The flexibility of context maximizing, allows us to describe efficient methods for producing minimum description length classification structures. Combinations of our maximizing methods for the different classes lead to interesting classification procedures, even for attributes that take values in "large" alphabets. Note that the algorithm for Class III selects the positions (attributes) which gives the highest reduction of the description length and produces a decision tree (see Quinlan and Rivest [4]), while Class II methods can be used to find the most effective thresholds in large attribute alphabets. The fact that there exist elegant context weighting methods to treat missing attributes demonstrates once more the flexibility of context weighting (maximizing).

ACKNOWLEDGMENT

The authors wish to thank the reviewers and the Associate Editor M. Feder for valuable comments.

REFERENCES


Nearest Neighbor Decoding for Additive Non-Gaussian Noise Channels

Amos Lapidoth, Member, IEEE

Abstract—We study the performance of a transmission scheme employing random Gaussian codebooks and nearest neighbor decoding over a power limited additive non-Gaussian noise channel. We show that the achievable rates depend on the noise distribution only via its power and thus coincide with the capacity region of a white Gaussian noise channel with signal and noise power equal to those of the original channel. The results are presented for single-user channels as well as multiple-access channels, and are extended to fading channels with side information at the receiver.

Index Terms—Nearest neighbor decoding, Euclidean distance, mismatched decoding, multiple-access channel, generalized cutoff rate, fading.

I. INTRODUCTION

Due to the simplicity of their implementation using matched filters and due to their robustness, nearest neighbor decoders (minimum Euclidean distance decoders) are often used on additive noise channels even if the noise is not a white Gaussian process. For such channels, the nearest neighbor decoding rule is usually suboptimal and a loss in performance is thus incurred. In this correspondence we quantify this loss in terms of the achievable rates, i.e., the rates at which reliable communication is possible with a nearest neighbor decoder.

Stated as above, the problem is a special case of the general mismatch problem that has been studied in [1]-[5] for the single-user channel, and in [6] for the multiple access channel. However, the techniques that we use in order to study the problem are quite different from those used in the above references and rely heavily on Euclidean geometry. This allows us to deal with the infinite input and output alphabets and with the memory that the noise may exhibit.

It should be noted that we only study the case where the transmitter uses random Gaussian codebooks. While this is the optimal input distribution for white Gaussian channels, this may not necessarily be optimal for non-Gaussian channels and the decoder that we assume, see Example 1.

The motivation to assume this input distribution is that Gaussian codebooks are relatively well understood, and that, as we shall see, Gaussian codebooks and nearest neighbor decoders form a very robust communication scheme. Furthermore, for the power-limited additive noise single-user channel, Gaussian noise and Gaussian signaling constitute a saddle-point for the mutual information functional [7], given that the noise is Gaussian the input distribution that maximizes the mutual information between the channel input and output is the Gaussian distribution, and given that the input distribution is Gaussian, the worst noise, i.e., the noise that minimizes

Manuscript received May 9, 1995; revised Feb. 20, 1996. This work was supported in part by NSF under Grant NCR-9205663, JSEP under Contract DAAM04-94-G-0058, and ARPA under Contract J-FBI-94-218-2. The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Trondheim, Norway, June 1994.

The author is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139-3007 USA.

Publisher Item Identifier S 0018-9448/96$05.00 © 1996 IEEE
the mutual information, is the Gaussian noise. It thus seems that a robust design approach is to design for the worst case, i.e., assume that the noise is Gaussian, and use a Gaussian codebook and a decoder that is optimal for Gaussian noise, i.e., a nearest neighbor decoder. This reasoning seems to imply that a Gaussian codebook and a nearest neighbor decoder should allow one to transmit at rates arbitrarily close to the Gaussian capacity (corresponding to Gaussian inputs and Gaussian noise) if the noise is Gaussian, and at even higher rates if the noise is not the worst noise (i.e., not Gaussian).

One should, however, exercise great caution when following the above line of reasoning because if the noise is not Gaussian, then nearest neighbor decoding is no longer the optimal decoder and the mutual information between the input and output is no longer an appropriate measure of the achievable rates. If the noise is not Gaussian, then two opposite forces are at work: on the one hand the noise is less malevolent than the Gaussian noise and the mutual information is bigger than the Gaussian capacity, but on the other hand the receiver is no longer optimal, and some loss in performance might arise from the suboptimality of the decoder.

The robustness of Gaussian codebooks and nearest neighbor decoding for the single-user channel has been demonstrated before, see [8]-[10] and references therein. (For the case where the noise power is smaller than the transmitted power, the strongest version is apparently due to Csiszar and Narayan [11].) Our contribution to the single-user channel is in proving a random coding converse, i.e., that with random Gaussian codebooks and nearest neighbor decoding no rate above $(1/2) \log (1 + P/N)$ is achievable, where $P$ is the signal power and $N$ is the noise power, and in extending these results to the multiple-access channel and to fading channel with side information at the receiver.

Our results do not imply that if one is to use Gaussian codebooks and one is to guarantee that $(1/2) \log (1 + P/N)$ be achievable for any noise distribution of power $N$, then no rate above $(1/2) \log (1 + P/N)$ is achievable with a decoder that does not depend on the noise distribution; it is only with a nearest neighbor decoder that this holds.

At least for memoryless noise processes, one can in fact achieve the mutual information between the channel input and channel output using a universal decoding rule, see [12] and [13]. The shortcoming of the universal decoders is that their implementation is often too complex even for codebooks with strong algebraic structure.

The term "Gaussian codebook" requires some clarification. In this correspondence it will always refer to a random codebook whose codewords are chosen independently of each other. Sometimes we shall take it to imply that each codeword is chosen uniformly over the $n$-dimensional sphere of radius $\sqrt{n}P$, where $P$ is the average power and $n$ is the blocklength, and at other times we shall take it to mean that each codeword is chosen according to a product Gaussian distribution so that the coordinates of the codewords are i.i.d. $N(0, P)$, where $N(m, \sigma^2)$ denotes the Gaussian distribution of mean $m$ and variance $\sigma^2$. The different meaning will usually be clear from the context or unimportant; otherwise, we shall make the distinction explicit and refer to the former case by "equi-energy Gaussian codebook" and to the latter by "product Gaussian codebook." All the results stated in this correspondence hold for both cases, but we shall prove them for the case for which the proof is simplest. It should be noted that, in general, a direct part is slightly more impressive if it is shown for a product-Gaussian codebook rather than for an equi-energy Gaussian codebook, whereas a random coding converse is stronger if it applies to an equi-energy Gaussian codebook.

