$m \in M_{\hat{f}}$

$$\begin{aligned} &\text{i)} \quad \hat{f}(m) \in T^k_{[P_{Y_1}]} \subset \mathcal{Y}_1^k \\ &\text{ii)} \quad \hat{\varphi}^{-1}(m) \subset T^k_{[\hat{V}]}(\hat{f}(m)), \end{aligned} \tag{75}$$

and the code $(\hat{f}, \hat{\varphi})$ has no extension with the same properties. For each $m \in M_{\hat{f}}$ let $C(m) \subset \hat{\varphi}^{-1}(m)$, be such that

$$\hat{V}(C(m) \mid \hat{f}(m)) \geq \eta. \tag{76}$$

Then

$$P_X\left(\bigcup_{m \in M_{\hat{f}}} C(m)\right) \geq 2^{-k\epsilon_k} \tag{77}$$

where $\epsilon_k \to 0$ as $k \to \infty$.

Proof: Let $\boldsymbol{x} \in C(m) \subset T^k_{[V]}(\hat{f}(m))$. Then $\boldsymbol{x} \in T_{[P_X]_{2\delta_k}}$ where P_X is the distribution induced by P_{Y_1} and $\hat{V}$. This implies the existence of a sequence $\epsilon'_k \to 0$ such that

$$P_X^k(\boldsymbol{x}) \geq 2^{-k(H(P_X)+\epsilon'_k)}. \tag{78}$$

On the other hand, (76) implies [4, Lemma 1.2.14]

$$|C(m)| \geq 2^{k(H(\hat{V}|P_{\boldsymbol{y}^{(m)}})-\epsilon''_k)} \tag{79}$$

where $\epsilon''_k \to 0$ as $k \to \infty$ and $\boldsymbol{y}^{(m)} = \hat{f}(m) \in T^k_{[P_{Y_1}]}$. Since $\boldsymbol{y}^{(m)}$ is P_{Y_1}-typical

$$|H(\hat{V} \mid P_{\boldsymbol{y}^{(m)}}) - H(\hat{V} \mid P_{Y_1})| \leq \delta_k |\mathcal{Y}_1| \log |\mathcal{X}| \triangleq \epsilon'''_k \tag{80}$$

implying

$$|C(m)| \geq 2^{k(H(\hat{V}|P_{Y_1})-\epsilon''_k-\epsilon'''_k)}. \tag{81}$$

Combining (87) and (81) we obtain

$$P_X(C(m)) \geq 2^{-k(I(P_{Y_1}, \hat{V})+\epsilon'_k+\epsilon''_k+\epsilon'''_k)}. \tag{82}$$

Finally, since $C(m) \cap C(m') = \varnothing$, $m \neq m'$ and [4, Lemma 2.1.13]

$$|M_{\hat{f}}| \geq 2^{k(I(P_{Y_1}, \hat{V})+-\epsilon''''_k)}, \tag{83}$$

$$P_X\left(\bigcup_{m \in M_{\hat{f}}} C(m)\right) \geq |M_{\hat{f}}| 2^{k(I(P_{Y_1}, \hat{V})+\epsilon'_k+\epsilon''_k+\epsilon'''_k)} \tag{84}$$

$$\geq 2^{k(\epsilon'_k+\epsilon''_k+\epsilon'''_k+\epsilon''''_k)}. \tag{85}$$

□

ACKNOWLEDGMENT

The author is particularly indebted to Prof. T. Cover for the wonderful time spent with his group at Stanford where he learned about rate distortion theory and successive refinement. He also wishes to thank his colleague, M. Miller, whose questions about Equitz and Cover's result initiated this work. Finally, he would like to thank Prof. T. Berger, H. Yamamoto, and J. O'Sullivan for helpful comments and, in particular, an anonymous reviewer for pointing out the alternative proof of Theorem 1.

REFERENCES

[1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[2] R. E. Blahut, *Theory and Practice of Information Theory.* Reading, MA: Addison-Wesley, 1987.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

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

[5] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[6] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding.* New York: McGraw-Hill, 1979.

[7] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, pp. 269–274, Mar. 1991.

[8] H. S. Witsenhausen and A. D. Wyner, "Source coding for multiple description II: A binary source," *Bell Syst. Tech. J.*, vol. 60, no. 10, pp. 2281–2292, Dec. 1981.

[9] J. K. Wolf, A. D. Wyner, and J. Ziv, "Source coding for multiple description," *Bell Syst. Tech. J.*, vol. 60, no. 10, pp. 2281–2292, Dec. 1981.

[10] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, no. 10, pp. 1909–1921, Dec. 1980.

[11] A. El Gamal and T. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 851–857, Nov. 1982.

[12] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 502–521, July 1987.

[13] R. Ahlswede, "The rate-distortion region for multiple descriptions without excess rate," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 721–726, Nov. 1985.

[14] R. M. Gray and A. D. Wyner, "Source coding for a simple network," *Bell Syst. Tech. J.*, vol. 53, no. 9, pp. 1681–1721, Nov. 1974.

[15] H. Yamamoto, "Source coding theory for cascade and branching communication systems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 299–308, May 1981.

[16] B. Rimoldi, "Successive refinement of information: Characterisation of the achievable rates," Tech. Rep. #91-44, Electron. Syst. and Signals Res. Lab., Elec. Eng. Dept., Washington Univ., St. Louis, MO.

Relations Between Entropy and Error Probability

Meir Feder, *Senior Member, IEEE*, and
Neri Merhav, *Senior Member, IEEE*

***Abstract*— The relation between the entropy of a discrete random variable and the minimum attainable probability of error made in guessing its value is examined. While Fano's inequality provides a tight *lower* bound on the error probability in terms of the entropy, we derive a converse result—a tight *upper* bound on the minimal error probability in terms of the entropy. Both bounds are sharp, and can draw a relation, as well, between the error probability for the maximum *a posteriori* (MAP) rule, and the conditional entropy (equivocation), which is a useful uncertainty measure in several applications. Combining this relation and the classical channel coding theorem, we present a channel coding theorem for the equivocation which, unlike the channel coding theorem for error probability, is meaningful at all rates. This theorem is proved directly for DMC's, and from this proof it is further concluded that for $R \geq C$ the equivocation achieves its minimal value of $R - C$ at the rate of $n^{1/2}$ where n is the block length.**

***Index Terms*—Entropy, error probability, equivocation, predictability, Fano's inequality, channel coding theorem.**

Manuscript received July 20, 1992; revised March 22, 1993. This work was supported in part by the Wolfson Research Awards administrated by the Israel Academy of Science and Humanities, Tel-Aviv University, Tel-Aviv, Israel.

M. Feder is with the Department of Electrical Engineering—Systems, Tel-Aviv University, Tel-Aviv, 69978, Israel.

N. Merhav is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, 32000, Israel.

IEEE Log Number 9215124.

I. Introduction

Intuitively, the entropy H of a random variable measures its complexity, or its degree of randomness. It seems plausible that the higher the entropy the harder it is to predict the value taken by this random variable. If the money made in gambling on the predicted value is a criterion for good prediction, this intuitive notion is affirmed by the observation (see, e.g., [10], [3]) that the optimal capital growth rate achievable by gambling on the outcome of, say, a binary random variable is $1-H$, i.e., the smaller the entropy the larger is the achievable capital growth rate. However, the degree of difficulty in predicting the value of the random variable is more naturally assessed by the minimal possible error probability associated with any prediction procedure. As was observed in [5], this prediction error is not uniquely determined by the entropy, i.e., two random variables with the same entropy may have different minimal prediction error probabilities.

