Teoría de la Información y Codificación

Norman Abramson

Teoría de la Información y Codificación

QUINTA EDICION

1981

PARANINFO SA

MADRID

Traducida por
JUAN ANTONIO DE MIGUEL MENOYO
Ingéniero de Telecomunicación

Version española de la obra inglesa:
INFORMATION THEORY AND CODING
Publicado por McGRAW-HILL Book Company, Inc.

IMPRESO EN ESPAÑA
PRINTED IN SPAIN

ISBN: 84-283-0232-4
Depósito Legal: M-33268-1980

PARANINFO S.A. Magallanes, 25 - MADRID (15) (3-2988-5)

ALCO, artes gráficas. Jaspe, 34. MADRID-26

PROLOGO

Este libro ha surgido de un conjunto de notas sobre la teoría de la información y la codificación binaria, preparadas con motivo de un ciclo de conferencias en los ITT Federal Laboratories. Posteriormente, estas notas se ampliaron, utilizándose como texto en un curso trimestral de ingeniería eléctrica en la Universidad de Stanford. La mayor parte de la versión final ha sido elaborada en un curso semestral que sobre teoría de la información tuvo lugar en los laboratorios de desarrollo de IBM, de San José, California.

El objeto del libro es presentar los conceptos básicos de la teoría de la información en su aspecto, siempre que sea posible, no matemático. Existe la posibilidad de tratar la teoría en forma puramente matemática, estudiando las propiedades de ciertas magnitudes abstractas, a las que se aplica una medida de probabilidad definida previamente. Nuestro interés, sin embargo, se centrará en las relaciones de la teoría con el mundo real y la correspondencia entre las magnitudes sometidas a estudio y ciertos conceptos naturales que influyen en un vasto número de campos. Con objeto de desarrollar completamente esta correspondencia, se emplea el lenguaje de las matemáticas para expresar la teoría de la información. Los teoremas se enuncian y demuestran con todo rigor. Todas las afirmaciones son comprobadas, no requiriendo la ayuda de la intuición más que para interpretar los resultados deducidos.

A pesar de todo lo dicho, los conocimientos matemáticos exigidos son limitados. Simplemente es necesario conocer los logaritmos y el significado, cuando menos intuitivo, de las probabilidades y valores medios. El cálculo no se emplea en el texto. Aprovechamos para advertir al lector no familiarizado con las matemáticas que requerirá tiempo y esfuerzo para comprender ciertas transformaciones de algunas demostraciones. Sin embargo, las demostraciones en sí mismas y su significado no precisan ningún conocimiento matemático.

Se ha alcanzado esta simplicidad limitando la generalidad matemática de los desarrollos. Se han considerado únicamente las fuentes de información de un número finito de símbolos de salida, así como los canales de información de un número finito de entradas y salidas y memoria nula. Estas restricciones nos han permitido tratar todos los conceptos fundamentales de la teoría de la información, no siguiendo, en contrapartida, una línea matemática demasiado elegante. Para los que estén interesados en esto, sin embargo, es fácil generalizar volviendo a plantear las demostraciones en el campo de Borel.

La materia de que trata el libro ha pasado ya tanto por la Universidad como por cursos industriales. Estudiantes de ingeniería, cálculo operacional y calculadoras pueden desarrollarla cómodamente en un semestre. Aquellos con una mayor formación matemática o especial interés pueden analizar más en detalle ciertos temas avanzados contenidos en las notas del final de cada capítulo. Estas notas sugieren un cierto número de áreas de desarrollo interesantes en el dominio de la teoría de la información. Al final de cada capítulo se han incluido algunos problemas; los precedidos por un asterisco requieren un cálculo más o menos complicado para su resolución.

Me encuentro en deuda con cierto número de personas que contribuyeron a la preparación del libro. El Dr. Richard Hamming hizo un detallado análisis de una primera versión. Los profesores David Braverman, Thomas Kailath y Wesley Peterson, aportaron otros interesantes comentarios. El manuscrito se benefició con las correcciones y aclaraciones sugeridas por los estudiantes de Stanford. Algunas de ellas, realizadas por Thomas Cover, Arthur Geoffrion y David Howell se han incorporado al texto. Errores y erratas, finalmente, surgieron de los apuntes tomados por los alumnos de dos cursos en Stanford y uno en los IBM Research Laboratories, a los que deseo hacer ahora una confesión: Las calificaciones no estaban basadas en las listas de correcciones que Vds. presentaron.

INDICE DE MATERIAS

GLOSARIO DE SIMBOLOS Y EXPRESIONES DE LA ENTROPIA

G-1. Figuras

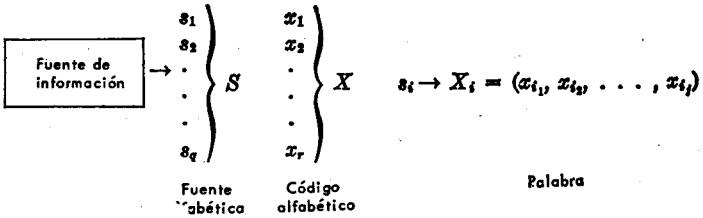


FIGURA G-1.

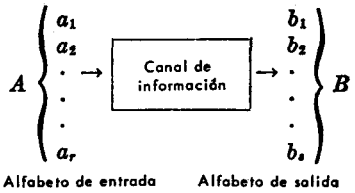


FIGURA G-2

G-2. Símbolos generales

S alfabeto de una fuente
s_i símbolo de una fuente
q número de símbolos de una fuente
S^n extensión de orden n del alfabeto de una fuente
σ_i símbolo de la extensión de orden n del alfabeto de una fuente
$\bar{S}$ afín de la fuente S
P_i probabilidad del símbolo s_i de una fuente

X alfabeto código
x_i símbolo de un código
r número de símbolos de un código (también, número de símbolos de entrada de un canal)
X_i palabra de un código (una secuencia de x_i's) correspondiente a s_i
l_i número de símbolos de la palabra código X_i correspondiente a s_i
λ_i número de símbolos de la palabra código correspondiente a σ_i
L longitud media de una palabra de un código del alfabeto S
L_n longitud media de la palabra de un código del alfabeto S^n

A alfabeto de entrada de un canal
a_i símbolo de entrada de un canal
A^n número de símbolos de entrada de un canal (también, número de símbolos de un código)
α_i extensión de orden n del alfabeto de entrada de un canal
r símbolo de la extensión de orden n del alfabeto de entrada de un canal

B alfabeto de salida de un canal
b_j símbolo de salida de un canal
s número de símbolos de salida de un canal
B^n extensión de orden n del alfabeto de salida de un canal
β_j símbolo de la extensión de orden n del alfabeto de salida de un canal

m orden de una fuente de Markov
P_{ij} elemento de la matriz de un canal; probabilidad de recibir b_j al enviar a_i
p probabilidad de error de un BSC ($\bar{p} = 1 - p$)
C capacidad de un canal
P_E probabilidad de error
M número de mensajes de un código
R velocidad de información
D distancia de Hamming
$d(b_j)$ regla de decisión

G-3. Expresiones de la entropía

$$I(s_i) = \log \frac{1}{P(s_i)}$$ cantidad de información obtenida al recibir s_i (fuente de memoria nula)

$$I(s_i/s_j) = \log \frac{1}{P(s_i/s_j)}$$ cantidad de información obtenida al recibir sucesivamente s_j y s_i (fuente de Markov de primer orden)

$$H(S) = \sum_{S} P(s_i) \log \frac{1}{P(s_i)}$$ entropía de la fuente de memoria nula S

$$H(S/s_j) = \sum_{i=1}^{q} P(s_i/s_j) \log \frac{1}{P(s_i/s_j)}$$ entropía condicional de una fuente de Markov de primer orden S

$$H(S) = \sum_{S^2} P(s_i, s_j) \log \frac{1}{P(s_i/s_j)}$$ entropía de la fuente de Marcov de primer orden S

$$H_r(S) = \frac{H(S)}{\log r}$$ entropía, medida en unidades r-arias

$$H(\omega) = \omega \log \frac{1}{\omega} + (1 - \omega) \log \frac{1}{1 - \omega}$$ función entropía (fig. 2-3)

$$H(A) = \sum_{A} P(a) \log \frac{1}{P(a)}$$ entropía condicional de A (entropía a posteriori)

$$H(A/b_j) = \sum_{A} P(a/b_j) \log \frac{1}{P(a/b_j)}$$ entropía del alfabeto de entrada A (entropía a priori)

$$H(A/B) = \sum_{A,B} P(a, b) \log \frac{1}{P(a/b)}$$ equivocación de A con respecto a B

$$H(A, B) = \sum_{A,B} P(a, b) \log \frac{1}{P(a, b)}$$ entropía afín de A y B

$$I(A; B) = H(A) - H(A/B)$$ información mutua de A y B

$$I(a; B) = \sum_{B} P(b/a) \log \frac{P(b/a)}{P(b)}$$ información mutua condicional

$$H(A, B/C) = \sum_{A,B,C} P(a, b, c) \log \frac{1}{P(a, b/c)}$$

equivocación de A y B con respecto a C

$$H(A/B, C) = \sum_{A,B,C} P(a, b, c) \log \frac{1}{P(a/b, c)}$$

equivocación de A con respecto a B y C

$$I(A; B/C) = H(A/C) - H(A/B, C)$$

información mutua de A y B. conocido C

$$I(A; B; C) = I(A; B) - I(A; B/C)$$

información mutua de A, B y C

CAPITULO 1

INTRODUCCION

1-1. Lo que no es la teoría de la información.

Teoría de la información es un nombre muy significativo para designar una disciplina científica; al aplicarse, sin embargo, al tema de que trata este libro puede resultar algo decepcionante. Los orígenes de la teoría de la información datan de la publicación, por Claude E. Shannon, de un artículo en el *Bell System Technical Journal* en 1948 (Shannon, 1948) *. Shannon, dándose quizás cuenta de las cualidades poco atractivas de la palabra *información*, tituló su artículo «Una teoría matemática de la comunicación». Si nos referimos al significado usual de la palabra información, el artículo de Shannon trata de sus soportes, los símbolos, y no de la información misma. Estudia más bien la comunicación y los medios de comunicación que el, llamémosle, producto final de ella, la información.

Deseamos hacer una aclaración importante. Comenzando en el capítulo 2, deduciremos un cierto número de propiedades fundamentales de los símbolos empleados para transmitir la información. Aprenderemos que los símbolos deben obedecer ciertas leyes si han de ser capaces de transmitir información; relacionaremos las propiedades de los símbolos con la cantidad de información que pueden contener. El que un símbolo determinado contenga información, sin embargo, dependé de una serie de factores que no estudiaremos en este libro. Por ejemplo, «le soleil brille», suministra información a solamente algunos lectores. Un lenguaje común facilita la transmisión de información. Los factores sicológicos afectan también, de manera menos evidente, a la información. La frase «el sol brilla» puede tener para un

* Las referencias indicadas entre paréntesis pueden encontrarse en la lista de referencias del final del libro

sicópata un sentido más amplio que el meteorológico. Factores semánticos pueden dar lugar a que un mismo conjunto de palabras contenga información diferente para distintos interlocutores. Shannon (1948) ha comentado que «los aspectos semánticos de la comunicación son inaplicables al problema de ingeniería propiamente dicho». Weaver (1949) sostiene, sin embargo, que la inversa no es necesariamente cierta, que los aspectos de ingeniería (o técnicos) de la comunicación están relacionados con los aspectos semántico, sicológico y lingüístico. En el apartado 2-8 mostraremos la aplicación de la teoría desarrollada en este libro de lingüística. Salvo en este apartado 2-8, y en algunas notas del final de cada capítulo, no se estudiarán las aplicaciones específicas de la teoría de la información a otros dominios.

Desarrollaremos las ideas fundamentales de la teoría de la información, haciendo hincapié sobre su medida e interpretación. El lector puede estar interesado en investigar más en detalle la posible aplicación de la teoría de la información a algún otro campo. En este sentido las posibilidades son ilimitadas. El tema estudiado en el libro puede relacionarse con la información suministrada por un experimento estadístico (Lindley, 1956; Kullback, 1959; Grettenberg, 1962). Veremos que el concepto de entropía, fundamental en la teoría de la información tal como se desarrolla aquí, tiene al menos una semejanza de forma con la entropía de la termodinámica (Brillouin, 1956; Jaynes, 1959). Se ha considerado la aplicación de la teoría de la información a la sicología (Quastler, 1956), al arte (Pierce, 1961, págs. 260-267) y semántica (Bar-Hillel y Carnap, 1952). Finalmente, daremos referencia al lector de una interesante interpretación de algunos aspectos teológicos de la teoría de la información Elías, 1958).

I-2. Lo que es la teoría de la información.

El primer paso en nuestro estudio de la información consistirá en la definición de una medida de la información, investigando sus propiedades. Estas propiedades darán un sentido más práctico a la medida y ayudarán a relacionar la teoría matemática con el modelo físico que la motivó. Es importante resaltar, sin embargo, que la justificación de la definición de la medida de información no puede basarse estrictamente en la validad de las relaciones contenidas en su estructura. Está claro que podríamos establecer una estructura de la teoría de la in-

formación que, en sí misma, fuera consistente y demostrable. Aún así, tal estructura sin una ulterior justificación práctica, constituiría simplemente una disciplina matemática. Es solamente en las relaciones contenidas en la estructura, elevadas a magnitudes completamente independientes de ella misma, que puede encontrarse justificación a esta teoría. Así, deduciremos una definición de información y un grupo de relaciones que en sí mismas resulten válidas. La definición de información, sin embargo, no estará justificada por su consistencia interna, sino demostrando que las relaciones definen unas magnitudes que no están implicadas en la estructura misma de la teoría de la información. Con objeto de insistir sobre la necesidad de la existencia de una correspondencia entre el modelo matemático y el mundo físico, dedicaremos este capítulo de introducción a plantear algunas importantes cuestiones, que pueden ser formuladas independientemente de una medida particular de información. En los capítulos 2, 3 y 4 veremos como nuestra definición de información da respuesta, cuantitativa y matemáticamente, a esas cuestiones.

1-3. Codificación de la información.

Con objeto de exponer las ideas básicas de la teoría de la información, consideremos algunos ejemplos de transmisión de información. Nos limitaremos, en principio, a considerar un tipo particular pero importante de información, la información binaria.

La información contenida en las tarjetas perforadas, los mensajes transmitidos mediante sistemas de teletipo todo-nada o la información almacenada en los elementos biestables de las calculadoras electrónicas, constituyen unos cuantos ejemplos de esta clase de información. Con esta limitación se simplifican notablemente las consideraciones que deseamos hacer en el resto del capítulo.

Es interesante resaltar que, contrariamente a la creencia general, la representación binaria de la información no es relativamente reciente, sino conocida desde hace no poco tiempo. En efecto, una primera referencia fue dada por Matthew 5:37. «Sea tu comunicación Sí, Sí; No, No; ya que cualquiera es más que las cometas del infierno». Este punto de vista puede resultar en cierto modo extremo, por lo que a partir del capítulo 2 consideraremos la teoría de la información en función de ambas, información binaria y no binaria.

La tabla 1-1 muestra un ejemplo sencillo de representación de información no binaria en función de los dígitos binarios 0 y 1.

TABLA 1-1. CODIFICACIÓN BINARIA DE LOS DÍGITOS DECIMALES

Dígito decimal	*Representación binaria*
1	0001
0	0000
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001

La correspondencia entre los dígitos decimales y binarios definida por la tabla 1-1 constituye un ejemplo de *código.* Las 10 secuencias binarias se denominan *palabras código* y los 10 dígitos decimales *símbolos mensaje.* En el apartado 3-1 definiremos más cuidadosamente el concepto de código y palabra código. Por el momento, sin embargo, admitiremos cierta ambigüedad en la discusión. Es evidente que mediante el código de la tabla 1-1 podremos deducir la secuencia de dígitos binarios correspondiente a cualquier secuencia de dígitos decimales (símbolos mensaje). Recíprocamente, de una secuencia de dígitos binarios perteneciente a este código, podremos obtener una única secuencia de dígitos decimales.

La posibilidad de establecer a partir de una serie de palabras código binarias los correspondientes símbolos mensaje no es siempre una operación inmediata. Consideremos, por ejemplo, el código definido en la tabla 1-2.

TABLA 1-2. CÓDIGO BINARIO

Símbolos mensaje	*Palabras código*
s_1	0
s_2	01
s_3	001
s_4	111

Dada una secuencia de palabras código de la tabla, bien podemos no estar en situación de deducir un único conjunto de símbolos mensaje. La secuencia binaria

$$111001 \tag{1-1}$$

puede provenir de

$$s_4\, s_3 \tag{1-2}$$

o de

$$s_4\, s_1\, s_2 \tag{1-3}$$

El lector puede objetar indicando que la simple inclusión de una coma (o espacio) es suficiente para eludir el compromiso. Naturalmente, esto es cierto; sin embargo, el empleo de una coma (o espacio) no está de acuerdo con la definición de código binario. Si utilizamos una coma para separar las palabras, estamos empleando realmente tres símbolos diferentes, cero, uno y coma.

Resulta sencillo encontrar un código que no presente los inconvenientes del de la tabla 1-2. A una secuencia de palabras código perteneciente a la tabla 1-3 puede asociarse un conjunto único de símbolos mensaje. En este capítulo nos ocuparemos exclusivamente de esta clase de códigos.

TABLA 1-3. CÓDIGO BINARIO

Símbolos mensaje	*Palabras código*
s_1	0
s_2	10
s_3	110
s_4	1110

1-4. Un problema en la transmisión de información.

Con objeto de exponer algunas ideas sobre codificación y su relación con la medida de la información, consideraremos el problema siguiente. Se desea establecer un sistema de comunicación entre San Francisco y New York. El sistema debe transmitir, a intervalos regula-

res, datos sobre el estado del tiempo, debiendo hacer uso únicamente de un equipo de funcionamiento todo-nada (binario). Para simplificar la cuestión, clasificaremos el estado del tiempo en San Francisco dentro de una de las cuatro condiciones siguientes: soleado, nublado, lluvia o niebla. Estas cuatro condiciones constituyen los símbolos mensaje de la tabla 1-4. En esta tabla se indica asimismo la probabilidad de cada condición. Supongamos los cuatro estados equiprobables.

TABLA 1-4. ESTADO DEL TIEMPO EN SAN FRANCISCO

Mensajes	*Probabilidades*
Soleado	1/4
Nublado	1/4
Lluvia	1/4
Niebla	1/4

La siguiente correspondencia, llamada código A, muestra uno de los métodos posibles de codificar estos mensajes en secuencias de símbolos binarios.

Código A

Soleado	00	
Nublado	01	
Lluvia	10	(1-4)
Niebla	11	

Así, utilizando el código A, «soleado, niebla, niebla, nublado», se codificaría en la forma «00111101».

Es evidente que el código A es aceptable para transmitir esta información en el sentido de que, dada una secuencia de palabras código, podremos deducir una secuencia de mensajes que se corresponde biunívocamente con ella.

Está claro, asimismo, que con el empleo del código A es necesario enviar dos dígitos binarios (binits) * por mensaje. El lector podrá de-

* En el resto del libro emplearemos la contracción *binit* para designar un dígito binario. Es importante establecer una distinción entre binit (dígito binario) y bit (unidad de información que definiremos en el capítulo 2). Como veremos, en algunas circunstancias, un binit puede contener un bit de información.

mostrar fácilmente que no es posible encontrar otro código válido que haga uso de *menos* de 2 binits por mensaje.

Consideremos ahora el mismo problema presentado a un ingeniero de Los Angeles. Es decir, se desea establecer un sistema de comunicación semejante para transmitir el estado del tiempo de Los Angeles a New York. Sabemos que existen importantes diferencias meteorológicas entre el tiempo en San Francisco y Los Angeles. Una de ellas puede tenerse en cuenta clasificando el estado del tiempo en Los Angeles en soleado, nublado, lluvia y *bruma.* Aun cuando la diferencia entre niebla y bruma es notoria para un residente en una de esas ciudades, no interviene como factor fundamental en el diseño del sistema de comunicación. Ya que los cuatro estados se codifican en secuencias binarias, el significado real de una secuencia en particular no tiene influencia alguna desde el punto de vista de la comunicación.

Puede existir, sin embargo, una diferencia meteorológica que sí interviene en el planteamiento del problema de la comunicación. En justicia deberemos asignar probabilidades diferentes a cada uno de los cuatro estados posibles del clima de Los Angeles. Estas probabilidades aparecen en la tabla 1-5.

TABLA 1-5. ESTADO DEL TIEMPO EN LOS ANGELES

Mensajes	*Probabilidades*
Soleado	1/4
Nublado	1/8
Lluvia	1/8
Bruma	1/2

Si utilizamos el mismo código A para transmitir esta información, la solución será igual, pero no mejor, que en el sistema de comunicación de San Francisco. Esto es, usando el código A, enviaremos dos binits por mensaje, independientemente del estado del tiempo. Consideremos, sin embargo, la posibilidad de emplear para transmitir la información el siguiente código, denominado código B:

Código B

Soleado	10	
Nublado	110	
Lluvia	1110	(1-5)
Bruma	0	

En este caso, el mensaje «soleado, bruma, bruma, nublado», se transmitiría como «1000110».

Igual que antes, cualquier secuencia binaria establecida a partir de este código daría lugar a una secuencia única de mensajes. Es cierto, ya que la secuencia binaria correspondiente a un mensaje termina en 0, pudiendo interpretarse el 0 como referencia de fin de palabra código. Utilizando el código B, la longitud media L (en binits) de una palabra código tiene por valor

$$L = 2 \text{ Pr. (soleado)} + 3 \text{ Pr. (nublado)} + 4 \text{ Pr. (lluvia)} + 1 \text{ Pr. (bruma)}$$
$$= 2(1/4) + 3(1/8) + 4(1/8) + 1(1/2) = 15/8$$
$$1\,7/8 \text{ binits/mensaje.} \qquad (1\text{-}6)$$

Es decir, en el sistema de comunicación de Los Angeles a New York, hemos encontrado un procedimiento para transmitir información sobre el estado del tiempo que exige una media de 1 7/8 binits por mensaje, en lugar de 2 binits por mensaje. El lector puede comprobar que la aplicación del código B para transmitir desde San Francisco (tabla 1-4) conduciría a un valor medio $L = 2\,1/2$ binits por mensaje. De esta forma, hemos demostrado que es posible transmitir el mismo tipo de información desde Los Angeles, con una economía media por mensaje de aproximadamente un 6 por ciento. Una reducción de un 6 por ciento en el número de dígitos binarios a transmitir en un sistema de comunicación representa una ganancia realmente importante, aún más si tenemos en cuenta que se ha logrado por el simple hecho de modificar la forma de los mensajes enviados.

1-5. Algunas preguntas importantes.

El ejemplo del apartado anterior plantea varios problemas de naturaleza fundamental. En primer lugar, el hecho de obtener una ganancia de un 6 por ciento de manera tan simple incita nuestro apetito a una ulterior mejora. ¿Podremos obtener una nueva ganancia adoptando un código más ingenioso? Si tal es posible (y en nuestro ejemplo particular lo es). ¿Hasta dónde podremos llegar? Es decir, ¿cuál es el menor número de binits por mensaje necesarios para transmitir esta información? Una vez que hayamos calculado el valor mínimo de L, el problema práctico consistirá en construir el código a que

corresponde. ¿Cuáles son los métodos prácticos de síntesis de tal código?

La última de las preguntas sugeridas por nuestro ejemplo es «¿Por qué?» ¿Qué diferencia existe entre la situación del estado del tiempo en Los Angeles y San Francisco que nos ha permitido transmitir desde Los Angeles con un número menor de binits? Esta última cuestión es ciertamente fundamental. En otros términos, la pregunta puede plantearse en la forma siguiente: «¿Cuál es la naturaleza de la información?» El hecho de necesitar menos binits para especificar el estado del tiempo en Los Angeles implica que, en cierto sentido, el conocimiento del estado del tiempo en Los Angeles contiene una información menor que el conocimiento del estado del tiempo en San Francisco. Más adelante veremos que esta vaga noción de *cantidad de información* se concretará en la propia definición de medida de la información. En el ejemplo del apartado 1-4 es evidente que la definición de información está relacionada con la probabilidad de presencia de los diferentes mensajes.

En los tres siguientes capítulos iremos respondiendo a estas preguntas, definiendo una medida de la información basada en la probabilidad de los mensajes. Esto es, obtendremos el valor mínimo del número medio de binits por mensaje que debe utilizarse; deduciremos los métodos de síntesis de códigos que nos permitan alcanzar este mínimo, y, finalmente, discutiremos la naturaleza intrínseca de la información.

NOTAS

Nota 1. Existe un artículo introducción de McMillan (1953) en donde expone, de forma fácilmente accesible, la interpretación matemática de la teoría de la información. McMillan hace también una divertida descripción del aspecto matemático de la teoría de la información realizada por un ingeniero en comunicaciones.

Nota 2. Puede alcanzarse una idea del tremendo alcance de la teoría de la información (en su concepto más amplio) pasando revista a las aproximadamente cuatro mil referencias contenidas en la bibliografía de Stumper sobre teoría de la información (1953, 1955, 1957, 1960).

PROBLEMA

1-1. En el apartado 1-4 se definieron dos códigos, $\mathscr{A}$ y $\mathscr{B}$, utilizados en la transmisión del estado del tiempo en Los Angeles. La longitud media del código $\mathscr{A}$ era de dos binits por mensaje, y en la del código $\mathscr{B}$, 17/8 binits por mensaje. En el capítulo 4 demostraremos que la menor longitud media posible de un código en el problema de la tabla 1-5 es de 1 3/4 binits por mensaje. Asimismo se describirá un procedimiento para inducir tal código.

Sin estudiar el capítulo 4, intentar encontrar el código que corresponde a esta longitud mínima. Téngase presente que una secuencia de palabras de dicho código debe representar una secuencia única de mensajes.

CAPITULO 2

LA INFORMACION Y SUS FUENTES

2-1. Definición de información.

En el capítulo 1 se formularon una serie de preguntas fundamentales sobre la naturaleza de la información. Con objeto de darles respuesta, comenzaremos definiendo una medida de la información, demostrando a continuación que posee ciertas propiedades imputables a cualquier otra definición. Notemos, sin embargo, que el hecho de que sea posible demostrar lo razonable y la consistencia intrínseca de la definición, no es suficiente para justificarla. Se hará solamente dando respuesta a las preguntas del capítulo 1 (preguntas que no dependen de ninguna definición particular de información) basándose en la propia definición.

> *Definición.* Sea E un suceso que puede presentarse con probabilidad $P(E)$. Cuando E tiene lugar, decimos que hemos recibido
>
> $$I(E) = \log \frac{1}{P(E)} \qquad (2\text{-}1)$$
>
> unidades de información.

La elección de la base del logaritmo que interviene en la definición equivale a elegir una determinada unidad, ya que,

$$\log_a x = \frac{1}{\log_b a} \log_b x \qquad (2\text{-}2)$$

Si introducimos el logaritmo de base 2, la unidad correspondiente se denomina *bit* *

$$I(E) = \log_2 \frac{1}{P(E)} \quad \text{bits} \qquad (2\text{-}3a)$$

Empleando logaritmos naturales, la unidad de información recibe el nombre de *nat* **.

$$I(E) = \ln \frac{1}{P(E)} \quad \text{nats} \qquad (2\text{-}3b)$$

En el caso de logaritmos de base 10, la unidad de información es el Hartley. R. V. Hartley fue quien primero sugirió la medida logarítmica de la información (Hartley, 1928).

$$I(E) = \log_{10} \frac{1}{P(E)} \quad \text{Hartleys} \qquad (2\text{-}3c)$$

En general, empleando logaritmos de base r,

$$I(E) = \log_r \frac{1}{P(E)} \quad \text{unidades de orden } r \qquad (2\text{-}3d)$$

De la relación (2-2), vemos que

$$1 \text{ Hartley} = 3{,}32 \text{ bits} \qquad (2\text{-}4a)$$

$$1 \text{ nat} = 1{,}44 \text{ bits} \qquad (2\text{-}4b)$$

Notemos, también, que si $P(E) = 1/2$, será $I(E) = 1$ bit. Es decir, *un bit es la cantidad de información obtenida al especificar una de dos posibles alternativas igualmente probables.* Esta situación se presenta al lanzar una moneda al aire o al examinar la salida de un sistema de comunicación binario.

Con objeto de hacernos una idea de la cantidad de información transmitida por un moderno sistema de comunicación, consideremos una imagen de televisión. Puede imaginarse formada por una estructura de puntos negros, blancos y grises, dispuestos en 500 filas y 600 columnas aproximadamente. Admitiremos que cada uno de esos $500 \times 600 = 300.000$ puntos puede adoptar uno de 10 niveles de brillo diferentes, de manera que puede haber $10^{300.000}$ imágenes distintas

* *N. del T.*: Contracción de *binary unit*, en español, «unidad binaria».
** *N. del T.*: Contracción de *natural unit*, «unidad natural».

de T. V. Si todas son igualmente probables, la probabilidad de una imagen cualquier es igual a $1/10^{300.000}$ y la cantidad de información que contiene *

$$I(E) = 300.000 \log 10$$
$$\approx 10^6 \text{ bits}$$

Puede compararse la información contenida en una imagen de televisión, calculada anteriormente, con la información contenida en 1.000 palabras emitidas por un locutor de radio. Supongamos que el locutor tiene un vocabulario de 10.000 palabras y que ha elegido entre ellas 1.000 completamente al azar (cifras que pueden considerarse aproximadas a la realidad, en algunos casos). La probabilidad de una secuencia de 1.000 palabras es $1/(10.000)^{1.000}$ y la cantidad de información contenida

$$I(E) = 1.000 \log 10.000$$
$$\approx 1{,}3 \times 10^4 \text{ bits}$$

Así, pues, una imagen de TV equivale a 100 palabras (radio).

2-2. Fuente de información de memoria nula.

Es interesante y útil describir matemáticamente un mecanismo generador de información. En este capítulo en consecuencia, definiremos una fuente de información discreta, tal como la mostrada en la figura 2-1.

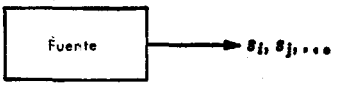


FIG. 2-1. Fuente de información.

Imaginemos la fuente emitiendo una secuencia de símbolos pertenecientes a un alfabeto finito y fijo, $S = \{s_1, s_2, \ldots, s_q\}$. Los símbolos emitidos sucesivamente se eligen de acuerdo con una ley fija de pro-

* En adelante escribiremos el logaritmo en base 2 de x simplemente como $\log x$, omitiendo el subíndice 2 del «log». Asimismo, expresaremos el logaritmo natural como $\ln x$. En todos los demás casos indicaremos la base mediante un subíndice (p. e., $\log_{10} x$).

babilidad. Ocasionalmente nos referimos a la fuente misma como S, sin que esto deba dar lugar a confusión. En la fuente más sencilla admitiremos que los símbolos emitidos son estadísticamente independientes. Tal fuente de información se conoce como *fuente de memoria nula* y puede describirse completamente mediante el alfabeto fuente S y las probabilidades con que los símbolos se presentan:

$$P(s_1),\ P(s_2),\ \ldots,\ P(s_q)$$

Puede calcularse la información media suministrada por una fuente de información de memoria nula en la forma siguiente: La presencia de un símbolo s_i corresponde a una cantidad de información igual a

$$I(s_i) = \log \frac{1}{P(s_i)} \quad \text{bits}$$

La probabilidad de que aparezca es precisamente $P(s_i)$, de modo que la cantidad *media* de información por símbolo de la fuente es

$$\sum_S P(s_i)\, I(s_i) \quad \text{bits}$$

donde $\sum_S$ indica la suma extendida a q símbolos de la fuente S. Esta magnitud, cantidad media de información por símbolo de la fuente, recibe el nombre de *entropía* $H(S)$ de la fuente de memoria nula *.

$$H(S) \triangleq \sum_S P(s_i) \log \frac{1}{P(s_i)} \quad \text{bits} \qquad (2\text{-}5a)$$

Ejemplo 2-1. Consideremos la fuente $S = \{s_1, s_2, s_3\}$ con $P(s_1) = 1/2$ y $P(s_2) = P(s_3) = 1/4$. Entonces

$$H(S) = 1/2 \log 2 + 1/4 \log 4 + 1/4 \log 4$$
$$= 3/2 \text{ bits}$$

Si medimos $I(s_i)$ en unidades de orden r, $H(S)$ vendrá dada en la misma unidad, y tendremos

$$H_r(S) = \sum_S P(s_i) \log_r \frac{1}{P(s_i)} \quad \text{unidades de orden } r \qquad (2\text{-}5b)$$

* La relación existente entre la entropía manejada en la teoría de la información y la entropía de la termodinámica ha sido analizada por Brillouin (1956).

De la ecuación (2-2) se deduce

$$H_r(S) = \frac{H(S)}{\log r} \tag{2-5c}$$

Nótese que de la definición dada en (2-1), $I(s_i)$ puede interpretarse como la información necesaria para que la presencia de s_i sea cierta. Asimismo $H(S)$ puede ser *bien* el valor medio de la información por símbolo suministrada por la fuente, *o* el valor medio de la incertidumbre de un observador antes de conocer la salida de la fuente. En la continuación usaremos ambas interpretaciones. En primer lugar, sin embargo, demostraremos algunas propiedades sencillas de la entropía de una fuente.

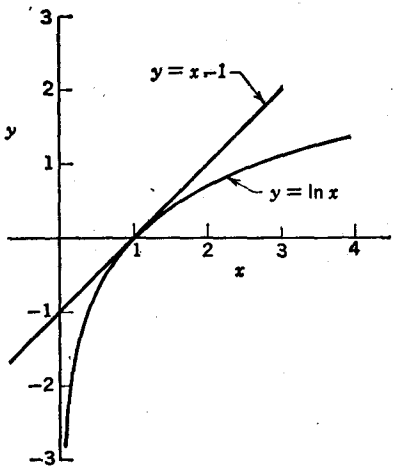


FIG. 2-2. Logaritmo natural de x y $x - 1$.

2-3. Propiedades de la entropía.

Con objeto de deducir algunas propiedades de la entropía consideraremos una propiedad particular del logaritmo. La figura 2-2 representa la curva de variación del logaritmo natural de x, así como la recta definida por la ecuación $y = x - 1$.

Fácilmente puede demostrarse que la recta se mantiene siempre por debajo de la curva $y = ln\, x$. Así, pues, podemos escribir la inecuación

$$\ln x \leqq x - 1 \tag{2-6}$$

que será una igualdad si, y solamente si, $x = 1$.

Multiplicando (2-6) por -1, deducimos una nueva inecuación

$$\ln \frac{1}{x} \geqq 1 - x \tag{2-7}$$

igualdad solamente si $x = 1$.

Deduciremos finalmente una última inecuación a partir de (2-6). Sean $x_1, x_2, \ldots, x_q$ e $y_1, y_2, \ldots, y_q$ dos conjuntos de probabilidades. Es decir

$$x_i \geqq 0 \qquad y_j \geqq 0, \qquad \text{para cualquier } i \text{ y } j$$

$$\sum_{i=1}^{q} x_i = \sum_{j=1}^{q} y_j = 1$$

Haciendo uso de (2-2), escribiremos

$$\sum_{i=1}^{q} x_i \log \frac{y_i}{x_i} = \frac{1}{\ln 2} \sum_{i=1}^{q} x_i \ln \frac{y_i}{x_i}$$

y aplicando la inecuación (2-6) a cada término de la suma,

$$\begin{aligned} \sum_{i=1}^{q} x_i \log \frac{y_i}{x_i} &\leqq \frac{1}{\ln 2} \sum_{i=1}^{q} x_i \left(\frac{y_i}{x_i} - 1 \right) \\ &\leqq \frac{1}{\ln 2} \left(\sum_{i=1}^{q} y_i - \sum_{i=1}^{q} x_i \right) \\ &\leqq 0 \end{aligned} \tag{2-8a}$$

o

$$\sum_{i=1}^{q} x_i \log \frac{1}{x_i} \leqq \sum_{i=1}^{q} x_i \log \frac{1}{y_i} \tag{2-8b}$$

que será una igualdad para cualquier valor de i, solamente si $x_i = y_i$.

Como se dijo anteriormente, la entropía de una fuente podía interpretarse como la información media por símbolo emitida por la fuente. Es lógico, por lo tanto, analizar en qué modo la entropía depende de la probabilidad de los diferentes símbolos de la fuente. En particular, sería interesante conocer cuanta información puede suministrar una fuente de información de memoria nula.

Supongamos una fuente de memoria nula, definida por su alfabeto $S = \{s_i\}$, $i = 1, 2, \ldots, q$, y sus probabilidades $P(s_i)$, $i = 1, 2, \ldots, q$. La $H(S)$ viene dada por

$$H(S) = \sum_{i=1}^{q} P_i \log \frac{1}{P_i} \tag{2-9}$$

Consideremos la expresión

$$\log q - H(S) = \sum_{i=1}^{q} P_i \log q - \sum_{i=1}^{q} P_i \log \frac{1}{P_i}$$

$$= \sum_{i=1}^{q} P_i \log q P_i$$

$$= \log e \sum_{i=1}^{q} P_i \ln q P_i \tag{2-10}$$

El último miembro se dedujo haciendo intervenir la relación (2-2). Aplicando las inecuaciones (2-7) a (2-10), se llega a la expresión

$$\log q - H(S) \geqq \log e \sum_{i=1}^{q} P_i \left(1 - \frac{1}{q P_i}\right)$$

$$\geqq \log e \left(\sum_{i=1}^{q} P_i - \frac{1}{q} \sum_{i=1}^{q} \frac{P_i}{P_i}\right)$$

$$\geqq 0 \tag{2-11}$$

Así, pues, $H(S)$ es siempre menor o igual que $\log q$. De la condición que transforma (2-7) en una igualdad se deduce la igualdad de (2-11) si, y solamente si, $P_i = 1/q$. Es decir, hemos demostrado que en una fuente de información de memoria nula con un alfabeto de q símbolos, *el valor máximo de la entropía es precisamente log q, alcanzándose solamente si todos los símbolos de la fuente son equiprobables.*

Un ejemplo partícularmente importante de fuente de información de memoria nula corresponde a una fuente binaria de memoria nula. En tal fuente, el alfabeto se reduce a $\{0, 1\}$. La probabilidad de un 0 es ω y la de un 1, $1 - \omega$. Llamaremos ϖ a $1 - \omega$. Calcularemos la entropía a partir de la fórmula (2-5)

$$H(S) = \omega \log \frac{1}{\omega} + \varpi \log \frac{1}{\varpi} \quad \text{bits} \tag{2-12}$$

La función ω (2-12), aparece con frecuencia en los problemas de la teoría de la información. Por esta razón se acostumbra a representar con un símbolo especial. Por definición

$$H(\omega) = \omega \log \frac{1}{\omega} + \varpi \log \frac{1}{\varpi} \tag{2-13}$$

que llamaremos *función entropía.* Hay que señalar la diferencia existente entre (2-12) y (2-13). $H(S)$ determina la entropía de una fuente particular S, mientras $H(\omega)$ es una función de la variable ω definida en el intervalo [0, 1]. El significado del símbolo $H(\cdot)$ depende, en definitiva, de la variable. Otro punto importante es que

$$\lim_{\omega \to 0} \omega \log \omega = 0$$

y así por definición

$$0 \log 0 = 0$$

En la Fig. 2-3 se ha representado la curva de variación $H(\omega)$ en función de ω, en el intervalo [0, 1] de la variable.

Nótese que si la salida de la fuente binaria es ciertá (bien $\omega = 0$ u $\omega = 1$), la fuente no suministra ninguna información. El valor medio de la información aportada por un símbolo de la fuente binaria alcanza su máximo en el caso en que ambos, 0 y 1, sean igualmente probables, siendo este valor máximo igual a log 2, es decir, 1 bit.

La salída de una fuente binaria está constituida por dígitos binarios o binits. Así una secuencia de binits producida por una fuente de información binaria de memoria nula, de 0s y 1s equiprobables, suministra un bit de información por binit. Si 0s y 1s no son igualmente probables, la cantidad de información dada por un binit será menor

o mayor de 1 bit dependiendo de los valores de las probabilidades [apartado (2-1)]. La cantidad *media* de información suministrada por un binit de tal fuente, sin embargo, será siempre menor o igual a 1 *bit* por *binit* (fig. 2-3).

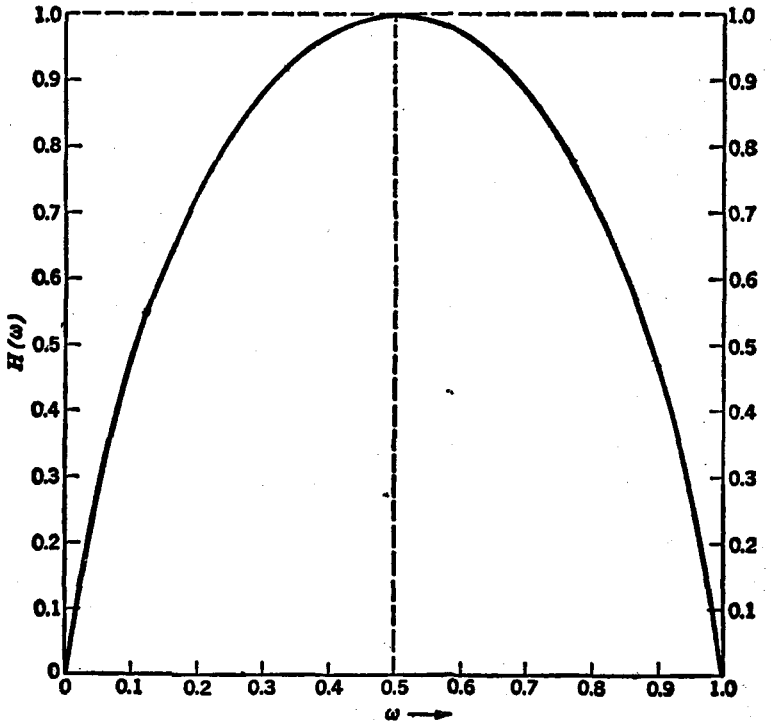


FIG. 2-3. $H(\omega)$, función entropía.

Hay que notar que la cantidad máxima de información dada por una fuente de memoria nula de q símbolos, crece lentamente al aumentar q. De hecho, la cantidad máxima de información crece con el logaritmo del número de símbolos de la fuente, de modo que para duplicar la cantidad máxima de información por símbolo en una fuente de q símbolos, sería necesaria una fuente de q^2 símbolos.

2-4. Extensiones de una fuente de memoria nula.

A lo largo de la discusión que mantendremos en los capítulos siguientes sobre las propiedades de las fuentes y los canales de información, veremos el interés que presenta el tratamiento de grupos de símbolos en lugar de símbolos aislados. Por ejemplo, en el caso de la fuente binaria considerada en el apartado anterior, puede imaginarse que los binits son emitidos en grupos de dos. De esta forma, puede considerarse como equivalente a una fuente de cuatro símbolos, 00, 01, 10 y 11. Esta idea puede generalizarse más aún. Si se imagina la fuente original emitiendo grupos de tres binits. Entonces, puesto que hay ocho secuencias binarias posibles de longitud 3, sería equivalente a una fuente con un alfabeto de ocho símbolos.

En general, si tenemos una fuente de memoria nula, S, con un alfabeto $\{s_1, s_2, \ldots, s_q\}$, podemos agrupar las salidas en paquetes de n símbolos. Tendremos, pues, q^n secuencias de salidas distintas. Formalizaremos este concepto con la siguiente definición.

> *Definición.* Sea S una fuente de información de memoria nula, con un alfabeto $\{s_1, s_2, \ldots, s_q\}$. Sea P_i la probabilidad correspondiente a s_i. La extensión de orden n de S, S^n, es una fuente de memoria nula de q^n símbolos, $\{\sigma_1, \sigma_2, \ldots, \sigma_{qn}\}$. El símbolo σ_i corresponde a una secuencia de n de los sq símbolos. La probabilidad de σ_i, $P(\sigma_i)$, es precisamente la probabilidad de la secuencia correspondiente. Es decir, si σ_i representa la secuencia $(s_{i_1}, s_{i_2}, \ldots s_{i_n})$, $P(\sigma_i) = P_{i_1} \cdot P_{i_2} \ldots P_{i_n}$.

Puesto que un símbolo de la extensión de orden n, S^n, de la fuente de la memoria nula S, corresponde a n símbolos de S, es de suponer * que la entropía por símbolo de S^n sea n veces mayor que la de S. La demostración no es difícil. Sea σ_i el símbolo de S^n que corresponde a la secuencia $(s_{i_1}, s_{i_2}, \ldots, s_{i_n})$ de S. Según esto

$$H(S^n) = \sum_{S^n}^{q^n} P(\sigma_i) \log \frac{1}{P(\sigma_i)} \qquad (2\text{-}14)$$

donde la suma se extiende a los q^n símbolos de S^n. Al tratar de una

* Es importante recordar que, de acuerdo con nuestra definición, la extensión de *primer* orden de S es la fuente misma.

fuente y sus extensiones, emplearemos la notación $\sum_{S^n}$ para indicar la suma generalizada a todos los símbolos de la extensión de orden n.

La suma aplicada a los q^n símbolos de S^n es equivalente a n sumandos, cada uno de los cuales se extiende a los q símbolos de S, por ser $\sigma_i = (s_{i_1}, s_{i_2}, \ldots, s_{i_n})$. Por ejemplo, puesto que en una fuente de memoria nula $P(\sigma_i) = P_{i_1} \cdot P_{i_2} \ldots P_{i_n}$,

$$\begin{aligned}\sum_{S} P(\sigma_i) &= \sum_{S^n} P_{i_1} P_{i_2} \ldots P_{i_n} \\ &= \sum_{i_1=1}^{q} \sum_{i_2=1}^{q} \ldots \sum_{i_n=1}^{q} P_{i_1} P_{i_2} \ldots P_{i_n} \\ &= \sum_{i_1=1}^{q} P_{i_1} \sum_{i_2=1}^{q} P_{i_2} \ldots \sum_{i_n=1}^{q} P_{i_n} \\ &= 1 \end{aligned} \tag{2-15}$$

La ecuación (2-14) puede escribirse en la forma

$$\begin{aligned}H(S^n) &= \sum_{S^n} P(\sigma_i) \log \frac{1}{P_{i_1} P_{i_2} \ldots P_{i_n}} \\ &= \sum_{S^n} P(\sigma_i) \log \frac{1}{P_{i_1}} + \sum_{S^n} P(\sigma_i) \log \frac{1}{P_{i_2}} \\ &\quad + \ldots + \sum_{S^n} P(\sigma_i) \log \frac{1}{P_{i_n}} \end{aligned} \tag{2-16}$$

Los n sumandos son similares; tomando el primero de ellos

$$\begin{aligned}\sum_{S} P(\sigma_i) \log \frac{1}{P_{i_1}} &= \sum_{S^n} P_{i_1} P_{i_2} \ldots P_{i_n} \log \frac{1}{P_{i_1}} \\ &= \sum_{i_1=1}^{q} P_{i_1} \log \frac{1}{P_{i_1}} \sum_{i_2=1}^{q} P_{i_2} \ldots \sum_{i_n=1}^{q} P_{i_n} \\ &= \sum_{i_1=1}^{q} P_{i_1} \log \frac{1}{P_{i_1}} \\ &= \sum_{S} P_{i_1} \log \frac{1}{P_{i_1}} \\ &= H(S) \end{aligned} \tag{2-17}$$

introduciendo esta relación en (2-16), se llega a la expresión

$$H(S^n) = n\,H(S) \tag{2-18}$$

Ejemplo 2-2. Consideremos la extensión de segundo orden de la fuente del ejemplo 2-1. Recordemos que la fuente tenía un alfabeto $S = \{s_1, s_2, s_3\}$, con $P(s_1) = 1/2$ y $P(s_2) = P(s_3) = 1/4$. Así la fuente S^2 tendrá los nueve símbolos siguientes:

Símbolos de S^2	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9
Secuencia correspondiente a los símbolos de S	s_1s_1	s_1s_2	s_1s_3	s_2s_1	s_2s_2	s_2s_3	s_3s_1	s_3s_2	s_3s_3
Probabilidad P (σ_i)	1/4	1/8	1/8	1/8	1/16	1/16	1/8	1/16	1/16

$$H(S^2) = \sum_{S^2} P(\sigma_i) \log \frac{1}{P(\sigma_i)}$$

$$= 1/4 \log 4 + 4 \times 1/8 \log 8 + 4 \times 1/16 \log 16$$

$$= 3 \text{ bits/símbolo}$$

2-5. Fuente de información de Markov.

La fuente de memoria nula considerada hasta aquí resulta demasiado limitada en algunas aplicaciones. Un tipo de fuente de información de q símbolos, más general que la de memoria nula, ya estudiada, consiste en aquella en que la presencia de un determinado símbolo s_i depende de un número finito m de símbolos precedentes. Tal fuente (llamada fuente de Markov de orden m) viene definida por su alfabeto, S, y el conjunto de probabilidades condicionales *.

$$P(s_i/s_{i_1}, s_{i_2}, \ldots, s_{i_m}) \quad \text{para} \quad i = 1, 2, \ldots, q; \; i_0 = 1, 2, \ldots, \tag{2-19}$$

En una fuente de Markov de orden m, la probabilidad de un símbolo cualquiera viene determinada por los m símbolos que lo preceden. En cualquier momento, por lo tanto, definiremos el *estado* de la fuente de Markov de orden m por los m símbolos precedentes. Puesto que existen q símbolos distintos, una fuente de Markov de orden m admitirá q^m estados posibles. Al emitir la fuente nuevos símbolos, el estado cambia. Un procedimiento simple de estudiar el comportamiento de la fuente consiste en utilizar un *diagrama de estados*. En este diagrama cada uno de los q^m estados posibles de la fuente se repre-

* La secuencia de símbolos implicada por la probabilidad condicional $P(s_i/s_{j1}, s_{j2}, \ldots, s_{jm})$ es $s_{j1}, s_{j2}, \ldots, s_{jm}, s_i$. Es decir, s_i va detrás de s_{im}.

senta por un punto, indicándose mediante flechas las transiciones entre estados.

Ejemplo 2-3. Consideremos una fuente de Markov de segundo orden con un alfabeto binario $S = \{0,1\}$. Supongamos que las probabilidades condicionales son

$$P(0/00) = P(1/11) = 0.8$$

$$P(1/00) = P(0/11) = 0.2$$

$$P(0/01) = P(0/10) = P(1/01) = P(1/10) = 0.5$$

Por ser q igual a 2 y haber supuesto la fuente de Markov de *segundo* orden, tendremos *cuatro* estados diferentes, 00, 01, 10, 11. La figura 2-4 representa el diagrama de estados de la fuente. Los cuatro estados vienen representados por cuatro puntos. Las transiciones posibles, mediante flechas entre estado y estado, indicándose sobre cada una de ellas la probabilidad asociada. Por ejemplo, si nos encontramos en el estado 00 podremos pasar al 01 ó al 00, pero nunca a los estados 10 y 11. La probabilidad de permanecer en el estado 00 es 0,8 y la de pasar al 01, según puede verse, 0,2.

