The Capacity of Discrete-Time Memoryless Rayleigh-Fading Channels

Ibrahim C. Abou-Faycal, Student Member, IEEE, Mitchell D. Trott, Member, IEEE, and Shlomo Shamai (Shitz), Fellow, IEEE

Abstract—We consider transmission over a discrete-time Rayleigh fading channel, in which successive symbols face independent fading, and where neither the transmitter nor the receiver has channel state information. Subject to an average power constraint, we study the capacity-achieving distribution of this channel and prove it to be discrete with a finite number of mass points, one of them located at the origin. We numerically compute the capacity and the corresponding optimal distribution as a function of the signal-to-noise ratio (SNR). The behavior of the channel at low SNR is studied and finally a comparison is drawn with the ideal additive white Gaussian noise channel.

Index Terms—Capacity, fading channels, memoryless fading, Rayleigh fading, time-varying channels.

I. INTRODUCTION

CHANNELS exhibiting fading and dispersion are often used for communication purposes, especially on a wireless medium. Perhaps the best known examples of such channels are radio links. The fading may be modeled as Rayleigh-distributed whenever there are a large number of independent scatterers and no line-of-sight path. One would like to understand how these time-varying channels differ from time-invariant ones, and perhaps obtain some qualitative results, for example, a characterization of the fading conditions under which capacity is close to the capacity of a Gaussian channel.

In this paper we study the memoryless discrete-time Rayleigh-fading channel

$$V_i = A_i U_i + N_i$$

(1)

where $U_i$ is the channel input, $V_i$ is the output, and $A_i$ and $N_i$ are independent complex circular Gaussian random variables with mean zero and variance $\sigma_A^2$ and $\sigma_N^2$, respectively. Equivalently, the amplitude of the fading coefficient $A_i$ is Rayleigh-distributed and its phase is uniform. The discrete time index is designed by $i$ and we assume that $\{A_i\}$ and $\{N_i\}$ are sequences of independent and identically distributed (i.i.d.) random variables. The input $U_i$ is average power limited: $E[U_i^2] \leq P$.

Neither the transmitter nor the receiver knows the value of $A_i$ or $N_i$, but both are assumed to know their statistics exactly.

Perhaps surprisingly, this basic channel is not as well understood as the discrete-time additive white Gaussian noise channel. In a 1969 technical report [1], Richters conjectured that the capacity-achieving input distribution for this channel is discrete, a rather unexpected result for a continuous-alphabet channel under an average power constraint. The conjecture was motivated by analytical arguments but not rigorously proved. The main result of the present paper is a proof of Richters’ conjecture. As is the case for the peak-amplitude-constrained Gaussian channels studied by Smith [2] and later by Shamai and Bar-David [4], the input distribution that achieves capacity for the fading channel is discrete with a finite number of mass points. The same result holds for the $L$th-order Rayleigh diversity channel with independent branches, as proved in Appendix III. Another relevant reference to this study is Telatar’s work [5] on the capacity and error exponent of this channel for infinite bandwidth.

The model (1) is appropriate in a number of scenarios. For example, a memoryless fading model is reasonable when a narrow-band signal is hopped rapidly over a large set of frequencies, one symbol per hop. The communication security gained by fast hopping comes at the price of decreased capacity; the results developed here help quantify this tradeoff. The model is also appropriate for a slowly time-varying channel in which successive symbols are widely separated in time, as might arise when transmitting opportunistically during guard times in a packet-based system, or when using information-bearing pilot tones for both channel identification and communication. A final motivation for studying this model is to complement the results that apply when fading information is available to the receiver only [6], [7] or to both receiver and transmitter [8], [9].

Let us note, however, that understanding the memoryless fading channel is only a small step toward understanding fading channels in their most general form. In particular, symbol-rate sampling of a rapidly varying continuous-time fading channel does not lead to the model (1). If the time variation in a continuous-time channel is fast enough to cause independent fading from symbol to symbol, there will be significant variation within each symbol, and the output bandwidth will be much
larger than the input bandwidth. In such cases, the conversion from continuous to discrete time must be done with care, for example, using several samples per symbol at the channel output. We do not consider this problem here.

The outline of the paper is as follows. In Section II, we derive a simpler mathematical model for this channel. In Section III, we write the Kuhn–Tucker condition for the capacity-achieving input random variable. We prove, in Section IV, the discrete character of the optimal input, and in Section V, we prove that it has necessarily a mass point at zero. In Section VI, numerical results for the fading channel are compared to the ideal additive white Gaussian channel. Discussions and conclusions in Section VII terminate the paper, and detailed proofs are given in Appendixes I–III.

II. THE MODEL

Since the channel (1) is stationary and memoryless, the capacity achieving statistics of $\{U_t\}$ are also memoryless (i.i.d.), and hence with no loss of generality we suppress the time index.

Conditioned on the input $U$, the output $V$ of the channel (1) is a complex circular Gaussian random variable (since $A$ and $N$ are independent, and consequently jointly Gaussian) with density

$$p_{W}(v | u) = \frac{1}{\pi (\sigma_A^2 |v|^2 + \sigma_N^2)} \exp \left[ -\frac{|v|^2}{\sigma_A^2 |v|^2 + \sigma_N^2} \right].$$  

(2)

Because the phase of the fading parameter $A$ is uniform, the conditional output density involves only the squared amplitude of $V$. Conditioned on the input, the sufficient statistic $|V|^2$ has a central chi-square distribution with two degrees of freedom

$$p_{|V|^2}(v | u) = \frac{1}{\sigma_A^2 |v|^2 + \sigma_N^2} \exp \left[ -\frac{\alpha}{\sigma_A^2 |v|^2 + \sigma_N^2} \right].$$  

(3)

Letting $Y = |V|^2 / \sigma_A^2$ and $X = |U| / \sigma_A$, we obtain an equivalent channel with nonnegative input $X$, nonnegative output $Y$, transition probability

$$p(y | x) = \frac{1}{1+x^2} \exp \left( -\frac{y}{1+x^2} \right)$$  

(4)

and average power constraint $E[X^2] \leq a$, where $a = P \sigma_A^2 / \sigma_N^2$.

In other words, the output density is conditionally exponential with mean modulated by the channel input.

Since $x$ appears in the above equations only via its square, it is convenient to make the invertible change of variables $s = 1/(1+x^2)$, so that

$$p(y | s) = s e^{-ys}, \quad s \in (0, 1]$$  

(5)

with the constraint $E[1/s - 1] \leq a$. Proving that the optimal input $X^*$ is a discrete random variable is equivalent to proving that $S^* = 1/(1+X^*2)$ is discrete.

