How Many Information Bits Does a Decoder Need About the Channel Statistics? *

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL

July 23, 1998

Abstract

We investigate the minimum amount of side information about the channel statistics, that must be provided to the decoder in order to guarantee reliable communication in the random coding sense, for certain classes of channels.

Index Terms: universal decoding, AEP, channel simulation.

1 Introduction

Consider a rate R, length n block code $\mathcal{E}$ for transmission over a finite alphabet channel W, and a decoder $\mathcal{D}$. The encoder, that does not know the statistics of the channel, selects at random (and shares with the decoder) a codebook, where each codeword is generated independently by some probability distribution Q^n. The question that we address is: how many information bits about the statistics of the channel must be provided to the decoder so as to guarantee reliable communication? Later on, we shall formalize precisely this problem. Clearly, the question is meaningful only if the channel at hand is 'good' in the sense that a random codebook generated by Q^n gives, with high probability, reliable communication at least for the optimum maximum likelihood (ML) decoder.

It is well-known that for certain parametric families of channels (e.g., memoryless channels, finite-state channels, etc.) there exist universal decoders that do not require *any* statistical side information, and yet, not only maintain reliable communication in the random

*This research was supported by the Wolfson Research Awards administered by the Israel Academy of Science and Humanities. Part of this work was done while the author was on leave at Hewlett-Packard Laboratories, Palo Alto, CA.

coding sense [3], but also achieve the optimum random coding error exponent [1], [2], [6], [11]. However, here we do not confine ourselves only to channels that are characterizable by a parameter set of fixed dimension, but rather to a much wider set of channels that includes a certain subclass of the class of stationary and ergodic channels. We shall elaborate later on the relation between earlier work on universal decoding and the present work.

Generally speaking, our main result is that exponentially $N = 2^{H(Y^n|X^n)+nR}$ bits are necessary and sufficient for describing the channel to the decoder, where $H(Y^n|X^n)$ is the nth order conditional output entropy given an input X^n governed by Q^n. More precisely, for a given $H_0 > 0$, if $N \geq 2^{n(H_0+R+\delta)}$, for some small $\delta > 0$, then there is an N-bit description that enables decoding with small average error probability, w.r.t. the ensemble of codes, for every 'good' channel whose conditional output entropy $H(Y^n|X^n)$ does not exceed nH_0. If, however, $N \leq 2^{n(H_0+R-\delta)}$, then regardless of the method of describing the channel, and the decoder used for this description, there is at least one 'good' channel for which $H(Y^n|X^n) \leq nH_0$, and yet the average error probability is high. Furthermore, this argument remains true even if the code is optimized for the given channel and decoder, rather than chosen at random. The intuition behind this expression of N is that the decoder must essentially know what are the $2^{H(Y^n|X^n)}$ conditionally typical output sequences given each one of the 2^{nR} channel input messages.

The significance of our results is primarily in characterizing the richness of the class of channels, or the "effective number of distinct channels" from the viewpoint of decoding, given certain parameter values R, n, and Q^n, of the encoder. An important conclusion is that training by independent output samples for each codeword, is an efficient (randomized) description of the channel in the sense that it achieves the above minimum description exponent with small average error probability.

Recently, a few similar problems have been addressed in the context of minimum statistical description of sources, for tasks like classification [10], lossless compression [5] and vector quantization [7]. In [5] and [10], the conclusion was that it must take roughly 2^{H_n} bits to describe a source, where H_n is a quantity related to the nth order entropy, and again, the intuition is that the typical sequences of the source must be conveyed in some way. In [7], however, the behavior appeared to be different: rather than describing the source itself, it turns out to be more efficient to describe the optimum 'device' (in that case, the vector quantizer) for the given source. This reduces N from 2^{nH_n} to essentially 2^{nR}

bits that are needed to describe the centroids of the rate R vector quantizer. In the channel decoding problem considered here, we have a mixed situation. The number N factorizes into the product of $2^{H(Y^n|X^n)}$ and 2^{nR}, where the former depends only on the channel (and the random coding distribution), and the latter depends only on the size of the 'device', namely, the encoder-decoder in this case.

Finally, it should be emphasized that, similarly as in [5],[7], and [10], our results are *non-asymptotic* in the sense that limits as $n \to \infty$ are never taken. Rather than that, we consider a *fixed* and finite block length n, which is assumed to be at least as large as some integer $n_0(\delta, \epsilon)$, where δ and ϵ are (arbitrarily small, but prescribed) positive reals, which are parameters of the problem. This is important since the convergence of $\{H(Y^n|X^n)/n\}_{n\geq 1}$ (if at all, a limit exists [8, Lemma 1]) might be arbitrarily slow for certain channels and input processes. Thus, for a certain $n \geq n_0(\delta, \epsilon)$, where our results are already valid, $H(Y^n|X^n)/n$ might be still far away from its limit.

The outline of the paper is as follows. In Section 2, the problem is defined along with notation conventions and the basic assumptions are described and discussed. In Section 3, some examples of channel descriptions are provided and the direct theorem, stating that $2^{nH(Y^n|X^n)+nR}$ bits are sufficient, is formalized and proved. Finally, in Section 4, the converse theorem, that tells that $2^{H(Y^n|X^n)+nR}$ bits are necessary, is stated and proved.

2 Notation, Problem Formulation, and Assumptions

We adopt the following notation conventions. Scalar random variables will be denoted by capital letters (e.g., X), specific values they may take will be denoted by the respective lower case letter (x), and alphabets will be denoted by the respective script letters ($\mathcal{X}$). The probability mass function (PMF) that governs a scalar random variable will be also denoted by a lower case letter (e.g., q). Random vectors will be denoted by capital letters with a superscript that denotes the dimension, e.g., $X^n = (X_1, ..., X_n)$. The same convention applies to specific vector values ($x^n = (x_1, ..., x_n)$), and the corresponding superalphabet ($\mathcal{X}^n$). The PMF that governs a random vector will be denoted by a capital letter with a superscript that denotes the dimension (e.g., Q^n). Thus, $Q^1 = q$. In a similar manner, processes (or sources) will be denoted by boldface capital letters, e.g., $\boldsymbol{X} = (X_1, X_2, ...)$, specific infinite strings will be denoted by boldface lower case letters (e.g., $\boldsymbol{x} = (x_1, x_2, ...)$), and probability measures that govern processes will be denoted by capital letters (e.g.,

Q). The same conventions will apply to conditional measures and conditional probability distributions associated with channels. The cardinality of a finite set will be denoted by $|\cdot|$, e.g., $|\mathcal{X}|$ is the size of the alphabet of X. The Cartesian product of two sets $\mathcal{A}$ and $\mathcal{B}$ will be denoted by $\mathcal{A} \times \mathcal{B}$.

The problem is defined as follows. A transmitter wishes to send information across some finite input-output alphabet channel by using a rate R block encoder $\mathcal{E}$ of block length n. Since the n-th order transition probabilities of the channel $W^n(y^n|x^n) = \Pr\{Y^n = y^n|X^n = x^n\}$, $x^n \in \mathcal{X}^n$, $y^n \in \mathcal{Y}^n$, are unknown to the transmitter, the $M = 2^{nR}$ codewords $x^n(1), x^n(2), ..., x^n(M)$, $(x^n(i) \in \mathcal{X}^n,\ 1 \leq i \leq M)$, that together form the codebook $\mathcal{E}$, are randomly drawn independently according to some PMF Q^n on $\mathcal{X}^n$. Once chosen, the codebook is then provided to the decoder $\mathcal{D}$ as well.