II. THE SCALAR SINGLE-USER CASE

The single-user additive noise scalar channel is a channel whose input and output are real valued and satisfy

$$Y^{(k)} = X^{(k)} + Z^{(k)}. \quad (1)$$

Here $X^{(k)}$ and $Y^{(k)}$ denote the input and output of the channel at time $k$, and $Z^{(k)}$ is the noise sample. We assume that the process $\{Z^{(k)}\}$ is independent of the input process, and that it is ergodic with second moment $N$, i.e.,

$$E[Z^{(k)}]^2 = N.$$ 

In fact, our results still hold if rather than requiring ergodicity of the noise process we require that its empirical average power converge in probability to $N$, i.e.,

$$\lim_{n \to \infty} Pr\left(\frac{1}{n} \sum_{k=1}^{n} (Z^{(k)})^2 - N > \delta \right) = 0, \forall \delta > 0. \quad (2)$$

**Theorem 1:** For a single-user scalar additive noise channel with nearest neighbor decoding we have that, irrespective of the noise distribution, the average probability of error, averaged over the ensemble of Gaussian codebooks of power $P$, approaches zero as the blocklength $n$ tends to infinity for code rates below $\frac{1}{2} \log (1 + P/N)$, and approaches one for rates above $\frac{1}{2} \log (1 + P/N)$.

**Proof:**

**Step 1:** Consider a blocklength $n$ equi-energy Gaussian codebook of average power $P$ and rate $R$

$$C = \{z(1), \ldots, z(2^nR)\}.$$ 

We are interested in $P$, the average probability of error corresponding to a codebook drawn from the ensemble, averaged over the ensemble of codebooks. Since the different codewords are chosen independently and according to the same distribution, the ensemble averaged probability of error corresponding to the $i$th message $P(i)$, does not depend on $i$, and we thus conclude that $P = \bar{P}(1)$. We can thus...
uniformly over the sphere, it follows that the conditional distribution of error given that the transmitted message is message 1 and that the noise vector is $z$. We thus have

$$P = P(1) = E_x(e(z))$$

where $E_x$ denotes expectation with respect to the noise sequence.

Step 2: We claim that $e(z)$ depends on $z$ only via its Euclidean norm $\|z\|$. To see this note that the spherical symmetry of the codebook distribution implies that given $z$ and $x(1)$, the probability of error depends only on $\|z + x(1)\|$ and $\|z\|$. Since $x(1)$ is drawn uniformly over the sphere, it follows that the conditional distribution of $\|z + x(1)\|$ depends on $z$ only via its norm. To stress that $e(z)$ depends on $z$ only via its norm (and dimension), we shall denote $e(z)$ by $e_n(\|z\|^2/n)$.

The monotonicity now follows from the observation that if for some $i 
eq 1$ we have $\|y - z(i)\| \leq \|y - z(1)\|$ where $y = z(1) + \|z\|\hat{a}$, then $\|y' - z(i)\| \leq \|y' - z(1)\|$ where $y' = z(1) + \|z\|\hat{a} + \delta\hat{a}$ and $\delta > 0$. Indeed

$$\|x(1) + \|z\|\hat{a} - z(i)\| = \|x(1) + \|z\|\hat{a} - z(1)\| + \delta \|\hat{a}\| - \|z\| \leq \|x(1) + \|z\|\hat{a} - x(i)\| + \delta \|\hat{a}\|$$

where the first inequality follows from the triangle inequality and the positivity of $\delta$, and the second from the assumption that for the noise vector $\|z\|\hat{a}$ the codeword $x(1)$ is at least as close to the received sequence as the transmitted codeword $x(1)$ is.

Step 3: We can now conclude the proof by a comparison argument. Consider first the case where the code’s rate satisfies

$$R < \frac{1}{2} \log \left(1 + \frac{P}{N}\right)$$

so that there exists an $N' > N$ such that

$$R < \frac{1}{2} \log \left(\frac{P}{N'}\right).$$

Compare now the ensemble average probability of error corresponding to a Gaussian codebook transmitted over the additive i.i.d. Gaussian channel of variance $N'$ with that over the channel under consideration. Let $z$ denote the noise vector in the original (non-Gaussian) channel, and let $z'$ denote the noise vector in the Gaussian channel. By the law of large numbers (LLN) $\|z\|/n$ approaches $N'$ as $n$ tends to infinity, and since $N' > N$ we conclude from (2) that

$$\Pr (\|z\| > \|z'\|) \to 1.$$
of events bound we obtain that the probability of a message error is upper-bounded by

$$2^{nR} \operatorname{Pr} (q(x', y) \geq q(x, y))$$

where $R$ is the code rate, and $n$ is its blocklength. Using the Chernoff bound we obtain

$$\Pr (q(x', y) \geq q(x, y)) = \inf_{\lambda > 0} \left( \sum_{(x, x') \in X \times X} p(x)p(x') e^\lambda (q(x', y) - q(x, y))\right)^n$$

and we can thus conclude that the generalized cutoff rate, defined as

$$\text{GR}_0 = -\log \inf_{\lambda > 0} \sum_{(x, x') \in X \times X} p(x)p(x') e^\lambda (q(x', y) - q(x, y))$$

is achievable, i.e., a lower bound on the mismatch capacity $C_M$.

Let us now focus once again on the scalar additive noise channel with nearest neighbor decoding, noise power $N$, and Gaussian codebook of power $P$. For this channel we know that irrespective of the noise distribution the rate $\frac{1}{2} \log (1 + \frac{N}{P})$ is achievable (Theorem 1). We shall, however, show that by a proper choice of the noise distribution (of power $N$), we can make $\text{GR}_0$ arbitrarily small.