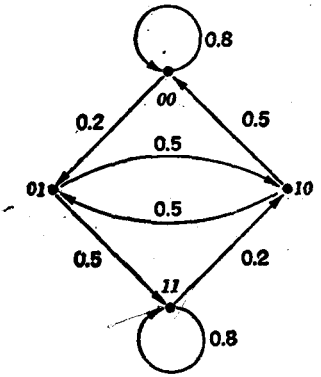


FIG. 2-4. Diagrama de estados de una fuente de Markov de segundo orden.

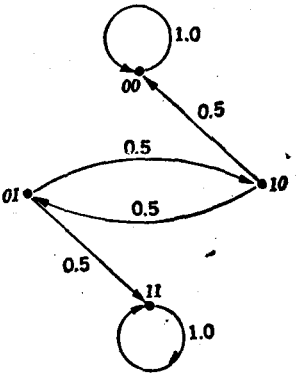


FIG. 2-5. Diagrama de estados de una fuente de Markov de segundo orden no ergódica.

En el estudio de las fuentes de información de Markov de orden m nos limitaremos a considerar las denominadas fuentes *ergódicas*. Para el matemático y el estadista matemático, el concepto de ergodicidad y las condiciones bajo las cuales una fuente es ergódica están en cierto modo relacionados. Para nuestros fines, sin embargo, el

concepto de fuente ergódica es la sencillez misma. Una fuente ergódica es meramente aquella que, observada durante un tiempo suficientemente largo, emite (con probabilidad 1) una secuencia «típica» de símbolos. En realidad la existencia de fuentes con esta probabilidad es tan natural que algunos lectores encontrarán dificultades en describir una fuente que *no* sea ergódica. Daremos a continuación un ejemplo de fuente de información *no ergódica.*

Ejemplo 2-4. Consideremos una fuente de Markov de segundo orden con un alfabeto binario $S = \{0,1\}$. Supongamos que las probabilidades condicionales son

$$P(0/01) = P(0/10) = P(1/01) = P(1/10) = 0.5$$

$$P(0/00) = P(1/11) = 1.0$$

$$P(1/00) = P(0/11) = 0$$

Existirán cuatro estados —00, 01, 10, 11—, como en el ejemplo anterior. La figura 2-5 representa el diagrama de estados de la fuente. De ahí se deduce que si en un momento dado se alcanza uno de los estados 00 ó 11, se permanecerá en él indefinidamente. Asimismo, si seleccionamos al azar uno de los cuatro estados posibles (cada estado será elegido con una probabilidad 1/4), comenzando en él las observaciones, veremos que después de que un número elevado de transiciones de estado haya tenido lugar, nos encontraremos en el 00 con probabilidad 0,5. Es decir, cuando la fuente está emitiendo símbolos, y a partir de un tiempo suficientemente largo, emitirá un 0 ó un 1 con la misma probabilidad, 0,5. *Dada* una secuencia cualquiera de la fuente, sin embargo, después de una espera suficiente, encontraremos con casi absoluta seguridad *todo* ceros o *todo* unos. En otras palabras (con probabilidad 1), no existe ninguna secuencia típica; no se trata de una fuente ergódica.

La discusión anterior señala la rareza, desde nuestro punto de vista, de las fuentes no ergódicas. Si seleccionamos el estado inicial de una fuente de Markov (de acuerdo con el conjunto de probabilidades propias a cada estado) y dejamos transcurrir un gran número de transiciones de estado, sabemos que existirá una probabilidad finita de que se presente cada uno de ellos.

Además, como se sugirió en el párrafo anterior con el empleo de la palabra *típica,* en una fuente ergódica los estados que realmente aparecen en una secuencia larga lo harán (con probabilidad 1) con las mismas probabilidades. Una propiedad más de las fuentes de Markov ergódicas que hay que destacar (Feller, 1950) es que la distribución de probabilidades de un conjunto de estados que se presentan después de producirse un gran número de transiciones (o, análogamente, la

distribución de estados en una secuencia de salida típica) *no* depende de la distribución inicial con que son elegidos los diferentes estados. Existe una distribución de probabilidades única para un conjunto de estados de una fuente de Markov *ergódica,* y los estados en cualquier secuencia suficientemente larga, se presentarán (con probabilidad 1) de acuerdo con esa distribución. Esta distribución única recibe el nombre de *distribución estacionaria* del proceso ergódico de Markov; *puesto que la distribución estacionaria no depende de la distribución inicial con que los estados son escogidos, puede calcularse directamente a partir de las probabilidades condicionales de los símbolos.* Por ejemplo, en la fuente de Markov definida en la figura 2-4, puede demostrarse que la distribución estacionaria es

$$\begin{aligned} P(00) &= P(11) = 5/14 \\ P(01) &= P(10) = 2/14 \end{aligned} \tag{2-20}$$

Cuando definimos las probabilidades condicionales de los símbolos $P(s_i/s_{j_1}, s_{j_2}, \ldots, s_{j_m})$ de un proceso ergódico de Markov de orden m, implícitamente definimos también las q^m probabilidades de estado $P(s_{j_1}, s_{j_2}, \ldots, s_{j_m})$. Combinando estas dos probabilidades se obtiene la probabilidad del *suceso simultáneo,* «fuente en el estado definido por $(s_{j_1}, s_{j_2}, \ldots s_{j_m})$ y s_i presente». Esta probabilidad es precisamente

$$P(s_{j_1}, s_{j_2}, \ldots, s_{j_m}, s_i) = P(s_i/s_{j_1}, s_{j_2}, \ldots, s_{j_m})\, P(s_{j_1}, s_{j_2}, \ldots, s_{j_m}) \tag{2-21}$$

Hay que notar que el problema de calcular las probabilidades de estado de una fuente ergódica de Markov a partir de las probabilidades condicionales de la fuente, no se ha tratado realmente. En general es una labor complicada que el lector puede encontrar detallada en los artículos de Feller (1950) o Bharucha-Reid (1960). Todo lo que que aquí podemos decir es que las probabilidades de estado *pueden* calcularse conociendo las probabilidades condicionales de los símbolos.

La información media suministrada por una fuente de Markov de orden m * puede calcularse de la forma siguiente: Si nos encontramos en el estado definido por $(s_{j_1}, s_{j_2}, \ldots, s_{j_m})$ (es decir, los m símbolos emitidos anteriormente fueron $s_{j_1}, s_{j_2}, \ldots, s_{j_m}$), la probabilidad condicional de recibir el símbolo s_i es $P(s_i/s_{j_1}, s_{j_2}, \ldots, s_{j_m})$. La información ob-

* En adelante omitiremos la palabra *ergódico* al hablar de tales fuentes. El resto del libro tratará exclusivamente del caso ergódico.

tenida si s_i se presenta cuando estamos en el estado $(s_{j_1}\ s_{j_2},\ \ldots,\ s_{j_m})$, según (2-1), es

$$I(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m}) = \log \frac{1}{P(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m})} \tag{2-22}$$

Por lo tanto, la cantidad media de información por símbolo *cuando nos encontramos en el estado* $(s_{j_1}\ s_{j_2}\ \ldots,\ s_{j_m})$ viene dada por la ecuación (2-23):

$$H(S/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m}) = \sum_{S} P(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m})\ I(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m}) \tag{2-23}$$

La cantidad media de información o entropía de la fuente de *Markov* de orden m, se obtendrá calculando el valor medio de esta cantidad, extendida a los q^m estados posibles.

$$H(S) = \sum_{S^m} P(s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m})\ H(S/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m}) \tag{2-24a}$$

Al escribir (2-24*a*) hemos supuesto que el estado $(s_{j_1},\ s_{j_2},\ \ldots\ s_{j_m})$ es equivalente a un símbolo * de S^m. Sustituyendo (2-23) en (2-24*a*), se llega a

$$\begin{aligned} H(S) &= \sum_{S^m} P(s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m}) \sum_{S} P(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m}) \\ &\qquad \times \log \frac{1}{P(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m})} \\ &= \sum_{S^{m+1}} P(s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m})\ P(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m}) \\ &\qquad \times \log \frac{1}{P(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m})} \\ &= \sum_{S^{m+1}} P(s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m},\ s_i) \\ &\qquad \times \log \frac{1}{P(s_i/s_{j_1},\ s_{j_2},\ \ldots,\ s_{j_m})} \end{aligned} \tag{2-24b}$$

habiendo hecho uso de la relación (2-21) en la última transformación.

* En términos rigurosos no se ha definido aún S^m, extensión de orden m de una fuente de *Markov*. La introducción de S^m en (2-24) no presenta, sin embargo, ninguna ambigüedad. La definición completa se dará en el apartado 2-7.

Nótese que si S fuese de memoria nula en lugar de Markov,

$$P(s_i/s_{j_1}, s_{j_2}, \ldots, s_{j_m}) = P(s_i)$$

y (2-24*b*) se reduce a (2-5*a*).

Ejemplo 2-5. Consideremos la fuente de Markov de la fig. 2-4. Su distribución estacionaria viene definida por (2-20). Las probabilidades más significativas están resumidas en la tabla 2-1.

TABLA 2-1. PROBABILIDADES DE LA FUENTE DE MARKOV DE LA FIG. 2-4

s_j, s_k, s_j	$P(s_i/s_j, s_k)$	$P(s_j, s_k)$	$P(s_j, s_k, s_i)$
000	0,8	5/14	4/14
001	0,2	5/14	1/14
010	0'5	2/14	1/14
011	0,5	2/14	1/14
100	0,5	2/14	1/14
101	0,5	2/14	1/14
110	0,2	5/14	1/14
111	0,8	5/14	4/14

La entropía se calculará a partir de (2-24*b*).

$$H(S) = \sum_{S^2} P(s_j, s_k, s_i) \log \frac{1}{P(s_i/s_j, s_k)}$$

$$= 2 \times 4/14 \log 1/0.8 + 2 \times 1/14 \log 1/0.2 + 4 \times 1/14 \log 1/0.5$$

$$= 0.81 \text{ bit/binit}$$

2-6. Fuente afín.

Dada una fuente de Markov de orden m, se puede, en principio, calcular su distribución estacionaria; la distribución de estados de la fuente de Markov. En una fuente de *primer orden,* el conjunto de estados es idéntico al conjunto de símbolos de la fuente, y la distribución estacionaria corresponde directamente a la distribución de probabilidades (incondicionales) de primer orden de los símbolos de la fuente. En una fuente de Markov de orden superior, la distribución de probabilidades de los símbolos de primer orden puede obtenerse fácilmente a partir de la distribución estacionaria. En la fuente de Markov a que corresponde la distribución estacionaria (2-20), por

ejemplo, puede demostrarse que las probabilidades de los símbolos de primer orden son $P(0) = P(1) = 1$. Utilizando las probabilidades de los símbolos de primer orden, puede definirse otra fuente.

> *Definición.* Supongamos que el alfabeto de una fuente de Markov de orden m es $S = \{s_1, s_2, \ldots, s_q\}$ y sean $P_1, P_2, \ldots, P_q$ las probabilidades de los símbolos de primer orden de la fuente. La *fuente afín* de S, llamada $\bar{S}$, es la fuente de información de memoria nula de alfabeto idéntico al de S, y de símbolos de probabilidades $P_1, P_2, \ldots, P_q$.

Por ejemplo, dada la simetría del diagrama de estados mostrado en la figura 2-4, 0 y 1 son igualmente probables. Así, pues, la fuente afín de la fuente de la figura 2-4 es una fuente binaria de memoria nula, de símbolos de entrada equiprobables y $H(\bar{S}) = 1$. Hay que destacar que la afín de una fuente S de *memoria nula*, es S *misma*. Demostraremos que la entropía de la fuente afín $\bar{S}$ nunca es inferior a la entropía de S. Este hecho tiene un significado importante. Las dos fuentes, S y $\bar{S}$, tienen las mismas probabilidades de primer orden. Difieren solamente en el hecho de que S cumple un requisito suplementario, consistente en las probabilidades *condicionales* de los símbolos impuestas a sus secuencias de salida. Esta limitación, por lo tanto, hace decrecer la cantidad media de información que fluye de la fuente.

Con objeto de simplificar el cálculo, probaremos que $H(\bar{S})$ es mayor o igual que $H(S)$ cuando S es una fuente de Markov de *primer orden*. La prueba para una fuente de orden m se deduce directamente por extensión.

Sea S una fuente de Markov de primer orden, de símbolos $s_1, s_2, \ldots, s_q$, con probabilidades condicionales $P(s_i/s_j), i, j = 1, 2, \ldots, q$. Supongamos que $P_1, P_2, \ldots, P_q$ son las probabilidades de primer orden de los símbolos de S y sea $\bar{S}$ la fuente afín. Si definimos $P(s_j, s_i)$ como la probabilidad afín de que estando la fuente en el estado especificado s_j se presente s_i, podremos escribir de acuerdo con (2-21)

$$P(s_j, s_i) = P(s_i/s_j)\, P_j \tag{2-25}$$

Examinaremos a continuación la doble suma

$$\sum_{S^2} P(s_j, s_i) \log \frac{P_j P_i}{P(s_j, s_i)} \tag{2-26}$$

Según (2-8*a*) comprobamos que esta suma es menor o igual a 0, siendo igual únicamente si

$$P(s_j, s_i) = P_j P_i \quad \text{para cualquier valor de } i \text{ y } j \tag{2-27}$$

Combinando (2-25) y (2-26) y escribiendo desigualdad, tendremos

$$\sum_{S^2} P(s_j, s_i) \log \frac{P_i}{P(s_i/s_j)} \leq 0$$

o

$$\sum_{S^2} P(s_j, s_i) \log \frac{1}{P(s_i/s_j)} \leq \sum_{S^2} P(s_j, s_i) \log \frac{1}{P_i}$$

$$\leq \sum_{i=1}^{q} \sum_{j=1}^{q} P(s_j, s_i) \log \frac{1}{P_i} \tag{2-28}$$

La suma extendida a todos los valores de j puede calcularse inmediatamente notando que el logaritmo es independiente de j y que

$$\sum_{j=1}^{q} P(s_j, s_i) = P_i$$

de modo que

$$\sum_{S^2} P(s_j, s_i) \log \frac{1}{P(s_i/s_j)} \leq \sum_{S} P_i \log \frac{1}{P_i}$$

o

$$H(S) \leq H(\bar{S}) \tag{2-29}$$

La condición de igualdad expresada en (2-27) es simplemente que s_i y s_j sean estadísticamente independientes, es decir, que S sea realmente una fuente de memoria nula. Ya se consideró un ejemplo que permitía comprobar la relación (2-29). Recordemos que, en la fuente de Markov de la figura 2-4, $H(S) = 0{,}81$ bits, mientras $H(\bar{S}) = 1$ bit.

2-7. Extensiones de una fuente de Markov.

En el apartado 2-4 se definió la extensión de una fuente de memoria nula. De forma análoga puede definirse la extensión de una

fuente de Markov, sin más que considerar que un bloque de n símbolos de la fuente constituye un nuevo símbolo σ_i.

> *Definición.* Sea S una fuente de información de Markov de orden m, de alfabeto $(s_1, s_2, \ldots, s_q)$ y probabilidades condicionales $P(s_i/s_{j_1}, s_{j_2}, \ldots, s_{j_m})$. La extensión de orden n de S, S^n, es una fuente de Markov de orden μ, con q^n símbolos, $\sigma_1, \sigma_2, \ldots, \sigma_{q^n}$. Cada σ_i corresponde a una secuencia de n de los S_i símbolos, y las probabilidades condicionales de σ_i son $P(\sigma_i/\sigma_{j_1}\ \sigma_{j_2}\ \ldots,\ \sigma_{j\mu})$. Estas probabilidades, así como μ, se definen a continuación.

Con objeto de describir completamente el comportamiento estadístico de la extensión de orden n de una fuente de Markov de orden m, deberemos definir las probabilidades condicionales

$$P(\sigma_i/s_{j_1}, s_{j_2}, \ldots, s_{j_m}) \tag{2-30}$$

donde σ_i representa un símbolo de la extensión de orden n, una secuencia de n símbolos. La secuencia $(s_{j_1}, s_{j_2}, \ldots, s_{j_m})$ es equivalente a alguna secuencia de σ_i, digamos $\sigma_{j_1}\ \sigma_{j_2}\ \ldots,\ \sigma_{j\mu}$ donde $\mu = [m/n]$ el menor número entero igual o superior a m/n.

Las probabilidades condicionales de los símbolos S de (2-30), por lo tanto, pueden escribirse en la forma

$$P(\sigma_i/\sigma_{j_1}, \sigma_{j_2}\ \ldots,\sigma_{j\mu}) \tag{2-31}$$

Por ejemplo, la tercera extensión de una fuente de Markov de quinto orden con q símbolos sería una fuente de Markov de segundo orden con q^3 símbolos. Hay que destacar que tomando al menos m extensiones de una fuente de Markov de orden m puede siempre obtenerse una fuente de Markov de *primer orden*. Deduciremos finalmente la expresión de $P(\sigma_i/\sigma_{j_1}, \sigma_{j_2}, \ldots, \sigma_{j\mu})$ en función de las probabilidades condicionales de los símbolos de la fuente original S.

Sea $\sigma_i = (s_{i_1}, s_{i_2}, \ldots, s_{i_n})$. Entonces

$$\begin{aligned} P(\sigma_i/\sigma_{j_1}\ \sigma_{j_2}\ \ldots,\ \sigma_{j\mu}) &= P(s_{i_1}, s_{i_2}, \ldots, s_{i_n}/s_{j_1}, s_{j_2}, \ldots, s_{j_m}) \\ &= P(s_{i_1}/s_{i_1}, s_{i_2}, \ldots, s_{i_m})\, P(s_{i_2}/s_{i_2}, s_{i_3}, \ldots, s_{i_m}, s_{i_1}) \\ &\quad \ldots P(s_{i_n}/s_{i_{n-m}}, s_{i_{n-m+1}}, \ldots, s_{i_{n-1}}) \end{aligned} \tag{2-32}$$

En el último término del producto se ha supuesto que $n > m$. Si $n \leq m$, este término sería $P(s_{i_n}/s_{j_n}, s_{j_{n-1}}, \ldots, s_{i_{n-1}})$.

Se demostró que la entropía de la extensión n de una fuente de memoria nula era igual a n veces la entropía de la fuente original. Es sencillo demostrar la aplicación de esta propiedad a las fuentes de Markov. Lo haremos en el caso de una fuente de Markov de primer orden.

Consideremos una fuente de Markov de primer orden, con un alfabeto $\{S_1, S_2, \ldots, S_q\}$, probabilidades de transición $P(s_i/s_j)$ y una distribución estacionaria $P_1, P_2, \ldots, P_q$. Sea S^n su extensión de orden n, de símbolos σ_i, $i = 1, 2, \ldots, q^n$. S^n es una fuente de Markov de primer orden (según nuestra definición de μ).

$$H(S^n) = \sum_{S^n} \sum_{S^n} P(\sigma_j, \sigma_i) \log \frac{1}{P(\sigma_i/\sigma_j)} \qquad (2\text{-}33a)$$

El segundo miembro de (2-33*a*), desde el punto de vista de la fuente S^n, es un sumando doble, donde tanto i como j varían de 1 a q^n. Por otra parte, podemos considerar esas sumas desde el punto de vista de la extensión de segundo orden de la fuente original S. En este caso,

$$H(S^n) = \sum_{S^{2n}} P(\sigma_j, \sigma_i) \log \frac{1}{P(\sigma_i/\sigma_j)} \qquad (2\text{-}33b)$$

Escribiendo la ecuación (2-32) para $m = 1$, encontramos

$$\begin{aligned} P(\sigma_i/\sigma_j) &= P(s_{j_1}, s_{i_2}, \ldots, s_{i_n}/s_j) \\ &= P(s_{i_1}/s_j)\, P(s_{i_2}/s_{i_1}) \ldots P(s_{i_n}/s_{i_{n-1}}) \end{aligned} \qquad (2\text{-}34)$$

El segundo miembro de (2-33*b*) puede descomponerse en n sumandos:

$$H(S^n) = \sum_{S^{2n}} P(\sigma_j, \sigma_i) \log \frac{1}{P(s_{i_1}/s_j)} + \ldots + \sum_{S^{2n}} P(\sigma_j, \sigma_i) \log \frac{1}{P(s_{i_n}/s_{i_{n-1}})} \qquad (2\text{-}35)$$

Puede simplificarse cada uno de estos sumandos. Por ejemplo, calculando en primer lugar $2n-2$ de ellos

$$\sum_{S^{2n}} P(\sigma_j, \sigma_i) \log \frac{1}{P(s_i/s_j)} = \sum_{S^2} P(s_j, s_i) \log \frac{1}{P(s_{i_1}/s_j)}$$

$$= H(S) \tag{2-36}$$

Por lo tanto

$$H(S^n) = n H(S) \tag{2-37}$$

Pueden deducirse otras propiedades interesantes de la entropía de una fuente de Markov considerando $\overline{S^n}$, la fuente afín de S^n. Supongamos que $P(\sigma_1)$, $P(\sigma_2)$, ..., $P(\sigma_{qn})$, son las probabilidades de los símbolos de primer orden σ_i, símbolos de la extensión de orden n de la fuente de Markov de primer orden considerada anteriormente. Puesto que σ_i corresponde a la secuencia $(s_{i_1}, s_{i_2}, \ldots, s_{i_n})$ vemos que $P(\sigma_i)$ puede interpretarse como la probabilidad afín de orden n de s_{i_k}.

$$H(\overline{S^n}) = \sum_{S} P(\sigma_i) \log \frac{1}{P(\sigma_i)}$$

$$= \sum_{S^n} P(s_{i_1}, s_{i_2}, \ldots, s_{i_n}) \log \frac{1}{P(s_{i_1}, s_{i_2}, \ldots, s_{i_n})} \tag{2-38}$$

Sin embargo S es una fuente de Markov de *primer orden*, por lo que

$$P(s_{i_1}, s_{i_2}, \ldots, s_{i_n}) = P(s_{i_1}) P(s_{i_2}/s_{i_1}) \ldots (s_{i_n}/s_{i_{n-1}}) \tag{2-39}$$

Introduciendo la relación (2-39) en (2-38), resulta

$$H(\overline{S^n}) = \sum_{S} P(s_{i_1}, s_{i_2}, \ldots, s_{i_n}) \left[\log \frac{1}{P(s_{i_1})} + \log \frac{1}{P(s_{i_2}/s_{i_n})} + \ldots + \log \frac{1}{P(s_{i_n}/s_{i_{n-1}})} \right]$$

$$= H(\bar{S}) + (n-1) H(S) \tag{2-40}$$

o

$$H(\overline{S^n}) = n H(S) + [H(\bar{S}) - H(S)] \tag{2-41}$$

Nótese que el término que aparece entre corchetes en (2-41) es una constante de valor positivo e independiente de n. Si S fuera una fuente de Markov de orden m (en lugar de primer orden), la expresión (2-41) quedaría en la forma

$$H(\overline{S^n}) = n\,H(S) + \epsilon_m \tag{2-42}$$

donde ϵ_m es una constante positiva que (siempre que $n > m$) depende únicamente del comportamiento estadístico de S (problema 2-1).

Dividiendo ambos miembros de (2-42) por n, resulta

$$\frac{H(\overline{S^n})}{n} = H(S) + \frac{\epsilon_m}{n} \tag{2-43}$$

Anteriormente, en la expresión (2-29), se vio que

$$H(\overline{S^n}) \geqq H(S^n) = n\,H(S) \tag{2-44}$$

Sin embargo, la ecuación (2-43) demuestra que esta desigualdad es menos importante al crecer n. De forma más precisa, a partir de (2-43) puede escribirse que

$$\lim_{n\to\infty} \frac{H(\overline{S^n})}{n} = H(S) \tag{2-45}$$

En otras palabras, para valores de n grandes, las limitaciones de Markov sobre los símbolos de S^n son cada vez menos importantes.

Al llegar a este punto, procede hacer la siguiente indicación. La fuente afín de la extensión de orden n de S *no* coincide con la extensión de orden n de la fuente afín de S.

$$H(\overline{S^n}) \neq H(\bar{S}^n) \tag{2-46}$$

Efectivamente, puesto que $\bar{S}$ es una fuente de memoria nula,

$$H(\bar{S}^n) = n\,H(\bar{S}) \tag{2-47}$$

que puede compararse con la expresión (2-44).

Ejemplo 2-6. Resumiremos algunos de los resultados obtenidos en los ejemplos anteriores en el caso de la fuente de la figura 2-4.

$$H(S) = 0.81 \text{ bit}$$
$$H(\bar{S}) = 1.00 \text{ bit}$$

De (2-37),

$$H(S^2) = 2H(S) = 1.62 \text{ bits}$$

Puede calcularse

$$H(\overline{S^2}) = \sum_{S^2} P(s_j, s_k) \log \frac{1}{P(s_j, s_k)}$$

$$= 1.86 \text{ bits}$$

Un cálculo más largo y complicado permite deducir los valores siguientes:

$$H(\overline{S^3}) = 2.66 \text{ bits}$$

$$H(\overline{S^4}) = 3.47 \text{ bits}$$

Hay que destacar cómo la secuencia

$$H(\overline{S}) = 1.00 \text{ bit}$$

$$\frac{H(\overline{S^2})}{2} = 0.93 \text{ bit}$$

$$\frac{H(\overline{S^3})}{3} = 0.89 \text{ bit}$$

$$\frac{H(\overline{S^4})}{4} = 0.87 \text{ bit}$$

se aproxima al valor $H(S)$.

2-8. Estructura del lenguaje.

En los apartados anteriores de este capítulo se ha definido un modelo de fuente de información, deduciéndose algunas de sus propiedades más simples. Es de indudable interés investigar las analogías que tal modelo presenta en relación con el proceso físico de generación de información. Un caso particularmente importante de generación es la creación de un mensaje compuesto de palabras de la lengua inglesa. Demostraremos en este apartado cómo podremos aproximarnos a un mensaje de este tipo mediante una secuencia de fuentes de información cada vez más complicadas.

Limitémonos a un conjunto de 27 símbolos, las 26 letras del alfabeto inglés, más un espacio. La fuente más simple de este alfabeto

sería aquella de memoria nula, con todos los símbolos igualmente probables. La entropía de esta fuente sería

$$H(S) = \log 27$$
$$= 4{,}75 \text{ bits/símbolos} \qquad (2\text{-}48)$$

La figura (2-6) muestra una secuencia típica de símbolos emitidos por la fuente. Definiremos esta secuencia como aproximación cero al inglés.

ZEWRTZYNSADXESYJRQY_WGECIJJ_OBVKRBQPOZB
YMBUAWVLBTQCNIKFMP_KMVUUGBSAXHLHSIE_M

FIG. 2-6. Aproximación cero al inglés.

En esta secuencia no se advierte ninguna estructura característica, ni se puede indentificar como perteneciente a un lenguaje particular que tenga el mismo alfabeto. Haciendo intervenir las probabilidades reales de los símbolos, expresados en la tabla 2-2, puede conseguirse una aproximación más exacta del idioma inglés. La entropía de una fuente de memoria nula, cuyas probabilidades sean las de esa tabla, tiene el valor

$$H(S) = \sum_S P_i \log \frac{1}{P_i}$$
$$= 4{,}03 \text{ bits/símbolos} \qquad (2\text{-}49)$$

TABLA 2-2. PROBABILIDADES DE LOS SÍMBOLOS EN INGLÉS (REZA 1961)

Símbolos	*Probabilidad*	*Símbolos*	*Probabilidad*
Espacio	0.1859	N	0.0574
A	0.0642	O	0.0632
B	0.0127	P	0.0152
C	0.0218	Q	0.0008
D	0.0317	R	0.0484
E	0.1031	S	0.0514
F	0.0208	T	0.0796
G	0.0152	U	0.0228
H	0.0467	V	0.0083
I	0.0575	W	0.0175
J	0.0008	X	0.0013
K	0.0049	Y	0.0164
L	0.0321	Z	0.0005
M	0.0198		

La figura 2-7 representa una secuencia típica de símbolos emitidos por esta fuente.

AI_NGAE_ _ITF_NNR_ASAEV_OIE_BAINTHA_HYR
OO_POER_SETRYGAIETRWCO_ _EHDUARU_EU_C_F
T_NSREM_DIY_EESE_ _F_O_SRIS_R_ _UNNASHOR

FIG. 2-7. Primera aproximación al inglés.

Aun cuando no puede calificarse de buen inglés, esta secuencia presenta la estructura propia del lenguaje (compárese con la aproximación cero): Las «palabras» de esta aproximación son, en su mayor parte, de longitud apropiada, y la proporción entre vocales y consonantes más real. Puede mejorarse la solución que dio lugar a la primera aproximación, utilizando una fuente de *Markov* de primer orden, con símbolos de probabilidades condicionales bien elegidas. Estas probabilidades fueron definidas por Pratt (1942).

$$H(S) = \sum_{S^2} P(i, j) \log \frac{1}{P(i/j)}$$

$$= 3{,}32 \text{ bits/símbolos} \qquad (2\text{-}50)$$

Usando las probabilidades enunciadas por Pratt es posible generar una secuencia típica de símbolos a partir de una fuente de Markov de primer orden. Shannon, sin embargo, indicó un método mucho más ingenioso. Las probabilidades de un texto ordinario inglés son las que se desprenden directamente de él. Por lo tanto, podemos abrir un libro y seleccionar una letra al azar, por ejemplo la U. A continuación saltamos unas cuan as líneas, leyendo hasta encontrar la primera U, eligiendo la letra que la sigue (en este caso fue una R). Se repite de nuevo la operación, saltando varias líneas, leyendo hasta la primera R y eligiendo la letra siguiente. Con este procedimiento se construyó la segunda aproximación al inglés (figura 2-8).

URTESHETHING_AD_E_AT_FOULE_ITHALIORT_W
ACT_D_STE_MINTSAN_OLINS_TWID_OULY_TE_T
HIGHE_CO_YS_TH_HR_UPAVIDE_PAD_CTAVED

FIG. 2-8. Segunda aproximación al inglés.

La secuencia obtenida en la segunda aproximación ya deja trascender un regusto a inglés. Es más lógico identificarla como una aproximación al inglés que, digamos, al francés.

El método de Shannon puede aplicarse a la construcción de mejores aproximaciones al inglés. En efecto, pueden elegirse las letras precedentes, construyendo así una secuencia típica de una fuente de Markov, aproximación del inglés, de segundo orden.

Shannon (1951) estimó que la entropía de la fuente correspondiente a la figura 2-9 es de 3,1 bits por símbolo. Por otros procedimientos dedujo que la entropía del idioma inglés, teniendo en cuenta todo el texto anterior, está comprendida entre 0,6 y 1,3 bits por símbolo.

IANKS_CAN_OU_ANG_RLER_THATTED_OF_TO_S
HOR_OF_TO_HAVEMEM_A_I_MAND_AND_BUT_
WHISSITABLY_THERVEREER_EIGHTS_TAKILLIS_TA

FIG. 2-9. Tercera aproximación al inglés.

Puede ampliarse el procedimiento anterior, para generar secuencias típicas de probabilidades idénticas. Sin embargo, es prácticamente imposible para m mayor de 2. En su lugar, Shannon, utilizó una fuente de información de memoria nula que emite *palabras* inglesas en lugar de letras. Las probabilidades de ocurrencia de las diferentes palabras son aproximadamente las mismas que en un texto inglés. Shannon (1948) obtuvo la aproximación mostrada en la figura 2-10.

REPRESENTING AND SPEEDILY IS AN GOOD APT
OR COME CAN DIFFERENT NATURAL HERE HE
THE A IN CAME THE TO OF TO EXPERT
GRAY COME TO FURNISHES THE LINE MES-
SAGE HAD BE THESE

FIG. 2-10. Cuarta aproximación al inglés.

Aún puede llegarse a una aproximación más compleja haciendo depender de la palabra precedente la probabilidad de que una palabra sea elegida. La fuente correspondiente sería una fuente de Markov de primer orden, con palabras inglesas como símbolos. Shannon (1948)

construyó una secuencia típica a partir de una fuente de este tipo (figura 2-11).

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

FIG. 2-11. Quinta aproximación al inglés.

Es interesante destacar cómo esta secuencia se aproxima al discurso incoherente emitido por un interlocutor que estuviera muy excitado. Resulta un estímulo comprobar cómo se puede simular (al menos en una cierta medida) una fuente de información tan compleja como un individuo hablando inglés, mediante unos sencillos modelos consistentes en fuentes de Markov de memoria nula. Muchas de las fuentes de información tratadas en relación con los problemas reales planteados por la comunicación tienen una naturaleza más simple, por lo que podemos imaginar que en esos casos nuestros modelos constituirán aún una aproximación más cercana a la realidad.

Pueden estudiarse las diferencias entre varios idiomas occidentales construyendo distintas secuencias basadas en sus estadísticas. Las figuras 2-12 a 2-14 muestran los resultados obtenidos en tres idiomas

R_EPTTFVSIEOISETE_TTLGNSSSNLN_UNST_FSNST
F_E_IONIOILECMPADINMEC_TCEREPTTFLLUMGLR
ADBIUVDCMSFUAISRPMLGAVEAI_MILLUO

a) Primera aproximación al francés

ITEPONT_JENE_IESEMANT_PAVEZ_L_BO_S_PAS
E_LQU_SUIN_DOTI_CIS_NC_MOUROUNENT_FUI
T_JE_DABREZ_DAUIETOUNT_LAGAUVRSOUT_MY

b) Segunda aproximación al francés

JOU_MOUPLAS_DE_MONNERNAISSAINS_DEME_U
S_VREH_BRE_TU_DE_TOUCHEUR_DIMMERE_LL
ES_MAR_ELAME_RE_A_VER_IL_DOUVENTS_SO

c) Tercera aproximación al francés.

FIG. 2-12. Serie de aproximaciones al francés.

diferentes. Como antes, la primera aproximación corresponde a una secuencia emitida por una fuente de memoria nula; la segunda de una fuente de Markov de primer orden; y la tercera de una fuente de Markov de segundo orden.

NNBNNDOETTNIIIAD_TSI_ISLEENS_LRI_LDRRBNF
REMTDEEIKE_U_HBF_EVSN_BRGANWN_IENEEHM
EN_RHN_LHD_SRG_EITAW_EESRNNCLGR

a) Primera aproximación al alemán.

AFERORERGERAUSCHTER_DEHABAR_ADENDERG
E_E_UBRNDANAGR_ETU_ZUBERKLIN_DIMASO
N_DEU_UNGER_EIEIEMMLILCHER_WELT_WIERK

b) Segunda aproximación al alemán.

BET_EREINER_SOMMEIT_SINACH_GAN_TURHATT
ER_AUM_WIE_BEST_ALLIENDER_TAUSSICHELLE
_LAUFURCHT_ER_BLEINDESEIT_UBER_KONN_

c) Tercera aproximación al alemán.

FIG. 2-13. Serie de aproximaciones al alemán.

UOALNAO_NEL_D_NIS_ETR_TEGATUEOEC_S_ASU
DU_ZELNNTSSCASOSED_T_I_R_EIS_TAMMO_TII
UOEDEO_UEI_EOSEELA_NMSLAANTEC

a) Primera aproximación al español.

CINDEUNECO_PE_CAL_PROS_E_LAS_LABITEJAS
TE_ONTOMECITRODRESIO_PAY_EN_SPUSEL_LA
_S_UTAJARETES_OLONDAMIVE_ESA_S_CLUS_

b) Segunda aproximación al español.

RAMA_DE_LLA_EL_GUIA_IMO_SUS_CONDIAS_S
U_E_UNCONDADADO_DEA_MARE_TO_BUERBALI
A_NUE_Y_HERARSIN_DE_SE_SUS_SUPAROCEDA

c) Tercera aproximación al español.

FIG. 2-14. Serie de aproximaciones al español.

Como ejemplo final, damos a continuación una serie de aproximaciones (figura 2-15) a otro idioma occidental, dejando al lector que determine su identidad.

SETIOSTT_NINN_TUEEHHIUTIAUE_N_IREAISRI_M
INRNEMOSEPIN_MAIPSAC_SES_LN_ANEIISUNTINU
_AR_TM_UMOECNU_RIREIAL_AEFIITP

a) Primera aproximación a ?

CT_QU_VENINLUM_UA_QUREO_ABIT_SAT_FIUMA
GE_ICAM_MESTAM_M_QUM_CUTAT_PAM_NOND
QUM_O_M_FIT_NISERIST_E_L_ONO_IHOSEROCO

b) Segunda aproximación a ?

ET_LIGERCUM_SITECI_LIBEMUS_ACERELEN_TE
_VICAESCERUM_PE_NON_SUM_MINUS_UTERNE
_UT_IN_ARION_POPOMIN_SE_INQUENEQUE_IRA

c) Tercera aproximación a ?

FIG. 2-15. Serie de aproximación a ?

NOTAS

Nota 1. La palabra entropía fue creada por Clausius en 1876, a partir de las palabras griegas εν y τρεπειν. Esas palabras juntas tienen el sentido de «replegarse hacia el interior». Clausius empleó entropía para designar la parte de la energía de un sistema que no puede transformarse en trabajo mecánico sin transmitir calor a algún otro cuerpo, o modificar su volumen. Bolzmann, en 1896, fue el primero en demostrar que la entropía de un sistema podía expresarse en función del valor medio del logaritmo de las probabilidades de los estados del sistema. Shannon (1948) introdujo la palabra en la teoría de la información.

Entropía es, seguramente, el más importante pero no el único punto de contacto entre la teoría de la información y la mecánica estadística. Jaynes (1959) consideró el empleo de la *función de partición* de la mecánica estadística en la teoría de la información.

Nota 2. A lo largo del texto supondremos que todas las fuentes tienen símbolos de probabilidades conocidas. Cuando son desconocidas (e incluso quizá cuando lo es el número de símbolos), podría evaluarse el valor de la entropía

de una fuente de memoria nula observando un número K de salidas. Miller y Madow (1954) calcularon el valor más probable de la entropía basándose en tal observación.

Basharin (1959) utilizó las probabilidades experimentales $\hat{p}_i$ para llegar al valor natural

$$\hat{H}(S) = \sum_{S} \hat{p}_i \log \frac{1}{\hat{p}_i}$$

Demostró que $\hat{H}(S)$ constituye una estimación consecuente, asintóticamente normal de $H(S)$, con

$$E[\hat{H}(S)] = H(S) - \frac{q-1}{2N} \log e + 0\left(\frac{1}{N^2}\right)$$

donde q es el número de símbolos de la fuente y N la dimensión de la muestra observada. Blyth (1958) definió otras estimaciones, demostrando además que no existe ninguna estimación de $H(S)$ no consecuente.

Nota 3. En el apartado 2-2 se definió la entropía de una variable al azar, variable que puede tomar uno entre un número finito de valores. (La función de distribución de una variable al azar está formada por un número finito de escalones.)

Con objeto de estudiar una variable al azar, s, que pueda tomar valores de forma continua (es decir, de función de distribución continua), puede considerarse una secuencia de funciones de distribución de escalón finito que se aproximan a la función de distribución continua. Supongamos que $[s]$ indica la parte entera de s. Puede definirse la nueva variable al azar.

$$s_n = \frac{1}{n}[ns]$$

y suponer

$$P_{nk} = \Pr\left\{ s_n = \frac{k}{n} \right\}$$

Sea S_n la fuente correspondiente a la variable al azar s_n. Al crecer n, la variable al azar s_n se aproxima más y más a s. ¡Desgraciadamente, $H(S_n)$ no se mantiene finita!

Renyi (1959) definió la dimensión de una variable al azar como

$$d(s) = \lim_{n \to \infty} \frac{H(S_n)}{\log n}$$

y la *entropía d-dimensional* de s como

$$H_d(S) = \lim_{n \to \infty} [H(S_n) - d \log n]$$

cuando estas cantidades existen. En el caso en que la función de distribución conste de un número finito de escalones, la dimensión es evidentemente nula y $H_0(S)$ se reduce a $H(S)$. Renyi demostró que cuando la función de distribución es continua y $H(S_1)$ finita, $d(s)$ es igual a la unidad, y suponiendo que $p(s)$ es la función densidad

$$H_1(S) = \int_{-\infty}^{\infty} p(s) \log \frac{1}{p(s)}\, ds$$

si la integral existe. Csiszar (1961) aportó una serie de transformaciones a estas conclusiones.

Nota 4. A partir de una fuente artificial, además de generar palabras pertenecientes a un idioma cualquiera, como se vió en el apartado 2-8, es posible la creación de composiciones musicales. Pinkerton (1936) utilizó este procedimiento. Pierce (1961) dedicó varias páginas a la generación de tal tipo de música; tal vez la última palabra sobre la teoría de la información aplicada al arte esté puesta de manifiesto en algunos pasajes de la «Suite Illiac» para cuarteto de cuerda, reproducida por Pierce (1957, p. 260).

Nota 5. El empleo de la entropía de una fuente como medida de la cantidad de información que suministra, ya se dijo anteriormente, se demuestra por el primer teorema de Shannon (capítulo 4). Es posible también justificarlo (Feinstein, 1958) basándose en que la entropía es la única función de las probabilidades de los símbolos de una fuente que cumple ciertos requisitos. Definamos tres fuentes de memoria nula y las probabilidades de sus símbolos en la forma siguiente: $\bar{\alpha} = 1 - \alpha$):

S		S_1		S_2	
s_1	P_1	s_1	P_1	s_1	α
s_2	P_2	s_2	P_2	s_1	$\bar{\alpha}$
.	.	.	.		
.	.	.	.		
.	.	.	.		
s_{q-1}	P_{q-1}	s_{q-1}	P_{q-1}		
s_q	P_q	s_q	αP_q		
		s_{q+1}	$\bar{\alpha} P_q$		

La entropía es la única función (excepto en el caso de una constante multiplicativa) de las probabilidades de los símbolos de una fuente que satisface:

a) $H(S_1) = H(S) + P_q H(S_2)$.

b) $H(S_2)$ es una función continua de α.

c) $H(S)$ es una función simétrica de $P_1, P_2, \ldots, P_q$.

PROBLEMAS

2-1. Demostrar la ecuación (2-42).

2-2. El diagrama de estados de una fuente de Markov de primer orden, con un alfabeto $S = \{0, 1, 2\}$, está representado en la figura P 2-2. Por definición $\bar{p} = 1 - p$. Por simetría, vemos que la distribución estacionaria es $P(0) = P(1) = P(2) = 1/3$.

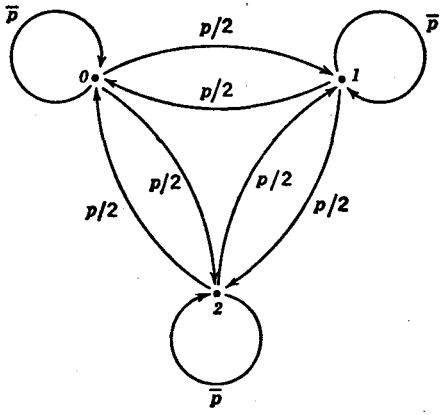


FIG. P 2-2.

a) Calcular $H(\bar{S})$.

b) Calcular $H(S)$. ¿Es correcta la respuesta para $p = 0$, $p = 1$?

**c*) ¿Cuál es el valor de p correspondiente al máximo de $H(S)$?

d) Analizar el comportamiento de $H(S)$ para $p = \varepsilon$, siendo $\varepsilon \approx 0$.

e) Analizar el comportamiento de $H(S)$ para $p = 1 - \delta$, siendo $\delta \approx 0$.

f) Dibujar el diagrama de estados correspondiente a S^2.

g) Calcular $H(S^2)$ y $H(\bar{S}^2)$.

2-3. Dos fuentes de memoria nula, S_1 y S_2, tienen q_1 y q_2 símbolos, respectivamente. Los símbolos de S_1 se representan con probabilidades P_i, $i = 1, 2, \ldots, q_1$; los de S_2 con Q_i, $i = 1, 2, \ldots, q_2$; las entropías de ambas son H_1 y H_2, respectivamente. Una nueva fuente de memoria nula $S(\lambda)$, denominada compuesta de S_1 y S_2, está formada con $q_1 + q_2$ símbolos. Los q_1 primeros símbolos de $S(\lambda)$ tienen probabilidades λP_i, $i = 1, 2, \ldots, q_1$, y los últimos q_2 probabilidades $\bar{\lambda} Q_i$, $i = 1, 2, \ldots, q_2$. ($\bar{\lambda} = 1 - \lambda$).

a) Demostrar que

$$H[S(\lambda)] = \lambda H_1 + \bar{\lambda} H_2 + H(\lambda)$$

dando una interpretación a esta igualdad.

* *b*) Expresar λ_0, valor de λ que hace máximo a $H[S(\lambda)]$, en función de H_1 y H_2. Calcular $H[S(\lambda_0)]$.

2-4. Generalizar la parte *a*) del problema 2-3, al caso de n fuentes de memoria nula, S_1, S_2, ..., S_n.

2-5. Hacer uso de las siguientes identidades (para $0 \leq \alpha < 1$)

$$\sum_1^\infty \alpha^n = \frac{\alpha}{1-\alpha} \qquad \text{and} \qquad \sum_1^\infty n\alpha^n = \frac{\alpha}{(1-\alpha)^2}$$

en el problema.

a) Una fuente de información de memoria nula posee un alfabeto enumerable infinito $S = \{s_1, s_2, \ldots\}$ con $P_i = a\alpha^i$, para cualquier valor de i. Expresar a en función de α.

b) Calcular y dibujar $H(S)$ en función de α. Estudiar en particular el comportamiento de $H(S)$ para $\alpha \approx 0$ y $\alpha \approx 1$.

2-6. El diagrama de estados de una fuente de información de Markov binaria de primer orden viene dado en la figura P 2-6. Demostrar que las probabilidades estacionarias de la fuente son $P(0) = q/(p+q)$, $P(1) = p/(p+q)$.

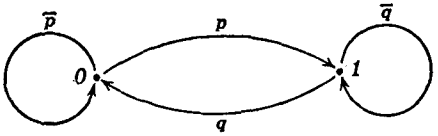


FIG. P 2-6.

a) Calcular $H(S)$.

b) Calcular $H(\bar{S})$.

c) Sea $p = q$. Calcular y dibujar $H(S)$ en función de p.

d) Calcular $H(\bar{S})$ cuando $p = q$.

2-7. *a*) Supongamos $q = 1$ en la fuente del problema 2-6 (cuando p no es igual a q). Calcular y dibujar $H(S)$ en función de p.

b) Con $q = 1$, calcular $H(S/0)$ y $H(S/1)$, información por símbolo cuando la fuente está en los estados 0 y 1, respectivamente.

2-8. *a*) Consideremos una fuente de Markov binaria de tercer orden en que la probabilidad de emitir un 0 ó un 1 no depende de los dos símbolos anteriores, sino del tercero. La probabilidad de que un símbolo coincida con el emitido tres lugares antes es igual a 0.9; la probabilidad de que sea distinto, 0.1. Dibujar el diagrama de estados de esta fuente.

b) Calcular la entropía de la fuente. (El método consistente en calcular las probabilidades estacionarias, etc., no es el más apropiado en este caso.)

2-9. Sea S_0 la extensión de tercer orden de una fuente binaria de memoria nula, cuya probabilidad de emitir un 0 es igual a p. Otra fuente, S, observa las salidas de S_0, emitiendo un 0, 1, 2 ó 3 según que la salida de S_0 contenga 0, 1, 2 ó 3 ceros.

a) Calcular $H(S_0)$.

b) Calcular $H(S)$.

c) Calcular $H(S_0) - H(S)$. Interpretar el significado de esta diferencia de entropías.

2-10. Generalizar la parte *c*) del problema 2-9 al caso en que S_0 es la extensión de orden n de una fuente binaria y S emite un 0, 1, 2, ..., ó n. SUGERENCIA: ¿Cuál es el valor medio de la información que se pierde al recibir un símbolo de S en lugar de S_0?

2-11. Consideremos una fuente de información binaria de memoria nula, S_0, cuya probabilidad de emitir un 0 sea igual a $p \approx 1$. Debido a la elevada probabilidad de presentarse una serie de ceros, puede admitirse que se emiten simplemente las longitudes de las series de ceros. Es decir, se considera una nueva fuente S, con símbolos s_1, s_2, s_3, ..., donde, por ejemplo, la secuencia s_3, s_2, s_4, s_1 y s_8 correspondería a la secuencia binaria.

$$\underbrace{001}_{s_3}\ \underbrace{01}_{s_2}\ \underbrace{0001}_{s_4}\ \underbrace{1}_{s_1}\ \underbrace{00000001}_{s_8}\ \ldots$$

a) Calcular la entropía de S, haciendo uso de la igualdad

$$H(S) = \sum_{S} P(s_i) \log \frac{1}{P(s_i)}$$

No expresar el resultado en forma de serie infinita.

b) Calcular $H(S)/H(S_0)$.

c) Calcular el número medio de binits de la fuente original dados por un símbolo de S.

2-12. La fuente S del problema 2-11 tiene un número infinito de mensajes posibles s_i. Consideremos la fuente S_n, aproximación de S con los $n + 1$ símbolos siguientes:

s_1	1
s_2	01
s_3	001
.	.
.	.
.	.
s_n	0000 · · · 01
s_{n+1}	0000 · · · 00

$\underbrace{\qquad\qquad}_{n \text{ binits}}$

a) Calcular $H(S_n)$.

b) Dibujar $H(S_n)$ en función de n para $p = 0.9$.

2-13. La figura P 2-13 representa el diagrama de estados de una fuente de Markov de primer orden de alfabeto $S = \{0,1,2\}$. Por simetría, la distribución estacionaria es

$$P(0) = P(1) = P(2) = 1/3$$

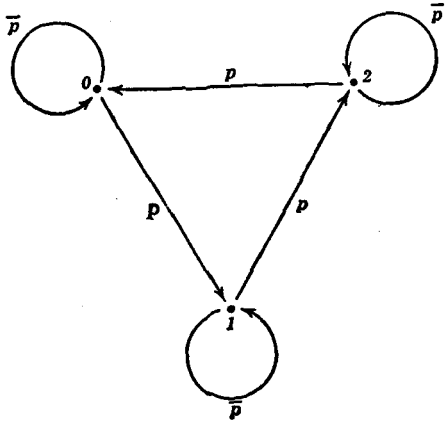


FIG. P 2-13.

a) Calcular $H(\bar{S})$.

b) Calcular $H(S)$. Comprobar el resultado para $p = 0$ y $p = 1$.

c) Calcular $H(S^2)$.

