Una teoría matemática de la comunicación

Claude E. Shannon

1948

Resumen

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 apxrendizaje *Teoría de la información y métodos de codificación*, impartida por Dra. Elisa Schaeffer en la primavera del 2013.

Hello world!

$$8 \leq 10 \approx \text{True}\, \pi \Lambda \tag{1}$$

0.1. Representacion grafica de procesos Markovianos

Los procesos estocasticos del tipo descrito arriba son matematicamente 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 numero finito de posibles .$^{\text{es}}$tados"de un sistema; $S_1, S_2, \ldots, S_n$. Además existen un conjunto de probabilidades de transicion; $p_i(j)$ 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 transicion desde un estado a otro. Los estados corresponderán al "residuo de influencia" de letras precedentes.

La situacion puede ser representada graficamente como se muestra en las figuras 3, 4 y 5. Los "estados" son los puntos de union en la grafica y las probabilidades y letras son producidas para una transicion son dadas ademas de la linea correspondiente. La figura 3 es para el ejemplo B en la seccion 2, mientras que la figura 4 corresponde al ejemplo C. En la figura 3 solamente hay un estado ya que letras sucesivas son independientes. En la figura 4 hay tantos estados como letras.

Si un ejemplo de un triagrama 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 es un grafo para el caso de estructura de palabras en el ejemplo D. Aqui S corresponde a el simbolo "espacio".

0.2. Fuentes ergódicas y mixtas

Como se ha indicado anteriormente, una fuente discreta para nuestros propositos puede ser considerada representada por un proceso Markoviano. Entre los posibles procesos discretos Markovianos existe un grupo con propiedades especiales con importancia en la teoria de la comunicacion. Esta clase especial consiste en los procesos "ergodicos" y deberiamos de llamar a las fuentes correspondientes, fuentes ergodicas. Aunque una definicion rigurosa de los procesos ergodicos es algo complicada, la idea general es simple. En un proceso ergodico cada secuencia producida por el proceso permanece igual en sus propiedades estadisticas. Por lo tanto las frecuencias de letras, las frecuencias de bigramas, etc., obtenidos de una secuencia en particulas, se acercaran a un limite definido conforme la longitud de las secuencua 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. Practicamente, la propiedad ergodica significa homogeneidad estadistica.

Todos los ejemplos de lenguaje artificial dados anteriormente son ergodicos. Esta propiedad está relacionada a la estructura de los grafos correspondientes. Si el grafo tiene las siguientes dos propiedades 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 union en la parte A a los puntos de union en la parte B a traves de las lineas del grafo en la direccion de las flechas y tambien es imposible ir desde las uniones en la parte B a las uniones en la parte A.
2.

$$b = a - 2 \text{ (ejemplo)} \tag{2}$$

[1] AQUI

Capítulo 1

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.

1.1. 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:
$$f_\theta(t) = \operatorname{sen}(t + \theta). \tag{1.1}$$
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,

$$f_\theta(t) = \operatorname{sen}(t+\theta), \tag{1.2}$$

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 $f_k(t)(k = 1, 2, \ldots, n)$ con la probabilidad de f_k siendo p_k.

2. Una familia de dimensión finita de funciones

   $$f(\alpha_1, \alpha_2, \ldots, \alpha_n; t) \tag{1.3}$$

   con una distribución de probabilidad sobre los parámetros α_i:

   $$p(\alpha_1, \ldots, \alpha_n). \tag{1.4}$$

   Por ejemplo podemos considerar la familia definida por

   $$f(a_1, \ldots, a_n, \theta_1, \ldots, \theta_n; t) = \sum_{i=1}^{n} a_i \operatorname{sen} i(\omega t + \theta_i) \tag{1.5}$$

   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

   $$f(a_i, t) = \sum_{n=-\infty}^{+\infty} a_n \frac{\operatorname{sen} \pi(2Wt - n)}{\pi(2Wt - n)} \tag{1.6}$$

   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

   $$\sum_{k=-\infty}^{\infty} f(t + t_k) \tag{1.7}$$

   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

$$f_\theta(t) = \operatorname{sen}(t + \theta) \tag{1.8}$$

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

[2] Esta 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).