Let $d(z)$ denote the probability that a randomly chosen codeword $z'$ is closer to $y = z + z$ than the correct codeword $x$ is. By the spherical symmetry of the Gaussian codebook $d(z)$ depends on the noise sequence only via its norm $\|z\|$. We shall hence denote $d(z)$ by $d_n(\|z\|/n)$. Following the reasoning outlined in Step 3 in the proof of Theorem 1 we conclude that $d_n(\|z\|/n)$ is monotonically nondecreasing.

Let $\epsilon > 0$ be arbitrary, and let $\eta_0$ be sufficiently large so that $\epsilon > \frac{1}{2} \log (\frac{\eta_0}{N})$. Consider now a deterministic noise $Z = \eta_0$. By the random coding converse of Theorem 1 we conclude that for sufficiently large blocklength $n$ the average probability of error, averaged over an ensemble of rate $\epsilon$ Gaussian codebooks, is greater than $\frac{1}{2}$. It thus follows from the union of events bound that for sufficiently large $n$, we have $2^{-n} \epsilon_n(\eta_0) > 1/2$. By the monotonicity of $\epsilon_n(\eta)$ we conclude that for sufficiently large $n$

$$\epsilon_n(\eta) > \frac{1}{2} 2^{-n\epsilon}, \quad \forall \eta \geq \eta_0. \quad (3)$$

For any noise distribution we thus have that for a sufficiently large blocklength $n$

$$\Pr (\|x' - y\| \leq \|z\|) \geq \frac{1}{2} \epsilon_n(\eta) \Pr (\|z\|/n) > \eta_0) \geq \frac{1}{2} 2^{-n\epsilon} \Pr (\|z\|/n) > \eta_0$$

where the first inequality follows from the nonnegativity of $\epsilon_n(\cdot)$, and the next from its monotonicity, and the last by (3). If the noise is i.i.d. with second moment $N$ then by the LLN

$$\lim_{n \to \infty} \Pr (\|z\|^2/n > \eta_0) = 0$$

provided that $\eta_0 > N$. However, the LLN does not have universal convergence rates, and we can find i.i.d. noise distributions of second moment $N$ such that

$$\lim_{n \to \infty} \frac{1}{n} \log \Pr (\|z\|^2/n > \eta_0) < \epsilon. \quad (5)$$

For such noise distributions, if we combine (5) and (4), we obtain that for sufficiently large $n$

$$\Pr (\|x' - y\| \leq \|z\|) > 2^{-2n\epsilon}$$

and the generalized cutoff rate is smaller than $2\epsilon$, i.e., arbitrarily small.

An example of such a noise distribution is given by

$$Z = \begin{cases} +\sqrt{A}, & \text{w.p. } \frac{N}{2A} \\ 0, & \text{w.p. } 1 - \frac{N}{A} \\ -\sqrt{A}, & \text{w.p. } \frac{N}{2A} \end{cases}$$

where $A$ is chosen sufficiently large. This distribution has second moment $N$ and the corresponding distribution of $Z^2$ has log-moment generating function $\Lambda_2(\theta)$

$$\Lambda_2(\theta) = \ln \left( 1 - \frac{N}{A} (1 - e^{\theta A}) \right)$$

Computing the Legendre–Fenchel transform at $\eta_0$ leads to

$$\Lambda_2^*(\eta_0) = \sup_{\theta < 0} \left[ \eta_0 \theta - \ln \left( 1 - \frac{N}{A} (1 - e^{\theta A}) \right) \right] \leq -\ln \left( 1 - \frac{N}{A} \right) \xrightarrow{A \to \infty} 0.$$

Notice that this noise distribution is discrete, so that for a Gaussian input distribution the matched capacity as well as the computational cutoff rate are infinite whereas the mismatch capacity corresponding to nearest neighbor decoding is lower-bounded by $\frac{1}{2} \log (1 + \frac{N}{P})$, and the generalized cutoff rate is smaller than $2\epsilon$. We thus see that the generalized cutoff rate and $C_{LM}$ can behave very differently even when the computational cutoff rate behaves much like the matched capacity. (The fact that the computational cutoff rate behaves much like the matched capacity in this example should not be taken for granted [16].)

The generalized cutoff rate behaves so poorly in our example because a noise vector of large norm may cause an exponential number of incorrect codewords to be closer to the received sequence than is the correct codeword.

IV. THE VECTOR SINGLE-USER CASE

The inputs and outputs of the single-user vector additive-noise channel are vectors in the $m$-dimensional Euclidean space $\mathbb{R}^m$. At any time $k$ we have

$$X(k) = X(k) + N(k)$$

where $X(k)$ and $N(k)$ denote the channel input and output at time $k$ and where $N(k)$ corresponds to the added noise at time $k$. The noise is assumed to be independent of the channel input, and ergodic with second moment (positive-definite) matrix $\Lambda \in \mathbb{R}^{m \times m}$, i.e.,

$$\mathbb{E}[N^2] = \Lambda.$$

Here $\cdot^t$ denotes the matrix transpose operation. An average power constraint is imposed on the channel input, so that

$$\mathbb{E}[\|X_k\|^2] \leq P$$

for some positive average power $P$. 
If the added noise were i.i.d. Gaussian, the channel capacity would be given by the water pouring solution, i.e.,
\[ C_G = \sum_{\nu=1}^{m} \frac{1}{2} \log \left( 1 + \frac{P_{\nu}}{\lambda_{\nu}} \right) \tag{6} \]
where \( \{\lambda_{\nu}\}_{\nu=1}^{m} \) are the eigenvalues of the matrix \( \Lambda \), and
\[ P_{\nu} = (\mu - \lambda_{\nu})^+, \quad \nu = 1, \ldots, m \tag{7} \]
where \( x^+ = \max\{x, 0\} \), and \( \mu \) is selected so that
\[ \sum_{\nu=1}^{m} P_{\nu} = P. \tag{8} \]

Let \( C \) be a blocklength \( n \), rate \( R \), codebook for the vector channel, i.e., a set of \( 2^{nR} \) \( n \)-tuples of elements from \( \mathbb{R}^m \). Given an output sequence \( y = (y^{(1)}, \ldots, y^{(n)}) \) the nearest neighbor decoder decides that the transmitted codeword is the codeword \( i \) that achieves
\[ \min_{1 \leq i' \leq 2^{nR}} \|y - z(i')\|^2 \]
where
\[ z(i') = (z^{(1)}(i'), \ldots, z^{(n)}(i')) \]
and
\[ \|y - z(i')\|^2 = \sum_{k=1}^{n} \|y^{(k)} - z^{(k)}(i')\|^2. \]

As before, we shall assume that if two codewords achieve the minimum then a decoding error is declared. Notice that even if the noise process is i.i.d. Gaussian, nearest neighbor decoding is in general not equivalent to the maximum-likelihood decoding rule unless the matrix \( \Lambda \) is a scalar matrix, i.e., the identity matrix multiplied by a scalar. Nevertheless we have the following result:

**Theorem 2:** The capacity of the vector additive-noise channel with average power \( P \), ergodic noise of second moment \( \Lambda \), and a nearest neighbor decoder is lower-bounded by the (matched) capacity of the corresponding vector i.i.d. Gaussian channel (6).

