UMA DEMONSTRAÇÃO SIMPLES E ELEGANTE PARA O TEOREMA FUNDAMENTAL DA TEORIA DE INFORMAÇÃO E SEU DUAL

FRANCISCO VENANCIO MOURA

ORIENTADOR

PROF. DR. GUR DIAL

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação da Universidade Estadual de Campinas, como requisito parcial para obtenção do Título de Mestre em Estatística.

Abril - 1981.

Aos meus pais, manos e tios,

pela minha formação

À minha esposa VERA

e a minha filha CAROLINA,

pelos estímulo, compreensão,

dedicação e amor.

O meu reconhecimento mais profundo.

ESPÍRITO-MAU BOM, PIOR

O homem nasce e vive em sociedade.
E perante ela tem inúmeros direitos,
Como também inúmeros deveres.
Tem o direito, por exemplo, de construir seu caráter,
De moldar seu espírito.
E seu espírito poderá ser bom, neutro ou,
Infelizmente, mau.
Mas sendo seu espírito mau tem o dever de
Jamais se camuflar de bom.
Para não ser pior.

Kamões

AGRADECIMENTOS

- Ao Prof. Dr. GUR DIAL pela sugestão do problema, bem como pela orientação segura e acima de tudo amiga;

- Ao DMPA da UFRN, Depto. de Estatística da UNICAMP e ao PICD pelos incentivos;

- Ao Casal PAULO-SÊ pelas pesquisas ao longo do curso , bem como pela convivência cordial;

- À minha esposa VERA, bem como ao PEREIRA, pelos eficientes trabalhos de datilografia e desenho dos gráficos respectivamente;

- Enfim, a todos "espíritos-bons"...

APRESENTAÇÃO

É fato notório que a informação existe desde épocas mais remotas. No entanto, o primeiro trabalho publicado sobre transmissão de informação data de 1928, realizado por Nyquist e Hartley [18]. Mas a "Teoria da Informação" propriamente dita está alicerçada, ou diríamos até, desabrochou, com um trabalho publicado por Shannon [24] em 1948. A partir de então esta teoria passou a ser pesquisada com mais afinco e intensidade em todos os sentidos e por consequência suas aplicações se estenderam a diversos ramos da ciência tais como Física [3], Estatística [22], Cibernética [30], Literatura [4], Ciências Físicas [7], Economia [29], Contabilidade [5] e muitos outros. Consequentemente a literatura sobre esta teoria passou a ficar deveras ampla e diversificada.

Seguindo a linha utilizada por Shannon [24] desenvolveremos nossa dissertação estudando "sistemas de comunicação" através de canais com ruído. No Capítulo I faremos um levantamento de resultados introdutórios, bem como vários desenvolvimentos necessários aos capítulos seguintes. No Capítulo II apresentaremos uma demonstração simples e elegante do teorema fundamental dado por R. G. Gallager em 1965. Esta demonstração dá um limite superior para a probabilidade de erro, que é de natureza exponencial. Utilizaremos, posteriormente, estes resultados em

alguns canais simples. No Capítulo III apresentaremos o dual do teorema fundamental e o limite inferior para a probabilidade de erro. Por fim, os Apêndices A e B contendo resultados úteis ao desenvolvimento da dissertação.

RESUMO

O tema "TEORIA DA INFORMAÇÃO" desenvolveu-se principalmente do Teorema Fundamental de Shannon, publicado em 1948. Dentre os magníficos resultados estabelecidos por Shannon o principal seria o teorema de codificação em canais com ruído, o qual estabelece que a transmissão da informação através destes canais pode ser feita com probabilidade de erro arbitrariamente pequena. O problema de obter limites para a probabilidade de erro surgiu com o teorema de codificação porque a avaliação exata desta probabilidade é, em geral, muito difícil de ser conseguida.

SUMMARY

The subject Information Theory mainly developed from Shannon's fundamental paper in 1948. Among the glorious achievements established by Shannon the most essential would be the theorem on noisy channels, which establishes that the transmission of information through noisy channels can be performed with arbitrary small probability of error. The problem of obtaining bounds on the probability of error arose with the coding theorem because the exact evaluation of the probability of error is very difficult to carry out in general.

ÍNDICE

Página

CAPÍTULO I

CANAIS DE COMUNICAÇÃO COM RUÍDO E RESULTADOS INTRODUTÓRIOS

1.1. INTRODUÇÃO

Inicialmente apresentamos o esquema de transmissão de mensagens através do "sistema de comunicação". E rigidamente ele se apresenta como no diagrama da Fig. (1.1.1)

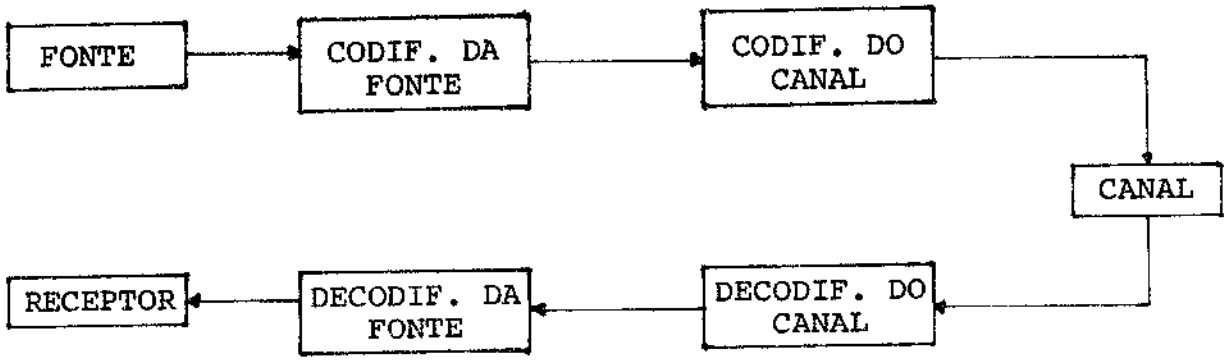

Fig. (1.1.1)

Façamos então a descrição do esquema no sentido de conhecermos seu mecanismo interno, o que nos será de grande valia. A fonte emite portanto qualquer mensagem m, $1 \leq m \leq M$. O codificador da fonte transforma esta mensagem numa palavra código, digamos de tamanho L, com elementos da base. Se a base possui D elemen

tos então haverá $M = D^L$ diferentes sequências de comprimento L e o codificador providencia uma palavra código para cada uma. A palavra código formada por elementos da base entra no codificador do canal a uma velocidade, ou taxa, de um dígito em cada τ_s segundos. O codificador do canal traduz então a palavra código para uma N-sequência, linguagem do canal, que passará a ter condições de transmití-la. Consideremos então o canal discreto no tempo e transmitindo cada dígito de seu alfabeto de entrada em τ_c segundos. O comprimento N do bloco codificador será tomado como a parte inteira de $L.\tau_s/\tau_c$, ou seja: $N = [L.\tau_s/\tau_c]$. E a velocidade ou taxa R de um código bloco é definida como sendo $R=(\log M)/N$. Então R bits (isto é, usando $\log_2$) é o número de dígitos binários entrando no codificador por dígito do canal transmitido. Em alguns casos, inclusive ao longo deste trabalho, é conveniente usar logarítmo natural e R, em unidades naturais (nats), será: $R=(\ln M)/N$. Notaremos que R não é uma entropia (embora ela possa ser interpretada como tal se os dígitos binários da fonte são independentes e igualmente prováveis). Como também R não é, em geral, a informação mútua média por canal.

Mas voltando ao estudo do mecanismo interno do esquema de transmissão vemos que, chegando a mensagem ao decodificador do canal, ele terá a missão de transformá-la em palavra código para o decodificador da fonte. Finalmente, o decodificador da fonte decifra essa palavra código que chegou a ele para uma mensagem, entregando-a ao receptor.

Vejamos agora o que irá captar o receptor. Se o canal não tem ruído o alfabeto de saída coincidirá sempre com o de entrada e mais: a mensagem recebida será sempre a mesma emitida m. Quando isto ocorre precisamos apenas conseguir um esquema de codificação que torne o comprimento médio das palavras código o menor possível, a fim de pouparmos tempo e custo de um modo geral. Entretanto o canal pode apresentar ruído sob qualquer forma ou aspecto. Este fato depende inclusive de seu estado. Como consequência desse ruído, possivelmente o alfabeto de saída diferirá do de entrada e mais fortemente a mensagem recebida m' diferirá de m.

A nossa preocupação principal, bem como um dos mais fortes objetivos da teoria da informação, será formular métodos que estude, discuta e controle os efeitos do ruído. O problema fundamental da comunicação é o de reproduzir, em um ponto qualquer, exatamente ou aproximadamente a mensagem de um outro ponto. Este problema fundamental (Berger [6]) pode ser separado em dois a saber:

i) Quanta informação será transmitida?

ii) Qual informação será transmitida?

O trabalho desenvolvido por Shannon [24] está mais ligado ao primeiro problema, qual seja o de selecionar codificadores de um conjunto de possíveis mensagens. Esta seleção tem que ser feita de uma maneira que as mensagens possam ser transmitidas corretamente sobre um canal de comunicação com ruído. O segundo problema delineado acima permaneceu esquecido algum tempo. Mas vol-

tou a ser abordado por Shannon [25] em 1959, além de outros pesquisadores.

Antes ainda de darmos outras definições essenciais do esquema da Fig. (1.1.1) trataremos agora de simplificá-lo, para um melhor entendimento. O que está nos interessando realmente é o problema do ruído no canal. Então consideraremos que a fonte já entrega a mensagem m ao canal perfeitamente codificada numa N-sequência $\vec{x}$ formada por elementos do alfabeto de entrada. E que o canal já transmite ao receptor uma N-sequência $\vec{y}$ formada por elementos do alfabeto de saída, ou seja:

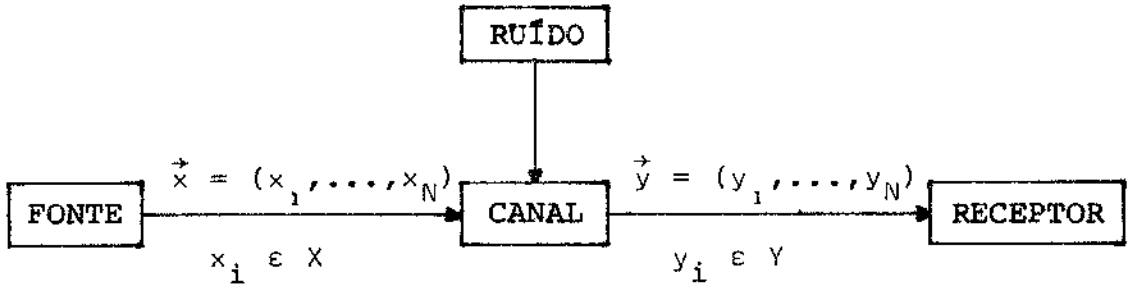

Fig. (1.1.2)

Como vemos, fonte, canal e receptor passam a falar a mesma linguagem no esquema da Fig. (1.1.2). Mas devido ao ruído no canal o receptor poderá decodificar a mensagem $\vec{y}$ que chega a ele como sendo uma $\vec{x}$ diferente da que foi enviada. Isto então se caracteriza como sendo um erro. Nos Capítulos II e III estudaremos esquemas de decodificação e condições no sentido de controlar com limites superior e inferior a probabilidade de cometermos este erro. Isto porque a probabilidade exata de erro é mui

to difícil, diríamos até impossível, de ser conseguida.

Shannon, em 1948, sugeriu a idéia de codificação aleatória e ele mesmo conseguiu limitar a probabilidade de erro em 1957 usando o fato de que ela é exponencial através da técnica de Chernoff (1952). Em 1961 Wolfowitz deu uma nova demonstração para o teorema fundamental através da formulação combinatória. Estas demonstrações clássicas podem ser vistas no Ash [2]. Nosso objetivo é dar uma apresentação simples e elegante para o Teorema Fundamental (Gallager [15]) e seu dual (Arimoto [1]).

1.2. OUTRAS DEFINIÇÕES

Na seção anterior apresentamos de uma maneira geral nosso problema. Entretanto precisamos formalizar as definições de certos entes aos quais já nos referimos e de outros que ainda serão citados, tais como:

- Entropia de Shannon e Informação mútua média
- Canal
- Capacidade do canal e o Teorema Fundamental
- Esquemas de decodificação
- Probabilidade de erro

1.2.A. ENTROPIA DE SHANNON E INFORMAÇÃO MÚTUA MÉDIA

Consideremos uma variável aleatória X discreta assumindo um número finito de valores

$$X = (x_1, \ldots, x_I)$$

e associemos a esta variável aleatória uma distribuição de probabilidade

$$\vec{p} = (p(x_1),\ldots,p(x_I));\ p(x_i) \geq 0\ ;\quad \sum_{i=1}^{I} p(x_i) = 1$$

Então a entropia de Shannon [16] da distribuição de probabilidade $\vec{p}$ é dada por

$$H(X) = H(p(x_1),\ldots,p(x_I)) = -\sum_{i=1}^{I} p(x_i)\ \log\ p(x_i)$$

onde a base do logarítmo será "2" se estivermos usando a base binária na codificação das possíveis mensagens que a variável pode emitir. Como a base do sistema de codificação pode ser arbitrária então a base do logarítmo no cálculo da entropia também o será. Na nossa dissertação utilizaremos sempre a base natural e a entropia será dada em nats.

