A Parallel Systems Approach to Universal Receivers

Upamanyu Madhow and Michael B. Pursley, *Fellow, IEEE*

***Abstract*—The problem of communication over a channel with unknown characteristics is addressed. The true channel is from a known set of channels, but the transmitter and receiver do not know which of these channels is actually in effect. The goal of a universal receiver is to provide nearly optimal demodulation regardless of the channel that is actually in effect. A parallel receiver implementation is proposed for a universal scheme to cope with such uncertainty. The parallel system consists of a finite number of receivers with the property that, for each channel in the set, the performance of at least one of the receivers is within a specified performance range. Data verification, the process of selecting the receiver output sequences that are "good" for the channel that is in effect, is accomplished by an appropriate coding system. Sufficient conditions for the existence of such a universal receiver for a prescribed set of channels are established, procedures are outlined for the receiver design, and an example is given to illustrate the applicability of the theory. For M-ary signaling it is shown that, from an information-theoretic viewpoint, the data verification can be achieved at no extra cost by use of the intrinsic side information that is provided by an appropriate coding scheme that also provides error correction. For practical codes, there is a cost in code rate for data verification, but Reed–Solomon codes with bounded distance decoding work well in providing both data verification and error correction.**

***Index Terms*—Universal communication receivers, error-control coding, demodulation.**

I. Introduction

The GENERAL model for a communication system is shown in Fig. 1. In the classical theory, the complete statistical characterization of the physical channel is assumed to be known to the system designer. For a given signaling scheme, it is possible to design a receiver that provides demodulation of the channel output in an optimal way, such as minimizing the probability of error. In terms of the operation of the system, the classical theory assumes that the characteristics of the physical channel are known to the transmitter and receiver, and that the receiver knows, or can easily learn, all of the parameters of the transmitted signal that it needs in order to demodulate the signal that is sent over the physical channel.

In many practical applications, these assumptions are not valid. In fact, the statistical characterization of the physical channel often changes with time (although the changes may take place slowly), and the receiver may not know all of the parameters of the transmitted signal. As a result, the receiver is faced with the task of demodulating the output of a physical channel with unknown characteristics when the input is a signal with unknown parameters. From the receiver's point of view, it does not make much difference whether the uncertainty is in the transmitted signal or the physical channel. As a result, we choose to group the transmitter and physical channel together, and we refer to the cascade of these two as the *channel*, as shown in Fig. 1.

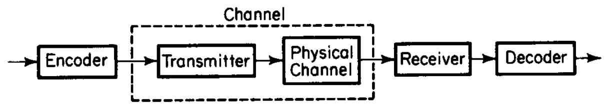

Fig. 1. Digital communication system.

An appropriate model for communication in the presence of uncertainties of the type previously mentioned is as follows. The channel over which the communication system must operate at a given time, referred to as the *channel in effect*, is from a known set of channels called the *channel class*. During the operation of the communication system, the encoder and receiver do not know which member of the channel class is the channel in effect. The encoder will therefore not be able to match the code to the channel, and it will not even be able to use the optimum rate for the channel in effect.

The appropriate model then is the *compound channel* for which neither the source nor the destination know the statistical characterization of the channel. For most applications, however, the compound channel of interest is not the classical discrete compound channel [7, p. 33]. Although the input to the channel is an encoded discrete-time data sequence from a fixed alphabet, the output is typically a sequence of real numbers or vectors, or even a continuous-time waveform. Furthermore, our interest lies primarily in the demodulation of the output of the channel, rather than in the encoding and decoding, as is the case for past research on compound channels. However, error-control coding does play an important role in our approach to dealing with uncertainty in the channel, as will be seen in what follows.

In this paper, the term *receiver* refers to the subsystem that performs the demodulation process; that is, it estimates the encoded data sequence based on the output of the channel. For our purposes in this paper, the decoding task is considered to be a separate operation to be performed on the receiver output, and this viewpoint is consistent with the

Manuscript received September 15, 1989; revised September 20, 1990. This work was supported by the Army Research Office (DAAL-03-87-K-0097). This work was presented in part at the International Symposium on Information Theory, San Diego, CA, January 14–19, 1990; and in part at the 1990 Bilkent International Conference on New Trends in Communication, Control and Signal Processing, Bilkent University, Ankara, Turkey, July 2–5, 1990.

The authors are with the Coordinated Science Laboratory, University of Illinois, 1101 W. Springfield Ave., Urbana, IL 61801.

IEEE Log Number 9041019.

0018-9448/91/0300-0291$01.00

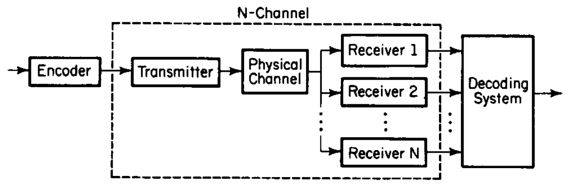

Fig. 2. Parallel receiver scheme.

block codes that we employ. The goal of a *universal receiver* is to provide nearly optimal demodulation regardless of the channel that is actually in effect. This goal is considerably more ambitious than attempting to design a receiver whose performance is relatively constant over the channel class. This latter approach is usually based on the existence of a receiver that optimizes the worst-case performance over the channel class. However, for most channel classes of interest, such a minimax receiver performs much worse on a given channel than the optimal receiver. The universal approach strives for optimal performance for each channel in the class. An adaptive receiver is an alternative to a universal receiver, provided that adaptive algorithms can be found that converge quickly enough to deal with the changes in the channel characteristics. The universal approach avoids the need for such algorithms.

The particular realization of the universal receiver introduced in this paper is based on a parallel receiver configuration, as shown in Fig. 2. The proposed system consists of a finite number of receivers with the property that, for each channel in the channel class, the performance of at least one of the receivers deviates from the optimal performance for that channel by no more than some prescribed amount. The determination of the receivers that are good for the channel that is in effect is accomplished by means of the intrinsic side information generated by an appropriate coding scheme. For a given transmitted sequence, the input to the decoding system of Fig. 2 is a set of N sequences, one from each of the parallel receivers. By virtue of the parallel design, at least one of the receivers is good for the channel in effect. The decoding system should be such that it can exploit this fact, even though the identity of the good receiver is not known, and the other receivers may perform very poorly for the channel in effect.

We refer to the single-input, N-output discrete channel formed by the transmitter, physical channel, and the bank of N parallel receivers as the *N-channel*. The receivers are required to be memoryless for our development, and so the N-channel is both discrete and memoryless.

One possible implementation of a decoding system for our parallel receiver configuration is as follows. For each transmitted codeword, the received words at the output of the N parallel receivers are decoded separately. The goal is that the received word for a good receiver will, with high probability, be correctly decoded, and the received word for a mismatched receiver will, with high probability, lead to decoding failure. Thus, it is necessary to employ a code for which the decoding is much more likely to fail to decode than to decode into an incorrect codeword, even if the symbol error probability out of the receiver is very high. Reed–Solomon (RS) codes are therefore a natural choice, due to their maximal distance properties and their low probability of decoding error when bounded-distance decoding is used.

The research problems that arise here can be summarized as follows: a) characterization of the channel classes for which it is possible to design a finite bank of receivers as previously described, and development of procedures to carry out such a design; b) design and performance analysis for suitable coding schemes; and c) design tradeoffs between the number of receivers in the bank and the specified degradation from the optimal performance. The last problem is best addressed by working out an example with a specific channel class. This has been done for Rician fading channels, and the results will be given in a later paper. A general formulation which attempts to address the first problem is given in the next section, together with a general example that illustrates how the theory applies to the problem of M-ary signaling under unknown or time-varying conditions. In Sections III and IV, we consider the performance of a coding system for the universal receiver. An information-theoretic characterization is given for the resulting N-channel in Section III. This is followed by results for practical codes in Section IV, in which bounds are developed for the probability of decoding failure and the probability of decoding error for the decoding system. It is demonstrated that Reed–Solomon codes with bounded distance decoding perform well for our application. Finally, Section V contains a discussion of the results of this paper as well as of some possibilities for future research.

II. General Formulation

In this section we consider what may be called the detection aspect of designing the universal receiver. We are restricted to a class of available receivers in our design, and the goal is to select a finite number of these receivers with the property that, for each channel in the class, at least one of the receivers performs almost as well as the optimal receiver (within the class of available receivers) for that channel. Any such set of receivers is called a *universal set of receivers*, since it is proposed to place these receivers in parallel to implement the universal receiver.

Let C be the class of channels and R the class of available receivers. The performance functional $f\colon C \times R \to [0,\infty)$ gives the value $f(x,y)$ of the performance measure when receiver y is used and channel x is in effect. Smaller values of $f(x,y)$ correspond to better performance, a convention that is adopted to conform to such common performance measures such as the probability of error or the mean-squared error. Define $g\colon C \to [0,\infty)$ by

$$g(x) = \min_{y \in R} f(x,y). \tag{1}$$

The number $g(x)$ is thus the best performance possible using receivers from R when channel x is in effect. It is implicitly assumed in (1) that, for each $x_0 \in C$, there exists an optimal receiver $y_0 \in R$ such that $g(x_0) = f(x_0, y_0)$.

The allowable deviation from optimality is specified by means of a continuous function $h\colon [0,\infty) \to [0,\infty)$, which satisfies $h(s) > s$ for all $s \in [0,\infty)$. Any such function will henceforth be called a *degradation function*. Given a channel $x \in C$, it is required to attain a performance of at most

$h[g(x)]$ by means of the universal set of receivers. As an example, consider the degradation function h defined by

$$h(s) = \max[s+\alpha, \beta s], \qquad s \in [0,\infty),$$

where $\alpha > 0$ and $\beta \geq 1$. For this example, the allowed degradation $h[g(x)]$ is such that the multiplicative degradation $\beta g(x)$ dominates for large values of $g(x)$, so that, for channels on which even the optimal receiver does not give good performance, it is required that the performance attained be within a fixed factor β of the optimal performance $g(x)$. For good channels with a small value of $g(x)$, it is more reasonable to permit an additive degradation. In fact, if C is such that g can take on the value zero, it is necessary that $\alpha > 0$ to ensure that $h[g(x)] > g(x)$ for all $x \in C$. However, the function h previously given is a valid degradation function even if the multiplicative degradation is dispensed with (i.e., β is taken to be 1). In order to relate the previous discussion to a communications application, suppose that the performance of interest is the probability of error, and that error probabilities less than 10^{-6} are not required at the demodulator output. A reasonable choice for the degradation function for such a situation may be $h(s) = \max[s+10^{-6}, 3s]$, $s \in [0,1]$. If, for instance, the optimal error probability is 10^{-7}, the maximum allowable error probability is 1.1×10^{-6}, but if the optimal error probability is 10^{-5}, the maximum allowable error probability is 3×10^{-5}. This type of specification is consistent with the goals in practical communication systems.

We first give some sufficient conditions on the channel class and on the performance functional for the existence of a universal set of receivers. We include examples that show that it may not be possible to obtain a universal set of receivers if any one of the given conditions is violated. Next, we develop procedures to design such a set of receivers. Finally, it is shown that our results apply to M-ary signaling over a channel whose statistical characteristics are known to lie in an appropriately chosen class.

A. Existence Results

It is seen in this section that the existence of a universal set of receivers hinges on finding a topology on the channel class relative to which both compactness requirements on the channel class and continuity requirements on the performance functional are satisfied.

