \sum_{x=1}^{L} Q^1(x).$$

(10)

Moreover, both inequalities are sharp in the sense that there exists $X \times \mathcal{Y}$-valued r.v.'s $(X, Y)$ achieving the equalities while respecting the constraint $P_X = Q$.

**Proof of Proposition 3:** See Appendix A-C.

Denote the minimum average error probability for list decoding concerning $X \sim Q$ without any side-information as

$$P_e^{(L)}(Q) := 1 - \sum_{x=1}^{L} Q^1(x).$$

(11)
Then, the second inequality in (10) is obvious, and it is similar to the property that *conditioning reduces uncertainty* (cf. [9, Theorem 2.8.1]). Proposition 3 ensures that when we have to consider the constraints \( P_e^{(L)}(X \mid Y) \leq \varepsilon \) and \( P_X = Q \), it suffices to consider a system \((Q, L, \varepsilon, Y)\) fulfilling
\[
1 - \sum_{x=1}^{\#(Y)L} Q^j(x) \leq \varepsilon \leq 1 - \sum_{x=1}^{L} Q^j(x).
\] (12)

**C. Main Results**

Given a system \((Q, L, \varepsilon)\) fulfilling (12) with \#(Y) = \infty, define the *Fano-distribution*\(^4\) of type-1 by the following \( X \)-marginal:
\[
P_{Fano-type1}^{(Q, L, \varepsilon)}(x) := \begin{cases} 
Q^j(x) & \text{if } 1 \leq x < J \text{ or } K_1 < x < \infty, \\
\mathcal{V}(J) & \text{if } J \leq x \leq L, \\
\mathcal{W}(K_1) & \text{if } L < x \leq K_1,
\end{cases}
\] (13)

where the weight \( \mathcal{V}(j) \) is defined by
\[
\mathcal{V}(j) := \begin{cases} 
(1 - \varepsilon) - \sum_{x=1}^{j-1} Q^j(x) & \text{if } 1 \leq j \leq L, \\
1 & \text{if } j > L
\end{cases}
\] (14)

for each \( j \geq 1 \); the weight \( \mathcal{W}(k) \) is defined by
\[
\mathcal{W}(k) := \begin{cases} 
-1 & \text{if } k = L, \\
\sum_{x=1}^{k} Q^j(x) - (1 - \varepsilon) & \text{if } L < k < \infty, \\
0 & \text{if } k = \infty
\end{cases}
\] (15)

for each \( k \geq L \); the integer \( J \) is chosen so that
\[
J := \min\{1 \leq j < \infty \mid Q^j(j) < \mathcal{V}(j)\};
\] (16)

and \( K_1 \) is chosen so that
\[
K_1 := \sup\{L \leq k < \infty \mid \mathcal{W}(k) < Q^j(k)\}.
\] (17)

A graphical representation of the Fano-type distribution of type-1 is shown in Fig. 1. We say that \( Y \) satisfies the condition \( \blacklozenge \) if
\[
|Y| \geq \min \left\{ \left( \frac{K_1 - J + 1}{L - J + 1} \right), (K_1 - J)^2 + 1 \right\}.
\] (18)

Denote by \( N_0 \) and \( \varepsilon \) the cardinalities of the natural numbers \( \mathbb{N} \) and the continuum \( \mathbb{R} \), respectively. The following theorem presents one of our Fano-type inequalities.

**Theorem 1.** Suppose that the system \((Q, L, \varepsilon, Y)\) fulfills (12). Then, it holds that
\[
\sup_{(X,Y);P_e^{(L)}(X \mid Y) \leq \varepsilon, P_X = Q} h_\phi(X \mid Y) \leq \phi \left( P_{Fano-type1}^{(Q, L, \varepsilon)} \right),
\] (19)

where the supremum is taken over the \( X \times Y \)-valued r.v.'s \((X,Y)\) satisfying \( P_e^{(L)}(X \mid Y) \leq \varepsilon \) and \( P_X = Q \). Inequality (19) holds with equality if any one of the following four conditions holds:
1) \( \varepsilon = P_e^{(L)}(Q) \);
2) \( Y \) satisfies the condition \( \blacklozenge \) and \( 0 < \varepsilon < P_e^{(L)}(Q) \);
3) \( Y \) satisfies the condition \( \blacklozenge \) and \( \supp(Q) \) is finite; or
4) \( J = L \) and \( |Y| = N_0 \),

where \( \binom{a}{b} := \frac{a!}{b!(a-b)!} \) stands for the binomial coefficient for two integers \( 0 \leq b \leq a \). Moreover, if \( |Y| \geq \varepsilon \), then there exists a \( \sigma \)-algebra on \( Y \) satisfying (19) with equality. A pair \((X,Y)\) achieves the supremum in (19) if
\[
P_{X|Y}^j(x) = P_{Fano-type1}^{(Q, L, \varepsilon)}(x) \quad \text{(a.s.)}
\] (20)

for every \( x \in X \). In particular, if the concavity of \( \phi \) is strict, then (20) is the necessary and sufficient condition of achieving the supremum in (19).

**Proof of Theorem 1:** See Appendix B-A.

\(^{4}\)Note that the term “Fano-distribution” was already used by Ahlswede [1] in a different definition.
Fig. 1: Plot of the Fano-type distribution of type-1 \((13)\) from an \(X\)-marginal \(Q\) with \(L = 3\). Each bar represents a probability mass with decreasing rearrangement \(Q\) (cf. (2)).

Fig. 2: Plot of making the Fano-distribution of type-2 \((21)\) from an \(X\)-marginal \(Q\) with \(L = 3\) and \(#(Y) = 2\). Each bar represents a probability mass of the decreasing rearrangement \(Q\).

**Remark 1.** The Fano-type inequality \((19)\) of Theorem 1 is not sharp in general, i.e., there is a system \((Q, L, \varepsilon, Y)\) satisfying \((19)\) with strict inequality. On the other hand, the conditions 1)–4) of Theorem 1 and the existence of a \(\sigma\)-algebra on \(Y\) are sufficient conditions to ensure sharpness of the Fano-type inequality \((19)\). Moreover, Equation \((20)\) implies that \((19)\) holds with equality, which implies the existence of a pair \((X, Y)\) achieving the supremum in \((19)\).

As a refinement of Theorem 1, whenever \(Y\) is finite, the Fano-type inequality \((19)\) can be tightened as follows: Given a system \((Q, L, \varepsilon, Y)\) fulfilling \((12)\), we define the Fano-distribution of type-2 as the following \(X\)-marginal:

\[
p^{(Q, L, \varepsilon, Y)}_{\text{Fano-type2}}(x) := \begin{cases} Q^i(x) & \text{if } 1 \leq x < J \text{ or } K_2 < x < \infty, \\ V(J) & \text{if } J \leq x \leq L, \\ W(K_2) & \text{if } L < x \leq K_2, \end{cases} \tag{21}\]

where \(V(j)\), \(W(k)\), and \(J\) are defined in \((14)\), \((15)\), and \((16)\), respectively; and the integer \(K_2\) is chosen so that

\[
K_2 := \max\{L \leq k \leq #(Y)L \mid W(k) < Q^i(k)\}. \tag{22}\]

A graphical representation of the Fano-distribution of type-2 is illustrated in Fig. 2. The following theorem is a refinement of Theorem 1 for finite \(Y\).
**Theorem 2.** Suppose that $\mathcal{Y}$ is finite, and the system $(Q, L, \varepsilon, \mathcal{Y})$ fulfills (12). Then, it holds that

$$
\sup_{(X,Y): P_e^{(L)}(X|Y) \leq \varepsilon, P_X = Q} \mathfrak{b}_\phi(X | Y) \leq \phi \left( P^{(Q,L,e)}_{\text{Fano-type1}} \right)
$$

where the supremum is taken over the $X \times \mathcal{Y}$-valued r.v.'s $(X,Y)$ satisfying $P_e^{(L)}(X | Y) \leq \varepsilon$ and $P_X = Q$. Inequality (23) holds with equality if either one of the following two conditions holds:

1) $\varepsilon = P_e^{(L)}(Q)$ and $|\mathcal{Y}| \geq 1$; or
2) the following inequality holds:

$$
|\mathcal{Y}| \geq \min \left\{ \left( \frac{K_2 - J + 1}{L - J + 1} \right), (K_2 - J)^2 + 1 \right\}.
$$

A pair $(X,Y)$ achieves the supremum in (23) if

$$
P_{X|Y}^{(1)}(x) = P^{(Q,L,e)}_{\text{Fano-type1}}(x) \quad (\text{a.s.})
$$

for every $x \in X$. In particular, if the concavity of $\phi$ is strict, then (25) is the necessary and sufficient condition of achieving the supremum in (23).

**Proof of Theorem 2:** See Appendix B-B.

**Remark 2.** By Lemma 5 in Appendix B-A, it can be verified that $P^{(Q,L,e,\mathcal{Y})}$ majorizes $P^{(Q,L,e)}_{\text{Fano-type1}}$; that is, it follows from Proposition 1 that the Fano-type inequality in (23) of Theorem 2 is tighter than inequality (19) of Theorem 1 if $\mathcal{Y}$ is finite.

Another benefit of Theorem 2 is that it is sharp if $L = 1$ (see (24)).

The main techniques of proving Theorems 1 and 2 will be summarized in Section IV-B.

### D. Special Cases: Fano-Type Inequality on Rényi’s Information Measures

In this subsection, we specialize Theorems 1 and 2 to Fano-type inequalities involving Shannon’s and Rényi’s information measures. We recover several known results such as those in [19, 28, 45, 47] along the way.

The conditional Shannon entropy [48] of an $X$-valued r.v. $X$ given a $\mathcal{Y}$-valued r.v. $Y$ is defined by

$$
H(X | Y) := \mathbb{E}[H(P_X|Y)],
$$

where the (unconditional) Shannon entropy of an $X$-marginal $P$ is defined by

$$
H(P) := \sum_{x \in X} P(x) \log \frac{1}{P(x)}.
$$

The following proposition is a well-known fact of Shannon’s information measures.

**Proposition 4 ([57]).** The Shannon entropy $H : \mathcal{P}(X) \to [0, \infty]$ is symmetric, concave, and lower semicontinuous.

Namely, the conditional Shannon entropy $H(X | Y)$ is a special case of $\mathfrak{b}_\phi(X | Y)$ with $\phi = H$.

While the choices of Shannon’s information measures are unique based on a set of axioms (see, e.g., [11, Theorem 3.6] and [60, Chapter 3]), there are several different definitions of conditional Rényi entropies [22, 34, 55]. Among them, this study focuses on Arimoto’s and Hayashi’s conditional Rényi entropies [4, 26]. Arimoto’s conditional Rényi entropy of $X$ given $Y$ is defined by

$$
H_\alpha^{(X,Y)} := \frac{\alpha}{1 - \alpha} \log \mathbb{E}[\|P_X|Y\|_\alpha]
$$

for each order $\alpha \in (0, 1) \cup (1, \infty)$, where the $\ell_\alpha$-norm of an $X$-marginal $P$ is defined by

$$
\|P\|_\alpha := \left( \sum_{x \in X} P(x)^\alpha \right)^{1/\alpha}.
$$

Here, note that the (unconditional) Rényi entropy of an $X$-marginal $P$ is defined by

$$
H_\alpha(P) := \frac{1}{1 - \alpha} \log \sum_{x \in X} P(x)^\alpha = \frac{\alpha}{1 - \alpha} \log \|P\|_\alpha,
$$

i.e., it is a monotone function of the $\ell_\alpha$-norm. The following proposition summarizes properties of the $\ell_\alpha$-norm.

**Proposition 5.** The $\ell_\alpha$-norm $\| \cdot \| : \mathcal{P}(X) \to [0, \infty]$ is symmetric and lower semicontinuous. Moreover, it is concave and convex if $0 < \alpha < 1$ and $\alpha \geq 1$, respectively.
Proof of Proposition 5: The symmetry is obvious. The lower semicontinuity can be verified by the same way as [57, Theorem 3.2]. The concavity and convexity can be verified by the reverse and forward Minkowski inequalities respectively.

Hence, Arimoto’s conditional Rényi entropy $H^H_{\alpha}(X \mid Y)$ is a monotone function of $h_\phi(X \mid Y)$ with $\phi = \| \cdot \|_\alpha$, i.e.,

$$H_{\alpha}(X \mid Y) = \frac{1}{1 - \alpha} \log \left( h_{\| \cdot \|_\alpha}(X \mid Y) \right). \quad (31)$$

On the other hand, Hayashi’s conditional Rényi entropy $H^H_{\alpha}(X \mid Y)$ is defined by

$$H^H_{\alpha}(X \mid Y) := \frac{1}{1 - \alpha} \log \mathbb{E}[\|P_X|Y\|_\alpha^\alpha] \quad (32)$$

for each order $\alpha \in (0, 1) \cup (1, \infty)$. It is easy to see that $\| \cdot \|_\alpha : \mathcal{P}(X) \to [0, \infty]$ also admits the same properties as those stated in Proposition 5. Therefore, Hayashi’s conditional Rényi entropy $H^H_{\alpha}(X \mid Y)$ is also a monotone function of $h_\phi(X \mid Y)$ with $\phi = \| \cdot \|_\alpha$, i.e.,

$$H^H_{\alpha}(X \mid Y) = \frac{1}{1 - \alpha} \log \left( h_{\phi}(X \mid Y) \right). \quad (33)$$

By convention of Rényi’s information measures, define $H^H_{\alpha}(X \mid Y) := H(X \mid Y)$ for each $\alpha \in \{A, H\}$.

The following corollary is a direct consequence of Theorem 1.

Corollary 1. Suppose that the system $(Q, L, \varepsilon, \mathcal{Y})$ satisfies (12). Then, it holds that for each $\alpha \in \{A, H\}$,

$$\sup_{(X, Y) : P^L_{x}(X|Y) \leq \varepsilon, P_X = Q} H^H_{\alpha}(X \mid Y) \leq H_{\alpha}\left( p^{(Q, L, \varepsilon)}_{\text{Fano-type1}} \right) \quad (34)$$

where the supremum is taken over the $X \times \mathcal{Y}$-valued r.v.’s $(X, Y)$ satisfying $P^L_X(X \mid Y) \leq \varepsilon$ and $P_X = Q$; and the mapping $\eta : [0, 1] \to [0, 1/\varepsilon]$ is defined by $\eta : u \mapsto -u \log u$ satisfying $\eta(0) = 0$.

Proof of Corollary 1: Suppose that $\hat{\alpha} = A$. If $\alpha = 1$, then it immediately follows from (19) of Theorem 1 and Proposition 4 that

$$\sup_{(X, Y) : P^L_{x}(X|Y) \leq \varepsilon, P_X = Q} H(X \mid Y) \leq H\left( p^{(Q, L, \varepsilon)}_{\text{Fano-type1}} \right), \quad (35)$$

which is indeed (34) with $\alpha = 1$. Similarly, It follows from (19) of Theorem 1 and Proposition 5 that

$$0 < \alpha \leq 1 \quad \Rightarrow \quad \sup_{(X, Y) : P^L_{x}(X|Y) \leq \varepsilon, P_X = Q} \mathbb{E}[\|P_X|Y\|_\alpha] \leq \left\| p^{(Q, L, \varepsilon)}_{\text{Fano-type1}} \right\|_{\alpha}, \quad (36)$$

$$\alpha \geq 1 \quad \Rightarrow \quad \inf_{(X, Y) : P^L_{x}(X|Y) \leq \varepsilon, P_X = Q} \mathbb{E}[\|P_X|Y\|_\alpha] \geq \left\| p^{(Q, L, \varepsilon)}_{\text{Fano-type1}} \right\|_{\alpha}. \quad (37)$$

Since the mapping $u \mapsto (\alpha/(1 - \alpha)) \log u$ is strictly increasing if $0 < \alpha < 1$; and is strictly decreasing if $\alpha > 1$, it follows from (30)–(31) and the above two inequalities that

$$\sup_{(X, Y) : P^L_{x}(X|Y) \leq \varepsilon, P_X = Q} H^H_{\alpha}(X \mid Y) \leq H_{\alpha}\left( p^{(Q, L, \varepsilon)}_{\text{Fano-type1}} \right) \quad (38)$$

for every $\alpha \in (0, 1) \cup (1, \infty)$. Combining (35) and (38), we have Corollary 1 with $\hat{\alpha} = A$. The proof with $\hat{\alpha} = H$ is the same as the proof with $\hat{\alpha} = A$, proving Corollary 1.

In contrast to Corollary 1, the following corollary is a direct consequence of Theorem 2.
Corollary 2. Suppose that \( Y \) is finite, and the system \((Q, L, \varepsilon, Y)\) satisfies (12). Then, it holds that for any \( \bar{\varepsilon} \in \{A, H\} \),

\[
\sup_{(X,Y):P^L_e(X|Y)\leq \varepsilon, P_X=Q} H^\alpha_\varepsilon(X \mid Y) \leq H^\alpha_\varepsilon \left( p^{(Q,L,\varepsilon,Y)}_{\text{Fano-type}2} \right)
\]

\[
= \begin{cases} 
\frac{1}{1-\alpha} \log \left( (J-L+1)V(J) + K_2(W(K_2) + \sum_{x<J \text{ or } x\geq K_2} \eta(Q(x)) \right) & \text{if } \alpha \in (0, 1) \cup (1, \infty), \\
(J-L+1)\eta(V(J)) + K_2\eta(W(K_2)) + \sum_{x \in \{1, \ldots, 8\}} \eta(Q(x)) & \text{if } \alpha = 1,
\end{cases}
\]

(39)

where the supremum is taken over the \( X \times Y \)-valued r.v.'s \((X,Y)\) satisfying \( P_e(X \mid Y) \leq \varepsilon \) and \( P_X = Q \).