In this work we further explore the relationship between the entropy of a random variable and the minimal error probability in guessing its value. While the well-known Fano inequality provides a tight *lower* bound on the error probability in terms of the entropy, we derive a converse result—a tight *upper* bound on the minimal error probability in terms of the entropy. This converse result is known in the binary case, see, e.g., [1] and [8], but we derive here the bound for the general case and show that it is tight. Since both Fano's inequality and the new bound are sharp, they determine the region of all allowable pairs of entropy and minimal error probability. These bounds are also applied to conditional entropies and the error probabilities obtained in the maximum *a posteriori* (MAP) rule: thus they also draw a relation between the entropy rate of a process (the process compressibility) and the minimal expected fraction of errors made by predicting its future outcome (the process predictability). Similar relations exist between the minimal average fraction of errors made in sequential prediction of sequences from a given set, and the size of the set.

While the entropy is the basic measure of uncertainty used in information theory, the channel coding theorems are usually stated in terms of the error probability. The relation between entropy and error probability allows us to state these theorems in terms of the entropy. In this work we prove directly the channel coding theorem for discrete memoryless channels (DMC's) using the conditional entropy of the channel input given the channel output (equivocation) as the desired error measure. Unlike the standard coding theorem, this coding theorem is relevant in describing the behavior of information transmission at rates below capacity, at capacity, and above the channel capacity.

Let us first recall the definitions of the entropy and the minimal error probability. Let X be a random variable over the alphabet $\{1,\cdots,M\}$, and suppose its probability distribution $\{p(x)\}_{x=1}^{M}$ is given. The entropy of the random variable is

$$H(X) = -\sum_{x=1}^{M} p(x)\log p(x) \tag{1}$$

where throughout the paper $\log = \log_2$ and the entropy is measured in bits. In the absence of any other knowledge regarding X, the estimator of X that minimizes the error probability is the value $\hat{x}$ with the highest probability. Let $\hat{p} = p(\hat{x}) = \max_x p(x)$. The minimal error probability in guessing the value of X is thus,

$$\pi(X) = \sum_{x\neq\hat{x}} p(x) = 1-\hat{p}. \tag{2}$$

The maximal entropy over an alphabet of size M is $\log M$, while the highest possible minimal error probability is $(M-1)/M$, both attained by a uniform random variable. On the other extreme, a random variable for which the entire probability mass is concentrated on a single value, has both a zero entropy and a zero minimal error probability.

The "uncertainty" in X given another random variable Y is usually assessed by the conditional entropy, or the *equivocation*. Let Y be a random variable (or vector) over an arbitrary sample space $\mathcal{Y}$ with a well-defined probability distribution $P(y)$, such that for each $y\in\mathcal{Y}$ (with a possible exception of a zero measure set), a probability mass function $p(\cdot|y)$ is well defined. Then we define the equivocation as

$$H(X\,|\,Y) = -\int_{\mathcal{Y}}\sum_{x} p(x|y)\log p(x|y)\,dP(y). \tag{3}$$

The minimum probability of error in estimating X given an observation y of Y is attained by the maximum *a posteriori* (MAP) estimator, i.e., by $\hat{x}(y) = \arg\max_x p(x|y)$. Thus, the expected minimal error probability is

$$\pi(X|Y) = \int_{\mathcal{Y}}\left[1-\max_x p(x|y)\right]dP(y). \tag{4}$$

Let $\mathcal{X} = \{X_t\}_{t=-\infty}^{\infty}$ be a stationary ergodic random process. The *entropy rate* of this process is given by

$$\mathcal{H}(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1},\cdots,X_1). \tag{5}$$

Similarly, we define the *predictability* of the process as

$$\Pi(\mathcal{X}) = \lim_{n\to\infty} \pi(X_n|X_{n-1},\cdots,X_1) \tag{6}$$

where this quantity is the expected minimal error probability in predicting the future value of the process given its past. The limits in (5) and (6) exist since both the conditional entropy and the predictability are positive and monotonically nonincreasing with n.

In the next section we present the bounds and the relation between the entropy and the minimal error probability. Despite the fact [5] that there is no one-to-one relation between the entropy and the minimal error probability, the bounds affirm that a variable is totally random (i.e., its entropy $\log M$) iff it is totally unpredictable (i.e., its minimal error probability is $((M-1)/M)$ and conversely, a variable is totally redundant (i.e., its entropy is zero) iff it is fully predictable (its minimal probability of error is zero). In Section III, the relations are applied to derive a bound on the fraction of errors made by arbitrary predictors over a set of arbitrary sequences. Finally, in the last section, we present a channel coding theorem in terms of the equivocation. This theorem could have been derived by combining the classical coding theorem which deals with the error probability and the relations presented here. We chose to develop in this work a direct proof, which we believe provides more insight on the behavior of the equivocation at rates equal or greater than the channel capacity.

II. The Bounds

Consider first a discrete random variable X taking values in the set $\{1,\cdots,M\}$ with probabilities $p(1), p(2),\cdots,p(M)$, and assume without loss of generality that

$$p(1)\geq p(2)\geq\cdots\geq p(M).$$

We define the probability vector $\boldsymbol{p} = [p(1),\cdots,p(M)]$, and use interchangeably the notation $H(\boldsymbol{p})$ or $H(X)$ for the entropy and similarly we use interchangeably $\pi(\boldsymbol{p})$ or $\pi(X)$ for the minimal error probability (or the predictability). Note that $p(1) = 1-\pi(\boldsymbol{p})$.

Clearly given π we can bound the entropy as

$$\max_{\boldsymbol{p}\in P_\pi} H(\boldsymbol{p}) \geq H(X) \geq \min_{\boldsymbol{p}\in P_\pi} H(\boldsymbol{p}) \tag{7}$$

where P_π is the set of all vectors $\boldsymbol{p}$ such that $p(i)\geq 0\,\forall i$, $\sum_i p(i) = 1$ and $p(1) = 1-\pi$. As shown in the following two lemmas, the maximization and minimization in (7) can be solved explicitly.

