Relative Entropy at the Channel Output of a Capacity-Achieving Code

Yury Polyanskiy and Sergio Verdú

***Abstract*—In this paper we establish a new inequality tying together the coding rate, the probability of error and the relative entropy between the channel and the auxiliary output distribution. This inequality is then used to show the strong converse, and to prove that the output distribution of a code must be close, in relative entropy, to the capacity achieving output distribution (for DMC and AWGN). One of the key tools in our analysis is the concentration of measure (isoperimetry).**

***Index Terms*—Shannon theory, strong converse, information measures, empirical output statistics, concentration of measure, general channels, discrete memoryless channels, additive white Gaussian noise.**

I. Introduction

The problem of constructing capacity achieving channel codes has been one of the main focuses of information and coding theories. In this paper we demonstrate some of the properties that such codes must necessarily posses. Such characterization facilitates the search for the good codes; leads to strong converses; may prove useful for establishing converse bounds in multi-user communication problems where frequently the code used at one terminal creates interference for others [1]; helps in the context of secure communication, where output statistics of the code is required to resemble the white noise; and also becomes crucial in the problem of asynchronous communication where the output statistics of the code imposes the limits on the quality of synchronization [2], [3].

Specifically, this paper focuses on the properties of the output distribution induced by a capacity achieving code. In this regard, [4] showed that capacity achieving codes with *vanishing* probability of error, satisfy [4, Theorem 2]:

$$\frac{1}{n}D(P_{Y^n}||P^*_{Y^n}) \to 0\,, \tag{1}$$

where P_{Y^n} denotes the output distribution of the code and P^*_Y the unique capacity achieving output distribution. As will be explained below, bounding the relative entropy $D(P_{Y^n}||P^*_{Y^n})$ leads to precision guarantees for the approximation of expectations

$$\int f(y^n)dP_{Y^n} \approx \int f(y^n)dP^*_{Y^n}\,.$$

In this paper we extend (1) to the case of *non-vanishing* probability of error. The motivation comes from the fact that the analysis of fundamental limits in the regime of fixed probability of error proves to be quite fruitful for non-asymptotic characterization of attainable performance over a given channel [5]. It turns out that extension of (1) only holds under the maximal probability of error criterion and inherently relies on the phenomenon of concentration of measure (isoperimetry).

The organization of the paper is as follows. Section II contains the main definitions and notation. In Section III a key inequality is derived upon which all of the results of the rest of the paper are based. Section IV presents a sufficient condition for the strong converse which simultaneously captures most of the cases considered in the literature. Sections V and VI prove (1) for a class of discrete memoryless channels (DMCs) and the additive white Gaussian noise (AWGN) channel. Section VII discusses a number of useful implications of the convergence (1). Finally, Section VIII demonstrates a technique for extending some of the results to channels for which no estimate (1) is known.

II. Notation

A random transformation $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ is a Markov kernel acting between a pair of measurable spaces. An $(M,\epsilon)_{avg}$ code for the random transformation $P_{Y|X}$ is a pair of random transformations $f : \{1,\dots,M\} \to \mathcal{X}$ and $g : \mathcal{Y} \to \{1,\dots,M\}$ such that

$$\mathbb{P}[\hat{W} \neq W] \leq \epsilon\,, \tag{2}$$

where the underlying probability space is

$$W \xrightarrow{f} X \overset{P_{Y|X}}{\to} Y \xrightarrow{g} \hat{W} \tag{3}$$

with W equiprobable on $\{1,\dots,M\}$. An $(M,\epsilon)_{max}$ code is defined similarly except that (2) is replaced with a more stringent maximal probability of error criterion:

$$\max_{1\leq j\leq M} \mathbb{P}[\hat{W} \neq W|W=j] \leq \epsilon\,. \tag{4}$$

A code is called deterministic, denoted $(M,\epsilon)_{det}$, if the encoder f is a functional (non-random) mapping.

For each random transformation $P_{Y|X}$ we define:

- maximal mutual information:

$$\mathcal{C} = \sup_{P_X} I(X;Y)\,, \tag{5}$$

  which we assume to be finite.
- a set of capacity achieving input distributions, *caid*'s:

$$\Pi = \{P_X : I(X;Y) = \mathcal{C}\}\,. \tag{6}$$

Y. Polyanskiy is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139 USA, e-mail: yp@mit.edu. S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544 USA. e-mail: verdu@princeton.edu.

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

Under the assumption $\mathcal{C} < \infty$, the set is non-empty [4].

- capacity achieving output distribution, *caod*:

$$P_Y^*(\cdot) = \int_{\mathsf{B}} P_{Y|X}(\cdot|x) P_X^*(dx)\,, \tag{7}$$

where $P_X^* \in \Pi$.

The important fact is that despite non-uniqueness of caid, caod is in fact unique [4]. Moreover, we have the following estimates [4]

$$\begin{aligned} D(P_{Y|X}||P_Y^*|P_X) &\le \mathcal{C} && (8)\\ D(P_Y||P_Y^*) &\le \mathcal{C} - I(X;Y)\,, && (9)\end{aligned}$$

where P_X is an arbitrary input distribution. In particular (9) shows that P_Y^* dominates all possible output distributions:

$$\begin{aligned} P_Y &\ll P_Y^* \qquad \forall P_X && (10)\\ P_{Y|X=x} &\ll P_Y^*, \qquad \forall x \in \mathcal{X}\,. && (11)\end{aligned}$$

