Universidad de Guanajuato

División de Ciencias Naturales y Exactas

Teoría de la Información: Portafolios, Redes y Axiomatización de Entropías

por

Saúl Toscano Palmerín

Tesis para obtener el título de
Licenciado en Matemáticas

Director de Tesis
Dr. Víctor Manuel Pérez Abreu Carrión

Guanajuato, Guanajuato, México. Junio 19, 2013

A mi padre Miguel Ángel Toscano Medina, por quererme, apoyarme, interesarse en mis pasiones, y ser mi ejemplo.

A mi madre Marisol Palmerín Cerna, por quererme y mostrarme siempre el lado humano de las cosas.

Agradecimientos

Agradezco a mis padres que si no hubieran asumido su vocación con tanto amor y dedicación, yo no hubiera acabado mis estudios universitarios.

Quiero agradecer al Dr. Víctor Manuel Pérez Abreu Carrión por su tiempo dedicado a formarme como matemático y su enorme apoyo. Así como por sus invaluables consejos que no fueron únicamente sobre matemáticas.

Agradezco a mis sinodales Dr. Víctor Rivero Mercado, Dr. Joaquín Ortega Sánchez, y Dr. Leonel Ramón Pérez Hernández, por los comentarios y el tiempo dedicado a revisar mi tesis.

Agradezco a mi profesor Lawrence D. Nash por enseñarme la lengua inglesa.

Doy las gracias al CIMAT por permitirme estar en una institución que cuenta con matemáticos de excelente calidad. Asimismo por el esfuerzo que hizo la institución por otorgarme una beca durante mis estudios de licenciatura. También agradezco a la Universidad de Guanajuato por el apoyo y la educación que me proporcionó.

Finalmente, agradezco al Sistema Nacional de Investigadores por el apoyo como ayudante de investigador nacional (SNI-4337).

Introducción

La teoría de la información es una teoría matemática que surge en el contexto de sistemas de comunicación que transmiten información o datos de un punto a otro. Esta se origina en los trabajos del matemático e ingeniero electrónico estadounidense Claude Elwood Shannon en 1948, en los artículos pioneros [30, 31]. Una de sus características es el énfasis en la teoría de probabilidad, así como en la teoría ergódica.

En el contexto de la teoría de comunicaciones, la teoría de la información tiene dos objetivos principales. El primero es el desarrollo de los límites teóricos fundamentales en el rendimiento que se puede alcanzar cuando se comunica una fuente de información a través de los canales de comunicación usando códigos que han sido previamente desarrollados. El segundo objetivo es el desarrollo de códigos que den lugar a un buen rendimiento comparado con el rendimiento óptimo que indica la teoría.

A principio de los 40's se pensaba que era imposible enviar información con una probabilidad de error pequeña. Shannon impresionó a la comunidad que estudiaba la teoría de la comunicación de esa época cuando probó que la probabilidad de error podía ser casi cero para todas las tasas de comunicación que estuvieran por debajo de la capacidad del canal de transmisión, la cual puede ser calculada a partir de las características del ruido del mismo y de la dinámica del canal. Shannon también afirmó que algunos procesos aleatorios tenían una complejidad (lo que llamó entropía) irreducible de tal forma que debajo de dicha complejidad la señal no podía ser comprimida. Además afirmó que si la entropía de la fuente es menor que la capacidad del canal, se puede estar excento de error de manera asintótica, es decir que si el número de códigos y el número de letras que constituyen dichos códigos tiende a infinito, entonces la probabilidad de error tiende a cero.

Shannon usó dos diferentes nociones de medidas de información las cuales están relacionadas. La primera fue la entropía, un concepto que surgió de la termodinámica y que fue previamente propuesto como una medida de la información en una señal aleatoria (la distribución de los eventos tienen la misma probabilidad) por el ingeniero electrónico estadounidense Hartley [21] en 1928. Shannon definió la entropía de un proceso aleatorio discreto de variables aleatorias discretas finitas $X = \{X_n\}$, la cual se denota por $H(X)$ y expresa la idea que la entropía de un proceso es la cantidad de información en el proceso. Cabe mencionar que este concepto fue considerado por Shannon en el contexto de codificación, otro de los aspectos característicos de la teoría de la información. Específicamente, Shannon mostró que si se desea codificar el proceso dado en una sucesión de símbolos binarios, entonces para que el receptor que recibe la sucesión binaria pueda reconstruir el proceso original perfectamente (o casi), se necesitan al menos $H(X)$ símbolos binarios o bits y esto puede hacerse con un número de bits muy cercano a $H(X)$. Este resultado es conocido como el teorema de codificación para una fuente sin ruido.

La segunda noción de información considerada por Shannon fue la información mutua. La entropía es en realidad una noción de información de si misma (la información que da un proceso aleatorio acerca de si mismo). La información mutua

es una medida de información contenida en un proceso acerca de otro proceso. Mientras que la entropía es suficiente para estudiar la reproducción de un proceso único a través de un ambiente sin ruido, es más frecuente que uno tenga dos o más procesos aleatorios distintos, por ejemplo, uno que representa una fuente de información y otro la salida de un medio de comunicación en donde la fuente de codificación ha sido distorsionada por otro proceso aleatorio conocido como ruido. En tales casos las observaciones son hechas en uno de los procesos con el objetivo de tomar decisiones acerca del otro. Shannon introdujo la noción del promedio de información mutua entre dos procesos, definiéndolo como la suma de cada una de las entropías de los procesos menos su entropía conjunta. Este concepto ha sido relevante en teoremas de codificación que involucran más de un proceso aleatorio.

La teoría de la información se centra en el estudio de dos puntos extremos del conjunto de todos los posibles esquemas de comunicación. El mínimo de la compresión de datos está en un extremo de dicho conjunto. Todos los esquemas de compresión de datos requieren de una tasa de al menos igual que su mínimo. En el otro extremo está el máximo de datos enviados conocido como la capacidad del canal. Entonces, todos los esquemas de modulación y compresión de datos están entre esos dos límites. La teoría también considera formas de alcanzar estos límites. Sin embargo, dichos esquemas, que pueden ser teóricamente óptimos, pueden resultar computacionalmente imprácticos. Un ejemplo de estas ideas es el uso de códigos que son correctores de errores en los discos compactos y DVDs. En particular, en años recientes la teoría de información ha sido fundamental para optimizar las formas en que la información se transmite por medio de los teléfonos, celulares, el internet, y también en las estrategias de inversión y la neurología, por mencionar algunos campos de aplicación.

Respecto a las caracterizaciones de las entropías, Shannon dio una en su trabajo pionero. Posteriormente el matemático ruso Alexander Khinchin [22] en 1953 da una caracterización más general, que después fue generalizada por el matemático ruso Dmitrii Konstantinovich Faddeev [16] en 1956 quien debilita las suposiciones de Khinchin.

Actualmente la teoría de la información sigue siendo importante tanto desde el punto de vista matemático como el de sus aplicaciones en diversas áreas. Dentro de lo primero, además de lo descrito en los párrafos anteriores, se han propuesto varias axiomatizaciones de las medidas de información, algunas de las cuales se presentan en un artículo de revisión del tema de Imre Csiszár [11] en 2008. Asimismo, de importancia fundamental en los teoremas de codificación es el uso de las medidas de información. Con respecto a las aplicaciones, la teoría de la información sigue jugando un papel relevante en áreas de vanguardia como la economía [8], redes [8], neurología [28] , física [4], criptología [4], teoría de la complejidad [4], compresión de datos [4], los hoyos negros [7], y el genoma [26] por mencionar algunas.

Esta tesis de licenciatura tiene dos objetivos principales. Primero, presentar de manera rigurosa dos axiomatizaciones de las medidas de información, las cuales dan lugar a otras caracterizaciones de la entropía que se obtienen en base a las relaciones existentes entre las propiedades de las mismas. Segundo, exponer dos modelos donde la entropía surge de manera natural en el contexto de temas contemporáneos: uno

en economía (inversiones) y otro en redes. En el modelo en inversiones se presentan todas las ideas de manera detallada. Algunas demostraciones de la literatura se hicieron de manera distinta debido a que no eran claras las mismas. En el capítulo de redes se exponen las ideas principales por ser el tema muy vasto.

Una de las axiomatizaciones de medidas de información que se considera se basa en la caracterización de la función logaritmo con dominio en los naturales, la cual fue la propuesta en [22, 16]; ver también [1]. La otra es la axiomatización de Lee [25] que usa funciones de informaciones, cuya idea intuitiva es medir la cantidad de información como función de probabilidades de eventos.

El modelo en economía que se considera se utiliza en los mercados de acciones. Cuando se invierte repetidamente en un mercado de acciones estacionario se obtiene un crecimiento exponencial de la riqueza. La razón de crecimiento de la riqueza (llamada la razón de duplicación) es dual en cierta manera de la razón de la entropía del mercado.

La teoría de información en redes resulta en el estudio de varias fuentes de información en la presencia de ruido e interferencia. Dicha teoría también pudiera describirse como la teoría de tasas simultáneas de comunicación de muchos remitentes a muchos receptores en presencia de interferencia y ruido. Algunos de los intercambios entre receptores y remitentes son inesperados, y todos tienen cierta simplicidad matemática. Sin embargo todavía no existe una teoría unificadora. Al no tener dicha teoría, un área que ha sido de interés es el análisis de sistemas MIMO (múltiples-entradas y múltiples-salidas), que usan múltiples antenas en el remitente y receptor, donde se usan herramientas de la teoría de matrices aleatorias; para una referencia en español sobre el tema se sugiere consultar Díaz [14] y las referencias que ahí se mencionan.

La organización de la tesis es la siguiente. En el capítulo 1 se presentan los preliminares sobre la teoría de la información tales como propiedades y relaciones algebraicas básicas de la entropía de Shannon, entropía relativa e información mutua. El material presentado aquí es clásico, y para una mayor y detallada exposición se remite al lector a los libros [1, 8].

En los capítulos 2 y 3 se consideran caracterizaciones de la entropía de Shannon. En el capítulo 2 nos centramos en una caracterización que se basa en las propiedades que determinan a la función logaritmo cuando está definida en los naturales. En el capítulo 3 se introduce el concepto de función de información, y se desarrolla la teoría correspondiente que permite tener otra caracterización de la entropía. Con el objetivo de dar una presentación completa del tema en esta parte de la tesis se construye un número no contable de funciones de información para lo cual es necesario usar el axioma de elección. Dichas funciones son no Lebesgue medibles y la demostración se obtiene como una consecuencia de un teorema de las caracterizaciones de las entropías.

En el capítulo 4 se presenta un modelo del mercado de acciones, donde se muestran las relaciones que existen entre la teoría de la información y las inversiones. Se basa principalmente en una serie de artículos y un libro escritos por Cover y coautores entre 1988 y 2006 [2, 8, 9, 10], así como en trabajos más recientes en 2009 por Twichpongtorn [32], y en 2011 por Bean y Singer [6]. El teorema 4.10

de esta tesis se demostró de manera detallada dado que Cover sólo presenta ideas generales en su libro, además de que fue necesario incluir y probar el teorema 4.9 donde se usaron propiedades de semi-continuidad. En la última sección se introduce brevemente lo que es el portafolio universal.

En el capítulo 5 se desarrolla el tema de la teoría de la información en redes, que es el estudio de flujos de información alcanzables en presencia de ruido e interferencia. Se presentan los conceptos de interferencia, retroalimentación, capacidad de un canal y el canal gaussiano. Este capítulo está basado en el libro de Cover [8], y en una serie de artículos entre 1970 y 2010 [3, 15, 17, 20, 24, 27, 31, 33, 34].

El lector que sólo esté interesado en estudiar las caracterizaciones de la entropía de Shannon, pudiera leer solamente el capítulo 1 y pasar a los modelos de los capítulos 4 y 5. Asimismo, los capítulos 4 y 5 son independientes entre sí.

Índice general

Capítulo 1

Entropía de Shannon y entropía relativa

En este capítulo se introducen la mayoría de las definiciones básicas que se usarán en los siguientes capítulos. Para cualquier distribución de probabilidad se define una cantidad llamada entropía, la cual tiene muchas propiedades que coinciden con la noción intuitiva de lo que una medida de información debiera satisfacer. Esta noción se extiende para definir la información mutua, que es una medida de la cantidad de información que una variable aleatoria contiene acerca de la otra. La información mutua es un caso especial de una cantidad más general llamada entropía relativa, la cual es una medida de la distancia entre dos distribuciones de probabilidad. Todas esas cantidades están muy relacionadas y comparten varias propiedades. Después se establecen reglas de la cadena, y se demuestra que la información mutua es no negativa.

1.1. Entropía

Primero se introducirá el concepto de entropía, la cual es una medida de incertidumbre de una variable aleatoria.

Definición. Sean

$$\Delta_n := \left\{ (p_1, p_2, \ldots, p_n) : 0 < \sum_{k=1}^{n} p_k \leq 1, p_k \geq 0, k = 1, 2, \ldots, n \right\} (n = 1, 2, \ldots,) \tag{1.1}$$

los conjuntos de medidas finitas que suman menos que uno. La entropía de Shannon es la sucesión de funciones $H_n : \Delta_n \to \mathbb{R} (n = 1, 2, \ldots)$ definidas por

$$H_n(p_1, \ldots, p_n) = \sum_{k=1}^{n} L(p_k) / \sum_{k=1}^{n} p_k, \tag{1.2}$$

donde

$$L(x) = \begin{cases} -x \log_2 x & \text{si } x \in (0,1] \\ 0 & \text{si } x = 0. \end{cases} \tag{1.3}$$

En este trabajo se supondrá que los logaritmos son considerados en base dos a menos que se indique lo contrario. Por esto en adelante se quita el 2 de $\log_2$. Las unidades de la entropía son bits. También se supondrá que $0 \cdot log0 := 0$.

Notemos que hay dos posibles interpretaciones de la entropía:

- Medida de ruido que se produce antes de que el experimento finalice.
- Medida de la información esperada de un experimento.

A continuación definiremos los siguiente conjuntos

$$\Gamma_n := \left\{ (p_1, p_2, \ldots, p_n) : \sum_{k=1}^{n} p_k = 1, p_k \geq 0; k = 1, 2, \ldots, n \right\} (n = 2, 3, \ldots) \tag{1.4}$$

En los siguientes conjuntos se excluyen las probabilidades 0,

$$\Delta_n^o := \left\{ (p_1, p_2, \ldots, p_n) : 0 < \sum_{k=1}^{n} p_k \leq 1, p_k > 0, k = 1, 2, \ldots, n \right\} (n = 1, 2, \ldots,) \tag{1.5}$$

$$\Gamma_n^o := \left\{ (p_1, p_2, \ldots, p_n) : \sum_{k=1}^{n} p_k = 1, p_k > 0; k = 1, 2, \ldots, n \right\} (n = 2, 3, \ldots) \tag{1.6}$$

Es evidente que Γ_n^o es el interior del conjunto Γ_n.

En el siguiente teorema se presentan ocho de las propiedades más importantes de la entropía de Shannon.

Teorema 1.1. La entropía H_n satisface:

1) *Simetría.* La información es invariante bajo el cambio en el orden de los eventos.
$$H_n(p_1, \ldots, p_n) = H_n(p_{\sigma(1)}, \ldots, p_{\sigma(n)})$$
para $(p_1, \ldots, p_n) \in \triangle_n$.

2) *Normalidad.* $H_2(1/2, 1/2) = 1$.