2-14. Sea S una fuente de memoria nula, de alfabeto $S = \{s_i\}$, $i = 1, 2, \ldots, q$, cuyos símbolos tienen probabilidades $P_1, P_2, \ldots, P_q$. Crear una nueva fuente de memoria nula, S', de doble número de símbolos, $S' = \{s'_i\}$, $i = 1, 2, \ldots, 2q$, con símbolos de probabilidades definidas por

$$\begin{aligned} P'_i &= (1 - \varepsilon)P_i \qquad i = 1, 2, \ldots, q \\ P'_i &= \varepsilon P_{i-q} \qquad i = q + 1, q + 2, \ldots, 2q \end{aligned}$$

Expresar $H(S')$ en función de $H(S)$.

CAPITULO 3

PROPIEDADES DE LOS CODIGOS

3-1. Introducción.

Con objeto de estudiar la relación existente entre codificación y la medida de información explicada en el capítulo 2, creemos necesario definir ciertas subclases de códigos. Hemos introducido ya la idea general de *código* (apartado 1-3) y las nociones de *alfabeto código y alfabeto fuente.*

> *Definición.* Denominemos $S = \{s_1, s_2, \ldots, s_q\}$ al conjunto de símbolos de un alfabeto dado. Se define un código como la correspondencia de todas las secuencias posibles de símbolos de S a secuencias de símbolos de algún otro alfabeto $X = \{x_1, x_2, \ldots, x_r\}$. S recibe el nombre de *alfabeto fuente* y X *alfabeto código.*

Esta definición de código es demasiado general para presentar interés al tratar de síntesis de códigos. Por lo tanto, limitaremos nuestra atención a aquellos códigos que poseen ciertas propiedades suplementarias. La primera propiedad exigida es que el código constituya un *código bloque.*

> *Definición.* Un *código bloque* es aquel que asigna cada uno de los símbolos del alfabeto fuente S a una secuencia fija de símbolos del alfabeto código X. Esas secuencias fijas (secuencias de x_j) reciben el nombre de *palabras código.* Denominaremos X_i a la palabra código que corresponde al símbolo s_i. Hay que notar que X_i constituye una *secuencia* de $x'_j s$ *.

* Algunos autores [p. ej., Peterson (1961)] definen los códigos bloque como aquéllos en que todas las palabras contienen un mismo número de símbolos.

Ejemplo 3-1. La tabla 3-1 da un ejemplo de código bloque binario.

TABLA 3-1. CÓDIGO BLOQUE BINARIO

Símbolos de la fuente	*Código*
s_1	0
s_2	11
s_3	00
s_4	11

A primera vista el requisito de codificar uno por uno los símbolos de la fuente en secuencias fijas de símbolos código resulta demasiado riguroso. Hay que destacar, sin embargo, que si un código hace corresponder todas las secuencias de longitud n de símbolos de la fuente con secuencias fijas de símbolos código, el código hace también corresponder cada símbolo de la extensión de orden n de la fuente original con una secuencia fija de símbolos código, constituyendo realmente un código bloque del alfabeto fuente S^n. Un conjunto de reglas que determinen la transformación de un alfabeto fuente en un alfabeto código puede cumplir la definición de código bloque solamente al tener en cuenta los símbolos de la extensión de orden n de la fuente. En gran parte de las discusiones que siguen trataremos precisamente de este tipo de códigos bloque.

3-2. Códigos unívocamente decodificables.

Es evidente, según se desprende del ejemplo anterior, que si se desea utilizar los códigos bloque han de imponerse ciertas restricciones; una restricción natural es que todas las palabras código X_i sean distintas. Nótese que las X_2 y X_4 del código dado en la tabla 3-1 *no* lo eran.

> *Definición.* Un código bloque se denomina *no singular* si todas sus palabras son distintas.

Ejemplo 3-2. La tabla 3-2 muestra un ejemplo de código bloque no singular.

TABLA 3-2. CÓDIGO BLOQUE NO SINGULAR

Símbolos de la fuente	*Código*
s_1	0
s_2	11
s_3	00
s_4	01

Aun cuando todas las palabras del código del ejemplo anterior son diferentes, es posible encontrar algún caso en que una secuencia dada puede tener un origen indefinido. Por ejemplo, la secuencia 0011 puede corresponder a $s_3\, s_2$ o $s_1\, s_1\, s_2$. Es decir, el código de la tabla 3-2, aun cuando es *no singular* en su detalle, es *singular* considerado de forma más general. Este ejemplo nos dice que, para definir códigos utilizables, debemos enunciar una condición más restrictiva que la no singularidad.

Supongamos un código bloque que hace corresponder los símbolos de un alfabeto fuente S con secuencias fijas de símbolos de un alfabeto código X. (La fuente S puede ser una extensión de otra fuente). Puesto que nos limitamos a considerar códigos bloque, tendremos una unidad natural y elemental de código; es decir, el símbolo de S y una palabra, compuesta de letras, del alfabeto código. Podemos colocar juntos estos bloques elementales, de la misma forma que hacíamos con los símbolos de una fuente, para constituir una extensión.

> *Definición.* La extensión de orden n de un código bloque que hace corresponder los símbolos s_i con las palabras código X_i, es el código bloque que hace corresponder las secuencias de símbolos de la fuente $s_{i_1}, s_{i_2}, \ldots, s_{i_n}$) con las secuencias de las palabras código $(X_{i_1}, X_{i_2}, \ldots, X_{i_n})$.

Según esta definición, la extensión de orden n de un código bloque es también un código bloque.

Ejemplo 3-3. La tabla 3-3 representa la extensión de segundo orden del código bloque de la tabla 3-2.

TABLA 3-3. SEGUNDA EXTENSIÓN DE UN CÓDIGO BLOQUE

Símbolos de la fuente	*Código*	*Símbolos de la fuente*	*Código*
s_1s_1	00	s_3s_1	000
s_1s_2	011	s_3s_2	0011
s_1s_3	000	s_3s_3	0000
s_1s_4	001	s_3s_4	0001
s_2s_1	110	s_4s_1	010
s_2s_2	1111	s_4s_2	0111
s_2s_3	1100	s_4s_3	0100
s_2s_4	1101	s_4s_4	0101

Definición. Un código bloque se dice *unívocamente decodificable* si, y solamente si, su extensión de orden n es no singular para cualquier valor finito de n.

Esta definición asegura que dos secuencias cualquiera de símbolos de la fuente *de la misma longitud* dan lugar a secuencias de símbolos códigos distintas. Es evidente que también será necesario que dos secuencias cualesquiera de símbolos de la fuente, incluso de diferente longitud, correspondan a secuencias de símbolos código distintas. Esta propiedad se deduce fácilmente de la definición. Admitamos, por ejemplo, lo contrario. Es decir, que existen dos secuencias, S_1 y S_2, de símbolos de la fuente que dan lugar a una misma secuencia de símbolos código, X_0. Hay que destacar que S_1, S_2 y X_0 representan *secuencias* de símbolos y no símbolos aislados. S_1 y S_2, además, pueden ser secuencias de diferente longitud. Formemos ahora dos nuevas secuencias de símbolos de la fuente, S_1' y S_2'. S_1' se define como la secuencia formada por S_2 seguida de S_1. S_2' es la secuencia formada por S_1 seguida de S_2. Vemos inmediatamente que tanto S_1' como S_2' dan lugar a una secuencia de símbolos código que es simplemente X_0. S_1' y S_2' tienen además la misma longitud. Por lo tanto, el código no satisface la condición de decodificación unívoca enunciada anteriormente.

Sardinas y Patterson (1953) encontraron las condiciones necesarias y suficientes que hacían un código unívocamente decodificable. Puesto que no tenemos interés más que en una subclase de este tipo de códigos, no mencionaremos sus resultados.

3-3. Códigos instantáneos.

En la tabla 3-4 * aparecen dos ejemplos de códigos unívocamente decodificables.

TABLA 3-4. DOS CÓDIGOS UNÍVOCAMENTE DECODIFICABLES

Símbolos de la fuente	*Código* $\mathscr{A}$	*Código* $\mathscr{B}$
s_1	00	0
s_2	01	10
s_3	10	110
s_4	11	1110

El código $\mathscr{A}$ da ejemplo del procedimiento más sencillo de generar códigos unívocamente decodificables. Todas sus palabras tienen la misma longitud, y además, $\mathscr{A}$ es evidentemente no singular. Puede comprobarse que estas dos propiedades son suficientes para garantizar la decodificación unívoca.

El código $\mathscr{B}$ de la tabla 3-4 es unívocamente decodificable puesto que no es singular y, además, constituye lo que se llama un *código coma.* Esto es, en $\mathscr{B}$, el 0 actúa como una coma que separa una palabra de la siguiente. Al observar una secuencia de símbolos, puede interpretarse la coma como lugar donde termina una palabra y comienza la siguiente.

La capacidad de reconocer cuando una palabra código, inmersa en una secuencia finita de símbolos, llega a su final, podría considerarse como propia de la configuración de los dos códigos particulares considerados. En realidad esta propiedad está íntimamente asociada con el concepto de código unívocamente decodificable. Consideremos aún otro nuevo código de esta clase (Tabla 3-5).

TABLA 3-5. OTRO CÓDIGO UNÍVOCAMENTE DECODIFICABLE

Símbolos de la fuente	*Código* $\mathscr{C}$
s_1	0
s_2	01
s_3	011
s_4	0111

* En adelante designaremos los códigos con letra cursiva.

El código $\mathscr{C}$ difiere de $\mathscr{A}$ y $\mathscr{B}$ en un aspecto importante. Si recibimos una secuencia binaria compuesta de palabras del código $\mathscr{C}$, no seríamos capaces de decodificar la sentencia en sus palabras, *según las vamos recibiendo.* Al recibir 01, por ejemplo, no podremos asegurar que corresponde al símbolo s_2 en tanto no hayamos recibido el símbolo siguiente. Si éste es un 0, sabemos que 01 corresponde verdaderamente a s_2. Si, por el contrario, es un 1, tendremos que analizar un símbolo más antes de afirmar si se trata de s_3 (011) o s_4 (0111). Este retraso es inherente al proceso de decodificación si se utiliza el código $\mathscr{C}$, en cambio con los códigos $\mathscr{A}$ y $\mathscr{B}$ podemos decodificar las palabras según van llegando.

> *Definición.* Un código unívocamente decodificable se denomina *instantáneo* cuando es posible decodificar las palabras de una secuencia sin precisar el conocimiento de los símbolos que las suceden.

Los códigos $\mathscr{A}$ y $\mathscr{B}$ vistos, son códigos instantáneos. El código $\mathscr{C}$ constituye un ejemplo de código unívoco, *no* instantáneo. En estos tres casos ha resultado sencillo comprobar si lo eran o no. Es interesante, sin embargo, disponer de una regla general que permita decir cuándo un código es instantáneo; la enunciaremos a continuación.

> *Definición.* Sea $X_i = x_{i_1}\ x_{i_2}\ \ldots\ x_{i_m}$ una palabra de un código. Se denomina *prefijo* de esta palabra a la secuencia de símbolos $(x_{i_1}\ x_{i_2}\ \ldots\ x_{i_j})$, donde $j \leq m$.

Ejemplo 3-4. La palabra código 0111 tiene cuatro prefijos, 0111, 011, 01 y 0.

Puede enunciarse la regla siguiente:

> La condición necesaria y suficiente para que un código sea instantáneo es que ninguna palabra del código coincida con el prefijo de otra.

La condición *suficiente* se deduce inmediatamente de la propia definición de código instantáneo. Si ninguna palabra es prefijo de otra, podrá decodificarse directamente a su recepción cualquier secuencia de símbolos formada por palabras código. Para ello se observa una secuencia hasta reconocer una subsecuencia formada por una palabra

código completa. La subsecuencia debe ser precisamente la palabra código, puesto que hemos admitido que no puede ser el prefijo de otra palabra. De esta manera puede procederse a decodificar las palabras, una por una, sin pérdida de tiempo en la operación.

La condición *necesaria* de la regla se demuestra por reducción al absurdo. Supongamos que existe una palabra del código, por ejemplo X_i, que es al mismo tiempo prefijo de otra X_j. Si observamos una secuencia recibida y encontramos la subsecuencia X_i, podrá ser bien una palabra completa o solamente la primera parte de X_j. No podremos decir cuál de las dos alternativas es la verdadera en tanto no hayamos examinado más símbolos de la secuencia principal. Según eso, el código no sería instantáneo.

En este punto es interesante resumir las distintas clases de códigos tratadas en las páginas precedentes. La figura 3-1 muestra la ramificación seguida en el árbol de subclases de códigos para llegar finalmente a la subclase correspondiente a los códigos instantáneos.

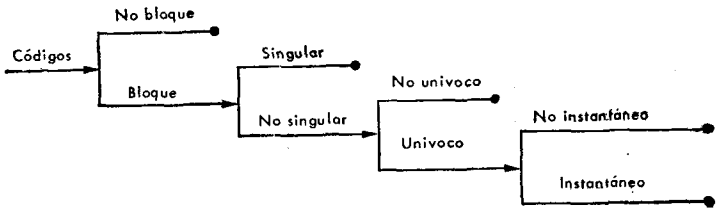


FIG. 3-1. Subclases de códigos.

3-4. Síntesis de un código instantáneo.

La naturaleza de los requisitos que debe cumplir un código instantáneo puede apreciarse más claramente analizando el procedimiento de síntesis de códigos en su forma más sencilla. Intentaremos sintetizar un código instantáneo binario a partir de una fuente de cinco símbolos. Comenzaremos asignando un 0 al símbolo s_1:

$$s_1 \rightarrow 0$$

Según esto, todos los demás símbolos de la fuente *deben* asociarse a palabras que comiencen por 1. De no ser así, se contradiría

la regla (3-1). El símbolo s_2 no debe asociarse a la palabra formada por el símbolo aislado 1; de hacerse, no quedaría ningún símbolo con el que pudieran comenzar las tres restantes palabras. Puede adoptarse

$$s_2 \rightarrow 10$$

lo que exige que los demás códigos comiencen por 11. Si

$$s_3 \rightarrow 110$$

el único prefijo de tres binits sin utilizar es 111, por lo que puede hacerse

$$s_4 \rightarrow 1110$$

y

$$s_5 \rightarrow 1111$$

Hay que destacar que por el hecho de asignar a s_1 el valor 0, automáticamente se limita el número de posibles palabras código. Dado este paso deberemos concretarnos a palabras que empiezan por 1. Puede imaginarse, por lo tanto, que si seleccionamos una palabra de 2 binits para representar a s_1, tendríamos un mayor grado de libertad en la elección de las siguientes, y llegaríamos incluso a evitar palabras tan largas como las dos últimas del código anterior. Probaremos esta sugerencia sintentizando un nuevo código binario instantáneo con los mismos cinco símbolos; comenzaremos por hacer

$$s_1 \rightarrow 00$$

Por lo que podremos elegir

$$s_2 \rightarrow 01$$

Quedan aun dos prefijos de longitud 2 sin utilizar, prefijos que podremos emplear en la siguiente forma:

$$s_3 \rightarrow 10$$

$$s_4 \rightarrow 110$$

$$s_5 \rightarrow 111$$

La pregunta de cuál de los dos códigos elaborados es mejor no puede contestarse simplemente con los criterios vistos hasta aquí. Este ejemplo se limita a demostrar que en la construcción de un código

instantáneo, cuanto más cortas son las primeras palabras, más largas tienen que ser las últimas. En el primer código, por elegir un 0, todas las demás palabras deberán ser secuencias que empiecen por 1. En el segundo código, la primera palabra es 00. En este caso podremos elegir todas las que empiezan por 1 y las que lo hacen por 01.

3-5. Inecuación de Kraft. Definición y discusión.

En el apartado 3-4 se discutieron cualitativamente algunas limitaciones del tamaño de las palabras de un código instantáneo, requisitos que pueden también expresarse en forma cuantitativa. El resto del capítulo tratará precisamente de estas limitaciones cuantitativas.

Consideremos un código instantáneo con un alfabeto fuente

$$S = \{s_1, s_2, \ldots, s_q\}$$

y un alfabeto código $X = \{x_1, x_2, \ldots, x_r\}$. Sean $X_1, X_2, \ldots, X_q$ las palabras del código y, por definición, l_i la longitud (es decir, el número de símbolos del código) de la palabra X_i. Normalmente es interesante que las longitudes de las palabras del código sean lo más cortas posible. La condición necesaria y suficiente para que exista un código instantáneo con palabras de longitud $l_1, l_2, \ldots, l_q$, viene definida por la *inecuación de Kraft* (Kraft, 1949).

> La condición necesaria y suficiente para la existencia de un código instantáneo de longitudes $l_1, l_2, \ldots, l_q$ es que
>
> $$\sum_{i=1}^{q} r^{-l_i} \leq 1$$
>
> donde r es el número de símbolos diferentes que constituyen el alfabeto código.

En el caso de alfabeto binario, la inecuación de Kraft se transforma en

$$\sum_{i=1}^{b} 2^{-l_i} \leq 1 \qquad (3\text{-}3)$$

donde la suma se extiende a todas las palabras del código bloque. Antes de probar esta inecuación, es interesante ver en qué forma puede

utilizarse para determinar si las l_i de una secuencia dada de l_i pueden constituir las longitudes de las palabras de un código instantáneo.

Tomemos una fuente de información con cuatro símbolos posibles, s_1, s_2, s_3 y s_4. En la tabla 5-6 se exponen los cinco códigos que pueden adoptarse para codificar estos símbolos en alfabeto binario.

TABLA 3-6. CINCO CÓDIGOS BINARIOS

Símbolos de la fuente	*Código* $\mathscr{A}$	*Código* $\mathscr{B}$	*Código* $\mathscr{C}$	*Código* $\mathscr{D}$	*Código* $\mathscr{E}$
s_1	00	0	0	0	0
s_2	01	100	10	100	10
s_3	10	110	110	110	110
s_4	11	111	111	11	11

Calcularemos el valor de $\sum_{i=1}^{4} 2^{-l_i}$ para cada uno de estos códigos.

Vemos, para el código $\mathscr{A}$, que

$$\sum_{i=1}^{4} 2^{-l_i} = 2^{-2} + 2^{-2} + 2^{-2} + 2^{-2}$$
$$= 1$$

Por lo tanto, las longitudes de las palabras de $\mathscr{A}$ son aceptables para un código instantáneo. Hay que resaltar, sin embargo, que la inecuación de Kraft *no* asegura que el código $\mathscr{A}$ sea un código instantáneo. La inecuación condiciona nuevamente las *longitudes* de las palabras y *no* las palabras mismas. En particular, en este ejemplo, la inecuación dice que puede existir un código binario instantáneo con cuatro palabras de longitud 2. En este caso está claro que, no sólo las longitudes del código $\mathscr{A}$ son aptas, sino también que las palabras mismas constituyen un código instantáneo.

Para el código $\mathscr{B}$

$$\sum_{i=1}^{4} 2^{-l_i} = 2^{-1} + 2^{-3} + 2^{-3} + 2^{-3}$$
$$= 7/8$$
$$\leq 1$$

Vemos nuevamente que las longitudes de sus palabras pueden constituir un código instantáneo. Analizándolas seguidamente, comprobamos que forman realmente un código instantáneo, por satisfacer la condición (3-1). El código $\mathscr{C}$ es idéntico al $\mathscr{B}$, excepto la segunda palabra de la que se ha suprimido un binit. Calculando

$$\sum_{i=1}^{4} 2^{-l_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3}$$
$$= 1$$

vemos que las longitudes de $\mathscr{C}$ satisfacen la inecuación de Kraft. Se confirma, además, que constituye un código instantáneo. El código $\mathscr{D}$ se deduce también del $\mathscr{B}$ suprimiendo un binit (esta vez de la cuarta palabra). Se comprueba que sus longitudes satisfacen la inecuación de Kraft. Como ya hemos dicho, esto no constituye condición suficiente para que el código $\mathscr{D}$ sea instantáneo, y, efectivamente, en este caso puede apreciarse que la cuarta palabra es un prefijo de la tercera. La condición (3-1) no se cumple, luego el código $\mathscr{D}$ *no* es instantáneo.

Finalmente, calculamos para el código $\mathscr{E}$ de la tabla 3-6 el valor de la suma

$$\sum_{i=1}^{4} 2^{-l_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-2}$$
$$= 1\,1/8$$

Este código no requiere más análisis. Las longitudes de sus palabras no satisfacen la inecuación de Kraft y, en consecuencia, no puede ser un código bloque instantáneo.

Consideremos un nuevo ejemplo, antes de proceder a la demostración de la inecuación de Kraft. Supongamos que deseamos codificar las salidas de una fuente decimal, $S = \{0, 1, 2, \ldots, 9\}$, en un código instantáneo binario. Admitamos además, que existe una razón que aconseja codificar lo símbolos 0 y 1 de la fuente decimal en palabras binarias relativamente cortas. Este requisito se presenta realmente en una fuente que emite muchos más 0s y 1s que 2s, 3s, etc. Si codificamos los 0s y 1s de la fuente en la forma siguiente *

$$\begin{aligned} 0 &\rightarrow 0 \\ 1 &\rightarrow 10 \\ 1 &\rightarrow 10 \end{aligned} \qquad (3\text{-}4)$$

* No puede hacerse $0 \rightarrow 0$ y $1 \rightarrow 1$, puesto que consumiremos todos los prefijos de 1 binit antes de codificar los ocho símbolos de la fuente, resultando por lo tanto imposible construir un código instantáneo.

podemos preguntarnos a continuación cuán cortas pueden hacerse las ocho palabras restantes. Si se exige que tengan la misma longitud,

TABLA 3-7. UN CÓDIGO BINARIO PARA DÍGITOS DECIMALES

Dígitos decimales	*Codigó binario*
0	0
1	10
2	11000
3	11001
4	11010
5	11011
6	11100
7	11101
8	11110
9	11111

digamos l, la inecuación de Kraft responde directamente a esta pregunta. Efectivamente, sabemos que debe cumplirse que

$$\sum_{i=0}^{9} 2^{-l_i} \leq 1 \tag{3-5}$$

Por hipótesis $l_0 = 1$, $l_1 = 2$ y $l_2 = l_3 = \ldots = l_9 = l$. Introduciendo estos valores en (3-5), encontramos

$$1/2 + 1/4 + 8\,(2^{-l}) \leq 1$$

o

$$l \geq 5 \tag{3-6}$$

Por lo tanto, no es posible encontrar un código instantáneo que cumpla nuestra condición con $l < 5$. La inecuación de Kraft dice que tal código existe para $l = 5$, pero no determina el procedimiento de síntesis a seguir. Sin embargo, no es difícil de encontrar. Corresponde precisamente al definido en la tabla 3-7.

3-6. Inecuación de Kraft. Demostración.

En el apartado anterior se introdujo la inecuación de Kraft, junto con algunos ejemplos que sirvieron para ayudar al lector a comprender la naturaleza de las limitaciones que impone. A continuación maneja-

remos estas limitaciones de manera que constituyan la misma demostración de la inecuación. Probaremos, en primer lugar, que la inecuación es condición suficiente para la existencia de un código instantáneo; lo haremos construyendo realmente un código instantáneo que satisfaga la expresión

$$\sum_{i=1}^{q} r^{-l_i} \leq 1 \tag{3-7}$$

Supongamos que deseamos formar un código instantáneo con palabras cuyas longitudes, $l_1, l_2, \ldots, l_q$ satisfacen la desigualdad (3-7). Estas longitudes pueden ser o no iguales. En principio es interesante considerar grupos de palabras de la misma longitud. Definamos, por lo tanto, n_1 como el número de palabras de longitud 1; n_2, las de longitud 2; etc. Si la más larga de las l_i es igual a l, tendremos

$$\sum_{i=1}^{l} n_i = q \tag{3-8}$$

Puede introducirse n_i en la expresión (3-7). La suma de (3-7) contiene n_i, términos de la forma r^{-1}, n_2 de la forma r^{-2}, etc. Por lo que puede escribirse como sigue

$$\sum_{i=1}^{l} n_i r^{-i} \leq 1 \tag{3-9}$$

o, multiplicando por r^l

$$\sum_{i=1}^{l} n_i r^{l-i} \leq r^l \tag{3-10}$$

Operando, obtendremos

$$n_l \leq r^l - n_1 r^{l-1} - n_2 r^{l-2} - \ldots - n_{l-1} r \tag{3-11a}$$

Dividiendo por r, se deduce una secuencia interesante de desigualdades

$$n_{l-1} \leq r^{l-1} - n_1 r^{l-2} - n_2 r^{l-3} - \ldots - n_{l-2} r \tag{3-11b}$$

.....

$$n_3 \leq r^3 - n_1 r^2 - n_2 r \tag{3-11c}$$

$$n_2 \leq r^2 - n_1 r \tag{3-11d}$$

$$n_1 \leq r \tag{3-11e}$$

Este conjunto de desigualdades constituye la clave de la construcción del código buscado. Hemos de formar n_1 palabras de longitud 1. Utilizando un alfabeto código de r símbolos, existirán r palabras posibles. Puesto que $n_1 \leq r$ pueden elegirse esos n_1 símbolos código arbitrariamente. Hagámoslo así; quedan entonces $r - n_1$ prefijos de longitud 1 permitidos, exactamente aquellos que no han sido elegidos al final de cada uno de esos prefijos; pueden formarse hasta

$$(r - n_1)\, r = r^2 - n_1 r \tag{3-12}$$

palabras de longitud 2. La ecuación (3-11*d*), sin embargo, asegura que el número de palabras de longitud 2 no debe exceder de esta cantidad. Como antes, seleccionamos arbitrariamente nuestras n_2 palabras de entre las r^2 ($n_1 r$ posibilidades); quedan entonces

$$(r^2 - n_1 r) - n_2$$

prefijos de longitud 2 sin utilizar, con los que pueden formarse

$$(r^2 - n_1 r - n_2)\, r = r^3 - n_1 r^2 - n_2 r \tag{3-13}$$

prefijos de longitud 3. La expresión (3-11*c*) asegura que este número es suficiente, y seleccionaremos a continuación las palabras de longitud 3 de entre ellos. Puede continuarse de esta manera, hasta formar todas las palabras del código. Las ecuaciones (3-11) aseguran, en cada etapa, que queda aún un número suficiente de prefijos.

Después de demostrar que la relación (3-7) [o su equivalente (3-9)] es suficiente para formar un código instantáneo de longitudes $l_1, l_2, \ldots, l_q$, es relativamente sencillo demostrar que la ecuación es también una condición necesaria. Habrá que invertir los argumentos empleados. En lugar de llevar adelante en todo detalle este proceso, llegaremos a una conclusión más definitiva.

3-7. Inecuación de McMillan

En el apartado anterior se demostró que

$$\sum_{i=1}^{q} r^{-l_i} \leq 1 \tag{3-14}$$

constituye una condición suficiente que deben cumplir las longitudes de las palabras de un código *instantáneo,* construyendo un código con

tales longitudes. Puesto que los códigos instantáneos son una subdivisión de los códigos unívocos, la condición *suficiente* se aplica también a ellos; es decir, si las longitudes $l_1, l_2, \ldots, l_q$ satisfacen la relación (3-14), puede construirse con ellas un código *unívoco.*

La demostración de la necesidad de la inecuación de Kraft, por el contrario, no puede extenderse a los códigos unívocos. Realmente la condición necesaria de la inecuación sugiere el análisis de los requisitos que deben cumplir las longitudes de las palabras de los códigos unívocos. Se sabe que (3-14) expresa una condición necesaria para los códigos instantáneos. ¿Es válida la misma condición para los códigos unívocamente decodificables, de carácter general?

El hecho de que la relación (3-14) sea condición necesaria para los códigos unívocos, así como para los códigos instantáneos, fue probado primeramente por McMillan (1956). Karush (1961) simplificó posteriormente la demostración. Consideremos la expresión

$$\left(\sum_{i=1}^{q} r^{-l} \right)^n = (r^{-l_1} + r^{-l_2} + \ldots + r^{-l_q})^n \tag{3-15}$$

Su desarrollo tendrá q^n términos, de la forma

$$r^{-l_{i_1}-l_{i_2}-\cdots-l_{i_n}} = r^{-k} \tag{3-16}$$

donde, por definición

$$l_{i_1} + l_{i_2} + \ldots + l_{i_n} = k \tag{3-17}$$

Como en el apartado anterior, sea l la mayor de las longitudes l_i. k puede entonces tomar un conjunto de valores comprendido entre n y nl. Definamos N_k como el número de términos de la forma r^{-k} existente en (3-15). Entonces

$$\left(\sum_{i=1}^{q} r^{-l} \right)^n = \sum_{k=n}^{nl} N_k r^{-k} \tag{3-18}$$

Ahora bien, teniendo en cuenta (3-17), vemos que N_k representa también el número de porciones de n palabras código que pueden formarse de modo que cada porción tenga una longitud de exactamente k símbolos. Si el código es unívocamente decodificable, N_k no debe

ser mayor de r^k, número de secuencias de orden r distintas de longitud k. Por tanto

$$\left(\sum_{i=1}^{q} r^{-l_i} \right)^n \leq \sum_{k=n}^{nl} r^k r^{-k}$$

$$\leq nl - n + 1$$

$$\leq nl \qquad (3\text{-}19)$$

La ecuación (3-19) es la prueba buscada, ya que si $x > 1$, $x^n > nl$, con tal de tomar un valor de n suficientemente grande. La expresión (3-19) se cumple para cualquier valor entero de n; de modo que tendremos

$$\sum_{i=1}^{q} r^{-l_i} \leq 1 \qquad (3\text{-}20)$$

3-8. Ejemplos.

Finalizaremos el capítulo tratando sobre las propiedades de los códigos en dos aplicaciones de la inecuación de Kraft y la construcción de un código instantáneo.

Supondremos primeramente que deseamos codificar una fuente de 10 símbolos en un código instantáneo trinario, de palabras de longitudes 1, 2, 2, 2, 2, 2, 3, 3, 3, 3. Aplicando la prueba de la inecuación de Kraft, obtenemos

$$\sum_{i=1}^{10} 3^{-l_i} = 1/3 + 5\,(1/9) + 4\,(1/27)$$

$$= 28/27 > 1$$

No es posible, por lo tanto, encontrar un código trinario instantáneo con palabras de esas longitudes.

En un segundo ejemplo, supongamos que deseamos codificar los símbolos de una fuente de nueve símbolos en un código instantáneo trinario con palabras de longitudes 1, 2, 2, 2, 2, 2, 2, 3, 3, 3. Esta vez, aplicando la prueba, encontramos

$$\sum_{i=1}^{9} 3^{-l_i} = 1/3 + 5\,(1/9) + 3\,(1/27)$$

$$= 1$$

En consecuencia, el código es aceptable. Se define como sigue:

$$s_1 \rightarrow 0$$

$$s_2 \rightarrow 10$$

$$s_3 \rightarrow 11$$

$$s_4 \rightarrow 12$$

$$s_5 \rightarrow 20$$

$$s_6 \rightarrow 21$$

$$s_7 \rightarrow 220$$

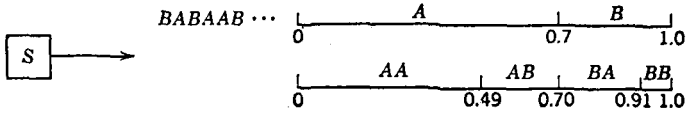
$$s_8 \rightarrow 221$$

$$s_9 \rightarrow 222$$

Hay que notar que la construcción del código anterior puede servir de ejemplo del método de codificación utilizado en la demostración de la inecuación de Kraft. Se elige un prefijo de longitud 1 (el 0), quedando obligados a adoptar uno de los dos restantes prefijos de longitud 1 para las demás palabras. Esto limita a dos veces tres, es decir seis, las palabras permitidas de longitud 2. Se emplean únicamente cinco de ellas, conservando la sexta (22) como prefijo de las tres últimas palabras.

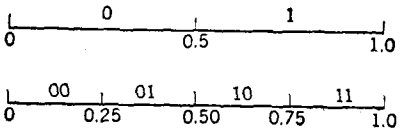
NOTAS

Nota 1. Un método de síntesis de códigos no bloque y unívocamente decodificables, en sentido familiar (no en el definido en el apartado 3-2), se debe a Elías. Consideremos, por ejemplo, una fuente de memoria nula de dos símbolos A y B, que se presentan con probabilidades 0,7 y 0,3, respectivamente.



Una secuencia arbitraria de longitud infinita, se representa mediante un punto del intervalo [0, 1], como se indica en la figura. (Las secuencias que comienzan por A se encuentran en el intervalo [0, 0.7]; las que comienzan por AB en el intervalo [0.49, 0.70]; etc.). Para codificar una secuencia de esta fuente, se pro-

cede simplemente a realizar la expansión binaria de cada uno de los puntos del intervalo [0, 1], en la forma siguiente:



Nótese que *no* es necesario recibir la secuencia binaria completa para poder comenzar la codificación. Por ejemplo, si la secuencia binaria empieza con 011..., sabemos que el punto representado debe estar situado entre 0.375 y 0.50; por lo tanto, el primer símbolo deberá ser una *A*. Si la secuencia comienza con 0110, el punto estará entre 0.375 y 0.4375; en consecuencia, los tres primeros símbolos serán *AAB*.

Esta idea es la base de una nueva demostración del primer teorema de Shannon (apartado 4-3, capítulo 4, nota 1) debida a Billingsley, quién representó las secuencias emitidas por una fuente de información por un punto del intervalo unidad e hizo uso de la teoría de la dimensión de Hausdorff para demostrar que la expansión natural de orden r define este punto de manera óptima.

Nota 2. Una de las *más interesantes aplicaciones* de las ideas presentadas en el capítulo 3 es la codificación genética (Golomb, 1961, 1962). Se ha comcomprobado que la *cantidad de información necesaria para especificar la estructura* de un sistema biológico está contenida en los cromosomas del sistema original. De forma más precisa, el ácido deoxiribonucléico (DNA) es quien transmite la información genética. En 1953 Crick y Watson demostraron que el DNA presenta el aspecto de una doble hélice. Estas hélices pueden imaginarse relacionadas por secuencias de cuatro nucleótidas que contienen el mensaje genético. Las nucleótidas, generalmente designadas A, C, G y T (adenina, citosina, guanina y timina) corresponden a los símbolos presentados en el capítulo 3. Experimentalmente, por otra parte, se ha puesto en evidencia que la naturaleza opera con un alfabeto de cuatro símbolos. Estos símbolos se combinan de diferente manera para representar alrededor de veinte aminoácidos que deben ser fabricados por el nuevo sistema biológico. La forma en que los nucleótidos (A, C, G, T) se codifican para representar los diferentes aminoácidos, constituye el problema fundamental de la codificación genética.

PROBLEMAS

3-1. Las palabras de un código instantáneo tienen longitudes $l_1, l_2, \ldots, l_q$, que satisfacen la inecuación

$$\sum_{i=1}^{q} r^{-l_i} < 1$$

Su alfabeto es $X = \{x_1, x_2, \ldots, x_r\}$. Demostrar que existen secuencias de símbolos $x_{i1}, x_{i2}, x_{i3}\ldots$ que no pueden ser decodificadas en secuencias de palabras.

3-2. Una fuente tiene seis salidas posibles, cuyas probabilidades se especifican en la tabla P 3-2. La tabla define también los códigos $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, $\mathscr{D}$, $\mathscr{E}$, $\mathscr{F}$.

a) ¿Cuál de los códigos es unívocamente decodificable?

b) ¿Cuál es instantáneo?

c) Calcular la longitud media de todos los códigos unívocos.

TABLA P 3-2

Salida	$P(s_i)$	$\mathscr{A}$	$\mathscr{B}$	$\mathscr{C}$	$\mathscr{D}$	$\mathscr{E}$	$\mathscr{F}$
s_1	1/2	000	0	0	0	0	0
s_2	1/4	001	01	10	10	10	100
s_3	1/16	010	011	110	110	1100	101
s_4	1/16	011	0111	1110	1110	1101	110
s_5	1/16	100	01111	11110	1011	1110	111
s_6	1/16	101	011111	111110	1101	1111	001

3-3. *a*) ¿Cuál de los conjuntos de longitudes de la tabla P 3-3 es válido para un código unívoco, cuando el alfabeto es $X = \{0, 1, 2\}$?

TABLA P 3-3

	Longitud de palabra				
	1	*2*	*3*	*4*	*5*
Número de palabras de longitud l_i por código:					
Código $\mathscr{B}$	2	2	2	3	1
Código $\mathscr{A}$	2	1	2	4	1
Código $\mathscr{C}$	1	4	6	0	0
Código $\mathscr{D}$	2	2	2	2	3

b) Construir un código instantáneo con cada uno de los conjuntos de longitudes válidos de la tabla.

3-4. Una fuente de memoria nula posee un alfabeto $S = \{a, b, c\}$ con probabilidades respectivas 0.6, 0.3 y 0.1. La fuente se codifica en un código binario, no bloque según el método descrito en la Nota 1. Encontrar todos los posibles binits con que pueden comenzar la palabra código correspondiente a *acbcaab*.

CAPITULO 4

CODIFICACION DE FUENTES DE INFORMACION

4-1. Longitud media de un código.

En el capítulo 3 se han estudiado los procedimientos para construir códigos instantáneos que asocien los símbolos de un alfabeto fuente con palabras formadas por símbolos de un alfabeto código cualquiera. Para un alfabeto fuente y un alfabeto código dados, es posible, sin embargo, elaborar más de un código instantáneo o unívocamente decodificable. Esta abundancia de códigos válidos obliga a adoptar un criterio que permita elegir uno de entre ellos. Un criterio natural de selección, aún cuando no el único posible, podría ser el de su longitud. Realmente se ha aplicado ya explícitamente en el ejemplo del capítulo 1, e implícitamente en varias de las últimas discusiones sobre codificación.

Desde el punto de vista de la mera economía de expresión y la consecuente economía en el equipo de comunicación, sin tener en cuenta otras consideraciones, es preferible un código formado por muchas palabras cortas a uno con palabras de gran longitud. Definiremos, por tanto, la longitud media de un código.

> *Definición.* Sea un código bloque que asocia los símbolos de una fuente $s_1, s_2, \ldots, s_q$ con las palabras $X_1, X_2, \ldots, X_q$. Supongamos que las probabilidades de los símbolos de la fuente son $P_1, P_2, \ldots, P_q$ y las longitudes de las palabras $l_1, l_2, \ldots, l_q$. Definiremos la *longitud media* del código, L, por la ecuación

$$L = \sum_{i=1}^{q} P_i l_i \qquad (4\text{-}1)$$

Será interesante encontrar códigos unívocos de longitud media mínima.

> *Definición.* Consideremos un código unívoco que asocia los símbolos de una fuente S con palabras formadas por símbolos de un alfabeto r-ario. Este código será *compacto* (respecto a S) si su longitud media es igual o menor que la longitud media de todos los códigos unívocos que pueden aplicarse a la misma fuente y el mismo alfabeto.

Una vez enunciadas estas dos definiciones, puede formularse el problema fundamental de la codificación de fuentes de información como aquel de la búsqueda de códigos compactos. Hay que destacar que ambas definiciones se refieren exclusivamente a las *longitudes* de las palabras de los códigos, y no a las palabras mismas. Por esta razón, puede reducirse la búsqueda a los códigos pertenecientes a la clase de códigos instantáneos (apartado 3-7). La inecuación de McMillan garantiza que cualquier conjunto de longitudes de palabras válido para un código unívoco, lo es también para un código instantáneo. Como primer paso se procederá a calcular el valor mínimo posible de L con un código instantáneo.

La definición de L es válida tanto para las fuentes de memoria nula como de Markov. Con objeto de simplificar la discusión, limitaremos por el momento las consideraciones a las fuentes de memoria nula. Más adelante, en el apartado 4-4, se suprimirá esta restricción.

Consideremos una fuente de memoria nula, cuyos símbolos, s_1, s_2, ..., s_q tienen respectivamente las probabilidades P_1, P_2, ..., P_q. Supongamos un código bloque que codifica estos símbolos en un alfabeto de r símbolos, y definimos por l_i la longitud de la palabra correspondiente a s_i. La entropía de esta fuente de memoria nula será, entonces

$$H(S) = -\sum_{i=1}^{q} P_i \log P_i \tag{4-2}$$

Sean Q_1, Q_2, ..., Q_q números tales que $Q_i \geq 0$ para cualquier valor de i, y $\sum_{i=1}^{q} Q_i = 1$. Debido a (2-8), sabemos que

$$\sum_{i=1}^{q} P_i \log \frac{1}{P_i} \leq \sum_{i=1}^{q} P_i \log \frac{1}{Q_i} \tag{4-3}$$

igualdad solamente cuando $P_i = Q_i$, para todo valor de i. Por lo tanto

$$H(S) \leq -\sum_{i=1}^{q} P_i \log Q_i \tag{4-4}$$

con signo igual en el mismo caso.

La ecuación (4-4) será válida para cualquier conjunto de números positivos, Q_i, cuya suma sea la unidad. En consecuencia, se podrá elegir

$$Q_i = \frac{r^{-l_i}}{\sum_{i=1}^{q} r^{-l_i}} \tag{4-5}$$

de donde

$$\begin{aligned} H(S) &\leq -\sum_{i=1}^{q} P_i (\log r^{-l_i}) + \sum_{i=1}^{q} P_i \left(\log \sum_{j=1}^{q} r^{-l_j} \right) \\ &\leq \log r \sum_{i=1}^{q} P_i l_i + \log \left(\sum_{j=1}^{q} r^{-l_j} \right) \\ &\leq L \log r + \log \sum_{j=1}^{q} r^{-l_j} \end{aligned} \tag{4-6}$$

Si exigimos que el código sea instantáneo, la inecuación de Kraft impone que el argumento del segundo logaritmo del segundo miembro de (4-6) sea igual o menor que la unidad. Por lo tanto, su logaritmo deberá ser igual o menor que cero, y

$$H(S) \leq L \log r \tag{4-7a}$$

o bien

$$\frac{H(S)}{\log r} \leq L \tag{4-7b}$$

$H(S)$ viene medida en bits en la ecuación (4-7*b*). Recordemos que L es el número medio de símbolos utilizados para codificar S. Expresando la entropía asimismo en unidades r-arias, como en (2-5*c*), la relación (4-7*b*) podría escribirse en la forma

$$H_r(S) \leq L \tag{4-7c}$$

4-2. Método de codificación de fuentes especiales.

Es importante destacar que la relación (4-7) marca un hito en el estudio de la teoría de la información. Esta ecuación constituye el primer indicio demostrativo de la relación existente entre la definición de información y una cantidad (en este caso L) *que no depende de la definición.* Con esta ecuación se comienza a desarrollar la justificación de nuestra medida de información.

A primera vista (4-7) no parece estar relacionada más que con L, longitud media de un código instantáneo. En ciertos casos, sin embargo, es posible deducir más consecuencias sin más que fijarse en los argumentos que han conducido a su definición. Examinemos con atención las condiciones que transforman (4-7) en una igualdad. La desigualdad se introdujo en dos puntos, primeramente en (4-4) y después al suprimir el segundo término de (4-6). De esta última se deduce como condición de igualdad, aplicable también a (4-7)

$$\sum_{j=1} r^{-l_j} = 1 \tag{4-8}$$

Volviendo atrás en el cálculo, hasta (4-4), se ve que la condición necesaria y suficiente para la igualdad es,

$$\begin{aligned} P_i &= Q_i \\ &= \frac{r^{-l_i}}{\sum_{j=1}^{q} r^{-l_j}} \\ &= r^{-l_i} \quad \text{para todo } i \end{aligned} \tag{4-9a}$$

o, de otra manera

$$\log_r \frac{1}{P_i} = l_i \quad \text{para todo } i \tag{4-9b}$$

Resumiendo estas consideraciones, puede decirse que, con un código instantáneo y una fuente de memoria nula, L debe ser igual o mayor que $H_r(S)$. Además L alcanzará su valor mínimo *si, y solamente si,* pueden elegirse las longitudes de las palabras, l_i, iguales a $\log_r(1/P_i)$. La condición de igualdad es, por consiguiente, que $\log_r(1/P_i)$ sea un número entero para cualquier valor de i.

En otras palabras, la condición de igualdad es que las probabilidades de los símbolos, P_i, sean de la forma $(1/r)^{\alpha_i}$, donde α_i es un número entero. Por feliz coincidencia se descubre además, que *si esas condiciones se cumplen, se habrán encontrado las longitudes de las palabras que constituyen un código compacto.* Bastará con elegir l_i igual a α_i. Una vez deducidas las longitudes, la construcción del código deberá hacerse siguiendo el procedimiento indicado en el apartado 3-8.

Ejemplo 4-1. Se ha alcanzado un punto en que pueden contestarse algunas de las preguntas sobre codificación planteadas en el capítulo 1. La tabla 4-1 reproduce la fuente de memoria nula definida anteriormente en la tabla 1-4.

TABLA 4-1. FUENTE DE INFORMACIÓN

Símbolo de la fuente	*Probabilidad del símbolo* P_i
s_1	1/4
s_2	1/4
s_3	1/4
s_4	1/4

La entropía de la fuente es:

$$H = \sum_{i=1}^{4} P_i \log \frac{1}{P_i}$$

$$= 2 \text{ bits/símbolo}$$

De (4-7*c*) se desprende que es *imposible* codificar los símbolos de esta fuente mediante un código binario unívoco, de longitud media inferior a 2 binits por símbolo. Cada símbolo de la fuente tiene una probabilidad de $1/4 = (1/2)^2$, luego, según (4-9*b*), un código compacto deberá tener cuatro palabras de longitud 2. Tal código fue definido en el capítulo 1. Es el siguiente

$$s_1 \rightarrow 00$$

$$s_2 \rightarrow 01$$

$$s_3 \rightarrow 10$$

$$s_4 \rightarrow 11$$

La longitud media por palabra es de 2 binits por símbolo, no existiendo ningún código unívoco de esta fuente con longitud media inferior.

En la tabla 1-5 se definió la fuente de memoria nula de la tabla 4-2.

TABLA 4-2. FUENTE DE INFORMACIÓN

Símbolo de la fuente	*Probabilidad del símbolo* P_i
s_1	1/2
s_2	1/4
s_3	1/8
s_4	1/8

La entropía de esta fuente tiene por valor

$$H = \sum_{i=1}^{4} P_i \log \frac{1}{P_i}$$

$$= 1/2 \log 2 + 1/4 \log 4 + 1/8 \log 8 + 1/8 \log 8$$

$$= 1\ 3/4 \text{ bits/símbolo}$$

La menor longitud media que se podrá obtener en un código instantáneo es, por consiguiente, de 1 3/4 binits por símbolo. En el capítulo 1 se llegó a 1 7/8 binits por símbolo, en el caso más favorable. Sin embargo, las probabilidades de los símbolos de la fuente eran de la forma $(1/2)^{\alpha_i}$, con α_i entero, por lo que podrá alcanzarse un mínimo de 1 3/4 binits por símbolo. Con la ayuda de (4-9*b*) vemos que se consigue adoptando palabras de longitudes iguales respectivamente a 1, 2, 3 y 3. El código es el siguiente

$$s_1 \rightarrow 0$$

$$s_1 \rightarrow 10$$

$$s_3 \rightarrow 110$$

$$s_4 \rightarrow 111$$

Como comprobación, calcularemos directamente el valor de L:

$$L = \sum_{i=1}^{4} P_i l_i = 1\ 3/4 \text{ binits/símbolo}$$

Ejemplo 4-2. Como ejemplo final en que puede alcanzarse el mínimo definido por las ecuaciones (4-7), consideremos la fuente de memoria nula de la tabla 4-3.

TABLA 4-3. FUENTE DE INFORMACIÓN

Símbolo de la fuente	*Probabilidad dtl símbolo* P_i
s_1	1/3
s_2	1/3
s_3	1/9
s_4	1/9
s_5	1/27
s_6	1/27
s_7	1/27

Supongamos que se desea construir un código *trinario* instantáneo. Calcularemos, en primer lugar, la entropía de la fuente (empleando unidades trinarias para simplificar el cálculo):

$$H_3 = \sum_{i=1}^{7} P_i \log_3 \frac{1}{P_i}$$

$$= 13/9 \text{ unidades trinarias/símbolo}$$

Por consiguiente, no se podrá construir para esta fuente un código instantáneo trinario con una media inferior a 13/9 símbolos trinarios por símbolo. Este código existe, ya que las probabilidades P_i de la fuente son de la forma $(1/3)^{\alpha_i}$, con α_i número entero. Haciendo uso de la ecuación (4-9*b*) para calcular las longitudes de las palabras, obtenemos finalmente el código

$$s_1 \rightarrow 0$$

$$s_2 \rightarrow 1$$

$$s_3 \rightarrow 20$$

$$s_4 \rightarrow 21$$

$$s_5 \rightarrow 220$$

$$s_6 \rightarrow 221$$

$$s_7 \rightarrow 222$$

Como comprobación, calcularemos directamentee l valor de L:

$$L = \sum_{i=1}^{7} P_i l_i = 13/9 \text{ símbolos trinarios/símbolo de la fuente}$$

4-3. Primer teorema de Shannon.

En el apartado anterior se ha resuelto el problema de la codificación de una fuente de memoria nula con símbolos cuyas probabilidades tienen la forma $(1/r)^{a_i}$. Dedicaremos a continuación nuestra atención a las fuentes de memoria nula cuyos símbolos tienen probabilidades arbitrarias.

La ecuación (4-9*b*) dice que si $\log_r (1/P_i)$ es un número entero, l_i debe hacerse igual a este valor. Si no lo es, parece lógico formar un código compacto eligiendo un l_i igual al número entero inmediatamente superior a $\log_r (1/P_i)$. De hecho esta conjetura no es correcta, pero seleccionando l_i de acuerdo con esta regla se obtendrán algunos resultados interesantes. Por lo tanto, se hará l_i igual al número entero que satisface la relación

$$\log_r \frac{1}{P_i} \leq l_i < \log_r \frac{1}{P_i} + 1 \tag{4-10}$$

En primer lugar, se comprobará que las longitudes definidas por este procedimiento cumplen la inecuación de Kraft y son, en consecuencia, aceptables para constituir un código instantáneo. Hallando el antilogaritmo de la primera inecuación de (4-10) se encuentra

$$\frac{1}{P_i} \leq r^{l_i}$$

o bien

$$P_i \geq r^{-l_i} \tag{4-11}$$

Sumando esta expresión, extendida a todos los valores de i, se obtiene

$$1 \geq \sum_{i=1}^{q} r^{-l_i}$$

Que demuestra que (4-10) define un conjunto de l_i válido para un código instantáneo.

Multiplicando (4-10) por P_i y sumando para todos los valores de i se obtiene

$$H_r(S) \leq L < H_r(S) + 1 \tag{4-12}$$

Antes de continuar es interesante destacar la diferencia fundamental que existe entre (4-12) y el valor mínimo de L definido por (4-7). Las ecuaciones (4-7) determinan el valor mínimo de la longitud media L, independientemente del sistema de codificación empleado. El único requisito exigido es que el código sea instantáneo. La ecuación (4-12), por otra parte, se dedujo admitiendo el procedimiento de codificación definido en (4-10). En definitiva, ambas ecuaciones definen los valores máximo y mínimo de L, *válidos al utilizar el método de codificación enunciado en* (4-10).