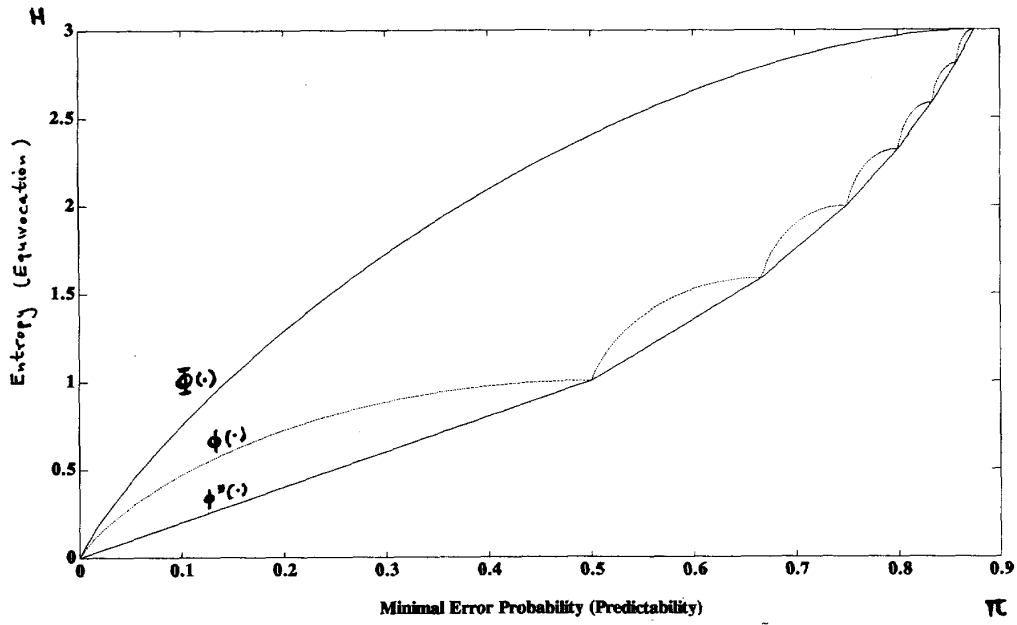

Fig. 1. The functions $\Phi(\cdot)$, $\phi(\cdot)$, and $\phi^*(\cdot)$, and the region $\tilde{A}$.

Lemma 1: The maximum in (7) is achieved by

$$\boldsymbol{p}_{\max}(\pi) = \left[1-\pi, \frac{\pi}{M-1}, \cdots, \frac{\pi}{M-1}\right], \tag{8}$$

and the corresponding maximum entropy is

$$\Phi(\pi) = H(\boldsymbol{p}_{\max}(\pi)) = h(\pi) + \pi \log (M-1) \tag{9}$$

where $h(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ is the binary entropy function.

Note that for any random variable X over an alphabet of size M, Eq. (9) implies that

$$H(X) \le h(\pi) + \pi \log (M-1), \tag{10}$$

which is a special case of Fano's inequality. The proof of Lemma 1 is straightforward and is given, for example, in [4, p. 39 and p. 48]. In fact, the proof in [4] was provided to show that Fano's inequality is sharp.

Lemma 2: The minimum in (7) is achieved by $\boldsymbol{p}_{\min}(\pi) = [p(1), \cdots, p(M)]$ where

$$\begin{array}{ll} p(1) = 1-\pi, \quad p(2) = \pi, & \\ p(3) = \cdots = p(M) = 0, & 0 \le \pi \le \frac{1}{2} \\ & \\ p(1) = p(2) = 1-\pi, \quad p(3) = 2\pi - 1, & \\ p(4) = \cdots = p(M) = 0, & \frac{1}{2} \le \pi \le \frac{2}{3} \\ \vdots & \vdots \\ p(1) = \cdots = p(M-1) = 1-\pi, & \\ p(M) = 1-(M-1)(1-\pi), & \frac{M-2}{M-1} \le \pi \le \frac{M-1}{M} \end{array} \tag{11}$$

and the corresponding minimum entropy $\phi(\pi)$ is shown in (12) at the bottom of the page.

This lemma is easily shown by straightforward verification of the Kuhn–Tucker conditions. Note that $\phi(\pi)$ is a continuous function with a piecewise continuous derivative, composed of $M-1$ concave segments where the ith segment is composed of a linear term with a slope $-i \log i$ and a concave binary entropy function whose argument takes values in the interval $[0, 1/(i+1)]$. The lower and upper bounds on the entropy for any value of minimum error probability are depicted in Fig. 1, for $M = 8$.

Since both $\phi(\cdot)$ and $\Phi(\cdot)$ are strictly monotonically increasing continuous functions, they have well defined inverses. Thus, if the

$$\phi(\pi) = H(\boldsymbol{p}_{\min}(\pi)) = \begin{cases} h(\pi), & 0 \le \pi \le \frac{1}{2} \\ 2(1-\pi) + h(2\pi - 1), & \frac{1}{2} \le \pi \le \frac{2}{3} \\ \vdots & \vdots \\ i \log i (1-\pi) + h(i\pi - (i-1)), & \frac{i-1}{i} \le \pi \le \frac{i}{i+1} \\ \vdots & \vdots \\ (M-1)\log(M-1)(1-\pi) + h((M-1)\pi - M + 2), & \frac{M-2}{M-1} \le \pi \le \frac{M-1}{M}. \end{cases} \tag{12}$$

entropy of the random variable is known, say H, we can find upper and lower bounds on the minimal error probability $\pi(X)$, i.e.,

$$\phi^{-1}(H) \geq \pi(X) \geq \Phi^{-1}(H). \tag{13}$$

Consider now the relation between the conditional entropy (equivocation) $H(X|Y)$ and the MAP error probability, $\pi(X|Y)$. It will be useful to define the following function

$$\phi^*(\pi) = \begin{cases} a_1\pi + b_1 & 0 \leq \pi \leq \frac{1}{2} \\ a_2\left(\pi - \frac{1}{2}\right) + b_2 & \frac{1}{2} \leq \pi \leq \frac{2}{3} \\ \quad\vdots & \quad\vdots \\ a_i\left(\pi - \frac{i-1}{i}\right) + b_i & \frac{i-1}{i} \leq \pi \leq \frac{i}{i+1} \\ \quad\vdots & \quad\vdots \\ a_{M-1}\left(\pi - \frac{M-2}{M-1}\right) + b_{M-1} & \frac{M-2}{M-1} \leq \pi \leq \frac{M-1}{M}, \end{cases} \tag{14}$$

where $a_i = i(i+1)\log((i+1)/i)$, that is $a_1 = 2 < a_2 = 6\log(3/2) < \cdots < a_{M-1}$, and $b_i = \log i$. This function, composed of $M-1$ piecewise linear segments is continuous and convex. It is the largest convex function that is still smaller than or equal to $\phi(x)$, for $0 \leq x \leq (M-1)/M$. It coincides with $\phi(x)$ at $x = 0, 1/2, 2/3, \cdots, (M-1)/M$ where it takes the values $0, 1, \log 3, \cdots, \log M$. It also coincides with $\Phi(x)$ at $x = 0$ and $x = (M-1)/M$ where both functions take the values 0 and $\log M$, respectively. Actually, if $\mathcal{A}$ is the set of all points in the $\pi - H$ plane that satisfy (13), the convex hull of $\mathcal{A}$, denoted $\tilde{\mathcal{A}}$, is the set of all points for which

$$\Phi(\pi) \geq H \geq \phi^*(\pi). \tag{15}$$

Having these definitions, we present the following theorem:

Theorem 1: The equivocation and the MAP error probability lie in the set $\tilde{\mathcal{A}}$ in the $\pi - H$ plane, i.e., the equivocation can be bounded in terms of the MAP error probability as

$$\Phi(\pi(X|Y)) \geq H(X|Y) \geq \phi^*(\pi(X|Y)). \tag{16}$$

Proof: We may write the equivocation as

$$H(X|Y) = \int_{\mathcal{Y}} H(X|y)\, dP(y)$$

and the MAP error probability as,

$$\pi(X|Y) = \int_{\mathcal{Y}} \pi(X|y)\, dP(y)$$

where for each $y \in \mathcal{Y}$, $H(X|y) = H(X|Y = y)$ and $\pi(X|y) = \pi(X|Y = y)$ are the entropy and the predictability, respectively, of a discrete random variable that can take M values. Thus, the points $\{c(y) = (\pi(X|y), H(X|y)), y \in \mathcal{Y}\}$ lie in the region $\mathcal{A}$ in the $\pi - H$ plane. Clearly, the point $c = (\pi(X|Y), H(X|Y))$ is a convex combination of the points $c(y)$ where the weights of this combination are given by the distribution $P(y)$. Thus, the point c must lie in $\tilde{\mathcal{A}}$, the convex hull of $\mathcal{A}$. □