The decoder operation model is as follows. Given a code $\mathcal{E} = \{x^n(i)\}_{i=1}^M$ and a received vector y^n, the decoder estimates the transmitted message as the integer which minimizes over i a certain function $D(x^n(i), y^n)$, henceforth referred to as the *decoding metric*, where ties are broken arbitrarily and counted as errors.

For a given code $\mathcal{E}$ of block length n, a channel W^n, and a decoding metric D, let $P_e(\mathcal{E}, W^n, D)$ denote the probability of error, where the prior probability distribution over the message set is uniform, i.e.,

$$P_e(\mathcal{E}, W^n, D) = \frac{1}{M} \sum_{i=1}^{M} \sum_{y^n \in \Lambda_i^c} W^n(y^n|x^n(i)), \tag{1}$$

and where Λ_i^c is complementary to the ith decision region

$$\Lambda_i = \{y^n :\ D(x^n(i), y^n) < \min_{j \neq i} D(x^n(j), y^n)\}. \tag{2}$$

For a random code $\mathcal{E}$ drawn according to Q^n, the *average probability of error* $\bar{P}_e(Q^n, W^n, D)$ is the expectation of $P_e(\mathcal{E}, W^n, D)$ w.r.t. the product measure $\prod_{i=1}^M Q^n(x^n(i))$.

Obviously, ML decoding can be carried out if the channel W^n is perfectly known to the decoder. Suppose that the decoder is provided with *partial* knowledge of W^n, which is summarized in an N-bit binary string z^N. This description of W^n by z^N may take on many forms, e.g., finite precision approximations of the transition probabilities $\{W^n(y^n|x^n)\}$, training samples of the channel output for certain inputs, and so on. Quite clearly, if N is very large, there are many ways to describe W^n sufficiently accurately such that the average error probability can be made essentially as small as that of the optimal ML decoder for W^n.

On the other extreme, it is also obvious that if N is too small, then regardless of the method of the describing W^n, the vector z^N cannot contain enough information about W^n so as to guarantee small error probability for every channel in a large class (even if the encoder is optimized). The questions that we investigate here are: Where is the transition between these two situations? What is the minimum N such that there still exists a description of W^n that keeps the average error probability small?

More precisely, let R, n, and Q^n be the parameters of the random code, and let $\mathcal{C}_n$ be a certain class of conditional PMF's $\{W^n : \mathcal{X}^n \to \mathcal{Y}^n\}$. An *$N$-bit description* for $\mathcal{C}_n$ is a deterministic mapping $F : \mathcal{C}_n \to \{0,1\}^N$. Associated with every $z^N \in \{0,1\}^N$, there is a decoding metric $D_{z^N}(\cdot,\cdot)$. For a given $\epsilon > 0$, let $N(n)$ be the smallest positive integer N for which there exists an N-bit description $z^N = F(W^n)$ for $\mathcal{C}_n$, and a set of 2^N decoding metrics $\{D_{z^N},\ z^N \in \{0,1\}^N\}$ such that for every $W^n \in \mathcal{C}_n$,

$$\bar{P}_e(Q^n, W^n, D_{F(W^n)}) \leq \bar{P}_e(Q^n, W^n, D^{W^n}) + \epsilon, \tag{3}$$

where D^{W^n} is the optimal ML decoding metric for W^n, i.e., $D^{W^n}(x^n, y^n) = -\log W^n(y^n|x^n)$.

Clearly, the problem is meaningful only for classes of *good* channels in the sense that $\bar{P}_e(Q^n, W^n, D^{W^n})$ is small for the given choice of R, n, and Q^n. For such classes of channels, we will be interested in characterizing the exponential growth rate of the function $N(n)$ for large n.

We next describe and discuss the basic assumptions. Consider a channel W with a finite input alphabet $\mathcal{X}$ and a finite output alphabet $\mathcal{Y}$. For a given channel input process $\boldsymbol{X}$ governed by Q, let $P = Q \times W$ denote the probability measure that governs the joint input-output process $(\boldsymbol{X}, \boldsymbol{Y}) = \{(X_t, Y_t)\}_{t \geq 1}$, and let V denote the marginal probability measure corresponding to the output process $\boldsymbol{Y}$. For a given positive integer n, let Q^n, P^n, and V^n denote the respective nth order marginals associated with (X^n, Y^n), and let $W^n(y^n|x^n) = P^n(x^n, y^n)/Q^n(x^n)$ denote the nth order restriction of W w.r.t. Q, where $W^n(y^n|x^n) \stackrel{\Delta}{=} 0$ for $Q^n(x^n) = 0$. The nth order conditional output entropy is defined as

$$H(Y^n|X^n) = -\sum_{x^n \in \mathcal{X}^n} \sum_{y^n \in \mathcal{Y}^n} P^n(x^n, y^n) \log W^n(y^n|x^n), \tag{4}$$

and the nth order output entropy is defined as

$$H(Y^n) = -\sum_{y^n \in \mathcal{Y}^n} V^n(y^n) \log V^n(y^n). \tag{5}$$

Finally, let $I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n)$.

For given $\delta > 0$ and $\epsilon > 0$, a positive integer n, and an input process Q with an nth order marginal Q^n, let $\mathcal{W}_n(Q^n, \delta, \epsilon)$ denote the class of conditional PMF's W^n such that

$$\Pr\{(x^n, y^n) : \log W^n(y^n|x^n) < -H(Y^n|X^n) - n\delta\} \leq \epsilon, \tag{6}$$

and

$$\Pr\{y^n : \ \log V^n(y^n) > -H(Y^n) + n\delta\} \leq \epsilon, \tag{7}$$

where the probabilities are defined w.r.t. P^n.

We assume that the channel W at hand is a member of the class $\mathcal{W}(Q, \delta, \epsilon)$ of all channels, such that for *some* $n \geq n_0(\delta, \epsilon)$, we have $W^n \in \mathcal{W}_n(Q^n, \delta, \epsilon)$. It should be stressed that $n_0(\delta, \epsilon)$ is a certain function that depends *solely* on δ and ϵ and not on the particular channel within $\mathcal{W}(Q, \delta, \epsilon)$. We next discuss the relationship between conditions (6), (7) and certain asymptotic properties of channels that are commonly assumed.