Retornando ao nosso sistema discreto de comunicação da Fig. (1.1.2) observamos que as mensagens entram no canal codificadas em função do alfabeto de entrada $X = (x_1,\ldots,x_I)$ com uma distribuição $\vec{p}$ conhecida, saindo dele codificada em função do alfabeto de saída $Y = (y_1,\ldots,y_J)$ com uma distribuição $\vec{q}$ digamos desconhecida. Se conseguirmos a distribuição de $Y|X$ (i.e. Y dado X), ou seja, a matriz de transição do canal $[P(y_j|x_i)]$, $i = 1,\ldots,I$ e $j = 1,\ldots,J$, automaticamente teremos condições de determinar as probabilidades $p(x_i, y_j)$ da variável aleatória bidimensional (X, Y), $q(y_j)$ da variável aleatória Y, e $P(x_i|y_j)$ da variável aleatória $X|Y$, como se segue:

$$P(x_i, y_j) = p(x_i) \cdot P(y_j | x_i) \quad \text{ou} \quad p(i, j) = p(i) \cdot P(j|i)$$

$$q(y_j) = \sum_{i=1}^{I} p(x_i, y_j) \quad \text{ou} \quad q(j) = \sum_{i} p(i, j)$$

$$P(x_i|y_j) = \frac{p(x_i, y_j)}{q(y_j)} \quad \text{ou} \quad P(i|j) = \frac{p(i, j)}{q(j)}$$

Vemos então que podemos associar cinco diferentes entropias ao esquema, quais sejam:

i) Entropias marginais de X e Y por

$$H(X) = - \sum_{i=1}^{I} p(x_i) \log p(x_i) = - \sum_{i} p(i) . \log p(i)$$

$$H(Y) = - \sum_{j=1}^{J} q(y_j) \log q(y_j) = - \sum_{j} q(j) . \log q(j)$$

ii) Entropia conjunta de (X, Y)

$$H(X, Y) = - \sum_{i=1}^{I} \sum_{j=1}^{J} p(x_i, y_j) \log p(x_i, y_j) =$$

$$= - \sum_{i} \sum_{j} p(i,j) \log p(i,j)$$

iii) Entropias condicionais X|Y e Y|X

$$H(X|Y) = -\sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i,y_j) \log P(x_i|y_j) =$$

$$= -\sum_{i}\sum_{j} p(i,j) \log P(i|j)$$

$$H(Y|X) = -\sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i,y_j) \log P(y_j|x_i) =$$

$$= -\sum_{i}\sum_{j} p(i,j) \log P(j|i)$$

A função do receptor é extrair, apesar do ruído, todas possíveis informações sobre o sinal transmitido. A informação que Y providencia sobre X pode ser averiguada pela incerteza que Y remove sobre X, ou seja:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Isto é simétrico em X e Y e portanto definida como informação mútua mëdia, Ash [2]. É fácil verificar que I(X; Y) é sempre não negativa.

1.2.B. CANAL

Intuitivamente um canal é um ente capaz de transmitir informação. Já vimos que restringiremos nossa atenção a canais discretos onde, devido ao ruído, será transmitida informação pela

emissão de símbolos mensagem escolhidos de um alfabeto $X = \{X_1, \ldots, X_I\}$ e chegada ao receptor de símbolos pertencentes ao alfabeto de saída $Y = \{Y_1, \ldots, Y_J\}$. O canal de transmissão pode ser definido em termos destes conjuntos de entrada e saída além da matriz de probabilidade de transição $[P(j|i)]$, $j = 1,\ldots,J$ e $i = 1,\ldots,I$, onde $P(j|i)$ denota a probabilidade de receber y_j quando x_i é emitido. A taxa R do canal é a quantidade de informação que pode ser processada através dele por letra ou unidade de tempo, conforme o caso. Então $R=I(X;Y)$.

O canal discreto mais simples é o simétrico binário. Seus alfabetos de entrada e saída consistem dos dígitos binários "0" e "1" e a matriz de transição é dada por

$$[P(j|i)] = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix} \end{array}$$

Um outro canal simples é o binário com rasura (erasure). Neste canal temos $X = \{0, 1\}$, $Y = \{0, 1, e\}$ e a matriz de transição

$$[P(j|i)] = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & e \end{array} \\ \begin{bmatrix} 1-p-q & p & q \\ p & 1-p-q & q \end{bmatrix} \end{array}$$

A teoria de codificação em canais trata das possibilidades de transmissão correta de mensagens através de canais com ruído . Intuitivamente estas mensagens podem ser refletidas como a representação da fonte de informação em alguma forma padrão (codificação da fonte). Elas podem ser sequências binárias corretas representando blocos de N saídas consecutivas de uma fonte de informação transmitidas sobre o canal. Se $M = 2^{NR}$ é o número de possíveis mensagens então isto significa que em média R digitos binários podem ser transmitidos por um canal, sendo R a taxa de transmissão.

1.2.C. CAPACIDADE DO CANAL E O TEOREMA FUNDAMENTAL

A capacidade do canal é, por definição, a taxa máxima sobre ele, ou seja:

$$C = \max_{\vec{p}(x)} I(X;Y)$$

Como vemos acima a maximização é feita com respeito a todas possíveis escolhas das distribuições de entrada. A principal significância da capacidade aparece no teorema de codificação em canal com ruído de Shannon (veja [16]).

Este teorema, que é o melhor resultado conseguido por Shannon, estabelece que a transmissão de informação através de canais com ruído pode ser conseguida com probabilidade de erro

arbitrariamente pequena quando a taxa de transmissão R é menor que a capacidade C do canal.

Trabalhos adicionais neste sentido foram feitos por Dobrushin [8], Feinstein [12], Wolfowitz [31] e outros.

1.2.D. ESQUEMAS DE DECODIFICAÇÃO

Em muitos casos práticos, (Feinstein [12]), alguém que está recebendo uma mensagem fica com a responsabilidade de decidir, com base no sinal recebido, qual mensagem símbolo foi transmitida. O receptor fica então com um clássico problema de inferência estatística. Ele terá, com alguma base essencialmente subjetiva, que determinar a importância relativa dos vários tipos de erro que pode tomar. Somente então, em geral, pode ele fixar um esquema para decodificar, para cada símbolo recebido, qual mensagem símbolo ele terá para melhor concluir o que foi emitido. Uma regra de decodificação pode ser definida como uma aplicação do conjunto de N-sequências de saída do canal no conjunto consistindo das mensagens emitidas. O objetivo da operação de decodificação é identificar a mensagem transmitida pela evidência verificada na N-sequência de saída. Para formular o problema mais precisamente definamos que o canal tem um alfabeto de entrada $x_1,\ldots,x_I$, um alfabeto de saída $y_1,\ldots,y_J$ e uma matriz de transição $[P(j|i)]$. Suponhamos que alguma mensagem é transmitida. A decodificação, ou esquema de decisão, é uma atribuição

a cada símbolo de saída y_j de um símbolo de entrada x_j'. Sejam X_N e Y_N denotando, respectivamente, o conjunto de todas N-se - quências de entrada e saída do canal. Vários critérios (veja [16]) podem ser usados na decodificação e alguns são dados a baixo.

i) Mínima Probabilidade de Erro

Seja $P(\vec{y}|\vec{x}_m)$ a probabilidade de receber uma N-sequência $\vec{y} \in Y_N$ dado que a N-sequência de entrada $\vec{x}_m \in \vec{X}_N$ é transmitida. A regra de decodificação de mínima probabilidade de erro é justamente aquela que minimiza a probabilidade de erro de decodificação para um dado conjunto de mensagens, conjunto de palavras código, e canal. Ela será então definida por: decodificar a N-sequência recebida $\vec{y}$ em $\vec{x}_m$, para a qual

$$P(\vec{x}_m,|\vec{y}) > P(\vec{x}_m|\vec{y}) \text{ , para todo } m' \neq m$$

Este esquema de decisão, que sempre escolhe $\vec{x}$ cuja probabilidade condicional, no momento, é a maior, é sempre chamado de "observador ideal".

ii) Decodificação de Máxima-Verossimilhança

O decodificador de máxima-verossimilhança é um tipo alternativo de regra definido por: dado $\vec{y}$, escolher $\vec{x}_m$, tal que

$$P(\vec{y}|\vec{x}_{m'}) > P(\vec{y}|\vec{x}_{m}) \quad \text{para todo} \quad m' \neq m$$

A óbvia vantagem do decodificador de máxima-verossimilhança é que ele pode ser usado quando as probabilidades a priori das mensagens são desconhecidas. Se as mensagens têm probabilidades a priori iguais então a regra de decodificação de mínima probabilidade de erro é equivalente à de máxima-verossimilhança.

iii) Decodificação de Custo Mínimo

Um outro tipo de regra, usual quando custos desiguais são associados com diferentes espécies de erro, é a decodificação de custo mínimo. Aqui $\vec{y}$ é decodificado numa mensagem m que minimiza o custo médio.

iv) Decodificação em Lista

Algumas vezes é conveniente considerar decodificação em listagem, onde o decodificador, ao contrário da aplicação das sequências recebidas em um simples inteiro, ou uma simples mensagem, o faz em uma lista de mensagens, ou uma lista de inteiros m, $1 \leq m \leq M$, sendo M o número total de palavras código. Decodificação em lista, ou em rol, foi considerada primeiramente por Elias [10] para o canal simétrico binário. Shannon, Gallager e Berlekamp [26], Ebert [9] e Forney [14] utilizaram o esquema de

decodificação em rol na obtenção de limites na probabilidade de erro e de rasura.

1.2.E. PROBABILIDADE DE ERRO

Como tomado por Shannon [24], Gallager [16] e outros, consideremos um canal discreto com alfabeto de entrada X, alfabeto de saída Y e matriz de transição $[P(j|i)]$ como em (1.2.B). Sejam X_N e Y_N os conjuntos de todas sequências de comprimento N que podem ser transmitidas e recebidas respectivamente sobre um dado canal. E seja $P(\vec{y}|\vec{x})$ a probabilidade de receber $\vec{y} \in Y_N$ dado que $\vec{x} \in X_N$ foi transmitida.

Para um código bloco de comprimento N e taxa R tal canal possui um conjunto de e^{NR} N-sequências de entrada ou palavras código (Forney [13], Gallager [16], Fano [11]) como vemos abaixo

$$\vec{x}_m = (x_{m1}, x_{m2}, \ldots, x_{mN}) \quad , \quad 1 \leq m \leq e^{NR}$$

onde R é a taxa de código natural por símbolo do canal.

Um codificador é um esquema mecânico que admite um dos e^{NR} comandos de uma fonte de dados e gera a correspondente N-sequência de entrada a ser transmitida pelo canal. Um decodificador é um ente que observa uma N-sequência de saída, processa esta sequência e apresenta o resultado para usar na forma desejada.

O evento no qual o estimador não é idêntico à palavra código de entrada é chamado um erro e a probabilidade deste evento é a probabilidade de erro.

A probabilidade de erro P_e depende do código, do canal e da estratégia usada pelo decodificador. Se o decodificador é determinístico então sua estratégia é descrita como uma aplicação do conjunto de todas N-sequências recebidas $\vec{y}$ na palavra código $\vec{x}_m$ e é especificada pela listagem do conjunto Y_m de N-sequências $\vec{y}$ que resultam no estimador decodificado de $\vec{x}_m$. Se assumimos que o código, canal e decodificador são todos especificados, então a probabilidade de erro (Forney [13]) será dada por

$$P_e = \sum_{m=1}^{M} p(\vec{x}_m) \cdot P(\vec{y} \notin Y_m | \vec{x}_m \text{ é transmitido}) =$$

$$= \sum_{m=1}^{M} \sum_{\vec{y} \notin Y_m} p(\vec{x}_m) \cdot P(\vec{y} | \vec{x}_m)$$

Suponhamos que, se o ruído é particularmente nocivo e o decodificador tem a opção de não decidir sobre todos estimadores, então o resultado da saída que o receptor não estima é chamado uma rasura (Forney [13]) e a probabilidade deste evento a probabilidade de rasura.

Nos Capítulos II e III iremos justamente determinar limites superior e inferior para a probabilidade de erro no caso de canal discreto sem memória. Mas podemos generalizar os resulta -

dos destes capítulos para Canais Contínuos (sem e com restrições sobre a entrada), para uma classe de canais desconhecidos e para Teoria de Codificação.

CAPÍTULO II

LIMITE SUPERIOR PARA A PROBABILIDADE DE ERRO E O TEOREMA FUNDAMENTAL DA INFORMAÇÃO

2.1. INTRODUÇÃO

Introduzimos o conceito de probabilidade de erro utilizando um C.D.S.M($X, Y, [P(j|i)], X_N, Y_N$) na seção (1.2.E) do capítulo anterior, onde:

X = alfabeto de entrada

Y = alfabeto de saída

$[P(j|i)]$ = matriz de transição

X_N = conjunto de todas N-sequências de entrada

Y_N = conjunto de todas N-sequências de saída

Também na seção (1.2.D) nos referimos ao esquema de decodificação de máxima-verossimilhança, ou seja: a N-sequência recebida $\vec{y} \in Y_N$ é decodificada como $\vec{x}_m, \in X_N$ se

$$P(\vec{y}|\vec{x}_{m'}) > P(\vec{y}|\vec{x}_m), \text{ para todo } m' \neq m \quad e \quad 1 \leq m' \leq M \quad (2.1.1)$$

Utilizaremos estes conceitos nas seções seguintes deste capítulo.

2.2. LIMITE SUPERIOR DA PROBABILIDADE DE ERRO

Seja P_{em} a probabilidade de erro quando $\vec{x}_m \in X_N$ é transmitida e escolhemos a regra de decodificação dada por (2.1.1) . Então:

$$P_{em} = \sum_{\vec{y} \in Y_N} P(\vec{y}|\vec{x}_m)\ \psi_m\ (\vec{y}) \qquad (2.2.1)$$

onde

$$\psi_m\ (\vec{y}) = \begin{cases} 1, \text{ se } P(\vec{y}|\vec{x}_m) \leq P(\vec{y}|\vec{x}_{m'}) , \quad m' \neq m \\ 0, \text{ caso contrário} \end{cases}$$

Observamos que P_{em} acima é uma variável aleatória em $[0, 1]$ cuja medida de probabilidade desconhecemos. Isto nos dificulta calcularmos probabilidades exatas para P_{em}. Se definirmos no entanto a probabilidade média de erro $\overline{P}_e$ como a média de P_{em} sobre todas palavras código, então um limite superior exponencial para $\overline{P}_e$ pode ser obtido se limitarmos superiormente de uma maneira conveniente a função $\psi_m(\vec{y})$.