Puesto que (4-12) puede aplicarse a cualquier fuente de memoria nula, lo haremos a la extensión de orden n de la fuente original

$$H_r(S^n) \leq L_n < H_r(S^n) + 1 \tag{4-13}$$

L_n representa la longitud media de las palabras correspondientes a los *símbolos de la extensión de orden n* de la fuente S. Esto es, si λ_i es la longitud de la palabra correspondiente al símbolo σ_i y $P(\sigma_i)$ la probabilidad de σ_i, entonces

$$L_n = \sum_{i=1}^{q^n} P(\sigma_i)\,\lambda_i \tag{4-14}$$

L_n/n, por lo tanto, es el número medio de símbolos * empleados en cada símbolo simple de S. Según (2-16), la entropía de S^n es igual a n veces la entropía de S. La ecuación (4-13) puede, entonces, escribirse en la forma

$$\boxed{H_r(S) \leq \frac{L_n}{n} < H_r(S) + \frac{1}{n}} \tag{4-15a}$$

de modo que siempre será posible encontrar un valor de L_n/n tan próximo a $H_r(S)$ como queramos, sin más que codificar la extensión de orden n de S, en lugar de S:

$$\boxed{\lim_{n\to\infty} \frac{L_n}{n} = H_r(S)} \tag{4-15b}$$

* Los símbolos L_n/n y L no deben confundirse. Ambos se refieren al número medio de símbolos empleados por símbolo de la fuente. L_n/n, sin embargo, indica que con objeto de alcanzar este valor medio los símbolos s_i de la fuente se han codificado en grupos de n, en lugar de independientemente.

La ecuación (4-15*a*) se conoce como *primer teorema de Shannon* o *teorema de la codificación sin ruido.* Constituye uno de los dos teoremas fundamentales de la teoría de la información. La ecuación (4-15*a*) dice que el número medio de símbolos *r*-arios correspondientes a un símbolo de la fuente puede hacerse tan pequeño, pero no inferior, a la entropía de la fuente expresada en unidades de orden *r*. El precio que se paga por la disminución de L_n/n es un aumento en la complejidad de la codificación debido al gran número (q^n) de símbolos de la fuente que hay que manejar.

4-4. Aplicación del primer teorema de Shannon a las fuentes de Markov.

Los resultados obtenidos en los apartados anteriores pueden generaralizarse con objeto de incluir también las fuentes de Markov. Se harán las pruebas necesarias aplicando los límites máximo y mínimo de la longitud media obtenidos a una fuente afín adecuada, la fuente de memoria nula. Bastará hacer uso de las propiedades de las fuentes afines, deducidas en los apartados 2-6 y 2-7, para completar la demostración.

Definiremos una fuente de Markov de primer orden S, de símbolos $s_1, s_2, \ldots, s_q$, y probabilidades $P(s_i/s_i)$. Definiremos también S^n, extensión de orden n de S, de símbolos $\sigma_1, \sigma_2, \ldots, \sigma_{qn}$, y probabilidades condicionales $P(\sigma_i/\sigma_j)$. Llamaremos a las probabilidades de primer orden (incondicionales) de S y S^n, P_i y $P(\sigma_i)$, respectivamente. El proceso de codificación de los símbolos $s_1, s_2, \ldots, s_q$ en un código bloque instantáneo es el mismo tanto para S como para su fuente adjunta $\bar{S}$. Si la longitud de la palabra correspondiente a s_i es l_i, la longitud media del código será [(4-1)].

$$L = \sum_{i=1}^{q} P_i l_i \tag{4-16}$$

Por otra parte, las longitudes medias correspondientes a S y $\bar{S}$ son iguales, ya que P_i, probabilidad de primer orden de s_i, es la misma en ambas fuentes. Sin embargo, $\bar{S}$ es una fuente de *memoria nula*, por lo que podemos aplicar la relación (4-7*c*), deduciendo

$$H_r(\bar{S}) \leq L \tag{4-17}$$

Inecuación que puede generalizarse en la forma siguiente

$$H_r(S) \leq H_r(\bar{S}) \leq L \tag{4-18}$$

Escribiendo la misma inecuación para S^n y $\overline{S^n}$, obtenemos

$$H_r(S^n) \leq H_r(\overline{S^n}) \leq L_n \tag{4-19}$$

L_n es la longitud media de la palabra correspondiente a un símbolo σ_i, tal como se definió en (4-14).

Nuevamente, como en el apartado 4-3, resulta evidente el carácter general de estas inecuaciones, en el sentido de que no dependen del sistema de codificación empleado. Eligiendo l_i de acuerdo con (4-10), se podrá acotar L, superior e inferiormente

$$H_r(\bar{S}) \leq L < H_r(\bar{S}) + 1 \tag{4-20}$$

o, para la fuente extensión,

$$H_r(\overline{S^n}) \leq L_n < H_r(\overline{S^n}) + 1 \tag{4-21}$$

Haciendo intervenir la relación (2-41) y dividiendo por n, se encuentra

$$H_r(S) + \frac{H_r(\bar{S}) - H_r(S)}{n} \leq \frac{L_n}{n} < H_r(S) + \frac{[H_r(\bar{S}) - H_r(S)] + 1}{n} \tag{4-22}$$

con lo que, de nuevo, puede conseguirse que L_n/n se acerque a $H_r(S)$ tanto como se quiera, sin más que elegir un valor de n suficientemente grande, es decir, codificando por grupos de suficiente longitud. Esta conclusión constituye la aplicación del primer teorema de Shannon a las fuentes de Markov de primer orden. La demostración correspondiente a las fuentes de Markov de orden m no difiere prácticamente de la que se acaba de realizar (Problema 4-1).

4-5. Codificación sin extensiones.

La demostración del primer teorema de Shannon (en los casos de memoria nula y Markov) ha resultado muy provechosa. Por una parte la relación (4-10) define un método para la determinación de las longitudes de las palabras. Utilizando este método para elegir las longitudes de un código bloque que codifique los símbolos de S^n *y tomando un valor de n suficientemente grande,* L_n/n puede tomar un valor tan cercano a $H_r(S)$ como se desee. ¿Qué ocurre, sin embargo, si n no es suficientemente grande? Para un valor fijo de n, el teorema

dice que eligiendo las longitudes de acuerdo con (4-10), la longitud media no será mayor que el segundo miembro de (4-15*a*) [o (4-22)]. El teorema, sin embargo, determina el valor exacto de L (o L_n/n). Aun más importante, no garantiza en modo alguno que eligiendo las longitudes de acuerdo con (4-10), el valor de L (o L_n/n) encontrado sea el más pequeño que puede obtenerse para ese valor de n.

Un simple ejemplo servirá para demostrar que (4-10) no es sino un procedimiento mediocre para elegir las longitudes de las palabras. Construyamos un código instantáneo binario para la fuente de memoria nula definida en la tabla 4-4. Supongamos que se desea codificar directamente, sin recurrir a la segunda, ni a una extensión de orden superior. ¿Cuál es la longitud media menor que puede obtenerse sin extensiones?

De acuerdo con (4-10) se comenzará calculando $\log(1/P_i)$, cuyos valores figuran en la tercera columna de la tabla 4-4. Se elige entonces,

TABLA 4-4. EJEMPLO DE CODIFICACIÓN

Símbolo de la fuente	P_i	$\log \frac{1}{P_i}$	l_i	*Código* $\mathscr{A}$	*Código* $\mathscr{B}$
s_1	2/3	0.58	1	0	0
s_2	2/9	2.17	3	100	10
s_3	1/9	3.17	4	1010	11

la longitud de la palabra correspondiente a S_i de forma que satisfaga la relación

$$\log \frac{1}{P_i} \leq l_i < \log \frac{1}{P_i} + 1$$

Las longitudes l_i se han enumerado en la cuarta columna de la tabla. El código $\mathscr{A}$ de la quinta columna, es uno de los códigos instantáneos que pueden formarse con esas longitudes. Su longitud media es

$$L_{\mathscr{A}} = 2/3 \times 1 + 2/9 \times 3 + 1/9 \times 4$$
$$= 1{,}78 \text{ binits/símbolos de la fuente}$$

Su entropía tiene el valor

$$H(S) = \sum_{i=1}^{3} P_i \log \frac{1}{P_i}$$
$$= 1{,}22 \text{ bits/símbolos de la fuente}$$

Recordemos que $L_{\mathscr{A}}$ está acotado en la forma siguiente

$$H(S) \leq L_{\mathscr{A}} < H(S) + 1 \qquad (4\text{-}23)$$

No es difícil encontrar un código instantáneo mejor que el código $\mathscr{A}$. Tal código ($\mathscr{B}$) figura en la última columna de la tabla. Su longitud media tiene el valor

$$L_{\mathscr{B}} = 2/3 \times 1 + 2/9 \times 2 + 1/9 \times 2$$
$$= 1{,}33 \text{ binits/símbolos de la fuente}$$

Este valor supone una notable mejora sobre la longitud media del código $\mathscr{A}$. Por otra parte, es evidente que no puede ganarse mucho más codificando la segunda (o superior) extensión de la fuente. Efectivamente, el mejor resultado que puede obtenerse es 1,22 binits por símbolo, habiéndose alcanzado ya 1,33 binits.

4-6. Construcción de códigos compactos binarios. Códigos de Huffman.

Un código instantáneo, correspondiente a una fuente de información cualquiera, tendrá una longitud media igual o mayor que la entropía de la fuente. En el ejemplo de la tabla 4-4 se ha visto, sin embargo, que el método de codificación empleado hasta aquí conduce a un código compacto solamente si n, orden de la extensión considerada, es suficientemente grande. ¿Qué procedimiento seguir para construir un código compacto correspondiente a una fuente dada?

En la definición de código compacto no interviene realmente el valor límite de L_n/n. El código compacto de una fuente S es el de menor longitud media que se obtiene al codificar los símbolos de la fuente de *uno en uno*. En este apartado se señalará un procedimiento para generar un código compacto en el caso de alfabeto binario. El caso general, generación a partir de un alfabeto *r-ario*, será tratado en el apartado 4-8. Ambos problemas fueron resueltos por Huffman (1952).

Consideremos una fuente S, de símbolos $s_1, s_2, \ldots, s_q$ y probabilidades $P_1, P_2, \ldots, P_q$. Supongamos los símbolos ordenados de tal forma que $P_1 \geq P_2 \geq \ldots \geq P_q$. Imaginando que los dos últimos símbolos de

S se confunden en uno solo, se obtiene una nueva fuente * de $q-1$ símbolos. La denominaremos fuente *reducida* de S. Los símbolos de la reducida pueden reordenarse, agrupando de nuevo los dos de menor probabilidad para formar una nueva fuente reducida. Continuando de esta forma, se obtendrá una secuencia de fuentes, cada una con un símbolo menos que la anterior, hasta llegar a una fuente de solamente dos símbolos.

Ejemplo 4-3. En la fig. 4-1 aparece una fuente de seis símbolos iniciales, junto con sus reducciones sucesivas.

Fuente original		Fuentes reducidas			
Símbolos	Probabilidades	S_1	S_2	S_3	S_4
s_1	0.4	0.4	0.4	0.4	0.6
s_2	0.3	0.3	0.3	0.3	0.4
s_3	0.1	0.1	0.2		
s_4	0.1	0.1	0.1		
s_5	0.06	0.1			
s_6	0.04				

FIG. 4-1. Una fuente y sus reducciones.

La formación de la secuencia de fuentes reducidas constituye el primer paso en la creación del código compacto instantáneo correspondiente a la fuente original S. El segundo paso consiste simplemente en fijarse en que el código compacto instantáneo binario de la última reducida (fuente de solo dos símbolos) está formado por las palabras 0 y 1. Finalmente, demostraremos que el código instantáneo compacto de una de las fuentes de la secuencia se deduce fácilmente conocido el de la fuente inmediata siguiente. Una vez demostrado esto, comenzando por la última fuente y el código instantáneo compacto hallado, se irá ascendiendo hasta encontrar el código instantáneo compacto correspondiente a la fuente original.

Sea S_j el código instantáneo compacto correspondiente a una de las fuentes de secuencia. Uno de sus símbolos, digamos s_a, estará for-

* Por conveniencia se considerará esta fuente como de memoria nula. Puesto que debemos codificar los símbolos de S uno por uno, poco importa que S sea de memoria nula o de Markov.

mado por *dos* símbolos de la fuente precedente S_{j-1}. Sean $s_{\alpha 0}$ y $s_{\alpha 1}$. Todos los demás símbolos de S_j se identifican con uno solo de s_{j-1}. Según esto, código instantáneo compacto correspondiente a S_{j-1} se deduce del correspondiente a S_j de acuerdo con la regla siguiente:

> Se asigna a cada símbolo de S_{j-1} (excepto $s_{\alpha 0}$ y $s_{\alpha 1}$) la palabra asignada al símbolo de S_j. Las palabras correspondientes a $s_{\alpha 0}$ y $s_{\alpha 1}$ se forman añadiendo un 0 y un 1, respectivamente, a la palabra asignada a s_α. (4-24)

La demostración de que el código definido es instantáneo es inmediata [condición (3-1)]. Por el contrario, la demostración de que constituye un código compacto no lo es tanto, por lo que se aplazará hasta después de describir la construcción de un código compacto.

Ejemplo 4-4. La fig. 4-2 representa el proceso seguido en la síntesis del código compacto binario de la fuente de la fig. 4-1.

Fuente original			Fuentes reducidas							
Símbolos	Probabilidades	Código	S_1		S_2		S_3		S_4	
s_1	0.4	1	0.4	1	0.4	1	0.4	1	0.6	0
s_2	0.3	00	0.3	00	0.3	00	0.3	00	0.4	1
s_3	0.1	0100	0.1	011	0.2	010	0.3	01		
s_4	0.1	0101	0.1	0100	0.1	011				
s_5	0.06	0110	0.1	0101						
s_6	0.04	0111								

FIG. 4-2. Síntesis de un código compacto.

El código compacto de la columna izquierda se ha formado en los tres pasos explicados. Primero se construye una secuencia de fuentes reducidas de la fuente original S. Se asignan a continuación los códigos 0 y 1 a la última fuente de la secuencia (en nuestro caso, S_4) y, finalmente, se pasa de S_4 a S componiendo las secuencias fuentes reducidas. Al hacerlo, una palabra del código primitivo da lugar a dos palabras del nuevo código.

El procedimiento mismo pone de relieve algunas propiedades de los códigos compactos. Su multiplicidad es especialmente importante.

Nótese que el método seguido para pasar de una fuente reducida a la siguiente consiste simplemente en añadir un binit a la palabra descompuesta. Es indiferente cual de cada una de las dos palabras formadas se asigna a cada símbolo de la fuente, lo que significa que la asociación de los símbolos 0 y 1 a las distintas palabras del código compacto se hace de forma completamente arbitraria. Puede sustituirse el dígito de orden *j* de cada palabra por su complemento * y obtenerse un nuevo código compacto. Por ejemplo, sustituyendo por su complemento los dígitos primero y último del código de la figura 4-2, se obtiene el «nuevo» código compacto:

0
10
111
1100
11011
11010

Este procedimiento, sin embargo, da lugar a un código que no presenta más que diferencias menores respecto al anterior. Realmente está deducido de él sin más que modificar los «nombres» de ciertos dígitos. De la misma fuente pueden deducirse dos códigos compactos fundamentalmente diferentes. Para comprobarlo, se sintentizará un nuevo código para el ejemplo de la figura 4-2.

Ejemplo 4-5. La fig. 4-3 representa un código compacto diferente para la misma fuente del ejemplo 4-4.

Fuente original			Fuentes reducidas							
Símbolos	Probabilidades	Código	S_1		S_2		S_3		S_4	
s_1	0.4	1	0.4	1	0.4	1	0.4	1	0.6	0
s_2	0.3	00	0.3	00	0.3	00	0.3	00	0.4	1
s_3	0.1	011	0.1	011	0.2	010	0.3	01		
s_4	0.1	0100	0.1	0100	0.1	011				
s_5	0.06	01010	0.1	0101						
s_6	0.04	01011								

FIG. 4-3. Síntesis de un código compacto.

* El complemento de 0 es 1; el de 1, 0.

Hay que resaltar que el procedimiento seguido en la construcción de las figuras 4-2 y 4-3 es idéntico hasta el instante en que se procede a pasar del código correspondiente a S_1 al correspondiente a la fuente original S. En ese punto puede elegirse cualquiera de las tres palabras

011
0100
0101

Eligiendo la primera, se obtiene un código de longitudes

1, 2, 4, 4, 4, 4

Eligiendo una cualquiera de las otras dos, las palabras del código resultante tendrán las longitudes

1, 2, 3, 4, 5, 5

Las longitudes medias de los códigos son idénticas:

$$L = 1\,(0{,}4) + 2\,(0{,}3) + 4\,(0{,}1) + 4\,(0{,}1) + 4\,(0{,}06) + 4\,(0{,}04)$$
$$= 2{,}2\text{binits/símbolo}$$
$$L = 1\,(0{,}4) + 2\,(0{,}3) + 3\,(0{,}1) + 4\,(0{,}1) + 5\,(0{,}06) + 5\,(0{,}04)$$
$$= 2{,}2\text{binits/símbolo}$$

no pudiendo construirse un código instantáneo de longitud media menor para esta fuente.

Fuente original			Fuente reducida	
Símbolos	Probabilidades	Código	S_1	
s_1	0.5	0	0.5	0
s_2	0.25	10	0.25	10
s_3	0.125	110	0.125	110
s_4	0.100	1110	0.125	111
s_5	0.025	1111		

FIG. 4-4. Síntesis de un código compacto.

Otro punto puesto en evidencia por el procedimiento de síntesis es que en algunas ocasiones resulta innecesario continuar la secuencia de reducciones de la fuente original hasta la fuente de dos símbolos. Unicamente deberá reducirse hasta encontrar una reducción para la cual el código sea compacto. Una vez obtenido un código compacto, puede volverse hacia atrás siguiendo la regla definida en (4-2), comenzando en la fuente reducida correspondiente a ese código. La figura 4-4 constituye un ejemplo de esta solución.

Una vez formada la primera reducción de la fuente, puede verse que las probabilidades de los símbolos tienen la forma $(1/2)^{a_i}$, siendo a_i un número entero. Se puede, entonces, hacer uso de los resultados del apartado 4-2 para formar un código compacto a partir de esta reducción, volviendo después hacia atrás hasta definir el código compacto de la fuente original.

4-7. Conclusión de la demostración.

El apartado anterior expuso cómo construir el código compacto correspondiente a una fuente de información cualquiera. Se demostrará a continuación que el código construido de acuerdo con la regla (4-24) es un código compacto. Supóngase encontrado un código compacto $\mathscr{C}_j$ correspondiente a una reducción, digamos S_j, de una fuente original S. Sea L_j la longitud media de este código. Uno de los símbolos de S_j, s_α, está formado a partir de los dos símbolos menos probables de la reducción precedente S_{j-1}. Sean estos símbolos $s_{\alpha 0}$ y $s_{\alpha 1}$, y $P_{\alpha 0}$ y $P_{\alpha 1}$, sus probabilidades respectivas. La probabilidad de s_α será, entonces, $P_\alpha = P_{\alpha 0} + P_{\alpha 1}$. Llamemos $\mathscr{C}_{j-1}$ al código correspondiente a S_{j-1}, formado de acuerdo con la regla (4-24) y sea L_{j-1} su longitud media. La relación entre L_j y L_{j-1} se deduce inmediatamente ya que las palabras de $\mathscr{C}_j$ y $\mathscr{C}_{j-1}$ son idénticas, excepto $s_{\alpha 0}$ y $s_{\alpha 1}$ que son un binit más largas que la palabra s_α. Así pues

$$L_{j-1} = L_j + P_{\alpha 0} + P_{\alpha 1} \tag{4-25}$$

Se desea demostrar que si $\mathscr{C}_j$ es compacto, $\mathscr{C}_{j-1}$ también lo es. En otras palabras, si L_j es la menor longitud media posible de un código instantáneo correspondiente a S_j, L_{j-1} [dada por la ecuación (4-25)] es también la menor longitud media posible de un código instantáneo correspondiente a S_{j-1}. Se demostrará nuevamente por reducción al ab-

surdo. Supongamos encontrado un código compacto para S_{j-1} de longitud media $\tilde{L}_{j-1} < L_{j-1}$. Sean $\tilde{\tilde{X}}_1$, $\tilde{\tilde{X}}_2$, ... $\tilde{\tilde{X}}_{\alpha 1}$ las palabras de dicho código, de longitudes respectivas $\tilde{l}_1$, $\tilde{l}_2$, ..., $\tilde{l}_{\alpha 1}$. Admitamos los subíndices ordenados según el orden decreciente de las probabilidades de los símbolos respectivos, es decir

$$\tilde{l}_1 \leq \tilde{l}_2 \leq \ldots \leq \tilde{l}_{\alpha 1}$$

Una de las palabras (por ejemplo, $\tilde{\tilde{X}}_{\alpha 0}$) debe coincidir con $\tilde{\tilde{X}}_{\alpha 1}$ salvo su último dígito. Si esto no fuese así, podría suprimirse el último dígito de $\tilde{\tilde{X}}_{\alpha 1}$ y disminuir en una unidad la longitud media del código, sin que deje ser instantáneo. A continuación se construiría $\tilde{C}_j$, código correspondiente a S_j, combinanco $\tilde{\tilde{X}}_{\alpha 1}$ y $\tilde{\tilde{X}}_{\alpha 0}$ y suprimiendo el último binit sin alterar los demás. El resultado es un código instantáneo para S_j, de longitud media $\tilde{L}_j$, relacionado con $\tilde{L}_{j-1}$, según la expresión

$$\tilde{L}_{j-1} = \tilde{L}_j + P_{\alpha 0} + P_{\alpha 1} \qquad (4\text{-}26)$$

Puede compararse esta ecuación con (4-25), de donde se ve que nuestra hipótesis, $\tilde{L}_{j-1} < L_{j-1}$, implica que se pueda construir un código de longitud media $\tilde{L}_j < L_j$, lo que es absurdo ya que se ha supuesto que el código de longitud media L_j era compacto.

Se ha demostrado que (4-24) permite pasar de un código compacto a otro. Antes de considerar el caso general de codificación en un alfabeto de r símbolos puede ser de interés destacar un par de propiedades de los códigos compactos descubiertos durante la demostración. La primera se enuncia simplemente diciendo que si las probabilidades de los símbolos de una fuente están ordenadas en orden decreciente $P_1 \geq P_2 \geq \ldots \geq P_q$, las longitudes de las palabras asignadas a esos símbolos lo estarán en orden creciente, $l_1 \leq l_2 \leq \ldots \leq l_q$. No es nada sorprendente. Constituye sencillamente expresión del hecho de asignar las palabras más cortas a los símbolos más probables del código. La segunda propiedad es quizás algo menos evidente. Se ha demostrado que las longitudes de las dos últimas palabras (ordenadas por probabilidades decrecientes) de un código compacto eran idénticas:

$$l_q = l_{q-1} \qquad (4\text{-}27)$$

Si existen varios símbolos de probabilidades P_a, los subíndices se podrán elegir de forma que las palabras asignadas a los dos últimos símbolos difieran solo en su último dígito.

4-8. Códigos compactos *r*-arios.

El apartado 4-6 puso de relieve la construcción de un código binario compacto en tres pasos sucesivos. En primer lugar se forma una secuencia de fuentes reducidas de la fuente original. A continuación se busca un código compacto para cualquiera de las fuentes de la secuencia, y, finalmente, se procede a recorrer la secuencia, en sentido inverso, construyendo códigos compactos a partir del hallado, hasta formar el correspondiente a la fuente original *S*. En el presente apartado se comprobará que el procedimiento de generación de un código compacto, cuando el alfabeto consta de r símbolos, consta de las mismas tres etapas. Las dos últimas, además, no difieren fundamentalmente del caso binario.

La formación de las fuentes reducidas que preparan la síntesis de un código *binario* compacto se llevaba a cabo combinando en uno solo los dos símbolos menos probables de cada fuente. Cuando se desea formar un código compacto r-ario, se deberán combinar r símbolos de manera que constituyen uno solo de la fuente reducida. Sin embargo, aparece un inconveniente que no apareceía en el caso binario. Entonces, cada fuente de la secuencia de fuentes reducidas contenía un símbolo menos que la fuente anterior. En el caso r-ario, por combinar r símbolos en uno solo, cada fuente tendrá $r-1$ símbolos menos que la precedente, siendo de esperar que la última de la secuencia tenga exactamente r símbolos (lo que permitiría construir fácilmente un código compacto para la fuente). Ahora bien, la última fuente tendrá r símbolos solamente si la fuente original estaba formada por $r+\alpha(r-1)$ símbolos, siendo α un número entero. Por lo tanto, si la fuente original no tiene este número de símbolos, deberemos añadir unos cuantos «falsos» símbolos en número suficiente para alcanzarlo. A los falsos símbolos se atribuye probabilidad nula, de modo que pueden ser ignorados una vez que el código haya sido construido.

Ejemplo 4-6. Consideremos la fuente *S* de 11 símbolos de la fig. 4-5. Se desea formar una secuencia de fuentes reducidas antes de codificar la fuente en un código cuaternario (código de cuatro símbolos). Si la última fuente de esta

secuencia ha de tener cuatro símbolos, S deberá tener $4 + 3\alpha$, siendo α un número entero. Puesto que 11 no es de la forma $4 + 3\alpha$, añadiremos dos falsos símbolos, de modo que obtengamos un total de 13 símbolos. A continuación, reduciendo la fuente por grupos de cuatro símbolos, alcanzaremos una última fuente de exactamente cuatro símbolos.

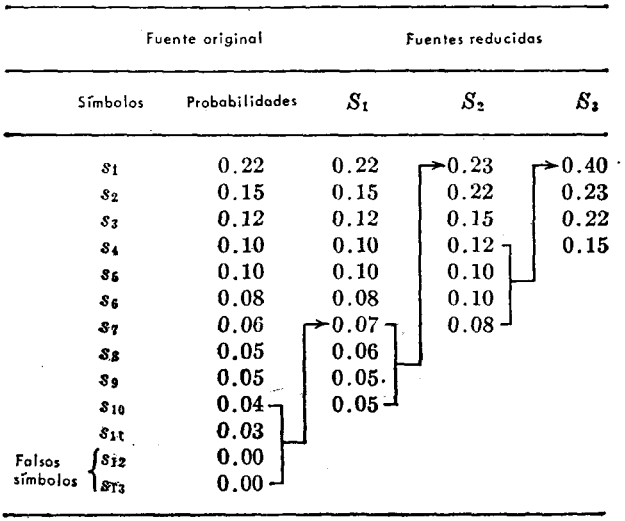


FIG. 4-5. Una fuente y sus reducciones.

Habiendo formado las reducciones de la figura 4-5 se procederá a sintetizar un código compacto según el método expuesto en el apartado 4-6. Se asignarán r palabras, de longitud unidad, a la última reducida con objeto de constituir un código compacto de esta fuente. Se alarga después este código, exactamente como en el caso binario, formando códigos compactos de cada una de las fuentes precedentes. Cada vez que se pasa de una fuente a la anterior se definen r nuevos símbolos a partir de uno solo, alcanzando un aumento neto de $r - 1$ símbolos. La demostración de que partiendo de un código compacto se llega finalmente a un código de la misma clase es análoga a la expuesta en el apartado 4-7 (problema 4-2).

Ejemplo 4-7. Con el fin de mostrar la aplicación práctica del procedimiento descrito anteriormente, la figura 4-6 representa la síntesis de un código compacto correspondiente a la fuente de la figura 4-5.

	Fuente original		Fuentes reducidas								
Símbolo	Probabilidad	Palabra	S_1		S_2			S_3			
s_1	0.22	2	0.22	2	0.23	1		0.40	0		
s_2	0.15	3	0.15	3	0.22	2		0.23	1		
s_3	0.12	00	0.12	00	0.15	3		0.22	2		
s_4	0.10	01	0.10	01	0.12	00		0.15	3		
s_5	0.10	02	0.10	02	0.10	01					
s_6	0.08	03	0.08	03	0.10	02					
s_7	0.06	11	0.07	10	0.08	03					
s_8	0.05	12	0.06	11							
s_9	0.05	13	0.05	12							
s_{10}	0.04	100	0.05	13							
s_{11}	0.03	101									
s_{12}	0.00	102									
s_{13}	0.00	103									

FIG. 4-6. Código compacto cuaternario.

4-9. Rendimiento y redundancia de un código.

El primer teorema de Shannon demostró la existencia de una unidad común con la que puede medirse cualquier fuente de información. El *valor* de un símbolo de una fuente S puede definirse en términos del número equivalente de dígitos binarios necesario para representarlo; el teorema establece que el valor medio de un símbolo de S es $H(S)$. De forma más general, el valor medio de un símbolo de S, en dígitos r-arios, es $H_r(S)$.

Supongamos que L es la longitud media de un código r-ario, unívoco, de la fuente S. L no puede ser inferior a $H_r(S)$. Según esto, se define η, *rendimiento* del código, como

$$\eta = \frac{H_r(S)}{L} \tag{4-28}$$

Igualmente, puede definirse la *redundancia* de un código

$$\text{Redundancia} = 1 - \eta$$
$$= \frac{L - H_r(S)}{L} \tag{4-29}$$

Ejemplo 4-8. Consideremos una fuente de memoria nula $S = \{s_1, s_2\}$, con $P(s_1) = 3/4$ y $P(s_2) = 1/4$. $H(S)$ valdrá

$$H(S) = 1/4 \log 4 + 3/4 \log 4/3$$
$$= 0.811 \text{ bits}$$

Un código compacto de esta fuente puede ser el siguiente:

s_i	$P(\sigma_i)$	*Código compacto*
s_1	3/4	0
s_2	1/4	1

La longitud media del código es 1 binit, de modo que el rendimiento tendrá el valor

$$\eta = 0.811$$

Para mejorarlo se codificará S^2, segunda extensión de S:

σ_i	$P(\sigma_i)$	*Código compacto*
s_1s_1	9/16	0
s_1s_2	3/16	10
s_2s_1	3/16	110
s_2s_2	1/16	111

La longitud media de este código es 27/16 binits. La entropía de S^2, $2H(S)$; así, pues,

$$\eta_2 = \frac{2 \times 0.811 \times 16}{27}$$

Codificando las extensiones de tercero y cuarto orden, se obtienen los rendimientos

$$\eta_3 = 0.985 \qquad \text{y} \qquad \eta_4 = 0.991$$

Según se codifiquen extensiones de mayor orden, el rendimiento se acerca a la unidad. En este ejemplo el crecimiento es bastante rápido, encontrando ya poca ventaja en ir más allá de la segunda extensión. Tal comportamiento es típico de los códigos de Huffman.

Ejemplo 4-9. Se dispone de una fuente S de memoria nula, de 13 símbolos, cuyas probabilidades se representan en la tabla 4-5. En ella se enumeran los códigos compactos (Huffman) correspondientes a alfabetos de 2 a 13 símbolos.

El ejemplo anterior mostró la mejora del rendimiento obtenida al aumentar el orden de las extensiones codificadas. Es también interesante estudiar la variación del rendimiento en función de r, número de símbolos del alfabeto.

TABLA 4-5. CÓDIGOS COMPACTOS PARA ALFABETOS DIFERENTES

		Códigos compactos para r =											
$P(s_i)$	s_i	13	12	11	10	9	8	7	6	5	4	3	2
1/4	s_1	0	0	0	0	0	0	0	0	0	0	0	00
1/4	s_2	1	1	1	1	1	1	1	1	1	1	1	01
1/16	s_3	2	2	2	2	2	2	4	2	2	20	200	1000
1/16	s_4	3	3	3	3	3	3	3	3	30	21	201	1001
1/16	s_5	4	4	4	4	4	4	4	4	31	22	202	1010
1/16	s_6	5	5	5	5	5	5	5	50	32	23	210	1011
1/16	s_7	6	6	6	6	6	6	60	51	33	30	211	1100
1/16	s_8	7	7	7	7	7	70	61	52	34	31	212	1101
1/16	s_9	8	8	8	8	80	71	62	53	40	32	220	1110
1/64	s_{10}	9	9	9	90	81	72	63	54	41	330	221	111100
1/64	s_{11}	A	A	A0	91	82	73	64	550	42	331	2220	111101
1/64	s_{12}	B	B0	A1	92	83	74	65	551	43	332	2221	111110
1/64	s_{13}	C	B1	A2	93	84	75	66	552	44	333	2222	111111
Longitud media L...		1	33/32	67/64	17/16	9/8	19/16	5/4	87/64	23/16	25/16	131/64	25/8

La entropía de la fuente de la tabla es de 3.125 bits por símbolo. Con este dato y (4-28) se puede representar la variación del rendimiento en función de r.

En la figura 4-7 se aprecia que el rendimiento tiende a crecer al disminuir r. Sin embargo, el crecimiento no es monótono; nótese los valores que adopta para $r = 2$ y $r = 4$. Las probabilidades de los símbolos tienen la forma $1/2^{\alpha}$ ó $1/4^{\alpha}$, donde α es un número entero. En estos casos, se sabe (apartado 4-2) que existe un código compacto de longitud media igual a la entropía.

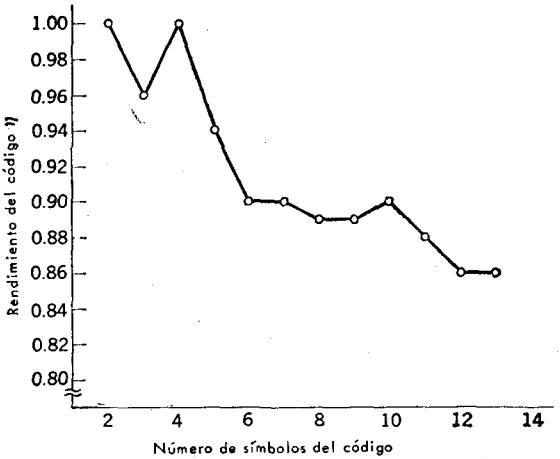


FIG. 4-7. Rendimiento del código en función del número de sus símbolos.

NOTAS

Nota 1. La demostración del primer teorema de Shannon realizada en este capítulo se aplica solamente a fuentes de Markov ergódicas, con un número finito de símbolos (es decir, de estados). Una demostración más elegante, válida para cualquier fuente ergódica, estacionaria, fue dada por McMillan (1953) en forma ligeramente diferente, llamada propiedad de equidistribución asintótica (AEP). En una fuente general S, sea

$$I(s_1, s_2, \ldots, s_n) = \log \frac{1}{P(s_1, s_2, \ldots, s_n)}$$

S tiene la propiedad de equidistribución asintótica si $I(s_1, s_2, \ldots, s_n)/n$ converge en probabilidad con $H(S)$. La importancia de esta propiedad reside en el hecho de que una fuente que la posee emite secuencias largas que pueden dividirse en dos clases:

1. Una clase tal que cada secuencia presenta una probabilidad aproximadamente igual a $2^{-nH(S)}$.

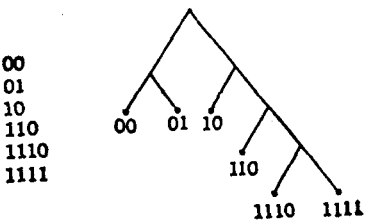
2. Una clase integrada por las secuencias que se presentan en raras ocasiones.

Thomasian (1960) demostró la propiedad de equidistribución asintótica por un método basado simplemente en la teoría combinatoria. Pérez (1959) la generalizó a fuentes más complejas.

Nota 2. En el apartado 4-6 se vió un ejemplo en que dos códigos binarios diferentes (de palabras de diferentes longitudes) podían ser códigos compactos de la misma fuente. Golomb ha estudiado las condiciones en que este fenóme-

no puede presentarse y el número de códigos compactos diferentes que admite una fuente dada.

La construcción de un código puede determinarse mediante un árbol de códigos (Fano, 1961). Consideremos, por ejemplo, el código y el árbol asociado siguientes

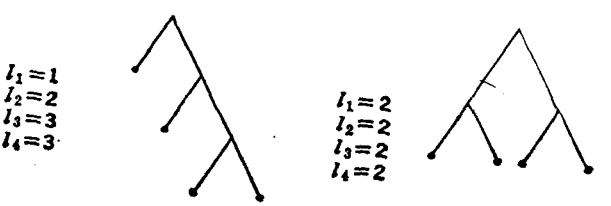


El número de códigos diferentes correspondientes a una fuente de q símbolos puede cifrarse con ayuda de árboles.

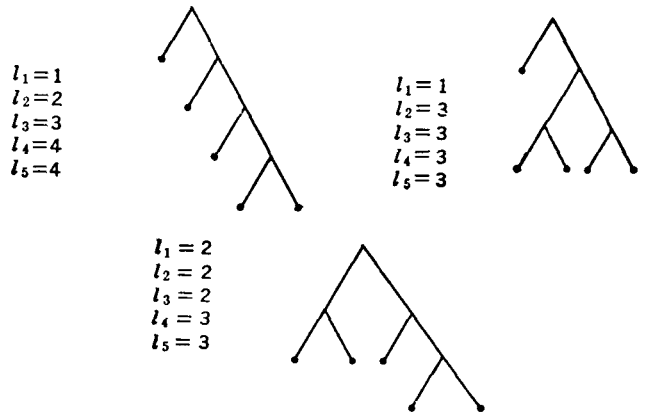
Para $q = 2$ existe solamente un árbol, que corresponde a las longitudes

Para $q = 3$ de nuevo no existe más que un solo árbol, de longitudes

Para $q = 4$ existen dos posibilidades



En el caso de $q = 5$ el número de árboles posibles es de tres



Para $q = 6$ y 7 los árboles posibles son cinco y nueve, respectivamente.

Golomb estableció también las condiciones que deben cumplir las probabilidades de los símbolos para que exista más de un código compacto. En el caso de $q = 4$, por ejemplo, será necesario que $P_1 = P_3 + P_4$. Un análisis más detallado exige, para la existencia de dos códigos compactos, que $1/3 \leq P_1 \leq 2/5$.

Nota 3. El problema de la codificación de fuentes de información se ha tratado admitiendo que la duración (u otro criterio de coste) de cada símbolo es la misma. Si no ocurre así, algunas de las conclusiones del capítulo 4 deben modificarse. Supóngase el alfabeto

$$X = \{x_1, x_2, \ldots, x_r\}$$

y sea t_i la duración del código x_i. Entonces, si $N(T)$ es el número de secuencias de duración T,

$$N(T) = N(T - t_1) + N(T - t_2) + \ldots + N(T - t_r)$$

Resuelta esta ecuación, se comprueba que para valores grandes de T, $N(T)$ crece con AR_0^T, donde A es una constante y R_0 la mayor de las raíces reales de la ecuación característica

$$z^{-t_1} + z^{-t_2} + \ldots + z^{-t_r} - 1 = 0$$

El número asintótico de binits por unidad de tiempo es

$$\lim_{T \to \infty} \frac{\log N(T)}{T} = R_0$$

resultado que puede emplearse para modificar la expresión del primer teorema de Shannon.

Karp (1961) trató el problema de codificación del mismo alfabeto en tiempo finito (equivalente a la codificación de Huffman).

PROBLEMAS

4-1. Demostrar la aplicación de la ecuación (4-22) a las fuentes de Markov de orden m.

4-2. Demostrar que, partiendo de un código compacto r-ario y pasando de una fuente reducida a otra, como se describió en el apartado 4-8, se genera un nuevo código compacto r-ario.

4-3. Una secuencia de símbolos de S^n se codifica en un alfabeto $X = \{x_1, x_2, \ldots, x_r\}$, según el método de Huffman. El resultado puede considerarse una nueva fuente de información de alfabeto X. Demostrar que la probabilidad de los símbolos x_i de la nueva fuente tiende a $1/r$ al crecer n.

4-4. Una fuente binaria de memoria nula tiene las probabilidades $P(0) = 0.1$ y $P(1) = 0.9$.

a) Calcular $H(S)$.

b) Calcular L, longitud media de las palabras de un código compacto de S, cuando $X = \{0, 1\}$.

c) Calcular L_n/n para $n = 2, 3, 4$ y $n \to \infty$ al codificar S^n en un código, siempre con $X = \{0, 1\}$.

d) Calcular el rendimiento de los cuatro dígitos.

4-5. En el problema anterior se codificaron S, S^2, S^3 y S^4 en X. Estos códigos dan lugar a secuencias de ceros y unos, que pueden imaginarse como emitidas por una nueva fuente, S_0, tal como se muestra en la figura P 4-5. Calcular $H(S_0)$ para $n = 1, 2, 3, 4$.

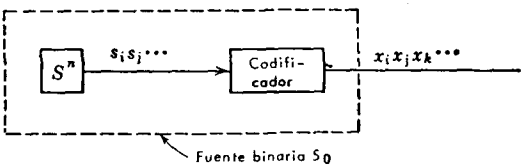


FIG. P 4-5

4-6. Dada la tabla

S	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$P(s_i)$	1/3	1/3	1/9	1/9	1/27	1/27	1/27

a) Calcular $H(S)$ y $H_3(S)$.

b) Encontrar códigos compactos de $H(S)$ cuando $X = \{0, 1\}$ y $X = \{0, 1, 2\}$.

c) Calcular el valor de L en ambos casos.

4-7. Dada la tabla

S	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
$P(s_i)$	0.4	0.2	0.1	0.1	0.05	0.05	0.05	0.05

a) Encontrar un código compacto de la fuente con $X = \{0, 1, 2\}$.

b) A esa fuente y ese alfabeto corresponde más de un código compacto (es decir, formados por palabras de longitudes diferentes). Enumerar todos los conjuntos de longitudes que pueden encontrarse.

4-8. En el problema 2-14, sea $\varepsilon = 1/2$. Existe un código binario correspondiente a S, con $L = H(S)$. Calcular el valor de L', longitud media del código compacto correspondiente a S'.

4-9. La fuente S consta de nueve símbolos, cada uno de probabilidad 1/9.

a) Encontrar un código compacto de alfabeto $X = \{0, 1\}$.

b) Lo mismo con un alfabeto $X = \{0, 1, 2\}$.

c) Idem con el alfabeto $X = \{0, 1, 2, 3\}$.

4-10. Una fuente S tiene seis símbolos de probabilidades respectivas P_1 a P_6. Suponiendo que las probabilidades están ordenadas en la forma $P_1 \geq P_2 \geq \ldots \geq P_6$, encontrar un código compacto de esta fuente de alfabeto $X = \{0, 1, 2, 3\}$. Definir unos conjuntos de longitudes de las palabras de tal código cuando $P_6 = 1/64$.

4-11. Encontrar todos los códigos binarios compactos posibles de la fuente de la tabla siguiente

S	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
$P(s_i)$ . . .	0.20	0.18	0.12	0.10	0.10	0.08	0.06	0.06	0.06	0.04

Se considerarán códigos «diferentes» solamente aquéllos que están formados por palabras de longitudes l_i distintas.

4-12. *a*) Encontrar los cinco árboles diferentes que corresponden a $q = 6$ en la Nota 2.

b) Encontrar los nueve árboles correspondientes a $q = 7$.

4-13. Este problema constituye una generalización de la Nota 2. Encontrar todos los árboles diferentes correspondientes a códigos compactos *trinarios* con $q = 3, 4, 5, 6, 7, 8, 9$.

CAPITULO 5

CANALES E INFORMACION MUTUA

5-1. Introducción.

Los cuatro primeros capítulos trataron de las propiedades de las fuentes de información y las transformaciones de secuencias de símbolos de una fuente en secuencias de símbolos de un código. Es posible establecer una relación entre nuestra medida de la información y las propiedades de las fuentes. En particular, se demostró que la entropía de una fuente (expresada en unidades adecuadas) definía el valor mínimo del número medio de símbolos necesarios para codificar cada símbolo de la fuente. Este mínimo permitió definir en el apartado (4-9) el rendimiento y la redundancia de un código. Realmente, repasando los capítulos anteriores, se pone en evidencia que la casi totalidad de la primera parte del libro se dedicó a sentar las bases para la definición de rendimiento y redundancia, así como a la síntesis de códigos con la menor redundancia posible.

Debido a la preocupación en la minimización de redundancia mostrada hasta aquí, el lector se sorprenderá al apercibir que los capítulos 5 y 6 principalmente estudian diversos procedimientos para volver a introducir redundancia en los códigos. No en todos los casos el código más apropiado es el que contiene poca o ninguna redundancia. En este capítulo nuestra atención se desviará de las *fuentes* de información, orientándose hacia los *canales* de información, esto es, de la generación de información a su transmisión.

La introducción del concepto de canal de información nos lleva inmediatamente a considerar la posibilidad de cometer errores durante el proceso de transmisión. Estudiaremos el efecto de tales errores sobre la transmisión misma, lo que conduce a considerar también la posibili-

dad de codificar tendiendo a minimizar este efecto. El lector no deberá sorprenderse al comprobar que nuestra definición de medida de información sirve asimismo para analizar este tipo de codificación, además del ya considerado. Realmente, a pesar del considerable progreso hecho hasta aquí, la conclusión fundamental de la teoría de la información y la aplicación más decisiva del concepto de entropía no se han encontrado todavía. Este resultado, el segundo teorema de Shannon, utilizará el concepto de entropía para definir la manera en que un canal *no confiable* puede transmitir información *confiable*.

5-2. Canales de información.

El resto del libro versa principalmente sobre los canales de información.

> *Definición. Un canal de información* * viene determinado por un alfabeto de entrada $A = \{a_i\}$, $i = 1, 2, \ldots, r$; un alfabeto de salida $B = \{b_j\}$, $j = 1, 2, \ldots, s$; y un conjunto de probabilidades condicionales $P(b_j/a_i)$. $P(b_j/a_i)$ es la probabilidad de recibir a la salida el símbolo b_j cuando se envía el símbolo de entrada a_i.

Un canal de gran importancia teórica es el binario simétrico (BSC) **.

La figura 5-2 representa el diagrama del BSC. Como es habitual, $\bar{p} = 1 - p$. Este canal posee dos símbolos de entrada ($a_1 = 0$, $a_2 = 1$)

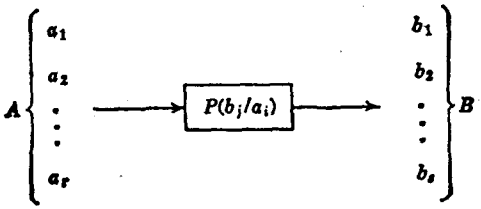


FIG. 5-1. Canal de información.

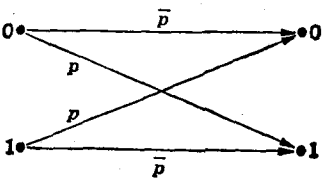


FIG. 5-2. Canal binario simétrico (BSC).

---

* El canal definido en esta forma se denomina, en algunas ocasiones, canal de información de memoria nula. Es posible establecer una definición más general, donde la probabilidad de una salida dada y_i puede depender de varios símbolos precedentes e incluso de los símbolos de salida. Tales canales se reconocen como canales con memoria.

** *N. del T.*: BSC, del inglés, binary symmetric channel.

y dos de salida ($b_1 = 0$, $b_2 = 1$). Es simétrico por ser iguales las probabilidades de recibir un 0 al enviar un 1 y viceversa; esta probabilidad, probabilidad de que tenga lugar un error es p.

La descripción del canal se hace de forma más conveniente disponiendo las probabilidades condicionales como en la figura 5-3.

Salidas

		b_1	b_2	$\cdots$	b_s
	a_1	$P(b_1/a_1)$	$P(b_2/a_1)$	$\cdots$	$P(b_s/a_1)$
Entradas	a_2	$P(b_1/a_2)$	$P(b_2/a_2)$	$\cdots$	$P(b_s/a_2)$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	a_r	$P(b_1/a_r)$	$P(b_2/a_r)$	$\cdots$	$P(b_s/a_r)$

FIG. 5-3. Descripción de un canal de información.

Cada fila corresponde a una entrada determinada siendo sus términos las probabilidades de obtener a la salida las diferentes b_j para una entrada fija. Esta descripción de canal de información se manejará con tanta frecuencia en el futuro que es interesante emplear una notación más reducida. Así, pues, se define

$$P_{ij} = P(b_j/a_i) \tag{5-1}$$

Con lo que la figura 5-3 se transforma en la matriz **P**

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \ldots & P_{1s} \\ P_{21} & P_{22} & \ldots & P_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ P_{r1} & P_{r2} & \ldots & P_{rs} \end{bmatrix} \tag{5-2}$$

Un canal de información está completamente definido por su matriz. Por lo tanto, usaremos indistintamente **P** para representar un canal o su matriz.

Cada fila de la matriz corresponde a una entrada del canal y cada columna a una salida. Hay que destacar una propiedad fundamental de la matriz de un canal; la suma de los términos de una fila cualquiera es igual a la unidad *. Se deduce teniendo en cuenta que si se

* Estas matrices reciben el nombre de matrices *estocásticas* o de *Markov*.

envía un símbolo de entrada a_i, debe obtenerse *algún* símbolo a la salida. Esta condición puede expresarse como sigue

$$\sum_{j=1}^{s} P_{ij} = 1 \qquad i = 1, 2, \ldots, r \tag{5-3}$$

La matriz del canal BSC es

$$\begin{bmatrix} \bar{p} & p \\ p & \bar{p} \end{bmatrix} \tag{5-4}$$

De la misma forma en que se procedió en el caso de las fuentes de información, pueden considerarse bloques de n símbolos de entrada y salida, en lugar de símbolos aislados. Así, la extensión de orden n de un canal se define como sigue.

Definición. Consideremos un canal de información, de alfabeto de entrada $A = \{a_i\}$, $i = 1, 2, \ldots, r$, alfabeto de salida $B = \{b_j\}$, $j = 1, 2, \ldots, s$; y matriz

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \ldots & P_{1s} \\ P_{21} & P_{22} & \ldots & P_{2s} \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ P_{r1} & P_{r2} & \ldots & P_{rs} \end{bmatrix}$$

La extensión de orden n del canal tiene un alfabeto de entrada $A^n = \{\alpha_i\}$, $i = 1, 2, \ldots, r^n$; alfabeto de salida $B^n = \{\beta_j\}$, $j = 1, 2, \ldots, s^n$; y matriz

$$\mathbf{\Pi} = \begin{bmatrix} \Pi_{11} & \Pi_{12} & \ldots & \Pi_{1s^n} \\ \Pi_{21} & \Pi_{22} & \ldots & \Pi_{2s^n} \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ \Pi_{r^n 1} & \Pi_{r^n 2} & \ldots & \Pi_{r^n s^n} \end{bmatrix}$$

Cada una de las entradas α_i consiste en una secuencia de n símbolos elementales de entrada $(a_{i_1}, a_{i_2}, \ldots, a_{i_n})$ y cada salida β_j en una secuencia de n símbolos de salida $(b_{j_1}, b_{j_2}, \ldots, b_{j_n})$. La probabilidad $\Pi_{ij} = P(\beta_j/\alpha_i)$ es igual al producto de las probabilidades elementales correspondientes.