Proof of Corollary 2: The proof is totally the same as the proof of Corollary 1.

Remark 3. If \( \alpha = 1 \) and \( L = 1 \) hold simultaneously, then Corollaries 1 and 2 coincide with the Ho–Verdú–Fano inequality [28, Theorem 1] and its refined inequality [28, Theorem 4], respectively. Note that both conditions for sharpness and equalities of Corollaries 1 and 2 are the same as Theorems 1 and 2, respectively (see also Remark 1); and we have omitted these conditions in the corollaries for the sake of brevity.

Finally, we shall reduce Corollary 1 to the forward Fano inequality for Rényi’s information measures on a finite alphabet [45], [47]. Given two integers \( 1 \leq L < M \) and a real number \( 0 \leq \varepsilon \leq 1 - L/M \), define the Fano-distribution of type-0 by the following \( X \)-marginal:

\[
p^{(M,L,\varepsilon)}_{\text{Fano-type0}}(x) := \begin{cases} 
\frac{1 - \varepsilon}{L} & \text{if } 1 \leq x \leq L, \\
\frac{\varepsilon}{M-L} & \text{if } L < x \leq M, \\
0 & \text{if } M < x < \infty.
\end{cases}
\]

(40)

A graphical representation of the Fano-distribution of type-0 is plotted in Fig. 3. The following corollary is a direct consequence of Corollary 1.

Corollary 3. For two integers \( 1 \leq L < M < \infty \) and a real number \( 0 \leq \varepsilon \leq 1 - L/M \) it holds that

\[
\max_{(X,Y):P^L_e(X|Y)\leq \varepsilon} H^\alpha_\varepsilon(X \mid Y) = H^\alpha_\varepsilon \left( p^{(M,L,\varepsilon)}_{\text{Fano-type0}} \right)
\]

\[
= \begin{cases} 
\frac{1}{1-\alpha} \log \left( (1-\varepsilon)^{\alpha}L^{1-\alpha} + \varepsilon^{\alpha}(M-L)^{1-\alpha} \right) & \text{if } \alpha \in (0, 1) \cup (1, \infty), \\
\eta_2(\varepsilon) + (1-\varepsilon) \log L + \varepsilon \log(M-L) & \text{if } \alpha = 1,
\end{cases}
\]

(41)

Fig. 3: Each bar represents a probability mass of the Fano-distribution of type-0 (40) with \( M = 8 \) and \( L = 3 \).
for each $\hat{x} \in \{A, H\}$, where the maximum is taken over the $\{1, \ldots, M\} \times \mathcal{Y}$-valued r.v.'s $(X,Y)$ satisfying $P_e^{(L)}(X \mid Y) \leq \epsilon$. A pair $(X,Y)$ achieves the maximum in (41) if and only if
\[ P_{X \mid Y}(x) = P_{\text{Fano-type}(\epsilon)}(x) \]
for every $x \in X$.

**Proof of Corollary 3:** It is known that every discrete probability distribution on $\{1, \ldots, M\}$ majorizes the uniform distribution on $\{1, \ldots, M\}$. Therefore, combining Proposition 1, Corollary 1, and Lemma 5 of Appendix B-A, we can obtain Corollary 3.

The condition for equality in (42) can be verified by constructing a $\{1, \ldots, M\} \times \mathcal{Y}$-valued r.v. $(X,Y)$ so that $P_X = P_{\text{Fano-type}(\epsilon)}$ and ensuring that the independence condition $X \independent Y$ hold. This completes the proof of Corollary 3.

**Remark 4.** Corollary 3 is indeed the Sason–Verdú–Fano inequality [47, Theorem 8]. If $L = 1$, then Corollary 3 coincides with both [45, Corollary 2 in arXiv] and [47, Theorem 3]. Ultimately, if $L = 1$ and $\alpha = 1$, then Corollary 3 coincides with the original Fano inequality (1).

### III. Asymptotic Estimates on Equivocations

In information theory, the **equivocation** or the remaining **uncertainty** of an r.v. $X$ relative to a correlated r.v. $Y$ has an important role in establishing fundamental limits of the optimal transmission ratio and/or rate in several communication models. Shannon’s equivocation $H(X \mid Y)$ is a well-known measure in the formulation of the notion of perfect secrecy of symmetric-key encryption in information-theoretic cryptography [49]. Iwamoto–Shikata [34] considered the extension such a secrecy criteria by generalizing Shannon’s equivocation to Rényi’s equivocation by showing various desired properties of the latter. Recently, Hayashi–Tan [27] and Tan–Hayashi [54] studied the asymptotics of Shannon’s and Rényi’s equivocations when the side-information about the source is given via a various class of random hash functions with a fixed rate.

In this section, we assume that certain error probability vanish and we establish asymptotic estimates on Shannon’s, or sometimes on Rényi’s, equivocation via the Fano-type inequalities we established in Section II-D. Throughout this section, we use the standard asymptotic notations (cf. [7, Chapter 3]).

#### A. Fano’s Inequality meets the AEP

We consider a general form of the asymptotic equipartition property (AEP) as follows.

**Definition 7 ([58]).** We say that a sequence of $X$-valued r.v.’s $X = \{X_n\}_{n=1}^{\infty}$ satisfies the AEP if
\[ \lim_{n \to \infty} P\left\{ \log \frac{1}{P_{X_n}(X_n)} \leq (1 - \delta) H(X_n) \right\} = 0 \]  
for every fixed $\delta > 0$.

In the literature the r.v. $X_n$ is commonly represented as a random vector $X_n = (Z_1, \ldots, Z_n)$. The formulation without reference to random vectors means that $X = \{X_n\}_{n=1}^{\infty}$ is a general source in the sense of [23, p. 100].

Let $\{L_n\}_{n=1}^{\infty}$ be a sequence of positive integers, $\{Y_n\}_{n=1}^{\infty}$ a sequence of nonempty alphabets, and $\{(X_n,Y_n)\}_{n=1}^{\infty}$ a sequence of r.v.’s, where $(X_n,Y_n)$ is $\mathcal{X} \times \mathcal{Y}$-valued for $n \geq 1$. Since $\mathbb{P}\{X_n \not\in f_n(Y_n)\} = o(1)$ implies $P_{e^{(L_n)}}(X_n \mid Y_n) = o(1)$ for any sequence of list decoders $\{f_n : \mathcal{Y} \rightarrow \mathcal{X}^{(L_n)}\}_{n=1}^{\infty}$, it suffices to consider $P_{e^{(L_n)}}(X_n \mid Y_n) = o(1)$ in our analysis. The following theorem is a novel characterization of the AEP via Fano’s inequality.

**Theorem 3.** Suppose that a general source $X = \{X_n\}_{n=1}^{\infty}$ satisfies the AEP, and $H(X_n) = \Omega(1)$. Then, it holds that
\[ P_{e^{(L_n)}}(X_n \mid Y_n) = o(1) \implies |H(X_n) \mid Y_n) - \log L_n| = o(H(X_n)), \]  
where $|u| := \max\{0, u\}$ for $u \in \mathbb{R}$. Consequently,
\[ P_{e^{(L_n)}}(X_n \mid Y_n) = o(1) \quad \text{and} \quad \log L_n = o(H(X_n)) \implies H(X_n \mid Y_n) = o(H(X_n)). \]

**Proof of Theorem 3:** See Appendix C-A.

The following three examples are particularizations of Theorem 3 in increasing order of complexities.

**Example 1.** Let $\{Z_n\}_{n=1}^{\infty}$ be an i.i.d. source on a countably infinite alphabet with finite Shannon entropy $H(Z_1) < \infty$. Suppose that $X_n = (Z_1, \ldots, Z_n)$ and $Y_n = Z^n$ for each $n \geq 1$. Then, Theorem 3 states that $\mathbb{P}\{X_n \not\in Y_n\} = o(1)$ implies that $H(X_n) = o(n)$. This result is commonly referred to as the weak converse property of the source $\{Z_n\}_{n=1}^{\infty}$ in the unique decoding setting.

**Example 2.** Let $X = \{X_n\}_{n=1}^{\infty}$ be a general source as described in Example 1. Suppose that $L_n = \exp[o(n)]$. Then, even if the list decoding setting, Theorem 3 states that $P_{e^{(L_n)}}(X_n \mid Y_n) = o(1)$ implies that $H(X_n \mid Y_n) = o(n)$, similarly to Example 1.

**Example 3.** Let $X = \{X_n\}_{n=1}^{\infty}$ be a general source as described in Example 1. Suppose that $L_n = \exp[o(n)]$. Then, even if the list decoding setting, Theorem 3 states that $P_{e^{(L_n)}}(X_n \mid Y_n) = o(1)$ implies that $H(X_n \mid Y_n) = o(n)$, similarly to Example 1.
This is a key observation in Ahlswede–Gács–Körner’s proof of the strong converse property of degraded broadcast channels [2, Chapter 5] (see also [41, Section 3.6.2]).

**Example 3.** Consider the Poisson source $X = \{X_n\}_{n=1}^{\infty}$ with growing mean $\lambda_n = \omega(1)$, i.e.,

$$P_{X_n}(k) = \frac{\lambda_n^k e^{-\lambda_n}}{(k-1)!} \quad \text{for } k \in X = \{1, 2, \ldots\}. \quad (46)$$

It is known that $(H(X_n)/\log \sqrt{n}) \to 1$ as $n \to \infty$, and the Poisson source $X$ satisfies the AEP (see [58]). Therefore, it follows from Theorem 3 that $P^{(L_n)}(X_n | Y_n) = o(1)$ implies that $|H(X_n | Y_n) - \log L_n|^+ = o(\log \lambda_n)$.

The following example is a counterexample of a general source that does not satisfy both AEP and (44).

**Example 4.** Let $L \geq 1$ be an integer, $\gamma > 0$ a positive real, and $(\delta_n)_{n=1}^{\infty}$ a sequence of reals satisfying $\delta_n = o(1)$ and $0 < \delta_n < 1$ for each $n \geq 1$. Since $p \mapsto h_2(p)/p$ is continuous on $(0,1]$ and $h_2(p)/p \to \infty$ as $p \to 0^+$, one can find a sequence of reals $(p_n)_{n=1}^{\infty}$ satisfying $0 < p_n \leq \min\{1,(1-\delta_n)/(\delta_n L)\}$ for each $n \geq 1$ and

$$\frac{\delta_n h_2(p_n)}{p_n} = \gamma \quad \text{for sufficiently large } n. \quad (47)$$

Consider a general source $X = \{X_n\}_{n=1}^{\infty}$ whose component distributions are given by

$$P_{X_n}(x) = \begin{cases} \frac{1 - \delta_n}{L} & \text{if } 1 \leq x \leq L, \\ \delta_n p_n(1 - p_n)^{x-(L+1)} & \text{if } x \geq L + 1 \end{cases} \quad (48)$$

for each $n \geq 1$. Suppose that $X_n \perp Y_n$ for each $n \geq 1$. After some algebra, we have

$$P^{(L)}(X_n | Y_n) = P^{(L)}(X_n) = \delta_n, \quad (49)$$

$$H(X_n | Y_n) = H(X_n) = h_2(\delta_n) + (1 - \delta_n) \log L \frac{\delta_n h_2(p_n)}{p_n} \quad (50)$$

for each $n \geq 1$. Therefore, we observe that $P^{(L)}(X_n | Y_n) = o(1)$ holds, but $|H(X_n | Y_n) - \log L|^+ = o(H(X_n))$ does not hold. In fact, since $H(X_n) \to \gamma + \log L$ as $n \to \infty$ and

$$\lim_{n \to \infty} P_{X_n}(x) = \begin{cases} \frac{1}{L} & \text{if } 1 \leq x \leq L, \\ 0 & \text{if } x \geq L. \end{cases} \quad (51)$$

Consequently, we also see that $X = \{X_n\}_{n=1}^{\infty}$ does not satisfy the AEP.

**B. Vanishing Unnormalized Equivocation**

Let $X$ be an $X$-valued r.v. satisfying $H(X) < \infty$, $(L_n)_{n=1}^{\infty}$ a sequence of positive integers, $(Y_n)_{n=1}^{\infty}$ a sequence of nonempty alphabets, and $(\{X_n,Y_n\})_{n=1}^{\infty}$ a sequence of $X \times Y$-valued r.v.s. The following theorem provides four conditions on a general source $X = \{X_n\}_{n=1}^{\infty}$ such that vanishing error probability implies vanishing unnormalized Shannon’s and Rényi’s equivocation.

**Theorem 4.** Let $\alpha \geq 1$ be an order. Suppose that any one of the following four conditions holds:

(a) the order $\alpha$ is strictly larger than 1, i.e., $\alpha > 1$;
(b) the sequence $\{X_n\}_{n=1}^{\infty}$ satisfies the AEP and $H(X_n) = O(1)$;
(c) there exists an $n_0 \geq 1$ such that $P_{X_n}$ majorizes $P_{X}$ for every $n \geq n_0$;
(d) the sequence $\{X_n\}_{n=1}^{\infty}$ converges in distribution to $X$ and $H(X_n) \to H(X)$ as $n \to \infty$.

Then, it holds that for each $\alpha \in \{A, H\}$,

$$P^{(L_n)}(X_n | Y_n) = o(1) \quad \implies \quad |H_{\alpha}^n(X_n | Y_n) - \log L_n|^+ = o(1). \quad (52)$$

**Proof of Theorem 4:** See Appendix C-B.\[ \]

In contrast to condition (b) of Theorem 4, conditions (a), (c), and (d) of Theorem 4 do not require the AEP to hold. Interestingly, condition (a) of Theorem 4 states that vanishing error probability $P^{(L_n)}(X_n | Y_n) = o(1)$ always implies vanishing Rényi’s equivocation $|H_{\alpha}^n(X_n | Y_n) - \log L_n|^+ = o(1)$ for every $\alpha > 1$ and $\alpha \in \{A, H\}$ without any other conditions on the general source $X = \{X_n\}_{n=1}^{\infty}$.

**Remark 5.** If $L_n = 1$ for each $n \geq 1$, then conditions (c) and (d) of Theorem 4 coincide with Ho–Verdu’s result [28, Theorem 18]. Moreover, if $L_n = 1$ for each $n \geq 1$, and if $X_n$ is $(1, \ldots, M^n)$-valued for each $n \geq 1$, then condition (a) of Theorem 4 coincides with Sason–Verdu’s result [47, assertion a) of Theorem 4].
C. Under the Symbol-Wise Error Criterion

Let $L = \{L_n\}_{n=1}^{\infty}$ be a sequence of positive integers, $\{\mathcal{Y}_n\}_{n=1}^{\infty}$ a sequence of nonempty alphabets, and $\{(X_n,Y_n)\}_{n=1}^{\infty}$ a sequence of $X \times \mathcal{Y}_n$-valued r.v.’s satisfying $H(X_n) < \infty$ for every $n \geq 1$. In this subsection, we focus on the minimum arithmetic-mean probability of symbol-wise list decoding error defined as

$$P_{L}(X^n \mid Y^n) = \frac{1}{n} \sum_{i=1}^{n} P_{L}(X^n_{L_i} \mid Y_i),$$

where $X^n = (X_1, X_2, \ldots, X_n)$ and $Y^n = (Y_1, Y_2, \ldots, Y_n)$. Now, let $X$ be an $X$-valued r.v. satisfying $H(X) < \infty$. Under this symbol-wise error criterion, the following theorem holds.

**Theorem 5.** Suppose that $P_{X_n}$ majorizes $P_X$ for sufficiently large $n$. Then, it holds that

$$P_{L}(X^n \mid Y^n) = o(1) \implies \limsup_{n \to \infty} \frac{1}{n} H(X^n \mid Y^n) \leq \limsup_{n \to \infty} \log L_n.$$  

**Proof of Theorem 5:** See Appendix C-C.

It is known that the classical Fano inequality (1) can be extended from the average error criterion $P(X^n \neq Y^n)$ to the symbol-wise error criterion $(1/n)\mathbb{E}[d_H(X^n,Y^n)]$ (see [11, Corollary 3.8]), where $d_H(x^n,y^n) := \{1 \leq i \leq n \mid x_i \neq y_i\}$ stands for the Hamming distance between $x^n = (x_1, \ldots, x_n)$ and $y^n = (y_1, \ldots, y_n)$. In fact, Theorem 5 states that $\mathbb{E}[d_H(X^n,Y^n)] = o(n)$ implies that $H(X^n \mid Y^n) = o(n)$, provided that $P_{X_n}$ majorizes $P_X$ for sufficiently large $n$.

However, in the list decoding setting, we observe that $P_{L}(X^n \mid Y^n) = o(1)$ does not imply $H(X^n \mid Y^n) = o(n)$ in general. A counterexample can be readily constructed.

**Example 5.** Let $\{(X_n)_{n=1}^{\infty}\}$ be uniformly distributed Bernoulli r.v.’s, and $\{Y_n\}_{n=1}^{\infty}$ arbitrary r.v.’s. Suppose that $\{X_n, Y_n\} \nsubseteq \{X_m, Y_m\}$ if $n \neq m$, $X_n \nsubseteq Y_n$ for each $n \geq 1$, and $L_n = 2$ for each $n \geq 1$. Then, we observe that $P_{L}(X^n \mid Y^n) = 0$ for every $n \geq 1$, but $H(X^n \mid Y^n) = n \log 2$ for every $n \geq 1$.

IV. Concluding Remarks

In this study, we established generalizations of Fano’s inequality. Asymptotic estimates on Shannon’s and Rényi’s equivocations form important consequences of these generalizations.