3) *Expansibilidad.* Agregar resultados cuya probabilidad es cero no cambia la incertidumbre del experimento.
$$\begin{aligned} H_n(p_1, \ldots, p_n) &= H_{n+1}(0, p_1, \ldots, p_n) = H_{n+1}(p_1, \ldots, p_k, 0, p_{k+1}, \ldots, p_n) \\ &= H_{n+1}(p_1, \ldots, p_n, 0) \, (k = 1, 2, \ldots, n-1) \, . \end{aligned}$$

4) *Decisivo.* No hay incertidumbre en un experimento con dos resultados cuyas probabilidades son uno y cero, respectivamente.

$$H_2(1,0) = H_2(0,1) = 0.$$

5) *Aditividad fuerte.* Describe la información esperada de dos experimentos que no son independientes.

$$\begin{aligned} &H_{mn}(p_1q_{11}, p_1q_{12}, \ldots, p_1q_{1n}, p_2q_{21}, p_2q_{22}, \ldots, \\ &\quad p_2q_{2n}, \ldots, \ldots, p_mq_{m1}, p_mq_{m2}, \ldots, p_mq_{mn}) \\ &= H_m(p_1, \ldots, p_m) + \sum_{j=1}^{m} p_j H_n(q_{j1}, q_{j2}, \ldots, q_{jn}) \end{aligned}$$

para $(p_1, \ldots, p_m) \in \Gamma_m$ y $(q_{j1}, \ldots, q_{jn}) \in \Gamma_n$, $j = 1, 2, \ldots, m$. Esta propiedad tiene la siguiente interpretación: si $A_1, \ldots, A_m$ y $B_1, \ldots, B_n$ son los posibles resultados de dos experimentos, entonces q_{jk} es la probabilidad condicional de B_k dado que ocurre A_j, y el segundo miembro de la igualdad es la entropía condicional del segundo experimento dado que ocurre el primero. Entonces, la información esperada de los dos experimentos es la información esperada del primero más la entropía condicional del segundo experimento dado que ocurre el primero, es decir $\sum_{j=1}^{m} p_j H_n(q_{j1}, \ldots, q_{jn})$ se interpreta usando $p_j = \mathbb{P}(A_j)$ y $q_{jk} = \mathbb{P}(B_k \mid A_j)$.

6) *Aditividad.* La información esperada de dos experimentos independientes es la suma de la información esperada de los experimentos independientes.

$$\begin{aligned} H_{mn}(p_1q_1, p_1q_2, \ldots, p_1q_n, p_2q_1, p_2q_2, \ldots, p_2q_n, \ldots, \ldots, p_mq_1, p_mq_2, \ldots, p_mq_n) \; &= \\ H_m(p_1, \ldots, p_m) + H_n(q_1, \ldots, q_n) & \end{aligned}$$

para $(p_1, \ldots, p_m) \in \Delta_m$ y $(q_1, \ldots, q_n) \in \Delta_n$.

7) *Recursividad.*

$$H_n(p_1, \ldots, p_n) = H_{n-1}(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

para $(p_1, \ldots, p_n) \in \Gamma_n$, y $p_1 + p_2 > 0$.

8) *Positiva.* $H(P) \geq 0$.

Demostración. Notemos que (1)-(4) son triviales. Así pues, demostremos (5). Primero veamos que $L(xy) = xL(y) + yL(x)$ si $x, y \in [0, 1]$. Si $xy > 0$, $L(xy) = -xy \log(xy) = -xy \log x - xy \log y = xL(y) + yL(x)$. Supongamos, sin pérdida de

generalidad que $x = 0$, $L(0y) = 0 = 0L(y) + L(0)y$. Por tanto, $L(xy) = xL(y) + yL(x)$. Ahora notemos que $H_n(p_1, \ldots, p_n) = \sum_{k=1}^{n} L(p_k)$. Por ende

$$\begin{aligned} H_{mn}(p_1q_{11}, \ldots, p_mq_{mn}) &= \sum_{j=1}^{m}\sum_{k=1}^{n} L(p_jq_{jk}) \\ &= \sum_{j=1}^{m}\sum_{k=1}^{n} [p_jL(q_{jk}) + L(p_j)q_{jk}] \\ &= \sum_{j=1}^{m} L(p_j)\sum_{k=1}^{n} q_{jk} + \sum_{j=1}^{m} p_j \sum_{k=1}^{n} L(q_{jk}) \\ &= \sum_{j=1}^{m} L(p_j) + \sum_{j=1}^{m} p_j H_n(q_{j1}, \ldots, q_{jn}) \left(\text{ pues} \sum_{k=1}^{n} q_{jk} = 1 \right) \\ &= H_m(p_1, \ldots, p_m) + \sum_{j=1}^{m} p_j H_n(q_{j1}, \ldots, q_{jn}). \end{aligned}$$

Para (6), notemos que en el caso que $(p_1, \ldots, p_m) \in \Gamma_m, (q_1, \ldots, q_n) \in \Gamma_n$, poniendo $q_{1k} = q_{2k} = \cdots = q_{mk} = q_k \, (k = 1, \ldots, n)$ en (5) se obtiene el resultado buscado. Para el otro caso

$$\begin{aligned} \frac{\sum_{j=1}^{m}\sum_{k=1}^{n} L(p_jq_k)}{\sum_{j=1}^{m}\sum_{k=1}^{n} p_jq_n} &= \frac{\sum_{j=1}^{m} p_j \sum_{k=1}^{n} L(q_k) + \sum_{j=1}^{m} L(p_j)\sum_{k=1}^{n} q_k}{\sum_{j=1}^{m} p_j \sum_{k=1}^{n} q_k} \\ &\equiv \frac{\sum_{k=1}^{n} L(q_k)}{\sum_{k=1}^{n} q_k} + \frac{\sum_{j=1}^{m} L(p_j)}{\sum_{j=1}^{m} p_j}. \end{aligned}$$

Demostremos (7). Sea $p = p_1 + p_2$,$q = \frac{p_2}{p_1+p_2}$. Por ende,$1 - q = \frac{p_1}{p_1+p_2}$, $p_1 = p(1-q), p_2 = pq$. Por ende, debemos probar que

$$H_n(p_1, \ldots, p_n) = H_n(p(1-q), pq, p_3, \ldots, p_n) = H_{n-1}(p, p_3, \ldots, p_n) + pH_2(1-q, q)$$

pero la aditividad fuerte implica lo anterior.

∎

Teorema 1.2 La función L es continua en $[0, 1]$ y es diferenciable y cóncava en $(0, 1)$.

Demostración. Como $(xlogx)'' = \frac{1}{x \ln 2} > 0$ si $x > 0$ entonces L es cóncava. Es continua pues $\lim_{x \to 0^+} (xlogx) = 0$.

∎

Por tanto H_n es continua en Δ_n.

La siguiente desigualdad para funciones diferenciables y cóncavas, permitirá demostrar una desigualdad (llamada la propiedad de subaditividad) de la entropía dado que la función L es diferenciable y cóncava en $(0, 1)$.

Lema 1.1 Si g es diferenciable y cóncava en (a,b), entonces para todo $x_k \in (a,b)(k=1,\ldots,n)$ y $(q_1,\ldots,q_n) \in \Gamma_n (n=2,3,\ldots)$ se cumple

$$g\left(\sum_{k=1}^{n} q_k x_k\right) \geq \sum_{k=1}^{n} q_k g(x_k) \tag{1.7}$$

Si de hecho, g está definida en $[a,b),(a,b]$, y si $g(a) \leq \lim_{x\to a^+} g(x)$ y/o $g(b) \leq \lim_{x\to b^-} g(x)$, entonces la igualdad (1.7) vale si alguna de las $x_k's$ son a o b, respectivamente.

Demostración. Sea $\bar{x} = \sum_{k=1}^{n} q_k x_k$. Entonces $\bar{x} \in (a,b)$, y por el teorema de Taylor

$$g(x_i) = g(\bar{x}) + g'(\bar{x})(x_k - \bar{x}) + \frac{1}{2} g''(\epsilon_k)(x_k - \bar{x})^2 (k = 1,2,\ldots,n)$$

con ϵ_k entre x_k y $\bar{x}$. Multiplicando las ecuaciones anteriores por q_k y sumando, se obtiene que (usando que $g''(x) \leq 0$ con $x \in (a,b)$ y $\sum_{k=1}^{n} q_k = 1$)

$$\begin{aligned} \sum_{k=1}^{n} q_k g(x_k) &= g(\bar{x}) + \frac{1}{2}\sum_{k=1}^{n} g''(\epsilon_k) q_k (x_k - \bar{x})^2 \\ &\leq g(\bar{x}) = g\left(\sum_{k=1}^{n} q_k x_k\right). \end{aligned}$$

∎

Teorema 1.3 Si $(p_1,\ldots,p_m) \in \Gamma_m$ y $(q_{j1},\ldots,q_{jn}) \in \Gamma_n$, entonces

$$\sum_{j=1}^{m} p_j H_n(q_{j1},\ldots,q_{jn}) \leq H_n\left(\sum_{j=1}^{m} p_j q_{j1},\ldots,\sum_{j=1}^{m} p_j q_{jn}\right)$$

Demostración. Por el Lema 1.1

$$L\left(\sum_{j=1}^{m} p_j q_{jk}\right) \geq \sum_{j=1}^{m} p_j L(q_{jk}) (k = 1,2,\ldots,n).$$

Por ende

$$\sum_{k=1}^{n} L\left(\sum_{j=1}^{n} p_j q_{jk}\right) \geq \sum_{j=1}^{m} p_j \sum_{k=1}^{n} L(q_{jk}).$$

∎

Teorema 1.4. La entropía H_n satisface la subaditividad. Es decir

$$\begin{aligned} H_{mn}(p_{11},p_{12},\ldots,p_{1n},p_{21},p_{22},\ldots,p_{2n},\ldots,p_{m1},p_{m2},\ldots,p_{mn}) &\leq \\ H_m\left(\sum_{k=1}^{n}p_{1k},\ldots,\sum_{k=1}^{n}p_{mk}\right)+H_n\left(\sum_{j=1}^{m}p_{j1},\ldots,\sum_{j=1}^{n}p_{jn}\right)& \end{aligned}$$

si $[(p_{11},\ldots,p_{mn})\in\Gamma_{mn};m,n=2,3,\ldots]$. Notemos que si

$$p_{j,k}=\mathbb{P}\left(A_j\bigcap B_k\right)=\mathbb{P}\left(A_j\right)\mathbb{P}\left(B_k\mid A_j\right)=r_jq_{jk}$$

entonces

$$\begin{aligned} H_{mn}\left(\mathbb{P}(A_1\bigcap B_1),\ldots,\mathbb{P}(A_1\bigcap B_n),\ldots,\mathbb{P}(A_m\bigcap B_1),\mathbb{P}(A_m\bigcap B_2),\ldots,\mathbb{P}(A_m\bigcap B_n)\right) &\leq \\ H_m\left(\mathbb{P}(A_1),\mathbb{P}(A_2),\ldots,\mathbb{P}(A_m)\right)+H_n\left(\mathbb{P}(B_1),\mathbb{P}(B_2),\ldots,\mathbb{P}(B_n)\right)& \end{aligned}$$

con $(\mathbb{P}(A_1),\ldots,\mathbb{P}(A_m))\in\Gamma_m$ y $(\mathbb{P}(B_1),\ldots,\mathbb{P}(B_n))\in\Gamma_n$. Es decir, la información esperada de dos experimentos no es mayor que la suma de las informaciones esperadas de cada uno de los experimentos.

Demostración. Usando el Teorema 1.3 y aditividad fuerte

$$\begin{aligned} H_{mn}\left(p_{11},p_{12},\ldots,p_{1n},p_{21},p_{22},\ldots,p_{2n},\ldots,p_{m1},p_{m2},\ldots.p_{mn}\right) &= \\ H_{mn}\left(\left(\sum_{k=1}^{n}p_{1k}\right)\frac{p_{11}}{\sum_{k=1}^{n}p_{1k}},\ldots,\left(\sum_{k=1}^{n}p_{mk}\right)\frac{p_{mn}}{\sum_{k=1}^{n}p_{mk}}\right) &= \\ H_m\left(\sum_{k=1}^{n}p_{1k},\ldots,\sum_{k=1}^{n}p_{mk}\right)+\sum_{j=1}^{m}\left(\sum_{k=1}^{n}p_{jk}\right)H_n\left(\frac{p_{j1}}{\sum_{k=1}^{n}p_{jk}},\ldots,\frac{p_{jn}}{\sum_{k=1}^{n}p_{jn}}\right) &\leq \\ H_m\left(\sum_{k=1}^{n}p_{1k},\ldots,\sum_{k=1}^{n}p_{mk}\right)+H_n\left(\sum_{j=1}^{m}p_{j1},\ldots,\sum_{j=1}^{m}p_{jn}\right)\quad. & \end{aligned}$$

∎

A continuación se demostrarán otras seis propiedades básicas de la entropía de Shannon.

Teorema 1.5. La entropía H_n satisface que:

1) Es maximal, es decir

$$H_n\left(p_1,\ldots,p_n\right)\leq H_n\left(\frac{1}{n},\ldots,\frac{1}{n}\right)$$

para todo n y $(p_1,\ldots,p_n)\in\Gamma_n$.

2) Es acotada por arriba, es decir existe k tal que

$$H_2\left(1-q,q\right) \leq k$$

para todo $(1-q,q) \in \Gamma_2$.

3) Es monótona, es decir la función $q \to H_2(1-q,q)$ es no decreciente en $\left[0,\frac{1}{2}\right]$.

4) Es medible, es decir $q \to H_2(1-q,q)$ es Lebesgue-medible en $[0,1]$.

5) Es estable en p_o, es decir

$$\lim_{q\to 0^+} H_2(p_o,q) = H_1(p_o)$$

para $p_o \in [0,1]$.

6) Es pequeña para probabilidades pequeñas, es decir

$$\lim_{q\to 0^+} H_2(1-q,q) = 0$$

por tanto si un resultado de un experimento es muy probable, entonces el otro va a ser muy improbable y por tanto la incertidumbre es pequeña.

Demostración. Notemos que las afirmaciones (4)-(6) son triviales. Demostremos (1)

$$H_n\left(p,\ldots,p\right) = \frac{-\sum_{k=1}^{n} plog(p)}{np} = -\log p$$

por ende $H_n\left(\frac{1}{n},\ldots,\frac{1}{n}\right) = -\log\left(\frac{1}{n}\right) = \log(n)$. Usando que L es cóncava y el Lema 1

$$\begin{aligned} H_n\left(p_1,\ldots,p_n\right) &= \frac{\sum_{k=1}^{n} L(p_k)}{\sum_{k=1}^{n} p_k} \\ &= \left(\frac{\sum_{k=1}^{n} L(p_k)}{n}\right)\frac{n}{\sum_{k=1}^{n} p_k} \\ &\leq n\frac{L\left(\frac{\sum_{k=1}^{n} p_k}{n}\right)}{\sum_{k=1}^{n} p_k} \\ &= H_n\left(\frac{\sum_{k=1}^{n} p_k}{n},\ldots,\frac{\sum_{k=1}^{n} p_k}{n}\right). \end{aligned}$$

Esto indica que la entropía, vista como la medida de incertidumbre, alcanza su máximo cuando todos los resultados tienen la misma probabilidad.

Ahora demostremos (2). Como $H_n\left(p_1,\ldots,p_n\right) \leq \log(n)$ para $(p_1,\ldots,p_n) \in \Gamma_n$, entonces $H_2\left(1-q,q\right) \leq 1$.

Para demostrar (3) veamos que

$$\begin{aligned} f(q) &:= H_2(1-q,q) \\ &= -(1-q)\log(1-q) - q\log(q) \end{aligned}$$

si $q \in (0,1)$. Entonces

$$\begin{aligned} f'(q) &= \log(1-q) + (1-q)\frac{\log e}{1-q} - \log q - \frac{q \log e}{q} \\ &= \log(1-q) - \log q = \log\left(\frac{1-q}{q}\right) \end{aligned}$$

por lo que

$$\begin{aligned} & f'(q) \geq 0 \\ \iff & \tfrac{1-q}{q} \geq 1 \\ \iff & \tfrac{1}{2} \geq q \end{aligned}$$

por lo tanto f es no decreciente en $(0, \frac{1}{2}]$, y $f(0) = 0$, de aquí que f es no decreciente en $[0, \frac{1}{2}]$.

∎

1.2. Entropía conjunta y condicional

Ahora se definirán la entropía conjunta y la entropía condicional.

Definición. La entropía conjunta $H(X,Y)$ de un par de variables aleatorias discretas (X,Y) con distribución conjunta $p(x,y)$ está definida como

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y).$$

Notar que $H(X,Y) = -\mathbb{E} \log p(X,Y)$.

