Arimoto Channel Coding Converse and Rényi Divergence

Yury Polyanskiy and Sergio Verdú

***Abstract*—Arimoto [1] proved a non-asymptotic upper bound on the probability of successful decoding achievable by any code on a given discrete memoryless channel. In this paper we present a simple derivation of the Arimoto converse based on the data-processing inequality for Rényi divergence. The method has two benefits. First, it generalizes to codes with feedback and gives the simplest proof of the strong converse for the DMC with feedback. Second, it demonstrates that the sphere-packing bound is strictly tighter than Arimoto converse for all channels, blocklengths and rates, since in fact we derive the latter from the former. Finally, we prove similar results for other (non-Rényi) divergence measures.**

***Index Terms*—Shannon theory, strong converse, information measures, Rényi divergence, feedback.**

I. Introduction

In [1], Arimoto has shown a simple non-asymptotic bound, that implies a (strengthening of the) strong converse to the channel coding for the DMC. Moreover, his bound is exponentially tight for rates above the capacity.

To state Arimoto's bound, recall that Gallager's $E_0(\rho, P_X, P_{Y|X})$, $\rho \neq 1$ function is defined for a pair of random variables $X \in \mathsf{A}$ and $Y \in \mathsf{B}$ as follows:

$$\begin{aligned} & E_0(\rho, P_X, P_{Y|X}) \\ &= -\log \sum_{y\in\mathsf{B}} \left(\sum_{x\in\mathsf{A}} P_X(x) P_{Y|X}^{\frac{1}{1+\rho}}(y|x) \right)^{1+\rho} \quad (1) \\ &= -\log \mathbb{E}\left[\left(\mathbb{E}\left[\exp\left\{ \frac{i(\bar{X};Y)}{1+\rho} \right\} \middle| Y \right] \right)^{1+\rho} \right], \quad (2) \end{aligned}$$

where the second expression is a generalization to the case of infinite alphabets, where

$$i(x;y) \stackrel{\triangle}{=} \log \frac{dP_{XY}}{d(P_X \times P_Y)}(x,y)\,, \quad (3)$$

and the joint distribution of $(\bar{X}, Y)$ is given by

$$P_{\bar{X}Y}(\bar{x}, y) = P_X(\bar{x}) P_Y(y)\,. \quad (4)$$

A random transformation is defined by a pair of measurable spaces of inputs A and outputs B and a conditional probability measure $P_{Y|X} : \mathsf{A} \mapsto \mathsf{B}$. An (M, ϵ) *code* for the random transformation $(\mathsf{A}, \mathsf{B}, P_{Y|X})$ is a pair of (possibly randomized) maps $f : \{1, \ldots, M\} \to \mathsf{A}$ (the encoder) and $g : \mathsf{B} \to \{1, \ldots, M\}$ (the decoder), satisfying

$$\frac{1}{M} \sum_{m=1}^{M} P[g(Y) \neq m | X = f(m)] \leq \epsilon\,. \quad (5)$$

Without loss of generality we assume that $\epsilon \leq 1 - \frac{1}{M}$. In applications, we will take A and B to be n-fold Cartesian products of alphabets $\mathcal{A}$ and $\mathcal{B}$, and a channel to be a sequence of random transformations $\{P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n\}$ [2]. An (M, ϵ) code for $\{\mathcal{A}^n, \mathcal{B}^n, P_{Y^n|X^n}\}$ is called an (n, M, ϵ) code. For the statement and proof of the main converse bounds, it is preferable not to assume that A and B have any structure such as a Cartesian product. This has the advantage of avoiding the notational clutter that results from explicitly showing the dimension (n) of the random variables taking values on A and B.

Arimoto has shown the following result:

Theorem 1 ([1]): The probability of error ϵ of any (M, ϵ) code for the random transformation $(\mathsf{A}, \mathsf{B}, P_{Y|X})$ satisfies for any $-1 < \rho < 0$

$$\epsilon \geq 1 - M^{\rho} \exp\{-E_0(\rho, P_X, P_{Y|X})\}\,, \quad (6)$$

where P_X is the distribution induced on A by the encoder.

Note that the bound (6) applies to an arbitrary codebook. To obtain a universal bound (i.e. the one whose right side depends only on M) one needs to take the infimum over all distributions P_X. When the blocklength of the code is large, a direct optimization becomes prohibitively complex. However, the following result resolves this difficulty and makes Theorem 1 especially useful:

Theorem 2 (Gallager-Arimoto): Consider the product channel $P_{Y^2|X^2}$ given by

$$P_{Y^2|X^2}(y_1 y_2 | x_1 x_2) = P_{Y_1|X_1}(y_1|x_1) P_{Y_2|X_2}(y_2|x_2)\,. \quad (7)$$

Then for all $-1 < \rho < 0$ we have [1]

$$\begin{aligned} & \min_{P_{X^2}} E_0(\rho, P_{X^2}, P_{Y^2|X^2}) \\ &= \min_{P_{X_1}} E_0(\rho, P_{X_1}, P_{Y_1|X_1}) + \min_{P_{X_2}} E_0(\rho, P_{X_2}, P_{Y_2|X_2}) \quad (8) \end{aligned}$$

Similarly, for $\rho > 0$ we have [3]

$$\begin{aligned} & \max_{P_{X^2}} E_0(\rho, P_{X^2}, P_{Y^2|X^2}) \\ &= \max_{P_{X_1}} E_0(\rho, P_{X_1}, P_{Y_1|X_1}) + \max_{P_{X_2}} E_0(\rho, P_{X_2}, P_{Y_2|X_2}) \quad (9) \end{aligned}$$

Or in other words, the extremum in the left-hand sides of (8) and (9) is achieved by the product distributions.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544 USA. e-mail: {ypolyans,verdu}@princeton.edu.