A channel is a sequence of random transformations, $\{P_{Y^n|X^n}, n = 1, \ldots\}$ indexed by the parameter n, referred to as the blocklength. In this paper we assume that $\mathcal{C}_n$, the maximal unnormalized mutual informations associated to $P_{Y^n|X^n}$, are finite for all $n = 1, \ldots,$ and

$$\mathcal{C}_n \to \infty\,, \qquad n \to \infty\,.$$

A channel (used without feedback) is called memoryless if

$$P_{Y^n|X^n=x^n} = \prod_{i=1}^n P_{Y|X=x_i}\,, \tag{12}$$

where $P_{Y|X}$ is a single-letter kernel. For a memoryless channel with no input constraints $\mathcal{C}_n = nC$, where $C = \mathcal{C}_1$ is the capacity of the channel. An (M, ϵ) code for the n-th random transformation is called an (n, M, ϵ) code. A sequence of (n, M_n, ϵ) codes is called capacity achieving if

$$\log M_n = \mathcal{C}_n + o(\mathcal{C}_n)\,. \tag{13}$$

We also need to introduce the performance of an optimal binary hypothesis test, which was one of the main tools in our previous treatment [5]. 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[1]

$$\beta_\alpha(P, Q) = \min \sum_{w \in \mathsf{W}} Q(w) P_{Z|W}(1|w)\,, \tag{14}$$

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

$$P_{Z|W} : \sum_{w \in \mathsf{W}} P(w) P_{Z|W}(1|w) \ge \alpha\,. \tag{15}$$

The minimum in (14) 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 $1 - \alpha$.

[1]We sometimes write summations over alphabets for simplicity of exposition; in fact, the definition holds for arbitrary measurable spaces.

III. Key inequality

Theorem 1: Consider a random transformation $P_{Y|X}$, a distribution P_X induced by an $(M, \epsilon)_{max,det}$ code and an auxiliary output distribution Q_Y. Assume that for all $x \in \mathcal{X}$ we have

$$d(x) \stackrel{\triangle}{=} D(P_{Y|X=x}||Q_Y) < \infty \tag{16}$$

and

$$\sup_x P_{Y|X=x}\left[\log \frac{dP_{Y|X=x}}{dQ_Y}(Y) \ge d(x) + \Delta\right] \le \delta'\,, \tag{17}$$

for some pair of constants $\Delta \ge 0$ and $0 \le \delta' < 1 - \epsilon$. Then we have

$$D(P_{Y|X}||Q_Y|P_X) \ge \log M - \Delta + \log \frac{1 - \epsilon - \delta'}{e}\,. \tag{18}$$

Proof: Fix arbitrary t and choose an (M', ϵ) subcode by including only codewords belonging to the set

$$A_t \stackrel{\triangle}{=} \{x : d(x) \le t\}\,. \tag{19}$$

Note that

$$M' = M\mathbb{P}[d(X) \le t]\,. \tag{20}$$

By the meta-converse [5, Theorem 31] we have

$$\inf_{x \in A_t} \beta_{1-\epsilon}(P_{Y|X=x}, Q_Y) \le \frac{1}{M'}\,. \tag{21}$$

On the other hand, using the standard lower bound on β [5, (102)]

$$\begin{aligned} &\beta_{1-\epsilon}(P_{Y|X=x}, Q_Y)\\ &\quad \ge \frac{1}{\gamma(x)}\left(1 - \epsilon - P_{Y|X=x}\left[\frac{P_{Y|X=x}}{Q_Y} \ge \gamma(x)\right]\right)\,, \end{aligned} \tag{22}$$

where $\gamma(x) = \exp\{d(x) + \Delta\}$. According to (17) we have

$$P_{Y|X=x}\left[\frac{P_{Y|X=x}}{Q_Y} \ge \gamma(x)\right] \le \delta'\,, \tag{23}$$

which applied to (22) implies

$$\beta_{1-\epsilon}(P_{Y|X=x}, Q_Y) \ge \frac{1}{\gamma(x)}\left(1 - \epsilon - \delta'\right)\,. \tag{24}$$

Plugging this back into (21) we get

$$\begin{aligned} \frac{1}{M'} &\ge \exp\{-\Delta - \sup_{x \in A_t} d(x)\}\left(1 - \epsilon - \delta'\right) && (25)\\ &\ge \exp\{-\Delta - t\}\left(1 - \epsilon - \delta'\right) && (26)\end{aligned}$$

But then from (20) we have for all t

$$\mathbb{P}[d(X) \le t] \le \frac{1}{1 - \epsilon - \delta'} \exp\{t + \Delta - \log M\}\,. \tag{27}$$

In other words,

$$\mathbb{P}[d(X) > t] \ge 1 - \frac{1}{1 - \epsilon - \delta'} \exp\{t + \Delta - \log M\}\,. \tag{28}$$

Integrating (28) over t we obtain

$$\begin{aligned} &\mathbb{E}\left[d(X)\right]\\ &\quad \ge \int_0^{t_m} \left(1 - \tfrac{1}{1-\epsilon-\delta'} \exp\{t + \Delta - \log M\}\right) dt \end{aligned} \tag{29}$$

where t_m is found by solving

$$1 - \frac{1}{1-\epsilon-\delta'} \exp\{t_m + \Delta - \log M\} = 0\,, \tag{30}$$

which yields

$$t_m = \log M(1-\epsilon-\delta') - \Delta\,. \tag{31}$$

Continuing from (29) we have

$$\begin{aligned}
\mathbb{E}\left[d(X)\right] &\geq \int_0^{t_m} (1 - \exp\{t - t_m\})\, dt && (32)\\
&= t_m - \int_{-t_m}^{0} \exp\{x\} dx && (33)\\
&\geq t_m - \int_{-\infty}^{0} \exp\{x\} dx && (34)\\
&= t_m - \log e && (35)\\
&= \log M(1-\epsilon-\delta') - \Delta - \log e\,. && (36)
\end{aligned}$$

■

One way to estimate the upper deviations in (17) is using Chebyshev's inequality. As an example, we obtain

Corollary 2: If in the conditions of Theorem 1 we replace (17) with[2]

$$\sup_x \mathrm{Var}\left[\log \frac{dP_{Y|X=x}}{dQ_Y}(Y)\middle| X = x\right] \leq S_m \tag{37}$$

for some constant $S_m \geq 0$, then we have

$$D(P_{Y|X}||Q_Y|P_X) \geq \log M - \sqrt{\frac{2S_m}{1-\epsilon}} + \log\frac{1-\epsilon}{2e}\,. \tag{38}$$