III. THE KUHN–TUCKER CONDITION

The capacity of an input-constrained memoryless channel is the supremum

$$C = \sup_{F \in \mathcal{F}} \int \int p(y | x) \ln \frac{p(y | x)}{p(y | F)} \, dy \, dF(x)$$  

(6)

of the mutual information between the input $X$ and output $Y$ over all input distribution functions $F$ that meet the constraint $F$, where

$$p(y; F) = \int p(y | x) \, dF(x)$$  

(7)

is the marginal output density induced by $F$. The input may include a mixture of discrete and continuous parts. However, (6) applies only when there exists a conditional density $p(y | x)$ for the channel output given the input, a requirement automatically met for channels with absolutely continuous additive noise.

For a channel with a continuous alphabet, the supremum in (6) need not be achievable. A sufficient condition for achievability is that there exist a topology for which i) mutual information is continuous in the input distribution function, and ii) the set of input distribution functions that meet the constraint is compact. We establish continuity and compactness for the average-power-constrained fading channel (1) in Appendix I. Using strict convexity, we further show that capacity is achieved by a unique input distribution (ignoring phase).

The maximization (6) is not readily solved using the calculus of variations, a methodology better suited to spaces of continuous functions than distribution functions. Instead, in Appendix II (following Smith [2]) we use the theory of convex optimization to show that an input random variable $X^*$ with distribution function $F^*$ achieves the capacity $C$ of an average power-limited channel if and only if there exists a $\gamma \geq 0$ such that

$$\gamma (\alpha^2 - \alpha) + C - \int p(y | x) \ln \left[ \frac{p(y | x)}{p(y; F^*)} \right] \, dy \geq 0$$  

(8)

for all $x$, with equality if $x$ is in the support of $X^*$. Equation (8) applies to any average power constrained channel for which mutual information is weak* continuous and weakly differentiable.

For channels over finite alphabets, (8) reduces to [10, Theorem 4.5.1], with the addition of a Lagrange multiplier arising from the power constraint. Similar generalizations from discrete to continuous alphabets exist for rate distortion theory. (See, for example, [11] or the summary in [12, Sec. 4.5].) A virtue of (8) is its exclusive use of densities; the input distribution function $F^*$ appears only indirectly through the output density $p(y; F^*)$ and the support of $X^*$.

Using the conditional density (5), (8) becomes

$$\gamma \left[ \frac{1}{s} - 1 - a\right] + C - \int_0^{\infty} sc^{-sy} \ln \left[ \frac{sc^{-sy}}{p(y; F^*)} \right] \, dy \geq 0$$  

(9)

for all $s \in (0, 1]$. Expanding the term inside the integral yields

$$\gamma \left[ \frac{1}{s} - 1 - a\right] + C - \ln s + 1 + \int_0^{\infty} sc^{-sy} \ln p(y; F^*) \, dy \geq 0$$  

(10)

for all $s \in (0, 1]$, with equality if $s$ is in the support of $S^*$. Equation (10) will be referred to as the “Kuhn–Tucker condition.”

IV. THE DISCRETE CHARACTER OF $X^*$

The optimal input $X^*$ must have exactly one the following properties:

1) its support contains an interval;
it is discrete, with an infinite number of mass points on some bounded interval;
3) it is discrete and infinite, but with only a finite number of mass points on any bounded interval; or
4) it is discrete with a finite number of mass points.

We do not have a direct proof that case 4) prevails. Instead, following Smith [2] and Shamai and Bar-David [4], we use the Kuhn–Tucker condition to prove that the first two cases are impossible. We then rule out the third possibility (which does not arise for the peak limited channels considered in [2], [4]) using a lower bounding argument. The existence and uniqueness of \( X^\ast \) consequently proves that \( X^\ast \) is discrete with a finite number of mass points.

A. A Positive Accumulation Point

Assume now that case 1) or 2) holds. Then the support of \( X^\ast \) includes a bounded infinite set of distinct points \( S_X \subseteq [0, A] \).

Equivalently, the support of \( S^\ast \) includes an infinite set of distinct points \( S_S \subseteq [1/(1 + A^2), 1] \). The interval \([1/(1 + A^2), 1]\) is compact, hence by the Bolzano–Weierstrass theorem the set \( S_S \) has an accumulation point in \([1/(1 + A^2), 1]\) \( \subseteq (0, 1] \).

Let us extend the left-hand term of (10) to the complex domain. Write \( p(y, F^\ast) \) as \( p(y) \), and define

\[
h(z) = \gamma \left[ \frac{1}{z} - 1 - a \right] + C - \ln z + 1 + \int_0^\infty ze^{-y \ln p(y)} dy,
\]

where \( \ln z \) is the principal branch of the logarithm [13]. Given this choice, the function \( h(z) \) is analytic over the domain \( D \) defined by \( \Re(e^z) > 0 \).

On the interval \((0, 1]\) of the real axis, the Kuhn–Tucker condition (10) states that the function \( h(z) \) is zero if \( z \) is in the support of \( S^\ast \). Therefore, \( h(z) \) is zero on \( S_S \). We have thus an analytic function over a domain \( D \) that is zero over a set having a point of accumulation in \( D \). The identity theorem [14] states that \( h(z) \) is zero over the whole domain \( D \). In other words, the Kuhn–Tucker condition (10) holds with equality for all real \( s \in (0, 1] \), and more generally for all complex \( s \in D \).

Let us examine carefully the consequences of this result.

Equation (10) can be rewritten as

\[
\int_0^{\infty} e^{-y \ln p(y)} dy = -\frac{1}{s} \left[ \gamma \left[ \frac{1}{s} - 1 - a \right] + C - \ln s + 1 \right],
\]

for all \( s \in D \). The left-hand side (LHS) is the unilateral Laplace transform of the function \( \ln p(y) \), while the right-hand side (RHS) can be recognized as the Laplace transform of \( \exp \left[ \gamma(1 + a) - C - 1 - C_E \right] \ln y \).

The uniqueness of the Laplace transform for continuous functions of bounded variations [13] implies that

\[
\ln p(y) = -\gamma y + \left[ \gamma(1 + a) - C - 1 - C_E \right] \ln y.
\]

Therefore, the only function that satisfies (12) is

\[
p(y) = K e^{\gamma y} y
\]

where

\[
K = \exp \left[ \gamma(1 + a) - C - 1 - C_E \right].
\]

But for any value of \( \gamma \), the integral over \((0, \infty)\) of (15) is infinite, hence it cannot be a probability density.

By contradiction, the original assumption on \( X^\ast \) is wrong. The only possibilities remaining are cases 3) and 4).

B. An Accumulation Point at Zero

Assume now that case 3) holds, so that \( X^\ast \) is discrete with infinitely many mass points, but with only a finite number in any bounded interval. Equivalently, assume that \( S^\ast \) has discrete support with an accumulation point only at zero. A random variable has at most a countable number of mass points, therefore, the support of \( S^\ast \) can be written as a sequence \( \{s_i\} \) converging to 0.

