

The Empirical Distribution of Good Codes

Shlomo Shamai (Shitz), *Fellow, IEEE*, and Sergio Verdú, *Fellow, IEEE*

Abstract—Let the k th-order empirical distribution of a code be defined as the proportion of k -strings anywhere in the codebook equal to every given k -string. We show that for any fixed k , the k th-order empirical distribution of any good code (i.e., a code approaching capacity with vanishing probability of error) converges in the sense of divergence to the set of input distributions that maximize the input/output mutual information of k channel uses. This statement is proved for discrete memoryless channels as well as a large class of channels with memory. If k grows logarithmically (or faster) with blocklength, the result no longer holds for certain good codes, whereas for other good codes, the result can be shown for k growing as fast as a certain fraction of blocklength.

Index Terms—Approximation of output statistics, channel capacity, discrete memoryless channels, divergence, error-correcting codes, Gaussian channels, Shannon theory.

I. INTRODUCTION

FINDING the input distribution that maximizes mutual information leads not only to the capacity of the channel, but to engineering insights on the behavior of *good* codes (approaching capacity with vanishing error probability). For example, it is widely accepted that in order to approach the capacity of a nonwhite Gaussian channel, a good code must be such that the channel input resembles a Gaussian process with spectral density “close” to the water-filling capacity-achieving solution.

Consider the easier special case of the binary-symmetric channel (BSC). The unique n -dimensional distribution that maximizes the n -block input-output mutual information of a BSC puts equal mass on all 2^n binary n -strings. Common wisdom in information theory indicates that a good code for the BSC must contain asymptotically equal proportion of 0's and 1's, and more generally, that all strings of a given length k occur asymptotically in the same proportion throughout the codebook. This paper formalizes and proves that statement. One may be tempted to carry these expectations further and hope that the ensemble of the equiprobable codewords of a good code for the BSC must appear to be generated by a source of independent equally likely bits. However, the entropy of the capacity-achieving n -dimensional distribution is equal to n bits, whereas the entropy of the codeword at the input of

the channel is (at most) equal to the logarithm of the number of codewords, which is equal to n times the rate of the code. Thus unless the BSC is noiseless, there is no hope that as $n \rightarrow \infty$, the channel input process may converge to a source of pure bits in any reasonable sense.

A good deal of the intuition on which the above common wisdom is grounded arises from the consideration of the input distributions of *random coding*, where not only do we average over equiprobable codewords, but over codebooks generated randomly according to the distribution maximizing mutual information. Then, the averaged input distributions of a random code are trivially equal to the capacity-achieving input distributions. However, as we just saw in the case of the BSC, the behavior of the averaged empirical distribution of random codes is quite different from the empirical distribution of good codes. Thus in this context, random coding reasonings not only lead to trivial results but may actually be misleading.

The fact that there exist good codes for discrete memoryless channels whose empirical distributions converge to the capacity-achieving distributions follows immediately from the optimality of constant-composition codes [1]. However, the present paper proves that such convergence must hold for *any good code*, thereby confirming an assertion in [2]. This proof provides a solid justification for eliminating from the search for good codes any codebooks whose empirical distributions are not close to the capacity-achieving distributions.

In Section II, we give the main definitions including that of the k th-order empirical distribution: the proportion of k -strings anywhere in the codebook equal to every given k -string. We show that the n th-order empirical distribution attains capacity asymptotically. A general result on the interplay between the deficit from maximal mutual information and the divergence of output distributions leads to a generalization of a key result of [3]: the *output* distribution induced by any good code sequence converges (in normalized divergence) to the (unique) output distribution induced by a capacity-achieving input distribution. In certain cases (such as discrete memoryless channels with full-rank transition matrices [4]), such a result implies convergence of the input statistics. However, in general, such convergence does not follow directly from the convergence of output statistics.

Section III deals with discrete memoryless channels. It shows that if the channel is such that mutual information is maximized by a unique distribution, then (for any k) the k th-order empirical distribution converges (in divergence) to the k -product of that capacity-achieving distribution. If the maximal mutual-information distribution is not unique, then the k th-order empirical distribution need not converge, or even when it does converge, it may converge to a nonproduct distribution.

Manuscript received November 5, 1995; revised July 20, 1996. This work was supported in part by the U.S.–Israel Binational Science Foundation under Grant 92-000202-2. The material in this paper was presented in part at the 1995 IEEE International Symposium on Information Theory, Whistler, BC, Canada, September 18–22, 1995.

S. Shamai (Shitz) is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel 32000.

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

Publisher Item Identifier S 0018-9448(97)02329-8.

In any case, it is shown that the “distance” from the k th-order empirical distribution to the set of distributions that maximize the k -mutual information vanishes asymptotically. In addition, we consider the case where instead of being fixed, the empirical distribution dimension, k , grows with the blocklength n . We show that exceeding a certain linear growth surely prevents the $k(n)$ th-order empirical distribution from converging, whereas at slower growths of $k(n)$, we construct a code (operating arbitrarily close to capacity) for which approximation of the capacity-achieving distribution occurs. In addition, we construct a code for which the $O(\log n)$ th-order empirical distribution does not converge.

Section IV deals with channels with memory, and gives an approximation result which generalizes the result proved in Section III for discrete memoryless channels with unique capacity-achieving input distributions.

Section V deals with additive-noise channels. In that case, the divergence approximation measure is not useful. However, convergence in distribution is particularly easy to show in the case of memoryless channels. In the case of nonwhite Gaussian channels, we show that the Ornstein distance between the k th-order empirical distribution and the k th Gaussian variate obtained from the water-filling solution vanishes with blocklength.

II. PRELIMINARIES

This section gives the main definitions and notation along with general results on the asymptotic maximization of mutual information by empirical distributions and on the approximation (in the sense of divergence) of output distributions induced by good codes. We consider channels with input alphabet A and output alphabet B . The random transformation operating on n -tuples is denoted by $W^{(n)}: A^n \rightarrow B^n$.

A. Good Codes

In order to formalize the idea of codes whose rate approaches capacity and whose error probability vanishes, we introduce the following definition.

Definition 1: A *good code-sequence* for a channel with capacity C is a sequence of codes with vanishing error probability whose rate satisfies

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log M = C. \quad (1)$$

Note that the existence of good code-sequences is guaranteed by the definition of channel capacity [1].

B. Empirical Distributions

In this subsection we define the empirical distributions of any code composed of M codewords of blocklength n

$$\{z_{im} \in A \quad i = 1, \dots, n \quad m = 1, \dots, M\}. \quad (2)$$

If the M codewords are equiprobable, then the distribution of the n -tuple input to the channel is defined on A^n as

$$P_{\hat{X}^n}(a_1, \dots, a_n) = \frac{1}{M} \sum_{m=1}^M \prod_{i=1}^n 1\{z_{im} = a_i\}, \quad (3)$$

In other words, $P_{\hat{X}^n}$ puts mass m/M on (a_1, \dots, a_n) if there are m codewords equal to (a_1, \dots, a_n) .

In addition to the joint distribution of the whole n -tuple, it is interesting to deal with the joint distribution of subcodewords $(a_j, \dots, a_l), 1 \leq j \leq l \leq n$

$$P_{\hat{X}_j^l}(a_j, \dots, a_l) = \frac{1}{M} \sum_{m=1}^M \prod_{i=j}^l 1\{z_{im} = a_i\} \quad (4)$$

which can be obtained from $P_{\hat{X}^n}$ by summing out all the components outside (a_j, \dots, a_l) . Note that $\hat{X}^n = \hat{X}_1^n$.