IV. Application: general channels

Our first application of Theorem 1 is in proving a general strong converse. Recall that a channel, e.g. [6, Definition 1], is a sequence of random transformations $P_{Y^n|X^n} : \mathcal{X}^n \to \mathcal{Y}^n$. Let $\mathcal{C}_n$ be the associated sequence of maximal mutual informations. Then a sequence of output distributions Q_{Y^n} is said to be *quasi-caod* if

$$\sup_{P_{X^n}} D(P_{Y^n|X^n}||Q_{Y^n}|P_{X^n}) \leq \mathcal{C}_n + o(\mathcal{C}_n)\,, \tag{39}$$

where the supremum is over all distributions on $\mathcal{X}^n$. Note that under the assumption of measurability of singletons in $\mathcal{X}^n$, (39) is equivalent to

$$\sup_{x\in\mathcal{X}^n} D(P_{Y^n|X^n=x}||Q_{Y^n}) \leq \mathcal{C}_n + o(\mathcal{C}_n)\,. \tag{40}$$

By taking P_{X^n} to be a capacity achieving input distribution, the $o(\mathcal{C}_n)$ term in (39) (and thus in (40)) is shown to be non-negative; it is zero precisely for those n for which Q_{Y^n} is the caod. For completeness, we notice that requiring $D(P^*_{Y^n}||Q_{Y^n}) = o(\mathcal{C}_n)$ and $D(Q_{Y^n}||P^*_{Y^n}) = o(\mathcal{C}_n)$ is not sufficient for Q_{Y^n} to be quasi-caod[3].

[2]Of course, variance in (37) is computed with Y distributed according to $P_{Y|X=x}$.

[3]For a counter-example, consider the sequence of n-ary symmetric channels with a fixed crossover probability δ (so that $\mathcal{C}_n = (1-\delta)\log n + o(\log n)$). Then set Q_{Y^n} equiprobable on $n-1$ elements and equal to $\frac{1}{n^2}$ on the remaining one.

The motivation for introducing quasi-caods is the following. For memoryless channels without input constraints, one can easily see that the caod $P^*_{Y^n}$ for blocklength n is simply an n-th power of a single-letter caod P^*_Y:

$$P^*_{Y^n} = (P^*_Y)^n\,. \tag{41}$$

At the same time, in the presence of input constraints finding the n-letter caod maybe problematic. For example, for the AWGN channel with SNR P and blocklength n the input space is

$$\mathcal{X}^n = \{x^n \in \mathbb{R}^n : \sum x_i^2 \leq nP\}\,. \tag{42}$$

Thus finding the caod involves solving a maximization problem for the mutual information over the input distributions supported on the ball, whose solution may not be straightforward. It is easy to show, however, that for the present channel $\mathcal{C}_n = nC + o(n)$. Thus the product-Gaussian distribution

$$Q^*_{Y^n} = \mathcal{N}(0, (1+P)\mathbf{I}_n) \tag{43}$$

is readily seen to be a quasi-caod. $Q^*_{Y^n}$ can be found by the following method: every input distribution over the ball is an element of a wider family of distributions satisfying

$$\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \leq nP\,. \tag{44}$$

Maximization of the mutual information over this wider family is easy and the corresponding output distribution is the product Gaussian (43).

Definition 1: Consider a channel $\{P_{Y^n|X^n}, n = 1,\ldots\}$ with a sequence of maximal mutual informations $\mathcal{C}_n$. A sequence of codes $\{\mathcal{F}_n, n = 1,\ldots\}$ for the channel is called strongly information stable if there exist sequences of numbers $\Delta_n = o(\mathcal{C}_n)$ and $\delta_n \to 0$ and a quasi-caod sequence $\{Q_{Y^n}, n = 1,\ldots\}$ such that

$$\sup_{x\in\mathcal{F}_n} P_{Y^n|X^n=x}\left[\log\frac{dP_{Y^n|X^n=x}}{dQ_{Y^n}} \geq nd_n(x) + \Delta_n\right] \leq \delta_n\,, \tag{45}$$

where

$$d_n(x) \stackrel{\triangle}{=} \frac{1}{n} D(P_{Y^n|X^n=x}||Q_{Y^n})\,. \tag{46}$$

The channel is called *strongly information stable* if (45) holds with supremum extended to the whole of $\mathcal{X}^n$.

Note that Definition 1 places no constraint on how reliable the code is, nor on its rate. Note also that a channel is Dobrushin information stable if for a sequence of caod's $\{P^*_{Y^n}, n = 1,\ldots\}$ one has for some $C \geq 0$

$$\frac{1}{n}\log\frac{dP_{Y^n|X^n}}{dP^*_{Y^n}}(Y^n|X^n) \to C \tag{47}$$

in probability, where X^n is distributed according to a capacity achieving input distribution. Thus, our definition is stronger in requiring concentration for each X^n as opposed to taking the average with respect to a capacity achieving distribution.

Theorem 3 (Strong converse): If channel $\{P_{Y^n|X^n}, n = 1,\ldots\}$ is strongly information stable then for any $0 < \epsilon < 1$ and any sequence of $(n, M_n, \epsilon)_{avg}$ codes we have

$$\log M_n \le \mathcal{C}_n + o(\mathcal{C}_n)\,. \tag{48}$$

Remark: Typically $\mathcal{C}_n = nC + o(n)$, in which case the right-hand side of (48) becomes

$$\log M_n \le nC + o(n)\,. \tag{49}$$

Proof: Since the probability of error ϵ is in the average sense, we can assume without loss of generality that the encoder is deterministic. Then standard expurgation shows that for any $\epsilon' > \epsilon$ there is a sequence of $(n, M_n', \epsilon')_{max,det}$ subcodes with

$$M_n' \ge cM_n\,, \tag{50}$$

for a certain constant $0 < c \le 1$. Then by Theorem 1 and (40) we have

$$\log M_n' \le \mathcal{C}_n + o(\mathcal{C}_n) + \Delta_n - \log\frac{1-\epsilon-\delta_n}{e}\,, \tag{51}$$

where (Δ_n, δ_n) are from (45). Together (50) and (51) prove (48). ∎

The sufficient conditions of Theorem 3 are quite general and capture many of the cases considered previously in the literature, including most memoryless and ergodic channels. One exception is the scalar fading channel [7] with memoryless fading process, where unfortunately the multiplicative random factor disables the estimate (45). In that case, however, one can show that every code must necessarily have a large, information stable subcode to which in turn Theorem 1 can be applied precisely as in the preceding proof.

