Empirical Distribution of Good Channel Codes With Nonvanishing Error Probability

Yury Polyanskiy, *Member, IEEE*, and Sergio Verdú, *Fellow, IEEE*

***Abstract*—This paper studies several properties of channel codes that approach the fundamental limits of a given (discrete or Gaussian) memoryless channel with a nonvanishing probability of error. The output distribution induced by an ϵ-capacity-achieving code is shown to be close in a strong sense to the capacity achieving output distribution. Relying on the concentration of measure (isoperimetry) property enjoyed by the latter, it is shown that regular (Lipschitz) functions of channel outputs can be precisely estimated and turn out to be essentially nonrandom and independent of the actual code. It is also shown that the output distribution of a good code and the capacity achieving one cannot be distinguished with exponential reliability. The random process produced at the output of the channel is shown to satisfy the asymptotic equipartition property.**

***Index Terms*—Additive white Gaussian noise, asymptotic equipartition property, concentration of measure, discrete memoryless channels, empirical output statistics, relative entropy, Shannon theory.**

I. Introduction

A RELIABLE channel codebook (or code, for the purposes of this paper) is a collection of codewords of fixed duration distinguishable with small probability of error when observed through a noisy channel. Such a code is optimal if it possesses the maximal cardinality among all codebooks of equal duration and probability of error. In this paper, we characterize several properties of optimal and close-to-optimal channel codes indirectly, i.e., without identifying the best code explicitly. This characterization provides theoretical insight and ultimately may facilitate the search for new good code families by providing a necessary condition they must satisfy.

Shannon [1] was the first to recognize, in the context of the additive white Gaussian noise channel, that to maximize information transfer across a memoryless channel codewords must be "noise-like," i.e., resemble a typical sample of a memoryless random process with marginal distribution that maximizes mutual information. Specifically, in [1, Sec. 25] Shannon states[1]:

> To approximate this limiting rate of transmission the transmitted signals must approximate, in statistical properties, a white noise.

A general and formal statement of this property of optimal codes was put forward by Han and Verdú [2, Th. 15]:

Theorem 1: Fix an arbitrary $\gamma > 0$. For any channel with finite input alphabet and capacity C that satisfies the strong converse, and sufficiently large n,

$$\frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) \le \gamma \tag{1}$$

where $P^*_{Y^n}$ is the maximal mutual information output distribution and P_{Y^n} is the output distribution induced by the codebook (assuming equiprobable codewords) of any (n, M_n, λ_n) code such that $\lambda_n \to 0$ as $n \to \infty$, and

$$\frac{1}{n} \log M_n \ge C - \frac{\gamma}{2}. \tag{2}$$

Therefore, for a finite-input memoryless channel, any capacity-achieving sequence of codes with *vanishing* probability of error must satisfy

$$\frac{1}{n} D(P_{Y^n} \| P^*_Y \times \cdots \times P^*_Y) \to 0, \tag{3}$$

where P^*_Y is the single-letter capacity achieving output distribution. Furthermore, Shamai and Verdú [3] show that (under regularity conditions) the empirical frequency of input letters (or sequential k-letter blocks) inside the codebook approaches the capacity achieving input distribution (or its kth power) in the sense of vanishing relative entropy.

In this paper, we focus attention on memoryless channels and we extend the result in Theorem 1 to the case of *nonvanishing* probability of error. Studying this regime as opposed to *vanishing* probability of error has recently proved to be fruitful for the nonasymptotic characterization of the maximal achievable rate [4]. Although for the memoryless channels considered in this paper the ϵ-capacity C_ϵ is independent of the probability of error ϵ, it does not immediately follow that a C_ϵ-achieving code necessarily satisfies the empirical distribution property (3). In fact, we will show that (3) fails to be necessary under the average probability of error criterion.

To illustrate the delicacy of the question of approximating P_{Y^n} with $P^*_{Y^n}$, consider a good, capacity-achieving k-to-n code for the binary symmetric channel (BSC) with crossover probability $\delta < \frac{1}{2}$ and capacity C. The probability of the codebook under P_{Y^n} is larger than the probability that no errors occur: $(1 - \delta)^n$. Under $P^*_{Y^n}$ the probability of the codebook is

Manuscript received December 21, 2012; revised August 30, 2013; accepted September 19, 2013. Date of publication October 04, 2013; date of current version December 20, 2013. This work was supported in part by the National Science Foundation (NSF) under Grants CCF-1016625, CCF-1253205, and in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370. This paper was presented at the 49th and 50th Allerton Conferences on Communication, Control, and Computing, 2011–2012.

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).

Communicated by A. B. Wagner, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2013.2284506

[1] In [1], "white noise" means white Gaussian noise.

2^{k-n}—which is exponentially smaller asymptotically since for a reliable code $k \leq n - nh(\delta) < n \log 2(1-\delta)$. On the other hand, consider a set E consisting of a union of small Hamming balls surrounding each codeword, whose radius $\approx \delta n$ is chosen such that $P_{Y^n}[E] = \frac{1}{2}$, say. Assuming that the code is decodable with small probability of error, the union will be almost disjoint and hence $P^*_{Y^n}[E] \approx 2^{k-nC}$—the two becoming exponentially comparable (provided $k \approx nC$). Thus, for certain events, P_{Y^n} and $P^*_{Y^n}$ differ exponentially, while on other, for less delicate events, they behave similarly. We will show that as long as the error probability is strictly less than one, the normalized relative entropy in (3) is upper bounded by the difference between capacity and code rate.

Studying the output distribution P_{Y^n} also becomes important in the context of secure communication, where the output due to the code is required to resemble white noise; and in the problem of asynchronous communication where the output statistics of the code imposes limits on the quality of synchronization [5]. For example, in a multiterminal communication problem, the channel output of one user may create interference for another. Assessing the average impairment caused by such interference involves the analysis of the expectation of a certain function of the channel output $\mathbb{E}[F(Y^n)]$. We show that under certain regularity assumptions on F not only one can approximate the expectation of F by substituting the unknown P_{Y^n} with $P^*_{Y^n}$, as in

$$\int F(y^n) dP_{Y^n} \approx \int F(y^n) dP^*_{Y^n}, \tag{4}$$

but one can also prove that in fact the distribution of $F(Y^n)$ will be tightly concentrated around its expectation. Thus, we are able to predict with overwhelming probability the random value of $F(Y^n)$ without any knowledge of the code used to produce Y^n (but assuming the code is ϵ-capacity-achieving).

Besides (3) and (4), we will show

1) an upper bound on relative entropy $D(P_{Y^n} \| P^*_{Y^n})$ in terms of the cardinality of the employed code;
2) the hypothesis testing problem between P_{Y^n} and $P^*_{Y^n}$ has zero Stein exponent;
3) the output process Y^n enjoys an asymptotic equipartition property.

Throughout the paper, we will observe a number of connections with the concentration of measure (isoperimetry) and optimal transportation, which were introduced into the information theory by the seminal works [6]–[8]. Although some key results are stated for general channels, most of the discussion is specialized to discrete memoryless channels (DMC) (possibly with a (separable) input cost constraint) and to the AWGN channel.

The organization of the paper is as follows. Section II contains the main definitions and notation. Section III proves a sharp upper bound on the relative entropy $D(P_{Y^n} \| P^*_{Y^n})$. In Section IV, we discuss various implications of the bounds on relative entropy and in particular prove approximation (4). Section V considers the hypothesis testing problem of discriminating between P_{Y^n} and $P^*_{Y^n}$. The asymptotic equipartition property of the channel output process is established in Section VI.

II. Definitions and Notation

A. Codes and Channels

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 $\mathsf{f}: \{1, \ldots, M\} \to \mathcal{X}$ and $\mathsf{g}: \mathcal{Y} \to \{1, \ldots, M\}$ such that

$$\mathbb{P}[\hat{W} \neq W] \leq \epsilon, \tag{5}$$

where in the underlying probability space $X = \mathsf{f}(W)$ and $\hat{W} = \mathsf{g}(Y)$ with W equiprobable on $\{1, \ldots, M\}$, and $W, X, Y, \hat{W}$ forming a Markov chain:

$$W \xrightarrow{\mathsf{f}} X \xrightarrow{P_{Y|X}} Y \xrightarrow{\mathsf{g}} \hat{W}. \tag{6}$$

In particular, we say that P_X (resp., P_Y) is the input (resp., output) distribution induced by the encoder f. An $(M, \epsilon)_{max}$ code is defined similarly except that (5) is replaced with the more stringent maximal probability of error criterion:

$$\max_{1 \leq j \leq M} \mathbb{P}[\hat{W} \neq W | W = j] \leq \epsilon. \tag{7}$$

A code is deterministic if the encoder f is a functional (nonrandom) mapping. We will frequently specify that a code is deterministic with the notation $(M, \epsilon)_{max,det}$ or $(M, \epsilon)_{avg,det}$. 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. An (M, ϵ) code for the nth random transformation is called an (n, M, ϵ) code, and the foregoing notation specifying average/maximal error probability and deterministic encoder will also be applied to that case. The nonasymptotic fundamental limit of communication is defined as[2]

$$M^*(n, \epsilon) = \max\{M: \exists (n, M, \epsilon)\text{-code}\}. \tag{8}$$

B. Capacity-Achieving Output Distribution

To the three types of channels considered below we also associate a special sequence of output distributions $P^*_{Y^n}$, defined as the nth power of a certain single-letter distribution P^*_Y[3]:

$$P^*_{Y^n} \triangleq (P^*_Y)^n = P^*_Y \times \cdots \times P^*_Y, \tag{9}$$

where P^*_Y is a distribution on the output alphabet defined as follows.

1) A DMC (without feedback) is built from a single letter transformation $P_{Y|X}: \mathcal{X} \to \mathcal{Y}$ acting between finite spaces by extending the latter to all $n \geq 1$ in a memoryless way. Namely, the input space of the nth random transformation $P_{Y^n|X^n}$ is given by[4]

$$\mathcal{X}_n = \mathcal{X}^n \triangleq \mathcal{X} \times \cdots \times \mathcal{X} \tag{10}$$

[2] Additionally, one should also specify which probability of error criterion, (5) or (7), is used.

[3] For general channels, the sequence $\{P^*_{Y^n}\}$ is required to satisfy a *quasi-caod* property, see [9, Sec. IV].

[4] To unify notation, we denote the input space as $\mathcal{X}_n$ (instead of the more natural $\mathcal{X}^n$) even in the absence of cost constraints.

and similarly for the output space $\mathcal{Y}^n = \mathcal{Y} \times \cdots \times \mathcal{Y}$, while the transition kernel is set to be

$$P_{Y^n|X^n}(y^n|x^n) = \prod_{j=1}^{n} P_{Y|X}(y_j|x_j)\,. \tag{11}$$

The capacity C and P_Y^*, the unique capacity-achieving output distribution (*caod*), are found by solving

$$C = \max_{P_X} I(X;Y)\,. \tag{12}$$

2) A DMC with input constraint (c, P) is a generalization of the previous construction with an additional restriction on the input space $\mathcal{X}_n$:

$$\mathcal{X}_n = \left\{ x^n \in \mathcal{X}^n : \sum_{j=1}^{n} \mathsf{c}(x_j) \le nP \right\}. \tag{13}$$

In this case, the capacity C and the caod P_Y^* are found by restricting the maximization in (12) to those P_X that satisfy

$$\mathbb{E}[\mathsf{c}(X)] \le P\,. \tag{14}$$

3) The AWGN(P) channel has input space[5]

$$\mathcal{X}_n = \left\{ \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \le \sqrt{nP} \right\} \tag{15}$$

output space $\mathcal{Y}^n = \mathbb{R}^n$ and transition kernel

$$P_{Y^n|X^n=\mathbf{x}} = \mathcal{N}(\mathbf{x}, \mathbf{I}_n), \tag{16}$$

where $\mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$ denotes a (multidimensional) normal distribution with mean $\mathbf{x}$ and covariance matrix $\mathbf{\Sigma}$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Then,[6]

$$C = \frac{1}{2}\log(1+P) \tag{17}$$

$$P_Y^* = \mathcal{N}(0, 1+P)\,. \tag{18}$$

As shown in [10] and [11] in all three cases P_Y^* is unique and $P_{Y^n}^*$ satisfies the key property

$$D(P_{Y^n|X^n=x} \| P_{Y^n}^*) \le nC, \tag{19}$$

for all $x \in \mathcal{X}_n$. Since $I(U;V) = D(P_{V|U}\|Q|P_U) - D(P_V\|Q)$, Property (19) implies that for every input distribution P_{X^n} the induced output distribution P_{Y^n} satisfies

$$D(P_{Y^n} \| P_{Y^n}^*) \le nC - I(X^n;Y^n)\,. \tag{20}$$

$$P_{Y^n} \ll P_{Y^n}^* \tag{21}$$

$$P_{Y^n|X^n=x^n} \ll P_{Y^n}^* \qquad \forall x^n \in \mathcal{X}_n\,. \tag{22}$$

As a consequence of (22) the information density is well defined

$$\imath_{X^n;Y^n}^*(x^n;y^n) \triangleq \log \frac{dP_{Y^n|X^n=x^n}}{dP_{Y^n}^*}(y^n)\,. \tag{23}$$

Moreover, for every channel considered here there is a constant $a_1 > 0$ such that[7]

$$\sup_{x^n \in \mathcal{X}_n} \operatorname{Var}\left[\imath_{X^n;Y^n}^*(X^n;Y^n) \,\middle|\, X^n = x^n \right] \le n a_1\,. \tag{24}$$

In all three cases, the ϵ-capacity C_ϵ equals C for all $0 < \epsilon < 1$, i.e.,

$$\log M^*(n,\epsilon) = nC + o(n), \qquad n \to \infty\,. \tag{25}$$

In fact, see [4]

$$\log M^*(n,\epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n), \qquad n \to \infty, \tag{26}$$

for any $0 < \epsilon < \frac{1}{2}$, a certain constant $V \ge 0$, called the channel dispersion, and Q^{-1} is the inverse of the standard complementary normal cdf.