The research was supported by the National Science Foundation under Grants CCF-06-35154 and CCF-07-28445.

Application of Theorems 1 and 2 to the DMC of blocklength n, i.e. the channel $(\mathcal{A}^n, \mathcal{B}^n, (P_{Y|X})^n)$, one obtains that any $(n, \exp\{nR\}, \epsilon)$ code over the DMC satisfies

$$\epsilon \geq 1 - \exp\left\{-n \sup_{-1<\rho<0}\left[\min_{P_X} E_0(\rho, P_X, P_{Y|X}) - \rho R\right]\right\}, \tag{10}$$

The bound (10) has a number of very useful properties:

1) it is non-asymptotic (i.e. valid for any $n \geq 1$),
2) it is universal (i.e., the only data about the code appearing in the right-hand side is the code rate R),
3) it is single-letter (i.e., its computational complexity is independent of the blocklength n),
4) a further analysis, see [1], shows that the exponent is negative for all $R > C$, thus proving a (strengthening of the) strong converse, which shows that above capacity the minimum probability of error goes to 1 exponentially fast with the blocklength. Moreover, it is known that this lower bound is exponentially tight in the sense that there exist a sequence of codes of rate R achieving the exponent [4, Problem 2.5.16b]. The counterpart in data compression is given by [4, Problem 1.2.6].

A drawback of the bound (10), severely limiting its use for finite blocklength analysis, is that the right-hand side vanishes for any $R \leq C$. In this paper we present a strengthening of the Arimoto bound which overcomes this drawback while retaining all the mentioned advantages. In particular, for $R < C$ it yields a non-trivial exponential lower bound on the probability of error, which although results in a weaker bound on the error exponent than the Shannon-Gallager-Berlekamp's sphere-packing bound, is much simpler and is applicable to non-discrete channels. We give two different proofs of this result, each having its own benefits. The first proof demonstrates that Arimoto's result is implied by the minimax converse shown in [5]. In particular this implies that for symmetric channels, the sphere-packing bound is always tighter than (10) for all rates and blocklengths. The second proof demonstrates that Arimoto's result is a simple consequence of the data-processing inequality for an asymmetric information measure introduced by Sibson [6] and Csiszár [7]. The proof parallels the standard derivation of the Fano's inequality and appears to be the simplest known proof of the strong converse for memoryless channels. In particular, no measure concentration inequalities are employed.

The second proof admits an important generalization to the case of codes with feedback. Namely, we show that (10) holds in this exact form for (block) codes with feedback. Although, this result is known asymptotically [4, Problem 2.5.16c], the non-asymptotic bound appears to be proven here for the first time[1]. A converse bound valid for all DMCs might prove to be helpful in the ongoing effort of establishing the validity of the sphere-packing exponent for codes with feedback over a general DMC [9], [10].

[1]A similar result can be extracted with a circuitous route from [8] whose proof contains gaps as pointed out in [9].

Finally, we conclude by showing which of the results generalize to other divergence measures, and which are special to Rényi divergence. A family of bounds obtained by fixing an arbitrary f-divergence includes Fano's inequality (corresponding to relative entropy), Arimoto converse (corresponding to Rényi divergence) and Wolfowitz strong converse (e.g., [5, Theorem 9]).

II. Arimoto converse: a proof via meta-converse

A. Preliminaries

One of the main tools in our treatment [5] is the performance of an optimal binary hypothesis test defined as follows. Consider a W-valued random variable W which can take probability measures P or Q. A randomized test between those two distributions is defined by a random transformation $P_{Z|W} : \mathsf{W} \mapsto \{0, 1\}$ where 0 indicates that the test chooses Q. The best performance achievable among those randomized tests is given by[2]

$$\beta_\alpha(P, Q) = \min \sum_{w\in\mathsf{W}} Q(w) P_{Z|W}(1|w)\,, \tag{11}$$

where the minimum is over all probability distributions $P_{Z|W}$ satisfying

$$P_{Z|W} : \sum_{w\in\mathsf{W}} P(w) P_{Z|W}(1|w) \geq \alpha\,. \tag{12}$$

The minimum in (11) is guaranteed to be achieved by the Neyman-Pearson lemma. Thus, $\beta_\alpha(P, Q)$ gives the minimum probability of error under hypothesis Q if the probability of error under hypothesis P is not larger than $1 - \alpha$.

In [5] we have shown that a number of classical converse bounds, including Fano's inequality, Shannon-Gallager-Berlekamp, Wolfowitz strong converse and Verdú-Han information spectrum converse, can be obtained in a unified manner as a consequence of the meta-converse theorem [5, Theorem 26]. One of such consequences is the following minimax converse [5]:

Theorem 3 (minimax converse): Every (M, ϵ) code satisfies

$$M \leq \sup_{P_X} \inf_{Q_Y} \frac{1}{\beta_{1-\epsilon}(P_{XY}, P_X \times Q_Y)}\,, \tag{13}$$

where P_X ranges over all input distributions on A, and Q_Y ranges over all output distributions on B.

The traditional sphere-packing bound for symmetric channels follows from Theorem 3 by choosing Q_Y to be equiprobable on the finite output alphabet. For this reason, Theorem 3 can be viewed as a natural generalization of the sphere-packing bound.