Theorem 4: Consider a sequence of $(n, M_n, \epsilon)_{max,det}$ codes which is both capacity-achieving and information stable. Then

$$I(X^n;Y^n) = \mathcal{C}_n + o(\mathcal{C}_n) \iff D(P_{Y^n}||Q_{Y^n}) = o(\mathcal{C}_n)\,, \tag{52}$$

where P_{X^n} and P_{Y^n} are the input and output distributions induced by the n-th code, and Q_{Y^n} is the quasi-caod sequence from Definition 1.

Remark: For memoryless channels $\mathcal{C}_n = nC + o(n)$ and $Q_{Y^n} = (P_Y^*)^n$ and thus (52) can be restated as

$$\frac{1}{n}I(X^n;Y^n) \to C \iff \frac{1}{n}D(P_{Y^n}||(P_Y^*)^n) \to 0\,. \tag{53}$$

Proof: The direction $\Rightarrow$ is trivial from the definition of quasi-caod and the identity

$$I(X^n;Y^n) = D(P_{Y^n|X^n}||Q_{Y^n}|P_{X^n}) - D(P_{Y^n}||Q_{Y^n})\,. \tag{54}$$

For the direction $\Leftarrow$ we have from (13), Definition 1 and Theorem 1

$$D(P_{Y^n|X^n}||Q_{Y^n}|P_{X^n}) \ge \mathcal{C}_n + o(\mathcal{C}_n)\,.$$

Then the conclusion follows from (54) and the fact that by definition $I(X^n;Y^n) \le \mathcal{C}_n$. ∎

We remark that Theorems 3 and 4 can also be derived from a simple extension of the Wolfowitz converse [8], see also [5, Theorem 9], to an arbitrary output distribution Q_Y.

V. Application: DMC

Theorem 5: Consider a DMC $P_{Y|X}$ with capacity $C > 0$. Then for any sequence of $(n, M_n, \epsilon)_{max,det}$ codes achieving capacity, i.e.

$$\lim_{n\to\infty}\frac{1}{n}\log M_n = C\,, \tag{55}$$

we have

$$\frac{1}{n}D(P_{Y^n}||P_{Y^n}^*) \to 0\,, \tag{56}$$

where P_{Y^n} is the output distribution induced by the code and $P_{Y^n}^* = (P_Y^*)^n$ is the multi-letter caod, which is an n-th power of the single-letter caod P_Y^*. The claim need not hold if the maximal probability of error is replaced with the average of if the encoder is allowed to be random.

Remark: If P_Y^* is equiprobable on $\mathcal{Y}$ (such as for some symmetric channels), (56) is equivalent to

$$H(Y^n) = nH(Y^*) + o(n)\,. \tag{57}$$

In any case (56) always implies (57) as (104) applied to $f(y) = \log\frac{1}{P_Y^*(y)}$ shows. Note also that traditional combinatorial methods, e.g. [9], are not helpful in dealing with quantities like $H(Y^n)$, $D(P_{Y^n}||P_{Y^n}^*)$ or P_{Y^n}-expectations of functions which are not of the form of cumulative average.

Proof: Here we only present a proof under an additional assumption that the transition matrix does not contain zeros: $P_{Y|X}(\cdot|\cdot) > 0$. Fix $y^n \in \mathcal{Y}^n$, $1 \le j \le n$ and denote

$$y^n(b)_j = (y_1,\ldots,y_{j-1},b,y_{j+1},\ldots,y_n)\,. \tag{58}$$

Then,

$$\begin{aligned}
&|\log P_{Y^n}(y^n) - \log P_{Y^n}(y^n(b)_j)| \\
&\quad= \left|\log\frac{P_{Y_j|Y_{\hat{j}}}(y_j|y_{\hat{j}})}{P_{Y_j|Y_{\hat{j}}}(b|y_{\hat{j}})}\right| && (59)\\
&\quad\le \max_{a,b,b'}\log\frac{P_{Y|X}(b|a)}{P_{Y|X}(b'|a)} && (60)\\
&\quad\stackrel{\triangle}{=} a_1 < \infty\,. && (61)
\end{aligned}$$

Therefore, the discrete gradient (see definition of $D(f)$ in [10, Section 4]) of the function $\log P_{Y^n}(y^n)$ on $\mathcal{Y}^n$ is bounded by $n|a_1|^2$ and thus by the discrete Poincaré inequality [10, Theorem 4.1f] we have

$$\operatorname{Var}\left[\log P_{Y^n}(Y^n)|X^n = x^n\right] \le n|a_1|^2\,. \tag{62}$$