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

$$f_\theta(t + t_1) = \operatorname{sen}(t + t_1 + \theta) \tag{1.9}$$

$$f_\theta(t + t_1) = \operatorname{sen}(t + \varphi) \tag{1.10}$$

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

$$\operatorname{sen}(t + \theta) \tag{1.11}$$

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

$$a \operatorname{sen}(t + \theta) \tag{1.12}$$

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 típico 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 $g_\alpha(t)$:

$$g_\alpha(t) = T f_\alpha(t). \tag{1.13}$$

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

$$g_\alpha(t) = T f_\alpha(t) \tag{1.14}$$

implica

$$g_\alpha(t + t_1) = T f_\alpha(t + t_1) \tag{1.15}$$

[3] Este 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.

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 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.

1.2. 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 1.2.1. *No dejar que $f(t)$ contenga frecuencias por encima de W. Entonces*

$$f(t) = \sum_{-\infty}^{\infty} X_n \frac{\operatorname{sen} \pi(2Wt - n)}{\pi(2Wt - n)} \tag{1.16}$$

donde

$$X_n = f\left(\frac{n}{2W}\right). \tag{1.17}$$

En esta expansión $f(t)$ es representada como una suma de funciones ortogonales. El coeficiente X_n de los diversos términos puede considerar 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 $p(x_1, \ldots x_n)$ en el correspondiente espacio n dimensional. Si la familia no está limitada en el tiempo se

[4] La 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.

[5] Para 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.

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 $p(x_1, \ldots, x_n)$ para dar la estructura estadística de la familia para intervalos de esa duración.

1.3. Entropía de una distribución continua

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

$$H = -\sum p_i \log p_i. \tag{1.18}$$

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

$$H = -\int_{-\infty}^{\infty} p(x) \log p(x)\, \mathrm{d}x. \tag{1.19}$$

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

$$H = -\int \cdots \int p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n. \tag{1.20}$$

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

$$H(x, y) = -\iint p(x, y) \log p(x, y)\, \mathrm{d}x\, \mathrm{d}y \tag{1.21}$$

y

$$H_x(y) = -\iint p(x, y) \log \frac{p(x, y)}{p(x)}\, \mathrm{d}x\, \mathrm{d}y \tag{1.22}$$

$$H_y(x) = -\iint p(x, y) \log \frac{p(x, y)}{p(y)}\, \mathrm{d}x\, \mathrm{d}y \tag{1.23}$$

donde

$$p(x) = \int p(x, y)\, \mathrm{d}y \tag{1.24}$$

$$p(y) = \int p(x, y)\, \mathrm{d}x. \tag{1.25}$$

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:

$$H\left(x, y\right) \leq H\left(x\right) + H\left(y\right) \tag{1.26}$$

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

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

$$p'\left(y\right) = \int a\left(x, y\right) p\left(x\right) \mathrm{d}x \tag{1.27}$$

con

$$\int a(x,y)\,\mathrm{d}x = \int a(x,y)\,\mathrm{d}y = 1,\ a(x,y) \geq 0 \tag{1.28}$$

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

4. Tenemos:

$$H(x,y) = H(x) + H_x(y) = H(y) + H_y(x) \tag{1.29}$$

y

$$H_x(y) \leq H(y) \tag{1.30}$$

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

$$H(x) = -\int p(x)\log p(x)\,\mathrm{d}x \tag{1.31}$$

con

$$\sigma^2 = \int p(x)\,x^2\,\mathrm{d}x \quad \text{y} \quad 1 = \int p(x)\,\mathrm{d}x \tag{1.32}$$

