# Hypothesis Testing and Information Theory

RICHARD E. BLAHUT, MEMBER, IEEE

***Abstract*—The testing of binary hypotheses is developed from an information-theoretic point of view, and the asymptotic performance of optimum hypothesis testers is developed in exact analogy to the asymptotic performance of optimum channel codes. The discrimination, introduced by Kullback, is developed in a role analogous to that of mutual information in channel coding theory. Based on the discrimination, an error-exponent function $e(r)$ is defined. This function is found to describe the behavior of optimum hypothesis testers asymptotically with block length. Next, mutual information is introduced as a minimum of a set of discriminations. This approach has later coding significance. The channel reliability-rate function $E(R)$ is defined in terms of discrimination, and a number of its mathematical properties developed. Sphere-packing-like bounds are developed in a relatively straightforward and intuitive manner by relating $e(r)$ and $E(R)$. This ties together the aforementioned developments and gives a lower bound in terms of a hypothesis testing model. The result is valid for discrete or continuous probability distributions. The discrimination function is also used to define a source code reliability-rate function. This function allows a simpler proof of the source coding theorem and also bounds the code performance as a function of block length, thereby providing the source coding analog of $E(R)$.**

## I. Introduction

An OPTIMAL channel block code of rate $R$ and block length $n$ has a probability of error, which is asymptotically exponential as a function of $n$, provided that $R$ is less than the channel capacity $C$. This was established by Feinstein [1]. Fano [2] explicitly found an exponentially decreasing function as an upper bound to this probability of error. The exponential decay coefficient $E(R)$ is nonzero only for rates below the channel capacity. By means of this latter observation, Fano provides a satisfying proof of the channel coding theorem of Shannon [3]. Fano further showed that $E(R)$ also provides a lower bound to code performance for high rates, and hence describes the performance of optimum high-rate codes.

The development of these results later was considerably refined and simplified by Gallager [4] and by Shannon *et al.* [5].

The function $E(R)$ usually appears during the proof of the coding theorem as the outcome of a series of manipulations, and so attains little intuitive appeal. One purpose of the present paper is to provide a stronger geometrical interpretation for the role of $E(R)$.

A second purpose of the paper is to establish stronger ties between the subjects of decision theory and information theory. The Neyman–Pearson theorem provides the structure for the optimum testing of binary hypotheses. This has been used by Forney [6] to study the behavior of optimum list decoders of channel block codes. Csiszar and Longo [7] showed that the probability of error of optimum hypothesis testers based on blocks of measurements is exponentially decreasing with block length. The exponential decay coefficient is related to the discrimination, a function first introduced by Kullback [8]. Using a similar development, we present various performance exponents in new forms and discuss the close relationships between them.

Our starting point in this paper is the discrimination, and its application to the behavior of hypothesis testers. This treatment is in part tutorial, and serves to develop in accord with our style and purpose a structure that underlies the remainder of the paper. The approach taken emphasizes the fundamental role played by the discrimination in the performance bounds of information theory.

The study of hypothesis testing is found to be a miniature version of the study of channel block codes and the discrimination an embryonic form of the mutual information. A function $e(r)$ is defined that describes achievable hypothesis testers and their associated performance. The reliability-rate function $E(R)$ is then defined by direct analogy with $e(r)$, and a number of properties are developed.

The sphere-packing bound is developed in terms of a lower bound to the hypothesis testing problem. This provides a simpler and more powerful version of the sphere-packing bound. The result is valid for discrete or for continuous probability distributions.

Finally, we turn to the subject of source coding, developing an exponential upper bound to the performance of optimum compressors of source output data in terms of a reliability-rate function $F(R,D)$. This function is defined in terms of the discrimination. The resulting bound is used to provide a stronger statement of the source compression coding theorem.

## II. The Error-Exponent Function

The concept of discrimination was introduced by Kullback and plays a fundamental role in the behavior of optimum hypothesis testers. The basic structure consists of two hypotheses $H_1$ and $H_2$ and their associated probability distributions $q_1,q_2$ on a discrete measurement space. The log likelihood ratio, $\log (q_{1k}/q_{2k})$, is a random variable, and its mean is called the discrimination (in favor of $H_1$ against $H_2$).

*Definition 1:* Let $\mathscr{P}^K$ be the space of discrete probability distributions on a set of $K$ elements. The *discrimination* is a function $J\colon \mathscr{P}^K \times \mathscr{P}^K \rightarrow \mathscr{R}$ defined by

$$J(q_1; q_2) = \sum_k q_{1k} \log \frac{q_{1k}}{q_{2k}}.$$

We will assume for later convenience that $q_1$ and $q_2$ have strictly positive components.

Manuscript received November 21, 1972; revised December 31, 1973. This paper was presented at the 1973 International Symposium on Information Theory, Ashkelon, Israel.

The author is with IBM Inc., Owego, N.Y., and Cornell University, Ithaca, N.Y.

The discrimination has the usual properties expected of an information measure. It is nonnegative and strictly positive if and only if its arguments are unequal. The discrimination is convex in each of its arguments.

Further, the discrimination is additive for independent measurements as a consequence of its logarithmic nature. Therefore, the discrimination for $n$ independent identically distributed (i.i.d.) measurements is $n$ times the discrimination for an individual measurement. If we have $n$ nonidentical measurements drawn from a set of $J$ measurement types indexed by $j$ ($j = 0,\cdots,J - 1$), and $n_j$ is the number of measurements of type $j$, we define the average discrimination by

$$J(Q_1; Q_2) = \sum_j p_j \sum_k Q_{1k|j} \log \frac{Q_{1k|j}}{Q_{2k|j}}$$

where $Q_1 = \{Q_{1k|j}\}$, $Q_2 = \{Q_{2k|j}\}$ give the probabilities of measurement outcome $k$ given measurement type $j$ under hypotheses $H_1$ and $H_2$, respectively, and $p_j = n_j/n$. The total discrimination for the $n$ measurements is then $n$ times this average discrimination.

This form is related to the mutual information, which can be defined as

$$I(p; Q) = \min_q \sum_j \sum_k p_j Q_{k|j} \log \frac{Q_{k|j}}{q_k}.$$

The minimum is achieved when $q_k = \sum_j p_j Q_{k|j}$. This version of the mutual information will be used later to pass from hypothesis testing to channel encoding.

An hypothesis-testing procedure is a partition of the measurement space into two disjoint sets $\mathcal{U}$ and $\mathcal{U}^c$. If the measurement $k$ is an element of $\mathcal{U}$, we decide that $H_1$ is true; if $k$ is an element of $\mathcal{U}^c$, we decide $H_2$ is true.

The probability of accepting hypothesis $H_1$ when $H_2$ actually is true is called the type 1 error probability $\alpha$. The probability of accepting hypothesis $H_2$ when $H_1$ actually is true is called the type 2 error probability $\beta$. Obviously

$$\alpha = \sum_{k \in \mathcal{U}} q_{2k} \qquad \beta = \sum_{k \in \mathcal{U}^c} q_{1k}.$$

The problem is to specify $(\mathcal{U},\mathcal{U}^c)$ so that $\alpha$ and $\beta$ are as small as possible. This is not yet a well-defined problem, since $\alpha$ generally can be made smaller by reducing $\mathcal{U}$, although $\beta$ thereby increases. The Neyman–Pearson point of view assumes that a maximum value of $\beta$ is specified, say $\beta_m$, and $(\mathcal{U},\mathcal{U}^c)$ must be determined so as to minimize $\alpha$ subject to this constraint on $\beta$ ($\beta \leq \beta_m$).

A method for finding these decision regions is given by the following well-known theorem.

*Theorem 1:* (Neyman–Pearson) For any real number $T$, let

$$\mathcal{U}(T) = \{k \mid q_{2k} \leq q_{1k}e^{-T}\}$$

$$\mathcal{U}(T)^c = \{k \mid q_{2k} > q_{1k}e^{-T}\}$$

and let $\alpha^*$, $\beta^*$ be the type 1 and type 2 error probabilities corresponding to this choice of decision regions. Suppose $\alpha,\beta$ are the type 1 and type 2 error probabilities corresponding to some other choice of decision regions and suppose $\beta < \beta^*$. Then $\alpha > \alpha^*$.

In the next section we study the behavior of type 1 and type 2 error probabilities for these optimum decision regions. This will be facilitated by introducing the error-exponent function.

*Definition 2:* Let the distributions $q_1$ and $q_2$ be given. Then the *error-exponent function* $e(r)$ is given by

$$e(r) = \min_{\hat{q} \in \mathcal{P}_r} J(\hat{q}; q_2)$$

where

$$\mathcal{P}_r = \{\hat{q} \in \mathcal{P}^K \mid J(\hat{q}; q_1) \leq r\}.$$

If $J(q_1; q_2)$ is finite, then $e(r)$ is finite since $q_1 \in \mathcal{P}_r$. Since we assume $q_1$ and $q_2$ have strictly positive components, $J(q_1; q_2)$ is always finite for finite sample spaces.

The definition is given an intuitive interpretation as follows. We introduce a dummy distribution $\hat{q}$ and then select $\hat{q}$ so that $J(\hat{q}; q_2)$ is smallest given that $J(\hat{q}; q_1) \leq r$. That is, in some sense, select $\hat{q}$ "between" $q_1$ and $q_2$ and at a distance $r$ from $q_1$.