Conditions (6) and (7) guarantee that $\bar{P}_e(Q^n, W^n, D^{W^n})$ is small for all $R \leq I(X^n;Y^n) - O(\delta)$, provided that ϵ is small and $n \geq n_0(\delta, \epsilon)$. At first glance, eqs. (6) and (7) seem similar to the asymptotic equipartition property (AEP). However, a more careful inspection reveals a few differences. First, while the common definition of a stationary ergodic channel (that satisfies the AEP) requires a stationary ergodic input-output process $(\boldsymbol{X}, \boldsymbol{Y})$ for *any* stationary and ergodic input $\boldsymbol{X}$, here the parallel requirement applies only to a process Q whose nth order marginal Q^n serves as the selected random coding distribution. Secondly, in contrast to the AEP, the deviations of the random variables $n^{-1}\log W^n(Y^n|X^n)$ and $n^{-1}\log V^n(Y^n)$ are defined w.r.t. the normalized nth order entropies $H(Y^n|X^n)/n$ and $H(Y^n)/n$, and *not* their limits as $n \to \infty$, namely, the entropy rates $\bar{H}(\boldsymbol{Y}|\boldsymbol{X})$ and $\bar{H}(\boldsymbol{Y})$, respectively. This is an important difference since the convergence of the sequences of normalized entropies might be arbitrarily slow, if at all these sequences converge. Therefore, for every given $n \geq n_0(\delta, \epsilon)$, there exist channels that satisfy (6) and (7), yet the probabilities of the events $\{(x^n, y^n) : \log W^n(y^n|x^n) < -n\bar{H}(\boldsymbol{Y}|\boldsymbol{X}) - n\delta\}$ and $\{y^n : \ \log V^n(y^n) > -n\bar{H}(\boldsymbol{Y}) + n\delta\}$ are still large even if the AEP is eventually satisfied. As an example, consider a memoryless channel $W^n(y^n|x^n) = \prod_{i=1}^n w_i(y_i|x_i)$ where $w_i(\cdot|x)$ is uniform on $\mathcal{Y}_i(x) \subseteq \mathcal{Y}$ (depending on i and x), with $|\mathcal{Y}_i(x)| = K_i$, and $\{K_i\}_{i\geq 1}$ is an *arbitrary* sequence of integers in $\{1, ..., |\mathcal{Y}|\}$. Clearly, this channel satisfies (6) for all n, even for $\delta = \epsilon = 0$. However, $H(Y^n|X^n)/n = n^{-1}\sum_{i=1}^n \log K_i$ may converge arbitrarily slowly to $\bar{H}(\boldsymbol{Y}|\boldsymbol{X})$, or

may not converge at all.

Obviously, this does not mean that conditions (6) and (7) are more general than the AEP. It demonstrates, however, that some situations allowed by these conditions are not covered by the AEP. Of course, the AEP is not the most general condition for the feasibility of reliable communication at positive rates. In [9], it has been shown the coding capacity is always given by $\underline{I}(\boldsymbol{X};\boldsymbol{Y})$, defined as the liminf in probability[1] of the sequence of normalized information densities $\{n^{-1}\log W^n(Y^n|X^n)/V^n(Y^n)\}_{n\geq 1}$. Again, this means that for every given n, there are channels that satisfy (6) and (7), yet the probability that $\{\log W^n(y^n|x^n)/V^n(y^n) \leq n(\underline{I}(\boldsymbol{X};\boldsymbol{Y})-\delta)\}$ is still large, and so might be the probability of error (see [9, eq. (2.1)]). This concludes our discussion regarding eqs. (6) and (7).

Finally, we describe our assumptions on the coding rate R. We mentioned earlier that eq. (3) would be interesting only for good channels in the sense that given R, Q^n, and n, the average error probability is small. As mentioned earlier, this is the case when $I(X^n;Y^n) > n(R+O(\delta))$. Also, since we expect the description length of a channel to increase with $H(Y^n|X^n)$ (see Introduction), then it will be natural to restrict $\mathcal{C}_n$ to channels for which $H(Y^n|X^n)$ is uniformly upper bounded by nH_0 for some constant $H_0 > 0$. Therefore, we define the class of channels as

$$\begin{aligned}\mathcal{C}_n &= \mathcal{C}_n(Q^n, R, H_0, \delta, \epsilon) \\ &= \mathcal{W}_n(Q^n,\delta,\epsilon) \cap \{W^n:\ I(X^n;Y^n) \geq n(R+5\delta),\ H(Y^n|X^n) \leq nH_0\}, \qquad (8)\end{aligned}$$

where the factor of 5 in front of δ is immaterial and introduced for technical reasons only. Note, that a necessary condition for $I(X^n;Y^n) \geq n(R+5\delta)$ to hold uniformly for every $W^n \in \mathcal{C}_n$, is that $R \leq \log|\mathcal{Y}| - H_0 - 5\delta$, where $H_0 < \log|\mathcal{Y}| - 5\delta$. Also, since $I(X^n;Y^n)$ never exceeds $n\log|\mathcal{X}|$, it is also necessary that $R \leq \log|\mathcal{X}| - 5\delta$. Thus, from the combination of these two requirements, it will be assumed that $R \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}| - H_0\} - 5\delta$.

3 Efficient Channel Descriptions and the Direct Theorem

In the Introduction, we mentioned that $2^{H(Y^n|X^n)+nR}$ bits are sufficient for describing a channel without much loss in average error probability. Before we establish this claim formally, let us begin with two informal examples of deterministic descriptions, and then turn to a randomized description for which we prove achievability.

[1]The liminf in probability $\underline{A}$ of a sequence of random variables $\{A_n\}$ is defined [9] as $\underline{A} = \sup\{a:\ \limsup_{n\to\infty}\Pr\{A_n \leq a\} = 0\}$.

Example 1 - description of conditional type classes. For each codeword $x^n(i)$, let $T(x^n(i)) = \{y^n : \log W^n(y^n|x^n(i)) \geq -H(Y^n|X^n) - n\delta\}$, for some small $\delta > 0$, and let z^N consist of the binary representations of all $y^n \in T(x^n(i))$ using $n \log |\mathcal{Y}|$ bits per vector. Since $|T(x^n(i))| \leq 2^{H(Y^n|X^n)+n\delta}$, N is upper bounded by $\sum_{i=1}^{M} |T(x^n(i))| \cdot n \log |\mathcal{Y}| \leq n \log |\mathcal{Y}| \cdot 2^{H(Y^n|X^n)+nR+n\delta}$, which has the desired exponential order. Consider now a decoder that estimates i as the transmitted message if $x^n(i)$ is the only codeword for which $y^n \in T(x^n(i))$, and declares an error otherwise. This decoder gives small error average error probability as long as the ML decoder does so. The intuition is that in view of eq. (6), the average probability that Y^n would fall outside $T(x^n(i))$ given that i is the transmitted message, is small. Thus, the error probability can be large only for codebooks with large intersections among $\{T(x^n(i))\}$, but then the error probability would be large even for the ML decoder.

