

Una teoría matemática de la comunicación

Claude E. Shannon

1948

Traducción al español

Esta es una traducción al español del artículo publicado por Shannon en *The Bell System Technical Journal*, realizado a base del PDF disponible en <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf> como un esfuerzo colectivo de estudiantes de octavo semestre del ITS de la FIME de la UANL por puntos extra en la unidad de aprendizaje *Teoría de la información y métodos de codificación*, impartida por Dra. Elisa Schaeffer en la primavera del 2013.

Índice general

Introducción	5
I Sistemas discretos silenciosos	8
1 El canal discreto silencioso	9
2 La fuente discreta de la información	12
3 La serie de aproximaciones al inglés	14
4 Representacion grafica de procesos Markovianos	15
5 Fuentes ergódicas y mixtas	17
6 Selección, incertidumbre y entropia	20
7 La entropía de una fuente de información	24
7.1 El teorema fundamental para un canal sin ruido	27
7.2 Discusión y ejemplos	29
II El canal discreto con ruido	31
8 Representación de un canal discreto con ruido	32
9 Equivocación y capacidad de canal	33
10 El teorema fundamental para un canal discreto con ruido	34
11 Discussion	35
12 Ejemplo de un canal discreto y su capacidad	36
13 La capacidad del canal en ciertos casos especiales	38
14 Un ejemplo de codificación eficiente	40
Anexos	41
Anexo A El crecimiento del número de bloques de símbolos con una condición de estado finito	41
Anexo B Derivación de $H = -\sum p_i \log p_i$	42
Anexo C Teoremas sobre fuentes ergódicas	44
Anexo D Maximizar la tasa para un sistema de restricciones	46
III Preliminares matemáticos	48
15 Conjuntos y familias de funciones	50
16 Funciones de familias con banda limitada	54

17 Entropía de una distribución continua	55
18 Entropía en un conjunto de funciones	59
19 Pérdida de entropía en filtros lineales	61
20 Entropía de la suma de dos conjuntos	64
 IV El canal continuo	 66
21 La capacidad de un canal continuo	67
22 Capacidad de la señal con una limitación de potencia media	70
 V La tasa para una fuente continua	 76
23 Las funciones de evaluación de fidelidad	77
24 La tasa para una fuente relativa a una evaluación de fidelidad	80
25 El cálculo de las tasas	82
Agradecimientos	84
Anexos	85
Anexo E	85
Anexo F	86
Anexo G	88

Índice de figuras

1	Diagrama esquemático de un sistema de comunicaciones general.	6
1.1	Una representación gráfica de las restricciones en los símbolos telegráficos.	10
4.1	Un grafo correspondiente a la fuente del ejemplo 2.2.	15
4.2	Un grafo correspondiente a la fuente en ejemplo 2.3	16
5.1	Un grafo correspondiente a la fuente en ejemplo 2.4.	18
6.1	Descomposición de una decisión de tres posibilidades.	21
6.2	Entropía en el caso de dos posibilidades con probabilidades p y $1 - p$	21
9.1	Un diagrama esquemático de un sistema de corrección.	33
10.1	La equivocación posible para una entropía de entrada dada a un canal.	34
10.2	Una representación esquemática de las relaciones entre las entradas y salidas en un canal.	34
12.1	Un ejemplo de un canal discreto.	36
13.1	Ejemplos de canales discretos con las mismas probabilidades de transición para cada entrada y cada salida.	39

Índice de cuadros

19.1 Ganancia, factor de potencia de la entropía y ganancia en decibels, y la respuesta impulso.	63
---	----

Introducción

El reciente desarrollo de varios métodos de modulación como PCM y PPM el cual intercambio el ancho de banda por señal a ruido ha intensificado el interés general de la teoría de la comunicación. El fundamento de esta teoría se encuentra en los artículos importantes por Nyquist [1924] y Hartley [1928] sobre este tema.

En el presente artículo se extiende la teoría a incluir un número de nuevos factor, en particular el efecto del sonido en un canal, y los ahorros posibles debido a la estructura estadística del mensaje original y a la naturaleza del destino final de la información. El problema fundamental de la comunicación es el de reproducir en un momento exacto o aproximado un mensaje seleccionado en otro punto.

Frecuentemente los mensajes tienen un significado; es decir que refieren o están correlacionados de acuerdo con algún sistema con ciertas entidades físicas o conceptuales. Estos aspectos semanticos de la comunicación son irrelevantes a el problema de ingeniería.

El significativo aspecto es que el actual mensaje es uno de los seleccionados desde un conjunto de posibles mensajes. El sistema debe estar diseñado para operar por cada selección posible, no solamente por el único que ha sido escogido ya que es desconocido en el momento del diseño.

Si el número de mensajes en el conjunto es finito entonces el numero o cualquier función monótono de este número puede ser considerado como una medida de la información producida cuando un mensaje es escogido de un conjunto, todas las opciones son igualmente probables.

Como se ha señalado por Hartley la opción mas natural es la función logarítmica. Aunque esta definición debe ser generalizado consideradamente cuando se cuenta la influencia de las estadísticas del mensaje y cuando tenemos un rango continuo de mensajes, vamos a utilizar en todos los casos una medida esencialmente logarítmica.

La medida logarítmica es más conveniente por varias razones:

1. Es practicamente muy útil. Parámetros importantes utilizados en la ingeniería como el tiempo, ancho de banda, número de pasos, etc., tiende a variar linealmente con el logaritmo de número de posibilidades. Por ejemplo, agregando un paso a un grupo dobla el número de posible estado de pasos. Eso agrega 1 a la base 2 del logaritmo de este número. Duplicando el tiempo aproximadamente hace que el número de cuadrados posibles, o duplicando el logaritmo, etc.
2. Es más cerca a nuestro sentimiento intuitivo en cuanto a la medida adecuada. Esto es cerca-namente relacionado desde que medimos intuitivamente entidades por una comparación lineal con estándares comunes Nyquist [1924, 1928]. Un sentimiento, por ejemplo, cuando perforas dos cartas deberían tener el doble de la capacidad de uno para la transmisión de información Hartley [1928].

3. Es matemáticamente más adecuado. Muchas de las operaciones que limitan son simples en términos de el logaritmo pero requeriría reemplantamiento descuidado en términos del número de posibilidades.

La selección de la base del logaritmo corresponde a la selección de la unidad para medir la información. Si la base 2 es usado, las unidades resultantes podrían ser llamados dígitos binarios, o más brevemente bits, una palabra sugerida por J. W. Tukey. Un dispositivo con dos posiciones estables, como lo es un número de paso o un circuito flip-flop, pueden almacenar un bit de información. N dispositivos pueden almacenar N bits, desde que el número total de posibles estados es 2^N y $\log_2 2^N = N$. Si la base 10 es usada las unidades podrían ser llamados dígitos decimales. Desde

$$\log_2 M = \log_1 0M / \log_1 02 = 3,32 \log_1 0M, \quad (1)$$

un dígito decimal es alrededor de $3 \frac{1}{3}$ bits. Una rueda digital en una computadora de escritorio tiene 10 posiciones estables y por lo tanto, tiene una capacidad de almacenamiento de un dígito decimal. En trabajo analítico donde la integración y la diferenciación están envueltos en base e es algunas veces conveniente. Las unidades resultantes de información serían llamados unidades naturales. El cambio de base a a base b simplemente requiere la multiplicación por $\log_b a$.

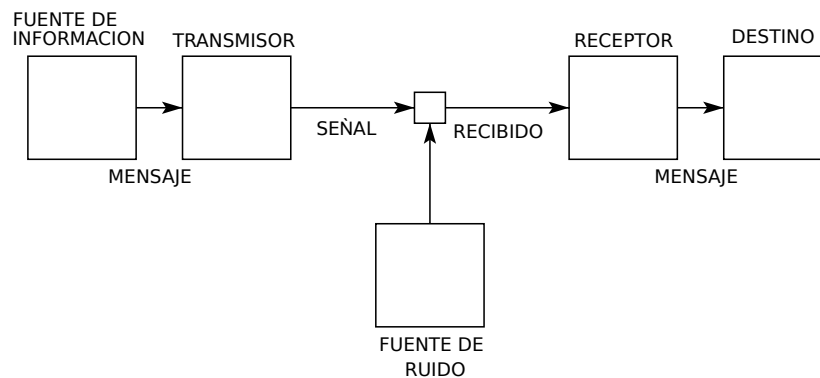


Figura 1: Diagrama esquemático de un sistema de comunicaciones general.

Por un sistema de comunicación se quiere decir que es un sistema de un tipo indicado esquemáticamente en la figura 1. Eso consiste de cinco partes esenciales:

1. La fuente de información el cual produce el mensaje o secuencia de mensajes a ser comunicado a la terminal que recibe. El mensaje puede ser de varios tipos:
 - (a) Una secuencia de letras como un sistema de telegrafo de teletipo;
 - (b) Una sola función de tiempo $f(t)$ como un radio o telefonía;
 - (c) Una función de tiempo y otras variables como televisión en blanco y negro — aquí el mensaje podría pasar a través de una función $f(x, y, t)$ de dos espacios de coordenadas y tiempo, la intensidad de la luz en el punto (x, y) y el tiempo t sobre un tubo en la placa;
 - (d) Dos o más funciones de tiempo, digamos $f(t)$, $g(t)$, $h(t)$ — este es el caso en “tres dimensional” transmisión de sonido o si el sistema está diseñado para dar servicio a varios canales individuales en multiplex;

- (e) Varias funciones de varias variables — en la televisión a color el mensaje consiste en tres funciones $f(x, y, t)$, $g(x, y, t)$, $h(x, y, t)$ definidas en tres dimensiones continuas — también se podría pensar que esas tres funciones como componentes de un campo vectorial definido en una región — de manera similar, distintas fuentes de televisiones en blanco y negro podrían producir “mensajes” que consisten en un número de funciones de tres variables;
 - (f) Diversas combinaciones también se producen, por ejemplo, en la televisión con un canal de audio asociado.
2. Un transmisor que opera en el mensaje de alguna manera para producir una señal adecuada para la transmisión sobre el canal. En la telefonía esta operación consiste simplemente en el cambio de presión de sonido en una corriente eléctrica proporcional. En la telegrafía tenemos una operación de codificación que produce una secuencia de puntos, guiones y espacios en el canal correspondiente al mensaje. En un sistema multiplex PCM las diferentes funciones de voz deben tomar muestras, compresiones, cuantificada y codificada, y finalmente intercalados adecuadamente para construir la señal. Sistemas de vocoder, la televisión y la modulación de frecuencia son ejemplos de operaciones complejas aplicadas a los mensajes para obtener la señal.
 3. El canal es simplemente el medio usado para transmitir la señal desde un transmisor a un receptor. Eso podría ser un par de cables, un cable coaxial, una radio frecuencia, un rayo de luz, etc.
 4. El receptor ordinalmente hace la operación inversa de lo que ya está hecho por el transmisor, reconstruye el mensaje desde la señal.
 5. La destinación es la persona (o cosa) hacia quien el mensaje es enviado.

Se desearía considerar ciertos problemas generales envueltos en el sistema de comunicación. Para hacer esto es necesario representar varios elementos envueltos como lo son las entidades matemáticas, adecuadamente idealizada desde sus contrapartes físicas.

A grandes rasgos se pueden clasificar los sistemas de comunicación en tres categorías principales: los discretos, continuos y mixtos. Por un sistema discreto se quiere decir que en tanto el mensaje como la señal son una secuencia de símbolos discretos. Un caso típico es la telegrafía donde el mensaje es una secuencia de letras y la señal de una secuencia de puntos, guiones y espacios. Un sistema continuo es aquel en el que están tanto el mensaje como señal tratadas como funciones continuas, por ejemplo, la radio o la televisión. Un sistema mixto es una en la que tanto las variables discretas y continuas aparecen, por ejemplo, la transmisión PCM de voz. Consideremos en primer lugar el caso discreto. Este caso tiene aplicaciones no sólo en teoría de la comunicación, sino también en la teoría de las máquinas de computación, el diseño de las centrales telefónicas y otros campos. Además el caso discreto crea una base para los casos continuos mixtos que serán tratadas en la segunda mitad del papel.

Parte I

Sistemas discretos silenciosos

Capítulo 1

El canal discreto silencioso

Teletipo y telegrafía son dos simples ejemplos de un canal discreto para la transmisión de información. Generalmente, con un canal discreto queremos decir que es un sistema, por lo cual una secuencia de selecciones desde un conjunto de símbolos elementarios S_1, \dots , puede ser transmitido desde un punto a otro. Cada uno de los símbolos S_i es asumido a tener una cierta duración en tiempo t_i segundos (no es necesariamente el mismo por cada S_i , por ejemplo los puntos y guiones. Es no requerido que todas las posibles secuencias de S_i sean capaces de hacer transmisión en el sistema; ciertas secuencias deben ser permitidas. Esos serían posibles señales por canal. Supongamos que en el telegrafo que los simbolos son:

1. Un punto, que consta de un cierre de línea para una unidad de tiempo y, a continuación de línea abierta para una unidad de tiempo;
2. Un guión, consta de tres unidades de tiempo de cierre y una unidad abierta;
3. Un espacio de palabra de seis unidades de línea abierta. Se podría colocar la restricción de secuencias permisibles que no hay espacios siguen uno a otro (por si dos espacios de letras son adyacentes, es idéntico con un espacio de palabra). La pregunta que ahora cuenta es cómo se puede medir la capacidad de un canal de transmitir información.

En el caso del teletipo donde todos los simbolos son de la misma duración, y cualquier secuencia de 32 simbolos se permite y la respuesta es fácil. Cada simbolo representa cinco bits de información. Si transmite el sistema de N simbolos por segundo, es natural decir que el canal tiene una capacidad de bits por segundo $5n$. Esto no significa que el canal del teletipo siempre será la transmisión de información a este ritmo - esta es la tasa máxima posible y si o no la tasa real alcanza este máximo depende de la fuente de información que alimenta el canal, como se verá más adelante. En el caso más general, con diferentes longitudes de los simbolos y las limitaciones en las secuencias permitidas, hacemos la siguiente definición.

Definición 1.1. La capacidad C de un canal discreto es dado por

$$C = \lim_{T \rightarrow \infty} \frac{\log N(T)}{T}, \quad (1.1)$$

donde $N(T)$ es el numero de señales permitidas de duración T .

Es fácilmente visto que en el caso de teletipo es reducido hacia el resultado previo. Puede ser demostrado que el limite en cuestion existe como un número finito en la mayoría de los casos de

interés. Supongamos que todas las secuencias de símbolos S_1, \dots, S_n son permitidas y esos símbolos tienen una duración t_1, \dots, t_n . ¿Cuál es la capacidad del canal? Si $N(t)$ representa el número de secuencias de duración t , nosotros tenemos:

$$N(t) = N(t - t_1) + N(t - t_2) + \dots + N(t - t_n). \quad (1.2)$$

El número total es equivalente a la suma de número de secuencias terminando en S_1, S_2, \dots, S_n y esos son $N(t - t_1), N(t - t_2), \dots, N(t - t_n)$, respectivamente. Acorde a un buen resultado en diferencias finitas, $N(t)$ es entonces asintótica para un largo t a X_{t_0} donde X_0 es la solución real de la ecuación característica:

$$X - t_1 + X - t_2 + \dots + X - t_n = 1. \quad (1.3)$$

En caso de que existan restricciones en secuencias permitidas todavía se puede obtener una ecuación diferencial de este tipo y encontrar C de la ecuación característica. En el caso mencionado anteriormente de la telegrafía

$$N(t) = N(t - 2) + N(t - 4) + N(t - 5) + N(t - 7) + N(t - 8) + N(t - 10), \quad (1.4)$$

como se ve por el cálculo de secuencia de los símbolos de acuerdo al último o al siguiente último símbolo ocurriendo. Por lo tanto $-\log \mu_0$ donde μ_0 es la raíz positiva de $1 = \mu^2 + \mu^4 + \mu^5 + \mu^7 + \mu^8 + \mu^{10}$. Resolviendo eso encontramos que $C = 0,539$.

Una restricción general se puede colocar en secuencias como la siguiente: Imaginemos un número de estados posibles a_1, a_2, \dots, m . Por cada estado solo ciertos símbolos del conjunto S_1, \dots, S_n pueden ser transmitidos (diferentes subconjuntos por cada estado). Cuando uno de esos han sido transmitidos el estado cambia a un nuevo estado dependiendo de el viejo estado y del símbolo en particular transmitido. Un ejemplo simple de esto es el caso del telégrafo. Hay dos estados en función de si o no un espacio era el último símbolo transmitido. Si es así, entonces sólo un punto o una raya puede ser enviado al lado y el estado cambia siempre. Si no, cualquier símbolo puede ser transmitido y los cambios de estado si un espacio es enviado, de lo contrario sigue siendo el mismo. Las condiciones pueden ser indicadas en una gráfica lineal como se ve en la figura 1.1.



Figura 1.1: Una representación gráfica de las restricciones en los símbolos telegráficos.

Los puntos de unión corresponden a los estados y las líneas indicando los símbolos posibles en un estado y el estado resultante. En el Anexo ?? se muestran que si las condiciones en las secuencias permitidas puede ser descrita en la forma C puede existir y se puede calcular de acuerdo con el siguiente resultado:

Teorema 1.1. Siendo $b_{ij}^{(s)}$ la duración del s -ésimo símbolo, el cual es permisible en el estado i y conduce al estado j . Después la capacidad del canal C es igual a $\log W$ donde W es la raíz real más

grande de la ecuación:

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0, \quad (1.5)$$

donde $\delta_{ij} = 1$ si $i = j$ y es cero en cualquier otro caso

Por ejemplo en el caso del telégrafo el determinante es:

$$\begin{vmatrix} -1 & (W^{-2} + W^{-2}) \\ (W^{-3} + W^{-6}) & (W^{-2} + W^{-2} - 1) \end{vmatrix} = 0 \quad (1.6)$$

Esta expansión conduce a la ecuación dada anteriormente para este caso.

Capítulo 2

La fuente discreta de la información

Hemos visto que bajo condiciones muy generales el logaritmo del número de señales posibles en un canal discreto aumenta linealmente con el tiempo. La capacidad de transmisión de información se puede especificar dando a esta tasa de aumento, el número de bits por segundo necesarios para especificar la señal particular utilizada. Consideremos ahora la fuente de información. ¿Cómo es una fuente de información a ser descrita matemáticamente, y la cantidad de información en bits por segundo se produce en una fuente dada? El principal punto en cuestión es la efecto de los datos estadísticos sobre la fuente en la reducción de la capacidad requerida de la canal, por el uso de una correcta codificación de la información. En la telegrafía, por ejemplo, los mensajes para ser transmitidos consisten en secuencias de letras. Estas secuencias, sin embargo, no son completamente aleatorias.

En general, las oraciones tienen una estructura estadística de, por ejemplo, inglés. La letra E se produce con más frecuencia que Q , la secuencia TH con más frecuencia que XP , etc. La existencia de esta estructura permite hacer un ahorro en el tiempo (o capacidad de canal) para codificar correctamente las secuencias de mensajes en una secuencia de señales. Esto ya se hace en una medida limitada en telegrafía utilizando el símbolo de canal más corto, un punto, por la más letra E que es la más común en inglés, mientras que las letras infrecuentes, Q , X , Z están representados por secuencias más largas de puntos y rayas.

Esta idea se lleva aún más lejos en ciertos códigos comerciales donde las palabras y frases comunes están representados por cuatro o cinco grupos de código de letra con un considerable ahorro de tiempo promedio. Los telegramas estándar de saludo y aniversario se han usado tanto hasta el punto de que se codifican una o dos frases en una secuencia relativamente corta de números.

Podemos pensar en una fuente discreta de generar el mensaje, símbolo por símbolo. Se elegirá una sucesión de símbolos de acuerdo con ciertas probabilidades dependiendo, en general, de las opciones anteriores, así como los símbolos en cuestión. Un sistema físico, o un modelo matemático de un sistema que produce una secuencia de símbolos gobernadas por un conjunto de probabilidades, se le conoce como un proceso estocástico. Podemos considerar por lo tanto que una fuente discreta puede ser representada por un proceso estocástico. A la inversa, cualquier proceso estocástico que produce una secuencia discreta de símbolos elegidos a partir de un conjunto finito puede ser considerado una fuente discreta. Esto incluye casos como:

1. Idiomas naturales escritos como el inglés, alemán y chino.
2. Fuentes de información continuos que se han vuelto discreta por algún proceso de cuantificación. Por ejemplo, la señal de voz cuantizada desde un transmisor PCM, o una señal de televisión cuantizada.
3. Casos matemáticos en los que simplemente se definen abstractamente procesos estocásticos que generan una secuencia de símbolos. Los siguientes son ejemplos de este último tipo de fuente.