Let \( \Pr[S^\ast = s_i] = p_i \neq 0 \). Then the output probability density is

\[
p(y) = \sum_{i=0}^{\infty} p_i K(y/s_i) = \sum_{i=0}^{\infty} p_i s_i e^{-y s_i};
\]

which implies the obvious lower bound \( p(y) > p_i s_i e^{-y s_i} \) for all \( y \geq 0 \) and \( i = 1, 2, \ldots \), Therefore,

\[
\int_0^{\infty} s e^{-y \ln p(y)} dy > \ln (p_i s_i) - \frac{s_i}{s};
\]

for all \( i \).

Using (17), the LHS of (10) may be lower-bounded as

\[
\text{LHS} > \gamma \left[ \frac{1}{s} - 1 - a \right] + C - \ln s + 1 + \ln p_i s_i - \frac{s_i}{s};
\]

where the \( \cdot \) term applies as \( s \to 0 \) for any fixed \( i \). If \( \gamma > s_i \), then this lower bound diverges to \( \infty \) as \( s \to 0 \). But the LHS of (10) equals zero on the support of \( S^\ast \), which, by assumption, contains a point of accumulation at 0. By contradiction, \( \gamma \leq s_i \).

The above analysis is valid for all \( i \); since \( s_i \to 0 \) this implies \( \gamma \leq 0 \). The Lagrange multiplier in (10) is nonnegative (see Appendix II), hence we conclude that if the support of \( S^\ast \) has an accumulation point at zero then \( \gamma = 0 \).

A Lagrange multiplier is zero when the corresponding constraint is inactive, or, more precisely, when the sensitivity to a change in the constraint is zero. For the fading channel, it means that the power constraint is inactive, which is not sensible.

The impossibility of \( \gamma = 0 \) can be reasoned more precisely as follows. The capacity \( C(a) \) of a channel increases monotonically with the power constraint \( a \); the concavity of mutual information in the input distribution (and the linearity of the power constraint) implies that \( C(a) \) is also concave. The Lagrange multiplier \( \gamma \) corresponding to a particular capacity-achieving input distribution \( X^\ast \) with power \( E[X^\ast]^2] = a \) may be interpreted as the slope of a line tangent to \( C(a) \) at \( a \). Thus, by convexity and monotonicity, if \( \gamma = 0 \) for some power constraint \( a \), then \( C(d) = C(a) \) for all \( d \geq a \). To demonstrate the impossibility of \( \gamma = 0 \) it, therefore, suffices to find a family of input
distributions with strictly monotonically increasing mutual information.

Such a family may be constructed in any number of ways. One approach is to use a discrete uniform input distribution with \(N\) input levels located at 0, \(K, K^2, \ldots, K^{2N-2}\). By taking \(K\) sufficiently large it can be shown that the probability of error for a simple minimum-probability-of-error receiver goes to zero, hence, by Fano’s inequality, the mutual information approaches \(\ln N\).

We have assumed that \(X^*\) is discrete with an infinite number of mass points, but with only finitely many mass points in any bounded interval. We proved that this is possible only if the Lagrange multiplier in the Kuhn–Tucker condition (10) is zero. But this is impossible, consequently, the optimal distribution is discrete with a finite set of mass points.

V. The Existence of an Impulse at Zero

The next natural task is to locate the mass points of \(X^*\). A closed-form solution seems unlikely. However, in this section, we will prove by contradiction that the optimal input random variable \(X^*\) has necessarily a mass point at zero. This is not surprising: a zero-level input is “good” for the power constraint and results in the smaller variance at the output and should be preferred to other levels.

Because \(X^*\) is discrete, it has a distribution function

\[
F(x) = \sum_{i=0}^{N} p_i u(x - x_i)
\]

where \(0 \leq x_0 < x_1 < \cdots < x_N\) and \(u(x)\) is the unit step. Assume now that \(X^*\) contains no mass point at zero, i.e., \(x_0 > 0\). Let us fix the \(p_i\)’s and \(x_1, x_2, \ldots\) and move \(x_0\) downwards. Clearly, the power constraint becomes looser. Proving that the mutual information increases is therefore sufficient to prove that the original density is suboptimal. To establish this result rigorously is a matter of algebra.

For a discrete input \(X\), the mutual information between \(X\) and \(Y\) is

\[
I(X; Y) = \sum_{i=0}^{N} p_i \int_0^{\infty} p(y|x_i) \ln \left[ \frac{p(y|x_i)}{\sum_j p_j p(y|x_j)} \right] dy. \tag{20}
\]

Differentiating \(I(X; Y)\) with respect to \(x_0\) yields

\[
\frac{\partial}{\partial x_0} I(X; Y) = p_0 \int_0^{\infty} \frac{\partial}{\partial x_0} p(y|x_0) \ln \left[ \frac{p(y|x_0)}{\sum_j p_j p(y|x_j)} \right] dy. \tag{21}
\]

To simplify (21) further, differentiate (4)

\[
\frac{\partial}{\partial x_0} p(y|x_0) = 2x_0 \left[ \frac{1}{1 + x_0^2} + \frac{y}{(1 + x_0^2)^2} \right] p(y|x_0)
\]

and define \(f(y) = \ln[p(y|x_0)/p(y)]\). Then

\[
\frac{\partial}{\partial x_0} I(X; Y) = \frac{2x_0 p_0}{1 + x_0^2} \int_0^{\infty} \left[y - (1 + x_0^2) f(y)\right] dy. \tag{22}
\]

Lemma 1: Let \(p(y)\) be a probability density function with mean \(m\). If \(f(y)\) is strictly monotonically decreasing then

\[
\int (y - m) f(y) f(y) dy < 0, \tag{24}
\]

Proof: The function \(f(y)\) is strictly decreasing, hence \((y - m)(f(m) - f(y))\) is a product of two positive terms for \(y > m\) and is a product of two negative terms for \(y < m\). Thus,

\[
(y - m)f(y) < (y - m)f(m) \quad \text{for } y \neq m,
\]

and

\[
\int (y - m) f(y) f(y) dy < \int (y - m)f(y) f(m) dy = 0 \tag{25}
\]

because \(m\) is the mean of \(p(y)\).

The optimal input is discrete with a finite set of mass points needed as a function of the power constraint.

VI. Numerical Results

A. Capacity Curves

Having characterized the optimal probability distribution, we can now compute numerically the capacity of an i.i.d. Rayleigh-fading channel as a function of the power constraint. This was attempted previously by Richters [1], but more powerful numerical techniques and computational tools are now available and more precise results may be obtained. Moreover, the conjectured capacity-achieving distribution was not provided in Richters’ work, and how this distribution varies as a function of the power constraint at low and high signal-to-noise ratios (SNRs) is of some interest.