Example 2 - description via channel simulation. The channel description problem is intimately related to the following simulation problem [8]: Given an input process Q, a realization X^n of Q^n, and a channel W, construct $\hat{Y}^n = \phi(X^n, U^k)$, where ϕ is a deterministic map, and $U^k = (U_1, ..., U_k)$ is an independent vector of k purely random bits, such that the PMF $\hat{P}^n = Q^n \times \hat{W}^n$ of $(X^n, \hat{Y}^n)$ would be close to $P^n = Q^n \times W^n$ for large n. What is the minimum number of random bits k so that such a mapping ϕ exists? The answer in [8] is given in full generality for arbitrary channels. Confining it to stationary and ergodic channels, it tells that for large enough n, essentially $k = H(Y^n|X^n)$ bits suffice to keep

$$d(P^n, \hat{P}^n) = \max_{A \subseteq \mathcal{X}^n \times \mathcal{Y}^n} |P^n(A) - \hat{P}^n(A)| \tag{9}$$

arbitrarily small. Now, define z^N as a description of ϕ in the following manner: For each one of the $2^{nR} \times 2^k$ possible input pairs (X^n, U^k), use $n \log |\mathcal{Y}|$ bits to describe the corresponding value of $\phi(X^n, U^k)$. Thus, N is again, exponentially $2^{H(Y^n|X^n)+nR}$ bits. As a decoding metric, we shall use $D^{\hat{W}^n}$, i.e., the ML decoder w.r.t. $\hat{W}^n$. Following eq. (9) and the optimality of $D^{\hat{W}^n}$ w.r.t. $\hat{W}^n$,

$$\begin{aligned} \bar{P}_e(Q^n, W^n, D^{\hat{W}^n}) &\leq \bar{P}_e(Q^n, \hat{W}^n, D^{\hat{W}^n}) + \frac{\epsilon}{2} \\ &\leq \bar{P}_e(Q^n, \hat{W}^n, D^{W^n}) + \frac{\epsilon}{2} \\ &\leq \bar{P}_e(Q^n, W^n, D^{W^n}) + \epsilon, \end{aligned} \tag{10}$$

which is equivalent to (3).

These two deterministic description methods suffer from the same problem: In reality, it is inconceivable that while the channel is unknown, one would have full information about

all the conditionally typical sequences or the optimum channel simulator. In practice, a common way to learn an unknown channel is carried out using random training examples. Intuitively, if we have sufficiently many independent channel-output training examples for each input $x^n(i)$, $1 \leq i \leq M$, such that the conditional output type classes are 'well-covered', then this should suffice for reasonably good training of the decoder. Another reason for confining attention to description by training examples (see also, [5],[7],[10]) is that it is a stronger setting for proving achievability. To see this, note that the N-bit description corresponding to (a binary representation of) a training database is given by a *random* rather than a deterministic mapping F. Nonetheless, if we can show the existence of a good random mapping as such, this would imply that a good deterministic mapping also exists, by a simple 'random coding' argument: If the average error probability over the ensemble of training databases of length N is small, there must be a deterministic database of the same length, whose performance is at least as good.

For these two reasons, stating the achievability result in terms of random training data is more desirable, although it does not provide a constructive description strategy. Indeed, we next show that if one has at least $2^{H(Y^n|X^n)+n\delta}$ independent random training examples for each code word (and thus a training database of total size N exponentially at least $2^{H(Y^n|X^n)+n(R+2\delta)}$ bits), then the average error probability, w.r.t. both the ensemble of codes and training data, is small for every good channel.

Theorem 1 *Let δ and ϵ be fixed positive reals. Let $n \geq n_0(\delta)$, where $n_0(\delta)$ is an integer depending only on δ. Let Q^n be an arbitrary PMF on $\mathcal{X}^n$, let $H_0 \in (0, \log|\mathcal{Y}| - 5\delta)$, $R \in (0, \min\{\log|\mathcal{X}|, \log|\mathcal{Y}| - H_0\} - 5\delta]$, and let $\mathcal{C}_n$ be defined in eq. (8). Let $\mathcal{E}$ be a rate R, length n, random block code with $M = 2^{nR}$ codewords drawn independently w.r.t. Q^n. For a given randomly chosen codebook $\mathcal{E} = \{x^n(1), ..., x^n(M)\}$, let $Z^{MK} = \{Y_{ij}^n,\ i = 1, ..., M,\ j = 1, ..., K\}$, ($K$ positive integer) be a training set of random vectors in $\mathcal{Y}^n$, where each Y_{ij}^n is drawn independently according to $W^n(\cdot|x^n(i))$. Let $D_{Z^{MK}}(x^n(i), y^n) = -\log \hat{W}^n(y^n|x^n(i))$ be the decoding metric associated with Z^{MK}, where $\hat{W}^n(y^n|x^n(i)) = K^{-1}\sum_{j=1}^{K} I\{Y_{ij}^n = y^n\}$, $I\{\cdot\}$ being the indicator function. If $K \geq 2^{n(H_0+3\delta)}$, then for every $W^n \in \mathcal{C}_n$,*

$$E\{P_e(\mathcal{E}, W, D_{Z^{MK}})\} \leq 2\epsilon + 2^{-n\delta} + \exp_2[nR - 2^{n\delta}], \tag{11}$$

where the expectation is taken w.r.t. the ensemble of random codebooks $\mathcal{E}$ and the ensemble of training sets Z^{MK} given $\mathcal{E}$.

Since $2^{n(H_0+3\delta)}$ training vectors per code word are sufficient for the assertion of the theorem to hold, and since each training vector Y_{ij}^n can be described by $n\log|\mathcal{Y}|$ bits, then the theorem tells us that for $\mathcal{C}_n$ defined as above, $N(n) \leq n\log|\mathcal{Y}| \cdot 2^{n(H_0+R+3\delta)}$, provided that n is sufficiently large.

The remaining part of this section is devoted to the proof of Theorem 1.

Proof of Theorem 1. For a given $W^n \in \mathcal{C}_n$, let $\lambda_n \stackrel{\Delta}{=} H(Y^n|X^n)+2n\delta$, $\mathcal{G} = \{y^n : \ \log V^n(y^n) \leq -H(Y^n)+n\delta\}$, $\mathcal{G}_i = \{y^n : \ \log \hat{W}^n(y^n|x^n(i)) \geq -\lambda_n\}$, and consider an auxiliary threshold decoder $D'_{Z^{MK}}$ that operates as follows.

1. If $y^n \in \mathcal{G}^c$ or $y^n \in \cap_{i=1}^M \mathcal{G}_i^c$ or $y^n \in \mathcal{G}_i$ for two or more indices i, then declare an error.

2. If an error was not declared in Step 1 and hence $y^n \in \mathcal{G}_i$ for exactly one index i, then declare that i was the index of the transmitted message.

Obviously, one must know $V^n(\cdot)$, $H(Y^n|X^n)$, and $H(Y^n)$ (which are not assumed to be known) in order to implement this threshold decoder. Nevertheless, this is not an obstacle for the purpose of deriving an upper bound on $EP_e(\mathcal{E}, W^n, D)$, from the following consideration: Whenever the threshold decoder $D'_{Z^{MK}}$ does not declare an error (that is, it reaches Step 2), it estimates the same transmitted message as the decoder corresponding to $D_{Z^{MK}}$, defined in Theorem 1. Therefore, the error probability of the decoder of Theorem 1 is upper bounded by the error probability of the threshold decoder for every given codebook $\mathcal{E}$ and training set Z^{MK}. A-fortiori, this inequality relation is maintained after taking ensemble averages over $\mathcal{E}$ and Z^{MK}. It will therefore suffice to upper bound the average error probability $E\{P_e(\mathcal{E}, W^n, D'_{Z^{MK}})\}$ of the threshold decoder.