The Rényi divergence for $\lambda > 0$, $\lambda \neq 1$ is [11]

$$D_\lambda(P||Q) = \frac{1}{\lambda - 1} \log \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^\lambda\right]. \tag{14}$$

[2]We write summations over alphabets for simplicity; however, all of our general results hold for arbitrary probability spaces.

Normalization ensures that

$$\lim_{\lambda\to 1} D_\lambda(P||Q) = D(P||Q)\,, \tag{15}$$

where $D(P||Q)$ is the relative entropy:

$$D(P||Q) \stackrel{\triangle}{=} \mathbb{E}_Q\left[\frac{dP}{dQ}\log\frac{dP}{dQ}\right]. \tag{16}$$

Although Rényi divergence is not an f-divergence in the sense of [12], it is a monotone transformation of the Hellinger divergence of order λ, see [13]. Since Hellinger divergence is an f-divergence for all $\lambda > 0, \lambda \neq 1$, the data-processing inequality automatically follows for Rényi divergence as well.

Additionally, we define a conditional Rényi divergence as follows:

$$\begin{aligned} &D_\lambda(P_{A|B}||Q_{A|B}|P_B) \\ &\stackrel{\triangle}{=} \frac{1}{\lambda-1}\log\sum_{b\in\mathcal{B}} P_B(b)\exp\{(\lambda-1)D_\lambda(P_{A|B=b}||Q_{A|B=b})\} &(17)\\ &= \frac{1}{\lambda-1}\log\sum_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}} P_B(b)P^\lambda_{A|B}(a|b)Q^{1-\lambda}_{A|B}(a|b) &(18)\\ &= D_\lambda(P_B\times P_{A|B}||P_B\times Q_{A|B})\,. &(19)\end{aligned}$$

Two obvious consequences of the definition are identities

$$\sup_{P_B} D_\lambda(P_{A|B}||Q_{A|B}|P_B) = \sup_{b\in\mathcal{B}} D_\lambda(P_{A|B=b}||Q_{A|B=b}) \tag{20}$$

and

$$D_\lambda(P_{AB}||Q_{AB}) = D_\lambda(P_B||Q_B) + D_\lambda(P_{A|B}||Q_{A|B}|P_B^{(\lambda)})\,, \tag{21}$$

where $P_B^{(\lambda)}$ is the λ-tilting of P_B towards Q_B given by

$$P_B^{(\lambda)}(b) \stackrel{\triangle}{=} P_B^\lambda(b)Q_B^{1-\lambda}(b)\exp\{-(\lambda-1)D_\lambda(P_B||Q_B)\}\,. \tag{22}$$

The binary Rényi divergence is given by

$$\begin{aligned} d_\lambda(p||q) &\stackrel{\triangle}{=} D_\lambda([p\ 1-p]\,||\,[q\ 1-q]) &(23)\\ &= \tfrac{1}{\lambda-1}\log\left(p^\lambda q^{1-\lambda} + (1-p)^\lambda(1-q)^{1-\lambda}\right) &(24)\end{aligned}$$

B. Main result

Theorem 4: Any (M,ϵ) code satisfies for $\lambda > 0, \lambda \neq 1$:

$$d_\lambda(1-\epsilon||\tfrac{1}{M}) \le \frac{\lambda}{1-\lambda}E_0(\lambda^{-1}-1, P_X, P_{Y|X}) \tag{25}$$

and in particular for any $-1 < \rho < 0$ we obtain (6) (letting $\lambda = \frac{1}{1+\rho} > 1$).

Proof: The key observation is that Gallager's $E_0(\rho, P_X, P_{Y|X})$ for $\rho > -1$ is given by a Rényi divergence of order $\lambda = \frac{1}{1+\rho}$:

$$\frac{\lambda}{1-\lambda}E_0(\lambda^{-1}-1, P_X, P_{Y|X}) = D_\lambda(P_{XY}||P_X\times Q_Y^*)\,, \tag{26}$$

where the auxiliary output distribution Q_Y^* is defined implicitly via

$$\begin{aligned} &\frac{dQ_Y^*}{dP_Y}(Y) \stackrel{\triangle}{=} \\ &\left(\mathbb{E}\left[\exp\{\lambda i(\bar{X};Y)\}|Y\right]\right)^{\frac{1}{\lambda}}\exp\{E_0(\tfrac{1-\lambda}{\lambda}, P_X, P_{Y|X})\} &(27)\end{aligned}$$

where we adopted the convention (3), (4).

Now from Theorem 3 we know that any (M,ϵ) code satisfies

$$\beta_{1-\epsilon}(P_{XY}, P_X\times Q_Y^*) \le \frac{1}{M}\,. \tag{28}$$

Applying the data-processing for Rényi divergence we get

$$d_\lambda(1-\epsilon||\beta_{1-\epsilon}(P_{XY}, P_X\times Q_Y^*)) \le D_\lambda(P_{XY}, P_X\times Q_Y^*)\,. \tag{29}$$

In view of (28) and $1-\epsilon \ge \frac{1}{M}$, (29) implies

$$d_\lambda(1-\epsilon||\tfrac{1}{M}) \le D_\lambda(P_{XY}, P_X\times Q_Y^*)\,. \tag{30}$$

Application of (26) completes the proof of (25).