Theorem 1: Let h be an arbitrary degradation function, and let C be a compact Hausdorff space (e.g., a compact metric space). Let the family of functions $\{f(x,y),\ y \in R\}$ be equicontinuous on C. Then there exists an integer N and receivers $y_1, \cdots, y_N$ in R such that, given any x in C,

$$\min_{1 \leq i \leq N} f(x, y_i) \leq h[g(x)]. \tag{2}$$

Proof: It is first shown that g is continuous on C. Let $\epsilon > 0$ be given. For any $x_0 \in C$, there is a neighborhood U such that for all $x \in U$,

$$|f(x,y) - f(x_0,y)| < \frac{\epsilon}{2}, \qquad \text{for all } y \in R, \tag{3}$$

using the equicontinuity of $\{f(x,y),\ y \in R\}$ on C. We also note that

$$\begin{aligned} g(x) - g(x_0) &= \min_{y \in R} f(x,y) - \min_{y \in R} f(x_0,y) \\ &= \min_{y \in R} f(x,y) + \max_{y \in R}[-f(x_0,y)] \\ &= \max_{y \in R}\left\{\left[\min_{y \in R} f(x,y)\right] - f(x_0,y)\right\} \\ &\leq \sup_{y \in R}\{f(x,y) - f(x_0,y)\} \\ &\leq \sup_{y \in R}|f(x,y) - f(x_0,y)|. \end{aligned}$$

Interchanging the roles of x_0 and x,

$$|g(x) - g(x_0)| \leq \sup_{y \in R}|f(x,y) - f(x_0,y)|. \tag{4}$$

For $x \in U$, therefore, we have, using (3) and (4), that

$$|g(x) - g(x_0)| \leq \epsilon/2 < \epsilon,$$

proving the continuity of g on C.

Define, for each $y \in R$, $B_y = \{x \in C\colon f(x,y) < h[g(x)]\}$. B_y is open for each $y \in R$, by the continuity of f, g, and h. Also $\bigcup_{y \in R} B_y = C$ since for each $x_0 \in C$, there exists $y_0 \in R$ such that $f(x_0, y_0) = g(x_0) < h[g(x_0)]$, which implies that $x_0 \in B_{y_0}$. Thus, $\{B_y,\ y \in R\}$ is an open cover for the compact space C. Therefore, there exists a finite subcover, say $B_{y_1}, \cdots, B_{y_N}$. Clearly, $y_1, \cdots, y_N$ satisfy (2), completing the proof. □

The compactness and equicontinuity conditions in Theorem 1 are both critical in guaranteeing the existence of a universal set of receivers for a given channel class, receiver class, and degradation function. This is demonstrated in the following by two examples for which it is *not* possible to find such a set of receivers. The compactness condition is violated in the first example, and the equicontinuity condition is violated in the second.

Example 1) Violation of the Compactness Condition: Consider binary equiprobable on–off signaling over a discrete time memoryless channel with additive Gaussian noise. The performance of interest is the bit error probability at the receiver output. An unknown signal level m is sent when a 1 is transmitted, and nothing is sent when a 0 is transmitted. In each case, the channel corrupts the transmitted signal by adding a noise sample drawn from a standard normal distribution. The noise samples are independent for different bits. The receiver must choose between the hypotheses

$$H_1\colon Y = m + W$$

and

$$H_0\colon Y = W,$$

where $m \geq 0$ is unknown and W is a standard normal random variable. The channel is identified with the signal level m, so that the channel class C is the nonnegative real line, the range of possible values of m. Note that this space is not compact in the standard topology of the real line.

If m were known, the minimum probability of error receiver would be given by

$$Y \underset{H_0}{\overset{H_1}{\gtrless}} m/2.$$

Since $m \in [0,\infty)$, the receiver class R is identified with the set of possible thresholds t. Thus, $R = [0,\infty)$, and we are said to use the receiver $t \in R$ when we use the following decision rule:

$$Y \underset{H_0}{\overset{H_1}{\gtrless}} t.$$

If the receiver t is used when the channel m is in effect, the performance functional of interest is the error probability $P(m,t)$, which is given by

$$P(m,t) = \tfrac{1}{2}Q(m-t) + \tfrac{1}{2}Q(t), \qquad (5)$$

where $Q(x)$ is the complementary standard normal distribution function defined by

$$Q(x) = (2\pi)^{-1/2}\int_x^{\infty} \exp(-x^2/2)\,dx. \qquad (6)$$

The minimum error probability $P^*(m)$ for the channel m is given by $P^*(m) = P(m, m/2) = Q(m/2)$.

It is easy to see that the equicontinuity requirements of Theorem 1 are satisfied here. For channels $m_1, m_2 \in C$, we have, using (5) and (6), that

$$\begin{aligned}|P(m_1,t) - P(m_2,t)| &= \tfrac{1}{2}(2\pi)^{-1/2}\left|\int_{m_1-t}^{m_2-t} \exp(-x^2/2)\,dx\right| \\ &\le \tfrac{1}{2}(2\pi)^{-1/2}|m_2 - m_1|.\end{aligned}$$

This proves that the family of functions $\{P(m,t),\ t \in R\}$ is equicontinuous on C. However, since C is not compact, the existence of a universal set of receivers for an arbitrary degradation function is not guaranteed by Theorem 1. Indeed, for a degradation function of the form $h(s) = \beta s$, for $\beta > 1$, it is not possible to have a set of receivers as described in Theorem 1. The proof is by contradiction. Assume that there are receivers $t_1, \cdots, t_N$ such that, for any $m \ge 0$,

$$\min_{1 \le i \le N} P(m, t_i) \le \beta P^*(m). \qquad (7)$$

It is known that $Q(x) \to 0$ as $x \to \infty$. Letting $m \to \infty$ in (7), we have $0 < \min_{1 \le i \le N} \tfrac{1}{2}Q(t_i) \le 0$, which is a contradiction. This implies that for any finite set of receivers $t_1, \cdots, t_N$, (7) is violated for sufficiently large m. It can be shown quite easily, however, that a universal set of receivers can be found if we allow an additive degradation from the optimal performance. But the point is, if the compactness requirement is removed from Theorem 1, other requirements must be imposed.

Example 2) Violation of the Equicontinuity Condition: We again consider binary equiprobable signaling over a discrete-time memoryless channel. The channel output alphabet is the interval $I = [0,1]$, and the output Y is a random variable with density p_i if symbol i $(i = 0,1)$ is sent, which corresponds to the hypothesis H_i in the related hypothesis testing problem. The class of available receivers of the following form: Decide on H_1 if $Y \in A$, and decide on H_0 otherwise, where A is any measurable subset of I. Henceforth, the receiver is identified with the corresponding subset A, so that the receiver class R is a subset of Σ, where Σ is the σ-algebra of Lebesgue-measurable subsets of I. The channel is determined by the pair of densities (p_1, p_0), which is denoted by $\boldsymbol{p}$. The channel class C is therefore a subset of the Banach space $B = L_1(I) \times L_1(I)$. The performance functional of interest is the probability of error $P(\boldsymbol{p}, A)$ that results if the receiver A is used when the channel $\boldsymbol{p}$ is in effect. It is given by

$$P(\boldsymbol{p}, A) = \tfrac{1}{2}\int_{A^c} p_1 + \tfrac{1}{2}\int_A p_0, \qquad (8)$$

where A^c denotes the complement of A in I, and where the integrals are with respect to the Lebesgue measure μ. Given $\boldsymbol{p}$, the optimal, or minimum probability of error, receiver is given by $A = \{x \in I:\ p_1(x) > p_0(x)\}$, and we denote the resulting optimal error probability by $P^*(\boldsymbol{p})$.

Consider the following sequence $\boldsymbol{p}_n = (p_1^{(n)}, p_0^{(n)})$ in B.

$$p_1^{(n)}(x) = \begin{cases} 2, & x \in [2k/2^n, (2k+1)/2^n), \quad k = 0,1,\cdots,(2^{n-1}-1), \\ 0, & \text{otherwise}, \end{cases} \qquad (9a)$$

and

$$p_0^{(n)}(x) = 2 - p_1^{(n)}(x), \qquad x \in I. \qquad (9b)$$

The optimal receiver A_n corresponding to $\boldsymbol{p}_n$ is given by

$$A_n = \bigcup_{k=0}^{(2^{n-1}-1)} [2k/2^n, (2k+1)/2^n),$$

and the corresponding probability of error is $P^*(\boldsymbol{p}_n) = P(\boldsymbol{p}_n, A_n) = 0$.

Consider the weak topology on B. The pair of densities $\boldsymbol{p} = (p_1, p_0)$ is said to converge weakly to $\boldsymbol{q} = (q_1, q_0)$ if, for any subset A of I, $\int_A p_1 \to \int_A q_1$, and $\int_A p_0 \to \int_A q_0$. The channel class C is defined to be the weak closure of $\{\boldsymbol{p}_n\}$ in B. The sequence $\{\boldsymbol{p}_n\}$ converges weakly to $\boldsymbol{u} = (u_1, u_0)$, where u_1 and u_0 are uniform densities on I, given by $u_1(x) = u_0(x) = 1$, $x \in I$. Thus, $C = \{\boldsymbol{p}_n\} \cup \{\boldsymbol{u}\}$. Since the sequences $(p_i^{(n)})$ $(i = 0,1)$ are uniformly integrable and are bounded in $L_1(I)$, we have, using Theorems IV.8.9 and V.6.1 in [3], that C is weakly compact in B. The receiver class R is taken to be the set of optimal receivers for C. It is shown in Appendix A that the equicontinuity requirement is violated under the same topology for which the compactness condition that is required to apply Theorem 1 is satisfied.

Consider, now, a degradation function of the form $h(s) = s + \epsilon$, $s \ge 0$, where $\epsilon > 0$. It is shown in Appendix A that, if $\epsilon < 1/8$, it is *not* possible to find a finite set of receivers $D_1, \cdots, D_N$ such that, for any $\boldsymbol{p} \in C$, $\min_{1 \le i \le N} P(\boldsymbol{p}, D_i) \le P^*(\boldsymbol{p}) + \epsilon$. Thus, it is not possible to find a universal set of receivers for this example, either.

We note as an aside that it is somewhat artificial to single out one of the two conditions as having been violated. It is always possible to find a topology in which at least one of the two conditions is satisfied. The conditions of Theorem 1 are not satisfied when there is no topology for which both conditions are satisfied. That this is indeed the case for the two examples above follows from the fact that the conclusion of Theorem 1 is shown not to hold for each example. For simplicity, we choose to consider topologies that seem to be natural for the examples, and show that one of the two conditions is violated.

Theorem 1 imposes no requirement for a topology on the receiver class R. Often, however, such a topology is available. The following result gives a sufficient condition for the existence of a finite set of receivers satisfying (2) for a given degradation function h when R is compact and Hausdorff.

Theorem 2: Let C and R be compact Hausdorff spaces. Let $f: C \times R \to [0,\infty)$ be continuous in the product topology. Then there exist an integer N and receivers $y_1,\cdots,y_N$ in R such that (2) holds for any given x in C.

Proof of Theorem 2: The statement can be proved as a corollary of Theorem 1 by showing that the family $\{f(x,y), y \in R\}$ is equicontinuous on C under the given hypotheses. However, we give an easier proof based on establishing directly the continuity of g on C. It is assumed, for simplicity, that C and R are metric spaces with metrics d_C and d_R, respectively, so that it suffices to work with sequences. Virtually the same proof works for arbitrary Hausdorff spaces, except that it is necessary to work with nets instead of sequences.