TEOREMA (2.2.1) - Seja um C.D.S.M$(X, Y, [P(j|i)], X_N, Y_N)$ com uma taxa R e uma distribuição de entrada $\vec{p} = (p(x_1), \ldots, (x_I))$. Então, para qualquer bloco de comprimento N e para a fonte com $M = \lceil e^{NR} \rceil$ palavras código, existe um código para o qual a probabilidade média de erro $\overline{P}_e$, sob o esquema (2.1.1), é limitada por

$$\overline{P_e} \leq \exp. -N \left[-\rho R + E_0(\rho, \vec{p}) \right] , \quad 0 \leq \rho \leq 1$$

com

$$E_0(\rho, \vec{p}) = -\ln \sum_j \left[\sum_i p(i) \left(P(j|i) \right)^{1/1+\rho} \right]^{1+\rho} \qquad (2.2.2)$$

onde ρ é um número arbitrário.

PROVA: É fácil de se observar que

$$\psi_m(\vec{y}) \leq \left[\frac{\sum_{m' \neq m} \left(P(\vec{y}|\vec{x}_{m'}) \right)^{1/1+\rho}}{\left(P(\vec{y}|\vec{x}_m) \right)^{1/1+\rho}} \right]^{\rho} , \quad \rho > 0$$

Portanto, de (2.2.1) temos que

$$P_{em} \leq \sum_{\vec{y} \in Y_N} \left(P(\vec{y}|\vec{x}_m) \right)^{1/1+\rho} \cdot \left[\sum_{m' \neq m} \left(P(\vec{y}|\vec{x}_{m'}) \right)^{1/1+\rho} \right]^{\rho} ,$$

para algum $\rho > 0$ ... (2.2.3)

A inequação (2.2.3) assegura um limite para P_{em} num código particular. Simplificamos o limite em P_{em} pela averiguação, sobre uma escolha apropriada, do conjunto de códigos. Definamos uma medida de probabilidade $\vec{p}(\vec{x})$ no conjunto X_N das possíveis N-sequências de entrada do canal. Geraremos um conjunto de códigos pela escolha de cada palavra código independentemente, de acordo com a medida de probabilidade $\vec{p}(\vec{x})$, tal que a probabilidade asso-

ciada com o código consistindo das palavras código $\vec{x}_1,\ldots,\vec{x}_M$ é $\prod_{m=1}^{M}\vec{p}(\vec{x}_m)$. É evidente que pelo menos um código no conjunto terá a probabilidade de erro que será menor que a probabilidade média de erro do conjunto. Usando uma barra para representar a média do conjunto de códigos temos

$$\overline{P_{em}} \leq \sum_{\vec{y}\in Y_N} \overline{\left(P(\vec{y}|\vec{x}_m)\right)^{1/1+\rho}\left[\sum_{m'\neq m}\left(P(\vec{y}|\vec{x}_{m'})\right)^{1/1+\rho}\right]^{\rho}}$$

Se impusermos agora a restrição adicional de que $\rho \leq 1$ então ficaremos com

$$\overline{P_{em}} \leq \sum_{\vec{y}\in Y_N} \overline{\left(P(\vec{y}|\vec{x}_m)\right)^{1/1+\rho}\left[\sum_{m'\neq m}\left(P(\vec{y}|\vec{x}_{m'})\right)^{1/1+\rho}\right]^{\rho}} =$$

$$= \sum_{\vec{y}\in Y_N} \overline{\left(P(\vec{y}|\vec{x}_m)\right)^{1/1+\rho}}\,\overline{\left[\sum_{m'\neq m}\left(P(\vec{y}|\vec{x}_{m'})\right)^{1/1+\rho}\right]^{\rho}} =$$

$$\leq \sum_{\vec{y}\in Y_N} \overline{\left(P(\vec{y}|\vec{x}_m)\right)^{1/1+\rho}}\left[\overline{\sum_{m'\neq m}\left(P(\vec{y}|\vec{x}_{m'})\right)^{1/1+\rho}}\right]^{\rho} =$$

$$= \sum_{\vec{y}\in Y_N} \overline{\left(P(\vec{y}|\vec{x}_m)\right)^{1/1+\rho}}\left[\sum_{m'\neq m}\overline{\left(\left(P(\vec{y}|\vec{x}_{m'})\right)^{1/1+\rho}\right)}\right]^{\rho} \qquad (2.2.4)$$

Mas desde que as palavras código são escolhidas com probabilidade $\vec{p}(\vec{x})$

$$\overline{\left(P(\vec{y}|\vec{x}_m)\right)^{1/1+\rho}} = \sum_{\vec{x}\in X_N} \vec{p}(\vec{x}) \left(P(\vec{y}|\vec{x})\right)^{1/1+\rho} \qquad (2.2.5)$$

Portanto, em vista de (2.2.5) e (2.2.4) temos que

$$\overline{P_{em}} \leq (M-1)^{\rho} \cdot \sum_{\vec{y}\in Y_N} \left[\sum_{\vec{x}\in X_N} \vec{p}(\vec{x}) \cdot \left(P(\vec{y}|\vec{x})\right)^{1/1+\rho} \right]^{1+\rho} \qquad (2.2.6)$$

para qualquer $0 < \rho \leq 1$

O limite em (2.2.6) é válido para todas escolhas de $\vec{p}(\vec{x})$ e todos os ρ's, $0 < \rho \leq 1$, e se aplica em algum canal discreto. Desde que o canal seja sem memória e tivermos $\vec{x} = (x_1, \ldots, x_I)$ e $\vec{y} = (y_1, \ldots, y_J)$ então

$$P(\vec{y}|\vec{x}) = \prod_{n=1}^{N} P(y_n|x_n) \qquad (2.2.7)$$

para todos $\vec{x} \in X_N$, $\vec{y} \in Y_N$, e para todo N.

Consideraremos agora somente a classe de conjunto de códigos na qual cada letra de cada palavra código é escolhida independentemente de todas as outras letras com a medida de probabilidade $\vec{p}(\vec{x})$, $\vec{x} \in X_N$. Então

$$\vec{p}(\vec{x}) = \prod_{n=1}^{N} p(x_n) \qquad (2.2.8)$$

Usando agora (2.2.7) e (2.2.8) obtemos

$$\overline{P_{em}} \leq (M-1)^{\rho} \sum_{\vec{y}\in Y_N} \left[\sum_{\vec{x}\in X_N} \prod_n p(x_n) \left(P(y_n|x_n)\right)^{1/1+\rho} \right]^{1+\rho} \quad (2.2.9)$$

$$= (M-1)^{\rho} \sum_{\vec{y}\in Y_N} \left[\prod_n \sum_{x_n\in X} p(x_n) \left(P(y_n|x_n)\right)^{1/1+\rho} \right]^{1+\rho} \quad (2.2.10)$$

Vemos que o resultado (2.2.10) segue de (2.2.9) porque o termo entre colchetes em (2.2.10) é o produto de somas e é igual ao termo entre colchetes em (2.2.9) pela regra aritmética usual para multiplicar somas de produtos. Tomando então o produto fora dos colchetes em (2.2.10) aplicamos a mesma regra novamente e obtemos

$$\overline{P_{em}} \leq (M-1)^{\rho} \prod_n \sum_{y_n\in Y} \left[\sum_{x_n\in X} p(x_n) \left(P(y_n|x_n)\right)^{1/1+\rho} \right]^{1+\rho} \quad \ldots (2.2.11)$$

Notando que todos os termos no produto são idênticos, e incluindo o caso trivial $\rho = 0$, podemos agora simplificar a notação de (2.2.11) do seguinte modo

$$\overline{P_{em}} \leq (M-1)^{\rho} \left[\sum_j \left(\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho} \right)^{1+\rho} \right]^N , \quad 0 \leq \rho \leq 1 \quad \ldots (2.2.12)$$

Podemos agora majorar $(M-1)$ por $M = \left[e^{NR}\right]$ e reescrever (2.2.12)

como sendo

$$\overline{P_B} \leq \exp - N\left[-\rho R - \ln \sum_j \left(\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho}\right)^{1+\rho}\right] =$$

$$= \exp - N\left[-\rho R + E_0(\rho, \vec{p})\right] \qquad (2.2.13)$$

usando a expressão (2.2.2).

Desde que o lado direito de (2.2.13) é independente de m ele é um limite para o conjunto das probabilidades médias de erro e é independente das probabilidades com as quais as palavras código são usadas. O resultado agora segue do fato de que pelo menos um código no conjunto terá uma probabilidade de erro menor que a média.

O teorema (2.2.1) é verdadeiro para qualquer ρ, $0 \leq \rho \leq 1$, e todos vetores de probabilidade $\vec{p} = (p(x_1), \ldots, p(x_I))$. Entretanto podemos obter um limite mais refinado pela maximização do expoente sobre ρ e $\vec{p}$. Isto nos dá o seguinte corolário simples.

COROLÁRIO (2.2.2) - Sob as condições do teorema (2.2.1) existe um código para o qual

$$\overline{P_e} \leq \exp - NE(R) \qquad (2.2.14a)$$

onde

$$E(R) = \max_{\rho, \vec{p}} \left[-\rho R + E_0(\rho, \vec{p}) \right] \qquad (2.2.14b)$$

e a maximização é tomada sobre todo ρ, $0 \leq \rho \leq 1$, e sobre todos vetores de probabilidade $\vec{p}$.

Podemos modificar este resultado, o que é imediato, para ter um limite na probabilidade de erro aplicado a cada palavra código separadamente, ao contrário do limite da média, no seguinte corolário.

COROLÁRIO (2.2.3) - Sob as condições do teorema (2.2.1) existe um código tal que, para todo m, $1 \leq m \leq M$, a probabilidade de erro quando a m-ésima palavra código é transmitida é limitada por

$$P_{em} \leq 4 \exp - NE(R) \qquad (2.2.15)$$

onde E(R) é dado por (2.2.14b)

PROVA: Escolhamos um código com $M' = 2M$ palavras código que satisfaça ao colorário (2.2.2) enquanto a fonte usa as 2M palavras código com probabilidades iguais. Se removemos as M palavras no código para as quais P_{em} é grande, então, desde que seja impossível sobre a metade das palavras no código termos uma probabilidade de erro maior que o dobro da média, as palavras código restan

tes satisfarão

$$P_{em} \leq 2 \exp - NE(R') \qquad (2.2.16)$$

onde R', a nova taxa, é dada por

$$R' = \frac{\ln 2M}{N} = \frac{\ln M}{N} + \frac{\ln 2}{N} = R + \frac{\ln 2}{N}$$

Agora, desde que $0 \leq \rho \leq 1$, (2.2.14) nos dá

$$E(R') \geq E(R) - \frac{\ln 2}{N} \qquad (2.2.17)$$

então o resultado em (2.2.15) segue de (2.2.16) pelo uso de (2.2.17).

Na próxima seção estudaremos as propriedades da função "Confiança" E(R).

2.3. PROPRIEDADES DO EXPOENTE DE CODIFICAÇÃO ALEATÓRIA E(R)

No sentido de entendermos o comportamento de E(R), primeiro analisaremos $E_0(\rho, \vec{p})$ como função de ρ. O teorema que veremos a seguir nos assegurará sempre o aspecto do gráfico da Fig. (2.3.1)

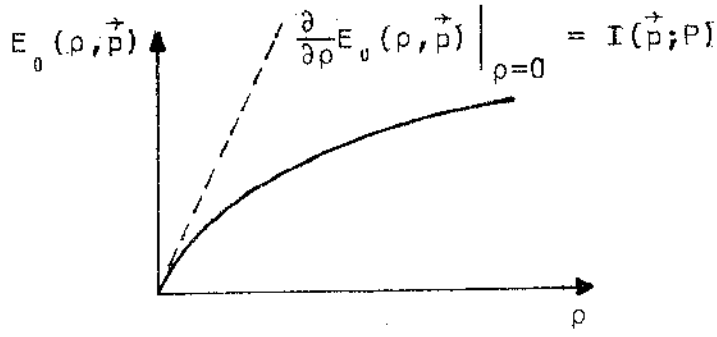

Fig. (2.3.1) - Gráfico de $E_0(\rho, \vec{p})$

A informação mútua mêdia $I(\vec{p}; P)$, como função da distribuição $\vec{p}$ e da probabilidade de transição do canal P, será dada por

$$I(\vec{p};\ P) = \sum_i \sum_j p(i)\ P(j|i).\ \ln \frac{P(j|i)}{\sum_k p(k)\ P(j|k)}$$

e simplesmente interpretado como a informação mútua mêdia por dígito no conjunto dos códigos.

TEOREMA (2.3.1) - Seja a probabilidade de entrada $\vec{p}$ e o canal discreto sem memória tal que $I(\vec{p};\ P) > 0$. Então $E_0\ (\rho,\ \vec{p})$ em (2.2.2) tem as seguintes propriedades:

$$E_0(\rho,\ \vec{p}) \geq 0\ , \qquad \rho \geq 0 \tag{2.3.1}$$

$$I(\vec{p};\ P) \geq \frac{\partial}{\partial \rho} E_0(\rho,\ \vec{p}) > 0 \quad ; \quad \rho \geq 0 \tag{2.3.2}$$

$$\frac{\partial^2}{\partial \rho^2} E_0\ (\rho,\ \vec{p}) \leq 0 \quad ; \quad \rho \geq 0 \tag{2.3.3}$$

onde em (2.3.1) e (2.3.2) as igualdades são asseguradas para $\rho=0$ e em (2.3.3) a igualdade é assegurada se para todo (i, j) tal que $p(i).P(j|i) > 0$ temos que

$$\ln \frac{P(j|i)}{\sum_k p(k).P(j|k)} = I(\vec{p};\ P) \tag{2.3.3a}$$

Isto é, se a variável aleatória informação mútua tem variância zero.