Definición. Si $(X,Y) \sim p(x,y)$, la entropía condicional $H(Y \mid X)$ está definida como

$$\begin{aligned} H(Y \mid X) &= \sum_x p(x) H(Y \mid X = x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) \\ &= -\sum_x \sum_y p(x,y) \log p(y \mid x) \\ &= -\mathbb{E} \log p(Y \mid X). \end{aligned}$$

Las definiciones anteriores resultan naturales cuando se demuestra que la entropía de un par de variables aleatorias es la suma de la entropía de una variable más la entropía condicional de la otra. Esto es probado en el siguiente teorema.

Teorema 1.6 (La regla de la cadena). Si $(X, Y) \sim p(x, y)$, entonces

$$H(X, Y) = H(X) + H(Y \mid X).$$

Demostración. Veamos que

$$\begin{aligned} H(X,Y) &= -\sum_x \sum_y p(x,y) \log p(x,y) \\ &= -\sum_x \sum_y p(x,y) \log p(x) p(y \mid x) \\ &= -\sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y \mid x) \\ &= H(X) + H(Y \mid X), \end{aligned}$$

donde $H(X)$ es la entropía de Shannon. Equivalentemente,

$$\mathbb{E} \log p(X, Y) = \mathbb{E} \log p(X) + \mathbb{E} \log p\left(Y \mid X\right).$$

∎

Corolario 1.1. Si $(X, Y, Z) \sim p(x, y, z)$, entonces

$$H(X, Y \mid Z) = H(X \mid Z) + H\left(Y \mid X, Z\right).$$

La prueba del corolario es un argumento similar al que se da en el Teorema 1.6.

Veamos un ejemplo donde se cumpla que $H\left(Y \mid X\right) \neq H(X \mid Y)$. Consideremos la siguiente distribución conjunta de (X, Y):

$Y \diagdown X$	1	2	3	4
1	1/8	1/16	1/32	1/32
2	1/16	1/8	1/32	1/32
3	1/16	1/16	1/16	1/16
4	1/4	0	0	0

La distribución marginal de X es $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$ y la distribución marginal de Y es $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ y por tanto $H(X) = \frac{7}{4}$ y $H(Y) = 2$. También $H(X \mid Y) = \frac{11}{8}$ y $H\left(Y \mid X\right) = \frac{13}{8}$ y $H\left(X, Y\right) = \frac{27}{8}$.

Nota. Aunque puede pasar que $H\left(Y \mid X\right) \neq H(X \mid Y)$, siempre se tiene que $H(X) - H\left(X \mid Y\right) = H(Y) - H\left(Y \mid X\right)$.

1.3. Entropía relativa e información mutua

Ahora veamos las definiciones de entropía relativa e información mutua. La entropía relativa $D\left(p \parallel q\right)$ es una medida de la ineficiencia asumiendo que la distribución es q cuando la distribución verdadera es p.

Definición. La entropía relativa (o distancia de Kullback-Leibler) entre dos funciones de probabilidad discretas $p(x)$ y $q(x)$ está definida como

$$\begin{aligned} D(p \parallel q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \mathbb{E}_p \log \frac{p(x)}{q(x)}. \end{aligned}$$

Usaremos la convención de que $p \log \frac{p}{0} = \infty$ si $p > 0$. Entonces si existe x tal que $p(x) > 0$ y $q(x) = 0$, entonces $D(p \parallel q) = \infty$.

En lo que sigue se demostrará que la entropía relativa es siempre no negativa y que es cero si y sólo si $p = q$. Sin embargo, no es una distancia entre distribuciones pues no es simétrica y no satisface la desigualdad del triángulo.

Ahora definiremos la información mutua, que es una medida de la cantidad de información que hay en una variable aleatoria de otra variable aleatoria.

Definición. Consideremos dos variables aleatorias X y Y con probabilidad conjunta $p(x, y)$ y probabilidades marginales $p(x)$ y $p(y)$. La información mutua $I(X; Y)$ es la entropía relativa entre la distribución conjunta y la distribución producto $p(x)p(y)$:

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y) \parallel p(x)p(y)) \\ &= \mathbb{E}_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}. \end{aligned}$$

Ejemplo. Sea $\mathcal{X} = \{0, 1\}$ un alfabeto y consideremos dos distribuciones p y q en $\mathcal{X}$. Sea $p(0) = 1 - r, p(1) = r$, y sea $q(0) = 1 - s, q(1) = s$. Entonces

$$D(p \parallel q) = (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s}$$

y

$$D(q \parallel p) = (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r}.$$

Si $r = s$, entonces $D(p \parallel q) = D(q \parallel p) = 0$. Si $r = 1/2$,$s = 1/4$, se obtiene que $D(p \parallel q) = 0{,}2075$ y $D(q \parallel p) = 0{,}1887$.

Notemos que en general $D(p \parallel q) \neq D(q \parallel p)$.

Ahora daremos la relación existente entre la entropía y la información mutua. Notemos que

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p\left(x \mid y\right)}{p(x)} \\ &= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p\left(x \mid y\right) \\ &= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p\left(x \mid y\right) \right) \\ &= H(X) - H(X \mid Y). \end{aligned}$$

Entonces la información mutua $I(X;Y)$ es la reducción de la incertidumbre de X debido al conocimiento que se tiene de Y.

Por simetría, se sigue que

$$I(X;Y) = H(Y) - H(Y \mid X).$$

Entonces, X dice tanto de Y como Y dice de X.

Usando que $H(X,Y) = H(X) + H\left(Y \mid X\right)$, se concluye que

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

Finalmente, notemos que

$$I(X;X) = H(X) - H(X \mid X) = H(X).$$

Entonces, la información mutua de una variable aleatoria consigo misma es la entropía de la variable aleatoria.

Juntando los resultados anteriores, se tiene el siguiente teorema.

Teorema 1.7. (Información Mutua y Entropía) Sean X, Y variables aleatorias discretas, entonces

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ I(X;Y) &= H(Y) - H\left(Y \mid X\right) \\ I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ I(X;Y) &= I(Y;X) \\ I(X;X) &= H(X). \end{aligned}$$

1.4. Reglas de la cadena

Ahora demostraremos que la entropía de una colección de variables aleatorias es la suma de las entropías condicionales.

Teorema 1.8. (La regla de la cadena para la entropía) Sean $X_1, \dots X_n$ variables aleatorias con densidad conjunta $p(x_1, \dots, x_n)$. Entonces

$$H\left(X_1, \dots, X_n\right) = H(X_1) + \sum_{i=2}^{n} H\left(X_i \mid X_{i-1}, \dots, X_1\right).$$

Demostración. Aplicando repetidamente la regla de expansión de dos variables para entropías, tenemos que

$$\begin{aligned}
H(X_1, X_2) &= H(X_1) + H\left(X_2 \mid X_1\right), \\
H(X_1, \dots, X_3) &= H(X_1) + H(X_2, X_3 \mid X_1) \\
&= H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_2, X_1), \\
&\vdots \\
H(X_1, X_2, \dots, X_n) &= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_{n-1}, \dots, X_1) \\
&= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \dots, X_1).
\end{aligned}$$

∎

A continuación definiremos la información mutua condicional como la reducción en la incertidumbre de X debida al conocimiento de Y dado que se conoce Z.

Definición. La información mutua condicional de las variables aleatorias X y Y dada Z está definida por

$$\begin{aligned}
I(X; Y \mid Z) &= H(X \mid Z) - H(X \mid Y, Z) \\
&= \mathbb{E}_{p(x,y,z)} \log \frac{p(X, Y \mid Z)}{p(X \mid Z) p\left(Y \mid Z\right)}.
\end{aligned}$$

La información mutua también satisface la regla de la cadena.

Teorema 1.9 (La regla de la cadena para la información) Sean $X_1, \dots, X_n, Y$ variables aleatorias discretas, entonces

$$I(X_1, X_2, \dots, X_n; Y) = I(X_1; Y) + \sum_{i=2}^{n} I\left(X_i; Y \mid X_{i-1}, \dots, X_1\right).$$

Demostración.

$$\begin{aligned}
I(X_1, \dots, X_n; Y) &= H(X_1, \dots, X_n) - H(X_1, \dots, X_n \mid Y) \\
&= H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_{i-1}, \dots, X_1) - H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_{i-1}, \dots, X_1, Y) \\
&= \sum_{i=1}^{n} I(X_i; Y \mid X_1, \dots, X_{i-1}).
\end{aligned}$$

Lo cual concluye la demostración del teorema.

∎

Definiremos una versión condicional para la entropía relativa.

Definición. Para densidades conjuntas de probabilidad $p(x,y)$ y $q(x,y)$, la entropía relativa condicional $D(p(y \mid x) \parallel q(y \mid x))$ es el promedio de las entropías relativas entre las densidades condicionales de probabilidad $p(y \mid x)$ y $q(y \mid x)$ promediada sobre la densidad de probabilidad $p(x)$. Con mayor precisión

$$\begin{aligned} D\left(p\left(y \mid x\right) \parallel q\left(y \mid x\right)\right) &= \sum_{x} p(x) \sum_{y} p\left(y \mid x\right) \log \frac{p(y \mid x)}{q\left(y \mid x\right)} \\ &= \mathbb{E}_{p(x,y)} \log \frac{p(Y \mid X)}{q(Y \mid X)}. \end{aligned}$$

La entropía relativa entre dos distribuciones conjuntas de un par de variables aleatorias puede ser expandida como la suma de una entropía relativa y una entropía relativa condicional.

Teorema 1.10 (La regla de la cadena para la entropía) Sean $p(x,y)$ y $q(x,y)$ dos probabilidades conjuntas de probabilidad, entonces

$$D\left(p(x,y) \parallel q(x,y)\right) = D\left(p(x) \parallel q(x)\right) + D\left(p(y \mid x) \parallel q(y \mid x)\right).$$

Demostración.

$$\begin{aligned} D\left(p(x,y) \parallel q(x,y)\right) &= \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_{x} \sum_{y} p(x,y) \log \frac{p(x)p(y \mid x)}{q(x)q\left(y \mid x\right)} \\ &= \sum_{x} \sum_{y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x} \sum_{y} p(x,y) \log \frac{p(y \mid x)}{q(y \mid x)} \\ &= D\left(p(x) \parallel q(x)\right) + D\left(p(y \mid x) \parallel q(y \mid x)\right). \end{aligned}$$

∎

1.5. Consecuencias de la desigualdad de Jensen

A continuación demostraremos una desigualdad de positividad de la entropía relativa.

Teorema 1.11 Sean $p(x), q(x), x \in \mathcal{X}$, dos densidades de probabilidad definidas en $\mathcal{X}$ ($\mathcal{X}$ es contable). Entonces

$$D\left(p \parallel q\right) \geq 0$$

con igualdad si y sólo si $p(x) = q(x)$ para todo x.

Demostración. Sea $A = \{x : p(x) > 0\}$ el soporte de $p(x)$. Entonces

$$\begin{aligned}
-D\left(p \parallel q\right) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \qquad (1.8) \\
&= \log \sum_{x \in A} q(x) \\
&\leq \log \sum_{x \in \mathcal{X}} q(x) \qquad (1.9) \\
&= \log 1 \\
&= 0.
\end{aligned}$$

Como $\log t$ es una función estrictamente cóncava de t, se tiene igualdad en (1.8) si y sólo si $q(x)/p(x)$ es constante en todos los puntos. Entonces, $\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c$. Se tiene igualdad en (1.9) si y sólo si $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, lo que implica que $c = 1$. Entonces, se tiene que $D\left(p \parallel q\right) = 0$ si y sólo si $p(x) = q(x)$ para todo x.

∎

Corolario 1.2. Para dos variables aleatorias discretas, X y Y

$$I(X;Y) \geq 0,$$

con igualdad si y sólo si X y Y son independientes.

Demostración. $I(X;Y) = D(p(x,y) \parallel p(x)p(y)) \geq 0$, con igualdad si y sólo si $p(x,y) = p(x)p(y)$.

∎

Corolario 1.3. Para dos variables aleatorias discretas X y Y

$$D\left(p(y \mid x) \parallel q\left(y \mid x\right)\right) \geq 0,$$

con igualdad si y sólo si $p(y \mid x) = q(y \mid x)$ para todo y y x tal que $p(x) > 0$.

Corolario 1.4. Para tres variables aleatorias discretas X, Y y Z

$$I(X;Y \mid Z) \geq 0,$$

con igualdad si y sólo si X y Y son condicionalmente independientes dado Z.

El siguiente teorema establece que condicionando se reduce la entropía, es decir que si se tiene más inormación no se puede tener más ruido.

Teorema 1.12 Para dos variables aleatorias discretas X y Y

$$H\left(X \mid Y\right) \leq H(X)$$

con igualdad si y sólo si X y Y son independientes.

Demostración. $0 \leq I(X;Y) = H(X) - H(X \mid Y)$.

∎

Intuitivamente, el teorema dice que conociendo otra variable aleatoria Y sólo puede reducir la incertidumbre en X. Notemos que esto sólo es cierto en el promedio. Específicamente, $H(X \mid Y = y)$ puede ser mayor que o menor que o igual que $H(X)$, pero en el promedio $H(X \mid Y) = \sum_y p(y)H(X \mid Y = y) \leq H(X)$. Por ejemplo, en un juicio, nueva evidencia podría incrementar la incertidumbre, pero en promedio la evidencia decrece la incertidumbre.

Ejemplo. Sea (X, Y) un vector aleatorio con la siguiente distribución conjunta:

$Y \setminus X$	1	2
1	0	3/4
2	1/8	1/8

Entonces $H(X) = 0{,}544$, $H(X|Y = 1) = 0$ y $H(X|Y = 2) = 1$. Haciendo cálculos $H(X \mid Y) = 0{,}25$. Entonces, la incertidumbre de X crece si $Y = 2$ y decrece si $Y = 1$, pero la incertidumbre decrece en promedio.

Teorema 1.13 Sean $X_1, \ldots, X_n$ con densidad conjunta $p(x_1, \ldots, x_n)$. Entonces

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

con igualdad si y sólo si las variables aleatorias X_i son independientes.

Demostración. Por la regla de la cadena para entropías

$$\begin{aligned} H(X_1, \ldots, X_n) &= H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_{i-1}, \ldots, X_1) \\ &\leq \sum_{i=1}^{n} H(X_i), \end{aligned}$$

donde la desigualdad se obtiene del Teorema 1.12. Tenemos igualdad si y sólo si X_i es independiente de $X_{i-1}, \ldots, X_1$ para todo i.

∎

1.6. Desigualdad del procesamiento de datos

A continuación demostraremos la desigualdad del procesamiento de datos que se usará en el capítulo de redes. Empecemos con una definición de cadenas de Markov en ese orden.

Definición. Las variables aleatorias X, Y, Z se dicen que forman una *cadena de Markov en ese orden* (denotada por $X \to Y \to Z$) si la distribución condicional de Z depende solamente de Y y es condicionalmente independiente de X. Específicamente, X, Y, Z forma una cadena de Markov $X \to Y \to Z$ si la densidad de la probabilidad conjunta se puede escribir como

$$p(x, y, z) = p(x)p(y \mid x)p(z \mid y).$$

Algunas consecuencias inmediatas son las siguientes:

- $X \to Y \to Z$ si y solamente si X y Z son condicionalmente independientes dado Y. La propiedad de Markov implica independencia condicional porque

$$p(x, z \mid y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z \mid y)}{p(y)} = p(x \mid y)p(z \mid y).$$

- $X \to Y \to Z$ implica que $Z \to Y \to X$. Entonces, a veces la condición se escribe como $X \leftrightarrow Y \leftrightarrow Z$.

- Si $Z = f(Y)$, entonces $X \to Y \to Z$.

Ahora demostremos la desigualdad del procesamiento de datos que dice que si no se procesa Y, aleatoriamente o de manera determinista, puede incrementar la información que Y contiene de X.

Teorema 1.14 Si $X \to Y \to Z$, entonces $I(X; Y) \geq I(X; Z)$.

Demostración. Por la regla de la cadena, podemos expandir la información mutua en dos maneras

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y \mid Z) \\ &= I(X;Y) + I(X;Z \mid Y). \end{aligned}$$

Como X y Z son condicionalmente independientes dado Y, tenemos que $I(X;Z \mid Y) = 0$. Como $I(X;Y \mid Z) \geq 0$, tenemos que

$$I(X;Y) \geq I(X;Z).$$

Tenemos igualdad si y solo si $I(X;Y \mid Z) = 0$ (es decir, $X \to Y \to Z$ forma una cadena de Markov). Similarmente, se puede probar que $I(Y;Z) \geq I(X;Z)$.