Therefore, for some $0 < a_2 < \infty$ and all $x^n \in \mathcal{X}^n$ we have

$$\begin{aligned}
&\operatorname{Var}\left[\log\frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)}\middle|X^n = x^n\right] \\
&\quad\le 2\operatorname{Var}\left[\log P_{Y^n|X^n}(Y^n|X^n)\middle|X^n = x^n\right] \\
&\qquad + 2\operatorname{Var}\left[\log P_{Y^n}(Y^n)|X^n = x^n\right] && (63)\\
&\quad\le 2na_2 + 2n|a_1|^2\,, && (64)
\end{aligned}$$

where (64) follows from the fact that $\log P_{Y^n|X^n}$ is a sum of independent random variables and (62). Applying Corollary 2 with $S_m = 2na_2 + 2n|a_1|^2$ and $Q_Y = P_{Y^n}$ we obtain:

$$D(P_{Y^n|X^n}||P_{Y^n}|P_{X^n}) \ge \log M_n + O(\sqrt{n})\,. \tag{65}$$

We can now complete the proof:

$$\begin{aligned}
&D(P_{Y^n}||P^*_{Y^n})\\
&= D(P_{Y^n|X^n}||P^*_{Y^n}|P_{X^n}) - D(P_{Y^n|X^n}||P_{Y^n}|P_{X^n}) && (66)\\
&\le nC - D(P_{Y^n|X^n}||P_{Y^n}|P_{X^n}) && (67)\\
&\le nC - \log M_n + O(\sqrt{n}) && (68)\\
&\le o(n)\,, && (69)
\end{aligned}$$

where (67) is because $P^*_{Y^n}$ is the caod and (8), (68) follows from (65) and (69) is because the considered sequence of codes is capacity achieving (55). Clearly, (69) is equivalent to (56).

Next we show that (56) cannot hold if the maximal probability of error is replaced with the average. To that end, consider a sequence of $(n, M'_n, \epsilon'_n)_{max,det}$ codes with $\epsilon'_n \to 0$ and

$$\frac{1}{n}\log M'_n \to C\,. \tag{70}$$

For all n such that $\epsilon'_n < \frac{1}{2}$ this code cannot have repeated codewords and we can additionally assume (perhaps by reducing M'_n by one) that there is no codeword equal to $(x_0, \dots, x_0) \in \mathcal{X}^n$, where x_0 is some fixed letter in $\mathcal{X}$ such that

$$D(P_{Y|X=x_0}||P^*_Y) > 0 \tag{71}$$

(existence of such x_0 relies on the assumption $C > 0$). Denote the output distribution induced by this code by P'_{Y^n}.

Next, extend this code by adding $\frac{\epsilon-\epsilon_n}{1-\epsilon}M'_n$ codewords which all coincide and are equal to $(x_0, \dots, x_0) \in \mathcal{X}^n$. Then the average probability of error of the extended code is easily seen to be not larger than ϵ. Denote the output distribution induced by the extended code by P_{Y^n} and define a binary random variable

$$S = 1\{X^n = (x_0, \dots, x_0)\} \tag{72}$$

with distribution

$$P_S(1) = 1 - P_S(0) = \frac{\epsilon - \epsilon'_n}{1 - \epsilon'_n}\,. \tag{73}$$

We have then

$$\begin{aligned}
&D(P'_{Y^n}||P^*_{Y^n})\\
&= D(P_{Y^n|S}||P^*_{Y^n}|P_S) - D(P_{S|Y^n}||P_S|P_{Y^n}) && (74)\\
&\ge D(P_{Y^n|S}||P^*_{Y^n}|P_S) - a_1 && (75)\\
&= nD(P_{Y|X=x_0}||P^*_Y)P_S(1) + D(P'_{Y^n}||P^*_{Y^n})P_S(0) - a_1 && (76)\\
&= nD(P_{Y|X=x_0}||P^*_Y)P_S(1) + o(n)\,, && (77)
\end{aligned}$$

where (74) is by the usual chain-rule for the relative entropy, (75) follows since S is binary and therefore for all sufficiently large n and any binary distribution Q_S we have

$$\begin{aligned}
D(Q_S||P_S) &\le \max\left\{\log\frac{1}{P_S(0)}, \log\frac{1}{P_S(1)}\right\} && (78)\\
&\le 2\max\left\{\log\frac{1}{\epsilon}, \log\frac{1}{1-\epsilon}\right\} && (79)\\
&\triangleq a_1 < \infty\,; && (80)
\end{aligned}$$

(76) is by noticing that $P_{Y^n|S=0} = P'_{Y^n}$, and (77) is by [4, Theorem 2]. It is clear that (71) and (77) show the impossibility of (56).

Similarly, one shows that (56) cannot hold if the assumption of the deterministic encoder is dropped. Indeed, then we can again take the very same (n, M'_n, ϵ'_n) code and make its encoder randomized so that with probability $\frac{\epsilon-\epsilon'_n}{1-\epsilon'_n}$ it outputs $(x_0, \dots, x_0) \in \mathcal{X}^n$ and otherwise it outputs the original codeword. The same analysis shows that (77) holds again and thus (56) fails. ∎

Note that the counter-examples constructed above also demonstrate that in Theorem 1 the assumptions of maximal probability of error and deterministic encoders are not superfluous.

VI. Application: AWGN

Recall that the $AWGN(P)$ channel is a sequence of random transformations $P_{Y^n|X^n} : \mathcal{X}^n \to \mathbb{R}^n$, where $\mathcal{X}^n$ is defined in (42) and

$$P_{Y^n|X^n=x} = \mathcal{N}(x, \mathbf{I}_n)\,. \tag{81}$$

Theorem 6: Consider a sequence of $(n, M_n, \epsilon)_{max,det}$ codes achieving the capacity of the $AWGN(P)$ channel. Then we have

$$\frac{1}{n}D(P_{Y^n}||\mathcal{N}(0, (1+P)\mathbf{I}_n) \to 0\,, \tag{82}$$

where P_{Y^n} is the output distribution induced by the code. The claim need not hold if the maximal probability of error is replaced with the average of if the encoder is allowed to be random.

Remark: As explained in Section II, $\mathcal{N}(0, (1+P)\mathbf{I}_n)$ is a quasi-caod sequence. Note also that Theorem 6 cannot hold if the power-constraint is understood in the average-over-the-codebook sense; see [6, Section 4.3.3].