Por inspeção de (2.2.2) vemos que $E_0(0,\ \vec{p}) = 0$, e pela diferenciação encontramos facilmente que

$$\frac{\partial}{\partial \rho} E_0(\rho,\ \vec{p}) \Big|_{\rho=0} = I(\vec{p};\ P)$$

Como necessitamos do resultado do teorema (2.3.1) para tornar fácil maximizar $E_0\ (\rho,\ \vec{p}) - \rho R$ sobre ρ para um dado $\vec{p}$, apresentaremos o restante de sua prova no Apêndice (A.1). Definamos

agora

$$E(R, \vec{p}) = \max_{0 \leq \rho \leq 1} \left[E_0(\rho, \vec{p}) - \rho R \right] \qquad (2.3.4)$$

A equação para o ponto estacionário de $E_0(\rho, \vec{p}) - \rho R$ com respeito a ρ é

$$\frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) - R = 0 \qquad (2.3.5)$$

Desde que $\frac{\partial^2}{\partial \rho^2} E_0(\rho, \vec{p}) \leq 0$, alguma solução de (2.3.5) no domínio $0 \leq \rho \leq 1$ maximiza (2.3.4). Adicionalmente, desde que $\frac{\partial}{\partial \rho} E_0(\rho, \vec{p})$ é contínua e decrescente com respeito a ρ, a solução de (2.3.5) com $0 \leq \rho \leq 1$ existe se

$$\left. \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \right|_{\rho=1} \leq R \leq \left. \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \right|_{\rho=0} = I(\vec{p}; P) \qquad (2.3.6)$$

O ponto $\left. \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \right|_{\rho=1}$ é chamado a **razão crítica** R_c para $\vec{p}$ dado.

Para R no domínio acima será mais conveniente usar (2.3.5) para relatar R e $E(R, \vec{p})$ parametricamente em termos de ρ

$$R = \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \quad ; \quad 0 \leq \rho \leq 1 \tag{2.3.7a}$$

$$E(R, \vec{p}) = E_0(\rho, \vec{p}) - \rho \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \tag{2.3.7b}$$

Diferenciando cada equação em (2.3.7) obtemos

$$\frac{\partial R}{\partial \rho} = \frac{\partial^2}{\partial \rho^2} E_0(\rho, \vec{p}) \tag{2.3.8a}$$

$$\frac{\partial}{\partial \rho} E(R, \vec{p}) = -\rho \frac{\partial^2}{\partial \rho^2} E_0(\rho, \vec{p}) \tag{2.3.8b}$$

Então, como $0 \leq \rho \leq 1$, R decresce monotonamente de $I(\vec{p}; P)$ a $\frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \Big|_{\rho=1}$ e $E(R, \vec{p})$ cresce monotonamente de zero a $E_0(1, \vec{p}) - \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \Big|_{\rho=1}$. Diferenciando (2.3.7b) em relação a R obtemos

$$\frac{\partial}{\partial R} E(R, \vec{p}) = -\rho \tag{2.3.9}$$

Então o parâmetro ρ é interpretado como a magnitude da inclinação do gráfico de $E(R, \vec{p})$ x R

Para $R < \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \Big|_{\rho=1}$ então $E_0(\rho, \vec{p}) - \rho R$ é maximizado (sobre $0 \leq \rho \leq 1$) por $\rho = 1$, assegurando

$$E(R, \vec{p}) = E_0(1, \vec{p}) - R \tag{2.3.10}$$

Finalmente, no interessante caso onde $R > I(\vec{p}; P)$, $E_0(\rho, \vec{p}) - \rho R$ é maximizado por $\rho = 0$, assegurando $E(R, \vec{p}) = 0$.

Resumindo, para R no domínio dado por (2.3.6), $E(R, \vec{p})$ e R são relatados por (2.3.7). Para pequenos valores de R, $E(R, \vec{p})$ e R são relacionados pela equação linear (2.3.10), e para grandes valores $E(R, \vec{p}) = 0$. Como uma função de R, $E(R, \vec{p})$ é estritamente decrescente e positiva para todo $R < I(\vec{p}; P)$.

Consideraremos agora o caso especial onde $\frac{\partial^2}{\partial \rho^2} E_0(\rho, \vec{p}) = 0$. De (2.3.3a), no teorema (2.3.1), vemos que esta relação será satisfeita para todo $\rho \geq 0$ se ela é satisfeita para qualquer $\rho \geq 0$. Neste caso, $\frac{\partial}{\partial \rho} E_0(\rho, \vec{p})$ é uma constante e o domínio sobre o qual (2.3.6) é satisfeita é um ponto, veja figura (2.3.2)

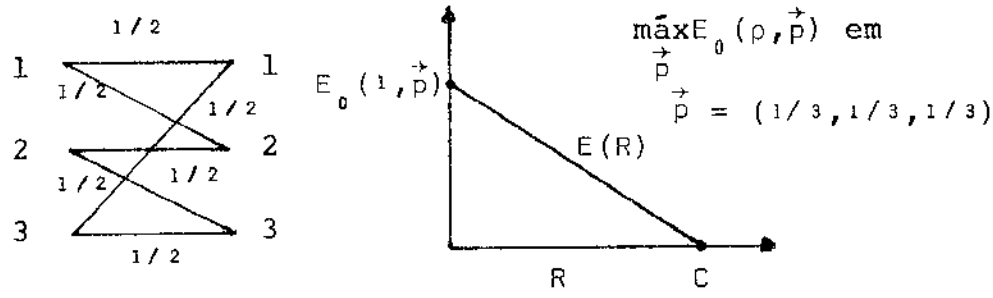

Fig. (2.3.2)

Caso especial onde $\frac{\partial^2}{\partial \rho^2} E_0 = 0$

Este caso especial é sem dúvida patológico e ocorre somente para canais "noiseless" [para o qual $H(X|Y) = 0$] e para alguns canais peculiares tais como na figura (2.3.2).

Para o caso usual, no qual $\frac{\partial^2}{\partial \rho^2} E_0(\rho, \vec{p}) < 0$, as equações paramétricas de (2.3.7) se aplicam sobre um domínio não zero de taxas. De (2.3.9) e (2.3.7) temos $\frac{\partial^2}{\partial R^2} E(R, \vec{p}) = - \left[\frac{\partial^2}{\partial \rho^2} E_0(\rho, \vec{p}) \right]^{-1} > 0$ e, então, $E(R, \vec{p})$ é estritamente convexa U em R sobre este domínio de R. Desde que $\frac{\partial^2}{\partial \rho^2} E(R, \vec{p}) = 0$ fora deste domínio temos que $E(R, \vec{p})$ é convexa U em R para todo $R \geq 0$.

O expoente de codificação aleatório E(R) pode agora ser relacionado a $E(R, \vec{p})$ por

$$E(R) = \max_{\vec{p}} E(R, \vec{p})$$

Isto é, um máximo sobre o conjunto de funções que são convexas U e decrescentes em R. É fácil de ver que a função maximização é também convexa U e decrescente em R. Também, para a distribuição de probabilidade $\vec{p}$ que assegura capacidade no canal, $E(R, \vec{p})$ é positivo para $R < I(\vec{p}; P) = C$ e, portanto, $E(R)$ é positivo para $R < C$. Isto prova o Teorema Fundamental, que se segue.

TEOREMA (2.3.2) - (Teorema de Codificação em Canal com Ruído) - Para qualquer canal discreto sem memória o expoente $E(R)$ [veja (2.2.14a) e (2.2.14b)] é convexa U, decrescente, função positiva de R para $0 \leq R < C$.

Uma interessante interpretação gráfica da maximização de $E_0(\rho, \vec{p}) - \rho R$ sobre ρ e $\vec{p}$ pode ser obtida por observação que, para ρ e $\vec{p}$ fixados, $E_0(\rho, \vec{p}) - \rho R$ é uma função linear de R com inclinação $-\rho$. Então $E(R, \vec{p})$, como vemos na figura (2.3.3), é o menor limite superior da família de retas geradas por diferentes valores de ρ, $0 \leq \rho \leq 1$. Desta construção, vemos que $E_0(\rho, \vec{p})$ é a taxa zero interseção da tangente a $E(R, \vec{p})$ de inclinação $-\rho$. A convexidade U de $E(R, \vec{p})$ também segue imediatamente desta construção.

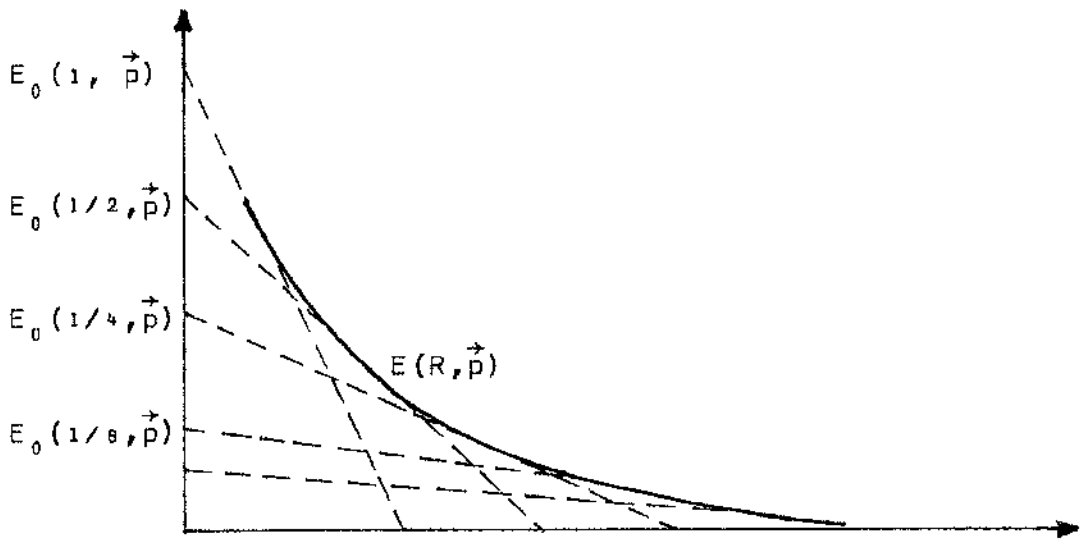

Fig. (2.3.3)

$E(R, \vec{p})$ como limite superior das funções lineares $E_0(\rho, \vec{p}) - \rho R$ para valores de ρ fixados.

Para maximizar $E_0(\rho, \vec{p}) - \rho R$ analiticamente sobre ρ e $\vec{p}$ é mais conveniente fazê-lo primeiro sobre $\vec{p}$.

$$E(R) = \max_{0 \leq \rho \leq 1} \left[-\rho R + \max_{\vec{p}} E_0(\rho, \vec{p}) \right]$$

A função $E_0(\rho, \vec{p})$ não é convexa U de $\vec{p}$ mas, felizmente, se definirmos $F(\rho, \vec{p})$ como sendo

$$F(\rho, \vec{p}) = \exp\left[-E_0(\rho, \vec{p})\right] = \sum_j \left[\sum_i p(i) \left(p(j|i)\right)^{1/1+\rho} \right]^{1+\rho}$$

... (2.3.11)

então o $\vec{p}$ que minimiza $F(\rho, \vec{p})$ maximizará $E_0(\rho, \vec{p})$.

TEOREMA (2.3.3) - Para qualquer $\rho \geq 0$, $F(\rho, \vec{p})$ como dada por (2.3.11) é uma função convexa U de $\vec{p}$ na região onde $\vec{p}$ é um vetor de probabilidade. Condições necessárias e suficientes ao vetor de probabilidade $\vec{p}$ que minimiza $F(\rho, \vec{p})$, e portanto maximiza $E_0(\rho, \vec{p})$, são

$$\sum_j (P(j|i))^{1/1+\rho} (\alpha_j(\vec{p}))^{\rho} \geq \sum_j (\alpha_j(\vec{p}))^{1+\rho} \quad , \text{ para todo } i \qquad \ldots (2.3.12)$$

com igualdade para todo i tal que $p(i) > 0$. A função $\alpha_j(\vec{p})$ é dada por

$$\alpha_j(\vec{p}) = \sum_i p(i) \; (P(j|i))^{1/1+\rho} \qquad (2.3.13)$$

PROVA: Para $\rho \geq 0$, $(\alpha_j(\vec{p}))^{1+\rho}$ é uma função convexa U se $\alpha(\vec{p}) \geq 0$, ou seja, sua segunda derivada é não-negativa. Portanto, se $\alpha_j(\vec{p})$ é uma função linear de $\vec{p}$, então se segue, da definição de convexidade, que $(\alpha_j(\vec{p}))^{1+\rho}$ é convexa U de $\vec{p}$. Disto, temos que

$$F(\rho, \vec{p}) = \sum_j (\alpha_j(\vec{p}))^{1+\rho}$$

é uma função convexa U de $\vec{p}$.

Usando o teorema do Apêndice (B-1) vemos que as condições necessárias e suficientes ao vetor de probabilidade $\vec{p}$ para minimizar $F(\rho, \vec{p})$ são

$$\frac{\partial}{\partial p(i)} F(\rho, \vec{p}) \geq A \ , \quad \text{para todo i, com igualdade se } p(i)>0.$$

Avaliando $\frac{\partial}{\partial p(i)} F(\rho, \vec{p})$ e dividindo por $(1 + \rho)$ obtemos (2.3.12). A constante no lado direito de (2.3.12) é avaliada pela multiplicação de cada equação por $p(i)$ e somando sobre i. |

Realmente, o problema de resolver (2.3.12) e (2.3.13) para encontrar o máximo de $E_0(\rho, \vec{p})$ é quase idêntico ao problema de encontrar capacidade. A maximização em $\vec{p}$ para alguns canais pode ser conjeturada e observada por (2.3.12). Para qualquer canal simétrico (veja um exemplo na seção (1.2.B)) podemos constatar que $E_0(\rho, \vec{p})$ é maximizado tomando os $p(i)$ todos idênticos. A seguir, se o número de entradas e saídas são iguais, algumas vezes é possível resolver (2.3.12) como um conjunto de equações lineares em $(\alpha_j(\vec{p}))^{\rho}$ e então resolver (2.3.13) para $p(i)$. Finalmente, usando a convexidade de $F(\rho, \vec{p})$, será fácil maximizar $E_0(\rho, \vec{p})$ com o computador.