La extensión de un canal de información, como en el caso en que se definió la extensión de una fuente, no constituye un concepto nue-

vo, sino solamente una nueva forma de un concepto antiguo. La extensión de orden n de un canal se obtiene meramente considerando bloques de símbolos de longitud n.

Ejemplo 5-1. La segunda extensión del BSC es un canal con cuatro símbolos de entrada y cuatro de salida. Su matriz está representada en la fig. 5-4.

$$\mathbf{II} = \begin{bmatrix} \bar{p}^2 & \bar{p}p & p\bar{p} & p^2 \\ \bar{p}p & \bar{p}^2 & p^2 & p\bar{p} \\ p\bar{p} & p^2 & \bar{p}^2 & \bar{p}p \\ p^2 & p\bar{p} & \bar{p}p & \bar{p}^2 \end{bmatrix}$$

FIG. 5-4. Matriz del canal $(BSC)^2$.

Puede apreciarse que la matriz del $(BSC)^2$ se expresa como una matriz de matrices. Sea **P**, igual que antes, la matriz del canal BSC. Entonces, la matriz del $(BSC)^2$ puede escribirse en la forma

$$\mathbf{II} = \begin{bmatrix} \bar{p}\mathbf{P} & p\mathbf{P} \\ p\mathbf{P} & \bar{p}\mathbf{P} \end{bmatrix}$$

Esta matriz se conoce como *cuadrado de Kronecker* (Bellman, 1960) (o cuadrado tensorial) de la matriz **P**. En un caso más general, la *matriz de la extensión* de orden n de un canal es la potencia *enésima* de Kronecker de la matriz del canal original.

En la primera parte del libro, se utilizó el concepto de medida de información para medir la cantidad media de información suministrada por una fuente. La función de un canal de información, sin embargo, no es generar información sino transmitirla de la entrada a la salida. Es de esperar, por lo tanto, que nuestra definición de medida de información permita evaluar la habilidad de un canal para transportar información. Este será el caso; a continuación se procederá a definir la cantidad de información que un canal puede transmitir.

5-3. Relaciones entre las probabilidades de un canal.

Consideremos un canal de r símbolos de entrada y s de salida. Lo definiremos por su matriz **P**:

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1s} \\ P_{21} & P_{22} & \dots & P_{2s} \\ \dots & \dots & \dots & \dots \\ P_{r1} & P_{r2} & \dots & P_{rs} \end{bmatrix}. \tag{5-5}$$

Los símbolos de entrada se eligen de acuerdo con sus probabilidades $P(a_1)$, $P(a_2)$, ..., $P(a_r)$. Los símbolos de salida aparecerán de acuerdo con otro conjunto de probabilidades: $P(b_1)$, $P(b_2)$, ..., $P(b_s)$. La relación entre las probabilidades de los diferentes símbolos de entrada y de salida puede deducirse con facilidad. Por ejemplo, el símbolo b_1 puede recibirse en r casos distintos. Enviado a_1 se presentará b_1 con una probabilidad P_{11}; si se envía a_2, se recibirá b_1 con una probabilidad P_{21}, etc. En consecuencia, escribiremos

$$P(a_1)P_{11} + P(a_2)P_{21} + \ldots + P(a_r)P_{r1} = P(b_1) \qquad (5\text{-}6a)$$

$$P(a_1)P_{12} + P(a_2)P_{22} + \ldots + P(a_r)P_{r2} = P(b_2) \qquad (5\text{-}6b)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$P(a_1)P_{1s} + P(a_2)P_{2s} + \ldots + P(a_r)P_{rs} = P(b_s) \qquad (5\text{-}6c)$$

Las ecuaciones (5-6) constituyen la expresión de la probabilidad de los distintos símbolos de salida, conocidas las probabilidades de entrada $P(a_i)$ y la matriz del canal, es decir la matriz de las probabilidades condicionales $P(b_j/a_i)$. En el resto del capítulo supondremos conocidos $P(a_i)$ y $P(b_j/a_i)$, de modo que $P(b_j)$ podrá calcularse a partir de (5-6). Hay que notar, sin embargo, que dadas las probabilidades de *salida* $P(b_j)$ y $P(b_j/a_i)$ no puede calcularse $P(a_i)$ invirtiendo el sistema de ecuaciones lineales (5-6).

Por ejemplo, en un BSC con $p = 1/2$, cualquier conjunto de probabilidades de entrada dará lugar a unos símbolos de salida equiprobables. En general, existirán muchas distribuciones de entrada que determinarán la misma distribución de salida. Dada una distribución de entrada, por otra parte, con ayuda de (5-6) puede determinarse siempre una distribución de salida única.

Además de $P(b_j)$, existen otros dos conjuntos de probabilidades relativas a un canal que pueden calcularse a partir de $P(a_i)$ y $P(b_j/a_i)$. Según la ley de Bayes, la probabilidad condicional de una entrada a_i, cuando se recibe una salida b_j, viene dada por la fórmula

$$P(a_i/b_j) = \frac{P(b_j/a_i)\,P(a_i)}{P(b_j)} \qquad (5\text{-}7a)$$

que, teniendo en cuenta (5-6), se transforma en

$$P(a_i/b_j) = \frac{P(b_j/a_i)\,P(a_i)}{\sum_{i=1}^{r} P(b_j/a_i)\,P(a_i)} \qquad (5\text{-}7b)$$

Las probabilidades $P(a_i/b_j)$ se denominan en algunas ocasiones *probabilidades hacia atrás*, para distinguirlas de las *probabilidades hacia adelante* $P(b_j/a_i)$.

El numerador del segundo miembro de (5-7) es la probabilidad del suceso (a_i, b_j)

$$P(a_i, b_j) = P(b_j/a_i)\,P(a_i) \tag{5-8a}$$

que puede también escribirse en la forma

$$P(a_i, b_j) = P(a_i/b_j)\,P(b_j) \tag{5-8b}$$

Ejemplo 5-2. Expondremos el cálculo de las probabilidades asociadas a un canal de información. Consideremos un canal binario; es decir, $A = \{0, 1\}$ y $B = \{0, 1\}$. Los valores de $P(b_j/a_i)$ están definidos por la matriz del canal

$$\mathbf{P} = \begin{bmatrix} 2/3 & 1/3 \\ 1/10 & 9/10 \end{bmatrix}$$

Las filas y columnas de esta matriz se relacionan con los símbolos de entrada y salida en orden natural. Por lo tanto, $P_x\{b = 0/a = 0\} = 2/3$, $P_x\{b = 1/a = 0\} = 1/3$, etc. Supongamos, finalmente, que $P_x\{a = 0\} = 3/4$ y $P_x\{a = 1\} = 1/4$. Todas estas informaciones se resumen en la figura 5-5.

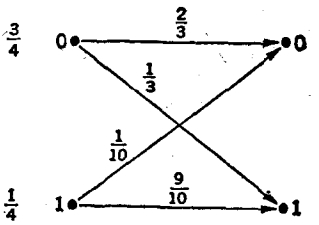


FIG. 5-5. Canal de información con ruidos.

La ecuación (5-6) permite calcular las probabilidades de los símbolos de salida

$$\Pr\{b = 0\} = (3/4)(2/3) + (1/4)(1/10) = 21/40 \tag{5-9a}$$

$$\Pr\{b = 1\} = (3/4)(1/3) + (1/4)(9/10) = 19/40 \tag{5-9b}$$

Vemos que $\Pr\{b = 0\} + \Pr\{b = 1\} = 1$, lo que sirve de comprobación. A partir de la ecuación (5-7) se calculan las probabilidades condicionales de entrada

$$\Pr\{a = 0/b = 0\} = \frac{(3/4)(2/3)}{(21/40)} = 20/21 \tag{5-10a}$$

$$\Pr\{a = 0/b = 1\} = \frac{(3/4)(1/3)}{(19/40)} = 10/19 \tag{5-10b}$$

Las otras probabilidades hacia atrás se calculan por el mismo procedimiento. Un método más simple, sin embargo, consiste en hacer uso del hecho de que $\Pr\{a = 0/b = 0\} + \Pr\{a = 1/b = 0\} = 1$

y $\Pr\{a = 0/b = 1\} + \Pr\{a = 1/b = 1\} = 1$. Según esto,

$$\Pr\{a = 1/b = 0\} = 1/21 \qquad (5\text{-}10c)$$

$$\Pr\{a = 1/b = 1\} = 9/19 \qquad (5\text{-}10d)$$

Las probabilidades de varios sucesos simultáneos se deduce a partir de (5-8). Nos limitaremos a calcular una de ellas:

$$\begin{aligned}\Pr\{a = 0,\ b = 0\} &= \Pr\{a = 0/b = 0\}\Pr\{b = 0\} \\ &= (20/21)(21/40) \\ &= 1/2 \end{aligned} \qquad (5\text{-}11)$$

5-4. Entropías a priori y a posteriori.

Los diferentes símbolos de salida de un canal se presentan de acuerdo con un conjunto de probabilidades $P(b_j)$. Hay que notar que la probabilidad de que se presente un símbolo de salida determinado, p. e., b_i, es igual a $P(b_i)$ solamente *si* se desconoce el símbolo de entrada enviado. En caso contrario, si el símbolo de entrada es a_i, la probabilidad de que el símbolo de salida sea b_j es $P(b_j/a_i)$. De la misma forma, recordaremos que la elección del símbolo de entrada a_i, se efectúa con una probabilidad $P(a_i)$. Sin embargo, si el símbolo de salida es b_j, la probabilidad de que el símbolo de entrada correspondiente sea a_i es $P(a_i/b_j)$ [(5-7)]. Centraremos nuestra atención en el cambio que sufre el valor de la probabilidad de los distintos símbolos de entrada por el hecho de recibir a la salida el símbolo b_j.

Denominaremos $P(a_i)$ la probabilidad *a priori* de los símbolos de entrada, es decir antes de recibir un símbolo de salida determinado. $P(a_i/b_j)$ recibirá el nombre de probabilidad *a posteriori*, probabilidad despues de la recepción de b_j. Según se explicó en el apartado 2-2 puede calcularse la entropía del conjunto de los símbolos de entrada teniendo en cuenta ambas probabilidades. La *entropía a priori* de A es *

$$H(A) = \sum_A P(a) \log \frac{1}{P(a)} \qquad (5\text{-}12)$$

y la *entropía a posteriori* de A, recibido b_j

$$H(A/b_j) = \sum_A P(a/b_j) \log \frac{1}{P(a/b_j)} \qquad (5\text{-}13)$$

* En el resto del libro omitiremos los subíndices de a_i y b_j al escribir sumas de términos extendidos a todos los símbolos de los alfabetos A y B.

La interpretación de estas dos relaciones puede hacerse basándose en el primer teorema de Shannon. $H(A)$ es el número medio de binits necesarios para representar un símbolo de una fuente con una probabilidad a priori $P(a_i)$; $i = 1, 2, \ldots, r$; $H(A/b_j)$ representa el número medio de binits necesarios para representar un símbolo de una fuente con una probabilidad a posteriori $P(a_i/b_j)$, $i = 1, 2, \ldots, r$.

Ejemplo 5-3. La figura 5-6 es repetición de la del ejemplo 5-2. La entropía a priori de los símbolos de entrada tiene el valor

$$H(A) = 3/4 \log 4/3 + 1/4 \log 4 = 0.811 \text{ bit} \tag{5-14}$$

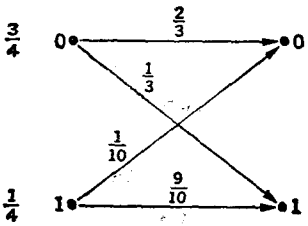


FIG. 5-6. Canal de información con ruidos.

Recibido el símbolo 0 a la salida del canal, las probabilidades a posteriori vienen dadas por (5-10*a*) y (5-10*b*). La entropía a posteriori será

$$H(A/0) = 20/21 \log 21/20 + 1/21 \log 21 = 0.276 \text{ bit} \tag{5-15}$$

Por el contrario, recibido el símbolo 1, la entropía a posteriori tiene el valor

$$H(A/1) = 9/19 \log 19/9 + 10/19 \log 19/10 = 0.998 \text{ bit} \tag{5-16}$$

Así pues, al recibir un 0, la entropía, incertidumbre sobre la entrada enviada, disminuye, aumentando al recibir un 1.

5-5. Generalización del primer teorema de Shannon.

Según el primer teorema de Shannon, la entropía de un alfabeto se interpreta como el número medio de binits necesarios para representar un símbolo de ese alfabeto. Apliquemos esta interpretación al concepto de entropía a priori y a posteriori (figura 5-7).

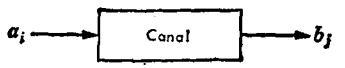


FIG. 5-7. Canal de información.

Antes de recibir un símbolo a la salida de un canal, se asocian las probabilidades a priori $P(a_i)$ con el alfabeto de entrada A. $H(A)$ es el número medio de binits necesarios para representar un símbolo de este alfabeto. Recibido un símbolo, por ejemplo b_j, se asocian al alfabeto de entrada las probabilidades a posteriori $P(a_i/b_j)$. $H(A/b_j)$ es el número medio de binits necesarios para representar un símbolo de ese alfabeto a partir de las probabilidades a posteriori. Puesto que los símbolos se presentan a la salida con probabilidades $P(b_j)$, es de esperar que el número medio de binits necesarios (valor medio extendido también a b_j) para representar un símbolo de entrada a_i, dado un símbolo de salida determinado, sea igual a la entropía media a posteriori

$$\sum_{b} P(b) H(A/b) \tag{5-17}$$

Este resultado es, de hecho, cierto. Sin embargo no se deduce directamente del primer teorema de Shannon. El teorema trata únicamente de la codificación de fuentes de comportamiento estadístico fijo y definido, y no de fuentes cuyo comportamiento estadístico variable se define después de la recepción de cada símbolo de salida. Generalizaremos el teorema de Shannon de forma que cubra también este caso.

La cuestión que se plantea con este objeto es idéntica a la que se planteó al deducir el primer teorema de Shannon, es decir: «¿Cuál es el procedimiento más eficaz para codificar una fuente?» (En este caso, la fuente es A). En esta ocasión, sin embargo, la estadística de la fuente a codificar varía de un símbolo a otro. Su comportamiento estadístico vendrá definido por cada símbolo de salida del canal, b_j. Puesto que a un conjunto de probabilidades corresponde un código compacto que, en general, no lo será para cualquier otro conjunto, aprovecharemos el conocimiento de b_j para formar s códigos binarios *, uno para uno de los b_j símbolos posibles. Cuando el símbolo de salida del canal es b_j, se utiliza el código binario j-imo para codificar el símbolo a_i transmitido. Supongamos que las longitudes de las palabras de los s códigos son las de la tabla 5-1.

---

* No es necesario que sean binarios, pero admitiéndolo se simplifica el desarrollo subsiguiente.

TABLA 5-1. LONGITUDES DE LAS PALABRAS DE s CÓDIGOS

Símbolo de entrada	*Código* 1	*Código* 2	...	*Código* s
a_1	l_{11}	l_{12}	...	l_{1s}
a_2	l_{21}	l_{22}	...	l_{2s}
...	...	...	...	...
a_r	l_{r1}	l_{r2}	...	l_{rs}

Si los códigos han de ser instantáneos, se aplicará a cada código la primera parte del teorema de Shannon (4-7), obteniendo

$$H(A/b_j) \leq \sum_A P(a_i/b_j)\, l_{ij} \triangleq L_j \tag{5-18}$$

Donde L_j es la longitud media del código j. Para calcular L_j deben utilizarse las probabilidades condicionales $P(a_i/b_j)$, en lugar de las marginales $P(a_i)$, ya que el código j-imo se aplica únicamente al recibir el símbolo b_j. El número medio de binits correspondiente a cada término del alfabeto A al codificar de esta forma, se obtiene hallando la media con respecto a todos los símbolos b_j. Multiplicando la ecuación (5-18) por $P(b_j)$ y calculando la suma extendida a B, se encuentra

$$\sum_B H(A/b_j)\, P(b_j) \leq \sum_{A,B} P(a_i,\ b_j)\, l_{ij} \triangleq L \tag{5-19}$$

donde L es el número medio de binits por símbolo del alfabeto A, valor medio respecto a los símbolos de entrada y salida. Es interesante destacar la semejanza entre las ecuaciones (5-19) y (4-7).

Con objeto de demostrar que puede alcanzarse el valor límite impuesto por la relación (5-19), describiremos un procedimiento adecuado de codificación. Si la salida del canal es b_j, seleccionaremos l_{ij}, longitud de la palabra de entrada a_i correspondiente, como el único número entero que satisface la ecuación

$$\log \frac{1}{P(a_i/b_j)} \leq l_{ij} < \log \frac{1}{P(a_i/b_j)} + 1 \tag{5-20}$$

Las longitudes definidas de esta guisa cumplen la inecuación de Kraft para cualquier valor de j*. Las diferentes l_{ij}, por lo tanto, definen

* La demostración es semejante a la del primer teorema de Shannon (apartado 4-3).

s conjuntos de longitudes aptas para s códigos instantáneos. Multiplicando a continuación (5-20) por $P(a_i, b_j) = P(a_i/b_j)\,P(b_j)$

$$P(b_j)\,P(a_i/b_j)\log\frac{1}{P(a_i/b_j)} \leq l_{ij}\,P(a_i,\ b_j) < P(b_j)\,P(a_i/b_j)\log\frac{1}{P(a_i/b_j)} + P(a_i,\ b_j) \qquad (5\text{-}21)$$

y sumando esta ecuación extendida a todos los miembros de los alfabetos A y B:

$$\sum_B P(b)\,H(A/b) \leq \overline{L} < \sum_B P(b)\,H(A/b) + 1 \qquad (5\text{-}22)$$

Esta ecuación es válida para cualquier canal del tipo considerado. En particular lo es para la extensión de orden n del canal original,

$$\sum_{B^n} P(\beta)\,H(A^n/\beta) \leq \overline{L_n} < \sum_{B^n} P(\beta)\,H(A^n/\beta) + 1 \qquad (5\text{-}23)$$

donde $\overline{L_n}$ es la longitud media de palabras de un símbolo de A^n, o, lo que es equivalente, la longitud media de las palabras de n símbolos de A. Las entropías a posteriori $H(A^n/\beta)$ de la ecuación (5-23) son iguales a $n\,H(A/b)$, por lo que la ecuación puede transformarse en

$$\boxed{\sum_B P(b)\,H(A/b) \leq \frac{\overline{L_n}}{n} < \sum_B P(b)\,H(A/b) + \frac{1}{n}} \qquad (5\text{-}24)$$

que constituye la generalización del primer teorema de Shannon. Hay que destacar la semejanza de (5-24) y (4-15*a*). Aumentando n, puede hacerse I_n/n tan próximo a

$$\sum_B P(b)\,H(A/b) \qquad (5\text{-}25)$$

como se desee. $\overline{L_n}/n$ es el número medio de binits necesarios para codificar un símbolo del alfabeto A, dado el símbolo del alfabeto B, correspondiente.

En la ecuación (5-24) $\overline{L_n}/n$ está expresado en binits y $H(A/b_j)$ en bits; la generalización al caso en que $\overline{L_n}$ venga medida en símbolos

r-arios y $H(A/b_j)$ en unidades de información *r*-arias es sumamente sencilla.

Hasta aquí no se ha intentado simplificar

$$\sum_B P(b)\, H(A/b)$$

con objeto de resaltar el hecho de que se trata del valor medio de las entropías a posteriori. Definamos, ahora

$$\begin{aligned} H(A/B) &= \sum_B P(b)\, H(A/b) \\ &= \sum_B P(b) \sum_A P(a/b) \log \frac{1}{P(a/b)} \\ &= \sum_{A,B} P(a,b) \log \frac{1}{P(a/b)} \end{aligned} \tag{5-26}$$

$H(A/B)$ recibe el nombre de *equivocación* de A con respecto a B, o *equivocación* del canal. La ecuación (5-24) puede expresarse en función de la equivocación, en la forma siguiente

$$\boxed{\lim_{n\to\infty} \frac{\overline{L_n}}{n} = H(A/B)} \tag{5-27}$$

Se ha insistido tanto en destacar la semejanza entre la demostración de la relación (5-24) y la del primer teorema de Shannon que el lector puede muy bien no haber apreciado una diferencia fundamental. Los sucesivos símbolos de entrada a_i (o bloques de ellos) se codifican empleando códigos distintos para cada símbolo de salida (o bloques de símbolos) b_j. Aún cuando cada uno de los códigos es unívoco, no es en general cierto que una secuencia de palabras código de una secuencia determinada de códigos unívocos sea también unívocamente decodificable. No es suficiente, por lo tanto, seleccionar un conjunto de códigos *unívocos* cuyas palabras tengan longitudes que satisfagan la relación (5-20); los códigos deberán ser *instantáneos*. En suma, la ecuación (5-24) se aplica solamente a códigos instantáneos, mientras que el primer teorema de Shannon se hace indistintamente a cualquier código unívoco, instantáneo o no.

5-6. Información mutua.

Consideremos nuevamente un canal de información con r entradas y s salidas (figura 5-8).

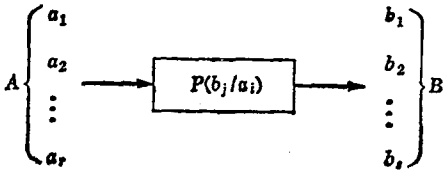


FIG. 5-8. Canal de información.

Seleccionando las entradas de acuerdo con las probabilidades $P(a_i)$, $i = 1, 2, \ldots, r$, la entropía del alfabeto de entrada será

$$H(A) = \sum_{A} P(a) \log \frac{1}{P(a)} \tag{5-28}$$

Conocidas las probabilidades de entrada y las probabilidades hacia adelante $P(b_j/a_i)$, pueden calcularse (apartado 5-3) las probabilidades hacia atrás $P(a_i/b_j)$, las probabilidades afines $P(a_i, b_j)$ y, finalmente, la equivocación

$$H(A/B) = \sum_{A,B} P(a,b) \log \frac{1}{P(a/b)} \tag{5-29}$$

Según el primer teorema de Shannon la determinación de un símbolo de entrada a_i exige una media de $H(A)$ binits. De acuerdo con la generalización del apartado 5-5, será solamente necesaria una media de $H(A/B)$ binits para definirlo, si *se puede conocer el símbolo de salida producido por esa entrada.* Es normal decir, en consecuencia, que, como media, la observación de un símbolo de salida proporciona $[H(A) - H(A/B)]$ bits de información. Esta diferencia se denomina *información mutua* (de A y B), o información mutua del canal. Se escribe

$$I(A\ ;\ B) = H(A) - H(A/B) \tag{5-30}$$

La información mutua puede expresarse de diferentes maneras,

$$I(A\,;\,B) = H(A) - H(A/B)$$

$$= \sum_{A} P(a) \log \frac{1}{P(a)} - \sum_{A,B} P(a,b) \log \frac{1}{P(a/b)}$$

$$= \sum_{A,B} P(a,b) \log \frac{1}{P(a)} - \sum_{A,B} P(a,b) \log \frac{1}{P(a/b)}$$

$$= \sum_{A,B} P(a,b) \log \frac{P(a/b)}{P(a)} \qquad (5\text{-}31a)$$

o, puesto que

$$P(a_i, b_j) = P(a_i/b_j)\, P(b_j)$$

$$I(A\,;\,B) = \sum_{A,B} P(a,b) \log \frac{P(a,b)}{P(a)\, P(b)} \qquad (5\text{-}31b)$$

La información mutua de la extensión de orden n de un canal se calcula a partir de la relación (5-31a). Si los símbolos de A^n se eligen de acuerdo con $P(\alpha_i) = P(a_{i_1})\, P(a_{i_2}) \ldots P(a_{i_n})$, la información mutua de la extensión de orden n es precisamente n veces la información mutua del canal original (problema 5-4):

$$I(A^n\,;\,B^n) = n\, I(A\,;\,B) \qquad (5\text{-}32)$$

5-7. Propiedades de la información mutua.

Se ha demostrado que la información mutua es el número medio de binits necesarios para determinar un símbolo de entrada antes de conocer el símbolo de salida correspondiente, menos el número medio de binits necesarios para especificar un símbolo de entrada después de conocer el símbolo de salida. Es decir,

$$I(A\,;\,B) = H(A) - H(A/B) \qquad (5\text{-}33)$$

Según esta interpretación se plantea inmediatamente la cuestión del signo de la información mutua. En el apartado 5-4 se indicó que $H(A) - H(A/b_j)$ puede tener un signo negativo; la entropía del alfabeto de entrada puede ser mayor cuando está determinado el símbolo de salida b_j. Sin embargo, la información mutua es el valor medio de $H(A) - H(A/b_j)$ (extendido a todos los símbolos de salida). ¿Puede

ser negativo? Para aclarar esta pregunta se escribirá (5-31*b*) de otra forma:

$$I(A;B) = \sum_{A,B} P(a,b) \log \frac{P(a,b)}{P(a)\,P(b)} \tag{5-31b}$$

Haciendo uso de la desigualdad (2-8*a*) encontramos

$$I(A;B) \geqq 0 \tag{5-34}$$

que será una igualdad cuando

$$P(a_i, b_j) = P(a_i)\,P(b_j) \text{ para cualquier } i,\ j \tag{5-35}$$

Esta conclusión es terminante. Dice que la información media recibida por un canal ha de ser siempre positiva. No se pierde en absoluto información por el hecho de observar la salida del canal. Además *la condición para que la información mutua sea nula es que los símbolos de entrada y salida sea estadísticamente independientes* (5-35).

De la relación (5-31*b*) puede deducirse otra importante propiedad de la información mutua. Esta ecuación, que puede interpretarse como definición de $I(A;\ B)$, es simétrica respecto a las variables a_i y b_j. Sustituyendo la entrada por la salida y viceversa, $I(A;\ B)$ no se altera. Por lo tanto, podrá escribirse

$$I(A;B) = I(B;A) \tag{5-36}$$

relación que pone de relieve la reciprocidad de la información mutua. Llevando más lejos este argumento, puede escribirse la ecuación (5-33) en la forma

$$I(A;B) = H(B) - H(B/A) \tag{5-37}$$

donde

$$H(B) = \sum_{B} P(b) \log \frac{1}{P(b)} \tag{5-38}$$

y

$$H(B/A) = \sum_{A,B} P(a,b) \log \frac{1}{P(b/a)} \tag{5-39}$$

Cantidad que se denomina equivocación de B con respecto a A.

Además de las entropías $H(A)$ y $H(B)$, puede definirse la entropía afín, que mide la incertidumbre del suceso simultáneo $(a_i,\ b_j)$. La pro-

babilidad de este suceso es $P(a_i, b_j)$, de modo que la entropía afín valdrá

$$H(A, B) = \sum_{A,B} P(a, b) \log \frac{1}{P(a, b)} \tag{5-40}$$

La relación entre $H(A, B)$ y $H(A)$ y $H(B)$ se deduce fácilmente

$$\begin{aligned} H(A, B) &= \sum_{A,B} P(a, b) \log \frac{P(a)\,P(b)}{P(a, b)} + \sum_{A,B} P(a, b) \log \frac{1}{P(a)\,P(b)} \\ &= -I(A; B) + \sum_{A,B} P(a, b) \log \frac{1}{P(a)} \\ &\qquad + \sum_{A,B} P(a, b) \log \frac{1}{P(b)} \\ &= -I(A; B) + \sum_{A} P(a) \log \frac{1}{P(a)} + \sum_{B} P(b) \log \frac{1}{P(b)} \\ &= H(A) + H(B) - I(A; B) \end{aligned} \tag{5-41}$$

La entropía afín de A y B, $H(A, B)$ es, como es lógico, simétrica respecto a A y B.

El diagrama de la figura 5-9 representa, en forma fácil de recordar, las diferentes relaciones deducidas hasta aquí.

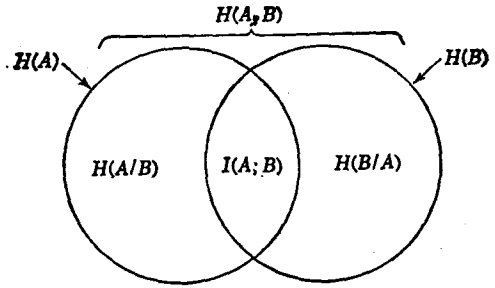


FIG. 5-9. Relaciones entre las diferentes magnitudes de un canal.

El círculo de la izquierda representa la entropía de A, y la de B el de la derecha. La zona común corresponde a la información mutua,

mientras que el resto de $H(A)$ y $H(B)$ representa las equivocaciones respectivas

$$H(A/B) = H(A) - I(A; B) \tag{5-42}$$

$$H(B/A) = H(B) - I(A; B) \tag{5-43}$$

La entropía afín $H(A, B)$ es la suma de $H(A)$ y $H(B)$, con la salvedad de que la zona común se ha incluido por partida doble, de modo que

$$H(A, B) = H(A) + H(B) - I(A; B) \tag{5-44}$$

También

$$H(A, B) = H(A) + H(B/A) \tag{5-45a}$$

y

$$H(A, B) = H(B) + H(A/B) \tag{5-45b}$$

Todas estas ecuaciones se deducen directamente a partir de (5-42), (5-43) y (5-44), o bien analizando la figura 5-9. Pueden interpretarse en el sentido de que la incertidumbre total de A y B es la suma de la incertidumbre de A más la de B una vez conocido A, o viceversa.

Finalmente, aun cuando nuestro interés se centra sobre los canales de información, es evidente que los argumentos empleados en este apartado no dependen del hecho de que A y B sean los alfabetos de entrada y salida de un canal de información. Las diferentes medidas de información expuestas en la figura 5-9 pueden aplicarse a dos conjuntos cualesquiera de variables. La información mutua tendrá signo positivo siempre que ambos conjuntos no sean estadísticamente independientes.

Ejemplo 5-4. Calcularemos la información mutua de un BSC. Su matriz es

$$\begin{bmatrix} p & \bar{p} \\ \bar{p} & p \end{bmatrix}$$

donde $\bar{p} = 1 - p$. Admitiendo que las probabilidades de transmitir un 0 y un 1 sean respectivamente ω y $\bar{\omega}$, la información mutua puede escribirse en la forma

$$\begin{aligned} I(A; B) &= H(B) - H(B/A) \\ &= H(B) - \sum_A P(a) \sum_B P(b/a) \log \frac{1}{P(a/b)} \\ &= H(B) - \sum_A P(a) \left(p \log \frac{1}{p} + \bar{p} \log \frac{1}{\bar{p}} \right) \\ &= H(B) - \left(p \log \frac{1}{p} + \bar{p} \log \frac{1}{\bar{p}} \right) \end{aligned} \tag{5-46}$$

Fácilmente se comprueba que las probabilidades de que $b_j = 0$ y $b_j = 1$ son iguales a $\omega\bar{p} + \bar{\omega}p$ y $\omega p + \bar{\omega}\bar{p}$, respectivamente. Por lo tanto,

$$I(A; B) = \left[(\omega\bar{p} + \bar{\omega}p) \log \frac{1}{\omega\bar{p} + \bar{\omega}p} + (\omega p + \bar{\omega}\bar{p}) \log \frac{1}{\omega p + \bar{\omega}\bar{p}} \right] - \left(p \log \frac{1}{p} + \bar{p} \log \frac{1}{\bar{p}} \right) \quad (5\text{-}47)$$

$I(A; B)$ se expresa como sigue, en función de la entropía (figura 2-3)

$$I(A; B) = H(\omega p + \bar{\omega}\bar{p}) - H(p) \quad (5\text{-}48)$$

Ecuación que tiene una sencilla interpretación geométrica. Puesto que $\omega p + \bar{\omega}p$ debe estar comprendido entre p y $\bar{p}$, $H(\omega p + \bar{\omega}p) \geqslant H(p)$, con lo que la figura 5-10 prueba que la información mutua debe ser positiva.

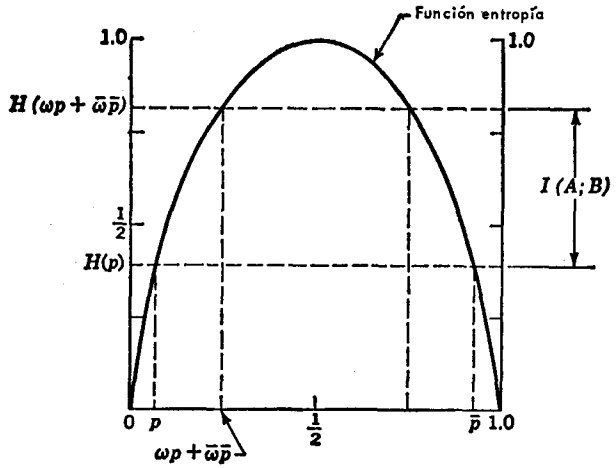


FIG. 5-10. Interpretación geométrica de la información mutua de un BSC.

La figura 5-10 permite apreciar ciertas condiciones límites interesantes. Por ejemplo, para un valor constante de p, puede observarse la conducta de $I(A; B)$ al variar ω. $I(A; B)$ alcanza un valor máximo para $\omega = 1/2$, siendo éste $1 - H(p)$. Por otra parte, la información mutua se anula para $\omega = 0$ y $\omega = 1$.

5-8. Canales sin ruido y canales determinantes.

En este apartado se definirán dos tipos de canales especiales, deduciendo las expresiones simplificadas de sus informaciones mutuas. Admitamos que al menos un elemento de cada columna de la matriz

del canal es distinta de cero. La probabilidad de que aparezca un símbolo de salida correspondiente a una columna de ceros es nula, independientemente de la distribución de símbolos de entrada. En consecuencia, el caso no presenta ningún interés y puede ser ignorado.

> *Definición.* Un canal definido por una matriz con un elemento, y solamente uno, distinto de cero en cada columna se denomina *canal sin ruido.*

Ejemplo 5-5. La matriz de un canal sin ruido es

$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/5 & 3/10 & 1/10 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

La figura 5-11 representa el diagrama de este canal.

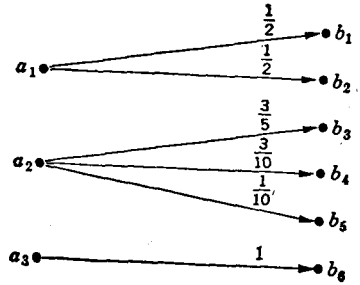


FIG. 5-11. Canal sin ruidos.

Un BSC en que la probabilidad p es igual a 0, es un canal sin ruido. Nótese, sin embargo, que un BSC cuya probabilidad de error es igual a la unidad, es también un canal sin ruido. Esto constituye la expresión del hecho de que un canal de este tipo coherente en el error es tan eficaz como un canal coherentemente correcto.

> *Definición.* Un canal definido por una matriz con un elemento, y solo uno, distinto de cero en cada fila, recibe el nombre de *canal determinante.*

Ejemplo 5-6. La matriz de un canal determinante es

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

El diagrama de este canal es el mostrado en la figura 5-12.

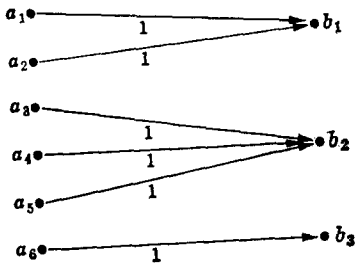


FIG. 5-12. Canal determinante.

Puesto que no hay más que un elemento distinto de cero en cada fila de la matriz de un canal determinante, y la suma de los de cada fila es igual a la unidad, los elementos son exclusivamente 0 y 1.

La información mutua de los canales definidos puede calcularse fácilmente. Consideremos, en primer lugar, un canal sin ruido. En este caso, al observar una salida b_j, se conoce con certeza el símbolo a_i transmitido, es decir las probabilidades condicionales $P(a_i/b_j)$ son 0 y 1. La equivocación $H(A/B)$ puede escribirse en la forma

$$H(A/B) = \sum_B P(b_j) \sum_A P(a_i/b_j) \log \frac{1}{P(a_i/b_j)} \qquad (5\text{-}49)$$

donde todos los términos del último sumando son nulos (bien $1 \times \log 1$ ó $0 \times \log 1/0$). Por tanto, en un canal sin ruido

$$H(A/B) = 0 \qquad (5\text{-}50)$$

Esta conclusión es también evidente si se considera la generalización del primer teorema de Shannon (apartado 5-5). Las salidas de un canal sin ruido son suficientes por sí mismas para determinar las en-

tradas del canal. Por lo tanto, el número medio del binits necesarios para definir la entrada, una vez conocida la salida, es nulo. Según (5-30), en un canal sin ruido se verifica que

$$I(A;B) = H(A) \tag{5-51}$$

La cantidad de información transmitida por este canal es igual a la incertidumbre total del alfabeto de entrada.

En los canales determinantes puede llegarse a una serie de conclusiones análogas. Efectivamente, el símbolo de *entrada* a_i es suficiente para determinar, con probabilidad 1, el símbolo de *salida* b_j. Por lo tanto las probabilidades $P(b_j/a_i)$ han de ser 0 ó 1, y

$$H(B/A) = \sum_A P(a_i) \sum_B P(b_j/a_i) \log \frac{1}{P(b_j/a_i)}$$

$$= 0 \tag{5-52}$$

O, introduciendo la relación (5-37),

$$I(A;B) = H(B) \tag{5-53}$$

5-9. Canales en serie.

El análisis de dos canales en serie pone de relieve algunas propiedades interesantes de la entropía y la información mutua (figura 5-13) [Silverman realizó (1955) una detallada investigación de las series de canales binarios].

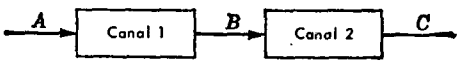


FIG. 5-13. Dos canales en serie.

Consideremos un canal con un alfabeto de entrada A de r símbolos y un alfabeto de salida B de s símbolos, conectado en serie con un segundo canal, como indica la figura anterior. El alfabeto de entrada de este segundo canal es idéntico a B y el de salida, de t símbolos se reconoce como C.

El hecho de conectarlos en serie implica ciertas relaciones entre las probabilidades. Cuando se transmite a_i, un símbolo de A, la salida

del primer canal es un símbolo de B, digamos b_j. A su vez b_j da lugar a una salida, c_k en el segundo canal. El símbolo c_k depende de la entrada a_i a través de b_j. Realmente, conocido el símbolo intermedio b_j, la probabilidad de obtener c_k depende solamente de b_j, y no del símbolo inicial a_i que dio lugar a b_j. Esta propiedad puede expresarse como

$$P(c_k/b_j, a_i) = P(c_k/b_j) \quad \text{para cualquier } i, j, k \tag{5-54}$$

La relación (5-54) define el significado de una serie de dos canales. La aplicación de la regla de Bayes a (5-54) da lugar a una ecuación semejante, en sentido inverso:

$$P(a_i/b_j, c_k) = P(a_i/b_j) \tag{5-55}$$

Hay que destacar que las relaciones (5-54) y (5-55) se cumplen únicamente cuando A, B y C son los alfabetos de dos canales en serie, conectados tal como indica en la figura 5-13.

Al transmitir una información a través de dos canales en serie parece lógico que la equivocación aumente, es decir que $H(A/C)$ sea mayor que $H(A/B)$. Intentaremos demostrarlo a continuación

$$\begin{aligned} H(A/C) - H(A/B) &= \sum_{A,C} P(a,c) \log \frac{1}{P(a/c)} \\ &\quad - \sum_{A,B} P(a,b) \log \frac{1}{P(a/b)} \\ &= \sum_{A,B,C} P(a,b,c) \log \frac{1}{P(a/c)} \\ &\quad - \sum_{A,B,C} P(a,b,c) \log \frac{1}{P(a/b)} \\ &= \sum_{A,B,C} P(a,b,c) \log \frac{P(a/b)}{P(a/c)} \end{aligned} \tag{5-56}$$

Sustituyendo (5-55) en (5-56), resulta

$$\begin{aligned} H(A/C) - H(A/B) &= \sum_{A,B,C} P(a,b,c) \log \frac{P(a/b,c)}{P(a/c)} \\ &= \sum_{B,C} P(b,c) \sum_{A} P(a/b,c) \log \frac{P(a/b,c)}{P(a/c)} \end{aligned} \tag{5-57}$$

La inecuación (2-8*a*) demuestra que la suma de todos los términos de (5-57), extendida al alfabeto A, es positiva. Por lo tanto,

$$H(A/C) - H(A/B) \geqq 0 \tag{5-58}$$

o

$$H(A/C) \geqq H(A/B) \tag{5-59}$$

Una consecuencia inmediata de esta relación es

$$I(A; B) \geqq I(A; C) \tag{5-60}$$

Woodward fue quien primero probó (1955) estas interesantes desigualdades, que demuestran que los canales tienden a «perder» información. La información que emerge finalmente de varios canales en serie no puede ser mayor que la que emergía de un punto intermedio de la serie, si se pudiera extraer de él.

La condición que define la igualdad de (5-59) y (5-60) presenta cierto interés. Repasando la demostración, vemos que corresponde a

$$P(a/b, c) = P(a/c) \tag{5-61a}$$

para cualquier símbolo a, b y c, siempre que $P(b, c) \neq 0$. Esta condición puede escribirse también en la forma

$$P(a/b) = P(a/c) \tag{5-61b}$$

aplicable a cualquier a, b y c, siempre que $P(b, c) \neq 0$.

La condición de igualdad merece algún comentario. A primera vista parece cumplirse solamente si el segundo canal de la figura 5-13 fuese sin ruidos. En caso contrario, es difícil demostrar la aplicación de la relación (5-61 *b*). Sin embargo, como se verá en el ejemplo siguiente, puede hacerse también en otras circunstancias.

Ejemplo 5-7. Consideremos el canal

$$\begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \end{bmatrix}$$

en serie con un segundo canal

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{bmatrix}$$

La figura 5-14 representa un diagrama del canal resultante.

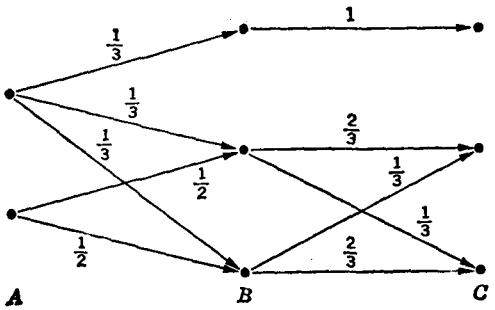


FIG. 5-14. Canales en serie.

Como puede apreciarse, la relación (5-61*b*) se cumple a pesar de ser dos dos canales con ruido, por lo tanto

$$I(A; B) = I(A; C)$$

En este ejemplo, (5-61*b*) se cumple cualquiera que sean las probabilidades asociadas al alfabeto *A*. Existen otros casos en que se aplica era relación únicamente para una distribución de entrada particular. Entraremos en esta cuestión en el siguiente apartado.

Un conjunto de personas, imaginadas como canales, constituye un ejemplo significativo de la pérdida que sufre la información al progresar a través de varios canales en serie. Un mensaje, escrito originalmente en inglés, se traduce a otra lengua y de nuevo al inglés, por un traductor que no conoce el mensaje original. El resultado de este proceso será una versión degenerada que puede considerarse como el mensaje emitido por un canal con ruido. Con objeto de simular una serie de canales repetiremos la operación, empezando, esta vez, por la versión degenerada.

Esta experiencia se realizó con un poema de cuatro líneas de Ogden Nash, «La tortuga». El poema se tradujo sucesivamente del inglés al francés, al inglés, al alemán, al inglés, al español y finalmente al inglés. No se intentó conservar la rima ni el metro de la obra original *.

The turtle lives 'twixt plated decks
Which practically conceal its sex.
I think it clever of the turtle
In such a fix to be so fertile.

* *N. del T.*: Es realmente complicado, y tal vez poco significativo, efectuar la traducción al español, conservando los matices que distinguen las sucesi-

La salida del canal inglés-francés-inglés fue

> The turtle lives in a scaled caparace which in fact hides its sex. I find that it is clever for the turtle to be so fertile in such a tricky situation.

Análogamente, la del canal inglés-alemán-inglés

> The turtle lives in a enclosed shell under which, in reality, it hides its sex. I find that the turtle must be very clever, indeed, to be so fertile in such a tight situation.

Finalmente, la salida del canal inglés-español-inglés

> The turtle lives inside a closed shell, under which, really, it hides its sex. I fell the turtle had to be certainly clever to be so fertile in a so tight situation.

La falta de ruido de los canales de comunicación humanos y la pérdida de información que introducen ha sido reconocida desde hace largo tiempo. Tucídides, en el libro I de «La guerra del Peloponeso», dice:

> De los sucesos de una guerra no me aventuro a hablar basándome en cualquier información, ni en mi propia opinión [es decir, probabilidades a priori]; no he descrito nada que yo mismo no haya presenciado o recogido de otros después de una cuidadosa encuesta [es decir, canales sin ruido]. La labor fue trabajosa, pues testigos del mismo acontecimiento lo narran de modo distinto, según su memoria o el bando en que hubieran participado [es decir, canales con ruido].

Como ejemplo final (más cuantitativo) de las pérdidas existentes en una serie de canales, consideraremos dos BSC idénticos.

Ejemplo 5-8. Dos canales BSC, de matriz

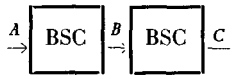
$$\begin{bmatrix} \bar{p} & p \\ p & \bar{p} \end{bmatrix}$$

---

vas versiones. Por esta razón se ha preferido mantener la versión inglesa del texto original, dejando al lector el cuidado de su apreciación.

> La tortuga vive en una coraza
> que prácticamente encubre su sexo.
> Creo que la tortuga es muy inteligente
> Para ser tan fértil en tan curiosa situación.

se conectan en la forma siguiente:



Las dos posibilidades de entrada del primer canal se eligen con la misma probabilidad. Así pues, de (5-48), tendremos

$$I(A; B) = 1 - H(p) \tag{5-62}$$

Es fácil demostrar que la combinación de estos dos canales en serie es equivalente a un solo BSC con probabilidad de error $2p\bar{p}$. Por lo tanto,

$$I(A; C) = 1 - H(2p\bar{p}) \tag{5-63}$$

Si se añadiera un tercer BSC (de alfabeto de salida D), obtendríamos

$$I(A; D) = 1 - H(3\bar{p}^2p + p^3) \tag{5-64}$$

La figura 5-15 representa estas curvas.

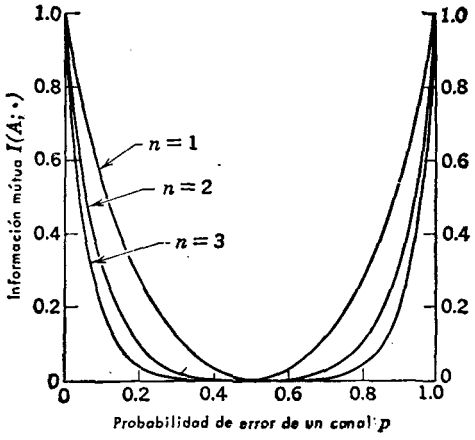


FIG. 5-15. Información mutua de una serie de n BSCs. (Los símbolos de entrada se suponen equiprobables.)

5-10. Canales reducidos y reducciones suficientes.

En la mayor parte de los canales encontrados en la vida real el conjunto de salidas es mayor de lo que sería de desear. Por ejemplo, los datos de carácter científico transmitidos por un satélite por vía de

un canal telemétrico binario contienen información que no tiene ningún significado en relación con el fenómeno sometido a observación. La antena terrestre de tal sistema recibe una secuencia de impulsos de diferentes amplitudes. El receptor toma cada impulso y, si su amplitud es superior a un valor humbral, lo interpreta como un «1». En caso contrario, si es inferior, como un «0». Pueden imaginarse dos canales distintos. En primer lugar, un canal de entradas binarias (el conjunto transmitido por el satélite) y un gran número de salidas (tantas como amplitudes puede distinguir el receptor). El segundo, un canal de entradas y salidas binarias (salidas correspondientes a las del receptor). Este último canal es una simplificación del primero, por lo que recibe el nombre de *reducción* suya.

Definición. Sea un canal de r entradas y s salidas, definido por la matriz **P**.

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1i} & P_{1,i+1} & \dots & P_{1s} \\ P_{21} & P_{22} & \dots & P_{2i} & P_{2,i+1} & \dots & P_{2s} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ P_{r1} & P_{r2} & \dots & P_{ri} & P_{r,i+1} & \dots & P_{rs} \end{bmatrix}$$

Se define un nuevo canal de r entradas y $s-1$ salidas asociando y sumando dos de las columnas de **P**. La matriz del nuevo canal es **P'**.

$$\mathbf{P}' = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1i} & P_{1,i+1} & \dots & P_{1s} \\ P_{21} & P_{22} & \dots & P_{2i} & P_{2,i+1} & \dots & P_{2s} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ P_{r1} & P_{r2} & \dots & P_{ri} & P_{r,i+1} & \dots & P_{rs} \end{bmatrix}$$

El nuevo canal es una *reducción elemental* de **P**. El proceso puede repetirse un cierto número de veces, formando la reducción elemental de **P**, etc. El canal resultante, después de efectuada más de una reducción elemental, recibe simplemente el nombre de *reducción* del canal original **P**.

Ejemplo 5-9. En el ejemplo 5-1 se formó la matriz de $(BSC)^2$

$$\mathbf{P} = \begin{bmatrix} \bar{p}^2 & \bar{p}p & p\bar{p} & p^2 \\ \bar{p}p & \bar{p}^2 & \bar{p}p & \bar{p}^2 \\ p\bar{p} & p^2 & p^2 & p\bar{p} \\ p^2 & p\bar{p} & \bar{p}^2 & \bar{p}p \end{bmatrix}$$

Una de las reducciones elementales de **P** se obtiene sumando la primera y segunda columnas

$$\mathbf{P}' = \begin{bmatrix} \bar{p} & p\bar{p} & p^2 \\ \bar{p} & p^2 & p\bar{p} \\ p & \bar{p}^2 & \bar{p}p \\ p & \bar{p}p & \bar{p}^2 \end{bmatrix}$$

Una reducción de **P** se obtendría sumando las columnas dos y tres de **P'**:

$$\mathbf{P}'' = \begin{bmatrix} \bar{p} & p \\ \bar{p} & p \\ p & \bar{p} \\ p & \bar{p} \end{bmatrix}$$

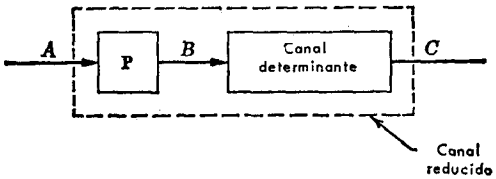


FIG. 5-16. Un canal reducido.