In addition to averaging over equiprobable codewords, it is important (when dealing with stationary channels) to define k th-order empirical distributions averaged over time. For example, for every codeword we can find its first-order empirical distribution by computing the fraction of symbols in the codeword equal to each input letter. Averaging those empirical distributions over equiprobable codewords we obtain the *first-order empirical distribution of the code*. Analogously, the k th-order empirical distribution can be defined by computing for each k -string $(\alpha_1, \dots, \alpha_k) \in A^k$ the fraction of k -strings anywhere within the codeword equal to $(\alpha_1, \dots, \alpha_k)$. Again, averaging over equiprobable codewords results in the k th-order empirical distribution of the code. Naturally, the order of averaging over time and codewords can be interchanged and we can give the following definition.

Definition 2: The *k th-order empirical distribution* ($1 \leq k \leq n$) of the code

$$\{z_{im}, i = 1, \dots, n, m = 1, \dots, M\}$$

is

$$\begin{aligned} Q_{\hat{X}^n}^{(k)}(\alpha_1, \dots, \alpha_k) &= \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} P_{\hat{X}_i^{i+k-1}}(\alpha_1, \dots, \alpha_k) \\ &= \frac{1}{n-k+1} \frac{1}{M} \sum_{i=1}^{n-k+1} \\ &\quad \cdot \sum_{m=1}^M 1\{z_{im} = \alpha_1\} \cdots 1\{z_{i+k-1,m} = \alpha_k\}. \end{aligned} \quad (5)$$

When a code with empirical distribution $P_{\hat{X}^n}$ is input to a random transformation $W^{(n)}: A^n \rightarrow B^n$, the output distribution induced on B^n is denoted by $P_{\hat{Y}^n}$. Analogously to (5), from $P_{\hat{Y}^n}$ we can define the k th-order empirical output distribution induced by the code:

$$Q_{\hat{Y}^n}^{(k)} = \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} P_{\hat{Y}_i^{i+k-1}} \quad (6)$$

where $P_{\hat{Y}_j^l}$ is obtained from $P_{\hat{Y}^n}$ by integrating out those components preceding j and succeeding l .

C. Asymptotic Maximization of Mutual Information

We can readily prove that the empirical distribution $P_{\hat{X}^n}$ of a good code sequence maximizes mutual information asymptotically by using the same reasoning that leads to the converse

coding theorem. This can be done in full generality (ruling out channels whose capacity is not given by the limit of maximal mutual informations (cf. [5]) as we can see in the following result (implicit in [3]):

Theorem 1: Suppose that the channel is such that its capacity satisfies

$$C = \lim_{n \rightarrow \infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n). \quad (7)$$

Then, the empirical distribution of any good code-sequence satisfies

$$\lim_{n \rightarrow \infty} \frac{1}{n} I(\hat{X}^n; \hat{Y}^n) = C. \quad (8)$$

Proof: By assumption (7) we have

$$\overline{\lim}_{n \rightarrow \infty} \frac{1}{n} I(\hat{X}^n; \hat{Y}^n) \leq C. \quad (9)$$

On the other hand, Fano's inequality implies that if S and R denote the message transmitted and decoded, respectively, by an (n, M, λ_n) code,¹ then

$$H(S|R) \leq \lambda_n \log M + \log 2.$$

The distribution at the output of the encoder when driven by equiprobable S is $P_{\hat{X}^n}$, whereas the distribution at the input of the decoder is $P_{\hat{Y}^n}$. Thus the data-processing lemma implies that

$$\begin{aligned} I(\hat{X}^n; \hat{Y}^n) &\geq I(S; R) \\ &= \log M - H(S|R) \\ &\geq (1 - \lambda_n) \log M - \log 2. \end{aligned}$$

Using the definition of good code-sequence we get

$$\underline{\lim}_{n \rightarrow \infty} \frac{1}{n} I(\hat{X}^n; \hat{Y}^n) \geq C$$

which together with (9), completes the proof. \square

It is straightforward to generalize Theorem 1 to channels with cost constraints such that given the cost functions

$$d_n: A^n \rightarrow [0, +\infty)$$

only those codebooks satisfying

$$\frac{1}{M} \sum_{m=1}^M d_n(z_{1m}, \dots, z_{nm}) \leq \beta \quad (10)$$

are allowed. The conclusion is that the empirical distribution of any good code sequence satisfying (10) must achieve the capacity-cost function asymptotically

$$\lim_{n \rightarrow \infty} \frac{1}{n} I(\hat{X}^n; \hat{Y}^n) = C(\beta) \quad (11)$$

provided

$$C(\beta) = \lim_{n \rightarrow \infty} \sup_{X^n: E[d_n(X^n)] \leq \beta} \frac{1}{n} I(X^n; Y^n). \quad (12)$$

It may seem that Theorem 1 is close to achieving our objective, namely, showing that empirical distributions of good

¹In the usual notation n stands for blocklength, M for the size of the code, and the third argument is an upper bound to the error probability.

codes converge to the capacity-achieving input distributions. However, the geometry of the mutual information as a function of the input n -variate distribution is working against that goal. Even if that function is strictly concave (with a unique maximum) for each n , its peak becomes increasingly flatter with n . In fact, as we saw in the case of a binary-symmetric channel the n -dimensional empirical distribution \hat{X}^n of any good code is quite far from the capacity-achieving distribution. Fortunately, the behavior of the fixed-length empirical distribution introduced in Definition 2 will be shown to satisfy our objective.

D. Output Approximation

Han and Verdú [3] show that for channels with finite input-alphabet the empirical output distribution converges to the maximal mutual-information output distribution in normalized divergence. This output approximation result is crucial for the remainder of this paper. In this subsection, we will give a proof that holds in complete generality for both discrete and continuous channels.

The modern tools based on the interplay between mutual information and conditional divergence play a central role in the technical development. If X and Y are connected by a random transformation W , i.e., $W(\cdot|a)$ is the probability measure $P_{Y|X=a}$, and the probability measure Q is defined on the same space as Y , denote the conditional divergence

$$D(W||Q|P_X) = \iint \log \frac{dW(b|a)}{dQ} dW(b|a) dP_X(a).$$

The following well-known fact will be used repeatedly in the sequel. If $P_Y \ll Q$, then

$$I(X; Y) = D(W||Q|P_X) - D(P_Y||Q) \quad (13)$$

which in the special case $Q = P_Y$ results in

$$I(X; Y) = D(W||P_Y|P_X). \quad (14)$$

Lemma 1: Consider measurable spaces (A, \mathcal{F}) and (B, \mathcal{G}) and an arbitrary Markov kernel $W: A \rightarrow B$, which defines a probability measure on (B, \mathcal{G}) for every $a \in A$, and a measurable function on (A, \mathcal{F}) for every member of \mathcal{G} . Let \mathcal{D} be a convex set of probability distributions on (A, \mathcal{F}) . Let $P_{\bar{X}} \in \mathcal{D}$ be such that

$$I(\bar{X}; \bar{Y}) = \max_{P_X \in \mathcal{D}} I(X; Y) < \infty. \quad (15)$$

Then, for all $P_X \in \mathcal{D}$,

$$a) \quad P_Y \ll P_{\bar{Y}} \quad (16)$$

$$b) \quad I(\bar{X}; \bar{Y}) - I(X; Y) \geq D(P_Y||P_{\bar{Y}}) \quad (17)$$

c) If $P_X \ll P_{\bar{X}}$, then (17) holds with equality

d) If X achieves $I(X; Y) = I(\bar{X}; \bar{Y})$, then $P_Y = P_{\bar{Y}}$.