Since g is the minimum of a family of continuous functions, it is automatically upper semicontinuous. It suffices, therefore, to show that g is lower semicontinuous on C. If g is not lower semicontinuous, then there exists a real number b such that the set $B = \{x \in C: g(x) > b\}$ is not open. Thus, there exists $x_0 \in B$ such that, for each positive integer n, $A_n = \{x \in C: d_C(x,x_0) < 1/n\}$ has a nonempty intersection with B^c. Let $x_n \in A_n \cap B^c$. Then

$$b \geq g(x_n) = \min_{y \in R} f(x_n, y) = f(x_n, y_n),$$

where $y_n \in R$ is any minimizing point. Since R is compact, (y_n) has a convergent subsequence (y_{n_k}), which converges to some $y_0 \in R$. We also have that $x_n \to x_0$, so that $(x_{n_k}, y_{n_k}) \to (x_0, y_0)$ in the product topology. By the continuity of f, therefore, we have

$$b \geq f(x_{n_k}, y_{n_k}) \to f(x_0, y_0) \geq g(x_0) > b.$$

This provides the required contradiction, proving the continuity of g on C. The rest of the proof is identical to the second half of the proof of Theorem 1. □

Theorems 1 and 2 are existence results. So far, no procedure has been given for actually determining the number N of receivers or the points $y_1,\cdots,y_N$ that satisfy (2) for a given degradation function. We now consider some special situations for which such procedures can be developed.

B. Design Procedures

Two situations are considered here: a) the performance functional f satisfies the equicontinuity requirement of Theorem 1, and the channel class C is a compact subset of a finite-dimensional normed linear space (NLS), and b) the performance functional f is jointly continuous on $C \times R$, and the receiver class R is a compact subset of a finite-dimensional NLS. In each case, the appropriate norm is denoted by $\|\cdot\|$. Note that the conditions in a) are stronger than the hypotheses of Theorem 1, and those in b) are stronger than the hypotheses of Theorem 2. This makes it possible to give constructive procedures for finding a finite number of receivers satisfying (2), as opposed to the results of the previous section, which guarantee only the existence of such a finite set of receivers. For a), the channel class C is partitioned into a finite number of sets with the property that for each set there is one receiver that provides the desired performance for all the channels in that set. The finite set of receivers obtained in this manner satisfies (2). For b), the receiver class is partitioned. For a given subset S of R, let C_S be the set of channels in C for which some receiver in S is optimal, that is,

$$C_S = \left\{x \in C: g(x) = \min_{y \in S} f(x,y)\right\}.$$

The partition of R is such that for each set S in the partition, there is a receiver y_S that provides the desired performance for all channels in C_S. Since each channel has an associated optimal receiver, the finite set of receivers y_S satisfies (2). The sets of the partition need not be disjoint for either a) or b).

Consider the situation described in a). For $x \in C$, let

$$R_x = \{y \in R: f(x,y) = g(x)\}.$$

R_x is the set of receivers that are optimal for the channel x. Usually, the optimal receiver is unique and R_x is a singleton set, but this might not be true, for instance, in a singular detection problem. Define, for $x \in C$,

$$m(x) = \sup\{r > 0: \text{for all } u \in \overline{B(x,r)},\; f(u,y) \leq h[g(u)], \text{for all } y \in R_x\},$$

where $B(x,r)$ is the ball of radius r centered at x. Thus, if a channel u lies in a ball of radius $r < m(x)$ around the channel x, then *any receiver that is optimal for the channel x* performs within the specified degradation for the channel u. Let $\delta_C = \inf_{x \in C} m(x)$. The channel class c is compact in a finite-dimensional NLS and is therefore a bounded set. If δ_C is positive, we can partition the channel class C and thereby obtain receivers $y_1,\cdots,y_N$ that satisfy (2). The resulting procedure is discussed below, followed by a proposition that states that δ_C is in fact positive under the given conditions.

To illustrate the procedure, consider the following one-dimensional example. Let C be the finite closed interval $[a,b]$ on the real line with the usual norm. Assume that $m(x)$ can be computed for any $x \in C$. Then $y_1,\cdots,y_N$ satisfying (2) can be chosen by means of the following algorithm.

1) Set $x_1 = a$, $i = 1$.
2) Choose $y_i \in R_{x_i}$ (that is, choose a receiver that is optimal for x_i).
3) If $x_i + m(x_i) > b$, stop. If not, set $x_{i+1} = x_i + m(x_i)$, increment i by one, and go to Step 2). (The chosen receiver y_i provides the desired performance for all channels in the interval $[x_i, x_{i+1})$.)

This procedure terminates in at most $\delta_C^{-1}(b-a)$ steps, because $m(x_i) \geq \delta_C$ for all i. Thus, the number of receivers $N \leq 1 + \delta_C^{-1}(b-a)$. It is easy to see how to modify the above algorithm for channel classes C of higher (but finite) dimensions. The termination of the algorithm depends crucially on the assumption that δ_C is strictly positive. This assumption is satisfied under the hypotheses of a), and this is stated formally in the following. The proof is given in Appendix A.

Proposition 1: If f satisfies the equicontinuity property and C is a compact subset of a NLS, then

$$\delta_C = \inf_{x \in C} m(x) > 0.$$

We now consider b), in which the hypotheses of Theorem 2 hold, and the receiver class R is assumed to be a compact subset of a NLS. Define, for $y \in R$,

$$C_y = \{x \in C: f(x,y) = g(x)\}.$$

C_y is thus the set of channels for which receiver y is optimal. Let

$$n(y) = \sup\{r > 0: \text{for all } \nu \in \overline{B(y,r)},\quad f(x,y) \le h[g(x)], \text{ for all } x \in C_\nu\},$$

where $B(y,r)$ is the ball of radius r centered at y. Thus, if a receiver ν lies in a ball of radius $r < n(y)$ around the receiver y, then the receiver y performs within the required degradation levels for *all channels for which receiver ν is optimal*. The function n on R previously defined is thus analogous to the function m defined on C in case a). The class of receivers R is finite-dimensional and compact, and hence it is a bounded set. Let $\delta_R = \inf_{y \in R} n(y)$. If δ_R is positive, the receiver class R can be partitioned to obtain $y_1, \cdots, y_N$ satisfying (2). As before, the discussion of the procedure is given first, followed by a proposition that states that δ_R is positive under the given conditions.

As in a), consider a one-dimensional example. Let R be the finite closed interval $[c,d]$ on the real line with the usual norm. Assume that $n(y)$ can be computed for any $y \in R$. Then $y_1, \cdots, y_N$ satisfying (2) can be chosen by means of the following algorithm.

1) Set $y_1 = c$, $i = 1$.
2) If $y_i + n(y_i) > d$, stop. If not, set $y_{i+1} = y_i + n(y_i)$, increment i by one, and repeat Step 2). (The receiver y_i provides the desired performance for all channels for which any receiver in the interval $[y_i, y_{i+1})$ is optimal.)

This procedure terminates in at most $\delta_R^{-1}(d-c)$ steps, since $n(y_i) \ge \delta_R$ for all i. Thus, the number of receivers $N \le 1 + \delta_R^{-1}(d-c)$. As before, the algorithm can be modified to handle higher dimensions. The termination of the algorithm depends on the fact that δ_R is positive under the given conditions. As stated in the next proposition, this result does hold under the hypotheses of b). The proof is quite similar to that of Proposition 1, and is therefore omitted.

Proposition 2: If f is jointly continuous on $C \times R$, C is compact Hausdorff, and R is a compact subset of a NLS, then $\delta_R = \inf_{y \in R} n(y) > 0$.

The design procedures previously outlined in the two situations of interest involve computation of the functions m and n, respectively, which may not always be an easy task. For instance, the computation of the function n in case b) involves a fairly complicated optimization. For any $y \in R$, define

$$s_y(r) = \min_{\nu \in \overline{B(y,r)}} \min_{x \in C_\nu} \{h[g(x)] - f(x,y)\}, \qquad r > 0. \tag{10}$$

Then the value of $n(y)$ is given by the following constrained optimization problem.

$$n(y) = \sup\{r: s_y(r) \ge 0\}.$$

The complexity of this problem is determined by that of the unconstrained optimization (10) involved in computing the function $s_y(r)$. We have imposed no restrictions on the channel class C thus far, apart from the compactness hypothesis. Since the inner minimization in (10) is over a subset of C, however, the properties of C and the variations of the functions f and g over C are major factors in determining the overall complexity (and hence the feasibility) of the design procedure suggested here. A detailed examination of these issues in various specific instances is required to evaluate the utility of the approach proposed here. This paper, however, is concerned with the general principles of the universal approach, and does not consider applications that are specified in sufficient detail to explore these issues.

C. Parallel Receivers for M-ary Signaling

The purpose of this section is to provide an example of a channel class, receiver class, and performance functional that may be encountered in a communications application. Consider the problem of M-ary signaling over a channel whose statistical characteristics are known to lie in a given class. It is shown here that, for certain channel classes, it is possible to design a finite number of receivers such that, given any channel in the class, at least one of the receivers performs almost as well as the minimum probability of error receiver for that channel.

The input alphabet is $A_X = \{1, \cdots, M\}$. The output alphabet A_Y can be continuous and is assumed to have a measure μ on it. The output Y has density $p_i(y)$ with respect to μ if symbol i is sent. Let α_i be the prior probability that the input X is i. If the densities p_i and the priors α_i are known, then the minimum probability of error receiver can be designed by solving the corresponding multiple hypothesis testing problem, with the ith hypothesis given by H_i: $Y \sim p_i$, $i = 1, \cdots, M$, where the prior probability of H_i is α_i. The decision space in this situation is the same as the input alphabet A_X.

Let λ be the counting measure on A_X. Then the input–output pair (X,Y) has joint density $p(x,y)$ with respect to the product measure $\lambda \times \mu$ on $A_X \times A_Y$, given by $p(i,y) = \alpha_i p_i(y)$ for each i in A_X. The minimum probability of error receiver corresponding to the joint density p is given by a function r^p: $A_Y \to A_X$ defined by

$$r^p(y) = \arg\max_{x \in A_X} p(x,y),$$

with ties being resolved arbitrarily but deterministically. Thus, $r^p(y) = j$ corresponds to deciding that the hypothesis H_j is correct.

Now, suppose that the joint density p is not known exactly, but it is known to lie in a parametric class $P = \{p^\theta: \theta \in \Theta\}$, where Θ is a subset of R^n. It is desired to specify conditions under which it is possible to obtain a finite set of receivers with the property that, given any $p \in P$, at least one receiver in the set performs within a specified level of degradation from the minimum probability of error for joint density p.

Any deterministic receiver for M-ary signaling is characterized by a function r: $A_Y \to A_X$, and $r(y)$ denotes the decision when the observed channel output is y. If this receiver is used and the joint density of (X,Y) is p, the performance functional of interest is the probability of error

$$Q(p,r) = \sum_{i=1}^{M} \int_{\{r \ne i\}} p(i,y)\mu(dy) = 1 - \sum_{i=1}^{M} \int_{\{r = i\}} p(i,y)\mu(dy). \tag{11}$$

The probability of error is minimized by using the receiver r^p, and the minimum probability of error is given by $Q^*(p)=Q(p,r^p)$.

Consider the following conditions on the class P: (C1) Θ is compact in R^k, and (C2) if $(\theta_n)\in\Theta$ converges to $\theta\in\Theta$, then $p^{\theta_n}\to p^{\theta}$ in $L_1(\lambda\times\mu)$. By Scheffe's theorem [1, p. 223], almost everywhere (a.e.) convergence of densities implies L_1 convergence, so that (C2) is implied by the following conditions: (C3) if $(\theta_n)\in\Theta$ converges to $\theta\in\Theta$, then $p^{\theta_n}\to p^{\theta}$ a.e. $(\lambda\times\mu)$. Condition (C3) is often easier to check than the weaker condition (C2). For instance, if the various conditional densities are Gaussian, and if θ parametrizes the prior probabilities of the different hypotheses, the means of the conditional densities, and the covariance of the conditional densities, then (C3) is satisfied.