The ensuing theory also holds virtually unchanged if instead we use the average discrimination. In this case we minimize over a set of transition matrices $\hat{Q} = \{\hat{Q}_{k|j}\}$ as follows:

$$e(r) = \min_{\hat{Q} \in \mathcal{Q}_r} J(\hat{Q}; Q_2)$$

where

$$\mathcal{Q}_r = \{\hat{Q} \mid J(\hat{Q}; Q_1) \leq r\}.$$

Since $e(r)$ is defined as the minimum of a convex function subject to a convex constraint, it possesses a number of important properties.

*Theorem 2:* $e(r)$ is a nonincreasing function defined for $r \geq 0$.

*Proof:* If $r > r'$, then $\mathcal{P}_r \supset \mathcal{P}_{r'}$, hence $e(r) \leq e(r')$. If $r < 0$, then $\mathcal{P}_r$ is empty and $e(r)$ is undefined.

*Theorem 3:* $e(r)$ is a convex function. That is, given $r',r''$ and $\lambda \in [0,1]$, then

$$e(\lambda r' + \bar{\lambda} r'') \leq \lambda e(r') + \bar{\lambda} e(r'')$$

where $\bar{\lambda} = (1 - \lambda)$.

*Proof:* Let $q',q''$ achieve $e(r'),e(r'')$, respectively, and let $q^* = \lambda q' + \bar{\lambda} q''$. Then $J(q^*; q_2) \leq \lambda J(q'; q_2) + \bar{\lambda} J(q''; q_2) \leq \lambda r' + \bar{\lambda} r''$. Therefore, $q^* \in \mathcal{P}_{\lambda r' + \bar{\lambda} r''}$ and $e(r) \leq J(q^*; q_1) \leq \lambda J(q'; q_1) + \bar{\lambda} J(q''; q_1) = \lambda e(r') + \bar{\lambda} e(r'')$.

Since $e(r)$ is convex, it is continuous on $(0,\infty)$ and is strictly decreasing prior to any interval on which it is constant.

*Theorem 4:* $e(r)$ can be expressed in terms of a parameter $s$ as

$$e(r_s) = -sr_s - \log \left(\sum_k q_{2k}^{1/1+s} q_{1k}^{s/1+s}\right)^{1+s}$$

where

$$r_s = \sum_k q_k{}^* \log \frac{q_k{}^*}{q_{1k}}$$

and

$$q_k{}^* = \frac{q_{2k}^{1/1+s} q_{1k}^{s/1+s}}{\sum_k q_{2k}^{1/1+s} q_{1k}^{s/1+s}}.$$

*Proof:* Temporarily ignore the constraints $q_k \geq 0$, and introduce the Lagrange multipliers $s$ and $\lambda$ so that

$$e(r) = -sr + \min_{\hat{q}} \left[\sum_k \hat{q}_k \log \frac{\hat{q}_k}{q_{2k}} + s \sum_k \hat{q}_k \log \frac{\hat{q}_k}{q_{1k}} + \lambda \left(\sum_k \hat{q}_k - 1\right)\right].$$

Equating the derivative to zero gives

$$q_k{}^* = \frac{q_{2k}^{1/1+s} q_{1k}^{s/1+s}}{\sum_k q_{2k}^{1/1+s} q_{1k}^{s/1+s}}$$

as the optimizing distribution, where $\lambda$ has been evaluated so that the $q_k{}^*$ sum to one. Notice that $q_k{}^*$ has nonnegative components and so is a probability distribution. Now $s$ should be selected so that $J(q^*; q_2) = r$. Since we cannot solve for $s$ as a function of $r$, we leave the result in parametric form, giving $r$ and $e(r)$ as functions of $s$.

*Theorem 5:*

a) As $s \to 0$, $e(r) \to 0$ and $r \to J(q_2; q_1)$.
b) As $s \to \infty$, $e(r) \to J(q_1; q_2)$ and $r \to 0$.

*Proof:* By writing

$$\frac{q_k{}^*}{q_{2k}} = \frac{(q_{1k}/q_{2k})^{s/1+s}}{\sum_k q_{2k}(q_{1k}/q_{2k})^{s/1+s}}$$

and

$$\frac{q_k{}^*}{q_{1k}} = \frac{(q_{2k}/q_{1k})^{1/1+s}}{\sum_k q_{1k}(q_{2k}/q_{1k})^{1/1+s}}$$

we see that $q_k{}^*$ goes to $q_{2k}$ and $q_{1k}$, respectively, as $s$ goes to zero and infinity. Evaluating the discriminations for these values of $q^*$ proves the theorem.

The Lagrange multiplier $s$ can be given an interpretation as a derivative.

*Theorem 6:* The parameter $s$ satisfies

$$\frac{de(r)}{dr} = -s.$$

*Proof:* Explicitly write $s(r)$ for the value of $s$ achieving $e(r)$. Then

$$e(r) = -rs(r) - (1 + s) \log \sum_k q_{2k}^{1/1+s} q_{1k}^{s/1+s}.$$

Therefore

$$\frac{de(r)}{dr} = -s(r) - r\frac{ds}{dr} - \frac{ds}{dr}\frac{d}{ds}\left[(1 + s) \log \sum_k q_{2k}^{1/1+s} q_{1k}^{s/1+s}\right].$$

Evaluating the last derivative gives

$$\frac{d}{ds}\left[(1 + s) \log \sum_k q_{2k}^{1/1+s} q_{1k}^{s/1+s}\right] = -\sum_k q_k{}^* \log \frac{q_k{}^*}{q_{1k}} = -r.$$

Therefore

$$\frac{de(r)}{dr} = -s(r).$$

We use Theorem 6 to find another representation of $e(r)$. Whereas Theorem 4 expresses $e(r)$ parametrically with $r$ depending on $s$, the following theorem expresses $e(r)$ with $r$ fixed and $s$ depending on $r$.

*Theorem 7:* A representation of $e(r)$ is

$$e(r) = \max_{s \geq 0} \left[-sr - \log \left(\sum_k q_{1k}^{s/1+s} q_{2k}^{1/1+s}\right)^{1+s}\right].$$

*Proof:* For each value of $s$, the bracketed term provides a linear function of $r$. This linear function is tangent to $e(r)$ at some point and is otherwise below $e(r)$ since $e(r)$ is convex. Take

$$s = -\frac{de(r)}{dr}.$$

This $s$ clearly achieves the maximum of the theorem and the right side equals $e(r)$ for this value of $s$.

The error-exponent function has been defined in terms of a single measurement. Its later usefulness in studying the testing of hypotheses by means of blocks of measurements is a consequence of the fact that the error exponent for a block of measurements has a simple relationship to the error exponent for single measurements. These measurements need not be identical provided the average discrimination is used in constructing the error-exponent function.

*Theorem 8:* Suppose a block of independent measurements is made. Let $Q_{1k|j}$ and $Q_{2k|j}$ be the probabilities of the $k$th outcome for a type $j$ experiment under $H_1$ and $H_2$, respectively, and let $e_n(r)$ be the error-exponent function defined on the block experiment. Then

$$e_n(nr) = ne(r).$$

*Proof:* Suppose that $y$ is the sequence of $n$ measurements and $q_1(y), q_2(y)$ are the probability of measurement $y$ under hypothesis $H_1$ and $H_2$. Then since the measurements are independent, these are product distributions

$$q_1(y) = \prod_{l=1}^{n} Q_{1k_l|j_l}$$

and

$$q_2(y) = \prod_{l=1}^{n} Q_{2k_l|j_l}.$$

The probability distribution $\hat{q}(y)$ achieving $e(r)$ is given by the tilted distribution of Theorem 4. Therefore

$$\hat{q}(y) = \frac{\prod_{l=1}^{n} Q_{1k_l|j_l}^{1/1+s} Q_{2k_l|j_l}^{s/1+s}}{\prod_{l=1}^{n} \sum_{k_l} Q_{1k_l|j_l}^{1/1+s} Q_{2k_l|j_l}^{s/1+s}}$$

which is a product distribution. The theorem then follows.

## III. Hypothesis-Testing Bounds

Although the Neyman–Pearson theorem specifies optimum decision regions, it does not directly specify the performance in terms of type 1 and type 2 error probabilities. The type 1 and type 2 error probabilities are given as sums possibly involving a large number of terms. It is generally not possible to reduce these sums into more useful expressions. However, approximations exist that are simple and often useful in interpreting the performance.

These approximations can be motivated by reference to the Neyman–Pearson theorem, which states that it is necessary to consider only threshold decision regions of the form

$$\mathcal{U} = \left\{ k \mid \log \frac{q_{1k}}{q_{2k}} \geq T \right\}$$

$$\mathcal{U}^c = \left\{ k \mid \log \frac{q_{1k}}{q_{2k}} < T \right\}.$$

The major purpose of this section is to develop an explicit relationship between $\alpha$, $\beta$, and $T$ in the form of asymptotically tight upper and lower bounds. The upper bound is given by the following theorem.