To obtain (6) observe a simple inequality:

$$d_\lambda(1-\epsilon||\tfrac{1}{M}) \ge \frac{1}{\lambda-1}\log\left((1-\epsilon)^\lambda M^{\lambda-1} + \epsilon^\lambda\right). \tag{31}$$

If we take $-1 < \rho < 0$ and let $\lambda = \frac{1}{1+\rho} > 1$ we can further lower-bound d_λ:

$$d_\lambda(1-\epsilon||\tfrac{1}{M}) \ge \frac{\lambda}{\lambda-1}\log(1-\epsilon) + \log M\,, \tag{32}$$

which together with (25) implies (6). ∎

As already mentioned, Theorem 4 extends Arimoto's result. First, as shown, inequality (25) is stronger than (6). Moreover, the family of bounds (25) includes Fano's inequality:

$$(1-\epsilon)\log M - h(\epsilon) \le I(X;Y)\,, \tag{33}$$

which is obtained by taking $\lambda \to 1$ (see (15)). This does not happen with (25) when $\rho \to 0$. Second, inequality (25) extends (6) to $\rho > 0$ as follows:

$$\epsilon \ge \left(\exp\left\{-\frac{1}{1+\rho}E_0(\rho, P_X, P_{Y|X})\right\} - M^{-\frac{\rho}{1+\rho}}\right)^{1+\rho}. \tag{34}$$

If we now apply (34) to codes over the DMC of blocklength n and take infimum over all P_{X^n} via Theorem 2 (similar to the derivation of (10)) we get an exponential lower bound on ϵ valid for all $(n, \exp\{nR\}, \epsilon)$ codes:

$$\epsilon \ge \exp\{-n\rho^* R + o(n)\}\,, \tag{35}$$

where ρ^* is found as a solution to

$$\max_{P_X} E_0(\rho, P_X, P_{Y|X}) = \rho R\,. \tag{36}$$

Compared to the derivation of the sphere-packing bound in [14], the bound (35) is much easier to obtain, but, alas, $\rho^* R$ is always larger than the sphere-packing exponent. Note also that for $R \ge C$ the solution $\rho^* = 0$ and (35) shows that exponentially small probabilities are impossible for such rates, a fact also clear from (10).

III. A second proof of Theorem 4

Observe that in the proof of Theorem 4 the inequality (31) holds even if Q_Y^* is replaced with an arbitrary measure Q_Y:

$$d_\lambda(1-\epsilon||\tfrac{1}{M}) \le D_\lambda(P_{XY}, P_X \times Q_Y)\,. \tag{37}$$

To obtain the best bound we may minimize the right-hand side over the choice of Q_Y. However, as noted in [7], the identity of Sibson [6]

$$D_\lambda(P_{XY}||P_XQ_Y) = D_\lambda(P_{XY}||P_XQ_Y^*) + D_\lambda(Q_Y^*||Q_Y) \tag{38}$$

shows that such method does not lead to any improvement since Q_Y^*, defined in (27), is in fact the minimizer:

$$\inf_{Q_Y} D_\lambda(P_{XY}||P_XQ_Y) = D_\lambda(P_{XY}||P_XQ_Y^*)\,. \tag{39}$$

This leads us naturally to the following asymmetric information measure, introduced by Csiszár in [7]:

$$K_\lambda(X;Y) \stackrel{\triangle}{=} \inf_{Q_Y} D_\lambda(P_{XY}||P_XQ_Y)\,. \tag{40}$$

In the special case of discrete P_X (40) was introduced by Sibson in [6]. Using Sibson identity (38) we obtain the following equivalent expressions for $K_\lambda(X;Y)$

$$\begin{aligned} K_\lambda(X;Y) &= D_\lambda(P_{XY}||P_XQ_Y^*) && (41)\\ &= \frac{\lambda}{1-\lambda}E_0(\lambda^{-1}-1, P_X, P_{Y|X}) && (42)\\ &= \frac{\lambda}{\lambda-1}\log\sum_{y\in\mathsf{B}}\left(\sum_{x\in\mathsf{A}}P_X(x)P_{Y|X}^\lambda(y|x)\right)^{\frac{1}{\lambda}} && (43)\end{aligned}$$

Notice that in [15] for the purpose of finding an efficient algorithm for computing $\sup_{P_X} E_0(\rho, P_X, P_{Y|X})$ Arimoto has shown a variational representation for E_0 (and therefore for K_λ; see (42)) different from (40).

An important property of $K_\lambda(X;Y)$ shown by Csiszár [7] is the following:

$$\sup_{P_X} K_\lambda(X;Y) = \inf_{Q_Y}\sup_x D_\lambda(P_{Y|X=x}||Q_Y)\,. \tag{44}$$

One application of (44) is a direct proof of Theorem 2:

$$\begin{aligned} &\sup_{P_{X_1X_2}} K_\lambda(X_1X_2;Y_1Y_2)\\ &= \inf_{Q_{Y_1Y_2}}\sup_{x_1,x_2} D_\lambda(P_{Y_1|X_1=x_1}P_{Y_2|X_2=x_2}||Q_{Y_1Y_2}) && (45)\\ &\le \inf_{Q_{Y_1}Q_{Y_2}}\sup_{x_1,x_2} D_\lambda(P_{Y_1|X_1=x_1}P_{Y_2|X_2=x_2}||Q_{Y_1}Q_{Y_2}) && (46)\\ &= \inf_{Q_{Y_1}}\sup_{x_1} D_\lambda(P_{Y_1|X_1=x_1})\\ &\quad + \inf_{Q_{Y_2}}\sup_{x_2} D_\lambda(P_{Y_2|X_2=x_2}) && (47)\\ &= \sup_{P_{X_1}} K_\lambda(X_1;Y_1) + \sup_{P_{X_2}} K_\lambda(X_2;Y_2)\,. && (48)\end{aligned}$$

Note that the more cumbersome original proofs of Theorem 2 relied on the Karush-Kuhn-Tucker conditions, which require additional justification in non-discrete settings.