C. Good Codes

We introduce the following increasing degrees of optimality for sequences of (n, M_n, ϵ) codes. A code sequence is called

1) $o(n)$ *-achieving* or ϵ-capacity-achieving if

$$\frac{1}{n}\log M_n \to C\,. \tag{27}$$

2) $O(\sqrt{n})$ *-achieving* if

$$\log M_n = nC + O(\sqrt{n})\,. \tag{28}$$

3) $o(\sqrt{n})$ *-achieving* or dispersion-achieving if

$$\log M_n = nC - \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n})\,. \tag{29}$$

4) $O(\log n)$ *-achieving* if

$$\log M_n = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)\,. \tag{30}$$

D. Binary Hypothesis Testing

We also need to introduce the performance of an optimal binary hypothesis test, which is one of the main tools in [4]. Consider an $\mathcal{A}$-valued random variable B which can take probability measures P or Q. A randomized test between those two distributions is defined by a random transformation $P_{Z|B}: \mathcal{A} \mapsto \{0,1\}$ where 0 indicates that the test chooses Q. The best performance achievable among those randomized tests is given by[8]

$$\beta_\alpha(P,Q) = \min \sum_{a \in \mathcal{A}} Q(a) P_{Z|B}(1|a), \tag{31}$$

[5] For convenience, we denote the elements of $\mathbb{R}^n$ as $\mathbf{x}, \mathbf{y}$ (for nonrandom vectors) and X^n, Y^n (for the random vectors).

[6] As usual, all logarithms $\log$ and exponents $\exp$ are taken to arbitrary fixed base, which also specifies the information units.

[7] For discrete channels (24) is shown, e.g., in [4, Appendix E].

[8] We sometimes write summations over alphabets for simplicity of exposition. For arbitrary measurable spaces $\beta_\alpha(P,Q)$ is defined by replacing the summation in (31) by an expectation.

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

$$P_{Z|B}\colon \sum_{a\in\mathcal{A}} P(a)P_{Z|B}(1|a) \ge \alpha\,. \tag{32}$$

The minimum in (31) 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 no larger than $1-\alpha$.

III. Upper Bound on the Output Relative Entropy

The main goal of this section is to establish (for each of the three types of memoryless channels introduced in Section II) that

$$D(P_{Y^n}\|P^*_{Y^n}) \le nC - \log M_n + o(n), \tag{33}$$

where P_{Y^n} is the sequence of output distributions induced by a sequence of (n, M_n, ϵ) codes, and $o(n)$ depends on ϵ. Furthermore, for all channels except DMCs with zeros in the transition matrix $P_{Y|X}$, $o(n)$ in (33) can be replaced by $O(\sqrt{n})$.

We start by giving a one-shot converse due to Augustin [12] in Section III-A. Then, we prove (33) for DMCs in Section III-B and for the AWGN in Section III-C.

A. Augustin's Converse

The following result first appeared as part of the proofs in [12, Satz 7.3 and 8.2] by Augustin and formally stated in [13, Sec. 2]. Note that particularizing Theorem 2 to a constant function ρ recovers the nonasymptotic converse bound that can be derived from Wolfowitz's proof of the strong converse [14].

Theorem 2 ([12], [13]): Consider a random transformation $P_{Y|X}$, a distribution P_X induced by an $(M,\epsilon)_{max,det}$ code, a distribution Q_Y on the output alphabet and a function $\rho\colon \mathcal{X}\to\mathbb{R}$. Then, provided the denominator is positive,

$$M \le \frac{\exp\{\mathbb{E}[\rho(X)]\}}{\inf_x P_{Y|X=x}\left[\log\frac{dP_{Y|X=x}}{dQ_Y}(Y) \le \rho(x)\right] - \epsilon}, \tag{34}$$

with the infimum taken over the support of P_X.

Proof: Fix a $(M,\epsilon)_{max,det}$ code Q_Y and the function ρ. Denoting by c_i the ith codeword, we have

$$Q_Y[\hat{W}(Y)=i] \ge \beta_{1-\epsilon}(P_{Y|X=c_i}, Q_Y), \qquad i=1,\dots,M, \tag{35}$$

since $\hat{W}(Y)=i$ is a suboptimal test to decide between $P_{Y|X=c_i}$ and Q_Y, which achieves error probability no larger than ϵ when $P_{Y|X=c_i}$ is true. Denoting the event

$$B_x(y) = \log\frac{dP_{Y|X=x}}{dQ_Y}(y) \le \rho(x) \tag{36}$$

we can bound

$$\frac{1}{M} \ge \frac{1}{M}\sum_{i=1}^{M} \beta_{1-\epsilon}(P_{Y|X=c_i}, Q_Y) \tag{37}$$

$$\ge \frac{1}{M}\sum_{i=1}^{M}\left(P_{Y|X=c_i}\left[B_{c_i}(Y)\right]-\epsilon\right)\exp\{-\rho(c_i)\} \tag{38}$$

$$\ge \left(\inf_x P_{Y|X=x}\left[B_x(Y)\right]-\epsilon\right)\frac{1}{M}\sum_{i=1}^{M}\exp\{-\rho(c_i)\} \tag{39}$$

$$\ge \left(\inf_x P_{Y|X=x}\left[B_x(Y)\right]-\epsilon\right)\exp\{-\mathbb{E}[\rho(X)]\}, \tag{40}$$

where (37) is by taking the arithmetic average of (35) over i, (40) is by Jensen's inequality, and (38) is by the standard estimate of β_α, e.g., [4, (102)],

$$\beta_{1-\epsilon}(P,Q) \ge \left(\mathbb{P}\left[\log\frac{dP}{dQ}(Z)\le\rho\right]-\epsilon\right)\exp\{-\rho\}, \tag{41}$$

with Z distributed according to P. ■

Remark 1: Following an idea of Poor and Verdú [15] we may further strengthen Theorem 2 in the special case of $Q_Y = P_Y$. The maximal probability of error ϵ for any test of M hypotheses $\{P_j, j=1,\dots,M\}$ satisfies

$$\epsilon \ge \left(1-\frac{\exp\{\bar{\rho}\}}{M}\right)\inf_{1\le j\le M}\mathbb{P}\left[\imath_{W;Y}(W;Y)\le\rho_j|W=j\right], \tag{42}$$

where the information density is as defined in (23), $\rho_j\in\mathbb{R}$ are arbitrary, $\bar{\rho}=\frac{1}{M}\sum_{j=1}^{M}\rho_j$, W is equiprobable on $\{1,\dots,M\}$ and $P_{Y|W=j}=P_j$. Indeed, since $\imath_{W;Y}(a;b)\le\log M$ we get from [16, Lemma 35]

$$\begin{aligned}&\exp\{\rho_j\}\beta_{1-\epsilon}(P_{Y|W=j},P_Y)\\&+\left(1-\frac{\exp\{\rho_j\}}{M}\right)\mathbb{P}\left[\imath_{W;Y}(W;Y)>\rho_j|W=j\right]\\&\ge 1-\epsilon\,.\end{aligned} \tag{43}$$

Multiplying by $\exp\{-\rho_j\}$ and using resulting bound in place of (41) we repeat steps (37)–(40) to obtain

$$\begin{aligned}&\frac{1}{M}\inf_{1\le j\le M}\mathbb{P}\left[\imath_{W;Y}(W;Y)\le\rho_j|W=j\right]\\&\ge\left(\inf_{1\le j\le M}\mathbb{P}\left[\imath_{W;Y}(W;Y)\le\rho_j|W=j\right]-\epsilon\right)\exp\{-\bar{\rho}\},\end{aligned} \tag{44}$$

which in turn is equivalent to (42).

Choosing $\rho(x)=D(P_{Y|X=x}\|Q_Y)+\Delta$ we can specialize Theorem 2 in the following convenient form.

Theorem 3: 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

$$d(x) \triangleq D(P_{Y|X=x}\|Q_Y) < \infty \tag{45}$$

and

$$\sup_x P_{Y|X=x}\left[\log\frac{dP_{Y|X=x}}{dQ_Y}(Y)\ge d(x)+\Delta\right]\le\delta', \tag{46}$$

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

$$D(P_{Y|X}\|Q_Y|P_X) \ge \log M - \Delta + \log(1-\epsilon-\delta')\,. \tag{47}$$

Remark 2: Note that (46) holding with a small δ' is a natural nonasymptotic embodiment of information stability of the underlying channel, cf. [9, Sec. IV].

A simple way to estimate the upper deviations in (46) is by using Chebyshev's inequality. As an example, we obtain

Corollary 4: If in the conditions of Theorem 3 we replace (46) with

$$\sup_x \mathrm{Var}\left[\log \frac{dP_{Y|X=x}}{dQ_Y}(Y)\,\middle|\, X=x\right] \le S_m \tag{48}$$

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

$$D(P_{Y|X}\|Q_Y|P_X) \ge \log M - \sqrt{\frac{2S_m}{1-\epsilon}} + \log\frac{1-\epsilon}{2}\,. \tag{49}$$

B. DMC

Notice that when Q_Y is chosen to be a product distribution, such as $P^*_{Y^n}$, $\log\frac{dP_{Y|X=x}}{dQ_Y}$ becomes a sum of independent random variables. In particular, (24) leads to a necessary and sufficient condition for (3).

Theorem 5: Consider a memoryless channel belonging to one of the three classes in Section II. Then, for any $0<\epsilon<1$ and any sequence of $(n, M_n, \epsilon)_{max,det}$ capacity-achieving codes we have

$$\frac{1}{n}I(X^n;Y^n)\to C \iff \frac{1}{n}D(P_{Y^n}\|P^*_{Y^n})\to 0, \tag{50}$$

where X^n is the output of the encoder.

Proof: The direction $\Rightarrow$ is trivial from property (20) of $P^*_{Y^n}$. For the direction $\Leftarrow$ we only need to lower bound $I(X^n;Y^n)$ since, asymptotically, it cannot exceed nC. To that end, we have from (24) and Corollary 4:

$$D(P_{Y^n|X^n}\|P^*_{Y^n}|P_{X^n}) \ge \log M_n + O(\sqrt{n})\,. \tag{51}$$

Then, the conclusion follows from (27) and the following identity applied with $Q_{Y^n} = P^*_{Y^n}$:

$$I(X^n;Y^n) = D(P_{Y^n|X^n}\|Q_{Y^n}|P_{X^n}) - D(P_{Y^n}\|Q_{Y^n}), \tag{52}$$

which holds for all Q_{Y^n} such that the unconditional relative entropy is finite. ■

We remark that Theorem 5 can also be derived from a simple extension of the Wolfowitz converse [14] to an arbitrary output distribution Q_{Y^n}, e.g., [16, Th. 10], and then choosing $Q_{Y^n} = P^*_{Y^n}$. Note that Theorem 1 allows us to conclude (50) but only for capacity-achieving codes with vanishing error probability, which are a subclass of those considered in Theorem 5.

Fano's inequality only guarantees the left side of (50) for code sequences with vanishing error probability. If there was a strong converse showing that the left side of (50) must hold for any sequence of (n, M_n, ϵ) codes, then the desired result (3) would follow. In the absence of such a result we will consider three separate cases in order to show (3), and, therefore, through Theorem 5, the left side of (50).

1) DMC With $C_1 < \infty$: For a given DMC denote the parameter introduced by Burnashev [17]

$$C_1 = \max_{a,a'} D(P_{Y|X=a}\|P_{Y|X=a'})\,. \tag{53}$$

Note that $C_1<\infty$ if and only if the transition matrix does not contain any zeros. In this section, we show (33) for a (regular) class of DMCs with $C_1<\infty$ by an application of the main inequality (47). We also demonstrate that (3) may not hold for codes with nondeterministic encoders or unconstrained maximal probability of error.