The region $\tilde{\mathcal{A}}$, for the case $M = 8$, is also depicted in Fig. 1. Observe that both inequalities in (16) are tight, i.e., both inequalities can be obtained with equality and so every point on the boundary of the region $\tilde{\mathcal{A}}$ can be attained. The upper bound in (16) is attained when the conditional distribution $p(x|y)$ is the same for all $y \in \mathcal{Y}$ with a non-zero measure, and is such that $H(X|y) = h(\pi(X|y)) + \pi(X|y)\log(M-1)$. The lower bound is attained with equality when for some $y \in \mathcal{Y}$, $p(x|y)$ has a uniform probability mass of $1/i$ over i values and so $\pi(X|y) = (i-1)/i$ and $H(X|y) = \log i$, while for the rest $y \in \mathcal{Y}$, $p(x|y)$ has a uniform probability mass of $1/(i+1)$ over $i+1$ values and so $\pi(X|y) = i/(i+1)$ and $H(X|y) = \log(i+1)$.

An immediate corollary of the theorem above is that the entropy rate $\mathcal{H}(\mathcal{X})$ and the predictability $\Pi(\mathcal{X})$ of any stationary process $\mathcal{X}$ over an alphabet of size M also lie in the region $\tilde{\mathcal{A}}$ in the $\pi - H$ plane, i.e.,

$$\Phi(\Pi(\mathcal{X})) \geq \mathcal{H}(X) \geq \phi^*(\Pi(\mathcal{X})). \tag{17}$$

The upper bound is tight whenever the process can be "whitened" by optimal prediction (i.e., the prediction error process is i.i.d.) and the probability distribution of the error process is such that the $M-1$ possible error values have the same probability of $\Pi/(M-1)$, while no error is made with probability $1 - \Pi$. The lower bound is attained when at some fraction μ of the time there is an ambiguity between i values, each has the same probability of $1/i$, and in the rest of the time, there is an ambiguity in predicting the next outcome between $i+1$ values. In this last case, the resulting fraction of error will be $\mu[(i-1)/i] + (1-\mu)[i/(i+1)]$, while the entropy rate will be $\mu\log i + (1-\mu)\log(i+1)$.

When we closely observe the points c_i where $\pi = [(i-1)/i]$ and $H = \log i$, $i = 1, \cdots, M-1$, i.e., the points which lie on the lower bound and at which $\phi(\pi) = \phi^*(\pi)$, we observe that at these points $H = \log[1/(1-\pi)]$. Define

$$\phi'(\pi) = \log\frac{1}{1-\pi} \tag{18}$$

and observe that $\phi'(\cdot)$ is a convex function that underbound $\phi(\cdot)$ and $\phi^*(\cdot)$. Thus, a lower bound on the entropy in terms of the predictability is $H \geq \phi'(\pi)$. Of course this bound is tight only at the points c_i. Nevertheless this bound is interesting, recognizing

$$R_\infty(X) \triangleq \lim_{q\to\infty} R_q(X) = \lim_{q\to\infty}\frac{1}{1-q}\log\sum_x p(x)^q = \log\frac{1}{\max_x p(x)} = \log\frac{1}{1-\pi(X)} \tag{19}$$

where $R_q(X)$ is the Rènyi entropy of order q of X. We recall that Shannon's entropy is the Rènyi entropy of order 1. The bound (18) thus follows from the known fact [9], asserting that for all $q > 1$, $R_q(X) \leq R_1(X) = H(X)$.

In this respect we further note that due to the one-to-one relationship between $\pi(X)$ and $R_\infty(X)$, given by (19) or its inverse $\pi(X) = 1 - 2^{-R_\infty(X)}$, Fano's inequality together with our Lemma 2 provide the region of allowable pairs of $R_\infty(X)$ and $H(X)$, i.e., for any value of $R_\infty(X)$

$$\phi(1 - 2^{-R_\infty(X)}) \leq H(X) \leq \Phi(1 - 2^{-R_\infty(X)}). \tag{20}$$

The left inequality tightens the well known relation $R_\infty(X) \leq H(X)$. Now, it should be observed that while $\pi(X)$ and $R_\infty(X)$ have a one-to-one relationship, this is no longer true for $\pi(X|Y)$ and $R_\infty(X|Y)$. Thus, while the convex hull of the region given by (20), which is

$$R_\infty(X|Y) \leq H(X|Y) \leq \Phi(1 - 2^{-R_\infty(X|Y)}) \tag{21}$$

provides all allowable pairs of $R_\infty(X|Y)$ and $H(X|Y)$, this convex hull is different from the one-to-one transformation of $\tilde{\mathcal{A}}$, given by

$$\phi^*(1 - 2^{-R_\infty(X|Y)}) \leq H(X|Y) \leq \Phi(1 - 2^{-R_\infty(X|Y)}). \tag{22}$$

Nevertheless, one may observe from Fig. 2 where both regions implied by (21) and (22) are also depicted, that for large, or even moderate, values of R_∞

$$R_\infty \approx \phi^*(1 - 2^{-R_\infty(X|Y)}) \tag{23}$$

and so in this case the bound $R_\infty \leq H$ at $R_\infty = -\log(1-\pi)$ is indeed a good lower bound on the entropy as a function of the error probability.

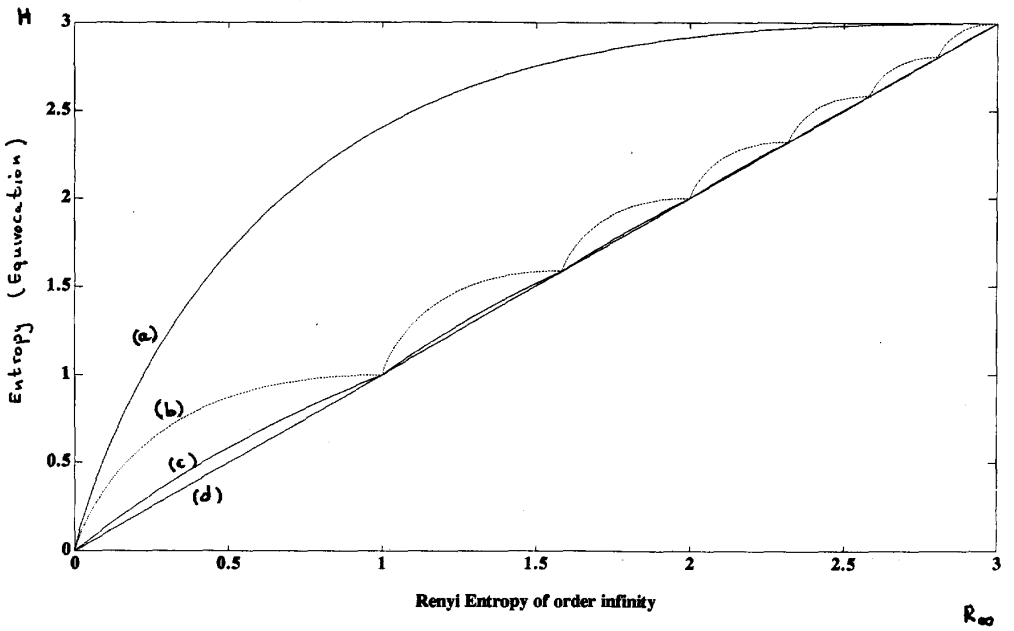

Fig. 2. The functions: (a) $H = \Phi(1 - 2^{-R_\infty})$, (b) $H = \phi(1 - 2^{-R_\infty})$, (c) $H = \phi^*(1 - 2^{-R_\infty})$, and (d) $H = R_\infty$, and the resulting regions in the $H - R_\infty$ plane.