A. Philosophy: How to Measure Conditional Information

Our Fano-type inequalities were stated in terms of the general conditional information $h_\phi(X \mid Y)$ defined in Section II-A. As shown in Section II-D, the quantity $h_\phi(X \mid Y)$ can be reduced to Shannon’s and Rényi’s information measures. Moreover, $h_\phi(X \mid Y)$ can be further reduced to the following quantities:

1) If $\phi = \| \cdot \|_{1/2}$, then $h_\phi(X \mid Y)$ coincides with the (unnormalized) **Bhattacharyya parameter** (cf. [39, Definition 17] and [52, Section 4.2.1]) defined by

$$B(X \mid Y) := \mathbb{E} \left[ \sum_{x,x' \in \mathcal{X}} \sqrt{P_{X|Y}(x) P_{X'|Y}(x')} \right].$$

When $X$ takes values in a finite alphabet with a certain algebraic structure, the Bhattacharyya parameter $B(X \mid Y)$ is useful in analyzing the speed of polarization for non-binary polar codes (cf. [39], [52]). Note that $B(X \mid Y)$ is a monotone function of Arimoto’s conditional Rényi entropy (28) of order $\alpha = 1/2$.

2) If $\phi = 1 - \| \cdot \|_{2}^{2}$, then $h_\phi(X \mid Y)$ coincides with the **conditional quadratic entropy** [8] defined by

$$H_\phi(X \mid Y) := \mathbb{E} \left[ \sum_{x \in \mathcal{X}} P_{X|Y}(x) \left( 1 - P_{X'|Y}(x) \right) \right],$$

which is used in the analysis of stochastic decoding (see, e.g., [40]). Note that $H_\phi(X \mid Y)$ is a monotone function of Hayashi’s conditional Rényi entropy (33) of order $\alpha = 2$.

3) If $X$ is $\{1, 2, \ldots, M\}$-valued, then one can define the following (variational distance-like) conditional quantity:

$$K(X \mid Y) := \mathbb{E} \left[ \frac{1}{2(M-1)} \sum_{x=1}^{M} \sum_{x'=1}^{M} \left| P_{X|Y}(x) - P_{X'|Y}(x') \right| \right].$$

5 Usually, the Bhattacharyya parameter (55) is defined so that $Z(X \mid Y) := (B(X \mid Y) - 1)/(M - 1)$, provided that $X$ is $\{0, 1, \ldots, M - 1\}$-valued. Then, it is normalized so that $0 \leq Z(X \mid Y) \leq 1$. 


Note that $0 \leq K(X \mid Y) \leq 1$. This quantity $K(X \mid Y)$ was introduced by Shuval–Tal [51] to analyze the speed of polarization of non-binary polar codes for sources with memory. When we define the function $\tilde{d} : \mathcal{P}([1,2,\ldots,M]) \to [0,1]$ by

$$\tilde{d}(P) := \frac{1}{2(M-1)} \sum_{x=1}^{M} \sum_{x'=1}^{M} |P(x) - P(x')|,$$

(58)

it holds that $K(X \mid Y) = b_{\tilde{d}}(X \mid Y)$. Clearly, the function $\tilde{d}$ is symmetric, convex, and continuous.

On the other hand, the quantity $b_{\phi}(X \mid Y)$ has the following properties that are appealing in information theory:

1) As $\phi$ is concave, lower bounded, and lower semicontinuous, it follows from Jensen’s inequality for an extended real-valued function on a closed, convex, and bounded subset of a Banach space [50, Proposition A-2] that

$$b_{\phi}(X \mid Y) \leq \phi(P_X).$$

(59)

This bound is analogous to the property that conditioning reduces entropy (cf. [9, Theorem 2.6.5]).

2) It is easy to check that for any (deterministic) mapping $g : X \to A$ with $A \subset X$, the regular conditional distribution $P_{g(X)|Y}$ majorizes $P_X|Y$ a.s. Thus, it follows from Proposition 1 that for any mapping $g : X \to A$,

$$b_{\phi}(g(X) \mid Y) \leq b_{\phi}(X \mid Y).$$

(60)

which is a counterpart of the data-processing inequality (cf. [27, Equations (26)–(28)]).

3) As shown in Section II-C, $b_{\phi}(X \mid Y)$ also satisfies an appropriate generalization of Fano’s inequality.

Therefore, similarly to the family of $f$-divergences [3], [10], the quantity $b_{\phi}(X \mid Y)$ is a generalization of various information-theoretic conditional quantities that also admit certain desirable properties. In addition, we can establish Fano-type inequalities based on $b_{\phi}(X \mid Y)$; this characterization provides insights on how to measure conditional information axiomatically.

B. Technical Contributions: A Novel Application of Majorization Theory

We proved our Fano-type inequalities via majorization theory [38]. Specifically, Theorems 1 and 2 were proved via infinite- and finite-dimensional majorization theories, respectively.

To further elaborate on our technical contributions, we would like to mention that we introduced an interesting class of $X \times Y$-valued r.v.’s $(X, Y)$. This is the so-called balanced regular conditional distributions; see the discussion above Lemma 3 in Appendix B-A. To show that the feasible region (78) satisfies this desirable property, we employed the infinite-dimensional Birkhoff’s theorem [43, Theorem 2] in Lemma 4 in Appendix B-A. This is the key idea to prove Theorem 1. While the finite-dimensional Birkhoff’s theorem [5] (see also [38, Theorem 2.A.2]) is well-known and widely-used, to the best of our knowledge, applications of the infinite-dimensional Birkhoff’s theorem are few and far between. Hence, a technical contribution of this study is an application of the infinite-dimensional Birkhoff’s theorem in the analysis of probability distributions on a countably infinite alphabet.

To prove Theorem 2 with finite $Y$, we established Lemma 6 in Appendix B-B, which is a reduction technique of our problem from a countably infinite to a finite alphabet. Based on this reduction technique, instead of using the infinite-dimensional Birkhoff’s theorem, we can employ the more common finite-dimensional Birkhoff’s theorem in Lemma 7 in Appendix B-B. This reduction technique is the key to prove Theorem 2. This reduction technique is also useful to prove Proposition 8 of Appendix D, which ensures that we can never establish any effective Fano-type inequality on $b_{\phi}(X \mid Y)$ if $\phi$ fulfills a certain postulate and $\phi(Q) = \infty$.

C. Does Vanishing Error Probability Imply Vanishing Equivocation?

In the list decoding setting, the rate of a block code with codeword length $n$, message size $M_n$, and list size $L_n$ can be defined as $(1/n) \log (M_n/L_n)$ (cf. [14]). Motivated by this, we established asymptotic estimates of this quantity in Theorems 3 and 4. We would like to emphasize that Example 2 shows that Ahlswede–Gács–Körner’s proof technique [2, Chapter 5] (see also [41, Section 3.6.2]) works for an i.i.d. source on a countably infinite alphabet, provided that the alphabets $(\mathcal{Y}_n)_{n=1}^{\infty}$ are finite.

Theorem 3 states that the asymptotic growth of $H(X_n \mid Y_n) - \log L_n$ is strictly slower than $H(X_n)$, provided that the general source $X = \{X_n\}_{n=1}^{\infty}$ satisfies the AEP and $P_e^{L_n}(X_n \mid Y_n) = o(1)$. This is a novel characterization of the AEP via Fano’s inequality. An instance of this characterization using the Poisson source (cf. [58, Example 4]) was provided in Example 3.
D. Future Works

1) While there are several studies of the reverse Fano inequalities [21], [35], [45], [47], [53], this study has focused only on the forward Fano inequality. Generalizing the reverse Fano inequality in the same spirit as was done in this study would be of interest.

2) The important technical tools used in our analysis include the finite- and infinite-dimensional Birkhoff’s theorem; they were employed to ensure that the constraint \( P_x = Q \) is satisfied. As a similar constraint is imposed in many information-theoretic problems, e.g., coupling problems (cf. [36], [46], [56]), finding further applications of these theorems would refine technical tools, and potentially results, when we are dealing with countably infinite alphabets.

3) We have described a novel connection between the AEP and Fano’s inequality in Theorem 3; its role in the classifications of sources and channels and its applications to other coding problems are of interest.

Acknowledgements

The author would like to thank Prof. Ken-ichi Iwata for his valuable comments on an earlier version of this paper. The author would also like to thank Prof. Vincent Y. F. Tan for his comments that greatly improved this paper.

Appendix A

Proof of Propositions

A. Proof of Proposition 1

In [38, Proposition 3.C.3], the assertion of Proposition 1 was proved in the case where the dimension of the domain of \( \phi \) is finite. Employing [37, Theorem 4.2] instead of [38, Corollary 2.B.3], the proof of [38, Proposition 3.C.3] can be directly extended to infinite-dimensional domains.

B. Proof of Proposition 2

The proposition is quite obvious, and we prove it to make this paper self-contained. For a given list decoder \( f : Y \rightarrow \mathcal{X}^{(L)} \) with list size \( 1 \leq L < \infty \), it follows that

\[
\mathbb{P}\{X \notin f(Y)\} = \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{X \notin f(Y)\}} | Y]] = \mathbb{E}[P_{X|Y}(X \setminus f(Y))]
\]

\[
\geq \mathbb{E}\left[ \sum_{x = L+1}^{\infty} P_{X|Y}^L(x) \right], \tag{61}
\]

where the equality of (a) can be achieved by an optimal list decoder \( f^* \) satisfying the two conditions: (i) \( |f^*(Y)| = L \) a.s.; and (ii) \( X \notin f^*(Y) \) implies \( P_{X|Y}(X) = P_{X|Y}^L(k) \) for some \( k \geq L + 1 \). This completes the proof of Proposition 2.

C. Proof of Proposition 3

The second inequality in (10) is indeed a direct consequence of Proposition 2 and (73). The sharpness of the second bound can be easily verified by setting that \( X \) and \( Y \) are statistically independent.

We next prove the first inequality in (10). When \( Y \) is infinite, the first inequality is an obvious one \( P_e^{(L)}(X | Y) \geq 0 \), and its equality holds by setting \( X \subset Y \) and \( X = Y \) a.s. Hence, it suffices to consider the case where \( Y \) is finite. Assume without loss of generality that \( Y = \{0, 1, \ldots, N - 1\} \) for some positive integer \( N \). Similar to the proof of Lemma 6, by the definition of cardinality, there exists a subset \( Z \subset X \) satisfying (i) \( |Z| = LN \), and (ii) for each \( x \in \{1, 2, \ldots, L\} \) and \( y \in \{0, 1, \ldots, N - 1\} \), there exists an element \( z \in Z \) satisfying \( P_{X|Y = y}(z) = P_{X|Y = y}^L(x) \). Then,

\[
P_e(X | Y) \geq_{(a)} \left[ 1 - \sum_{y \in Y} P_Y(y) \sum_{x = 1}^{L} P_{X|Y = y}^L(x) \right] \tag{62}
\]

\[
\geq_{(b)} 1 - \sum_{y \in Y} P_Y(y) \sum_{x \in Z} P_{X|Y = y}(x)
\]

\[
= 1 - \sum_{x \in Z} P_X(x)
\]

\[
\geq_{(c)} 1 - \sum_{x = 1}^{LN} Q^L(x),
\]
where (a) follows from Proposition 2; (b) follows from the construction of \( \mathcal{Z} \); and (c) follows from the facts that \( |\mathcal{Z}| = LN \) and \( P_X = Q \). This is indeed the first inequality in (10). Finally, the sharpness of the first inequality can be verified by the \( X \times \mathcal{Y} \)-valued r.v. \((U, V)\) determined by the joint probability distribution \( P_{U,V}\):

\[
P_{U|V=v}(u) = \begin{cases} 
\frac{\omega_2(Q, L)}{\omega_1(Q, v, L)} Q^1(u) & \text{if } vL < u \leq (1 + v)L, \\
\frac{\omega_1(Q, v, L)}{\omega_2(Q, L)} Q^1(x) & \text{if } LN < u < \infty, \\
0 & \text{otherwise,}
\end{cases}
\]

where \( \omega_1(Q, v, L) \) and \( \omega_2(Q, L) \) are defined by

\[
\omega_1(Q, v, L) := \sum_{u=1+vL}^{(1+v)L} Q^1(u),
\]

\[
\omega_2(Q, L) := \sum_{i=0}^{LN-1} \omega_1(Q, v, L).
\]

A direct calculation shows that \( P_U = Q^1 \) and

\[
P_e^{(L)}(U | V) = 1 - \sum_{x=1}^{LN} Q^1(x),
\]

which implies the sharpness of the first inequality. This completes the proof of Proposition 3.

\section*{APPENDIX B}

\textbf{PROOFS OF FANO-TYPE INEQUALITIES}

In this section, we prove our Fano-type inequalities presented in Section II-C via majorization theory [38].

\subsection*{A. Proof of Theorem 1}

In this subsection, we reduce the maximization problem of (19) in Theorem 1 via some useful lemmas, i.e., our preliminary results. We first give an elementary fact of the weak majorization on the finite-dimensional real vectors.

\textbf{Lemma 1.} Let \( \mathbf{p} = (p_i)_{i=1}^n \) and \( \mathbf{q} = (q_i)_{i=1}^n \) be \( n \)-dimensional real vectors satisfying \( 0 \leq p_1 \leq p_2 \leq \cdots \leq p_n \) and \( 0 \leq q_1 \leq q_2 \leq \cdots \leq q_n \), respectively, and \( k \in \{1, \ldots, n\} \) an integer satisfying \( q_k = q_i \) for every \( i = k, k+1, \ldots, n \). If

\[
\sum_{i=1}^{j} p_i \geq \sum_{i=1}^{j} q_i \quad \text{for } j = 1, 2, \ldots, k-1,
\]

\[
\sum_{i=1}^{n} p_i \geq \sum_{i=1}^{n} q_i,
\]

then it holds that

\[
\sum_{i=1}^{j} p_i \geq \sum_{i=1}^{j} q_i \quad \text{for } j = 1, 2, \ldots, n.
\]

\textbf{Proof of Lemma 1:} This lemma is quite trivial, but we prove it to make the paper self-contained. Actually, this can be directly proved by contradiction. Suppose that (68) and (69) hold, but (70) does not hold. Since (70) does not hold, there must exist an \( l \in \{k, k+1, \ldots, n-1\} \) satisfying

\[
\sum_{i=1}^{l} p_i < \sum_{i=1}^{l} q_i.
\]

Since \( q_j \) is constant for each \( j = k, k+1, \ldots, n \), it follows from (68) that \( p_j < q_j \) for every \( j = l, l+1, \ldots, n \). Then, we observe that

\[
\sum_{i=1}^{n} p_i < \sum_{i=1}^{n} q_i,
\]

which contradicts to the hypothesis (69); and therefore, Lemma 1 must hold.
For an alphabet $\mathcal{Z}$, we say that a $\mathcal{Z}$-valued r.v. $Z$ is *almost surely constant* if $Z = z$ a.s. for some $z \in \mathcal{Z}$. Then, we give the following simple lemma.

**Lemma 2.** If $P_{X|Y}^1$ is almost surely constant, then $P_{X|Y}^1$ majorizes $P_X$ a.s.

Note that $P_{X|Y}^1$ is a $\sigma(Y)$-measurable r.v. satisfying $P_{X|Y}^1 \in \mathcal{P}(X)$ a.s. While an almost surely constant $P_{X|Y}$ implies the independence $X \indep Y$, note also that an almost surely constant $P_{X|Y}^1$ does not imply the independence.

*Proof of Lemma 2:* Since $P_X(x) = \mathbb{E}[P_{X|Y}(x)]$ for each $x \in X$, it can be verified by induction that

$$
\sum_{x=1}^k P_{X}^1(x) \leq \mathbb{E}\left[\sum_{x=1}^k P_{X|Y}^1(x)\right]
$$

(73)

for every $k \geq 1$. If $P_{X|Y}^1$ is almost surely constant, then (73) implies that

$$
\sum_{x=1}^k P_{X}^1(x) \leq \sum_{x=1}^k P_{X|Y}^1(x) \text{ (a.s.)}
$$

(74)

for every $k \geq 1$, which is indeed the majorization relation of Definition 1, completing the proof of Lemma 2.

Consider a set $\mathcal{A}$ of $X \times Y$-valued r.v.’s. We say that $\mathcal{A}$ has *balanced regular conditional distributions* if it satisfies the following: if $(X, Y) \in \mathcal{A}$, then there exists $(U, V) \in \mathcal{A}$ such that

$$
P_{U|V}^1(x) = \mathbb{E}[P_{X|Y}^1(x)] \text{ (a.s.)}
$$

(75)

for every $x \in X$. For such a collection $\mathcal{A}$, the following lemma holds.

**Lemma 3.** Suppose that $\mathcal{A}$ has balanced regular conditional distributions. There exists a pair $(U, V) \in \mathcal{A}$ such that

$$
b_\phi(U \mid V) = \max_{(X, Y) \in \mathcal{A}} b_\phi(X \mid Y)
$$

(76)

and $P_{U|V}^1$ is almost surely constant. In particular, whenever the concavity of $\phi$ is strict, a pair $(U, V) \in \mathcal{A}$ satisfies (76) only if $P_{U|V}^1$ is almost surely constant.

*Proof of Lemma 3:* For any $(X, Y) \in \mathcal{A}$, it holds that

$$
b_\phi(X \mid Y) \overset{(a)}{=} \mathbb{E}\left[\phi(P_{X|Y}^1)\right]
\leq \phi(\mathbb{E}[P_{X|Y}^1]) 
\overset{(b)}{=} \phi(P_{U|V}^1) \text{ (a.s.)} 
\overset{(c)}{=} \mathbb{E}\left[\phi(P_{U|V}^1)\right] \text{ (a.s.)} 
= b_\phi(U \mid V),
$$

(77)

where (a) follows by the symmetry of $\phi$; (b) follows by Jensen’s inequality [50, Proposition A-2]; (c) follows by the existence of $(U, V) \in \mathcal{A}$ satisfying (75); and (d) follows by the symmetry of $\phi$ again. Since (75) implies that $P_{U|V}^1$ is almost surely constant, we have (76). The last assertion of Lemma 3 follows from the equality condition of Jensen’s inequality for a strict concave function $\phi$. This completes the proof of Lemma 3.