Como ocorre no cálculo da capacidade, a solução para $\alpha_j(\vec{p})$ em (2.3.12) e (2.3.13) é única, mas a solução para $p(i)$ não necessita ser única. Se o alfabeto de entrada tem tamanho I maior que o de saída J será sempre possível maximizar $E_0(\rho, \vec{p})$

com somente J dos p(i) não zero. A única diferença significante entre maximizar I(X;Y) e $E_0(\rho, \vec{p})$ é que as probabilidades de saída para a capacidade são sempre estritamente positivas onde alguns dos $\alpha_j(\vec{p})$ podem ser zero.

Dada a distribuição $\vec{p}$ que maximiza $E_0(\rho, \vec{p})$ para cada ρ, podemos usar a técnica gráfica da Fig. (2.3.3) para encontrar a curva E(R). Alternativamente podemos usar as equações (2.3.7) e (2.3.10) usando para cada ρ o $\vec{p}$ que maximiza $E_0(\rho, \vec{p})$. Para ver que estas equações geram todos os pontos na curva E(R) observamos que para cada R há algum ρ e algum $\vec{p}$ tal que $E(R) = E_0(\rho, \vec{p}) - \rho R$. Para este $\vec{p}$, $E(R) = E(R, \vec{p})$. Mas desde que as equações paramétricas asseguram $E(R, \vec{p})$ para ρ e $\vec{p}$ dados elas sempre asseguram E(R). Passaremos agora a ver alguns exemplos de aplicação.

EXEMPLO 1:

Para o canal simétrico binário da Fig. (2.3.4) $E_0(\rho, \vec{p})$ é maximizado sobre $\vec{p}$ por p(0) = p(1) = 1/2.

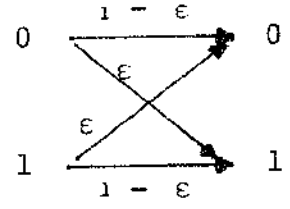

Fig. (2.3.4) - CSB

Sendo $p(0) = p(1) = 1/2$ temos

$$E_0(\rho, \vec{p}) = \rho \ln 2 - (1 + \rho) \ln \left[\varepsilon^{1/1+\rho} + (1 - \varepsilon)^{1/1+\rho} \right]$$

... (2.3.14)

As equações paramétrica (2.3.7) ficam

$$R = \ln 2 - \ln \left[\varepsilon^{1/1+\rho} + (1 - \varepsilon)^{1/1+\rho} \right] +$$

$$+ \frac{\varepsilon^{1/1+\rho}}{\varepsilon^{1/1+\rho} + (1 - \varepsilon)^{1/1+\rho}} \ln \varepsilon^{1/1+\rho} +$$

$$+ \frac{(1 - \varepsilon)^{1/1+\rho}}{\varepsilon^{1/1+\rho} + (1 - \varepsilon)^{1/1+\rho}} \ln (1 - \varepsilon)^{1/1+\rho} .$$

$$E(R, \vec{p}) = - \ln \left[\varepsilon^{1/1+\rho} + (1 - \varepsilon)^{1/1+\rho} \right] -$$

$$- \frac{\varepsilon^{1/1+\rho}}{\varepsilon^{1/1+\rho} + (1 - \varepsilon)^{1/1+\rho}} \ln \varepsilon^{\rho/1+\rho} -$$

$$- \frac{(1-\varepsilon)^{1/1+\rho}}{\varepsilon^{1/1+\rho} + (1-\varepsilon)^{1/1+\rho}} \ln (1-\varepsilon)^{\rho/1+\rho} .$$

Estas equações podem ser manipuladas da seguinte maneira

$$\begin{cases} R = \ln 2 - H(\delta) & (2.3.15a) \\ E(R) = T_{\varepsilon}(\delta) - H(\delta) & (2.3.15b) \end{cases}$$

onde o parâmetro δ é relacionado ao parâmetro ρ em (2.3.7) por

$$\delta = \frac{\varepsilon^{1/1+\rho}}{\varepsilon^{1/1+\rho} + (1-\varepsilon)^{1/1+\rho}}$$

e $H(\delta)$ e $T_{\varepsilon}(\delta)$ são dados por

$$\begin{cases} H(\delta) = -\delta \cdot \ln\delta - (1-\delta).\ln(1-\delta) & (2.3.16a) \\ T_{\varepsilon}(\delta) = -\delta \cdot \ln\varepsilon - (1-\delta).\ln(1-\varepsilon) & (2.3.16b) \end{cases}$$

Estas equações somente são válidas para δ tal que

$$\varepsilon \leq \delta \leq \sqrt{\varepsilon} \,/\, \left(\sqrt{\varepsilon} + \sqrt{1 - \varepsilon}\right)$$

Para $R < \ln 2 - H\left[\left(\sqrt{\varepsilon} \,/\, \left(\sqrt{\varepsilon} + \sqrt{1 - \varepsilon}\right)\right) \right]$ podemos combinar (2.3.10) com (2.3.14) para obtermos

$$E(R) = \ln 2 - 2 \ln \left(\sqrt{\varepsilon} + \sqrt{1 - \varepsilon}\right) - R$$

As equações (2.3.15) podem ser interpretadas graficamente como na Fig. (2.3.5) a seguir. Nela pode ser visto que $T_\varepsilon(\delta)$, como função de δ, é a equação da tangente no ponto ε da curva $H(\delta)$

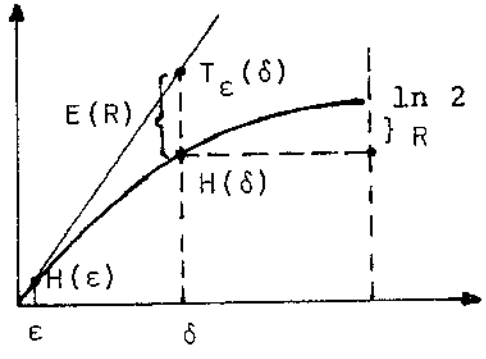

Fig. (2.3.5)

O expoente de codificação para o CSB

O mais significante ponto sobre este exemplo é que não há maneira simples para expressar E(R) exceto na forma paramétrica.

EXEMPLO 2:

Encontrar o expoente de codificação aleatório E(R) (na forma paramétrica) para o canal binário com rasura da Fig. (2.3.6), sendo a probabilidade de rasura ε. Para $\varepsilon = 1/2$ fazer o gráfico de E(R) especificando os valores numéricos de E(0), $R_c = \frac{\partial}{\partial \rho} E_0(\rho) \Big|_{\rho=1}$ e $E(R_c)$. Finalmente encontrar uma interpretação gráfica de E(R) semelhante à que foi feita no exemplo anterior

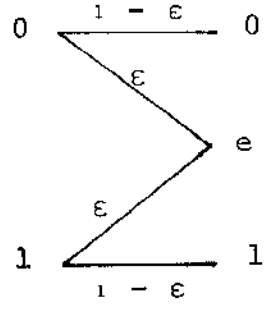

Fig. (2.3.6) - CSBR

Novamente, como no exemplo anterior, o máximo de $E_o(\rho, \vec{p})$ será atingido quando $p(0) = p(1) = 1/2$. Para este $\vec{p}$

temos então:

$$E_0(\rho, \vec{p}) = - \ln \left[\varepsilon + 2^{-\rho}(1 - \varepsilon) \right]$$

As equações paramétricas (2.3.7) ficam

$$R = \frac{2^{-\rho}(1 - \varepsilon) \ln 2}{\varepsilon + 2^{-\rho}(1 - \varepsilon)} \quad ; \quad 0 \leq \rho \leq 1$$

$$E(R) = - \ln \left[\varepsilon + 2^{-\rho}(1 - \varepsilon) \right] - \rho \frac{2^{-\rho}(1 - \varepsilon) \ln 2}{\varepsilon + 2^{-\rho}(1 - \varepsilon)}$$

Do resultado acima tiramos que $\frac{(1 - \varepsilon) \ln 2}{1 + \varepsilon} \leq R \leq (1 - \varepsilon) \ln 2.$

Para $R < \frac{(1 - \varepsilon) \ln 2}{1 + \varepsilon}$ temos que $E(R) = \ln 2 - \ln(1 + \varepsilon) - R,$

como na Fig. (2.3.7)

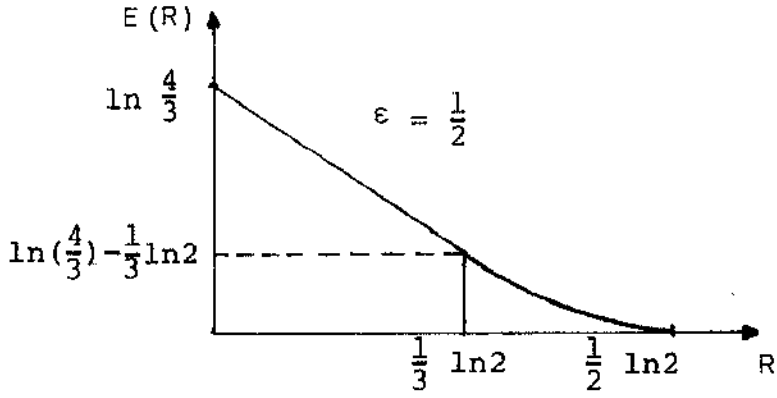

Fig. (2.3.7) - E(R) x R para o CSBR

Seja agora $\delta = \dfrac{\varepsilon}{\varepsilon + 2^{-\rho}(1 - \varepsilon)}$

Em termos do parâmetro δ as equações paramétricas são

$$\begin{cases} R = (1 - \delta)\ \ln 2 \\ E(R) = T_{\varepsilon}(\delta) - H(\delta) \quad , \quad \text{veja (2.3.16).} \end{cases}$$

Estas equações podem ser vistas graficamente como na Fig.(2.3.8)

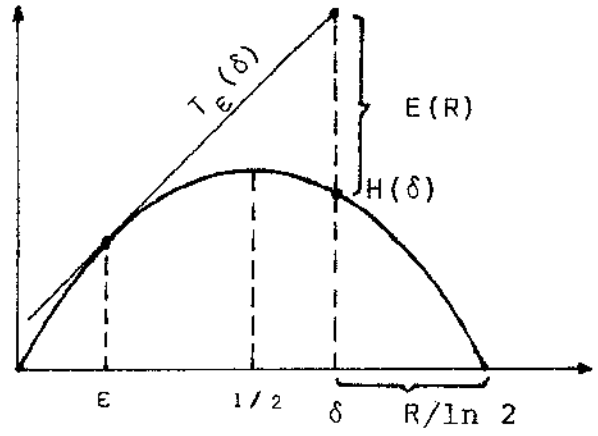

Fig. (2.3.8)

O expoente de codificação para o CSBR

EXEMPLO 3: (CANAIS COM MUITO RUÍDO)

Consideremos um canal com muito ruído no sentido de que a probabilidade de receber uma dada saída é quase independente da entrada. Encontraremos uma aproximação de $E(R)$ para tais canais que depende somente da capacidade. Seja w_j, $j = 1, \ldots, J$ um conjunto de probabilidades definidas para as saídas do canal e definamos ε_{ji} por

$$P(j|i) = w_j(1 + \varepsilon_{ji}) \qquad (2.3.17)$$

Façamos $|\varepsilon_{ji}| << 1$, para todo (i,j), de modo que o canal será de muito ruído no sentido acima. Se (2.3.17) é somado sobre j obtemos

$$\sum_j w_j \varepsilon_{ji} = 0 \quad \text{para todo } i. \qquad (2.3.18)$$

Agora calculamos $E_0(\rho, \vec{p})$ para o canal por expansão de E_0 como uma série de potências em ε_{ji} e abandonamos todos os termos maiores que a segunda ordem

$$E_0(\rho, \vec{p}) = - \ln \sum_j \left[\sum_i p(i)\, w_j^{1/1+\rho}\, (1 + \varepsilon_{ji})^{1/1+\rho} \right]^{1+\rho}$$

... (2.3.19)

Retirando w_j da soma interna e expandindo $(1 + \varepsilon_{ji})^{1/1+\rho}$ teremos

$$E_0(\rho,\vec{p}) \simeq -\ln \sum_j w_j \left\{ \sum_i p(i) \left[1 + \frac{\varepsilon_{ji}}{1+\rho} - \rho \frac{\varepsilon_{ji}^2}{2(1+\rho)^2} \right] \right\}^{1+\rho} \simeq$$

$$\simeq -\ln \sum_j w_j \left\{ 1 + \sum_i (1+\rho) p(i) \left[\frac{\varepsilon_{ji}}{1+\rho} - \frac{\rho \varepsilon_{ji}^2}{2(1+\rho)^2} \right] + \right.$$

$$\left. + \rho \frac{(1+\rho)}{2} \left[\sum_i p(i) . \frac{\varepsilon_{ji}}{1+\rho} \right]^2 \right\}$$

Usando (2.3.18) isto se torna

$$E_0(\rho,\vec{p}) \simeq -\ln \left\{ 1 - \frac{\rho}{2(1+\rho)} \sum_j w_j \left[\sum_i p(i) \varepsilon_{ji}^2 - \left(\sum_i p(i) \varepsilon_{ji} \right)^2 \right] \right\} \simeq$$

$$\simeq \frac{\rho}{2(1+\rho)} \sum_j w_j \left[\sum_i p(i) \varepsilon_{ji}^2 - \left(\sum_i p(i) \varepsilon_{ji} \right)^2 \right] =$$

$$= \frac{\rho}{1+\rho} f(\vec{p}) \qquad (2.3.20)$$

onde

$$f(\vec{p}) = \frac{1}{2} \sum_j w_j \left[\sum_i p(i)\ \epsilon_{ji}^2 - \left(\sum_i p(i)\ \epsilon_{ji} \right)^2 \right] .$$

As equações paramétricas ficarão

$$\begin{cases} R \simeq \dfrac{f(\vec{p})}{(1+\rho)^2} & (2.3.21a) \\ E(R) \simeq \rho^2 \dfrac{f(\vec{p})}{(1+\rho)^2} & (2.3.21b) \end{cases}$$

A informação mútua média usando a probabilidade de entrada $\vec{p}$ é dada por (2.3.21a) com $\rho = 0$. Então a capacidade do canal é dada por

$$C = \max_{\vec{p}} f(\vec{p}) \qquad (2.3.22)$$

Finalmente, resolvendo (2.3.21) para ρ e usando (2.3.22) obtemos

$$E(R) \simeq (\sqrt{C} - \sqrt{R})^2 \quad ; \quad \frac{C}{4} \leq R \leq C$$

Para $R < \frac{C}{4}$ podemos combinar (2.3.10), (2.3.20) e (2.3.22) e obteremos

$$E(R) \simeq \frac{C}{2} - R \quad , \quad 0 \leq R < \frac{C}{4}$$

O gráfico desta função será o da Fig. (2.3.9).