To find the capacity \(C(a)\) as a function of the power constraint \(a\), we introduce a Lagrange multiplier \(\gamma\) and maximize the functional \(I(X; Y) - \gamma E[X^2]\) over all input random variables \(X\). The multiplier \(\gamma\) is the slope of a line tangent to the curve \(C(a)\) at \(a\), so by varying \(\gamma\) between 0 and \(C'(0)\) (which equals 1 when the fading and noise variance are normalized to 1) the full curve \(C(a)\) may be found.

The optimal input is discrete with a finite set of mass points, hence for each \(\gamma\) we must maximize over the number of mass points \(N\), their probabilities \(p_1, \ldots, p_{N-1}\), and their locations \(x_1 < x_2 < \cdots < x_{N-1}\), subject to the constraints \(0 < x_1 < x_3 < \cdots < x_{N-1}, 0 < p_i\) for all \(i\), and \(\sum_{i=0}^{N-1} p_i = 1\). Unfortunately, though the optimization problem is convex over the space of all input distribution function, there is no reason to believe it remains convex when parameterized by the mass point probabilities and locations. Furthermore, we have no bounds on the number of mass points needed as a function of the power constraint.
As a practical matter, these potential difficulties do not arise. Because mutual information is continuous and strictly concave in the input distribution, the optimal input distribution function changes continuously (in the weak* topology) with the SNR \( \alpha \). Also, we have found empirically that two mass points are optimal for low SNR and the required number of mass points increases monotonically with SNR. Strong evidence for these conclusions comes from the Kuhn–Tucker condition (10), which, being necessary and sufficient for optimality, allows us to establish (up to the resolution of the numerical algorithms) that a local maximum found by a descent method is in fact global.

To apply the Kuhn–Tucker test to a postulated \( X^* \) and \( \gamma \), we first compute \( p(\gamma, F^*) \) and \( \alpha = E[(X^*)^2] \), then plot the LHS of (8) as a function of \( \gamma \). The resulting graph must be nonnegative and must touch zero at the atoms of \( X^* \), as in Fig. 1.

The curve \( C(\alpha) \) and the mass point locations and probabilities were computed using a gradient descent method, together with Gauss–Laguerre quadrature to evaluate the necessary integrals. Projected gradients were used to keep the mass point probabilities positive. An alternative optimization method, using a quantized version of Arimoto–Blahut, was too slow to be useful.

**Low SNR:** For low values of the SNR \( \alpha \) the capacity-achieving input distribution has only two mass points, and, therefore, amounts to “on–off” keying. One point is always located at zero; the other is easily found with standard one-dimensional optimization techniques, because the power constraint determines the mass point probability as a function of its location. The Kuhn–Tucker condition verifies that the optimized two-point distribution achieves capacity. The optimal mass point locations and probabilities are plotted as a function of \( \alpha \) in Fig. 2.

As \( \alpha \) decreases, the probability of the nonzero mass point approaches zero, while its amplitude increases, albeit quite slowly. That the amplitude increases to infinity as \( \alpha → 0 \) is proved by Gallager [10, Theorem 8.6.1].

**Higher SNR:** As SNR increases, does a new mass point appear apart from the others, or does an existing mass point split in two? Fix [15] considered an analogous problem in rate distortion theory, and concluded from plots of a quantity comparable to Fig. 1 that both possibilities arise. The question of splitting for rate distortion theory was also studied (but not solved) by Rose [16], whose results can also be shown by techniques similar to those in [2].

From a numerical standpoint this question is quite challenging, as mutual information is insensitive to the location of a new mass point with near-zero probability. Some insight can be gained from the Kuhn–Tucker condition, which for our problem suggests that mass points never split and that new mass points appear initially at \( \infty \). For our problem, when the SNR reaches a level where a new mass point is needed, the \( x → \infty \) asymptote of the quantity plotted in Fig. 1 switches suddenly from \( +\infty \) to \( -\infty \). We conjecture that the analysis that applies to the \( \alpha → 0 \) case can be modified to prove that all new mass points enter at \( \infty \).

Fig. 3 shows the capacity \( C(\alpha) \) (in nats per channel use) of the i.i.d. Rayleigh-fading channel as a function of the power constraint \( \alpha \). Figs. 4 and 5 show the locations and probabilities of the mass points in the capacity-achieving input distribution. The mass point with lowest probability has the highest amplitude, the mass point with the second lowest probability has the second highest amplitude, and so on. The dashed segments in the figures are conjectured curves; numerical optimization became unstable when the lowest mass point probability dropped below \( 10^{-4} \).

Interestingly, the location of a new mass point initially moves downward as \( \alpha \) increases, then moves upward. This peculiar behavior has little engineering consequence, as the probability of the mass point remains negligible until its location begins its upward trend.

The sensitivity of mutual information to the exact number and location of the mass points appears to be small. Fig. 6, for example, shows the maximum mutual information achievable when a distribution with two mass points is used in the SNR region where three or four mass points are optimal. At an SNR of 10.5 there is only a 4% gap to capacity. Though not shown in the figure, moving the nonzero mass point 10% from its optimal location yields a mutual information that is only 0.5% lower.

Finally, a reference that is worth mentioning is Taricco and Elia’s work [18] which provides bounds that reflect the asymptotic low and high SNR behavior.

**B. Comparison with the Ideal Gaussian Channel**

In Fig. 7, the ratio of the capacity of the ideal additive white Gaussian noise channel to the capacity of the i.i.d. Rayleigh-fading channel is plotted as a function of the power constraint, where the channels are normalized to have the same SNR at the receiver. The graph suggests that the capacity of the fading channel approaches the capacity of the Gaussian channel as \( \alpha → 0 \). This is indeed the case, as is shown for a continuous-time model in [10, Theorem 8.6.1] and for a discrete-time model in [17, Example 3]. Fig. 7 illustrates, however, that the asymptote is approached quite slowly.

At low SNR, on–off keying with a low duty cycle is optimal, and a capacity-achieving codebook for the fading channel resembles pulse-position modulation. Unlike a Gaussian channel, where energy is spread uniformly over all degrees of freedom, for the fading channel energy becomes more concentrated as...
bandwidth increases. At moderate SNR, the optimal input distribution for the fading channel resembles a uniform distribution over uniformly spaced levels. At high SNR, the loss in performance due to fading grows rapidly.

C. Comparison with the Fading Channel with Side Information Given to the Receiver

The case when fading information is available to the receiver was studied by Ericson [6], and later by Ozarow, Shamai, and Wyner [7]. The capacity of the channel with perfect channel state information (CSI) is

$$ C = -\exp\left(\frac{1}{\text{SNR}}\right) E_\nu\left(-\frac{1}{\text{SNR}}\right) $$

(27)
VII. SUMMARY AND DISCUSSION

Motivated by previous work done by Smith [2] and Shamai and Bar-David [4], we have proven what Richters [1] conjectured in his original report—that the capacity-achieving distribution for the discrete-time memoryless Rayleigh-fading channel is discrete with a finite set of mass points. The main results also hold for $L$th-order diversity with independent branches (see Appendix III).