For a system $(Q, L, \epsilon, \mathcal{Y})$ fulfilling (12), we now define the collection of r.v.’s

$$
\mathcal{R}(Q, L, \epsilon, \mathcal{Y}) := \left\{(X, Y) \mid (X, Y) \text{ is } X \times \mathcal{Y}\text{-valued,} \right. 
\left. P_{\epsilon}^{L}(X \mid Y) \leq \epsilon, 
\left. P_X = Q \right) \right\},
$$

(78)

which is indeed the feasible region of the maximum in the Fano-type inequality (19) of Theorem 1. The main idea of proving Theorem 1 is to apply Lemma 3 for this collection of r.v.’s. The correction $\mathcal{R}(Q, L, \epsilon, \mathcal{Y})$ does not, however, have balanced regular conditional distributions of (75) in general. Fortunately, the following lemma can avoid this issue by blowing-up the collection $\mathcal{R}(Q, L, \epsilon, \mathcal{Y})$ via infinite-dimensional Birkhoff’s theorem [43, Theorem 2].

**Lemma 4.** If $|\mathcal{Y}| \geq \epsilon$, then there exists a $\sigma$-algebra on $\mathcal{Y}$ such that the collection $\mathcal{R}(Q, L, \epsilon, \mathcal{Y})$ has balanced regular conditional distributions.

6This was conjectured by Birkhoff [6, p. 266], and was solved by Révész [43, Theorems 1–2].
Proof of Lemma 4: Firstly, we shall choose an appropriate alphabet \( \mathcal{Y} \) so that \( |\mathcal{Y}| = \mathfrak{c} \). Denote by \( \Psi \) the set of \( \infty \times \infty \) permutation matrices, where an \( \infty \times \infty \) permutation is a real matrix \( \Pi = \{\pi_{i,j}\}_{i,j=1}^{\infty} \) satisfying either \( \pi_{i,j} = 0 \) or \( \pi_{i,j} = 1 \) for each \( 1 \leq i, j < \infty \), and

\[
\sum_{j=1}^{\infty} \pi_{i,j} = 1 \quad \text{for each } 1 \leq i < \infty, \tag{79}
\]

\[
\sum_{i=1}^{\infty} \pi_{i,j} = 1 \quad \text{for each } 1 \leq j < \infty. \tag{80}
\]

For an \( \infty \times \infty \) permutation matrix \( \Pi = \{\pi_{i,j}\}_{i,j} \in \Omega \), define the permutation \( \psi_{\Pi} \) on \( X = \{1, 2, \ldots\} \) by

\[
\psi_{\Pi}(i) := \sum_{j=1}^{\infty} \pi_{i,j} j. \tag{81}
\]

It is known that there is a one-to-one correspondence between the permutation matrices \( \Pi \) and the bijections \( \psi_{\Pi} \); and thus, \( |\Psi| = \mathfrak{c} \). Therefore, in this proof, we may assume without loss of generality that \( \mathcal{Y} = \Psi \).

Secondly, we shall construct an appropriate \( \sigma \)-algebra on \( \mathcal{Y} \) by infinite-dimensional Birkhoff’s theorem \([43, \text{Theorem } 2]\) for \( \infty \times \infty \) doubly stochastic matrices, where an \( \infty \times \infty \) doubly stochastic matrix is a real matrix \( \mathbf{M} = \{m_{i,j}\}_{i,j=1}^{\infty} \) satisfying \( 0 \leq m_{i,j} \leq 1 \) for each \( 1 \leq i, j < \infty \), and

\[
\sum_{j=1}^{\infty} m_{i,j} = 1 \quad \text{for each } 1 \leq i < \infty, \tag{82}
\]

\[
\sum_{i=1}^{\infty} m_{i,j} = 1 \quad \text{for each } 1 \leq j < \infty. \tag{83}
\]

Similar to \( \Psi \), denote by \( \Psi_{i,j} \) the set of \( \infty \times \infty \) permutation matrices in which the entry in the \( i \)th row and the \( j \)th column is 1, where note that \( \Psi_{i,j} \subset \mathcal{Y} \). By infinite-dimensional Birkhoff’s theorem \([43, \text{Theorem } 2]\), there exists a \( \sigma \)-algebra \( \Gamma \) on \( \mathcal{Y} \) satisfying the two assertions: (i) \( \Psi_{i,j} \in \Gamma \) for each \( 1 \leq i, j < \infty \); and (ii) for any \( \infty \times \infty \) doubly stochastic matrix \( \mathbf{M} = \{m_{i,j}\}_{i,j=1}^{\infty} \), there exists a \( \mathcal{Y} \)-valued r.v. \( Z \) satisfying \( P_Z(\Psi_{i,j}) = m_{i,j} \) for every \( 1 \leq i, j < \infty \). In other words, this is a probabilistic description of an \( \infty \times \infty \) doubly stochastic matrix via a probability measure on the \( \infty \times \infty \) permutation matrices. We employ this \( \sigma \)-algebra \( \Gamma \) on \( \mathcal{Y} \) in the proof.

Thirdly, we shall show that under this measurable space \( (\mathcal{Y}, \Gamma) \), the collection \( \mathcal{R}(Q, L, \varepsilon, \mathcal{Y}) \) has balanced regular conditional distributions defined in \((75)\). In other words, for a given \( X \times \mathcal{Y} \)-valued \((X, Y)\) satisfying \( P^L_{\varepsilon}(X \mid Y) \leq \varepsilon \) and \( P_X = Q \), it suffices to construct another \( X \times \mathcal{Y} \)-valued r.v. \((U, V)\) satisfying the three conditions: (i) Equation \((75)\) holds; (ii) \( P^L_{\varepsilon}(U \mid V) \leq \varepsilon \); and (iii) \( P_U = Q \). At first, construct its regular conditional distribution \( P^\dagger_{U \mid V} \) by

\[
P_{U \mid V}(x) = \mathbb{E}\left[P^\dagger_{X \mid Y}(\psi_{V}(x)) \mid \sigma(V)\right] \quad \text{(a.s.)} \tag{84}
\]

for each \( x \in X \). Since \( P^\dagger_{U \mid V}(x) = P_{U \mid V}(\psi_{V}(x)) \) a.s. for every \( x \in X \), we readily see that \((75)\) holds, and \( P^\dagger_{U \mid V} \) is almost surely constant. Thus, it follows by \((73)\) and the hypothesis \( P_X = Q \) that \( P^\dagger_{U \mid V} \) majorizes \( Q \) a.s. Therefore, it follows from the characterization of the majorization relation via \( \infty \times \infty \) doubly stochastic matrices \([37, \text{Lemma } 3.1] \) (see also \([38, \text{p. } 25]\)) that one can find an \( \infty \times \infty \) doubly stochastic matrix \( \mathbf{M} = \{m_{i,j}\}_{i,j=1}^{\infty} \) satisfying

\[
Q(i) = \sum_{j=1}^{\infty} m_{i,j} P^\dagger_{U \mid V}(j) \quad \text{(a.s.)} \tag{85}
\]

for every \( i \geq 1 \). By infinite-dimensional Birkhoff’s theorem \([43, \text{Theorem } 2]\), we can construct an induced probability measure \( P_V \) so that \( P_V(\Psi_{i,j}) = m_{i,j} \) for each \( 1 \leq i, j < \infty \). It remains to verify that \( P^L_{\varepsilon}(U \mid V) \leq \varepsilon \) and \( P_U = Q \).
Since $\psi_{\Pi}$ is a permutation defined in (81), we have

\[
P_e^{(L)}(X \mid Y) \overset{(a)}{=} 1 - \mathbb{E}\left[ \sum_{x=1}^{L} P_{X|Y}^{1}(x) \right] = 1 - \mathbb{E}\left[ \sum_{x=1}^{L} \mathbb{E}[P_{X|Y}^{1}(x) \mid \sigma(V)] \right] \overset{(b)}{=} 1 - \mathbb{E}\left[ \sum_{x=1}^{L} P_{U|V}(\psi_{\Pi}^{-1}(x)) \right] \overset{(c)}{=} 1 - \mathbb{E}\left[ \sum_{x=1}^{L} P_{U|V}(x) \right] \overset{(d)}{=} P_e^{(L)}(U \mid V),
\]

where (a) and (d) follow from Proposition 2; and (b) and (c) follow from (84). Hence, we see that $P_e^{(L)}(X \mid Y) \leq \epsilon$ is equivalent to $P_e^{(L)}(U \mid V) \leq \epsilon$. Furthermore, we observe that

\[
Q(i) \overset{(85)}{=} \sum_{j=1}^{\infty} m_{i,j} P_{U|V}(j) \quad \text{(a.s.)}
\]

\[
\overset{(a)}{=} \sum_{j=1}^{\infty} \mathbb{E}\left[ \mathbf{1}_{V \in \Psi_{i,j}} \right] P_{U|V}(j) \quad \text{(a.s.)}
\]

\[
\overset{(b)}{=} \sum_{j=1}^{\infty} \mathbb{E}\left[ \mathbf{1}_{V \in \Psi_{i,j}} P_{U|V}(j) \right] \quad \text{(a.s.)}
\]

\[
\overset{(c)}{=} \sum_{j=1}^{\infty} \mathbb{E}\left[ \mathbf{1}_{V \in \Psi_{i,j}} P_{U|V}(\psi_{\Pi}^{-1}(j)) \mid \sigma(V) \right] \overset{(d)}{=} \sum_{j=1}^{\infty} \mathbb{E}\left[ \mathbf{1}_{V \in \Psi_{i,j}} P_{U|V}(\psi_{\Pi}^{-1}(j)) \mid \sigma(V) \right] \overset{(e)}{=} \sum_{j=1}^{\infty} \mathbb{E}\left[ \mathbf{1}_{V \in \Psi_{i,j}} P_{U|V} \left( \sum_{k=1}^{\infty} \mathbf{1}_{V \in \Psi_{k,j}} \right) \mid \sigma(V) \right] \overset{(f)}{=} \sum_{j=1}^{\infty} \mathbb{E}\left[ \mathbf{1}_{V \in \Psi_{i,j}} P_{U|V} \left( \sum_{k=1}^{\infty} \mathbf{1}_{V \in \Psi_{k,j}} \right) \right] \overset{(g)}{=} \sum_{j=1}^{\infty} \mathbb{E}\left[ \mathbf{1}_{V \in \Psi_{i,j}} P_{U|V} \left( \sum_{k=1}^{\infty} \mathbf{1}_{V \in \Psi_{k,j}} \right) \right] \overset{(h)}{=} \mathbb{E}\left[ P_{U|V}(i) \right] = P_U(i)
\]

for every $i \geq 1$, where (a) follows by the identity $m_{i,j} = \mathbb{P}(V \in \Psi_{i,j})$; (b) follows from the fact that $P_{U|V}$ is almost surely constant; (c) follows from the fact that (84) implies $P_{U|V}(j) = P_{U|V}(\psi_{\Pi}^{-1}(j))$ a.s. for every $j \geq 1$; (d) follows by the definition of $V_{i,j}$; (e) follows from the fact that the inverse of a permutation matrix is its transpose; and (f) follows by the Fubini–Tonelli theorem. Therefore, we have $P_U = Q$, and the assertion of Lemma 4 is proved in the case where $|\mathcal{Y}| = \epsilon$.

Finally, note that the assertion of Lemma 4 with $|\mathcal{Y}| > \epsilon$ can be immediately proved by considering the trace $\mathcal{Y} \cap \Psi_i$. This completes the proof of Lemma 4.

Finally, we show some majorization relations of Definition 1 to the Fano-type distribution of type-1 defined in (13). It is clear by the definition that $P_{\text{Fano-type}1}^{(Q,L,e)}$ majorizes $Q$ (see also Fig. 1). The following lemma is a final tool to prove Theorem 1.

**Lemma 5.** Suppose that $(Q,L,e)$ fulfills (12) with $|\mathcal{Y}| = \infty$, i.e., $0 \leq \epsilon \leq P_e^{(L)}(Q)$. For every $X$-marginal $R$ in which $R$ majorizes $Q$ and $P_e^{(L)}(R) \leq \epsilon$, it holds that $R$ majorizes $P_{\text{Fano-type}1}^{(Q,L,e)}$, as well, where $P_{\text{Fano-type}1}^{(Q,L,e)}$ is defined in (13).
Proof of Lemma 5: For simplicity, we write \( P_1 = P_{1,\text{Fano-type}}(Q,L,e) \) in this proof. By the definition (13), we readily see that \( P_{1,\text{Fano-type}} = P_1 \). Hence, it suffices to prove that for every \( k \geq 1 \),

\[
\sum_{x=1}^{k} P_1(x) \leq \sum_{x=1}^{k} R^i(x). \tag{88}
\]

Since \( P_1(x) = Q^i(x) \) for each \( 1 \leq x < J \), it follows by the hypothesis \( Q < R \) that (88) holds for each \( 1 \leq x < J \). Moreover, since \( P_e(L)(P_1) = e \), it follows by the hypothesis \( P_e(L)(R) \leq e \leq P_e(L)(Q) \) that

\[
\sum_{x=J}^{L} P_1(x) \leq \sum_{x=J}^{L} R^i(x). \tag{89}
\]

In addition, since (88) holds for each \( 1 \leq x < J \) and \( P_1(x) = \mathcal{W}(J) \) for each \( J \leq x \leq L \), it follows from Lemma 1 and (89) that (88) also holds for each \( 1 \leq k \leq L \). If \( K_1 = \infty \), then \( P_1(x) = \mathcal{W}(\infty) = 0 \) for each \( x \geq L + 1 \); and thus, Inequality (88) holds for every \( k \geq 1 \) if \( K_1 = \infty \). Now, it remains to prove the case where \( K_1 < \infty \).

Since \( P_1(x) = Q^i(x) \) for each \( x \geq K_1 + 1 \), it follows by the hypothesis \( Q < R \) that (88) holds for every \( k \geq K_1 \). Moreover, since (88) holds for every \( 1 \leq x \leq L \) and \( x \geq K_1 \), we observe that

\[
\sum_{x=L+1}^{K_1} P_1(x) \leq \sum_{x=L+1}^{K_1} R^i(x). \tag{90}
\]

Finally, since (88) holds for \( 1 \leq x \leq L \) and \( P_1(x) = \mathcal{W}(K_1) \) for \( L < x \leq K_1 \), it follows by Lemma 1 and (90) that (88) holds for every \( 1 \leq x \leq K_1 \). Therefore, Inequality (88) holds for \( k \geq 1 \), and Lemma 5 is just proved.

Using the above lemmas, we can prove Theorem 1 as follows:

Proof of Theorem 1: For short, we write \( \mathcal{R} = \mathcal{R}(Q,L,e,\mathcal{Y}) \) and \( P_1 = P_{1,\text{Fano-type}}(Q,L,e) \) in the proof. Let \( \mathcal{T} \) be a \( \sigma \)-algebra on \( \mathcal{Y} \), \( \Psi \) an alphabet satisfying \( |\Psi| = \epsilon \), and \( \Gamma \) a \( \sigma \)-algebra on \( \Psi \) so that \( \mathcal{R}(Q,L,e,\mathcal{Y}) \) has balanced regular conditional distributions (see Lemma 4). Now, we define the set \( \mathcal{R}^e := \mathcal{R}(Q,L,e,\mathcal{Y} \cup \Psi) \), where the \( \sigma \)-algebra on \( \mathcal{Y} \cup \Psi \) is given by the smallest \( \sigma \)-algebra \( \mathcal{T} \vee \Gamma \) containing \( \mathcal{T} \) and \( \Gamma \). It is clear that \( \mathcal{R} \subset \mathcal{R}^e \), and \( \mathcal{R}^e \) has balanced regular conditional distributions as well (see the last paragraph of the proof of Lemma 4). Then, we have

\[
\sup_{(X,Y):P_e(L)(X) \leq \epsilon, P_X = Q}(X,Y) = \sup_{(X,Y) \in \mathcal{R}} b_{\phi}(X \mid Y) \tag{a} \\
\leq \sup_{(X,Y) \in \mathcal{R}} b_{\phi}(X \mid Y) \tag{b} \\
= \sup_{(X,Y) \in \mathcal{R}} b_{\phi}(X \mid Y), \tag{c}
\]

where (a) follows by the definition (78); (b) follows by inclusion \( \mathcal{R} \subset \mathcal{R}^e \); (c) follows from Lemma 3 the fact that \( \mathcal{R}^e \) has balanced regular conditional distributions; (d) follows by the symmetric of \( \phi : \mathcal{P}(X) \rightarrow [0, \infty] \) and \( P_e(L) : \mathcal{P}(X) \rightarrow [0, 1] \); (e) follows from Lemma 2; and (f) follows from Proposition 1 and Lemma 5. Inequalities (91) are indeed the Fano-type inequality (19) of Theorem 1.

Henceforth, we verify the sharpness conditions 1)–4) of Theorem 1. If \( \epsilon = P_e(L)(Q) \), then it can be verified by the definition (13) that \( P_1 = Q \) (see also Fig. 1). In such a case, the maximum in (19) can be achieved by a pair \( (X,Y) \) satisfying \( P_X = Q \) and the independence \( X \perp Y \). This implies the sharpness condition 1).