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

$$\int \left[-p(x)\log p(x) + \lambda p(x)\,x^2 + \mu p(x)\right]\mathrm{d}x \tag{1.33}$$

La condición para ésto es

$$-1 - \log p(x) + \lambda x^2 + \mu = 0 \tag{1.34}$$

y consecuentemente (ajustando las constantes para satisfacer los límites)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(x^2/2\sigma^2\right)} \tag{1.35}$$

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

$$A_{ij} = \int\cdots\int x_i x_j p(x_1,\ldots,x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \tag{1.36}$$

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

$$H(x) = \log\sqrt{2\pi e\sigma} \tag{1.37}$$

a

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(x^2/2\sigma^2\right)} \tag{1.38}$$

$$-\log p(x) = \log\sqrt{2\pi\sigma} + \frac{x^2}{2\sigma^2} \tag{1.39}$$

$$H(x) = -\int p(x) \log p(x) \, \mathrm{d}x \tag{1.40}$$

$$= \int p(x) \log \sqrt{2\pi\sigma} \, \mathrm{d}x + \int p(x) \frac{x^2}{2\sigma^2} \, \mathrm{d}x \tag{1.41}$$

$$= \log \sqrt{2\pi\sigma} + \frac{\sigma}{2\sigma^2} \tag{1.42}$$

$$= \log \sqrt{2\pi\sigma} + \log \sqrt{e} \tag{1.43}$$

$$= \log \sqrt{2\pi e\sigma} \tag{1.44}$$

a

$$p(x_1, \ldots, x_n) = \frac{|a_{ij}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum a_{ij} x_i x_j\right) \tag{1.45}$$

a

$$H = \log (2\pi e)^{n/2} |a_{ij}|^{-\frac{1}{2}} \tag{1.46}$$

a

$$a = \int_0^\infty p(x) \, x \, \mathrm{d}x \tag{1.47}$$

a

$$p(x) = \frac{1}{a} e^{-(x/a)} \tag{1.48}$$

a

$$H(y) = \int \cdots \int p(x_1, \ldots, x_n) \, J\left(\frac{x}{y}\right) \log p(x_1, \ldots, x_n) \, J\left(\frac{x}{y}\right) \mathrm{d}y_1 \cdots \mathrm{d}y_n \tag{1.49}$$

a

$$H(y) = H(x) - \int \cdots \int p(x_1, \ldots, x_n) \log J\left(\frac{x}{y}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \tag{1.50}$$

a

$$y_j = \sum_i a_{ij} x_i \tag{1.51}$$

a

$$H\left(y\right) = H\left(x\right) + \log\left|a_{ij}\right| \tag{1.52}$$

a

$$p\left(x_1, \ldots, x_n\right) \tag{1.53}$$

a

$$H' = -\lim_{n\to\infty} \frac{1}{n} \int \cdots \int p\left(x_1, \ldots, x_n\right) \log p\left(x_1, \ldots, x_n\right) \mathrm{d}x_1 \ldots \mathrm{d}x_n \tag{1.54}$$

a

$$H' = \log\sqrt{2\pi e N} \tag{1.55}$$

$$H = W \log 2\pi e N \tag{1.56}$$

a

$$\left|\frac{\log p}{n} - H'\right| < \varepsilon \tag{1.57}$$

a

$$\lim_{n\to\infty} \frac{\log V_n\left(q\right)}{n} = H' \tag{1.58}$$

a

$$p\left(x_1, \ldots, x_n\right) = \frac{1}{\left(2\pi N\right)^{n/2}} \exp -\frac{1}{2N} \sum x_i^2 \tag{1.59}$$

a

$$N_1 = \frac{1}{2\pi e} \exp 2H' \tag{1.60}$$

a

$$H_2 = H_1 + \frac{1}{W} \int_W \log |Y(f)|^2 \, df \tag{1.61}$$

a

$$J = \prod_{i=1}^{n} |Y(f_i)|^2 \tag{1.62}$$

a

$$\exp \frac{1}{W} \int_W \log |Y(f)|^2 \, df \tag{1.63}$$

a

¡Hola, mundo!