An immediate direction for future work is the i.i.d. Ricean fading channel, modeled by giving the fading variable $A$ in (1) a nonzero mean. The Ricean model is appropriate when there is a line-of-sight path from transmitter to receiver, or when the receiver (and possibly the transmitter) have side information about the fading. We conjecture that the optimal input will again be discrete. Unlike the Rayleigh channel, the Ricean channel has two parameters: the SNR and the ratio of the fading standard deviation to the fading mean. The capacity of this channel when the input is restricted to be Gaussian has been studied by Zhou, Mei, Xu, and Yao [19]. A block constant fading was assumed, and no side information was available to either the transmitter or receiver. While the result gives some indication for the effect of the direct path, the optimal input distribution might not be Gaussian.

Good, easily computed bounds on the capacity of the i.i.d. Rayleigh-fading channel at low and high SNR would be useful for engineering design, as the exact capacity is tedious to compute. Some general guidelines on signal set design would also be helpful. The sensitivity of mutual information to the input distribution function is generally small, and the fading model is likely to be only marginally accurate in many scenarios. Is there a simple rule of thumb for approximately selecting the number of mass points, their locations, and their probabilities as a function of SNR?

Complementary to the case where the receiver has CSI is the case where the CSI is available at the transmitter only. We expect difficulties here, as transmission strategies rather than input distributions should be considered, as is concluded by extrapolating the results of Shannon [20] to the continuous state space.

An important and challenging extension is to generalize the study to non-i.i.d. Rayleigh-fading channels, modeled for example by a Gauss–Markov process. A classical receiver attempts to track the channel variations when the fading coefficients are correlated in time. Should an information-theoretic receiver do the same? We expect that the answer is effectively “yes” at high SNR with slow fading, and “no” when the fading is fast. The desired results will be difficult to obtain, as the optimal input process need not be i.i.d. in general. Some work has been done in this direction by Marzeta and Hochwald [21] who considered an $M$-transmitters and $N$-receivers channel, whose fading is constant for $T$ symbols. The capacity of this channel in some simple cases was computed numerically, and the conjecture of a discrete optimal distribution was found to be accurate.

A most challenging problem is to combine fading memory and input memory. Underwater acoustic channels, for example, combine rapid time variation with long intersymbol interference. Little is known about the capacity of such channels or how to achieve it.

APPENDIX I
THE OPTIMIZATION PROBLEM

In this appendix, we establish the existence and uniqueness of the capacity-achieving input distribution for the average power-limited Rayleigh-fading channel. Existence is automatic for finite-alphabet channels but not for continuous-alphabet ones. Uniqueness follows from a particular parameterization of the input space that disregards phase. The structure of the existence proof below follows Smith [2], [3], but the details are different; the fading channel has both multiplicative and additive noise, and compactness is more difficult to prove for an average power constraint than for a peak power constraint.

We establish existence using topological arguments found in optimization theory and probability theory. In optimization theory, one starts by defining the real normed linear space $X$ of all bounded continuous functions on $H$. The dual $X^*$ of $X$ includes the set of all probability measures. Optimization results are then obtained using the weak* topology on $X^*$ [22, Sec. 5.10]. In probability theory, one starts with the set of probability measures, and defines weak convergence—which is actually the weak* convergence in $X^*$—on this set. Next, a metric that metrizes weak convergence is defined (e.g., the Lévy metric [23, Sec. III.7]) and optimization is done in the metric topology.
Since the topology obtained in both theories is the same, we freely combine the approaches.

Existence and uniqueness follow from the following basic theorem of optimization. The remainder of the appendix establishes that the conditions of the theorem are met.

**Theorem 1:** If \( J \) is a real-valued, weak* continuous functional on a weak* compact set \( \Omega \subseteq X^* \), then \( J \) achieves its maximum on \( \Omega \). If furthermore \( \Omega \) is convex, and \( J \) is strictly concave, then the maximum

\[
C = \max_{F \in \Omega} J(F)
\]

is achieved by a unique \( F_0 \) in \( \Omega \).

**Proof:** The first statement is given in [22, Sec. 5.10]. The second follows from the definition of strict concavity: if the maximum were achieved at two points, then the evaluation of \( J \) along their convex combination would exceed that maximum, a contradiction. \( \square \)

**A. The Set \( \Omega \) is Convex and Compact**

Let \( F \) denote the set of all distribution functions, and let \( \Omega \subseteq F \) be the distribution functions of nonnegative random variables with a second moment constraint. That is, \( \Omega \) is the set of distribution functions \( F \) such that

\[
F(0^-) = 0
\]

and

\[
\int_0^\infty x^2 \, dF(x) \leq \alpha.
\]

The set \( \Omega \) is convex. Indeed, for any \( F_1, F_2 \in \Omega \) and \( \lambda \in [0, 1] \), the convex combination \( F = \lambda F_1 + (1 - \lambda) F_2 \) is a distribution function (nondecreasing, right continuous, \( F(\infty) = 0 \) and \( F(+\infty) = 1 \)) and is in \( \Omega \), because \( F(0^-) = 0 \) and because the second moment of \( F \) is the linear combination of the second moments of \( F_1 \) and \( F_2 \).

To prove that \( \Omega \) is weak* compact we first show that the average power constraint makes \( \Omega \) tight. Then, because the weak* topology on distribution functions is metrizable, Prokhorov’s Theorem [23, Sec. III.2] implies that \( \Omega \) is relatively compact. That is, for every sequence \( \{F_n\} \) of distribution functions in \( \Omega \) we can find a subsequence \( \{F_{n_k}\} \) and a distribution function \( F^* \), not necessarily in \( \Omega \), such that \( F_{n_k} \xrightarrow{w^*} F^* \). Finally, we prove that \( F^* \) is in \( \Omega \), establishing that \( \Omega \) is sequentially compact and hence compact (again because the topology is metrizable).

The set \( \Omega \) is tight if, for every \( \epsilon > 0 \), there is a \( K > 0 \) such that

\[
\sup_{F \in \Omega} [F(\infty) + (1 - F(K))] \leq \epsilon.
\]

This follows immediately from Markov’s inequality

\[
F(\infty) + (1 - F(K)) \leq E[X^2] / K^2 = \alpha / K^2
\]

for all \( F \in \Omega \), where \( X \) is a (nonnegative) random variable with distribution function \( F \).