Proof:

a) Let us assume that for some $P_{X^*} \in \mathcal{D}$, P_{Y^*} is not absolutely continuous with respect to $P_{\bar{Y}}$, i.e., there exists $G \in \mathcal{G}$ such that

$$P_{\bar{Y}}(G) = 0 < P_{Y^*}(G). \quad (18)$$

The mixture $P_{X_\alpha} = (1 - \alpha)P_{\bar{X}} + \alpha P_{X^*}$ attains the following mutual information:

$$I(X_\alpha; Y_\alpha) = D(W \| P_{Y_\alpha} | P_{X_\alpha}) \quad (19)$$

$$= (1 - \alpha)D(W \| P_{Y_\alpha} | P_{\bar{X}}) + \alpha D(W \| P_{Y_\alpha} | P_{X^*}) \quad (20)$$

$$= (1 - \alpha)D(W \| P_{\bar{Y}} | P_{\bar{X}}) + (1 - \alpha)D(P_{\bar{Y}} \| P_{Y_\alpha}) + \alpha D(W \| P_{Y_\alpha} | P_{X^*}) \quad (21)$$

$$\geq (1 - \alpha)I(\bar{X}; \bar{Y}) + (1 - \alpha)D(P_{\bar{Y}} \| P_{Y_\alpha}) + \alpha D(P_{Y^*} \| P_{Y_\alpha}) \quad (22)$$

$$\geq I(\bar{X}; \bar{Y}) + \alpha [D(P_{Y^*} \| P_{Y_\alpha}) - I(\bar{X}; \bar{Y})] \quad (23)$$

where (20) follows from the definition of conditional divergence; (21) follows from (13) and (14); (22) follows from (14) and the data-processing theorem for divergences [1]; and (23) follows from the positivity of the middle term in (22).

We will now use (23) to contradict (15) by showing that $D(P_{Y^*} \| P_{Y_\alpha})$ can be made as large as desired by appropriate choice of $0 < \alpha < 1$.

By definition of divergence as supremum over partitions [6]

$$D(P_{Y^*} \| P_{Y_\alpha}) \geq P_{Y^*}(G) \log \frac{P_{Y^*}(G)}{P_{Y_\alpha}(G)} + (1 - P_{Y^*}(G)) \log \frac{1 - P_{Y^*}(G)}{1 - P_{Y_\alpha}(G)} \quad (24)$$

but since $P_{Y_\alpha}(G) = \alpha P_{Y^*}(G)$ (cf. (18)), the right-hand side of (24) goes to $+\infty$ as $\alpha \downarrow 0$.

b) For any P_X we can write

$$I(\bar{X}; \bar{Y}) - I(X; Y) - D(P_Y \| P_{\bar{Y}}) = D(W \| P_{\bar{Y}} | P_{\bar{X}}) - D(W \| P_Y | P_X) - D(P_Y \| P_{\bar{Y}}) \quad (25)$$

$$= D(W \| P_{\bar{Y}} | P_{\bar{X}}) - D(W \| P_{\bar{Y}} | P_X) \quad (26)$$

where (25) follows from (14) and (26) follows from (13) and (16). To contradict the equation we want to show (17), let us assume now that there exists $P_{X^*} \in \mathcal{D}$ such that

$$D(W \| P_{\bar{Y}} | P_{\bar{X}}) < D(W \| P_{\bar{Y}} | P_{X^*}) \quad (27)$$

and construct the mixture

$$P_{X_\alpha} = (1 - \alpha)P_{\bar{X}} + \alpha P_{X^*} \quad (28)$$

which achieves

$$I(X_\alpha; Y_\alpha) = D(W \| P_{\bar{Y}} | P_{X_\alpha}) - D(P_{Y_\alpha} \| P_{\bar{Y}}) = (1 - \alpha)I(\bar{X}; \bar{Y}) + \alpha D(W \| P_{\bar{Y}} | P_{X^*}) - D(P_{Y_\alpha} \| P_{\bar{Y}}), \quad (29)$$

At least for small α , $I(X_\alpha; Y_\alpha)$ is strictly larger than $I(\bar{X}; \bar{Y})$ thereby contradicting the optimality of \bar{X} . To see this note that for $\alpha = 0$, $I(X_\alpha; Y_\alpha) = I(\bar{X}; \bar{Y})$, with derivative

$$\frac{d}{d\alpha} I(X_\alpha; Y_\alpha) = [D(W \| P_{\bar{Y}} | P_{X^*}) - I(\bar{X}; \bar{Y})] - \frac{d}{d\alpha} D(P_{Y_\alpha} \| P_{\bar{Y}}) \quad (30)$$

which is strictly positive at $\alpha = 0$, because of (27) and the following result due to Csiszár [7] particularized to $P = P_{Y^*}$, $Q = P_{\bar{Y}}$:

Lemma 2: Let $P \ll Q$, then

$$\frac{d}{d\alpha} D(\alpha P + (1 - \alpha)Q \| Q)|_{\alpha=0} = 0. \quad (31)$$

c) Define the random variable $f : (A, \mathcal{F}) \rightarrow (\mathcal{R}, \mathcal{B})$

$$f(a) = D(W(\cdot | a) \| P_{\bar{Y}}).$$

According to (26) we need to show that if $P_X \ll P_{\bar{X}}$, then

$$\int f dP_{\bar{X}} = \int f dP_X \quad (32)$$

but we have already shown in b) that

$$\int f dP_{\bar{X}} \geq \int f dP_X \quad (33)$$

for all P_X (regardless of whether it is absolutely continuous with respect to $P_{\bar{X}}$ or not). This requires that $P_{\bar{X}}$ put all its mass on a subset of $f^{-1}(I(\bar{X}; \bar{Y}))$. And the same must be true for any $P_X \ll P_{\bar{X}}$, thereby establishing (32).

d) Follows immediately from b). \square

In the particular context of discrete memoryless channels, properties a) and d) of Lemma 1 are known (cf. [8, pp. 95–96]).

The convergence of empirical output distributions (in normalized divergence) now follows directly from Theorem 1 and Lemma 1 (cf. [3, Theorem 15] in the special case of finite-input channels).

Theorem 2: Under the assumption of Theorem 1, the empirical output distribution converges to the maximal-mutual-information output distribution:

$$\lim_{n \rightarrow \infty} \frac{1}{n} D(P_{\hat{Y}^n} \| P_{\bar{Y}^n}) = 0 \quad (34)$$

where \bar{Y}^n is the unique output distribution such that for some $\bar{X}^{n,2}$

$$I(\bar{X}^n; \bar{Y}^n) = \max_{\bar{X}^n} I(X^n; Y^n).$$

²To express the theorem in complete generality, it is not necessary that the maximum in (35) be attained, only that the sequence $\{\bar{X}^n\}$ achieves C asymptotically.

III. DISCRETE MEMORYLESS CHANNELS

In this section we assume that the input and output alphabets A and B , respectively, are finite and that

$$P_{Y^n|X^n}(b_1, \dots, b_n|a_1, \dots, a_n) = \prod_{i=1}^n W(b_i|a_i)$$

where W is a stochastic matrix.

It is well known (cf. Lemma 1d) that for every channel matrix W , there exists a unique output distribution $P_{\bar{Y}}$ on B such that if

$$I(\tilde{X}^n; \tilde{Y}^n) = \max_{X^n} I(X^n; Y^n) \quad (35)$$

then

$$P_{\tilde{Y}^n} = P_{\bar{Y}} \times \dots \times P_{\bar{Y}}. \quad (36)$$

On the other hand, we should keep in mind that for some channels, the distribution that maximizes (cf. [1] for notation)

$$I(P, W) = \sum_{a \in A} \sum_{b \in B} P(a) W(b|a) \log \frac{W(b|a)}{\sum_{d \in A} W(b|d) P(d)}$$

need not be unique. If that distribution is not unique, say both P_{X_1} and P_{X_2} achieve $\max_P I(P, W)$, then $I(X^n; Y^n)$ is maximized not only by the product distributions

$$P_{X_1} \times \dots \times P_{X_1} \quad \text{and} \quad P_{X_2} \times \dots \times P_{X_2}$$

but also by nonproduct distributions such as

$$\begin{aligned} & \frac{1}{2} P_{X_1} \times P_{X_2} \times P_{X_1} \cdots \times P_{X_2} \times P_{X_1} \\ & + \frac{1}{2} P_{X_2} \times P_{X_1} \times P_{X_2} \cdots \times P_{X_1} \times P_{X_2} \end{aligned}$$

because of the concavity of $I(P, W)$.