1.4. 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 $p(x_1, ..., x_n)$ y el segundo $q(x_1, ..., x_n)$. Después la función de densidad para la adición es dada por la convolución

$$r(x_1, ..., x_n) = \int \cdots \int p(y_1, ..., yn) q(x_1 - y_1, ..., x_n - y_n) \, \mathrm{d}y_1 ... \, \mathrm{d}y_n \tag{1.64}$$

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

El siguiente resultado es derivado en el Apéndice 6

Teorema 1.4.1. *Deje que la potencia media de los 2 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*

$$\bar{N}_1 + \bar{N}_2 \leq \bar{N}_3 \leq N_1 + N_2. \tag{1.65}$$

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 aproximádamente 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($x_1, ..., x_n$) es la función de distribución de densidad

$$a_{ij} = \int \cdots \int p(x_i - \alpha_i)(x_j - \alpha_j dx_1 \cdots dx_n) \tag{1.66}$$

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 bii 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

$$N + b_{ii}. \tag{1.67}$$

El potencial de entropía de esta distribución es

$$[\Pi(N + b_{ii})]^{\frac{1}{n}} \tag{1.68}$$

o aproximadamente

$$= [(N)^n + \Sigma b_{ii}(N)^n - 1]^{\frac{1}{n}} \tag{1.69}$$

$$= N + \frac{1}{n}\Sigma b_{ii}. \tag{1.70}$$

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

Capítulo 2

El canal continuo

2.1. 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 seÃśales 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

$$P(x_1, \cdots, x_n) = P(x) \tag{2.1}$$

y los del ruido por la distribución de probabilidad condicional

$$P_{x1,\cdots,xn}(y_1, \cdots, y_n) = P_x(y). \tag{2.2}$$

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

$$R = H(x) - H_y(x), \tag{2.3}$$

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 $P(x) = P(x_1, ..., x_n)$ y maximizar

$$-\int P(x) \log P(x) dx + \int\int P(x, y) \log \frac{P(x, y)}{P(y)} dx \, \mathrm{d}y. \tag{2.4}$$

Esto puede ser escrito

$$\int\int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx \, \mathrm{d}y \tag{2.5}$$

usando el hecho que $\int\int P(x, y) \log P(x) dx \, \mathrm{d}y = \int P(x) \log P(x) dx$. La capacidad del canal se expresa así:

$$C = \limsup_{T \to \infty P(x)} \frac{1}{T} \int\int P(x, y) \log \frac{P(x, y)}{P(x)\ P(y)} dx \, \mathrm{d}y. \tag{2.6}$$

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 estáq 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 es sólo la capacidad continua definida anteriormente.

En el lado matemático se pede 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

$$H(x) - H_y(x) \leq H(u) - H_v(u). \tag{2.7}$$

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 $n = (y - x)$,

$$P_x(y) = Q(y - x). \tag{2.8}$$

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 16: 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

$$R = H(y) - H(n), \tag{2.9}$$

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

$$C = \lim_{P(x)} \sup H(y) - H(n) \tag{2.10}$$

Tenemos, desde $y = x + n$

$$H(x,y) = H(x,n). \tag{2.11}$$

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

$$H(y) + H_y(x) = H(x) + H(n). \tag{2.12}$$

Por lo tanto

$$R = H(x) - H_y(x) = H(y) - H(n). \tag{2.13}$$

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.