As noted above, in the binary case, the fact that the entropy (or entropy rate) and the predictability do not have a one-to-one relationship, and the relevant bounds

$$h(\pi) \geq H \geq 2\pi \tag{24}$$

or equivalently,

$$\frac{1}{2}H \geq \pi \geq h^{-1}(H) \tag{25}$$

have been mentioned in [5]. It turns out that the lower bound in (24) for the binary case has been previously derived in [1], see also [8]. Furthermore in [7, pp. 520–521], it has also been used for nonbinary discrete random variables. However, our lower bound in (17) is tighter since always $\phi^*(\pi) \geq 2\pi$, and for nonbinary variables the inequality is strict for $\pi > 1/2$.

We finally point out that the techniques presented here can be used to derive upper and lower bound on the average loss in terms of the entropy, for general loss functions. For example, the minimum mean square error in estimating a random variable is measured by its variance. Thus, one can find the maximum entropy and the minimum entropy of a random variable under a variance constraint, as a function of the variance value. By drawing the region between these two functions, and considering its convex hull, one obtains the entire set of achievable pairs of entropy and mean square error values.

III. Prediction of Deterministic Sequences from a Finite Set

We now confine our attention to sequential prediction of *arbitrary* deterministic sequences. To simplify the exposition we consider in this section binary sequences. Recall that a sequential predictor of a binary sequence is a procedure for producing at each instant t, upon observing the data $x_1, \cdots, x_t$, an estimate of the next outcome $\hat{x}_{t+1}$,

$$\hat{x}_{t+1} = f_t(x_t, \cdots, x_1). \tag{26}$$

In general $f_t(\cdot)$ can either be deterministic or stochastic. The performance of a deterministic sequential predictor is measured in terms of the fraction of prediction errors along the sequence, i.e., for a sequence $\boldsymbol{x} = x_1, \cdots, x_n$ of length n,

$$\pi_f(\boldsymbol{x}) = \frac{1}{n}\sum_{t=1}^{n}[1 - \delta(x_i, \hat{x}_i)] \tag{27}$$

where $\delta(a, b)$ is 1 for $a = b$ and 0, otherwise. For stochastic predictors, the performance is given by

$$\pi_f(\boldsymbol{x}) = E\left\{\frac{1}{n}\sum_{t=1}^{n}[1 - \delta(x_i, \hat{x}_i)]\right\} = \frac{1}{n}\sum_{t=1}^{n} Pr\{x_i \neq \hat{x}_i\} \tag{28}$$

where it should be kept in mind that the expectation is with respect to the predictor's randomness while the sequence $\boldsymbol{x}$ is fixed.

Now, as noted in [2], for any sequence there is a predictor that happens to guess correctly its future values, but this predictor may not perform well on other sequences. Thus, we consider the average performance of any predictor over a set of deterministic sequences. Interestingly, the relation between this average number of errors and the logarithm of the number of sequences in the set is the same as the relation between the predictability and the entropy derived in the previous section. An additional insight is gained by explicitly describing the structure of the sets of sequences that attain the resulting bounds.

Suppose we have a set $\mathcal{X}$ of N binary sequences $\{\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\}$ each of length n. The performance of any predictor over this set is

$$\overline{\pi}_f(\mathcal{X}) = \frac{1}{N}\sum_{j=1}^{N}\pi_f(\boldsymbol{x}^{(j)}), \tag{29}$$

and so the performance of the best predictor for this set is

$$\overline{\pi}(\mathcal{X}) = \min_f \overline{\pi}_f(\mathcal{X}) \tag{30}$$

where the minimization is over all predictors, deterministic or stochastic. We claim the following theorem.

Theorem 2: For any set $\mathcal{X}$ of N binary sequences of length n

$$\frac{1}{2}\cdot\frac{\log N}{n}\geq\overline{\pi}(\mathcal{X})\geq h^{-1}\left(\frac{\log N}{n}\right). \tag{31}$$

This theorem is related to the bounds in the previous section. To see this, construct a binary random process which emits blocks of length n, where each such block can be any of the N sequences in the set with equal probability. The entropy rate, which is the entropy-per-symbol of the block, is $\log N/n$. It can also be shown that the predictability of this process is given by (30). With that, the theorem can follow by applying the bounds in (25). We, however, prove below the theorem directly using combinatorial arguments since this proof provides an additional insight on the structure of the sets that attain the bounds.

Proof: We begin with the lower bound. As was observed in [2], Proposition I, when all 2^n binary sequences of length n are considered, any deterministic predictor makes exactly k errors over $\binom{n}{k}$ sequences, i.e., there is one sequence on which this predictor makes no error, n sequences with a single error, $\binom{n}{2}$ with two errors, etc. Thus, the best one can hope for, is to exhaust all possibilities of making i errors or less before making $i+1$ errors. Let m be the largest integer such that

$$\sum_{i=0}^{m}\binom{n}{i}\leq N.$$

The minimal total number of errors made by any deterministic predictor over N sequences of length n is lower bounded by

$$k_n(N)=\sum_{i=0}^{m}i\binom{n}{i}+(m+1)\cdot\left(N-\sum_{i=0}^{m}\binom{n}{i}\right) \tag{32}$$

and so

$$\overline{\pi}(\mathcal{X})\geq\frac{1}{nN}k_n(N). \tag{33}$$

Using linear interpolation and considering $k_n(\cdot)$ as a function of a continuous argument we observe that it is a piecewise linear, concave, monotonically increasing function, having a slope 0 between $x_0\leq x\leq x_1$, a slope 1 at $x_1\leq x\leq x_2$, a slope 2 at $x_2\leq x\leq x_3$, etc. where $x_0=0$, $x_1=1$, $x_2=x_1+n$ and in general $x_{i+1}=x_i+\binom{n}{i}$. Since $k_n(\cdot)$ is convex and since the performance of any stochastic predictor is a convex combination of deterministic predictors, the lower bound (33) holds for stochastic predictors as well.

It is easy to verify that $k_n(x)\geq nxh^{-1}(\log x/n)$, $1\leq x\leq 2^n$. Thus,

$$\frac{1}{nN}k_n(N)\geq h^{-1}\left(\frac{\log N}{n}\right), \tag{34}$$

and the lower bound is proved.