*Theorem 9:* Let $e(r)$ be the error-exponent function for a hypothesis testing problem. Then, for any $r > 0$, decision regions defined by

$$\mathcal{U} = \{k \mid q_{2k}e^{e(r)} \leq q_{1k}e^{r}\}$$

$$\mathcal{U}^c = \{k \mid q_{2k}e^{e(r)} > q_{1k}e^{r}\}$$

are such that the following are satisfied simultaneously:

$$\alpha \leq e^{-e(r)}$$

$$\beta \leq e^{-r}.$$

*Proof:* Let $s$ parametrize the point $(r,e(r))$. The characteristic functions of the decision regions satisfy

$$\phi_{\mathcal{U}}(k) \leq \left(\frac{q_{1k}}{q_{2k}} e^{r-e(r)}\right)^{s/1+s}$$

$$\phi_{\mathcal{U}^c}(k) \leq \left(\frac{q_{2k}}{q_{1k}} e^{e(r)-r}\right)^{1/1+s}$$

which are verified by examining the case where the characteristic function equals zero and the case where it equals one. Therefore, for all positive $s$

$$\beta = \sum_{k \in \mathcal{U}^c} q_{1k} = \sum_{k} q_{1k}\phi_{\mathcal{U}^c}(k)$$

$$\leq \sum_{k} q_{1k} \left(\frac{q_{2k}}{q_{1k}}\right)^{1/1+s} e^{(e(r)-r)/1+s}$$

$$= e^{-r}.$$

Similarly

$$\alpha = \sum_{k \in \mathcal{U}} q_{2k} = \sum_{k} q_{2k}\phi_{\mathcal{U}}(k)$$

$$\leq \sum_{k} q_{2k} \left(\frac{q_{1k}}{q_{2k}}\right)^{s/1+s} e^{s(r-e(r))/1+s}$$

$$= e^{-e(r)}.$$

*Corollary 1:* Given a set of $n$ independent measurements, then, for any $r > 0$, decision regions can be defined such that the following are satisfied.

$$\alpha \leq e^{-ne(r)}$$

$$\beta \leq e^{-nr}.$$

*Proof:* Follows from Theorem 7.

Theorem 9 and Corollary 1 give upper bounds on the probability of error and give a rapid means of specifying decision regions based upon a given error probability specification.

The error-exponent function, therefore, is important as a tool for describing simply the performance of hypothesis testers and studying the possible compromises between type 1 and type 2 error probabilities.

The error-exponent function will be even more useful if we can show that the bounds in Theorem 9 are in some sense tight. We next show that this tightness exists when the number of measurements is large by providing a lower bound in terms of the error-exponent function. This theorem is similar to a theorem of Shannon, Gallager, and Berlekamp. Our proof maintains closer contact with the discrimination and is easier to follow. The result also is algebraically tighter with block length. The variances appearing in the theorem are defined as follows:

$$\sigma_1^2 = \sum_{k} q_k^* \left(\log \frac{q_k^*}{q_{1k}}\right)^2 - \left(\sum_{k} q_k^* \log \frac{q_k^*}{q_{1k}}\right)^2$$

$$\sigma_2^2 = \sum_{k} q_k^* \left(\log \frac{q_k^*}{q_{2k}}\right)^2 - \left(\sum_{k} q_k^* \log \frac{q_k^*}{q_{2k}}\right)^2$$

where $q^*$ achieves $e(r)$ for the value of $r$ under consideration.

*Theorem 10:* Let $\varepsilon > 0$ be given, and let $\gamma \in (0,1)$ be arbitrary. Suppose

$$\beta \leq \gamma e^{-(r+\varepsilon)}.$$

Then

$$\alpha \geq \left(1 - \frac{\sigma_1^2 + \sigma_2^2}{\varepsilon^2} - \gamma\right) e^{-(e(r)+\varepsilon)}.$$

*Proof:* The Neyman–Pearson theorem states that only decision regions of the form described in terms of a threshold $T$ need be considered. Let $T = r - e(r)$ and let $q^*$ achieve $e(r)$. (Notice that this implies the restriction $T \in [-J(q_2; q_1), J(q_1; q_2)]$. We will later see in Theorem 11 that other values of the threshold are of little interest.) Thus we can write

$$\mathcal{U} = \left\{ k \,\middle|\, \frac{q_{2k}}{\hat{q}_k} e^{e(r)} \leq \frac{q_{1k}}{\hat{q}_k} e^r \right\}$$

$$\mathcal{U}^c = \left\{ k \,\middle|\, \frac{q_{2k}}{\hat{q}_k} e^{e(r)} > \frac{q_{1k}}{\hat{q}_k} e^r \right\}$$

where $\hat{q}_k$ achieves $e(r)$ and is included in the denominator for later convenience. The errors are bounded as follows. Define

$$\mathcal{U}_1(\varepsilon) = \left\{ k \mid e^{-\varepsilon} < \frac{q_{2k}}{\hat{q}_k} e^{e(r)} \leq \frac{q_{1k}}{\hat{q}_k} e^r \right\}$$

$$\mathcal{U}_2(\varepsilon) = \left\{ k \mid e^{-\varepsilon} < \frac{q_{1k}}{\hat{q}_k} e^r < \frac{q_{2k}}{\hat{q}_k} e^{e(r)} \right\}$$

so that $\mathcal{U}_1(\varepsilon) \subset \mathcal{U}$, $\mathcal{U}_2(\varepsilon) \subset \mathcal{U}^c$, and

$$\alpha = \sum_{k \in \mathcal{U}} q_{2k} \geq \sum_{k \in \mathcal{U}_1(\varepsilon)} q_{2k}$$
$$\geq \sum_{k \in \mathcal{U}_1(\varepsilon)} \hat{q}_k e^{-(e(r)+\varepsilon)}.$$

A similar estimate applies to $\beta$. Thus

$$\alpha \geq e^{-(e(r)+\varepsilon)} \sum_{k \in \mathcal{U}_1(\varepsilon)} \hat{q}_k$$
$$\beta \geq e^{-(r+\varepsilon)} \sum_{k \in \mathcal{U}_2(\varepsilon)} \hat{q}_k.$$

We now estimate the summations. Let

$$\mathcal{U}_A = \left\{ k \mid e^{-\varepsilon} < \frac{q_{1k}}{\hat{q}_k} e^{e(r)} \right\}$$

$$\mathcal{U}_B = \left\{ k \mid e^{-\varepsilon} < \frac{q_{2k}}{\hat{q}_k} e^r \right\}.$$

Then $\mathcal{U}_1(\varepsilon) \cup \mathcal{U}_2(\varepsilon) = \mathcal{U}_A \cap \mathcal{U}_B$ and

$$\sum_{k \in \mathcal{U}_A \cap \mathcal{U}_B} \hat{q}_k \geq 1 - \sum_{k \in \mathcal{U}_A{}^c} \hat{q}_k - \sum_{k \in \mathcal{U}_B{}^c} \hat{q}_k.$$

These can now be estimated by Chebyshev's inequality. Let

$$\mathcal{U}_T = \left\{ k\colon \left| \log \frac{\hat{q}_k}{q_{1k}} e^{-r} \right| \geq \varepsilon \right\}$$
$$= \left\{ k\colon \left| \log \frac{\hat{q}_k}{q_{1k}} - \sum_k \hat{q}_k \log \frac{\hat{q}_k}{q_{1k}} \right| \geq \varepsilon \right\}.$$

Then $\mathcal{U}_A{}^c \subset \mathcal{U}_T$ and

$$\sum_{k \in \mathcal{U}_A{}^c} \hat{q}_k \leq \sum_{k \in \mathcal{U}_T} \hat{q}_k \leq \frac{\sigma_1{}^2}{\varepsilon^2}.$$

Similarly

$$\sum_{k \in \mathcal{U}_B{}^c} \hat{q}_k \leq \frac{\sigma_2{}^2}{\varepsilon^2}.$$

Therefore

$$\sum_{k \in \mathcal{U}_A \cap \mathcal{U}_B} \hat{q}_k \geq 1 - \frac{\sigma_1{}^2 + \sigma_2{}^2}{\varepsilon^2}$$

and if

$$\sum_{k \in \mathcal{U}_1(\varepsilon)} \hat{q}_k \leq \gamma$$

then

$$\sum_{k \in \mathcal{U}_2(\varepsilon)} \hat{q}_k \geq 1 - \frac{\sigma_1{}^2 + \sigma_2{}^2}{\varepsilon^2} - \gamma.$$

So the proof is complete.

If the measurements are now replaced by blocks of independent measurements, we have the following corollary.

*Corollary 2:* Let $\varepsilon > 0$ be given and let $\gamma \in (0,1)$ be arbitrary. Suppose

$$\beta \leq \gamma e^{-n(r+\varepsilon)}$$

Then

$$\alpha \geq \left(1 - \frac{\sigma_1{}^2 + \sigma_2{}^2}{n\varepsilon^2} - \gamma\right) e^{-n(e(r)+\varepsilon)}.$$

*Proof:* Replace $\varepsilon$ by $n\varepsilon$ in the theorem, and use Theorem 8 together with the fact that variances add for independent random variables.

We now show that stronger than $\exp[-nJ(q_2; q_1)]$ dependence of the type 2 error probability on block length results in a type 1 error probability which approaches 1 with blocklength. This is a direct analog of a theorem of Wolfowitz [16] and is proven in the same way. (See, for example, Gallager [9].)