Ejemplo 2.1. Supongamos que tenemos cinco letras A, B, C, D, E , que se eligen cada una con probabilidad 0,2, elecciones sucesivas siendo independientes. Esto daría lugar a una secuencia, la siguiente es un típico ejemplo:

BDCBCECCCADCBDDAAECCEEA,
ABBAEDECACEEEEBACBCEAD.

Este fue construido con el uso de una tabla de azar.

Ejemplo 2.2. Utilizando las mismas cinco letras y siendo las probabilidades 0,4, 0,1, 0,2, 0,2, 0,1, respectivamente, con decisiones independientes sucesivas. Un mensaje típico de esta fuente es entonces:

AAACDCBDCEAADADACEDA,
EADCABEDADEDEDCCAAAAAD.

Ejemplo 2.3. Una estructura más complicada se obtiene si los símbolos no son elegidos de manera independiente pero sus probabilidades dependen de las letras anteriores. En el caso más simple este tipo de elección depende solamente de la letra anterior y no de las que estan antes. La estructura estadística puede entonces ser descrita por un conjunto de probabilidades de transición $p_i(j)$, la probabilidad de que la letra i es seguida por la letra j . Los índices de i y j se extienden sobre todos los símbolos posibles. Una segunda manera equivalente de especificar la estructura es dar el “diagrama” de probabilidades $p(i, j)$, la frecuencia relativa de el diagrama ij . La frecuencia de letras $p(i)$, (la probabilidad de la letra i), la transición de probabilidades $p_i(j)$ y el diagrama de probabilidades $p(i, j)$ son descritos en las siguientes formulas:

$$p(i) = \sum_j p(i, j) = \sum_j p(j, i) = \sum_j p(j)p_j(i) \quad (2.1)$$

falta terminar ejemplo C.

Ejemplo 2.4. Falta ejemplo D

Falta lo que está al final de la página 6 y al inicio de la 7, incluyendo la nota a pie de página¹.

¹falta esto

Capítulo 3

La serie de aproximaciones al inglés

falta todo esto

Ejemplo 3.1. Aproximación de orden cero

Ejemplo 3.2. Aproximación de orden cero

Ejemplo 3.3. Aproximación de primer orden

Ejemplo 3.4. Aproximación de segundo orden

Ejemplo 3.5. Aproximación de tercer orden

Ejemplo 3.6. Aproximación de palabra de primer orden

Ejemplo 3.7. Aproximación de palabra de segundo orden

faltan unos párrafos de texto

Capítulo 4

Representacion grafica de procesos Markovianos

Los procesos estocásticos del tipo descrito arriba son matemáticamente conocidos como Procesos Markovianos Discretos y han sido estudiados extensivamente en la literatura ¹. El caso general puede ser descrito de la siguiente manera: Existe un número finito de posibles “estados” de un sistema; S_1, S_2, \dots, S_n . Además existen un conjunto de probabilidades de transición; $p_i(j)$ la probabilidad que si el sistema esta en estado S_i entonces enseguida vaya al estado S_j . Para realizar este proceso Markoviano en una fuente de información solo necesitamos asumir que una letra es producida para cada transición desde un estado a otro. Los estados corresponderán al “residuo de influencia” de letras precedentes.

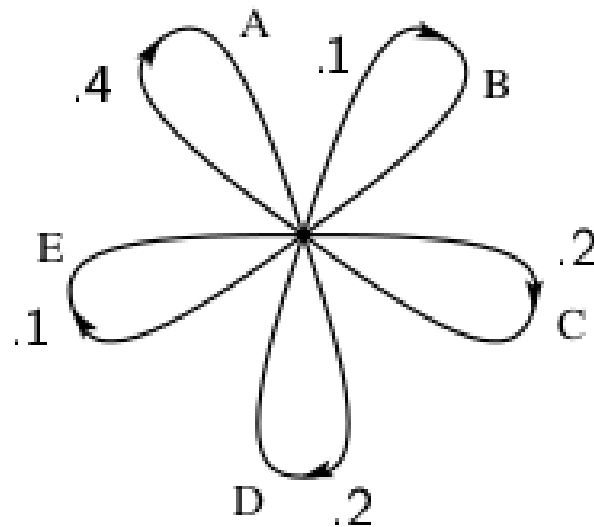


Figura 4.1: Un grafo correspondiente a la fuente del ejemplo 2.2.

La situación puede ser representada graficamente como se muestra en las figuras 4.1, 4.2 y 5.1. Los

¹Para una explicación detallada ver M. Fréchet, *Méthode des fonctions arbitraires. Théorie des événements en chaîne dans le cas d'un nombre fini d'états possibles*. Paris, Gauthier-Villars, 1938.

“estados” son los puntos de union en la grafica y las probabilidades y letras son producidas para una transicion son dadas a lado de la linea correspondiente. La figura 1.1 es para el ejemplo 2.2 en el capítulo 2, mientras que la figura 4.2 corresponde al ejemplo 2.3. En la figura 4.1 solamente hay un estado ya que letras sucesivas son independientes. En la figura 4.2 hay tantos estados como letras.

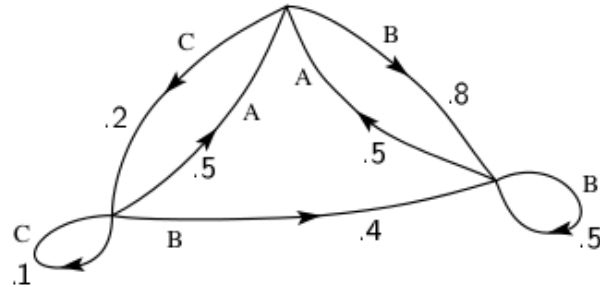


Figura 4.2: Un grafo correspondiente a la fuente en ejemplo 2.3

Si un ejemplo de un trigrama fuera construido, habría por maximo n^2 estados correspondiendo a los posibles pares de letras precediendo a uno que haya sido elegido. La figura 5.1 es un grafo para el caso de estructura de palabras en el ejemplo 2.4. Aqui S corresponde a el simbolo “espacio”.

Capítulo 5

Fuentes ergódicas y mixtas

Como se ha indicado anteriormente, una fuente discreta para nuestros propósitos puede ser considerada para ser representada por un proceso Markoviano. Entre los posibles procesos discretos Markovianos existe un grupo con propiedades especiales con importancia en la teoría de la comunicación. Esta clase especial consiste en los procesos “ergódicos” y deberíamos de llamar a las fuentes correspondientes, fuentes ergódicas. Aunque una definición rigurosa de los procesos ergódicos es algo complicada, la idea general es simple. En un proceso ergódico cada secuencia producida por el proceso permanece igual en sus propiedades estadísticas. Por lo tanto las frecuencias de letras, las frecuencias de bigramas, etc., obtenidos de una secuencia en partículas, se acercan a un límite definido conforme la longitud de la secuencia aumenta, independientemente de la secuencia en particular. En realidad esto no es meramente cierto para cada secuencia pero el grupo para el cual esto es falso tiene una probabilidad de cero. Prácticamente, la propiedad ergódica significa homogeneidad estadística.

Todos los ejemplos de lenguaje artificial dados anteriormente son ergódicos. Esta propiedad está relacionada a la estructura de los grafos correspondientes. Si el grafo tiene las siguientes dos propiedades

¹ el proceso correspondiente será ergódico:

1. El grafo no consiste de dos partes aisladas A y B dado que es imposible ir desde los puntos de unión en la parte A a los puntos de unión en la parte B a través de las líneas del grafo en la dirección de las flechas y también es imposible ir desde las uniones en la parte B a las uniones en la parte A .
2. Una serie de líneas cerradas en un grafo con todas sus flechas en las líneas apuntando en la misma dirección son llamados “circuitos”. La “longitud” de un circuito es el número de líneas en él. Por lo tanto en la figura 5.1, la serie $BEBES$ es un circuito de longitud cinco. La segunda propiedad requerida es que el máximo común divisor de la longitud de todos los circuitos en el grafo sea igual a uno.

Si la primera condición es satisfecha pero la segunda no por tener un máximo común divisor igual a $d > 1$, las secuencias tienen algún tipo de estructura periódica. Las diferentes secuencias caen dentro de d clases diferentes que son estadísticamente las mismas partiendo desde un cambio del origen (por ejemplo, que letra en la secuencia es llamada letra 1). Por un cambio de desde 0 hasta $d-1$ cualquier secuencia puede ser estadísticamente equivalente a cualquier otra. Un ejemplo simple con $d = 2$ es

¹Estas son reformulaciones en términos de la gráfica de las condiciones dadas en Fréchet.

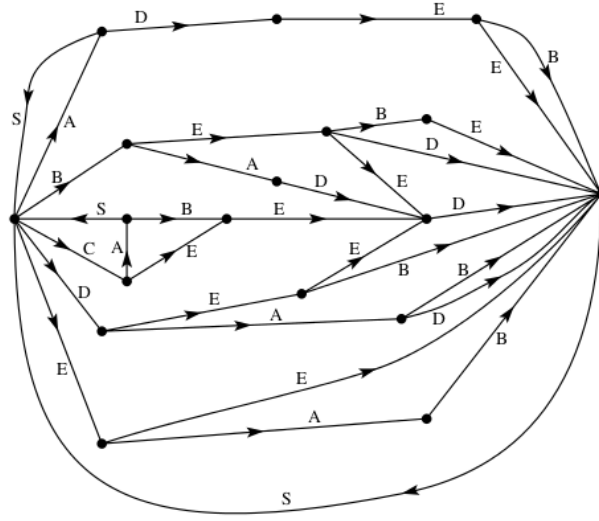


Figura 5.1: Un grafo correspondiente a la fuente en ejemplo 2.4.

el siguiente: Existen tres posibles letras a, b, c . La letra a es seguida con b ó c con probabilidades $\frac{1}{3}$ y $\frac{2}{3}$ respectivamente. Tanto b como c son siempre seguidas por una letra a . Por lo tanto una secuencia típica sería

abacacacabacababacac.

Este tipo de situación no es de mucha importancia para nuestro trabajo. Si la primera condición no se cumple, el grafo puede ser separado en un conjunto de subgrafos que satisfagan cada uno la primera condición. Nosotros asumiremos que la segunda condición es también satisfecha para cada subgrafo. Tenemos en este caso lo que se le llama una fuente “mezclada” hecha de un número de componentes puros. Estos componentes corresponden a los diversos subgrafos. Si L_1, L_2, L_3, \dots son los componentes fuente entonces podemos escribir

$$L = p_1 L_1 + p_2 L_2 + p_3 L_3 + \dots, \quad (5.1)$$

donde p_i es la probabilidad del componente fuente L_i . Físicamente la situación representada es esta: Hay diferentes fuentes L_1, L_2, L_3, \dots que son cada uno de una estructura estadística homogénea (por ejemplo, son ergódicos). No sabemos *a priori* cuál será utilizada, pero una vez que la secuencia empieza en un componente puro dado L_i , este continúa indefinidamente de acuerdo a la estructura estadística de ese componente. Como un ejemplo, uno puede tomar dos de estos procesos definidos arriba y asumir $p_1 = 0,2$ y $p_2 = 0,8$. Una secuencia de la fuente mezclada

$$L = 0,2L_1 + 0,8L_2 \quad (5.2)$$

sería obtenida escogiendo primero L_1 o L_2 con probabilidades 0,2 y 0,8 y después de esta elección generando una secuencia de lo que sea que haya sido elegido. Excepto cuando se exprese lo contrario nosotros debemos asumir que una fuente es ergódica. Esta suposición permite a uno identificar los promedios a lo largo de la secuencia con promedios encima del conjunto de posibles secuencias (la probabilidad de una discrepancia sea cero). Por ejemplo la frecuencia relativa de la letra A en una secuencia infinita particular sería, con la probabilidad uno, igual a su frecuencia relativa en el conjunto de secuencias. Si P_i es la probabilidad de un estado i y $p_i(j)$ la probabilidad de transición de un estado j , entonces para que el proceso sea estacionario es claro que P_i debe satisfacer condiciones de equilibrio:

$$P_j = \sum_i i P_i p_i(j). \quad (5.3)$$

En el caso ergódico este puede ser demostrado que con cualquier condición inicial las probabilidades $P_j(N)$ de estar en un estado j despues de N simbolos, se aproximan al equilibrio de valores conforme $N \rightarrow \infty$.

Capítulo 6

Selección, incertidumbre y entropía

Hemos presentado una fuente de información discreta como un proceso Markoviano. Podremos definir una cantidad que mida, en algún sentido, que tanta información es “producida” por estos procesos, o mejor aun, a que tasa la información es producida? Suponiendo que tenemos un conjunto de posibles eventos cuya probabilidad de ocurrir es p_1, p_2, \dots, p_n . Estas probabilidades son conocidas pero eso es todo lo que conocemos en cuanto a lo que concierne a que evento ocurrirá. Podremos encontrar una medida de cuanta “opcion” está involucrada en la seleccion de un evento o de que tan incierto pudiera ser la salida? En caso de que exista dicha medida, $H(p_1, p_2, \dots, p_n)$, es razonable el requerir de ésta las siguientes propiedades:

1. H debe ser continuo en p_i .
2. Si todas las p_i son iguales, $p_i = \frac{1}{n}$, entonces H debe ser una funcion de incremento monotona de n . Con eventos igualmente probables hay mas opciones, o incertidumbre, cuando hay mas eventos posibles.
3. Si una opcion se desglosara en dos opciones, la H original debería ser la suma ponderada de los valores individuales de H . El significado de esto es ilustrado en la figura 6.1. A la izquierda tenemos tres posibilidades $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$. A la derecha primero se escoge entre las dos posibilidades cada una con probabilidad $\frac{1}{2}$, y si la segunda ocurre hacer otra selección con probabilidades $\frac{2}{3}, \frac{1}{3}$. Los resultados finales tienen las mismas probabilidades que antes. Requerimos, en este caso en especial, que

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3}, \frac{1}{3}\right). \quad (6.1)$$

El coeficiente $\frac{1}{2}$ es porque esta segunda eleccion solo ocurre la mitad del tiempo.

En el Anexo ??, el siguiente resultado es establecido:

Teorema 6.1. *La unica H que satisface las tres suposiciones de arriba tiene la forma:*

$$H = -K \sum_i p_i \log p_i, \quad (6.2)$$

donde K es una constante positiva.

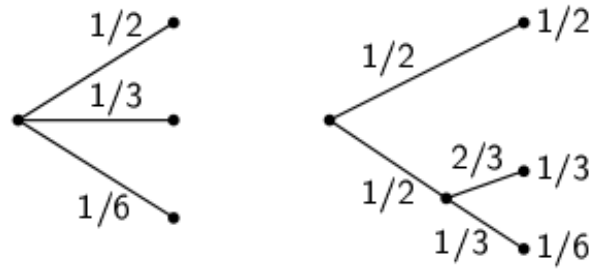


Figura 6.1: Descomposición de una decisión de tres posibilidades.

Este teorema, y las suposiciones requeridas para su demostración, no son en ninguna manera necesarias para la teoría de la que se está hablando. Se da principalmente para dar verosimilitud a algunas definiciones posteriores. La justificación real de estas definiciones, sin embargo, residirá en sus implicaciones. Cantidades de la forma $H = -\sum p_i \log p_i$ (la constante K viene a ser una opción de una unidad de medida) juegan un papel principal en la teoría de información como medidas de información, opciones e incertidumbre. La forma de H será reconocida como de entropía como se define en ciertas formulaciones de estadística mecánica¹ donde p_i es la probabilidad de un sistema de estar en una celda i de su espacio de fase. H es entonces, por ejemplo, la H en el famoso teorema H de Boltzmann. Debemos llamar $H = -\sum p_i \log p_i$ la entropía del conjunto de probabilidades p_1, \dots, p_n . Si x es una variable de probabilidad escribiríamos $H(x)$ para su entropía; por lo tanto x no es un argumento de una función sino una manera de representar un número, para diferenciar de $H(y)$ se diría la entropía de la variable de probabilidad y . La entropía en este caso de dos posibilidades con probabilidades p y $q = 1 - p$, es decir,

$$H = -(p \log p + q \log q), \quad (6.3)$$

es graficado en la figura 6.2 como una función de p .

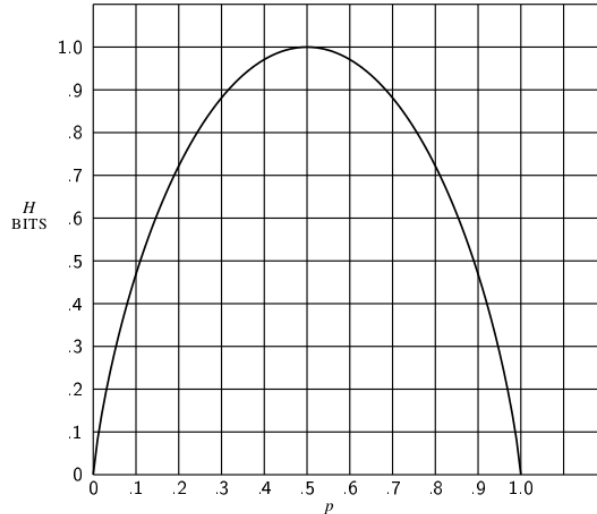


Figura 6.2: Entropía en el caso de dos posibilidades con probabilidades p y $1 - p$.

La cantidad H tiene un número interesante de propiedades que además se sustentan como una medida razonable de opción o información.

¹Ver, por ejemplo, R. C. Tolman, *Principles of Statistical Mechanics*, Oxford, Clarendon, 1938

1. $H = 0$ si y solo si todas las p_i excepto una son cero, esta teniendo el valor unitario. Así solamente cuando estemos seguros del resultado H desaparece. De lo contrario H es positivo.
2. Dado un n , H es un valor maximo igual a $\log n$ cuando todos los p_i son iguales (por ejemplo, $\frac{1}{n}$). Esto tambien es intuitivamente la situación más incierta.
3. Suponiendo que hay dos eventos, x y y , en cuestion con m posibilidades para el primero y n para el segundo. Sea $p(i, j)$ la probabilidad de la union de que ocurra i para el primero y j para el segundo. La entropía de la union de los eventos es

$$H(x, y) = - \sum i, j p(i, j) \log p(i, j), \quad (6.4)$$

mientras

$$\begin{aligned} H(x, y) &= - \sum i, j p(i, j) \log \sum j p(i, j), \\ H(x, y) &= - \sum i, j p(i, j) \log \sum i p(i, j). \end{aligned} \quad (6.5)$$

Es fácil demostrar que

$$H(x, y) \leq H(x) + H(y) \quad (6.6)$$

con igualdad solo si los eventos son independientes (por ejemplo, $p(i, j) = p(i)p(j)$). La incertidumbre de la union de eventos es menor que o igual a la suma de las incertidumbres individuales.

4. Cualquier cambio hacia la eculizacion de las probabilidades p_1, p_2, \dots, p_n incrementa H . Por lo tanto, si $p_1 < p_2$ y incrementamos p_1 , disminuyendo p_2 una cantidad igual de manera que p_1 y p_2 sean mas cercanos, entonces H incrementa. De una manera mas general, si ejecutamos de nuevo alguna operacion "promediante" en p_i de la forma

$$p'_i = \sum j a_{ij} p_j, \quad (6.7)$$

donde $\sum_i a_{ij} = \sum_j a_{ij} = 1$, y todo $a_{ij} \geq 0$, entonces H incrementa (excepto en el caso especial donde esta transformacion de cantidades a no mas de p_j permutaciones con H permaneciendo igual).

5. Suponiendo que hay dos oportunidades de eventos x y y como en 3, no necesariamente independiente. Para cada valor particular i que x puede tener hay una probabilidad condicional $p_i(j)$ que y tenga el valor j . Esto está dado por

$$p_i(j) = \frac{p(i, j)}{\sum_j p(i, j)}. \quad (6.8)$$

Definimos la *entropía condicional* de y , $H_x(y)$ como el promedio de entropia de y para cada valor de x , ponderado de acuerdo a la probabilidad de obtener ese x en particular. Esto es

$$H_x(y) = - \sum i, j p(i, j) \log p_i(j) \quad (6.9)$$

Esta cantidad mide que tan incierto se está de y en promedio cuando se conoce x . Substituyendo el valor de $p_i(j)$ obtenemos

$$H_x(y) = - \sum i, j p(i, j) \log p(i, j) + \sum i, j p(i, j) \log \sum j p(i, j) = H(x, y) - H(x), \quad (6.10)$$

o tambien

$$H(x, y) = H(x) + H_x(y). \quad (6.11)$$

La incertidumbre (o entropia) del evento conjunto x, y es la incertidumbre de x mas la incertidumbre de y cuando x es conocido.