∎

1.7. Desigualdad de Fano

Para finalizar este capítulo demostraremos la desigualdad de Fano cuya importancia es vital para demostrar el teorema de capacidad para el canal de múltiple acceso en el capítulo 5 de redes. Supongamos que conocemos una variable aleatoria Y y deseamos adivinar el valor de una variable aleatoria X correlacionada con Y. La desigualdad de Fano relaciona la probabilidad de error en adivinar la variable aleatoria X con la entropía condicional $H(X \mid Y)$.

Se espera que se puede estimar X con una probabilidad de error pequeña si la entropía condicional $H(X \mid Y)$ es pequeña. La desigualdad de Fano cuantifica esta idea. Supongamos que deseamos estimar una variable aleatoria X con una distribución $p(x)$ que toma valores en $\mathcal{X}$. Observamos una variable aleatoria Y que está relacionada con X por la distribución condicional $p(y \mid x)$. A partir de Y, calculamos una función $g(Y) = \hat{X}$, donde $\hat{X}$ es una estimación de X y toma valores en $\hat{\mathcal{X}}$. No restringiremos a que el alfabeto $\hat{\mathcal{X}}$ sea igual a $\mathcal{X}$, y también permitiremos que la función $g(Y)$ sea aleatoria. Deseamos acotar la probabilidad de que $X \neq \hat{X}$. Observemos que $X \to Y \to \hat{X}$ forma una cadena de Markov. Definamos la probabilidad de error

$$P_e = \mathbb{P}(\hat{X} \neq X).$$

Teorema 1.15 Para un estimador $\hat{X}$ tal que $X \to Y \to \hat{X}$, con $P_e = \mathbb{P}(\hat{X} \neq X)$, tenemos que

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X \mid \hat{X}) \geq H(X \mid Y). \tag{1.10}$$

Esta desigualdad se puede hacer más débil

$$1 + P_e \log |\mathcal{X}| \geq H(X \mid Y).$$

Notemos que de (1.10) $P_e = 0$ implica que $H(X \mid Y) = 0$.

Demostración. Primero probemos la primera desigualdad en (1.10). Luego usaremos la desigualdad de procesamiento de datos para probar la segunda desigualdad en (1.10). Definamos una variable aleatoria indicadora

$$E = \begin{cases} 1 & \text{si } \hat{X} \neq X, \\ 0 & \text{si } \hat{X} = X. \end{cases}$$

Entonces, usando la regla de la cadena para entropías podemos desarrollar $H(E, X \mid \hat{X})$ en dos diferentes maneras, tenemos que

$$\begin{aligned} H(E, X \mid \hat{X}) &= H(X \mid \hat{X}) + \underbrace{H(E \mid X, \hat{X})}_{=0} \\ &= \underbrace{H(E \mid \hat{X})}_{\leq H(P_e)} + \underbrace{H(X \mid E, \hat{X})}_{\leq P_e \log |\mathcal{X}|}. \end{aligned}$$

Dado que condicionando se reduce la entropía, $H(E \mid \hat{X}) \leq H(E) = H(P_e)$. Luego, dado que E es función de X y $\hat{X}$, la entropía condicional $H(E \mid X, \hat{X})$ es igual a 0. También, dado que E es una variable aleatoria binaria, $H(E) = H(P_e)$. El término que falta, $H(X \mid E, \hat{X})$, puede acotarse de la siguiente manera

$$\begin{aligned} H(X \mid E, \hat{X}) &= \mathbb{P}(E = 0) H(X \mid \hat{X}, E = 0) + \mathbb{P}(E = 1) H(X \mid \hat{X}, E = 1) \\ &\leq (1 - P_e) 0 + P_e \log |\mathcal{X}|, \end{aligned}$$

puesto que dado $E = 0, X = \hat{X}$, y dado $E = 1$, podemos acotar por arriba la entropía condicional por el logaritmo del número de los posibles resultados. Combinando esos resultados, obtenemos que

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X \mid \hat{X}).$$

Por la desigualdad del procesamiento de datos, tenemos que $I(X; \hat{X}) \leq I(X; Y)$ dado que $X \to Y \to \hat{X}$ es una cadena de Markov, y por tanto $H(X \mid \hat{X}) \geq H(X \mid Y)$. Entonces, tenemos (1.10).

∎

Capítulo 2

Propiedades deseables de las entropías

En este capítulo se darán varias caracterizaciones de la entropía de Shannon. Se considerarán funciones que satisfacen ciertas propiedades convenientes y se demostrará que dichas funciones tienen que ser la entropía de Shannon. De manera particular, se darán propiedades que determinan a la función logaritmo cuando está definida en los naturales, lo cual es natural ya que en la definición de entropía aparece el logaritmo.

2.1. Propiedades

Se comenzará dando una definición que abarca las propiedades más importantes de las entropías.

Definición. La sucesión I_n:

(I) Es n_o-simétrica si para $n_o \geq 2$,

$$I_{n_o}(p_1, \ldots, p_{n_o}) = I_{n_o}\left(p_{\sigma(1)}, \ldots, p_{\sigma(n_o)}\right)$$

para todo $(p_1, \ldots, p_{n_o}) \in \Delta_{n_o}$.

(II) Es simétrica si es n_o-simétrica para todo $n_o \geq 2$.

(III) Está normalizada si $I_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$.

(IV) Es n_o-expandible si

$$\begin{aligned} I_{n_o}(p_1, \ldots, p_{n_o}) &= I_{n_o+1}(0, p_1, \ldots, p_{n_o}) \\ &= I_{n_o+1}(p_1, \ldots, p_k, 0, p_{k+1}, \ldots, p_{n_o}) \\ &= I_{n_o}(p_1, \ldots, p_{n_o}, 0) \end{aligned}$$

donde $k \in \{1, \ldots, n_o - 1\}$ para todo $(p_1, \ldots, p_{n_o}) \in \Delta_{n_o}$ o Γ_{n_o}.

(V) Es expandible si es n_o-expandible para todo n_o.

(VI) Es decisiva si $I_2(1,0) = I_2(0,1) = 0$.

(VII) Es n_o-recursiva si dado $n_o \geq 3$

$$I_{n_o}(p_1,\ldots,p_{n_o}) = I_{n_o-1}(p_1+p_2,p_3,\ldots,p_{n_o}) + (p_1+p_2)I_2\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right)$$

para todo $(p_1,\ldots,p_{n_o}) \in \Gamma_{n_o}$, con la convención de que $0 \cdot I_2\left(\frac{0}{0},\frac{0}{0}\right) := 0$.

(VIII) Es recursiva si es n_o-recursiva para todo $n_o \geq 3$.

(IX) Es (m_o,n_o)-fuertemente aditiva si dados m_o,$n_o \geq 2$

$$I_{m_on_o}\left(p_1q_{11},p_1q_{12},\ldots,p_1q_{n_o},p_2q_{21},\ldots,p_2q_{2n_o},\ldots,p_{m_o}q_{m_o1},\ldots,p_{m_o}q_{m_on_o}\right) = I_{m_o}(p_1,\ldots,p_{m_o}) + \sum_{j=1}^{m_o} p_j I_{n_o}(q_{j1},\ldots,q_{jn_o}) \quad (2.1)$$

para todo $(p_1,\ldots,p_{m_o}) \in \Gamma_{m_o}$,$(q_{j1},q_{j2},\ldots,q_{jn_o}) \in \Gamma_{n_o}$ $(j=1,2,\ldots,m_o)$.

(X) Es fuertemente aditiva si es (m_o,n_o)-fuertemente aditiva para todo $m_o,n_o \geq 2$.

(XI) Es (m_o,n_o) $-$aditiva si dados m_o,n_o y $(p_1,\ldots,p_{m_o}) \in \Delta_{m_o}$,$(q_1,\ldots,q_{n_o}) \in \Delta_{n_o}$,

$$I_{m_on_o}\left(p_1q_1,p_1q_2,\ldots,p_1q_{n_o},p_2q_1,\ldots,p_2q_{n_o},\ldots,\ldots,p_{m_o}q_1,p_{m_o}q_2,\ldots,p_{m_o}q_{n_o}\right) = I_{m_o}(p_1,\ldots,p_{m_o}) + I_{n_o}(q_1,\ldots,q_{n_o}) \quad (2.2)$$

(XII) Es aditiva si es (m_o,n_o) $-$aditiva para todo (m_o,n_o).

(XIII) Es (m_o,n_o)-subaditiva si dados $m_o,n_o \geq 2$

$$I_{m_on_o}\left(p_{11},p_{12},\ldots,p_{1n_o},p_{21},\ldots,p_{2n_o},\ldots,p_{m_o1},p_{m_o2},\ldots,p_{m_on_o}\right) \leq I_{m_o}\left(\sum_{k=1}^{n_o}p_{1k},\ldots,\sum_{k=1}^{n_o}p_{m_ok}\right) + I_{n_o}\left(\sum_{j=1}^{m_o}p_{j1},\ldots,\sum_{j=1}^{m_o}p_{jn_o}\right)$$

para todo $(p_{11},\ldots,p_{m_on_o}) \in \Gamma_{m_on_o}$.

(XIV) Es subaditiva si es (m_o,n_o)-subaditiva para todo $m_o,n_o \geq 2$.

(XV) Es n_o-maximal si dado $n_o \geq 2$

$$I_{n_o}(p_1,\ldots,p_{n_o}) \leq I_{n_o}\left(\frac{1}{n_o},\ldots,\frac{1}{n_o}\right)$$

para todo $(p_1,\ldots,p_{n_o}) \in \Gamma_{n_o}$.

(XVI) Es maximal si es n_o-maximal para todo $n_o \geq 2$.

(XVII) Es acotada por arriba si existe k tal que

$$I_2(1-q,q) \leq k$$

para todo $(1-q,q) \in \Gamma_2$.

(XVIII) Es n_o-negativa si para $n_o \geq 1$

$$I_{n_o}(p_1,\ldots,p_{n_o}) \geq 0$$

para todo $(p_1,\ldots,p_{n_o}) \in \Delta_{n_o}$.

(XIX) Es no negativa si es n_o-negativa para todo n_o.

(XX) Es monótona si la función $q \to I_2(1-q,q)$ es no decreciente en $\left[0,\frac{1}{2}\right]$.

(XXI) Es medible si $q \to I_2(1-q,q)$ es Lebesgue-medible en $(0,1)$ o en $[0,1]$.

(XXII) Es n_o-continua si dado n_o entonces I_{n_o} es continua en Δ_{n_o}.

(XXIII) Es continua si es n_o-continua para todo n_o.

(XXIV) Es estable en p_o si

$$\lim_{q\to 0^+} I_2(p_o,q) = I_1(p_o)$$

para $p_o \in [0,1]$.

(XXV) Es pequeña para probabilidades pequeñas si

$$\lim_{q\to 0^+} I_2(1-q,q) = 0.$$

Notemos que toda entropía 2-continua en Δ_2 o en Γ_2 es estable si es 2-expandible; y es pequeña para probabilidades pequeñas si es decisiva.

Nota 1. Las condiciones anteriores con excepción de (IV),(V) y (VI), se pueden formular en Δ_n^o y Γ_n^o. Entonces, (VI) y (IV) pueden usarse como definiciones para extender I_n de Δ_n^o a Δ_n o de Γ_n^o a Γ_n. También se supondrá que las otras condiciones son ciertas cuando I_n se extiende a Δ_n o Γ_n.

Nota 2. La recursividad se puede escribir como

$$I_{n_o}\left(p(1-q),pq,p_3,\ldots,p_{n_o}\right) = I_{n_o-1}(p,p_3,\ldots,p_{n_o})$$

donde $(p(1-q),pq,p_3,\ldots,p_{n_o}) \in \Gamma_{n_o}$.

2.2. Funciones aditivas y completamente aditivas

A continuación se demostrarán algunos resultados de teoría de números para poder demostrar una primera caracterización de la entropía de Shannon. De hecho, los siguientes tres resultados dan caracterizaciones del logaritmo cuando su dominio es el conjunto de los naturales, lo cual es natural de presentar puesto que en la definición de la entropía de Shannon aparece el logaritmo.

Definición. Una función cuyo dominio es el conjunto de los naturales es aditiva si

$$\phi(mn) = \phi(m) + \phi(n)$$

para todos los primos relativos m y n. Si se satisface para todos los naturales, entonces se le llama completamente aditiva.

Lema 2.1. Sea ϕ una función definida en los naturales. Existe $c \geq 0$ tal que $\phi(n) = c \log n$ $(n = 1, 2, \ldots,)$ si y sólo si ϕ es completamente aditiva, es no decreciente para $n \geq 2$ y $\phi(1) = 0$.

Demostración. La ida es trivial. Demostremos, pues, el regreso. Se tiene que

$$\liminf_n [\phi(n+1) - \phi(n)] \geq 0.$$

Sea $\epsilon > 0$ y $p > 1$ un entero. Por lo de arriba, existe un entero no negativo k tal que

$$\phi(n+1) - \phi(n) \geq -\epsilon \text{ si } n > p^k.$$

Usando inducción

$$\phi(n+j) \geq \phi(n) - j\epsilon \tag{2.3}$$

para toda $j \in \mathbb{N}$ y $n > p^k$. Sea $n > p^k$, existe $m = m(n) \geq k$ tal que

$$p^m \leq n < p^{m+1}.$$

Sea $n = a_m p^m + a_{m-1}p^{m-1} + \cdots + a_1 p + a_0$ con $1 \leq a_m < p, 0 \leq a_i < p\,(i = 0, 1, \ldots, m-1)$ la representación p-ádica de n. Entonces, usando (2.3) y aditividad

$$\begin{aligned}
\phi(n) &= \phi(a_m p^m + \cdots + a_1 p + a_0) \\
&\geq \phi(a_m p^m + \cdots + a_1 p) - a_0\epsilon \\
&> \phi\left(p\left(a_m p^{m-1} + \cdots + a_1\right)\right) - p\epsilon \\
&= \phi(p) + \phi(a_m p^{m-1} + \cdots + a_1) - p\epsilon \\
&\geq \phi(p) + \phi(a_m p^{m-1} + \cdots + a_2 p) - a_1\epsilon - p\epsilon \\
&> \phi(p) + \phi\left(p\left(a_m p^{m-2} + \cdots + a_3 p + a_2\right)\right) - 2p\epsilon \\
&= 2\phi(p) + \phi(a_m p^{m-2} + \cdots + a_2) - 2p\epsilon \\
&\geq \cdots \\
&\geq (m-k+1)\phi(p) + \phi\left(a_m p^{k-1} + a_{m-1}p^{k-2} + \cdots + a_{m-k+1}\right) - (m-k+1)\epsilon.
\end{aligned}$$

Sea $M = \underset{n<p^k}{max}|\phi(n)|$. Por tanto, para toda m

$$\phi(n) > (m-k+1)\phi(p) - M - (m-k+1)\epsilon p \tag{2.4}$$

entonces

$$p^m \leq n < p^{m+1}$$

implica

$$\begin{aligned} & m\log p \leq \log n < (m+1)\log p \\ \Rightarrow \quad & \tfrac{m}{\log n} \leq \tfrac{1}{\log p} < \tfrac{m}{\log n} + \tfrac{1}{\log n} \\ \Rightarrow \quad & \tfrac{1}{\log p} - \tfrac{1}{\log n} < \tfrac{m}{\log n} \leq \tfrac{1}{\log p} \end{aligned}$$

entonces

$$\lim_n \frac{m(n)}{\log n} = \frac{1}{\log p}$$

y como k depende de ϵ y no de n, entonces

$$\lim_n \frac{m-k+1}{\log n} = \frac{1}{\log p}.$$

De aquí que si $A(n) = \frac{(m-k+1)\phi(p) - M - (m-k+1)\epsilon p}{\log n}$ entonces

$$\lim_n A(n) = \frac{\phi(p)}{\log p} - \frac{\epsilon p}{\log p}.$$

Por (2.4)

$$\liminf \frac{\phi(n)}{\log n} \geq \liminf A_n = \frac{\phi(p)}{\log p} - \frac{\epsilon p}{\log p}$$

por lo tanto

$$\liminf \frac{\phi(n)}{\log n} \geq \frac{\phi(p)}{\log p} \ \forall p > 1$$

por aditividad

$$\phi(m^l) = l\phi(m)\,(l = 1, 2, \ldots)$$

entonces

$$c = \liminf_n \frac{\phi(n)}{\log n} \leq \lim_t \frac{\phi(p^t)}{\log p^t} = \frac{\phi(p)}{\log p}.$$

Por tanto, $\phi(p) = c\log p$ si $p > 1$. Como $\phi(1) = 0$, $\phi(n) = c\log n$ para toda n.