The set R of admissible receivers is taken to be the set of minimum probability of error receivers corresponding to the channels in the uncertainty class; that is, $R=\{r^p\colon\ p\in P\}$. The following theorem gives conditions under which a universal set of receivers can be obtained.

Theorem 3: Suppose (C1), and either (C2) or (C3), are true. Let h be a degradation function. Then there is an integer N, and receivers $r_1,\cdots,r_N$ in R such that, for any $p\in P$,

$$\min_{1\le j\le N} Q(p,r_j)\le h[Q^*(p)].$$

Proof: P is a compact subset of $L_1(\lambda\times\mu)$, since it inherits the compactness of Θ via (C2) or (C3). We will now show the equicontinuity of the family of functions $\{Q(p,r),\ r\in R\}$ on P under the L_1 topology. Theorem 1 of the general formulation can then be applied to yield the desired result. For $p,q\in P$, and for any $r\in R$, we have, using (11), that

$$\begin{aligned}
&|Q(p,r)-Q(q,r)|\\
&\quad\le\left|\sum_{i=1}^{M}\int_{\{r=i\}}(p(i,y)-q(i,y))\mu(dy)\right|\\
&\quad\le\sum_{i=1}^{M}\int_{A_Y}|p(i,y)-q(i,y)|\mu(dy)\\
&\quad=\|p-q\|_1\to 0\ \text{ uniformly in } r \text{ as } p\to q \text{ in } L_1(\lambda\times\mu),
\end{aligned}$$

proving that $\{Q(p,r),\ r\in R\}$ is an equicontinuous family on P. Theorem 1 in Section II-A now applies, completing the proof. Since P is a parametric class, the algorithm suggested for case a) in Section II-B for partitioning the channel class can be used to obtain the finite set of receivers guaranteed by the theorem.

III. An Information-Theoretic Perspective for Degraded Subchannels

One aspect of the performance of the universal receiver concerns the capacity and reliability function of the discrete memoryless channel (DMC) that is formed by cascading the transmitter, physical channel, and receiver. As the physical channel changes, so does the DMC. The capacity specifies the maximum rate at which reliable communication can take place over the channel in effect when a particular receiver is employed, and the reliability function specifies the largest attainable error exponent. By use of the proper code for the DMC in effect, it is possible to transmit at any rate less than the capacity with an error probability that goes to zero exponentially as the code block length increases, and the exponential rate of convergence can be as close as desired to the reliability function for the DMC. There are several approaches one might take in the examination of this aspect of the universal receiver performance. The one followed here is to consider an idealized genie-based receiver that provides universal demodulation, and to compare the performance of the DMC corresponding to it with that of the DMC corresponding to the universal receiver.

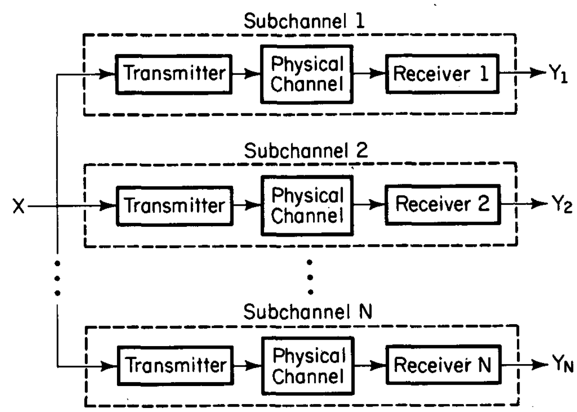

Fig. 3. Discrete N-channel resulting from parallel configuration.

Assume that we have obtained a set of N receivers with the property that, for each channel in the class, at least one of the receivers performs almost as well as the optimal receiver for that channel. It is known from the results of the previous section that, for many channel classes of interest, it is possible to find a finite set of receivers with this property. The two receivers to be compared are the universal receiver, which utilizes all N component receivers at all times, and an idealized (and unrealizable) receiver in which a genie continually chooses the best of the N component receivers for the channel in effect. Because the best of the N receivers is nearly as good as the optimal receiver for any channel in the class, the genie-based receiver can be designed to be as close to optimal as desired over the entire channel class.

A block diagram for the DMC with the universal receiver is shown in Fig. 3. As illustrated, this DMC actually consists of N DMC's in parallel, one for each of the N receivers that make up the universal receiver. This gives a single-input, N-output DMC that we refer to as the N-channel. Each of the N component channels is referred to as a subchannel. The subchannel that corresponds to the best of the N receivers for the channel in effect is referred to as the *best subchannel*. For the universal receiver, the input to the decoder consists of the N outputs of the discrete N-channel. For the genie-based receiver, only the best subchannel is used, and the output of this subchannel is the input to the decoder.

For either the universal receiver or the genie-based receiver, there is the question of what code to use and how to decode it. It can be seen from classical compound channel theory [7] that to attain the capacity of a DMC, it suffices for the encoder to know the capacity of the DMC and the input distribution that achieves the capacity of the DMC. The minimum that the encoder must know for the genie-based receiver, therefore, is the capacity of the best subchannel and the input distribution that achieves capacity for the best subchannel. An important result of this section is that this is also the minimum information required by the encoder if the universal receiver is used. The decoder does not need any information about the channel in either situation.

It is easy to see how the capacity of the best subchannel can be achieved for the universal receiver. The encoder codes for the best subchannel and appends to the codeword some test symbols that enable the decoder to choose the best subchannel with a high probability of being correct. The decoder then focuses on the best subchannel alone. The number of test symbols required is asymptotically negligible for large code block lengths, and so this procedure does not affect the code rate. We also prove a more general result on error exponents for code rates below the capacity of the best subchannel. This result cannot be obtained by a simple test symbol approach previously described, since the use of test symbols affects the error exponent. Instead, the universal coding results of [2] are used. These results guarantee the existence of good codes that achieve the random coding error exponent [2] for any DMC when used in conjunction with a universal maximum mutual information decoding procedure. The encoder is assumed to have some additional information, but as before, the required information is the same for the universal receiver and the genie-based receiver.

Thus, the requirements for achieving good performance (in terms of either capacity or error exponent) on the channel with the universal receiver are no more stringent than the requirements on the channel with the genie-based receiver. The difference is as follows. For the genie-based receiver, the decoder's only input is the output of the best subchannel. For the universal receiver, the decoder does not know the identity of the best receiver (and hence of the best subchannel), and the other receivers may perform very poorly for the channel in effect. Our result states that, regardless of the possible confusion caused by the presence of the other subchannels, the capacity and the reliability function of the discrete channel with the universal receiver are the same as for the discrete channel with the genie-based receiver. This result defines another way in which the parallel configuration provides universal demodulation.

At this point, it is necessary to introduce some additional terminology and notation, largely taken from [2]. A DMC is characterized by its transition probability matrix W, and the transition probability matrix is used to refer to the corresponding DMC. Given an input X to the DMC W, the conditional distribution of the output Y is given by

$$P(Y=y|X=x)=W(y|x),$$

which is the (x,y)th entry of the stochastic matrix W. The capacity of a DMC W is denoted by $C(W)$. If the encoder is constrained to use input distribution P, the random coding exponent function $E_r(R,P,W)$ is a universally (with respect to W) attainable error exponent at rate R [2, p. 172]. Optimizing this error exponent over P for a given W yields [2] the random coding exponent function $E_r(R,W)=\max_P E_r(R,P,W)$. Thus, for a given W and R, if the encoder knows the input distribution P that achieves the maximum value of $E_r(R,P,W)$, then $E_r(R,W)$ is an attainable error exponent at rate R for the DMC W. Let $E(R,W)$ denote the reliability function of W. It is known that $E(R,W)\geq E_r(R,W)$, with equality if and only if $R\geq R_{cr}(W)$, where $R_{cr}(W)$ is the critical rate for W.

Define the following partial order $<_c$ on DMCs [6]: for DMCs V and W, we say $V<_c W$ if there is a stochastic matrix Q such that $V=WQ$. Thus $V<_c W$ if and only if W is a *degraded* version [2, p. 115] of V. It follows that $V<_c W$ implies $C(V)\leq C(W)$.

Consider the N-channel illustrated in Fig. 3, and let $\boldsymbol{W}$ denote the $M\times M^N$ transition probability matrix for the N-channel. The transition probability $p(Y_1=y_1,\cdots,Y_N=y_N|X=x)$ is the entry of $\boldsymbol{W}$ in the row corresponding to input $x\in\{1,\cdots,M\}$ and the column corresponding to output vector $(y_1,\cdots,y_N)\in\{1,\cdots,M\}^N$. The columns are indexed by $\nu=f(y_1,\cdots,y_N)$, where f is a one-to-one mapping of $\{1,\cdots,M\}^N$ onto $\{1,\cdots,M^N\}$. Let W_i, $i=1,\cdots,N$, denote the corresponding $M\times M$ transition probability matrix for the ith subchannel; that is, the (x,y)th entry of W_i is the transition probability $p_i(y|x)=P(Y_i=y|X=x)$. It is important to distinguish between $\boldsymbol{W}$, the *joint* transition probability matrix for the N-channel, which has entries of the form $p(Y_1=y_1,\cdots,Y_N=y_N|X=x)$ for $(y_1,\cdots,y_N)\in\{1,\cdots,M\}^N$, and W_i, the *marginal* transition probability matrix for the ith subchannel, which has entries of the form $p_i(y|x)$ for $y\in\{1,\cdots,M\}$.

Because only degraded subchannels are considered in this section, there is a best subchannel (not necessarily unique) in the sense that all of the subchannels can be considered as degraded versions of one of the subchannels. To be more precise, for a given transition probability matrix W^* for a best subchannel, the transition probability matrix for each of the other subchannels is worse than W^* with respect to the partial order $<_c$. The identity of the best subchannel is not known, nor are the transition probabilities for the other subchannels. The class of joint transition probability matrices $\boldsymbol{W}$ for which these conditions hold and for which W^* is a best subchannel is a compound N-channel. This compound N-channel, denoted by Λ, is defined formally as follows. For any joint transition probability $\boldsymbol{W}$, consider the corresponding marginal transition probabilities W_i, $i=1,\cdots,N$. The N-channel $\boldsymbol{W}$ is in Λ if

$$W_j=W^*, \qquad \text{for some } j, \tag{12}$$

and

$$W_i=W^*Q_i, \qquad i=1,\cdots,N, \tag{13}$$

for some unknown stochastic matrices Q_i.

For our application, the lack of knowledge of the identity of the best subchannel is reflected in (12); the lack of knowledge of the transition probabilities for the other subchannels is reflected in (13), in which the Q_i are assumed unknown. Notice also that the constraints on the compound channel Λ apply only to the marginal transition probabilities for the subchannels: the joint transition probabilities are otherwise unconstrained. This corresponds to the fact that the joint transition probabilities for the outputs of the subchannels in the universal receiver are unknown.

Consider the genie-based receiver. If the encoder knows the input distribution P^{**} that achieves the capacity of the best subchannel as well as the value $C(W^*)$ of the capacity, then it follows from results in [7] that the capacity of the best subchannel can be achieved. The first part of the following proposition states that if the encoder for the N-channel corresponding to the universal receiver has the same information, then the capacity of the N-channel equals $C(W^*)$.