6. Desde 3 y 5 tenemos

$$H(x) + H(y) \geq H(x, y) = H(x) + H_x(y) \quad (6.12)$$

Por lo tanto

$$H(y) \leq H_x(y) \quad (6.13)$$

La incertidumbre de y nunca es incrementada conociendo x . Esta será decrementada a menos que x y y sean eventos independientes, en cuyo caso no se cambia.

Capítulo 7

La entropía de una fuente de información

Considere una fuente discreta del tipo estado finito considerado arriba. Para cada posible estado i existirá un conjunto de probabilidades $p_i(j)$ resultado de producir los diferentes simbolos j . Por lo tanto existe una entropía H_i para cada estado. La entropía de la fuente se definirá como el promedio de estas H_i ponderadas de acuerdo con la probabilidad de ocurrencia de los estados en cuestión:

$$H = \sum_i P_i H_i = - \sum_{i,j} P_i p_i(j) \log p_i(j). \quad (7.1)$$

Esta es la entropía de la fuente por simbolo de texto. Si el proceso Markoviano está procediendo a una tasa de tiempo definido hay tambien una entropía por segundo,

$$H' = \sum_i f_i H_i, \quad (7.2)$$

donde f_i es la frecuencia promedio (sucesos por segundo) del estado i . Claramente

$$H' = mH, \quad (7.3)$$

donde m es el numero promedio de simbolos producidos por segundo. H o H' mide la cantidad de información generada por la fuente por simbolo o por segundo. Si la base logaritmica es 2, representarán bits por simbolo o por segundo. Si simbolos sucesivos son independientes entonces H es simplemente $-\sum p_i \log p_i$ donde p_i es la probabilidad de simbolo i . Suponiendo que en este caso se considere un mensaje largo de N simbolos. Con mucha probabilidad este contendría alrededor de $p_1 N$ ocurrencias del primer simbolo, $p_2 N$ ocurrencias del segundo, etc. Por lo tanto la probabilidad en particular de este mensaje sería aproximadamente

$$p = p_1^{p_1 N} p_2^{p_2 N} \dots p_n^{p_n N} \quad (7.4)$$

o

$$\begin{aligned} \log p &\doteq N \sum_i p_i \log p_i \\ \log p &\doteq -NH \\ H &\doteq \frac{\log 1/p}{N} \end{aligned} \quad (7.5)$$

Entonces H es aproximadamente el logaritmo del reciproco de la probabilidad de una longitud de secuencia tipica dividido por el numero de simbolos en la secuencia. El mismo resultado permanece para cualquier fuente. Dicho de otra forma mas precisa tenemos (ver apéndice ??):

Teorema 7.1. Dado cualquier $\epsilon > 0$ y $\delta > 0$, podemos encontrar un N_0 de tal manera que las secuencias de cualquier longitud $N \geq N_0$ caen dentro de dos clasificaciones:

1. Un conjunto cuya probabilidad total es menor que ϵ .
2. El residuo, todos aquellos miembros cuyas probabilidades satisfacen la desigualdad

$$\left| \frac{\log p^{-1}}{N} - H \right| < \delta \quad (7.6)$$

En otras palabras estamos casi seguros de tener $\frac{\log p^{-1}}{N}$ muy cerca a H cuando N es grande. Un resultado estrechamente relacionado trata con el numero de secuencias de varias probabilidades. Considerando de nuevo el numero de secuencias de longitud N y dejando que sean acomodados en orden decreciente de probabilidad. Definimos $n(q)$ para ser el numero que debemos de tomar de este conjunto empezando con el mas probable a fin de acumular una probabilidad total q para aquellos tomados.

Teorema 7.2.

$$\lim_{N \rightarrow \infty} \frac{\log n(q)}{N} = H \quad (7.7)$$

cuando q no es igual a 0 o 1.

Se puede interpretar $\log n(q)$ como el numero de bits requeridos para especificar la secuencia cuando consideramos solo la secuencia mas probable con una probabilidad q . Entonces $\frac{\log n(q)}{N}$ es el numero de bits por simbolo para la especificación. Este teorema dice que para un N grande este será independiente de q e igual a H . La tasa de crecimiento del logaritmo del numero de secuencias razonablemente probables está dado por H , independientemente de nuestra interpretacion de "razonablemente probable". Debido a estos resultados, que son probados en el apendice 3, es posible para la mayoría de nuestros propósitos el tratar a las secuencias largas como si solo fueran 2^{HN} de si mismas, cada una con una probabilidad 2^{-HN} . Los siguientes dos teoremas muestran que H y H' pueden ser determinados al limitar las operaciones directamente de las estadísticas de las secuencias de mensajes, sin referencia a los estados y probabilidades de transición entre los estados.

Teorema 7.3. Sea $p(B_i)$ la probabilidad de una secuencia B_i de símbolos de una fuente. Sea

$$G_N = -\frac{1}{N} \sum i p(B_i) \log p(B_i), \quad (7.8)$$

donde la suma está por encima de todas las secuencias B_i que contienen N simbolos. Entonces G_N es una funcion de decremento monotono de N y

$$\lim_{N \rightarrow \infty} G_N = H. \quad (7.9)$$

Teorema 7.4. Sea $p(B_i, S_j)$ la probabilidad de secuencia B_i seguida por el simbolo S_j y $p_{B_j}(S_j) = p(B_i, S_j)/p(B_i)$ sea la probabilidad condicional de S_j después de B_i . Sea

$$F_N = -\sum i, j p(B_i, S_j) \log p_{B_i}(S_j), \quad (7.10)$$

donde la suma está por encima de todos los bloques B_i de $N-1$ simbolos y sobre todos los simbolos S_j . Entonces F_N es una funcion monótona decreciente de N ,

$$\begin{aligned} F_N &= NG_N - (N-1)G_{N-1}, \\ G_N &= \frac{1}{N} \sum n = 1NF_N, \\ F_N &\leq G_N, \end{aligned} \quad (7.11)$$

y $\lim_{N \rightarrow \infty} F_N = H$.

Estos resultados se derivan del Anexo ???. Muestran que unas series de aproximaciones a H pueden ser obtenidas considerando solo la estructura estadística de las secuencias sobre $1, 2, \dots, N$ símbolos. F_N es la mejor aproximación. De hecho F_N es la entropía de la aproximación de N^{th} orden a la fuente del tipo discutido arriba. Si no existen influencias estadísticas se extienden sobre mas N símbolos, eso si la probabilidad condicional de el siguiente símbolo sabiendo el $(N - 1)$ anterior no es cambiado por conocimiento de cualquiera anterior a el, entonces $F_N = H$. F_N por supuesto es la entropía condicional del siguiente símbolo cuando los $(N - 1)$ anteriores son conocidos, mientras G_N es la entropía por símbolo de bloques de N símbolos. El radio de entropía de una fuente a el valor máximo valor que puede tener mientras aun está restringido a los mismos símbolos será llamado su *entropía* relativa. Este es la compresión máxima posible cuando codificamos dentro del mismo alfabeto. Uno menos la entropía relativa es la *redundancia*. La redundancia de el idioma inglés ordinario, no teniendo en cuenta a las estructuras estadísticas sobre distancias mayores a ocho letras, es aproximadamente 50 %. Esto significa que cuando se escribe en ingles la mitad de lo que se escribe es determinado por la estructura del lenguaje y la mitad se escoge libremente. La cifra de 50 % se encontró por diversos metodos independientes que todos arrojaron resultados similares. Uno es por calculo de la entropía de las aproximaciones al inglés. Un segundo metodo trata sobre borrar una cierta fraccion de letras de un ejemplo de texto en inglés y despues permitir a alguien el tratar de restaurarlo. Pueden ser restaurados cuando el 50 % son borrados y la redundancia es mayor a 50 %. Un tercer metodo depende de ciertos resultados conocidos en criptografia.

Dos extremos de la redundancia en prosa inglesa son representados por inglés básico y por el libro de James Joyce "Finnegans Wake". El vocabulario del inglés basico está limitado a 850 palabras y la redundancia es muy alta. Esto se refleja en la expansion que ocurre cuando un texto es traducido a inglés basico. Por otro lado Joyce agranda el vocabulario y alega lograr una compresión del contenido semántico. La redundancia de un lenguaje está relacionada a la existencia de crucigramas. Si la redundancia es cero, cualquier secuencia de letras es un texto razonable en el lenguaje y cualquier arreglo bidimensional de letras forma un crucigrama. Si la redundancia es muy alta, el lenguaje impone demasiadas restricciones para que los crucigramas grandes existan. Un análisis mas detallado muestra que si se asume que las restricciones impuestas por el lenguaje son de una naturaleza caótica y aleatoria, crucigramas grandes son posibles de realizar cuando la redundancia es 50 %. Si la redundancia es 33 %, crucigramas tridimensionales deberían ser realizables.

Supongamos que tenemos un sistema de restricciones sobre las posibles secuencias del tipo que puede ser representada por un gráfico lineal como la Figura ???. Si las probabilidades de $p_{ij}^{(s)}$ fueron asignados a las distintas líneas de conexión del estado i al estado j esto se convertiría en una fuente. Existe un trabajo en particular que maximiza la entropía resultante (véase Apéndice D).

Teorema 7.5. *El sistema de restricciones consideradas como un canal tiene una capacidad $C = \log W$. Si nosotros asignamos*

$$p_{ij}^{(s)} = \frac{B_j}{B_i} W^{-l_{ij}^{(s)}} \quad (7.12)$$

donde $l_{ij}^{(s)}$ es la duración de s -ésimo símbolo que va desde el estado i al estado j y satisface B_i

$$B_i = \sum_{s,j} B_j W^{-l_{ij}^{(s)}} \quad (7.13)$$

entonces H es maximizada e igual a C .

Por asignación adecuada de las probabilidades de transición de la entropía de los símbolos en un canal puede ser maximizada a la capacidad del canal.

7.1. El teorema fundamental para un canal sin ruido

Ahora vamos a justificar nuestra interpretación de H , tal como la tasa de generación de información por demostrar que H determina la capacidad de canal requerida con la codificación más eficiente.

Teorema 7.6. *Vamos a tener una fuente de entropía H (bits por símbolo) y un canal con una capacidad C (bits por segundo). Entonces es posible codificar la salida de la fuente de tal manera que se transmita a la media $\frac{C}{H} - \epsilon$ símbolos por segundo, sobre el canal donde es arbitrariamente pequeño. No es posible transmitir a una tasa promedio mayor que $\frac{C}{H}$.*

La parte inversa del teorema, que $\frac{C}{H}$ no puede ser excedido, puede probarse por senalar que la entropía de la entrada del canal por segundo es igual a la de la fuente, ya que el transmisor debe ser no singular, y también esta entropía no puede exceder la capacidad del canal. Por lo tanto $H' \leq C$ y el número de símbolos por segundo $= H'/H \leq C/H$.

La primera parte del teorema se demostró de dos maneras diferentes. El primer método es considerar el conjunto de todas las secuencias de N símbolos producidos por la fuente. Para N grandes podemos dividir en dos grupos, uno que contiene menos de $2^{(H+\eta)N}$ miembros N y la segunda contiene menos de 2^{RN} miembros (donde R es el logaritmo del número de símbolos diferentes) y que tiene una probabilidad total menos que μ . A medida que aumenta N , η y μ se aproximan a cero. El número de señales de duración T en el canal es mayor que $2^{(C-\theta)T}$ con θ pequeña cuando T es grande. Si nosotros elegimos

$$T = \left(\frac{H}{C} + \lambda\right)N \quad (7.14)$$

entonces no habra un numero suficiente de secuencias de simbolos de canal para el grupo de alta probabilidad cuando N y T son lo suficientemente grandes (por pequeño λ) y también algunos adicionales. El grupo de alta probabilidad se codifica en forma arbitraria de uno a una dirección dentro de este conjunto. Las restantes secuencias están representadas por secuencias no utilizadas para el grupo de alta probabilidad. Esta secuencia especial actúa como una señal de inicio y parada

para un código diferente. En el medio se permite un tiempo suficiente para dar suficientes secuencias diferentes a todos los mensajes de baja probabilidad. Esto requerirá

$$T_1 = \left(\frac{R}{C} + \varphi\right)N \quad (7.15)$$

cuando φ es pequeño. La tasa media de transmisión de símbolos de mensajes por segundo será entonces mayor que

$$\left[(1 - \delta)\frac{T}{N} + \delta\frac{T_1}{N}\right]^{-1} = [(1 - \delta)\left(\frac{H}{C} + \lambda\right) + \delta\left(\frac{R}{C} + \varphi\right)]^{-1}. \quad (7.16)$$

Como N incrementa δ , λ y φ se aproximan a cero y la tasa se aproxima a $\frac{C}{H}$. Otro método de realizar esta codificación y demostrando así el teorema se puede describir de la siguiente manera: Organizar los mnsajes de longitud N en orden decreciente de probabilidad y suponiendo que sus probabilidades son $p_1 \geq p_2 \geq p_3 \dots \geq p_n$.

Vamos a $P_s = \sum_{i=1}^{s-1} p_i$; que es P_s es la probabilidad acumulada hasta, pero no incluyendo, P_s . En primer lugar, codificar en un sistema binario. El código binario para el mensaje de s se obtiene mediante la expansión de P_s como un número binario. La expansión se lleva a cabo para m_s lugares, donde m_s es el número entero que satisface:

$$\log_2 \frac{1}{P_s} \leq m_s \times 1 + \log_2 \frac{1}{P_s}. \quad (7.17)$$

Por lo tanto los mensajes de alta probabilidad son representados por los códigos de acceso y los de baja probabilidad de códigos largos. A partir de estas desigualdades que tenemos:

$$\frac{1}{2^{m_s}} \leq P_s < \frac{1}{2^{m_s-1}}. \quad (7.18)$$

El código por P_s será diferente de todos los subsiguientes en uno o más de sus lugares m_s , ya que todos los restantes P_i es al menos $\frac{1}{2^{m_s}}$ grande y por lo tanto sus expansiones binarios difieren en los primeros lugares de m_s . En consecuencia, todos los códigos son diferentes, y es posible recuperar el mensaje de su código. Si las secuencias de los canales no son ya secuencias de dígitos binarios, que se pueden atribuir números binarios de manera arbitraria y por lo tanto el código binario traducido en señales adecuadas para el canal.

El número promedio de H' de dígitos binarios utilizados por ímbolo de mensaje original se calcula fácilmente. Nosotros tenemos

$$H' = \frac{1}{N} \sum m_s p_s, \quad (7.19)$$

pero

$$\frac{1}{N} \sum (\log_2 \frac{1}{p_s}) p_s \leq \frac{1}{N} \sum (1 + \log_2 \frac{1}{p_s}) p_s, \quad (7.20)$$

y por lo tanto,

$$G_N \leq H' < G_N + \frac{1}{N}. \quad (7.21)$$

Como N aumenta G_N se aproxima a H , la entropía de la fuente y H' se acerca H . Vemos de esto que la ineficiencia en la codificación, cuando se utiliza sólo un retardo finito de N símbolos, no tiene que ser mayor que N más la diferencia entre la verdadera entropía de H y la entropía G_N calculada para secuencias de longitud N . El exceso de tiempo por ciento necesario sobre el ideal es por lo tanto, menos

$$\frac{G_N}{H} + \frac{1}{HN} - 1. \quad (7.22)$$

Este método de codificación es sustancialmente el mismo que uno que encontraron de forma independiente por TODO [1949]. Su método es el de organizar los mensajes de longitud N en orden decreciente de probabilidad. Dividir esta serie en dos grupos de tan iguales probabilidades sea posible. Si el mensaje está en el primer grupo su primer dígito binario será 0, en caso contrario 1. Los grupos están divididos de manera similar en subconjuntos de casi la misma probabilidad y el subgrupo dígitos determina el segundo dígito binario. Este proceso continúa hasta que cada subconjunto contiene sólo un mensaje. Es fácil ver que, aparte de pequeñas diferencias (generalmente en el último dígito) esto equivale a lo mismo que el proceso de cálculo descrito anteriormente.

7.2. Discusión y ejemplos

Con el fin de obtener la máxima transferencia de energía de un generador a una carga, un transformador debe en general ser introducido de modo que el generador como se ve desde la carga tiene la resistencia de carga. La situación aquí es más o menos similar. El transductor hace que la codificación debe coincidir con la fuente para el canal en un sentido estadístico. La fuente como se ve desde el canal a través del transductor debe tener la misma estructura estadística como la fuente que maximiza la entropía en el canal. El contenido del teorema 9 es que, a pesar de una coincidencia exacta no es posible en general, podemos aproximar la mayor fidelidad si lo deseas. La relación de la velocidad real de transmisión de la capacidad C se puede llamar la eficiencia del sistema de codificación. Esto es, por supuesto, igual a la relación de la entropía real de los símbolos de canal a la entropía máxima posible. En general, la codificación ideal o casi ideal requiere un largo retraso en el transmisor y el receptor. En el caso silenciosos que hemos estado considerando, la función principal de este retraso es permitir razonablemente buena adaptación de las probabilidades correspondientes a longitudes de secuencias. Con un buen código el logaritmo de la probabilidad recíproco de un mensaje largo debe ser proporcional a la duración de la señal correspondiente, de hecho

$$\left| \frac{\log p^{-1}}{T} - C \right| \quad (7.23)$$

debe ser pequeña para todos, pero una pequeña fracción de los mensajes largos. Si una fuente puede producir sólo un mensaje particular, su entropía es cero, y no se requiere ningún canal. Por ejemplo, una máquina de computación configurado para calcular los dígitos sucesivos de produce una secuencia definida con ningún elemento de azar. No se requiere ningún canal de "transmitir" este a otro punto. Se podría construir una segunda máquina para calcular la misma secuencia en el punto. Sin embargo, esto puede ser poco práctico. En tal caso se puede optar por ignorar algunos o todos de los conocimientos estadísticos que tenemos de la fuente. Podríamos considerar los dígitos de π son una secuencia aleatoria en que se construye un sistema capaz de enviar cualquier secuencia de dígitos.

De manera similar, podemos optar por utilizar algunos de nuestro conocimiento estadístico de Inglés en la construcción de un código, pero no todos de la misma. En este caso se considera la fuente de la máxima entropía sujeta a las condiciones estadísticas que deseamos conservar. La entropía de esta fuente determina la capacidad del canal que es necesaria y suficiente. En π el ejemplo de la única información retenida es que todos los dígitos se eligen desde el conjunto $0, 1, \dots, 9$. En el caso de Inglés que uno podría desear utilizar el ahorro estadística posible debido a las frecuencias de la letra, pero nada más. La fuente de entropía máxima es entonces la primera aproximación al Inglés y su entropía determina la capacidad del canal deseado.

Como un simple ejemplo de algunos de estos resultados consideran una fuente que produce una secuencia de letras elegidas de entre A, B, C, D con probabilidades $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$, símbolos sucesivos

se eligen independientemente. Nosotros tenemos

$$H = -\left(\frac{1}{2} \log \frac{1}{2} + \frac{1}{4} \log \frac{1}{4} + \frac{2}{8} \log \frac{1}{8}\right) = \frac{7}{4} \quad (7.24)$$

bits por símbolo.

Por lo tanto se puede establecer un sistema de codificación para codificar los mensajes aproximada de esta fuente en dígitos binarios con un promedio de $\frac{7}{4}$ dígito binario por símbolo. En este caso podemos realmente alcanzar el valor límite por el siguiente código (obtenida por el método de la segunda prueba del teorema 7.6):

A	0
B	10
C	110
D	111

El número medio de dígitos binarios utilizados en la codificación de una secuencia de N símbolos es:

$$N\left(\frac{1}{2}x_1 + \frac{1}{4}x_2 + \frac{2}{8}x_3\right) = \frac{7}{4}N. \quad (7.25)$$

Es fácil ver que los dígitos binarios 0, 1 tienen probabilidades de $\frac{1}{2}$, $\frac{1}{2}$ por lo que la H para las secuencias codificadas es un bit por símbolo. Puesto que, en promedio, tenemos $\frac{7}{4}$ símbolos binarios por carta original, las entropías sobre una base de tiempo son los mismos. La entropía máxima posible para el conjunto original es $\log_2 4 = 2$, que se producen cuando A , B , C , D tienen probabilidades de $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$. Por lo tanto, la entropía relativa es $\frac{7}{8}$. Podemos traducir las secuencias binarias en el conjunto original de los símbolos en un dos-a-uno en la siguiente tabla:

00	A'
01	B'
10	C'
11	D'

Parte II

El canal discreto con ruido

Capítulo 8

Representación de un canal discreto con ruido

FALTA

Capítulo 9

Equivocación y capacidad de canal

FALTA

Teorema 9.1. *Si el canal de corrección...*

FALTA



Figura 9.1: Un diagrama esquemático de un sistema de corrección.