By symmetry of the random coding mechanism, we may assume without loss of generality, that the transmitted message is $i = 1$, and hence the average error probability associated with the threshold decoder is bounded as follows.

$$\begin{aligned} E\{P_e(\mathcal{E}, W^n, D'_{Z^{MK}})\} &= \Pr\left\{\mathcal{G}^c \bigcup \mathcal{G}_1^c \bigcup \left[\bigcup_{i=2}^M \mathcal{G}_i\right]\right\} \\ &\leq \Pr\{\mathcal{G}^c\} + \Pr\{\mathcal{G}_1^c\} + \sum_{i=2}^M \Pr\left\{\mathcal{G}_i \bigcap \mathcal{G}\right\} \\ &\leq \epsilon + \Pr\{\mathcal{G}_1^c\} + (M-1)\cdot \Pr\left\{\mathcal{G}_2 \bigcap \mathcal{G}\right\}, \end{aligned} \tag{12}$$

where the first inequality follows from the union bound, and the second inequality follows from the fact that $W^n \in \mathcal{W}_n(Q^n, \delta, \epsilon)$, and the fact that the average probability of $\mathcal{G}_i \cap \mathcal{G}$ is

the same for all $i \geq 2$, provided that $i = 1$ is the transmitted message. Let us focus on the term $\Pr\{\mathcal{G}_2 \cap \mathcal{G}\}$ first.

$$\begin{aligned}\Pr\left\{\mathcal{G}_2 \bigcap \mathcal{G}\right\} &= \sum_{\mathcal{X}^n \times \mathcal{G}} Q^n(x^n(2))V^n(y^n)\Pr\left\{\log \hat{W}^n(y^n|x^n(2)) \geq -\lambda_n | x^n(2), y^n\right\} \\ &= \sum_{\mathcal{X}^n \times \mathcal{G}} Q^n(x^n(2))V^n(y^n)\Pr\left\{\frac{1}{K}\sum_{j=1}^{K} 1\{Y_{2j}^n = y^n\} \geq 2^{-\lambda_n} | x^n(2), y^n\right\} \quad (13)\end{aligned}$$

where $\Pr\{\cdot|x^n(2), y^n\}$ is w.r.t. the distribution of $\{Y_{2j}^n \ j = 1, ..., K\}$, whereas $x^n(2)$ and y^n are held fixed. Let us classify the pairs $(x^n(2), y^n)$ in $\mathcal{X}^n \times \mathcal{G}$ into two complementary subsets T and T^c, where $T = \{(x^n(2), y^n) : \ W^n(y^n|x^n(2)) < 2^{-\lambda_n - n\delta}\}$. For every $(x^n(2), y^n) \in T$, the event $\frac{1}{K}\sum_{j=1}^{K} 1\{Y_{2j}^n = y^n\} > 2^{-\lambda_n}$ henceforth denoted by $\mathcal{F}$, is a large deviations event associated with the empirical mean of i.i.d. Bernoulli random variables being significantly larger than their expectation. Thus, the probability of $\mathcal{F}$ is upper bounded by [1]

$$\Pr\{\mathcal{F}|x^n(2), y^n\} \leq \exp_2\left[-KD(2^{-\lambda_n}||2^{-\lambda_n - n\delta})\right], \qquad (x^n(2), y^n) \in T, \quad (14)$$

where

$$\begin{aligned}D(2^{-\lambda_n}||2^{-\lambda_n - n\delta}) &= 2^{-\lambda_n}\log\frac{2^{-\lambda_n}}{2^{-\lambda_n - n\delta}} + (1 - 2^{-\lambda_n})\log\frac{1 - 2^{-\lambda_n}}{1 - 2^{-\lambda_n - n\delta}} \\ &\geq (n\delta - \log e)2^{-\lambda_n} \quad (15)\end{aligned}$$

and where we have used the fact that $\log x \geq (1 - 1/x)\log e$. Thus,

$$\Pr\{\mathcal{F}|x^n(2), y^n\} \leq \exp_2\left[K(n\delta - \log e)2^{-\lambda_n}\right], \qquad (x^n(2), y^n) \in T. \quad (16)$$

We then have

$$\begin{aligned}\Pr\{\mathcal{G}_2 \cap \mathcal{G}\} &= \sum_{(\mathcal{X}^n \times \mathcal{G}) \cap T} Q^n(x^n(2))V^n(y^n)\Pr\{\mathcal{F}|x^n(2), y^n\} + \\ &\quad \sum_{(\mathcal{X}^n \times \mathcal{G}) \cap T^c} Q^n(x^n(2))V^n(y^n)\Pr\{\mathcal{F}|x^n(2), y^n\} \\ &\leq \exp_2\left[K(n\delta - \log e)2^{-\lambda_n}\right] + \sum_{(\mathcal{X}^n \times \mathcal{G}) \cap T^c} Q^n(x^n(2))V^n(y^n). \quad (17)\end{aligned}$$

As for the second term on the right-most side of eq. (17), we have

$$\begin{aligned}\sum_{(\mathcal{X}^n \times \mathcal{G}) \cap T^c} Q^n(x^n(2))V^n(y^n) &= \sum_{(\mathcal{X}^n \times \mathcal{G}) \cap T^c} Q^n(x^n(2))W^n(y^n|x^n(2))2^{\log[V^n(y^n)/W^n(y^n|x^n(2))} \\ &\leq \sum_{(\mathcal{X}^n \times \mathcal{G}) \cap T^c} Q^n(x^n(2))W^n(y^n|x^n(2))2^{H(Y^n|X^n) - H(Y^n) + 4n\delta} \\ &\leq \sum_{\mathcal{X}^n \times \mathcal{Y}^n} Q^n(x^n(2))W^n(y^n|x^n(2))2^{-[I(X^n;Y^n) - 4n\delta]} \\ &= 2^{-[I(X^n;Y^n) - 4n\delta]}. \quad (18)\end{aligned}$$

In a similar manner, it is readily seen that

$$\begin{aligned}\Pr\{\mathcal{G}_1^c\} \quad &\leq \quad \exp_2[-KD(2^{-n\lambda_n}||2^{-\lambda_n+n\delta})] + \\ &\qquad \Pr\left\{(x^n(1), y^n):\ W^n(y^n|x^n(1)) < 2^{-\lambda_n+n\delta}\right\}, \end{aligned} \tag{19}$$

where $\Pr\{\cdot\}$ is w.r.t. $P^n = Q^n \times W^n$. Since the second term is less than ϵ (by definition of $\mathcal{W}_n(Q^n, \delta, \epsilon)$) and since $\lambda_n \geq 2n\delta$, this can be further upper bounded for $n \geq n_0(\delta)$ by

$$\Pr\{\mathcal{G}_1^c\} \leq \exp_2[-0.125K2^{-\lambda_n+n\delta}] + \epsilon. \tag{20}$$

Combining eqs. (12), (17), (18), and (20), we get, for all large n

$$EP_e(\mathcal{E}, W^n, D'_{ZMK}) \leq 2\epsilon + 2^{nR}\exp_2[-K2^{-\lambda_n}] + 2^{nR-I(X^n;Y^n)+4n\delta}. \tag{21}$$

Now, since $K \geq 2^{n(H_0+3\delta)} \geq 2^{\lambda_n+n\delta}$ for every $W^n \in \mathcal{C}_n$, the second term in (21) is upper bounded by $\exp_2[nR - 2^{n\delta}]$, and since $I(X^n;Y^n) \geq n(R+5\delta)$, the last term in (21) does not exceed $2^{-n\delta}$. This completes the proof of the Theorem 1.