The proof follows from classical compound channel theory [7] and is outlined in Appendix B. Suppose now, that for a given rate R, the encoder for the genie-based receiver knows the input distribution $P^*(R)$ that maximizes the random coding exponent function $E_r(R, P, W^*)$ for the best subchannel. Results in universal coding [2, p. 172] imply that the $E_r(R, W^*)$ is an attainable error exponent at rate R for the genie-based receiver. Note that $E_r(R, W^*) = E(R, W^*)$ for $R \geq R_{cr}(W^*)$, so that the reliability function of W^* is attainable for rates higher than the critical rate. The second part of the following proposition states that if the encoder for the N-channel has the same information, then $E_r(R, W^*)$ is an attainable error exponent for the N-channel. The proof is given in Appendix B, and is a simple consequence of results in [2].

Proposition 3:

1) If the encoder knows P^{**} and $C(W^*)$, then $C(\Lambda) = C(W^*)$.
2) For a given rate R, if the encoder knows $P^*(R)$, then $E(R, \Lambda) \geq E_r(R, W^*)$. If $R \geq R_{cr}(W^*)$, then $E(R, \Lambda) = E(R, W^*)$.

From a practical point of view, the important feature of Proposition 3 is that it implies the performance of the universal receiver is just as good as that of the genie-based receiver in the following sense. If an appropriate code of rate less than the capacity of the best subchannel is used, arbitrarily small error probabilities can be achieved with either the universal receiver or the genie-based receiver, and the random coding error exponent for the best subchannel is an attainable error exponent for each of these receivers. Furthermore, the encoder for the system with the universal receiver does not need any additional information beyond that required by the encoder for the system with the genie-based receiver. It is worth noting that for most situations of practical interest, the best subchannel is symmetric (in the broad sense defined by Gallager [4, p. 94]), so that P^{**}, as well as $P^*(R)$ for any given R, is the uniform distribution. The encoder needs no further information about the N-channel in such a case.

Proposition 3 implies that, from an information-theoretic point of view, there is no penalty in terms of either the rate or the error exponent for not knowing the identity of the best subchannel. However, these results require the transition probability matrix for the N-channel to remain unchanged for arbitrarily long time periods, corresponding to large block lengths for the codes. For time-varying channels, the codeword lengths must be finite and small enough that the channel conditions do not change appreciably over the duration of a codeword. For practical codes, therefore, there may be a penalty for not knowing which subchannel is the best, as we shall see from the example given in the next section.

IV. Performance of Parallel Receivers with Coding—An Example

In this section we consider the performance of practical coding schemes for a given parallel configuration of receivers. In order to analyze the performance of the coding scheme for a given channel in effect, consider as in the previous section the corresponding N-channel (see Fig. 3). We restrict attention to a simple example in which each subchannel is an M-ary symmetric channel (MSC), which is characterized completely by its crossover probability. Let ϵ_1 denote the smallest of these crossover probabilities; that is, ϵ_1 is the crossover probability for the best subchannel. For a given channel class, ϵ_1 will in general range over a set of possible values. It is assumed, therefore, that the crossover probability ϵ_1 for the best subchannel is known to lie in a range $[\epsilon_{\min}, \epsilon_{\max}]$. The goal of this section is to design a coding scheme that will give good performance over the entire range of possible values of ϵ_1.

In the following, bounds on the performance of a given coding scheme are obtained. These bounds depend only on the crossover probability ϵ_1 for the best subchannel, and the number N of subchannels. No assumptions are made about the joint transition probabilities for the N-channel. In particular, the identity of the best subchannel and the values of the crossover probabilities for the other subchannels are not known. In evaluating the performance of a coding scheme, therefore, we derive worst-case results over all possible N-channels for which the subchannels are MSC's, and the best subchannel has crossover probability ϵ_1.

In the analysis, notation is also needed for the crossover probabilities for the other subchannels. Number the subchannels from 1 through N, and let ϵ_i denote the crossover probability for the ith subchannel. According to this convention, the first subchannel is the best, in the sense that

$$\epsilon_1 \leq \epsilon_i, \qquad i = 2, \cdots, N. \tag{14}$$

The receiver does not know which subchannel is the best, and the identity of the best subchannel may change as the channel in effect changes. The convention (14) is established merely for notational convenience in the analysis. For the purpose of the analysis, therefore, if the identity of the best subchannel changes, an appropriate renumbering of the subchannels (in which the first subchannel is the best) is implicit in our formulation. The N-channel considered here is a special case of the N-channel W of the previous section: the best subchannel $W^* = W_1$ is an MSC with crossover probability ϵ_1, and the stochastic matrices Q_i correspond to MSCs with (unknown) crossover probabilities q_i given by

$$q_i = (1 - M^{-1}) \frac{\epsilon_i - \epsilon_1}{(1 - M^{-1}) - \epsilon_1}.$$

After deriving the performance bounds, we give numerical results for the performance of coding schemes based on extended Reed–Solomon codes, followed by a procedure for designing such coding schemes to attain specified levels of performance. For extended Reed–Solomon codes, the performance bounds turn out to be monotone in ϵ_1. Thus, in order to design the code for ϵ_1 in the range of interest, it is enough to design the code for $\epsilon_1 = \epsilon_{\max}$.

The coding scheme used for the N-channel is shown in Fig. 4. The same codeword is sent over all the subchannels, and the corresponding outputs are decoded separately in the usual manner. Each decoder produces one of three possible outcomes: a) correct decoding, in which the number of symbol errors is within the error-correcting capability of the code, b) decoding error, in which the number of symbol errors is beyond the error-correcting capability of the code, and the decoder decodes into a wrong codeword, or c) decoding failure, in which the number of symbol errors is

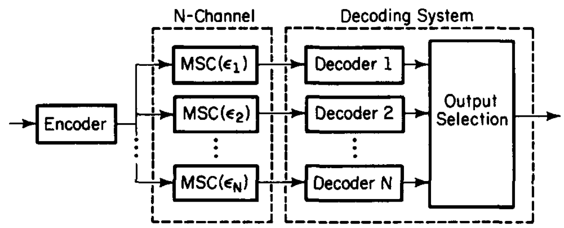

Fig. 4. Model for analysis of performance with coding.

beyond the error-correcting capability of the code, but the decoder realizes this and reports a failure to decode instead of decoding into a wrong codeword.

The results of the individual decodings are processed using the following output selection algorithm. All outputs that fail to decode are ignored. If two or more of the outputs that are decoded fail to match, or if all outputs fail to decode, a failure is declared for the decoding system. If all outputs that decode are in agreement, then the resulting decoded word is assumed to be correct, and a successful decoding is declared for the decoding system. Note that if all the decoded outputs match, but are wrong, the result is a decoding error for the decoding system.

The performance measures of interest are the probability of decoding failure Q_F and the probability of decoding error Q_E for the decoding system. It is clear that the performance of the decoding system will depend on the joint transition probabilities for the N-channel. In what follows, we develop bounds on the performance that depend on the numbers ϵ_1 and N only.

A word on the choice of the code is in order at this point. Several of the receivers may be badly matched to a given channel, which translates to several of the discrete subchannels having high symbol error probabilities. The code should be such that, with high probability, such bad outputs fail to decode rather than lead to a decoding error. Reed–Solomon codes are particularly good in this context, especially if the full error-correcting capability of the code is not utilized. Thus, bounded distance decoding rather than maximum likelihood decoding is used in this application.

An (n,k) code with minimum distance d is used, and the decoders in the decoding system attempt to correct up to t errors, where $2t+1 \le d$. Let $P_E(u)$ denote the conditional probability of decoding error for a particular decoder given that there are u errors in the decoder input. Note that a decoding error can occur only if the number of errors exceeds $(d-t)$, since the received word must be within a Hamming distance of t from a wrong codeword in order to produce a decoding error. Thus, by decreasing t (i.e., decoding below the full error-correcting capability of the code), the probability of decoding error for the decoders can be reduced. For a maximum distance separable code, such as a Reed–Solomon code, the minimum distance is $d = n-k+1$, and upper bounds on $P_E(u)$ have been developed in [5]. The expression for $P_B(u)$, one such bound, is

$$P_B(u) = \begin{cases} 0, & u \le d-t-1, \\ (M-1)^{-(n-k)} \displaystyle\sum_{j=\min(0,d-u)} \binom{n}{j}(M-1)^j, & u \ge d-t. \end{cases} \tag{15}$$

This bound is substituted for $P_E(u)$ in the expression derived later in order to get our numerical results for schemes using extended Reed–Solomon codes.

Simple union bounds are first developed for Q_F and Q_E. The bound on Q_F is sufficiently tight for the applications that arise in this paper, and we proceed no further in this case. The bound on Q_E, however, is quite loose, so a better bound is developed.

The performance of the decoder at the output of a given subchannel depends only on the symbol error probability for that subchannel. Let $P_{DF}(\epsilon)$ be the probability that a decoder at the output of an MSC with crossover probability ϵ does not decode correctly. Clearly $P_{DF}(\epsilon)$ is an upper bound for the probability of decoding failure for that decoder. Because the decoder is far more likely to report a decoding failure than decode into a wrong codeword, this upper bound is very tight. The expression for $P_{DF}(\epsilon)$ is

$$P_{DF}(\epsilon) = \sum_{j=t+1}^{n} \binom{n}{j} \epsilon^j (1-\epsilon)^{(n-j)},$$

whereas the probability of decoding error for a decoder at the output of an MSC with crossover probability ϵ is given by

$$P_{DE}(\epsilon) = \sum_{j=d-t}^{n} P_E(j) \binom{n}{j} \epsilon^j (1-\epsilon)^{(n-j)}.$$

These are used to get bounds on Q_F and Q_E as follows.

Note that if the received word for the best subchannel is decoded correctly, a decoding failure for the decoding system can occur if there is an incorrect decoding for one of the other subchannels. This observation is used to obtain the following bound on Q_F. We have

$$Q_F \le \Pr\begin{bmatrix} \text{decoder for best subchannel} \\ \text{does not decode correctly} \end{bmatrix} + \Pr\begin{bmatrix} \text{decoder for best subchannel decodes} \\ \text{correctly and one of the other} \\ \text{decoders decodes incorrectly} \end{bmatrix},$$

The first term $P_{DF}(\epsilon_1)$, and the second term is bounded above by the probability that one of the other decoders decodes incorrectly, which in turn is bounded by $\sum_{i=2}^{N} P_{DE}(\epsilon_i)$ using a union bound. We thus have

$$Q_F \le P_{DF}(\epsilon_1) + \sum_{i=2}^{N} P_{DE}(\epsilon_i).$$

This yields the upper bound

$$Q_F \le P_{DF}(\epsilon_1) + (N-1) \max_{\epsilon_1 \le \epsilon_2 \le 1} P_{DE}(\epsilon_2), \tag{16}$$

in which the worse subchannels are taken to be as bad as possible while satisfying the constraint (14). For Q_E, we have

$$Q_E \le \Pr[\text{at least one decoder decodes incorrectly}],$$

so that, using a union bound, we get $Q_E \le \sum_{i=1}^{N} P_{DE}(\epsilon_i)$. This yields the worst-case bound

$$Q_E \le P_{DE}(\epsilon_1) + (N-1) \max_{\epsilon_1 \le \epsilon_2 \le 1} P_{DE}(\epsilon_2). \tag{17}$$

The bound (16) for Q_F is tight in the sense that there are joint transition probabilities for the N-channel that satisfy the given constraints, and for which the bound (16) is close to the actual value of Q_F. As an example, consider an N-channel for which, whenever there is an error at the output of the best subchannel, there is an error at the output of all the other subchannels. In this case, the best subchannel produces, with probability one, the fewest errors in a given transmitted codeword, and $P_{DF}(\epsilon_1)$ is a very good bound for the probability that none of the decoders decodes correctly. The latter is, in turn, a good upper bound for the probability that all the decoders fail to decode (which leads to decoding failure for the decoding system), since for each decoder, the probability of decoding failure is much larger than that of decoding error. If the output of the best subchannel decodes correctly, decoding failure for the decoding system can still result from a decoding error at one of the decoders. The second term in (16) is an upper bound for the probability of this event, and it is typically much smaller than $P_{DF}(\epsilon_1)$. As a result, the first term in (16) is usually the dominant one. Thus, for the N-channel in the example above, the bound (16) is a good approximation for the actual value of Q_F. On the other hand, the upper bound (17) for the probability of decoding error for the decoding system is not very good, being governed by the probability of decoding error for the worst subchannel. In deriving the upper bound, we have ignored the possibility that even if the output of a bad subchannel decodes incorrectly, a correct decoding at the output of a good subchannel will result in decoding failure rather than decoding error for the decoding system. It is desirable, therefore, to develop a bound that exploits the presence of the good subchannels, and this is done in the following.