We next verify the sharpness conditions 2) and 3). If \( P_e(L)(Q) > 0 \) or \( \#(\text{supp } Q) < \infty \), then it can be verified by the definition (17) that \( K_1 < \infty \). This implies that, in other words, there are at most finitely-many elements \( x \in X \) satisfying \( P_1(x) \neq Q^i(x) \). Moreover, it is clear by the definition (13) that \( P_1 \) majorizes \( Q \) (see also Fig. 1). Therefore, roughly speaking, we can prove the sharpness conditions 2) and 3) by applying both Hardy–Littlewood–Pólya theorem [25, Theorem 8] (see also...
Farahat–Mirska’s refinement of finite-dimensional Birkhoff’s theorem [20, Theorem 3] (see also [38, Theorem 2.6.2]) to the $X$-marginals $P_1$ and $Q$. We shall do it as follows: Supposing that $\mathcal{Y} = \{1, 2, \ldots, (K_1 - J) + 1\}$, we now construct a pair $(X, Y)$ satisfying (i) $P^1_X = Q_1^1$; (ii) $P_e^{(L)}(X \mid Y) = \varepsilon$; and (iii) $b_\phi(X \mid Y) = \phi(p_1)$. Since $\mathcal{Y}$ is countable, note that $P_{X \mid Y}$ can be handled as an (ordinary) conditional distribution, rather than a regular conditional distribution. By the definition (13), we observe that

$$
\begin{align*}
\sum_{x \in J} Q^1(x) & \leq \sum_{x \in J} P_1(x) \quad \text{for } J \leq k \leq K_1, \\
\sum_{x \in J} Q^1(x) & = \sum_{x \in J} P_1(x).
\end{align*}
$$

Equations (92) and (93) are indeed a majorization relation between two $(K_1 - J + 1)$-dimensional real vectors; and thus, it follows from the Hardy–Littlewood–Pólya theorem that there exists a $(K_1 - J + 1) \times (K_1 - J + 1)$ doubly stochastic matrix $M = \{m_{i,j}\}_{i,j=1}^{K_1}$ satisfying

$$
Q^1(i) = \sum_{j=1}^{K_1} m_{i,j} P_1(j)
$$

for each $J \leq i \leq K_1$. Moreover, it follows from Farahat–Mirska’s refinement of finite-dimensional Birkhoff’s theorem that for such a doubly stochastic matrix $M = \{m_{i,j}\}_{i,j=1}^{K_1}$, there exists a pair of a $|\mathcal{Y}|$-dimensional probability vector $\lambda = (\lambda_y)_{y \in \mathcal{Y}}$ and a collection of $(K_3 - J + 1) \times (K_1 - J + 1)$ permutation matrices $\{\pi_{i,j}^{(y)}\}_{i,j=1}^{K_1}$ satisfying

$$
m_{i,j} = \sum_{y \in \mathcal{Y}} \lambda_y \pi_{i,j}^{(y)},
$$

for every $1 \leq i, j \leq K_1 - J + 1$. Using them, we construct an $X \times \mathcal{Y}$-valued r.v. $(X, Y)$ via the following distributions:

$$
P_{X \mid Y = y}(x) = \begin{cases} P_1(x) & \text{if } 1 \leq x < J \text{ or } K_1 < x < \infty, \\ P_1(\tilde{\psi}_y(x)) & \text{if } J \leq x \leq K_1, \end{cases}
$$

$$
P_Y(y) = \lambda_y.
$$

where the permutation $\tilde{\psi}_y$ on $\{J, J + 1, \ldots, K_1\}$ is defined by

$$
\tilde{\psi}_y(i) = \sum_{j=1}^{K_1} \pi_{i,j}^{(y)} j
$$

for each $y \in \mathcal{Y}$. Then, it follows from (94) and (95) that $P_X = Q^1$. Moreover, it is easy to see that $P_{X \mid Y = y}^1 = P_1$ for every $y \in \mathcal{Y}$. Therefore, we observe that $P_e^{(L)}(X \mid Y) = P_e^{(L)}(P_1) = \varepsilon$ and $b_\phi(X \mid Y) = \phi(p_1)$, which implies that $(X, Y)$ satisfies the Fano-type inequality (19) with equality. Furthermore, since $P_1(x) = \mathcal{W}(J)$ for $J \leq x \leq L$ and $P_1(x) = \mathcal{W}(K_1)$ for $L < x \leq K_1$, it follows that the distributions $\{P_{X \mid Y = y}\}_{y \in \mathcal{Y}}$ are at most $(K_1 - J + 1)$ distinct distributions. Namely, the number $|\mathcal{Y}| \geq (K_1 - J + 1)$ is also sufficient, which yields the sharpness conditions 2) and 3) of Theorem 1.

Moreover, we shall prove the sharpness condition 4). If $K_1 < \infty$, then the sharpness condition 4) is an immediate consequence from the previous paragraph proving the sharpness conditions 2) and 3). Therefore, it suffices to consider the case where $K_1 = \infty$. In this case, by the definition (17), it must be satisfied that $\#(\text{supp}(\mathcal{Q})) = \infty$, $\varepsilon = 0$, $\mathcal{W}(J) > 0$, and $\mathcal{W}(K_1) = 0$. Suppose that $\mathcal{Y} = \{L, L + 1, L + 2, \ldots\} \subset X$, i.e., $|\mathcal{Y}| = \aleph_0$. We then construct an $X \times \mathcal{Y}$-valued r.v. $(X, Y)$ via the following distributions:

$$
P_{X \mid Y = y}(x) = \begin{cases} Q^1(x) & \text{if } 1 \leq x < L, \\ \mathcal{W}(J) & \text{if } L \leq x < \infty \text{ and } x = y, \\ 0 & \text{if } L \leq x < \infty \text{ and } x \neq y,
\end{cases}
$$

$$
P_Y(y) = \frac{Q^1(y)}{\mathcal{W}(J)}.
$$

We readily see that $P_{X \mid Y = y}^1 = P_1$ for every $y \in \mathcal{Y}$, and $P_X = Q^1$. Therefore, we observe that $P_e^{(L)}(X \mid Y) = P_e^{(L)}(P_1) = \varepsilon$ and $b_\phi(X \mid Y) = \phi(p_1)$, which implies that $(X, Y)$ satisfies the Fano-type inequality (19) with equality. This is the sharpness condition 4) of Theorem 1.

Furthermore, supposing that $|\mathcal{Y}| \geq \varepsilon$, we shall show the existence of a $\sigma$-algebra on $\mathcal{Y}$ satisfying (19) with equality. Similar to the proof of Lemma 4, it suffices to consider the case where $\mathcal{Y}$ is the set of $\infty \times \infty$ permutation matrices, and its corresponding
σ-algebra is given by infinite-dimensional Birkhoff’s theorem [43, Theorem 2]. Suppose that $P_{X|Y}^t = P_t$ a.s. In this case, it is easy to see that $P_e^{(L)}(X|Y) = P_e^{(L)}(P_t) = \varepsilon$ and $b_\phi(X|Y) = \phi(P_t)$. Moreover, since $P_t$ majorizes $Q$, it follows from infinite-dimensional Birkhoff’s theorem [43, Theorem 2] and the characterization of the majorization relation via $\infty \times \infty$ doubly stochastic matrices [37, Lemma 3.1] (see also [38, p. 25]), we can find an induced probability measure $P_Y$ satisfying $P_X = Q$ (see also the proof of Lemma 4). Hence, the assertion holds.

Finally, we shall verify that the sharpness (20) is a sufficient condition on $(X,Y)$ achieving the supremum in (19). In fact, it is easy to see that if (20) holds, then $P_e^{(L)}(X|Y) = P_e^{(L)}(P_t) = \varepsilon$ and $b_\phi(X|Y) = \phi(P_t)$, as we have seen so far. That is, our Fano-type inequality (19) tells us that (20) is indeed a sufficient condition. Furthermore, it follows from Lemma 3 that whenever the concavity of $\phi$ is strict, an r.v. $(X,Y)$ achieves the supremum in (19) only if $P_{X|Y}^t$ is almost surely constant. Therefore, whenever the concavity of $\phi$ is strict, Equation (20) is the necessary and sufficient condition. This completes the proof of Theorem 1.

### B. Proof of Theorem 2

To prove Theorem 2, we need some more preliminary results. Throughout this subsection, assume that the alphabet $\mathcal{Y}$ is finite and nonempty. In this case, note that for any $X \times \mathcal{Y}$-valued r.v. $(X,Y)$, one can consider $P_{X|Y}$ as an (ordinary) conditional distribution rather than a regular conditional distribution. Namely, suppose throughout this subsection that $P_{X|Y = y}(x) \geq 0$ for every $(x,y) \in X \times \mathcal{Y}$ and $\sum_{x \in X} P_{X|Y = y}(x) = 1$ for every $y \in \mathcal{Y}$.

For a subset $Z \subset X$, define

$$P_e^{(L)}(X | Y \parallel Z) := \min_{f : Y \rightarrow Z^{(L)}} \mathbb{P}(X \notin f(Y)),$$

(101)

where $Z^{(L)}$ is defined in the same manner as (6). Note that the difference between $P_e^{(L)}(X | Y)$ and $P_e^{(L)}(X | Y \parallel Z)$ is the restriction of the decoding range $Z \subset X$, and the inequality $P_e^{(L)}(X | Y) \geq P_e^{(L)}(X | Y \parallel Z)$ is trivial from those definitions (5) and (101). The following propositions are easy consequences of the proofs of Propositions 2 and 3, and so we omit those proofs in this paper.

**Proposition 6.** It holds that

$$P_e^{(L)}(X | Y \parallel Z) = 1 - \mathbb{E} \left[ \min_{D \in Z^{(L)}} \sum_{x \in D} P_{X|Y}(x) \right].$$

(102)

**Proposition 7.** Let $\beta : \{1, \ldots, \#(Z)\} \rightarrow Z$ be a bijection satisfying $P_X(\beta(i)) \geq P_X(\beta(j))$ if $i < j$. It holds that

$$1 - \sum_{x \in Z} P_X(x) \leq P_e^{(L)}(X | Y) \leq 1 - \sum_{x \in Z} P_X(\beta(x)).$$

(103)

For a finite subset $Z \subset X$, denote by $\Psi(Z)$ the set of $\#(Z) \times \#(Z)$ permutation matrices in which both rows and columns are indexed by the elements in $Z$. The main idea of proving Theorem 2 is the following lemma.

**Lemma 6.** For any $X \times \mathcal{Y}$-valued r.v. $(X,Y)$, there exist a subset $Z \subset X$ satisfying $\#(Z) = \#(\mathcal{Y})L$ and an $X \times \Psi(Z)$-valued r.v. $(U,W)$ such that

$$P_U(x) = P_X(x) \quad \text{for } x \in X,$$

(104)

$$P_e^{(L)}(U | W \parallel Z) = P_e^{(L)}(X | Y),$$

(105)

$$h_\phi(U | W) \geq b_\phi(X | Y),$$

(106)

$$P_{U|W=w}(x) = P_X(x) \quad \text{for } x \in X \setminus Z \text{ and } w \in \Psi(Z).$$

(107)

**Proof of Lemma 6:** Suppose without loss of generality that $\mathcal{Y} = \{0, 1, \ldots, N - 1\}$ for some positive integer $N$. By the definition of cardinality, one can find a subset $Z \subset X$ satisfying (i) $\#(Z) = LN$, and (ii) for each $x \in \{1, 2, \ldots, L\}$ and $y \in \{0, 1, \ldots, N - 1\}$, there exists an element $z \in Z$ satisfying $P_{X|Y = y}(z) = P_{X|Y = y}(x)$. For each $\Pi = \{\pi_{i,j}\}_{i,j \in \mathbb{Z}} \in \Psi(Z)$, one can define the permutation $\varphi_\Pi : Z \rightarrow Z$ by

$$\varphi_\Pi(z) := \sum_{w \in Z} \pi_{z,w} w,$$

(108)

as in (81) and (98). It is clear that for each $y \in \mathcal{Y}$, there exists at least one $\Pi \in \Psi(Z)$ such that $P_{X|Y = y}(\varphi_\Pi(x_1)) \leq P_{X|Y = y}(\varphi_\Pi(x_2))$ for every $x_1, x_2 \in Z$ satisfying $x_1 \leq x_2$, which implies that the permutation $\varphi_\Pi$ plays a role of a decreasing rearrangement of $P_{X|Y = y}$ on $Z$. To denote such a correspondence between $\mathcal{Y}$ and $\Psi(Z)$, one can choose an injection $i : \mathcal{Y} \rightarrow \Psi(Z)$ appropriately. In other words, one can find an injection $i$ so that $P_{X|Y = y}(\varphi_{i(y)}(x_1)) \leq P_{X|Y = y}(\varphi_{i(y)}(x_2))$ for every $y \in \mathcal{Y}$
and \( x_1, x_2 \in \mathcal{Z} \) satisfying \( x_1 \leq x_2 \). We now construct an \( \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \)-valued r.v. \((U, V, W)\) as follows: The conditional distribution \( P_{U \mid V, W} \) is given by

\[
P_{U \mid V = v, W = w}(u) = \begin{cases} P_{X \mid Y = v}(\varphi_{t(v)} \circ \varphi_w(u)) & \text{if } u \in \mathcal{Z}, \\ P_X(u) & \text{if } u \in \mathcal{X} \setminus \mathcal{Z}, \end{cases}
\]

\( (109) \)

where \( \varphi_1 \circ \varphi_2 \) stands for the composition of two bijections \( \varphi_1 \) and \( \varphi_2 \); the induced probability distribution \( P_V \) of \( V \) is given by \( P_V = P_Y \); and the independence \( V \perp W \) holds. Since \( V \perp W \) implies \( P_{U, V, W} = P_{U \mid V, W} P_V P_W \), it remains to determine the induced probability distribution \( P_W \) of \( W \), and we defer to determine it until the last paragraph of this proof. A direct calculation shows

\[
P_{U \mid W = w}(u) = \sum_{v \in \mathcal{Y}} P_{V \mid W = w}(v) P_{U \mid V = v, W = w}(u)
\]

\( (110) \)

where (a) follows by \( V \perp W \) and \( P_V = P_Y \); and (b) follows by \( (109) \) and defining \( \omega(u, w) \) so that

\[
\omega(u, w) := \sum_{v \in \mathcal{Y}} P_Y(v) P_X(v \mid \varphi_{t(v)} \circ \varphi_w(x))
\]

\( (111) \)

for each \( x \in \mathcal{Z} \) and \( w \in \Psi(\mathcal{Z}) \). Now, we readily see from \( (110) \) that \( (107) \) holds for any induced probability distribution \( P_W \) of \( W \). Therefore, to complete the proof, it suffices to show that \((U, W)\) satisfies \( (105) \) and \( (106) \) with an arbitrary choice of \( P_W \), and \((U, W)\) satisfies \( (104) \) with an appropriate choice of \( P_W \).

Firstly, we shall prove \( (105) \). For each \( w \in \Psi(\mathcal{Z}) \), we denote by \( \mathcal{D}(w) \in {\mathcal{Z}} \) the set satisfying \( \varphi_w(k) < \varphi_w(x) \) for every \( k \in \mathcal{D}(w) \) and \( x \in \mathcal{Z} \setminus \mathcal{D}(w) \), i.e., it stands for the set of first \( L \) elements in \( \mathcal{Z} \) under the permutation rule \( w \in \Psi(\mathcal{Z}) \). Then, we have

\[
P_e^{(L)}(U \mid W) \leq P_e^{(L)}(U \mid W \parallel \mathcal{Z})
\]

\( (a) \)

\[
= 1 - \sum_{w \in \Psi(\mathcal{Z})} P_W(w) \sum_{u \in \mathcal{D}(w)} P_{U \mid W = w}(u)
\]

\( (b) \)

\[
= 1 - \sum_{w \in \Psi(\mathcal{Z})} P_W(w) \sum_{u \in \mathcal{D}(w)} \omega(u, w)
\]

\( (c) \)

\[
= 1 - \sum_{w \in \Psi(\mathcal{Z})} P_W(w) \sum_{u=1}^{L} \sum_{v \in \mathcal{Y}} P_Y(v) P_{X \mid Y = v}(u)
\]

\( (d) \)

\[
= 1 - \sum_{y \in \mathcal{Y}} P_Y(y) \sum_{x=1}^{L} P_{X \mid Y = y}(x)
\]

\( (e) \)

\[
= P_e^{(L)}(X \mid Y),
\]

\( (112) \)

where (a) is an obvious inequality (see the definitions \( (5) \) and \( (101) \)); (b) follows from Proposition 6 and the definition of \( \mathcal{D}(w) \); (c) follows from \( (110) \) and the fact that \( \mathcal{D}(w) \subset \mathcal{Z} \) for each \( w \in \Psi(\mathcal{Z}) \); (d) follows from the constructions of the subset \( \mathcal{D}(w) \subset \mathcal{Z} \), the injection \( t : \mathcal{Y} \to \Omega(\mathcal{Z}) \), and the subset \( \mathcal{Z} \subset \mathcal{X} \); and (e) follows from Proposition 2. Hence, we obtain \( (105) \).

Secondly, we shall prove \( (106) \). We get

\[
b_{\phi}(X \mid Y) = \sum_{y \in \mathcal{Y}} P_Y(y) \phi(P_{X \mid Y = y})
\]

\[
= \sum_{w \in \Psi(\mathcal{Z})} P_W(w) \sum_{y \in \mathcal{Y}} P_Y(y) \phi(P_{X \mid Y = y})
\]

\( (a) \)

\[
= \sum_{w \in \Psi(\mathcal{Z})} P_W(w) \sum_{y \in \mathcal{Y}} P_Y(y) \phi(P_{U \mid V = y, W = w})
\]

\( (b) \)

\[
= \sum_{w \in \Psi(\mathcal{Z})} P_W(w) \sum_{v \in \mathcal{Y}} P_Y(v) \phi(P_{U \mid V = v, W = w})
\]

\( (c) \)
\[
_{(d)} \sum_{w \in \Psi(Z)} P_W(w) \phi(P_{U|W=w}) = b_{\phi}(U | W),
\]
where (a) follows by the symmetry of \( \phi \) and (109); (b) follows by \( P_V = P_Y \); (c) follows by Jensen’s inequality; and (d) follows by \( U \perp W \). Hence, we obtain (106).