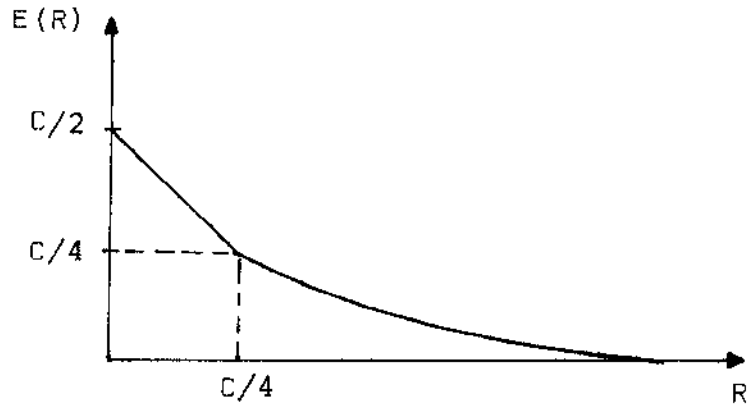

Fig. (2.3.9)

E(R) para canais com muito ruído.

EXEMPLO 4: (CANAIS PARALELOS)

Sejam $[P_1(y_j|x_K)]$ e $[P_2(y_\ell|x_i)]$ as probabilidades de transição de dois C.D.S.M. Usaremos estes canais em paralelo, isto é, em cada unidade de tempo o transmissor emite um símbolo x_K sobre o primeiro canal e um x_i sobre o segundo. Os canais serão considerados independentes, isto é, a probabilidade de receber um símbolo y_j no primeiro canal e um y_ℓ no segundo canal, dado que o par (x_K, x_i) foi emitido, é $P_1(y_j|x_K) \cdot P_2(y_\ell|x_i)$. Estes

canais paralelos podem ser considerados como um simples canal com entradas consistindo em (x_k, x_i) pares e saídas consistindo em (y_j, y_ℓ) pares. Podemos aplicar o teorema de codificação nesta combinação de canais usando sequências de pares de entradas como palavras código. Sendo $p(x_k, x_i)$ a probabilidade atribuída aos pares de entrada, temos

$$E_0(\rho,\vec{p}) = -\ln \sum_{y_j, y_\ell} \left[\sum_{x_k, x_i} p(x_k, x_i) \left(P_1(y_j|x_k) \cdot P_2(y_\ell|x_i)\right)^{1/1+\rho} \right]^{1+\rho}$$

Se restringimos $p(x_k, x_i) = p_1(x_k) \cdot p_2(x_i)$,

onde $\vec{p}_1$ e $\vec{p}_2$ são probabilidades de entrada atribuídas aos canais separados, então $E_0(\rho, \vec{p})$ simplifica do seguinte modo

$$E_0(\rho, \vec{p}) = -\ln \sum_{y_j, y_\ell} \left\{ \left[\sum_{x_k} p_1(x_k) \cdot \left(P_1(y_j|x_k)\right)^{1/1+\rho} \right]^{1+\rho} \cdot \right.$$

$$\left. \cdot \left[\sum_{x_i} p_2(x_i) \left(P_2(y_\ell|x_i)\right)^{1/1+\rho} \right]^{1+\rho} \right\} =$$

$$= E_0^1(\rho, \vec{p}_1) + E_0^2(\rho, \vec{p}_2)$$

onde $E_0^1(\rho, \vec{p}_1) = -\ln \sum_{y_j} \left[\sum_{x_k} p_1(x_k) \left(P_1(y_j|x_k)\right)^{1/1+\rho} \right]^{1+\rho}$

$$E_0^2(\rho, \vec{p}_2) = -\ln \sum_{y_\ell} \left[\sum_{x_i} p_2(i) \left(P_2(y_\ell|x_i)\right)^{1/1+\rho} \right]^{1+\rho}$$

Se escolhermos $\vec{p}_1$ que maximiza $E_0^1(\rho, \vec{p}_1)$ e $\vec{p}_2$ que maximiza $E_0^2(\rho, \vec{p}_2)$ para um dado ρ, então se segue de (2.3.12) que $E_0(\rho, \vec{p})$ é maximizado por $p(x_k, x_i) = p_1(x_k) \cdot p_2(x_i)$ e então

$$\max_{\vec{p}} E_0(\rho, \vec{p}) = \max_{\vec{p}_1} E_0^1(\rho, \vec{p}_1) + \max_{\vec{p}_2} E_0^2(\rho, \vec{p}_2)$$

Este resultado tem a interessante interpretação geométrica da Fig. (2.3.10)

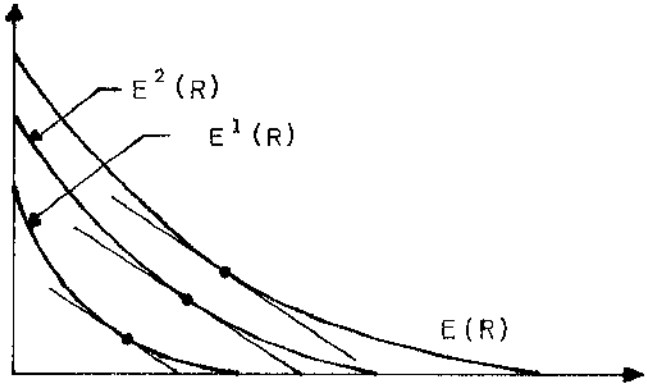

Fig. (2.3.10)

E(R) para canais paralelos

Sejam $E(\rho)$ e $R(\rho)$ os expoente e taxa para a combinação paralela como parametricamente relatados por (2.3.7) usando a otimização sobre $\vec{p}$. Sejam $E^1(\rho)$, $R^1(\rho)$, $E^2(\rho)$ e $R^2(\rho)$ as quantidades anã

logas para os canais individuais. Então

$$E(\rho) = E^1(\rho) + E^2(\rho)$$

$$R(\rho) = R^1(\rho) + R^2(\rho)$$

E daí a combinação paralela é formada pelo vetor adição de pontos com a mesma inclinação nas curvas individuais como vemos na Fig. (2.3.10). É claro que isto pode ser generalizado para quando tivermos "n" canais paralelos.

CAPÍTULO III

O DUAL DO TEOREMA FUNDAMENTAL DA INFORMAÇÃO E O LIMITE INFERIOR DA PROBABILIDADE DE ERRO

3.1. INTRODUÇÃO

Já vimos no capítulo anterior que Shannon [24], em 1948, descobriu que se a taxa de transmissão R de um canal é menor que sua capacidade C então é sempre possível se escolher um código bloco, de comprimento N suficientemente grande, tal que a probabilidade de erro de decodificação seja arbitrariamente pequena. Mas somente em 1965 Gallager [15] conseguiu dar uma demonstração simples e elegante para o fato, estabelecendo um limite superior para a probabilidade de erro, como no Teorema (2.2.1). Iniciaram se investigações sobre o que poderia acontecer se a taxa de transmissão R superasse a capacidade C do canal. Então Feinstein [12] provou, utilizando a desigualdade de Fano [11], que a probabilidade de erro tende para um limite longe de zero quando N cresce muito. Este bom resultado, entretanto, passou a ser considerado a inversa fraca do teorema de codificação depois que Wolfowitz [31] deu uma recíproca forte para o mesmo. Em 1973 Arimoto [1] apresentou o verdadeiro dual do teorema de Gallager.

Neste Capítulo mostraremos o dual do teorema de codificação para R maior do que C no problema de determinar o limite inferior para a probabilidade de erro.

3.2. LIMITE INFERIOR DA PROBABILIDADE DE ERRO

Novamente utilizaremos o C.D.S.M$(X,Y,[P(j|i)],X_N,Y_N)$ definido na seção (2.1). Um código com M palavras de comprimento N é uma aplicação do conjunto de m mensagens da fonte, $1\leq m\leq M$, no conjunto de M palavras código $\vec{x}_1,\ldots,\vec{x}_M$, onde $\vec{x}_i \in X_N$, $1 \leq i \leq M$. A decodificação será uma aplicação do conjunto das N-sequências de saída Y_N nos inteiros $1,\ldots, M$. Partiremos do princípio de que:

i) As mensagens são igualmente prováveis

ii) Será adotado o decodificador de máxima-verossimilhança.

Então a N-sequência de saída $\vec{y}$ será decodificada no inteiro m', através da palavra código $\vec{x}_{m'}$, se

$$P(\vec{y}|\vec{x}_{m'}) > P(\vec{y}|\vec{x}_m) \quad , \quad 1 \leq m' \leq M, \quad m' \neq m$$

No capítulo anterior definimos a probabilidade de erro de uma palavra código através de uma função indicadora. No presente caso será interessante e possível darmos uma expressão para a proba

bilidade de erro de decodificação para um dado conjunto de M palavras código $C = \{\vec{x}_1, \ldots, \vec{x}_M\}$ na seguinte forma:

$$P_e(C) = 1 - \sum_{\vec{y}\in Y_N} \frac{1}{M} \left[\max_{m'} P(\vec{y}|\vec{x}_{m'}) \right] \quad .$$

É claro que, para qualquer $\vec{y}\in Y_N$ e qualquer constante fixada $\beta<0$,

$$\max_{m'} P(\vec{y}|\vec{x}_{m'}) = \left\{ \left[\max_{m'} P(\vec{y}|\vec{x}_{m'}) \right]^{1/\beta} \right\}^{\beta} \leq$$

$$\leq \left[\sum_{m'} (P(\vec{y}|\vec{x}_{m'}))^{1/\beta} \right]^{\beta} \quad .$$

Usando este resultado vemos que a probabilidade de erro de decodificação satisfará à desigualdade

$$P_e(C) \geq 1 - \frac{1}{M} \sum_{\vec{y}\in Y_N} \left[\sum_{m'} (P(\vec{y}|x_{m'}))^{1/\beta} \right]^{\beta} \qquad (3.2.1)$$

A desigualdade (3.2.1) acima dá um limite inferior para um código particular quando se usa o esquema de decodificação de máxima verossimilhança. Entretanto, quando N cresce muito, este limite torna-se complicado ao mesmo tempo em que não recai em alguma propriedade de convergência. Então será necessário simplificar a desigualdade (3.2.1) pela avaliação sobre o conjunto de todos

códigos possíveis.

Para isso vamos supor que $C^0 = \{\vec{x}_1^0, \ldots, \vec{x}_M^0\}$ seja o código que minimiza a probabilidade de erro, que denotaremos por P_e. Denotaremos ainda por $\vec{p}(\vec{x}_1, \ldots, \vec{x}_M)$ a probabilidade associada à escolha do código $C = \{\vec{x}_1, \ldots, \vec{x}_M\}$ e por E a esperança baseada na medida de probabilidade $\vec{p}(\vec{x}_1, \ldots, \vec{x}_M)$. Então podemos ver que existe pelo menos uma medida de probabilidade $\vec{p}^0(\vec{x}_1, \ldots, \vec{x}_M)$ tal que $E\{P_e(C)\} = P_e$. Por exemplo

$$\vec{p}^0(\vec{x}_1, \ldots, \vec{x}_M) = \begin{cases} \dfrac{1}{M!}, & \text{se } C=(\vec{x}_1, \ldots, \vec{x}_M) \text{ coincide com} \\ & C^0=(\vec{x}_1^0, \ldots, \vec{x}_M^0) \text{ em alguma permutação} \\ 0, & \text{noutro caso} \end{cases}$$

Esta medida de probabilidade assegura que $E\{P_e(C)\} = P_e$.

Seja $\vec{p}(\vec{x}_1, \ldots, \vec{x}_M)$ uma medida de probabilidade arbitrária no conjunto de todos os possíveis códigos. Então, voltando à desigualdade (3.2.1) e aplicando o valor esperado obtemos

$$E\{P_e(C)\} \geq 1 - \frac{1}{M} \sum_{\vec{y} \in Y_N} \left\{ \sum_{m'} E\left[\left(P(\vec{y}|\vec{x}_{m'})\right)^{1/\beta} \right] \right\}^{\beta},$$

$$0 < \beta \leq 1 \qquad (3.2.2)$$

Isto porque, para uma constante fixada $0 < \beta \leq 1$, a função escalar $f(t) = -t^{\beta}$ é convexa U para $t \geq 0$.