La figura 5-16 muestra una representación muy interesante de un canal reducido. El canal determinante combina los símbolos del alfabeto *B* formando un número menor pertenecientes a un alfabeto *C*. Así, pues, el canal de alfabeto de entrada *A* y de salida *C*, marcado por una línea de trazos, constituye una reducción del canal **P**. Este método de representación permite aplicar los resultados obtenidos en el apartado anterior, que trataba de canales en serie, a los canales reducidos. En particular, tendremos (en relación con la figura 5-16)

$$H(A/C) \geq H(A/B) \tag{5-65}$$

y

$$I(A; C) \leq I(A; B) \tag{5-66}$$

La reducción de un canal disminuye (o a lo sumo mantiene constante) la información mutua entre los alfabetos de entrada y salida. Es el precio que hay que pagar por su simplificación.

Las observaciones anteriores sugieren una cuestión más importante. «¿En qué condiciones puede simplificarse un canal sin que cueste nada

el hacerlo?» Es decir, «¿Cuándo la información mutua de un canal reducido es igual a la del original?» Para responder bastará considerar el caso de reducciones elementales, extendiendo los resultados al caso general por el método de inducción.

Formemos una partición elemental del canal

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1s} \\ P_{21} & P_{22} & \dots & P_{2s} \\ \dots & \dots & \dots & \dots \\ P_{r1} & P_{r2} & \dots & P_{rs} \end{bmatrix} \tag{5-67}$$

Sin pérdida de generalidad puede suponerse que la partición elemental se ha formado combinando las dos primeras columnas de **P**, situación representada en la figura 5-17.

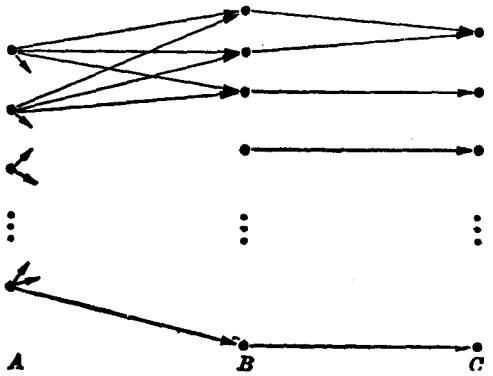


FIG. 5-17. Reducción de un canal mediante una serie.

El apartado 5-9 enunció la condición necesaria y suficiente para que una serie de canales no perdiera información. Esta condición era [(5-61 *b*)].

$$P(a/b) = P(a/c) \tag{5-68}$$

para todos los símbolos *a*, *b*, y *c*, tales que

$$P(b, c) \neq 0$$

En el caso de una reducción elemental esta condición se cumple para todos los símbolos de *B*, excepto los dos que se han combinado

b_1 y b_2. Sea c_1 el símbolo de C formado con b_1 y b_2. Aplicando la relación (5-68) a b_1 y b_2 se encuentra, como condición necesaria y suficiente

$$P(a/b_1) = P(a/c_1) = P(a/b_2) \quad \text{para cualquier } a \tag{5-69}$$

Que es equivalente a *

$$P(a/b_1) = P(a/b_2) \quad \text{para cualquier } a \tag{5-70}$$

En otras palabras, *los símbolos de salida b_1 y b_2 se combinan sin pérdida de información solamente si las probabilidades hacia atrás, $P(a/b_1)$ y $P(a/b_2)$ son iguales para cualquier valor de a.* Esta conclusión reviste gran importancia, tanto para ayudar a comprender mejor el concepto de información, como desde un punto de vista práctico. Define en qué condiciones puede simplificarse un canal sin pérdida de información. Ahora bien, las probabilidades hacia atrás dependen de las probabilidades a priori $P(a_i)$; es decir, son función de la forma en que se utilice el canal. Presenta aún mayor interés determinar cuándo pueden combinarse las salidas de un canal independientemente de su utilización; es decir, para un conjunto cualquiera de probabilidades a priori. Aplicando la ley de Bayes, la expresión (5-70) puede escribirse como sigue

$$\frac{P(b_1/a)\,P(a)}{\sum_A P(b_1/a)\,P(a)} = \frac{P(b_2/a)\,P(a)}{\sum_A P(b_2/a)\,P(a)} \quad \text{para cualquier } a \tag{5-71}$$

o

$$\frac{P(b_1/a)}{P(b_2/a)} = \frac{\sum_A P(b_1/a)\,P(a)}{\sum_A P(b_2/a)\,P(a)} \quad \text{para cualquier } a \tag{5-72}$$

Si esta relación se cumple para todas las probabilidades a priori posibles, $P(a)$, tendremos

$$P(b_1/a) = \text{const} \times P(P(b_2/a)) \quad \text{para cualquier } a \tag{5-73}$$

que es la condición buscada. Si la matriz de un canal satisface la relación (5-73), dos cualquiera de sus columnas pueden combinarse, obteniendo una matriz tan buena como la anterior. Más concretamente, cualquiera que sean las probabilidades del alfabeto de entrada, las in-

* La condición correspondiente a $P(a/c_1)$ se deduce automáticamente de (5-70).

formaciones mutuas de un canal y el reducido serían idénticas. Un canal reducido con esta propiedad se denomina *reducción suficiente.*

Ejemplo 5-10. El canal

$$\begin{bmatrix} 1/6 & 1/3 & 1/2 & 0 \\ 1/12 & 1/6 & 1/4 & 1/2 \end{bmatrix}$$

se reduce a

$$\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$$

y finalmente a

$$\begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}$$

Este último canal es una reducción suficiente del canal original.

5-11. Propiedad asociativa de la información mutua.

La *asociabilidad* es otra importante propiedad de la información mutua. La estudiaremos en este apartado, considerando la cantidad media de información suministrada por una sucesión de símbolos de salida de un canal que tiene un conjunto de símbolos de entrada bien definido. Esto es, se estudiará la ganancia de información obtenida al considerar varias observaciones en lugar de una sola. Un ejemplo típico de esta situación lo constituye un canal con ruidos en el que los símbolos de entrada se repiten un cierto número de veces. Tal procedimiento aumenta la confiabilidad del mensaje transmitido a través de un canal no confiable. Otro ejemplo interesante sería un canal donde la respuesta que corresponde a un símbolo de entrada es una secuencia de símbolo de salida, en lugar de uno solo.

Estudiaremos la propiedad aditiva de la información mutua en el caso en que a un símbolo de entrada corresponden dos de salida. El caso general, n símbolos de salida por uno de entrada, puede tratarse seguidamente por inducción.

El modelo de canal de información definido tendrá que modificarse, de modo que a un símbolo de entrada correspondan dos de salida, b_j y c_k. Estos símbolos pertenecen a los alfabetos $B = \{b_j\}$, $j = 1, 2, \ldots, s$, y $C = \{c_k\}$, $k = 1, 2, \ldots, t$.

Sin perder generalidad puede admitirse que los símbolos se reciben en el orden b_j, ck. Las probabilidades a priori de los símbolos de entrada, $P(a_i)$, se transforman, entonces, después de recibido el primer símbolo de salida, en las probabilidades a posteriori $P(a_i/b_j)$; una vez recibido el segundo se convierten en las probabilidades «aún más a posteriori» $P(a_i/b_j, c_k)$.

Al recibir los símbolos b_j y c_k, la incertidumbre o entropía del conjunto de símbolos de entrada pasa de

$$H(A) = \sum_{A} P(a) \log \frac{1}{P(a)} \tag{5-74a}$$

a la entropía a posteriori

$$H(A/b_j) = \sum_{A} P(a/b_j) \log \frac{1}{P(a/b_j)} \tag{5-74b}$$

y a la entropía «aún más a posteriori»

$$H(A/b_j, c_k) = \sum_{A} P(a/b_j, c_k) \log \frac{1}{P(a/b_j, c_k)} \tag{5-74c}$$

Como en el apartado 5-5, se calculará el valor medio de $H(A/b_j)$ extendido a b_j, para encontrar el valor de la entropía media a posteriori, o equivocación de A respecto a B:

$$\sum_{B} P(b) H(A/b) = H(A/B) \tag{5-75a}$$

Del mismo modo, para encontrar la equivocación de A respecto a B y C, se calcula la media de $H(A/b_j, c_k)$ extendida a

$$\sum_{B,C} P(b, c) H(A/b, c) = H(A/B, C) \tag{5-75b}$$

Según la generalización del primer teorema de Shannon (apartado 5-5), $H(A/B, C)$ es el número medio de binits necesario para codificar un símbolo del alfabeto A después de conocidos los símbolos de B y C correspondientes.

Las ecuaciones (5-75*a*) y (5-75*b*) sugieren dos procedimientos diferentes para medir la información mutua de (B, C) y A. El primero consiste en definir la información mutua exactamente de la misma manera que en un canal cuya salida es un solo símbolo. Es decir,

$$I(A; B, C) = H(A) - H(A/B, C) \tag{5-76}$$

En segundo lugar, puede estimarse la cantidad de información que sobre A suministra B, y después la que suministra C *después de conocido* B. Estas cantidades son

$$H(A) - H(A/B) \tag{5-77a}$$

y

$$H(A/B) - H(A/B, C) \tag{5-77b}$$

La primera de las cuales ha sido ya definida como

$$I(A; B) = H(A) - H(A/B) \tag{5-78a}$$

La relación (5-77*b*) puede definirse también como

$$I(A; C/B) = H(A/B) - H(A/B, C) \tag{5-78b}$$

llamada información mutua de A y C, conocido B. Sumando (5-78*a*) y (5-78*b*), se encuentra

$$\begin{aligned} I(A; B) + I(A; C/B) &= H(A) - H(A/B, C) \\ &= I(A; B, C) \end{aligned} \tag{5-79}$$

Ecuación que expresa la propiedad asociativa de la información mutua. Dice que la cantidad media de información obtenida en una observación no depende de que sea un todo o esté descompuesto en varias partes. La ecuación puede generalizarse

$$\begin{aligned} I(A; B, C, \ldots, D) = I(A; B) + I(A; C/B) + \ldots \\ + I(A; D/B, C, \ldots) \end{aligned} \tag{5-80}$$

El primer término es la cantidad media de información que sobre A suministra una observación de los alfabetos $B, C, \ldots, D$. El primer término del segundo miembro es la cantidad media de información debida a la observación del alfabeto B. El segundo término, la cantidad media de información debida a la observación de C *después* de haber observado el alfabeto B, etc. ...

El orden en que se recibe la información es indiferente. Podía haberse escrito [correspondiendo a (5-79)], por ejemplo.

$$I(A; B, C) = I(A; C) + I(A; B/C) \tag{5-81}$$

Las cantidades de información calculadas anteriormente pueden expresarse de forma diferente. De (5-76), se deduce

$$I(A;\ B,\ C) = H(A) - H(A/B,\ C)$$

$$= \sum_{A} P(a) \log \frac{1}{P(a)} - \sum_{A,B,C} P(a, b, c) \log \frac{1}{P(a/b, c)}$$

$$= \sum_{A,B,C} P(a, b, c) \log \frac{1}{P(a)} - \sum_{A,B,C} P(a, b, c) \log \frac{1}{P(a/b, c)}$$

$$= \sum_{A,B,C} P(a, b, c) \log \frac{P(a/b, c)}{P(a)} \qquad (5\text{-}82a)$$

Multiplicando numerador y denominador del logaritmo de la ecuación anterior por $P(b, c)$, se obtiene otra interesante expresión

$$I(A;\ B,\ C) = \sum_{A,B,C} P(a, b, c) \log \frac{P(a, b, c)}{P(a)\, P(b, c)} \qquad (5\text{-}82b)$$

Hay que hacer notar al lector la semejanza de (5-82*a*) y (5-82*b*) con (5-31*a*) y (5-31*b*), que serían iguales sin más que reemplazar b por (b, c). Esta semejanza sugiere la definición

$$H(B,\ C/A) = \sum_{A,B,C} P(a, b, c) \log \frac{1}{P(b, c/a)} \qquad (5\text{-}83)$$

Se comprueba fácilmente que

$$I(A;\ B, C) = H(B, C) - H(B, C/A) \qquad (5\text{-}84)$$

Ejemplo 5-11. Sea el BSC

$$\begin{bmatrix} p & \bar{p} \\ \bar{p} & p \end{bmatrix}$$

donde $\bar{p} = 1 - p$. Para comprobar la asociabilidad de la información mutua se supondrá que el símbolo de entrada (0 ó 1) se repite, de manera que la salida del canal consta de dos símbolos binarios, b_j, c_k, por cada símbolo a_i de entrada. Para mayor sencillez se supondrá que las probabilidades de las dos entradas son iguales. Por lo tanto, haciendo $\omega = 1/2$ en (5-48), se encuentra

$$I(A; B) = 1 - H(p) \qquad (5\text{-}85)$$

La ecuación (5-82*b*) permite calcular $I(A; B, C)$. La tabla 5-2 contiene las

TABLA 5-2. PROBABILIDADES DE UN BSC REPETITIVO

a_i	b_j	c_k	$P(a_i)$	$P(a_i, b_j, c_k)$	$P(b_j, c_k)$
0	0	0	1/2	$1/2\,\bar{p}$	$p^2 + \bar{p}^2$
0	0	1	1/2	$1/2\,p\bar{p}$	$p\bar{p}$
0	1	0	1/2	$1/2\,p\bar{p}$	$p\bar{p}$
0	1	1	1/2	$1/2\,p^2$	$p^2 + \bar{p}^2$
1	0	0	1/2	$1/2\,p^2$	$p^2 + \bar{p}^2$
1	0	1	1/2	$1/2\,p\bar{p}$	$p\bar{p}$
1	1	0	1/2	$1/2\,p\bar{p}$	$p\bar{p}$
1	1	1	1/2	$1/2\,\bar{p}$	$p^2 + \bar{p}^2$

probabilidades necesarias. Sustituyéndolas en (5-82*b*), se obtiene

$$I(A; B, C) = p^2 \log \frac{2p^2}{p^2 + \bar{p}^2} + \bar{p}^2 \log \frac{2\bar{p}^2}{p^2 + \bar{p}^2}$$

$$= (p^2 + \bar{p}^2) \left[1 - H \left(\frac{p^2}{p^2 + \bar{p}^2} \right) \right] \tag{5-86}$$

La interpretación de (5-86) es inmediata. Al encontrar las salidas 10 ó 01, el significado es ambiguo; las dos entradas son igualmente probables y la observación no habrá añadido ninguna información. Si, por el contrario, se encuentra 00 ó 11, la información obtenida sobre la entrada equivale a la que se habría obtenido observando una sola salida de un BSC de probabilidad de error

$$\frac{p^2}{p^2 + \bar{p}^2}$$

Según (5-85), la información correspondiente a esa observación es

$$1 - H \left(\frac{p^2}{p^2 + \bar{p}^2} \right) \tag{5-87}$$

La probabilidad de encontrar 00 ó 11 es $p^2 + \bar{p}^2$; luego, de ahí, se deduce (5-86).

Estos argumentos pueden generalizarse fácilmente al caso de un BSC utilizado con más de una repetición. Por ejemplo, si cada entrada da lugar a tres salidas binarias, se encuentra

$$I(A; B, C, D) = (p^3 + \bar{p}^3) \left[1 - H \left(\frac{p^3}{p^3 + \bar{p}^3} \right) \right] + 3p\bar{p}[1 - H(p)] \tag{5-88}$$

Las curvas que representan las ecuaciones (5-85) aparecen en la figura 5-18

5-12. Información mutua de alfabetos diferentes

El estudio realizado en el apartado anterior sobre la propiedad asociativa de la información mutua dio lugar a la consideración de la secuencia de entropías.

$$\begin{aligned} &H(A) \\ &H(A/B) \\ &H(A/B, C) \\ &\cdot \\ &\cdot \\ &\cdot \end{aligned} \tag{5-89}$$

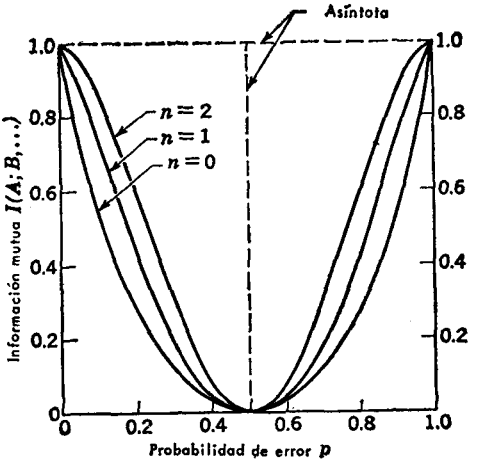


FIG. 5-18. Información mutua de un BSC con n repeticiones.

Cada término no es mayor que el precedente. La diferencia entre dos consecutivos podía interpretarse como la información media suministrada por una nueva observación.

$$I(A; B) = H(A) - H(A/B) \tag{5-90a}$$

$$I(A; C/B) = H(A/B) - H(A/B, C) \tag{5-90b}$$

.....

$I(A; B)$ es la información mutua de A y B; $I(A; C/B)$ la información mutua de A y C, después de conocido B. Ambas cantidades,

sin embargo, se refieren a la información mutua entre dos únicos alfabetos, pudiendo extenderse el concepto a un número mayor (McGill, 1954). La información mutua entre A, B y C se define por

$$I(A;\ B;\ C) = I(A;\ B) - I(A;\ B/C) \tag{5-91a}$$

Esta definición implica que la expresión de $I(A;\ B;\ C)$ sea simétrica respecto a A, B y C. Si es así, (5-91a) puede escribirse también en la forma

$$I(A;\ B;\ C) = I(B;\ C) - I(B;\ C/A) \tag{5-91b}$$

$$= I(C;\ A) - I(C;\ A/B) \tag{5-91c}$$

La simetría de (5-91a) se demuestra como sigue:

$$\begin{aligned} I(A;\ B;C) &= \sum_{A,B} P(a,b) \log \frac{P(a,b)}{P(a)\,P(b)} \\ &\quad - \sum_{A,B,C} P(a,b,c) \log \frac{P(a,b/c)}{P(a/c)\,P(b/c)} \\ &= \sum_{A,B,C} P(a,b,c) \log \frac{P(a,b)\,P(a/c)\,P(b/c)}{P(a)\,P(b)\,P(a,b/c)} \\ &= \sum_{A,B,C} P(a,b,c) \log \frac{P(a,b)\,P(a,c)\,P(b,c)}{P(a)\,P(b)\,P(c)\,P(a,b,c)} \\ &= H(A,B,C) - H(A,B) - H(A,C) \\ &\quad - H(B,C) + H(A) + H(B) + H(C) \end{aligned} \tag{5-92}$$

La expresión final es simétrica, como quería demostrarse, recordando además la expresión correspondiente a la información mutua de dos alfabetos:

$$I(A;\ B) = H(A,B) - H(A) - H(B) \tag{5-93}$$

La generalización a más de tres alfabetos se lleva a cabo fácilmente. La información mutua de A, B, C y D, por ejemplo, es

$$\begin{aligned} I(A;\ B;\ C;\ D) &= I(A;\ B;\ C) - I(A;\ B;\ C/D) \\ &= [H(A,\ B,\ C,\ D)] - [H(A,\ B,\ C) \\ &\quad + H(A,\ B,\ D) + H(A,\ C,\ D) \\ &\quad + H(B,\ C,\ D)] + [H(A,\ B) + H(A,\ C) \\ &\quad + H(A,\ D) + H(B,\ C) + H(B,\ D) \\ &\quad + H(C,\ D)] - [H(A) + H(B) \\ &\quad + H(C) + H(D)] \end{aligned} \tag{5-94}$$

Blachman (1961) sugirió la posibilidad de generalizar la representación de la figura 5-9 para interpretar las expresiones anteriores. La figura 5-19 muestra las relaciones existentes en el caso de tres alfabetos.

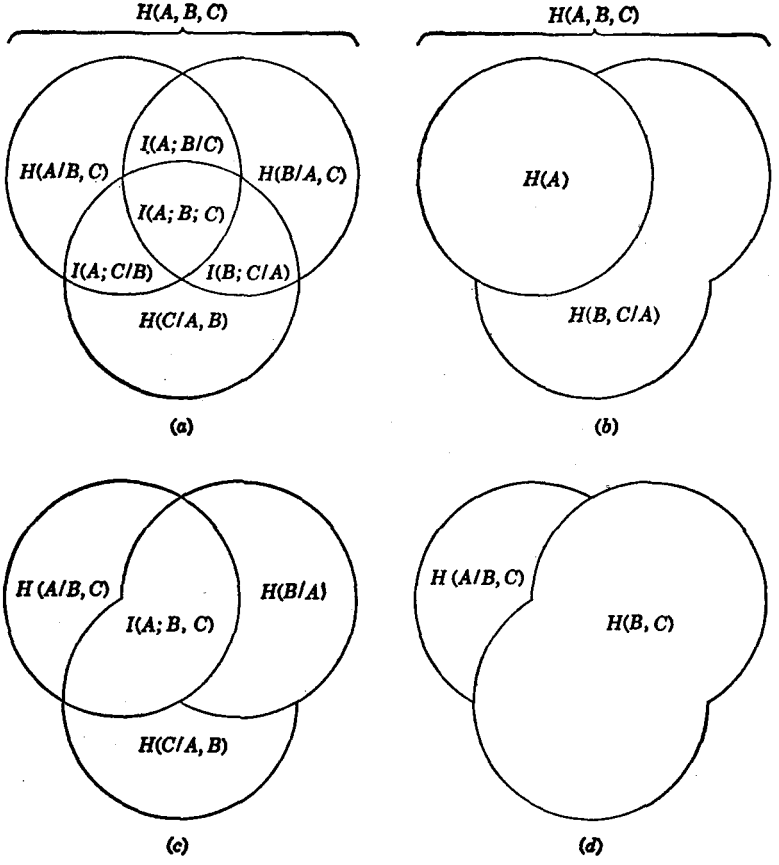


FIG. 5-19. Algunas relaciones de la información.

La figura 5-19 permite memorizar con facilidad estas relaciones; sin embargo presenta ciertas limitaciones. La información mutua $I(A; B)$, como ya se dijo, es una cantidad positiva. $I(A; B; C)$, por el contrario, puede ser negativa. Esto significa que la intersección de los

tres círculos de la figura 5-19*a* puede tomar *este signo.* Se comprobará con el ejemplo siguiente.

Ejemplo 5-12. Dados tres alfabetos binarios A, B y C, se elige un 0 ó un 1 para a_i y b_j con probabilidad 1/2 e independientemente del otro. Finalmente, supóngase que se asigna un 0 a c_k si a_i es igual a b_j y un 1 si no lo son. Algunas de las probabilidades de esas tres variables están determinadas en la tabla 5-3.

TABLA 5-3. PROBABILIDADES DE TRES VARIABLES AL AZAR

$a_ib_jc_k$	$P(a_i, b_j, c_k)$	$P(a_i/b_j, c_k)$	$P(a_i, b_j/c_k)$	$P(a_i, b_j)$	$P(a_i)$
000	1/4	1	1/2	1/4	1/2
001	0	0	0	1/4	1/2
010	0	0	0	1/4	1/2
011	1/4	1	1/2	1/4	1/2
100	0	0	0	1/4	1/2
101	1/4	1	1/2	1/4	1/2
110	1/4	1	1/2	1/4	1/2
111	0	0	0	1/4	1/2

A partir de esta tabla, se calcula:

$$I(A; B) = 0 \text{ bits}$$

$$I(A; B/C) = 1 \text{ bit}$$

$$I(A; B; C) = I(A; B) - I(A; B/C) = -1 \text{ bit}$$

La razón de haber llegado a este resultado es evidente. Puesto que A y B son estadísticamente independientes, $I(A; B) = 0$ y B no proporciona ninguna información sobre A. Sin embargo, si se conoce C, el conocimiento posterior de B dice cuál es el A elegido, y, por lo tanto, suministra un bit de información.

5-13. Capacidad de un canal.

Supóngase un canal de alfabeto de entrada A, alfabeto de salida B, y probabilidades condicionales $P(b_j/a_i)$. El cálculo de la información mutua

$$I(A; B) = \sum_{A,B} P(a, b) \log \frac{P(a, b)}{P(a)\,P(b)} \tag{5-95}$$

exige el conocimiento de las probabilidades de los símbolos de entrada, $P(a_i)$. La información mutua según eso, depende no solamente del

canal sino de la forma en que se emplea, es decir, de las probabilidades con que se eligen los símbolos de entrada. Es interesante examinar la variación de $I(A; B)$ al variar esas probabilidades.

Ejemplo 5-13. En un BSC de probabilidad de error p, se tiene [(5-48)]

$$I(A; B) = H(\omega p + \bar{\omega}\bar{p}) - H(p) \tag{5-96}$$

donde ω es la probabilidad de elegir un 0 de entrada y además $\bar{\omega} = 1 - \omega$, $\bar{p} = 1 - p$. La figura 5-20 representa la curva de variación de $I(A; B)$ con respecto a ω, para un valor de p constante.

La información mutua varía entre 0 y $1 - H(p)$. El mínimo, 0, se alcanza para $\omega = 0$ y $\omega = 1$. En estos casos, se conoce el símbolo enviado con probabilidad 1, incluso antes de recibir el símbolo de salida correspondiente. El valor mínimo, $1 - H(p)$, se obtiene para $\omega = 1/2$, es decir, cuando las dos entradas son equiprobables.

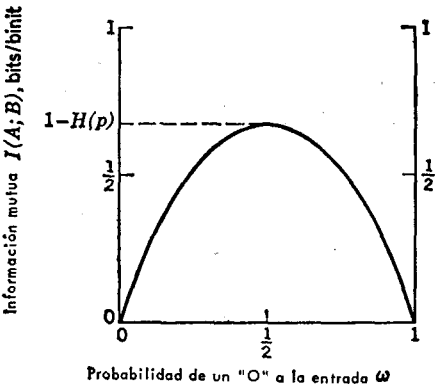


FIG. 5-20. Información mutua de un BSC.

La información mutua de un canal cualquiera puede hacerse igual a cero sin más que elegir uno de los símbolos de entrada con probabilidad 1. Puesto que la información mutua es positiva, esto responde a la pregunta de cuál es el valor mínimo de $I(A; B)$. El valor máximo, sin embargo, es más difícil de calcular. Se denomina C, *capacidad* del canal:

$$C = \max_{P(a_i)} I(A; B) \tag{5-97}$$

Hay que destacar que la capacidad de un canal de información es función exclusivamente de sus probabilidades condicionales. No depende en absoluto de las probabilidades de entrada, o sea de la forma en que se utiliza. Según la figura 5-20, la capacidad de un BSC de probabilidad de error p es $1 - H(p)$.

El cálculo de la capacidad de un canal es, en general, bastante complicado (Muroga, 1953; Shannon, 1957; Fano, 1961). En ciertos casos, sin embargo, puede simplificarse. Una de las clases más importantes en que esto es posible está constituida por los canales llamados *uniformes.*

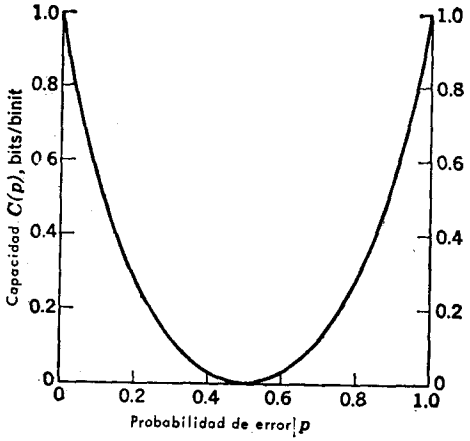


FIG. 5-21. Capacidad de un BSC.

Definición. Consideremos el canal definido por la matriz

$$\begin{bmatrix} P_{11} & P_{12} & \dots & P_{1s} \\ P_{21} & P_{22} & \dots & P_{2s} \\ \dots & \dots & \dots & \dots \\ P_{r1} & P_{r2} & \dots & P_{rs} \end{bmatrix}$$

Como antes $P_{ij} = P(a_i/b_j)$. El canal es uniforme si cada fila y cada columna de la matriz es una permutación arbitraria de los elementos de la primera fila.

Ejemplo 5-14. Se ha considerado ya, en otro ejemplo, un canal de información mutua uniforme, el BSC. La generalización del BSC, el canal simétrico r-ario (rSC), es un canal simétrico de r símbolos de entrada y r de salida. Su matriz aparece en la figura 5-22.

$$\begin{bmatrix} p & \frac{p}{r-1} & \frac{p}{r-1} & \cdots & \frac{p}{r-1} \\ \frac{p}{r-1} & p & \frac{p}{r-1} & \cdots & \frac{p}{r-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{p}{r-1} & \frac{p}{r-1} & \frac{p}{r-1} & \cdots & p \end{bmatrix}$$

FIG. 5-22. Matriz del canal rSC.

Como siempre, $\bar{p} = 1 - p$. La probabilidad de error del canal es p, pero existen $r - 1$ salidas incorrectas por cada símbolo de entrada.

Calcularemos a continuación la capacidad de un canal uniforme. La capacidad es el valor máximo de $I(A; B)$ al variar la distribución de entrada

$$\begin{aligned} I(A; B) &= H(B) - H(B/A) \\ &= H(B) - \sum_{A} P(a) \sum_{B} P(b/a) \log \frac{1}{P(b/a)} \end{aligned} \tag{5-98}$$

El último sumando representa la suma, para cada a_i, de los términos de la fila i de la matriz del canal. Sin embargo, si el canal es uniforme, esta suma es independiente de i. Por lo tanto,

$$I(A; B) = H(B) - \sum_{B} P(b/a) \log \frac{1}{P(b/a)} \tag{5-99}$$

cuyo último término no depende de la distribución de entrada. El valor máximo del segundo miembro de la ecuación (5-99) corresponderá al máximo de $H(B)$. Puesto que el alfabeto de salida consta de r símbolos, $H(B)$ no puede exceder de los r bits, valor que alcanzará si todos los símbolos de salida se presentan con la misma probabilidad. En general no se da la circunstancia de que la distribución de los símbolos de entrada sea tal que los símbolos de salida sean equiprobables. En un canal simétrico, sin embargo, es fácil comprobar que símbolos de entrada equiprobables dan lugar a símbolos de salida

equiprobables. Así, pues, el valor máximo de (5-99), capacidad del canal uniforme, será

$$C = \log r - \sum_{B} P(b/a) \log \frac{1}{P(b/a)}$$

$$= \log r + \sum_{B} P(b/a) \log P(b/a) \tag{5-100}$$

Ejemplo 5-15. Calcular la capacidad de un rSC haciendo uso de la fórmula (5-100):

$$C = \log r + \bar{p} \log \bar{p} + p \log \frac{p}{r-1}$$

$$= \log r - p \log (r-1) - H(p) \tag{5-101}$$

5-14. Información mutua condicional.

La capacidad de un canal es el valor máximo de

$$I(A; B) = \sum_{A,B} P(a, b) \log \frac{P(b/a)}{P(b)} \tag{5-102}$$

valor medio de $\log [P(b/a)/P(b)]$ extendido a los alfabetos de entrada y salida, A y B. La información mutua puede también escribirse en la forma

$$I(A; B) = \sum_{A} P(a) \sum_{B} P(b/a) \log \frac{P(b/a)}{P(b)}$$

$$= \sum_{A} P(a) I(a; B) \tag{5-103}$$

donde se define

$$I(a; B) = \sum_{B} P(b/a) \log \frac{P(b/a)}{P(b)} \tag{5-104}$$

$I(a; B)$ recibe el nombre de *información mutua condicional* (condicionada por a). Corresponde al valor medio de $\log [P(b/a)\ P(b)]$ con respecto a la probabilidad condicional $P(b/a)$.

En general $I(a; B)$ depende del símbolo de entrada a. No obstante, si los símbolos se eligen de acuerdo con un conjunto de probabilidades que dan lugar a la capacidad del canal, se demostrará que $I(a; B)$

no depende de a, siempre que para ese símbolo de entrada $P(a) \neq 0$. *Cuando las probabilidades de entrada son elegidas de forma que se alcance la capacidad del canal,*

$$I(a;\ B) = C \tag{5-105}$$

para cualquier a tal que $P(a) \neq 0$.

Este hecho es fundamental en el cálculo de la capacidad de un canal más general que un canal uniforme, tratado en el apartado anterior (Fano, 1961). Se mencionará también en el apartado 6-10, durante la demostración del segundo teorema de Shannon.

La relación (5-105) se demuestra por reducción al absurdo. Supongamos un conjunto de probabilidades * $P(a_1)$, $P(a_2)$, ..., $P(a_r)$ que dan lugar a la capacidad del canal, pero que no cumplen la relación (5-105). Ya que el valor medio de $I(a; B)$ es igual a la capacidad, debe existir al menos un valor de $I(a; B)$ superior y al menos otro inferior a C. Sin pérdida de generalidad, suponemos

$$I(a_1;\ B) > C \tag{5-106a}$$

$$I(a_2;\ B) < C \tag{5-106b}$$

Sustituyendo a continuación las probabilidades anteriores

$$P(a_1),\ P(a_2),\ P(a_3),\ \ldots,\ P(a_r) \tag{5-107a}$$

por

$$P(a_1) + \Delta,\ P(a_2) - \Delta,\ P(a_3),\ \ldots,\ P(a_r) \tag{5-107b}$$

donde Δ es un pequeño número positivo menor que $P(a_2)$, se demostrará que el valor de la información mutua aumenta. Puesto que el conjunto de probabilidades (5-107a) se supuso daba lugar a la capacidad del canal, este resultado es absurdo; por tanto, la hipótesis de que $I(a; B)$ no era constante es falsa.

Designaremos las nuevas probabilidades definidas en (5-107b) por $P_1(a_1)$, $P_1(a_2)$, ..., $P_1(a_r)$. $P_1(b)$ está dado por

$$\begin{aligned} P_1(b) &= \sum_A P_1(a)\, P(b/a) \\ &= P(b) + \Delta\,[P(b/a_1) - P(b/a_2)] \end{aligned} \tag{5-108}$$

* Se admite que ninguno de los $P(a_i)$ es nulo. Si $P(a_i)$ fuera igual a cero, se consideraría un nuevo canal derivado del anterior, eliminando la entrada a_i.

Sea $I_1(A;\ B)$ el valor de la información mutua calculado mediante las probabilidades $P_1(a)$; por hipótesis, la información mutua correspondiente a las probabilidades $P(a)$ es C, capacidad del canal. Según esto, calcularemos

$$I_1(A;\ B) - C = \sum_A P_1(a) \sum_B P(b/a) \log \frac{P(b/a)}{P_1(b)} - \sum_A P(a) \sum_B P(b/a) \log \frac{P(b/a)}{P(b)}$$

$$= \Delta \left[\sum_B P(b/a_1) \log P(b/a_1) - \sum_B P(b/a_2) \times \log P(b/a_2) \right] + \sum_B P_1(b) \log \frac{1}{P_1(b)} - \sum_B P(b) \log \frac{1}{P(b)} \qquad (5\text{-}109)$$

Sumando y restando la cantidad

$$\Delta \left[\sum_B P(b/a_1) \log \frac{1}{P(b)} - \sum_B P(b/a_2) \log \frac{1}{P(b)} \right] \qquad (5\text{-}110)$$

de ambos miembros de (5-109), se obtiene

$$I_1(A;\ B) - C = \Delta \left[\sum_B P(b/a_1) \log \frac{P(b/a_1)}{P(b)} - \sum_B P(b/a_2) \log \frac{P(b/a_2)}{P(b)} \right] + \sum_B P_1(b) \log \frac{P(b)}{P_1(b)}$$

$$= \Delta \left[I(a_1;\ B) - I(a_2;\ B) \right] + \sum_B P_1(b) \log \frac{P(b)}{P_1(b)} \qquad (5\text{-}111)$$

Para llegar a un absurdo, el segundo miembro de (5-111) ha de ser positivo. Según (5-106), su primer término es mayor que cero. El segundo, por otra parte, de acuerdo con la relación (2-8*a*) frecuentemente utilizada, es negativo. A primera vista, por lo tanto, no parece posible conocer el signo del segundo miembro de (5-111). Sin embargo

no hay razón para ser pesimistas; bastará con examinar en detalle el último término de la expresión

$$\sum_B P_1(b) \log \frac{P(b)}{P_1(b)} = \sum_B \left\{ P(b) + \Delta [P(b/a_1) - P(b/a_2)] \right\} \times \log \frac{1}{1 + \dfrac{\Delta [P(b/a_1) - P(b/a_2)]}{P(b)}} \qquad (5\text{-}112)$$

Para valores de x suficientemente pequeños, el $\log [1/(1 + x)]$ puede sustituirse por $-x/\ln 2$. Según esto, la relación (5-112), para valores pequeños de Δ, se transforma en

$$\begin{aligned} \sum_B P_1(b) \log \frac{P(b)}{P_1(b)} &\approx \frac{-1}{\ln 2} \sum_B \{P(b) + \Delta [P(b/a_1) - P(b/a_2)]\} \times \frac{\Delta [P(b/a_1) - P(b/a_2)]}{P(b)} \\ &\approx \frac{-\Delta}{\ln 2} \sum_B [P(b/a_1) - P(b/a_2)] - \Delta^2 \sum_B \frac{[P(b/a_1) - P(b/a_2)]^2}{P(b)} \\ &\approx \frac{-\Delta^2}{\ln 2} \sum_B \frac{[P(b/a_1) - P(b/a_2)]^2}{P(b)} \end{aligned} \qquad (5\text{-}113)$$

ya que $\sum_B P(b/a_1) = \sum_B P(b/a_2) = 1$. Así pues, el término negativo del segundo miembro de (5-111) es del mismo orden que Δ^2, para valores de Δ pequeños. En cambio el primer término (cantidad positiva) es del mismo orden que Δ; en definitiva, para un Δ suficientemente pequeño, el segundo miembro sería positivo, lo que significaría un absurdo.

Hemos demostrado que la hipótesis de que no todos los valores de la información mutua $I(a; B)$ son iguales a la capacidad del canal, es falsa.

NOTAS

Nota 1. Puede definirse un canal de memoria nula, con un número finito de entradas y salidas, más general que el del apartado 5-1. Un canal de memoria

nula consiste en un espacio A de entradas, un espacio B de salidas y una medida de la probabilidad $p(\cdot/a)$ de B para cada a de A.

Según esto, un canal de información es matemáticamente equivalente a un «experimento estadístico» (Kempthorne, 1952). Las «hipótesis» del experimento corresponden a los símbolos de entrada, y los «resultados» a las salidas del canal. La configuración de un experimento (lo mismo que la de un canal) se define mediante un conjunto de probabilidades condicionales respecto al espacio «resultados».

Gran parte de las cuestiones que presentan interés en el campo de los experimentos estadísticos no lo tienen en absoluto en el caso de los canales de información y viceversa. Un área de interés común es la comparación entre diversos experimentos, o canales de información (Blackwell, 1953; Lindley, 1956; Shannon, 1958). Como el lector puede haber intuido, la capacidad no es el único criterio para evaluar la calidad de un canal de información. Cuando el número de hipótesis (símbolos de entrada) es 2, pueden aplicarse otras conclusiones más concretas, descritas en la literatura estadística (Kullback, 1959; Grettenberg, 1962; Birnbaum, 1961). En algunos casos particulares los métodos estadísticos tradicionales conducen a resultados diametralmente opuestos a los de la teoría de la información (Abramson, 1960).

Nota 2. La capacidad de un ser humano, considerado como un canal de información, fue estudiada por Pierce y Karlin (1957). Realizaron el cálculo mediante un cierto número de experimentos de lectura, llegando a la siguiente conclusión:

La diferencia entre la capacidad de un canal humano (40-50 bits/segundo) y la capacidad de un canal telefónico o de televisión (alrededor de 50.000 bits/segundo y 50.000.000 bits/segundo, respectivamente) es definitiva.

Hay que destacar que Pierce y Karlin intentaron medir la información asimilada por sus sujetos de ensayo; es decir, la información recibida en un punto intermedio del sistema humano de elaboración correspondiente. Kelly (1962), por otra parte, midió la capacidad de información de la retina, cifrándola en 10^9 bits por segundo.

Nota 3. La relación de posibilidad (y muchas veces su logaritmo) juegan un papel importante en la demostración de dos hipótesis estadísticas. Si x y $1-x$ son las probabilidades respectivas de las hipótesis 1 y 2, el logaritmo de la relación de posibilidad es

$$\log \frac{x}{1-x}$$

Golomb (1961) se sirvió del hecho de que

$$\int_x^{\prime} \log \frac{u}{1-u}\, du = H(y) - H(x)$$

[donde $H(\cdot)$ es la función entropía] para identificar el logaritmo de la relación de posibilidad como una *densidad de información*. Si las probabilidades a priori

de las dos hipótesis son x y $1-x$ y las probabilidades a posteriori, después del i-imo experimento (o símbolo de salida), son y_i y $1-y_i$,

$$\int_x^{y_i} \log \frac{u}{1-u}\, du = H(y_i) - H(x)$$

que, según la notación del apartado 5-4, es $H(A/b_j) - H(A)$. Puede calcularse el valor medio de esta diferencia extendida a todas las salidas posibles, obteniéndose una cantidad que corresponde a la información mutua existente entre los resultados del experimento y las hipótesis, cambiando de signo.

Golomb generalizó este concepto al caso de n distinto del de dos hipótesis.

Nota 4. Shannon (1956) señaló la posibilidad de construir un álgebra de canales. La suma corresponde al caso en que se utiliza uno de los canales (pero no los dos). Los alfabetos de entrada y salida del nuevo canal son las reuniones de los alfabetos de entrada y salida originales (ver problema 5-13). El producto de dos canales corresponde a la utilización simultánea de ambos. Tanto la suma como el producto gozan de las propiedades asociativa y conmutativa; el producto, además, de la propiedad distributiva. ,

Nota 5. Kelly (1956) estudió otra interpretación de la capacidad de un canal, que presenta algún interés en ciertos problemas de economía (Murphy, 1962). Imaginemos un jugador observando las salidad de un BSC de probabilidad de error $p < 1/2$ y apostando sobre los símbolos transmitidos. Si desea reunir el capital máximo posible después de n apuestas, deberá jugar en cada observación todo lo que posee. Desgraciadamente, si n es muy grande, esta estrategia le llevará a la bancarrota con toda certeza. Ahora bien, si juega solamente una fracción fija de su capital (menor que 1) en cada observación, éste irá creciendo exponencialmente con el número de apuestas. Kelly sugiere la estrategia a seguir para que la pendiente del crecimiento sea máxima, encontrando que su valor es C, capacidad del canal. Diversas generalizaciones del problema han sido estudiadas por Kelly en este artículo.

Nota 6. La definición de información mutua de dos variables al azar, $I(A; B)$, no se limita estrictamente al caso en que A y B son los alfabetos de entrada y salida de un canal. a_i y b_j pueden ser dos variables al azar cualquiera, siendo $I(A; B)$ la cantidad de información que una de ellas suministra sobre la otra. Pinsker (1954), Powers (1956) y Gel'fand y Yaglom (1957) definieron la cantidad de información que sobre una función al azar contiene otra función semejante, generalización de la información mutua definida en este capítulo. Sea μ_{ab} la medida de la probabilidad de la variable al azar (a, b), y supongamos que μ_a y μ_b son las de a y b. Si μ_{ab} es absolutamente continua con respecto a $\mu_a \mu_b$, la definición de Gel'fond y Yaglom es equivalente a

$$I(A; B) = \int_{A,B} \left(\log \frac{d_{\mu_{ab}}}{d_{\mu_a} d_{\mu_b}} \right) d_{\mu_{ab}}$$

donde $d_{\mu_{ab}}/d_{\mu_a \mu_b}$ es la derivada de Radon-Nikodym de μ_{ab} respecto a $\mu_a \mu_b$. Si las variables al azar a y b toman únicamente un número finito de valores, la defi-

nición se reduce a la enunciada en este capítulo. Si a y b poseen unas densidades de probabilidad afín e indidividual de valor $p(a, b)$, $p(a)$ y $p(b)$,

$$I(A; B) = \int\int p(a, b) \log \frac{p(a, b)}{p(a)\, p(b)}\, da\, db$$

donde a y b son los vectores de Gauss; esta expresión se reduce a

$$I(A; B) = 1/2 \log \frac{|K_a|\ |K_b|}{|K_{ab}|}$$

donde $|K_{ab}|$, $|K_a|$ y $|K_b|$ son los determinantes de la matriz covariante de (a, b), a y b, respectivamente. Si a representa una función muestra de un proceso de Gauss al azar, definido en un intervalo (posiblemente infinito) y b es otra variable al azar,

$$I(A; B) = -1/2 \log \sigma_b^2$$

donde σ_b^2 es el error cuadrático medio obtenido al estimar el valor de b a partir de la observación de a. Finalmente, cuando a y b son funciones muestras de procesos de Gauss al azar, definidos en un intervalo infinito, la proporción media con la que uno de esos procesos al azar suministra información sobre el otro es

$$i(A; B) = 1/2 \int \log \frac{S_a(f) S_b(f)}{S_a(f) S_b(f) - |S_{ab}(f)|^2}\, df$$

donde $S_a(f)$ y $S_b(f)$ constituyen los espectros de densidades de los procesos al azar a y b y $S_{ab}(f)$ es la densidad espectral de intercesión.

PROBLEMAS

5-1. La matriz de un canal de información binario es

$$\begin{array}{c} \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{cc} b_1 & b_2 \end{array} \\ \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix} \end{array}$$

Los símbolos correspondientes a las filas y las columnas de una matriz han sido escogidos convenientemente. Sea $P(a_1) = P_1$, $P(a_2) = P_2$, $P(b_1) = Q_1$ y $P(b_2) = Q_2$.

a) Escribir las ecuaciones (5-6) aplicadas a este canal expresando los Q_i en función de P_i.

b) Resolver las ecuaciones de P_i en función de Q_i.

c) Calcular los valores de $P(a_i/b_j)$ y Q_j de este canal cuando $P_1 = P_2 = 0{,}5$.

d) Expresar P_i en función de Q_j, utilizando el valor de $P(a_i/b_j)$ obtenido en la parte *c*). Comparar las respuestas de las partes *b*) y *d*).

5-2. Cada vez que un símbolo de entrada se transmite sobre el canal 1, se repite simultáneamente sobre el canal 2 (ver la figura P 5-2), de forma que la salida puede considerarse como una pareja de símbolos (b_j, c_k).

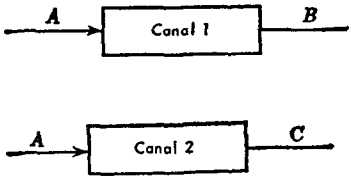


FIG. P 5-2.

Supongamos además que esta repetición se realiza independientemente de los resultados de la transmisión original, según eso

$$P(c_k/a_i, b_j) = P(c_k/a_i)$$

Hay que destacar que esto *no* quiere decir que b_j y c_k sean estadísticamente independientes.

$$P(c_k/b_j) \neq P(c_k)$$

a) Demostrar que

$$I(A; B, C) = I(A; B) + I(A; C) - I(B; C)$$

interpretándolo seguidamente.

b) Generalizar la parte *a*) al caso de n canales.

5-3. Haciendo uso del resultado del problema 5-2 *a*, comprobar la ecuación (5-86).

5-4. Demostrar la ecuación (5-32):

$$I(A^n; B^n) = nI(A; B)$$

5-5. Consideremos el canal de información mostrado en la figura P 5-5. El conjunto de números $R_i = \lambda P_i + \bar{\lambda} Q_i$, para dos conjuntos cualquiera de probabilidades de entrada P_i, $i = 1, 2, \ldots, r$, y Q_i, $i = 1, 2, \ldots, r$ y cualquier λ comprendido en el intervalo [0, 1], define también un conjunto de probabilidades de entrada. Ya que

$$R_i \geq 0 \quad \text{para todo } i$$

y

$$\sum_{i=1}^{r} R_i = 1$$

sean $I_P(A; B)$, $I_Q(A; B)$ e $I_R(A; B)$ la información mutua del canal susodicho, cuando las probabilidades de entrada son P_i, Q_i y R_i.

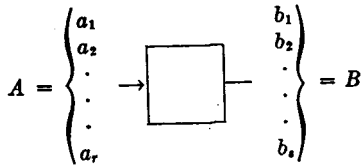


FIG. P 5-5.

a) Demostrar la «convexidad» de la información mutua. Es decir, que

$$I_R(A; B) \geqq \lambda I_P(A; B) + \bar{\lambda} I_Q(A; B)$$

b) Demostrar que

$$I_R(A; B) \leqq \lambda I_P(A; B) + \bar{\lambda} I_Q(A; B) + H(\lambda)$$

5-6. Generalizar las partes *a*) y *b*) del problema 5-5 al caso en que el conjunto de probabilidades R_i está formado por n conjuntos de probabilidades, en lugar de solamente dos.

5-7. Considérense dos canales de información con alfabetos de entrada A_1 y A_2 y alfabetos de salida respectivos B_1 y B_2 (fig. 5-7). Las probabilidades del canal 1 son $P_1(b/a)$ y las del canal 2 $P_2(b/a)$. Sean $P_1(a)$ y $P_2(a)$ las distribuciones de entrada de A_1 y A_2.

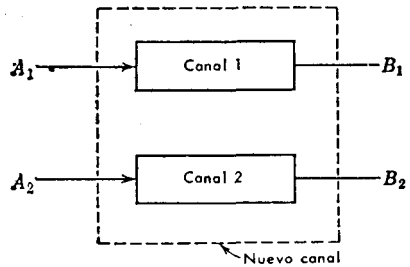


FIG. P 5-7.

a) Definir un nuevo canal con un alfabeto de entrada que comprende los símbolos de entrada A_1 y A_2. El nuevo alfabeto de salida B reúne los símbolos de B_1 y los de B_2. Una entrada del nuevo canal se selecciona eligiendo bien A_1 (con probabilidad λ) o A_2 (con probabilidad $1 - \lambda = \bar{\lambda}$) y seleccionando a continuación un símbolo de ese alfabeto de acuerdo con las probabilidades $P_1(a)$ o $P_2(a)$. Expresar $H(A)$ en función de $H(A_1)$, $H(A_2)$ y λ.

b) Las probabilidades del nuevo canal, $P(b/a)$, están dadas por $P_1(b/a)$ si a y b están en A_1 y B_1, por $P_2(b/a)$ si a y b están en A_2 y B_2, y son nulas si a está en A_1 y b en B_2, ó bien a en A_2 y b en B_1. Expresar $H(A/B)$ en función de $H(A_1/B_1)$, $H(A_2/B_2)$ y λ.

c) Expresar $I(A, B)$ en función de $I(A_1; B_1)$, $I(A_2; B_2)$ y λ.

5-8. Generalizar el problema 5-7 al caso de n canales de información.