Proof: Denote by lower-case $p_{Y^n|X^n=x}$ and p_{Y^n} densities of $P_{Y^n|X^n=x}$ and P_{Y^n}. Then an elementary computation shows

$$\nabla \log p_{Y^n}(y) = (y - \mathbb{E}\left[X^n|Y^n = y\right])\log e\,. \tag{83}$$

For convenience denote

$$\hat{X}^n = \mathbb{E}\left[X^n|Y^n\right] \tag{84}$$

and notice that since $\|X^n\| \le \sqrt{nP}$ we have also

$$\left\|\hat{X}^n\right\| \le \sqrt{nP}\,. \tag{85}$$

Then

$$\begin{aligned}
&\frac{1}{\log^2 e}\mathbb{E}\left[\|\nabla \log p_{Y^n}(Y^n)\|^2 \,|\, X^n\right] \\
&= \mathbb{E}\left[\left\|Y^n - \hat{X}^n\right\|^2 \,\middle|\, X^n\right] && (86)\\
&\le 2\mathbb{E}\left[\|Y^n\|^2 \,\middle|\, X^n\right] + 2\mathbb{E}\left[\left\|\hat{X}^n\right\|^2 \,\middle|\, X^n\right] && (87)\\
&\le 2\mathbb{E}\left[\|Y^n\|^2 \,\middle|\, X^n\right] + 2nP && (88)\\
&= 2\mathbb{E}\left[\|X^n + Z^n\|^2 \,\middle|\, X^n\right] + 2nP && (89)\\
&\le 4\|X^n\|^2 + 4n + 2nP && (90)\\
&\le (6P+4)n\,, && (91)
\end{aligned}$$

where (87) is by a simple Cauchy-Schwartz estimate for any $a, b \in \mathbb{R}^n$

$$\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2\,, \tag{92}$$

(88) is by (85), in (89) we introduced $Z^n \sim \mathcal{N}(0, \mathbf{I}_n)$ which is independent of X^n, (90) is by (92) and (91) is by the power-constraint for X^n.

According to (81), conditioned on X^n random vector Y^n is Gaussian. Thus, from Poincaré inequality for the Gaussian measure, e.g. [11, (2.16)], we have

$$\mathrm{Var}[\log p_{Y^n}(Y^n) \,|\, X^n] \le \mathbb{E}\left[\|\nabla \log p_{Y^n}\|^2 \,|\, X^n\right] \tag{93}$$

and together with (91) this yields the required estimate

$$\mathrm{Var}[\log p_{Y^n}(Y^n) \,|\, X^n] \le a_1 n \tag{94}$$

for some $a_1 > 0$. The argument then proceeds step by step as in the proof of Theorem 5 with (94) taking the place of (62) and invoking the following (quasi-caod) property of $P^*_{Y^n}$ for (67):

$$\max_{x:||x||\le\sqrt{nP}} D(P_{Y^n|X^n=x}||P^*_{Y^n}) = nC\,, \tag{95}$$

where $C = \frac{1}{2}\log(1+P)$ and $P^*_{Y^n} = \mathcal{N}(0, (1+P)\mathbf{I}_n)$.

Counter-examples are constructed similarly to those in Theorem 5 with $x_0 = 0$. ∎

Remark: Proofs of Theorems 5 and 6 can be shown to imply that the entropy density $\log\frac{1}{P_{Y^n}(Y^n)}$ concentrates up to $\sqrt{n}$ around the entropy $H(Y^n)$. Such questions are also interesting in other contexts and for other types of distributions, see [12].

VII. Convergence in relative entropy

We have shown, (56) and (82), that the distributions P_{Y^n} induced by capacity-achieving codes become close to the caod, $P^*_{Y^n}$ in the sense of (1). In this section we discuss some implications of such a convergence.

First, by convexity from (1) we have

$$D(\bar{P}_n||P^*_Y) \le \frac{1}{n}D(P_{Y^n}||P^*_{Y^n}) \to 0\,, \tag{96}$$

where $\bar{P}_n$ is the empirical output distribution

$$\bar{P}_n \triangleq \frac{1}{n}\sum_{j=1}^n P_{Y_i}\,. \tag{97}$$

More generally, we have [4, (41)]

$$D(\bar{P}^{(k)}_n||P^*_{Y^k}) \le \frac{k}{n-k+1}D(P_{Y^n}||P^*_{Y^n}) \to 0\,, \tag{98}$$

where $\bar{P}^{(k)}_n$ is a k-th order empirical output distribution

$$\bar{P}^{(k)}_n = \frac{1}{n-k+1}\sum_{j=1}^{n-k+1} P_{Y_j^{j+k-1}}\,. \tag{99}$$

Knowing that a sequence of distributions P_n converges in relative entropy to a distribution P, i.e.

$$D(P_n||P) \to 0 \tag{100}$$

implies convergence properties for the expectations of functions:

1) By the Csiszar-Pinsker inequality

$$||P_n - P||_{TV} \to 0\,, \tag{101}$$

or, equivalently, for all bounded functions f we have

$$\int f dP_n \to \int f dP\,. \tag{102}$$

2) In fact, (102) holds for a wider class of functions, namely those that satisfy Cramer condition under P, i.e.

$$\int e^{tf} dP < \infty \tag{103}$$

for all t in some neighborhood of 0; see [13, Lemma 3.1].

Together (102) and (96) show that for a wide class of functions $f : \mathcal{Y} \to \mathbb{R}$ empirical averages over distributions induced by good codes converge to the average over the caod:

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n f(Y_j)\right] \to \int f dP^*_Y\,. \tag{104}$$

From (98) a similar conclusion holds for k-th order empirical averages.