It is not possible for the decoding system to decode incorrectly if there is a correct decoding for the best subchannel. A decoding error for the decoding system may result only if either a) the decoder for the best subchannel decodes incorrectly, or b) the decoder for the best subchannel fails to decode, and some other decoder decodes incorrectly. For each of the decoders in the decoding system, if a correct decoding is not possible, decoding failure is much more likely than decoding error. Thus, the probability of the event described in b) is bounded quite tightly by the probability that the decoder for the best subchannel does not decode correctly and some other decoder decodes incorrectly. It is clear from the foregoing that the probability of decoding error for the decoding system satisfies

$$Q_E \le \Pr[\text{decoder for best subchannel decodes incorrectly}] + \Pr\begin{bmatrix}\text{decoder for best subchannel does not}\\ \text{decode correctly and one of the other}\\ \text{decoders decodes incorrectly}\end{bmatrix}.$$

The first term equals $P_{DE}(\epsilon_1)$. Application of a union bound to the second term gives

$$Q_E \le P_{DE}(\epsilon_1) + \sum_{i=2}^{N} \Pr\begin{bmatrix}\text{decoder for best subchannel}\\ \text{does not decode correctly and}\\ \text{decoder for } i\text{th subchannel}\\ \text{decodes incorrectly}\end{bmatrix}. \quad (18)$$

Some additional notation is required to make this bound more explicit. Let U_i be the number of code symbols in error at the output of the ith subchannel. The ith subchannel is memoryless and has crossover probability ϵ_i, and a codeword comprises n symbols. Thus, U_i is a binomial random variable with parameters n and ϵ_i. The distribution of the random vector $U = (U_1, \cdots, U_N)$ depends on the joint transition probabilities for the N-channel.

Define the indicator function $I_t(u)$ by

$$I_t(u) = \begin{cases} 0, & u \le t, \\ 1, & u \ge t+1. \end{cases}$$

The conditional probability corresponding to a typical term in the summation in (18) is

$$\Pr\left[\begin{matrix}\text{decoder for best subchannel does not}\\ \text{decode correctly and decoder for } i\text{th}\\ \text{subchannel decodes incorrectly}\end{matrix}\,\middle|\, U = u\right] = I_t(u_1) P_E(u_i).$$

Removing the conditioning, we have that the ith term in the summation in (18) equals

$$E\{I_t(U_1) P_E(U_i)\} \le E\{I_t(U_1)\}\left(\max_u P_E(u)\right). \quad (19)$$

But $E\{I_t(U_1)\} = P_{DF}(\epsilon_1)$, the probability that the output of the first subchannel does not decode correctly. Thus, each term in the summation in (18) is bounded by (19), so that

$$Q_E \le P_{DE}(\epsilon_1) + (N-1) P_{DF}(\epsilon_1)\left(\max_u P_E(u)\right). \quad (20)$$

The foregoing analysis gives bounds on the performance of the decoding system shown in Fig. 4. By an appropriate choice of the receivers in the parallel configuration, it is ensured that at least one of the receivers performs well for any channel in the channel class. However, it is not possible to differentiate between the good receivers and the bad for a given channel. The effect of this lack of side information on the performance of the coding scheme for the universal receiver is evaluated by means of a comparison with the performance of the same code for the genie-based receiver described in Section III. For the genie-based receiver, the performance is given by that of the decoder at the output of the best subchannel; that is, the probability of decoding failure is given by $P_{DF}(\epsilon_1)$, and the probability of decoding error by $P_{DE}(\epsilon_1)$. Thus, the bound (16) for Q_F is compared with $P_{DF}(\epsilon_1)$, and the union bound (17) and the bound (20) on Q_E are compared with $P_{DE}(\epsilon_1)$.

For the remainder of this section, we restrict ourselves to extended Reed–Solomon codes with block length $n = M$. The minimum distance d of the code is related to the number k of information symbols by $d = n - k + 1$. The decoders in the decoding system attempt to correct up to t errors. Given the alphabet size M, therefore, the coding schemes considered henceforth are completely specified by the values of k and t. The maximum error-correcting capability of the code is $e = \lfloor (d-1)/2 \rfloor$, and t may be chosen to be strictly less than e so that the outputs of bad subchannels are less likely to decode into a wrong codeword. In the numerical results, the upper bound $P_B(u)$ given by (15) is substituted for $P_E(u)$ in the expression of interest, leading to upper bounds for these expressions. The bound $P_B(u)$ is nondecreasing in u, so that the minimum probability of

TABLE I-A
DECODING ERROR AND DECODING FAILURE PROBABILITIES FOR $M=32$ WITH A (32,16) EXTENDED REED–SOLOMON CODE, $\epsilon_1=0.1$

		Decoding Error Probabilities			Decoding Failure Probabilities	
N	t	$P_{DE}(\epsilon_1)$	Bound (20) on Q_E	Union Bound (17) on Q_E	$P_{DF}(\epsilon_1)$	Bound (16) on Q_F
2	2	7.00E-26	4.16E-19	6.57E-19	0.633	0.633
	4	2.52E-19	9.70E-15	4.59E-14	0.211	0.211
	6	1.93E-13	4.01E-11	1.11E-09	3.58E-02	3.58E-02
	8	4.07E-08	8.18E-08	1.25E-05	3.30E-03	3.31E-03
5	2	7.00E-26	1.66E-18	2.63E-18	0.633	0.633
	4	2.52E-19	3.88E-14	1.83E-13	0.211	0.211
	6	1.93E-13	1.60E-10	4.45E-09	3.58E-02	3.58E-02
	8	4.07E-08	2.05E-07	4.99E-05	3.30E-03	3.35E-03
10	2	7.00E-26	3.74E-18	5.91E-18	0.633	0.633
	4	2.52E-19	8.73E-14	4.13E-13	0.211	0.211
	6	1.93E-13	3.59E-10	1.00E-08	3.58E-02	3.58E-02
	8	4.07E-08	4.10E-07	1.12E-04	3.30E-03	3.41E-03

TABLE I-B
DECODING ERROR AND DECODING FAILURE PROBABILITIES FOR $M=32$ WITH A (32,16) EXTENDED REED–SOLOMON CODE, $\epsilon_1=0.01$

		Decoding Error Probabilities			Decoding Failure Probabilities	
N	t	$P_{DE}(\epsilon_1)$	Bound (20) on Q_E	Union bound (17) on Q_E	$P_{DF}(\epsilon_1)$	Bound (16) on Q_F
2	2	3.16E-40	2.62E-21	6.57E-19	3.99E-03	3.99E-03
	4	1.33E-31	7.37E-19	4.59E-14	1.61E-05	1.61E-05
	6	1.18E-23	3.01E-17	1.11E-09	2.70E-08	2.81E-08
	8	2.81E-16	5.65E-16	1.25E-05	2.28E-11	1.25E-05
5	2	3.16E-40	1.05E-20	2.63E-18	3.99E-03	3.99E-03
	4	1.33E-31	2.95E-18	1.83E-13	1.61E-05	1.61E-05
	6	1.18E-23	1.20E-16	4.45E-09	2.70E-08	3.15E-08
	8	2.81E-16	1.42E-15	4.98E-05	2.28E-11	4.98E-05
10	2	3.16E-40	2.36E-20	5.91E-18	3.99E-03	3.99E-03
	4	1.33E-31	6.64E-18	4.13E-13	1.61E-05	1.61E-05
	6	1.18E-23	2.71E-16	1.00E-08	2.70E-08	3.71E-08
	8	2.81E-16	2.84E-15	1.12E-04	2.28E-11	1.12E-04

decoding error for the decoder for a bad subchannel is

$$\max_{\epsilon_1 \le \epsilon_2 \le 1} P_{DE}(\epsilon_2) = P_{DE}(1).$$

Also, we have

$$\max_u P_E(u) = P_B(n) = (M-1)^{-(n-k)} \sum_{j=0}^{t} \binom{n}{j} (M-1)^j.$$

The numerical results presented in Table I-A and Table I-B are for $M=32$ and a (32,16) extended Reed–Solomon code, so $n=32$, $k=16$, $d=17$, and $e=8$. The values of ϵ_1, N, and t are varied. It is found that the bound (20) for Q_E is always better than the union bound (17), but the difference is more significant for small values of ϵ_1. Both these bounds are significantly larger than $P_{DE}(\epsilon_1)$, the probability of decoding error for the genie-based receiver, but the probability of decoding error of the decoding system for the universal receiver is still low enough for a practical system design. The probability of decoding failure for the decoding system, on the other hand, is almost equal to $P_{DF}(\epsilon_1)$, the probability of decoding failure for the genie-based receiver, as long as we decode below the maximum error-correcting capability of the code.

Next, the problem of designing a suitable coding scheme for attaining specified levels of performance is considered. We restrict attention to the coding schemes based on extended Reed–Solomon codes described above, for which the design consists simply of specifying values for k and t. Suppose that p_f and p_e are the desired probability of decoding failure and decoding error, respectively, to be attained over a given class of channels. For the N-channels corresponding to these channels, the crossover probability ϵ_1 for the best subchannel is assumed to lie in a range $[\epsilon_{\min}, \epsilon_{\max}]$. Given $\epsilon_{\min}$, $\epsilon_{\max}$, N, and n, we are required to find k and t such that the information rate k/n is maximized subject to the constraints $Q_F \le p_f$ and $Q_E \le p_e$ for all possible channels in effect. This is equivalent to minimizing the minimum distance d subject to the same constraints, since $n=M$ and $d=n-k+1$.

The following is a design procedure that minimizes d under the given constraints. The design is based on the bound (16) on Q_F and the bound (20) on Q_E. For economy of notation, let Q_F denote the bound in (16) and Q_E the bound in (20). It is easy to see that each of these bounds is increasing in ϵ_1, so it suffices for the design to consider $\epsilon_1 = \epsilon_{\max}$. This is the value of ϵ_1 used in the following. Thus, a worst-case approach is used for the problem of code design.