We now prove the upper bound. Let $N(\nu)$ be the number of sequences in $\mathcal{X}$ that begin with the string ν. In this notation $N=N(\Lambda)$ where Λ is the empty string. The predictor that minimizes the total number of errors predicts "0," upon observing the string ν, if $N(\nu 0)>N(\nu 1)$ and "1," otherwise. Thus, the minimum total number of errors over all sequences is

$$n\cdot N\cdot\overline{\pi}(\mathcal{X})=\sum_{i=0}^{n-1}\sum_{\nu\in\{0,1\}^i}\min\{N(\nu 0),N(\nu 1)\}. \tag{35}$$

Since for $0\leq\alpha\leq 1$, $\min\{\alpha,1-\alpha\}\leq\frac{1}{2}h(\alpha)$,

$$\min\left\{\frac{N(\nu 0)}{N(\nu)},\frac{N(\nu 1)}{N(\nu)}\right\}\leq\frac{1}{2}\left(-\frac{N(\nu 0)}{N(\nu)}\log\frac{N(\nu 0)}{N(\nu)}-\frac{N(\nu 1)}{N(\nu)}\log\frac{N(\nu 1)}{N(\nu)}\right). \tag{36}$$

Multiplying both sides of (36) by $N(\nu)$, substituting in (35) and rearranging the summation, we get

$$n\cdot N\cdot\overline{\pi}(\mathcal{X})\leq-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}\log\frac{N(p_j(\boldsymbol{x}^{(i)}))}{N(p_{j-1}(\boldsymbol{x}^{(i)}))}=-\frac{1}{2}\sum_{i=1}^{N}\log\prod_{j=1}^{n}\frac{N(p_j(\boldsymbol{x}^{(i)}))}{N(p_{j-1}(\boldsymbol{x}^{(i)}))} \tag{37}$$

where $p_j(\boldsymbol{x}^{(i)})$ denotes the prefix of length j of the sequence $\boldsymbol{x}^{(i)}$. Now observe that due to telescopic multiplication

$$\prod_{j=1}^{n}\frac{N(p_j(\boldsymbol{x}^{(i)}))}{N(p_{j-1}(\boldsymbol{x}^{(i)}))}=\frac{N(p_n(\boldsymbol{x}^{(i)}))}{N(p_0(\boldsymbol{x}^{(i)}))}=\frac{1}{N} \tag{38}$$

for each sequence $\boldsymbol{x}^{(i)}$. Substituting in (37), proves the upper bound. □

An example where the lower bound in (33) [which is slightly better than (31)] is attained with equality is the set which contains all sequences of length n whose number of ones is less than or equal to k for some $k\leq n$. Clearly, a predictor that constantly predicts 0 will attain (33) on this set. The upper bound in (31) can also be attained with equality. For example, consider the set of 2^k sequences, $k\leq n$, each beginning with a different prefix of length k and continuing arbitrarily (e.g., for each sequence the last $n-k$ bits are zero). Since in the first k bits all 2^k possibilities appear, on the average any predictor will make $k/2$ errors in these initial bits. Now the first k bits determine the sequence and so the optimal predictor will not make any error over the remaining $n-k$ bits for any sequence. The total average fraction of errors is thus $k/(2n)=\log N/(2n)$, as the upper bound.

Note that when the set of sequences is a type Q, i.e., the set of all sequences with a given count of zeros and ones, then $N=2^{nH_e(Q)+O(\log n)}$ where $H_e(Q)$ is the empirical entropy, which is the same for all sequences in the type. In this case for large n the relation (31) becomes

$$\frac{1}{2}H_e(Q)\geq\overline{\pi}(Q)\geq h^{-1}(H_e(Q))$$

which is analogous to the probabilistic case with empirical probabilities replacing true probabilities. Also note that the bounds in (31) affirm the intuitive notion that the average fraction of errors over *all* possible sequences of some length cannot be better than prediction by coin-tossing, i.e., 50% errors, while if the number of sequences in the set grows less than exponentially fast with the sequence length the average fraction of prediction errors can be made arbitrarily small.

IV. A Channel Coding Theorem for the Equivocation

The coding theorems of information theory are usually stated in terms of error probability. However, it might be useful to state the theorems in terms of the equivocation, for the following reasons. First, the equivocation is a useful uncertainty measure with applications, e.g., in cryptology, see [11] and references therein. Second, the equivocation measures naturally the minimal residual uncertainty about the input, achievable in, e.g., observing the data via a noisy channel. Also, this statement is simpler; in transmitting information at rate R via a noisy channel, the equivocation of the input can be made $R-C$ if $R\geq C$, and 0 if $R\leq C$ where C is the channel capacity. Throughout this section, R and C are measured in bits per channel use.

The channel coding theorem for the equivocation can be easily proved, for $R<C$, by combining the regular channel theorem, in terms of error probability, and the fact, discussed below, that zero equivocation is achieved if and only if zero error probability is

achieved. Now, it turns out that the channel coding theorem, in terms of equivocation, can be proved directly, at least for DMC's. Although this proof is not simpler than the standard proof of the channel coding theorem, it provides some additional insight. As expected, when $R < C$ the equivocation approaches zero exponentially fast with the block length. The additional conclusion from this proof is that when the rate is exactly the capacity, the equivocation, normalized to bits per channel use (the equivocation rate), approaches zero as $O(n^{-1/2})$ where n is the block length. Furthermore, for $R > C$, the normalized equivocation approaches its minimal value of $R-C$, again, at a rate $O(n^{-1/2})$.

The following proof of the coding theorem in terms of equivocation, for DMC's, makes an essential use of random coding arguments and it resembles Gallager's well-known proof [6]. The usual scenario is assumed. There is a codebook of size $M = 2^{nR}$ codewords, where each codeword is a vector $\boldsymbol{x}$ of length n whose components are channel input symbols. To transmit the maximal information through the channel, the index of the codeword in the codebook, to be transmitted, is selected with a uniform distribution and so the codebook may be considered as a random vector $\boldsymbol{X}$, whose entropy is $H(\boldsymbol{X}) = \log M = nR$. In the random coding scenario, the codebook is constructed by randomly choosing codewords according to some distribution. Our interest is the equivocation $H(\boldsymbol{X}|\boldsymbol{Y})$ of the codebook. To utilize the random coding arguments, we consider the average of this equivocation, denoted $\overline{H}(\boldsymbol{X}|\boldsymbol{Y})$, where the average is with respect to an ensemble of randomly selected codebooks. We claim the following theorem.

Theorem 3: Consider a DMC with a transition distribution $p(y|x)$. Let a codebook be reconstructed by choosing randomly $M = 2^{nR}$ codewords, of length n, using an i.i.d. distribution $q(x)$. Then, for any $0 \le \rho \le 1$, the equivocation, averaged over all codebook selections, satisfies

$$\overline{H}(\boldsymbol{X}|\boldsymbol{Y}) \le \left(1+\frac{1}{\rho}\right)(\log e)2^{-n[E_0(\rho,\, q)-\rho R]} \tag{39}$$

where $E_0(\rho, q)$ is the random coding exponent,

$$E_0(\rho, q) = -\log \sum_y \left[\sum_x q(x)p(y|x)^{1/(1+\rho)}\right]^{1+\rho} \tag{40}$$

Proof: In the proof we bound from below the average mutual information between the codeword input and the channel output, and the bound then implies the desired upper bound on the equivocation. For a given codebook $C = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M\}$, define

$$J_C(\boldsymbol{x}_m; \boldsymbol{y}) = \log \frac{p(\boldsymbol{y}|\boldsymbol{x}_m)}{\frac{1}{M}\sum_{m'=1}^{M} p(\boldsymbol{y}|\boldsymbol{x}_{m'})}. \tag{41}$$