We first illustrate that for even the simplest channels the empirical distribution of good codes does not satisfy the kind of approximation shown for the output empirical distribution $P_{\bar{Y}^n}$ in Theorem 2.

Example 1: Consider a BSC with capacity $C < 1$. Denote the unique maximal-mutual-information input distribution by $P_{\bar{X}} = P_{\bar{X}} \times \dots \times P_{\bar{X}}$, where $P_{\bar{X}}$ is equally likely 0 or 1. Then, the empirical input distribution $P_{\hat{X}^n}$ of every (n, M, λ_n) -code satisfies

$$\frac{1}{n} D(P_{\hat{X}^n} \| P_{\bar{X}}) \geq 1 - \frac{\log M}{n} \text{ bit}. \quad (37)$$

Thus $(1/n)D(P_{\hat{X}^n} \| P_{\bar{X}})$ cannot converge to 0 unless the rate of the code exceeds capacity.

Let us turn our attention to finite-order empirical distributions, for which we will be able to prove the corresponding input approximation result.

The following result shows output approximation for finite-order empirical distributions.

Theorem 3: Consider an arbitrary discrete memoryless channel. For every k , the k th-order output empirical distribution (cf. (6)) of a good code satisfies

$$\lim_{n \rightarrow \infty} D(Q_{\hat{Y}^n}^{(k)} \| P_{\bar{Y}} \times \dots \times P_{\bar{Y}}) = 0 \quad (38)$$

where the second argument denotes the k -product of the unique maximal-mutual-information output distribution.

Proof: To show (38) we will block the codeword indices into subblocks each of which have k components or fewer.

$$1, \dots, l; l+1, \dots, l+k; l+k+1, \dots, l+2k; \dots; l+r k+1, \dots, n$$

where $l = 0, \dots, k-1$ and $r = \lfloor n/k \rfloor$. Since the second argument in the divergence of (38) is a product distribution, the following inequality holds [9]:

$$\begin{aligned} & D(P_{\hat{Y}^n} \| P_{\bar{Y}} \times \dots \times P_{\bar{Y}}) \\ & \geq D(P_{\hat{Y}^l} \| P_{\bar{Y}} \times \dots \times P_{\bar{Y}}) \\ & \quad + \sum_{q=0}^{r-1} D(P_{\hat{Y}_{qk+l+1}^{qk+l+k}} \| P_{\bar{Y}} \times \dots \times P_{\bar{Y}}) \\ & \quad + D(P_{\hat{Y}_{rk+l+1}^n} \| P_{\bar{Y}} \times \dots \times P_{\bar{Y}}) \end{aligned} \quad (39)$$

where each product distribution has a multiplicity dictated by the first argument of the corresponding divergence. If we sum the k inequalities (39) parametrized by $l = 0, \dots, k-1$ and we drop from consideration those divergences between distributions of fewer than k random variables we get

$$k D(P_{\hat{Y}^n} \| P_{\bar{Y}}) \geq \sum_{i=1}^{n-k+1} D(P_{\hat{Y}_i^{i+k-1}} \| P_{\bar{Y}} \times \dots \times P_{\bar{Y}}). \quad (40)$$

Dividing both sides by $n+k-1$ and using (6) along with the convexity of divergence we get

$$D(Q_{\hat{Y}^n}^{(k)} \| P_{\bar{Y}} \times \dots \times P_{\bar{Y}}) \leq \frac{k}{n-k+1} D(P_{\hat{Y}^n} \| P_{\bar{Y}}). \quad (41)$$

But the right side of (41) vanishes because Theorem 2 applies to any discrete memoryless (stationary) channel.

The main goal of this section is to obtain a result parallel to Theorem 3 for the k th-order empirical *input* distribution. Before stating and proving our main result, we will illustrate some of the pitfalls of the input approximation problem with some examples.

Example 2: Consider the discrete memoryless deterministic channel of Fig. 1. Any input distribution on $\{a, b, c, d\}$ which places mass $\frac{1}{2}$ on the subsets $\{a, b\}$ and $\{c, d\}$ maximizes mutual information. The nonuniqueness of the optimal input distribution leads to the existence of good code sequences with unexpected behavior. For example, construct a $(2n, 4^n, 0)$ code consisting of all sequences such that at odd times $\{a, c\}$ are forbidden and at even times $\{b, d\}$ are forbidden. Clearly, this is a good code sequence since it achieves capacity (equal to 1 bit) with zero error probability. Its second-order empirical distribution $Q_{\hat{X}^n}^{(2)}(w, z)$ is equal to $\frac{1}{8}$ if

$$(w, z) \in \{(a, b), (a, d), (b, a), (b, c), (c, d), (c, b), (d, a), (d, c)\}$$

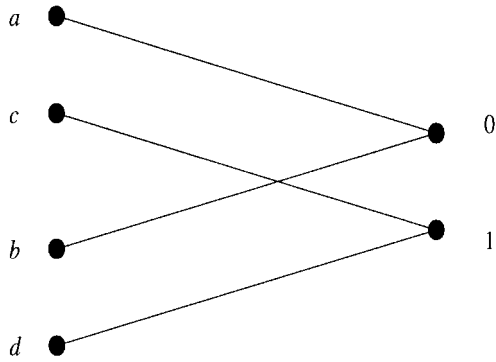


Fig. 1. Discrete memoryless channel with nonunique optimal input distribution.

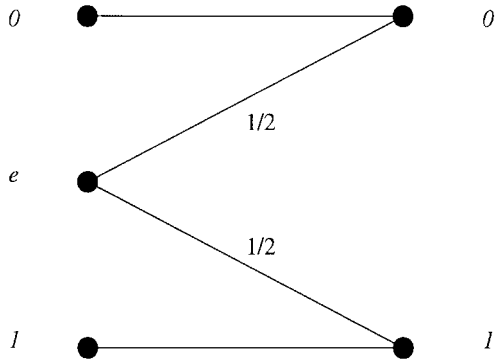


Fig. 2. Discrete memoryless channel in Example 3.

and 0 otherwise. This is not a product distribution; however, it does maximize the mutual information $I(\cdot, W^{(2)})$ with

$$W^{(2)}: \{a, b, c, d\}^2 \rightarrow \{0, 1\}^2.$$

Convergence in distribution of the k th-order empirical distribution does not take place for some codes operating on this channel. To see this consider an encoding procedure where at each time 2^i we switch between sending a string drawn from $\{a, c\}$ to a string drawn from $\{b, d\}$. Again, this code has rate equal to capacity and zero probability of error. Its first-order empirical distribution puts masses on (a, b, c, d) which oscillate between $(\frac{1}{6}, \frac{1}{6}, \frac{1}{3}, \frac{1}{3})$ and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6})$.

Example 3: Consider the channel of Fig. 2. The maximal mutual-information input $P_{\bar{X}}$ puts mass $(\frac{1}{2}, 0, \frac{1}{2})$ on $(0, e, 1)$. Suppose we construct a code $(n, 2^{n-1}, 0)$ which consists of all codewords whose first symbol is e and whose other symbols are forbidden to be e . This is a good code sequence for which

$$D(Q_{\hat{X}^n}^{(1)} \| P_{\bar{X}}) = +\infty \quad (42)$$

for all n . The reason for this ill behavior is that the code uses a symbol which has zero mass under the maximal mutual-information input distribution.

From a design viewpoint there is little incentive to use input symbols which are not used by any input distribution that maximizes mutual information as those symbols lead to “noisier” conditional output distributions. While good code sequences may include those symbols, as we saw in Example 3, only a vanishing percentage of those symbols is allowed, for otherwise \hat{X}^n could not maximize mutual information

asymptotically (Theorem 1). This motivates the following definition (useful in the context of discrete channels).