Finally, we shall prove that there exists an induced probability distribution \( P_W \) satisfying (104). If we denote by \( I \in \Psi(Z) \) the identity matrix, then it follows from (111) that
\[
P_{U|W=I}(u) = P_{U|W=w}(\varphi^{-1}_w(u))
\]
for every \((u, w) \in Z \times \Psi(Z)\). It follows from (110) that
\[
\sum_{x \in Z} P_X(x) = \sum_{u \in Z} P_{U|W=I}(u).
\]
Now, denote by \( \beta_1 : \{1, 2, \ldots, LN\} \to Z \) and \( \beta_2 : \{1, 2, \ldots, LN\} \to Z \) two bijections satisfying \( P_X(\beta_1(i)) \geq P_X(\beta_1(j)) \) and \( \beta_2(i) < \beta_2(j) \), respectively, provided that \( i < j \). That is, the bijection \( \beta_1 \) and \( \beta_2 \) play roles of decreasing rearrangements of \( P_X \) and \( P_{U|W=I} \), respectively, on \( Z \). Using those bijections, one can rewrite (115) as
\[
\sum_{i=1}^{LN} P_X(\beta_1(i)) = \sum_{i=1}^{LN} P_{U|W=I}(\beta_2(i)).
\]
In the same way as (73), it can be verified from (111) by induction that
\[
\sum_{i=1}^{k} P_X(\beta_1(i)) \leq \sum_{i=1}^{k} P_{U|W=I}(\beta_2(i))
\]
for each \( k = 1, 2, \ldots, LN \). Equations (116) and (117) are indeed a majorization relation between two finite-dimensional real vectors, because \( \beta_1 \) plays a role of a decreasing rearrangement of \( P_X \) on \( Z \). Combining (114) and this majorization relation, it follows from the Hardy–Littlewood–Pólya theorem [25, Theorem 8] and finite-dimensional Birkhoff’s theorem [5] (see also [28, Theorem 2.1.2]) that there exists an induced probability distribution \( P_W \) satisfying \( P_U = P_X \), i.e., Equation (104) holds, as in (92)–(97). This completes the proof of Lemma 6.

It is worth mentioning that Lemma 6 is a useful reduction from infinite to finite-dimensional settings in the sense of (107). In other words, if \( Y \) is finite, it suffices to vary at most \( \#(Z) = \#(Y)L \) many probability masses \( P_X|_{Y=g}(x) \), \( x \in Z \), for every \( y \in Y \); and otherwise \( P_X|_{Y=g}(x) = P_X(x) \), \( x \in X \setminus Z \), for every \( y \in Y \). Fortunately, Lemma 6 is useful not only to prove Theorem 2 but also to prove Proposition 8 (see Appendix D).

As with (78), for a subset \( Z \subseteq X \), we define
\[
R(Q, L, \varepsilon, Y, Z) := \{(X, Y) : (X, Y) \text{ is an } X \times Y \text{-valued r.v.,}
\]
\[
P^{(L)}_e(X | Y \parallel Z) \leq \varepsilon,
\]
\[
P_{X|Y=g}(x) = Q(x) \quad \forall (x, y) \in (X \setminus Z) \times Y
\]
provided that \( Y \) is finite. It is clear that (118) coincides with (78) if \( Z = X \). It follows from Lemma 6 that for each system \((Q, L, \varepsilon, Y) \) satisfying (12), there exists a subset \( Z \subseteq X \) such that \( \#(Z) = \#(Y)L \) and \( R(Q, L, \varepsilon, Y, Z) \) is nonempty, provided that \( Y \) is finite.

Another important idea of proving Theorem 2 is to apply Lemma 3 for this collection of r.v.’s. The correction \( R(Q, L, \varepsilon, Y, Z) \) does not, however, have balanced regular conditional distributions of (75) in general, as with (78). Fortunately, similar to Lemma 4, the following lemma can avoid this issue by blowing-up the collection \( R(Q, L, \varepsilon, Y, Z) \) via finite-dimensional Birkhoff’s theorem [5].

**Lemma 7.** Suppose that \( Z \subseteq X \) is finite and \( R(Q, L, \varepsilon, Y, Z) \) is nonempty. If \( \#(Z)! \leq \#(Y) < \infty \), the collection \( R(Q, L, \varepsilon, Y, Z) \) has balanced regular conditional distributions.

**Proof of Lemma 7:** Lemma 7 can be proven in a similar fashion to the proof of Lemma 4. Since this proof is slightly long as with Lemma 4, we only give a sketch of the proof as follows.

Since \( \#(\Psi(Z)) = \#(Z)! \), we may assume without loss of generality that \( Y = \Psi(Z) \). For short, we write \( \tilde{R} = R(Q, L, \varepsilon, Y, Z) \) in this proof. Here, note the \( \sigma \)-algebra on \( Y \) is discrete. For an \( X \times Y \)-valued r.v. \( (X, Y) \in \tilde{R} \), we construct another \( X \times Y \)-valued r.v. \( (U, V) \), as in (84), so that \( P_{U|V=g}(x) = Q(x) \) for every \( (x, y) \in (X \setminus Z) \times Y \). By such a construction (84), the condition (75) is obviously satisfied. In the same way as (86), we can verify that \( P^{(L)}_e(U | V \parallel Z) = P^{(L)}_e(X | Y \parallel Z) \). Moreover, employing finite-dimensional Birkhoff’s theorem [5] instead of infinite-dimensional Birkhoff’s theorem [43, Theorem 2], we can also
verify the existence of induced probability distributions $P_V$ satisfying $P_V = Q$ in the same way as (87). Therefore, for any $(X, Y) \in \mathcal{R}$, one can find $(U, V)$ so that (75) holds and $(U, V) \in \mathcal{R}$ as well. This completes the proof of Lemma 7.

Let $\mathcal{Z} \subset \mathcal{X}$ be a subset. Similar to Proposition 7 and the proof of Lemma 6, consider a bijection $\beta : \{1, 2, \ldots, \#(\mathcal{Z})\} \to \mathcal{Z}$ satisfying $Q(\beta(i)) \geq Q(\beta(j))$ whenever $i < j$, i.e., it plays a role of a decreasing rearrangement of $Q$ on $\mathcal{Z}$. Suppose that $(Q, L, \varepsilon, Y, \mathcal{Z})$ fulfills

$$1 - \sum_{x \in \mathcal{Z}} Q(x) \leq \varepsilon \leq 1 - \sum_{x = 1}^{L} Q(\beta(x)).$$

(119)

Then, define Fano-distribution of type-3 by the following $\mathcal{X}$-marginal:

$$p_{Fano-type3}(\alpha, \gamma, \varepsilon, \mathcal{Z}) := \begin{cases} 
  \mathcal{V}_3(J_3) & \text{if } x \in \mathcal{Z} \text{ and } J_3 \leq \beta_1^{-1}(x) \leq L, \\
  \mathcal{W}_3(K_3) & \text{if } x \in \mathcal{Z} \text{ and } L < \beta_1^{-1}(x) \leq K_3, \\
  Q(x) & \text{otherwise},
\end{cases}$$

(120)

where the weight $\mathcal{V}_3(j)$ is defined by

$$\mathcal{V}_3(j) := \frac{(1 - \varepsilon - \sum_{x = 1}^{j - 1} Q(\beta_1(x))}{L - j + 1}$$

(121)

for each integer $1 \leq j \leq L$; the weight $\mathcal{W}_3(k)$ is defined by

$$\mathcal{W}_3(k) := \begin{cases} 
  -1 & \text{if } k = L, \\
  \frac{\sum_{x = 1}^{k} Q(\beta_1(x)) - (1 - \varepsilon)}{k - L} & \text{if } k > L
\end{cases}$$

(122)

for each integer $L \leq k \leq \#(\mathcal{Y})L$; the integer $J_3$ is chosen so that

$$J_3 := \min\{1 \leq j \leq L \mid Q(\beta_1(j)) \leq \mathcal{V}_3(j)\};$$

(123)

and the integer $K_3$ is chosen so that

$$K_3 := \max\{L \leq k \leq \#(\mathcal{Y})L \mid \mathcal{W}_3(k) \leq P_X(\beta_1(k))\}.$$

(124)

It is worth pointing out that Fano-distribution of type-3 can be reduced to the Fano-distribution of type-2 defined in (21) and Ho–Verdú’s truncated distribution [28, Equation (17)] by setting $\mathcal{Z} = \mathcal{X}$ and $L = 1$, respectively. In fact, the following lemma shows a relation between the type-2 and type-3.

**Lemma 8.** Suppose that $\#(\mathcal{Z}) = \#(\mathcal{Y})L$ and $(Q, L, \varepsilon, Y, \mathcal{Z})$ fulfills (119). Then, $p_{Fano-type3}(\alpha, \gamma, \varepsilon, \mathcal{Z})$ majorizes $p_{Fano-type2}(\alpha, \gamma, \varepsilon, \mathcal{Z})$.

**Proof of Lemma 8:** For simplicity, we write $P_2 = p_{Fano-type2}(\alpha, \gamma, \varepsilon, \mathcal{Z})$ and $P_3 = p_{Fano-type3}(\alpha, \gamma, \varepsilon, \mathcal{Z})$ in this proof. We readily see that $P_2 = P_3$ if $\mathcal{Z} = \{1, 2, \ldots, \#(\mathcal{Y})L\}$ and $Q = Q_1$, because $\beta : \{1, 2, \ldots, \#(\mathcal{Z})\} \to \mathcal{Z}$ used in (120) is the identity mapping in this case. Actually, we may assume without loss of generality that $Q = Q_1$.

While $P_2 = P_3$ does not hold in general, we can see from the definition (21) that $P_2(x) = P_3(x)$ for each $x = 1, 2, \ldots, L$. Hence, since $P_2(x) = Q(x) \leq P_3(x)$ for each $x = 1, 2, \ldots, J - 1$, it follows that

$$\sum_{x = 1}^{k} P_2(x) \leq \sum_{x = 1}^{k} P_3(x)$$

(125)

for each $k = 1, 2, \ldots, J - 1$. By the definitions (14), (16), (121), and (123), it can be verified that $J \geq J_3$ and $\mathcal{V}(J) \leq \mathcal{V}_3(J_3)$. Thus, as $P_2(x) = \mathcal{V}(J)$ for each $x = J, J + 1, \ldots, L$, it follows that $P_2(x) \geq \mathcal{V}_3(J_3)$ for each $x = J, J + 1, \ldots, L$; which implies that (125) also holds for each $k = J, J + 1, \ldots, L$. Therefore, we observe that $P_3$ majorizes $P_2$ over the subset $\{1, 2, \ldots, L\} \subset \mathcal{X}$.

We prove the rest of the majorization relation by contradiction. Namely, assume that

$$\sum_{x = 1}^{l} P_2(x) > \sum_{x = 1}^{l} P_3(x)$$

(126)

for some integer $l \geq L + 1$. Recall that $J \geq J_3$ and $\mathcal{V}(J) \leq \mathcal{V}_3(J_3)$. Moreover, by the definitions (15), (22), (122), and (124), it can be verified that $K_3 \leq K_3$ and $\mathcal{W}(K_2) \geq \mathcal{W}_3(K_3)$. Thus, since (i) $P_3(x) = \mathcal{W}(K_2) \leq Q(x)$ for each $x = L + 1, L + 2, \ldots, K_2$ and (ii) $P_3(x) = \mathcal{W}_3(K_3) \leq Q(x)$ for each $x = \beta_1(L + 1), \beta_1(L + 2), \ldots, \beta_1(K_3)$, it follows that $P_2(x) \geq P_3(x)$ for every $x = l, l + 1, \ldots$, which implies together with the hypothesis (126) that

$$\sum_{x = l}^{\infty} P_2(x) > \sum_{x = l}^{\infty} P_3(x).$$

(127)
This, however, contradicts to the definition of probability distributions. This completes the proof of Lemma 8. ■

Similar to (101), we now define

$$P_e^{(L)}(X \parallel Z) := \min_{D \in Z^{(L)}} P\{X \in D\}. \quad (128)$$

As with Proposition 7, we can verify that

$$P_e^{(L)}(X \parallel Z) = 1 - \min_{D \in Z^{(L)}} \sum_{x \in D} P_X(x) = 1 - \sum_{x=1}^L P_X(\beta(x)). \quad (129)$$

Hence, the restriction (119) comes from the same observation as (12) (see Propositions 3 and 7). In view of (129), we write $P_e^{(L)}(Q \parallel Z) = P_e^{(L)}(X \parallel Z)$ if $P_X = Q$. As in Lemma 5, the following lemma holds.

**Lemma 9.** Suppose that the system $(Q, L, \varepsilon, Y, Z)$ fulfills (119). In addition, suppose that an $X$-marginal $R$ satisfies the following: (i) $R$ majorizes $Q$; (ii) $P_e^{(L)}(R \parallel Z) \leq \varepsilon$; and (iii) $R(k) = Q(k)$ for each $k \in X \setminus Z$. Then, it holds that $R$ majorizes $P^{(Q, L, \varepsilon, Y, Z)}_{\text{Fano-type3}}$ as well, where $P^{(Q, L, \varepsilon, Y, Z)}_{\text{Fano-type3}}$ is defined in (120).

**Proof of Lemma 9:** For simplicity, we write $P_3 = P^{(Q, L, \varepsilon, Y, Z)}_{\text{Fano-type3}}$ in this proof. Since $R(x) = P_3(x) = Q(x)$ for every $x \in X \setminus Z$, it suffices to verify the majorization relation over $Z$. Denote by $\beta_1 : \{1, 2, \ldots, \#(Y)L\} \to Z$ and $\beta_2 : \{1, 2, \ldots, \#(Y)L\} \to Z$ two bijections satisfying $R(\beta_1(i)) \geq R(\beta_2(i))$ and $R(\beta_2(i)) \leq R(\beta_1(j))$, respectively, whenever $i < j$. In other words, two bijections $\beta_1$ and $\beta_2$ play roles of decreasing rearrangements of $R$ and $P_3$, respectively, on $Z$. That is, we shall prove that

$$\sum_{x=1}^k P_3(\beta_2(x)) \leq \sum_{x=1}^k R(\beta_1(x)) \quad (130)$$

for every $k = 1, 2, \ldots, \#(Z)$.

As $R$ majorizes $Q$, it follows from (120) that (130) holds for each $k = 1, 2, \ldots, J_3 - 1$. Moreover, we readily see from (120) that

$$\sum_{x=1}^L P_3(\beta_2(x)) = 1 - \varepsilon; \quad (131)$$

hence, it follows from Lemma 1 and the hypothesis $P_e^{(L)}(R \parallel Z) \leq \varepsilon$ that (130) holds for each $k = J_3, J_3 + 1, \ldots, L$. Similarly, since (130) holds with equality if $k = \#(Z)$, it also follows from Lemma 1 that (130) holds for each $k = L + 1, L + 2, \ldots, \#(Z)$. Therefore, we observe that $R$ majorizes $P_3$. This completes the proof of Lemma 9. ■

Finally, we can prove Theorem 2 by using the above lemmas.

**Proof of Theorem 2:** For short, we write $P_2 = P^{(Q, L, \varepsilon)}_{\text{Fano-type3}}$ in the proof. Moreover, we define

$$\mathcal{R}_1 := \mathcal{R}(Q, L, \varepsilon, Y), \quad (132)$$

$$\mathcal{R}_2 := \bigcup_{Z \subset X \setminus \#(Z) = \#(Y)L} \mathcal{R}(Q, L, \varepsilon, Y, Z), \quad (133)$$

$$\mathcal{R}_3 := \bigcup_{Z \subset X \setminus \#(Z) = \#(Y)L} \mathcal{R}(Q, L, \varepsilon, Y \cup \Psi(Z), Z), \quad (134)$$

$$\mathcal{P}_3 := \left\{ R \in \mathcal{P}(X) \mid \exists Z \subset X \text{ s.t. } \#(Z) = \#(Y)L, \quad \begin{array}{c} P_e^{(L)}(R \parallel Z) \leq \varepsilon, \quad P_e^{(L)}(R \parallel Z) \leq \varepsilon, \\ R(x) = Q(x) \text{ for } x \in X \setminus Z \end{array} \right\}. \quad (135)$$

Then, we have

$$\sup_{(X, Y) : P_e^{(L)}(X \parallel Y) \leq \varepsilon, P_X = Q} b_\phi(X \mid Y) \overset{(a)}{=} \sup_{(X, Y) \in \mathcal{R}_1} b_\phi(X \mid Y) \overset{(b)}{=} \sup_{(X, Y) \in \mathcal{R}_2} b_\phi(X \mid Y) \overset{(c)}{=} \sup_{(X, Y) \in \mathcal{R}_3} b_\phi(X \mid Y) \overset{(d)}{=} \sup_{(X, Y) : P_e^{(L)}(X \parallel Y) \leq \varepsilon, P_X = Q} b_\phi(X \mid Y) \overset{(e)}{=} \sup_{R \in \mathcal{P}_3} \phi(R) \text{ is almost surely constant}$$

where $b_\phi(X \mid Y)$ is the Bhattacharyya function.
\[ \begin{align*}
(\text{f}) & \leq \sup_{\mathcal{Z} \in \mathcal{X} : \mathbb{P}(\mathcal{Z}) \neq \emptyset, \mathbb{P}(\mathcal{Y} \cup \mathcal{Z})} \phi \left( P^{(Q, L, e, \mathcal{Y}, \mathcal{Z})}_{\text{Fano-type3}} \right) \\
(\text{g}) & \leq \phi(P_2),
\end{align*} \]

where (a) follows from the definition (78); (b) follows from Lemma 6 and the definition (118); (c) follows from the inclusions \( \mathcal{R}(Q, L, e, \mathcal{Y}, \mathcal{Z}) \subset \mathcal{R}(Q, L, e, \mathcal{Y} \cup \Psi(\mathcal{Z}), \mathcal{Z}) \); (d) follows from Lemmas 3 and 7; (e) follows from Lemma 2; (f) follows from Lemma 9; and (g) follows from Proposition 1 and Lemma 8. Inequalities (136) is indeed the Fano-type inequality (23) of Theorem 2. The sharpness conditions 1) and 2) of Theorem 2 can be proved in the same way as the proof of Theorem 1. Analogously, the equality condition (25) can be verified in the same way as the proof of Theorem 1. This completes the proof of Theorem 2.