**Proof:** The achievability of (6) will be demonstrated by treating the channel as a MAC with \( m \) transmitters. Let \( \{\psi_{\nu}\}_{\nu=1}^{m} \) denote an orthonormal set of real eigenvectors for the matrix \( \Lambda \), where \( \psi_{\nu} \) is an eigenvector corresponding to the eigenvalue \( \lambda_{\nu} \). Let \( \chi_{1} = \cdots = \chi_{m} = \mathbb{R} \) denote the input alphabets of \( m \) users, and consider the mapping
\[ \phi: \chi_{1} \times \cdots \times \chi_{m} \to \mathbb{R}^m \]
\[ \phi: (x_{1}, \ldots, x_{m}) \mapsto \sum_{\nu=1}^{m} x_{\nu} \psi_{\nu}. \]

For an arbitrarily small \( \delta > 0 \), assign to each user \( \nu \) a codebook \( C_{\nu} \) of rate
\[ R_{\nu} = \frac{1}{2} \log \left( 1 + \frac{P_{\nu}}{\lambda_{\nu}} \right) - \delta \]
and average power \( P_{\nu} \) as determined by (7). Note that since the we have chosen \( \{\psi_{\nu}\} \) to be orthonormal, the average power of \( \phi(x_{1}, \ldots, x_{m}) \) is by (8) equal to \( P \).

We now show that when the vector channel is used as a MAC and nearest neighbor decoding is employed, then each user experiences a scalar additive noise channel with nearest neighbor decoding. To be more precise, we claim that the resulting MAC decodes the received sequence \( y \) to the \( m \)-tuple of codewords \( (i_{1}, \ldots, i_{m}) \) where \( z_{\nu}(i_{\nu}) \) is the message in user \( \nu \)'s codebook that achieves
\[ \min_{(x^{(1)}, \ldots, x^{(m)}) \in C_{\nu}} \sum_{k=1}^{n} (\bar{y}_{\nu}^{(k)} - y^{(k)})^2 \]
and \( \bar{a} \cdot \bar{b} \) denotes the standard Euclidean inner product between \( \bar{a} \) and \( \bar{b} \). The above claim follows immediately from the orthogonality of the basis \( \{\psi_{\nu}\} \) since
\[ \arg \min_{(x_{1}, \ldots, x_{m}) \in \mathbb{C}^{m}} \sum_{k=1}^{n} (\bar{y}_{\nu}^{(k)} - x^{(k)})^2 = \left( \arg \min_{x_{1} \in C_{1}} \sum_{k=1}^{n} (\bar{y}_{1}^{(k)} - x_{1}^{(k)})^2, \ldots, \right) \]
\[ \arg \min_{x_{m} \in C_{m}} \left( \bar{y}_{m}^{(k)} - x_{m}^{(k)})^2 \right). \]

Since \( \{\bar{N}^{(k)}\} \) is ergodic with second moment \( \Lambda \), it follows that for any \( \nu = 1, \ldots, m \), the process \( \{\bar{y}_{\nu} \cdot \bar{N}^{(k)}\} \) is ergodic and of second moment \( \lambda_{\nu} \). It now follows from Theorem 1 that the codebooks \( C_{1}, \ldots, C_{m} \) can be chosen so that each user is decoded with arbitrarily small probability of error. We can conclude that the sum of the rates of the different users is achievable, and since \( \delta \) was chosen arbitrarily small, the rate (6) is achievable.

The above argument is very similar to the argument that we used in [17] and subsequently in [5] to prove the result for i.i.d. multivariate Gaussian noise.

It follows as a corollary of Theorem 2 that for i.i.d. vector Gaussian noise and nearest neighbor decoding, the mismatch capacity is equal to the matched capacity, even though nearest neighbor decoding is clearly different from maximum-likelihood decoding rule which requires a whitening filter. It is interesting to note that this cannot in general be demonstrated using a standard random coding argument. Indeed, if under the water pouring solution (7), (8) there exist two distinct eigenvalues \( \lambda_{\nu}, \lambda_{\nu}' \) such that \( P_{\nu} \) and \( P_{\nu}' \) are both positive, then random coding under no input distribution will demonstrate that the mismatch capacity is equal to the matched capacity. This is shown in Appendix I.

**V. ADDITIVE-NOISE MAC**

We now consider the additive-noise multiple-access channel with joint minimum Euclidean distance decoding. The channel has two real-valued inputs and one real-valued output. The channel output at time \( k \), \( Y^{(k)} \), is given by
\[ Y^{(k)} = X_{1}^{(k)} + X_{2}^{(k)} + Z^{(k)} \]
where \( X_{\nu}^{(k)}, \nu = 1, 2 \), is the signal transmitted at time \( k \) by user \( \nu \), and \( Z^{(k)} \) is the noise sample. The noise is assumed to be independent of the inputs and is further assumed to be ergodic of second moment \( N \) or to satisfy (2). We further assume that the average power transmitted by user \( \nu \) is limited by \( P_{\nu} \) for \( \nu = 1, 2 \). We are interested in the achievable rates with Gaussian codebooks and joint nearest neighbor decoding. Given a received sequence \( y \), such a decoder declares that the transmitted codewords were
\[ \arg \min_{x_{1} \in C_{1}, x_{2} \in C_{2}} \|y - x_{1} - x_{2}\| \]
and declares decoding failure if the minimum is not achieved uniquely. Let \( K_{G} \) denote the achievable region corresponding to...
to a Gaussian noise [7], i.e., \( \mathcal{R}_G \) is the set of all rate pairs \((R_1, R_2)\) that satisfy
\[
I \leq \frac{1}{2} \log \left( 1 + \frac{P_1}{N} \right), \quad \nu = 1, 2
\] and
\[
R_1 + R_2 \leq \frac{1}{2} \log \left( \frac{P_1 + P_2}{N} \right).
\]

**Theorem 3:** For the additive noise MAC with joint nearest neighbor decoding we have that irrespective of the noise distribution, the average probability of error, averaged over the ensemble of Gaussian codebooks pairs of powers \(P_1\) and \(P_2\), approaches zero as the blocklength \(n\) tends to infinity, if \((R_1, R_2)\) is in the interior of \( \mathcal{R}_G \), and approaches one if \((R_1, R_2) \notin \mathcal{R}_G \).