*Theorem 11:* Let $c = J(q_2; q_1)$. Suppose $\beta \leq e^{-nr}$ where $r > c$. Then

$$\alpha > 1 - \frac{4\sigma^2}{n(r-c)^2} - e^{-n(r-c)/2}$$

where

$$\sigma^2 = \sum_k q_{2k} \left(\log \frac{q_{2k}}{q_{1k}}\right)^2 - \left(\sum_k q_{2k} \log \frac{q_{2k}}{q_{1k}}\right)^2.$$

*Proof:* Let $\mathcal{U}, \mathcal{U}^c$ be the decision regions, and let

$$\mathcal{U}^* = \{ y \mid q_1(y) \geq q_2(y) e^{-n(c+\varepsilon)} \}$$

$$\mathcal{U}^{*c} = \{ y \mid q_1(y) < q_2(y) e^{-n(c+\varepsilon)} \}$$

where $\varepsilon > 0$ is arbitrary and $y$ denotes a sequence of $n$ measurements. We now have

$$1 - \alpha = \sum_{y \in \mathcal{U}^c} q_2(y)$$
$$= \sum_{y \in \mathcal{U}^c \cap \mathcal{U}^*} q_2(y) + \sum_{y \in \mathcal{U}^c \cap \mathcal{U}^{*c}} q_2(y)$$
$$\leq \sum_{y \in \mathcal{U}^c \cap \mathcal{U}^*} q_1(y) e^{n(c+\varepsilon)} + \sum_{y \in \mathcal{U}^c \cap \mathcal{U}^{*c}} q_2(y).$$

This is further bounded by summing the first term over all of $\mathscr{U}^c$ and the second over all of $\mathscr{U}^{*c}$

$$1 - \alpha \leq \sum_{y \in \mathscr{U}^c} q_1(y) e^{n(c+\varepsilon)} + \sum_{y \in \mathscr{U}^{*c}} q_2(y)$$
$$\leq e^{-nr} e^{n(c+\varepsilon)} + \sum_{y \in \mathscr{U}^{*c}} q_2(y).$$

The second term can be written

$$\sum_{y \in \mathscr{U}^{*c}} q_2(y) = P\left[\left\{ y \mid \log \frac{q_2(y)}{q_1(y)} > n(c + \varepsilon)\right\}\right]$$

where

$$nc = \sum_y q_2(y) \log \frac{q_2(y)}{q_1(y)}$$

and this can be bounded by Chebyshev's inequality

$$\sum_{y \in \mathscr{U}^{*c}} q_2(y) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Therefore

$$1 - \alpha \leq e^{n(c - r + \varepsilon)} + \frac{\sigma^2}{n\varepsilon^2}.$$

This holds for any $\varepsilon > 0$. Hence, pick $\varepsilon = (r - c)/2$, thereby proving the theorem.

## IV. The Reliability-Rate Function

The error-exponent function $e(r)$ was found to be a useful measure of the performance of optimum hypothesis testers. In this section the analogous function for the channel coding problem is defined as a discrimination; it is called the reliability-rate function $E(R)$. Fano [2] notes that $E(R)$ can be put in the form of a discrimination. We, on the other hand, define $E(R)$ as an extremization problem over all such discriminations. Starting from this definition, the known properties of $E(R)$ are more simply developed and new such properties are found.

*Definition 3:* Let the conditional probability matrix $Q$ be given. Then the reliability-rate function $E(R)$ is given by

$$E(R) = \max_p \min_{\hat{Q} \in \mathscr{Q}_R} \sum_j \sum_k p_j \hat{Q}_{k|j} \log \frac{\hat{Q}_{k|j}}{Q_{k|j}}$$

where

$$\mathscr{Q}_R = \left\{ \hat{Q} \,\middle|\, \sum_j p_j \sum_k \hat{Q}_{k|j} \log \frac{\hat{Q}_{k|j}}{\sum_j p_j \hat{Q}_{k|j}} \leq R \right\}.$$

*Theorem 12:* $E(R)$ is a decreasing, convex, and hence a continuous function defined for $R \geq 0$. It is strictly decreasing in some interval $0 \leq R \leq R_{\max}$, and in this interval the solution satisfies the constraint with equality.

*Proof:* Consider the inner minimum, letting $E(p,R)$ denote this minimum. Suppose $R > R'$; then $\mathscr{P}_R \supset \mathscr{P}_{R'}$, hence $E(p,R) \leq E(p,R')$ and the constraint is satisfied with equality in the region where $E(p,R)$ is strictly decreasing. For each $p$, convexity follows since $E(p,R)$ is the minimum of a convex function subject to a convex constraint. Finally, $E(R)$ is convex since it is the pointwise maximum of a set of convex functions.

*Theorem 13:* $E(R)$ is positive for $R < C$ and zero for $R \geq C$, where $C$ is the channel capacity.

*Proof:* $E(R)$ is zero if and only if $\hat{Q} = Q$, which occurs if and only if

$$Q \in \mathscr{Q}_R(p).$$

Hence $E(R)$ is zero if and only if

$$\sum_j \sum_k p_j Q_{k|j} \log \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}} \leq R$$

for every value of $p$. This is true if and only if $R \geq C$.

We now obtain the properties of $E(R)$ by reference to the properties of $e(r)$. Consider the minimization problem

$$\min_q \min_{\hat{Q} \in \mathscr{Q}_R} J(\hat{Q}; Q)$$

where

$$\mathscr{Q}_R = \{\hat{Q} \mid J(\hat{Q}; q) \leq R\}$$

and $q$ is a dummy argument of a minimization operator. The inner minimum is of the form of the definition of $e(r)$ and so defines a function $e(R)$. Therefore, it can be solved in terms of a Lagrange multiplier $s$, which is the negative of the slope of $e(R)$. The Lagrange multiplier is, therefore, positive and

$$\min_q \min_{\hat{Q}} [J(\hat{Q}; Q) + sJ(\hat{Q}; q)]$$
$$= \min_{\hat{Q}} [J(\hat{Q}; Q) + sI(p; \hat{Q})]$$

since $q_k = \sum_j p_j \hat{Q}_{k|j}$ minimizes

$$J(\hat{Q}; q) = \sum_j \sum_k p_j \hat{Q}_{k|j} \log \frac{\hat{Q}_{k|j}}{q_k}$$

as a simple consequence of the positiveness of discrimination. Therefore, we have proved the following.

*Theorem 14:* $E(R)$ can be expressed parametrically as follows:

$$E(R_s) = \max_p [-sR_s + \min_q \min_{\hat{Q}} [J(\hat{Q}; Q) + sJ(\hat{Q}; q)]]$$

where

$$R_s = I(p; \hat{Q})$$

for the values of $p,\hat{Q}$, which achieve the solution.

Theorem 3 immediately gives the following corollary.

*Corollary 3:* $E(R)$ can be expressed as follows:

$$E(R_s) = \max_p \min_{\hat{q}} \left[ -sR_s - \sum_j p_j \log \left( \sum_k Q_{k|j}^{1/1+s} \hat{q}_k^{s/1+s} \right)^{1+s} \right]$$

where $R_s$ is given by

$$R_s = J(Q^*; q^*)$$

$p_j, q_k{}^*$ achieve the solution, and

$$Q^*_{k|j} = \frac{Q_{k|j}^{1/1+s} q_k^{*s/1+s}}{\sum_k Q_{k|j}^{1/1+s} q_k^{*s/1+s}}.$$

Now let $E(R,p)$ be the argument of the maximum over $p$ in Theorem 14, and let $Q^*(s)$ and $q^*(s)$ achieve the minima. Except for the fact that $q^*$ depends on $s$, Theorem 6 provides the derivative with respect to $R$. However, we have for the missing term in the derivative

$$\frac{d}{ds} \sum_j \sum_k p_j Q_{k|j} \log \frac{Q_{k|j}}{q_k(s)} = -\sum_k \sum_j p_j Q_{k|j} \frac{dq_k(s)/ds}{q_k(s)}$$

which is evaluated at the point $q_k(s) = \sum_j p_j Q_{k|j}$. But then the derivative becomes

$$\sum_k \frac{dq_k(s)}{ds} = \frac{d}{ds} \sum_k q_k(s) = 0.$$

Therefore, the dependence of $q_k$ on $s$ does not affect the derivative. We have then for each $p$

$$\frac{dE(R,p)}{dR} = -s$$

and since $E(R,p)$ is convex, this proves the following theorem.

*Theorem 15:*

$$E(R) = \max_p \max_{s>0} \min_{\hat{q}} \left[ -sR - \sum_j p_j \log \left( \sum_k Q_{k|j}^{1/1+s} \hat{q}_k^{s/1+s} \right)^{1+s} \right].$$

The two maximizations can be interchanged. Comparing the result with Corollary 3 and noting that $E(R)$ is convex, we find that

$$\frac{dE(R)}{dR} = -s$$

provided the derivative exists, since the maximum $s$ is over all straight lines lying below $E(R)$.

The Kuhn–Tucker theorem (9) can be applied to provide useful conditions on the distribution $p^*$ achieving $E(R)$. This is formally the same as the Kuhn–Tucker condition on the distribution achieving capacity of a constrained channel [10] if an expense schedule

$$e_j = \sum_k Q^*_{k|j} \log \frac{Q^*_{k|j}}{Q_{k|j}}$$

is defined. The Kuhn–Tucker conditions for a constrained channel then gives

$$\sum_k Q^*_{k|j} \log \frac{Q^*_{k|j}}{\sum_j p_j{}^* Q^*_{k|j}} + \frac{1}{s} \sum_k Q^*_{k|j} \log \frac{Q^*_{k|j}}{Q_{k|j}} \le \gamma$$

where $\gamma$ is a constant and the inequality is satisfied with equality if $p_j{}^* \neq 0$.

*Theorem 16:* Suppose $p^*$ achieves $E(R)$ and $q_k{}^* = \sum_j p_j{}^* Q^*_{k|j}$. Then

$$\max_p \left[ J(Q^*; q^*) + \frac{1}{s} J(Q^*; Q) \right] = \max_p \left[ I(p; Q^*) + \frac{1}{s} J(Q^*; Q) \right].$$

*Proof:* Multiplying the Kuhn–Tucker condition by $p_j{}^*$ and summing over $j$ gives

$$\gamma = I(p^*; Q^*) + \frac{1}{s} J(Q^*; Q) = \max_p \left[ I(p; Q^*) + \frac{1}{s} J(Q^*; Q) \right].$$

Multiplying the Kuhn–Tucker conditions by $p_j$ and summing over $j$ gives

$$J(Q^*; q^*) + \frac{1}{s} J(Q^*; Q) \le \gamma$$

for all $p$. Since equality is achieved by choosing $p = p^*$, the theorem follows.