Definition 3: An (n, M, ϵ) code is *regular* if its empirical distribution $P_{\hat{X}^n} \ll P_{\bar{X}^n}$ for some \bar{X}^n which maximizes mutual information

$$I(\bar{X}^n; \bar{Y}^n) = \max_{\bar{X}^n} I(X^n; Y^n). \quad (43)$$

From the proof of Lemma 1c), note that in the context of discrete memoryless channels, regular codes are those that avoid any input symbol $a \in A$ for which

$$D(W(\cdot|a) \| P_{\bar{Y}}) < C. \quad (44)$$

Example 3 shows that for the purposes of proving approximation results using the divergence between the empirical input distribution and a maximal-mutual-information distribution it is necessary to restrict attention to regular good codes. As we have argued, this entails no essential loss of generality in the context of discrete channels.

Our main result in this section is

Theorem 4: For any discrete memoryless channel with capacity C , the k th-order empirical distribution of any regular good code sequence satisfies

$$\lim_{n \rightarrow \infty} \min_{\bar{X}^k: I(\bar{X}^k; \bar{Y}^k) = kC} D(Q_{\hat{X}^n}^{(k)} \| P_{\bar{X}^k}) = 0. \quad (45)$$

Proof: We will prove first the case $k = 1$. Our starting point is Theorem 3 which proves the corresponding result for the empirical output distribution. The output of W due to $Q_{\hat{X}^n}^{(1)}$ is $Q_{\hat{Y}^n}^{(1)}$. Our goal is to show that if $Q_{\hat{Y}^n}^{(1)}$ is close to $P_{\bar{Y}}$, then $Q_{\hat{X}^n}^{(1)}$ must be close to an input distribution that maximizes $I(X; Y)$. Note that as we saw in Example 3, the closest optimum input distribution may not converge as $n \rightarrow \infty$.

Recall that the optimal output distribution is unique and is denoted by $P_{\bar{Y}}$. Define

$$r(P) = \min_{\bar{X}: I(\bar{X}; \bar{Y}) = C} D(P \| P_{\bar{X}}) \quad (46)$$

and

$$w(\delta) = \max_{X: D(P_X \| P_{\bar{Y}}) \leq \delta} r(P_X). \quad (47)$$

Since the code is regular, we can restrict attention to channels all of whose input symbols have nonzero probability under some capacity-achieving input distribution. This entails no loss of generality because it does not change the set of distributions under which the minimum is taken in (45). For those channels, every input distribution is absolutely continuous with respect to at least one optimal input distribution. Thus Lemma 1c) implies that

$$w(0) = 0. \quad (48)$$

Upon showing that

$$\lim_{\delta \downarrow 0} w(\delta) = 0 \quad (49)$$

it will follow that (45) holds for $k = 1$, because by Theorem 3, $D(Q_{\hat{Y}^n}^{(1)} \| P_{\bar{Y}}) \rightarrow 0$.

To show (49), let us first check that r is a convex function.

$$\begin{aligned}
& r(\alpha P_1 + (1 - \alpha)P_2) \\
&= \min_{Q: I(Q, W) = C} D(\alpha P_1 + (1 - \alpha)P_2 \| Q) \quad (50) \\
&= \min_{Q_1: I(Q_1, W) = C} \min_{Q_2: I(Q_2, W) = C} \\
&\quad \cdot D(\alpha P_1 + (1 - \alpha) \cdot P_2 \| \alpha Q_1 + (1 - \alpha)Q_2) \quad (51) \\
&\leq \min_{Q_1: I(Q_1, W) = C} \min_{Q_2: I(Q_2, W) = C} \alpha D(P_1 \| Q_1) \\
&\quad + (1 - \alpha)D(P_2 \| Q_2) \quad (52) \\
&= \alpha r(P_1) + (1 - \alpha)r(P_2) \quad (53)
\end{aligned}$$

where (52) is a consequence of the convexity of divergence and (51) follows from the fact that

$$\begin{aligned}
& \{Q: I(Q, W) = C\} \\
&= \{Q = \alpha Q_1 + (1 - \alpha)Q_2, I(Q_1, W) = I(Q_2, W) = C\}
\end{aligned}$$

because $I(\cdot, W)$ is concave with a maximum value of C .

The set \mathcal{K} of all distributions on the finite set A is a compact subset of a Euclidean space. Furthermore, the function

$$g(P) = D(PW \| P_{\bar{Y}}) \quad (54)$$

is convex and continuous, where PW denotes the output distribution induced by P [1]. Therefore, the feasible set in (47) $\mathcal{K}_\delta = \{P \in \mathcal{K}, g(P) \leq \delta\}$ is compact. Thus

$$w(\delta) = r(P_{X_\delta}) \quad (55)$$

for some P_{X_δ} , such that

$$g(P_{X_\delta}) \leq \delta. \quad (56)$$

The compactness of \mathcal{K}_δ dictates that any sequence in that set will contain a subsequence that converges in the set. In particular, there must exist a decreasing sequence $\delta_n \rightarrow 0$ such that $P_{X_{\delta_n}}$ converges to an element which we shall denote by Q . By the continuity of r and g we can conclude that

$$w(\delta_n) = r(P_{X_{\delta_n}}) \rightarrow r(Q) \quad (57)$$

and

$$g(Q) = \lim_{n \rightarrow \infty} g(P_{X_{\delta_n}}) \leq \lim_{n \rightarrow \infty} \delta_n = 0. \quad (58)$$

Therefore, $Q \in \mathcal{K}_0$, which implies that

$$w(0) \geq r(Q) = \lim_{n \rightarrow \infty} w(\delta_n) \quad (59)$$

and (49) follows because w is monotone nondecreasing.

Having shown the desired result for $k = 1$, we will proceed to argue that it holds for arbitrary k . The essential part of the argument is that we can view k consecutive uses of a discrete memoryless channel $W: A \rightarrow B$ as one use of the discrete memoryless channel $W^{(k)}: A^k \rightarrow B^k$. This would prove the desired result had the time-averaged k th-order empirical distribution (5) been defined by averaging distributions in consecutive nonoverlapping blocks. At any rate, we can view $Q_{\hat{X}^n}^{(k)}$ as the mixture of such averaged

distributions over nonoverlapping blocks

$$Q_{\hat{X}^n}^{(k)}(\alpha_1, \dots, \alpha_k) = \frac{1}{k} \sum_{j=1}^k Q_{\hat{X}^n}^{(j,k)} \quad (60)$$

with

$$Q_{\hat{X}^n}^{(j,k)}(\alpha_1, \dots, \alpha_k) = \frac{k}{n - k + 1} \sum_l P_{\hat{X}_{j+lk}^{j+lk+k-1}}(\alpha_1, \dots, \alpha_k). \quad (61)$$

By symmetry, it is clear that for every offset $j = 1, \dots, k$

$$\lim_{n \rightarrow \infty} \min_{\bar{X}^k: I(\bar{X}^k; \bar{Y}^k) = kC} D(Q_{\hat{X}^n}^{(j,k)} \| P_{\bar{X}^n}) = 0 \quad (62)$$

and (45) follows immediately from (60) and the convexity of divergence. \square

Whenever capacity is achieved by a unique input distribution, the minimizing set in (45) is a singleton and Theorem 4 can be simplified as follows.

Corollary: Consider any discrete memoryless channel with capacity C such that \bar{X} is the unique input distribution achieving

$$C = I(\bar{X}; \bar{Y}).$$

For any $k = 1, 2, \dots$ the k th-order empirical distribution of any regular code sequence satisfies

$$\lim_{n \rightarrow \infty} D(Q_{\hat{X}^n}^{(k)} \| P_{\bar{X}} \times \dots \times P_{\bar{X}}) = 0. \quad (63)$$

Note that (63) implies convergence in variational distance and in distribution of the k th-order empirical distribution.