**Proof:** The average probability of error, averaged over the ensemble of codebook pairs, is equal to the probability that the message \((1, 1)\) is incorrectly decoded, averaged over the above ensemble. We shall thus assume that the transmitted codewords are \(x_1(1)\) and \(x_2(1)\). Let \(E_{i,j}\) denote the event
\[
\|x_1(i) + x_2(j) - y\| \leq \|x_1(1) + x_2(1) - y\| = \|z\|
\]
where \(y\) denotes the output sequence, and \(z\) is the noise sequence. It follows from the single-user result (Theorem 1) that
\[
\Pr\left( \bigcup_{i=2}^{2^n R_1} E_{i,1} \right) \xrightarrow{n \to \infty} \begin{cases} 
1, & \text{if } R_1 > \frac{1}{2} \log \left( \frac{P_1}{N} \right) \\
0, & \text{if } R_1 < \frac{1}{2} \log \left( \frac{P_1}{N} \right)
\end{cases}
\]
and similarly
\[
\Pr\left( \bigcup_{j=2}^{2^n R_2} E_{1,j} \right) \xrightarrow{n \to \infty} \begin{cases} 
1, & \text{if } R_2 > \frac{1}{2} \log \left( \frac{P_2}{N} \right) \\
0, & \text{if } R_2 < \frac{1}{2} \log \left( \frac{P_2}{N} \right)
\end{cases}
\]
To conclude the proof we need to study the asymptotic behavior of the probability of the event
\[
E = \bigcup_{i=2}^{2^n R_1} \bigcup_{j=2}^{2^n R_2} E_{i,j}.
\]
We first note that by the spherical symmetry of distribution on the codebooks, the conditional probability of the event \(E\) given the noise sequence \(z\) depends only on the norm \(\|z\|\). We shall consider this conditional probability by \(e_n(\|z\|/n)\). Next note that the functions \(e_n(\cdot)\) are monotonic nondecreasing. To see this assume that \(z = \|z\|\hat{a}\) where \(\hat{a}\) is some deterministic unit vector, and note that for \(\delta > 0\)
\[
\|x_1(1) + x_2(1) + \|z\|\hat{a} - x_1(i) - x_2(j)\| \leq \|z\|\]
implies
\[
\|x_1(1) + x_2(1) + (\|z\| + \delta)\hat{a} - x_1(i) - x_2(j)\| \leq \|z\| + \delta.
\]
We can now use a comparison argument similar to the one used in Step 4 in the proof of Theorem 1 to conclude from the above that the asymptotic probability of \(E\) depends only on the second moment of the noise. By considering a Gaussian noise of equal second moment we can conclude from the direct part to the coding theorem for the Gaussian MAC that if (10) is satisfied with strict inequality then
\[
\Pr(E) \to 0.
\]
Similarly, if (10) is violated then
\[
\Pr(E) \to 1
\]
as follows from the random coding converse to the Gaussian MAC, which can be proved using arguments very similar to those that were used in the proof of [6, Theorem 3].

**VI. FAADING CHANNELS**

We next consider fading additive noise channels. We assume throughout that the receiver has side information about the fading process, but that this information is not available at the transmitter. We shall treat both the single-user case as well as the MAC.

Consider first the single-user channel described by
\[
y(k) = A(k)x(k) + Z(k), \quad k \in \mathbb{Z}
\]
where \(Y(k)\) and \(X(k)\) are the received and transmitted symbols at time \(k\), \(A(k)\) is the fading process sampled at time \(k\), and \(Z(k)\) is the noise sample. We assume that the processes \(\{A(k)\}_{k=0}^{\infty}\) and \(\{Z(k)\}_{k=-\infty}^{\infty}\) are ergodic, independent of each other, and jointly independent of the input sequence. The second moment of \(Z(k)\) is denoted by \(N\). We shall focus on a modified minimum Euclidean distance decoder that, given the received sequence \(y\) and the fading sequence \(a\), decided that the transmitted codeword is
\[
\arg\min_{x \in \mathcal{C}} \sum_{k=1}^{n} (y(k) - a(k)x(k))^2
\]
and declares a decoding failure if the minimum is not unique. This decoding rule is equivalent to the maximum-likelihood decoding rule for the case where the additive noise is i.i.d. Gaussian. We now show that the random coding capacity of the fading channel with Gaussian codebooks and modified nearest neighbor decoding depends on the noise process only via its second moment \(N\). To simplify the proof we assume that the fading process \(\{A(k)\}\) takes value in some finite set \(\mathcal{A} \subseteq \mathbb{R}\). A proof of the more general result is sketched in Appendix II.

**Theorem 4:** Under the above assumptions, the random coding capacity of the fading channel with Gaussian codebooks and modified nearest neighbor decoding is given by
\[
E \left( \frac{1}{2} \log \left( 1 + \frac{A^2 P}{N} \right) \right)
\]
where \(P\) is the average power available at the transmitter, and \(N\) is the second moment of the noise.