4 The Converse Theorem

In this section, we state and prove the converse theorem, which tells us that under the conditions of Theorem 1, if $N < 2^{n(H_0+R-\delta)}$ the average probability of error must be large for some $W \in \mathcal{C}_n$, and so, $N(n)$ is at least as large as $2^{n(H_0+R-\delta)}$.

Our converse theorem is slightly more restrictive than the direct theorem in that it is confined to the definition of $\mathcal{C}_n$ w.r.t. the uniform[2] random coding distribution Q^n on $\mathcal{X}^n$ or on an arbitrary subset $\mathcal{A}_n$ of $\mathcal{X}^n$. On the other hand, it is stronger than the strict converse to Theorem 1 in two important aspects. First, it is stated for deterministic rather than randomized channel descriptions. Clearly, the nonexistence of a good deterministic mapping F for small N, implies that a good randomized mapping (whose performance is given by the expectation over deterministic ones) cannot exist either. Secondly, it claims that if N is not large enough, then not only the average error probability (w.r.t. the ensemble of codes) must be large for some $W^n \in \mathcal{C}_n$, but moreover, the error probability for *any deterministic code* (including the one optimized to the actual channel and a given decoder) must be large as well.

[2]In the absence of knowledge of W^n at the transmitter side, this is a natural choice of Q^n.

Theorem 2 *Let δ and ϵ be arbitrary positive reals, and let $n \geq n_0(\delta, \epsilon)$, where $n_0(\delta, \epsilon)$ is an integer that depends only on δ and ϵ. Let Q^n be the uniform distribution on $\mathcal{A}_n \subseteq \mathcal{X}^n$, where $|\mathcal{A}_n| = 2^{nA}$, $A \leq \log|\mathcal{X}|$. Fix $H_0 \in (0, \log|\mathcal{Y}| - 6\delta)$, $R \in (0, \min\{A, \log|\mathcal{Y}| - H_0\} - 6\delta]$, and let $\mathcal{C}_n$ be as in eq. (8). If $N \leq 2^{n(H_0+R-\delta)}$, then for any rate R block code $\mathcal{E}$ of block length n, any N-bit description $z^N = F(W^n)$, and any set of 2^N decoding metrics $\{D_{z^N}\}$, there exists $W^n \in \mathcal{C}_n$, such that $P_e(\mathcal{E}, W^n, D_{F(W^n)}) \geq 1 - \epsilon$.*

Discussion

Before we turn to the formal proof of Theorem 2, we discuss the intuition behind this result. We make an attempt to explain why there are 'complicated' channels whose description is so long, and what is the difference between these channels and the channels in [1], [2], [6], and [11], for which statistical side information is not needed, as explained in the Introduction.

The first important point is that the description length N is not due to the complexity of the actual channel W^n, but due to the richness of the class of allowed channels $\mathcal{C}_n$. A rich class corresponds to little prior knowledge on the variety of channels to be encountered. Obviously, if $\mathcal{C}_n$ contains one channel only, there is no need for statistical side information since the decoder can be designed optimally for this channel.

On the other hand, a rich class of channels might contain also 'simple' channels, yet the full price of description must be payed if it is not known in advance that the underlying channel is such. Consider the class of conditional PMF's defined by

$$W^n(y^n|x^n) = \begin{cases} 2^{-nH_0} & y^n \in \mathcal{B}(x^n) \\ 0 & \text{elsewhere} \end{cases} \tag{22}$$

where $\mathcal{B}(x^n)$, $x^n \in \mathcal{X}^n$, are subsets of $\mathcal{Y}^n$, with $|\mathcal{B}(x^n)| = 2^{nH_0}$ for all x^n. This can be thought of as an idealization of a certain stationary and ergodic channel that distributes evenly all the probability on the set of conditionally typical sequences $\mathcal{B}(x^n)$. Specifically, had we known ahead of time that the channel at hand is memoryless, then $\mathcal{B}(x^n)$ would be the set of all channel-output sequences for which the relative frequencies $\{\hat{p}(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\}$ are close to the joint probabilities $\{p(x, y)\}$. Because of this simple structure of $\mathcal{B}(x^n)$, if the decoder knew $\hat{p}$ within a reasonable accuracy, then all conditionally typical sets $\mathcal{B}(x^n(i))$ would have been essentially available by the appropriate permutations. Thus, in our context, $N(n)$ is some constant, and so the exponential order of $N(n)$ is zero. Moreover, as it turns out from [1] and [3], $N(n) = 0$ for the class of memoryless channels, because

universal decoders for memoryless channels are implicitly jointly estimating the channel and the transmitted message from y^n and $\mathcal{E}$.

This remains essentially true even for wider parametric families of channels, such as finite-state channels [11]. Feder and Lapidoth [2] show that for general parametric families, the price of universality is in multiplying $\bar{P}_e(Q^n, W^n, D^{W^n})$ by the 'effective number of distinct channels' in the class. This is because universal decoding can be carried out by interlacing optimum decoders of finitely many 'representative' channels in the class, and in the parametric case, the number of such channels is fairly small.

In contrast, as will be shown in the proof of Theorem 2 below, if $\mathcal{C}_n$ contains the set of *all* channels of the form (22) with *arbitrary* subsets $\mathcal{B}(x^n)$, then $N(n)$ must be at least of the exponential order of $2^{n(H_0+R)}$, since for most of the channels in this class, there is no simple structure that is explainable in a short message. This is because the number of degrees of freedom of this class of channels grows rapidly with n.

At this point, there is again a relation with channel simulation. In the proof of Theorem 2 below, we show that if N is not large enough, then for most of the channels of the form (22) the probability of error must be larger than $1-\epsilon$. Note that the channels of the form (22) can be represented as $Y^n = \phi(X^n, U^k)$ for some ϕ, where U^k is a vector of $k \leq nH_0$ independent fair coin tosses. This is exactly the set of all channel simulators with at most H_0 random bits per symbol as discussed in Example 2 above. In other words, the set of channels given in (22) covers, within variational distance less than ϵ, the class of all channels for which $H(Y^n|X^n)$ is essentially less than nH_0, and hence covers also $\mathcal{C}_n$. Note that for every representative channel (22) with average error probability larger than $1-\epsilon$, all channels in the ϵ-neighborhood of this representative would yield average probability of error larger than $1-2\epsilon$. This follows again from the fact that small variational distance corresponds to uniform closeness of probabilities of events. Thus, in a certain sense we can say that the converse it strong in that it holds for 'most' channels in $\mathcal{C}_n$ at the same time.

In summary, while in Example 2 we have demonstrated that essentially $N = 2^{n(H_0+R)}$ are *sufficient* for describing channels that are simulateable by H_0 bits per symbol, here we see that this description length is also *necessary* for these channels.

The remaining part of this section is devoted to the proof of Theorem 2.

Proof of Theorem 2. Similarly as in [5], [7], and [10], the proof employs a 'sphere covering' argument. Since there are finitely many possible encoders and a finitely decoders $\{D_{z^N}\}$,

the number of encoder-decoder pairs is obviously finite as well. We will first show that almost all channels of the form (22) are in $\mathcal{C}_n$. Then, we upper bound the number of such channels that can be 'covered' by a single encoder-decoder pair in the sense that the probability of error is less than $1-\epsilon$. Finally, we show that if N is not large enough then the overall number of covered channels is smaller than the total number of channels (22) in $\mathcal{C}_n$. Therefore, there must be channels for which the probability of error is larger than $1-\epsilon$.