∎

Corolario 2.1 Sea ϕ una función definida en los naturales. Si (y sólo si) ϕ es (completamente) aditiva y

$$\lim_n [\phi(n+1) - \phi(n)] = 0$$

entonces existe una constante c tal que $\phi(n) = c\log n$.

Corolario 2.2 ϕ es una función aditiva y

$$\lim_n \left[\phi\left(n+1\right) - \frac{n}{n+1}\phi(n)\right] = 0$$

si y solamente si existe una constante c tal que $\phi(n) = c\log n$.

Demostración. El regreso es trivial. Probemos la ida. Sean

$$a_n = \phi(n+1) - \frac{n}{n+1}\phi(n) = \phi(n+1) - \phi(n) + \frac{1}{n+1}\phi(\mathrm{n}).$$

Por hipótesis $\lim a_n = 0$, por el Corolario 2.1, sólo tenemos que probar que

$$\lim_n \frac{\phi(n)}{n+1} = 0.$$

Ahora bien

$$\left(n+1\right)a_n = \left(n+1\right)\phi(n+1) - n\phi(n)\left(n = 1, 2, \ldots\right).$$

Sumando de $n = 1$ a $n = N-1$ se obtiene que (notando que $\phi(1) = 0$, lo cual se ve por la aditividad de la función)

$$\sum_{n=1}^{N-1}\left(n+1\right)a_n = N\phi(N).$$

Como $a_n \to 0$, para toda $\epsilon > 0$ existe N_o tal que

$$|a_n| < \epsilon \text{ si } n > N_o.$$

Entonces

$$\begin{aligned} 0 &\leq \left|\frac{\phi(N)}{N+1}\right| = \left|\frac{1}{N(N+1)}\sum_{n=1}^{N-1}\left(n+1\right)a_n\right| \\ &\leq \frac{1}{N(N+1)}\left|\sum_{n=1}^{N_o}\left(n+1\right)a_n\right| + \frac{\epsilon}{N\left(N+1\right)}\sum_{n=N_o+1}^{N-1}\left(n+1\right). \end{aligned}$$

Ahora el primer término del lado derecho de la desigualdad tiende a cero cuando $N \to \infty$. Para el segundo término notemos que

$$\begin{aligned} \frac{\epsilon}{N(N+1)}\sum_{n=N_o+1}^{N-1}\left(n+1\right) &\leq \frac{\epsilon}{N(N+1)}\sum_{n=0}^{N-1}\left(n+1\right) = \frac{\epsilon}{N(N+1)}\frac{N(N+1)}{2} \\ &= \frac{\epsilon}{2}. \end{aligned}$$

Entonces para N suficientemente grande

$$0 \leq \left|\frac{\phi(N)}{N+1}\right| < \epsilon.$$

Como ϵ puede ser arbitrariamente pequeña, se concluye que

$$\lim_n \frac{\phi(n)}{n+1} = 0.$$

∎

2.3. Relaciones y consecuencias

El siguiente resultado da una forma de escribir $f\left(\frac{n_1}{n}\right) = I_2\left(\frac{n-n_1}{n}, \frac{n_1}{n}\right)$ en términos de $I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$.

Teorema 2.1. Si $\{I_n\}$ es expandible, decisiva, y fuertemente aditiva, y si $\phi(n) = I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$, entonces, para todos los racionales $r = n_1/n \in [0,1]\,(0 \leq n_1 \leq n)$, se tiene

$$\begin{aligned} f(r) &:= f\left(\frac{n_1}{n}\right) = I_2\left(\frac{n-n_1}{n}, \frac{n_1}{n}\right) \\ &= \frac{-n_1}{n}\left[\phi(n_1) - \phi(n)\right] - \left(1 - \frac{n_1}{n}\right)\left[\phi(n-n_1) - \phi(n)\right] \end{aligned} \tag{2.5}$$

suponemos que $0 \cdot \phi(0) := 0$.

Demostración. Usando la propiedad de que $\{I_n\}$ son fuertemente aditivas (2.1) con $m_o = 2, n_o = n, p_1 = 1 - \frac{n_1}{n} = \frac{n-n_1}{n}, p_2 = 1 - p_1 = \frac{n_1}{n} (0 < n_1 < n)$, y

$$\begin{aligned} q_{1k} &= \begin{cases} \frac{1}{n_1 - n} & \text{si } k = 1, 2, \ldots, n - n_1 \\ 0 & \text{si } k = n - n_1 + 1, n - n_1 + 2, \ldots, n \end{cases} \\ q_{2k} &= \begin{cases} \frac{1}{n_1} & \text{si } k = 1, 2, \ldots, n_1 \\ 0 & \text{si } k = n_1 + 1, n_1 + 2, \ldots, n \end{cases} \end{aligned}$$

se tiene que

$$\begin{aligned} I_{2n}\Bigg(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n-n_1 \text{ veces}}, \underbrace{0, \ldots, 0}_{n_1 \text{ veces}}, \underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n_1 \text{ veces}}, \underbrace{0, \ldots, 0}_{n-n_1 \text{ veces}}\Bigg) &= \\ I_2\left(1 - \frac{n_1}{n}, \frac{n_1}{n}\right) + \left(1 - \frac{n_1}{n}\right) I_n\Bigg(\underbrace{\frac{1}{n-n_1}, \ldots, \frac{1}{n-n_1}}_{n-n_1 \text{ veces}}, \underbrace{0, \ldots, 0}_{n_1 \text{ veces}}\Bigg) & \\ + \frac{n_1}{n} I_n\Bigg(\underbrace{\frac{1}{n_1}, \ldots, \frac{1}{n_1}}_{n_1 \text{ veces}}, \underbrace{0, \ldots, 0}_{n-n_1 \text{ veces}}\Bigg) & \end{aligned}$$

usando que $\{I_n\}$ es expandible

$$\phi(n) = I_n\left(\frac{1}{n},\ldots,\frac{1}{n}\right) = I_2\left(1-\frac{n_1}{n},\frac{n_1}{n}\right) + \left(1-\frac{n_1}{n}\right)\phi(n-n_1) + \frac{n_1}{n}\phi(n_1).$$

Como $\{I_n\}$ es decisiva, la igualdad también es cierta cuando $n_1 = 0$ y $n_1 = n$.

∎

Los siguientes cuatro resultados tienen como intención dar relaciones existentes entre las propiedades deseables de la entropía de Shannon, lo cual permitirá obtener una caracterización de la entropía de Shannon.

Proposición 2.1. Si $\{I_n\}$ es aditiva y expandible, entonces es decisiva.

Demostración. En (2.2) ponemos $n_o = m_o = 2, p_1 = q_1 = 1.p_2 = q_2 = 0$, obteniendo que

$$\begin{aligned} I_2(1,0) &= I_4(1,0,0,0) = I_2(1,0) + I_2(1,0) \\ \Rightarrow I_2(1,0) &= 0. \end{aligned}$$

Similarmente se obtiene $I_2(0,1) = 0$.

∎

Proposición 2.2. Si $I_n : \Gamma_n \to \mathbb{R}\,(n = 2,3,\ldots)$ es expandible y fuertemente aditiva, entonces es recursiva.

Demostración. Poniendo $m_o = n_o - 1 = n-1, p_1 = \tilde{p_1}+\tilde{p_2}, p_j = \tilde{p}_{j+1}\,(j = 2,\ldots,n-1)$ y

$$\begin{aligned} q_{11} &= \frac{\tilde{p}_1}{\tilde{p}_1+\tilde{p}_2}, q_{12} = \frac{\tilde{p}_2}{\tilde{p}_1+\tilde{p}_2}, q_{1k} = 0\,(k = 3,4,\ldots,n) \\ q_{jj} &= 1\,(j = 2,3,\ldots,n-1) \\ q_{jk} &= 0\,(j = 2,3,\ldots,n-1; j \neq k = 1,\ldots,n) \end{aligned}$$

en (2.1), se obtiene que

$$\begin{aligned} I_{(n-1)n}(\tilde{p}_1,\tilde{p}_2,0,\ldots,0,0,\tilde{p}_3,0,0,\ldots,\ldots,0,\ldots,0,\tilde{p}_k,0) &= \\ I_{n-1}(\tilde{p}_1+\tilde{p}_2,\tilde{p}_3,\ldots,\tilde{p}_n) + (\tilde{p}_1+\tilde{p}_2)\, I_n\left(\frac{\tilde{p}_1}{\tilde{p}_1+\tilde{p}_2},\frac{\tilde{p}_2}{\tilde{p}_1+\tilde{p}_2},0,\ldots,0\right) \\ +\tilde{p}_3 I_n(0,1,0,\ldots,0) + \cdots + \tilde{p}_n I_n(0,\ldots,0,1,0) \end{aligned}$$

aplicando varias veces que es expandible y decisiva (es decisiva por la Proposición 2.1), se obtiene la recursividad.

∎

Proposición 2.3. Si $\{I_n\}$ es 3-recursiva y 3-simétrica, entonces también es 2-simétrica y es decisiva.

Demostración. Como es 3-recursiva y 3-simétrica entonces

$$\begin{aligned}(p_1+p_2)\,I_2\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right)&\\ =I_3\left(p_1,p_2,p_3\right)-I_2\left(p_1+p_2,p_3\right) &= I_3\left(p_2,p_1,p_3\right)-I_2\left(p_2+p_1,p_3\right)\\ =(p_2+p_1)\,I_2\left(\frac{p_2}{p_1+p_2},\frac{p_1}{p_1+p_2}\right)&\text{ si } p_1+p_2>0,\end{aligned}$$

lo cual da la simetría de I_2 en Γ_2.

Usando que es 3-simétrica con $p_1=p_2=\frac{1}{2}$ y $p_3=0$

$$I_3\left(\frac{1}{2},\frac{1}{2},0\right)=I_2\left(1,0\right)+\frac{1}{2}I_2\left(1,0\right);$$

ahora poniendo $p_1=p_3=\frac{1}{2}$ y $p_2=0$ llegamos a que

$$I_3\left(\frac{1}{2},0,\frac{1}{2}\right)=I_2\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{2}I_2\left(1,0\right).$$

Por la 3-simetría, el lado izquierdo de las dos igualdades anteriores son iguales, entonces también los lados derechos son iguales, lo cual da que $\{I_n\}$ es decisiva $I_2\left(1,0\right)=0$. Por la 2-simetría, $I_2\left(0,1\right)=0$.

∎

Proposición 2.4. Si $\{I_n\}$ es recursiva y 3-simétrica, entonces es simétrica y expandible.

Demostración. Por la proposición anteriores sabemos que es simétrica para $n_o=2$. Demostremos que es simétrica si $n_o\geq 4$ usando inducción. Por n_o-recursividad

$$I_{n_o}\left(p_1,\ldots,p_{n_o}\right)=I_{n_o-1}\left(p_1+p_2,p_3,\ldots,p_{n_o}\right)+(p_1+p_2)\,I_2\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right).$$

Usando que es 2-simétrica y (n_o-1)-simétrica, se obtiene que I_{n_o} es invariante bajo permutaciones de $(p_3,p_4,\ldots,p_{n_o})$ y de (p_1,p_2). Entonces, para probar que es invariante bajo permutaciones de $(p_1,\ldots,p_{n_o})$ es suficiente demostrar, por ejemplo, que I_{n_o} es invariante bajo el intercambio de p_2 y p_3. Para evitar fórmulas innecesarias, primero excluiremos los casos $p_1+p_2=0$ y $p_1+p_3=0$. Por recursividad se tiene que

$$\begin{aligned} I_{n_o}(p_1,\ldots,p_{n_o}) &= I_{n_o-1}(p_1+p_2,p_3,\ldots,p_{n_o}) + (p_1+p_2)\, I_2\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right) \\ &= I_{n_o-2}(p_1+p_2+p_3,p_4,\ldots,p_{n_o}) \\ &\quad + (p_1+p_2+p_3)\, I_2\left(\frac{p_1+p_2}{p_1+p_2+p_3},\frac{p_3}{p_1+p_2+p_3}\right) \\ &\quad + (p_1+p_2)\, I_2\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right), \end{aligned} \tag{2.6}$$

$$\begin{aligned} I_{n_o}(p_1,\ldots,p_{n_o}) &= I_{n_o-1}(p_1+p_3,p_2,p_4,\ldots,p_{n_o}) + (p_1+p_3)\, I_2\left(\frac{p_1}{p_1+p_3},\frac{p_3}{p_1+p_3}\right) \\ &= I_{n_o-2}(p_1+p_2+p_3,p_4,\ldots,p_{n_o}) \\ &\quad + (p_1+p_2+p_3)\, I_2\left(\frac{p_1+p_3}{p_1+p_2+p_3},\frac{p_2}{p_1+p_2+p_3}\right) \\ &\quad + (p_1+p_3)\, I_2\left(\frac{p_1}{p_1+p_3},\frac{p_3}{p_1+p_3}\right). \end{aligned}$$

Sin embargo,

$$\left(\frac{p_1}{p_1+p_2+p_3},\frac{p_2}{p_1+p_2+p_3},\frac{p_3}{p_1+p_2+p_3}\right) \in \Gamma_3,$$

y, dado que es 3-recursiva y 3-simétrica, lo cual se ha supuesto,

$$\begin{aligned} I_2\left(\frac{p_1+p_2}{p_1+p_2+p_3},\frac{p_3}{p_1+p_2+p_3}\right) + \frac{p_1+p_2}{p_1+p_2+p_3} I_2\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right) &= \\ I_3\left(\frac{p_1}{p_1+p_2+p_3},\frac{p_2}{p_1+p_2+p_3},\frac{p_3}{p_1+p_2+p_3}\right) &= \\ I_3\left(\frac{p_1}{p_1+p_2+p_3},\frac{p_3}{p_1+p_2+p_3},\frac{p_2}{p_1+p_2+p_3}\right) &= \\ I_2\left(\frac{p_1+p_3}{p_1+p_2+p_3},\frac{p_2}{p_1+p_2+p_3}\right) & \\ + \frac{p_1+p_3}{p_1+p_2+p_3} I_2\left(\frac{p_1}{p_1+p_3},\frac{p_3}{p_1+p_3}\right). & \end{aligned}$$

Entonces los lados derechos de (2.6) son iguales, y la n_o-simetría está probada en todos los casos excepto cuando $p_1 = p_2 = 0$ o $p_1 = p_3 = 0$.

Fácilmente se ve que si $p_1 = p_2 = 0$,

$$I_{n_o}(0,0,p_3,\ldots,p_{n_o}) = I_{n_o-1}(0,p_3,\ldots,p_{n_o})$$

y I_{n_o} es simétrica por la hipótesis de inducción.

Para el caso $p_1 = p_3 = 0$, podemos suponer que $p_2 > 0$; de otra forma, tendremos el caso $p_1 = p_2 = 0$, que ya se ha analizado. Si $p_1 = 0$ pero $p_2 > 0$, entonces, por la n_o-recursivdad y por ser decisiva (por la proposición anterior)

$$\begin{aligned} I_{n_o}(0, p_2, p_3, \ldots, p_{n_o}) &= I_{n_o-1}(p_2, p_3, \ldots, p_{n_o}) + p_2 I_2(0,1) \\ &= I_{n_o-1}(p_2, p_3, \ldots, p_{n_o}) \end{aligned}$$

y

$$\begin{aligned} I_{n_o}(p_2, 0, p_3, \ldots, p_{n_o}) &= I_{n_o-1}(p_2, p_3, \ldots, p_{n_o}) + p_2 I_2(1,0) \\ &= I_{n_o-1}(p_2, p_3, \ldots, p_{n_o}). \end{aligned}$$

Entonces, en éste caso, I_{n_o} es invariante bajo permutaciones de (p_1, p_2), y también, por la hipótesis de inducción, bajo permutaciones de $(p_2, p_3, \ldots, p_{n_o})$; entonces, I_{n_o} es simétrica. Notar que, en ambos casos

$$I_n(0, p_2, p_3, \ldots, p_n) = I_{n-1}(p_2, p_3, \ldots, p_n). \tag{2.7}$$

Entonces, hemos probado la simetría en todos los casos. Es fácil ver que la simetría junto con (2.7), implica expansibilidad.