We prove the theorem for a "product Gaussian codebook." Assume that the transmitted codeword is \(x(1) = (x^{(1)}(1), \ldots, x^{(n)}(1))\) and let the fading sequence be given by \(a = (a^{(1)}, \ldots, a^{(n)})\). The received sequence \(y\) is then given by \(y = o \circ x(1) + z\) where \(o\) denotes the Schur product (coordinate-wise product) and \(z\) is the added noise sequence. Let
\[
\sum_{k \in \mathcal{A}} (z(k))^2 = \sum_{k \in \mathcal{A}} (y(k))^2 = \kappa(a)
\]
where \(\kappa(a) = \{1 \leq k \leq n: a(k) = a\}\), \(a \in \mathcal{A}\).

**Step 1:** We claim that conditioned on \(x, \ a, \) and \(y, \) the probability of error is a function of \(\{\kappa(a)\}_{a \in \mathcal{A}}, \ \{y_\delta\}_{a \in \mathcal{A}}, \) and \(a\) only. To see this note that the metric accumulated by codeword \(x(i)\), for \(i \neq 1\), is given by
\[
\sum_{a \in \mathcal{A}} \sum_{k \in \kappa(a)} (y(k) - ax(x(i))^2
\]
which is distributed as a sum of $|A|$ independent random variables that are noncentral chi-square distributed with $\kappa(a)$ degrees of freedom and a noncentrality parameter $y^2_a$. Since the metric accumulated by the correct codeword $x(1)$ is

$$\sum_{a \in A} z_a^2$$

we conclude that the probability of error depends only on $\{z_a\}_{a \in A}$, $\{y_a\}_{a \in A}$, and $a$.

Step 2: We claim that conditioned on $z$ and $a$, the probability of error is a function of $\{z_a\}_{a \in A}$ and $a$ only. This follows easily from the previous step by integrating the conditional probability of error given $z$, $a$, and $y$ with respect to $y$ and noting that since $x(1)$ is i.i.d. Gaussian, the conditional distribution of $y_a$ given $z$ depends only on $z_a$.

Step 3: We show that the conditional probability of error given $\{z_a\}_{a \in A}$ and $a$ is monotonic in $z_a$ for any $a \in A$. Let $a^* \in A$ be fixed. By Step 2 we may drop loss in generality assume that the noise vector, projected to $\kappa(a^*)$, is aligned with some deterministic unit vector $\hat{a}$. We now use a coupling argument to compare the probability of error conditioned on $a$ and $\{z_a\}_{a \in A}$ with the probability of error conditioned on $a$ and $\{z_a\}_{a \in A}$ where

$$z_a^* = \begin{cases} z_a^* + \delta, & \text{if } a = a^* \\ z_a, & \text{otherwise} \end{cases}$$

and $\delta > 0$. For the purpose of computing the two conditional probabilities we choose the random codebooks used in the computation of the two conditional probabilities to be identical. We further choose the noise samples to be identical for all coordinates that are not in $\kappa(a^*)$. Let $\kappa(a^*)$ be chosen so as to align the projection of the noise vector on $\kappa(a^*)$ with $\hat{a}$, and so as to have the appropriate norm. In the system with smaller noise, an error will occur if

$$\exists j \neq 1: \sum_{a \neq a^*} \sum_{k \in \kappa(a^*)} (y^{(k)}_j - a^{(k)}_j z_j)^2 + \sum_{k \in \kappa(a^*)} (y^{(k)}_j - a^{(k)}_j z_j)^2 \leq \sum_{a \in A} z_a^2$$

or

$$\exists j \neq 1: \sum_{a \neq a^*} \sum_{k \in \kappa(a^*)} (y^{(k)}_j - a^{(k)}_j z_j)^2 \leq \sum_{k \in \kappa(a^*)} (z^{(k)}_j)^2.$$

The right-hand side of this equation is just

$$z_a^2 - \|z_a^* - \hat{a}^* \Pi a^* x(1) - a^* \Pi a^* z(1)\|^2$$

where $\Pi a^* x(1)$ denotes the projection of $x(1)$ to the coordinates for which $a^{(k)} = a^*$. By the triangle inequality this term is monotonic in $z_a^*$, and the monotonicity of the conditional probability of error is thus established.

Step 4: The theorem can now be proved using the LLN and by comparing the general noise to a Gaussian noise, as in the final step in the proof of Theorem 1.

Similar arguments can be made to show that for the additive-noise fading MAC, Gaussian codebooks and joint modified nearest neighbor decoding give rise to an achievable region that depends on the noise probability law only via its second moment. The output of the additive-noise fading MAC is given by

$$y^{(k)} = A^{(k)} X_1^{(k)} + B^{(k)} X_2^{(k)} + Z^{(k)}$$

where $X_1^{(k)}$ and $X_2^{(k)}$ are the signals transmitted by user 1 and user 2, respectively, $Z^{(k)}$ is the additive noise sample, and $\{A^{(k)}\}$ and $\{B^{(k)}\}$ are the fading processes that are assumed jointly ergodic and independent of the inputs and of the noise. The noise is assumed ergodic of second moment $N$, and the two users’ signals are assumed average-power limited by $P_1$ and $P_2$, respectively. The modified joint nearest neighbor decoder declares that the transmitted codewords are

$$\arg \min_{x_1, x_2 \in C_1 \times C_2} \sum_{k=1}^n \left( y^{(k)} - a^{(k)}_1 x_1^{(k)} - a^{(k)}_2 x_2^{(k)} \right)^2$$

and declares an error if the minimum is not unique. Once again we assume that the fading processes $\{A^{(k)}\}$ and $\{B^{(k)}\}$ take on values in finite subsets $A$ and $B$ of the real line. This assumption is not essential and merely simplifies the proof.

Theorem 3: Subject to the above conditions, the achievable region for additive noise fading MAC with Gaussian codebooks of powers $P_1$ and $P_2$, noise power $N$, and modified joint nearest neighbor decoding is given by the set of all rate pairs $(R_1, R_2)$ that satisfy

$$R_1 \leq E \left( \frac{1}{2} \log \left( 1 + \frac{A^2 P_1}{N} \right) \right)$$

$$R_2 \leq E \left( \frac{1}{2} \log \left( 1 + \frac{B^2 P_2}{N} \right) \right)$$

and

$$R_1 + R_2 \leq E \left( \frac{1}{2} \log \left( 1 + \frac{A^2 P_1 + B^2 P_2}{N} \right) \right).$$

Proof: The proof is given for a "product Gaussian codebook." To analyze the probability that only one of the codewords is correctly decoded one can use Theorem 4. It remains to check for the asymptotic probability that both codewords are incorrectly decoded. The analysis of this event is very similar to the analysis that was used in the proof of Theorem 4.