25. 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

$$H(y) = W \log 2\pi e(P+N), \tag{2.14}$$

Y la entropía de ruido es

$$H(n) = W \log 2\pi eN. \tag{2.15}$$

La capacidad del canal es

$$C = H(y) - H(n) = W \log \frac{P+N}{N}. \tag{2.16}$$

Resumiendo tenemos lo siguiente:

Teorema 17: 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

$$C = W \log \frac{P+N}{N}. \tag{2.17}$$

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 $M = 2^s$

Meh meh

$$x^2 + y^2 = h^2$$

$$p\left(x_1, ..., x_n\right) \tag{2.18}$$

$$H' = -\lim_{n\to\infty} \frac{1}{n} \int \cdots \int p\left(x_1, ..., x_n\right) \log p\left(x_1, ..., x_n\right) \mathrm{d}x_1 ... \mathrm{d}x_n \tag{2.19}$$

$$H' = \log \sqrt{2\pi e N} H = W \log 2\pi e N \tag{2.20}$$

Capítulo 3

La tasa para una fuente continua

En el caso de una fuente continua 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 númericamente valuada.

$$v(P(x, y)) \tag{3.1}$$

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:

$$v(P(x,y)) = \int\int P(x,y)\rho(x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.2}$$

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 tipica entrada y salida x_1 y y_1, formar la evaluación tentativa en la base de esas muestra; 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 $T \to \infty$) a una constante para la mayoría (x,y) los cuales están en la región altamente probable correspondiente al sistema:

$$\rho(x,y) \to v(P(x,y)) \tag{3.3}$$

y también podemos escribir

$$\rho(x,y) \to \int\int P(x,y)\rho(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{3.4}$$

debido a que

$$\int\int P(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1 \tag{3.5}$$

Esto establece el resultado deseado. La función $\rho(x,y)$ tiene la naturaleza general de una "distancia.^{en}tre 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.

   $$v = (x(t) - y(t))^2 \tag{3.6}$$

   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.

   $$\rho(x,y) = \frac{1}{T}\int_0^T [x(t) - y(t)]^2\,\mathrm{d}t \tag{3.7}$$

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

   $$e(t) = x(t) - y(t) \tag{3.8}$$

   y

   $$f(t) = \int_{-\infty}^{\infty} \epsilon(\Theta)k(t-\Theta)\,\mathrm{d}\Theta \tag{3.9}$$

   entonces

   $$\rho(x,y) = \frac{1}{T}\int_0^T f(t)^2\,\mathrm{d}t \tag{3.10}$$

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

3. Criterio del error absoluto

$$\rho(x,y) = \frac{1}{T}\int_0^T |x(t) - y(t)|\,\mathrm{d}t. \tag{3.11}$$

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.

3.1. 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

$$v = \int\int \rho(x,y)P(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{3.12}$$

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

$$R = \int\int P(x,y)\log\frac{P(x,y)}{P(x)P(y)}\,\mathrm{d}x\,\mathrm{d}y \tag{3.13}$$

Definimos que la tasa R_1 de generación de información para una calidad de reproducción dada v1 sea el mínimo R cuando mantenemos v fija en v_1 y $P_x(y)$ variable. Esto es:

$$R_1 = \min_{P_x(y)} \int\int P(x,y)\log\frac{P(x,y)}{P(x)P(y)}\,\mathrm{d}x\,\mathrm{d}y \tag{3.14}$$

sujeto a la restricción:

$$v_1 = \int\int P(x,y)\rho(x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.15}$$

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 ultima 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 3.1.1. *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

$$2^{(R1+E)T} \tag{3.16}$$

miembros donde $E \to 0$ como $T \to \infty$. Con una T grande, cada punto escogido será conectado por una linea 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 $T \to \infty$). 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 pequeÃśa 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.