It is worth mentioning that Step (b) in (136) is indeed the reduction step from infinite to finite-dimensional settings via Lemma 6 (see also the paragraph below the proof of Lemma 6), i.e., it is a key of our analysis in this subsection. Moreover, it is worth pointing out that the proof of Theorem 2 does not work if \( \mathcal{Y} \) is infinite, while the proof of Theorem 1 works well for any nonempty alphabet \( \mathcal{Y} \).

**APPENDIX C**

**PROOFS OF ASYMPTOMATIC ESTIMATES OF EQUIVOCATIONS**

A. **Proof of Theorem 3**

Defining the *variational distance* between two \( X \)-marginals \( P \) and \( Q \) by

\[
d(P, Q) := \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|,
\]

we now introduce the following lemma, which is useful to prove Theorem 3.

**Lemma 10** ([32, Theorem 3]). *Let \( Q \) be an \( X \)-marginal, and \( 0 \leq \delta \leq 1 - Q(1) \) a real number. Then, it holds that*

\[
\min_{R \in \mathcal{P}(\mathcal{X}) : d(Q, R) \leq \delta} H(R) = H(S(Q, \delta)),
\]

*where the \( X \)-marginal \( S(Q, \delta) \) is defined by*

\[
S(Q, \delta)(x) := \begin{cases} 
Q(x) + \delta & \text{if } x = 1, \\
Q(x) & \text{if } 1 < x < B, \\
\sum_{k=0}^{\infty} Q(k) - \delta & \text{if } x = B, \\
0 & \text{if } x > B,
\end{cases}
\]

*and the integer \( B \) is chosen so that*

\[
B := \sup \left\{ b \geq 1 \mid \sum_{k=b}^{\infty} Q(k) \geq \delta \right\}.
\]

*Proof of Theorem 3:* For short, in this proof, we write \( \varepsilon_n := P_{e_n}^{(L_n)}(X_n \mid Y_n) \), \( P_n = P_{X_n}^{(e_n)} \), and \( P_{1,n} := P_{\text{Fano-type1}}^{(P_n, L_n, \varepsilon_n)} \) for each \( n \geq 1 \). Suppose that \( \varepsilon_n = o(1) \). By Corollary 1, instead of (44), it suffices to consider the following:

\[
|H(P_{1,n}) - \log L_n| = o(H(X_n)).
\]

Since \( \text{supp}(P_{1,n}) = \{1, \ldots, L_n\} \) if \( \varepsilon_n = 0 \), we may assume without loss of generality that \( 0 < \varepsilon_n < 1 \).

Define two \( X \)-marginals \( Q_{n}^{(1)} \) and \( Q_{n}^{(2)} \) by

\[
\begin{align*}
Q_{n}^{(1)}(x) & = \begin{cases} 
P_{1,n}(x) & \text{if } 1 \leq x \leq L_n, \\
1 - \varepsilon_n & \text{if } x = L_n + 1, \\
0 & \text{if } x \geq L_n + 2,
\end{cases} \\
Q_{n}^{(2)}(x) & = \begin{cases} 
P_{1,n}(x) & \text{if } 1 \leq x \leq L_n, \\
0 & \text{if } x = L_n + 1, \\
\varepsilon_n & \text{if } x \geq L_n + 2
\end{cases}
\end{align*}
\]
for each \( n \geq 1 \). Since \( Q_n^{(1)} \) majorizes the uniform distribution on \( \{1, 2, \ldots, L_n\} \), it is clear from the Schur-concavity of the Shannon entropy that \( H(Q_n^{(1)}) \leq \log L_n \). Thus, since \( P_{1,n} = (1 - \varepsilon_n)Q_n^{(1)} + \varepsilon_n Q_n^{(2)} \), it follows by the strong additivity of the Shannon entropy that

\[
H(P_{1,n}) = h_2(\varepsilon_n) + (1 - \varepsilon_n) H(Q_n^{(1)}) + \varepsilon_n H(Q_n^{(2)}) \\
\leq h_2(\varepsilon_n) + (1 - \varepsilon_n) \log L_n + \varepsilon_n H(Q_n^{(2)}). \tag{144}
\]

Thus, since \( h_2(\varepsilon_n) = o(1) \), it suffices to verify the asymptotic estimate of the third term in the right-hand side of (144), i.e., whether

\[
\varepsilon_n H(Q_n^{(2)}) = o(H(X_n)) \tag{145}
\]

holds or not.

Consider the \( X \)-marginal \( Q_n^{(3)} \) given by

\[
Q_n^{(3)}(x) = \frac{P_n(x) - \varepsilon_n Q_n^{(2)}(x)}{1 - \varepsilon_n} \tag{146}
\]

for each \( n \geq 1 \). Since \( P_n = \varepsilon_n Q_n^{(2)} + (1 - \varepsilon_n) Q_n^{(3)} \), it follows by the concavity of the Shannon entropy that

\[
H(X_n) \geq \varepsilon_n H(Q_n^{(2)}) + (1 - \varepsilon_n) H(Q_n^{(3)}) \tag{147}
\]

for each \( n \geq 1 \). A direct calculation shows

\[
d(P_n, Q_n^{(3)}) = \frac{1}{2} \sum_{x=1}^{\infty} \left| P_n(x) - Q_n^{(3)}(x) \right| \\
= \frac{1}{2} \sum_{x=1}^{\infty} \left| P_n(x) - \frac{P_n(x) - \varepsilon_n Q_n^{(2)}(x)}{1 - \varepsilon_n} \right| \\
= \frac{\varepsilon_n}{1 - \varepsilon_n} \sum_{x=1}^{\infty} \left| P_n(x) - Q_n^{(2)}(x) \right| \\
\leq \frac{\varepsilon_n}{1 - \varepsilon_n} d(P_n, Q_n^{(2)}) \\
= \delta_n \tag{148}
\]

for each \( n \geq 1 \), where note that \( \varepsilon_n = o(1) \) implies \( \delta_n = o(1) \) as well. Thus, it follows from Lemma 10 that

\[
H(Q_n^{(3)}) \geq H(S(P_n, \delta_n)) \\
= (a) \eta(P_n(1) + \delta_n) + \sum_{x=2}^{B_n - 1} \eta(P_n(x)) + \eta \left( \sum_{k=1}^{\infty} P_n(k) - \delta_n \right) \\
\geq \sum_{x=1}^{B_n} \eta(P_n(x)) - 2 \gamma_n \\
= \sum_{x=1}^{B_n} P_n(x) \log \frac{1}{P_n(x)} - 2 \gamma_n \tag{149}
\]

for every \( \varepsilon > 0 \) and each \( n \geq 1 \), where (a) follows by the definitions \( \eta(u) := -u \log u \) and

\[
B_n := \sup \left\{ b \geq 1 \left| \sum_{k=b}^{\infty} P_n(k) \geq \delta_n \right. \right\} \tag{150}
\]
for each $n \geq 1$; (b) follows by the continuity of $\eta$ and the fact that $\delta_n = o(1)$, i.e., there exists a sequence $\{\gamma_n\}_{n=1}^{\infty}$ of positive reals satisfying $\gamma_n = o(1)$ and

$$\left| \eta(P_n(1)) - \eta(P_n(1) + \delta_n) \right| \leq \gamma_n, \quad (151)$$

$$\left| \eta(P_n(B_n)) - \eta\left( \sum_{k=B_n}^{\infty} P_n(k) - \delta_n \right) \right| \leq \gamma_n \quad (152)$$

for each $n \geq 1$; (c) follows by constructing the subset $B^{(n)} \subset X$ so that

$$|B^{(n)}| = \min_{B \supseteq X, \|B\| \geq 1 - \delta_n} |B| \quad (153)$$

for each $n \geq 1$; (d) follows by defining the typical set $\mathcal{A}_\epsilon^{(n)} \subset X$ so that

$$\mathcal{A}_\epsilon^{(n)} := \{ x \in X \mid \log \frac{1}{P_{X_n}(x)} \leq (1 - \epsilon) H(X_n) \} \quad (154)$$

with some $\epsilon > 0$ for each $n \geq 1$; and (e) follows by the definition of $\mathcal{A}_\epsilon^{(n)}$. Since $\{X_n\}_{n=1}^{\infty}$ satisfies the AEP of Definition 7, since $\mathbb{P}\{X_n \in B^{(n)}\} \geq 1 - \delta_n$, and since $\delta_n = o(1)$, it is clear that $\mathbb{P}\{X_n \notin \mathcal{A}^{(n)} \cap B^{(n)}\} = o(1)$ (see, e.g., [9, Problem 3.11]).

Thus, since $\epsilon > 0$ can be arbitrarily small and $\epsilon_n = o(1)$, it follows from (149) that there exists a sequence $\{\lambda_n\}_{n=1}^{\infty}$ of positive real numbers satisfying $\lambda_n = o(1)$ and

$$(1 - \epsilon_n)H(Q_n^{(3)}) \geq (1 - \lambda_n)H(X_n) - \frac{2 \gamma_n}{1 - \epsilon_n} \quad (155)$$

for each $n \geq 1$. Combining (147) and (155), we observe that

$$\lambda_n H(X_n) + \frac{2 \gamma_n}{1 - \epsilon_n} \geq \epsilon_n H(Q_n^{(3)}) \quad (156)$$

for each $n \geq 1$. Therefore, Equation (145) is indeed valid, which proves (141) together with (144). This completes the proof of Theorem 3.

B. Proof of Theorem 4

Proof of Theorem 4: The condition (b) is a direct consequence of Theorem 3; and we shall verify the conditions (a), (c), and (d) in the proof. For short, in the proof, we write $\epsilon_n := P_{X_n}(Y_n \mid Y_n)$, $P_n := P_{X_n}$, $P = P_{X}$, and $P_{1,n} := P_{f_{\alpha}(P_{X_n},L_n,\epsilon_n)}$ for each $n \geq 1$. By Corollary 1, instead on (52), it suffices to consider the following:

$$\lim_{n \to \infty} \left| H_{\alpha}(P_{1,n}) - \log L_n \right|^+ = 0 \quad (157)$$

under any one of the conditions (a)–(c). Similar to the proof of Theorem 3, we may assume without loss of generality that $0 < \epsilon_n < 1$.

Firstly, we shall verify the condition (a). Let $Q_n$ be an $X$-marginal given by

$$Q_n(x) = \begin{cases} \frac{1 - \epsilon_n}{L_n} & \text{if } 1 \leq x \leq L_n, \\ P_{\text{type}X_n}(x) & \text{if } x \geq L_n + 1 \end{cases} \quad (158)$$

for each $n \geq 1$. Since $P_{1,n}$ majorizes $Q_n$, it follows by the Schur-concavity of the Rényi entropy that

$$H_{\alpha}(P_{1,n}) \leq H_{\alpha}(Q_n)$$

$$= \frac{1}{1 - \alpha} \log \left( (1 - \epsilon_n)^{\alpha} L_n^{1 - \alpha} + \sum_{x = L_n}^{\infty} P_{1,n}(x)^{\alpha} \right)$$

$$\leq \frac{1}{1 - \alpha} \log \left( (1 - \epsilon_n)^{\alpha} L_n^{1 - \alpha} \right)$$

$$= \log L_n + \frac{\alpha}{1 - \alpha} \log(1 - \epsilon_n), \quad (159)$$

where the second inequality follows by the hypothesis $\alpha > 1$, i.e., the condition (a). These inequalities immediately ensure (157) under the condition (a).
Secondly, we shall verify the condition (d) of Theorem 4. Since \(X\) and \(\{X_n\}_n\) are discrete r.v.'s, note that the convergence in distribution \(X_n \xrightarrow{d} X\) is equivalent to \(P_n(x) \rightarrow P(x)\) as \(n \rightarrow \infty\) for each \(x \in X\), i.e., the pointwise convergence \(P_n \rightarrow P\). It is well-known that the Rényi entropy \(H_\alpha(P)\) is nonincreasing for \(\alpha \geq 0\); hence, it suffices to verify (157) with \(\alpha = 1\), i.e.,
\[
\lim_{n \rightarrow \infty} \left| H(P_{1,n}) - \log L_n \right| = 0.
\] (160)

We now define two \(X\)-marginals \(Q_n^{(1)}\) and \(Q_n^{(2)}\) in the same ways as (142) and (143), respectively, for each \(n \geq 1\). By (144), it suffices to verify whether the third term in the right-hand side of (144) approaches to zero, i.e.,
\[
\lim_{n \rightarrow \infty} \epsilon_n H(Q_n^{(2)}) = 0.
\] (161)

This can be verified in a similar fashion to the proof of [28, Lemma 3] as follows: Consider the \(X\)-marginal \(Q_n^{(3)}\) defined in (146) for each \(n \geq 1\). Since \(Q_n^{(2)}(1) = 0\) and \(\epsilon_n Q_n^{(2)}(x) \leq \epsilon_n\) for each \(x \geq 2\), we observe that \(\epsilon_n Q_n^{(2)}(x) = o(1)\) for every \(x \geq 1\); hence,
\[
\lim_{n \rightarrow \infty} Q_n^{(3)}(x) = \lim_{n \rightarrow \infty} P_\perp X_n(x)
\] (162)
for every \(x \geq 1\). Therefore, since \(P_n\) converges pointwise to \(P\) as \(n \rightarrow \infty\), we see that \(Q_n^{(3)}\) also converges pointwise to \(P_\perp X\) as \(\epsilon_n\) vanishes. Hence, by the lower semicontinuity of the Shannon entropy, we observe that
\[
\liminf_{n \rightarrow \infty} H(Q_n^{(3)}) \geq H(X),
\] (163)
and we then have
\[
H(X) = \lim_{n \rightarrow \infty} H(X_n) \geq \limsup_{n \rightarrow \infty} \left( \epsilon_n H(Q_n^{(2)}) + (1 - \epsilon_n) H(Q_n^{(3)}) \right)
\geq \limsup_{n \rightarrow \infty} \left( \epsilon_n H(Q_n^{(2)}) + \liminf_{n \rightarrow \infty} \left( (1 - \epsilon_n) H(Q_n^{(3)}) \right) \right)
= \limsup_{n \rightarrow \infty} \left( \epsilon_n H(Q_n^{(2)}) + \liminf_{n \rightarrow \infty} H(Q_n^{(3)}) \right)
\geq \limsup_{n \rightarrow \infty} \left( \epsilon_n H(Q_n^{(2)}) + H(X) \right).
\] (164)

Thus, it follows from (164), the hypothesis \(H(X) < \infty\), and the nonnegativity of the Shannon entropy that (161) is valid, which proves (160) together with (144).

Finally, we shall verify the condition (c). Define the \(X\)-marginal \(\bar{Q}_n^{(2)}\) by
\[
\bar{Q}_n^{(2)}(x) = \begin{cases} 0 & \text{if } 1 \leq x \leq L_n, \\ \frac{P_{1,n}(x)}{\epsilon_n} & \text{if } x \geq L_n + 1, \end{cases}
\] (165)
for each \(n \geq 1\), where \(P_{1,n} = P_{\text{Fano-type}}(P, L_n, \epsilon_n)\). Note that the difference between \(Q_n^{(2)}\) and \(\bar{Q}_n^{(2)}\) is the difference between \(P_n\) and \(P\). It can be verified by the same way as (164) that
\[
\lim_{n \rightarrow \infty} \left( \epsilon_n H(\bar{Q}_n^{(2)}) \right) = 0.
\] (166)

It follows by the same manner as [28, Lemma 1] that if \(P_n\) majorizes \(P\), then \(Q_n^{(2)}\) majorizes \(\bar{Q}_n^{(2)}\) as well. Therefore, it follows from the Schur-concavity of the Shannon entropy that if \(P_n\) majorizes \(P\) for sufficiently large \(n\), then
\[
H(Q_n^{(2)}) \leq H(\bar{Q}_n^{(2)})
\] (167)
for sufficiently large \(n\). Combining (166) and (167), Equation (161) also holds under the condition (c). This completes the proof of Theorem 4.

\[\blacksquare\]
C. Proof of Theorem 5

To prove Theorem 5, we now give the following lemma.

Lemma 11. If \( H(Q) < \infty \), then the mapping \( \varepsilon \mapsto H(P^{Q,L,\varepsilon}_{\text{Fano-type1}}) \) is concave in the interval (12) with \( \#(Y) = \infty \).