Let $B \stackrel{\Delta}{=} \log|\mathcal{Y}|$, fix $\epsilon > 0$, $\delta > 0$, $H_0 \in (0, B-5\delta)$, $R \in (0, \min\{A, B-H_0\} - 6\delta]$, let $n \geq n_0(\delta, \epsilon)$, and consider the class $\mathcal{L}_n$ of all channels of the form (22). Clearly, every channel in $\mathcal{L}_n$ satisfies $H(Y^n|X^n) = nH_0$ as well as eq. (6). Thus, for such a channel to be in $\mathcal{C}_n$, the only additional requirements are eq. (7) and

$$I(X^n; Y^n) \geq n(R + 5\delta). \tag{23}$$

Although not all channels in $\mathcal{L}_n$ satisfy these requirements, we now demonstrate that for large n, most of them do, and so they are members of $\mathcal{C}_n$.

First, observe that for every channel in $\mathcal{L}_n$, $I(X^n; Y^n) = H(Y^n) - nH_0$, where $H(Y^n)$ is defined w.r.t. the uniform input distribution Q^n over $\mathcal{A}_n$. Thus, eq. (23) is equivalent to

$$H(Y^n) \geq n(H_0 + R + 5\delta). \tag{24}$$

To show that most channels of $\mathcal{L}_n$ satisfy eqs. (7) and (24), we consider the uniform probability distribution on $\mathcal{L}_n$, and show that a randomly chosen $W^n \in \mathcal{L}_n$ satisfies both requirements with high probability. At this point, we make a distinction between two cases according to (i) $H_0 \geq B - A - \delta/4$ or (ii) $H_0 < B - A - \delta/4$.

Consider case (i) first. We will show that for large n, most channels of $\mathcal{L}_n$ satisfy $V^n(y^n) \leq 2^{-n(B-\delta/2)}$ *simultaneously* for all $y^n \in \mathcal{Y}^n$. Since $R \leq B - H_0 - 6\delta$, this implies that eq. (24) holds, and since $H(Y^n) \leq nB$, it would guarantee also that eq. (7) is met. For a given $W^n \in \mathcal{L}_n$, it is straightforward to see that $V^n(y^n) = 2^{-n(A+H_0)} J_W(y^n)$, where $J_W(y^n)$ is the number of input vectors $x^n \in \mathcal{A}_n$ for which $\mathcal{B}(x^n)$ includes y^n. Thus, it will be sufficient to show that for most channels of $\mathcal{L}_n$, $J_W(y^n) \leq 2^{n(A+H_0-B+\delta/2)}$ simultaneously for all $y^n \in \mathcal{Y}^n$. Let $J_W(x^n, y^n)$ denote the indicator function of the event $\{W^n \in \mathcal{L}_n : \ y^n \in \mathcal{B}(x^n)\}$, and so, $J_W(y^n) = \sum_{x^n \in \mathcal{A}_n} J_W(x^n, y^n)$. Clearly,

$$\Pr\{W^n \in \mathcal{L}_n : \ y^n \in \mathcal{B}(x^n)\} = E J_W(x^n, y^n) = \frac{2^{nH_0}}{2^{nB}} = 2^{-n(B-H_0)} \tag{25}$$

for every x^n and y^n. Since the subsets $\{B(x^n)\}$ are drawn independently and equiprobably under the above defined probability distribution on $\mathcal{L}_n$, then $J_W(y^n)$ is the sum of i.i.d. Bernoulli random variables $\{J_W(x^n, y^n)\}$. Thus, the event $\{W^n \in \mathcal{L}_n : \ J_W(y^n) \geq 2^{n(A+H_0-B+\delta/2)}\}$ for a given y^n, is a large deviations event whose probability is upper bounded by

$$\Pr\{W^n : \ J_W(\boldsymbol{y}) \geq 2^{n(A+H_0-B+\delta/2)}\} \ \leq \ \exp_2[-2^{nA} D(2^{-n(B-H_0-\delta/2)}||2^{-n(B-H_0)})], \quad (26)$$

which decays double-exponentially rapidly like $\exp_2[-2^{n(A+H_0+\delta/2-B)}] \leq \exp_2[-2^{n\delta/4}]$ under the assumption of case (i) (see eqs. (14), (15) for a similar derivation). Because of this double-exponential decay rate and by the union bound, the probability continues to go to zero, even if the above event is extended and defined for *some* $y^n \in \mathcal{Y}^n$ rather than for a *fixed* y^n. Thus, we have shown that most channels in $\mathcal{L}_n$ give $J_W(y^n) \leq 2^{n(A+H_0-B+\delta/2)}$ simultaneously for all y^n.

Turning now to case (ii), observe that only subsets of size $2^{n(H_0+A)}$ can be obtained at the channel output space. Using the same technique, it is sufficient to prove and easy to see, that with high probability w.r.t. the random choice of $W^n \in \mathcal{L}_n$, each one of the nonzero-probability output vectors y^n satisfies $V^n(y^n) \leq 2^{-n(H_0+A-\delta/2)}$, and so, $H(Y^n) \geq n(H_0 + A - \delta/2) > n(H_0 + R + 5\delta)$. At the same time, since $H(Y^n) \leq n(A + H_0)$, eq. (7) is again satisfied.

We have seen that in both case (i) and case (ii), most of the channels in $|\mathcal{L}_n|$, are in $\mathcal{C}_n$ for large n. A conservative estimate in either case would be

$$|C_n \cap \mathcal{L}_n| \geq \frac{|\mathcal{L}_n|}{2} = \frac{1}{2} S_1^{2^{nA}} \quad (27)$$

where

$$S_1 \triangleq \left(\begin{array}{c} 2^{nB} \\ 2^{nH_0} \end{array} \right). \quad (28)$$

We next upper bound the number of channels N_W in $\mathcal{L}_n$ for which a given encoder-decoder pair provides error probability less than $1 - \epsilon$. Fix an encoder $\mathcal{E} = \{x^n(1), ..., x^n(M)\}$, and a decoder $\mathcal{D}$ that corresponds to a certain partition of $\mathcal{Y}^n$ into $M = 2^{nR}$ decision regions $\Lambda_1, ..., \Lambda_M$. For $P_e(\mathcal{E}, W^n, D)$ to be less than $1 - \epsilon$, at least $0.5\epsilon 2^{nR}$ decision regions $\{\Lambda_i\}$ must satisfy $\Pr\{\Lambda_i^c | x^n(i)\} \leq 1 - \epsilon/2$, where $\Pr\{\cdot\}$ is defined w.r.t. W^n. Also, since $\sum_{i=1}^{M} |\Lambda_i| = 2^{nB}$, the number of decision regions for which $|\Lambda_i| \leq 2^{n(B-\epsilon)}$ is at least $2^{nR} - 2^{n\epsilon}$. It follows that the number of decision regions that satisfy both $\Pr\{\Lambda_i^c | x^n(i)\} \leq 1 - \epsilon/2$ and