3.2. 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:

$$R = \min_{P_x(y)} \int\int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, \mathrm{d}x \, \mathrm{d}y \tag{3.17}$$

con $P(x)$ y $v_1 = \int\int P(x, y)\rho(x, y) \, \mathrm{d}x \, \mathrm{d}y$ fija. En la segunda:

$$C = \max_{P(x)} \int\int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, \mathrm{d}x \, \mathrm{d}y \tag{3.18}$$

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

Usando el método de Lagrange, consideramos:

$$\int\int [P(x, y) \log \frac{P(x, y)}{P(x)P(y)} + \mu P(x, y)\rho(x, y) + v(x)P(x, y)] \, \mathrm{d}x \, \mathrm{d}y \tag{3.19}$$

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

$$Py(x) = B(x)\epsilon^{-\lambda\rho(x,y)} \tag{3.20}$$

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

$$\int B(x)\epsilon^{\lambda\rho(x,y)} \, \mathrm{d}x = 1. \tag{3.21}$$

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 epecial donde la función de distancia $\rho(x, y)$ depende solo en la diferencia (vector) entre x y y,

$$\rho(x, y) = \rho(x - y) \tag{3.22}$$

tenemos

$$\int B(x)\epsilon^{-\lambda p(x-y)}\,\mathrm{d}x = 1 \tag{3.23}$$

Por lo tanto $B(x)$ es contante, digamos alpha, y

$$P_y(x) = \alpha\epsilon^{-\lambda\rho(x-y)}. \tag{3.24}$$

Desafortunadamente etas soluciones formales son difíciles de evaluar en casos particulares y parece ser de poco valor. De hecho, el calculo actual de las tasas ha sido llevado a cabo en solo alguno casos muy simples. Si la función de ditancia 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 cao tenemos

$$R = \text{mín}[H(x) - H_y(x)] = H(x) - \text{máx}\, H_y(x) \tag{3.25}$$

con $N = \overline{(x-y)^2}$. Pero el $\text{máx}\, H_y(x)$ 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

$$R = W_1 \log 2\pi\epsilon Q - W_1 \log 2\pi\epsilon N \tag{3.26}$$

$$= W_1 \log \frac{Q}{N} \tag{3.27}$$

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

Teorema 3.2.1. *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:*

$$R = W_1 \log \frac{Q}{N} \tag{3.28}$$

donde N es es el cuadradio 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 3.2.2. *La tasa para cualquier fuente de banda W_1 es delimitada por:*

$$W_1 \log Q1/N <= R <= W_1 \log Q/N \tag{3.29}$$

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

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

Reconocimientos

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.

Apéndice A

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

$$S_1 = TS_2. \tag{A.1}$$

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

$$H^\lambda S_1 = H^\lambda T S_2 = T H^\lambda S_2. \tag{A.2}$$

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,

$$\begin{array}{rcl} m[H^\lambda S_1] & = & m[TH^\lambda S_2] = m[H^\lambda S_2] \\ & = & m[S_2] = m[S_1] \end{array} \tag{A.3}$$

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 ergÃşdica 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

$$H^\lambda S_1 = H^\lambda T S_2 = T H^\lambda S_2 = S_1 \tag{A.4}$$

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

$$m[H^\lambda S_2] = m[S_1] \tag{A.5}$$

esto implica

$$H^\lambda S_2 = S_2 \tag{A.6}$$

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

Apéndice B

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 $r(x_i)$:

$$r(x_i) = \int p(y_i) q(x_i - y_i)\, \mathrm{d}y_i. \tag{B.1}$$

La entropía H_3 de r es dada por:

$$H_3 = -\int r(x_i) \log r(x_i)\, \mathrm{d}x_i. \tag{B.2}$$

Deseamos minimizar esto sujeto a las restricciones:

$$H_1 = -\int p(x_i) \log p(x_i)\, \mathrm{d}x_i \tag{B.3}$$

$$H_2 = -\int q(x_i) \log q(x_i)\, \mathrm{d}x_i. \tag{B.4}$$

Consideramos entonces

$$U = -\int [r(x) \log r(x) + \lambda p(x) \log p(x) + \mu q(x) \log q(x)]\, \mathrm{d}x \tag{B.5}$$

$$\delta U = -\int [[1 + \log r(x)]\delta r(x) + \lambda[1 + \log p(x)\delta p(x) + \mu[1 + \log q(x)]\delta q(x)]\, \mathrm{d}x \tag{B.6}$$