The probability of decoding failure for the decoding system is given by

$$Q_F = Q_F(\epsilon_1; d, t) = P_{DF}(\epsilon_1; t) + (N-1) P_{DE}(1; d, t), \quad (21)$$

where the dependence on ϵ_1, d, and t has been made explicit. The first term on the extreme right-hand side is usually the dominant one, hence the second term is ignored

TABLE II-A
CODE DESIGN FOR $M = 32$ USING EXTENDED REED–SOLOMON CODES WITH BLOCKLENGTH $n = 32$
$\epsilon_1 = 0.1$

	Code Parameters			Performance Attained	
N	k^*	t^*	% Loss in Rate	Q_F	Q_E
Required Performance: $p_f = 1.0\text{E-}3$, $p_e = 1.0\text{E-}13$					
1	12	9	0	8.1E-4	3.6E-14
2	11	9	8.3	8.1E-4	2.9E-14
5	11	9	8.3	8.1E-4	5.9E-14
10	10	9	16.7	8.1E-4	8.5E-15
Required Performance: $p_f = 1.0\text{E-}6$, $p_e = 1.0\text{E-}16$					
1	6	13	0	8.1E-7	1.2E-17
2	6	13	0	8.1E-7	2.3E-17
5	6	13	0	8.1E-7	5.9E-17
10	5	13	16.7	8.1E-7	3.5E-18

TABLE II-B
CODE DESIGN FOR $M = 32$ USING EXTENDED REED–SOLOMON CODES WITH BLOCKLENGTH $n = 32$
$\epsilon_1 = 0.01$

	Code Parameters			Performance Attained	
N	k^*	t^*	% Loss in Rate	Q_F	Q_E
Required Performance: $p_f = 1.0\text{E-}6$, $p_e = 1.0\text{E-}16$					
1	19	5	0	7.2E-7	5.4E-18
2	16	5	15.8	7.2E-7	5.8E-18
5	16	5	15.8	7.2E-7	2.3E-17
10	16	5	15.8	7.2E-7	5.2E-17
Required Performance: $p_f = 1.0\text{E-}10$, $p_e = 1.0\text{E-}20$					
1	14	8	0	2.3E-11	1.4E-22
2	12	8	14.3	3.6E-11	3.1E-22
5	12	8	14.3	7.7E-11	1.2E-21
10	11	8	21.4	2.7E-11	8.9E-23

in making a preliminary choice of t that is likely to satisfy the constraint on Q_F, as follows:

$$t^* = \min\{t\colon P_{DF}(\epsilon_1; t) \le p_f\}.$$

Given that $t = t^*$, minimizing d subject to the constraint on Q_E yields

$$d^* = \min\{d\colon d \ge 2t^* + 1 \text{ and } Q_E(\epsilon_1; d, t^*) \le p_e\}. \quad (22)$$

Finally, it is checked whether the constraint on Q_F is satisfied when the second term on the extreme right-hand side of (21) is taken into account; that is, whether

$$Q_F(\epsilon_1; d^*, t^*) = P_{DF}(\epsilon_1; t^*) + (N-1)P_{DE}(1; d^*, t^*) \le p_f.$$

If the condition is satisfied, the design is complete. It not, we have to increase either t^* or d^* (or both) so that the performance criteria are met while minimizing the value of d^*. The value of t^* affects the decoding algorithm but does not affect the code rate; it is, moreover, an important intermediate variable in determining d^*, as is seen from (22). The number of information symbols is given by $k^* = n - d^* + 1$. Numerical examples of designs obtained are given in Tables II-A and II-B.

It is of interest to compare the design above to a coding scheme (based on an appropriate extended Reed–Solomon code) that attains the same performance levels for the genie-based receiver. The probability of decoding failure, $P_{DF}(\epsilon_1)$, and the probability of decoding error, $P_{DE}(\epsilon_1)$, for the genie-based receiver are both increasing in ϵ_1, so that, as before, it suffices to consider $\epsilon_1 = \epsilon_{\max}$ in designing for ϵ_1 in the given range. Using this value of ϵ_1, the design is simply given by

$$t^* = \min\{t\colon P_{DF}(\epsilon_1; t) \le p_f\}$$

and

$$d^* = \min\{d\colon d \ge t^* + 1 \text{ and } P_{DE}(\epsilon_1; d, t^*) \le p_e\}.$$

In most of the numerical examples in Tables II-A and II-B, the value of t^* obtained is the same that was given previously, but the value of d^* is in general smaller. This means that transmission at a higher rate is possible if the identity of the best subchannel is unknown, since the outputs of the bad subchannels can then be ignored. This corresponds to the results for $N = 1$ in Tables II-A and II-B, which display the loss in rate, as the number of subchannels increases, in order to maintain the same performance levels.

The foregoing results show that the universal receiver of Section II does achieve universal demodulation, and that it is possible to design a practical coding scheme that exploits the presence of the good receiver in the parallel configuration, despite the fact that the identity of the good receiver is not known. Unlike the information-theoretic results of the previous section, however, the presence of receivers that may perform very poorly for the channel in effect results in transmission at a lower rate in order to achieve the same performance as the genie-based receiver.

V. CONCLUSION

We have introduced a universal approach to dealing with uncertainty in the knowledge of channel characteristics. We have given conditions and procedures for finding a finite set of receivers which, when placed in a parallel configuration, provide universal demodulation. It is established that the resulting universal receiver does provide universal demodulation by comparing its performance with that of an idealized and unrealizable genie-based receiver.

Some topics for further investigation are listed in the following.

1) The detection aspect of the design has been stated in a rather abstract setting in this paper. In our following work, we intend to consider specific channel classes and carry out a universal design for specific degradation functions. This will enable us to explore the trade-offs between the number of receivers and the allowable deviation from optimality.

2) The results of this paper are for memoryless channels and symbol-by-symbol demodulation, and the performance functional is typically the symbol error probability. It would be interesting to apply the universal approach to channels with memory, such as channels with intersymbol interference. The performance functional for such an application must reflect the accuracy with which a *sequence* of symbols is reproduced at the receiver.

3) In the coding aspect of our design, we have considered codes with nonbinary alphabets. Due to the widespread use of binary signaling, it is also of interest to design schemes using binary codes. Our results here indicate that concatenated codes with a binary inner code and a Reed–Solomon outer code might perform well, but there may be other good choices.

4) In a specific application, some additional information about the joint transition probabilities for the N-channel may be available. It may be possible to exploit this information to obtain tighter bounds on the performance of the coding system of Section IV, as well as to design a better coding system that takes advantage of the additional information. In particular, side information concerning the identity of the good receiver could be used to improve the performance of the decoding system. It is especially important to enhance the performance of binary coding systems using such information, because the common binary codes do not have the good distance properties of Reed–Solomon codes.

5) The concept that is applied to communication systems in this paper may be applicable to other systems as well, but several of the issues that arise differ considerably from the communications application. In a radar-type application, for instance, there is no single technique to identify the best subsystem. Trying to make the probability of a miss smaller using a parallel configuration can lead to a high probability of false alarm. However, there are certain applications for which the latter is not a big drawback (e.g., if false alarms can be rejected by other means).

Appendix A

In this Appendix, we supply some details for Example 2 and give the proof of Proposition 1.

Details of Example 2: To see that the equicontinuity requirement is violated under the weak topology on C, note that although $\boldsymbol{p}_n$ converges weakly to $\boldsymbol{u}$,

$$\lim_n |P(\boldsymbol{u}, A_n) - P(\boldsymbol{p}_n, A_n)| = 1/2.$$

Thus, the family of functions $\{P(\boldsymbol{p}, A),\ A \in R\}$ is not equicontinuous on C.

Next, it is shown that for the degradation function considered, it is not possible to obtain a universal set of receivers. In particular, we show that, if $\epsilon < 1/8$, then for any finite set of receivers $D_1, \cdots, D_N$, there is an integer M such that, for all $n > M$,

$$\min_{1 \le i \le N} P(\boldsymbol{p}_n, D_i) > P^*(\boldsymbol{p}_n) + \epsilon = \epsilon. \tag{A.1}$$

To this end, consider any given receiver A. Using (8) and (9) from Section II-A, it is easy to show that

$$P(\boldsymbol{p}_n, A) = \mu(A_n \cap A^c) + \mu(A_n^c \cap A). \tag{A.2}$$

Suppose that, for some n, $P(\boldsymbol{p}_n, A) < \epsilon$. From (A.2), we have $\mu(A_n^c \cap A) < \epsilon$. Since $\mu(A_n^c) = 1/2$, this implies that

$$\mu(A_n^c \cap A^c) > 1/2 - \epsilon. \tag{A.3}$$

For any $m > n$, we have

$$\begin{aligned} \mu(A_n^c \cap A^c) &= \mu(A_m^c \cap A_n^c \cap A^c) + \mu(A_m \cap A_n^c \cap A^c) \\ &\le \mu(A_m^c \cap A_n^c) + \mu(A_m \cap A^c). \end{aligned} \tag{A.4}$$

Since $\mu(A_m^c \cap A_n^c) = 1/4$, we have using (A.3) and (A.4) that, for $\epsilon < 1/8$,

$$\mu(A_m \cap A^c) \ge (1/2 - \epsilon) - 1/4 = 1/4 - \epsilon > \epsilon.$$

This implies, using (A.1), that $P(\boldsymbol{p}_m, A) > \epsilon$. Thus, a receiver that provides the desired performance for one of the $\boldsymbol{p}_n$ does not do so for $\{\boldsymbol{p}_m,\ m > n\}$. These considerations imply (A.1), which in turn implies that a universal set of receivers cannot be obtained for the given degradation function.

Proof of Proposition 1: The proof is by contradiction. Assume that $\inf_{x \in C} m(x) = 0$. Then there is a sequence (x_k) in C such that, for each k,

$$m(x_k) < \frac{1}{k}. \tag{A.5}$$

Without loss of generality, assume that (x_k) converges to $x_0 \in C$ (passing to a convergent subsequence if necessary, using the compactness of C). By (A.5), for each k, we have $u_k \in C$ and $y_k \in R_{x_k}$ satisfying

$$\|x_k - u_k\| < \frac{1}{k}, \tag{A.6}$$

and

$$f(u_k, y_k) > h[g(u_k)]. \tag{A.7}$$

We see from (A.6) that the sequence (u_k) also converges to x_0. Also, we have

$$\begin{aligned} f(u_k, y_k) = {} & [f(u_k, y_k) - f(x_0, y_k)] \\ & + [f(x_0, y_k) - f(x_k, y_k)] + f(x_k, y_k). \end{aligned}$$

The first two terms go to zero as $k \to \infty$ because of the equicontinuity of the family $\{f(x, y),\ y \in R\}$ on C. The third term equals $g(x_k)$, since $y_k \in R_{x_k}$. We note that g is continuous under the given hypotheses (see the proof of Theorem 1). Also, the degradation function h is continuous by definition. Hence, letting $k \to \infty$ on both sides of (A.7) yields

$$g(x_0) \ge h[g(x_0)].$$

This contradicts the fact that $h(s) > s$ for all $s \in [0, \infty)$, and proves that $\delta_C > 0$.

Appendix B

We prove Proposition 3 in Section III here. Before doing so, it is necessary to develop some additional notation and terminology not introduced in the main text for the sake of clarity of presentation. As before, the reader is referred to [2] for details. For distributions P_1 and P_2, let $D(P_1\|P_2)$ be the divergence of P_1 relative to P_2. If the input X of a DMC W is distributed according to a probability vector P (considered to be a row vector), the output Y is distributed according to the probability vector PW. The mutual information $I(X;Y)$ between X and Y is denoted by $I(P,W)$ in this case.

$$I(P,W)=\sum_x P(x)D(W(\cdot|x)\|PW(\cdot)).$$

Given another DMC V, the conditional divergence $D(V\|W|P)$ is given by

$$D(V\|W|P)=\sum_x P(x)D(V(\cdot|x)\|W(\cdot|x)).$$

For a DMC W with input distribution P, the sphere packing exponent function $E_{sp}(R,P,W)$ for constant composition codes of rate R is given by

$$E_{sp}(R,P,W)=\min_{V:\,I(P,V)\le R} D(V\|W|P).$$

Maximizing over P, we get the sphere packing exponent function $E_{sp}(R,W)$. For the same situation as before, the random coding exponent functions $E_r(R,P,W)$ and $E_r(R,W)$ have been defined in Section III. From [2],

$$E_r(R,P,W)=\min_{R'\ge R}\{E_{sp}(R',P,W)+R'-R\},$$

which yields $E_r(R,W)=\min_{R'\ge R}\{E_{sp}(R',W)+R'-R\}$ on maximizing over P. The reliability function $E(R,W)$ is upperbounded by $E_{sp}(R,W)$ and lowerbounded by $E_r(R,W)$. The critical rate $R_{cr}=R_{cr}(W)$ for the DMC W is the smallest R at which the convex curve $E_{sp}(R,W)$ meets its supporting line of slope -1. For rates above the critical rate, the sphere packing exponent function and the random coding exponent function are equal, and are therefore equal to the reliability function of the channel.