*Theorem 17:* Suppose $p^*$ achieves $E(R)$; then

$$\sum_k Q_{k|j}^{1/1+s} \left( \sum_j p_j{}^* Q_{k|j}^{1/1+s} \right)^s = \lambda, \qquad \text{if } p_j{}^* \neq 0$$

$$\sum_k Q_{k|j}^{1/1+s} \left( \sum_j p_j{}^* Q_{k|j}^{1/1+s} \right)^s \ge \lambda, \qquad \text{if } p_j{}^* = 0$$

where

$$\lambda = \sum_j p_j{}^* \sum_k Q_{k|j}^{1/1+s} \left( \sum_j p_j{}^* Q_{k|j}^{1/1+s} \right)^s = \sum_k \left( \sum_j p_j{}^* Q_{k|j}^{1/1+s} \right)^{1+s}.$$

*Proof:* The Kuhn–Tucker condition can be rewritten

$$\frac{1+s}{s} \sum_k Q^*_{k|j} \log \frac{Q^*_{k|j}}{q_k^{*s/1+s} Q_{k|j}^{1/1+s}} \le \gamma$$

or

$$\frac{1+s}{s} \sum_k Q^*_{k|j} \log \frac{1}{\sum_k Q_{k|j}^{1/1+s} q_k^{*s/1+s}} \le \gamma$$

since from Corollary 3

$$Q^*_{k|j} = \frac{Q_{k|j}^{1/1+s} q_k^{*s/1+s}}{\sum_k Q_{k|j}^{1/1+s} q_k^{*s/1+s}}.$$

Carrying out the outer sum, this reduces to

$$\sum_k Q_{k|j}^{1/1+s} q_k^{*s/1+s} \ge A$$

with equality for all $j$ with $p_j \neq 0$, where $A$ is a constant. Now $q_k{}^*$ also satisfies

$$q_k{}^* = \sum_j p_j{}^* Q^*_{k|j} = \sum_j p_j{}^* \frac{Q_{k|j}^{1/1+s} q_k^{*s/1+s}}{\sum_k Q_{k|j}^{1/1+s} q_k^{*s/1+s}} = \frac{1}{A} q_k^{*s/1+s} \sum_j p_j{}^* Q_{k|j}^{1/1+s}.$$

This gives

$$q_k^{*s/1+s} = \left(\frac{1}{A}\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^s.$$

Using this in the preceeding inequality completes the proof.

*Corollary 4:* Suppose $p^*,Q^*$ achieve $E(R)$, and let $q^*$ be given by

$$q_k^* = \sum_j p_j^* Q_{k|j}^*.$$

Then the following are satisfied:

$$1) \qquad q_k^* = \frac{\left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^{1+s}}{\sum_k \left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^{1+s}}$$

$$2) \qquad Q_{k|j}^* = \frac{Q_{k|j}^{1/1+s}\left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^s}{\sum_k Q_{k|j}^{1/1+s}\left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^s}.$$

*Proof:* The proof of Theorem 17 shows that

$$q_k^* = \left(\frac{1}{A}\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^{1+s}.$$

Selecting $A$ so that $\sum_k q_k^* = 1$ gives the first condition. The second condition is obtained by substituting the first condition into

$$Q_{k|j}^* = \frac{Q_{k|j}^{1/1+s} q_k^{*s/1+s}}{\sum_k Q_{k|j}^{1/1+s} q_k^{*s/1+s}}.$$

We now represent $E(R)$ as a maximization problem. This is done in terms of three similar expressions, including the simple form most frequently used.

*Theorem 18:* $E(R)$ can be expressed by either of the following three expressions:

$$1) \quad E(R) = \max_{s\ge 0} \max_p \left[-sR - (1+s) \cdot \sum_j p_j \log\left(\sum_k Q_{k|j}^{1/1+s}\left(\sum_j p_j Q_{k|j}^{1/1+s}\right)^s\right) + s \log \sum_k \left(\sum_j p_j Q_{k|j}^{1/1+s}\right)^{1+s}\right]$$

$$2) \quad E(R) = \max_{s\ge 0} \max_p \left[-sR - \sum_j p_j \cdot \log\left(\sum_k Q_{k|j}^{1/1+s}\left(\sum_j p_j Q_{k|j}^{1/1+s}\right)^s\right)\right]$$

$$3) \quad E(R) = \max_{s\ge 0} \max_p \left[-sR - \log \sum_k \left(\sum_j p_j Q_{k|j}^{1/1+s}\right)^{1+s}\right].$$

*Proof:* Write

$$E(R) = \max_{s\ge 0}\left[-sR + \max_p \min_{\hat{Q}} \min_q \cdot \left[\sum_j\sum_k p_j \hat{Q}_{k|j} \log \frac{\hat{Q}_{k|j}}{Q_{k|j}} + s \sum_j \sum_k p_j \hat{Q}_{k|j} \log \frac{\hat{Q}_{k|j}}{q_k}\right]\right]$$

and take the optimizing values of $Q,q$ given by Corollary 4. Substitution gives

$$E(R) = \max_{s\ge 0} \max_p \left[-sR + \sum_j p_j \cdot \log \frac{\left(\sum_k \left(\sum_j p_j Q_{k|j}^{1/1+s}\right)^{1+s}\right)^s}{\left(\sum_k Q_{k|j}^{1/1+s}\left(\sum_j p_j Q_{k|j}^{1/1+s}\right)^s\right)^{1+s}}\right]$$

which is just the first of the preceeding expressions. It remains to prove the equivalence of the three expressions. Denote these temporarily by $E_1(R)$, $E_2(R)$, and $E_3(R)$. Since the log is concave, we can use Jensen's inequality to write

$$E(R) = E_1(R) \ge E_2(R) \ge E_3(R).$$

Equality will hold if we can also show that $E_3(R) \ge E_1(R)$. Let $p^*$ achieve $E(R)$. Then

$$E_3(R) \ge \max_{s\ge 0}\left[-sR - \log \sum_k \left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^{1+s}\right]$$

$$= \max_{s\ge 0}\left[-sR - (1+s)\log \sum_k \left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^{1+s} + s \log \sum_k \left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^{1+s}\right].$$

However, $p^*$ achieves $E(R)$, so by Theorem 15 this becomes

$$E_3(R) \ge \max_{s\ge 0}\left[-sR - (1+s) \cdot \sum_j p_j^* \log\left(\sum_k Q_{k|j}^{1/1+s}\left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^s\right) + s \log \sum_k \left(\sum_j p_j^* Q_{k|j}^{1/1+s}\right)^{1+s}\right]$$

and the right side is $E_1(R)$. Therefore, we have

$$E(R) = E_1(R) \ge E_2(R) \ge E_3(R) \ge E_1(R)$$

and the theorem is proved.

## V. Lower Bounds to Block Code Performance

The application of $E(R)$ to the upper bounds for channel block codes is well known [4]. Similarly, the application to lower bounds is provided by the sphere-packing bound developed in Shannon, Gallager, and Berlekamp [5]. The present section provides a simplified approach to the sphere-packing bound by posing the channel decoding problem in terms of a simple binary hypothesis testing problem. The present treatment also results in the minor improvement of an algebraically tighter bound.

The probability of error is to be bounded over the set of all block codes of a given rate and block length. A codeword is a sequence of symbols from the source alphabet. The hypothesis that a given codeword was transmitted, therefore, can be restated as the hypothesis that a particular sequence of symbols was transmitted, which is of the form of hypothesis testing with measurements of different type as has been discussed. Accordingly, we shall lowerbound the performance of block codes via considerations which allow us to appeal to Theorem 10.

An important feature of the lower bound, which will be derived, is that no assumption of constant-composition codewords is made, not even as an intermediate step. This is an improvement to previous proofs of the sphere-packing bound since it permits us to generalize the lower bound to continuous channels with no change in technique.

Under the preceding restrictions, the most general coding scheme considered here is a set of $M$ $n$-tuples each used with probability $1/M$, together with a partition of the space of output words into $M$ decode regions $\{\mathcal{U}_m \mid m = 1,\cdots,M\}$. Reception of an output in $\mathcal{U}_m$ causes a decode of the $m$th codeword. Therefore, a lower bound must underestimate the probability of error under the assumption that the $M$ codewords and the decode regions have been selected optimally.

The performance of the code can be characterized by either 1) maximum probability over the set of codewords, or 2) average probability over the set of codewords. We will bound the first of these by identifying a suitable hypothesis testing problem so that the theory of Section III can be applied.

We first prove a simple lemma.

*Lemma 1:* Suppose $\{\mathcal{U}_m \mid m = 1,\cdots,M\}$ is a partition of a set and $\hat{q}$ is a probability distribution on this same set. Then for some $\mathcal{U}_m$

$$\sum_{y\in\mathcal{U}_m} \hat{q}(y) \leq \frac{1}{M}.$$

*Proof:*

$$\sum_{m=1}^{M} \sum_{y\in\mathcal{U}_m} \hat{q}(y) = \sum_{y} \hat{q}(y) = 1.$$

The lower bound theorem is then as follows.