FALTA

Ejemplo 9.1. Suponga que los errores suceden al azar...

FALTA

Teorema 9.2. *Que un canal discreto tenga*

FALTA

Capítulo 10

El teorema fundamental para un canal discreto con ruido



Figura 10.1: La equivocación posible para una entropía de entrada dada a un canal.

FALTA



Figura 10.2: Una representación esquemática de las relaciones entre las entradas y salidas en un canal.

FALTA

Capítulo 11

Discussion

FALTA

En el caso sin ruido un retraso general requiere aproximarse a una codificación ideal. Ahora tiene una función adicional que permite una muestra amplia de ruido para afectar a la señal antes de cualquier juicio, se hace en el punto de recepción para el mensaje original. Incrementando el tamaño del ejemplo siempre agudiza las posibles afirmaciones estadísticas.

El contenido del teorema 9.2 y su prueba se pueden formular de una manera diferente que muestra una conexión sin ruido de una manera mas clara. Tenga en cuenta las posibles señales de duración T y supongamos que un subconjunto de ellos es seleccionado para ser usado. Los que estén en el subconjunto se utilizan todos con igual probabilidad, y suponiendo que el receptor está construido para seleccionar, como la señal original, la causa más probable del subconjunto, cuando una señal perturbada es recibida. Nosotros definimos $N(T, q)$ siendo el numero máximo de señales que podemos elegir para el subconjunto tal que la probabilidad de una interpretación incorrecta sea menor o igual a q .

Teorema 11.1.

$$\lim_{T \rightarrow \infty} \frac{\log N(T, q)}{T} = C, \quad (11.1)$$

donde C es la capacidad del canal, siempre que q no sea igual a 0 o 1

En otras palabras, no importa la forma en que se establece los límites de fiabilidad, podemos distinguir de forma fiable mensajes en tiempo T para corresponder a CT bits, cuando T es suficientemente grande. En el teorema ?? podemos comparar la capacidad de un canal sin ruido dado en la sección uno ??.

Capítulo 12

Ejemplo de un canal discreto y su capacidad

Un ejemplo simple de un canal discreto se indica en la figura 12.1. Hay tres posibles símbolos. El primero nunca se vera afectado por el ruido. El segundo y el tercero tienen cada uno una probabilidad p de llegar inalterado y q de ser cambiado en otro elemento par. Tenemos (permitiendo $\alpha = -[p \log p + q \log q]$ y P y Q son probabilidades de estar usando los símbolos primero y segundo).

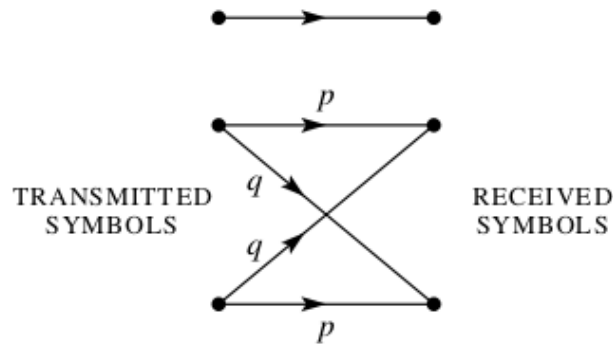


Figura 12.1: Un ejemplo de un canal discreto.

$$\begin{aligned} H(x) &= -P \log P - 2Q \log Q \\ H_y(x) &= 2Q\alpha \end{aligned} \tag{12.1}$$

Nosotros deseamos elegir P y Q , de tal manera que se maximice $H(x) - H_y(x)$ sujeto a la restricción $P + 2Q = 1$.

Por lo tanto consideremos:

$$\begin{aligned}
 U &= -P \log P - 2Q \log Q - 2Q\alpha + \lambda(P + 2Q) \\
 \frac{\partial U}{\partial P} &= -1 - \log P + \lambda = 0 \\
 \frac{\partial U}{\partial Q} &= -2 - 2 \log Q - 2\alpha + 2\lambda = 0
 \end{aligned} \tag{12.2}$$

Eliminando λ

$$\begin{aligned}
 \log P &= \log Q + \alpha \\
 P &= Qe^\alpha = Q\beta
 \end{aligned} \tag{12.3}$$

$$P = \frac{\beta}{\beta + 2} Q = \frac{1}{\beta + 2} \tag{12.4}$$

La capacidad del canal es de:

$$C = \log \frac{\beta + 2}{\beta} \tag{12.5}$$

Notese como esto comprueba los valores obvios en los casos: $p = 1$ y $p = \frac{1}{2}$. En primero, $\beta = 1$ y $C = \log 3$, el cual es correcto debido a que el canal es entonces sin ruido con tres posibles símbolos. Si $p = \frac{1}{2}$, $\beta = 2$ y $C = \log 2$. Aquí el segundo y el tercer símbolo, no se puede distinguir en absoluto y actúan conjuntamente como un solo símbolo. El primer símbolo se utiliza con una probabilidad $P = \frac{1}{2}$ y el segundo junto al tercero con probabilidad $\frac{1}{2}$. Esto puede ser distribuido entre ellos de cualquier modo deseado y todavía alcanzar la máxima capacidad. Para los valores intermedios de la capacidad del canal p estara entre $\log 2$ y $\log 3$. Esta distinción entre el segundo y tercer símbolo transmite alguna información, pero no tanto como en el caso sin ruido. El primer símbolo se utiliza tanto mas frecuentemente que los otros dos, debido a su ausencia de ruido.

Capítulo 13

La capacidad del canal en ciertos casos especiales

Si el ruido afecta símbolos sucesivos del canal de forma independiente pueden ser descritos por un conjunto de transición de probabilidades $p_{i,j}$. Esta es la probabilidad, si el símbolo i es enviado, que j será recibido. La tasa de canal máximo viene dado por el máximo de:

$$\sum_{i,j} P_i p_{i,j} \log \sum_{i,j} P_i p_{i,j} + \sum_{i,j} P_i p_{i,j} \log p_{i,j} \quad (13.1)$$

Donde variamos P_i sujeto a $\sum P_i = 1$. Esto se conduce por el método de Lagrange para las ecuaciones,

$$\sum_j P_{sj} \log \frac{P_{sj}}{\sum_i P_i p_{ij}} = us = 1, 2, \dots \quad (13.2)$$

Multiplicando por P_s y sumando en s muestra que $\mu = C$. Se hace la inversa de p_{sj} (si existe) en h_{st} de modo que $\sum_s h_{st} p_{sj} = \delta_{tj}$. Entonces:

$$\sum_{s,j} h_{st} p_{s,j} \log p_{s,j} - \log \sum_i P_i p_{i,t} = C \sum_s h_{s,t}. \quad (13.3)$$

Por lo tanto:

$$\sum_i P_i p_{i,t} = \exp[-C \sum_s h_{s,t} + \sum_{s,j} h_{s,t} p_{s,j} \log p_{s,j}] \quad (13.4)$$

o

$$P_i = \sum_t h_{i,t} \exp[-C \sum_s h_{s,t} + \sum_{s,j} h_{s,t} p_{s,j} \log p_{s,j}]. \quad (13.5)$$

Este es el sistema de ecuaciones para determinar el valor máximo de P_i , con C se determina de manera que $\sum P_i = 1$. Cuando esto está hecho, C sera la capacidad del canal y P_i las probabilidades para los símbolos del canal para lograr esta capacidad. Si cada símbolo de entrada tiene el mismo conjunto

de probabilidades en las líneas que emergen de esto y lo mismo sucede a cada símbolo de salida, la capacidad puede ser calculada fácilmente. Los ejemplos se muestran en la figura 12 13.1. En tal caso $H_x(y)$ es independiente de la distribución de probabilidades de los símbolos de entrada, y esta dada por $-\sum p_i \log p_i$. Cuando p_i son los valores de probabilidad de transición de cualquier símbolo de entrada. La capacidad del canal es:

$$\text{Max}[H(y) - H_x(y)] = \text{Max} H(y) + \sum p_i \log p_i. \quad (13.6)$$

El valor máximo de $H(y)$ es claramente $\log m$ donde m es el numero de símbolos de salida, ya que es posible que se den igualmente probables haciendo los símbolos de entradas igualmente probables. La capacidad del canal es por lo tanto

$$C = \log m + \sum p_i \log p_i. \quad (13.7)$$

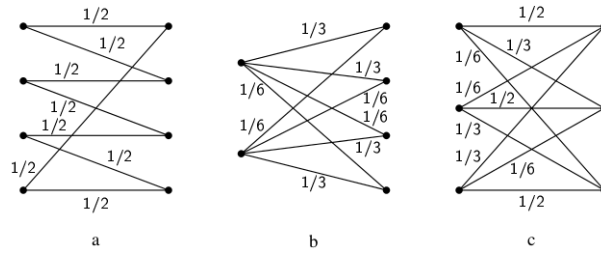


Figura 13.1: Ejemplos de canales discretos con las mismas probabilidades de transición para cada entrada y cada salida.

En la figura 12a 13.1 sería

$$C = \log 4 - \log 20 \log 2. \quad (13.8)$$

Esto se podría lograr mediante el uso solo el primer y el tercer símbolo. En la Figura 12b13.1:

$$C = \log 4 - \frac{2}{3} \log 3 - \frac{1}{3} \log 6 = \log 4 - \log 3 - \frac{1}{3} \log 2 = \log \frac{1}{3} 2^{\frac{5}{3}} \quad (13.9)$$

En la figura 12c 13.1 tenemos:

$$C = \log 3 - \frac{1}{2} \log 2 - \frac{1}{3} \log 3 - \frac{1}{6} \log 6 = \log \frac{3}{2^{\frac{1}{2}} 3^{\frac{1}{3}} 6^{\frac{1}{6}}}. \quad (13.10)$$

Supongamos que los símbolos se dividen en varios grupos tal que el ruido causa a un símbolo en un grupo a ser confundido con un símbolo de otro grupo. Deja la capacidad de un grupo n -ésimo ser C_n (en bits por segundo) donde solo utilizamos los símbolos de este grupo. Entonces es fácil desmotrar que para un mejor uso de todo el conjunto, la probabilidad P_n total de todos los símbolos del grupo n -ésimo debería ser:

$$P_n = \frac{2^{C_n}}{\sum 2^{C_n}} \quad (13.11)$$

En un grupo la probabilidad se distribuye tal como sería si estos eran los únicos símbolos que se utilizan. La capacidad del canal es:

$$C = \log \sum 2^{C_n}. \quad (13.12)$$

Capítulo 14

Un ejemplo de codificación eficiente

El siguiente ejemplo, aunque es un poco irrealista, es un caso en que la coincidencia exacta para un canal con ruido, es posible. Hay dos símbolos de canal 0 y 1, y el ruido les afecta en bloques de siete símbolos. Un bloque de siete o se transmite sin error, o exactamente un símbolo de los siete es incorrecta. Estas ocho posibilidades son igualmente probables. Tenemos:

$$C = \text{Max}[H(y) - H_x(y)] = \frac{1}{7} \left[7 + \frac{8}{8} \log \frac{1}{8} \right] = \frac{4}{7} \text{ bits/símbolos} \quad (14.1)$$

Un código eficiente, permite la corrección de todos los errores y transmitir a la tasa C , es el siguiente (encontrado por un método de R. Hamming):

Deje un bloque de siete símbolos que X_1, X_2, \dots, X_7 . De estos X_3, X_5, X_6 y X_7 son los mensajes de símbolos y elegidos arbitrariamente por la fuente. Los otros tres son redundantes y se calculan como sigue:

$$\begin{array}{llll} X_4 & \text{es elegido para hacer} & \alpha = X_4 + X_5 + X_6 + X_7 & \text{par,} \\ X_2 & \text{" " "} & \beta = X_2 + X_3 + X_6 + X_7 & \text{" " "}, \\ X_1 & \text{" " "} & \gamma = X_1 + X_3 + X_5 + X_7 & \text{" " "}. \end{array}$$

Cuando un bloque de siete es recibido α, β y γ son calculados y si incluso llamó a cero, si un extraño llamado. El número binario $\alpha\beta\gamma$ entonces da el subíndice de la X_i que no es correcto (si es 0 no hay error).

Anexo A

El crecimiento del número de bloques de símbolos con una condición de estado finito

Siendo $N_i(L)$ el número de bloques de símbolos de largo L terminando en estado i . Entonces tenemos

$$N_j(L) = \sum_{i,s} N_i(L - b_{ij}^{(s)}) \quad (\text{A.1})$$

donde $b_{ij}^1, b_{ij}^2, \dots, b_{ij}^m$ el largo de los símbolos los cuales pueden ser elegidos en estado i y pasar a estado j . Esas son ecuaciones diferenciales lineales y el comportamiento como $L \rightarrow \infty$ debe ser de tipo

$$N_j = A_j W^L. \quad (\text{A.2})$$

Sustituyendo en la ecuación diferencial

$$A_j W^L = \sum_{i,s} A_i W^{L-b_{ij}^{(s)}} \quad (\text{A.3})$$

o

$$\sum_i \left(\sum_s W^{b_{ij}^{(s)}} - \delta_{ij} \right) A_i = 0. \quad (\text{A.4})$$

Para hacer posible esto el determinante

$$D(W) = |a_{ij}| = \left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| \quad (\text{A.5})$$

debe desaparecer y esto determina W , que es, por supuesto, la mayor raíz real de $D = 0$. La cantidad C está dada entonces por

$$C = \lim_{L \rightarrow \infty} \frac{\log \sum A_j W^L}{L} = \log W \quad (\text{A.6})$$

y observamos también que las mismas propiedades de crecimiento resultan si nosotros requerimos que todos los bloques comiencen en el mismo (elegido arbitrariamente) estado.

Anexo B

Derivación de $H = - \sum p_i \log p_i$

Siendo $H(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}) = A(n)$. Desde la condición (3) se puede descomponer una elección de s^m con las mismas probabilidades en una serie de m elecciones de s de posiblemente igualmente probabilidades y obtener

$$A(s_m) = mA(s). \quad (\text{B.1})$$

Igualmente

$$A(t^n) = nA(t). \quad (\text{B.2})$$

Podemos elegir n arbitrariamente grande y encontrar una m para satisfacer

$$s^m \leq t^n < s^{(m+1)}. \quad (\text{B.3})$$

Así, tomando logaritmos y dividiendo en $n \log s$,

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n} \text{ o } \left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \epsilon \quad (\text{B.4})$$

donde ϵ es arbitrariamente pequeña. Ahora de la propiedad monotonía de $A(n)$,

$$A(s^m) \leq A(t^n) \leq A(s^{m+1}) \quad (\text{B.5})$$

$$mA(s) \leq nA(t) \leq (m+1)A(s) \quad (\text{B.6})$$

Por lo tanto

$$H = K \left[\sum p_i \log \sum n_i - \sum p_i \log n_i \right] \quad (\text{B.7})$$

$$= -K \sum p_i \log \frac{n_i}{\sum n_i} = -K \sum p_i \log p_i. \quad (\text{B.8})$$

Si el p_i es inconmensurables, que se puede aproximar por racionales y la misma expresión debe contener por nuestra suposición de continuidad. Así, la expresión se mantiene en general. La elección del coeficiente K es un asunto de conveniencia y asciende a la elección de una unidad de medida.

Anexo C

Teoremas sobre fuentes ergódicas

Si es posible pasar de un estado con $P > 0$ a cualquier otro a lo largo de un camino de probabilidad $p > 0$, el sistema es ergódico y una fuerte ley de numeros largos es aplicada. Así, el número de veces que un camino dado p_{ij} en una red se desplaza en una larga secuencia de longitud N es aproximadamente proporcional a la probabilidad de estar en i , dice P_i , y escoge la ruta, $P_i p_{ij} N$. Si N es lo suficientemente larga la probabilidad de porcentaje de error $\pm \delta$ es esto es menor que ϵ así que para todos, pero un conjunto de baja probabilidad de los números reales se encuentran dentro de los límites

$$(P_i p_{ij} \pm \delta) N. \quad (C.1)$$

Por lo tanto casi todas las secuencias tienen una probabilidad p dada por

$$p = \prod p_{ij}^{(P_i p_{ij} \pm \delta) N} \quad (C.2)$$

y $\frac{\log p}{N}$ es limitada por

$$\frac{\log p}{N} = \sum (P_i p_{ij} \pm \delta) \log p_{ij} \quad (C.3)$$

o

$$\left| \frac{\log p}{N} - \sum (P_i p_{ij} \pm \delta) \log p_{ij} \right| < \eta \quad (C.4)$$

Esto demuestra el teorema 3 PONER REF. Teorema PONER REF sigue inmediatamente de esto en el cálculo de los límites superior e inferior para $n(q)$ basado en el rango posible de valores de p en el Teorema PONER REF. En el mixto (no ergódico) caso si

$$L = \sum p_i L_i \quad (C.5)$$

y las entropías de los componentes son $H_1 \leq H_2 \leq \dots \leq H_n$ tenemos :

Teorema C.1. $\lim_{N \rightarrow \infty} \frac{\log n(q)}{N} = \varphi(q)$ es una función de paso decreciente $\varphi(q) = H_s$ en el intervalo $\sum_1^{s-1} \alpha_i < q < \sum_1^s \alpha_i$.

Para demostrar teoremas 7.3 y 7.4, primero note que F_N es monótona decreciente porque el aumento de N agrega un subíndice a la entropía condicional. Una sencilla sustitución de $p_{B_i}(S_j)$ en la definición de F_N muestra que

$$F_N = N G_N - (N - 1) G_N - 1, \quad (C.6)$$

y sumando esto por todas las N dadas $G_N = \frac{1}{N} \sum F_n$. Por lo tanto $G_N \leq F_N$ y G_N son monótonas decrecientes. También se debe aproximar al mismo límite. Usando el teorema 7.1 se observa que $\lim_{N \rightarrow \infty} G_N = H$.

Anexo D

Maximizar la tasa para un sistema de restricciones

Supongamos que tenemos un conjunto de restricciones sobre secuencias de símbolos que es del tipo de estado finito y puede ser representado por una gráfica lineal. Siendo $l_{ij}^{(s)}$ el largo de varios símbolos que pueden pasar de estado i a estado j . ¿Qué distribución de probabilidades P_i para los estados diferentes y $p_{ij}^{(s)}$ para elegir un símbolo s en estado i e ir a estado j maximiza la tasa de generación de información en virtud de estas limitaciones? Las limitaciones definen un canal discreto y la velocidad máximo debe ser menor o igual a la capacidad C de este canal, ya que si todos los bloques de longitud grande eran igualmente probables, esta tasa daría lugar, y si es posible que esto sería mejor. Se mostrará que esta tasa se puede lograr mediante la elección adecuada de P_i y $p_{ij}^{(s)}$. La tasa es:

$$\frac{-\sum P_i p_{ij}^{(s)} \log p_{ij}^{(s)}}{\sum P_i p_{ij}^{(s)} l_{ij}^{(s)}} = \frac{N}{M}. \quad (D.1)$$

Siendo $l_{ij} = \sum_s l_{ij}^{(s)}$. Evidentemente para un máximo de $p_{ij}^{(s)} = \exp l_{ij}^{(s)}$. Las restricciones a la maximización son $\sum P_i = 1$, $\sum P_i (p_{ij} - \delta_{ij}) = 0$. Por lo tanto podemos maximizar

$$U = \frac{-\sum P_i p_{ij} \log p_{ij}}{\sum P_i p_{ij} l_{ij}} + \lambda \sum_i P_i + \sum \mu_i p_{ij} + \sum \eta_j P_i (p_{ij} - \delta_{ij}) \quad (D.2)$$

$$U = \frac{-\sum P_i p_{ij} \log p_{ij}}{\sum P_i p_{ij} l_{ij}} + \lambda \sum_i P_i + \sum \mu_i p_{ij} + \sum \eta_j P_i (p_{ij} - \delta_{ij}) \quad (D.3)$$

$$\frac{\partial U}{\partial p_{ij}} = -\frac{MP_i(1 + \log p_{ij}) + NP_i l_{ij}}{M^2} + \lambda + \mu_i + \eta_j P_i = 0. \quad (D.4)$$

Resolviendo para p_{ij}

$$p_{ij} = A_i B_j D^{-l_{ij}} \quad (D.5)$$

Desde

$$\sum_j p_{ij} = 1, \quad A_i^{-1} = \sum_j B_j D^{-l_{ij}} \quad (\text{D.6})$$

$$p_{ij} = \frac{B_j D^{-l_{ij}}}{\sum_s B_s D^{-l_{is}}} \quad (\text{D.7})$$

para luego

$$p_{ij} = \frac{B_j}{B_i} C^{-l_{ij}} \quad (\text{D.8})$$

$$\sum p_i \frac{B_j}{B_i} C^{-l_{ij}} = P_j \quad (\text{D.9})$$

o

$$\sum \frac{P_i}{B_i} C^{-l_{ij}} = \frac{P_j}{B_j}. \quad (\text{D.10})$$

Entonces si λ_i satisface

$$\sum \gamma_i C^{-l_{ij}} = \gamma_j \quad (\text{D.11})$$

$$P_i = B_i \gamma_i. \quad (\text{D.12})$$

Tanto los conjuntos de ecuaciones para B_i and γ_i puede ser satisfecha ya que C es

$$|C^{-l_{ij}} - \delta_{ij}| = 0. \quad (\text{D.13})$$

En este caso la tasa es

$$\sum \frac{P_i p_{ij} \log \frac{B_j}{B_i} C^{-l_{ij}}}{\sum P_i p_{ij} l_{ij}} = C - \frac{P_i p_{ij} \log \frac{B_j}{B_i}}{\sum P_i p_{ij} l_{ij}} \quad (\text{D.14})$$

pero

$$\sum P_i p_{ij} (\log B_j - \log B_i) = \sum_j P_j \log B_j - \sum_i P_i \log B_i = 0 \quad (\text{D.15})$$

Por lo tanto la tasa es C y como esto nunca se podría superar este es el máximo, lo que justifica la solución supuesta.