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

$$\delta r(x) = q(x_i - s_i) \tag{B.7}$$

y

$$\delta U = -\int q(x_i - si_i) \log r(x_i)\, \mathrm{d}x_i - \lambda \log p(s_i) = 0 \tag{B.8}$$

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

$$\int q(x_i - s_i) \log r(x_i)\, \mathrm{d}x_i = -\lambda \log p(s_i) \tag{B.9}$$

$$\int p(x_i - s_i) \log r(x_i) \, \mathrm{d}x_i = -\mu \log q(s_i) \tag{B.10}$$

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

$$H_3 = -\lambda H_1 \tag{B.11}$$

$$H_3 = -\mu H_2 \tag{B.12}$$

o resolviendo λ y μ, y reemplazando en las ecuaciones

$$H_1 \int q(x_i - s_i) \log r(x_i) \, \mathrm{d}x_i = -H_3 \log p(s_i) \tag{B.13}$$

$$H_2 \int p(x_i - s_i) \log r(x_i) \, \mathrm{d}x_i = -H_3 \log q(s_i) \tag{B.14}$$

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

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

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

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

$$c_{ij} = a_{ij} + b_{ij}. \tag{B.17}$$

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

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

$$\int q(x_i - s_i) \log r(x_i) \, \mathrm{d}x_i = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2}\Sigma(C_{ij}S_i S_j) - \frac{1}{2}\Sigma(C_{ij}B_{ij}) \tag{B.19}$$

Esto deberia ser equivalente a

$$\frac{H_3}{H_1}[\frac{n}{2} \log \frac{1}{2\pi} |A_{ij}| - \frac{1}{2}\Sigma(A_{ij}S_i S_j)] \tag{B.20}$$

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

Apéndice C

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, y serán identificadas como de todos los puntos cuyos x pertenecen al sub conjunto Si las posibles seÃśales transmitidas y recibidas en una larga duración T. Llamaremos al conjunto de todos los puntos cuyas 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

$$R_1 = \frac{1}{T}\sum_i (P(X_i, Y_i) \log \frac{P(X_i, Y_i)}{P(x_i)P(Y_i)} \tag{C.1}$$

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

$$\begin{array}{rclrcl} P(Y_1) & = & a & P(X_1) & = & b+c \\ P(X_1') & = & b & P(X_1', Y_1) & = & d \\ P(X_1'') & = & c & P(X_1'', Y_1) & = & e \\ P(X_1, Y_1) & = & d+e & & & \end{array} \tag{C.2}$$

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

$$(d+e)\log\frac{(d+e)}{a(b+c)} pord \log\frac{d}{ab} + e\log\frac{e}{ac}. \tag{C.3}$$

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

$$[\frac{d+e}{b+c}]^{d+e} \leq \frac{ddee}{b^d c^e} \tag{C.4}$$

y consecuentemente la suma es incrementada. Por lo tanto las varias formas posibles de subdivisión formanun conjunto dirigido, con R incrementandose monotónicamente con el refinamiento de la subdivisión.

Podemos definir R sin ambigÃijedad como el menor límite superior para R_1 y escribirlo

$$R = \frac{1}{T} \int \int P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \, \mathrm{d}x \, \mathrm{d}y \tag{C.5}$$

La integral, entendida en el sentido anterior, incluye ambos los casos continus 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 escencialmente 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 seÃśal 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 $\rho_\Theta \left[f_\alpha(t), f_\beta(t) \right]$ 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 $N(\varepsilon, \delta, \Theta)$ 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

$$\lambda = \lim_{\delta \to \infty} \lim_{\varepsilon \to \infty} \lim_{\Theta \to \infty} \frac{\log N(\varepsilon, \delta, \Theta)}{\Theta \log \varepsilon} \tag{C.6}$$

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.