For notational convenience we introduce a random variable Y^{*n} which has distribution $P^*_{Y^n}$ so that

$$\mathbb{E}\left[F(Y^{*n})\right] = \int_{\mathcal{Y}^n} F(y^n) dP^n_{Y^*}\,. \tag{105}$$

Regarding general functions of Y^n we have the following:

Lemma 7: Suppose that $F : \mathcal{Y}^n \to \mathbb{R}$ is such that for some $c > 0$ we have

$$\log \mathbb{E}\left[\exp\{tF(Y^{*n})\}\right] \le t\mathbb{E}\left[F(Y^{*n})\right] + ct^2 \tag{106}$$

for all $t \in \mathbb{R}$ with $Y^{*n} \sim P^*_{Y^n}$. Then

$$|\mathbb{E}\left[F(Y^n)\right] - \mathbb{E}\left[F(Y^{*n})\right]| \le 2\sqrt{cD(P_{Y^n}||P^*_{Y^n})}\,. \tag{107}$$

Proof: The key tool for obtaining estimates on expectations of functions from the estimates of relative entropy is the Donsker-Varadhan inequality [14, Lemma 2.1]: For any probability measures P and Q with $D(P||Q) < \infty$ and a

measurable function g such that $\int \exp\{g\}dQ < \infty$ we have that $\int gdP$ exists (but perhaps is $-\infty$) and moreover

$$\int gdP - \log \int \exp\{g\}dQ \le D(P||Q)\,. \tag{108}$$

Since by (106) the moment generating function of F under $P^*_{Y^n}$ exists, from (108) applied to tF we get

$$t\mathbb{E}\left[F(Y^n)\right] - \log \mathbb{E}\left[\exp\{tF(Y^{*n})\}\right] \le D(P_{Y^n}||P^*_{Y^n})\,. \tag{109}$$

From (106) we have then

$$ct^2 - t\mathbb{E}\left[F(Y^n)\right] + t\mathbb{E}\left[F(Y^{*n})\right] + D(P_{Y^n}||P^*_{Y^n}) \ge 0 \tag{110}$$

for all t. Thus discriminant of this quadratic polynomial (in t) must be non-positive which is precisely (107). ∎

Estimates of the form (106) are known as the Gaussian concentration of measure and are available for various classes of functions F and measures $P^*_{Y^n}$; see [11] for a survey[4]. As an example, we have

Corollary 8: For any $0 < \epsilon < 1$ there exist two constants $a_1, a_2 > 0$ such that for any $(n, M, \epsilon)_{max,det}$ code for the $AWGN(P)$ channel and for any function $F : \mathbb{R}^n \to \mathbb{R}$ with Lipschitz constant not exceeding 1 we have

$$\left|\mathbb{E}\left[F(Y^n)\right] - \mathbb{E}\left[F(Y^{*n})\right]\right| \le a_1\sqrt{nC - \log M_n + a_2\sqrt{n}}\,, \tag{111}$$

where we remind that $Y^{*n} \sim \mathcal{N}(0, (1+P)\mathbf{I}_n)$ and $C = \frac{1}{2}\log(1+P)$ is the capacity.

Proof: In the proof of Theorem 6 we obtained an upper bound

$$D(P_{Y^n}||P^*_{Y^n}) \le nC - \log M_n + a_2\sqrt{n}\,. \tag{112}$$

Then, since $P^*_{Y^n} = \mathcal{N}(0, (1+P)\mathbf{I}_n)$ is Gaussian, any 1-Lipschitz function satisfies (107); see [15, Proposition 2.1], for example. Then Lemma 7 completes the proof. ∎

Note that in the proof of the corollary concentration of measure was used twice: once for $P_{Y^n|X^n}$ in the form of Poincaré inequality (proof of Theorem 6) and once in the form of (106) (proof of Lemma 7).

As a closing remark, we notice that convergence of output distributions can often be propagated to statements about the input distributions. For example, this is obvious for the case of the AWGN, since convolution with a Gaussian kernel is an injective map of measures (e.g., by a simple Fourier argument), and a DMC with a non-singular (more generally, injective) matrix $P_{Y|X}$. For other DMCs, the following argument complements that of [4, Theorem 4]. By Theorem 4 and 5 we know that

$$\frac{1}{n}I(X^n; Y^n) \to C\,.$$

By concavity of mutual information, we must necessarily have

$$I(\bar{X}; \bar{Y}) \to C\,,$$

where $P_{\bar{X}} = \frac{1}{n}\sum_{j=1}^n P_{X_j}$. By compactness of the simplex of input distributions and continuity of the mutual information on that simplex the distance to the (compact) set of capacity achieving distributions Π must vanish:

$$d(P_{\bar{X}}, \Pi) \to 0\,.$$

VIII. Extension to other channels

As discussed above, statements of the form (1) are quite strong and imply all sorts of weaker results, such as convergence of empirical distributions and estimates for the expectations of functions. In this section we demonstrate a technique showing how to prove such corollary results directly from Theorem 1.

To illustrate the technique we start with a weaker (Fano-like) estimate. Fix a random transformation $P_{Y|X}$ with the caod P^*_Y and a function $F : \mathcal{Y} \to \mathbb{R}$ such that

$$Z_F = \log \mathbb{E}\left[\exp\{F(Y^*)\}\right] < \infty\,, \tag{113}$$

where as before $Y^* \sim P^*_Y$. Denote by $Q^{(F)}$ an F-tilting of P^*_Y:

$$Q^{(F)} = P^*_Y \exp\{F - Z_F\}\,. \tag{114}$$

Consider an $(M, \epsilon)_{avg}$ code for $P_{Y|X}$. Following the meta-converse principle [5, Section III.E], we consider a pair of measures on the probability space (3): one induced by the code and another induced by replacing the kernel $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ with $Q^F : \mathcal{X} \to \mathcal{Y}$ (the latter is oblivious to the input). Then applying data-processing for relative entropy to the random variable $1\{W \ne \hat{W}\}$ we obtain

$$d(1-\epsilon||\tfrac{1}{M}) \le D(P_{Y|X}||Q^{(F)}|P_X)\,, \tag{115}$$

where $d(x||y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ is a binary relative entropy and P_X is the input distribution induced by the code. Expanding both sides we get

$$\begin{aligned}
&(1-\epsilon)\log M + h(\epsilon) \\
&\le d(1-\epsilon||\tfrac{1}{M}) && (116)\\
&\le D(P_{Y|X}||Q^{(F)}|P_X) && (117)\\
&= D(P_{Y|X}||P^*_Y|P_X) - \mathbb{E}\left[F(Y)\right] + \log\mathbb{E}\left[\exp\{F(Y^*)\}\right] && (118)\\
&\le \mathcal{C} - \mathbb{E}\left[F(Y)\right] + \log\mathbb{E}\left[\exp\{F(Y^*)\}\right], && (119)
\end{aligned}$$

where (119) follows by (8). If $P_{Y|X}$ corresponds to a block-length n random transformation of a memoryless channel we have $\mathcal{C} = nC$. As a result, we directly obtain both the Donsker-Varadhan inequality and the estimate for $D(P_Y||P^*_Y)$:

$$\mathbb{E}\left[F(Y)\right] - \log\mathbb{E}\left[\exp\{F(Y^*)\}\right] \le nC - (1-\epsilon)\log M - h(\epsilon)\,. \tag{120}$$