Parte III

Preliminares matemáticos

En esta entrega final del documento se aborda el caso donde las señales o los mensajes, o ambos, son variables continuas, en contraste con la naturaleza discreta asumida hasta ahora. En gran medida, el caso continuo puede obtenerse a través de un proceso limitado del caso discreto dividiendo la continuidad de mensajes y señales en un número elevado pero finito de pequeñas regiones y calculando los diferentes parámetros que intervienen en una base discreta. A medida que el tamaño de las regiones se disminuye, desde el enfoque general, estos parámetros limitan los valores adecuados para el caso continuo. Sin embargo, hay algunos efectos nuevos que aparecen y también un cambio general del énfasis en la dirección de la especialización de los resultados generales a casos particulares.

No vamos a intentar, en el caso continuo, obtener nuestros resultados con la mayor generalidad, o con el rigor extremo de la matemática pura, ya que esto implicaría una gran cantidad de teoría de la medida abstracta y oscurecería el hilo principal del análisis. Sin embargo, un estudio preliminar indica que la teoría puede ser formulada de una manera completamente axiomática y rigurosa que incluye tanto los casos continuos y discretos, y muchos otros.

Capítulo 15

Conjuntos y familias de funciones

Tendremos que hacer frente en el caso continuo con conjuntos de funciones y familias de funciones. Un conjunto de funciones, como el nombre implica, es simplemente una clase o colección de funciones, generalmente de una variable, el tiempo. Puede ser especificado dando una representación explícita de las diversas funciones en el conjunto, o implícitamente, dando una propiedad cuya función en el conjunto poseen y otros no lo hacen. Algunos ejemplos son:

1. El conjunto de funciones:

$$f_{\theta}(t) = \text{sen}(t + \theta). \quad (15.1)$$

Cada valor particular de θ determina una función particular en el conjunto.

2. El conjunto de todas las funciones de tiempo no conteniendo frecuencias sobre W ciclos por segundo.
3. El conjunto de todas las funciones limitadas en banda a W y amplitud en A .
4. El conjunto de todas las señales de habla inglesa como funciones de tiempo.

Una familia de funciones es un conjunto de funciones junto con una medida de probabilidad mediante el cual se puede determinar la probabilidad de una función en el conjunto que tiene ciertas propiedades.¹ Por ejemplo con el conjunto,

$$f_{\theta}(t) = \text{sen}(t + \theta), \quad (15.2)$$

podemos tener una distribución de probabilidad para θ , $P(\theta)$. El conjunto se convierte en una familia. Algunos otros ejemplos de familias de funciones son:

1. Un conjunto finito de funciones $f_k(t)$ ($k = 1, 2, \dots, n$) con la probabilidad de f_k siendo p_k .
2. Una familia de dimensión finita de funciones

$$f(\alpha_1, \alpha_2, \dots, \alpha_n; t) \quad (15.3)$$

con una distribución de probabilidad sobre los parámetros α_i :

$$p(\alpha_1, \dots, \alpha_n). \quad (15.4)$$

¹En terminología matemática, las funciones pertenecen a un espacio de medida cuya medida total es la unidad.

Por ejemplo podemos considerar la familia definida por

$$f(a_1, \dots, a_n, \theta_1, \dots, \theta_n; t) = \sum_{i=1}^n a_i \sin i(\omega t + \theta_i) \quad (15.5)$$

con las amplitudes a_i distribuidas normalmente e independientemente, y las fases θ_i distribuidas uniformemente (desde 0 a 2π) e independientemente.

3. La familia

$$f(a_i, t) = \sum_{n=-\infty}^{+\infty} a_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)} \quad (15.6)$$

con la a_i normal e independiente todas con la misma desviación estándar \sqrt{N} . Esta es una representación de ruido “blanco”, banda limitada a la banda de 0 a W ciclos por segundo y con potencia media de N .²

4. Los puntos se distribuirán en el eje t de acuerdo con una distribución de Poisson. En cada punto seleccionado la función $f(t)$ es colocada y las diferentes funciones agregadas, dando la familia

$$\sum_{k=-\infty}^{\infty} f(t + t_k) \quad (15.7)$$

donde los t_k son los puntos de la distribución Poisson. Esta familia puede ser considerada como un tipo de impulso o disparo de ruido donde todos los impulsos son idénticos.

5. El conjunto de todas las funciones de habla inglesa con la medida de probabilidad dada por la frecuencia de ocurrencia en el uso ordinario.

Una *familia* de funciones $f_\alpha(t)$ es *estacionaria* si la misma familia resulta cuando todas las funciones son desplazadas una cantidad fija de tiempo. La familia

$$f_\theta(t) = \sin(t + \theta) \quad (15.8)$$

es estacionario si θ es distribuido uniformemente desde 0 a 2π . Si desplazamos cada función en t_1 obtenemos

$$f_\theta(t + t_1) = \sin(t + t_1 + \theta) \quad (15.9)$$

$$f_\theta(t + t_1) = \sin(t + \varphi) \quad (15.10)$$

con φ distribuida uniformemente desde 0 a 2π . Cada función ha cambiado, pero la familia como un todo es invariante por el desplazamiento. Los otros ejemplos dados anteriormente son también estacionarios.

Una familia es *ergódica* si es estacionaria, y no existe un subconjunto de las funciones en el conjunto con una probabilidad distinta de 0 y 1 que es estacionaria. La familia

$$\sin(t + \theta) \quad (15.11)$$

es ergódica. Ningún subconjunto de estas funciones de probabilidad $\neq 0, 1$ se transforma en sí mismo bajo todos los desplazamientos en el tiempo. Por otra parte la familia

$$a \sin(t + \theta) \quad (15.12)$$

²Esta representación puede ser utilizada como una definición de banda de ruido blanco limitada. Esto tiene ciertas ventajas que implican un menor número de operaciones limitantes que usan definiciones que se han utilizado en el pasado. El nombre de “ruido blanco”, ya firmemente arraigada en la literatura, es tal vez un poco desafortunado. En óptica, luz blanca significa cualquier espectro continuo en contraste con un espectro de punto, o un espectro que es plano con una *longitud de onda* (que no es el mismo que un espectro plano con frecuencia).

con a distribuida normalmente y θ uniformemente, es estacionaria pero no ergódica. El subconjunto de estas funciones con a entre 0 y 1, por ejemplo, es estacionario.

De los ejemplos dados, el 3 y 4 son ergódicos, y el 5 puede quizás ser considerado así. Si una familia es ergódica, podemos decir que aproximadamente cada función en el conjunto es típica de la familia.

Más precisamente, se sabe que con un conjunto ergódico un promedio de cualquier estadística sobre el conjunto es igual (con una probabilidad de 1) a un promedio sobre los desplazamientos de tiempo de una función particular del conjunto.³ En términos generales, en cada función se puede esperar que, a medida que avanza el tiempo, pase con la frecuencia adecuada todas las convoluciones de cualquiera de las funciones en el conjunto.

Del mismo modo que se pueden realizar diversas operaciones sobre los números o funciones para obtener nuevos números o funciones, podemos realizar operaciones sobre familias para obtener nuevas familias. Supongamos por ejemplo que tenemos una familia de funciones $f_\alpha(t)$ y un operador T que nos da por cada función $f_\alpha(t)$ una función resultante $g_\alpha(t)$:

$$g_\alpha(t) = Tf_\alpha(t). \quad (15.13)$$

La medida de probabilidad se define por el conjunto de $g_\alpha(t)$ por medio del conjunto $f_\alpha(t)$. La probabilidad de un cierto subconjunto de las funciones $g_\alpha(t)$ es igual a la del subconjunto de las funciones $f_\alpha(t)$ que producen los miembros del subconjunto dado de funciones g bajo la operación T . Físicamente esto corresponde al pasar el conjunto a través de algún dispositivo, por ejemplo, un filtro, un rectificador o un modulador. Las funciones de salida del dispositivo forman la familia $g_\alpha(t)$.

Un dispositivo u operador T se llama invariante si desplazando la entrada simplemente se desplaza la salida, es decir, si

$$g_\alpha(t) = Tf_\alpha(t) \quad (15.14)$$

implica

$$g_\alpha(t + t_1) = Tf_\alpha(t + t_1) \quad (15.15)$$

para toda $f_\alpha(t)$ y toda t_1 . Esto es fácilmente demostrado (ver Apéndice 5) que si T es invariante y la familia de entrada es estacionaria, luego la familia de salida es estacionaria. Del mismo modo, si la entrada es ergódica, la salida también será ergódica.

Un filtro o un rectificador es invariante bajo todos los desplazamientos de tiempo. La operación de modulación no es desde la fase portadora que proporciona una estructura de tiempo determinado. Sin embargo, la modulación es invariante bajo todos los desplazamientos que son múltiplos del período del portador.

Wiener ha señalado la íntima relación entre la invariancia de dispositivos físicos en desplazamientos en tiempo y la teoría de Fourier.⁴ De hecho, él ha demostrado que si un dispositivo es lineal, así

³Este es el famoso teorema ergódico o más bien un aspecto de este teorema que fue demostrado en formulaciones algo diferentes por Birkoff, von Neumann, y Koopman, y posteriormente generalizada por Wiener, Hopf, Hurewicz y otros. La literatura sobre la teoría ergódica es bastante extensa y se remite al lector a los trabajos de estos autores para formulaciones precisas y generales; por ejemplo, E. Hopf, "Ergodentheorie," *Ergebnisse der Mathematik und ihrer Grenzgebiete*, v. 5; "On Causality Statistics and Probability," *Journal of Mathematics and Physics*, v. XIII, No. 1, 1934; N. Wiener, "The Ergodic Theorem," *Duke Mathematical Journal*, v. 5, 1939.

⁴La teoría de la comunicación está muy en deuda con Wiener por gran parte de su filosofía y teoría. Su artículo clásico NDRC, *The Interpolation, Extrapolation and Smoothing of Stationary Time Series* (Wiley, 1949), contiene la primera formulación clara de la teoría de la comunicación como un problema estadístico, el estudio de las operaciones en series de tiempo. Este trabajo, aunque ocupa principalmente de la predicción lineal y el problema de filtración, es una referencia colateral importante en relación con el presente documento. También podemos referirnos aquí a *Wiener's Cybernetics* (Wiley, 1948), que trata de los problemas generales de comunicación y control.

como el invariante análisis de Fourier, es entonces la herramienta matemática adecuada para tratar con el problema.

Una familia de funciones es la representación matemática adecuada de los mensajes producidos por una fuente continua (por ejemplo, el habla), de las señales producidas por un transmisor, y el del ruido perturbador. La teoría de la comunicación se interesa específicamente, como se ha enfatizado por Wiener, no con las operaciones en funciones particulares, pero con las operaciones sobre familias de funciones. Un sistema de comunicación no está diseñado para una función de habla particular y menos aún para una onda sinusoidal, pero si para la familia de las funciones del habla.

Capítulo 16

Funciones de familias con banda limitada

Si la función de tiempo $f(t)$ es limitada a la banda de 0 a W ciclos por segundo, este es completamente determinado dando sus ordenadas en una serie de puntos discretos espaciados $\frac{1}{2W}$ segundos aparte de la manera indicada por el siguiente resultado.⁵

Teorema 16.1. *No dejar que $f(t)$ contenga frecuencias por encima de W . Entonces*

$$f(t) = \sum_{-\infty}^{\infty} X_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)} \quad (16.1)$$

donde

$$X_n = f\left(\frac{n}{2W}\right). \quad (16.2)$$

En esta expansión $f(t)$ es representada como una suma de funciones ortogonales. El coeficiente X_n de los diversos términos puede considerarse como coordenadas en una dimensión infinita “espacio funcional”. En este espacio cada función corresponde precisamente un punto y cada punto a una función.

Una función puede ser considerada para estar sustancialmente limitada a un tiempo T si todas las ordenadas X_n fuera de este intervalo de tiempo es cero. En este caso todo pero $2TW$ de las coordenadas serían cero. Así funciones limitadas a una banda W y duración T corresponden a puntos en un espacio de $2TW$ dimensiones.

Un subconjunto de funciones de banda W y duración T corresponde a una región en este espacio. Por ejemplo, las funciones cuya energía total es menor que o igual a E corresponden a puntos en una esfera dimensional $2WT$ con radio $r = \sqrt{2WE}$.

Una familia de funciones de duración limitada y la banda representada por una distribución de probabilidad $p(x_1, \dots, x_n)$ en el correspondiente espacio n dimensional. Si la familia no está limitada en el tiempo se pueden considerar las coordenadas $2TW$ en un intervalo T para representar sustancialmente la parte de la función en el intervalo T y la distribución de probabilidad $p(x_1, \dots, x_n)$ para dar la estructura estadística de la familia para intervalos de esa duración.

⁵Para una demostración de este teorema y discusión adicional, véase el artículo del autor “Communication in the Presence of Noise” publicado en *Proceedings of the Institute of Radio Engineers*, v. 37, No. 1, Enero, 1949, pp. 10-21.

Capítulo 17

Entropía de una distribución continua

La entropía de un conjunto discreto de probabilidades p_1, \dots, p_n ha sido definido como:

$$H = - \sum p_i \log p_i. \quad (17.1)$$

De manera análoga se define la entropía de una distribución continua con la función de la distribución de densidad $p(x)$ por:

$$H = - \int_{-\infty}^{\infty} p(x) \log p(x) dx. \quad (17.2)$$

Con una distribución n dimensional $p(x_1, \dots, x_n)$ tenemos

$$H = - \int \dots \int p(x_1, \dots, x_n) \log p(x_1, \dots, x_n) dx_1 \dots dx_n. \quad (17.3)$$

Si tenemos dos argumentos x y y (que pueden ser ellos mismos multidimensionales) las entropías conjuntas y condicionales de $p(x, y)$ están dadas por

$$H(x, y) = - \iint p(x, y) \log p(x, y) dx dy \quad (17.4)$$

y

$$H_x(y) = - \iint p(x, y) \log \frac{p(x, y)}{p(x)} dx dy \quad (17.5)$$

$$H_y(x) = - \iint p(x, y) \log \frac{p(x, y)}{p(y)} dx dy \quad (17.6)$$

donde

$$p(x) = \int p(x, y) dy \quad (17.7)$$

$$p(y) = \int p(x, y) dx. \quad (17.8)$$

Las entropías de distribuciones continuas tienen la mayoría (pero no todos) las propiedades del caso discreto. En particular, tenemos lo siguiente:

1. Si x es limitado a un cierto volumen v en su espacio, entonces $H(x)$ es un máximo e igual a $\log v$ cuando $p(x)$ es constante ($1/v$) en el volumen.
2. Con cualesquiera dos variables x, y tenemos:

$$H(x, y) \leq H(x) + H(y) \quad (17.9)$$

con igualdad si (y solo si) x y y son independientes, por ejemplo, $p(x, y) = p(x)p(y)$ (además posiblemente un conjunto de puntos de probabilidad cero).

3. Considere una operación de promedio generalizada del siguiente tipo:

$$p'(y) = \int a(x, y) p(x) dx \quad (17.10)$$

con

$$\int a(x, y) dx = \int a(x, y) dy = 1, a(x, y) \geq 0 \quad (17.11)$$

Entonces la entropía de la distribución promediada $p'(y)$ es igual o mayor a la distribución original $p(y)$.

4. Tenemos:

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x) \quad (17.12)$$

y

$$H_x(y) \leq H(y). \quad (17.13)$$

5. Dejamos $p(y)$ ser una distribución unidimensional. La forma de $p(y)$ dando una máxima entropía sujeto a la condición de que la desviación estándar de x es fija en σ es Gaussiana. Para demostrar ésto debemos maximizar:

$$H(x) = - \int p(x) \log p(x) dx \quad (17.14)$$

con

$$\sigma^2 = \int p(x) x^2 dx \quad y \quad 1 = \int p(x) dx \quad (17.15)$$

como límites. Ésto requiere, por el cálculo de variaciones, maximizar:

$$\int [-p(x) \log p(x) + \lambda p(x) x^2 + \mu p(x)] dx \quad (17.16)$$

La condición para ésto es

$$-1 - \log p(x) + \lambda x^2 + \mu = 0 \quad (17.17)$$

y consecuentemente (ajustando las constantes para satisfacer los límites)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x^2/2\sigma^2)} \quad (17.18)$$

Del mismo modo en n dimensiones, supongamos que los momentos de segundo orden de $p(x_1, \dots, x_n)$ son fijos en A_{ij} :

$$A_{ij} = \int \dots \int x_i x_j p(x_1, \dots, x_n) dx_1 \dots dx_n \quad (17.19)$$

Entonces la máxima entropía ocurre (por un cálculo similar) cuando $p(x_1, \dots, x_n)$ es la distribución Gaussiana n dimensional con los momentos de segundo orden A_{ij} .

6. La entropía de una distribución Gaussiana undimensional cuya desviación estandar es σ está dada por

$$H(x) = \log \sqrt{2\pi e \sigma} \quad (17.20)$$

Ésta es calculada de la siguiente forma

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x^2/2\sigma^2)} \\ -\log p(x) &= \log \sqrt{2\pi}\sigma + \frac{x^2}{2\sigma^2} \\ H(x) &= -\int p(x) \log p(x) dx \\ &= \int p(x) \log \sqrt{2\pi}\sigma dx + \int p(x) \frac{x^2}{2\sigma^2} dx \\ &= \log \sqrt{2\pi}\sigma + \frac{\sigma}{2\sigma^2} \\ &= \log \sqrt{2\pi}\sigma + \log \sqrt{e} \\ &= \log \sqrt{2\pi e \sigma} \end{aligned} \quad (17.21)$$

Asimismo, la distribución Gaussiana n-dimensional con una forma cuadrática asociada A_{ij} está dada por:

$$p(x_1, \dots, x_n) = \frac{|a_{ij}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum a_{ij} x_i x_j\right) \quad (17.22)$$

y la entropía puede se calculada como:

$$H = \log (2\pi e)^{n/2} |a_{ij}|^{-\frac{1}{2}}, \quad (17.23)$$

donde $|a_{ij}|$ es la determinante cuyos elementos son a_{ij} .

7. Si x es limitada a una media línea ($p(x) = 0$ para $x \leq 0$ y el primer momento de x es fijado en a :

$$a = \int_0^\infty p(x) x dx, \quad (17.24)$$

entonces la máxima entropía ocurrirá cuando

$$p(x) = \frac{1}{a} e^{-(x/a)} \quad (17.25)$$

y es igual a $\log ea$.

8. Hay una importante diferencia entre la entropía continua y discreta. En el caso discreto la entropía mide de un modo absoluto la aleatoriedad de la variable de oportunidad. En el caso continuo la medición es relativa al sistema de coordenadas. Si cambiamos las coordenadas la entropía cambiará de forma general. En efecto, si cambiamos las coordenadas $y_1 \dots y_n$ la nueva entropía estará dada por:

$$H(y) = \int \dots \int p(x_1, \dots, x_n) J\left(\frac{x}{y}\right) \log p(x_1, \dots, x_n) J\left(\frac{x}{y}\right) dy_1 \dots dy_n, \quad (17.26)$$

donde $J\left(\frac{x}{y}\right)$ es el Jacobiano de la transformación de las coordenadas. En la ampliación de los logaritmos y cambiando las variables a $x_1 \dots x_n$, obtendremos:

$$H(y) = H(x) - \int \dots \int p(x_1, \dots, x_n) \log J\left(\frac{x}{y}\right) dx_1 \dots dx_n \quad (17.27)$$

Por consiguiente, la nueva entropía es la vieja entropía menos el logaritmo esperado del Jacobiano. En el caso continuo la entropía puede considerarse una medida de aleatoriedad relativa

a un estándar asumido, es decir, el sistema de coordenadas elegido con cada pequeño elemento de volumen $dx_1 \dots dx_n$ dado el mismo peso. Cuando cambiamos el sistema de coordenadas, la entropía en el nuevo sistema mide la aleatoriedad cuando a elementos de volumen iguales $dy_1 \dots dy_n$ en el nuevo sistema se les da el mismo peso.