∎

2.4. Caracterizaciones de Sahnnon-Khinchin y Faddeev

En el siguiente teorema se presenta la caracterización de la entropía de Shannon que dió Khinchin en 1953.

Teorema 2.2. Tenemos que $I_n(p_1, \ldots, p_n) = H_n(p_1, \ldots, p_n)$ para $(p_1, \ldots, p_n) \in \Gamma_n$ y toda $n \geq 2$ si y sólo si I_n está normalizada, es expandible, decisiva, fuertemente aditiva, maximal y 2-continua.

Demostración. La ida ya está demostrada. Así pues, demostremos el regreso.

Sea $\phi(n) = I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ $(n \geq 2)$ y $\phi(1) = 0$. Entonces si $m, n \geq 2$, usando que es aditiva se obtiene que

$$\begin{aligned} \phi(mn) &= I_{mn}\left(\frac{1}{mn}, \ldots, \frac{1}{mn}\right) \\ &= I_m\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) + I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) \\ &= \phi(m) + \phi(n). \end{aligned}$$

Si $m = 1$,

$$\phi(n) = I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \phi(1) + \phi(n)$$

por tanto, $\phi(mn) = \phi(m) + \phi(n)$ para todo $m, n \in \mathbb{Z}$. También es no decreciente para $n \geq 2$, dado que usando que $\{I_n\}$ es maximal

$$\begin{aligned} \phi(n) &= I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = I_{n+1}\left(\frac{1}{n}, \ldots, \frac{1}{n}, 0\right) \\ &\leq I_{n+1}\left(\frac{1}{n+1}, \ldots, \frac{1}{n+1}\right) = \phi(n+1). \end{aligned}$$

Usando el Lema 2.1 se concluye que existe $c > 0$ tal que $\phi(n) = c \log n$. Usando el Teorema 2.1 con $r = \frac{n_1}{n} \in [0, 1]$

$$\begin{aligned} I_2(1-r, r) &= I_2\left(\frac{n-n_1}{n}, \frac{n_1}{n}\right) \\ &= -\frac{n_1}{n}(c\log n_1 - c\log n) - \left(1 - \frac{n_1}{n}\right)(c\log(n-n_1) - c\log n) \\ &= c\left[-\frac{n_1}{n}\log\frac{n_1}{n} - \left(1-\frac{n_1}{n}\right)\log\left(1-\frac{n_1}{n}\right)\right] \\ &= c\left(-(1-r)\log(1-r) - r\log r\right) \end{aligned}$$

como $I_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$, entonces

$$1 = I_2\left(\frac{1}{2}, \frac{1}{2}\right) = c\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) = c$$

por tanto

$$I_2(1-r, r) = -(1-r)\log(1-r) - r\log r \text{ si } r \in [0,1] \bigcap \mathbb{Q}.$$

Usando que I_2 es continua en Γ_2, entonces

$$I_2(1-x, x) = H_2(1-x, x)\ \forall x \in [0, 1]$$

y por tanto la igualdad es cierta para $n = 2$. Por la Proposición 2.2, $\{I_n\}$ es recursiva, y la recursividad determina únicamente a I_n para $n = 3, 4, \ldots$ cuando I_2 es dado, y como $\{H_n\}$ es recursiva entonces se concluye la prueba.

∎

Ahora se busca dar una caracterización de la entropía de Shannon donde no se pida que $\{I_n\}$ sea maximal. Para esto se van a demostrar tres resultados que muestran relaciones entre las propiedades deseables de la entropía.

Lema 2.2. Si $\{I_n\}$ es simétrica y recursiva, entonces

$$I_{m-1+n}(p_1q_{11}, p_1q_{12}, \ldots, p_1q_{1n}, p_2, p_m) = I_m(p_1, \ldots, p_m) + p_1 I_n(q_{11}, \ldots, q_{1n}) \quad (2.8)$$

siempre que $(p_1, \ldots, p_m) \in \Gamma_m, (q_{11}, q_{12}, \ldots, q_{1n}) \in \Gamma_n\,(m, n = 2, 3, \ldots)$.

Demostración. Cuando $n = 2$ es trivial ver que se cumple (2.8) usando la recursividad. Probemos que se cumple para n usando inducción. Supongamos que es cierta para $n \geq 2$. Entonces para $n+1$ y algunos $(q_{11}, q_{12}, \ldots, q_{1,n+1}) \in \Gamma_{n+1}$, usando la recursividad y simetría, se tiene que

$$\begin{aligned}
I_{m-1+n+1}\left(p_1q_{11}, \ldots, p_1q_{1,n-1}, p_1q_{1n}, p_1q_{1,n+1}, p_2, \ldots, p_m\right) &= \\
I_{m-1+n}\left(p_1q_{11}, \ldots, p_1q_{1,n-1}, p_1\left(q_{1n}+q_{1,n+1}\right), p_2, \ldots, p_m\right) & \\
+p_1\left(q_{1n}+q_{1,n+1}\right) I_2\left(\frac{q_{1n}}{q_{1n}+q_{1,n+1}}, \frac{q_{1,n+1}}{q_{1n}+q_{1,n+1}}\right) &= \\
I_m\left(p_1, p_2, \ldots, p_m\right) & \\
+p_1\left[I_n\left(q_{11}, \ldots, q_{1,n-1}, q_{1n}+q_{1,n+1}\right) + \left(q_{1n}+q_{1,n+1}\right) I_2\left(\frac{q_{1n}}{q_{1n}+q_{1,n+1}}, \frac{q_{1,n+1}}{q_{1n}+q_{1,n+1}}\right)\right] &= \\
I_m\left(p_1, \ldots, p_m\right) + p_1 I_{n+1}\left(q_{11}, \ldots, q_{1,n-1}, q_{1n}, q_{1,n+1}\right) &
\end{aligned}$$

lo cual concluye la demostración.

∎

Teorema 2.3. Si $\{I_n\}$ es recursiva y 3-simétrica, entonces es fuertemente aditiva.

Demostración. Por el lema anterior y la Proposición 2.4 las condiciones implican (2.8) y simetría. Usando repetidamente esas propiedades obtenemos que

$$\begin{aligned}
I_{mn}\left(p_1q_{11}, p_1q_{12}, \ldots, p_1q_{1n}, p_2q_{21}, \ldots, p_2q_{2n}, \ldots, p_mq_{m1}, p_mq_{m2}, \ldots, p_mq_{mn}\right) & \\
= I_{mn-(n-1)}\left(p_1, p_2q_{21}, p_2q_{22}, \ldots, p_2q_{2n}, \ldots, p_mq_{m1}, p_mq_{m2}, \ldots, p_mq_{mn}\right) & \\
+p_1 I_n\left(q_{11}, q_{12}, \ldots, q_{1n}\right) & \\
= \cdots = I_m\left(p_1, \ldots, p_m\right) + \sum_{j=1}^{m} p_j I_n\left(q_{j1}, \ldots, q_{jn}\right), &
\end{aligned}$$

lo cual termina la demostración.

Usando las Proposiciones 2.3,2.4; y los Teoremas 2.1 y 2.3 se obtiene el siguiente corolario.

Corolario 2.3. Las fórmulas $\phi(mn) = \phi(m) + \phi(n)$ y (2.5) son válidas para ϕ definida por $\phi(n) = I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$, si se asume solamente que $\{I_n\}$ 3-simétrica y recursiva.

Usando las Proposiciones 2.3 y 2.4; el Teorema 2.3 y el Corolario 2.3 se obtiene una caracterización de la entropía de Shannon que no requiere que $\{I_n\}$ sea maximal.

Teorema 2.4. Tenemos que $I_n(p_1,\ldots,p_n) = H_n(p_1,\ldots,p_n)$ para $(p_1,\ldots,p_n) \in \Gamma_n$ y toda $n \geq 2$ si y sólo si I_n es 3-simétrica, normalizada, recursiva, 2-continua, y maximal.

El siguiente resultado nos permitirá dar otra caracterización de la entropía de Shannon.

Proposición 2.5. Si $\{I_n\}$ es recursiva, 3-simétrica y pequeña para probabilidades pequeñas, entonces $\phi(n) = I_n\left(\frac{1}{n},\ldots,\frac{1}{n}\right)$ satisface que

$$\lim_n \left[\phi(n+1) - \phi(n)\frac{n}{n+1}\right] = 0$$

Demostración. Por el Corolario 2.3, tenemos el Teorema 2.1, y entonces

$$\begin{aligned} I_2\left(\frac{n}{n+1},\frac{1}{n+1}\right) &= -\frac{1}{n+1}\left[\phi(1) - \phi(n+1)\right] - \frac{n}{n+1}\left[\phi(n) - \phi(n+1\right] \\ &= \phi(n+1) - \frac{n}{n+1}\phi(n). \end{aligned}$$

Y usando que es pequeña para probabilidades pequeñas

$$\lim_n I_2\left(\frac{n}{n+1},\frac{1}{n+1}\right) = 0$$

y se concluye la prueba.

∎

Teorema 2.5. Tenemos que $I_n(p_1,\ldots,p_n) = H_n(p_1,\ldots,p_n)$ para $(p_1,\ldots,p_n) \in \Gamma_n$ y toda $n \geq 2$ si y sólo si I_n es 3-simétrica, normalizada, recursiva, y 2-continua.

Demostración. En la prueba del Teorema 2.3 solamente se uso que es maximal en la prueba de $\phi(n) = c\log n$ cuando se buscaba demostrar que $\phi(mn) = \phi(m)+\phi(n)$ usando el hecho de que $\phi(n)$ es no decreciente. Ahora bien, como I_2 es continua, entonces I_2 es pequeña para probabilidades pequeñas (es decisiva por la Proposición 2.3), y entonces por la Proposición 2.5

$$\lim_n \left[\phi(n+1) - \frac{n}{n+1}\phi(n)\right] = 0.$$

Entonces, por el Corolario 2.2 se concluye que $\phi(n) = c\log n$. Por tanto, se concluye la demostración del teorema.

∎

Capítulo 3

La ecuación fundamental de la información

Las caracterizaciones de la entropía de Shannon presentadas en el capítulo 2 no manifiestan la característica de que una entropía mide la cantidad de información. En este capítulo se presentarán caracterizaciones donde aparece esta propiedad, la cual está basada en la cantidad de información que hay en los eventos a través de su probabilidad. De hecho se obtendrá una caracterización donde sólo se pide que las funciones $\{I_n\}$ sean recursivas, 3−simétrica, y medibles, lo cual es notablemente diferente a las presentadas en el capítulo anterior.

3.1. Funciones de información

Sea $(X, \mathbb{F}, \mathbb{P})$ un espacio de probabilidad. Si $A \in \mathbb{F}$, buscamos saber cuánta información contiene A. Así que le asociaremos una función $I(A)$ que es la información contenida en A y que depende de $\mathbb{P}(A)$. Para lo anterior, supondremos que existe una función f tal que $I(A) = f(\mathbb{P}(A))$.

Supondremos que nuestro espacio es no atómico, entonces si $A \in \mathbb{F}$ es tal que $\mathbb{P}(A) > 0$ y $z \in [0, \mathbb{P}(A)]$, existe $C \in \mathbb{F}$ tal que $C \subset A$ con $\mathbb{P}(C) = z$. Si $B \in \mathbb{F}$ con $\mathbb{P}(B) > 0$, definimos

$$\mathbb{F}_B := \{A : A \subset B, A \in \mathbb{F}\}$$

y

$$\mathbb{P}_B(A) := \mathbb{P}(A/B) = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} \text{ si } A \in \mathbb{F}_{\mathbb{B}}.$$

Consideraremos el espacio $(X, \mathbb{F}_B, \mathbb{P}_B)$. Ahora definiremos la *información relativa.*

Definición. La información relativa $I(A/B)$ contenida en $A \in \mathbb{F}_B$ con respecto a B es

$$I(A/B) = \mathbb{P}(B) f(\mathbb{P}(A/B)) = \mathbb{P}(B) f\left(\frac{\mathbb{P}(A)}{\mathbb{P}(B)}\right) \text{ si } \mathbb{P}(B) > 0$$

y $I(A/B) = 0$ si $\mathbb{P}(B) = 0$.

Definición. La información $I(A,B)$ contenida en (A,B) es

$$I(A,B) = I(A) + I(B/A^c) \text{ si } A \bigcap B = \emptyset \text{ con } \mathbb{P}(A), \mathbb{P}(B) < 1.$$

Nuestro primer postulado es que $I(A,B) = I(B,A)$, es decir que

$$\begin{aligned} I(B,A) &= I(B) + I(A/B^c) = I(B) + \mathbb{P}(B^c) f\left(\frac{\mathbb{P}(A)}{\mathbb{P}(B^c)}\right) \\ &= f(\mathbb{P}(B)) + \mathbb{P}(B^c) f\left(\frac{\mathbb{P}(A)}{\mathbb{P}(B^c)}\right) \end{aligned}$$

y

$$\begin{aligned} I(A,B) &= I(A) + \mathbb{P}(A^c) f\left(\frac{\mathbb{P}(B)}{\mathbb{P}(A^c)}\right) \\ &= f(\mathbb{P}(A)) + \mathbb{P}(A^c) f\left(\frac{\mathbb{P}(B)}{\mathbb{P}(A^c)}\right). \end{aligned}$$

El segundo postulado es que $I(A) = 1$ si $\mathbb{P}(A) = \frac{1}{2}$. Y los eventos con probabilidad 1 y 0 tienen la misma cantidad de información.

Nota 1. Si $A \bigcap B = \emptyset$ con $x = \mathbb{P}(A) \in [0,1]$ y $y = \mathbb{P}(B) \in [0,1]$. Sea $D := \{(x,y) : x \in [0,1], y \in [0,1], x + y \leq 1\}$, entonces

$$I(A,B) = f(x) + (1-x) f\left(\frac{y}{1-x}\right) \text{ si } (x,y) \in D$$

la simetría implica

$$f(x) + (1-x) f\left(\frac{y}{1-x}\right) = f(y) + (1-y)\, f\left(\frac{x}{1-y}\right) \text{ si } (x,y) \in D$$

a la ecuación anterior se le llama la *ecuación fundamental de la información.*

Nota 2. Si $(x,y) \in [0,1]$ existen A y B independientes tales que $\mathbb{P}(A) = x$ y $\mathbb{P}(B) = y$. Puesto que, $P(X) = 1$ y $x \leq 1$, entonces existe A tal que $\mathbb{P}(A) = x$; y $y \leq 1$, entonces $xy \leq x$, por tanto existe $C \subset A$ tal que $\mathbb{P}(C) = xy$. Luego, $\mathbb{P}(A^c) = 1 - x$, y $(1-x)y \leq 1 - x$, entonces existe $D \subset A^c$ tal que $\mathbb{P}(D) = (1-x)y$. Sea $B = C \bigcup D$, entonces $\mathbb{P}(A) = x$, $\mathbb{P}(B) = \mathbb{P}(C) + \mathbb{P}(D) = y$ y $\mathbb{P}(A \bigcap B) = \mathbb{P}(C) = xy = \mathbb{P}(A)\mathbb{P}(B)$.

Por tanto si $(x,y) \in D$ se satisface la ecuación de la información. Notemos además que por el segundo postulado $f(1/2) = 1$ y $f(1) = f(0)$.

Definición. Una función $f : [0,1] \to \mathbb{R}$ que satisface la ecuación fundamental de la información, y es tal que $f(1/2) = 1$ y $f(1) = f(0)$, se le llama función de información.

En el siguiente ejemplo se presenta una función de información.

Ejemplo. Consideremos $S(x) = L(x) + L(1-x) = H_2(1-x, x)$, dicha función es función de información.