$|\Lambda_i| \leq 2^{n(B-\epsilon)}$ is at least $0.5\epsilon 2^{nR} - 2^{n\epsilon} \geq 0.5\epsilon(1-\epsilon)2^{nR}$ for all large enough n. Since the channels in $\mathcal{L}$ induce a uniform distribution on each $\mathcal{B}(x^n)$ given x^n, the requirement $\Pr\{\Lambda_i^c | x^n(i)\} \leq 1 - \epsilon/2$ is equivalent to the requirement that a fraction at least as large as $\epsilon/2$ of the vectors in $\mathcal{B}(x^n(i))$ would fall in Λ_i. Now, for every decision region Λ_i of size less than $2^{n(B-\epsilon)}$ the number of combinations S_2 of choosing an output subset $\mathcal{B}(x^n(i))$ with a fraction of vectors at least $\epsilon/2$ in Λ_i, is upper bounded by

$$\begin{aligned}
S_2 &\leq \sum_{i=0.5\epsilon 2^{nH_0}}^{2^{nH_0}} \left(\begin{array}{c} 2^{n(B-\epsilon)} \\ i \end{array} \right) \left(\begin{array}{c} 2^{nB} \\ 2^{nH_0} - i \end{array} \right) \\
&\leq 2^{nH_0} \left(\begin{array}{c} 2^{n(B-\epsilon)} \\ 0.5\epsilon 2^{nH_0} \end{array} \right) \left(\begin{array}{c} 2^{nB} \\ (1-0.5\epsilon)2^{nH_0} \end{array} \right) \\
&\leq 2^{nH_0} \exp_2\{0.5\epsilon 2^{nH_0}[n(B-\epsilon-H_0) + \log e]\} \exp_2\{(1-0.5\epsilon)2^{nH_0}[n(B-H_0) + \log e]\} \\
&= \exp_2\{nH_0 + 2^{nH_0}[n(B - H_0 - 0.5\epsilon^2) + \log e]\}, \qquad (29)
\end{aligned}$$

where for the second inequality we have used the fact that for large n, the greatest summand corresponds to $i = 0.5\epsilon 2^{nH_0}$ and the following step follows from the inequality (see, e.g., [7])

$$\log \left(\begin{array}{c} n \\ m \end{array} \right) \leq m \log \left(\frac{en}{m} \right). \qquad (30)$$

For each one of the remaining $2^{nA} - 0.5\epsilon(1-\epsilon)2^{nR}$ channel input vectors that are not corresponding to decision regions of cardinality less than $2^{n(B-\epsilon)}$ and conditional probability of error less than $1 - \epsilon/2$, we allow free choice of the output subset, resulting in at most S_1 combinations, where S_1 is defined as in eq. (28). Finally, the total number of channels N_W that satisfy the necessary conditions for error probability less than $1 - \epsilon$, for a given encoder-decoder pair is upper bounded by

$$\begin{aligned}
N_W &\leq \left(\begin{array}{c} 2^{nR} \\ 0.5\epsilon(1-\epsilon)2^{nR} \end{array} \right) S_2^{0.5\epsilon(1-\epsilon)2^{nR}} \cdot S_1^{2^{nA} - 0.5\epsilon(1-\epsilon)2^{nR}} \\
&\leq 2^{2^{nR}} S_2^{0.5\epsilon(1-\epsilon)2^{nR}} \cdot S_1^{2^{nA} - 0.5\epsilon(1-\epsilon)2^{nR}}, \qquad (31)
\end{aligned}$$

where for an upper bound we have ignored the fact that the counted channels should be restricted to $\mathcal{C}_n \cap \mathcal{L}_n$. Let $N_E = 2^{nA2^{nR}}$ denote the number of different encoders and let $N_D = 2^N \leq 2^{2^{n(H_0+R-\delta)}}$ denote the number of different decoders. Clearly, if we show that

$$|\mathcal{L}_n \cap \mathcal{C}_n| > N_W \cdot \exp_2[2^{n(H_0+R-\delta)}] \cdot \exp_2(nA2^{nR}) \geq N_W N_E N_D \qquad (32)$$

this will imply that there must be channels that are not 'covered' by any encoder-decoder pair in the sense of yielding error probability less than $1-\epsilon$, and hence for these channels the

error probability must be larger than $1-\epsilon$. This follows from the following consideration.

$$\frac{|\mathcal{L}_n \cap \mathcal{C}_n|}{N_W N_E N_D} \geq \frac{1}{2}\exp_2[-(nA2^{nR} + 2^{n(H_0+R-\delta)} + 2^{nR})] \cdot \left[\frac{S_1}{S_2}\right]^{0.5\epsilon(1-\epsilon)2^{nR}}. \quad (33)$$

But following the inequality

$$\left(\begin{array}{c} n \\ m \end{array} \right) \geq 2^{m\log(n/m)}, \quad (34)$$

we have

$$S_1 = \left(\begin{array}{c} 2^{nB} \\ 2^{nH_0} \end{array} \right) \geq 2^{2^{nH_0} n(B-H_0)}, \quad (35)$$

and therefore by eq. (29)

$$\frac{S_1}{S_2} \geq \exp_2[2^{nH_0}(0.5\epsilon^2 n - \log e) - nH_0], \quad (36)$$

which in turn implies that

$$\frac{|\mathcal{L}_n \cap \mathcal{C}_n|}{N_W N_E N_D} \geq \frac{1}{2}\exp_2\left[(0.5\epsilon^2 n - \log e)0.5\epsilon(1-\epsilon)2^{n(H_0+R)} - 2^{n(H_0+R-\delta)} - (nA + nH_0 + 1)2^{nR}\right]. \quad (37)$$

Not only the last expression is larger than 1 for large enough n, it grows double-exponentially as fast as $2^{2^{n(H_0+R)}}$. In other words, not only there exists a channel in $\mathcal{L}_n \cap \mathcal{C}_n$, but for *most* members of $\mathcal{L}_n \cap \mathcal{C}_n$, the probability of error is larger than $1-\epsilon$. This completes the proof of Theorem 2.

Acknowledgement

The very useful comments of the reviewers are greatly appreciated.

References

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

[2] M. Feder and A. Lapidoth, "Universal decoders for channels with memory," preprint 1996.

[3] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Contr. Inform. Theory*, vol. 4, pp. 97-102, 1975.

[4] R. M. Gray, *Entropy and Information Theory*, Springer Verlag 1990.

[5] Y. Hershkovits and J. Ziv, "On fixed-database universal data compression with limited memory," submitted for publication.

[6] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. IT-39, no. 4, pp. 1261-1269, July 1993.

[7] N. Merhav and J. Ziv, "On the amount of statistical side information required for lossy data compression," to appear in *IEEE Trans. Inform. Theory.*

[8] Y. Steinberg and S. Verdú, "Channel simulation and coding with side information," *IEEE Trans. Inform. Theory*, vol. IT-40, no. 3, pp. 634-646, May 1994.

[9] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, Vol. IT-40, no. 4, pp. 1147-1157, July 1994.

[10] A. D. Wyner and J. Ziv, "Classification with finite memory," *IEEE Trans. Inform. Theory*, Vol. IT-42, no. 2, pp. 337-347, March 1996.

[11] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 4, pp. 453-460, July 1985.