It remains to show that the limiting distribution function \( F^* \) is in \( \Omega \). That \( F^*(0^-) = 0 \) follows from [23, Theorem 1, Sec. III.1]. To show that \( F^* \) satisfies the second moment constraint (31), observe that \( x^2 \) is continuous and bounded below, hence

\[
\int x^2 \, dF^n(x) \leq \liminf_{n \to \infty} \int x^2 \, dF_n(x) \leq \alpha.
\]

**B. Mutual Information is Continuous and Strictly Concave**

The weak* topology on distribution functions is metrizable, hence weak* continuity of a function \( f \) is equivalent to

\[
F_n \xrightarrow{w^*} F \implies f(F_n) \to f(F).
\]

We prove that this property is satisfied by the mutual information for the Rayleigh-fading channel. For simplicity, denote the mutual information resulting from a specific input distribution function \( F \) by

\[
I(F) = h_Y(F) - h_{Y|X}(F)
\]

where

\[
h_Y(F) = -\int_0^\infty p(y; F) \ln p(y; F) \, dy
\]

and

\[
h_{Y|X}(F) = -\int \int p(y|x) \ln p(y|x) \, dy \, df(x).
\]

We first show that \( h_Y(F) \) is weak* continuous. Let \( F_n \xrightarrow{w^*} F \). We show that \( h_Y(F_n) \to h_Y(F) \) by establishing the following chain of equalities:

\[
\lim_n h_Y(F_n) = -\lim_n \int p(y; F_n) \ln p(y; F_n) \, dy = -\int p(y; F_n) \ln p(y; F_n) \, dy = -\int p(y; F) \ln p(y; F) \, dy = h_Y(F).
\]

Equations (40) and (43) are definitions. To establish (42), note that \( p(y|x) \), defined in (4), is a bounded continuous function of \( x \). Hence, by definition of the weak* topology

\[
p(y; F) = \int p(y|x) \, df(x)
\]

is a continuous function of \( F \) for all \( y \geq 0 \). Now, \( x \ln x \) is continuous, hence \( p(y; F) \ln p(y; F) \) is also continuous in \( F \). By (36)

\[
\lim_{n \to \infty} p(y; F_n) \ln p(y; F_n) = p(y; F) \ln p(y; F).
\]

To establish the interchange of the integral and limit in (41), it suffices, by the Lebesgue dominated convergence theorem, to find an integrable function \( g \) such that

\[
|p(y; F_n) \ln p(y; F_n)| \leq g(y)
\]

for all \( F_n \).
Let \( g(y) = \min \{ 1, 1/y^2 \} \). Note that \(|x \ln x|^2 \leq x\) for \( x \geq 0 \), and, by inspection of (4), \( p(y|x) \leq g^2(y) \) for all \( y \) and \( x \). Therefore, for all \( y \) and \( F_n \),

\[
\left[ p(y; F_n) \ln p(y; F_n) \right]^2 \leq p(y; F_n) \leq \int p(y|F_n) dF_n(x) 
\]

(46)

(47)

(48)

(49)

Since \( g(y) \) is integrable, (41) is established, and \( h_Y(F) \) is weak* continuous.

We now show that \( h_{Y|X}(F) \) is weak* continuous in \( F \). (This result is obvious for channels with purely additive noise, as in [2], [4].)

The innermost integral in (39) is equal to \( 1 + \ln(1 + x^2) \). Let \( F_n \to F \). We must show that

\[
\int \ln(1 + x^2) dF_n \to \int \ln(1 + x^2) dF. 
\]

(50)

This follows if the function

\[
g(x) = \begin{cases} 
0, & \text{for } x < 0 \\
\ln(1 + x^2), & \text{for } x \geq 0 
\end{cases} 
\]

is uniformly integrable in \( F_n \) [25, Sec. 11.4], that is, if \( g \) is continuous, if \( \int |g| dF_n < \infty \), and if

\[
\lim_{b \to \infty} \int_b^\infty \ln(1 + x^2) dF_n(x) = 0 
\]

(51)

uniformly in \( n \).

The function \( g \) is clearly continuous. Moreover, since \( \ln(1 + x^2) \leq x^2 \) for \( x \geq 0 \)

\[
\int |g| dF_n = \int_0^\infty \ln(1 + x^2) dF_n(x) \leq a
\]

for every \( n \). To verify (51) note that \( \ln(1 + x^2) \leq x \) for \( x \geq 0 \), therefore, for \( b \geq 1 \)

\[
\int_b^\infty \ln(1 + x^2) dF_n(x) \leq \int_b^\infty x dF_n(x) 
\]

(52)

\[
\leq \frac{1}{b} \int_b^\infty x^2 dF_n(x) 
\]

(53)

\[
\leq \frac{a}{b}
\]

(54)

a bound independent of \( n \) that converges to 0 as \( b \to \infty \). Hence, \( g \) is uniformly integrable in \( F_n \), and \( h_{Y|X}(F) \) is a weak* continuous function of \( F \).

Being the difference of two weak* continuous functions, the mutual information \( I \) is consequently a weak* continuous function of \( F \).

We now prove that \( I(F) \) is a strictly concave function over \( \Omega \), and in fact over \( F \).

**Lemma 2:** The operator that associates to a given \( F \) an output density \( p(y) \) is injective.

**Proof:** Equation (4) describes a multiplicative channel

\[
y = (1 + x^2) \omega 
\]

(55)

where \( \omega \) is independent of \( x \) and has a probability density function \( p(\omega) = e^{-\omega} \). To simplify notation, define \( z = (1 + x^2) \).

Since we have an invertible relationship between \( x \) and \( z \), when \( x \) is nonnegative, to prove injectivity it is sufficient to show that

\[
p(y; F_z) = p(y; F_z^*) \Rightarrow F_z = F_z^*.
\]

Assume \( p(y; F_z) = p(y; F_z^*) \). Then \( y = z \omega \) and \( y' = z' \omega \) are equal in distribution. Therefore,

\[
\ln(y) = \ln(z) + \ln(\omega)
\]

and

\[
\ln(y') = \ln(z') + \ln(\omega)
\]

are equal in distribution. Equivalently, \( \ln(y) \) and \( \ln(y') \) have equal characteristic functions

\[
\Phi_{\ln(y)}(f) \Phi_{\ln(\omega)}(f) = \Phi_{\ln(z)}(f) \Phi_{\ln(\omega)}(f).
\]

It can be shown that \( \Phi_{\ln(\omega)}(f) \) is well-defined and analytic for complex values of \( f \) on the band \(-1 < \Im(f) < 1 \) [13], and, therefore, \( \Phi_{\ln(\omega)}(f) \) has isolated zeros on the real axis. This implies that \( \Phi_{\ln(\omega)}(f) = \Phi_{\ln(z)}(f) \) everywhere except at potentially isolated points, but given that characteristic functions are continuous, they are equal everywhere. Thus, \( \ln(z) \) and \( \ln(z') \) are equal in distribution and hence so are \( z \) and \( z' \).