A pesar de esta dependencia en el sistema de coordenadas, el concepto de entropía es tan importante en el caso continuo como en el caso discreto. Ésto se debe al hecho de que los conceptos derivados de la tasa de información y la capacidad del canal dependen de la diferencia de ambas entropías y esta diferencia no depende del marco de coordenadas, cada uno de los dos terminos serán cambiados por la misma cantidad. La entropía de una distribución continua puede ser negativa. La escala de medición establece un zero arbitrario correspondiente a una distribución uniforme en una unidad de volumen. Una distribución que es más limitada que ésto tiene menos entropía y será negativa. Las tasas y capacidades serán, sin embargo, siempre no negativas.

9. Un caso particular de cambio de coordenadas es la transformación lineal

$$y_j = \sum_i a_{ij} x_i. \quad (17.28)$$

En este caso el Jacobiano es simplemente la determinante $|a_{ij}|^{-1}$, y

$$H(y) = H(x) + \log |a_{ij}|. \quad (17.29)$$

En el caso de la rotación de coordenadas (o cualquier medición preservando la transformación) $J = 1$ y $H(y) = H(x)$.

Capítulo 18

Entropía en un conjunto de funciones

Considere un conjunto de funciones ergódico limitado a una cierta banda de ancho W ciclos por segundo. Dejar

$$p(x_1, \dots, x_n) \quad (18.1)$$

ser la función de distribución de densidad para amplitudes $x_1 \dots x_n$ en n puntos sucesivos de muestra. Definimos la entropía del conjunto por grado de libertad como:

$$H' = - \lim_{n \rightarrow \infty} \frac{1}{n} \int \dots \int p(x_1, \dots, x_n) \log p(x_1, \dots, x_n) dx_1 \dots dx_n \quad (18.2)$$

También definiremos la entropía H por segundo dividiendo, no por n , sino por el tiempo T en segundos para n muestras. Desde que $n = 2TW$, $H = 2WH'$.

Con un blanco ruido térmico p es Gaussiano, tenemos:

$$H' = \log \sqrt{2\pi eN} \quad (18.3)$$

$$H = W \log 2\pi eN \quad (18.4)$$

Para una determinada potencia media N , el ruido blanco tiene la máxima entropía posible. Esto se deduce de las propiedades de maximizar la distribución Gaussiana señalada arriba.

La entropía de un proceso estocástico continuo tiene muchas propiedades análogas a ello para procesos discretos. En el caso discreto, la entropía está relacionada al logaritmo de la probabilidad de largas secuencias, y al número de secuencias razonablemente probables de larga longitud. En el caso continuo se relaciona de una manera similar al logaritmo de la densidad de probabilidad de una larga serie de muestras, y el volumen de una probabilidad razonablemente alta en el espacio de la función.

Más precisamente, si asumimos que $p(x_1, \dots, x_n)$ continua en todas las x_i y para toda n , entonces para una suficientemente larga n

$$\left| \frac{\log p}{n} - H' \right| < \varepsilon \quad (18.5)$$

para todas las opciones de (x_1, \dots, x_n) aparte de un conjunto cuya probabilidad total es menor, arbitrariamente pequeña. Lo siguiente forma la propiedad ergódica si dividimos el espacio en un gran número de pequeñas celdas. La relación de H a un volumen puede enunciarse como sigue: Bajo las mismas suposiciones consideré el espacio n -dimensional correspondiente a $p(x_1, \dots, x_n)$. Dejar $V_n(q)$ ser el menor volumen en este espacio que incluye en su interior una probabilidad total de q . Entonces:

$$\lim_{n \rightarrow \infty} \frac{\log V_n(q)}{n} = H' \quad (18.6)$$

dado q no es igual a 0 o 1.

Estos resultados muestran que para una gran n existe un volumen bien definido (al menos en el sentido logarítmico) de alta probabilidad, y que dentro de este volumen la densidad de probabilidad es relativamente uniforme (de nuevo en el sentido logarítmico).

En el caso del ruido blanco, la función de distribución está dada por:

$$p(x_1, \dots, x_n) = \frac{1}{(2\pi N)^{n/2}} \exp -\frac{1}{2N} \sum x_i^2 \quad (18.7)$$

Dado que éste solo depende de x , entonces, las superficies de igual densidad de probabilidad son esferas y la distribución entera tiene simetría esférica. La región de alta probabilidad es una esfera de radio \sqrt{nN} . Como $n \rightarrow \infty$ la probabilidad de caer fuera de una esfera de radio $\sqrt{n(N + \varepsilon)}$ se aproxima a cero y $\frac{1}{n}$ veces el logaritmo del volumen de la esfera se aproxima a $\log \sqrt{2\pi e N}$. En el caso continuo es conveniente no trabajar con la entropía H de un conjunto pero con una cantidad derivada la cual llamaremos potencia de entropía. Esto es definido como el poder en el ruido blanco limitado a la misma banda que el original y teniendo la misma entropía. En otras palabras, si H' es la entropía de un conjunto, su potencia de entropía es:

$$N_1 = \frac{1}{2\pi e} \exp 2H'. \quad (18.8)$$

En la imagen geométrica esta cantidad de medir el volumen de mayor probabilidad por el radio al cuadrado de una esfera teniendo el mismo volumen. Desde que el ruido blanco tiene la máxima entropía para una potencia dada, la potencia de entropía de cualquier ruido es menor o igual a su actual potencia.

Capítulo 19

Pérdida de entropía en filtros lineales

Teorema 19.1. *Si un conjunto tiene una entropía H_1 por grado de libertad en la banda W , y es transmitido a través de un filtro con características $Y(f)$, el conjunto de salida tendrá la entropía*

$$H_2 = H_1 + \frac{1}{W} \int_W \log |Y(f)|^2 df. \quad (19.1)$$

La operación de un filtro es esencialmente una transformación lineal de coordenadas. Si pensamos en los diferentes componentes frecuentes como el sistema original de coordenadas, los nuevos componentes frecuentes son meramente los viejos multiplicados por factores. La matriz de transformación de coordenadas es entonces esencialmente diagonalizada en términos de estas coordenadas. El Jacobiano de las transformaciones es (para los componentes n seno y n coseno):

$$J = \prod_{i=1}^n |Y(f_i)|^2, \quad (19.2)$$

donde la f_i esta igualmente espaciada a través de la banda W . Esto ocurre en el límite:

$$\exp \frac{1}{W} \int_W \log |Y(f)|^2 df \quad (19.3)$$

Desde que J es constante, su valor promedio es la misma cantidad y aplicando el teorema en el cambio de entropía con un cambio en las coordenadas, resultará lo siguiente. Podremos redactarlo en términos de la potencia de entropía. Así si la potencia de entropía del primer conjunto es N_1 entonces la del segundo será

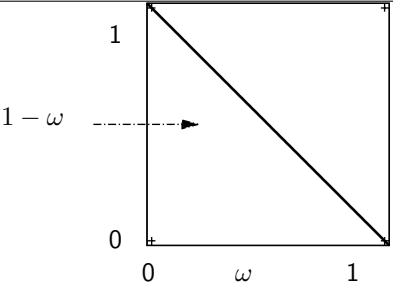
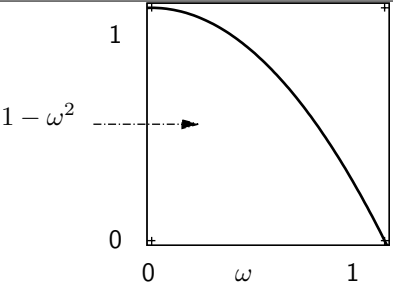
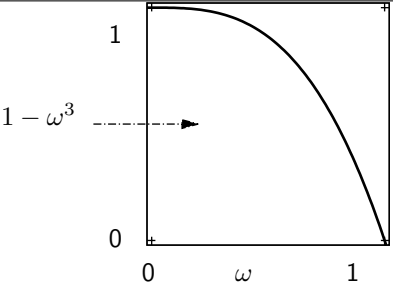
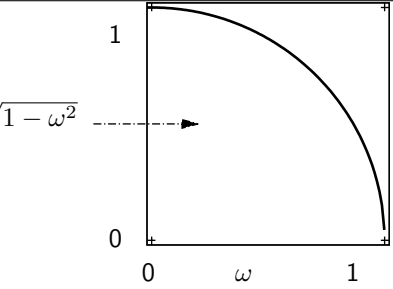
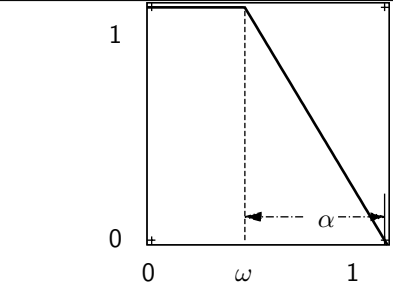
$$N_1 \exp \frac{1}{W} \int_W \log |Y(f)|^2 df. \quad (19.4)$$

La potencia de entropía final es la entropía inicial multiplicada por la ganancia media geométrica del filtro. Si la ganancia es medida en db , entonces la potencia de entropía de salida será incrementada por la ganancia media aritmética db sobre W .

En la tabla 1, la pérdida de potencia de entropía ha sido calculada (también expresada en db) para un número ideal de ganancias características. Las respuestas impulsivas de estos filtros estan dadas también por $W = 2\pi$, con fase en cero.

La pérdida de entropía para otros muchos casos puede ser obtenida desde estos resultados. Por ejemplo, la del factor de potencia de entropía $\frac{1}{e^2}$ para el primer caso también aplica para cualquier ganancia característica obtenida desde $1 - \omega$ por una medida preservando la transformación del eje ω . En una particular ganancia que incrementa linealmente $G(\omega) = \omega$, o un “diente de sierra” característico entre cero y uno tienen la misma pérdida de entropía. La ganancia recíproca tiene el factor recíproco. Entonces, $\frac{1}{\omega}$ tiene el factor e^2 . Aumentando la ganancia a cualquier potencia incrementa el factor a esta potencia.

Cuadro 19.1: Ganancia, factor de potencia de la entropía y ganancia en decibeles, y la respuesta impulso.

GANANCIA	FACTOR DE POTENCIA DE LA ENTROPÍA	GANANCIA DE ENERGÍA DE LA ENTROPÍA EN DECIBELES	RESPUESTA IMPULSO
	$\frac{1}{e^2}$	-8,69	$\frac{\sin^2(t/2)}{t^2/2}$
	$\left(\frac{2}{e}\right)^4$	-5,33	$2 \left[\frac{\sin t}{t^3} - \frac{\cos t}{t^2} \right]$
	0,411	-3,87	$6 \left[\frac{\cos t - 1}{t^4} - \frac{\cos t}{2t^2} + \frac{\sin t}{t^3} \right]$
	$\left(\frac{2}{e}\right)^2$	-2,67	$\frac{\pi}{2} \frac{J_1(t)}{t}$
	$\frac{1}{e^{2\alpha}}$ 63	-8,69α	$\frac{1}{\alpha t^2} [\cos(1 - \alpha)t - \cos t]$

Capítulo 20

Entropía de la suma de dos conjuntos

Si tenemos dos conjuntos de funciones $f_\alpha(t)$ y $g_\beta(t)$ podemos formar un nuevo conjunto por “adición”. Supongamos que el primer conjunto tiene la función de densidad de probabilidad $p(x_1, \dots, x_n)$ y el segundo $q(x_1, \dots, x_n)$. Después la función de densidad para la adición es dada por la convolución

$$r(x_1, \dots, x_n) = \int \cdots \int p(y_1, \dots, y_n) q(x_1 - y_1, \dots, x_n - y_n) dy_1 \cdots dy_n \quad (20.1)$$

Físicamente esto corresponde a sumar los ruidos o señales representados por los conjuntos originales de las funciones

El siguiente resultado es derivado en el Anexo F

Teorema 20.1. *Deje que la potencia media de los dos conjuntos sea N_1 y N_2 y deje que sus poderes de entropía sean \bar{N}_1 y \bar{N}_2 . Entonces, el poder de entropía de la suma, N_3 , está delimitado por*

$$\bar{N}_1 + \bar{N}_2 \leq \bar{N}_3 \leq N_1 + N_2. \quad (20.2)$$

El ruido blanco Gaussiano tiene la peculiar propiedad que puede absorber cualquier otro ruido o conjunto de señales que puede ser añadido a la misma con una potencia de entropía resultante aproximadamente igual a la suma de a potencia ruido blanco y la potencia de la señal (medida apartir del valor promedio de la señal, que es normalmente cero), siempre que la potencia de la señal es pequeña, en cierto sentido, en comparación con el ruido.

Considere el espacio de la función asociada con estos conjuntos que tienen dimensiones n . El ruido blanco corresponde a la distribución Gaussiana esférica en este espacio. El conjunto de la señal corresponde a otra distribución de probabilidad, no necesariamente Gaussiana o esférica. Deje que los segundos momentos de esta distribución alrededor de su centro de gravedad sea a_{ij} . Es decir, si $p(x_1, \dots, x_n)$ es la función de distribución de densidad

$$a_{ij} = \int \cdots \int p(x_i - \alpha_i)(x_j - \alpha_j) dx_1 \cdots dx_n, \quad (20.3)$$

donde α_i son las coordenadas del centro de gravedad. Ahora a_{ij} es una forma cuadrática positiva, y podemos rotar nuestro sistema de coordenadas para alinearla con las direcciones principales de

esta forma. a_{ij} es entonces reducido a la forma diagonal b_{ii} . Se requiere que cada b_{ii} sea pequeño comparado con N , el cuadrado del radio de la distribución esférica.

En este caso, la convolución del ruido y señal producen aproximadamente una distribución Gaussiana cuya forma cuadrática correspondiente es

$$N + b_{ii}. \quad (20.4)$$

El potencial de entropía de esta distribución es

$$[\Pi(N + b_{ii})]^{\frac{1}{n}} \quad (20.5)$$

o aproximadamente

$$\begin{aligned} &\approx [(N)^n + \Sigma b_{ii}(N)^n - 1]^{\frac{1}{n}} \\ &= N + \frac{1}{n}\Sigma b_{ii}. \end{aligned} \quad (20.6)$$

El último término es la potencia de la señal, mientras que el primero es la potencia del ruido.

Parte IV

El canal continuo

Capítulo 21

La capacidad de un canal continuo

En un canal continuo de las señales de entrada o transmitidas serán funciones continuas de tiempo $f(t)$ pertenecientes a un determinado conjunto, y la señales de salida o recibidas serán versiones perturbadas de estas. Vamos a considerar solo el caso en que ambas señales transmitidas y recibidas se limitan a una determinada banda W . Pueden ser después identificadas por un tiempo T , por los números $2TW$, y su estructura estadística de las funciones de distribución finitos tridimensionales. Así las estadísticas de la señal transmitida será determinada por

$$P(x_1, \dots, x_n) = P(x) \quad (21.1)$$

y los del ruido por la distribución de probabilidad condicional

$$P_{x_1, \dots, x_n}(y_1, \dots, y_n) = P_x(y). \quad (21.2)$$

La tasa de transmisión de información por un canal continuo se define de una manera análoga a la de un canal separado, esto es

$$R = H(x) - H_y(x), \quad (21.3)$$

donde $H(x)$ es la entropía de la entrada y $H_y(x)$ el equívoco. La capacidad del canal C se define como el máximo de R cuando varían la entrada para todos los conjuntos posibles. Esto significa que en una aproximación dimensional finita debemos variar $P(x) = P(x_1, \dots, x_n)$ y maximizar

$$- \int P(x) \log P(x) dx + \int \int P(x, y) \log \frac{P(x, y)}{P(y)} dx dy. \quad (21.4)$$

Esto puede ser escrito

$$\int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy \quad (21.5)$$

usando el hecho que

$$\int \int P(x, y) \log P(x) dx dy = \int P(x) \log P(x) dx. \quad (21.6)$$

La capacidad del canal se expresa así:

$$C = \limsup_{T \rightarrow \infty} \frac{1}{T} \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy. \quad (21.7)$$

Es obvio que en esta forma R y C son independientes del sistema de coordenadas dado que el numerador y denominador en $\log \frac{P(x,y)}{P(x)P(y)}$ será multiplicado por los mismos factores cuando x y y son transformados en cualquier forma uno a uno. Esta expresión integral para C es más general que $H(x) - H_y(x)$. Correctamente interpretada (ver Anexo 7) que siempre existirá mientras $H(x) - H_y(x)$ puede asumir una forma indeterminada $\infty - \infty$ en algunos casos. Esto ocurre, por ejemplo, si x está limitada a una superficie de menos dimensiones que n en su aproximación n dimensional.

Si la base logarítmica utilizada en computación $H(x)$ y $H_y(x)$ es dos entonces C es el número máximo de dígitos binarios que pueden ser enviados por segundo a través del canal con equivocación arbitrariamente pequeña, al igual que en el caso discreto. Esto se puede ver físicamente al dividir el espacio de señales en un gran número de celdas pequeñas y suficientemente pequeño para que la densidad de probabilidad $P_x(y)$ de la señal x que está siendo perturbado hasta el punto de que y es sustancialmente constante en una celda (ya sea de x o y). Si las celdas son consideradas como puntos distintos, la situación es esencialmente la misma que un canal discreto y las pruebas usadas se aplicarán allá. Pero está claro que físicamente esta cuantificación del volumen en puntos individuales no puede de ninguna manera práctica alterar significativamente la respuesta final, siempre que las regiones sean suficientemente pequeñas. Así la capacidad será el límite de las capacidades de las subdivisiones discretas y esto es sólo la capacidad continua definida anteriormente.

En el lado matemático se puede demostrar primero (ver el Apéndice 7) que si u es el mensaje, x es la señal, y es la señal recibida (perturbada por el ruido) y v es el mensaje recuperado, entonces

$$H(x) - H_y(x) \leq H(u) - H_v(u). \quad (21.8)$$

Independientemente de lo que las operaciones realizan en u para obtener x o en y para obtener v . No importa como modificamos los dígitos binarios para obtener la señal, o como decodificamos la señal recibida para recuperar el mensaje, la tasa discreta para los dígitos binarios no excede la capacidad del canal que tenemos definida. Por otra parte, es posible bajo condiciones muy generales encontrar un sistema de codificación para transmitir dígitos binarios en la tasa C con pequeña equivocación o frecuencia de errores como se desee. Este es el caso, por ejemplo, si, cuando tomamos un espacio finito de aproximación para las funciones de las señales, $P(x, y)$ es continuo tanto en x como en y , excepto en un conjunto de puntos de probabilidad cero.

Un caso especial se produce cuando el ruido se añade a la señal y es independiente de ello (en el sentido de la probabilidad). Entonces $P_x(y)$ es una función sólo de la diferencia $n = (y - x)$,

$$P_x(y) = Q(y - x) \quad (21.9)$$

y nosotros podemos asignar una entropía definida al ruido (independientemente de las estadísticas de la señal), es decir, la entropía de la distribución $Q(n)$. Esta entropía se denotará por $H(n)$.

Teorema 21.1. Si la señal y el ruido son independientes y la señal recibida es la suma de señal transmitida y el ruido entonces la tasa de transmisión es

$$R = H(y) - H(n), \quad (21.10)$$

es decir, la entropía de la señal recibida menos la entropía del ruido.

La capacidad del canal es

$$C = \limsup_{P(x)} H(y) - H(n) \quad (21.11)$$

Tenemos, desde $y = x + n$

$$H(x, y) = H(x, n). \quad (21.12)$$

Expandiendo el lado izquierdo y utilizando el hecho de que x y n son independientes

$$H(y) + H_y(x) = H(x) + H(n). \quad (21.13)$$

Por lo tanto

$$R = H(x) - H_y(x) = H(y) - H(n). \quad (21.14)$$

Puesto que $H(n)$ es independiente de $P(x)$, maximizando R requiere maximizar $H(y)$, la entropía de la señal recibida. Si existen ciertas restricciones en el conjunto de las señales transmitidas, la entropía de la señal recibida debe ser maximizada sujeto a esas restricciones.