Theorem 6: Consider a DMC $P_{Y|X}$ with $C_1<\infty$ and capacity $C>0$ (with or without an input constraint). Then, for any $0\le\epsilon<1$ there exists a constant $a=a(\epsilon)>0$ such that any $(n, M_n, \epsilon)_{max,det}$ code satisfies

$$D(P_{Y^n}\|P^*_{Y^n}) \le nC - \log M_n + a\sqrt{n}, \tag{54}$$

where P_{Y^n} is the output distribution induced by the code. In particular, for any capacity-achieving sequence of such codes we have

$$\frac{1}{n}D(P_{Y^n}\|P^*_{Y^n})\to 0\,. \tag{55}$$

Proof: Fix $y^n, \bar{y}^n \in \mathcal{Y}^n$ which differ in the jth letter only. Then, denoting $y_{\setminus j} = \{y_k, k\ne j\}$ we have

$$\begin{aligned}
|\log P_{Y^n}(y^n) - \log P_{Y^n}(\bar{y}^n)| &= \left|\log\frac{P_{Y_j|Y_{\setminus j}}(y_j|y_{\setminus j})}{P_{Y_j|Y_{\setminus j}}(\bar{y}_j|y_{\setminus j})}\right| && (56)\\
&\le \max_{a,b,b'}\log\frac{P_{Y|X}(b|a)}{P_{Y|X}(b'|a)} && (57)\\
&\triangleq a_1 < \infty, && (58)
\end{aligned}$$

where (57) follows from

$$P_{Y_j|Y_{\setminus j}}(b|y_{\setminus j}) = \sum_{a\in\mathcal{X}} P_{Y|X}(b|a)P_{X_j|Y_{\setminus j}}(a|y_{\setminus j})\,. \tag{59}$$

Thus, the function $y^n \mapsto \log P_{Y^n}(y^n)$ is a_1-Lipschitz in Hamming metric on $\mathcal{Y}^n$. Its discrete gradient (absolute difference of values taken at consecutive integers) is bounded by $n|a_1|^2$ and thus by the discrete Poincaré inequality (the variance of a function with countable support is upper bounded by (a multiple of) the second moment of its discrete gradient) [18, Th. 4.1f] we have

$$\mathrm{Var}\left[\log P_{Y^n}(Y^n)|\, X^n = x^n\right] \le n|a_1|^2\,. \tag{60}$$

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

$$\begin{aligned}
&\mathrm{Var}\left[\imath_{X^n;Y^n}(X^n;Y^n)|\, X^n = x^n\right]\\
&\quad\le 2\mathrm{Var}\left[\log P_{Y^n|X^n}(Y^n|X^n)\middle|\, X^n = x^n\right]\\
&\qquad + 2\mathrm{Var}\left[\log P_{Y^n}(Y^n)|\, X^n = x^n\right]\\
&\quad\le 2na_2 + 2n|a_1|^2, && (61)
\end{aligned}$$

where (61) follows from

$$\mathrm{Var}\left[\sum_{i=1}^K Y_i\right] \le K\sum_{i=1}^K \mathrm{Var}[Y_i] \tag{62}$$

and (61) follows from (60) and the fact that the random variable in the first variance in (61) is a sum of n independent terms. Applying Corollary 4 with $S_m = 2na_2 + 2n|a_1|^2$ and $Q_Y = P_{Y^n}$ we obtain

$$D(P_{Y^n|X^n}\|P_{Y^n}|P_{X^n}) \ge \log M_n + O(\sqrt{n})\,. \tag{63}$$

We can now complete the proof:

$$\begin{aligned}D(P_{Y^n}\|P^*_{Y^n}) &= D(P_{Y^n|X^n}\|P^*_{Y^n}|P_{X^n})\\ &\quad - D(P_{Y^n|X^n}\|P_{Y^n}|P_{X^n}) \qquad (64)\\ &\le nC - D(P_{Y^n|X^n}\|P_{Y^n}|P_{X^n}) \qquad (65)\\ &\le nC - \log M_n + O(\sqrt{n}) \qquad (66)\end{aligned}$$

where (65) is because $P^*_{Y^n}$ satisfies (19) and (66) follows from (63). This completes the proof of (54). ■

Remark 3: As we will see in Section IV-A, (55) implies

$$H(Y^n) = nH(Y^*) + o(n) \qquad (67)$$

[by (132) applied to $f(y) = \log P^*_Y(y)$]. Note also that traditional combinatorial methods, e.g., [19], 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 that are not of the form of cumulative average.

Remark 4: Note that any (n, M, ϵ) code is also an (n, M, ϵ') code for all $\epsilon' \ge \epsilon$. Thus $a(\epsilon)$, the constant in (54), is a non-decreasing function of ϵ. In particular, (54) holds uniformly in ϵ on compact subsets of $[0, 1)$. In their follow-up to this paper, Raginsky and Sason [20] use McDiarmid's inequality to derive a tighter estimate for a.

Remark 5: Equation (55) need not hold if the maximal probability of error is replaced with the average or if the encoder is allowed to be random. Indeed, for any $0 < \epsilon < 1$ we construct a sequence of $(n, M_n, \epsilon)_{avg}$ capacity-achieving codes which do not satisfy (55) can be constructed as follows. Consider a sequence of $(n, M'_n, \epsilon'_n)_{max,det}$ codes with $\epsilon'_n \to 0$ and

$$\frac{1}{n}\log M'_n \to C\,. \qquad (68)$$

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 $(x_0, \ldots, x_0) \in \mathcal{X}_n$, where x_0 is some fixed letter in $\mathcal{X}$ such that

$$D(P_{Y|X=x_0}\|P^*_Y) > 0 \qquad (69)$$

(the 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 $\frac{\epsilon-\epsilon'_n}{1-\epsilon}M'_n$ identical codewords: $(x_0, \ldots, x_0) \in \mathcal{X}_n$. Then, the minimal average probability of error achievable with the extended codebook of size

$$M_n \triangleq \frac{1-\epsilon_n}{1-\epsilon}M'_n \qquad (70)$$

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

$$S = 1\{X^n = (x_0, \ldots, x_0)\} \qquad (71)$$

with distribution

$$P_S(1) = 1 - P_S(0) = \frac{\epsilon - \epsilon'_n}{1 - \epsilon'_n} \qquad (72)$$

which satisfies $P_S(1) \to \epsilon$. We have then

$$\begin{aligned}D(P_{Y^n}\|P^*_{Y^n}) &= D(P_{Y^n|S}\|P^*_{Y^n}|P_S) - I(S;Y^n) \qquad (73)\\ &\ge D(P_{Y^n|S}\|P^*_{Y^n}|P_S) - \log 2 \qquad (74)\\ &= nD(P_{Y|X=x_0}\|P^*_Y)P_S(1)\\ &\quad + D(P'_{Y^n}\|P^*_{Y^n})P_S(0) - \log 2 \qquad (75)\\ &= nD(P_{Y|X=x_0}\|P^*_Y)P_S(1) + o(n), \qquad (76)\end{aligned}$$

where (73) is by (52), (74) follows since S is binary, (75) is by noticing that $P_{Y^n|S=0} = P'_{Y^n}$, and (76) is by (55). It is clear that (69) and (76) show the impossibility of (55) for this code.

Similarly, one shows that (55) 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 $(x_0, \ldots, x_0) \in \mathcal{X}_n$ and otherwise it outputs the original codeword. The same analysis shows that (76) holds again and thus (55) fails.

The counterexamples constructed above can also be used to demonstrate that in Theorem 3 (and hence Theorem 2) the assumptions of maximal probability of error and deterministic encoders are not superfluous, contrary to what is claimed by Ahlswede [13, Remark 1].

2) DMC With $C_1 = \infty$: Next, we show an estimate for $D(P_{Y^n}\|P^*_{Y^n})$ differing by a $\log^{\frac{3}{2}} n$ factor from (33) for the DMCs with $C_1 = \infty$.

Theorem 7: For any DMC $P_{Y|X}$ with capacity $C > 0$ (with or without input constraints), $C_1 = \infty$, and $0 \le \epsilon < 1$ there exists a constant $b > 0$ with the property that for any sequence of $(n, M_n, \epsilon)_{max,det}$ codes we have for all $n \ge 1$

$$D(P_{Y^n}\|P^*_{Y^n}) \le nC - \log M_n + b\sqrt{n}\log^{\frac{3}{2}} n\,. \qquad (77)$$

In particular, for any such sequence achieving capacity, we have

$$\frac{1}{n}D(P_{Y^n}\|P^*_{Y^n}) \to 0\,. \qquad (78)$$

Proof: Let c_i and $D_i, i = 1, \ldots M_n$ denote the codewords and the decoding regions of the code. Denote the sequence

$$\ell_n = b_1\sqrt{n\log n} \qquad (79)$$

with $b_1 > 0$ to be further constrained shortly. According to the isoperimetric inequality for Hamming space [19, Corollary I.5.3], there is a constant $a > 0$ such that for every $i = 1, \ldots, M_n$

$$1 - P_{Y^n|X^n=c_i}[\Gamma^{\ell_n}D_i] \le Q\left(Q^{-1}(\epsilon) + \frac{\ell_n}{\sqrt{n}}a\right) \qquad (80)$$

$$\begin{aligned} &\le \exp\left\{-b_2\frac{\ell_n^2}{n}\right\} && (81)\\ &= n^{-b_2} && (82)\\ &\le \frac{1}{n}, && (83)\end{aligned}$$

where the ℓ-blowup of D is defined as

$$\Gamma^\ell D = \{\bar{y}^n \in \mathcal{Y}^n : \exists y^n \in D \text{ s.t. } |\{j : y_j \neq \bar{y}_j\}| \le \ell\} \quad (84)$$

denotes the ℓth Hamming neighborhood of a set D and we assumed that b_1 was chosen large enough so there is $b_2 \ge 1$ satisfying (83).

Let

$$M'_n = \frac{M_n}{n\binom{n}{\ell_n}|\mathcal{Y}|^{\ell_n}} \quad (85)$$

and consider a subcode $F = (F_1, \ldots, F_{M'_n})$, where $F_i \in \mathcal{C} = \{c_1, \ldots, c_M\}$ and note that we allow repetition of codewords. Then, for every possible choice of the subcode F we denote by $P_{X^n(F)}$ and $P_{Y^n(F)}$ the input/output distribution induced by F, so that for example

$$P_{Y^n(F)} = \frac{1}{M'_n}\sum_{j=1}^{M'_n} P_{Y^n|X^n=F_j}\,. \quad (86)$$

We aim to apply the random coding argument over all equally likely $M_n^{M'_n}$ choices of a subcode F. Random coding among subcodes was originally invoked in [7] to demonstrate the existence of a good subcode. The expected (over the choice of F) induced output distribution is

$$\begin{aligned} \mathbb{E}[P_{Y^n}(F)] &\triangleq \frac{1}{M_n^{M'_n}}\sum_{F_1\in\mathcal{C}}\cdots\sum_{F_{M'_n}\in\mathcal{C}} P_{Y^n(F)} && (87)\\ &= \frac{1}{M_n^{M'_n}}\frac{1}{M'_n}\sum_{j=1}^{M'_n}\sum_{F_1\in\mathcal{C}}\cdots\sum_{F_{M'_n}\in\mathcal{C}} P_{Y^n|X^n=F_j} \\ &= \frac{M_n^{M'_n-1}}{M_n^{M'_n}}\sum_{c\in\mathcal{C}} P_{Y^n|X^n=c} && (88)\\ &= P_{Y^n}\,. && (89)\end{aligned}$$

Next, for every F we denote by $\epsilon'(F)$ the minimal possible average probability of error achieved by an appropriately chosen decoder. With this notation we have, for every possible value of F:

$$\begin{aligned} D(P_{Y^n(F)}\|P^*_{Y^n}) &= D(P_{Y^n|X^n}\|P^*_{Y^n}|P_{X^n(F)}) \\ &\quad - I(X^n(F);Y^n(F)) && (90)\\ &\le nC - I(X^n(F);Y^n(F)) && (91)\\ &\le nC - (1-\epsilon'(F))\log M'_n + \log 2 && (92)\\ &\le nC - \log M'_n + n\epsilon'(F)\log|\mathcal{X}| + \log 2 && (93)\\ &\le nC - \log M_n + n\epsilon'(F)\log|\mathcal{X}| \\ &\quad + b_3\sqrt{n}\log^{\frac{3}{2}} n && (94)\end{aligned}$$

where (90) is by (52), (91) is by (19), (92) is by Fano's inequality, (93) is because $\log M'_n \le n\log|\mathcal{X}|$, and (94) holds for some $b_3 > 0$ by the choice of M'_n in (85) and by

$$\log\binom{n}{\ell_n} \le \ell_n \log n\,. \quad (95)$$

Taking the expectation of both sides of (94), applying convexity of relative entropy and (89) we get

$$\begin{aligned} D(P_{Y^n}\|P^*_{Y^n}) &\le nC - \log M_n + n\mathbb{E}[\epsilon'(F)]\log|\mathcal{X}| \\ &\quad + b_3\sqrt{n}\log^{\frac{3}{2}} n\,. && (96)\end{aligned}$$