Recall that \( I(F) = h_Y(F) - h_{Y|X}(F) \). The function \( h_Y \) is clearly a strictly concave function of \( p(y) \), and since \( p(y) \) is an injective linear function of \( F \), \( h_Y \) is a strictly concave function of \( F \). Since the other term \( h_{Y|X} \) is linear in \( F \), \( I \) is a strictly concave function of \( F \).

**APPENDIX II**

**THE KUHN–TUCKER THEOREM**

The result of this appendix can be obtained using either the local theory of constrained optimization (the generalized Kuhn–Tucker Theorem), or the global one. For simplicity, we have chosen the latter, while the end result can still be called a Kuhn–Tucker condition.

The following results closely parallel Smith [2], and use the idea of weak differentiability defined by Smith [2].

**A. The Lagrangian Theorem**

**Theorem 2** [22, Sec. 8.3]: Let \( X \) be a linear vector space, \( Z \) a normed space, \( F \) a convex subset of \( X \), and \( P \) the positive cone in \( Z \). Assume that \( P \) contains an interior point.

Let \( f \) be a real-valued concave functional on \( F \) and \( g \) a convex mapping from \( F \) to \( Z \). Assume the existence of a point \( F_1 \in F \) for which \( g(F_1) < 0 \).

Let

\[
C = \sup_{F_1 \in F, g(F_1) < 0} f(F_1)
\]

(56)
and assume $C$ is finite. Then there is an element $z_0^* \geq 0$ in $\mathbb{Z}^*$ such that
\[
C = \sup_{F \in \mathcal{F}} \{J(F) - \langle g(F), z_0^* \rangle\}. 
\tag{57}
\]
Furthermore, if the supremum is achieved in (56) at $F_0$, it is achieved by $F_0$ in (57) and
\[
\langle g(F_0), z_0^* \rangle = 0. 
\tag{58}
\]
To apply Theorem 2 to our problem, the functional $f$ is the mutual information $I(F)$ and $\mathcal{F}$ the set of distribution functions of nonnegative random variables. It is straightforward to establish that $\mathcal{F}$ is convex. We have established in Appendix I that $I(F)$ is (strictly) concave over $\mathcal{F}$. Define the mapping
\[
g(F) = \int_0^\infty x^2 dF(x) - a \tag{59}\]
from $\mathcal{F}$ to $R$, where $a > 0$. The mapping $g(F)$ is linear and hence convex. The positive cone in $R$ has an interior point, and if $F_1$ is the unit step, $g(F_1) < 0$.

By Theorem 2, there exists $\gamma > 0$ in $R$ such that
\[
C = \sup_{F \in \mathcal{F}} \{I(F) - \gamma g(F)\}. \tag{60}
\]
Moreover, since capacity is achieved for some $F_0$, the supremum is achieved in (60) by $F_0$ and
\[
\gamma g(F_0) = 0. \tag{61}\]

### B. Weak Differentiability in a Convex Space

**Definition:** Let $f$ be a functional on a convex set $\mathcal{F}$. Let $F_0$ be a fixed element of $\mathcal{F}$, and $\theta$ a number in $[0, 1]$. Suppose there exists a map $f_{F_0}^{\theta}$, $\mathcal{F} \to R$ such that
\[
f_{F_0}^{\theta}(F) = \lim_{\theta \to 0} \frac{f((1-\theta)F_0 + \theta F) - f(F_0)}{\theta} \quad \forall F \in \mathcal{F}. \tag{62}\]
Then $f$ is said to be weakly differentiable in $\mathcal{F}$ at $F_0$, and $f_{F_0}^{\theta}$ is the weak derivative in $\mathcal{F}$ at $F_0$. If $f$ is weakly differentiable in $\mathcal{F}$ at $F_0$ for all $F_0 \in \mathcal{F}$, $f$ is said to be weakly differentiable in $\mathcal{F}$ or simply weakly differentiable.

**Theorem 3:** Assume a weakly differentiable functional $f$ in a convex set $\mathcal{F}$ achieves its maximum.

1) If $f$ achieves it maximum at $F_0$ then $f_{F_0}^{\theta}(F) \leq 0$ for all $F \in \mathcal{F}$.
2) If $f$ is concave, then $f_{F_0}^{\theta}(F) \leq 0$ for all $F \in \mathcal{F}$ implies that $f$ achieves its maximum at $F_0$.

**Proof:** See Smith [2].

Let us prove that the functionals $I$ and $g$ are weakly differentiable in $\mathcal{F}$. Define
\[
F_{\theta} = (1-\theta)F_0 + \theta F
\]
and
\[
i(x; F) = \int p(y | x) \ln[p(y | x) / p(y; F)] dy.
\]
Then
\[
I(F_{\theta}) - I(F_0) = \int p(y | x) \ln \frac{p(y; F_{\theta})}{p(y; F_0)} dy dF_0(x)
+ \theta \int i(x; F_{\theta}) dF(x)
- \theta \int i(x; F_0) dF_0(x).
\]
Since $p(y; F_{\theta}) = (1-\theta)p(y; F_0) + \theta p(y; F)$, we have
\[
I_{F_0}^{\prime}(F) = \lim_{\theta \to 0} \frac{I(F_{\theta}) - I(F_0)}{\theta}
= \int i(x; F_0) dF(x) - I(F_0). \tag{63}\]
This equality is valid as long as each of the terms in the difference is finite. In our case, this is guaranteed through the power constraint. As for $g$ we have
\[
g_{F_0}^{\prime}(F) = g(F) - g(F_0). \tag{64}\]
These expressions are valid for arbitrary $F_0$ and $F$ in $\mathcal{F}$, which implies that $I$ and $g$ are weakly differentiable, and, consequently, so is $I - g$.

Moreover, since $g(F)$ is linear in $F$, and $I$ is (strictly) concave, $I - g$ is concave. Hence, by Theorem 3, a necessary and sufficient condition for $F_0$ to achieve the supremum in (60) is
\[
I_{F_0}^{\prime}(F) - g_{F_0}^{\prime}(F) \leq 0, \quad \forall F \in \mathcal{F} \tag{65}\]
that is,
\[
\int [i(x; F_0) - \gamma x^2] dF(x) \leq I(F_0) - \gamma \int x^2 dF_0(x) \tag{66}\]
or, equivalently,
\[
\int [i(x; F_0) - \gamma x^2] dF(x) \leq C - \gamma a \tag{67}\]
because if $\int x^2 dF_0(x)$ is strictly less than $a$, the moment constraint is trivial and $\gamma$ is zero by (61), and (67) remains true.