*Theorem 19:* Let $(p_e)_{\max}$ be the maximum probability of error over all codewords of a block code having $M = e^{nR}$ codewords. Let $\varepsilon > 0$ and $\gamma \in (0,1)$ be arbitrary, and suppose $R^*$ satisfies

$$\frac{1}{M} \leq \gamma e^{-n(R^*+\varepsilon)}.$$

Then

$$(p_e)_{\max} \geq \left(1 - \frac{A}{n\varepsilon^2} - \gamma\right) e^{-n(E(R^*)+\varepsilon)}$$

where $A$ is an appropriate constant.

*Proof:* Let $\hat{q}$ be any arbitrary distribution on the output letters and $\hat{q}(y)$ the associated product distribution on the output sequence $y$. Select $m$ so that $\mathcal{U}_m$ is the decode region satisfying Lemma 1. Then

$$\sum_{y\in\mathcal{U}_m} \hat{q}(y) \leq \frac{1}{M} \leq \gamma e^{-n(R^*+\varepsilon)}.$$

We are now in a position to use the theory of Section III. The hypotheses are now characterized by $Q(y \mid x_m)$ and $\hat{q}(y)$, and these replace $Q_2(y \mid x)$ and $Q_1(y \mid x)$, respectively. The probability of error, $p_{e|m}$, replaces $\alpha$ and $1/M$ replaces $\beta$. Since the dummy alternate hypothesis characterized by $\hat{q}$ has probability of resulting in $\mathcal{U}_m$ of less than $1/M$, we can lowerbound $p_{e|m}$ using Theorem 10. This theorem holds for any pair of sets, and we can use the theorem directly without the necessity of developing additional structure.

Thus we consider the pair of sets $\{\mathcal{U}_m,\mathcal{U}_m{}^c\}$ and note that

$$p_{e|m} = \sum_{y\in\mathcal{U}_m{}^c} Q(y \mid x_m).$$

Hence by Corollary 2

$$p_{e|m} \geq \left(1 - \frac{\sigma_1{}^2 + \sigma_2{}^2}{n\varepsilon^2} - \gamma\right) e^{-n(e(R^*)+\varepsilon)}$$

and

$$(p_e)_{\max} \geq p_{e|m}$$

where $\sigma_1,\sigma_2$ are appropriate variances, and by definition

$$e(R^*) = \min_{\hat{Q}\in\mathcal{Q}_{R^*}} \sum_j p_j \sum_k \hat{Q}_{k|j} \log \frac{\hat{Q}_{k|j}}{Q_{k|j}}$$

where

$$\mathcal{Q}_{R^*} = \left\{\hat{Q} \,\middle|\, \sum_j p_j \sum_k \hat{Q}_{k|j} \log \frac{\hat{Q}_{k|j}}{\hat{q}_k} \leq R^*\right\}$$

and $p_j$ is the relative frequency of the composition of the $m$th codeword, and the constraint is satisfied with equality unless $e(R^*) = 0$.

Since the composition of the $m$th word is unknown, we bound $e(R^*)$ by

$$e(R^*) \leq \max_{p} \min_{\hat{Q}\in\mathcal{Q}_{R^*}} \sum_j p_j \sum_k \hat{Q}_{k|j} \log \frac{\hat{Q}_{k|j}}{Q_{k|j}}.$$

Now since $\mathcal{Q}_{R^*}$ depends on $\hat{q}$, the right side depends on $\hat{q}$, and bounds $e(R^*)$ for every $\hat{q}$.

We now wish to choose $\hat{q}_k = \sum_j p_j \hat{Q}_{k|j}$. However, the set $\mathcal{U}_m$ depends on $\hat{q}$, and if $\hat{q}$ depends in turn on the composition of the $m$th codeword, the maximization over $p$ has no apparent meaning. Instead, let $p^*$ achieve $E(R)$ and choose $\hat{q} = q^*$, where $q_k{}^* = \sum_j p_j{}^* Q_{k|j}^*$ and $Q^*$ achieves $E(R)$. This choice does not depend on the composition of the $m$th codeword, so there is no possibility of circularity in the argument. Next attach the constraint $J(\hat{Q};\, q^*) \leq R^*$ by means of a Lagrange multiplier. Then

$$e(R^*) \leq \max_{p} \min_{\hat{Q}} \left[J(\hat{Q};\, Q) + sJ(\hat{Q};\, q^*)\right].$$

Without violating the inequality, we set $\hat{Q} = Q^*$ and drop the minimum on $\hat{Q}$. Then

$$e(R^*) \leq \max_{p} \left[J(Q^*,Q) + sJ(Q;\, q^*)\right].$$

Now use Theorem 16 to replace $J(Q^*;\, q^*)$ by $I(p;\, Q^*)$. That is,

$$\sum_j \sum_k p_j Q_{k|j}^* \log \frac{Q_{k|j}^*}{\sum_j p_j{}^* Q_{k|j}^*} \Rightarrow \sum_j \sum_k p_j Q_{k|j}^* \log \frac{Q_{k|j}^*}{\sum_j p_j Q_{k|j}^*}$$

and

$$e(R^*) \leq \max_{p} \left[J(Q^*;\, Q) + sI(p;\, Q^*)\right].$$

Now recall the definition of $E(R)$. Since $E(R^*)$ occurs at a saddle point in $(p,\hat{Q})$, the right side of the preceding equation can be replaced to give

$$e(R^*) \leq \max_{p} \min_{\hat{Q}} [J(\hat{Q}; Q) + sI(p; \hat{Q})]$$

so that

$$e(R^*) \leq E(R^*).$$

Hence

$$(p_e)_{\max} \geq \left(1 - \frac{\sigma_1{}^2 + \sigma_2{}^2}{n\varepsilon^2} - \gamma\right) e^{-n(E(R^*)+\varepsilon)}.$$

Since $\sigma_1{}^2 + \sigma_2{}^2$ is constant, the theorem is proved.

The constant $A$ is only of slight interest. It can be evaluated by evaluating $\sigma_1{}^2 + \sigma_2{}^2$. Since

$$\sigma_1{}^2 + \sigma_2{}^2 = \sum_j p_j \left[\operatorname{var}\left(\log \frac{Q^*_{k|j}}{\hat{q}_k}\right) + \operatorname{var}\left(\log \frac{Q^*_{k|j}}{Q_{k|j}}\right)\right]$$

and since the composition $p_j$ has not been determined, we take $A$ as

$$A = \max_j \left[\operatorname{var}\left(\log \frac{Q^*_{k|j}}{\hat{q}_k}\right) + \operatorname{var}\left(\log \frac{Q^*_{k|j}}{Q_{k|j}}\right)\right].$$

We have now a bound on the maximum probability of error. This can be used to obtain a bound on the average probability of error. The standard technique is used in the following theorem.

*Theorem 20:* Let $p_e$ be the average probability of error of a block code having $M = e^{nR}$ codewords. Let $\varepsilon > 0$ and $\gamma \in (0,1)$ be arbitrary. Suppose $R^*$ satisfies

$$\frac{1}{M} \leq \gamma e^{-n(R^*+\varepsilon)}.$$

Then

$$p_e \geq \frac{1}{2}\left(1 - \frac{A}{n\varepsilon^2} - \gamma\right) e^{-n[E(R^*-(\log 2)/n)+\varepsilon]}.$$

*Proof:* Remove the $M/2$ codewords whose probability of error is largest. This results in a code having $M/2 = e^{nR-\log 2}$ codewords which must satisfy Theorem 19. But we are given that

$$\frac{1}{M} \leq \gamma e^{-n(R^*+\varepsilon)}.$$

Hence

$$\frac{2}{M} \leq \gamma e^{-n(R^*-(\log 2)/n+\varepsilon)}.$$

Let $\mathscr{C}_p$ denote the set of codewords in the purged code and $(p_e)^p_{\max}$ denote the maximum error of this purged code. Then the average error of the original code is bounded as follows:

$$p_e = \frac{1}{M}\sum_m p_{e|m} \geq \frac{1}{M}\sum_{m \in \mathscr{C}_p{}^c} p_{e|m}$$

$$\geq \frac{1}{M}\sum_{m \in \mathscr{C}_p{}^c} (p_e)^p_{\max} = \tfrac{1}{2}(p_e)^p_{\max}.$$

Invoking Theorem 19 completes the proof.

## VI. The Source Code Reliability-Rate Function

The optimum performance of distortionless source codes of block length $n$ and rate $R$ has been studied by Jelinek [11]. He bounds the probability that the source will generate a block which cannot be encoded. For fixed rate $R$, this probability decreases exponentially with block length, the rate of decay being given by a function of rate.

This section develops a similar bound for the case of source compression codes, which reduces to Jelinek's bound when the distortion is zero. We first develop the distortionless bound in terms of the discrimination.

*Definition 4:* Let a memoryless discrete source be characterized by a distribution $p$. The distortionless reliability-rate function is

$$F(R) = \min_{\hat{p} \in \mathscr{P}_R} \sum \hat{p}_j \log \frac{\hat{p}_j}{p_j}$$

$$\mathscr{P}_R = \left\{\hat{p}: -\sum_j \hat{p}_j \log \hat{p}_j \geq R\right\}.$$

This is similar to the form of Definition 3. Similar arguments show that $F(R)$ is a convex increasing function. It is strictly increasing in the interval $H(p) \leq R \leq \log J$ and is zero for $R \leq H(p)$, where $H(p)$ is the entropy and $J$ is the size of the source alphabet.