Una pregunta natural es si existen funciones de información diferentes de S. La respuesta a esta pregunta se da en el siguiente teorema. Dichas funciones existen y no son Lebesgue medibles y aparecen también en el contexto de ecuaciones funcionales.

Teorema 3.1. Sea $h : (0,1) \to \mathbb{R}$ tal que

$$\begin{aligned} \text{(i) } h(xy) &= h(x) + h(y) \\ \text{(ii) } h(1/2) &= 1 \end{aligned}$$

entonces f definida por $f(x) = k(x) + k(1-x)$, donde

$$k(x) = \begin{cases} xh(x) & \text{si } x \in (0,1) \\ 0 & \text{si } x = 0 \text{ o } x = 1 \end{cases}$$

es una función de información. Además, existe h que satisface $(i), (ii)$, y es diferente de $x \to -\log x$.

Demostración. Es trivial ver que $k(xy) = xk(y) + yk(x)$ si $x, y \in [0,1]$. Entonces

$$k(s) = k\left(\frac{s}{t}t\right) = \frac{s}{t}k(t) + tk\left(\frac{s}{t}\right)$$

entonces

$$k\left(\frac{s}{t}\right) = \frac{k(s)t - sk(t)}{t^2}$$

siempre que $t \neq 0$ y $0 \leq s \leq t \leq 1$. Veamos que f satisface la ecuación fundamental de la información

$$f(x) + (1-x)f\left(\frac{y}{1-x}\right) = k(x) + k(1-x) + (1-x)\left[k\left(\frac{y}{1-x}\right) + k\left(\frac{1-x-y}{1-x}\right)\right] =$$

$$k(x) + k(1-x) + \frac{(1-x)k(y) - yk(1-x) + (1-x)k(1-x-y) - (1-x-y)k(1-x)}{1-x} =$$

$$k(x) + k(y) + k(1-x-y) \text{ si } (x,y) \in D.$$

Como el lado derecho es simétrico se sigue que f satisface la ecuación fundamental de la información. Además

$$f\left(\frac{1}{2}\right) = \frac{1}{2}h\left(\frac{1}{2}\right) + \frac{1}{2}h\left(\frac{1}{2}\right) = h\left(\frac{1}{2}\right) = 1$$

y $f(0) = f(1) = 0$.

Ahora veamos la existencia de la función h que satisface $(i), (ii)$, y que es diferente de $x \to -\log x$. Consideremos el espacio vectorial $\mathbb{R}$ sobre $\mathbb{Q}$. Sea B una base de Hamel del espacio vectorial. Toda combinación lineal es determinada por un subconjunto finito de $\mathbb{Q} \times B$. Es claro que $\mathcal{N} \leq card\,(\mathbb{Q} \times B) = \mathcal{N}_\alpha$, donde $\mathcal{N}$ es la cardinalidad de los naturales. Entonces, hay $\mathcal{N}_\alpha^n = \mathcal{N}_\alpha$ subconjuntos de tamaño n. Por tanto $card\,(\text{combinaciones lineales}) \leq \mathcal{N} \cdot \mathcal{N}_\alpha = \mathcal{N}_\alpha$. Luego

$$\begin{aligned} card\,(B) &\leq card\,(\mathbb{R}) = c = card\,(\text{combinaciones lineales}) \\ &\leq card\,(\mathbb{Q} \times B) = max\,\{\mathcal{N}, \text{card } B\} \end{aligned}$$

por ende card $B = c$. Sea $B = \{b_a : a \in \mathbb{R}\}$ una base. De hecho, podemos pedir que $b_0 = 1$. Si $x = q_1 b_{a_1} + \cdots + q_n b_{a_n}$, definimos

$$f(x) = \sum_i q_i f\,(b_{a_i})$$

falta determinar los valores de $f(b_a)$ para todo $a \in \mathbb{R}$. Sea g una biyección sobre $\mathbb{R}$ tal que para todo $a \in \mathbb{R}$, $f\,(b_a) = b_{g(a)}$ donde g es distinta de la identidad; por ejemplo, podemos tomar $g = 2a$. Notemos que $f(b_0) = b_0 = 1$. Entonces

$$f(x) = \sum_i q_i b_{2a_i} \text{ si } x = \sum q_i b_{a_i}$$

por ende, $h(x) = f(-\log x)$ si $x \in (0,1)$ satisface que $h(xy) = h(x) + h(y)$ y $h(1/2) = 1$. Demostremos que f es discontinua en todos los puntos; si f fuera continua en un punto x_o entonces

$$\begin{aligned} \lim_{y \to x} f(y) &= \lim_{y \to x} f\,(y - x + x - x_0 + x_0) = \lim_{y \to x} (f(y - x + x_0) + f(x - x_0)) \\ &= \lim_{y \to x_0} (f(y) + f(x - x_0)) = f(x_0) + f(x - x_0) = f(x) \end{aligned}$$

es decir que f es continua en todos lados. Pero si $x \in \mathbb{R}$ sea $\{q_n\} \subset \mathbb{Q}$ tal que $q_n \to x$, entonces

$$\begin{aligned} f(x) &= f(\lim q_n) = \lim f(q_n) = \lim q_n f(1) \\ &= xf(1) = cx \text{ donde } c = f(1) \end{aligned}$$

es decir que $f(b_a) = cb_a$ para todo $a \in \mathbb{R}$. Pero, $f(b_0) = b_0 = cb_0$ por tanto $c = 1$ dado que $b_0 \neq 0$. Por tanto, $f(b_a) = b_a$ para todo $a \in \mathbb{R}$, y esto es una contradicción. Por tanto, f es discontinua en todos lados. Y de esta forma, h es distinta de $x \to -\log x$. El Teorema 3.8 justifica que además estas funciones no son Lebesgue medible.

∎

En lo que resta del capítulo se determinarán todas las funciones de información que satisfacen ciertas *condiciones de regularidad* que son naturales en el contexto de la teoría de la información, y también se darán otras caracterizaciones de la entropía de Shannon bajo condiciones más débiles. Para este propósito, estableceremos conexiones entre las funciones de información y las entropías. Primero presentaremos dos propiedades importantes de las funciones de información.

Proposición 3.1. Toda función de información satisface que

$$\begin{aligned} f(0) &= f(1) = 0, \\ f(x) &= f(1-x) \text{ si } x \in [0,1]. \end{aligned}$$

Demostración. Con $y = 0$ en la ecuación fundamental de información

$$f(x) + (1-x)f(0) = f(0) + f(x)$$

por tanto $f(0) = f(1) = 0$. Con $y = 1 - x$,

$$\begin{aligned} f(x) + (1-x)\, f(1) &= f(1-x) + xf(x) \text{ si } x \in [0,1] \\ \Rightarrow f(x) &= f(1-x). \end{aligned}$$

∎

Teorema 3.2. Si una entropía $\{I_n\}$ es normalizada, 3-simétrica, y 3-recursiva, entonces

$$f(x) = I_2(1-x, x)\ x \in [0,1]$$

es una función de información.

Demostración. Usando que $\{I_n\}$ es normalizada, recursiva y $f(1/2) = 1$

$$I_3\,(p_1, p_2, p_3) = (p_1 + p_2) f\left(\frac{p_2}{p_1 + p_2}\right) + f(p_3) \text{ si } (p_1, p_2, p_3) \in \Gamma_3.$$

Como es 3-simétrica, entonces

$$\begin{aligned} (p_1 + p_3)\, f\left(\frac{p_3}{p_1 + p_3}\right) + f(p_2) &= I_3\,(p_1, p_3, p_2) \\ &= I_3\,(p_1, p_2, p_3) \\ &= (p_1 + p_2)\, f\left(\frac{p_2}{p_1 + p_2}\right) + f(p_3), \qquad (3.1) \end{aligned}$$

si $(p_1, p_2, p_3) \in \Gamma_3$. Si $p_1 \in [0,1], p_2 \in [0,1], p_3 \in [0,1], p_1 + p_2 + p_3 = 1$ con $x = p_2, y = p_3$ en lo anterior se obtiene que

$$f(x) + (1-x) f\left(\frac{y}{1-x}\right) = f(y) + (1-y) f\left(\frac{x}{1-y}\right) \text{ si } (x,y) \in D.$$

Esto a su vez implica que $f(1) = f(0) = 0$.

∎

El teorema anterior también pudo haber servido como definición de función de información.

En el siguiente teorema se define una entropía a partir de una función de información que satisface todas las propiedades algebraicas que resultan deseables para las entropías.

Teorema 3.3. Si f es una función de información. Definimos $\{I_n\}$ por

$$I_n(p_1,\ldots,p_n)=\sum_{k=2}^{n}(p_1+\ldots+p_k)\,f\left(\frac{p_k}{p_1+\cdots+p_k}\right)\ \text{ si }\ (p_1,\ldots,p_n)\in\Gamma_n\ n=2,3,\ldots$$

suponiendo que $0\cdot f\left(\frac{0}{0}\right):=0$. Entonces $\{I_n\}$ es simétrica, normalizada, expandible, decisiva, recursiva, fuertemente aditiva y aditiva.

Demostración. Por definición $I_2\left(p_1,p_2\right)=f(p_2)$ si $(p_1,p_2)\in\Gamma_2$. Luego,

$$\begin{aligned}
I_n\left(p_1,\ldots,p_n\right)-I_{n-1}\left(p_1+p_2,p_3,\ldots,p_n\right) &= \\
\sum_{k=2}^{n}(p_1+\ldots+p_k)\,f\left(\frac{p_k}{p_1+\cdots+p_k}\right)-\sum_{k=3}^{n}(p_1+\ldots+p_k)\,f\left(\frac{p_k}{p_1+\cdots+p_k}\right) &= \\
(p_1+p_2)\,f\left(\frac{p_2}{p_1+p_2}\right) &= \\
(p_1+p_2)\,I_2\left(\frac{p_1}{p_2+p_1},\frac{p_2}{p_1+p_2}\right)
\end{aligned}$$

si $(p_1,\ldots,p_n)\in\Gamma_n$, y por ende $\{I_n\}$ es recursiva. Como $I_2\left(p_1,p_2\right)=f(p_2)$ y $f(1)=f(0)=0$, entonces es decisiva. Como $f(1/2)=1$ entonces también es normalizada, y, como $f(x)=f(1-x)$ es 2−simétrica. Ahora, usando la ecuación fundamental de la información tenemos que

$$I_3\left(p_1,p_2,p_3\right)=I_3\left(p_1,p_3,p_2\right) \tag{3.2}$$

excepto cuando $p_3=1$ o $p_2=1$. En cualquiera de esos casos, $p_1=0$, y entonces se cumple (3.1) porque $f(1)=f(0)=0$ y $f(x)=f(1-x)$. Entonces, (3.2) es cierta si $(p_1,p_2,p_3)\in\Gamma_3$, y junto con el hecho de que $\{I_n\}$ es 2-simétrica se obtiene que $\{I_n\}$ es 3-simétrica. Pero entonces por la Proposición 2.4, $\{I_n\}$ es simétrica y expandible, y también fuertemente aditiva (Teorema 2.3), por lo que también se tiene que $\{I_n\}$ es aditiva.

∎

El siguiente corolario que es consecuencia de los dos teoremas anteriores, conecta los resultados de dichos teoremas.

Corolario 3.1. Si la entropía $\{I_n\}$ es normalizada, 3-simétrica, y recursiva entonces f definida como en el Teorema 3.2 es una función de información, y $\{I_n\}$ también es decisiva, expandible, simétrica, fuertemente aditiva y aditiva.

Notar que aquí lo importante es que las entropías se pueden construir a partir de funciones de información.

3.2. Funciones de información continuas en el origen

La siguiente proposición nos permitirá obtener una nueva caracterización de la entropía de Shannon: sólo se pedirá que $\{I_n\}$ sea pequeña para probabilidades pequeñas, normalizada y 3-simétrica.

Proposición 3.2. La entropía de Shannon es la única función de información continua en $[0, 1]$.

Demostración. Por el Teorema 3.3, las entropías que se construyen a partir de una función de la información son simétricas, normalizadas, y recursivas. También, de la demostración del Teorema 3.3

$$I_2(p_1, p_2) = f(p_2) \text{ si } (p_1, p_2) \in \Gamma_2.$$

Como f es continua en $[0, 1]$, entonces I_2 es continua en Γ_2. Usando el Teorema 2.5

$$f(q) = I_2(1 - q, q) = H_2(1 - q, q) = S(q)$$

con $q \in [0, 1]$. Lo cual concluye la demostración.

∎

Teorema 3.4. La función de información f es continua en 0 (por la derecha) si y sólo si $f \equiv S$ $(S(x) = H_2(1 - x, x))$ en $[0, 1]$.

Demostración. La entropía definida en el teorema 3.3 es 3-simétrica, normalizada y recursiva, por el teorema 2.5 sólo hace falta ver que es 2-continua. Como $\{I_n\}$ es simétrica y recursiva entonces es aditiva. Además, por la proposición 2.5, se concluye del corolario 2.2 que $\phi(n) = \log n$, donde $\phi(n) := I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. Luego, por el teorema 2.1,

$$f(r) = f\left(\frac{n_1}{n}\right) = -\frac{n_1}{n}\log\frac{n_1}{n} - \left(1 - \frac{n_1}{n}\right)\log\left(1 - \frac{n_1}{n}\right) = S(r) \tag{3.3}$$

para todos los racionales $r \in (0, 1)$.

Ahora, sea $y \in (0, \frac{1}{2})$ arbitrario, y $\{r_n\}$ una sucesión de racionales tales que $r_n \in (, y, 2y) \subset (0, 1)$ $(n = 1, 2, \ldots)$ y con

$$\lim_n r_n = y.$$

Entonces, la nueva sucesión definida por

$$x_n = 1 - \frac{y}{r_n} \in \left(0, \frac{1}{2}\right)(n = 1, 2, \ldots), \tag{3.4}$$

satisface que

$$\lim_n x_n = 0. \tag{3.5}$$

Poniendo $x = x_n$ en la ecuación fundamental ($x_n + y \leq 1$) se obtiene que

$$f(y) = f(x_n) + (1-y)\, f\left(\frac{x_n}{1-y}\right) + (1-x_n)\, f\left(\frac{y}{1-x_n}\right).$$

Pasando al límite cuando $n \to \infty$, se obtiene, por (3.3),(3.4),(3.5), la continuidad por la derecha de 0 de f y usando que $f(0) = f(1) = 0$,

$$\begin{aligned} f(y) &= \lim_{n\to\infty} (1-x_n)\, f\left(\frac{y}{1-x_n}\right) = \lim_{n\to\infty} (1-x_n)\, f(r_n) \\ &= \lim_{n\to\infty} S(r_n) = S(y), \end{aligned}$$

puesto que S es continua. Y con esto se concluye la prueba de la ida en $\left(0, \frac{1}{2}\right)$. Usando que $f(1) = f(0) = 0$, $f(x) = f(1-x)$,$f\left(\frac{1}{2}\right) = 1$ extiende la validez a $[0,1]$, y la recursividad demuestra la ida. El regreso es fácil, pues $S(x) = H_2(1-x, x)$ es continua por la derecha en 0, y esto concluye la demostración del teorema.

∎

Corolario 3.2. $\{I_n\}$ es 3-simétrica, normalizada, recursiva y pequeña para probabilidades pequeñas si y sólo si

$$I_n(p_1, \ldots, p_n) = H_n(p_1, \ldots, p_n)$$

si $(p_1, \ldots, p_n) \in \Gamma_n$ y $n \geq 2$.

3.3. Funciones de información medibles y entropías

Ahora se busca demostrar que toda función de información que es Lebesgue medible en $(0,1)$ es la función de información de Shannon. Para dicho objetivo seguiremos los siguientes pasos. Mostraremos que toda función de información medible es acotada en un intervalo y, entonces acotada e integrable en todo intervalo cerrado de $(0,1)$. De aquí, demostraremos que f es diferenciable y que satisface una ecuación diferencial, de donde se ve que $f = S$.

Necesitaremos la siguiente definición de cierto conjunto que depende de funciones arbitrarias en $[0,1]$.