Let $a$ and $b$ be the realizations of the fading processes, and define $\kappa(a, b) = \{1 \leq k \leq n : (a^{(k)}, b^{(k)}) = (a, b)\}$. $\kappa(a, b)$ is a conditional probability that both codewords are incorrectly decoded. Further define $\Pi_{a, b} y = \Pi_{a, b} y_1(x(1)) = \Pi_{a, b} y_2(x(1))$ where $\Pi_{a, b} y$ denotes the projection of the vector $y$ to the coordinates at which the fading processes take on the value $(a, b)$, i.e., $\kappa(a, b)$.

Step 1: We claim that for any $(a, b) \in A \times B$ the joint distribution of $\{V_{i, j}\}$ $i = 2, \ldots, 2nR_1$, $j = 2, \ldots, 2nR_2$, when conditioned on $a$, $b$, and $x$ on a, $b$, and $x$ are independent for $(a, b) \neq (a', b')$.

Step 2: We conclude that the probability of error depends only on $\{z_{a, b}\}$ and $\kappa(a, b)$. $\kappa(a, b)$ is a conditional probability that both codewords are incorrectly decoded. Further define

$$V_{i, j} = \|\Pi_{a, b} y - a \Pi_{a, b} x(1) - b \Pi_{a, b} x(2)\|^2$$

where $\Pi_{a, b} y$ denotes the projection of the vector $y$ to the coordinates at which the fading processes take on the value $(a, b)$, i.e., $\kappa(a, b)$.

Step 1: We claim that for any $(a, b) \in A \times B$ the joint distribution of $\{V_{i, j}\}$ $i = 2, \ldots, 2nR_1$, $j = 2, \ldots, 2nR_2$, when conditioned on $a$, $b$, and $x$ on a, $b$, and $x$ are independent for $(a, b) \neq (a', b')$.
We now integrate the conditional distribution with respect to the conditional distribution of \( y_{z,b} \) given \( z \) and note that the latter depends only on \( z_{a,b} \), since it is a noncentral chi-square distribution with \(|\kappa(a,b)|\) degrees of freedom and noncentrality parameter \( z_{a,b} \). (The random quantity \( y_{z,b} \) is the sum of \(|\kappa(a,b)|\) Gaussian random variables of mean \( z_{a,b} \) and variance \( a^2P_1 + b^2P_2 \).

We now show that given \( a, b, \) and \( (z_{a,b}) \), the probability of error is a monotonic function of \( z_{a,b} \). In fact, proceeding as in Step 3 in the proof of Theorem 4, we see that it suffices to show that

\[
z^2_{a,b} - \|z_{a,b} - \bar{\bar{\alpha}} + a^*\Pi_{a,b} x_1(i) + b^*\Pi_{a,b} x_2(1) - b^*\Pi_{a,b} x_2(j)\|^2
\]

is monotonic in \( z_{a,b} \), which follows from the triangle inequality.

We then have that for a codebook of \( M+1 \) codewords the ensemble averaged probability of correct decoding, \( P_c \), is given by

\[
P_c = E[(1 - h(x_1, z(i) + z))^{(M+1)}].
\]

Notice that by the construction of the random codebook, \( \{x^{(k)}(i) : \bar{\bar{\nu}}, k, \nu \} \) are independent and \( \mathcal{N}(0, P_\nu) \) distributed. Let \( 0 < \delta < 1/2 \) be arbitrarily small, and let

\[
A_\delta = \left\{ (x_1, y) : \frac{1}{n}\|y - x_1\|^2 - \text{Tr}(A) < \delta, \right. \\
\left. \left( \frac{1}{n} \sum_{k=1}^{n} y_{k,\nu}^2 \right)^{1/2} - \sqrt{P_\nu + \lambda_\nu} < \delta, \right. \\
\left. \forall 1 \leq \nu \leq m \right\}
\]

where \( \text{Tr}(A) \) denoted the trace of a matrix \( A \). By the LLN the probability of the typical set \( A_\delta \) approaches one as the blocklength goes to infinity. By (11) we have

\[
P_c \leq \int_{A_\delta} \left( 1 - h(x_1, z(i) + z) \right)^{M} dP(x, y) + \text{Pr}(A_\delta^c)
\]

(12)

Let us now lower-bound \( h(x_1, z(i) + z) \) for \( (x_1, z(i) + z) \in A_\delta \). Noting that for \( (x_1, z(i) + z) \in A_\delta \), we have that

\[
\frac{1}{n}\|y - x\|^2 > \text{Tr}(A) - \delta
\]

we can conclude that

\[
h(x_1, y) \geq \text{Pr}\left( \frac{1}{n} \sum_{k=1}^{n} \sum_{\nu=1}^{m} \|y - x(i)\| \leq \text{Tr}(A) - \delta \right)
\]

\[
= \text{Pr}\left( \frac{1}{n} \sum_{k=1}^{n} \sum_{\nu=1}^{m} W_{k,\nu} \leq \text{Tr}(A) - \delta \right)
\]

\[
= \text{Pr}\left( \frac{1}{n} \sum_{k=1}^{n} \sum_{\nu=1}^{m} W_{k,\nu} \leq \text{Tr}(A) - \delta \right)
\]

(13)

**APPENDIX I**

In this appendix we briefly outline how one can show that for the additive i.i.d. multivariate Gaussian noise channel with nearest neighbor decoding, the probability of error is a monotonic function of \( z_{a,b} \). (The random quantity \( y_{z,b} \) is the sum of \(|\kappa(a,b)|\) Gaussian random variables of mean \( z_{a,b} \) and variance \( a^2P_1 + b^2P_2 \).)
where $W_{k,\nu}$ are independent random variables that are $\mathcal{N}(y_{k,\nu}, P_{\nu})$ distributed, and $\{W_{k,\nu}\}$ are a set of independent random variables.