We notice as a side remark that the maximum of $K_\lambda(X;Y)$ over P_X is known as the capacity of order λ. It was defined in [16] as a maximization of a different information measure (based on Rényi entropy). A simple algorithm for its computation is derived in [15]. In [7] it was shown that the same value for the capacity of order λ is obtained by maximizing two other information measures (based on Rényi divergence), one of them $K_\lambda(X;Y)$; see also [17].

The next result relates $K_\lambda(X;Y)$ to a proof of Theorem 4 and also demonstrates a remarkable resemblance between the properties of $K_\lambda(X;Y)$ and the mutual information $I(X;Y)$.

Theorem 5: For $\lambda > 0, \lambda \ne 1$ the following holds.

1) The function $f_\lambda(K_\lambda(X;Y))$ is convex in P_X and concave in $P_{Y|X}$, where $f_\lambda(x) = \frac{1}{\lambda-1}\exp\{(1-\lambda^{-1})x\}$ is monotonically increasing.
2) For random variables $W-X-Y-Z$ forming a Markov chain the following holds
$$K_\lambda(W;Z) \le K_\lambda(X;Y)\,. \tag{49}$$
3) If X and Y take values in the same set $\{1,\dots,M\}$ and X is equiprobable, then
$$\min_{P_{Y|X}:\mathbb{P}[X\ne Y]\le\epsilon} K_\lambda(X;Y) = d_\lambda(1-\epsilon||\tfrac{1}{M}) \tag{50}$$
if $\epsilon \le 1-\frac{1}{M}$ and minimum is equal to zero otherwise.

Proof: Property 1 follows by noticing that

$$f_\lambda(K_\lambda(X;Y)) = \frac{1}{\lambda-1}\sum_{y\in\mathsf{B}}\left(\sum_{x\in\mathsf{A}}P_X(x)P_{Y|X}^\lambda(y|x)\right)^{\frac{1}{\lambda}} \tag{51}$$

and applying convexity (concavity) of $x^{\frac{1}{\lambda}}$ for $\lambda < 1$ ($\lambda > 1$). The concavity in $P_{Y|X}$ follows from Minkowski inequality.

To show Property 3, consider an arbitrary P_{XY} with $\mathbb{P}[X \ne Y] = s$. Then the data-processing for Rényi divergence applied to the transformation $(X,Y) \to 1\{X \ne Y\}$ shows

$$D_\lambda(P_{XY}||P_XQ_Y) \ge d_\lambda(1-s||\tfrac{1}{M})\,. \tag{52}$$

Since the function in the left-hand side is decreasing for all $s \le 1-\frac{1}{M}$, we find that

$$\min_{P_{Y|X}:\mathbb{P}[X\ne Y]\le\epsilon} K_\lambda(X;Y) \ge d_\lambda(1-\epsilon||\tfrac{1}{M})\,. \tag{53}$$

provided that $\epsilon \le 1-\frac{1}{M}$. On the other hand, the lower bound is achieved by the kernel $P_{Y|X}$ defined as:

$$P_{Y|X}(y|x) = \begin{cases} 1-\epsilon\,, & x = y\\ \frac{\epsilon}{M-1}\,, & x \ne y\,. \end{cases} \tag{54}$$

The proof of Property 2 is the key step. Notice because of the asymmetric nature of $K_\lambda(X;Y)$ we must prove two statements separately:

- "data post-processing": if $X-Y-Z$ form a Markov chain, then
$$K_\lambda(X;Z) \le K_\lambda(X;Y)\,. \tag{55}$$
This inequality follows from the following argument. For an arbitrary Q_Y denote
$$Q_Z(b) = \sum_{y\in\mathsf{B}} Q_Y(y)P_{Z|Y}(b|y)\,. \tag{56}$$

Then by the data-processing for Rényi divergence we have:

$$D_\lambda(P_{XZ}||P_XQ_Z) \le D_\lambda(P_{XY}||P_XQ_Y)\,. \tag{57}$$

Taking infimum over Q_Y and using the definition of $K_\lambda(X;Z)$ shows (55).

- "data pre-processing": if $W - X - Y$ form a Markov chain, then

$$K_\lambda(W;Y) \le K_\lambda(X;Y)\,. \tag{58}$$

Consider the computation of $D(P_{XY}||P_XQ_Y)$. For a fixed Q_Y the random variable (X,Y) is distributed either as P_{XY} or as P_XQ_Y. Observe that applying random transformation $P_{WY|XY}$ to (X,Y) we obtain (W,Y) distributed either as P_{WY} or as P_WQ_Y (the Markov property is needed to see that the distribution of W is P_W in the alternative hypothesis). Then by the data-processing for Rényi divergence:

$$D_\lambda(P_{WY}||P_WQ_Y) \le D_\lambda(P_{XY}||P_XQ_Y)\,, \tag{59}$$

which implies (58) after taking infimum over Q_Y. ∎

Proof of Theorem 4: Notice that an (M,ϵ) code defines four random variables forming a Markov chain $W - X - Y - \hat{W}$, where W is the message (equiprobable on $\{1,\dots,M\}$), X is the channel input, Y is the channel output and $\hat{W}$ is the decoder estimate of the message W. Then Properties 2 and 3 (Theorem 5) together imply Theorem 4. ∎