Accordingly, it remains to show that

$$n\mathbb{E}[\epsilon'(F)] \le 2\,. \quad (97)$$

To that end, for every subcode F define the suboptimal randomized decoder that chooses for $F_j \in L(y, F)$,

$$\hat{W}(y) = F_j \quad \text{with probability } \frac{1}{|L(y,F)|}, \quad (98)$$

where $L(y, F)$ is a list of those indices $i \in F$ for which $y \in \Gamma^{\ell_n} D_i$. Since the transmitted codeword F_W is equiprobable on F, averaging over the selection of F we have

$$\mathbb{E}[|L(Y^n,F)|\,|\,F_W \in L(Y^n,F)] \le 1 + \frac{\binom{n}{\ell_n}|\mathcal{Y}|^{\ell_n}}{M_n}(M'_n - 1), \quad (99)$$

because each $y \in \mathcal{Y}^n$ can belong to at most $\binom{n}{\ell_n}|\mathcal{Y}|^{\ell_n}$ enlarged decoding regions $\Gamma^{\ell_n} D_i$ and each F_j is chosen independently and equiprobably among all possible M_n alternatives. The average (over random decoder, F, and channel) probability of error can be upper bounded as

$$\begin{aligned} \mathbb{E}[\epsilon'(F)] &= \mathbb{P}[F_W \notin L(Y^n,F)] \\ &\quad + \mathbb{E}\left[\frac{|L(Y^n,F)|-1}{|L(Y^n,F)|}1\{F_W \in L(Y^n,F)\}\right] && (100)\\ &\le \mathbb{P}[F_W \notin L(Y^n,F)] + \frac{\binom{n}{\ell_n}|\mathcal{Y}|^{\ell_n}M'_n}{M_n} && (101)\\ &\le \frac{1}{n} + \frac{\binom{n}{\ell_n}|\mathcal{Y}|^{\ell_n}M'_n}{M_n} && (102)\\ &\le \frac{2}{n}, && (103)\end{aligned}$$

where (100) reflects the fact that a correct decision requires that the true codeword not only belong to $L(Y^n, F)$ but that it be the one chosen from the list; (101) is by Jensen's inequality applied to $\frac{x-1}{x}$ and (99); (102) is by (83); and (103) is by (85). Since (103) also serves as an upper bound to $\mathbb{E}[\epsilon'(F)]$ the proof of (97) is complete. ■

Remark 6: Claim (78) fails to hold if either the maximal probability of error is replaced with the average, or if we allow the encoder to be stochastic. Counterexamples are constructed exactly as in Remark 5.

Remark 7: Raginsky and Sason [20] give a sharpened version of (77) with explicitly computed constants but with the same $O(\sqrt{n\log^3 n})$ remainder term behavior.

C. Gaussian Channel

Theorem 8: For any $0 < \epsilon < 1$ and $P > 0$ there exists $a = a(\epsilon, P) > 0$ such that the output distribution P_{Y^n} of any $(n, M_n, \epsilon)_{max,det}$ code for the AWGN(P) channel satisfies[9]

$$D(P_{Y^n} \| P^*_{Y^n}) \le nC - \log M_n + a\sqrt{n}, \tag{104}$$

where $P^*_{Y^n} = \mathcal{N}(0, 1+P)^n$. In particular, for any capacity-achieving sequence of such codes we have

$$\frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) \to 0\,. \tag{105}$$

Proof: Denote by $p_{Y^n|X^n=\mathbf{x}}$ and p_{Y^n} the densities of $P_{Y^n|X^n=\mathbf{x}}$ and P_{Y^n}, respectively. The argument proceeds step by step as in the proof of Theorem 6 with (106) taking the place of (60) and recalling that property (19) holds for the AWGN channel too. Therefore, the objective is to show

$$\mathrm{Var}[\log p_{Y^n}(Y^n) \,|\, X^n] \le a_1 n \tag{106}$$

for some $a_1 > 0$. Poincaré's inequality for the Gaussian measure, e.g., [21, (2.16)] states that if Y is an N-dimensional Gaussian measure, then

$$\mathrm{Var}[f(Y)] \le \mathbb{E}[\|\nabla f(Y)\|^2]. \tag{107}$$

Since conditioned on X^n, the random vector Y^n is Gaussian, the Poincaré inequality ensures that the left side of (106) is bounded by

$$\mathrm{Var}[\log p_{Y^n}(Y^n) \,|\, X^n] \le \mathbb{E}[\|\nabla \log p_{Y^n}\|^2 \,|\, X^n]. \tag{108}$$

Therefore, the reminder of the proof is devoted to showing that the right side of (108) is bounded by $a_1 n$ for some $a_1 > 0$. An elementary computation shows

$$\begin{aligned}
\nabla \log p_{Y^n}(\mathbf{y}) &= \frac{\log e}{p_{Y^n}(\mathbf{y})} \nabla p_{Y^n}(\mathbf{y}) && (109)\\
&= \frac{\log e}{p_{Y^n}(\mathbf{y})} \sum_{j=1}^{M} \frac{1}{M(2\pi)^{\frac{n}{2}}} \nabla e^{-\frac{1}{2}\|\mathbf{y}-\mathbf{c}_j\|^2} && (110)\\
&= \frac{\log e}{M(2\pi)^{\frac{n}{2}} p_{Y^n}(\mathbf{y})} \sum_{j=1}^{M} (\mathbf{c}_j - \mathbf{y}) e^{-\frac{1}{2}\|\mathbf{y}-\mathbf{c}_j\|^2} \\
&= (\mathbb{E}[X^n | Y^n = \mathbf{y}] - \mathbf{y}) \log e\,. && (111)
\end{aligned}$$

For convenience, denote

$$\hat{X}^n = \mathbb{E}[X^n | Y^n] \tag{112}$$

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

$$\left\|\hat{X}^n\right\| \le \sqrt{nP}\,. \tag{113}$$

Then,

$$\begin{aligned}
&\frac{1}{\log^2 e} \mathbb{E}[\|\nabla \log p_{Y^n}(Y^n)\|^2 \,|\, X^n] \\
&= \mathbb{E}\left[\left\|Y^n - \hat{X}^n\right\|^2 \,\middle|\, X^n \right] && (114)\\
&\le 2\mathbb{E}\left[\|Y^n\|^2 \,\middle|\, X^n \right] + 2\mathbb{E}\left[\left\|\hat{X}^n\right\|^2 \,\middle|\, X^n \right] && (115)\\
&\le 2\mathbb{E}\left[\|Y^n\|^2 \,\middle|\, X^n \right] + 2nP && (116)\\
&= 2\mathbb{E}\left[\|X^n + Z^n\|^2 \,\middle|\, X^n \right] + 2nP && (117)\\
&\le 4\|X^n\|^2 + 4n + 2nP && (118)\\
&\le (6P+4)n, && (119)
\end{aligned}$$

where (115) is by

$$\|\mathbf{a} + \mathbf{b}\|^2 \le 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2\,. \tag{120}$$

Equation (116) is by (113), in (117) we introduced $Z^n \sim \mathcal{N}(0, \mathbf{I}_n)$ which is independent of X^n, (118) is by (120) and (119) is by the power-constraint imposed on the codebook. In view of (108), we have succeeded in identifying a constant a_1 such that (106) holds. ■

Remark 8: Equation (105) need not hold if the maximal probability of error is replaced with the average or if the encoder is allowed to be stochastic. Counterexamples are constructed similarly to those for Remark 5 with $x_0 = 0$. Note also that Theorem 8 need not hold if the power-constraint is in the average-over-the-codebook sense; see [16, Sec. 4.3.3].

IV. Implications

We have shown that there is a constant $a = a(\epsilon)$ independent of n and M_n such that

$$D(P_{Y^n} \| P^*_{Y^n}) \le nC - \log M_n + a\sqrt{n}, \tag{121}$$

where P_{Y^n} is the output distribution induced by an arbitrary $(n, M_n, \epsilon)_{max,det}$ code. Therefore, any $(n, M_n, \epsilon)_{max,det}$ necessarily satisfies

$$\log M_n \le nC + a(\epsilon)\sqrt{n} \tag{122}$$

as is classically known [22]. In particular, (121) implies that any ϵ-capacity-achieving code must satisfy (3). In this section, we discuss this and other implications of this result, such as

1) Equation (121) implies that the empirical marginal output distribution
$$\bar{P}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} P_{Y_i} \tag{123}$$
converges to P^*_Y in a strong sense (see Section IV-A).
2) Equation (121) guarantees estimates of the precision in the approximation (4) (see Sections IV-B and IV-E).
3) Equation (121) provides estimates for the deviations of $f(Y^n)$ from its average (see Section IV-C).
4) Relation to optimal transportation (see Section IV-D),
5) Implications of (3) for the empirical input distribution of the code (see Sections IV-G and IV-H).

A. Empirical Distributions and Empirical Averages

Considering the empirical marginal distributions, the convexity of relative entropy and (3) result in

$$D(\bar{P}_n \| P^*_Y) \le \frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) \to 0, \tag{124}$$

[9] More precisely, our proof yields the bound $nC - \log M_n + \sqrt{6n(3+4P)}\log e + \log \frac{2}{1-\epsilon}$.

where $\bar{P}_n$ is the empirical marginal output distribution (123).

More generally, we have [3, (41)]

$$D(\bar{P}_n^{(k)} \| P^*_{Y^k}) \le \frac{k}{n-k+1} D(P_{Y^n} \| P^*_{Y^n}) \to 0, \tag{125}$$

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

$$\bar{P}_n^{(k)} = \frac{1}{n-k+1} \sum_{j=1}^{n-k+1} P_{Y_j^{j+k-1}} . \tag{126}$$

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

$$D(P_n \| P) \to 0 \tag{127}$$

implies convergence properties for the expectations of functions

$$\int f dP_n \to \int f dP. \tag{128}$$

1) For bounded functions, (128) follows from the Csiszár–Kemperman–Kullback–Pinsker inequality (e.g., [23]):

$$\|P_n - P\|^2_{TV} \le \frac{1}{2 \log e} D(P_n \| P), \tag{129}$$

where

$$\|P - Q\|_{TV} \triangleq \sup_A |P(A) - Q(A)|. \tag{130}$$

2) For unbounded f, (128) holds as long as f satisfies Cramer's condition under P, i.e.,

$$\int e^{tf} dP < \infty \tag{131}$$

for all t in some neighborhood of 0; see [24, Lemma 3.1].

Together (128) and (124) 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 capacity achieving output distribution (caod):

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} f(Y_j)\right] \to \int f dP^*_Y . \tag{132}$$

From (125), a similar conclusion holds for kth order empirical averages.

B. Averages of Functions of Y^n

To go beyond empirical averages, we need to provide some definitions and properties (see [21]).

Definition 1: The function $F: \mathcal{Y}^n \to \mathbb{R}$ is called (b,c)-concentrated with respect to measure μ on $\mathcal{Y}^n$ if for all $t \in \mathbb{R}$

$$\int \exp\{t(F(Y^n) - \bar{F})\} d\mu \le b \exp\{ct^2\}, \qquad \bar{F} = \int F d\mu . \tag{133}$$

A function F is called (b,c)-concentrated for the channel if it is (b,c)-concentrated with respect to every $P_{Y^n|X^n=x}$ and $P^*_{Y^n}$ and all n.

A couple of simple properties of (b,c)-concentrated functions:

1) Gaussian concentration around the mean

$$\mathbb{P}[|F(Y^n) - \mathbb{E}[F(Y^n)]| > t] \le b \exp\left\{-\frac{t^2}{4c}\right\} . \tag{134}$$

2) Small variance

$$\begin{aligned} \mathrm{Var}[F(Y^n)] &= \int_0^\infty \mathbb{P}[|F(Y^n) - \mathbb{E}[F(Y^n)]|^2 > t] dt \\ &\le \int_0^\infty \min\left\{b \exp\left\{-\frac{t}{4c}\right\}, 1\right\} dt \qquad (135) \\ &= 4c \log(2be) . \qquad (136) \end{aligned}$$

Some examples of concentrated functions include:

1) A bounded function F with $\|F\|_\infty \le A$ is $(\exp\{A^2(4c)^{-1}\}, c)$-concentrated for any c and any measure μ. Moreover, for a fixed μ and a sufficiently large c any bounded function is $(1,c)$-concentrated.
2) If F is (b,c)-concentrated then λF is $(b, \lambda^2 c)$-concentrated.
3) Let $f: \mathcal{Y} \to \mathbb{R}$ be $(1,c)$-concentrated with respect to μ. Then, so is

$$F(y^n) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(y_j) \tag{137}$$

with respect to μ^n. In particular, any F defined in this way from a bounded f is $(1,c)$-concentrated for a memoryless channel (for a sufficiently large c independent of n).

4) If $\mu = \mathcal{N}(0,1)^n$ and F is a Lipschitz function on $\mathbb{R}^n$ with Lipschitz constant $\|F\|_{Lip}$ then F is $(1, \frac{\|F\|^2_{Lip}}{2\log e})$-concentrated with respect to μ, e.g., [25, Proposition 2.1]:

$$\int_{\mathbb{R}^n} \exp\{t(F(y^n) - \bar{F})\} d\mu(y^n) \le \exp\left\{\frac{\|F\|^2_{Lip}}{2\log e} t^2\right\} . \tag{138}$$

Therefore, any Lipschitz function is $(1, \frac{(1+P)\|F\|^2_{Lip}}{2\log e})$-concentrated for the AWGN channel.