Agora impomos a restrição adicional de que a distribuição de probabilidade $\vec{p}(\vec{x}_1, \ldots, \vec{x}_M)$ é invariante sob alguma permutação de seus argumentos. Então isto implica que as distribuições de probabilidade marginais

$$\vec{p}_i(\vec{x}_i) = \sum_{\vec{x}_1 \epsilon X_N} \cdots \sum_{\vec{x}_{i-1} \epsilon X_N} \sum_{\vec{x}_{i+1} \epsilon X_N} \cdots \sum_{\vec{x}_M \epsilon X_N} \vec{p}(\vec{x}_1, \ldots, \vec{x}_M)$$

são todas iguais e, portanto, todos $E\left[(P(\vec{y}|\vec{x}_m,))^{1/\beta}\right]$ são iguais em (3.2.2).

Então

$$E\{P_e(C)\} \geq 1 - M^{\beta-1} \sum_{\vec{y} \epsilon Y_N} \left[\sum_{\vec{x} \epsilon X_N} \vec{p}(\vec{x}) (P(\vec{y}|\vec{x}))^{1/\beta} \right]^{\beta} \qquad \ldots \qquad (3.2.3)$$

onde a distribuição de probabilidade marginal $\vec{p}(\vec{x})$ será definida da forma

$$\vec{p}(\vec{x}) = \sum_{\vec{x}_2 \epsilon X_N} \cdots \sum_{\vec{x}_M \epsilon X_N} P(\vec{x}, \vec{x}_2, \ldots, \vec{x}_M)$$

Desde que existe uma medida de probabilidade $\vec{p}^{\,0}(\vec{x}_1,\ldots,\vec{x}_M)$ que é invariante sob permutações de seus argumentos e satisfaz à condição $E\left[P_e(C)\right] = P_e$, como visto anteriormente, então é possível reescrever (3.2.3) como

$$P_e \geq \inf_{\vec{p}(\vec{x})} \left\{ 1 - M^{\beta-1} \sum_{\vec{y}\in Y_N} \left[\sum_{\vec{x}\in X_N} \vec{p}(\vec{x}) \left(P(\vec{y}|\vec{x})\right)^{1/\beta} \right]^{\beta} \right\} \qquad (3.2.4)$$

onde a operação ínfimo é tomada sobre o conjunto de todas possíveis distribuições de probabilidade em X_N.

Usando o fato de que o canal é discreto sem memória temos

$$P(\vec{y}|\vec{x}) = \prod_{i=1}^{N} p(y_i|x_i)$$

onde: $\begin{cases} \vec{x} = (x_1,\ldots,x_N) \in X_N \quad e \quad x_i \in X = \{x_1,\ldots,x_I\} \\ \vec{y} = (y_1,\ldots,y_N) \in Y_N \quad e \quad y_j \in Y = \{y_1,\ldots,y_J\} \end{cases}$

Precisaremos agora do seguinte lema.

LEMA (3.2.1): Para o canal discreto sem memória especificado pelas probabilidades de transição $P(\vec{y}|\vec{x})$ é assegurado que

$$\max_{\vec{p}(\vec{x})} \sum_{\vec{y}\in Y_N} \left[\sum_{\vec{x}\in X_N} \vec{p}(\vec{x}) \left(P(\vec{y}|\vec{x})\right)^{1/\beta} \right]^{\beta} =$$

$$= \left\{ \max_{p(x)} \left[\sum_{y\in Y} \left(\sum_{x\in X} p(x) \left(P(y|x)\right)^{1/\beta} \right)^{\beta} \right] \right\}^{N}, \quad 0 < \beta \leq 1$$

... (3.2.5)

PROVA: No capítulo anterior vimos que a maximização de $E_0(\rho, \vec{p})$ foi possível definindo uma função

$$F(\rho, \vec{p}) = \sum_j \left[\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho} \right]^{1+\rho}$$

onde $0 \leq p(i) \leq 1$, $\sum_i p(i) = 1$ e $0 \leq P(j|i) \leq 1$, $\sum_j P(j|i) = 1$.

No caso onde $0 \leq \rho \leq 1$ Gallager provou, usando a convexidade de $F(\rho, \vec{p})$ com respeito a $\vec{p}$, que condições necessárias e suficientes ao vetor de probabilidade $\vec{p}$ que minimizavam $F(\rho, \vec{p})$, maximizando $E_0(\rho, \vec{p})$, eram dadas pelas expressões (2.3.12) e (2.3.13) do teorema (2.3.3). Mas agora temos $-1 \leq \rho < 0$ e como consequência a função $F(\rho, \vec{p})$ será convexa $\cap$. As condições necessárias e suficientes então para que $\vec{p}$ maximize $F(\rho, \vec{p})$ serão

$$\sum_j (P(j|i))^{1/1+\rho} \cdot (\alpha_j(\vec{p}))^{\rho} \leq \sum_j (\alpha_j(\vec{p}))^{1+\rho} \quad \forall \quad i$$

com igualdade se $p(i) > 0$ onde $\alpha_j(\vec{p})$ é definido como em (2.3.13). Para um β fixado arbitrariamente, $0 < \beta \leq 1$, seja $\vec{p}^0 = (p^0(x_1), \ldots, p^0(x_I))$ o vetor de maximização tal que

$$\sum_{y \in Y} (P(y|x))^{1/\beta} \cdot (\alpha(y))^{\beta-1} \leq \sum_{y \in Y} (\alpha(y))^{\beta} \quad , \quad \forall \quad x \in X$$

com igualdade se $p^0(x) > 0$ onde

$$\alpha(y) = \sum_{x \in X} p^0(x) (P(y|x))^{1/\beta}$$

É evidente, então, que a distribuição de probabilidade

$$\vec{p}^0(\vec{x}) \equiv \vec{p}^0(x_1, \ldots, x_N) = p^0(x_1) \ldots p^0(x_N) \quad \forall \quad \vec{x}$$

satisfaz

$$\sum_{\vec{y} \in Y_N} (P(\vec{y}|\vec{x}))^{1/\beta} \cdot (\alpha(\vec{y}))^{\beta-1} \leq \sum_{\vec{y} \in Y_N} (\alpha(\vec{y}))^{\beta}$$

com igualdade se $\vec{p}^0(\vec{x}) > 0$ onde

$$\alpha(\vec{y}) = \sum_{\vec{x}\in X_N} \vec{p}^{\,0}(\vec{x})\,(P(\vec{y}|\vec{x}))^{1/\beta}$$

Portanto, $\vec{p}^{\,0}(\vec{x})$ é a distribuição de probabilidade que maximiza

$$\sum_{\vec{y}\in Y_N} \left[\sum_{\vec{x}\in X_N} \vec{p}(\vec{x})\,(P(\vec{y}|\vec{x}))^{1/\beta} \right]^{\beta}$$

Isto resulta na igualdade (3.2.5).

Retornando a (3.2.4) e aplicando o Lema (3.2.1) temos que

$$P_e \geq 1 - M^{\beta-1} \left\{ \max_{\vec{p}} \left[\sum_j \left(\sum_i p(i)\,(P(j|i))^{1/\beta} \right)^{\beta} \right]^{N} \right\}$$

... (3.2.6)

onde $p(i)$ e $P(j|i)$ são:

$$\begin{cases} p(i) = p(x_i), & i = 1,\ldots, I \\ P(j|i) = P(y_j|x_i), & j = 1,\ldots, J \quad e \quad i = 1,\ldots, I \end{cases}$$

Finalmente, sejam $M = \exp(NR)$, onde R é a taxa em nats, $\beta = 1+\rho$, e $E_0(\rho, \vec{p})$ como em (2.2.2). Então a inequação (3.2.6) pode ser

colocada na forma

$$P_e \geq 1 - \exp. -N \left[-\rho R + \min_{\vec{p}} E_0(\rho, \vec{p}) \right] , \text{ para } -1 \leq \rho < 0$$

... (3.2.7)

e onde o caso especial $\rho = -1$ será interpretado como

$$\min_{\vec{p}} E_0(-1, \vec{p}) = \lim_{\rho \to -1} \min_{\vec{p}} E_0(\rho, \vec{p})$$

que resultará equivalente a

$$-\ln \left[\sum_j \max_i P(j|i) \right]$$

Então provamos o seguinte Teorema.

TEOREMA (3.2.1): Suponha um C.D.S.M$(X, Y, [P(j|i)], X_N, Y_N)$. Então, para algum bloco de comprimento N, algum número de palavras código $M = \exp(NR)$, e alguma seleção de códigos, a probabilidade de erro de decodificação é limitada pela desigualdade (3.2.7).

A função expoente $\min_{\vec{p}} E_\alpha(\rho, \vec{p})$, para $-1 \leq \rho < 0$, tem propriedades interessantes semelhantes às da função $\max_{\vec{p}} E_0(\rho, \vec{p})$,

para $0 < \rho \leq 1$. A mais importante é a que se segue no seguinte Lema.

LEMA (3.2.2):

$$\lim_{\rho \uparrow o} \frac{1}{\rho} \min_{\vec{p}} E_o(\rho, \vec{p}) = \lim_{\rho \downarrow o} \frac{1}{\rho} \max_{\vec{p}} E_o(\rho, \vec{p}) = C \qquad (3.2.8)$$

onde C é a capacidade do canal, isto é, o máximo da informação mútua média

$$I(\vec{p}; P) = \sum_i \sum_j p(i)\, P(j|i) \ln \frac{P(j|i)}{\sum_k p(k)\, P(j|k)}$$

como já definida na seção (2.3)

PROVA: É fácil de ver que

$$E_o(\rho, \vec{p}) = 0, \quad \text{para} \quad \rho = 0$$

$$\frac{\partial}{\partial \rho} E_o(\rho, \vec{p}) \Big|_{\rho=0} = I(\vec{p}; P)$$

Entretanto, de acordo com o Teorema (2.3.1), se $I(\vec{p}, P)$ é positivo e $\rho > 0$ então

$$E_0(\rho, \vec{p}) > 0 \qquad (3.2.9a)$$

$$\frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) > 0 \qquad (3.2.9b)$$

$$\frac{\partial^2}{\partial \rho^2} E_0(\rho, \vec{p}) \leq 0 \qquad (3.2.9c)$$

Portanto, as desigualdades (3.2.9b) e (3.2.9c) são válidas sempre para o caso no qual $-1 \leq \rho < 0$. Desde que a desigualdade (3.2.9c) implica que $E_0(\rho, \vec{p})$ é uma função convexa $\cap$ com respeito a ρ para um $\vec{p}$ fixado, é evidente que para algum $\vec{p}$

$$E_0(\rho, \vec{p}) \geq \rho \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \qquad \text{para} \quad -1 \leq \rho < 0$$

$$E_0(\rho, \vec{p}) \leq \rho \left[\frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \Big|_{\rho=0} \right] \qquad \text{para} \quad 0 < \rho \leq 1$$

as quais asseguram, respectivamente,

$$\lim_{\rho \to 0} \frac{\min_{\vec{p}} E_0(\rho, \vec{p})}{\rho} \leq \lim_{\rho \to 0} \max_{\vec{p}} \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) = \max_{\vec{p}} I(\vec{p}; P) = C$$

... (3.2.10)

$$\lim_{\rho \to o} \frac{\max_{\vec{p}} E_0(\rho, \vec{p})}{\rho} \leq \max_{\vec{p}} \left[\frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \Big|_{\rho=0} \right] = C$$

... (3.2.11)

Por outro lado se segue que para algum $\vec{p}$

$$E_0(\rho, \vec{p}) \leq \rho \left[\frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \Big|_{\rho=0} \right] \quad \text{para} \quad -1 \leq \rho < 0$$

$$E_0(\rho, \vec{p}) \geq \rho \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \qquad \text{para} \quad 0 < \rho \leq 1$$

as quais asseguram, respectivamente

$$\lim_{\rho \to o} \frac{\min_{\vec{p}} E_0(\rho, p)}{\rho} \geq \max_{\vec{p}} \left[\frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \Big|_{\rho=0} \right] = C \quad (3.2.12)$$

$$\lim_{\rho \to o} \frac{\max_{\vec{p}} E_0(\rho, p)}{\rho} \geq \lim_{\rho \to o} \left[\max_{\vec{p}} \frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) \right] = C \quad (3.2.13)$$

As desigualdades (3.2.12) e (3.2.13) conjuntamente com (3.2.10) e (3.2.11) implicam em (3.2.8)

TEOREMA (3.2.2): Para qualquer canal discreto sem memória, se $R > C$, então

$$E(R) = \max_{-1 \leq \rho < 0} \left[-\rho R + \min_{\vec{p}} E_0(\rho, \vec{p}) \right] > 0 \qquad (3.2.14)$$

PROVA: Notamos que existe um número $\varepsilon > 0$ tal que $R > C + \varepsilon$. Então, é evidente do Lema (3.2.2), que existe um número ρ_0, $-1 \leq \rho_0 < 0$ tal que

$$\frac{1}{\rho_0} \min_{\vec{p}} E_0(\rho_0, \vec{p}) \leq C + \varepsilon$$

Substituindo isto em (3.2.14) temos

$$-\rho_0 R + \min_{\vec{p}} E_0(\rho_0, \vec{p}) \geq -\rho_0 (R - C - \varepsilon) > 0$$

Os teoremas (3.2.1) e (3.2.2) provam a inversa forte do teorema de codificação para canais discretos sem memória. O limite inferior dado para a probabilidade de erro por (3.2.7) é um resultado forte e que pode ser generalizado para outros tipos de canais.

APÊNDICE A

A-1: Antes de apresentarmos o restante da demonstração deste teorema (2.3.1) veremos o seguinte Lema.