Inequality (25) applied to an arbitrary (n,M,ϵ) code for the channel $P_{Y^n|X^n}$ states that

$$d_\lambda(1-\epsilon||\tfrac{1}{M}) \le K_\lambda(X^n;Y^n)\,, \tag{60}$$

where X^n has the distribution induced by the encoder. Maximizing the right-hand side of (60) over all P_{X^n} is particularly simple for memoryless channels since when $P_{Y^n|X^n} = (P_{Y|X})^n$, then by (48) we have

$$\sup_{P_{X^n}} K_\lambda(X^n;Y^n) = n\sup_{P_X} K_\lambda(X;Y) \tag{61}$$

and hence from (60) we get the following result:

Corollary 6: Every (n,M,ϵ) code for a memoryless channel $(\mathcal{A}^n,\mathcal{B}^n,(P_{Y|X})^n)$ satisfies

$$d_\lambda(1-\epsilon||\tfrac{1}{M}) \le n\sup_{P_X} K_\lambda(X;Y)\,. \tag{62}$$

As explained in Section II inequality (62) further simplifies to either (10) when $\lambda > 1$ or to (35) when $\lambda < 1$.

IV. Codes with feedback

In [18] Shannon showed that the capacity of a DMC does not increase even if we allow the encoder to use a full noiseless instantaneous feedback. In this Section we demonstrate that, moreover, the non-asymptotic bound in Corollary 6, continues to hold even in the setting of Shannon feedback. A precise definition of the feedback code can also bee found in [4, Problem 2.1.27], for example.

Theorem 7: Every (n,M,ϵ) feedback code for a memoryless channel $(\mathcal{A}^n,\mathcal{B}^n,(P_{Y|X})^n)$ satisfies (62) and in particular (10).

Proof: Take an arbitrary (n,M,ϵ) feedback code. Then it induces a certain joint distribution on (W,Y^n) according to

$$P_{WY^n}(w,y^n) = \frac{1}{M}\prod_{i=1}^n P_{Y|X}(y_i|f_i(w,y^{i-1}))\,, \tag{63}$$

where $f_i : \{1,\dots,M\}\times\mathcal{B}^{i-1} \to \mathcal{A}$, $i = 1,\dots,n$ are the encoder maps. The decoder estimate $\hat{W}$ is obtained as a (possibly randomized) function of Y^n and therefore $W - Y^n - \hat{W}$ form a Markov chain. By Theorem 5 (Property 2) we have

$$K_\lambda(W;\hat{W}) \le K_\lambda(W;Y^n)\,, \tag{64}$$

and by Theorem 5 (Property 3) we have

$$K_\lambda(W;\hat{W}) \ge \frac{1}{\lambda-1}\log\left((1-\epsilon)^\lambda M^{\lambda-1} + \epsilon^\lambda\right)\,. \tag{65}$$

To conclude the proof we need to show that

$$K_\lambda(W;Y^n) \le n\sup_{P_X} K_\lambda(X;Y)\,. \tag{66}$$

To that end consider the following chain:

$$\begin{aligned}
&K_\lambda(W;Y^n)\\
&\stackrel{\triangle}{=} \inf_{Q_{Y^n}} D_\lambda(P_{WY^n}||P_WQ_{Y^n}) && (67)\\
&= \inf_{Q_{Y^n}}\Big[D_\lambda(P_{WY^{n-1}}||P_WQ_{Y^{n-1}})\\
&\quad + D_\lambda(P_{Y_n|Y^{n-1}W}||Q_{Y_n|Y^{n-1}}|P^{(\lambda)}_{WY^{n-1}})\Big] && (68)\\
&= \inf_{Q_{Y^{n-1}}}\Big[D_\lambda(P_{WY^{n-1}}||P_WQ_{Y^{n-1}})\\
&\quad + \inf_{Q_{Y_n|Y^{n-1}}} D_\lambda(P_{Y_n|Y^{n-1}W}||Q_{Y_n|Y^{n-1}}|P^{(\lambda)}_{WY^{n-1}})\Big] && (69)\\
&\le \inf_{Q_{Y^{n-1}}}\Big[D_\lambda(P_{WY^{n-1}}||P_WQ_{Y^{n-1}})\\
&\quad + \inf_{Q_{Y_n}} D_\lambda(P_{Y_n|Y^{n-1}W}||Q_{Y_n}|P^{(\lambda)}_{WY^{n-1}})\Big] && (70)\\
&\le \inf_{Q_{Y^{n-1}}}\Big[D_\lambda(P_{WY^{n-1}}||P_WQ_{Y^{n-1}})\\
&\quad + \inf_{Q_{Y_n}}\sup_{x\in\mathcal{A}} D_\lambda(P_{Y_n|X_n=x}||Q_{Y_n})\Big] && (71)\\
&= \inf_{Q_{Y^{n-1}}} D_\lambda(P_{WY^{n-1}}||P_WQ_{Y^{n-1}})\\
&\quad + \sup_{P_X} K_\lambda(X;Y) && (72)\\
&= K_\lambda(W;Y^{n-1}) + \sup_{P_X} K_\lambda(X;Y)\,, && (73)
\end{aligned}$$

where (68) is by (21), (69) follows since the first term does not depend on $Q_{Y_n|Y^{n-1}}$, (70) follows by restricting the infimum to $Q_{Y_n|Y^{n-1}} = Q_{Y_n}$, (71) is by (20), (72) is by (44), and (73) is by the definition of K_λ in (40). The proof of (66) now follows from (73) by induction. ∎

V. Generalization to other divergence measures

Notice that the key Properties 2 and 3 of K_λ needed for the proof of Theorem 4 also hold (with the same proof) if the Rényi divergence D_λ in (40) is replaced by any other function of a pair of distributions, satisfying the data-processing inequality; for example, any f-divergence works as well. This section formalizes this idea.