5) For discrete $\mathcal{Y}^n$ endowed with the Hamming distance

$$d(y^n, z^n) = |\{i: y_i \ne z_i\}| \tag{139}$$

define Lipschitz functions in the usual way. In this case, a simpler criterion is: $F: \mathcal{Y}^n \to \mathbb{R}$ is Lipschitz with constant ℓ if and only if

$$\max_{y^n, b, j} |F(y_1, \ldots, y_j, \ldots, y_n) - F(y_1, \ldots, b, \ldots, y_n)| \le \ell. \tag{140}$$

Let μ be any product probability measure $P_1 \times \ldots \times P_n$ on $\mathcal{Y}^n$, then the standard Azuma–Hoeffding estimate shows that

$$\sum_{y^n \in \mathcal{Y}^n} \exp\{t(F(y^n) - \bar{F})\}\mu(y^n) \le \exp\left\{\frac{n\|F\|_{Lip}^2}{2\log e} t^2\right\} \tag{141}$$

and thus any Lipschitz function F is $(1, \frac{n\|F\|_{Lip}^2}{2\log e})$-concentrated with respect to any product measure on $\mathcal{Y}^n$.

Note that unlike the Gaussian case, the constant of concentration c worsens linearly with dimension n. Generally, this growth cannot be avoided as shown by the coefficient $\frac{1}{\sqrt{n}}$ in the exact solution of the Hamming isoperimetric problem [26]. At the same time, this growth does not mean that (141) is "weaker" than (138); for example, $F = \sum_{j=1}^n \phi(y_j)$ has Lipschitz constant $O(\sqrt{n})$ in Euclidean space and $O(1)$ in Hamming. However, for convex functions the concentration (138) holds for product measures even under Euclidean distance [27].

We now show how to approximate expectations of concentrated functions.

Proposition 9: Suppose that $F\colon \mathcal{Y}^n \to \mathbb{R}$ is (b, c)-concentrated with respect to $P^*_{Y^n}$. Then,

$$|\mathbb{E}[F(Y^n)] - \mathbb{E}[F(Y^{*n})]| \le 2\sqrt{cD(P_{Y^n}\|P^*_{Y^n}) + c\log b}, \tag{142}$$

where

$$\mathbb{E}[F(Y^{*n})] = \int F(y^n) dP^*_{Y^n} \,. \tag{143}$$

Proof: Recall the Donsker–Varadhan inequality [28, 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 $\int g dP$ exists (but perhaps is $-\infty$) and moreover

$$\int g dP - \log \int \exp\{g\} dQ \le D(P\|Q) \,. \tag{144}$$

Since by (133) the moment generating function of F exists under $P^*_{Y^n}$, applying (144) to tF we get

$$t\mathbb{E}[F(Y^n)] - \log \mathbb{E}[\exp\{tF(Y^{*n})\}] \le D(P_{Y^n}\|P^*_{Y^n}) \,. \tag{145}$$

From (133), we have

$$ct^2 - t\mathbb{E}[F(Y^n)] + t\mathbb{E}[F(Y^{*n})] + D(P_{Y^n}\|P^*_{Y^n}) + \log b \ge 0 \tag{146}$$

for all t. Thus, the discriminant of the parabola in (146) must be nonpositive which is precisely (142). ■

Note that for empirical averages $F(y^n) = \frac{1}{n}\sum_{j=1}^n f(y_i)$ we may either apply the estimate for concentration in the example (137) and then use Proposition 9, or directly apply Proposition 9 to (124); the result is the same:

$$\left|\frac{1}{n}\sum_{j=1}^n \mathbb{E}[f(Y_j)] - \mathbb{E}[f(Y^*)]\right| \le 2\sqrt{\frac{c}{n} D(P_{Y^n}\|P^*_{Y^n})} \to 0, \tag{147}$$

for any f which is $(1, c)$-concentrated with respect to P^*_Y.

For the Gaussian channel, Proposition 9 and (138) yield

Corollary 10: 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 Lipschitz function $F\colon \mathbb{R}^n \to \mathbb{R}$ we have

$$|\mathbb{E}[F(Y^n)] - \mathbb{E}[F(Y^{*n})]| \le a_1\|F\|_{Lip}\sqrt{nC - \log M_n + a_2\sqrt{n}}, \tag{148}$$

where $C = \frac{1}{2}\log(1 + P)$ is the capacity.

Note that in the proof of Corollary 10, concentration of measure is used twice: once for $P_{Y^n|X^n}$ in the form of Poincaré's inequality (proof of Theorem 8) and once in the form of (133) (proof of Proposition 9).

C. Concentration of Functions of Y^n

Not only can we estimate expectations of $F(Y^n)$ by replacing the unwieldy P_{Y^n} with the simple $P^*_{Y^n}$, but in fact the distribution of $F(Y^n)$ exhibits a sharp peak at its expectation.

Proposition 11: Consider a channel for which (121) holds. Then, for any F which is (b, c)-concentrated for such channel, we have for every $(n, M, \epsilon)_{max,det}$ code

$$\begin{aligned}\mathbb{P}[|F(Y^n) - \mathbb{E}[F(Y^{*n})]| > t]\\ \le 3b\exp\left\{nC - \log M + a\sqrt{n} - \frac{t^2}{16c}\right\}\end{aligned} \tag{149}$$

and,

$$\mathrm{Var}[F(Y^n)] \le 16c\left(nC - \log M + a\sqrt{n} + \log(6be)\right) \,. \tag{150}$$

Proof: Denote for convenience

$$\bar{F} \triangleq \mathbb{E}[F(Y^{*n})], \tag{151}$$

$$\phi(x^n) = \mathbb{E}[F(Y^n)|X^n = x^n] \,. \tag{152}$$

Then, as a consequence of F being (b, c)-concentrated for $P_{Y^n|X^n=x^n}$ we have

$$\mathbb{P}[|F(Y^n) - \phi(x^n)| > t | X^n = x^n] \le b\exp\left\{-\frac{t^2}{4c}\right\} \,. \tag{153}$$

Consider now a subcode $\mathcal{C}_1$ consisting of all codewords such that $\phi(x^n) > \bar{F} + t$ for $t > 0$. The number $M_1 = |\mathcal{C}_1|$ of codewords in this subcode is

$$M_1 = M\mathbb{P}[\phi(X^n) > \bar{F} + t] \,. \tag{154}$$

Let Q_{Y^n} be the output distribution induced by $\mathcal{C}_1$. We have the following chain:

$$\bar{F} + t \le \frac{1}{M_1}\sum_{x \in \mathcal{C}_1} \phi(x^n) \tag{155}$$

$$= \int F(Y^n) dQ_{Y^n} \tag{156}$$

$$\le \bar{F} + 2\sqrt{cD(Q_{Y^n}\|P^*_{Y^n}) + c\log b} \tag{157}$$

$$\le \bar{F} + 2\sqrt{c(nC - \log M_1 + a\sqrt{n}) + c\log b} \tag{158}$$

where (155) is by the definition of $\mathcal{C}_1$, (156) is by (152), (157) is by Proposition 9, and the assumption of (b, c)-concentration of F under $P^*_{Y^n}$, and (158) is by (121).

Together (154) and (158) imply

$$\mathbb{P}[\phi(X^n) > \bar{F} + t] \le b \exp\left\{ nC - \log M + a\sqrt{n} - \frac{t^2}{4c} \right\}. \tag{159}$$

Applying the same argument to $-F$, we obtain a similar bound on $\mathbb{P}[|\phi(X^n) - \bar{F}| > t]$ and thus

$$\begin{aligned} &\mathbb{P}[|F(Y^n) - \bar{F}| > t] \\ &\le \mathbb{P}[|F(Y^n) - \phi(X^n)| > t/2] + \mathbb{P}[|\phi(X^n) - \bar{F}| > t/2] \\ &\le b \exp\left\{ -\frac{t^2}{16c} \right\} \left(1 + 2\exp\{nC - \log M + a\sqrt{n}\}\right) \qquad (160) \\ &\le 3b \exp\left\{ -\frac{t^2}{16c} + nC - \log M + a\sqrt{n} \right\}, \qquad (161) \end{aligned}$$

where (160) is by (153) and (159); and (161) is by (122). Thus, (149) is proven. Moreover, (150) follows by (136). ■

Following up on Proposition 11, [20] gives a bound, which in contrast to (149), shows explicit dependence on ϵ.

D. Relation to Optimal Transportation

Since the seminal work of Marton [8], [29], optimal transportation theory has emerged as one of the major tools for proving (b, c)-concentration of Lipschitz functions. Marton demonstrated that if a probability measure μ on a metric space satisfies *a T_1 inequality*

$$W_1(\nu, \mu) \le \sqrt{c' D(\nu \| \mu)} \qquad \forall \nu \tag{162}$$

then any Lipschitz f is $(b, \|f\|_{Lip}^2 c)$-concentrated with respect to μ for some $b = b(c, c')$ and any $0 < c < \frac{c'}{4}$. In (162), $W_1(\nu, \mu)$ denotes the linear-cost transportation distance, or Wasserstein-1 distance, defined as

$$W_1(\nu, \mu) \triangleq \inf_{P_{YY'}} \mathbb{E}[d(Y, Y')], \tag{163}$$

where $d(\cdot, \cdot)$ is the distance on the underlying metric space and the infimum is taken over all couplings $P_{YY'}$ with fixed marginals $P_Y = \mu$, $P_{Y'} = \nu$. Note that according to [30], we have $\|\nu - \mu\|_{TV} = W_1(\nu, \mu)$ when the underlying distance on $\mathcal{Y}$ is $d(y, y') = 1\{y \neq y'\}$.

In this section, we show that (162) in fact directly implies the estimate of Proposition 9 without invoking either Marton's argument or Donsker–Varadhan inequality. Indeed, assume that $F\colon \mathcal{Y}^n \to \mathbb{R}$ is a Lipschitz function and observe that for any coupling $P_{Y^n, Y^{*n}}$ we have

$$|\mathbb{E}[F(Y^n)] - \mathbb{E}[F(Y^{*n})]| \le \|F\|_{Lip} \mathbb{E}[d(Y^n, Y^{*n})], \tag{164}$$

where the distance d is either Hamming or Euclidean depending on the nature of $\mathcal{Y}^n$. Now taking the infimum in the right-hand side of (164) with respect to all couplings we observe

$$|\mathbb{E}[F(Y^n)] - \mathbb{E}[F(Y^{*n})]| \le \|F\|_{Lip} W_1(P_{Y^n}, P_{Y^n}*) \tag{165}$$

and therefore by the transportation inequality (162) we get

$$|\mathbb{E}[F(Y^n)] - \mathbb{E}[F(Y^{*n})]| \le \sqrt{c' \|F\|_{Lip}^2 D(P_{Y^n} \| P_{Y^n}^*)} \tag{166}$$

which is precisely what Proposition 9 yields for $(1, \frac{c'\|F\|_{Lip}^2}{4})$-concentrated functions.

Our argument can be turned around and used to *prove* linear-cost transportation T_1 inequalities (162). Indeed, by the Kantorovich–Rubinstein duality [31, Ch. 1], we have

$$\sup_F |\mathbb{E}[F(Y^n)] - \mathbb{E}[F(Y^{*n})]| = W_1(P_{Y^n}, P_{Y^n}^*), \tag{167}$$

where the supremum is over all F with $\|F\|_{Lip} \le 1$. Thus, the argument in the proof of Proposition 9 shows that (162) must hold for any μ for which every 1-Lipschitz F is $(1, c')$-concentrated, demonstrating an equivalence between T_1 transportation and Gaussian-like concentration—a result reported in [32, Th. 3.1].

We also mention that unlike general i.i.d. measures, an i.i.d. Gaussian $\mu = \mathcal{N}(0, 1)^n$ satisfies a much stronger T_2-transportation inequality [33]

$$W_2(\nu, \mu) \le \sqrt{c' D(\nu \| \mu)} \qquad \forall \nu \ll \mu, \tag{168}$$

where remarkably c' does not depend on n and the Wasserstein-2 distance W_2 is defined as

$$W_2(\nu, \mu) \triangleq \inf_{P_{YY'}} \sqrt{\mathbb{E}[d^2(Y, Y')]}, \tag{169}$$

the infimum being over all couplings as in (163).

E. Empirical Averages of Non-Lipschitz Functions

One drawback of relying on the transportation inequality (162) in the proof of Proposition 9 is that it does not show anything for non-Lipschitz functions. In this section, we demonstrate how the proof of Proposition 9 can be extended to functions that do not satisfy the strong concentration assumptions.