The average mutual information, $\overline{I}(\boldsymbol{X}; \boldsymbol{Y})$, is the average of (41) using the distribution $p(\boldsymbol{x}_m, \boldsymbol{y}) = (1/M) \cdot p(\boldsymbol{y}|\boldsymbol{x}_m)$ and then averaging over all selections of codewords according to the i.i.d. distribution $q(x)$. It will be useful to interchange the order of averaging and use symmetry, as follows. Define

$$J(\boldsymbol{x}_m; \boldsymbol{y}) = E_C\{J_C(\boldsymbol{x}_m; \boldsymbol{y})\} \tag{42}$$

where the expectation is with respect to all codewords $\boldsymbol{x}_{m'} \neq \boldsymbol{x}_m$, each chosen with the i.i.d. distribution $q(\cdot)$. The desired average mutual information is

$$\overline{I}(\boldsymbol{X}; \boldsymbol{Y}) = \frac{1}{M}\sum_{m=1}^{M} E\{J(\boldsymbol{x}_m; \boldsymbol{y})\} = E\{J(\boldsymbol{x}_m; \boldsymbol{y})\} \tag{43}$$

where here the expectation is with respect to the measure $q(\boldsymbol{x}_m)p(\boldsymbol{y}|\boldsymbol{x}_m)$. Note that due to symmetry, this expectation is independent of m, and so the right equality in (43) follows.

Calculating $J(\boldsymbol{x}_m; \boldsymbol{y})$ explicitly we get,

$$J(\boldsymbol{x}_m; \boldsymbol{y}) = \log M - E_C\left\{\log \sum_{m'=1}^{M} \frac{p(\boldsymbol{y}|\boldsymbol{x}_{m'})}{p(\boldsymbol{y}|\boldsymbol{x}_m)}\right\}$$
$$= nR - E_C\left\{\log\left(1 + \sum_{m' \neq m} \frac{p(\boldsymbol{y}|\boldsymbol{x}_{m'})}{p(\boldsymbol{y}|\boldsymbol{x}_m)}\right)\right\}. \tag{44}$$

Now for any $0 \le \rho \le 1$ and nonnegative numbers $\{a_i\}$ we have both

$$\sum_i a_i \le \left(\sum_i (a_i)^{1/1+\rho}\right)^{1+\rho} \tag{45}$$

$$\sum_i a_i \le \left(\sum_i (a_i)^{\rho}\right)^{1/\rho}. \tag{46}$$

Thus, we can lower bound $J(x_m; \boldsymbol{y})$

$$J(\boldsymbol{x}_m; \boldsymbol{y})$$
$$\ge nR - (1+\rho)E_C\left\{\log\left(1 + \sum_{m' \neq m}\left[\frac{p(\boldsymbol{y}|\boldsymbol{x}_{m'})}{p(\boldsymbol{y}|\boldsymbol{x}_m)}\right]^{1/1+\rho}\right)\right\}$$
$$\ge nR - \left(1+\frac{1}{\rho}\right)E_C\left\{\log\left(1 + \left[\sum_{m' \neq m}\left[\frac{p(\boldsymbol{y}|\boldsymbol{x}_{m'})}{p(\boldsymbol{y}|\boldsymbol{x}_m)}\right]^{1/1+\rho}\right]^{\rho}\right)\right\}$$
$$\ge nR - \left(1+\frac{1}{\rho}\right)(\log e)E_C\left\{\left[\sum_{m' \neq m}\left[\frac{p(\boldsymbol{y}|\boldsymbol{x}_{m'})}{p(\boldsymbol{y}|\boldsymbol{x}_m)}\right]^{1/1+\rho}\right]^{\rho}\right\} \tag{47}$$

where for the first inequality we used (45), for the second inequality we used (46) and the third inequality follows from the relation $x \log e \ge \log(1+x)$. Now

$$J(\boldsymbol{x}_m; \boldsymbol{y})$$
$$\ge nR - \left(1+\frac{1}{\rho}\right)(\log e)\left[E_C\left\{\sum_{m' \neq m}\left[\frac{p(\boldsymbol{y}|\boldsymbol{x}_{m'})}{p(\boldsymbol{y}|\boldsymbol{x}_m)}\right]^{1/1+\rho}\right\}\right]^{\rho}$$
$$= nR - \left(1+\frac{1}{\rho}\right)(\log e)\left[(M-1)\sum_{\boldsymbol{x}'} q(\boldsymbol{x}')\left[\frac{p(\boldsymbol{y}|\boldsymbol{x}')}{p(\boldsymbol{y}|\boldsymbol{x}_m)}\right]^{1/1+\rho}\right]^{\rho}$$
$$\ge nR - \left(1+\frac{1}{\rho}\right)(\log e)M^{\rho}\left[\sum_{\boldsymbol{x}'} q(\boldsymbol{x}')\left[\frac{p(\boldsymbol{y}|\boldsymbol{x}')}{p(\boldsymbol{y}|\boldsymbol{x}_m)}\right]^{1/1+\rho}\right]^{\rho} \tag{48}$$

where the first line follows from Jensen's inequality the second line follows by writing the expectation over all $x_{m'} \neq x_m$ explicitly and observing that after taking the expectation all $M-1$ terms in the summation over $m' \neq m$ become equal, and the inequality in the third line follows since $M-1$ is replaced by M.

We now take the second expectation to get a bound for $\overline{I}(\boldsymbol{X}; \boldsymbol{Y})$,

$$\overline{I}(\boldsymbol{X}; \boldsymbol{Y})$$
$$\ge nR - \left(1+\frac{1}{\rho}\right)(\log e)2^{\rho n R}\sum_{\boldsymbol{y}}\sum_{\boldsymbol{x}} q(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})\left[\sum_{\boldsymbol{x}'} q(\boldsymbol{x}')\right.$$
$$\left.\cdot\left[\frac{p(\boldsymbol{y}|\boldsymbol{x}')}{p(\boldsymbol{y}|\boldsymbol{x})}\right]^{1/1+\rho}\right]^{\rho}$$
$$= nR - \left(1+\frac{1}{\rho}\right)(\log e)2^{\rho n R}\sum_{\boldsymbol{y}}\left[\sum_{\boldsymbol{x}} q(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})^{1/1+\rho}\right]$$
$$\cdot\left[\sum_{\boldsymbol{x}'} q(\boldsymbol{x}')p(\boldsymbol{y}|\boldsymbol{x}')^{1/1+\rho}\right]^{\rho}$$
$$= nR - \left(1+\frac{1}{\rho}\right)(\log e)2^{\rho n R}\sum_{\boldsymbol{y}}\left[\sum_{\boldsymbol{x}} q(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)}\right]^{1+\rho} \tag{49}$$

Now, since $q(\cdot)$ and $p(\cdot|\cdot)$ are i.i.d. distributions, the double summation over the vectors can be replaced by a summation over a single letter raised to the power of n. Similar manipulations have been performed in [6]. Using the definition of the random coding exponent (40), and recalling that $\overline{H}(\boldsymbol{X}|\boldsymbol{Y}) = \overline{H}(\boldsymbol{X}) - \overline{I}(\boldsymbol{X};\boldsymbol{Y})$ where $H(\boldsymbol{X}) = \overline{H}(\boldsymbol{X}) = nR$, the desired result (39) follows. □

The inequality (39) holds for any choice of $q(\cdot)$ and ρ. Clearly to get the tightest bound, at a given rate R, one has to minimize the RHS of (39) with respect to $q(\cdot)$ and ρ. Now, the exponent (40) is well investigated, and $\max_{0\le\rho\le 1, q(\cdot)}[E_0(\rho, q) - \rho R]$ is strictly positive as long as $R < C$, providing the random coding exponential decay of both the error probability *and* the equivocation. Note that as $R \to C$ the optimal $q(\cdot)$ is the distribution that achieves the channel capacity.