First, consider a measurable space W and, a pair of distributions P and Q on it and a transition probability kernel $P_{W'|W}$ from W to W. Applying $P_{W'|W}$ to P and Q we obtain a pair of distribution P' and Q':

$$P'(w') = \sum_{w\in\mathsf{W}} P_{W'|W}(w'|w)P(w) \quad (74)$$

$$Q'(w') = \sum_{w\in\mathsf{W}} P_{W'|W}(w'|w)Q(w)\,. \quad (75)$$

Definition 1: A function $\mathcal{D}(P||Q)$ assigning an extended real number to a pair of distributions is called a generalized divergence, or a *g-divergence*, if for any $P_{W'|W}$ we have

$$\mathcal{D}(P'||Q') \le \mathcal{D}(P||Q)\,. \quad (76)$$

Note that restricting transformations to those mapping W to W is made without loss of generality, as we can consider that the space W is rich enough to contain copies of any A and B considered in the given problem and therefore, the function $\mathcal{D}$ satisfies the data-processing inequality with respect to transformations from A to B as well.

Examples of g-divergences:

- All f-divergences [12], [17], in particular total variation, relative entropy and Hellinger divergence [13].
- Rényi divergence; note that it is a non-decreasing function of the Hellinger divergence.
- $-\beta_\alpha(P,Q)$ for any $0 \le \alpha \le 1$. This example shows that the class of g-divergences is larger than just non-decreasing functions of f-divergences, since $-\beta_\alpha(P,Q)$ cannot be obtained from any f-divergence[3].

To any g-divergence $\mathcal{D}(P||Q)$ we define a binary g-divergence $\delta(p||q)$ as the divergence between the distributions on $\{0,1\}$ given by $P(1)=p$ and $Q(1)=q$; formally,

$$\delta(p||q) \stackrel{\triangle}{=} \mathcal{D}([p\ 1-p]||[q\ 1-q])\,. \quad (78)$$

Following the approach of Sibson [6] and Csiszár [7] for any g-divergence we define an information measure

$$\mathcal{K}(X;Y) \stackrel{\triangle}{=} \inf_{Q_Y} \mathcal{D}(P_{XY}||P_XQ_Y)\,. \quad (79)$$

The following theorem summarizes the results that can be obtained by the same methods as above:

Theorem 8: Consider a g-divergence $\mathcal{D}(P||Q)$. Then all of the following hold:

1) Any (M,ϵ) code for the random transformation $(\mathsf{A},\mathsf{B},P_{Y|X})$ satisfies

$$\delta(1-\epsilon||\tfrac{1}{M}) \le \sup_{P_X}\inf_{Q_Y}\mathcal{D}(P_{XY}||P_XQ_Y) \quad (80)$$

$$= \sup_{P_X}\mathcal{K}(X;Y) \quad (81)$$

$$\le \inf_{Q_Y}\sup_{P_X}\mathcal{D}(P_{XY}||P_XQ_Y) \quad (82)$$

2) For random variables $W-X-Y-Z$ forming a Markov chain the following holds

$$\mathcal{K}(W;Z) \le \mathcal{K}(X;Y)\,. \quad (83)$$

3) If X and Y are taking values in the same set $\{1,\dots,M\}$ and X is equiprobable, then

$$\min_{P_{Y|X}:\mathbb{P}[X\neq Y]\le\epsilon}\mathcal{K}(X;Y) = \delta(1-\epsilon||\tfrac{1}{M}) \quad (84)$$

if $\epsilon \le 1-\frac{1}{M}$ and minimum is equal to $\delta(\frac{1}{M}||\frac{1}{M})$ otherwise.

4) If $\mathcal{D}(P||Q)$ is an f-divergence, then we have an equality in (82) and

$$\sup_{P_X}\mathcal{D}(P_{XY}||P_XQ_Y) = \sup_{x\in\mathsf{A}}\mathcal{D}(P_{Y|X=x}||Q_Y)\,. \quad (85)$$

In particular, for $\mathcal{K}(X;Y)$ we have

$$\sup_{P_X}\mathcal{K}(X;Y) = \inf_{Q_Y}\sup_{x\in\mathsf{A}}\mathcal{D}(P_{Y|X=x}||Q_Y)\,. \quad (86)$$

Remark: What this theorem shows is that many of the properties of D_λ are common to all g-divergences. However, what makes D_λ special is additivity under products:

$$D_\lambda(P_1P_2||Q_1Q_2) = D_\lambda(P_1||Q_1) + D_\lambda(P_2||Q_2)\,, \quad (87)$$

which results in identities like (38) and (21), and in turn in single-letter bounds like (10).