Proposition 12: Let $f\colon \mathcal{Y} \to \mathbb{R}$ be a (single-letter) function such that for some $\theta > 0$ we have $m_1 \triangleq \mathbb{E}[\exp\{\theta f(Y^*)\}] < \infty$ (one-sided Cramer condition) and $m_2 = \mathbb{E}[f^2(Y^*)] < \infty$. Then, there exists $b = b(m_1, m_2, \theta) > 0$ such that for all $n \ge \frac{16}{\theta^4}$ we have

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[f(Y_j)] \le \mathbb{E}[f(Y^*)] + \frac{1}{n^{\frac{3}{4}}} D(P_{Y^n} \| P_{Y^n}^*) + \frac{b}{n^{\frac{1}{4}}}. \tag{170}$$

Proof: It is clear that if the moment-generating function $t \mapsto \mathbb{E}[\exp\{t f(Y^*)\}]$ exists for $t = \theta > 0$ then it also exists for all $0 \le t \le \theta$. Notice that since

$$x^2 \exp\{-x\} \le 4e^{-2} \log e \qquad \forall x \ge 0 \tag{171}$$

we have for all $0 \le t \le \frac{\theta}{2}$:

$$\begin{aligned}
&\mathbb{E}[f^2(Y^*)\exp\{tf(Y^*)\}] \\
&\quad \le \mathbb{E}[f^2(Y^*)1\{f<0\}] \\
&\qquad + \frac{16e^{-2}\log e}{(\theta-t)^2}\mathbb{E}\left[\exp\{\theta f(Y^*)\}1\{f\ge 0\}\right] && (172)\\
&\quad \le m_2 + \frac{e^{-2}m_1\log e}{(\theta-t)^2} && (173)\\
&\quad \le m_2 + \frac{4e^{-2}m_1\log e}{\theta^2} && (174)\\
&\quad \triangleq b(m_1,m_2,\theta)\cdot 2\log e\,. && (175)
\end{aligned}$$

Then, a simple estimate

$$\log\mathbb{E}[\exp\{tf(Y^*)\}] \le t\,\mathbb{E}[f(Y^*)] + bt^2, \qquad 0 \le t \le \frac{\theta}{2}, \tag{176}$$

can be obtained by taking the logarithm of the identity

$$\begin{aligned}
&\mathbb{E}[\exp\{tf(Y^*)\}] \\
&\quad = 1 + \frac{t}{\log e}\mathbb{E}[f(Y^*)] \\
&\quad + \frac{1}{\log^2 e}\int_0^t ds\int_0^s \mathbb{E}[f^2(Y^*)\exp\{tf(Y^*)\}]du && (177)
\end{aligned}$$

and invoking (175) and $\log x \le (x-1)\log e$.

Next, we define $F(y^n) = \frac{1}{n}\sum_{j=1}^n f(y_i)$ and consider the chain

$$\begin{aligned}
&t\mathbb{E}[F(Y^n)] \\
&\quad \le \log\mathbb{E}[\exp\{tF(Y^{*n})\}] + D(P_{Y^n}\|P^*_{Y^n}) && (178)\\
&\quad = n\log\mathbb{E}[\exp\{\frac{t}{n}f(Y^*)\}] + D(P_{Y^n}\|P^*_{Y^n}) && (179)\\
&\quad \le t\mathbb{E}[f(Y^*)] + \frac{bt^2}{n} + D(P_{Y^n}\|P^*_{Y^n}), && (180)
\end{aligned}$$

where (178)–(180) follow from (145), $P^*_{Y^n} = (P^*_Y)^n$ and (176) assuming $\frac{t}{n} \le \frac{\theta}{2}$. The proof concludes by letting $t = n^{\frac{3}{4}}$ in (180). ■

A natural extension of Proposition 12 to functions such as

$$F(y^n) = \frac{1}{n-r+1}\sum_{j=1}^{n-r+1} f(y_j^{j+r-1}) \tag{181}$$

is made by replacing the step (179) with an estimate

$$\log\mathbb{E}[\exp\{tF(Y^*)\}] \le \frac{n-r+1}{r}\log\mathbb{E}\left[\exp\left\{\frac{tr}{n}f(Y^{*r})\right\}\right], \tag{182}$$

which in turn is shown by splitting the sum into r subsums with independent terms and then applying Holder's inequality:

$$\mathbb{E}[X_1\cdots X_r] \le \left(\mathbb{E}[|X_1|^r]\cdots\mathbb{E}[|X_1|^r]\right)^{\frac{1}{r}}. \tag{183}$$

F. Functions of Degraded Channel Outputs

Notice that if the same code is used over a channel $Q_{Y|X}$ which is stochastically degraded with respect to $P_{Y|X}$ then by the data-processing for relative entropy, the upper bound (121) holds for $D(Q_{Y^n}\|Q^*_{Y^n})$, where Q_{Y^n} is the output of the $Q_{Y|X}$ channel and $Q^*_{Y^n}$ is the output of $Q_{Y|X}$ when the input is distributed according to a capacity-achieving distribution of $P_{Y|X}$. Thus, in all the discussions the pair $(P_{Y^n}, P^*_{Y^n})$ can be replaced with $(Q_{Y^n}, Q^*_{Y^n})$ without any change in arguments or constants. This observation can be useful in questions of information theoretic security, where the wiretapper has access to a degraded copy of the channel output.

G. Input Distribution: DMC

As shown in Section IV-A, we have for every ϵ-capacity-achieving code

$$\bar{P}_n = \frac{1}{n}\sum_{j=1}^n P_{Y_j} \to P^*_Y\,. \tag{184}$$

As noted in [3], convergence of output distributions can be propagated to statements about the input distributions. This is obvious for the case of a DMC with a nonsingular (more generally, injective) matrix $P_{Y|X}$. Even if the capacity-achieving input distribution is not unique, the following argument extends that of [3, Th. 4]. By Theorems 5 and 6, we know that

$$\frac{1}{n}I(X^n;Y^n) \to C\,. \tag{185}$$

Denote the single-letter empirical input distribution by $P_{\bar{X}} = \frac{1}{n}\sum_{j=1}^n P_{X_j}$. Naturally, $I(\bar{X};\bar{Y}) \le C$. However, in view of (185) and the concavity of mutual information, we must necessarily have

$$I(\bar{X};\bar{Y}) \to C\,. \tag{186}$$

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

$$d(P_{\bar{X}},\Pi) \to 0\,. \tag{187}$$

If the capacity achieving distribution P^*_X is unique, then (187) shows the convergence of $P_{\bar{X}} \to P^*_X$ in the (strong) sense of total variation.

H. Input Distribution: AWGN

In the case of the AWGN, just like in the discrete case, (50) implies that for any capacity achieving sequence of codes we have

$$P^{(n)}_{\bar{X}} = \frac{1}{n}\sum_{j=1}^n P_{X_j} \to wP^*_X \triangleq \mathcal{N}(0,P), \tag{188}$$

however, in the sense of weak convergence of distributions only. Indeed, the induced empirical output distributions satisfy

$$P^{(n)}_{\bar{Y}} = P^{(n)}_{\bar{X}} * \mathcal{N}(0,1), \tag{189}$$

where $*$ denotes convolution. By (50), (189) converges in relative entropy and thus weakly. Consequently, characteristic functions of $P^{(n)}_{\bar{Y}}$ converge pointwise to that of $\mathcal{N}(0,1+P)$. By dividing out the characteristic function of $\mathcal{N}(0,1)$ (which is strictly positive), so do characteristic functions of $P_{\bar{X}}^{(n)}$. Then, Levy's criterion establishes (188).

We now discuss whether (188) can be claimed in a stronger topology than the weak one. Since $P_{\bar{X}}$ is purely atomic and P_X^* is purely diffuse, we have

$$\|P_{\bar{X}} - P_X^*\|_{TV} = 1, \tag{190}$$

and convergence in total variation (let alone in relative entropy) cannot hold.

On the other hand, it is quite clear that the second moment of $\frac{1}{n}\sum P_{X_j}$ necessarily converges to that of $\mathcal{N}(0,P)$. Together weak convergence and control of second moments imply [31, (12), p. 7]

$$W_2^2\left(\frac{1}{n}\sum_{j=1}^n P_{X_j}, P_X^*\right) \to 0\,. \tag{191}$$

Therefore, (188) holds in the sense of topology metrized by the W_2-distance.

Note that convexity properties of $W_2^2(\cdot,\cdot)$ imply

$$\begin{aligned} W_2^2\left(\frac{1}{n}\sum_{j=1}^n P_{X_j}, P_X^*\right) &\le \frac{1}{n}\sum_{j=1}^n W_2^2\left(P_{X_j}, P_X^*\right) && (192)\\ &\le \frac{1}{n}W_2^2\left(P_{X^n}, P_{X^n}^*\right), && (193) \end{aligned}$$

where we denoted

$$P_{X^n}^* \triangleq (P_X^*)^n = \mathcal{N}(0, PI_n)\,. \tag{194}$$

Comparing (191) and (193), it is natural to conjecture a stronger result. For any capacity-achieving sequence of codes

$$\frac{1}{\sqrt{n}}W_2(P_{X^n}, P_{X^n}^*) \to 0\,. \tag{195}$$

Another reason to conjecture (195) arises from considering the behavior of Wasserstein distance under convolutions. Indeed from the T_2-transportation inequality (168) and the relative entropy bound (121) we have

$$\frac{1}{n}W_2^2(P_{X^n} * \mathcal{N}(0,I_n), P_{X^n}^* * \mathcal{N}(0,I_n)) \to 0, \tag{196}$$

since by definition

$$\begin{aligned} P_{Y^n} &= P_{X^n} * \mathcal{N}(0,I_n) && (197)\\ P_{Y^n}^* &= P_{X^n}^* * \mathcal{N}(0,I_n), && (198) \end{aligned}$$

where $*$ denotes convolution of distributions on $\mathbb{R}^n$. Trivially, for any probability measures P, Q, and $\mathcal{N}$ on $\mathbb{R}^n$ we have (e.g., [31, Proposition 7.17])

$$W_2(P * \mathcal{N}, Q * \mathcal{N}) \le W_2(P, Q)\,. \tag{199}$$

Thus, overall we have

$$W_2(P_{X^n} * \mathcal{N}(0,I_n), P_{X^n}^* * \mathcal{N}(0,I_n)) \le W_2(P_{X^n}, P_{X^n}^*)$$

and (195) would imply that the convolution with the Gaussian kernel is unable to significantly decrease W_2.

Despite the foregoing intuitive considerations, conjecture (195) is false. Indeed, define $D^*(M,n)$ to be the minimum achievable average square distortion among all vector quantizers of the memoryless Gaussian source $\mathcal{N}(0,P)$ for blocklength n and cardinality M. In other words,

$$D^*(M,n) = \frac{1}{n}\inf_Q W_2^2(P_{X^n}^*, Q), \tag{200}$$

where the infimum is over all probability measures Q supported on M equiprobable atoms in $\mathbb{R}^n$. The standard rate-distortion (converse) lower bound dictates

$$\frac{1}{n}\log M \ge \frac{1}{2}\log\frac{P}{D^*(M,n)} \tag{201}$$

and hence

$$\begin{aligned} W_2^2(P_{X^n}, P_{X^n}^*) &\ge nD^*(n,M) && (202)\\ &\ge nP\exp\left\{-\frac{2}{n}\log M\right\}, && (203) \end{aligned}$$

which shows that for any sequence of codes with $\log M_n = O(n)$, the normalized transportation distance stays strictly bounded away from zero

$$\liminf_{n\to\infty}\frac{1}{\sqrt{n}}W_2(P_{X^n}, P_{X^n}^*) > 0\,. \tag{204}$$

Nevertheless, assertion (188) may be strengthened in several ways. For example, it can be shown that quadratic-forms and ℓ_p-norms of codewords X^n from good codes have very similar values (in expectation) to $\approx \mathcal{N}(0, P\cdot\mathbf{I}_n)$. Full details are available in [34]. Here we only give two sample statements.

1) Let $\mathbf{A} = \{a_{i,j}\}_{i,j=1}^n$ be a symmetric matrix satisfying $-\mathbf{I}_n \le \mathbf{A} \le \mathbf{I}_n$. Then, for any $O(\sqrt{n})$-achieving code we have

$$\mathbb{E}\left[\sum_{i,j=1}^n a_{i,j}X_iX_j\right] = P\,\mathrm{tr}\,\mathbf{A} + O(n^{\frac{3}{4}}). \tag{205}$$

2) Various upper bounds for the ℓ_q-norms, $\|\mathbf{x}\|_q \triangleq \left(\sum_{j=1}^n |x_j|^q\right)^{\frac{1}{q}}$, of codewords of good codes are presented in Table I. A sample result, corresponding to $q=4$, is as follows. For any $(n,M,\epsilon)_{max,det}$-code for the AWGN(P) channel at least *half* of the codewords satisfy

$$\|\mathbf{x}\|_4^2 \le \frac{2}{b_1}\left(nC + b_2\sqrt{n} - \log\frac{M}{2}\right), \tag{206}$$

where C is the capacity of the channel and b_1, b_2 are some code-independent constants.