Since F was arbitrary, as in Lemma 7 one concludes that $\mathbb{E}\left[F(Y)\right] \approx \mathbb{E}\left[F(Y^*)\right]$ provided that the right-hand side of (120) is small. Unfortunately, even for the code with $\log M \approx nC$ this is not the case unless $\epsilon \to 0$.

We can fix this problem by invoking Theorem 1 at the expense of restricting to $(M, \epsilon)_{max,det}$ codes and reducing the class of functions for which (120) is valid. As an example of

[4]E.g., consider $F(y^n) = \frac{1}{n}\sum_{j=1}^n f(y_i)$ and $P^*_{Y^n}$ – a product distribution; then (106) follows from a similar single-letter estimate for f, which is typically trivial (e.g., if f is bounded). The resulting estimate in this case can also be obtained by directly applying Lemma 7 to (96).

such an argument we provide an alternative prove of Corollary 8, which also illuminates relation to the concentration of measure.

Alternative proof of Corollary 8: Since F is 1-Lipschitz by Poincaré inequality for the Gaussian measure we have

$$\mathrm{Var}[F(Y^n)|X^n] \le 1 \tag{121}$$

and thus from the definition of $Q^{(F)}$ in (114) we have

$$\begin{aligned}
&\mathrm{Var}\left[\log \frac{dP_{Y^n|X^n}}{dQ^{(F)}} \,\middle|\, X^n\right] \\
&\le 2\,\mathrm{Var}\left[\frac{P_{Y^n|X^n}(Y^n|X^n)}{\prod_{j=1}^n P_Y^*(Y_j)} \,\middle|\, X^n\right] + \mathrm{Var}[F(Y^n)|X^n] \qquad (122)\\
&= O(n)\,. \qquad (123)
\end{aligned}$$

Then we have

$$\begin{aligned}
&\log M_n \\
&\le D(P_{Y^n|X^n}||Q^{(F)}|P_{X^n}) + O(\sqrt{n}) \qquad (124)\\
&= D(P_{Y^n|X^n}||P^*_{Y^n}|P_{X^n}) - \mathbb{E}\left[F(Y^n)\right] \\
&\quad + \log \mathbb{E}\left[\exp\{F(Y^{*n})\}\right] + O(\sqrt{n}) \qquad (125)\\
&\le nC - \mathbb{E}\left[F(Y^n)\right] + \log \mathbb{E}\left[\exp\{F(Y^{*n})\}\right] + O(\sqrt{n})\,, \qquad (126)
\end{aligned}$$

where (124) is by Corollary 2 with S_m estimated from (123), (125) is by the definition of $Q^{(F)}$ in (114) with $Y^{*n} \sim P^*_{Y^n}$; and (126) is by (95). From (126) the proof proceeds as in Lemma 7. ∎

The upshot of this section is that even if (1) does not hold (or is not known to hold), one frequently can derive explicit non-asymptotic bounds on the expectations of functions, such as (111), provided that the function satisfies concentration of measure under both $P_{Y^n|X^n}$ and the caod, $P^*_{Y^n}$. In view of the progress in log-Sobolev inequalities and optimal transportation (which are the main tools used to prove the concentration of measure) the approach of this section looks especially promising.

REFERENCES

[1] S. Nitinawarat, "On maximal error capacity regions of symmetric gaussian multiple-access channels," in *Proc. 2011 IEEE Int. Symp. Inf. Theory (ISIT)*, St. Petersburg, Russia, Aug. 2011.

[2] A. Tchamkerten, V. Chandar, and G. W. Wornell, "Communication under strong asynchronism," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4508–4528, Oct. 2009.

[3] Y. Polyanskiy, "Asynchronous communication: capacity, strong converse and dispersion," *IEEE Trans. Inf. Theory*, 2011, submitted. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/async.pdf

[4] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 836–846, 1997.

[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] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 2010, available: http://people.lids.mit.edu/yp/homepage/.

[7] Y. Polyanskiy and S. Verdú, "Scalar coherent fading channel: dispersion analysis," in *Proc. 2011 IEEE Int. Symp. Inf. Theory (ISIT)*, St. Petersburg, Russia, Aug. 2011.

[8] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois J. Math.*, vol. 1, pp. 591–606, 1957.

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

[10] S. Bobkov and F. Götze, "Discrete isoperimetric and Poincaré-type inequalities," *Probab. Theory Relat. Fields*, vol. 114, pp. 245–277, 1999.

[11] M. Ledoux, "Concentration of measure and logarithmic Sobolev inequalities," *Seminaire de probabilites XXXIII*, pp. 120–216, 1999.

[12] S. Bobkov and M. Madiman, "Concentration of the information in data with log-concave distributions," *Ann. Probab.*, vol. 39, no. 4, pp. 1528–1543, 2011.

[13] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, Feb. 1975.

[14] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. i. ii." *Comm. Pure Appl. Math.*, vol. 28, no. 1, pp. 1–47, 1975.

[15] M. Ledoux, "Isoperimetry and Gaussian analysis," *Lecture Notes in Math.*, vol. 1648, pp. 165–294, 1996.