5-9. El *canal multiplicativo binario* del dibujo posee dos entradas binarias y una salida binaria, $b = ac$. Este canal puede describirse como un canal ordi-

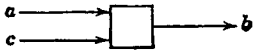


FIG. P 5-9.

nario de memoria nula, considerando las cuatro combinaciones de entrada posibles como partes de un nuevo alfabeto de entrada A:

$$A = \left\{ \begin{matrix} 00 \\ 01 \\ 10 \\ 11 \end{matrix} \right\}$$

a) Escribir la matriz del canal con alfabeto de entrada A y salida B.

b) Los símbolos de entrada a y c se seleccionan independientemente. $\Pr\{a = 0\} = \omega_1$ y $\Pr\{c = 0\} = \omega_1$. Sea $1 - \omega_1 = \overline{\omega}_1$ y $1 - \omega_2 = \overline{\omega}_2$. Calcular $I(A; B)$. Interpretar el resultado.

c) Encontrar el valor máximo de $I(A'; B)$ cuando ω_1 y ω_2 varían. Calcular todas las combinaciones posibles de ω_1 y ω_2 que dan lugar a este valor máximo.

5-10. Sea **P** la matriz de un canal con r entradas y s salidas. Supongamos que a es el número de columnas de la matriz que tienen todos sus elementos nulos.

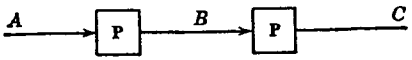


FIG. P 5-10.

a) Si el canal es determinante, calcular su capacidad.

b) Si (en lugar de la hipótesis de la parte *a*) suponemos que es un canal sin ruido, calcular su capacidad.

c) Admitamos simultáneamente las hipótesis de las partes *a*) y *b*). Dos canales de estas características se colocan en serie, tal como e representa en el dibujo. Calcular la capacidad del canal resultante, de entrada A y salida C.

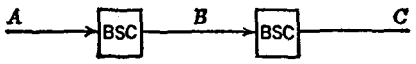


FIG. P 5-11.

5-11. Se conectan en serie dos BSCs, cada uno de probabilidad de error p, tal como se representa en el dibujo. Las entradas 0 y 1 de A se eligen con idéntica probabilidad. Calcular:

a) $H(A)$.
b) $H(B)$.
c) $H(C)$.
d) $H(A, B)$.
e) $H(B, C)$.
f) $H(A, C)$.
g) $H(A, B, C)$.
h) $I(A; B; C)$.

5-12. Sean a y b dos variables binarias al azar, independientes y de distribuciones idénticas, tales que la probabilidad de un 0 es igual a la probabilidad de un 1. Definir la variable binaria al azar

$$c = ab$$

Calcular:

a) $H(A)$, $H(B)$, $H(C)$.
b) $I(A; B)$, $I(A; C)$, $I(B; C)$.
c) $H(A, B)$, $H(A, C)$, $H(B, C)$.
d) $H(A, B, C)$.
e) $H(A/B)$, $H(A/C)$, $H(C/A, B)$.
f) $H(A/B, C)$, $H(B/A, C)$, $H(C/A, B)$.
g) $I(A; B/C)$, $I(B; A/C)$, $I(C; A/B)$.
h) $I(A; B; C)$.

5-13. Sean a y b dos variables binarias al azar, independientes y con distribuciones idénticas, tales que la probabilidad de un 0 es igual a la probabilidad de un 1. Definir la variable binaria al azar $c = a + b$, módulo 2. Es decir, c es 0 si a es igual a b, y c es 1 si a es distinta de b. Calcular.

a) $H(A)$, $H(B)$, $H(C)$.
b) $I(A; B)$, $I(A; C)$, $I(B; C)$.
c) $H(A, B)$, $H(A, C)$, $H(B, C)$.
d) $H(A, B, C)$.
e) $H(A/B)$, $H(A/C)$, $H(B/C)$.
f) $H(A/B, C)$, $H(B/A, C)$, $H(C/A, B)$.
g) $I(A; B/C)$, $I(B; A/C)$, $I(C; A/B)$.
h) $I(A; B; C)$.

5-14. Encontrar la capacidad de

$$\begin{bmatrix} 1-p-q & q & p \\ p & q & 1-p-q \end{bmatrix}$$

El caso particular de $p = 0$ se denomina *canal binario de borrado*. Dar una interpretación a esta capacidad.

5-15. Sean P_1 y P_2 las matrices de dos canales de alfabetos de entrada A_1 y A_2 y de salida B_1 y B_2, respectivamente. Formar una nueva matriz **P** de alfabeto de entrada $A = A_1 \cup A_2$ y de salida $B = B_1 \cup B_2$, como se muestra a continuación:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{P}_2 \end{bmatrix}$$

O representa una matriz de elementos nulos.

Sea $P(a_i)$ la probabilidad de que un símbolo de entrada $a_i \in A$. Supongamos $Q_1 = \sum_{A_1} P(a_i)$ y $Q_2 = \sum_{A_1} P(a_i)$. Q_1 es la probabilidad de que un símbolo de A_1 sea enviado. Sean C_1, C_2 y C las capacidades respectivas de P_1, P_2 y P.

a) Calcular los valores de Q_i (en función de C_1 y C_2) que dan lugar a la capacidad del canal **P**.

b) Calcular C en función de C_1 y C_2.

c) Generalizar los resultados de *a*) y *b*) al caso en que se combinan n canales, en lugar de dos.

5-16. *a*) Calcular la capacidad del canal

$$\begin{bmatrix} \bar{p} & p & 0 & 0 \\ p & \bar{p} & 0 & 0 \\ 0 & 0 & \bar{p} & p \\ 0 & 0 & p & \bar{p} \end{bmatrix}$$

Dibujar la variación de la capacidad en función de p.

b) Calcular la capacidad de

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \bar{p} & p \\ 0 & p & \bar{p} \end{bmatrix}$$

Dibujar la variación de la capacidad en función de p y comparar este resultado con el de la parte *a*).

5-17. Calcular la capacidad de los dos canales siguientes:

a)
$$\begin{pmatrix} \bar{p}-\varepsilon & p-\varepsilon & 2\varepsilon \\ p-\varepsilon & \bar{p}-\varepsilon & 2\varepsilon \end{pmatrix}$$

b)
$$\begin{pmatrix} \bar{p}-\varepsilon & p-\varepsilon & 2\varepsilon & 0 \\ p-\varepsilon & \bar{p}-\varepsilon & 0 & 2\varepsilon \end{pmatrix}$$

c) Aplicar la aproximación

$$\log(1+\varepsilon) \approx \frac{\varepsilon}{\ln 2} \qquad \text{para } \varepsilon \approx 0$$

para calcular y comparar la conducta de los dos canales anteriores cuando ε es muy pequeño.

CAPITULO 6

MENSAJES CONFIABLES TRANSMITIDOS POR CANALES NO CONFIABLES

6-1. Introducción.

En este capítulo se demostrará el segundo teorema de Shannon, la más sorprendente e importante conclusión de la teoría de la información. Debido al significado de este teorema sería conveniente volver atrás y resumir las principales conclusiones deducidas hasta aquí. Se ha podido justificar el empleo de la entropía y las medidas de información derivadas de ella, en dos ocasiones: Primer teorema de Shannon (apartado 4-3) y su generalización, que trataba de la equivocación (apartado 5-5). El primer teorema de Shannon facilitó una unidad práctica con la que medir la información emitida por una fuente. Este teorema hizo posible evaluar los símbolos de una fuente según los binits (o símbolos de orden r) necesarios para representarlos. La generalización del teorema mostró que podía utilizarse, como unidad con la que medir los resultados de la transmisión a través de un canal, una magnitud relacionada con la entropía (equivocación).

Para codificar los símbolos de un alfabeto fuente A deben emplearse, por término medio, $H(A)$ binits por símbolo. Sin embargo, si los símbolos de A se transmiten por un canal, y se observa los símbolos del alfabeto de salida B, se necesitarán solamente, para representar los símbolos de entrada, $H(A/B)$ binits por símbolo de A. Por lo tanto, en ese sentido, la salida del canal ha suministrado $H(A) - H(A/B)$ bits de información. La equivocación $H(A/B)$ puede variar entre cero (para un canal sin ruidos) y $H(A)$ (para un canal cuyas entradas y salidas son estadísticamente independientes). El número de binits recibidos por cada símbolo de A varía entre cero y $H(A)$.

La transmisión de $H(A) - H(A/B)$ binits es un logro importante. Sin embargo, la forma en que estos binits se presentan a la salida deja mucho que desear. Examinemos esta cuestión más en detalle. Supongamos la transmisión de un bloque n de símbolos, desde una fuente A a través de un canal de información. Si el canal no tiene ruidos, $H(A/B)$ es nula, y cada símbolo de salida contiene $H(A)$ bits de información; una secuencia de n salidas permite reconstruir la secuencia de n entradas emitidas, siendo evidente que los $H(A)$ bits de información recibidos están *libres de error.* Si el canal tiene ruidos, por el contrario, la equivocación no será en general nula, por lo que cada símbolo de salida no contendrá más que $H(A) - H(A/B)$ bits de información. Hay que destacar, además, la diferencia fundamental existente entre esta información y la proviniente de un canal sin ruidos. La secuencia de entrada no puede reconstruirse perfectamente por el mero conocimiento de la salida del canal. Todo lo que puede afirmarse, por el hecho de conocerla, es que las entradas se codifican empleando $H(A) - H(A/B)$ binits menos por símbolo. Por lo tanto, aun cuando se obtiene una cierta información, no se llega al conocimiento libre de error del mensaje transmitido. Esta dificultad va a resolverla el segundo teorema de Shannon.

El segundo teorema de Shannon, publicado por primera vez en 1848, fue un acontecimiento que marcó el nacimiento de la teoría de la información. No obstante, la demostración del teorema en su versión original contenía algunos puntos débiles (McMillan, 1953). La primera demostración rigurosa se debe a Feinstein (1957). Posteriormente aparecieron otras, obra de Shannon (*1957a*); Blackwell, Breiman y Thomasian (1939); y Fano (1961). La presentada en este capítulo es en cierto modo más sencilla que las mencionadas.

6-2. Probabilidad de error y reglas de decisión.

El segundo teorema de Shannon trata de la cantidad de información sin error que puede obtenerse de un cierto canal. Con objeto de apreciar más claramente el significado del teorema, estudiaremos el problema de la probabilidad de error de un canal. En algunos de los canales primarios vistos hasta aquí, tales como el BSC y el rSC, se intuye el concepto de probabilidad de error de un canal. No obstante, incluso en esos casos, como se verá a continuación, la probabilidad

de error depende de un factor que aún no se ha tenido en cuenta. Consideremos, por ejemplo, el BSC.

$$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \tag{6-1}$$

Normalmente diremos que la probabilidad de error de este canal es 0.1. Hay que destacar, sin embargo, que al afirmarlo, se ha supuesto que el canal se utiliza de forma «lógica y razonable». Si al examinar la salida se decidiera que a un *cero* recibido corresponde un *uno* enviado y viceversa, la probabilidad sería 0.9. Naturalmente esta forma de emplear el canal no es la indicada. No obstante hay que tener en cuenta esta posibilidad. La probabilidad de error depende de la forma en que el receptor interpreta los símbolos que salen del canal.

Considerando un caso más significativo, tomemos el canal

$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.3 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{bmatrix} \tag{6-2}$$

El canal tiene tres entradas a_1, a_2, a_3 y tres salidas b_1, b_2, b_3. ¿Qué símbolo de entrada corresponde a un símbolo de salida recibido? Esta pregunta da lugar a la siguiente definición.

> *Definición.* Consideremos un canal con un alfabeto de entrada r-ario $A = \{a_i\}$, $i = 1, 2, \ldots, r$, y un alfabeto de salida de s símbolos $B = \{b_j\}$, $j = 1, 2, \ldots, s$. Se denomina regla de decisión, $d(b_j)$ a la función que especifica el símbolo de entrada único que corresponde a cada símbolo de salida.

Ejemplo 5-1. Dos reglas de decisión del canal de (6-2) podrían ser

$$\begin{aligned} d(b_1) &= a_1 \\ d(b_2) &= a_2 \\ d(b_3) &= a_3 \end{aligned} \tag{6-3}$$

$$\begin{aligned} d(b_1) &= a_1 \\ d(b_2) &= a_2 \\ d(b_3) &= a_2 \end{aligned} \tag{6-4}$$

Un canal de r entradas y s salidas admite r^s reglas de decisión diferentes. La pregunta que sugirió la definición puede volverse a plan-

tear en la forma «¿Cuál de las r^s reglas de decisión debe escogerse en cada caso? La respuesta depende en general del objetivo perseguido, sin embargo una meta lógica es la minimización de la probabilidad de error del canal. Por lo tanto, se elegirá la regla de decisión que haga mínima la probabilidad de error. Para encontrarla definiremos en primer lugar la probabilidad de error P_E, que se expresa como el valor medio de $P(E/b_j)$ probabilidad condicional de error cuando la salida del canal es b_j.

$$P_E = \sum_B P(E/b) P(b) \tag{6-5}$$

Esta ecuación determina la probabilidad de error como suma de una serie de términos positivos. Según eso, la regla de decisión $d(b_j)$ que hace mínima a P_E será aquella que haga mínimo cada término de la suma. $P(b_j)$ es independiente de la regla de decisión empleada; así pues, la regla de decisión elegida, $d(b_j)$, deberá hacer mínima la probabilidad condicional $P(E/b_j)$.

Para una regla de decisión fija,

$$P(E/b_j) = 1 - P[d(b_j)/b_j] \tag{6-6}$$

donde, por ser la regla fija, $d(b_j) = a_i$ es la probabilidad hacia atrás $P(a_i/b_j)$. Finalmente, con objeto de que (6-6) sea mínimo para cada b_j, se elige

$$d(b_j) = a^* \tag{6-7a}$$

donde a^* está definida por

$$P(a^*/b_j) \geq P(a_i/b_j) \quad \text{para cualquier } i \tag{6-7b}$$

En otras palabras, *la probabilidad de error de un canal será mínima con la regla de decisión que asigna a cada símbolo de salida el símbolo de entrada de mayor probabilidad.* Esta regla de decisión recibe el nombre de regla de *máxima posibilidad condicional.* Depende de las probabilidades a priori $P(a_i)$. La ley de Bayes permite escribir la ecuación (6-7*b*) en la forma

$$\frac{P(b_j/a^*) P(a^*)}{P(b_j)} \geq \frac{P(b_j/a_i) P(a_i)}{P(b_j)} \quad \text{para cualquier } i \tag{6-8}$$

Así, pues, cuando todas las probabilidades a priori son idénticas, la regla de decisión de máxima posibilidad condicional se transforma en

$$d(b_j) = a^* \tag{6-9a}$$

donde

$$P(b_i|a^*) \geqq P(b_i|a_i) \quad \text{para cualquier } i \tag{6-9b}$$

La regla de decisión definida por esta relación se conoce como la de *máxima posibilidad*; es independiente de las probabilidades a priori. Cuando todas las probabilidades a priori son iguales, la regla de decisión de máxima posibilidad corresponde a la probabilidad de error mínima. Aun cuando no sean iguales (e incluso desconocidas), se empleará este método de decisión; en tales casos, como es natural, la probabilidad de error del canal no tiene por qué ser mínima.

Ejemplo 6-2. A partir de (6-9) puede definirse inmediatamente la regla de decisión de máxima posibilidad correspondiente al canal de (6-2). Esta regla es

$$d(b_1) = a_1$$
$$d(b_2) = a_2$$
$$d(b_3) = a_3$$

Hay que destacar que la regla no es única. Realmente pueden aplicarse tres reglas de decisión de máxima posibilidad a este canal.

El valor de la probabilidad de error que corresponde al empleo de una regla de decisión cualquiera puede calcularse fácilmente a partir de (6-5) y (6-6).

$$\begin{aligned} P_E &= \sum_B P(E/b)\, P(b) \\ &= \sum_B P(b) - \sum_B P[d(b)/b]\, P(b) \\ &= 1 - \sum_B P[d(b),\ b] \end{aligned} \tag{6-10}$$

Los términos de la suma son las probabilidades simultáneas de transmitir $d(b_j) = a^*$ y recibir b_j. Por lo tanto, siendo $\bar{P}_E = 1 - P_E$, (6-10) se convierte en

$$\bar{P}_E = \sum_B P(a^*,\ b) \tag{6-11}$$

Puesto que

$$\sum_{A,B} P(a,\ b) = 1 \tag{6-12}$$

(6-10) puede también escribirse como

$$P_E = \sum_{B,A-a^*} P(a,\ b) \tag{6-13}$$

El símbolo $\sum_{A-a^*}$ representa la suma extendida a todos los miembros del alfabeto A, excepto $d(b_j) = a^*$. Otra forma de expresar (6-13) es

$$P_E = \sum_{B,A-a^*} P(b/a)\,P(a) \tag{6-14}$$

Si las probabilidades a priori son iguales, la ecuación (6-14) se transforma en

$$P_E = \frac{1}{r} \sum_{B,A-a^*} P(b/a) \tag{6-15}$$

Esta ecuación presenta algún interés (en el caso de igualdad de las probabilidades a priori) ya que es la expresión de la probabilidad de error de un canal en función de una suma extendida a los elementos de la matriz del canal $P(b/a)$. La suma se extiende a todos ellos, omitiendo uno de cada columna [el correspondiente a $d(b_j)$].

Ejemplo 6-3. Calcularemos la probabilidad de error del canal utilizado en los ejemplos 6-1 y 6-2.

$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.3 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{bmatrix} \tag{6-16}$$

Supondremos que los tres símbolos de entrada se eligen con la misma probabilidad y que se aplica la regla de decisión de máxima posibilidad. (Recordemos que esta regla da lugar al mínimo de P_E si las probabilidades a priori son iguales).

$$P_E = 1/3\ [(0.2 + 0.3) + (0.3 + 0.3) + (0.2 + 0.4)]$$
$$= 0.567$$

6-3. Límite de Fano.

La probabilidad de error se ha definido en el apartado anterior sin mencionar el concepto de entropía, equivocación, o información

mutua. El objeto de este capítulo es establecer una conexión entre estos dos conjuntos de conceptos independientes. Como primer paso en este sentido, expresaremos los límites superior e inferior de la equivocación en función de la probabilidad de error.

Durante el cálculo siguiente se hará uso repetidas veces de las relaciones (6-11) y (6-13)

$$\bar{P}_E = \sum_{B} P(a^*, b) \tag{6-11}$$

$$P_E = \sum_{B, A-a^*} P(a, b) \tag{6-13}$$

A partir de ellas, se deduce la identidad

$$H(P_E) + P_E \log (r-1) = P_E \log \frac{r-1}{P_E} + \bar{P}_E \log \frac{1}{\bar{P}_E}$$

$$= \sum_{B, A-a^*} P(a, b) \log \frac{r-1}{P_E} + \sum_{B} P(a^*, b) \log \frac{1}{\bar{P}_E} \tag{6-17}$$

La equivocación $H(A/B)$ puede expresarse en función de las mismas sumas

$$H(A/B) = \sum_{B, A-a^*} P(a, b) \log \frac{1}{P(a/b)} + \sum_{B} P(a^*, b) \log \frac{1}{P(a^*/b)} \tag{6-18}$$

Restando (6-17) de (6-18) se encuentra

$$H(A/B) - H(P_E) - P_E \log (r-1)$$

$$= \sum_{B, A-a^*} P(a, b) \log \frac{P_E}{(r-1) P(a/b)} + \sum_{B} P(a^*, b) \log \frac{\bar{P}_E}{P(a^*/b)} \tag{6-19}$$

Mediante la relación (2-2), puede cambiarse la base de los logaritmos del segundo miembro, con lo que resulta

$$(\log e)^{-1}\,[H(A/B) - H(P_E) - P_E \log (r-1)]$$

$$= \sum_{B,A-a^*} P(a,b)\,\ln \frac{P_E}{(r-1)\,P(a/b)} + \sum_{B} P(a^*,b)\,\ln \frac{\bar{P}_E}{P(a^*/b)} \qquad (6\text{-}20)$$

Puede introducirse la relación

$$\ln x \leq x - 1 \qquad (6\text{-}21)$$

en cada uno de los términos de la suma. El segundo miembro de (6-20) es menor o igual que

$$\sum_{B,A-a^*} P(a,b)\left[\frac{P_E}{(r-1)\,P(a/b)} - 1\right] + \sum_{B} P(a^*,b)\left[\frac{P_E}{P(a^*/b)} - 1\right]$$

$$= \left[\frac{P_E}{r-1}\sum_{B,A-a^*} P(b)\right] - P_E + \left[\bar{P}_E \sum_{B} P(b)\right] - \bar{P}_E$$

$$= 0 \qquad (6\text{-}22)$$

Con lo que se llega a la desigualdad buscada,

$$H(A/B) \leq H(P_E) + P_E \log (r-1) \qquad (6\text{-}23)$$

Esta importante relación fue deducida en primer lugar por Fano. Tiene validez cualquiera que sea la regla de decisión aplicada, aun cuando la probabilidad de error dependa de ella. La desigualdad sugiere una interpretación interesante. Supongamos una regla de decisión dada. Al recibir un símbolo, se necesitan $H(P_E)$ bits de información para reconocer si la regla de decisión ha dado lugar a un error. Un error se produce con probabilidad P_E, pudiendo especificarse entonces, con un máximo de $\log (r-1)$ bits, cual de los $r-1$ restantes símbolos de entrada es el enviado. Desgraciadamente esta interpretación no prueba la relación (6-23), aun cuando constituye la base de una demostración diferente de la desarrollada.

Examinemos en qué caso el límite de Fano se transforma en una igualdad. La desigualdad

$$\ln x \leq x - 1 \qquad (6\text{-}21)$$

es una igualdad para $x = 1$. Sustituyendo esta condición en (6-23), encontramos que la relación de Fano es una igualdad solamente cuando

$$P(a/b) = \frac{P_E}{r-1} \quad \text{para } b \text{ y } a \neq a \tag{6-24a}$$

y

$$P(a^*/B) = \bar{P}_E \quad \text{para cualquier } b \tag{6-24b}$$

ya que

$$\sum_A P(a/b) = 1 \quad \text{para cualquier valor de } b$$

La condición (6-24*b*) se deduce directamente de la primera, (6-24*a*). La ecuación (6-24*a*) implica que todos los símbolos de entrada, excepto el elegido por la regla de decisión, sean igualmente probables. Esta condición refuerza aun más la interpretación del límite de Fano.

6-4. Mensajes confiables y canales no confiables.

La finalidad del segundo teorema de Shannon es definir las limitaciones fundamentales que un canal no confiable ofrece a la transmisión de mensajes sin error. Consideremos en primer lugar la transmisión de mensajes confiables a través de un BSC (figura 6-1).

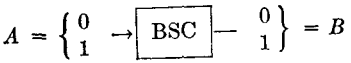


FIG. 6-1. Un BSC.

Para mayor precisión supongamos que p, probabilidad de error de un BSC, es igual a 0.01. Es decir, el 99 por ciento de los binits transmitidos es recibido correctamente. En gran parte de los modernos sistemas de transmisión, sin embargo, este nivel de confiabilidad está lejos de ser aceptable. Las probabilidades de error admitidas son del orden de 10^{-6}, 10^{-8} e incluso menores. Con objeto de aumentar la confiabilidad del canal, cada mensaje debe repetirse varias veces. Supongamos, por ejemplo, que se decide repetir *tres* veces cada uno de ellos (0 ó 1). La figura 6-2 representa este proceso.

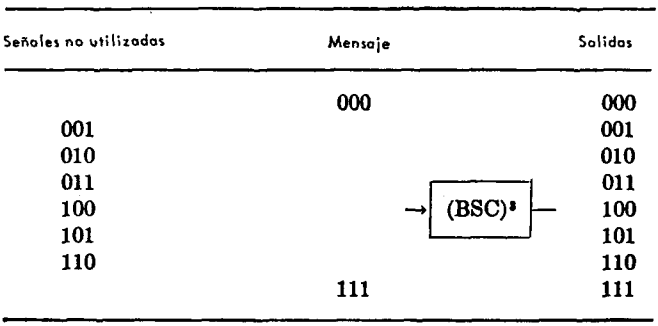


FIG. 6-2. Un método de aumentar la confiabilidad.

La salida del canal en estas circunstancias es un elemento de $(BSC)^3$, una secuencia binaria de longitud 3. La probabilidad de que no se presente ningún error en la transmisión de los tres dígitos es

$$(1 - p)^3 = (\bar{p})^3$$

La probabilidad de un error y solo uno

$$3\, p\, \bar{p}^2$$

La probabilidad de dos errores

$$3\, p^2\, \bar{p}$$

mientras que la probabilidad de que los tres binits recibidos sean erróneos es

$$p^3$$

Siempre que p sea menor que 1/2 (es decir, siempre que la probabilidad de recibir un binit correctamente sea mayor que recibirlo con error), parece razonable decidir que el mensaje emitido ha sido 000 ó 111 por mayoría entre los tres binits recibidos. Esta regla de decisión no precisa realmente justificación; es fácil comprobar que se trata de la regla de decisión de máxima posibilidad. En cualquier caso, tal regla da lugar a una probabilidad de interpretar el mensaje erróneamen-

te * P_E (igual a la suma de las probabilidades de que dos y tres binits sean erróneos).

$$P_E = p^3 + 3\,p^2\,\bar{p} \tag{6-25}$$

Para $p = 0.01$, obtenemos

$$P_E \approx 3 \times 10^{-4} \tag{6-26}$$

Así pues, la probabilidad de error ha pasado de 10^{-2} (enviando un 0 o un 1) a 3×10^{-4} (al enviar 000 ó 111). Continuando en la misma dirección no es difícil aumentar aún la confiabilidad. Pueden enviarse cinco binits por mensaje, tal como representa la figura 6-3.

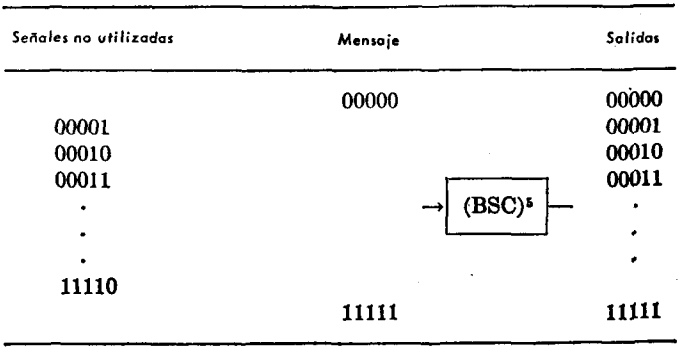


FIG. 6-3. Un método de aumentar la confiabilidad

Las probabilidades respectivas de que se produzcan en la transmisión cero, uno, dos, tres, cuatro o cinco binits erróneos son

$$\bar{p}^5$$
$$5\,p\,\bar{p}^4$$
$$10\,p^2\,\bar{p}^3$$
$$10\,p^3\,\bar{p}^2$$
$$5\,p^4\,\bar{p}$$
$$p^5$$

* La probabilidad de mensaje erróneo depende normalmente de las probabilidades a priori. No obstante, dada la simetría de la situación descrita, la probabilidad de error en este caso es independiente de dichas probabilidades.

Haciendo uso nuevamente de la regla de mayoría (es decir, máxima posibilidad) para decidir si el mensaje enviado fue 00000 ó 11111, la probabilidad de error tiene el valor

$$P_E = p^5 + 5\,p^4\,\bar{p} + 10\,p^3\,\bar{p}^2 \tag{6-27}$$

(o sea, suma de las probabilidades de tres, cuatro y cinco binits erróneos). Para $p = 0.01$, se encuentra

$$P_E \approx 10^{-5} \tag{6-28}$$

Con este procedimiento la confiabilidad puede crecer indefinidamente. La tabla 6-1 muestra la probabilidad de error al transmitir por un BSC de probabilidad de error $p = 0.01$, 1, 3, 5, 7, 9 y 11 binits por mensaje.

TABLA 6-1. PROBABILIDADES DE UN MENSAJE ERRÓNEO EN UN BSC

Binits por mensaje binario	*Probabilidad de mensaje erróneo*
1	10^{-2}
3	3×10^{-4}
5	10^{-5}
7	4×10^{-7}
9	10^{-8}
11	5×10^{-10}

La mejora que se aprecia en la tabla no se alcanza sin pagar un precio a cambio. El precio se cifra en el aumento de redundancia de los binits transmitidos. En otras palabras, aun cuando puede reducirse la probabilidad de error de 0.01 a 5×10^{-10}, al pasar de 1 binit a 11 binits por mensaje binario, la velocidad de mensaje disminuye, pasando de 1 mensaje por binit a 1/11 mensaje por binit. En general, el procedimiento repetitivo descrito plantea un compromiso entre la velocidad de los mensajes y su confiabilidad. La figura 6-4 representa la variación de estos valores.

6-5. Ejemplo de codificación con corrección de errores.

La figura 6-4 sugiere una pregunta importante. El esquema de codificación estudiado hasta aquí (simple repetición) constituye el pro-

cedimiento más directo de intercambiar velocidad de mensaje por confiabilidad. ¿Existe algún método más sofisticado y eficaz de efectuar este intercambio? Es decir, para un valor dado de probabilidad de mensaje erróneo. ¿Existe algún método de codificación que dé una velocidad de mensaje mayor que la obtenida por simple repetición, indicada en la figura 6-4? La respuesta es sencillamente: «¡Sí!»

El segundo teorema de Shannon responde precisamente a esa pregunta (apartado 6-10). No sólo afirma que pueden obtenerse resultados mejores que los de la figura 6-4, sino que dice en *cuánto* pueden mejorarse. La respuesta «cuánto mejor» aportada por el teorema es verdaderamente lo más sorprendente de lo que a continuación se dirá. La figura 6-5 representa la respuesta en su aspecto cuantitativo.

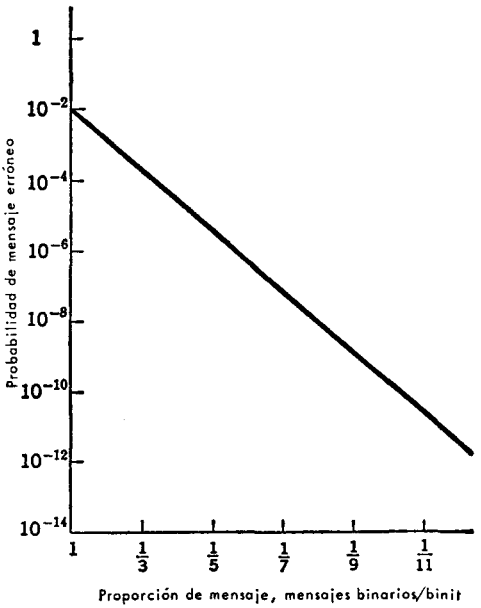


FIG. 6-4. Compromiso entre la proporción y la confiabilidad en un BSC con repetición.

El segundo teorema de Shannon dice que para *cualquier* velocidad de mensaje menor que la capacidad C del canal, existen códigos tales que la probabilidad de mensaje erróneo es menor que cualquier número positivo ε, tan pequeño como se quiera. El teorema concluye en

forma sorprendente diciendo que *no* es necesario hacer tender a 0 la velocidad del mensaje para que la confiabilidad del canal aumente indefinidamente.

En el apartado 6-4 se discutió la posibilidad de transmitir una información virtualmente libre de error, a través de un canal no confiable, un BSC. Examinaremos a continuación, con un poco más de atención, el compromiso existente entre velocidad de un mensaje y su confiabilidad. En el apartado anterior se vio que la velocidad disminuía por el hecho de repetir el mensaje binario transmitido. Tal

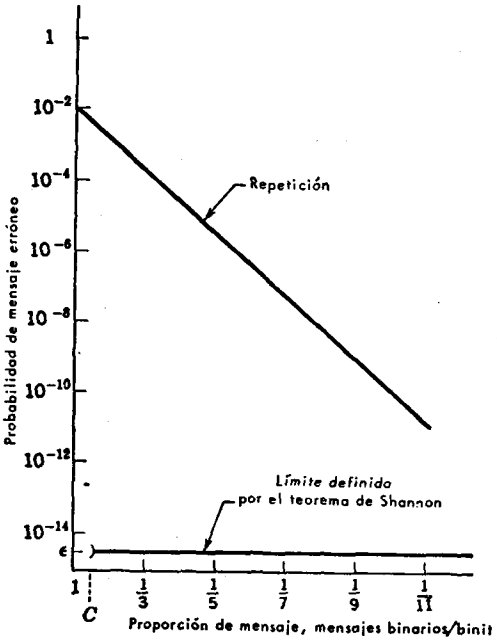


FIG. 6-5. Dos valores diferentes del intercambio de velocidad por confiabilidad en un BSC.

como indicaban las figuras 6-1 y 6-2, puede interpretarse como un aumento del orden de la extensión del canal y la selección de mensajes de dos de los posibles símbolos de entrada, a_i. Un procedimiento más eficaz para variar la velocidad de los mensajes (que se empleará para demostrar el segundo teorema de Shannon) consiste en fijar el orden de la extensión y variar el número de símbolos de entrada, a_i,

usados como mensajes. La figura 6-6 representa esta solución en el caso de un BSC.

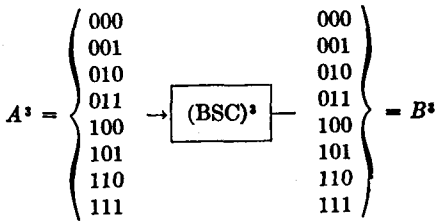


FIG. 6-6. Tercera extensión de un BSC.

Supongamos que los símbolos binarios pueden transmitirse a través de un BSC a la velocidad de uno por segundo. Entonces los α_i, consistentes en secuencias de 3 binits, se transmitirán a un ritmo de uno cada 3 segundos. Si se seleccionan como mensajes las dos secuencias 000 y 111, puede obtenerse una probabilidad de error

$$P_E = 3 \times 10^{-4} \tag{6-29}$$

mientras la velocidad es de 1/3 de binit por segundo. Si, por el contrario, los ocho α_i son mensajes, la probabilidad de que un mensaje (no un binit) se transmita correctamente es $\bar{p}^3$. La probabilidad de mensaje erróneo es, entonces, $1 - \bar{p}^3$. Para $p = 0.01$, se obtiene

$$P_E \approx 3 \times 10^{-2} \tag{6-30}$$

La velocidad que corresponde a esta probabilidad de error es de 1 binit por segundo. Naturalmente entre estos dos extremos existen otras posibilidades. Pueden seleccionarse cuatro de los α_i para representar cuatro mensajes equiprobables. Sean, por ejemplo,

$$\begin{matrix} 000 \\ 011 \\ 101 \\ 110 \end{matrix} \tag{6-31}$$

Elegidos los cuatro α_i, puede aplicarse la regla de máxima posibilidad * de la figura 6-7.

---

* Como se vio en el ejemplo 6 2, la regla de máxima posibilidad no es única. En aquel caso existían otras reglas además de la mostrada en la figura 6-7.

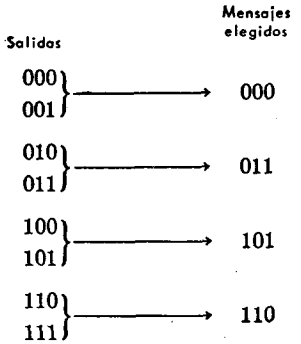


FIG. 6-7. Una regla de decisión de máxima posibilidad.

La probabilidad de interpretar correctamente un mensaje, $\bar{P}_E$, es precisamente igual a la probabilidad de transmitir sin error los dos primeros binits, es decir

$$\bar{P}_E = \bar{p}^2 \tag{6-32}$$

Para $p = 0.01$ se encuentra

$$P_E \approx 2 \times 10^{-2} \tag{6-33}$$

Puesto que las cuatro secuencias binarias utilizadas corresponden a dos mensajes binarios y se emplean 3 seg. en transmitir cada uno de ellos, la velocidad es de 2/3 binits por segundo. Comparando los resultados obtenidos al seleccionar dos, cuatro u ocho mensajes de las ocho entradas posibles del $(BSC)^3$ se comprueba que, en general, *la probabilidad de error aumenta con el número de mensajes utilizados.*

La extensión de orden n de una fuente de r símbolos tiene un total de r^n símbolos de entrada. Utilizando solamente M de ellos como mensajes, se disminuye la probabilidad de error. *El quid está en disminuir la probabilidad, y por tanto M, sin que la proporción o velocidad de los mensajes,* ($\log M/n$ *) *llegue a ser demasiado pequeña.* El

* La proporción o velocidad de los mensajes se mide por su equivalente, mensajes binarios por símbolo. Es decir, el envío de uno de los M mensajes posibles de n símbolos es equivalente a enviar M mensajes binarios de n símbolos a una velocidad de $(\log M)/n$ mensajes binarios por símbolo.

segundo teorema de Shannon dice que la probabilidad de error puede ser *tan pequeña como se quiera* en tanto que M sea inferior a 2^{nC}. Para este valor de M, la velocidad de mensaje es

$$\frac{\log M}{n} = C \tag{6-34}$$

Es decir, la capacidad de un canal coincide con el valor máximo de la *velocidad de mensaje sin error.*

6-6. Distancia de Hamming.

Los apartados 6-7 y 6-8 versan sobre la demostración del segundo teorema de Shannon en el caso particular de un BSC, donde se aprovecha la naturaleza binaria de los símbolos de entrada y salida para su simplificación. Hamming introdujo por primera vez (1950) el importante concepto de distancia entre dos secuencias binarias. La distancia de Hamming entre dos secuencias binarias, α_i y β_i, de la misma longitud, está definida por el número de lugares en que difieren. Sea, por ejemplo,

$$\alpha_i = 101111$$

$$\beta_i = 111100$$

y $D(\alpha_i, \beta_i)$ la distancia de Hamming entre α_i y β_i. Entonces $D(\alpha_i, \beta_i) = 3$.

Este concepto puede aplicarse a los tres códigos del $(BSC)^3$ tratados en el apartado anterior.

TABLA 6 2. TRES CÓDIGOS PARA UN $(BSC)^3$

	Código $\mathscr{A}$	*Código* $\mathscr{B}$	*Código* $\mathscr{C}$
	000	000	000
	001	011	111
	010	101	
	011	110	
	100		
	101		
	110		
	111		
Número de mensajes M:	8	4	2

Las palabras de los códigos dados en la tabla 6-2 pueden considerarse vértices de cubos tridimensionales. La distancia de Hamming entre dos palabras cualquiera, entonces, es igual al número de saltos que debe darse para pasar de uno a otro. Las distancias mínimas en los códigos $\mathscr{A}$, $\mathscr{B}$ y $\mathscr{C}$ son, respectivamente, 1, 2 y 3.

La distancia mínima entre palabras de un código está íntimamente relacionada con su probabilidad de error. En general, a mayor distancia mínima, la probabilidad de error será menor. Como es lógico, cuanto mayor es la distancia mínima, el número de palabras que puede alojarse en los vértices de un cubo de n dimensiones es menor, lo que no es sino expresión del resultado puesto de relieve en el apartado anterior. La ventaja de poder representar un gran número de mensajes con un código, por un lado, se equilibra, por el otro, con la de tener un canal de baja probabilidad de error.

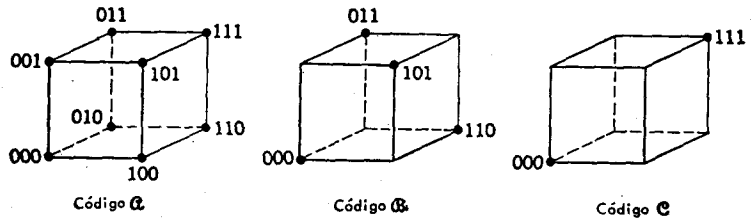


FIG. 6-8. Tres códigos diferentes de un $(BSC)^2$.

Los errores surgidos en la transmisión de una secuencia α_0 de n bits, a través de un $(BSC)^n$, dan lugar a que la secuencia recibida, β_j, sea distinta de ella. Si han aparecido D errores, la distancia de Hamming entre α_0 y β_j, será D.

$$D(\alpha_0, \beta_j) = D \tag{6-35}$$

El número *medio* de errores que se presentan en un grupo de n bits será np, siendo p la probabilidad de error del BSC. Así pues, la distancia media de Hamming entre una secuencia transmitida y una recibida será también np. Naturalmente la distancia que realmente existirá entre dos secuencias particulares raras veces coincidirá con la media. Según esto, deberá estudiarse el problema de determinar la secuencia transmitida que corresponde a una secuencia recibida, β_j; es decir, determinar la regla de decisión a aplicar.

A lo largo del capítulo se ha supuesto que todos los mensajes (y, por lo tanto, las palabras) son equiprobables. En el apartado 6-2 se demostró que cuando las entradas son equiprobables la probabilidad de error mínima corresponde a la aplicación de la regla de decisión de máxima posibilidad. A continuación se demostrará que esta regla admite una interpretación sencilla desde el punto de vista de la distancia de Hamming. Sea α_i la palabra transmitida y β_j una de las posibles secuencias de salida del canal. Supóngase, asimismo, que D es la distancia de Hamming entre esas dos secuencias binarias de longitud n. En ese caso α_i y β_j difieren exactamente en D lugares, y la probabilidad de recibir β_i al enviar α_i es precisamente la de que aparezca un error en cada uno de los D lugares en que difieren y no se produzca, en cambio, en los $n - D$ restantes.

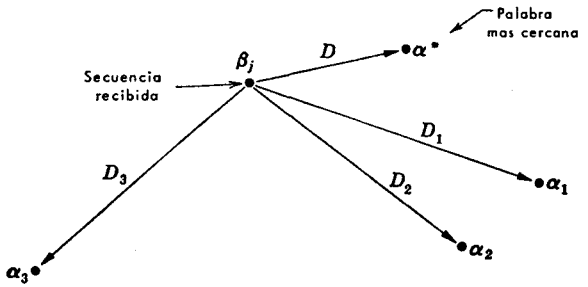


FIG. 6-9. Regla de máxima posibilidad de un $(BSC)^n$.

$$P(\beta_j/\alpha_i) = (p)^D\,(\bar{p})^{n-D} \tag{6-36}$$

Para $p < 1/2$ (único caso de interés), $P(\beta_j/\alpha_i)$ disminuye al aumentar D. Cuanto mayor sea la distancia entre β_j y la secuencia transmitida, menor será la probabilidad de recibirla. La regla de máxima posibilidad elige la palabra que hace máxima $P(\beta_j/\alpha_i)$; es decir, selecciona la palabra más cercana a β_j, según el concepto de distancia de Hamming.

6-7. El segundo teorema de Shannon aplicado a un BSC. Primer paso.

En este apartado se procederá a demostrar el segundo teorema de Shannon, en el caso particular de un BSC. La demostración general, válida para cualquier canal de información de memoria nula con un número finito de símbolos, se hará en el apartado 6-9.

Segundo teorema de Shannon (caso particular).

La probabilidad de error de un BSC es p, y en consecuencia su capacidad, $C = 1 - H(p)$. Sea ε un número positivo tan pequeño como se quiera, y $M = 2^{n(C-\epsilon)}$. Para n suficientemente grande puede formarse un subconjunto de M palabras (que representan M mensajes equiprobables) del conjunto de las 2^n posibles entradas del canal $(BSC)^n$, de manera que la probabilidad de error al decodificar la salida del canal puede ser tan pequeña como se quiera.

La figura 6-10 representa las 2^n entradas y salidas de la extensión de orden n de un BSC de probabilidad de error p.

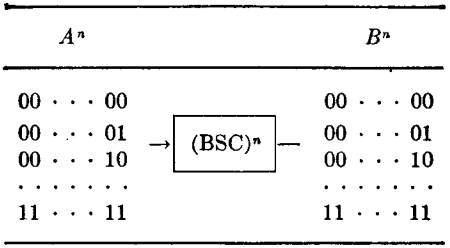


FIG. 6-10. Extensión de orden n de un BSC.

Las entradas y salidas del canal están constituidas por secuencias de n dígitos binarios. Con objeto de enviar M mensajes a través del canal, se seleccionan M de la 2^n entradas posibles. Según se dijo en el apartado 6-5, la probabilidad de mensaje erróneo, P_E, aumenta al crecer M. La pregunta a la que debe darse respuesta es «¿Cuántos mensajes es posible enviar manteniendo la probabilidad de error pequeña?»

La respuesta depende, como es natural, de la forma en que se seleccionen los símbolos que constituyen los mensajes. La probabilidad de error será mayor si las palabras elegidas están apiñadas, que si existe una distancia regular entre ellas. El procedimiento de codificación influye de manera notable sobre la probabilidad de error y, por lo tanto, sobre el número máximo de mensajes que pueden utilizarse. Sin embargo, se dejará por el momento a un lado esta importante cuestión, suponiendo que por un procedimiento cualquiera se ha seleccionado un código consistente en M palabras de n binits.

Al enviar a través del canal una de esas palabras, por ejemplo α_0, se recibe otra secuencia binaria de longitud n, β_j (figura 6-11).

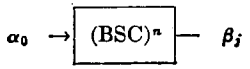


FIG. 6-11. El canal.

La regla de decisión de máxima posibilidad, descrita en los apartados anteriores, hace mínima la probabilidad de error si los mensajes se envían con la misma probabilidad. Sin embargo, esta regla es difícil de analizar, por lo que se hará uso de otra de similares características, que permite asimismo alcanzar una probabilidad de error tan pequeña como se quiera.

Se ha puesto ya de relieve que la distancia media entre las secuencias transmitidas y recibida, α_0 y β_j, es np, donde n es el orden de la extensión del BSC (o la longitud del bloque del código) y p su probabilidad de error. Al recibir un símbolo β_j, la inclinación natural tiende a buscar el símbolo transmitido entre aquellos (si es que existen) que se encuentran a una distancia np, o menor de β_j. Puede interpretarse en términos geométricos diciendo que se busca en el interior de una esfera de radio np trazada con centro en β_j. Ahora bien, np es solamente la distancia media entre α_0 y β_j, por lo que es prudente agrandar ligeramente la esfera para garantizar que α_0 se encontrará en su interior con gran probabilidad. Los matemáticos acostumbran a denominar ε a ese margen de seguridad, por lo que mantendremos este símbolo. Sea np_ε el radio de la esfera, donde $p_\varepsilon = p + \varepsilon$ (figura 6-12).

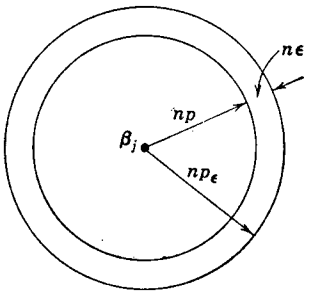


FIG. 6-12. Esfera con centro en el símbolo recibido.

El proceso de decisión consiste en dibujar la esfera de radio np_ε con centro en β_j, y, si no hay más que *un solo* punto (palabra) en su interior, decidir qué es el transmitido. Si no existe un solo punto (bien porque hay varios o ninguno) se elegirá simplemente al azar, cometiendo a ciencia cierta un error. El lector puede objetar que, en estas circunstancias, se ha procedido demasiado a la ligera. La observación es correcta. Sin embargo se demostrará que, aún siguiendo ese procedimiento, la probabilidad de error es despreciable.

Según el método descrito, al decodificar un símbolo recibido pueden presentarse dos casos de error. Se designará por $S(np_\varepsilon)$ la esfera de radio np_ε trazada alrededor del símbolo recibido, β_j (figura 6-13).

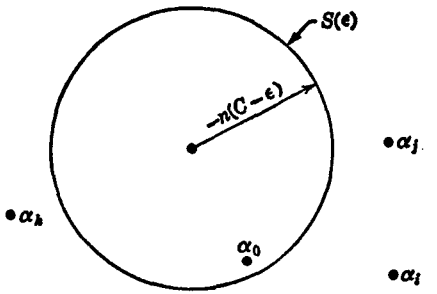


FIG. 6-13. Decodificación correcta de β_j.

El primer caso es aquel en que α_0, palabra transmitida, no se encuentra en $S(np_\varepsilon)$; el segundo, sí lo está, pero existe además otra palabra. La probabilidad de error puede escribirse en la forma

$$P_E = \Pr\{\alpha_0 \notin S(np_\varepsilon)\} + \Pr\{\alpha_0 \in S(np_\varepsilon)\} \times \Pr\{\text{al menos otra palabra} \in S(np_\varepsilon)\} \qquad (6\text{-}37)$$

donde $\in$ y $\notin$ significan «contenida en» y «no contenida en», respectivamente. Puesto que $\Pr\{\alpha_0 \notin S(np_\varepsilon)\} \leq 1$, la ecuación (6-37) implica que

$$P_E \leq \Pr\{\alpha_0 \in S(np_\varepsilon)\} + \Pr\{\text{al menos otra palabra} \in S(np_\varepsilon)\} \qquad (6\text{-}38)$$

La probabilidad de que ocurra al menos uno de los dos sucesos no es nunca mayor que la suma de las probabilidades de que ocurra cada

uno de ellos separadamente. Una generalización de esta ley conduce a $\Pr\{\text{al menos otra palabra} \in S(np_\varepsilon)\}$

$$\leq \sum_{\alpha_i \neq \alpha_0} \Pr\{\alpha_i \in S(np_\varepsilon)\} \qquad (6\text{-}39)$$

donde la suma del segundo miembro está extendida a las $M-1$ palabras no transmitidas. Sustituyendo (6-39) en (6-38) se encuentra la desigualdad buscada

$$P_E \leq \Pr\{\alpha_0 \notin S(np_\varepsilon)\} + \sum_{\alpha_i \neq \alpha_0} \Pr\{\alpha_i \in S(np_\varepsilon)\} \qquad (6\text{-}40)$$

La ecuación (6-40) define el límite de la probabilidad de error de un conjunto específico de M palabras. El primer término es la probabilidad de que las palabras transmitida y reciba *no se encuentren* a una distancia de Hamming inferior a $n(p+\varepsilon)$; el segundo, la suma de las probabilidades (una por palabra no transmitida) de que la palabra recibida y cada una de las *no* transmitidas *estén* a una distancia de Hamming inferior a $n(p+\varepsilon)$.

El primer término es fácil de evaluar. Es sencillamente la probabilidad de que se presenten más de $n(p+\varepsilon)$ errores en la transmisión de n binits a través de un BSC de probabilidad de error p. El número *medio* de errores en un grupo de n binits es np. Para cualquier valor finito de n existirá una probabilidad finita que el número de errores exceda del valor medio en $n\varepsilon$ o más. Al crecer n, sin embargo, la probabilidad disminuye. De forma más precisa, la ley de los números grandes (Parzen, 1961) dice que para dos números positivos cualesquiera, ε y δ, existe un n_0 tal que para cualquier $n > n_0$ la probabilidad de que el número de errores exceda a su valor medio en más de $n\varepsilon$ es menor que δ. Así, pues, tomando un n suficientemente grande, estaremos seguros de que

$$\Pr\{\alpha_0 \notin S(np_\varepsilon)\} < \delta \qquad (6\text{-}41)$$

con δ tan pequeño como queramos.

Esta ecuación reduce a la mitad el esfuerzo en la evaluación de la probabilidad de error (6-40), es decir el trabajo de demostrar el segundo teorema de Shannon. Sustituyendo (6-41) en (6-40) resulta

$$P_E \leq \delta + \sum_{\alpha_i \neq \alpha_0} \Pr\{\alpha_i \in S(np_\varepsilon)\} \qquad (6\text{-}42)$$

Hay que destacar que δ es independiente del conjunto de M pala-

bras elegido para representar los M mensajes. El último término de (6-42), por otro lado, depende fundamentalmente del código elegido. ¿En qué forma se hará uso de la relación (6-42) para encontrar el límite de la probabilidad de error, sin tener que afrontar el intrincado problema de qué código utilizar?