Proof: Notice that any hypothesis test between P_{XY} and P_XQ_Y is a random transformation from $\mathsf{A}\times\mathsf{B}$ to $\{0,1\}$. Applying the data-processing property for $\mathcal{D}$ we get that any test attaining probabilities of success $1-\epsilon$ and $1-\beta$ over P_{XY} and $P_X\times Q_Y$, respectively, must satisfy

$$\delta(1-\epsilon||\beta_{1-\epsilon}) \le \mathcal{D}(P_{XY},P_X\times Q_Y)\,. \quad (88)$$

Note that the data-processing property implies that whenever $p\le p'\le q$ we have

$$\delta(p'||q) \le \delta(p||q) \quad (89)$$

and a similar monotonicity in the second argument. Since by Theorem 3, $\beta_{1-\epsilon} \le \frac{1}{M}$ and $\frac{1}{M} \le 1-\epsilon$ by assumption, we have from (88):

$$\delta(1-\epsilon||\tfrac{1}{M}) \le \mathcal{D}(P_{XY},P_X\times Q_Y)\,. \quad (90)$$

Therefore, taking first infimum over all Q_Y and the supremum over all P_X we get (80). Then (81) is by definition (79) and (82) is obvious.

Proofs of (83) and (84) are exact repetition of the proofs of Properties 2 and 3 in Theorem 5, since there we have not

[3]Assume otherwise, then we would have (see Theorem 8, Property 4) that

$$\inf_{P_X}\beta_\alpha(P_{XY}||P_XQ_Y) = \inf_{x\in\mathsf{A}}\beta_\alpha(P_{Y|X=x},Q_Y)\,, \quad (77)$$

but it is easy to construct a counter-example where this does not hold.

used any special properties of the Rényi divergence, except the data-processing property.

Finally, when $\mathcal{D}(P||Q)$ is an f-divergence then $\mathcal{D}(P_{XY}||P_XQ_Y)$ is linear in P_X and convex in Q_Y. Thus the equality in (82) follows from the minimax theorem by interchanging $\sup$ and $\inf$ exactly as explained by Csiszár [7] in the proof of (44). (85) follows from linearity of $\mathcal{D}(P_{XY}||P_XQ_Y)$ in P_X. Finally, (86) follows from (85) and the equality in (82). ■

Remark: Examples of the application of Theorem 8 (Property 1) include:

- Fano's inequality: take $\mathcal{D}$ to be the relative entropy.
- Theorem 4: take $\mathcal{D}$ to be Rényi divergence D_λ.
- Wolfowitz strong converse, e.g. [5, Theorem 9]: take $\mathcal{D}$ to be an f-divergence appearing in the DT-bound [5, (78)], $f = |x-\gamma|^+$.

If we apply Theorem 8 with a g-divergence given by $-\beta_\alpha(P,Q)$, $0 \le \alpha \le 1$, we get the following (equivalent) form of Theorem 3:

Corollary 9: Every (M,ϵ) code satisfies for all $0 \le \alpha \le 1$:

$$\inf_{P_X}\sup_{Q_Y} \beta_\alpha(P_{XY}||P_XQ_Y) \le \frac{\alpha}{M(1-\epsilon)} + \left(\frac{1}{\epsilon} - \frac{1}{M(1-\epsilon)}\right)|\alpha - 1 + \epsilon|^+ \quad (91)$$

where P_X ranges over all input distributions on A, and Q_Y ranges over all out distributions on B.

Taking $\alpha = 1-\epsilon$ in (91) one recovers Theorem 3. The additional benefit of stating the minimax problem in this form is that it demonstrates that to bound the cardinality of a code for a given ϵ, it is not required to evaluate β_α for $\alpha = 1-\epsilon$. In fact, determining the value of β_α for any α sufficiently close to $1-\epsilon$ also works. This is useful when β_α is computed via a Neyman-Pearson lemma.

REFERENCES

[1] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 19, no. 3, pp. 357 – 359, May 1973.

[2] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, 1994.

[3] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inf. Theory*, vol. 11, no. 1, pp. 3–18, 1965.

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

[5] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[6] R. Sibson, "Information radius," *Z. Wahrscheinlichkeitstheorie und Verw. Geb.*, vol. 14, pp. 149–161, 1969.

[7] I. Csiszár, "Generalized cutoff rates and Renyi's information measures," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26 –34, Jan. 1995.

[8] A. Y. Sheverdyaev, "Lower bound for error probability in a discrete memoryless channel with feedback," *Prob. Peredachi Inform.*, vol. 18, no. 4, pp. 5–15, 1982.

[9] G. Como and B. Nakiboğlu, "Sphere-packing bound for block-codes with feedback and finite memory," in *Proc. 2010 IEEE Int. Symp. Inf. Theory (ISIT)*, Austin, TX, USA, Jun. 2010.

[10] H. Palaiyanur and A. Sahai, "An upper bound for the block coding error exponent with delayed feedback," in *Proc. 2010 IEEE Int. Symp. Inf. Theory (ISIT)*, Austin, TX, USA, Jun. 2010.

[11] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Mathematics, Statistics, and Probability*, vol. 1, Berkeley, CA, USA, 1961, pp. 547–561.

[12] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Sci. Math. Hungar.*, vol. 2, pp. 229–318, 1967.

[13] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.

[14] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels i," *Inf. Contr.*, vol. 10, pp. 65–103, 1967.

[15] S. Arimoto, "Computation of random coding exponent functions," *IEEE Trans. Inf. Theory*, vol. 22, no. 6, pp. 665 – 671, Nov. 1976.

[16] ——, "Information measures and capacity of order α for discrete memoryless channels," in *Topics in Information Theory*, ser. Colloq. Math. Soc. J. Bolyai 16, I. Csiszár and P. Elias, Eds. Amsterdam: North Holland, 1977, pp. 41–52.

[17] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–73, 2008.

[18] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.