By introducing a Lagrange multiplier $s$, we can solve the minimization problem defining $F(R)$. This multiplier is found to be the slope of $F(R)$, and by the convexity of $F(R)$, the solution can be written

$$F(R) = \max_{s \geq 0}\left[sR - \log\left(\sum_j p_j^{1/1+s}\right)^{1+s}\right].$$

This form shows that $F(R)$ is identical to the exponent of Jelinek.

We now introduce the reliability-rate function for source compression codes and develop some of its major properties. Later, a source coding theorem is proved that states that every source word can be encoded with distortion at most $nD$ except for a set of source words of probability $p_e \leq e^{-n[F(R,D)+o(n)]}$. This is a stronger form of the source coding theorem, which usually is concerned only with average distortion. Knowledge of where $F(R,D)$ is positive provides a sufficient condition for the existence of good codes for large $n$. Therefore, the present section also investigates the region of positive $F(R,D)$ and shows that it occurs for $R$ greater than the rate-distortion function, thereby providing the basis for an alternate proof of the source compression coding theorem.

*Definition 5:* Let $\rho_{jk}$ be a single-letter distortion function associated with a memoryless source. The reliability-rate function for associated source compression codes is given by

$$F(R,D) = \max_{Q} \min_{\hat{p} \in \mathscr{P}_{R,D}} \sum_j \hat{p}_j \log \frac{\hat{p}_j}{p_j}$$

where

$$\mathscr{P}_{R,D} = \left\{\hat{p}: \sum_j \sum_k \hat{p}_j Q_{k|j} \log \frac{Q_{k|j}}{\sum_j \hat{p}_j Q_{k|j}} \geq R,\right.$$

$$\left.\sum_j \sum_k \hat{p}_j Q_{k|j} \rho_{jk} \geq D\right\}.$$

Notice that the set $\mathscr{P}_{R,D}$ involves constraints that are related to the usual definition of the rate-distortion function[1]

$$B(D) = \min_{Q \in \mathscr{Q}_D} \sum_j \sum_k p_j Q_{k|j} \log \frac{Q_{k|j}}{\sum_j p_j Q_{k|j}}$$

$$\mathscr{Q}_D = \left\{ Q : \sum_j \sum_k p_j Q_{k|j} \rho_{jk} \leq D \right\}$$

so that $F(R,D)$ should have an intimate relationship with $B(D)$.

*Theorem 21:* $F(R,D)$ is convex in both variables. It is increasing in $R$ and in $D$.

*Proof:* The inner minimization is the minimum of a convex function subject to convex constraints and is convex in $\langle R,D \rangle$. The outer maximum is then over a set of convex functions and hence yields a convex function. Similarly, the inner minimum on $p$ is increasing in $R$ and in $D$ and hence $F(R,D)$ also behaves in this way.

*Theorem 22:* $F(R,D)$ is positive if $R > B(D)$ and is zero if $R \leq B(D)$.

*Proof:* $F(R,D)$ is zero if and only if $\hat{p} = p$, and it is otherwise positive. Thus, the minimization problem defining $F(R,D)$ will achieve its minimum at $\hat{p} = p$ and so equals zero if and only if $p \in \mathscr{P}_{R,D}$ for every value of $\hat{Q}$. In particular, $p \in \mathscr{P}_{R,D}(\hat{Q})$ for the value of $\hat{Q}$ achieving $B(D)$, which implies that $B(D) = I(p,\hat{Q}) \geq R$. Thus $F(R,D) = 0$ implies $R \leq B(D)$. Conversely, if $R > B(D)$, then a $\hat{Q}$ achieving $B(D)$ is such that $I(p,\hat{Q}) < R$, and hence $p \notin \mathscr{P}_{R,D}(\hat{Q})$. Therefore $F(R,D) > 0$.

In the range of interest, $F(R,D)$ is strictly increasing in both variables; hence, the constraints involving $R$ and $D$ are satisfied with equality. Therefore, we can evaluate $F(R,D)$ by means of Lagrange multipliers. This gives the parametric representation

$$F(R,D) = sR - stD + \max_Q \min_{\hat{p}} \left[ \sum_j \hat{p}_j \log \frac{\hat{p}_j}{p_j} - s \left( \sum_j \sum_k \hat{p}_j Q_{k|j} \log \frac{Q_{k|j}}{\sum_j \hat{p}_j Q_{k|j}} - t \sum_j \sum_k \hat{p}_j Q_{k|j} \rho_{jk} \right) \right]$$

where $R,D$ are given in terms of the optimizing distributions.

*Theorem 23:* If $p^*,Q^*$ achieve $F(R,D)$ and $q^*$ is given by

$$q_k^* = \sum_j p_j^* Q_{k|j}^*$$

then the following are satisfied

$$Q_{k|j}^* = \frac{q_k^* e^{t\rho_{jk}}}{\sum_k q_k^* e^{t\rho_{jk}}} \qquad p_j^* = \frac{p_j \left( \sum_k q_k^* e^{t\rho_{jk}} \right)^{-s}}{\sum_j p_j \left( \sum_k q_k^* e^{t\rho_{jk}} \right)^{-s}}.$$

*Proof:* The maximin occurs at a true saddle point and hence equals the minimax. The first relation is obtained in the same way as is the similar expression for rate-distortion functions. The second relation is obtained by first finding the result analogous to Theorem 17 and then simplifying by using the expression for $Q_{k|j}^*$.

We now make use of Theorem 23 in order to obtain a convenient representation of $F(R,D)$.

*Theorem 24:* $F(R,D)$ can be expressed as follows.

$$F(R,D) = sR - stD + \max_q \left[ -\log \sum_j p_j \left( \sum_k q_k e^{t\rho_{jk}} \right)^{-s} \right]$$

where $s \in [0,\infty]$, $t \in [-\infty,0]$ are arbitrary, and $R,D$ are given in terms of the optimizing $q$.

This theorem evaluates $F(R,D)$ parametrically at fixed $s$ and $t$. A representation, which evaluates $F(R,D)$ at fixed $R$ and $D$ and is analogous to the most common representation of $E(R)$, is given next.

*Theorem 25:* $F(R,D)$ can be expressed as follows:

$$F(R,D) = \max_{s \geq 0} \min_{t \leq 0} \max_q \left[ sR - stD - \log \sum_j p_j \left( \sum_k q_k e^{t\rho_{jk}} \right)^{-s} \right].$$

*Proof:* In the parametric representation following Theorem 22, the bracketed term is convex in $\hat{p}$ and concave in $Q$. Therefore, the solution is at a saddle point and the maximin equals the minimax. Taking the maximum first is equivalent to solving the parenthesized rate-distortion function, and we can write the equivalent form [10]

$$F(R,D) = sR_s + \min_{\hat{p}} \left[ \sum_j \hat{p}_j \log \frac{\hat{p}_j}{p_j} - s \left[ \max_{t \leq 0} \min_q \left[ tD + \sum_j p_j \log \left( \sum_k q_k e^{t\rho_{jk}} \right) \right] \right] \right].$$

Now substitute the optimizing $p^*$ from Theorem 23. This results in

$$F(R,D) = sR_s + \min_{t \leq 0} \max_q \left[ -stD - \log \sum_j p_j \left( \sum_k q_k e^{t\rho_{jk}} \right)^{-s} \right].$$

Finally, the value of $s$ that must be used to achieve $F(R,D)$ at a given $R$ is that positive value of $s$ that maximizes the entire right side, since $F$ is convex in $R$ and

$$\frac{\partial F}{\partial R} = s.$$

The channel code reliability-rate function provides an upper bound to the performance of channel codes only when the Lagrange multiplier $s$ is smaller than one. A similar limitation occurs in the following. We define a restricted version of $F(R,D)$ which will be used.

[1] We use the notation $B(D)$ rather than the usual $R(D)$ to avoid confusion with the code rate $R$ appearing in $F(R,D)$.

*Definition 6:* The restricted reliability-rate function for source codes $F_0(R,D)$ is given by

$$F_0(R,D) = \max_{s\in[0,1]} \min_{t\leq 0} \max_{q} \left[ sR - stD - \log \sum_j p_j \left( \sum_k q_k e^{t\rho_{jk}} \right)^{-s} \right].$$

Notice that this differs from $F(R,D)$ only in the restriction of the range of $s$, and that $F_0(R,D)$ is zero if and only if $F(R,D)$ is zero. This representation of $F_0(R,D)$ will be useful in proving the upper bound of Theorem 26. The output probability distribution $q$ will arise as a distribution on an ensemble of codes. Anticipating this result, we point out that the distribution $q$ achieving $F_0(R,D)$ is a good distribution with which to randomly select codewords for a code of rate $R$ and distortion $D$. As we shall see, such a random selection of codewords will usually give a good code. Roughly speaking, some good codes of rate $R$ and distortion $D$ will have a composition approximated by $q$. A lower bound is required before we can assert that all good codes have this distribution.

We now are ready to prove a source compression theorem. For a given block length $n$ and distortion $D$, we are interested in the probability of occurrence of a source word that cannot be encoded with distortion less than or equal to $nD$. We shall see that, for appropriate codes, an upper bound on this probability goes to zero exponentially with block length, with decay coefficient given by $F_0(R,D)$. The source coding theorem then follows by using Theorem 22 to note where $F_0(R,D)$ is positive.

*Theorem 26:* It is possible to select $M = e^{nR}$ codewords so that the probability that a source word occurs, which cannot be encoded with distortion $nD$ (or less), satisfies

$$p_e \leq e^{-n[F_0(R,D)+o(n)]}$$

where $o(n) \to 0$ as $n \to \infty$.