La respuesta a este último dilema fue aportada ingeniosamente por Shannon. En lugar de calcular (6-42) en el caso de un código particular, Shannon demostró la posibilidad de hallar su valor medio extendido a todos los códigos posibles. El primer término no depende del código. Los $M-1$ sumandos, sí. Calculando su valor medio extendido a todos los códigos posibles obtendremos la probabilidad media de error correspondiente a todos ellos. No es exactamente el procedimiento seguido, pero veremos que es suficiente para demostrar el teorema fundamental.

6-8. Codificación al azar. Segundo paso.

El razonamiento de Shannon, llamado algunas veces de la codificación al azar, es el siguiente. Las M palabras del código de entrada son elegidas al azar de un conjunto de 2^n. Puede imaginarse que los 2^n símbolos de entrada se han escrito sobre 2^n hojas de papel e introducidas en un sombrero. Con los ojos vendados se procede a elegir M papeles, teniendo buen cuidado de devolver al sombrero cada uno de ellos antes de la siguiente elección. Así, pues, los M papeles seleccionados definen las M palabras del código *. Al elegir una palabra existen 2^n posibilidades distintas. Puesto que seleccionamos n palabras consecutivas, el número total de códigos diferentes que pueden formarse es de 2^{nM}. Cada uno de ellos tiene una probabilidad 2^{nM} de ser elegido. La probabilidad de error que corresponde a cada uno de ellos viene determinada por la fórmula (6-42). La probabilidad media de error, $\bar{P}_E$, se obtendrá calculando el valor medio de (6-42) extendido a los 2^{Mn} códigos. Ya hemos indicado que δ, primer término del segundo miembro de (6-42), no depende del código elegido. Por lo tanto solamente será necesario hallar el valor medio extendido a $M-1$ términos de la forma $\Pr\{u_i \in S(np_\varepsilon)\}$, donde $u_i \neq u_0$. Si empleamos un

* Este procedimiento puede dar origen a un código singular; es decir, una hoja puede elegirse más de una vez, utilizándose, en definitiva, la misma palabra para mensajes diferentes. Para $M \ll 2^n$ tal cosa es posible pero improbable. Si $M > 2^n$, en cambio, resulta inevitable.

trazo ondulado para indicar el valor medio de los 2^{Mn} códigos, la ecuación (6-42) puede expresarse en la forma

$$\widetilde{P_E} \leq \delta + (M-1)\,\widetilde{\Pr\{\alpha_i \in S(np_\varepsilon)\}}$$
$$\leq \delta + M\,\widetilde{\Pr\{\alpha_i \in S(np_\varepsilon)\}} \quad \alpha_i \neq \alpha_0 \tag{6-43}$$

Emplearemos el mismo procedimiento de codificación utilizado para generar α_i, para evaluar $\widetilde{\Pr\{\alpha_i \in S(np_\varepsilon)\}}$, $\alpha_i \neq \alpha_0$. Los α_0 se eligieron al azar entre los 2^n códigos posibles; por lo tanto, la probabilidad de que α_i, una palabra distinta de la palabra transmitida α_0, esté contenida en una esfera de radio np_ε trazada alrededor de la secuencia recibida β_j, es igual al cociente entre $N(np_\varepsilon)$, número de secuencias binarias diferentes contenidas en la esfera, y 2^n, número de secuencias binarias de longitud n diferentes.

$$\widetilde{\Pr\{\alpha_i \in S(np_\varepsilon)\}} = \frac{N(np_\varepsilon)}{2^n} \quad \alpha_i \neq \alpha_0 \tag{6-44}$$

Finalmente, calcularemos el límite de $N(np_\varepsilon)$. El número de secuencias binarias de longitud n situadas a distancia k de β_j es precisamente igual al número de maneras posibles en que una secuencia binaria de longitud n puede diferir de β_j en k lugares, es decir $\binom{n}{k}$ Sumando para todos los valores de k menores o iguales a np_ε, se obtiene *

$$N(np_\varepsilon) = 1 + \binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{np_\varepsilon}$$
$$\sum_{k=0}^{(np_\varepsilon)} \binom{n}{k} \tag{6-45}$$

Esta suma puede acotarse introduciendo una desigualdad frecuentemente utilizada en la teoría de la información (Peterson, 1961, p. 246; Wozencraft and Reiffen, 1961, p. 71):

$$\sum_{k=0}^{(np_\varepsilon)} \binom{n}{k} \leq 2^{nH(p_\varepsilon)} \quad \text{para} \quad p_\varepsilon < \tfrac{1}{2} \tag{6-46}$$

---

* Naturalmente, np_ε no tiene por qué ser un número entero. Según esto, lo reemplazaremos por el mayor entero inferior a np_ε, sin que la demostración pierda valor en ningún aspecto.

Así, pues, combinando (6-44), (6-45) y (6-46), se obtiene

$$\widetilde{\Pr}\{\alpha_i \in \widetilde{S}(np_\varepsilon)\} \leq 2^{-n[1-H(p_\varepsilon)]} \qquad \alpha_i \neq \alpha_0 \tag{6-47}$$

que, llevando a (6-43), da lugar a la acotación

$$\widetilde{P_E} \leq \delta + M\, 2^{-n[1-H(p_\varepsilon)]} \tag{6-48}$$

La ecuación (6-48) contiene la esencia del segundo teorema de Shannon (en el caso particular de un BSC). El parámetro δ puede hacerse tan pequeño como se quiera aumentando la longitud n de los bloques. Por lo tanto, el segundo miembro de (6-48) puede hacerse tan pequeño como se quiera, siempre que

$$M < 2^{n[1-H(p_\varepsilon)]} < 2^{n[1-H(p)]} \tag{6-49}$$

que constituye la expresión buscada. Tomando un ε pequeño,

$$H(p_\varepsilon) = H(p + \varepsilon)$$

puede alcanzar un valor muy cercano a $H(p)$ y podrá elegirse un número de mensajes tan próximo a $2^{n[1-H(p)]}$ como se desee. Ahora bien, $1 - H(p)$ es la capacidad del BSC. Por lo tanto, podrán elegirse M mensajes, siendo M cualquier número inferior a 2^{nC}, y la probabilidad media de error ser inferior a cualquier valor predeterminado. Al menos existirá un código tan bueno como la media, de forma que puede afirmarse que hay un código de $M < {}^{nC}$ palabras y probabilidad de error arbitrariamente pequeña.

Este es el resultado anunciado al final del apartado 6-5. Si en un BSC se emplea una longitud de bloque n suficientemente larga, pueden elegirse M palabras ($M < 2^{nC}$), y tener aún una probabilidad de no identificar una palabra, tan pequeña como queramos. Por tanto, por BSC de capacidad C, puede enviarse por cada binit hasta [ver (6-34]

$$\frac{\log 2^{nC}}{n} = C \tag{6-50}$$

mensajes binarios sin error.

6-9. Segundo teorema de Shannon .Discusión.

El teorema demostrado en los apartados anteriores es válido en un caso muy particular. El canal considerado, un BSC, era de alcance muy

limitado. Sin embargo, la mayor parte de los conceptos necesarios para demostrar el teorema en su generalidad, así como las importantes consecuencias que de él se derivan, han aparecido de forma más o menos evidente en esos apartados. En este discutiremos esos conceptos, procediendo a la demostración del teorema en el apartado siguiente.

El primer concepto introducido por Shannon es el de *codificación al azar*. Si deseamos apreciar las limitaciones del teorema es necesario comprender antes tal procedimiento de codificación. Puesto que las palabras del código se eligen al azar, podrá aplicarse la ecuación (6-47) para acotar la probabilidad de que una palabra cualquiera se encuentre dentro de una esfera de centro en β_j y radio np_ε. Si las palabras se han elegido por algún otro procedimiento, no podrá hablarse de la probabilidad de que una palabra se encuentre a una distancia inferior a np_ε de la secuencia recibida β_j. Analizando esta cuestión más en detalle, el procedimiento de codificación descrito puede definirse como la falta absoluta de procedimiento.

Desde un punto de vista práctico, la codificación al azar deja mucho que desear. La probabilidad media de error puede hacerse tan pequeña como se quiera. Este valor medio, desgraciadamente, se refiere a la totalidad de códigos posibles. Así, una vez determinado un código, no puede afirmarse que se trate de un buen código. Como caso extremo citaremos la posibilidad de obtener un código en que los M mensajes correspondan a la misma palabra.

El segundo teorema de Shannon se caracteriza por ser algo más que una prueba de la existencia del código y algo menos que una regla práctica para encontrarlo. El teorema no dice exactamente cómo construir un buen código, por lo que realmente no define un método para su determinación. Por otro lado, sin embargo, el teorema enuncia un procedimiento que por *término medio* da lugar a buenos códigos; así, pues, no se limita a la mera demostración de su existencia.

En la versión generalizada del segundo teorema de Shannon, que demostraremos en el apartado siguiente, veremos que pueden seleccionarse $M = 2^{n(C-\varepsilon)}$ palabras, $\varepsilon > 0$, (donde C es la capacidad del canal), y ser aún la probabilidad de error tan pequeña como se quiera. Se demostrará también una transformación del teorema: si se eligen $M = 2^{n(C+\varepsilon)}$, $\varepsilon > 0$, palabras, no es posible encontrar una regla de decisión que dé lugar a una probabilidad de error arbitrariamente pequeña, P_E, aumentando n, longitud de bloque del código. Esta manera de expresar la transformación del teorema es suficiente para nuestro

propósito. Hay que añadir, no obstante, que pueden demostrarse otras expresiones más potentes. Wolfowitz (1959) demostró que eligiendo $M = 2^{n(C+\epsilon)}$ palabras (C, capacidad del canal y $\epsilon > 0$), la probabilidad de error óptima tiende a la *unidad,* al crecer n.

El teorema de la codificación dice que la probabilidad de interpretar mal una palabra, enviada a través de un canal con ruido, puede hacerse tan pequeña como se quiera. La importancia de esta frase reposa fundamentalmente en el hecho de que se refiere tanto a la probabilidad de error de los mensajes como de las *palabras* del código. En el caso de un BSC, por ejemplo, el teorema establece que la probabilidad de interpretar mal una secuencia de n ceros y unos es arbitrariamente pequeña. Argumento más eficaz que la mera afirmación de que la probabilidad de interpretar mal un simple *binit* es arbitrariamente pequeña. Esta distinción ha dado lugar a más de una mala interpretación de las conclusiones derivadas de las diversas formas de la transformación del segundo teorema de Shannon. Refiriéndose a un BSC, la transformación afirma que si el número de mensajes equiprobables M es superior a 2^{nC} (donde C es nuevamente la capacidad del BSC), la probabilidad de error en una palabra tiende a la unidad al crecer n. Esta conclusión es válida para cualquier conjunto de palabras (no solamente como término medio en un grupo de códigos) y cualquier regla de decisión. El teorema presenta un gran interés matemático y su importancia en relación con el problema de la comunicación ha sido puesta de manifiesto repetidas veces. El teorema no afirma que la comunicación sea imposible si $M > 2^{nC}$. Para aclarar este punto consideremos un BSC en que los binits 0 y 1 se eligen con la misma probabilidad y dibujemos la variación de la probabilidad de un *binit* erróneo en función de la cantidad de mensajes por unidad.

Supuesta una velocidad de mensaje cualquiera, R mensajes binarios por binit, menor que C, capacidad del canal, sabemos que la probabilidad de un binit erróneo puede hacerse tan pequeña como se quiera. Si R es mayor que C, puede imaginarse el siguiente proceder, consistente en emplear la extensión de orden n del BSC y hacer crecer n. Para que la velocidad sea de R mensajes por binit, deberán disponerse 2^{nR} mensajes para enviar a través de la extensión n del BSC. Se enviarán alternativamente nR binits. Pueden transmitirse hasta nC binits con probabilidad de error arbitrariamente pequeña. El receptor decidirá si los $nR - nC$ binits restantes son $0's$ ó $1's$ lanzando simplemente una moneda al aire. Cara será un 0, cruz un 1. La probabilidad

de error de *esos* binits será igual a 1/2. La probabilidad media de error, tanto para los binits confiables como para los demás, será ligeramente superior a $1/2\ (R\text{-}C)/R$. La figura 6-14 representa el resultado.

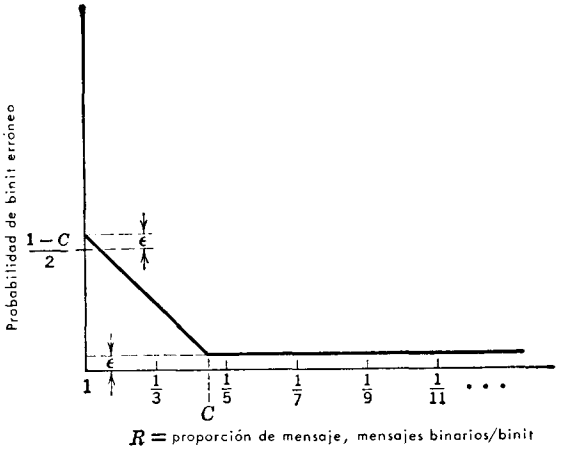


FIG. 6-14. Proporción de binits erróneos en función de la proporción de mensajes en un BSC.

La parte de la figura que corresponde a $R > C$ se ha obtenido por el procedimiento indicado. No se ha demostrado, sin embargo, que sea el mejor. De hecho, el cálculo de la probabilidad mínima de error cuando la proporción de mensajes es $R > C$ está aún por resolver. Hay que notar, además, que aunque se ha determinado la abscisa correspondiente a $R = 1$, puede alcanzarse una velocidad de mensaje mayor mediante el procedimiento de la moneda descrito. Consideremos, por ejemplo, un BSC sin ruido ($p = 0$). El procedimiento de la moneda da lugar a una posibilidad de binit erróneo de 0.25 con una proporción de 2 mensajes binarios por binit.

Antes de entrar en la demostración general del segundo teorema de Shannon analizaremos los límites de la probabilidad de error. Tanto en la demostración general como en la particular, aplicada a un BSC, estamos únicamente interesados en probar que la probabilidad de error puede hacerse arbitrariamente pequeña cuando $M \leq 2^{n(C-\epsilon)}$. Se han obtenido, no obstante, otros resultados, que definen la velocidad con la que la probabilidad de error tiende a 0 al aumentar n, orden de la ex-

tensión utilizada. Nos limitaremos simplemente a indicar que otros autores han demostrado que la probabilidad varía exponencialmente (o casi exponencialmente) con n. Las notas del final del capítulo contienen referencias de estos trabajos.

6-10. Segundo teorema de Shannon. Caso general.

El problema se reducirá a la demostración del teorema de Shannon en el caso de un canal discreto sin memoria. Conceptualmente la demostración no difiere casi en absoluto de la presentada en los apartados 6-7 y 6-8, correspondiente a un BSC.

Segundo teorema de Shannon.

Consideremos un canal de r entradas, s salidas y capacidad C. Sea ε un número arbitrariamente pequeño y $M = 2^{n(C-\varepsilon)}$. Para un n suficientemente grande, es posible seleccionar un conjunto de M palabras (que representarán M mensajes equiprobables) entre las r^n entradas posibles de la extensión n del canal, tales que la probabilidad de error al decodificar la salida sea tan pequeña como se quiera.

La figura 6-15 representa las r^n entradas y s^n salidas posibles de la extensión n del canal.

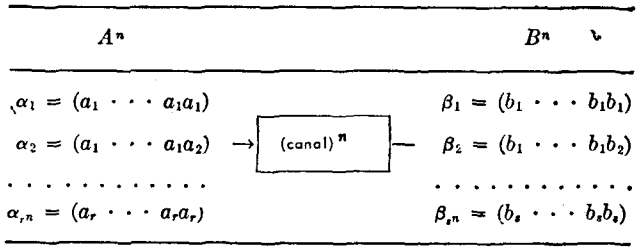


FIG. 6-15. Extensión de orden n de un canal.

Los M mensajes a enviar se seleccionan entre las r^n entradas. Nuevamente nos planteamos la pregunta, «¿Cuántos mensajes pueden enviarse manteniendo aún pequeña la probabilidad de error?»

Al enviar una palabra, α_0, a través del canal, se encuentra una salida, β_j (figura 6-16).

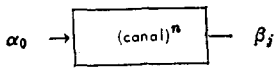


FIG. 6-16. El canal.

Puesto que los M mensajes se suponen equiprobables, la regla de decisión que da lugar a la probabilidad de error mínima es la de máxima posibilidad

$$d(\beta_j) = \alpha^* \tag{6-51a}$$

donde

$$P(\beta_j/\alpha^*) \geq P(\beta_j/\alpha_i) \quad \text{para cualquier } i \tag{6-51b}$$

Nuevamente encontramos conveniente calcular la probabilidad de error utilizando una regla de decisión íntimamente relacionada con la de máxima posibilidad, en lugar de ella misma. Para definir el parámetro a escribiremos una condición equivalente a (6-51*b*). Puesto que el logaritmo es una función monótona. (5-61*b*) puede sustituirse por

$$\log(P(\beta_j/\alpha^*) \geq \log P(\beta_j/\alpha_i) \quad \text{para cualquier } i \tag{6-52a}$$

o

$$\log \frac{1}{P(\beta_j/\alpha^*)} \leq \log \frac{1}{P(\beta_j/\alpha_i)} \quad \text{para cualquier } i \tag{6-52b}$$

Supongamos que $P_0(\beta_j)$ representa la distribución de probabilidades del conjunto de secuencias de salida que aparece si las secuencias de entrada se eligen de acuerdo con la ley de probabilidad que corresponde a la capacidad del canal. [Las entradas, como es natural, *no* se seleccionan de acuerdo con esta ley; por esta razón se ha introducido el subíndice para distinguir $P_0(\beta_j)$ de $P(\beta_j)$, distribución real de β_j]. Añadiendo $\log P_0(\beta_j)$ a ambos miembros de (6-52*b*)

$$\log \frac{P_0(\beta_j)}{P(\beta_j/\alpha^*)} \leq \log \frac{P_0(\beta_j)}{P(\beta_j/\alpha_i)} \quad \text{para cualquier } i \tag{6-53}$$

La cantidad

$$\log \frac{P_0(\beta_j)}{P(\beta_j/\alpha_i)}$$

juega el mismo papel que la distribución de Hamming en la demostración correspondiente al BSC. Para una secuencia transmitida dada, α_0, el valor medio de esta nueva «distancia» entre α_0 y la secuencia recibida es

$$\sum_{B^n} P(\beta_i/\alpha_0) \log \frac{P_0(\beta_j)}{P(\beta_j/\alpha_i)} \tag{6-54}$$

La razón de la introducción de $P_0(\beta_j)$ en (6-53) aparece ahora con claridad. La suma (6-54) es igual a la información mutua $I(\alpha_0; B^n)$, definida en el apartado 5-13, cambiada de signo. Puesto que $P_0(\beta_j)$ son las probabilidades de β_j cuando $I(A^n; B^n)$, la capacidad de la extensión de orden n, $I(\alpha_0; B^n)$, es independiente de α_0, por lo que tendremos

$$\sum_{B} P(\beta_j/\alpha_0) \log \frac{P_0(\beta_j)}{P(\beta_j/\alpha_i)} = -nC \quad \text{para cualquier } \alpha_0 \tag{6-55}$$

Por lo tanto, al recibir un símbolo β_j, la inclinación natural es de buscar el símbolo transmitido entre aquellos (si es que existen) que cumplen la condición

$$\log \frac{P_0(\beta_i)}{P(\beta_j/\alpha_i)} \approx -nC \tag{6-56}$$

Geométricamente se expresa trazando una esfera * alrededor de la secuencia β_j recibida. La esfera contiene todas las palabras α_i que satisfacen la condición

$$\log \frac{P_0(\beta_j)}{P(\beta_j/\alpha_i)} \leqq -nC \tag{6-57}$$

Así, pues, se procede a buscar la secuencia α_0 en el interior de esta esfera. Igual que antes, como margen de seguridad, se suma una cantidad ε, de forma que todas las palabras que cumplen

$$\log \frac{P_0(\beta_j)}{P(\beta_j/\alpha_i)} \leqq -nC + n\varepsilon = -n(C - \varepsilon) \tag{6-58}$$

estarán contenidas en ella.

---

* La palabra *esfera* se introduce únicamente para indicar la analogía existente con el caso del BSC. El *radio* de la «esfera» es negativo.

La regla de decisión consistirá en dibujar la esfera definida por la relación (6-58), decidiendo automáticamente cuál es el símbolo transmitido cuando no existe más que *un solo* punto en su interior. Si el punto no es único (bien porque no haya ninguno o más de uno), se elige al azar, cometiendo un error. La probabilidad de error resultante es despreciable.

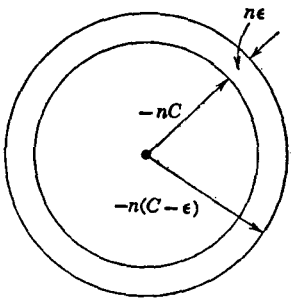


FIG. 6-17. Una esfera que incluye todos los puntos del código que satisfacen las ecuaciones (6-57) y (6-58).

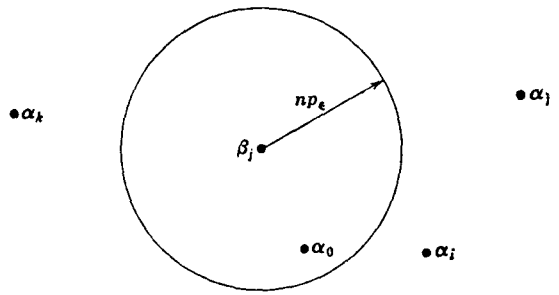


FIG. 6-18. Decodificación correcta de β_j.

Siguiendo este método, un error puede presentarse por dos caminos distintos. Llamemos $S(\varepsilon)$ al conjunto de puntos que satisfacen la relación (6-58) (a saber, los puntos contenidos en la esfera de la figura 6-17). El primero, si α_0, palabra transmitida, no está contenida en $S(\varepsilon)$; el otro, si α_0 pertenece a $S(\varepsilon)$, pero existe además alguna otra palabra que cumpla la misma condición (figura 6-18). La probabilidad

de error, por tanto, será

$$P_E = \Pr\{\alpha_0 \notin S(\varepsilon)\} + \Pr\{\alpha_0 \in S(\varepsilon)\} \times \Pr\{\text{al menos otra palabra} \in S(\varepsilon)\} \quad (6\text{-}59)$$

donde $\in$ y $\notin$ significan «está contenida en» y «no está contenida en», respectivamente. Con los mismos argumentos utilizados para llegar a (6-38), (6-39) y (6-40), se obtiene

$$P_E \leq \Pr\{\alpha_0 \notin S(\varepsilon)\} + \{\text{al menos otra palabra} \in S(\varepsilon)\} \quad (6\text{-}60)$$

$$\Pr\{\text{al menos otra palabra} \in S(\varepsilon)\} \leq \sum \Pr\{\alpha_i \in S(\varepsilon)\} \quad (6\text{-}61)$$

$$P_E \leq \Pr\{\alpha_0 \notin S(\varepsilon)\} + \sum \Pr\{\alpha_i \in S(\varepsilon)\} \quad (6\text{-}62)$$

La ecuación (6-62) es una acotación inmediata de la probabilidad de error de un conjunto específico de M palabras. El primer término representa la probabilidad de que el código transmitido, α_0, no satisfaga la relación (6-58); el segundo, la suma de las probabilidades de que cada una de las palabras no transmitidas satisfaga la misma condición.

Acotaremos el primer término por el mismo procedimiento que en la demostración anterior; el segundo término se evaluará aplicando el razonamiento de Shannon de la codificación al azar. Como ya se ha visto, el valor medio de

$$\log \frac{P_0(\beta_j)}{P(\beta_j/\alpha_0)}$$

es $-n^c$. Este logaritmo puede descomponerse en suma de n términos, cada uno de los cuales atañe a uno de los n símbolos que comprende β_j y a uno de los n que comprende α_0. Por lo tanto, tomando un valor de n suficientemente grande, la probabilidad de que la suma exceda de $-nC$ en más de $n\varepsilon$, puede hacerse menor que una cantidad δ, tan pequeña como queramos. La ecuación (6-62) se transforma en

$$P_E \leq \delta + \sum_{\alpha_i \neq \alpha_0} \Pr\{\alpha_i \in S(\varepsilon)\} \quad (6\text{-}63)$$

A continuación aplicaremos el razonamiento de la codificación del azar. Sean $P_0(\alpha_i)$ las probabilidades de entrada que corresponden a la capacidad del canal, de acuerdo con las cuales se seleccionan M pa-

labras (se admite la posibilidad de un código singular). Esta vez los r^{Mn} códigos posibles no son necesariamente equiprobables; la probabilidad de seleccionar un conjunto determinado de M palabras es el producto de las M probabilidades correspondientes.

El límite de la probabilidad media de error, $\widetilde{P_E}$, se obtiene calculando el valor medio de (6-63) extendido a los r^{Mn} códigos posibles. Empleando una línea ondulada para indicar el valor medio sobre los r^{Mn} códigos, se obtiene

$$P_E \leq \widetilde{\delta + \sum_{\alpha_i \neq \alpha_0} \Pr\{\alpha_i \in S(\varepsilon)\}}$$

$$\leq \delta + (M-1)\widetilde{\Pr\{\alpha_i \in S(\varepsilon)\}}$$

$$\leq \delta + M\widetilde{\Pr\{\alpha_i \in S(\varepsilon)\}} \qquad (6\text{-}64)$$

Hasta este punto ha existido un marcado paralelismo entre esta demostración y la correspondiente a un BSC. Sin embargo, para evaluar $\widetilde{\Pr\{\alpha_i \in S(\varepsilon)\}}$, es necesario introducir un nuevo argumento. $\widetilde{\Pr\{\alpha_i \in S(\varepsilon)\}}$ es la probabilidad media de que α_i esté contenida en $S(\varepsilon)$. Para un β_j dado, esta cantidad puede escribirse como $\sum_{S(\varepsilon)} P_0(\alpha_i)$. Ahora bien, $S(\varepsilon)$ depende de β_j, de modo que la expresión buscada [suponiendo que $P_0(\beta_j)$ representa las probabilidades de salida correspondientes a $P_0(\alpha_i)$] es

$$\widetilde{\Pr\{\alpha_i \in S(\varepsilon)\}} = \sum_{B^n} P_0(\beta_j) \sum_{S(\varepsilon)} P_0(\alpha_i)$$

$$\sum_{B^n, S(\varepsilon)} P_0(\beta_j) P_;(a_i) \qquad (6\text{-}65)$$

La suma del segundo miembro está extendida a todas las parejas α_i, β_j, tales que

$$\log \frac{P_0(\beta_j)}{P(\beta_j/\alpha_i)} \leq -n(C-\varepsilon) \qquad (6\text{-}66)$$

Para las que se cumple

$$P_0(\beta_j) P_0(\alpha_i) \leq P(\beta_j/\alpha_i) P_0(\alpha_i) 2^{-n(C-\varepsilon)} \qquad (6\text{-}67)$$

Sumando (6-67) para todas esas parejas, encontramos

$$\sum_{B^n,S(\varepsilon)} P_0(\beta_j)P_0(\alpha_i) \leq 2^{-n(C-\varepsilon)} \sum_{B^n,S(\varepsilon)} P(\beta_j/\alpha_i)\,P_0(\alpha_i)$$

$$\leq 2^{-n(C-\varepsilon)} \tag{6-68}$$

Sustituyendo (6-68) y (6-65) en (6-64), se obtiene

$$\boxed{\tilde{\tilde{P}}_E \leq \delta + M\,2^{-n(C-\varepsilon)}} \tag{6-69}$$

La ecuación (6-69) constituye la médula del segundo teorema de Shannon. El parámetro puede hacerse tan pequeño como se quiera aumentando n, longitud de bloque. En consecuencia, el segundo miembro de (6-69) puede hacerse tan pequeño como se quiera, siempre que

$$M \leq 2^{n(C-\varepsilon')} < 2^{n(C-\varepsilon)} \tag{6-70}$$

para cualquier valor de $\varepsilon' < \varepsilon < 0$. Esta es la expresión buscada. ε, y por tanto ε', puede elegirse arbitrariamente pequeño. Entonces, si M satisface la expresión (6-70), la probabilidad media de error, P_E, puede llegar a ser inferior que cualquier valor predeterminado. Al menos existirá un código tan bueno como la media; así, pues, puede asegurarse que hay un código de probabilidad de error arbitrariamente pequeña con casi 2^{nC} palabras. En consecuencia, por cada símbolo de un canal de capacidad C, pueden enviarse hasta

$$\frac{\log 2^{nC}}{n} = C \tag{6-71}$$

mensajes binarios sin error.

La transformación del segundo teorema de Shannon se demuestra mediante el límite de Fano (6-23). Se desea probar que si se usan $M = 2$ palabras para representar m mensajes equiprobables, la probabilidad no puede hacerse arbitrariamente pequeña al aumentar n. Supongamos que se emplean $M = 2^{n(C+\varepsilon)}$ palabras, con la misma probabilidad $1/M$. Entonces, puesto que

$$H(A^n) - H(A^n/B^n) \leq nC \tag{6-72}$$

tendremos

$$\log 2^{n(C+\varepsilon)} - H(A^n/B^n) \leq nC$$

o

$$n\varepsilon \leq H(A^n/B^n) \tag{6-73}$$

Pero, según la inecuación de Fano,

$$\begin{aligned} H(A^n/B^n) &\leq H(P_E) + P_E \log M \\ &\leq 1 + P_E(nC + n\varepsilon) \end{aligned} \tag{6-74}$$

Sustituyendo (6-74) en (6-73), se encuentra

$$P_E \geq \frac{n\varepsilon - 1}{nC + n\varepsilon} \tag{6-75}$$

Al crecer n, el límite inferior de la probabilidad de error de un código se aleja de 0. Así, pues, con una velocidad que exceda de la capacidad del canal no pueden transmitirse mensajes sin error.

6-11. Epílogo.

En el apartado anterior se ha demostrado que seleccionando al azar un cierto número de palabras de longitud n para transmitir por un canal de capacidad C, la probabilidad de error será pequeña siempre que su número sea inferior a 2^{nC}. Inmediatamente se plantea la cuestión: «¿Cómo encontrar el código que corresponde a la confiabilidad definida por el segundo teorema de Shannon?».

Naturalmente, puede recurrirse para elegir las palabras del código a una tabla de números al azar. Este método, sin embargo, no se presta en forma óptima al diseño de un sistema de comunicación. El equipo que requiere es prácticamente irrealizable. Por otra parte existe siempre la posibilidad (verdaderamente ínfima) de que el código resultante no dé lugar a una probabilidad de error pequeña. El segundo teorema de Shannon ha demostrado que casi todos, pero no todos, la tienen (en definitiva, se ha demostrado eligiendo un código al azar). ¿Puede, entonces, no existir un método para generar buenos códigos?

Este dilema persiste desde que Shannon publicó su artículo en 1948. A pesar del enorme esfuerzo desarrollado desde entonces (Peterson, 1961), en la aclaración de esta incógnita de la teoría de la información, aún no se ha encontrado el método definitivo para generar los códigos intuidos y anunciados por Shannon.

NOTAS

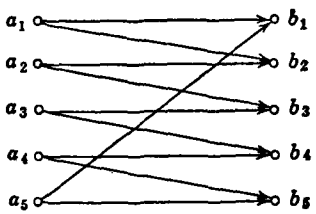
Nota 1. Continuando con la correspondencia existente entre los canales de información y los experimentos estadísticos, puesta de manifiesto en la Nota 1 del final del capítulo anterior, el segundo teorema de Shannon puede considerarse como una consecuencia de las propiedades asintóticas de dichos experimentos (Abramson, 1960).

Nota 2. Tal como se mencionó en el apartado 6-9, la bibliografía estadística contiene una buena cantidad de resultados que demuestran que para velocidades de mensaje inferiores a la capacidad del canal, la probabilidad de error tiende exponencialmente (o casi exponencialmente) a cero al aumentar la longitud n de bloque. Feinstein (1955) fue el primero en demostrar el límite de error exponencial. Además, su límite se aplicaba a la probabilidad máxima de error y no solamente a la probabilidad media. El límite más sencillo es quizá el debido a Blackwell, Breiman y Thomasian (1959). Utilizando una técnica debida a Chernoff (1952) y una ecuación equivalente a la (6-63), obtuvieron

$$P_E \leqq 2 \exp \left[-\frac{(C-R)^2}{16rs} n \right] \quad \text{para } 0 \leqq C - R \leqq 1/2$$

donde C es la capacidad del canal, R la velocidad de mensaje, r el número de símbolos de entrada, s el de salida y n la longitud de bloque.

Nota 3. La capacidad de un canal constituye el límite superior del conjunto de velocidades de mensaje con que puede enviarse una información con una probabilidad de error *aproximadamente* igual a cero. En ciertas circunstancias, la probabilidad puede ser *igual* a cero. Shannon (1956) definió el menor límite superior de velocidad de mensaje que permite transmitir con probabilidad de error nula. Consideremos, por ejemplo, el canal donde las probabilidades asociadas con cada flecha son arbitrarias y cumplen la condición $0 < P_{ij} < 1$. Entonces, puesto



que los símbolos a_1 y a_3 se transmiten con probabilidad de error nula, la capacidad sin error es de al menos un bit. El límite puede mejorarse empleando la segunda extensión del canal. En este caso, a_1a_1, a_2a_3, a_3a_5, a_4a_2 y a_5a_4 se transmiten con probabilidad de error nula, luego la capacidad sin error es al menos igual a 1/2 og 5 bits.

PROBLEMAS

6-1. Un canal uniforme tiene r entradas, que se eligen con la misma probabilidad. La regla de decisión de máxima posibilidad da lugar a una probabilidad de error p. Calcular el límite inferior de la equivocación $H(A/B)$ en función de r o p, o de ambos. El valor 0 no se tomará en cuenta.

6-2. Definir las tres reglas de máxima posibilidad del canal (6-2).

BIBLIOGRAFIA

Abramson, N. (1960): A Partial Ordering for Binary Channels, *IRE Trans. Inform. Theory*, vol. 6, no. 5, pp. 529-539, December.

Bar-Hillel, Y., and R. Carnap (1952): Semantic Information, in Willis Jackson (ed.), «Communication Theory», Academic Press Inc., New York.

Basharin, G. P. (1959): On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables, *Theory Probability Appl.*, vol. 4, no. 3, pp. 333-336.

Bell, D. A. (1953): «Information Theory and Its Engineering Applications», Sir Isaac Pitman & Sons, Ltd., London.

Bellman, R. (1960): «Introduction to Matrix Analysis», McGraw Hill Book Company, Inc., New York.

Bharucha-Reid, A. T. (1960): «Elements of the Theory of Markov Processes and Their Applications», McGraw-Hill Book Company, Inc., New York.

Billingsley, P. (1961): On the Coding Theorem for the Noiseless Channel, *Ann. Math. Statist.*, vol. 32, no. 2, pp. 576-601.

Birnbaum, A. (1961): On the Foundations of Statistical Inference: Binary Experiments, *Ann. Math. Statist.*, vol. 32, no. 2, pp. 414-435, June.

Blachman, N. M. (1961): A Generalization of Mutual Information, *Proc. IRE*, vol. 49, no. 8, pp. 1331-1332, August.

Blackwell, D. (1953): Equivalent Comparisons of Experiments, *Ann. Math. Statist.*, vol. 24, pp. 265-272, June.

———, L. Breiman, and A. J. Thomasian (1958): Proof of Shannon's Transmission Theorem for Finite-state Indecomposable Channels, *Ann. Math. Statist.*, vol. 29, no. 4, pp. 1209-1220, December.

———, ———, and ——— (1959): The Capacity of a Class of Channels, *Ann. Math. Statist.*, vol. 30, pp. 1229-1241, December.

———, ———, and ——— (1960): The Capacities of Certain Channel Classes under Random Coding, *Ann. Math. Statist.*, vol. 31, pp. 558-567, September.

Blyth, C. R. (1958): Note on Estimating Information, *Tech. Rept.* 17, Department of Statistics, Stanford University.

Breiman, L. (1957): The Individual Ergodic Theorem of Information Theory, *Ann. Math. Statist.*, vol. 28, no. 3, pp. 809-811; a correction to this paper is published in *Ann. Math. Statist.*, vol. 31, no. 3, pp. 809-810.

Brillouin, L. (1956): «Science and Information Theory», Academic Press Inc., New York.

Chernoff, H. (1952): A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations, *Ann. Math. Statist.*, vol. 23, pp. 493-507.

Cherry, C. (1957): «On Human Communication», John Wiley & Sons, Inc., New York.

Csiszar, I. (1961): Some Remarks on the Dimension and Entropy of Random Variables, *Acta Math. Acad. Sci. Hung.*, vol. 12, pp. 399-408.

Elias, P. (1953): Optics and Communication Theory, *J. Opt. Soc. Am.*, vol. 43, pp. 229-232, April.

——— (1958): Two Famous Papers, *IRE Trans. Inform. Theory*, vol. 4, n. 3, p. 99, September.

Fano, R. (1949): The Transmission of Information, I, *MIT Res. Lab. Electron. Tech. Rept.* 65.

——— (1950): The Transmission of Information, II, *MIT Res. Lab. Electron. Tech. Rept.* 149.

——— (1961): «Transmission of Information», John Wiley & Sons, Inc., New York.

Feinstein, A. (1955): Error Bounds in Noisy Channels without Memory, *IRE Trans. Inform. Theory*, vol. IT-1, no. 2, pp. 13-14, September.

——— (1958): «Foundations of Information Theory», McGraw-Hill Book Company, Inc., New York.

Feller, W. (1950): «Probability Theory and Its Applications», John Wiley & Sons, Inc., New York.

Gel'fand, I. M.. and A. M. Yaglom (1957): Computation of the Amount of Information about a Stochastic Function Contained in Another Such Function, *Usp. Mat. Nauk*, vol. 12, no. 1, pp. 3-52 (in Russian; a translation appears in *Am. Math. Soc. Transl.*, ser. 2, vol. 12, pp. 199-246).

Golomb, S. (1961a): A New Derivation of the Entropy Expressions, *IRE Trans. Inform. Theory*, vol. IT-7, no. 3, pp. 166-167, July.

——— (1961b): Efficient Coding for the Desoxyribonucleic Channel, *Proc. Symp. Appl. Math.*, vol. 14, *Mathematical Problems in the Biological Sciences*, American Mathematical Society, pp. 87-100.

——— (1962): Genetic Coding, *Eng. Sci. Mag.*, California Institute of Technology, April.

Grettenberg, T. L. (1962): The Ordering of Finite Experiments, *Trans. Third Prague Conf. Inform. Theory Statist. Decision Functions*, Publishing House of the Czechoslovak Academy of Sciences, Prague.

Hamming, R. W. (1950): Error Detecting and Error Correcting Codes, *Bell System Tech. J.*, vol. 29, pp. 147-150.

Harman, W. W. (1963): «Principles of the Statistical Theory of Communication», McGraw-Hill Book Company, Inc., New York.

Hartley, R. V. L. (1928): Transmission of Information, *Bell System Tech. J.*, vol. 7, pp. 535-563.

Huffman, D. A. (1952): A. Method for the Construction of Minimum Redundancy Codes, *Proc. IRE*, vol. 40, no. 10, pp. 1098-1101, September.

Jaynes, E. T. (1959): A Note on Unique Decipherability, *IRE Trans. Inform. Theory*, vol. 5, pp. 98-102, September.

Karp. R. M. (1961): Minimum-redundancy Coding for the Discrete Noiseless Channel, *IRE Trans. Inform. Theory*, vol. IT-7, pp. 27-38, January.

Karush, J. (1961): A Simple Proof of an Inequality of McMillan, *IRE Trans. Inform. Theory*, vol. IT-7, no. 2, p. 118, April.

Kelly, D. H. (1962): Information Capacity of a Single Retinal Channel, *IRE Trans. Inform. Theory*, vol. IT-8, no. 3, pp. 221-226, April.

Kelly, J. L., Jr. (1956): A New Interpretation of Information Rate, *Bell System Tech. J.*, vol. 35, pp. 917-927.

Kempthorne, O. (1952): «The Design and Analysis of Experiments», John Wiley & Sons, Inc., New York.

Khinchin. A. I. (1957): «Mathematical Foundations of Information Theory», Dover Publications, Inc., New York.

Kraft, L. G. (1949): «A Device for Quantizing, Grouping, and Coding Amplitude Modulated Pulses», M.S. thesis, Electrical Engineering Department, Massachusetts Institute of Technology, March.

Kullback, S. (1959): «Information Theory and Statistics», John Wiley & Sons, Inc., New York.

Lindley, D. (1956): On a Measure of the Information Provided by an Experiment, *Ann. Math. Statist.*, vol. 27, pp. 986-1005.

McGill, W. J. (1954): Multivariate Information Transmission, *IRE Trans. Inform. Theory*, vol. 4, pp. 93-111, September.

McMillan, B. (1953): The Basic Theorems of Information Theory, *Ann. Math. Statist.*, vol. 24, pp. 196-219.

McMillan, B. (1956): Two Inequalities Implied by Unique Decipherability, *IRE Trans. Inform. Theory*, vol. IT-2, pp. 115-116, December.

Miller, G. A., and W. G. Madow (1954): On the Maximum Likelihood Estimate of the Shannon-Wiener Measure of Information, *Air Force Cambridge Res. Center Rept.*, Cambridge, Mass.

Muroga, S. (1953): On the Capacity of a Discrete Channel I, *J. Phys. Soc. Japan*, vol. 8, pp. 484-494.

——— (1956): On the Capacity of a Discrete Channel, II, *J. Phys. Soc. Japan*. vol. 11, pp. 1109-1120.

Murphy, R. (1962): Adaptive Processes in Economic Systems, *Stanford Univ. Appl. Math. Statist. Lab. Tech. Rept.* 119, July.

Parzen, E. (1960): «Modern Probability Theory and Its Applications», John Wiley & Sons, Inc., New York.

——— (1961): «Stochastic Processes», Holden-Day, Inc., San Francisco.

Perez. A. (1959): Information Theory with an Abstract Alphabet, *Theory Probability Appl.*, vol. 4, no. 1, pp. 99-102.

Peterson, W. W. (1961): «Error-correcting Codes», John Wiley & Sons, Inc., New York.

Pierce, J. R. (1961): «Symbols, Signals and Noise», Harper & Row, Publishers, Incorporated, New York.

——— and J. E. Karlin (1957): Reading Rates and the Information Rate of a Human Channel, *Bell System Tech. J.*, vol. 36, pp. 497-516.

Pinkerton, R. C. (1956): Information Theory and Melody, *Sci. Am.*, PP. 77-87, February.

Pinsker, M. S. (1954): The Quantity of Information about a Gaussian Random Stationary Process, Contained in a Second Process Connected with It in a Stationary Manner, *Dokl. Akad. Nauk SSSR*, pp. 213-216 (in Russian).

Powers, K. H. (1956): A. Unified Theory of Informattion, *MIT Res. Lab. Electron. Tech. Rept.* 311, February.

Pratt, F. (1942): «Secret and Urgent», Doubleday & Company, Inc., Garden City, N. Y.

Quastler, H. (1956): «Information Theory in Psychology», The Free Press of Glencoe, New York.

Renyi, A. (1959): On the Dimension and Entropy of Probability Distributions, *Acta Math. Acad. Sci. Hung.*, vol. 10, pp. 193-215.

Reza, F. M. (1961): «An Introduction to Information Theory», McGraw-Hill Book Company, Inc., New York.

Sardinas, A. A., and G. W. Patterson (1953): A Necessary and Sufficient Condition for the Unique Decomposition of Coded Messages, 1953 *IRE Conv. Record*, pt. 8, pp. 104-108.

Shannon, C. E. (1951): Prediction and Entropy of Printed English, *Bell System Tech. J.*, vol. 30, no. 1, pp. 50-64, January.

——— (1956): The Zero Error Capaeity of a Noisy Channel, *IRE Trans. Inform. Theory*, vol. IT-2, no. 3, pp. 8-16, September.

——— (1957a): Certain Results in Coding Theory for Noisy Channels, *Inform. Control*, vol. 1, no. 1, pp. 6-25, September.

——— (1957b): Geometric Interpretation of Some Results of Channel Capacity Calculations, *Nachrichtentechnik*, vol. 10, pp. 1-4.

——— (1958): A note on a Partial Ordering for Communication Channels, *Inform. Control*, vol. 1, pp. 390-397, December.

——— and W. Weaver (1949): «The Mathematical Theory of Communication», The University of Illinois Press, Urbana, Ill. (The first part of this book is a reprint of Shannon's paper A Mathematical Theory of Communication, *Bell System Tech. J.*, vol. 27, pp. 379-423, 623-656, 1948.)

Silverman, R. A. (1955): On Binary Channels and Their Cascades, *IRE Trans. Inform. Theory*, vol. IT-1, pp. 19-27, December.

Stumpers, F. L. H. M. (1953): A Bibliography of Information Theory, *IRE Trans. Inform. Theory*, vol. PGIT-2, November.

——— (1955): A Bibliography of information Theory, First Supplement, *IRE Trans. Inform. Theory*, vol. IT-1, pp. 31-47, September.

——— (1957): A Bibliography of Information Theory, Second Supplement, *IRE Trans. Inform. Theory*, vol. IT-3, pp. 150-166, June.

——— (1960): A Bibliography of Information Theory, Third Supplement, *IRE Trans. Inform. Theory*, vol. IT-6, pp. 25-51, March.

Thomasian, A. J. (1960): An Elementary Proof of the AEP of Information Theory, *Ann. Math. Statist.*, vol. 31, pp. 452-456.

Wolfowitz, J. (1959): Strong Converse of the Coding Theorem for Semi-continuous Channels, *Illinois J. Math.*, vol 3, no. 4, pp. 477-489.

Woodward, P. M. (1955): «Probability and Information Theory, with Applications to Radar», Pergamon Press, New York.

Wocencraft, J. M., and B. Reiffen (1961): «Sequential Decoding», John Wiley & Sons, Inc., New York.

Yaglom, A. M., and I. M. Yaglom (1959): «Probabilité et Information», Dunod, Paris (in French, translated from the Russian).

TABLAS

TABLA A-1. LOGARITMO DE BASE 2

n	$\log n$	n	$\log n$
1	0.000000	26	4.700439
2	1.000000	27	4.754887
3	1.584962	28	4.807355
4	2.000000	29	4.857981
5	2.321928	30	4.906890
6	2.584962	31	4.954196
7	2.807355	32	5.000000
8	3.000000	33	5.044394
9	3.169925	34	5.087463
10	3.321928	35	5.129283
11	3.459431	36	5.169925
12	3.584962	37	5.209453
13	3.700440	38	5.247927
14	3.807355	39	5.285402
15	3.906890	40	5.321928
16	4.000000	41	5.357552
17	4.087463	42	5.392317
18	4.169925	43	5.426264
19	4.247927	44	5.459431
20	4.321928	45	5.491853
21	4.392317	46	5.523562
22	4.459431	47	5.554589
23	4.523562	48	5.584962
24	4.584962	49	5.614710
25	4.643856	50	5.643856

TABLA A-1. LOGARITMO DE BASE 2 (*Continuación*)

n	$\log n$	n	$\log n$
51	5.672425	76	6.247927
52	5.700439	77	6.266786
53	5.727920	78	6.285402
54	5.754887	79	6.303780
55	5.781359	80	6.321928
56	5.807355	81	6.339850
57	5.832890	82	6.357552
58	5.857981	83	6.375039
59	5.882643	84	6.392317
60	5.906890	85	6.409391
61	5.930737	86	6.426264
62	5.954196	87	6.442943
63	5.977280	88	6.459431
64	6.000000	89	6.475733
65	6.022367	90	6.491853
66	6.044394	91	6.507794
67	6.066089	92	6.523562
68	6.087462	93	6.539158
69	6.108524	94	6.554588
70	6.129283	95	6.569855
71	6.149747	96	6.584962
72	6.169925	97	6.599912
73	6.189824	98	6.614709
74	6.209453	99	6.629356
75	6.228818	100	6.643856

TABLA A-2. LA FUNCIÓN ENTROPÍA
$H(p) = -p \log p - \bar{p} \log \bar{p}$

p	$H(p)$	p	$H(p)$
0.005	0.045415	0.130	0.557438
0.010	0.080793	0.135	0.570993
0.015	0.112364	0.140	0.584239
0.020	0.141441	0.145	0.597185
0.025	0.168661	0.150	0.609840
0.030	0.194392	0.155	0.622213
0.035	0.218878	0.160	0.634310
0.040	0.242292	0.165	0.646138
0.045	0.264765	0.170	0.657705
0.050	0.286397	0.175	0.669016
0.055	0.307268	0.180	0.680077
0.060	0.327445	0.185	0.690894
0.065	0.346981	0.190	0.701471
0.070	0.365924	0.195	0.711815
0.075	0.384312	0.200	0.721928
0.080	0.402179	0.206	0.731816
0.085	0.419556	0.210	0.741483
0.090	0.436470	0.215	0.750932
0.095	0.452943	0.220	0.760167
0.100	0.468996	0.225	0.769193
0.105	0.484648	0.230	0.778011
0.110	0.499916	0.235	0.786626
0.115	0.514816	0.240	0.795040
0.120	0.529361	0.245	0.803257
0.125	0.543564	0.250	0.811278

TABLA A-2. LA FUNCIÓN ENTROPÍA (*Continuación*)
$H(p) = -p \log p - \bar{p} \log \bar{p}$

p	$H(p)$	p	$H(p)$
0.255	0.819107	0.380	0.958042
0.260	0.826746	0.385	0.961497
0.265	0.834198	0.390	0.964800
0.270	0.841465	0.395	0.967951
0.275	0.848548	0.400	0.970951
0.280	0.855451	0.405	0.973800
0.285	0.862175	0.410	0.976550
0.290	0.868721	0.415	0.979051
0.295	0.875093	0.420	0.981454
0.300	0.881291	0.425	0.983708
0.305	0.887317	0.430	0.985815
0.310	0.893173	0.435	0.987775
0.315	0.898861	0.440	0.989588
0.320	0.904381	0.445	0.991254
0.325	0.909736	0.450	0.992774
0.330	0.914925	0.455	0.994149
0.335	0.919953	0.460	0.995378
0.340	0.924819	0.465	0.996462
0.345	0.929523	0.470	0.997402
0.350	0.934068	0.475	0.998196
0.355	0.938454	0.480	0.998846
0.360	0.942683	0.485	0.999351
0.365	0.946755	0.490	0.999711
0.370	0.950672	0.495	0.999928
0.375	0.954434	0.500	0.100000

INDICE