**Theorem 4:** Let $E_0$ be the points of increase of a distribution function $F_0$. Then
\[
\int [i(x; F_0) - \gamma x^2] dF(x) \leq C - \gamma a \tag{68}\]
for all $F \in \mathcal{F}$ if and only if
\[
i(x; F_0) \leq C + \gamma(x^2 - a), \quad \forall x, \tag{69}\]
and
\[
i(x; F_0) = C + \gamma(x^2 - a), \quad \forall x \in E_0. \tag{70}\]

**Proof:** The implication from (69) to (68) is immediate. For the converse, assume (69) is false. Then there exists an $\tilde{x}$ such that
\[
i(\tilde{x}; F_0) > C + \gamma(\tilde{x}^2 - a). \tag{71}\]
If $F$ is a unit step function at $\tilde{x}$, then
\[
\int [i(x; F_0) - \gamma x^2] dF(x) = i(\tilde{x}; F_0) - \gamma \tilde{x}^2 > C - \gamma a \tag{72}\]
which contradicts (68).
Assume now that (69) is true but (70) is false, that is, there exists \( \tilde{x} \in E_0 \) such that
\[
\tilde{a}(\tilde{x}; F_0) < C + \gamma(\tilde{x}^2 - a).
\]  
(73)

Since all the functions in the above equation are continuous in \( x \), the inequality is satisfied strictly on a neighborhood \( E_\delta \) of \( \tilde{x} \). By definition of a point of increase, the set \( E_\delta \) necessarily has a nonzero measure \( \int_{E_\delta} dF_0(x) = \delta > 0 \). Hence,
\[
C + \gamma a = I(F_0) = \gamma a
\]
\[
= \int [\tilde{a}(x; F_0) - \tilde{x}^2] dF_0(x)
\]
\[
= \int_{E_\delta} [\tilde{a}(x; F_0) - \tilde{x}^2] dF_0(x)
\]
\[
+ \int_{E \setminus E_\delta} [\tilde{a}(x; F_0) - \tilde{x}^2] dF_0(x)
\]
\[
< \delta(C - \gamma a) + (1 - \delta)(C - \gamma a)
\]
\[
< C - \gamma a
\]
(74)
which is a contradiction. \( \square \)

APPENDIX III

\( L \)th-Order Diversity

Let us generalize our study to a real-valued scalar channel with \( L \) independent branches. If \( U \) is the real-valued channel input, \( V^k \) the real-valued output of the \( k \)th branch, and \( A^k \) and \( N^k \) the fading and additive noise of the \( k \)th branch, then the channel is described by
\[
V^k = A^k U + N^k, \quad \text{for } k = 1, \ldots, L.
\]
(75)

We assume that the \( A^k \)'s and \( N^k \)'s are mutually independent, and zero-mean real Gaussian random variables with variance \( \sigma_A^2 \) and \( \sigma_N^2 \), respectively.

When \( L = 2 \) we obtain the Rayleigh-fading channel studied in this paper. The case \( L = 1 \) describes a Gaussian fading channel encountered in the study of the voice channel [26]. When \( L \) is even and greater than 2, the model describes an i.i.d. Rayleigh-fading channel with \( L/2 \)-order diversity.

A sufficient statistic for \( U \) is \( \sum_{k=1}^{L} V^k \). Hence, denoting
\[
Y = \sum_{k=1}^{L} |V^k|^2/\sigma_N^2
\]
and
\[
X = [U]|\sigma_A/\sigma_N
\]
an equivalent channel is obtained with transition probability
\[
p(y|x) = \frac{y^{L/2-1}}{(1 + x^2)^{L/2} I(L/2)} \exp \left(-\frac{y}{1 + x^2}\right)
\]
(76)
which, as a function of the random variable \( S \), can be written
\[
p(y|S) = \frac{s^{L/2}y^{L/2-1}}{I(L/2)} \exp(-ys).
\]
(77)

For Theorem 1 of Appendix I to remain valid for the \( L \)th-order diversity channel, we need to prove that \( I(F) \) is weak* continuous and strictly concave. The chain of equations proving the weak* continuity of \( I(F) \) remains valid since \( p(y|x) \) can still be upper-bounded by a function that is initially constant and then decaying as \( 1/y^2 \). Furthermore, (76) describes a multiplicative channel \( y = (1 + x^2)\omega \), where \( \omega \) is independent of \( x \) and has a probability density function
\[
p(\omega) = \frac{\omega^{L/2-1}}{I(L/2)} \exp(-\omega).
\]
Following Appendix I, it can be shown that \( p(y|f) \) is well-defined and analytic for complex values of \( f \) on the band \(-L/2 < \text{Im}(f) < L/2 \) [13]. Hence, the same technique may be applied and \( I(F) \) is strictly concave.

We can thus apply the Kuhn–Tucker condition to yield
\[
\gamma \left[ \frac{1}{s} - 1 - a \right] + C - \ln s + \ln I(L/2)
\]
\[
+ (1 - L/2)\Psi(L/2) + L/2
\]
\[
+ \frac{1}{I(L/2)} \int_0^\infty s^{L/2} y^{L/2-1} \exp(-ys) p(y|F^*) dy \geq 0
\]
(78)
for \( s \in (0, 1] \), with equality if \( s \) is in the support of \( F^* \), and where \( \Psi(z) = d\ln I(z)/dz \).

The proof of the finite character of the optimal distribution follows what is presented in the paper. Having a positive accumulation point is absurd, since the only output density that satisfies (78) with equality everywhere has the form \( p(y) = Ke^{-\gamma y}/y \), which is not a valid probability density. The accumulation point at zero can then be ruled out using a similar technique. Indeed, \( p(y) \) can be lower-bounded by
\[
p(y) > p_n \frac{s_n^{L/2} y^{L/2-1}}{I(L/2)} \exp(-ys_n),
\]
(79)
Hence, (78) cannot be satisfied with equality for \( s \) arbitrarily close to zero unless the Lagrange multiplier \( \gamma \) is zero. But, by using a discrete uniform input distribution with \( N \) input levels located at \( 0, \frac{K}{2}, \ldots, \frac{K^{N-2}}{2} \), it can be proven that \( \gamma \neq 0 \).

The proof of the existence of an impulse at zero is identical to the one given in the paper. Given that
\[
\frac{\partial}{\partial \alpha_0} p(y|x_0) = \frac{-2x_0}{(1 + x_0^2)}[y - (1 + x_0^2)L/2]p(y|x_0)
\]
(80)
and the mean of \( p(y|x_0) \) is \( (1 + x_0^2)L/2 \), then Lemma 1 applies and the proof is the same.

ACKNOWLEDGMENT

The authors wish to thank Prof. R. Gallager and A. Lapidoth of the Laboratory for Information and Decision Systems at MIT for numerous helpful discussions.

REFERENCES