The inequality (39) also holds for any value of R, including $R \ge C$. Now, unlike the random coding bound on the error probability which becomes useless as it exceeds 1, this bound on the equivocation is always meaningful. When $R = C$ we find that the optimal ρ approaches zero. In this case by letting $\rho = \rho_n$ to vanish with n, and using a Taylor expansion of $E_0(\rho, q)$ about $\rho = 0$ we obtain

$$E_0(\rho_n, q) - \rho_n C = -\gamma\rho_n^2 + o(\rho_n^2), \tag{50}$$

leading to the bound

$$\overline{H}(\boldsymbol{X}|\boldsymbol{Y}) \le \left(1 + \frac{1}{\rho_n}\right)(\log e)2^{n\gamma\rho_n^2}. \tag{51}$$

Now, there are two conflicting goals. On one hand, ρ_n should vanish as quickly as possible to make the exponent the smallest. On the other hand, it should vanish slow, to make the term $1/\rho_n$ the smallest. It is clear that the optimal choice is $\rho_n = \beta/\sqrt{n}$, for some constant β, which cancels the exponential growth, with the smallest increase of the $1/\rho_n$ term. With this choice, at $R = C$ the average equivocation of the codebook satisfies

$$\overline{H}(\boldsymbol{X}|\boldsymbol{Y}) \le \alpha\sqrt{n} \tag{52}$$

where $\alpha > 0$ is some constant. Thus, the average equivocation rate decays at the rate of $n^{-1/2}$. The bound (52) for $\overline{H}(\boldsymbol{X}|\boldsymbol{Y})$ implies, of course, that there exists at least one code C^* whose equivocation rate vanishes as $O(n^{-1/2})$.

Using (52) it is easy to see that we can construct a random vector $\boldsymbol{X}$ whose entropy is R bits per input symbol where $R > C$, and whose equivocation satisfies

$$\frac{1}{n}H(\boldsymbol{X}|\boldsymbol{Y}) \le R - C + O(n^{-1/2}). \tag{53}$$

The idea is to take the codebook C^* described above, which contains 2^{nC} words, and just replicate each word $2^{n(R-C)}$ times to get in total a codebook of 2^{nR} words. The index of the word is chosen with a uniform distribution. The random variable representing the index is the encoded input $\boldsymbol{X}$, and we denote by $\boldsymbol{U}$ the random variable representing the codewords themselves, where $\boldsymbol{U}$ can take only 2^{nC} different values. Now, $\boldsymbol{X} \to \boldsymbol{U} \to \boldsymbol{Y}$ is a Markov chain and $\boldsymbol{U}$ is deterministically determined by $\boldsymbol{X}$. Thus,

$$H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{X}, \boldsymbol{U}|\boldsymbol{Y}) = H(\boldsymbol{X}|\boldsymbol{U}) + H(\boldsymbol{U}|\boldsymbol{Y}). \tag{54}$$

The result (53) follows since by construction $H(\boldsymbol{X}|\boldsymbol{U}) = n(R-C)$, and since from (52), $n^{-1}H(\boldsymbol{U}|\boldsymbol{Y}) \le O(n^{-1/2})$. Since always $H(\boldsymbol{X}|\boldsymbol{Y}) \ge H(\boldsymbol{X}) - \max_q I(\boldsymbol{X};\boldsymbol{Y}) = nR - nC$, we conclude that the equivocation, per input symbol, can be made exactly $R-C$, at a rate $O(n^{-1/2})$.

The relation between entropy and error probability implies that any bound concerning the error probability can be used for bounding the equivocation as well. For example, at low rates, a better bound for the error probability is provided by the expurgated error exponent. Using the standard techniques and similarly to the proof of Theorem 3, one can easily derive directly, as well, the expurgated bound in terms of equivocation.

We have shown above that at $R = C$ the equivocation rate of the codebook vanishes as $O(n^{-1/2})$. One may wonder whether the error probability per input symbol (the error probability rate) has a similar behavior. Unfortunately, this is not implied by the relations between the entropy and error probability. The reason is that the error probability rate is given by $n^{-1}\sum_{i=1}^n \Pi(X_i|\boldsymbol{Y})$, while applying the relations discussed in Section II to (52) yields only a trivial bound for $\Pi(\boldsymbol{X}|\boldsymbol{Y})$. Nevertheless, in a scenario where the correct symbols $X_1 \cdots X_t$ are revealed to the decoder before it decodes the next symbol X_{t+1}, a meaningful upper bound on the error probability rate can be derived as follows. The equivocation per input symbol can be written as

$$\frac{1}{n}H(\boldsymbol{X}|\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^n H(X_i|X_{i-1}\cdots X_1, \boldsymbol{Y}). \tag{55}$$

Thus,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n &\Pi(X_i|X_{i-1}\cdots X_1, \boldsymbol{Y}) \\ &\le \frac{1}{n}\sum_{i=1}^n \phi^{*-1}(H(X_i|X_{i-1}\cdots X_1, \boldsymbol{Y})) \\ &\le \phi^{*-1}\left(\frac{1}{n}H(\boldsymbol{X}|\boldsymbol{Y})\right)\end{aligned} \tag{56}$$

where the right inequality follows from the convexity of $\phi^*(\cdot)$. At $R = C$ this means that the error probability rate in this scenario approaches zero as $O(n^{-1/2})$.

ACKNOWLEDGMENT

We thank N. Shulman for a useful suggestion concerning the proof of Theorem 1. We also acknowledge S. Verdu for his suggestion to present the region of allowable pairs of $R_\infty(X)$ and $H(X)$.

REFERENCES

[1] J. Chu and J. Chueh, "Inequalities between information measures and error probability," *J. Franklin Inst.*, vol. 282, pp. 121–125, Aug. 1966.

[2] T. M. Cover, "Behavior of sequential predictors of binary sequences," in *Proc. 4th Prague Conf. Inform. Theory, Statist. Decision Functions, Random Processes, 1965*, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1967, pp. 263–272.

[3] ——, "Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin," Dep. of Statistics, Stanford Univ., Tech. Rep. 12, Oct. 1974.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, New York, 1991.

[5] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1270, July 1992.

[6] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 3–18, Jan. 1965.

[7] ——, *Information Theory and Reliable Communications*. Wiley, New York: 1968.

[8] M. E. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 368–372, July 1970.

[9] G. Jumarie, *Relative Information: Theory and Applications*. New York: Springer-Verlag, 1990.

[10] J. L. Kelly, Jr., "A new interpretation of information rate," *Bell Syst. Tech. J.*, vol. 35, pp. 917–926, 1956.

[11] H. Yamamoto, "Information theory in cryptology," *IEICE Trans.*, vol. E-74, no. 9, pp. 2456–2464, 1991.