I. Extension to Other Channels: Tilting

Let us review the scheme of investigating functions of the output $F(Y^n)$ that was employed in this paper so far. First, an

TABLE I
AWGN CHANNEL: ℓ_q NORMS $\|\mathbf{x}\|_q$ OF CODEWORDS, $q \in [1, \infty]$

Code	$[1, 2]$	$(2, 4]$	$(4, \infty)$	∞
random Gaussian	$n^{\frac{1}{q}}$	$n^{\frac{1}{q}}$	$n^{\frac{1}{q}}$	$\sqrt{\log n}$
any $O(\log n)$-achieving	$n^{\frac{1}{q}}$	$n^{\frac{1}{q}}$	$n^{\frac{1}{q}} \log^{\frac{q-4}{2q}} n$	$\sqrt{\log n}$
any dispersion-achieving	$n^{\frac{1}{q}}$	$n^{\frac{1}{q}}$	$o(n^{\frac{1}{4}})$	$o(n^{\frac{1}{4}})$
any $O(\sqrt{n})$-achieving	$n^{\frac{1}{q}}$	$n^{\frac{1}{q}}$	$n^{\frac{1}{4}}$	$n^{\frac{1}{4}}$
any capacity-achieving	$n^{\frac{1}{q}}$	$o(n^{\frac{1}{2}})$	$o(n^{\frac{1}{2}})$	$o(n^{\frac{1}{2}})$
any code	$n^{\frac{1}{q}}$	$n^{\frac{1}{2}}$	$n^{\frac{1}{2}}$	$n^{\frac{1}{2}}$

inequality (121) was shown by verifying that $Q_Y = P^*_{Y^n}$ satisfies the conditions of Theorem 3. Then, an approximation of the form

$$F(Y^n) \approx \mathbb{E}[F(Y^n)] \approx \mathbb{E}[F(Y^{*n})] \tag{207}$$

follows by Propositions 9 and 11 *simultaneously* for all concentrated (e.g., Lipschitz) functions. In this way, all the channel-specific work is isolated in proving (121). On the other hand, verifying conditions of Theorem 3 for $Q_Y = P^*_{Y^n}$ may be quite challenging even for memoryless channels. In this section, we show how Theorem 3 can be used to show (207) for a given function F in the absence of the universal estimate in (121).

Let $P_{Y|X}: \mathcal{X} \to \mathcal{Y}$ be a random transformation, Y' distributed according to auxiliary distribution Q_Y and $F: \mathcal{Y} \to \mathbb{R}$ a function such that

$$Z_F = \log \mathbb{E}[\exp\{F(Y')\}] < \infty\,. \tag{208}$$

Let $Q_Y^{(F)}$ an F-tilting of Q_Y, namely

$$dQ_Y^{(F)} = \exp\{F - Z_F\} dQ_Y\,. \tag{209}$$

The core idea of our technique is that if F is sufficiently regular and Q_Y satisfies conditions of Theorem 3, then $Q_Y^{(F)}$ also does. Consequently, the expectation of F under P_Y (induced by the code) can be investigated in terms of the moment-generating function of F under Q_Y. For brevity, we only present a variance-based version (similar to Corollary 4).

Theorem 13: Let Q_Y and F be such that (208) holds and

$$S = \sup_x \mathrm{Var}\left[\log \frac{dP_{Y|X=x}}{Q_Y}(Y) \middle| X = x\right] < \infty, \tag{210}$$

$$S_F = \sup_x \mathrm{Var}[F(Y)|X = x]\,. \tag{211}$$

Then, there exists a constant $a = a(\epsilon, S) > 0$ such that for any $(M, \epsilon)_{max,det}$ code we have for all $0 \le t \le 1$

$$\begin{aligned} t\mathbb{E}[F(Y)] - \log \mathbb{E}[\exp\{tF(Y')\}] \\ \le D(P_{Y|X}\|Q_Y|P_X) - \log M + a\sqrt{S + t^2 S_F}\,. \end{aligned} \tag{212}$$

Proof: Note that since

$$\log \frac{dP_{Y|X}}{dQ_Y^{(F)}} = \log \frac{dP_{Y|X}}{dQ_Y} - F(Y) + Z_F \tag{213}$$

we have for any $0 \le t \le 1$:

$$\begin{aligned} D(P_{Y|X}\|Q_Y^{(tF)}|P_X) = D(P_{Y|X}\|Q_Y|P_X) \\ - t\,\mathbb{E}[F(Y)] + \log \mathbb{E}[\exp\{tF(Y')\}], \end{aligned} \tag{214}$$

and from (62)

$$\mathrm{Var}\left[\log \frac{dP_{Y|X}}{dQ_Y^{(F)}} \middle| X = x\right] \le 2(S + t^2 S_F)\,. \tag{215}$$

We conclude by invoking Corollary 4 with Q_Y and S replaced by $Q_Y^{(tF)}$ and $2S + 2t^2 S_F$, respectively. ■

For example, Corollary 10 is recovered from (212) by taking $Q_Y = P^*_{Y^n}$, applying (19), estimating the moment-generating function via (138) and bounding S_F via Poincaré inequality

$$S_F \le b\|F\|^2_{Lip}\,. \tag{216}$$

V. Binary Hypothesis Testing P_{Y^n} Versus $P^*_{Y^n}$

We now turn to the question of distinguishing P_{Y^n} from $P^*_{Y^n}$ in the sense of binary hypothesis testing. First, a simple data-processing reasoning yields for any $0 < \alpha \le 1$,

$$d(\alpha\|\beta_\alpha(P_{Y^n}, P^*_{Y^n})) \le D(P_{Y^n}\|P^*_{Y^n}), \tag{217}$$

where we have denoted the binary relative entropy

$$d(x\|y) \triangleq x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y}\,. \tag{218}$$

From (121) and (217), we conclude: every $(n, M, \epsilon)_{max,det}$ code must satisfy

$$\beta_\alpha(P_{Y^n}, P^*_{Y^n}) \ge \left(\frac{M}{2}\right)^{\frac{1}{\alpha}} \exp\left\{-n\frac{C}{\alpha} - \sqrt{n}\frac{a}{\alpha}\right\} \tag{219}$$

for all $0 < \alpha \le 1$. Therefore, in particular we see that the hypothesis testing problem for discriminating P_{Y^n} from $P^*_{Y^n}$ has zero Stein's exponent $-\frac{1}{n}\log \beta_\alpha(P_{Y^n}, P^*_{Y^n})$, provided that the sequence of $(n, M_n, \epsilon)_{max,det}$ codes with output distribution P_{Y^n}, is capacity achieving.

The main result in this section gives a better bound than (219).

Theorem 14: Consider one of the three types of channels introduced in Section II. Then, every $(n, M, \epsilon)_{avg}$ code must satisfy

$$\beta_\alpha(P_{Y^n}, P^*_{Y^n}) \ge M \exp\{-nC - a_2\sqrt{n}\} \qquad \epsilon \le \alpha \le 1, \tag{220}$$

where $a_2 = a_2(\epsilon, a_1) > 0$ depends only on ϵ and the constant a_1 from (24).

To prove Theorem 14, we introduce the following converse whose particular case $\alpha = 1$ is [4, Th. 27].

Theorem 15: Consider an $(M, \epsilon)_{avg}$ code for an arbitrary random transformation $P_{Y|X}$. Let P_X be equiprobable on the codebook $\mathcal{C}$ and P_Y be the induced output distribution. Then, for any Q_Y and $\epsilon \le \alpha \le 1$ we have

$$\beta_\alpha(P_Y, Q_Y) \ge M\beta_{\alpha-\epsilon}(P_{XY}, P_X Q_Y)\,. \tag{221}$$

If the code is $(M,\epsilon)_{max,det}$ then additionally

$$\beta_\alpha(P_Y,Q_Y) \ge \frac{\delta}{1-\alpha+\delta} M \inf_{x\in C} \beta_{\alpha-\epsilon-\delta}(P_{Y|X=x},Q_Y) \tag{222}$$

when $\epsilon+\delta\le\alpha\le 1$.

Proof: For a given $(M,\epsilon)_{avg}$ code, define

$$Z = 1\{\hat{W}(Y)=W, Y\in E\}, \tag{223}$$

where W is the message and E is an arbitrary event of the output space satisfying

$$P_Y[E] \ge \alpha\,. \tag{224}$$

As in the original metaconverse [4, Th. 26] the main idea is to use Z as a suboptimal hypothesis test for discriminating P_{XY} against $P_X Q_Y$. Following the same reasoning as in [4, Th. 27] one notices that

$$(P_X Q_Y)[Z=1] \le \frac{Q_Y[E]}{M} \tag{225}$$

and

$$P_{XY}[Z=1] \ge \alpha-\epsilon\,. \tag{226}$$

Therefore, by definition of β_α we must have

$$\beta_{\alpha-\epsilon}(P_{XY},P_X Q_Y) \le \frac{Q_Y[E]}{M}\,. \tag{227}$$

To complete the proof of (221) we take the infimum in (227) over all E satisfying (224).

To prove (222), we again consider any set E satisfying (224). Denote the codebook $\mathcal{C}=\{c_1,\dots,c_M\}$ and for $i=1,\dots,M$

$$p_i = P_{Y|X=c_i}[E] \tag{228}$$
$$q_i = Q_Y[\hat{W}=i,E]. \tag{229}$$

Since the sets $\{\hat{W}=i\}$ are disjoint, the (arithmetic) average of q_i is upper-bounded by

$$\mathbb{E}[q_W] \le \frac{1}{M} Q_Y[E], \tag{230}$$

whereas because of (224) we have

$$\mathbb{E}[p_W] \ge \alpha\,. \tag{231}$$

Thus, the following lower bound holds:

$$\mathbb{E}\left[\frac{Q_Y[E]}{M} p_W - \delta q_W\right] \ge \frac{Q_Y[E]}{M}(\alpha-\delta) \tag{232}$$

implying that there must exist $i\in\{1,\dots,M\}$ such that

$$\frac{Q_Y[E]}{M} p_i - \delta q_i \ge \frac{Q_Y[E]}{M}(\alpha-\delta)\,. \tag{233}$$

For such i we clearly have

$$P_{Y|X=c_i}[E] \ge \alpha-\delta \tag{234}$$

$$Q_Y[\hat{W}=i,E] \le \frac{Q_Y[E]}{M}\frac{1-\alpha-\delta}{\delta}\,. \tag{235}$$

By the maximal probability of error constraint, we deduce

$$P_{Y|X=c_i}[E,\hat{W}=i] \ge \alpha-\epsilon-\delta \tag{236}$$

and thus by the definition of β_α:

$$\beta_{\alpha-\epsilon-\delta}(P_{Y|X=c_i},Q_Y) \le \frac{Q_Y[E]}{M}\frac{1-\alpha-\delta}{\delta}\,. \tag{237}$$

Taking the infimum in (237) over all E satisfying (224) completes the proof of (222). ■

Proof of Theorem 14: To show (220), we first notice that as a consequence of (19), (24) and [4, Lemma 59] (see also [16, (2.71)]) we have for any $x^n \in \mathcal{X}_n$:

$$\beta_\alpha(P_{Y^n|X^n=x^n},P^*_{Y^n}) \ge \frac{\alpha}{2}\exp\left\{-nC-\sqrt{\frac{2a_1 n}{\alpha}}\right\}\,. \tag{238}$$

From [16, Lemma 32] and the fact that the function of α in the right-hand side of (238) is convex, we obtain that for any P_{X^n}

$$\beta_\alpha(P_{X^nY^n},P_{X^n}P^*_{Y^n}) \ge \frac{\alpha}{2}\exp\left\{-nC-\sqrt{\frac{2a_1 n}{\alpha}}\right\}\,. \tag{239}$$

Finally, (239) and (221) imply (220). ■

VI. AEP for the Output Process Y^n

Conventionally, we say that a sequence of distributions P_{Y^n} on $\mathcal{Y}^n$ (with $\mathcal{Y}$ a countable set) satisfies the asymptotic equipartition property (AEP) if

$$\frac{1}{n}\left|\log\frac{1}{P_{Y^n}(Y^n)} - H(Y^n)\right| \to 0 \tag{240}$$

in probability. In this section, we will take the AEP to mean convergence of (240) in the stronger sense of L_2, namely,

$$\mathrm{Var}[\log P_{Y^n}(Y^n)] = o(n^2), \qquad n\to\infty\,. \tag{241}$$

A. DMC

Although the sequence of output distributions induced by a code is far from being (a finite chunk of) a stationary ergodic process, we will show that (240) is satisfied for ϵ-capacity-achieving codes (and other codes). Thus, in particular, if the channel outputs are to be almost-losslessly compressed and stored for later decoding, $\frac{1}{n}H(Y^n)$ bits per sample would suffice [cf. (67)]. In fact, $\log\frac{1}{P_{Y^n}(Y^n)}$ 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 [35], [36].

Theorem 16: Consider a DMC $P_{Y|X}$ with $C_1<\infty$ (with or without input constraints) and a capacity achieving sequence of $(n,M_n,\epsilon)_{max,det}$ codes. Then, the output AEP (240) holds.

Proof: In the proof of Theorem 6, it was shown that $\log P_{Y^n}(y^n)$ is Lipschitz with Lipschitz constant upper bounded by a_1. Thus, by (141) and Proposition 11 we find that for any capacity-achieving sequence of codes (241) holds. ■

For many practically interesting DMCs (such as those with additive noise in a finite group), the estimate (241) can be improved to $O(n)$ even without assuming the code to be capacity-achieving. The proof of the following result is more circuitous than that of (60) which capitalizes on the conditional independence of the output given the input.

Theorem 17: Consider a DMC $P_{Y|X}$ with $C_1 < \infty$ (with or without input constraints) and such that $H(Y|X = x)$ is constant on $\mathcal{X}$. Then, for any sequence of $(n, M_n, \epsilon)_{max,det}$ codes there exists a constant $a = a(\epsilon)$ such that for all n sufficiently large

$$\mathrm{Var}\left[\log P_{Y^n}(Y^n)\right] \le an\,. \tag{242}$$

In particular, the output AEP (241) holds.