*Proof:* We shall employ a random coding argument. First let

$$T = \{\langle x,y \rangle \mid \rho_n(x,y) \leq nD\}$$

$$T(x) = \{y \mid \rho_n(x,y) \leq nD\}$$

and let $s,t$ be Lagrange multipliers associated with $F_0(R,D)$ at $R,D$. Then

$$p_e = \sum_x p(x) \prod_{m=1}^{M} (1 - \phi(x,y_m))$$

where $\{y_m\}$ is the set of codewords and $\phi$ is the characteristic function of the set $T$. Now select the codewords at random with an independently identically distributed (i.i.d.) distribution $q(y)$. Then the expected value of $p_e$ is

$$\bar{p}_e = \sum_x p(x) \left( 1 - \sum_y q(y)\phi(x,y) \right)^M$$

$$= \sum_x p(x) \left( 1 - \sum_{y\in T(x)} q(y) \right)^M.$$

Now let

$$\mathscr{U} = \left\{ x: \sum_y q(y) e^{t(\rho(x,y)-nD)} \geq e^{-nR} \right\}$$

so that

$$\bar{p}_e \leq \sum_{\mathscr{U}^c} p(x) + \sum_{\mathscr{U}} p(x) \left( 1 - \sum_{T(x)} q(y) \right)^M$$

$$\leq \sum_{\mathscr{U}^c} p(x) \left( \frac{e^{-nR}}{\sum_y q(y) e^{t(\rho(x,y)-nD)}} \right)^s$$

$$+ \sum_{\mathscr{U}} p(x) \exp \left[ -M \sum_{T(x)} q(y) \right].$$

Replacing the first term by a sum over all $x$, noting the form of Theorem 24, and naming the second term $p_{e2}$ gives the following

$$\bar{p}_e \leq e^{-nF_0(R,D)} + p_{e2}$$

where

$$p_{e2} = \sum_{\mathscr{U}} p(x) \exp \left[ -M \sum_{T(x)} q(y) \right].$$

The summation in the exponent is a cumulative distribution function and can be upperbounded by means of Chernoff bounds [15]. Using Jelinek [11], Theorem 5.7, we have the following

$$\sum_{T(x)} q(y) \geq \tfrac{1}{2} \exp n \left[ \gamma_x(t) - t\gamma_x'(t) + t \frac{\sqrt{2\gamma_x''(t)}}{\sqrt{n}} \right]$$

where

$$\gamma_x(t) = n^{-1} \sum_j n(j \mid x) \log \sum_k q_k e^{t\rho_{jk}}$$

$$= n^{-1} \log \sum_y q(y) e^{t\rho(x,y)}.$$

Pick $t \leq 0$ so that $\gamma_x'(t) = D$. This gives

$$p_{e2} \leq \sum_{\mathscr{U}} p(x) \exp \left\{ -\exp \left[ nR + \log \sum_y q(y) e^{t(\rho(x,y)-nD)} + \log \xi \right] \right\}$$

where

$$\xi = \tfrac{1}{2}[\exp t\sqrt{2n\gamma_x''(t)}].$$

But, by the definition of $\mathscr{U}$

$$nR + \log \sum_y q(y) e^{t(\rho(x,y)-nD)} \geq 0.$$

Therefore, for $s \in [0,1]$

$$p_{e2} \leq \sum_{\mathscr{U}} p(x) \exp \left\{ -\exp \left[ s \left[ nR + \log \sum_y q(y) \cdot e^{t(\rho(x,y)-nD)} \right] + \log \xi \right] \right\}.$$

TABLE I

| Channel ($Q_{k\|j}$) | Source ($p_j$) |
|---|---|
| Capacity<br>$C = \max_{p} I(p,Q)$ | Entropy<br>$H = \min_{Q} I(p,Q)$ |
| Expense Schedule ($e_j$) | Distortion Measure ($\rho_{jk}$) |
| Capacity-Expense Function<br>$C(S) = \max_{p \in \mathscr{P}_S} I(p,Q)$<br>$\mathscr{P}_S = \{p \mid e(p) \leq S\}$ | Rate-Distortion Function<br>$B(D) = \min_{Q \in \mathscr{Q}_D} I(p,Q)$<br>$\mathscr{Q}_D = \{Q \mid d(Q) \leq D\}$ |
| Reliability-Rate Function<br>$E(R) = \max_{p} \min_{\hat{Q} \in \mathscr{Q}_R} J(\hat{Q},Q)$<br>$\mathscr{Q}_R = \{\hat{Q} \mid I(p,\hat{Q}) \leq R\}$ | Reliability-Rate Function<br>$F(R) = \max_{Q} \min_{\hat{p} \in \mathscr{P}_R} J(\hat{p},p)$<br>$\mathscr{P}_R = \{\hat{p} \mid I(\hat{p},Q) \geq R\}$<br>Note: max achieved when $Q$ is the identity |
| Constrained Reliability-Rate Function<br>$E(R,S) = \max_{p \in \mathscr{P}_S} \min_{\hat{Q} \in \mathscr{Q}_R} J(\hat{Q},Q)$<br>$\mathscr{Q}_R = \{\hat{Q} \mid I(p,\hat{Q}) \leq R\}$<br>$\mathscr{P}_S = \{p \mid e(p) \leq S\}$ | Constrained Reliability-Rate Function<br>$F(R,D) = \max_{Q} \min_{\hat{p} \in \mathscr{P}_{RD}} J(\hat{p},p)$<br>$\mathscr{P}_{RD} = \{\hat{p} \mid I(\hat{p},Q) \geq R,$<br>$\hat{d}(Q) \geq D\}$ |

Now use the inequality $e^x \geq 1 + x$

$$p_{e2} \leq e^{-1}\xi^{-1} \sum_{\mathscr{U}} p(x)e^{-snR} \left(\sum_{y} q(y)e^{t(\rho(x,y)-nD)}\right)^{-s}.$$

Hence, replacing the sum over $\mathscr{U}$ by a sum over all $x$ gives

$$p_{e2} \leq e^{-1}\xi^{-1}e^{-nF_0(R,D)}.$$

Finally, we have

$$\bar{p}_e \leq e^{-nF_0(R,D)} + e^{-1}\xi^{-1}e^{-nF_0(R,D)}$$
$$\leq (1 + 2e^{-1}e^{-t\sqrt{2n}\,\sigma})e^{-nF_0(R,D)}$$

where

$$\sigma^2 = \gamma_x''(t).$$

Moving the parenthesized term into the exponent completes the proof of the theorem.

Since $F(R,D)$ is positive for $R > B(D)$, we can find good codes in this region. We next state the source compression coding theorem as usual in terms of an average distortion rather than in terms of a maximum distortion.

*Theorem 27 (Source compression coding theorem):* Given a finite-alphabet memoryless source with bounded fidelity criterion, any $\varepsilon > 0$, and any positive $D$ it is possible to choose $M$ codewords such that the average distortion is less than $D$, provided that

$$M \geq e^{n[B(D)+\varepsilon]}$$

and $n$ is sufficiently large.

*Proof:* Follows immediately from Theorem 22 and Theorem 26.

## VII. Summary

The various performance bounds of information theory have been organized in a common setting. This was done in part by emphasizing the fundamental role played by the discrimination. These bounds are discriminations evaluated for probability distributions having certain desirable properties and the parameters $s,\rho$ etc., which appear in tilted probability distributions are nothing more than Lagrange multipliers.

This point of view is partly a matter of taste. Most of information theory can be developed without reference to the discrimination. This latter approach, however, fails to recognize the common mathematical structure that underlies many of the traditional results.

A summary of the various bounds is given in Table I. In order to strengthen the analogies, entropy is shown as the minimum of a mutual information although the minimum can be trivially evaluated. Similarly, $F(R)$ is expressed in terms of a maximum although the maximum can be trivially evaluated.

The channel reliability rate function $E(R,S)$ in the presence of an input constraint (expense schedule) is also shown although it has not been discussed. It is a straightforward generalization of $E(R)$.

## Acknowledgment

The author wishes to acknowledge the advice and criticism of Prof. Toby Berger of Cornell University which helped to shape this paper. Comments by G. D. Forney, J. Omura, R. M. Gray, and K. Marton were also helpful.

## References

[1] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inform. Theory*, vol. IT-4, pp. 2–22, Sept. 1954.

[2] R. M. Fano, *Transmission of Information.* Cambridge, Mass.: M.I.T. Press, 1961.

[3] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[4] 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.

[5] C. E. Shannon, R. G. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Inform. Contr.*, vol. 10, Feb. 1967.

[6] G. D. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 549–557, 1968.

[7] I. Csiszar and G. Longo, "On the error exponent for source coding and for testing simple statistical hypotheses," Hungarian Academy of Sciences, Budapest, Hungary, 1971.

[8] S. Kullback, *Information Theory and Statistics.* New York: Dover, 1968 and New York: Wiley, 1959.

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

[10] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.

[11] F. Jelinek, *Probabilistic Information Theory.* New York: McGraw-Hill, 1968.

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

[13] E. R. Berlekamp, "Block coding with noiseless feedback," Ph.D. thesis, Dep. Elec. Eng., M.I.T., Cambridge, Mass., Sept. 1964.

[14] R. E. Blahut, "An hypothesis testing approach to information theory," Ph.D. dissertation, Cornell Univ., Ithaca, N.Y., Aug. 1972.

[15] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 495–507, Apr. 1952.

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