Having shown that for all k the k th-order empirical distribution approaches the set of maximal-mutual-information input distributions, we will examine what happens if rather than keeping k fixed we let it grow with the blocklength, in which case the degree of approximation is gauged by the *normalized divergence* $d(k(n), n)$ defined as

$$d(k, n) = \min_{\bar{X}^k: I(\bar{X}^k; \bar{Y}^k) = kC} \frac{1}{k} D(Q_{\hat{X}^n}^{(k)} \| P_{\bar{X}^k}). \quad (64)$$

We first give the following extension to the negative result illustrated in Example 1.

Theorem 5: Consider a discrete memoryless channel with a unique capacity-achieving distribution

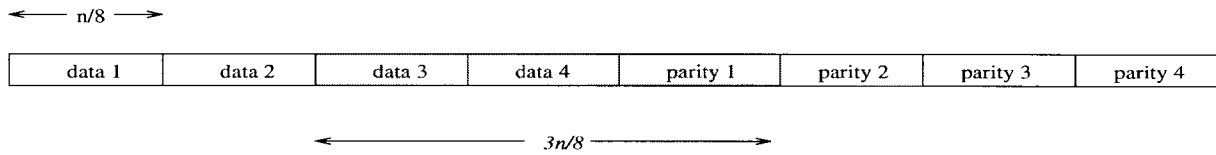
$$C = I(\bar{X}; \bar{Y}).$$

Suppose that

$$\frac{k(n)}{n} \geq \frac{C}{H(\bar{X})}. \quad (65)$$

Then, for every good code sequence

$$\lim_{n \rightarrow \infty} d(k(n), n) > 0. \quad (66)$$

Fig. 3. Juxtaposition of q systematic codes; $q = 4$.

Proof: We first claim that for any $k = 1, \dots, n$ and any code with M codewords, the entropy of the k th-order empirical distribution satisfies

$$H(Q_{\hat{X}^n}^{(k)}) \leq \log M + \log n. \quad (67)$$

To see this note that

$$H(Q_{\hat{X}^n}^{(k)}) = H(\hat{X}_J^{J+k-1}) \quad (68)$$

$$\leq H(\hat{X}_J^{J+k-1}, J) \quad (69)$$

$$= H(J) + \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} H(\hat{X}_i^{i+k-1}) \quad (70)$$

$$\leq \log n + \log M \quad (71)$$

where J is equiprobable on $1, \dots, n-k+1$. Equation (68) follows from Definition 2 (see (5)); and (71) follows from the fact that there can only be at most M different substrings $(i, \dots, i+k-1)$ if there are M codewords.

Let $Q_{\bar{X}}^\alpha$ denote the normalized version of $P_{\bar{X}}$ raised to the power α chosen to satisfy

$$H(Q_{\bar{X}}^\alpha) = \frac{1}{k}(\log M + \log n).$$

For any $k = 1, \dots, n$

$$\begin{aligned} d(k, n) &\geq \frac{1}{k} \min_{Q^k: H(Q^k) \leq \log M + \log n} D(Q^k \| P_{\bar{X}} \times \dots \times P_{\bar{X}}) \\ &= D(Q_{\bar{X}}^\alpha \| P_{\bar{X}}) \end{aligned} \quad (72)$$

where the minimum is attained by the product distribution [10]

$$Q^k = Q_{\bar{X}}^\alpha \times \dots \times Q_{\bar{X}}^\alpha.$$

A parametric solution of the optimization problem in (72) is shown in [10], from whose properties it can be concluded that if

$$\overline{\lim}_{n \rightarrow \infty} \frac{\log M + \log n}{k(n)} < H(\bar{X}) \quad (73)$$

then $d(k(n), n)$ is bounded away from 0. But (73) indeed holds because of (65) and the fact that for a good code and a discrete memoryless channel

$$\frac{\log M}{n} \rightarrow C = H(\bar{X}) - H(\bar{X}|\bar{Y}), \quad (74)$$

□

Theorem 5 shows that the empirical distribution of good codes cannot approximate the maximal-mutual-information distribution in a $k(n)$ -horizon with $k(n)$ growing faster than a certain constant times n . On the other hand, approximation is guaranteed for any fixed horizon that does not grow with n . What happens for horizons that do grow with n but not as

fast as the case considered in Theorem 5? The answer is: the maximum horizon at which approximation occurs depends on the code (for $k(n)$ growing logarithmically or faster). We will illustrate this with two codes for the BSC; in this case, the normalized divergence is simply

$$d(k, n) = 1 - \frac{1}{k} H(Q_{\hat{X}^n}^{(k)}) \quad \text{bit}. \quad (75)$$

In the first example, $k(n)$ grows almost as fast as (65) and yet convergence in normalized divergence occurs. In the second example, $k(n) = O(\log n)$ and convergence does not occur. For the sake of clarity both examples deal with a BSC whose capacity is arbitrarily close to the (fixed) rate of the constructed codes.

Example 4: Fix an arbitrary $\epsilon > 0$, and consider a BSC with capacity $C = \frac{1}{2} + \epsilon$. Let us construct a systematic code of rate $\frac{1}{2}$ and blocklength m by a) listing all the data strings of length $\frac{m}{2}$ and b) applying a permutation to that collection of $2^{m/2}$ strings which becomes the list of parity check $\frac{m}{2}$ strings.

From the optimality of systematic codes for the BSC [11], [12] we know that there exists at least one permutation (in fact, almost all permutations will work) such that the error probability of the code will vanish with blocklength when used with a BSC whose capacity is greater than $\frac{1}{2}$.

We will now fix an arbitrary integer q and will juxtapose the code chosen above (with blocklength $m = \frac{n}{q}$) with itself q times. The juxtaposition is arranged so that all the data bits of codes 1 through q appear consecutively and in that order, followed by the corresponding parity check bits in the same order. Thus the overall code has rate $\frac{1}{2}$, blocklength n , and vanishing error probability when used with the BSC.

Now let us examine $d(k(n), n)$ when

$$k(n) = \frac{q-1}{2q} n \quad (76)$$

which is equal to the length of $q-1$ blocks of data or parity checks. No matter which window of $k(n)$ consecutive bits we consider, it never includes data bits and parity-check bits of the same m -codeword. Whether the window includes only data bits, only parity checks, or both data bits and parity checks, there is never any correlation among the bits in the window, because the data consists of pure bits and every possible $\frac{m}{2}$ -string is a parity check string for the m -codeword. Thus for all $i = 1, \dots, k(n) + 1$

$$H(\hat{X}_i^{i+k(n)-1}) = k(n) \quad (77)$$

which together with (75) and the concavity of entropy implies that

$$d(k(n), n) = 0. \quad (78)$$

Note that for the channel considered in this example, the right side of (65) is equal to $\frac{1}{2} + \epsilon$, which is arbitrarily close to the factor in (76) for sufficiently large q and sufficiently small $\epsilon > 0$.

Example 5: Consider a BSC with capacity $C < 1$ bit. We will construct a code with rate arbitrarily close to C , vanishing error probability, and such that $d(k(n), n)$ is bounded away from 0 with $k(n) = O(\log n)$. Fix an arbitrary $R < C$. It is well known [8] that there exists a sequence of (n, M, λ_n) -codes such that

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log M \geq R \quad (79)$$

and

$$\lambda_n \leq \exp(-nE(R)) \quad (80)$$

where $E(R) > 0$ for $R < C$.

Let $m(n)$ be the solution to

$$m(n) = n \exp\left(-\frac{E(R)}{2} m(n)\right) \quad (81)$$

which implies that

$$\frac{m(n)}{\log n} = \frac{2}{E(R)} \left(1 - \frac{\log m(n)}{\log n}\right) \quad (82)$$

and, thus, $m(n) = O(\log n)$. Choose an $(m(n), M, \lambda_{m(n)})$ -code which satisfies

$$\liminf_{n \rightarrow \infty} \frac{\log M}{m(n)} \geq R \quad (83)$$

and

$$\lambda_{m(n)} \leq \exp(-E(R)m(n)) \quad (84)$$

and juxtapose it with itself $s(n) = n/m(n)$ times.