Capítulo 22

Capacidad de la señal con una limitación de potencia media

Una sencilla aplicación del teorema 16 es el caso cuando el ruido es un ruido térmico blanco y las señales transmitidas están limitadas a un cierto promedio de potencia P . Luego las señales recibidas tienen una potencia media $P + N$ donde N es la potencia media del ruido. La entropía máxima para las señales recibidas se produce cuando también forman un ruido blanco ya que es la entropía mayor posible para una potencia $P + N$ y puede obtenerse mediante una elección adecuada de las señales transmitidas, a saber, si forman un conjunto de ruido blanco de potencia P . La entropía (por segundo) del conjunto recibido es luego

$$H(y) = W \log 2\pi e(P + N), \quad (22.1)$$

Y la entropía de ruido es

$$H(n) = W \log 2\pi eN. \quad (22.2)$$

La capacidad del canal es

$$C = H(y) - H(n) = W \log \frac{P + N}{N}. \quad (22.3)$$

Resumiendo tenemos lo siguiente:

Teorema 22.1. *La capacidad de un canal de banda W de potencia perturbada por el ruido térmico blanco N cuando la potencia de transmisión media se limita a P viene dada por*

$$C = W \log \frac{P + N}{N}. \quad (22.4)$$

Esto significa que por sistemas de codificación suficientemente implicados se puede transmitir dígitos binarios a la tasa $W \log_2 \frac{P+N}{N}$ bits por segundo, con arbitrariamente pequeña frecuencia de errores. No es posible transmitir a una velocidad mayor por cualquier sistema de codificación sin una frecuencia definida positiva de errores.

Para aproximar esta limitación de la tasa de transmisión, las señales transmitidas deben aproximarse, en propiedades estadísticas, un ruido blanco. Un sistema que se aproxima a la tasa ideal puede ser descrito como sigue: Sea $M = 2^s$ de ruido blanco sean construidas con una duración T . Estos se

asignan números binarios desde 0 a $M - 1$. En el transmisor las secuencias de mensajes se dividen en grupos de s y para cada grupo de la muestra de ruido correspondiente se transmite como señal real recibida (perturbada por el ruido) se compara con cada uno de ellos. La muestra que tiene la menor discrepancia de R.M.S de la señal recibida se elige como la señal transmitida y el correspondiente número binario reconstruido. Este proceso equivale a elegir la más probable de la señal. El número M de muestras de ruido usadas dependerá de la frecuencia admisible de errores, pero para casi todas las selecciones de muestras tenemos

$$\lim_{\epsilon \rightarrow 0} \lim_{T \rightarrow \infty} \frac{\log M(\epsilon, T)}{T} = W \log \frac{P + N}{N}, \quad (22.5)$$

entonces no importa que tan pequeño valor de ϵ se escoja, podemos, tomando valores de T suficientemente grandes, transmitir tan cerca como queramos a $TW \log \frac{P+N}{N}$ dígitos binarios en el tiempo T .

Fórmulas similares a $C = W \log \frac{P+N}{N}$ para el ruido blanco han sido desarrolladas independientemente por otros escritores, también con diferentes interpretaciones. Podemos mencionar el trabajo de N. Wiener, W. G. Tuller, y H. Sullivan en esta conexión.

En el caso de un ruido arbitrario (no necesariamente ruido blanco térmico) no parece que el problema de maximización involucrado en la determinación de la capacidad del canal C se pueden resolver explícitamente. Sin embargo, los límites superior e inferior se pueden ajustar para C en términos de la potencia media de ruido N , la entropía de potencia de ruido N_1 . Estos límites están suficientemente próximos entre sí en la mayoría de los casos prácticos de proporcionar una solución satisfactoria a este problema.

Teorema 22.2. *La capacidad de un canal de banda W perturbado por un ruido arbitrario está limitada por las desigualdades*

$$W \log \frac{P + N_1}{N_1} \leq C \leq W \log \frac{P + N}{N_1}, \quad (22.6)$$

donde

$$\begin{aligned} P &= \text{Potencia de transmisión media} \\ N &= \text{Potencia de ruido media} \\ N_1 &= \text{Potencia del ruido de la entropía} \end{aligned} \quad (22.7)$$

Aquí de nuevo la potencia media de las señales perturbadas será $P + N$. La entropía máxima para esta potencia se produciría si la señal recibida no es ruido blanco y sería $W \log 2\pi\epsilon(P + N)$. Puede que no sea posible lograr esto, es decir, no puede ser cualquier conjunto de señales transmitidas que, añadido al ruido perturbador, produce un ruido blanco térmico en el receptor, pero al menos esto establece un límite superior para $H(y)$. Tenemos, por lo tanto

$$C = \text{Max} H(y) - H(n) \leq W \log 2\pi\epsilon(P + N) - W \log 2\pi\epsilon N_1. \quad (22.8)$$

Este es el límite superior dado en el teorema. El límite inferior puede ser obtenido considerando la tasa si hacemos la señal transmitida por un ruido blanco, de potencia P . En este caso, la potencia de la entropía de la señal recibida debe ser al menos tan grande como la de un ruido blanco de potencia $P + N_1$ ya que hemos mostrado en un teorema anterior que el poder de la entropía de la suma de dos conjuntos es mayor o igual que la suma de las potencias individuales de entropía. Por lo tanto

$$\text{máx } H(y) \leq W \log 2\pi\epsilon(P + N_1) \quad (22.9)$$

y

$$C \leq W \log 2\pi e(P + N_1) - W \log 2\pi e N_1 = W \log \frac{P + N_1}{N_1}. \quad (22.10)$$

A medida que aumenta P , los límites superior e inferior se aproximan entre sí, por lo que tenemos como una tasa asintótica

$$W \log \frac{P + N}{N_1}. \quad (22.11)$$

Si el ruido es en sí blanco, $N = N_1$ y el resultado se reduce a la fórmula presentada anteriormente:

$$C = W \log \left(1 + \frac{P}{N}\right). \quad (22.12)$$

Si el ruido es Gaussiano pero con un espectro que no es necesariamente plano, N_1 es la media geométrica de la potencia de ruido en las diferentes frecuencias en la banda W . Así

$$N_1 = \exp \frac{1}{W} \int_W \log N(f) df, \quad (22.13)$$

donde $N(f)$ es la potencia de ruido a la frecuencia f .

Teorema 22.3. Si fijamos la capacidad de un transmisor de potencia P dada igual a

$$C = W \log \frac{P + N - \eta}{N_1}, \quad (22.14)$$

entonces η es monótona decreciente a medida que P aumenta y se aproxima a cero como límite.

Supongamos que para una potencia P_1 la capacidad del canal es

$$W \log \frac{P_1 + N - \eta_1}{N_1}. \quad (22.15)$$

Esto significa que la mejor distribución de la señal, por ejemplo $p(x)$, cuando se añade a la distribución de ruido $q(x)$, da una distribución recibida $r(y)$ cuya potencia de la entropía es $(P_1 + N - \eta_1)$.

Vamos a aumentar la potencia a $P_1 + \Delta P$ añadiendo un ruido blanco de potencia ΔP a la señal. La entropía de la señal recibida es ahora por lo menos

$$H(y) = W \log 2\pi e(P_1 + N - \eta_1 + \Delta P) \quad (22.16)$$

por la aplicación del teorema de la potencia mínima de la entropía de una suma. Por lo tanto, ya que puede alcanzar la H indicada, la entropía de la distribución maximizada debe ser al menos tan grande y debe ser monotónica decreciente. Para demostrar eso $\eta \rightarrow 0$ como $P \rightarrow \infty$ se considera una señal que es de ruido blanco con una gran P .

Cualquiera que sea el ruido perturbador, la señal recibida será de aproximadamente un ruido blanco, si P es suficientemente grande, en el sentido de que tiene una potencia de entropía aproximada $P + N$.

Teorema 22.4. La capacidad del canal C para una banda W perturbada por el ruido térmico blanco de potencia N está limitada por

$$C \geq W \log \frac{2}{\pi e^3} \frac{S}{N}, \quad (22.17)$$

donde S es el pico permitido por el transmisor de potencia. Para $\frac{S}{N}$ suficientemente grande

$$C \leq W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon) \quad (22.18)$$

donde ϵ es arbitrariamente pequeño. Como $\frac{S}{N} \rightarrow 0$ (y siempre que la banda W inicia en 0)

$$\frac{C}{W \log \left(1 + \frac{S}{N} \right)} \rightarrow 1. \quad (22.19)$$

Queremos maximizar la entropía de la señal recibida. Si $\frac{S}{N}$ es grande, esto ocurrirá muy pronto cuando maximizamos la entropía de la familia transmitida.

La parte superior asintótica se obtiene mediante la relajación de las condiciones en la familia. Supongamos que el poder se limita a S no en cada instante de tiempo, pero sólo en los puntos de muestreo. La entropía máxima de la familia transmitida en estas condiciones debilitadas, es ciertamente mayor que o igual a la que en las condiciones originales. Este problema alterado se puede resolver fácilmente. La entropía máxima se produce si las diferentes muestras son independientes y tienen una función de distribución que es constante a partir de $-\sqrt{S}$ a $+\sqrt{S}$. La entropía se puede calcular como

$$W \log 4S. \quad (22.20)$$

La señal recibida tendrá entonces una entropía menor que

$$W \log(4S + 2\pi e N)(1 + \epsilon) \quad (22.21)$$

con $\epsilon \rightarrow 0$ y $\frac{S}{N} \rightarrow \infty$, la capacidad del canal se obtiene restando la entropía del ruido blanco, $W \log 2\pi e N$:

$$W \log(4S + 2\pi e N)(1 + \epsilon) - W \log(2\pi e N) = W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon). \quad (22.22)$$

Este es el límite superior deseado unido a la capacidad del canal.

Para obtener un límite inferior consideramos la misma familia de funciones. Permitir que estas funciones pasen a través de un filtro ideal con una característica de transferencia triangular. La ganancia es igual a la unidad en la frecuencia cero y disminuyendo linealmente hasta obtener cero en la frecuencia W . En primer lugar, demuestra que las funciones de salida de los filtros tienen una limitación de potencia pico S en todo momento (no sólo los puntos de muestreo). En primer lugar observamos que un pulso $\frac{\sin 2\pi W t}{2\pi W t}$ dentro del filtro produce

$$\frac{1}{2} \frac{\sin^2 \pi W t}{(\pi W t)^2} \quad (22.23)$$

en la salida. Esta función nunca es negativa. La función de entrada (en el caso general) se puede considerar como la suma de una serie de funciones desplazadas

$$a \frac{\sin 2\pi Wt}{2\pi Wt} \quad (22.24)$$

donde a , la amplitud de la muestra, no es mayor que \sqrt{S} . Por lo tanto, la salida es la suma de funciones desplazadas de la forma no negativa, anteriormente con los mismos coeficientes. Estas funciones son no negativas, el mayor valor positivo para cualquier t se obtiene cuando todos los coeficientes a tienen sus valores positivos máximos, es decir, \sqrt{S} . En este caso la función de entrada es una constante de amplitud \sqrt{S} y ya que el filtro tiene una unidad de ganancia para D.C., la salida es la misma. Por lo tanto el conjunto de salida tiene una potencia pico S .

La entropía de la familia de salida puede ser calculada a partir de la familia de entrada usando el teorema a tratar con dicha situación. La entropía de salida es igual a la entropía de entrada más la ganancia media geométrica del filtro:

$$\int_0^W \log G^2 df = \int_0^W \log \left(\frac{W-f}{W} \right)^2 df = -2W. \quad (22.25)$$

Por lo tanto la entropía de salida es

$$W \log 4S - 2W = W \log \frac{4S}{e^2} \quad (22.26)$$

y la capacidad del canal es mayor que

$$W \log \frac{2}{\pi e^3} \frac{S}{N}. \quad (22.27)$$

Ahora queremos demostrar que, para pequeños $\frac{S}{N}$ (potencia pico de la señal a través de la potencia media de ruido blanco), el canal de capacidad es aproximadamente

$$C = W \log \left(1 + \frac{S}{N} \right) \quad (22.28)$$

Más precisamente $\frac{C}{W \log \left(1 + \frac{S}{N} \right)} \rightarrow 1$ como $\frac{S}{N} \rightarrow 0$. Puesto que la señal de potencia media P

es menor o igual a el pico S , se deduce que para todos $\frac{S}{N}$

$$C \leq W \log \left(1 + \frac{P}{N} \right) \leq W \log \left(1 + \frac{S}{N} \right). \quad (22.29)$$

Por lo tanto, si podemos encontrar una familia de funciones tal que correspondan a la tasa cerca de $W \log \left(1 + \frac{S}{N} \right)$ y se limitan a la banda W y pico S el resultado será demostrado. Considérese la familia de funciones del siguiente tipo. Una serie de muestras t tienen el mismo valor, ya sea $+\sqrt{S}$ o $-\sqrt{S}$, entonces las siguientes muestras t tienen el mismo valor, etc. El valor de una serie se elige al azar, probabilidad $\frac{1}{2}$ para $+\sqrt{S}$ y $\frac{1}{2}$ para $-\sqrt{S}$. Si esta familia se pasa a través de un filtro con característica de ganancia triangular (unidad de ganancia en D.C.), la salida está limitada al pico $\pm S$. Además la potencia media es casi S y se puede hacer para acercarse a esto tomando t suficientemente grande. La entropía de la suma de este y el ruido térmico se encuentra aplicando el teorema de la suma de un ruido y una pequeña señal. Este teorema se aplicará si

$$\sqrt{t} \frac{S}{N} \quad (22.30)$$

es suficientemente pequeño. Esto puede garantizarse tomando $\frac{S}{N}$ suficientemente pequeño (después de que se elige t). La energía de la entropía será $S + N$ para acercarse a una aproximación como se desee, y por lo tanto la tasa de transmisión tan cerca como queremos

$$W \log \left(\frac{S + N}{N} \right). \quad (22.31)$$

Parte V

La tasa para una fuente continua

Capítulo 23

Las funciones de evaluación de fidelidad

En el caso de una fuente discreta de información nos fue posible determinar una definida tasa de generación de información, esta es la entropía del proceso estocástico subyacente. Con una continua fuente, la situación es más complicada. En primer lugar, una cantidad continuamente variable puede ser asumida como un número infinito de valores y por lo tanto requiere un número infinito de dígitos binarios para su especificación exacta. Esto significa que para transmitir la salida de una fuente continua con una *recuperación exacta* en el punto de recepción, requiere generalmente un canal de capacidad infinita (en bits por segundo). Debido a que, ordinariamente, los canales tienen una cierta cantidad de ruido, y por lo tanto una capacidad finita, la transmisión exacta es imposible.

Esto, aun así, evade el problema real. De forma práctica, nosotros no estamos interesados en transmisión exacta cuando tenemos una fuente continua, sino solamente en la transmisión dentro de una cierta tolerancia. La cuestión es si podemos asignar una tasa definida a una fuente continua cuando requerimos solamente una cierta fidelidad de recuperación, medida en una forma adecuada. Claro, a como los requerimientos de fidelidad sean incrementados la tasa se incrementará de igual manera. Será mostrado que podemos, en casos muy generales, definir tal tasa, teniendo la propiedad de que es posible, propiamente mediante la codificación de la información para transmitirla a otro canal cuya capacidad sea igual a la tasa en cuestión, y así satisfacer los requerimientos de fidelidad. Un canal de menor capacidad es insuficiente.

Primero es necesario dar la formulación matemática general de la idea de fidelidad de transmisión. Considera el conjunto de mensajes de larga duración, digamos T segundos. La fuente es descrita dando la densidad de probabilidad en el espacio asociado, así que la fuente seleccione el mensaje en cuestión $P(x)$. Un cierto sistema de comunicación es descrito (desde el punto de vista externo) dando la probabilidad condicional $P_x(y)$ así que si el mensaje x es producido por la fuente, el mensaje recuperado en el punto de recepción será y . El sistema como un todo (incluyendo la fuente y el sistema de transmisión) es descrito por la función de probabilidad $P(x, y)$, probabilidad de tener mensaje x y salida final y . Si esta función es conocida, las características completas del sistema desde el punto de vista de fidelidad son conocidas. Cualquier evaluación de fidelidad debe corresponder matemáticamente a una operación aplicada a $P(x, y)$. Esta operación debe tener por lo menos las propiedades de un simple ordenamiento de sistemas, por ejemplo, debe ser posible decir que dos sistemas representados por $P_1(x, y)$ Y $P_2(x, y)$ que, de acuerdo a nuestro criterio de fidelidad cumpla con ya sea (1) el primero tiene una fidelidad más alta, (2) el segundo tiene una fidelidad más

alta, o (3) cuentan con una fidelidad equivalente. Esto significa que el criterio de fidelidad puede ser representado mediante una función numéricamente valuada.

$$v(P(x, y)) \quad (23.1)$$

cuyos argumentos van más allá de las posibles funciones de probabilidad $P(x, y)$. Ahora mostraremos que bajo suposiciones muy generales y razonables, la función $v(P(x, y))$ puede ser escrita en una forma aparentemente mucho más especializada, esta siendo un promedio de una función $\rho(x, y)$ sobre el conjunto de valores posibles de x y y :

$$v(P(x, y)) = \int \int P(x, y) \rho(x, y) dx dy. \quad (23.2)$$

Para obtener esto necesitamos solamente asumir (1) que la fuente y el sistema son ergódicos así que una muestra muy larga será, probablemente cercana a 1, típicamente del conjunto, y (2) que la evaluación es razonable en el sentido que es posible, mediante la observación de una típica entrada y salida x_1 y y_1 , formar la evaluación tentativa en la base de esas muestras; y si estas muestras son incrementadas en duración la evaluación tentativa, con probabilidad 1, se acercará a la evaluación exacta basada en un total conocimiento de $P(x, y)$. Digamos que la evaluación tentativa es $\rho(x, y)$. Entonces la función $\rho(x, y)$ se acerca (como $T \rightarrow \infty$) a una constante para la mayoría (x, y) los cuales están en la región altamente probable correspondiente al sistema:

$$\rho(x, y) \rightarrow v(P(x, y)) \quad (23.3)$$

y también podemos escribir

$$\rho(x, y) \rightarrow \int \int P(x, y) \rho(x, y) dx dy \quad (23.4)$$

debido a que

$$\int \int P(x, y) dx dy = 1 \quad (23.5)$$

Esto establece el resultado deseado. La función $\rho(x, y)$ tiene la naturaleza general de una "distancia" entre x y y ⁹. Mide que tan indeseable es (de acuerdo a nuestro criterio de fidelidad) recibir y mientras x es transmitido. El resultado general dado anteriormente puede ser expresado como sigue: Cualquier evaluación razonable puede ser representada como un promedio de una función de distancia sobre el conjunto de mensajes y mensajes recuperados x y y ponderados de acuerdo a la probabilidad $P(x, y)$ de obtener el par en cuestión, siempre que la duración T de los mensajes sea suficientemente larga.

1

1. Criterio R.M.S.

$$v = (x(t) - y(t))^2 \quad (23.6)$$

En esta medida de fidelidad muy comúnmente usada, la función de distancia $\rho(x, y)$ es (aparte de un factor constante) el cuadrado de la distancia Euclidiana ordinaria entre los puntos x y y en la función espacio asociada.

$$\rho(x, y) = \frac{1}{T} \int_0^T [x(t) - y(t)]^2 dt \quad (23.7)$$

⁹No es "métrica" en el sentido estricto, ya que en general no satisface uno u otro ya sea: $\rho(x, y) = \rho(y, x)$ o: $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$.

2. Criterio R.M.S. con frecuencia ponderada. Más generalmente uno puede aplicar diferentes ponderaciones a los diferentes componentes de frecuencia antes de usar una medición de fidelidad R.M.S. Esto es el equivalente a pasar la diferencia $x(t) - y(t)$ a través de un filtro de conformación y entonces determinar la potencia promedio en la salida. Así, sea

$$e(t) = x(t) - y(t) \quad (23.8)$$

y

$$f(t) = \int_{-\infty}^{\infty} \epsilon(\tau) k(t - \tau) d\tau \quad (23.9)$$

entonces

$$\rho(x, y) = \frac{1}{T} \int_0^T f(t)^2 dt \quad (23.10)$$

3. Criterio del error absoluto

$$\rho(x, y) = \frac{1}{T} \int_0^T |x(t) - y(t)| dt. \quad (23.11)$$

4. La estructura de la oreja y el cerebro determina implícitamente una evaluación, o más bien un número de evaluaciones, apropiado en el caso de transmisión de música o habla. Hay, por ejemplo, un criterio de "inteligibilidad" en el cual $\rho(x, y)$ es equivalente a la frecuencia relativa de palabras incorrectamente interpretadas cuando el mensaje $x(t)$ es recibido como $y(t)$. Aunque no podemos dar una representación explícita de $\rho(x, y)$, en esos casos podría, en principio, ser determinada por suficiente experimentación. Algunas de sus propiedades hacen seguimiento a buenos experimentos conocidos sobre el oído, por ejemplo, la oreja es relativamente insensible a la fase y la sensibilidad de amplitud y frecuencia es aproximadamente logarítmica.
5. El caso discreto puede ser considerado como una especialización en la cual hemos asumido tácitamente una evaluación basada en la frecuencia de los errores. La función $\rho(x, y)$ es entonces definida como el número de símbolos en la secuencia y que difieren de símbolos correspondientes en x dividido por el total de número de símbolos en x .