LEMA: Seja $\vec{p} = (p(x_1), \ldots, p(x_I))$ um vetor de probabilidade e sejam $a_1, \ldots, a_I$ um conjunto de números não-negativos. Então a função

$$f(s) = \ln \left[\sum_{i=1}^{I} p(i)\ a_i^{1/s} \right]^s$$

é não-crescente e convexa U em s para $s > 0$. Entretanto, $f(s)$ é estritamente decrescente a menos que todos a_i's para os quais $p(i) > 0$ sejam iguais. A convexidade é estrita a menos que todos a_i's não-zero para os quais $p(i) > 0$ sejam iguais

PROVA: O fato de que $f(s)$ seja não-crescente e as condições para que ela seja estritamente decrescente seguem diretamente da desigualdade padrão

$$\left[\sum_i p(i)\ a_i^{1/s} \right]^s \geq \left[\sum_i p(i)\ a_i^{1/r} \right]^r$$

para $0 < s < r$. Para estabelecer a convexidade sejam s, r e θ números arbitrários, $0 < s < r$, $0 < \theta < 1$, e definamos

$$t = \theta s + (1 - \theta) r \tag{A.1}$$

Para mostrar que $f(s)$ é convexa U mostraremos que

$$f(t) \leq \theta f(s) + (1 - \theta)\ f(r) \tag{A.2}$$

Definamos o número λ por

$$\lambda = \frac{s\theta}{t} \quad , \quad 1 - \lambda = \frac{r(1 - \theta)}{t} \tag{A.3}$$

Estas expressões podem ser vistas serem consistentes pela adição delas e usando (A.1). Sempre se segue de (A.3) que

$$\frac{1}{t} = \frac{\theta}{t} + \frac{(1 - \theta)}{t} = \frac{\lambda}{s} + \frac{1 - \lambda}{r}$$

Portanto

$$\sum_i p(i).a_i^{\,1/t} = \sum_i p(i).a_i^{\,\lambda/s} a_i^{(1-\lambda)/r} \leq$$

$$\leq \left[\sum_i p(i).a_i^{\,1/s} \right]^{\lambda} . \left[\sum_i p(i).a_i^{\,1/r} \right]^{1-\lambda} \tag{A.4}$$

onde (A.4) segue da desigualdade de Hölder, ver Apêndice (B-3) . Elevando ambos os lados de (A.4) à potência t e usando (A.3) temos

$$\left[\sum_i p(i).a_i^{1/t} \right]^t \leq \left[\sum_i p(i).a_i^{1/s} \right]^{s\theta} . \left[\sum_i p(i).a_i^{1/r} \right]^{r(1-\theta)}$$

... (A.5)

tomando o logarítmo de (A.5) obtemos (A.2). A convexidade é estrita a menos que (A.4) seja satisfeita com a igualdade, o que ocorre se há uma constante k tal que

$$p(i).a_i^{1/s} = p(i).a_i^{1/r} . k, \text{ para todo i, veja Apêndice (B-3)}$$

Isto implica imediatamente a condição para convexidade estrita no lema.

Voltando então ao Teorema (2.3.1)

$$E_0(\rho, \vec{p}) = -\ln \sum_j \left[\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho} \right]^{1+\rho}$$

Sendo $P(j|i)$ correspondente a a_i no lema, e $1+\rho$ correspondente a s, vemos que

$$\left[\sum_{i} p(i) \left(P(j|i)\right)^{1/1+\rho} \right]^{1+\rho} \qquad \text{(A.6)}$$

é não-crescente com ρ para cada j. Por suposição, $I(\vec{p}; P) > 0$, e então $P(j|i)$ não é independente de i sobre aqueles i's para os quais $p(i) > 0$. Então a expressão (A.6) acima é estritamente decrescente para pelo menos um j; $E_0(\rho, \vec{p})$ é estritamente crescente com ρ, e $\frac{\partial}{\partial \rho} E_0(\rho, \vec{p}) > 0$ para $\rho \geq 0$. Desde que $E_0(0, \vec{p}) = 0$ isto implica sempre que $E_0(\rho, \vec{p}) > 0$ para $\rho > 0$. Proximamente, mostraremos que $E_0(\rho, \vec{p})$ é convexa $\cap$ em ρ. Sejam $\rho_1 > 0$ e $\rho_2 > 0$ arbitrários e θ tal que $0 < \theta < 1$. Definamos $\rho_3 = \rho_1 \theta + \rho_2 (1 - \theta)$. Do lema anterior, veja (A.5), temos que

$$\sum_{j} \left[\sum_{i} p(i) \left(P(j|i)\right)^{1/1+\rho_3} \right]^{1+\rho_3} \leq \sum_{j} \left[\sum_{i} p(i) \left(P(j|i)\right)^{1/1+\rho_1} \right]^{(1+\rho_1)\theta} \cdot$$

$$\cdot \left[\sum_{i} p(i) \left(P(j|i)\right)^{1/1+\rho_2} \right]^{(1+\rho_2)(1-\theta)} \qquad \text{(A.7)}$$

Aplicando agora a desigualdade de Hölder, veja Apêndice (B-2), ao lado direito de (A.7) obtemos

$$\sum_j \left[\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho_3} \right]^{1+\rho_3} \leq \left\{ \sum_j \left[\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho_1} \right]^{1+\rho_1} \right\}^{\theta} .$$

$$\left\{ \sum_j \left[\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho_2} \right]^{1+\rho_2} \right\}^{1-\theta} \qquad (A.8)$$

tomando o logarítmo de ambos os lados de (A.8) acima temos

$$-E_0(\rho_3, \vec{p}) \leq -\theta E_0(\rho_1, \vec{p}) - (1 - \theta) E_0(\rho_2, \vec{p}) .$$

Isto estabelece que $E_0(\rho, \vec{p})$ é convexa $\cap$ em ρ. A convexidade estrita falha se ambas (A.7) e (B-2) são satisfeitas com igualdade. Através do lema vemos que (A.7) é satisfeita com igualdade se $P(j|i)$ é independente de i para todo (i,j) satisfazendo $p(i)P(j|i) > 0$. A condição para igualdade em (B-2) é que haja uma constante k tal que, para todo j,

$$\left[\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho_1} \right]^{1+\rho_1} = k \left[\sum_i p(i) \left(P(j|i)\right)^{1/1+\rho_2} \right]^{1+\rho_2}$$

... (A.9)

Se (A.7) é satisfeita com igualdade então os $P(j|i)$ não zero podem ser fatorados da equação (A.9) acima restando

$$\left[\sum_{i:P(j|i)>0} p(i)\right]^{1+\rho_1} = k \left[\sum_{i:P(j|i)>0} p(i)\right]^{1+\rho_2}$$

para todo j. Isto implica que o termo entre colchetes acima é alguma constante α, independente de j, e então para qualquer(i,j) com $p(i)P(j|i) > 0$ temos:

$$\frac{\sum_i p(i)P(j|i)}{P(j|i)} = \alpha \quad .$$

APÊNDICE B

B-1: TEOREMA

Seja $f(\vec{\alpha})$ uma função convexa $\cap$ de $\vec{\alpha} = (\alpha_1, \ldots, \alpha_I)$ sobre uma região R quando $\vec{\alpha}$ é um vetor de probabilidade. Consideremos que as derivadas parciais $\frac{\partial}{\partial \alpha_i} f(\vec{\alpha})$ são definidas e contínuas sobre a região R com a possível excessão de que $\lim_{\alpha_i \to 0} \frac{\partial}{\partial \alpha_i} f(\vec{\alpha})$ pode ser $+\infty$. Então,

$$\frac{\partial}{\partial \alpha_i} f(\vec{\alpha}) = \lambda \quad ; \quad \text{para todo i tal que } \alpha_i > 0$$

$$\frac{\partial}{\partial \alpha_i} f(\vec{\alpha}) \leq \lambda \quad ; \quad \text{para todo i tal que } \alpha_i = 0$$

são condições necessárias e suficientes ao vetor de probabilidade $\vec{\alpha}$ para maximizar f sobre a região R.

Para B-2 e B-3 deste apêndice consideraremos a_i, b_i e p_i números não negativos definidos sobre um conjunto finito de i, $1 \leq i \leq I$ digamos. Consideraremos $0 < \lambda < 1$ e mais:

$$\sum_i p_i = 1$$

B-2: DESIGUALDADE DE HÖLDER

$$\sum_i a_i b_i \leq \left(\sum_i a_i^{1/\lambda} \right)^{\lambda} \left[\sum_i b_i^{1/(1-\lambda)} \right]^{1-\lambda}$$

com igualdade se, para algum c, $a_i^{1-\lambda} = b_i^{\lambda} \cdot c$ para todo i.

B-3: VARIANTE DA DESIGUALDADE DE HÖLDER

$$\sum_i p_i a_i b_i \leq \left(\sum_i p_i a_i^{1/\lambda} \right)^{\lambda} \left[\sum_i p_i b_i^{1/(1-\lambda)} \right]^{1-\lambda}$$

com igualdade se, para algum c, $p_i a_i^{1/\lambda} = p_i b_i^{1/(1-\lambda)} \cdot c$ para todo i.

REFERÊNCIAS

[1] ARIMOTO, S. (1973): On a Strong Converse to the Coding Theorem for Discrete Memoryless Channels, IEEE Trans. Inform. Theory, 19, 357-359.

[2] ASH, R. B. (1965): Information Theory, Interscience , New York.

[3] BAIERLEIN, R. (1968): Atoms and Information Theory, Freeman, San. Francisw, New York.

[4] BAR-HILLEY, Y. (1964): Language and Information Theory, Addison Wesley Pub. Co. Inc., The Jerusalem Acad. Press Ltd.

[5] BARUCH LEV (1969): Accounting and Information Theory, Americal Accounting Association Studies, 2, Menasha,Wisc. George Banta Pub. Co.

[6] BERGER, T. (1971): Rate Distortion Theory: a Mathematical Basic for Data Compression, Englewood Cliffs, Prentice Hall, New Jersey.

[7] BRILLOUIN, L. (1956): Science and Information Theory. Academic Press, Inc., New York.

[8] DOBRUSHIN, R. L. (1959): A General Formulation of the Fundamental Shannon Theorem in Information Theory, Uspehi Mat. Akad. Nauk, SSSR, 14, 3-104. (Also , Trans. Amer. Math. Soc., Serie 2, 33, 323-438).

[9] EBERT, P. M. (1968): An Extension of Rate Distortion Theory To Confusion Matrices, IEEE Trans. Inform. Theory, 14, 6-11.

[10] ELIAS, P. (1957): List Decoding for Noisy Channels, MIT Research Lab. of Electronics, Tech. Rept. 335.

[11] FANO, R. M. (1961): Transmission of Information, The MIT Press, Cambridge, Mass. and John Wiley and Sons, Inc., New York.

[12] FEINSTEIN, A. (1958): Foundation of Information Theory, McGraw Hill, New York.

[13] FORNEY, G. D. (1966): Concatenated Codes, MIT Press, Cambridge, Mass.

[14] FORNEY, G. D. (1968): Exponential Error Bounds for Erasure, List and Decision Feedback Schemes, IEEE Trans. Inform. Theory, 4, 206-220.

[15] GALLAGER, R. G. (1965): A Simple Derivation of the Coding Theorem and Some Applications, IEEE Trans. Inform. Theory, 11, 3-18.

[16] GALLAGER, R. G. (1968): Information Theory and Reliable Communication, New York: Wiley.

[17] HARDY, G. H., LITTLEWOOD, J. E. and POLYA, G. (1934): Inequalities, Camb. Univ. Press, London (2nd ed., 1952, 1959).

[18] HARTLEY, R. V. L. (1928): Transmission of Information, BSTJ, 7, 535.

[19] JELINEK, F. (1968): Probabilistic Information Theory, McGraw Hill, New York.

[20] KHINCHIN, A. I. (1958): Mathematical Foundations in Information Theory, Dover Pub. Inc., New York.

[21] KOLMOGOROV, A. N. (1956): On the Shannon Theory of Information Transmission in the Case of Continuous Signals, Trans. IRE, 2, 102-108.

[22] KULLBACK, S. (1959): Information Theory and Statistics, John Wiley Pub. New York.

[23] REIFFEN, B. (1966): A Per Letter Converse to the Channel Coding Theorem, IEEE Trans. Inform. Theory, 12 , 475-480.

[24] SHANNON, C. E. (1948): A Mathematical Theory of Communication, BSTJ, 27, 379-423, 623-656. Also reprinted in C. E. SHANNON and W. WEAVER: The Mathematical Theory of Communication, Univ. of Illinois Press, Urbana, Illinois (1949).

[25] SHANNON, C. E. (1959): Coding Theorems for a Discrete Source with a Fidelity Criterion, IRE Nat'l, Conv. Rec., part IV, 142-163 (Also in Information and Decision Processes, R. E. Machel, ed., McGraw Hill, Inc., (1960), 93-126, New York).

[26] SHANNON, C. E., GALLAGER, R. G., and BERLEKAMP, E. R. (1967): Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels part I and II, Inform. and Contr., 10, 65-103 (Part I), 522-552 (part II).

[27] STICLITZ, I. G. (1966): Coding for a Class of Unknown Channels, IEEE Trans. Inform. Theory, 12, 189-195.

[28] STICLITZ, I. G. (1967): A Coding Theorem for a Class of Unknown Channels, IEEE Trans. Inform. Theory, 13, 217-220.

[29] THEIL, H. (1967): Economics and Information Theory, North-Holland Pub. Co., Amsterdam.

[30] WEINER, N. (1961): Cybernetics, the MIT Press and John Wiley and Sons, Inc., New York.

[31] WOLFOWITZ, J. (1964): Coding Theorems of Information Theory, Springer-Verlag and Prentic Hall, Englewood.

[32] WYNER, A. D. (1970): Another look at the Coding Theorem of Information Theory, Procedings of the IEEE, 58 , 894-913.

[33] YUDKIN, H. (1967): On the Exponential Error Bound and Capacity for Finite State Channels, International Symposium of Information Theory, San Remo, Italy.