We have constructed an $(n, M^{s(n)}, s(n)\lambda_{m(n)})$ -code. Its asymptotic rate is at least R because of (83) and its error probability vanishes because

$$\begin{aligned} s(n)\lambda_{m(n)} &\leq \exp(-E(R)m(n)) \frac{n}{m(n)} \\ &= \exp\left(-E(R) \frac{m(n)}{2}\right). \end{aligned} \quad (85)$$

Now, select an integer q such that

$$q > \frac{R}{1-R} \quad (86)$$

and let $k(n) = qm(n)$. This implies that the window can intercept at most $q+1$ consecutive $m(n)$ -codewords, which means that $Q_{\hat{X}^n}^{(k(n))}$ is a distribution on a set of at most $qm(n)M^{q+1}$ elements. Thus

$$\begin{aligned} d(k(n), n) &= 1 - \frac{H(Q_{\hat{X}^n}^{(k(n))})}{k(n)} \\ &\geq 1 - \frac{q+1}{q} \frac{\log M}{m(n)} - \frac{\log(qm(n))}{qm(n)} \end{aligned} \quad (87)$$

which is bounded away from 0 in view of (83) and (86).

$k(n) =$	$O(1)$	must approximate
$k(n) =$	$o(\log n)$	no negative example known
$k(n) =$	$O(\log n) \leftrightarrow \alpha n$	depends on the code
$k(n) \geq$	αn	cannot approximate

Fig. 4. Approximation by $k(n)$ -order empirical distributions.

As evidenced by Theorems 4 and 5 and Examples 3 and 4, the situation is depicted in Fig. 4, where

$$\alpha = \frac{C}{H(\bar{X})}$$

for a DMC with a unique maximal-mutual-information input distribution.

IV. CHANNELS WITH MEMORY

In Section II we saw two approximation results that hold for channels with memory: empirical distributions achieve capacity asymptotically (Theorem 1), and approximation of empirical output statistics (Theorem 2). As in Section III, we now want to go one step further and show that the empirical distribution of a good code converges to a maximal-mutual-information distribution. As we saw in Section III, the nonuniqueness of maximal-mutual-information distributions is a source of difficulty, which is compounded with many other sources of ill behavior when dealing with channels with memory. For this reason, we prefer to restrict attention to channels with memory which have unique optimal distributions (in a certain sense stated below). This allows an easier statement of the result along with a simplified proof (which provides a shortcut to an independent proof of the Corollary to Theorem 4), valid for a wide class of approximation measures. Beyond the assumption of uniqueness, we just require that the channel be such that the normalized maximal-mutual informations grow with n and converge to capacity. This condition is frequently satisfied, for example, discrete channels with finite memory [13], and colored noise additive channels (Section V).

Theorem 6: Fix k . Consider the following assumptions on a channel.

- 1) For all distributions P on A^k , the function

$$C_n(P) = \frac{1}{n} \max_{X^n: Q_{X^n}^{(k)} = P} I(X^n; Y^n) \quad (88)$$

is increasing with n and converges; denote its limit by $C(P)$.

- 2) The supremum of $C(P)$ over all input distributions is attained by a unique distribution, which will be denoted by $P_{\bar{X}}^{(k)}$.

- 3) The channel capacity is equal to $C = C(P_{\bar{X}}^{(k)})$.

Consider a distance measure between distributions $d(Q, P)$ which is convex in Q , $d(Q, P) = 0$ if and only if $P = Q$, and such that $d(\cdot, P_{\bar{X}}^{(k)})$ is bounded in a compact neighborhood \mathcal{P} of $P_{\bar{X}}^{(k)}$. Any good code such that $Q_{\hat{X}^n}^{(k)} \in \mathcal{P}$ satisfies

$$\lim_{n \rightarrow \infty} d(Q_{\hat{X}^n}^{(k)}, P_{\bar{X}}^{(k)}) = 0. \quad (89)$$

Proof: We will prove the result for $k = 1$. The extension to general k by considering a new channel with k consecutive uses of the original channel follows the same lines of the proof of Theorem 4. Let us first check that the assumption of Theorem 1 is satisfied. If \bar{X}^n is such that it maximizes $I(X^n; Y^n)$, denote the equal mixture of its marginals by $Q_{\bar{X}^n}^{(1)}$. Using the assumptions of the present theorem and the general upper bound on capacity as the \liminf of maximal normalized mutual informations [5], we can write the chain of inequalities

$$C \leq \liminf_{n \rightarrow \infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n) \quad (90)$$

$$\leq \overline{\lim}_{n \rightarrow \infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n) \quad (91)$$

$$= \overline{\lim}_{n \rightarrow \infty} C_n(Q_{\bar{X}^n}^{(1)}) \quad (92)$$

$$\leq \overline{\lim}_{n \rightarrow \infty} C(Q_{\bar{X}^n}^{(1)}) \quad (93)$$

$$\leq C(P_{\bar{X}}) \quad (94)$$

$$= C \quad (95)$$

and thus all inequalities must hold with equality, and (7) holds. According to Theorem 1, this implies that for all $\gamma > 0$ and sufficiently large n

$$C - \gamma \leq \lim_{n \rightarrow \infty} \frac{1}{n} I(\hat{X}^n; \hat{Y}^n) \quad (96)$$

$$\leq C_n(Q_{\hat{X}^n}^{(1)}) \quad (97)$$

$$\leq C(Q_{\hat{X}^n}^{(1)}) \quad (98)$$

$$\leq C \quad (99)$$

which enables us to conclude that

$$C(Q_{\hat{X}^n}^{(1)}) \rightarrow C = C(P_{\bar{X}}). \quad (100)$$

Now we will use a continuity argument along with the uniqueness of the maximizing argument of $C(P)$ to complete the proof. Analogously to (47), define

$$w(\delta) = \max_{P \in \mathcal{P}: C - C(P) \leq \delta} d(P, P_{\bar{X}}). \quad (101)$$

By the assumption of uniqueness, $w(0) = 0$. Moreover, for a neighborhood of the origin, the function $w(\delta)$ is finite. Now, if we can show that

$$\lim_{\delta \downarrow 0} w(\delta) = 0 \quad (102)$$

then the proof will be complete in view of (100). To prove (102) we can use a simplified version of the argument that led to the analogous continuity result in the proof of Theorem 4. Following that argument, since $d(P, P_{\bar{X}})$ is convex in P , it is enough to show that the function $C(P)$ is concave. But $C(P)$ is the limit of the functions $C_n(P)$, which are easily shown to be concave (using the concavity of mutual information on the input distribution). \square

Corollary: For any discrete channel with memory satisfying the conditions of Theorem 6, any good code such that

$$Q_{\hat{X}^n}^{(k)} \ll P_{\bar{X}}^{(k)} \quad (103)$$

satisfies

$$\lim_{n \rightarrow \infty} D(Q_{\hat{X}^n}^{(k)} || P_{\bar{X}}^{(k)}) = 0. \quad (104)$$

Proof: In this case we are considering divergence as the approximation measure $d(Q, P) = D(Q || P)$. Note that the finiteness of the input alphabet implies that a neighborhood \mathcal{P} in which the distance measure is bounded includes the set of all distributions absolutely continuous with respect to $P_{\bar{X}}^{(k)}$ mass.