Proof: First, let X be a random variable and A some event (think $\mathbb{P}[A^c] \ll 1$) such that

$$|X - \mathbb{E}[X]| \le L \tag{243}$$

if $X \notin A$. Then, denoting $\mathrm{Var}[X|A] = \mathbb{E}[X^2|A] - \mathbb{E}^2[X|A]$,

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2 1_A] + \mathbb{E}[(X - \mathbb{E}[X])^2 1_{A^c}]\\ &\le \mathbb{E}[(X - \mathbb{E}[X])^2 1_A] + \mathbb{P}[A^c]L^2 &(244)\\ &= \mathbb{P}[A]\mathrm{Var}[X|A] + \mathbb{P}[A^c]L^2\\ &\quad + \frac{\mathbb{P}^2[A^c]}{\mathbb{P}[A]}(\mathbb{E}[X] - \mathbb{E}[X|A^c])^2 &(245)\\ &\le \mathrm{Var}[X|A] + \frac{\mathbb{P}[A^c]}{\mathbb{P}[A]}L^2, &(246)\end{aligned}$$

where (244) is by (243), (245) is because

$$\begin{aligned}&\mathbb{E}[(X - \mathbb{E}[X])^2|A]\\ &\quad = \mathrm{Var}[X|A] + (\mathbb{E}[X|A] - \mathbb{E}[X])^2 &(247)\\ &\quad = \mathrm{Var}[X|A] + \left(\frac{P[A^c]}{P[A]}\right)^2 (\mathbb{E}[X] - \mathbb{E}[X|A^c])^2 &(248)\end{aligned}$$

which in turn follows from identity

$$\mathbb{E}[X|A] = \frac{\mathbb{E}[X] - \mathbb{P}[A^c]\mathbb{E}[X|A^c]}{\mathbb{P}[A]} \tag{249}$$

and (246) is because (243) implies $|\mathbb{E}[X|A^c] - \mathbb{E}[X]| \le L$. Next, fix n and for any codeword $x^n \in \mathcal{X}_n$ denote for brevity

$$\begin{aligned}d(x^n) &= D(P_{Y^n|X^n=x^n} \| P_{Y^n}) &(250)\\ v(x^n) &= \mathbb{E}\left[\log \frac{1}{P_{Y^n}(Y^n)} \,\middle|\, X^n = x^n\right] &(251)\\ &= d(x^n) + H(Y^n|X^n = x^n)\,. &(252)\end{aligned}$$

If we could show that for some $a_1 > 0$

$$\mathrm{Var}[d(X^n)] \le a_1 n \tag{253}$$

the proof would be completed as follows:

$$\begin{aligned}&\mathrm{Var}\left[\log \frac{1}{P_{Y^n}(Y^n)}\right]\\ &\quad = \mathrm{Var}\left[\log \frac{1}{P_{Y^n}(Y^n)} \,\middle|\, X^n\right] + \mathrm{Var}[v(X^n)] &(254)\\ &\quad \le a_2 n + \mathrm{Var}[v(X^n)] &(255)\\ &\quad = a_2 n + \mathrm{Var}[d(X^n)] &(256)\\ &\quad \le (a_1 + a_2)n, &(257)\end{aligned}$$

where (255) follows for an appropriate constant $a_2 > 0$ from (60), (256) is by (252), and $H(Y^n|X^n = x^n)$ does not depend on x^n by assumption,[10] and (257) is by (253).

To show (253), first note the bound on the information density

$$\imath_{X^n;Y^n}(x^n; y^n) = \log \frac{P_{X^n|Y^n}(x^n|y^n)}{P_{X^n}(x^n)} \le \log M_n\,. \tag{258}$$

Second, as shown in (61) one may take $S_m = a_3 n$ in Corollary 4. In turn, this implies that one can take $\Delta = \sqrt{\frac{2a_3 n}{1-\epsilon}}$ and $\delta' = \frac{1-\epsilon}{2}$ in Theorem 3, that is

$$\inf_{x^n} \mathbb{P}\left[\log \frac{P_{Y^n|X^n=x^n}}{P_{Y^n}}(Y^n) < d(x^n) + \Delta \,\middle|\, X^n = x^n\right] \ge \frac{1+\epsilon}{2}. \tag{259}$$

Then, applying Theorem 2 with $\rho(x^n) = d(x^n) + \Delta$ to the $(M_n', \epsilon)_{max,det}$ subcode consisting of all codewords with $\{d(x^n) \le \log M_n - 2\Delta\}$ we get

$$\mathbb{P}[d(X^n) \le \log M_n - 2\Delta] \le \frac{2}{1-\epsilon}\exp\{-\Delta\}, \tag{260}$$

since $M_n' = M_n \mathbb{P}[d(X^n) \le \log M_n - 2\Delta]$ and

$$\mathbb{E}[\exp(\rho(X^n))|d(X^n) \le \log M_n - 2\Delta] \le M_n \exp(-\Delta)\,. \tag{261}$$

Now, we apply (246) to $d(X^n)$ with $L = \log M_n$ and $A = \{d(X^n) > \log M_n - 2\Delta\}$. Since $\mathrm{Var}[X|A] \le \Delta^2$ this yields

$$\mathrm{Var}[d(X^n)] \le \Delta^2 + \frac{2\log^2 M_n}{1-\epsilon}\exp\{-\Delta\} \tag{262}$$

for all n such that $\frac{2}{1-\epsilon}\exp\{-\Delta\} \le \frac{1}{2}$. Since $\Delta = O(\sqrt{n})$ and $\log M_n = O(n)$ we conclude from (262) that there must be a constant a_1 such that (253) holds. ■

B. AWGN

Following the argument of Theorem 17 step by step with (106) used in place of (60), we arrive at a similar AEP for the AWGN channel.

Theorem 18: Consider the *AWGN* channel. Then, for any sequence of $(n, M_n, \epsilon)_{max,det}$ codes there exists a constant $a = a(\epsilon)$ such that for all n sufficiently large

$$\mathrm{Var}\left[\log p_{Y^n}(Y^n)\right] \le an, \tag{263}$$

where p_{Y^n} is the density of Y^n.

[10] This argument also shows how to construct a counterexample when $H(Y|X = x)$ is nonconstant: merge two constant composition subcodes of types P_1 and P_2 such that $H(W|P_1) \neq H(W|P_2)$ where $W = P_{Y|X}$ is the channel matrix. In this case, one clearly has $\mathrm{Var}[\log P_{Y^n}(y^n)] \ge \mathrm{Var}[v(X^n)] = \mathrm{const}\cdot n^2$.

Corollary 19: If in the setting of Theorem 18, the codes are spherical (i.e., the energies of all codewords X^n are equal) or, more generally,

$$\mathrm{Var}[\|X^n\|^2] = o(n^2), \tag{264}$$

then

$$\frac{1}{n}\left|\log\frac{dP_{Y^n}}{dP^*_{Y^n}}(Y^n) - D(P_{Y^n}\|P^*_{Y^n})\right| \to 0 \tag{265}$$

in P_{Y^n}-probability.

Proof: To apply Chebyshev's inequality to $\log\frac{dP_{Y^n}}{dP^*_{Y^n}}(Y^n)$ we need, in addition to (263), to show

$$\mathrm{Var}[\log p^*_{Y^n}(Y^n)] = o(n^2), \tag{266}$$

where $p^*_{Y^n}(y^n) = (2\pi(1+P))^{-\frac{n}{2}}e^{-\frac{\|y^n\|^2}{2(1+P)}}$. Introducing i.i.d. $Z_j \sim \mathcal{N}(0,1)$ we have

$$\begin{aligned}&\mathrm{Var}[\log p^*_{Y^n}(Y^n)]\\ &\quad= \frac{\log^2 e}{4(1+P)^2}\mathrm{Var}\left[\|X^n\|^2 + 2\sum_{j=1}^{n} X_j Z_j + \|Z^n\|^2\right].\end{aligned} \tag{267}$$

The variances of the second and third terms are clearly $O(n)$, while the variance of the first term is $o(n^2)$ by assumption (264). Then, (267) implies (266) via (62). ■

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, Jul./Oct. 1948.

[2] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

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

[4] 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.

[5] 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.

[6] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Probability Theory Related Fields*, vol. 34, no. 2, pp. 157–177, 1976.

[7] R. Ahlswede and G. Dueck, "Every bad code has a good subcode: A local converse to the coding theorem," *Probability Theory Related Fields*, vol. 34, no. 2, pp. 179–182, 1976.

[8] K. Marton, "A simple proof of the blowing-up lemma," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 3, pp. 445–446, May 1986.

[9] Y. Polyanskiy and S. Verdú, "Relative entropy at the channel output of a capacity-achieving code," presented at the 49th Allerton Conf. Commun. Control Comput., Monticello, IL, USA, Oct. 2011.

[10] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Sci. Math. Hungar.*, vol. 2, pp. 291–292, 1967.

[11] J. H. B. Kemperman, "On the Shannon capacity of an arbitrary channel," *Indagationes Math.*, vol. 77, no. 2, pp. 101–115, 1974.

[12] U. Augustin, "Gedächtnisfreie kanäle für diskrete zeit," *Z. Wahrscheinlichkeitstheorie und Verw. Geb.*, vol. 6, pp. 10–61, 1966.

[13] R. Ahlswede, "An elementary proof of the strong converse theorem for the multiple-access channel," *J. Comb. Inf. Syst. Sci.*, vol. 7, no. 3, pp. 216–230, 1982.

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

[15] H. V. Poor and S. Verdú, "A lower bound on the error probability in multihypothesis testing," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1992–1993, Nov. 1995.

[16] Y. Polyanskiy, "Channel Coding: Non-Asymptotic Fundamental Limits," Ph.D. dissertation, Princeton Univ, Princeton, NJ, USA, 2010 [Online]. Available: http://people.lids.mit.edu/yp/homepage/

[17] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," *Probl. Peredachi Inf.*, vol. 12, no. 4, pp. 10–30, 1976.

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

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

[20] M. Raginsky and I. Sason, "Refined bounds on the empirical distribution of good channel codes via concentration inequalities," presented at the IEEE Int. Symp. Inf. Theory, Istanbul, Turkey, Jul. 2013.

[21] M. Ledoux, "Concentration of measure and logarithmic Sobolev inequalities," in *Seminaire de Probabilités de Strasbourg*, 1999, vol. 33, pp. 120–216.

[22] J. Wolfowitz, *Coding Theorems of Information Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1962.

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

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

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

[26] L. Harper, "Optimal numberings and isoperimetric problems on graphs," *J. Combin. Theory*, vol. 1, pp. 385–394, 1966.

[27] M. Talagrand, "An isoperimetric theorem on the cube and the Khintchine-Kahane inequalities," in *Amer. Math. Soc.*, 1988, vol. 104, pp. 905–909.

[28] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time. I. II," *Commun. Pure Appl. Math.*, vol. 28, no. 1, pp. 1–47, 1975.

[29] K. Marton, "Bounding $\bar{d}$-distance by information divergence: A method to prove measure concentration," *Ann. Probab.*, vol. 24, pp. 857–866, 1990.

[30] R. L. Dobrushin, "Prescribing a system of random variables by conditional distributions," *Theory Probability Appl.*, vol. 15, no. 3, pp. 458–486, 1970.

[31] C. Villani, *Topics in Optimal Transportation*. Providence, RI, USA: American Mathematical Society, 2003, vol. 58.

[32] S. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic Sobolev inequalities," *J. Funct. Anal.*, vol. 163, pp. 1–28, 1999.

[33] M. Talagrand, "Transportation cost for Gaussian and other product measures," *Geom. Funct. Anal.*, vol. 6, no. 3, pp. 587–600, 1996.

[34] Y. Polyanskiy, "ℓ_p-norms of codewords from capacity and dispersion-achieving Gaussian codes," presented at the 50th Allerton Conf., Monticello, IL, USA, Oct. 2012.

[35] S. Verdú and T. S. Han, "The role of the asymptotic equipartition property in noiseless source coding," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 847–857, May 1997.

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

Yury Polyanskiy (S'08–M'10) is an Assistant Professor of Electrical Engineering and Computer Science at MIT. He received the M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia in 2005 and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2010. In 2000–2005 he lead the development of the embedded software in the Department of Surface Oilfield Equipment, Borets Company LLC (Moscow). Currently, his research focuses on basic questions in information theory, error-correcting codes, wireless communication and fault-tolerant circuits. Over the years Dr. Polyanskiy won the 2013 NSF CAREER award, the 2011 IEEE Information Theory Society Paper Award and Best Student Paper Awards at the 2008 and 2010 IEEE International Symposia on Information Theory (ISIT).

Sergio Verdú (S'80–M'84–SM'88–F'93) is the Eugene Higgins Professor of Electrical Engineering at Princeton University. A member of the National Academy of Engineering, Verdú is the recipient of the 2007 Claude E. Shannon Award, and the 2008 IEEE Richard W. Hamming Medal.