We can now use the inequality

$$(a + \Delta)^{2} \leq (1 + 2|\Delta|)a^{2} + |\Delta|^{2}, \quad \forall |\Delta| < \frac{1}{2}, \quad a \in \mathbb{R}$$

with

$$a = W_{k,\nu} - \sqrt{\frac{1}{n} \sum_{k=1}^{n} y_{k,\nu}^{2} + \sqrt{P_{\nu} + \lambda_{\nu}}}$$

and

$$\Delta = \sqrt{\frac{1}{n} \sum_{k=1}^{n} y_{k,\nu}^{2} - \sqrt{P_{\nu} + \lambda_{\nu}}}$$

to obtain that since $(z(1), y)$ implies $|\Delta| < \delta$

$$\hat{W}_{k,\nu}^{2} \leq (1 + 2\delta)\hat{W}_{k,\nu}^{2} + \delta$$

where

$$W_{k,\nu} = W_{k,\nu} - \sqrt{\frac{1}{n} \sum_{k=1}^{n} y_{k,\nu}^{2} + \sqrt{P_{\nu} + \lambda_{\nu}}}$$

and is hence $\mathcal{N}(\sqrt{P_{\nu} + \lambda_{\nu}}, P_{\nu})$ distributed.

It now follows from (14) and (13) that for $(z(1), y) \in A_{n}^{2}$

$$h(z(1), y) > Pr\left(\frac{1}{n} \sum_{k=1}^{n} \sum_{\nu} \hat{W}_{k,\nu}^{2} < \frac{Tr(A) - (m + 1)\delta}{1 + 2\delta}\right)$$

$\overset{\text{def}}{=} a(n, \delta).$

Notice that $a(n, \delta)$ does not depend on $(z(1), y)$ provided that they are in $A_{n}^{2}$ and hence, by (12)

$$P_{e} \leq (1 - a(n, \delta))^{M} + Pr\left(A_{n}^{S}\right)$$

$$\leq e^{-Ma(n, \delta)}$$

where the second inequality follows from $1 - x \leq e^{-x}$. Recalling the definition of the code's rate we conclude that if $2^{nR}a(n, \delta)$ approaches infinity as $n$ tends to infinity, then the ensemble averaged probability of error tends to 1. It therefore remains to compute

$$C(\delta) = -\lim_{n \to \infty} \frac{1}{n} \log a(n, \delta).$$

This can be easily found using the theory of large deviations, and in particular, using Cramér's theorem [19]. The random variables $\{U_{k}\}_{k=1}^{m}$ where

$$U_{k} = \sum_{\nu=1}^{m} \hat{W}_{k,\nu}^{2}$$

are i.i.d. distributed with log-moment generating function

$$A_{U}(\theta) \overset{\text{def}}{=} \log E[e^{\theta U}]$$

$$= \sum_{\nu=1}^{m} \left(\frac{\theta(P_{\nu} + \lambda_{\nu})}{1 - 2\theta P_{\nu}} - \frac{1}{2} \log (1 - 2\theta P_{\nu})\right)$$

$$\theta < \min_{1 \leq \nu \leq m} \left\{\frac{1}{2P_{\nu}}\right\}.$$
statement can be made, namely that, except for a set of measure zero, (13) holds for every $\theta \leq 0$. To see this, note that by the ergodic theorem we can find a set of measure zero outside of which (15) holds for every nonpositive rational $\theta$. However, the functions $\Lambda_n(\theta)$ are convex [19], and hence, since $\Lambda(\theta)$ is continuous, convergence on the rationals implies convergence on the reals.

Let us denote by $G$ a set of probability 1 of realizations of the fading process, the received process, and noise process for which (15) holds for every nonpositive real $\theta$ and for which

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} (z(k))^{2} = N. \tag{16}$$

It now follows from the Gärtner–Ellis theorem [19] that for $\Lambda^*(\tau) = \frac{1}{\tau} \log \Pr(S_n < \tau) = -\Lambda^*(\tau)$ (17)

where $\Lambda^*(\tau)$ is the Fenchel–Legendre transform of $\Lambda(\cdot)$, i.e., $\Lambda^*(\tau) = \sup_{\theta \leq 0} (\theta \tau - \Lambda(\theta))$.

It now follows from (17), (18), and the union of events bound that for any $\delta > 0$ the rate $\Lambda^*(N + \delta)$ is achievable.

Similarly, for rates above $\Lambda^*(N - \delta)$, the expected number of incorrect codewords whose modified distance to the received sequence is smaller than that of the correct codeword is exponentially large, and hence, since the codewords are selected independently, the average probability of error averaged over the ensemble of codebooks tends to one.

It is straightforward to check that

$$\Lambda^*(N) = E \left[ \frac{1}{2} \log \left( 1 + \frac{A^2 P}{N} \right) \right] \tag{18}$$

and is achieved at $\theta = -1/(2N)$. The continuity of $\Lambda^*(\tau)$ (see [19]) now establishes the theorem.

ACKNOWLEDGMENT

Stimulating discussions with V. Balakirsky, T. Cover, E. Ordentlich, S. Shamai (Shitz), and A. Wyner are gratefully acknowledged. The author wishes to thank I. Csiszár and P. Narayan for an early preprint of their paper [18].

REFERENCES


On the Relative Entropy of Discrete-Time Markov Processes with Given End-Point Densities

Alessandro Beghi

Abstract—Given a Markov process $x(k)$ defined over a finite interval $I = [0, N]$, if $x \in R$ we construct a process $x^*(k)$ with the same initial density as $x$, but a different end-point density, which minimizes the relative entropy between $x$ and $x^*$. It is shown that $x^*$ is a Markov process in the same reciprocal class as $x$. In the Gaussian case, the minimum relative entropy problem is related to a minimum energy LQG optimal control problem.

Index Terms—Markov and reciprocal processes, relative entropy, minimum energy LQG control.

I. INTRODUCTION

The problem of building Markov processes with preassigned marginal probability densities at both ends of a given interval arises in both system identification and control. For instance, consider the situation where transition density and initial density $p_0(x)$ of

Manuscript received January 19, 1995; revised February 2, 1996. The material in this correspondence was presented in part at the 3rd European Control Conference, Rome, Italy, Sept. 1995.

The author is with the Dipartimento di Elettronica e Informatica, Università di Padova, I-35131 Padova, Italy.
Publisher Item Identifier S 0018-9448(96)03228-5.