Proof of Lemma 11: It is well-known that for a fixed \( P_X \), the conditional Shannon entropy \( H(X \mid Y) \) is concave in \( P_{Y \mid X} \) (cf. [9, Theorem 2.7.4]). Defining the distortion measure \( d : X \times X^{(L)} \rightarrow (0,1) \) by

\[
d(x, \hat{x}) = \begin{cases} 1 & \text{if } x \not\in \hat{x}, \\ 0 & \text{if } x \in \hat{x}, \end{cases}
\]

the average probability of list decoding error is equal to the average distortion, i.e.,

\[
P\{X \not\in f(Y)\} = \mathbb{E}[d(X, f(Y))]
\]

for any list decoder \( f : Y \rightarrow X^{(L)} \). Therefore, by following Theorem 1, the concavity of Lemma 11 can be proved by the same argument as the proof of the convexity of the rate-distortion function (cf. [9, Lemma 10.4.1]).

Proof of Theorem 5: For short, we write \( P = P_X, P_n = P_{X^n}, \varepsilon_n = P^{(L_n,\varepsilon_n)}_e(X_n \mid Y_n), P_{1,n} = P^{(L_n,\varepsilon_n)}_e, \) and \( \hat{P}_{1,n} = P^{(L_n,\varepsilon_n)}_{\text{Fano-type1}} \) in this proof. Define \( \hat{L} := \limsup_{n \to \infty} L_n \). If \( \hat{L} = \infty \), then (54) is a trivial inequality. Hence, it suffices to consider the case where \( \hat{L} < \infty \).

It is clear that there exists an integer \( n_0 \geq 1 \) such that \( L_n \leq \hat{L} \) for every \( n \geq n_0 \). Then, we can verify that \( P_{1,n} \) majorizes \( \hat{P}_{1,n} \) for every \( n \geq n_0 \) as follows: Let \( J_n \) and \( J_3 \) be given by (16) with \( (Q, L, \varepsilon) = (P_n, L_n, \varepsilon_n) \) and \( (Q, L, \varepsilon) = (P_n, L, \varepsilon_n) \), respectively. Similarly, let \( K_n \) and \( K_3 \) be given by (17) with \( (Q, L, \varepsilon) = (P_n, L_n, \varepsilon_n) \) and \( (Q, L, \varepsilon) = (P_n, L, \varepsilon_n) \), respectively. Since \( L_n \leq \hat{L} \) implies \( J_n \leq J_3 \) and \( K_n \leq K_3 \), it can be seen from (13) that

\[
P_{1,n}(x) = \hat{P}_{1,n}(x) \quad \text{for } 1 \leq x < J_n \text{ or } x \geq K_3, \\
P_{1,n}(x) \geq \hat{P}_{1,n}(x) \quad \text{for } J_n \leq x \leq L_n \text{ or } \hat{L} < x \leq K_3, \\
P_{1,n}(x) \leq \hat{P}_{1,n}(x) \quad \text{for } L_n < x \leq \hat{L}.
\]

Therefore, noting that

\[
\sum_{x=1}^{L_n} P_{1,n}(x) = \sum_{x=1}^{\hat{L}} \hat{P}_{1,n}(x) = 1 - \varepsilon_n,
\]

we obtain the majorization relation \( P_{1,n} \succeq \hat{P}_{1,n} \) for every \( n \geq n_0 \).

By hypothesis, there exists an integer \( n_1 \geq 1 \) such that \( P_n \) majorizes \( P \) for every \( n \geq n_1 \). Letting \( n_2 = \max\{n_0, n_1\} \), we observe that

\[
\frac{1}{n} H(X^n \mid Y^n) \leq \frac{1}{n} \sum_{i=1}^{n} H(X_i \mid Y_i)
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n-1} H(X_i \mid Y_i) + \frac{1}{n} \sum_{j=n_2}^{n} H(X_i \mid Y_i)
\]

\[
\leq \frac{n_2 - 1}{n} \left( \max_{1 \leq i < n_2} H(X_i) \right) + \frac{1}{n} \sum_{j=n_2}^{n} H(\hat{P}_{1,j}) \tag{a}
\]

\[
\leq \frac{n_2 - 1}{n} \left( \max_{1 \leq i < n_2} H(X_i) \right) + \frac{1}{n} \sum_{j=n_2}^{n} H\left( P^{(L_j,\varepsilon_j)}_{\text{Fano-type1}} \right) \tag{b}
\]

\[
\leq \frac{n_2 - 1}{n} \left( \max_{1 \leq i < n_2} H(X_i) \right) + \frac{n - n_2 + 1}{n} H\left( P^{(L,\varepsilon)}_{\text{Fano-type1}} \right) \tag{c}
\]

for every \( n \geq n_2 \), where (a) follows by Corollary 1 and \( P_{1,n} \succeq \hat{P}_{1,n} \); (b) follows by the condition (b) of Theorem 5 and the same manner as [28, Lemma 1]; (c) follows by Lemma 11 together with the definition:

\[
\tilde{\varepsilon}_n := \frac{1}{n - n_2 + 1} \sum_{j=n_2}^{n} \varepsilon_j - \frac{1}{n - n_2 + 1} \sum_{j=n_2}^{n} P^{(L_j)}_e(X_j \mid Y_j).
\]

Note that the Schur-convexity of the Shannon entropy is used in both (b) and (c) of (174). Since \( P^{(L,\varepsilon)}_{\text{Fano-type1}}(X^n \mid Y^n) = o(1) \) is equivalent to \( \tilde{\varepsilon}_n = o(1) \), it follows from (157) that there exists an integer \( n_3 \geq 1 \) such that

\[
H\left( P^{(L,\varepsilon)}_{\text{Fano-type1}} \right) \leq \log \hat{L}.
\]
for every \( n \geq n_3 \). Hence, it follows from (174) that

\[
\frac{1}{n} H(X^n \mid Y^n) \leq \frac{n_2 - 1}{n} \left( \max_{1 \leq i < n_2} H(X_i) \right) + \frac{n - n_2 + 1}{n} \log L
\]

(177)

for every \( n \geq \max(n_2, n_3) \). Therefore, letting \( n \to \infty \) in (177), we have (54). This completes the proof of Theorem 5. \( \blacksquare \)

APPENDIX D

AN IMPOSSIBILITY OF ESTABLISHING FANO-TYPE INEQUALITY

In Section II-C, we have considered Fano-type inequalities on \( h_\phi(X \mid Y) \) without any explicit form of \( \phi \) under the three postulates: \( \phi \) is symmetric, concave, and lower semicontinuous. If we impose another postulate on \( \phi \), then we can also avoid the (degenerate) case in which \( \phi(Q) = \infty \). The following proposition shows this fact.

Proposition 8. Let \( g_1 : [0, 1] \to [0, \infty) \) be a function satisfying \( g_1(0) = 0 \), and \( g_2 : [0, \infty] \to [0, \infty] \) a function satisfying \( g_2(u) = \infty \) only if \( u = \infty \). Suppose that \( \phi : \mathcal{P}(X) \to [0, \infty] \) is of the form

\[
\phi(Q) = g_2 \left( \sum_{x \in X} g_1(Q(x)) \right).
\]

(178)

the system \((Q, L, \varepsilon, Y)\) satisfies (12) and \( \varepsilon > 0 \). Then, it holds that

\[
\sup_{(X,Y):P_e^{(L)}(X|Y)\leq \varepsilon,P_X=Q} h_\phi(X \mid Y) < \infty \iff \phi(Q) < \infty,
\]

(179)

where the supremum is taken over the \( X \times Y \)-valued r.v.’s \((X,Y)\) satisfying \( P_e^{(L)}(X \mid Y) \leq \varepsilon \) and \( P_X = Q \).

Proof of Proposition 8: The “if” part \( \iff \) of Proposition 8 is quite obvious from Jensen’s inequality even if \( \phi : \mathcal{P}(X) \to [0, \infty] \) is not of the form (178). Hence, it suffices to prove the “only if” part \( \Rightarrow \). In other words, we shall prove

\[
\phi(Q) = \infty \implies \sup_{(X,Y):P_e^{(L)}(X|Y)\leq \varepsilon,P_X=Q} h_\phi(X \mid Y) = \infty.
\]

(180)

In the following, we show (180) by employing Lemma 6 of Section B-B.

Since \( g_2(u) = \infty \) only if \( u = \infty \), it is immediate from (178) that

\[
\phi(Q) = \infty \implies \sum_{x \in X} g_1(Q(x)) = \infty,
\]

(181)

where note that \( \phi(Q) = \infty \) implies \( g_2(\infty) = \infty \) as well. Moreover, since \( g_1(0) = 0 \), we get

\[
\sum_{x \in X} g_1(Q(x)) = \infty \implies \#(\text{supp}(Q)) = \infty.
\]

(182)

Due to (12), we can find a finite subset \( S \subset Y \) satisfying

\[
1 - \sum_{x \in \mathcal{X} \setminus S} Q(x) \leq \varepsilon
\]

(183)

by taking a finite but sufficiently large cardinality \( \#(S) < \infty \). This implies that the new system \((Q, L, \varepsilon, S)\) still fulfills (12); and thus, it follows from Proposition 3 that there exists an \( X \times S \)-valued r.v. \((X,Y)\) satisfying \( P_e^{(L)}(X \mid Y) \leq \varepsilon \) and \( P_X = Q \).

Therefore, the feasible region \( \mathcal{R}_2 = \mathcal{R}(Q, L, \varepsilon, S) \) defined in (78) is nonempty by this choice of \( S \). Since \( S \subset Y \), it is clear that \( \mathcal{R}_2 \subset \mathcal{R}_1 \), where \( \mathcal{R}_1 = \mathcal{R}(Q, L, \varepsilon, Y) \).

By Lemma 6, one can find \( Z \subset X \) so that \( \#(Z) = \#(Y) \) and \( \mathcal{R}_3 = \mathcal{R}(Q, L, \varepsilon, S, Z) \) defined in (118) is nonempty as well. Moreover, since \( P_e^{(L)}(X \mid Y) \leq P_e^{(L)}(X \mid Y \parallel Z) \), if follows that \( \mathcal{R}_3 \subset \mathcal{R}_2 \). Then, we have

\[
\sup_{(X,Y):P_e^{(L)}(X|Y)\leq \varepsilon,P_X=Q} h_\phi(X \mid Y) \overset{(a)}{=} \sup_{(X,Y)\in \mathcal{R}_1} h_\phi(X \mid Y)
\]

(184)
where (a) follows by the definition (78); (b) follows by the inclusions \( 0 \neq \mathcal{R}_3 \subset \mathcal{R}_2 \subset \mathcal{R}_1 \); (c) follows from the fact that \((X, Y) \in \mathcal{R}_3\) implies \(P_{X|Y=y}(x) = Q(x)\) for \(x \in \mathcal{X} \setminus \mathcal{Z}\) and \(y \in \mathcal{S}\); and (d) follows from the facts that (i) \(#(\text{supp}(Q) \setminus \mathcal{Z}) = \infty\), (ii) \(g_1(u) \geq 0\) for \(0 \leq u \leq 1\), and (iii) \(g_2(\infty) = \infty\). Inequalities (184) just ensure (180), completing the proof of Proposition 8.

As seen in Section II-D, the conditional Shannon and Rényi entropies can be expressed by \(h_\alpha(X \mid Y)\) while fulfilling the additional postulate (178) on \(\phi\). Proposition 8 shows that we cannot establish an effective Fano-type inequality based on the conditional information measure \(h_\alpha(X \mid Y)\) subject to our original postulates in Section II-A, provided that (i) \(\phi\) satisfies the additional postulate (178), (ii) \(\varepsilon > 0\), and (iii) \(\phi(Q) = \infty\).

REFERENCES

[1] R. Ahlswede, “An elementary proof of the strong converse theorem for the multiple-access channel,” J. Combinat., Inf. Syst. Sci., vol. 7, no. 3, pp. 216–230, 1982.
[2] R. Ahlswede, P. Gács, and J. Körner, “Bounds on conditional probabilities with applications in multi-user communication,” Z. Wahrschein. Verw. Geb., vol. 34, no. 3, pp. 157–177, Jan. 1976.
[3] M. S. Ali and D. Silvey, “A general class of coefficients of divergence of one distribution from another,” J. Roy. Statist. Soc., series B, vol. 28, no. 1, pp. 131–142, 1966.
[4] S. Arimoto, “Information measures and capacity of order \(\alpha\) for discrete memoryless channels,” in Topics Inf. Theory, 2nd Colloq. Math. Soc. J. Bolyai, Keszthely, Hungary, vol. 6, pp. 41–52, 1977.
[5] G. Birkhoff, “Tres observaciones sobre el algebra lineal,” Univ. Nac. Tucumán Rev. Ser. A, vol. 5, pp. 147–151, 1946.
[6] I. Csiszár and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, 2nd ed., Cambridge: Cambridge Univ. Press, 2011.
[7] T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[8] T. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed., New York: Wiley, 2006.
[9] M. Hayashi, “Exponential decreasing rate of leaked information in universal random privacy amplification,” IEEE Trans. Inf. Theory, vol. 60, no. 5, pp. 2598–2614, May 2014.
[10] A. El Gamal and Y.-H. Kim, Network Information Theory, Cambridge, UK: Cambridge University Press, 2011.
[11] T. Cover and J. Thomas, Elements of Information Theory, revised edition, Amer. Math. Soc., 1991.
[12] G. Dueck, “The strong converse to the coding theorem for the multiple-access channel,” J. Combinat., Inf. Syst. Sci., vol. 6, no. 3, pp. 187–196, 1981.
[13] R. Durrett, Probability: Theory and Examples, 4th ed., Cambridge: Cambridge Univ. Press, 2010.
[14] P. Elias, “List decoding for noisy channels,” in IRE WESCON Conv. Rec., vol. 2, 1957, pp. 94–104.
[15] T. van Erven and P. Harremoës, “Rényi divergence and Kullback–Leibler divergence,” IEEE Trans. Inf. Theory, vol. 56, no. 7, pp. 3797–3820, July 2014.
[16] A. Rényi, “On measures of entropy and information,” in Proc. 4th Berkeley Symp. Math. Statist. Prob., Berkeley, Calif., vol. 1, Univ. of Calif. Press, pp. 547–561, 1961.
[17] A. Rényi, “On measures of entropy and information,” in Proc. 4th Berkeley Symp. Math. Statist. Prob., Berkeley, Calif., vol. 1, Univ. of Calif. Press, pp. 547–561, 1961.
[18] A. Rényi, “On measures of entropy and information,” in Proc. 4th Berkeley Symp. Math. Statist. Prob., Berkeley, Calif., vol. 1, Univ. of Calif. Press, pp. 547–561, 1961.
[19] P. Révész, “A probabilistic solution of problem 111 of G. Birkhoff,” Acta Math. Hungar., vol. 3, nos. 1–2, pp. 188–198, Mar. 1962.
[44] Y. Sakai, “Generalized Fano-type inequality for countably infinite systems with list decoding,” in Proc. Int. Symp. Inf. Theory Appl. (ISITA), Singapore, Oct. 2018.
[45] Y. Sakai and K. Iwata, “Sharp bounds on Arimoto’s conditional Rényi entropies between two distinct orders,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Aachen, Germany, June 2017, pp. 2975–2979. [Online] Available: https://arxiv.org/abs/1702.00014.
[46] I. Sason, “Entropy bounds for discrete random variables via maximal coupling,” IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7118–7131, Nov. 2013.
[47] I. Sason and S. Verdú, “Arimoto–Rényi conditional entropy and Bayesian M-ary hypothesis testing,” IEEE Trans. Inf. Theory, vol. 64, no. 1, pp. 4–25, Jan. 2018.
[48] C. E. Shannon, “A mathematical theory of communication,” Bell Syst. Tech. J., vol. 27, nos. 3–4, pp. 379–423 and 623–656, July/Oct. 1948.
[49] ———, “Communication theory of secrecy systems,” Bell Syst. Tech. J., vol. 28, no. 4, pp. 656–715, Nov. 1949.
[50] M. E. Shirokov, “On properties of the space of quantum states and their application to the construction of entanglement monotones,” Izv. Math., vol. 74, no. 4, pp. 849–882, 2010.
[51] B. Shuval and I. Tal, “Fast polarization for processes with memory,” submitted to IEEE Trans. Inf. Theory, 2017. [Online]. Available: https://arxiv.org/abs/1710.02849.
[52] E. Şaşoğlu, “Polarization and polar codes,” Found. Trends Commun. Inf. Theory, vol. 8, no. 4, pp. 259–381, Oct. 2012.
[53] D. L. Tebbe and S. J. Dwyer III, “Uncertainty and probability of error,” IEEE Trans. Inf. Theory, vol. 14, no. 3, pp. 516–518, May 1968.
[54] V. Y. F. Tan and M. Hayashi, “Analysis of remaining uncertainties and exponents under various conditional Rényi entropies,” IEEE Trans. Inf. Theory, vol. 64, no. 5, pp. 3734–3755, May 2018.
[55] A. Teixeira, A. Matos, and L. Antunes, “Conditional Rényi entropies,” IEEE Trans. Inf. Theory, vol. 58, no. 7, pp. 4273–4277, Jul. 2012.
[56] H. Thorisson, Coupling, stationarity, and regeneration. vol. 14, New York: Springer, 2000.
[57] F. Topsøe, “Basic concepts, identities and inequalities—the toolkit of information theory,” Entropy, vol. 3, no. 3, pp. 162–190, Sept. 2001.
[58] S. Verdú and T. S. Han, “The role of the asymptotic equipartition property in noiseless coding theorem,” IEEE Trans. Inf. Theory, vol. 43, no. 3, pp. 847–857, May 1997.
[59] J. Wolfowitz, Coding Theorems of Information Theory. 3rd ed., New York: Springer, 1978.
[60] R. W. Yeung, Information Theory and Network Coding. New York: Springer, 2008.