Recall that the compound channel Λ of interest has a single M-ary input X and an M-ary N-vector $Y=(Y_1,\cdots,Y_N)$ as output. Assume, without loss of generality, that $W_1=W^*$, that is, the first subchannel is the best. This assumption is made only for notational convenience, and does not imply any knowledge of the identity of the best subchannel. Then, for any $\boldsymbol{W}\in\Lambda$, the transition probability matrix W_i for the ith subchannel is given by

$$W_i=W_1Q_i,\qquad i=2,\cdots,N,$$

where the Q_i are unknown stochastic matrices that depend on $\boldsymbol{W}$.

We show first that

$$\min_{\boldsymbol{W}\in\Lambda} I(P,\boldsymbol{W})=I(P,W_1). \tag{B.1}$$

If the encoder knows P^{**}, the input distribution that maximizes the right-hand side above, and the value of the resulting maximum, it is easy to show, considering the proof of Theorem 4.3.1 and 4.4.1 in [7], that

$$C(\Lambda)=\max_P\min_{\boldsymbol{W}\in\Lambda} I(P,\boldsymbol{W})=\max_P I(P,W_1)$$
$$=I(P^{**},W_1)=C(W_1).$$

This will prove the first part of the proposition. The only information required by the encoder here is the input distribution achieving the capacity of the best subchannel, and the value of that capacity.

To show (B.1), note that, for any $\boldsymbol{W}\in\Lambda$, we have by definition that

$$I(P,\boldsymbol{W})=I(X;Y_1,\cdots,Y_N),$$

where $\boldsymbol{Y}=(Y_1,\cdots,Y_N)$ is the output of the DMC $\boldsymbol{W}$ when the input X has distribution P. Also, we have

$$I(P,W_1)=I(X;Y_1).$$

Now,

$$I(X;Y_1,\cdots,Y_N)=I(X;Y_1)+I(X;Y_2,\cdots,Y_N|Y_1). \tag{B.2}$$

This proves that $I(X;Y_1,\cdots,Y_N)\ge I(X;Y_1)$, which implies that

$$I(P,\boldsymbol{W})\ge I(P,W_1),\qquad \text{for all } \boldsymbol{W}\in\Lambda. \tag{B.3}$$

Now, define a DMC $\boldsymbol{W}$ as

$$\boldsymbol{W}(y_1,\cdots,y_N|x)=W_1(y_1|x)\boldsymbol{Q}(y_2,\cdots,y_N|y_1), \tag{B.4}$$

where $\boldsymbol{Q}$ is an arbitrary stochastic matrix. Clearly, $\boldsymbol{W}\in\Lambda$. Also, for this DMC, $X\to Y_1\to(Y_2,\cdots,Y_N)$ is a Markov chain, so that $I(X;Y_2,\cdots,Y_N|Y_1)=0$, which implies, using (B.2), that $I(X;Y_1,\cdots,Y_N)=I(X;Y_1)$. In other words, there is a $\boldsymbol{W}\in\Lambda$ such that $I(P,\boldsymbol{W})=I(P,W_1)$. This, together with (B.3), implies (B.1), completing the proof of the first part of the proposition.

To prove the second part of the proposition, consider the sphere packing exponent function $E_{sp}(R,P,\boldsymbol{W})$ for any $\boldsymbol{W}\in\Lambda$. It is first shown that

$$E_{sp}(R,P,\boldsymbol{W})\ge E_{sp}(R,P,W_1). \tag{B.5}$$

Next, it is proved that, for $\boldsymbol{W}$ as in (B.4),

$$E_{sp}(R,P,\boldsymbol{W})=E_{sp}(R,P,W_1). \tag{B.6}$$

It follows from (B.5) and (B.6) that

$$\min_{\boldsymbol{W}\in\Lambda} E_{sp}(R,P,\boldsymbol{W})=E_{sp}(R,P,W_1). \tag{B.7}$$

The following equality holds for any DMC $\boldsymbol{W}$ [2]

$$E_r(R,P,\boldsymbol{W})=\min_{R'\ge R}\left[E_{sp}(R',P,\boldsymbol{W})+R'-R\right], \tag{B.8}$$

which implies, using (B.7), that

$$\min_{\boldsymbol{W}\in\Lambda} E_r(R,P,\boldsymbol{W})=E_r(R,P,W_1). \tag{B.9}$$

The universal coding results in [2, p. 161–173] show that, using codes of constant composition P with maximum mutual information decoding, it is possible to achieve error exponents of $E_r(R,P,\boldsymbol{W})$ for any DMC $\boldsymbol{W}$. Note that neither the codewords nor the decoder depend on $\boldsymbol{W}$ here. For the compound channel Λ, it is easy to see that, using codes based on the previous constant composition codes, it is possible to achieve an error exponent of

$$E_r(R,\Lambda)=\max_P\min_{\boldsymbol{W}\in\Lambda} E_r(R,P,\boldsymbol{W}),$$

which equals $E_r(R,W_1)$, maximizing both sides of (B.9) over P. The only information required by the encoder here is the input distribution $P^*(R)$ that achieves this maximum. This supplies a lower bound to the reliability function $E(R,\Lambda)$ of the compound channel Λ. It is also shown in [2, p. 173] that the reliability function is overbounded by the sphere packing exponent function for the compound channel Λ, given by

$$E_{sp}(R,\Lambda)=\max_{P}\min_{W\in\Lambda}E_{sP}(R,P,W),$$

which equals $E_{sp}(R,W_1)$, maximizing both sides of (B.7) over P. Thus, the reliability function $E(R,\Lambda)$ is bounded as follows:

$$E_r(R,W_1)\le E(R,\Lambda)\le E_{sp}(R,W_1).$$

For $R\ge R_{cr}(W_1)$, we have

$$E_{sp}(R,W_1)=E_r(R,W_1)=E(R,W_1),$$

which implies that $E(R,\Lambda)=E(R,W_1)$ for such rates. Thus, the proof of the second half of the proposition will be complete once we have proven (B.5) and (B.6).

To this end, consider any $W\in\Lambda$. Let V be another single input, N-output DMC such that $I(P,V)\le R$. If V_1 is the transition probability of the first subchannel of V, we know that $I(P,V_1)\le I(P,V)\le R$ (arguing as earlier, when (B.3) was proved). Also,

$$D(V\|W|P)\ge D(V_1\|W_1|P). \tag{B.10}$$

To see this, consider

$$\begin{aligned}D(V\|W|P)&=\sum_x P(x)\sum_{y_1,\cdots,y_N}V(y_1,\cdots,y_N|x)\\&\quad\cdot\log\frac{V(y_1,\cdots,y_N|x)}{W(y_1,\cdots,y_N|x)}\\&=\sum_x P(x)\sum_{y_1,\cdots,y_N}V(y_1,\cdots,y_N|x)\\&\quad\cdot\log\left(\frac{V_1(y_1|x)V'(y_2,\cdots,y_N|x,y_1)}{W_1(y_1|x)W'(y_2,\cdots,y_N|x,y_1)}\right)\\&=\sum_x P(x)\sum_{y_1}V_1(y_1|x)\log\frac{V_1(y_1|x)}{W_1(y_1|x)}\\&\quad+\sum_x P(x)\sum_{y_1}V_1(y_1|x)\\&\quad\cdot\sum_{y_2,\cdots,y_N}V'(y_2,\cdots,y_N|x,y_1)\\&\quad\cdot\log\frac{V'(y_2,\cdots,y_N|x,y_1)}{W'(y_2,\cdots,y_N|x,y_1)},\end{aligned}$$

where V',W' denote the conditional probability distributions of $(Y_2,\cdots,Y_N)$, given (X,Y_1), for the DMCs V and W, respectively. Hence,

$$D(V\|W|P)=D(V_1\|W_1|P)+D(V'\|W'|PV_1), \tag{B.11}$$

which yields (B.10), since the second term on the right hand side is nonnegative. Thus,

$$\begin{aligned}E_{sp}(R,P,W)&=\min_{\{I(P,V)\le R\}}D(V\|W|P)\\&\ge\min_{\{V_1:\,I(P,V)\le R\}}D(V_1\|W_1|P).\end{aligned}$$

We will show that the extreme right-hand side equals

$$E_{sp}(R,P,W_1)=\min_{\{V:\,I(P,V)\le R\}}D(V\|W_1|P),$$

thus proving (B.5). To show this, it suffices to prove that

$$\{V:\,I(P,V)\le R\}=\{V_1:\,I(P,V)\le R\}.$$

Since $I(P,V)\ge I(P,V_1)$, it is clear that the right-hand side is contained in the left-hand side. To show the reverse containment, let V be an $M\times M$ transition probability matrix such that $I(P,V)\le R$. Define V by

$$V(y_1,\cdots,y_N|x)=V(y_1|x)Q(y_2,\cdots,y_N|y_1), \tag{B.12}$$

where Q is an arbitrary stochastic matrix. Then $I(P,V)=I(P,V)\le R$, and the transition probability matrix V_1 of the DMC V is V, showing the reverse containment. This completes the proof of (B.5).

To show (B.6), let W be as in (B.4). Let V be such that $I(P,V)\le R$, and let V be as in (B.12), where the stochastic matrices Q are the same for both V and W. Then

$$\begin{aligned}V'(y_2,\cdots,y_N|x,y_1)&=W'(y_2,\cdots,y_N|x,y_1)\\&=Q(y_2,\cdots,y_N|y_1),\end{aligned}$$

which yields $D(V'\|W'|PV_1)=0$. Thus, from (B.11), we have $D(V\|W|P)=D(V\|W_1|P)$. Also, we have

$$I(P,V)=I(P,V)=R,$$

so that

$$E_{sp}(R,P,W)\le D(V\|W|P)=D(V\|W_1|P).$$

Minimizing the extreme right-hand side over $\{V:\,I(P,V)\le R\}$, we have $E_{sp}(R,P,W)\le E_{sp}(R,P,W_1)$ for W as in (B.4). This proves (B.6), and completes the proof of the second half of the proposition. □

Before concluding this Appendix, we point out that the worst-case (in terms of both capacity and error exponents) $W\in\Lambda$ is as in (B.4), which corresponds to the outputs $(Y_2,\cdots,Y_N)$ of the worse subchannels being a physically degraded version of the output Y_1 of the best subchannel. In this case, it is not possible for the performance of the universal receiver to be better than that of the best subchannel. What these results show is that the performance of the universal receiver is no worse than that of the best subchannel, despite the fact that the identity of the best subchannel is not known.

References

[1] P. Billingsley, *Convergence of Probability Measures.* New York: Wiley, 1968.

[2] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic Press, 1981.

[3] N. Dunford and J. T. Schwartz, *Linear Operators, Part 1: General Theory.* New York: Wiley, 1957.

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

[5] R. J. McEliece and L. Swanson, "On the decoder error probability for Reed–Solomon codes," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 5, pp. 701–703, Sept. 1986.

[6] C. E. Shannon, "A note on the partial ordering for communication channels," *Inform. Contr.*, vol. 1, pp. 390–397, 1958.

[7] J. Wolfowitz, *Coding Theorems of Information Theory.* New York: Springer-Verlag, 1975.