Capítulo 24

La tasa para una fuente relativa a una evaluación de fidelidad

Estamos ahora en una posición de definir la tasa de generación de información para una fuente continua. Se nos da $P(x)$ para la fuente y una evaluación v determinada por una función de distancia $\rho(x, y)$ la cual se asumirá continua en ambos x y y . Con un sistema particular $P(x, y)$ la calidad es medida por

$$v = \int \int \rho(x, y) P(x, y) dx dy \quad (24.1)$$

Más aún, la tasa de flujo de dígitos binarios correspondientes a $P(x, y)$ es

$$R = \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy \quad (24.2)$$

Definimos que la tasa R_1 de generación de información para una calidad de reproducción dada v_1 sea el mínimo R cuando mantenemos v fija en v_1 y $P_x(y)$ variable. Esto es:

$$R_1 = \min_{P_x(y)} \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy \quad (24.3)$$

sujeto a la restricción:

$$v_1 = \int \int P(x, y) \rho(x, y) dx dy. \quad (24.4)$$

Esto significa que consideramos, en efecto, todos los sistemas de comunicación que pueden ser usados y que transmiten con la fidelidad requerida. La tasa de transmisión en bits por segundo es calculada para cada uno y escogemos el que tiene la menor tasa. Ésta última tasa es la tasa que asignamos a la fuente para la fidelidad en cuestión.

La justificación de esta definición está en el siguiente resultado:

Teorema 24.1. *Si una fuente tiene una tasa R_1 para una valuación v_1 , es posible codificar la salida de la fuente y transmitirla sobre un canal de capacidad C con fidelidad tan cercana a v_1 como se desee, siempre que $R_1 \leq C$. Esto no es posible si $R_1 > C$.*

El último enunciado del teorema sigue inmediatamente de la definición de R_1 y resultados previos. Si esto no fuera cierto podríamos transmitir más de C bits por segundos sobre un canal de capacidad C .

La primera parte del teorema es comprobada por un método análogo al que fue usado en el Teorema ???. Podemos, en primer lugar, dividir el espacio (x, y) en un gran número de pequeñas celdas y representar la situación en un caso discreto. Esto no va a cambiar la función de evaluación por más que una pequeña cantidad arbitraria (cuando las celdas son muy pequeñas) debido a la continuidad asumida para $\rho(x, y)$. Suponga que $P_1(x, y)$ es el sistema particular el cual minimiza la tasa y da R_1 . Escogemos desde las y 's de alta probabilidad, un conjunto al azar que contenga

$$2^{(R_1+E)T} \quad (24.5)$$

miembros donde $E \rightarrow 0$ como $T \rightarrow \infty$. Con una T grande, cada punto escogido será conectado por una línea de alta probabilidad (como en la figura ??) a un conjunto de x 's. Un cálculo similar al usado para comprobar el Teorema ??? muestra que con una T grande la mayoría de las x 's son cubiertas por los *fans* de los puntos y escogidos, para la mayoría de las elecciones de y 's. El sistema de comunicación a ser usado opera como sigue: Los puntos seleccionados son números binarios asignados. Cuando un mensaje x es originado se encontrara dentro de al menos uno de los *fans* (con probabilidad acercándose uno ya que $T \rightarrow \infty$). El número binario correspondiente es transmitido (o uno de ellos es escogido arbitrariamente si existen múltiples) sobre el canal por modos de codificación adecuados para dar una pequeña probabilidad de error. Ya que $R_1 \leq C$, esto es posible. En el punto de recepción la y correspondiente es reconstruida y usada como el mensaje de recuperación.

La evaluación v'_1 para este sistema se puede hacer arbitrariamente cercana a v_1 tomando una T suficientemente grande. Esto es debido a el hecho de que para cada muestra larga de un mensaje $x(t)$ y un mensaje de recuperación $y(t)$, la evaluación se acerca a v_1 (con probabilidad 1). Es interesante notar que, en este sistema, el ruido en el mensaje recuperado es en realidad producido por un tipo de cuantificación general en el transmisor y no producido por el ruido en el canal. Es más o menos análogo al ruido de cuantificación en PCM.

Capítulo 25

El cálculo de las tasas

La definición de la tasa es similar en muchos aspectos a la definición de capacidad de canal. En la primera:

$$R = \min_{P_x(y)} \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy \quad (25.1)$$

con $P(x)$ y $v_1 = \int \int P(x, y) \rho(x, y) dx dy$ fija. En la segunda:

$$C = \max_{P(x)} \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy \quad (25.2)$$

con $P_x(y)$ fija y posibilidad de uno o más restricciones (por ejemplo, una limitación de potencia promedio) de la forma $K = \int \int P(x, y) \lambda(x, y) dx dy$. Una solución parcial del problema de maximización general para determinar la tasa de una fuente se puede dar.

Usando el método de Lagrange, consideramos:

$$\int \int [P(x, y) \log \frac{P(x, y)}{P(x)P(y)} + \mu P(x, y) \rho(x, y) + v(x) P(x, y)] dx dy \quad (25.3)$$

La equación variacional (cuando tomamos la primera variación de $P(x, y)$) lleva a:

$$P_y(x) = B(x) \epsilon^{-\lambda \rho(x, y)} \quad (25.4)$$

donde λ es determinada para dar la fidelidad requerida y $B(x)$ es elegida para satisfacer:

$$\int B(x) \epsilon^{\lambda \rho(x, y)} dx = 1. \quad (25.5)$$

Esto muestra que, con la mejor codificación, la probabilidad condicional de una cierta causa de variación recibida y , $P_y(x)$ estará en decline exponencialmente con la función de distancia $\rho(x, y)$ entre el x y y en cuestión. En el caso especial donde la función de distancia $\rho(x, y)$ depende solo en la diferencia (vector) entre x y y ,

$$\rho(x, y) = \rho(x - y) \quad (25.6)$$

tenemos

$$\int B(x) \epsilon^{-\lambda \rho(x - y)} dx = 1 \quad (25.7)$$

Por lo tanto $B(x)$ es constante, digamos α , y

$$P_y(x) = \alpha e^{-\lambda \rho(x-y)}. \quad (25.8)$$

Desafortunadamente estas soluciones formales son difíciles de evaluar en casos particulares y parece ser de poco valor. De hecho, el cálculo actual de las tasas ha sido llevado a cabo en solo algunos casos muy simples. Si la función de distancia $p(x,y)$ es el cuadrado medio de la discrepancia entre x y y , y el mensaje conjunto es ruido blanco, la tasa puede ser determinada. En ese caso tenemos

$$R = \min[H(x) - H_y(x)] = H(x) - \max H_y(x) \quad (25.9)$$

con $N = \overline{(x-y)^2}$. Pero el $\max H_y(x)$ ocurre cuando $y - x$ es un ruido blanco, y es equivalente a $W_1 \log 2\pi\epsilon N$ donde W_1 es el ancho de banda del mensaje conjunto. Por lo tanto

$$R = W_1 \log 2\pi\epsilon Q - W_1 \log 2\pi\epsilon N \quad (25.10)$$

$$= W_1 \log \frac{Q}{N} \quad (25.11)$$

donde Q es la potencia promedio del mensaje. Esto comprueba lo siguiente:

Teorema 25.1. *La tasa para la medición de fidelidad de una fuente de ruido blanco de potencia Q y banda W_1 relativa a un R.M.S. es:*

$$R = W_1 \log \frac{Q}{N} \quad (25.12)$$

donde N es el cuadrado medio del error permitido entre el mensaje original y el recuperado.

Más generalmente, con cualquier fuente de mensaje podemos obtener desigualdades delimitando la tasa a un criterio de cuadrado medio del error.

Teorema 25.2. *La tasa para cualquier fuente de banda W_1 es delimitada por:*

$$W_1 \log Q_1/N \leq R \leq W_1 \log Q/N \quad (25.13)$$

donde Q es la potencia promedio de la fuente, Q_1 la energía de entropía y N el cuadrado medio del error permitido.

El límite inferior sigue el hecho de que la $\max H_y(x)$ para un $\overline{(x-y)^2} = N$ dado ocurre en el caso de ruido blanco. El límite superior resulta si colocamos puntos (usados en la comprobación del Teorema ??) no en la mejor forma sino al azar en una esfera de radio $\sqrt{(Q-N)}$.

Agradecimientos

El escritor está en deuda con sus colegas en los laboratorio, particularmente al Dr. H. W. Bode, Dr. J. R. Pierce, Dr. B. McMillan y al Dr. B. M. Oliver, por muchas sugerencias y criticismos útiles durante el curso de su trabajo. Crédito debe también ser otorgado al Profesor N. Wiener, cuya solución elegante al problema de filtración y predicción de conjuntos estacionarios ha influido considerablemente la forma de pensar del escritor en este campo de estudio.

Anexo E

Sea S_1 cualquier subconjunto medible del conjunto g , y S_2 el subconjunto del conjunto f el cual da S_1 bajo la operación T . Entonces

$$S_1 = TS_2. \quad (\text{E.1})$$

Sea H^λ el operador que desplaza todas las funciones en un conjunto con tiempo λ . Entonces

$$H^\lambda S_1 = H^\lambda TS_2 = TH^\lambda S_2. \quad (\text{E.2})$$

debido a que T es invariante y por lo tanto conmuta con H^λ . Por lo tanto, si $m[S]$ es la probabilidad de medición del conjunto S ,

$$\begin{aligned} m[H^\lambda S_1] &= m[TH^\lambda S_2] = m[H^\lambda S_2] \\ &= m[S_2] = m[S_1] \end{aligned} \quad (\text{E.3})$$

donde la segunda igualdad es por definición la medición del espacio g , el tercero ya que el conjunto f es estacionario, y el último nuevamente por definición de la medición de g .

Para probar que la propiedad ergódica es preservada bajo operaciones invariantes, sea S_1 un subconjunto del conjunto g , el cual es invariante bajo H^λ , y sea S_2 el conjunto de todas las funciones f que se transforman en S_1 . Entonces

$$H^\lambda S_1 = H^\lambda TS_2 = TH^\lambda S_2 = S_1 \quad (\text{E.4})$$

así que $H^\lambda S_2$ es incluida en S_2 para todas las λ . Ahora, debido a que

$$m[H^\lambda S_2] = m[S_1] \quad (\text{E.5})$$

esto implica

$$H^\lambda S_2 = S_2 \quad (\text{E.6})$$

para todo λ con $m[S_2] \neq 0, 1$. Esta contradicción muestra que S_1 no existe.

Anexo F

El límite superior, $\overline{N_3} \leq N_1 + N_2$, se debe al hecho que la máxima entropía posible para la potencia $N_1 + N_2$ ocurre cuando tenemos ruido blanco de esta potencia. En este caso la energía de entropía es $N_1 + N_2$.

Para obtener el límite inferior, suponga que tenemos dos distribuciones en n dimensiones $p(x_i)$ y $q(x_i)$ con energías de entropía $\overline{N_1}$ y $\overline{N_2}$. Que forma debería p y q tener para poder minimizar la energía de entropía $\overline{N_3}$ de su convolución $r(x_i)$:

$$r(x_i) = \int p(y_i)q(x_i - y_i) dy_i. \quad (\text{F.1})$$

La entropía H_3 de r es dada por:

$$H_3 = - \int r(x_i) \log r(x_i) dx_i. \quad (\text{F.2})$$

Deseamos minimizar esto sujeto a las restricciones:

$$H_1 = - \int p(x_i) \log p(x_i) dx_i \quad (\text{F.3})$$

$$H_2 = - \int q(x_i) \log q(x_i) dx_i. \quad (\text{F.4})$$

Consideramos entonces

$$U = - \int [r(x) \log r(x) + \lambda p(x) \log p(x) + \mu q(x) \log q(x)] dx \quad (\text{F.5})$$

$$\delta U = - \int [[1 + \log r(x)]\delta r(x) + \lambda[1 + \log p(x)]\delta p(x) + \mu[1 + \log q(x)]\delta q(x)] dx \quad (\text{F.6})$$

Si $p(x)$ es variado en un argumento particular $x_i = s_i$, la variación en $r(x)$ es

$$\delta r(x) = q(x_i - s_i) \quad (\text{F.7})$$

y

$$\delta U = - \int q(x_i - s_i) \log r(x_i) dx_i - \lambda \log p(s_i) = 0 \quad (\text{F.8})$$

y similarmente cuando q es variado. Entonces las condiciones para un mínimo son:

$$\int q(x_i - s_i) \log r(x_i) dx_i = -\lambda \log p(s_i) \quad (\text{F.9})$$

$$\int p(x_i - s_i) \log r(x_i) dx_i = -\mu \log q(s_i) \quad (\text{F.10})$$

Si multiplicamos el primero por $p(s_i)$ y el segundo por $q(s_i)$ e integramos con respecto a s_i , obtenemos:

$$H_3 = -\lambda H_1 \quad (\text{F.11})$$

$$H_3 = -\mu H_2 \quad (\text{F.12})$$

o resolviendo λ y μ , y reemplazando en las ecuaciones

$$H_1 \int q(x_i - s_i) \log r(x_i) dx_i = -H_3 \log p(s_i) \quad (\text{F.13})$$

$$H_2 \int p(x_i - s_i) \log r(x_i) dx_i = -H_3 \log q(s_i) \quad (\text{F.14})$$

Ahora supongamos que $p(x_i)$ y $q(x_i)$ son normales

$$p(x_i) = \frac{|A_{ij}^{\frac{n}{2}}|}{(2\pi)^{\frac{n}{2}}} \exp - \frac{1}{2} \Sigma(A_{ij} x_i x_j) \quad (\text{F.15})$$

$$q(x_i) = \frac{|B_{ij}^{\frac{n}{2}}|}{(2\pi)^{\frac{n}{2}}} \exp - \frac{1}{2} \Sigma(B_{ij} x_i x_j) \quad (\text{F.16})$$

Entonces $r(x_i)$ puede también ser normal con la forma cuadrática C_{ij} . Si las inversas de éstas formas son a_{ij} , b_{ij} y c_{ij} , entonces

$$c_{ij} = a_{ij} + b_{ij}. \quad (\text{F.17})$$

Deseamos mostrar que estas funciones satisfacen las condiciones de minimización sí y solo si $a_{ij} = Kb_{ij}$ y por lo tanto da el mínimo H_3 bajo las restricciones. Primero tenemos

$$\log r(x_i) = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2} \Sigma(C_{ij} x_i x_j) \quad (\text{F.18})$$

$$\int q(x_i - s_i) \log r(x_i) dx_i = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2} \Sigma(C_{ij} S_i S_j) - \frac{1}{2} \Sigma(C_{ij} B_{ij}) \quad (\text{F.19})$$

Esto debería ser equivalente a

$$\frac{H_3}{H_1} \left[\frac{n}{2} \log \frac{1}{2\pi} |A_{ij}| - \frac{1}{2} \Sigma(A_{ij} S_i S_j) \right] \quad (\text{F.20})$$

lo cual requiere $A_{ij} = H_1/H_3 C_{ij}$. En este caso $A_{ij} = H_1/H_2 B_{ij}$ y ambas ecuaciones se reducen a identidades.

Anexo G

Lo siguiente indicará un acercamiento más general y riguroso a las definiciones centrales de teoría de la comunicación. Consideremos un espacio de medición de probabilidad cuyos elementos están ordenados en pares (x, y) . Las variables x_i y y_i serán identificadas como de todos los puntos cuyos x pertenecen al sub conjunto S_1 las posibles señales transmitidas y recibidas en una larga duración T . Llamaremos al conjunto de todos los puntos cuyos x pertenecen a un sub conjunt S_1 de puntos x : la tira sobre S_1 , y similarmente al conjunto cuyas y pertenecen a S_2 , la tira sobre S_2 . Dividimos x y y en una colección de subconjuntos medibles no superpuestos X_i y Y_i , aproximado a la tasa de transmisión R por

$$R_1 = \frac{1}{T} \sum_i (P(X_i, Y_i) \log \frac{P(X_i, Y_i)}{P(x_i)P(y_i)}) \quad (G.1)$$

donde

- $P(x_i)$ es la probabilidad de medición de la tira sobre X_i
- $P(y_i)$ es la probabilidad de medición de la tira sobre Y_i
- $P(X_i, Y_i)$ es la probabilidad de medición dela interseccion de las tiras.

Una subdivisión adicional no puede disminuir R_1 nunca. Dejemos que X_1 sea dividido en $X_1 = X'_1 + X''_1$ y sea

$$\begin{aligned} P(Y_1) &= a & P(X_1) &= b + c \\ P(X'_1) &= b & P(X'_1, Y_1) &= d \\ P(X''_1) &= c & P(X''_1, Y_1) &= e \\ P(X_1, Y_1) &= d + e \end{aligned} \quad (G.2)$$

Entonces en la suma hemos reemplazado (para la intersección X_1, Y_1)

$$(d + e) \log \frac{(d + e)}{a(b + c)} \text{ por } d \log \frac{d}{ab} + e \log \frac{e}{ac}. \quad (G.3)$$

Es fácilmente mostrado que con la limitación que tenemos en b, c, d, e ,

$$\left[\frac{d + e}{b + c} \right]^{d+e} \leq \frac{d d e e}{b^d c^e} \quad (G.4)$$

y consecuentemente la suma es incrementada. Por lo tanto las varias formas posibles de subdivisión forman un conjunto dirigido, con R incrementándose monotónicamente con el refinamiento de la subdivisión. Podemos definir R sin ambigüedad como el menor límite superior para R_1 y escribirlo

$$R = \frac{1}{T} \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy \quad (G.5)$$

La integral, entendida en el sentido anterior, incluye ambos los casos continuos y discretos, y por supuesto, muchos otros que no pueden ser representados en cualquiera de las formas. Es trivial en esta formulación que si x y u están en correspondencia uno a uno, la tasa de u a y es equivalente a aquella entre x y y . Si v es cualquier función de y (no necesariamente con una inversa) entonces la tasa desde x a y es mayor o igual a aquella entre x a v debido a que, en el cálculo de las aproximaciones, las subdivisiones de y son esencialmente subdivisiones más finas que aquellas para v . Más generalmente si y y v están relacionadas, no funcionalmente pero estadísticamente, por ejemplo si tenemos un espacio de medida de probabilidad (y, v) , entonces $R(x, v) \leq R(x, y)$. Esto significa que cualquier operación aplicada a la señal recibida, aunque involucre elementos estadísticos, no incrementa R .

Otra noción que debe ser definida precisamente en una formulación abstracta de la teoría, es "la tasa de dimensión", que es el número promedio de dimensiones por segundo requeridas para especificar a un miembro de un conjunto. En el caso de una banda limitada con $2W$ números por segundo son suficientes. Una definición general puede ser enmarcada como sigue. Sea $f_\alpha(t)$ un conjunto de funciones y sea $\rho_\tau[f_\alpha(t), f_\beta(t)]$ una métrica midiendo la forma de "distancia" desde f_α hasta f_β sobre el tiempo T (por ejemplo la discrepancia R.M.S. sobre éste intervalo). Sea $N(\varepsilon, \delta, \tau)$ el menor número de elementos f que pueden ser elegidos, así que todos los elementos del conjunto además de un conjunto de medición δ , están dentro de distancia ε de por lo menos uno de los escogidos.

Por lo tanto, estamos cubriendo el espacio dentro de ε separado de un conjunto de poca medida δ . Definimos la tasa de dimensión λ para el conjunto, por el triple límite

$$\lambda = \lim_{\delta \rightarrow \infty} \lim_{\varepsilon \rightarrow \infty} \lim_{\tau \rightarrow \infty} \frac{\log N(\varepsilon, \delta, \tau)}{\tau \log \varepsilon} \quad (G.6)$$

Esta es una generalización de las definiciones de tipo de medida de la dimensión en la topología, y está de acuerdo con la tasa de dimensión intuitiva para conjuntos simples donde los resultados deseados son obvios.

Bibliografía

R.V.L. Hartley. Transmission of information. *Bell System Technical Journal*, page 535, July 1928.

H. Nyquist. Certain factors affecting telegraph speed. *Bell System Technical Journal*, page 324, April 1924.

H. Nyquist. Certain topics in telegraph transmission theory. *A.I.E.E. Trans.*, 47:617, April 1928.

FALTA CASI TODO. Technical Report 65, 1949.