V. ADDITIVE-NOISE CHANNELS

In this section we consider channels where the outputs are

$$Y_i = X_i + N_i \quad (105)$$

and codebooks are constrained to satisfy (10) with

$$d_n(a_1, \dots, a_n) = \frac{1}{n} \sum_{i=1}^n a_i^2 \leq \beta. \quad (106)$$

Under general conditions on the noise sequence $\{N_i\}$ it is possible to prove that the capacity–cost function is given by (12). In such case, the empirical input distributions of good codes must achieve the capacity–cost function asymptotically, and as in Theorem 2, the output empirical distribution approaches in normalized divergence the unique optimal output distribution. The input empirical distribution is always discrete. Therefore, any attempt to show convergence in the sense of vanishing divergence for the *input* distributions would be futile if the maximal-mutual-information input distribution were continuous (e.g., if the noise were Gaussian). A useful alternative distance measure which does not suffer from such a drawback is the Ornstein distance [14]

$$\rho(P_{X^k}, P_{\bar{X}^k}) = \min_{P_{X^k \bar{X}^k}} E[||X^k - \bar{X}^k||^2] \quad (107)$$

where the minimum is over all joint distributions with marginals P_{X^k} and $P_{\bar{X}^k}$.

Note that convergence in Ornstein distance implies convergence in distribution. On the other hand, in certain cases convergence in (normalized) divergence implies convergence in (normalized) Ornstein distance [15].

Let us first consider the special case of white noise, for which a particularly simple argument leads to the convergence of input statistics. Note that the Gaussian-noise channel is a special case of the channels admissible in the following result.

Theorem 7: If the noise is independent and identically distributed (i.i.d.) and its characteristic function $\Psi_N(\omega)$ is nonzero for all ω , then for every k

$$Q_{\hat{X}^n}^{(k)} \xrightarrow{d} P_{\bar{X}} \times \dots \times P_{\bar{X}} \quad (108)$$

where \xrightarrow{d} denotes convergence in distribution and $P_{\bar{X}}$ is the unique distribution that maximizes $I(X; X + N)$.

Proof: First notice that the condition on the noise distribution implies that different input distributions result in different output distributions, and, thus, Lemma 2d) implies that the optimal input distribution is unique. If the noise in (105) is white, then the optimal output distribution is a product distribution and Theorem 3 can be shown to hold in this case, using an entirely analogous proof. Convergence of unnormalized divergence of the k -order empirical output distribution implies convergence in distribution

$$Q_{\hat{Y}^n}^{(k)} \xrightarrow{d} P_{\bar{Y}} \times \dots \times P_{\bar{Y}} \quad (109)$$

which, in turn, implies pointwise convergence of the corresponding characteristic functions [16]

$$\Psi_{Q_{\bar{X}^n}^{(k)}}(\omega_1, \dots, \omega_k) \rightarrow \prod_{i=1}^k \Psi_{P_{\bar{Y}}^{(k)}}(\omega_i). \quad (110)$$

We may divide both sides of (110) by

$$\Psi_N(\omega_1) \cdots \Psi_N(\omega_k) \quad (111)$$

yielding convergence of the characteristic function of the k -order empirical distribution to a function which is continuous at the origin, and thus [16] yielding the desired result. \square

Let us turn our attention to the case of nonwhite Gaussian noise. Among the n -input distributions such that

$$E[\|X^n\|^2] \leq n\beta \quad (112)$$

the one that maximizes mutual information is well known [8] to be Gaussian zero-mean with covariance matrix

$$\mathbf{K}_{\bar{X}}^{(n)} = \mathbf{U} \text{diag}\{(\nu_n - \lambda_1)^+, \dots, (\nu_n - \lambda_n)^+\} \mathbf{U}^t \quad (113)$$

where ν_n is chosen so that (112) is satisfied with equality and $(\lambda_1, \dots, \lambda_n)$ and \mathbf{U} are the eigenvalues and eigenvector matrix of the noise covariance matrix

$$\mathbf{K}_N^{(n)}(i, j) = R_N[i - j], \quad (114)$$

As $n \rightarrow \infty$, \bar{X}^n becomes a stationary Gaussian random process whose power spectral density is

$$S_{\bar{X}} = (\nu - S_N(f))^+ \quad (115)$$

with $S_N(f)$ equal to the Fourier transform of $R_N[i]$ and ν adjusted so that the input power is β . The convergence of the empirical input distribution to a Gaussian process with the water-filling spectral density (115) follows from the following result.

Theorem 8: Consider a channel (105) with stationary Gaussian noise with power spectral density which is nonzero except at most in a singular set of frequencies. Fix k . Denote by $\Phi^{(k)}$ the k -Gaussian variate corresponding to k consecutive samples of the optimal stationary input distribution (whose spectral density is given by (115)). Then, the Ornstein distance between the k th-order empirical distribution and $\Phi^{(k)}$ satisfies

$$\lim_{n \rightarrow \infty} d(Q_{\bar{X}^n}^{(k)}, \Phi^{(k)}) = 0. \quad (116)$$

Proof: First we note that the Ornstein distance is convex in each argument. Instead of working with the time-averaged statistics as defined in Definition 2, it is more convenient to work with time averages over nonoverlapping blocks such as (61). The analogous result to (116) is stronger for those empirical statistics, because in that case (60) and the convexity of Ornstein distance imply (116). It can be checked that in Theorem 6 we may replace $Q_{\bar{X}^n}^{(k)}$ by $Q_{\bar{X}^n}^{(j,k)}$. Having done that, we need to show that the conditions of Theorem 6 are satisfied.

To check that $C_n(P)$ is monotonically increasing note first that if n is a multiple of k , then

$$\frac{1}{n} I(X^n; Y^n) \leq \frac{1}{2n} I(X^{2n}; Y^{2n}) \quad (117)$$

where X^{2n} is constructed by juxtaposing two independent copies of X^n . If X^n is such that it satisfies $Q_{\bar{X}^n}^{(j,k)} = P$, so will X^{2n} . Thus $C_n(P) \leq C_{2n}(P)$. The more general monotonicity condition required by Theorem 6 can be obtained by partitioning X^n into independent blocks of different sizes. The convergence of $C_n(P)$ follows from its monotonicity and the fact that it is upper-bounded by the channel capacity. The uniqueness of the maximizing argument of $C(P)$ follows for additive Gaussian channels by the fact that the random process which attains capacity is Gaussian with power spectral density given by the water-filling solution. No process whose k -dimensional distribution $Q_{\bar{X}^n}^{(k)}$ is different from $\Phi^{(k)}$ (nonzero variational distance) can hope to achieve capacity. \square

ACKNOWLEDGMENT

Fruitful discussions with Prof. A. Dembo are acknowledged.

REFERENCES

- [1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, pp. 379–423, 623–656, July–Oct. 1948.
- [3] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. 39, pp. 752–772, May 1993.
- [4] T. S. Han and S. Verdú, "Spectrum invariance under output approximation for discrete memoryless channels with full rank," *Probl. Pered. Inform.*, vol. 2, pp. 9–27, 1993.
- [5] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147–1157, July 1994.
- [6] S. Kullback, J. C. Keegel, and J. H. Kullback, *Topics in Statistical Information Theory* (Lecture Notes in Statistics), vol. 42. Berlin, Germany: Springer, 1987.
- [7] I. Csiszár, "Sanov property, generalized i -projection and a conditional limit theorem," *Ann. Probab.*, pp. 768–793, Aug. 1984.
- [8] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [9] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.
- [10] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, pp. 63–86, Jan. 1996.
- [11] E. M. Gabidulin, "Limits for the decoding error probability when linear codes are used in memoryless channels," *Probl. Pered. Inform.*, vol. 2, pp. 55–62, 1967.
- [12] S. Shamai (Shitz) and S. Verdú, "Capacity of channels with side information," *European Trans. Telecomm.*, vol. 6, pp. 587–600, Sept.–Oct. 1995.
- [13] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. New York: Springer, 1978.
- [14] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's d -bar distance with applications to information theory," *Ann. Probab.*, pp. 315–328, 1975.
- [15] M. Talagrand, "Transportation cost for Gaussian and other product measures," preprint, 1995.
- [16] K. L. Chung, *A Course in Probability Theory*. New York: Academic, 1974.