Definición. Sea f una función real en $[0,1]$. $\lambda > 0$ pertenece a G_f cuando existen $\delta_\lambda \in (0,1)$ y $k_\lambda \in (0,\infty)$ tales que

$$\left| f(t) - \lambda f\left(\frac{t}{\lambda}\right) \right| < k_\lambda \ \forall t \in [0, \delta_\lambda).$$

Lema 3.1. G_f es un grupo respecto a la multiplicación.

Demostración. Es claro que $1 \in G_f$. Sean $\lambda_1, \lambda_2 \in G_f$, existen $\delta_{\lambda_i} = \delta_i > 0, k_i > 0$ con $i = 1, 2$ tales que

$$\left| f(t) - \lambda_i f\left(\frac{t}{\lambda_i}\right) \right| < k_i \text{ si } t \in [0, \delta_i).$$

Entonces,

$$\begin{aligned} \left| f(t) - \lambda_1\lambda_2 f\left(\frac{t}{\lambda_1\lambda_2}\right) \right| &\leq \\ \left| f(t) - \lambda_1 f\left(\frac{t}{\lambda_1}\right) \right| + \lambda_1 \left| f\left(\frac{t}{\lambda_1}\right) - \lambda_2 f\left(\frac{t}{\lambda_1\lambda_2}\right) \right| &< k_1 + \lambda_1 + k_2 \text{ si } t \in [0, \min\ \{\delta_1, \delta_2\}], \end{aligned}$$

por tanto $\lambda_1\lambda_2 \in G_f$.

Si $\lambda \in G_f$, sea $t* = \frac{t}{\lambda}$,

$$|f(t*)| = \left| \frac{1}{\lambda} f(t * \lambda) \right| < \frac{k_\lambda}{\lambda} \text{ si } t* \in \left[0, \frac{\delta_\lambda}{\lambda}\right]$$

por tanto $\frac{1}{\lambda} \in G_f$.

∎

Así que si f es una función de información, entonces G_f es un subconjunto del grupo multiplicativo de los reales positivos. La estructura del grupo G_f influencia el comportamiento de f. Esto se ve en el siguiente teorema.

Teorema 3.5. Si f es una función de información y $G_f = (0, \infty)$, entonces f es acotada en todo subconjunto cerrado de $(0, 1)$.

Demostración. Sea $y \in (0, 1)$. Como $1 - y \in G_f$, existen $\delta_y > 0$ y $k_y > 0$ tales que

$$\left| f(x) - (1 - y) f\left(\frac{x}{1 - y}\right) \right| < k_y \text{ si } x \in [0, \delta_y].$$

Entonces, por la ecuación fundamental

$$\left| f(y) - (1 - x)\, f(\frac{y}{1 - x}) \right| < k_y$$

si $x \in [0, \delta_y]$ y $x + y < 1$. Por tanto si $z = \frac{y}{1-x} \in [y, 1]$

$$|1 - x|\, f(z) - |f(y)| \leq |(1 - x)f(z) - f(y)| < k_y$$

o bien

$$\begin{aligned} |f(z)| &\leq \frac{k_y + |f(y)|}{1-x} = \frac{k_y + |f(y)|}{y} z \\ &< \frac{k_y + |f(y)|}{y} =: k \end{aligned}$$

siempre que $0 \leq x = 1 - \frac{y}{z} < \delta_y$, entonces

$$|f(z)| < k \text{ si } z \in [y, y+\delta]\,.$$

También lo anterior es cierto si $1-y \in (0,1)$, es decir existen $\delta' > 0$, $k' > 0$ tales que $|f(z)| < k'$ si $z \in [y-\delta', y]$. Sean $k^* = \max\{k, k'\}$, $\delta^* = \min\{\delta, \delta'\}$, entonces

$$|f(z)| < k^* \text{ si } z \in (y-\delta^*, y+\delta^*)$$

por tanto para toda y en $(0,1)$ existe una vecindad donde f es acotada.

Si $C \subset (0,1)$ es cerrado puede ser cubierto con un número finito de intervalos donde f es acotada. Si k_o es la cota más grande de f en esos intervalos, entonces

$$|f(z)| < k_o \text{ si } z \in C.$$

∎

Para una función de información f, el siguiente resultado da una condición suficiente para que $G_f = (0, \infty)$.

Teorema 3.6. Si la función de información es acotada en $(\alpha, \beta) \subset (0,1)$, entonces $G_f = (0, \infty)$, y, entonces, f es acotada en todo subconjunto cerrado de $(0,1)$.

Demostración. Si $x \in (\alpha, \beta)$, existe $\delta > 0$ tal que

$$\frac{x}{1-y} \in (\alpha, \beta) \text{ si } y \in [0, \delta),$$

entonces por la ecuación fundamental

$$\begin{aligned} \left| f(y) - (1-x) f\left(\frac{y}{1-x}\right) \right| &= \left| f(x) - (1-y) f\left(\frac{x}{1-y}\right) \right| \\ &\leq |f(x)| + \left| f\left(\frac{x}{1-y}\right) \right| < 2k \text{ si } y \in [0, \delta), \end{aligned}$$

entonces $1-x \in G_f$. Por tanto, $(1-\beta, 1-\alpha) \subset G_f$. Como G_f es un grupo se sigue que $(0, \infty) \subset G_f$.

∎

Como consecuencia del teorema anterior obtenemos el siguiente corolario.

Corolario 3.3. Si la función de información es acotada en $[0, \epsilon]$ entonces f es acotada en $[0, 1]$.

Nota. Un problema que sigue sin resolverse es saber si dado que la función de información es acotada en todo cerrado de $(0, 1)$, ¿dicha función será acotada en $[0, 1]$?

A continuación presentaremos unos resultados de las funciones reales.

Lema 3.2. Si $A, B, C \subset \mathbb{R}$ son acotados y medibles. Entonces, la función

$$F(u, v) = m\left(A \bigcap (1 - uB) \bigcap vC\right)$$

con $u, v \in \mathbb{R}$ es continua y m es la medida de Lebesgue.

Demostración. Notemos que

$$0 \leq F(u, v) \leq m(A \bigcap (1 - uB) \bigcap vC) \leq m(vC) = vm(C). \tag{3.6}$$

Demostremos que f es continua en $(u_0, 0)$ y $(0, v_o)$. Si $(u, v) \to (u_0, 0)$ entonces $v \to 0$ y usando (3.6) se concluye que

$$\lim_{(u,v)\to(u_o,0)} F(u, v) = 0.$$

Luego

$$\begin{aligned} 0 \quad &\leq \quad F(u_o, 0) = m\left(A \bigcap (1 - u_o B) \bigcap \{0\}\right) \leq m(\{0\}) = 0 \\ \Rightarrow F(u_o, 0) \quad &= 0 \quad = \lim_{(u,v)\to(u_o,0)} F(u, v) = 0. \end{aligned}$$

Ahora supongamos que $u_o \neq 0$ y $v_o \neq 0$, por tanto

$$F(u, v) = m\left[A \bigcap \left(1 - \frac{u}{u_o} u_o B\right) \bigcap \frac{v}{v_o} v_o C\right].$$

Para probar la continuidad de F en (u_o, v_o) es suficiente demostrar la continuidad de G en $(1, 1)$, donde G es

$$G(s, t) = m\left[A \bigcap (1 - sD) \bigcap tE\right] \text{ con } D = u_o B, E = v_o C.$$

Sea $\epsilon > 0$, existen f, g continuas en $\mathbb{R}$ que se anulan fuera de un compacto T, tal que

$$\begin{cases} |f| \leq 1, \int_{\mathbb{R}} |1_{1-D}(x) - f(1 - x)|\, dx < \epsilon \\ |g| \leq 1, \int_{\mathbb{R}} |1_E(x) - g(x)|\, dx < \epsilon. \end{cases}$$

Con esas funciones obtenemos que

$$\begin{aligned}
|G(s,t)-G(1,1)| &= \left|m(A\bigcap(1-sD)\bigcap tE)-m(A\bigcap(1-D)\bigcap E)\right| \\
&= \left|\int_{\mathbb{R}}\left[1_A(x)1_{1-sD}(x)1_{tE}(x)-1_A(x)1_{1-D}(x)1_E(x)\right]dx\right| \\
&\leq \int_{\mathbb{R}}\left|1_{1-sD}(x)1_{tE}(x)-1_{1-D}(x)1_E(x)\right|dx \\
&\leq \int_{\mathbb{R}}\left|\left[1_{1-sD}(x)1_{tE}(x)-f\left(\frac{1-x}{s}\right)1_{tE}(x)\right]\right|dx \\
&\quad+\int_{\mathbb{R}}\left|f\left(\frac{1-x}{s}\right)1_{tE}(x)-f\left(\frac{1-x}{s}\right)g\left(\frac{x}{t}\right)\right|dx \\
&\quad+\int_{\mathbb{R}}\left|f\left(\frac{1-x}{s}\right)g\left(\frac{x}{t}\right)-f(1-x)g(x)\right|dx \\
&\quad+\int_{\mathbb{R}}\left|f(1-x)g(x)-1_{1-D}(x)g(x)\right|dx \\
&\quad+\int_{\mathbb{R}}\left|1_{1-D}(x)g(x)-1_{1-D}(x)1_E(x)\right|dx \\
&\leq s\epsilon+t\epsilon+\int_{\mathbb{R}}\left|f\left(\frac{1-x}{s}\right)g\left(\frac{x}{t}\right)-f(1-x)g(x)\right|dx+2\epsilon
\end{aligned}$$

si s,t son suficientemente cercanos a 1, como f,g son continuas y se anulan afuera de T, entonces la integral es arbitrariamente pequeña.

∎

Lema 3.3. Si $E\subset\left(\frac{1}{2},1\right)$, entonces $m\left(E^{-1}\right)\leq 4m\left(E\right)$; donde $E^{-1}:=\{x^{-1}:x\in E\}$.

Demostración.

$$\begin{aligned}
m\left(E^{-1}\right) &= \int_1^2 1_{E^{-1}}(x)dx=\int_1^2 1_E\left(\frac{1}{x}\right)dx \\
&= \int_{\frac{1}{2}}^1\frac{1_E(t)}{t^2}dt\leq\left(\sup_{(\frac{1}{2},1)}\frac{1}{t^2}\right)\int_{1/2}^1 1_E(t)dt\leq 4m(E).
\end{aligned}$$

∎

Estamos listos para probar el siguiente teorema que muestra cuando una función de información es acotada.

Teorema 3.7. Si f es una función de información que es medible en $(0,1)$, entonces existe $[\alpha,\beta]\subset(0,1)$ donde f es acotada.

Demostración. Por la ecuación fundamental

$$\begin{aligned} |f(x)| &\leq |f(y)| + (1-y)\left|f\left(\frac{x}{1-y}\right)\right| + (1-x)\left|f\left(\frac{y}{1-x}\right)\right| \\ &\leq |f(y)| + \left|f\left(\frac{x}{1-y}\right)\right| + \left|f\left(\frac{y}{1-x}\right)\right|. \end{aligned} \tag{3.7}$$

Se busca encontrar un conjunto $B_N \subset (0,1)$ tal que f está acotada por N en dicho conjunto. Así pues definamos B_n como

$$B_n := \{x : 0 < x < 1, |f(x)| \leq n\} \, (n = 1, 2, \ldots).$$

Sea $\epsilon > 0$, existe N tal que $m(B_N) \geq 1 - \epsilon$ o bien $m((0,1) \setminus B_N) \leq \epsilon$, pues $\sup_n m(B_n) = 1$. Ahora

$$m\left(\left(0, \frac{1}{2}\right) \setminus B_N\right) \leq m((0,1) \setminus B_N) \leq \epsilon, \tag{3.8}$$

también

$$m\left(\left(\frac{1}{2}, 1\right) \setminus B_N\right) \leq m((0,1) \setminus B_N) \leq \epsilon,$$

usando el Lema 3.3 se obtiene que $m\left((1,2) \setminus B_N^{-1}\right) \leq 4\epsilon$, de donde

$$m\left(\left(\frac{1}{2}, 1\right) \setminus \frac{1}{2}B_N^{-1}\right) = \frac{1}{2}m\left((1,2) \setminus B_N^{-1}\right) \leq 2\epsilon, \tag{3.9}$$

lo cual implica que

$$m\left(\left(0, \frac{1}{2}\right) \setminus \left(1 - \frac{1}{2}B_N^{-1}\right)\right) = m\left(\left(\frac{1}{2}, 1\right) \setminus \frac{1}{2}B_N^{-1}\right) \leq 2\epsilon \tag{3.10}$$

y por (3.3)

$$m\left(\left(0, \frac{1}{2}\right) \setminus \frac{1}{2}B_N\right) = \frac{1}{2}m((0,1) \setminus B_N) \leq \frac{1}{2}\epsilon;$$

usando (3.8),(3.9) y (3.10), se obtiene que

$$\begin{aligned} m\left(\left(0, \frac{1}{2}\right) \setminus \left[B_N \bigcap \left(1 - \frac{1}{2}B_N^{-1}\right) \bigcap \frac{1}{2}B_N\right]\right) &\leq \\ m\left(\left(0, \frac{1}{2}\right) \setminus B_N\right) + m\left(\left(0, \frac{1}{2}\right) \setminus \left(1 - \frac{1}{2}B_N^{-1}\right)\right) + m\left(\left(0, \frac{1}{2}\right) \setminus \frac{1}{2}B_N\right) &\leq \\ \epsilon + 2\epsilon + \frac{1}{2}\epsilon &< 4\epsilon. \end{aligned}$$

Si $4\epsilon < \frac{1}{2} = m\left(\left(0, \frac{1}{2}\right)\right)$, entonces

$$m\left(B_N \bigcap \left(1 - \frac{1}{2}B_N^{-1}\right) \bigcap \frac{1}{2}B_N\right) > 0,$$

como $f(x) = m\left(B_N \bigcap \left(1 - xB_N^{-1}\right) \bigcap \frac{1}{2}B_N\right)$ es continua, entonces existe un intervalo $[\alpha, \beta]$ que contiene a $\frac{1}{2}$ tal que

$$m\left(B_N \bigcap \left(1 - \frac{1}{2}B\right) \bigcap \frac{1}{2}B_N\right) > 0,$$

entonces

$$B_N \bigcap \left(1 - xB_N^{-1}\right) \bigcap (1-x)\, B_N \neq \emptyset \text{ si } x \in [\alpha, \beta].$$

Si $x \in [\alpha, \beta]$, existe $y \in B_N$ tal que $\frac{x}{1-y}$ y $\frac{y}{1-x}$ están en B_N. Por tanto, $|f(x)| \leq 3N$ si $x \in [\alpha, \beta]$.

∎

Teorema 3.8. La función de información f es medible en el intervalo $(0,1)$ si y sólo si $f(x) = S(x)$ si $x \in [0,1]$.

Demostración. El regreso es trivial. Demostremos la ida. Como f es medible entonces existe $[\alpha, \beta] \subset [0,1]$ donde f es acotada. Entonces, f es acotada en todo intervalo cerrado de $(0,1)$, y, siendo medible f, entonces f es Lebesgue integrable en esos intervalos cerrados.

Sea $y \in (0,1)$ y sean λ, μ tales que

$$0 < y < y + \lambda < y + \mu < 1. \tag{3.11}$$

Entonces, si $x \in [\alpha, \mu]$, se tiene que $\frac{x}{1-y}$ y $\frac{y}{1-x}$ están en un intervalo cerrado donde f es acotada, y por ende, también es integrable. La razón de lo anterior es la siguiente, por (3.11)

$$\begin{aligned} 0 &< \lambda < \frac{\lambda}{1-y} \leq \frac{x}{1-y} \leq \frac{\mu}{1-y} < 1, \\ 0 &< \frac{y}{1-\lambda} \leq \frac{y}{1-x} \leq \frac{y}{1-\mu} < 1. \end{aligned} \tag{3.12}$$

Integrando la ecuación fundamental de λ a μ respecto de x, obtenemos

$$\begin{aligned} (\mu - \lambda) f(y) &= \int_\lambda^\mu f(y)dx \\ &= \int_\lambda^\mu f(x)dx + \int_\lambda^\mu (1-x) f\left(\frac{y}{1-x}\right) dx \\ &\quad -(1-y)\int_\lambda^\mu f\left(\frac{x}{1-y}\right) dx \\ &= \int_\lambda^\mu f(x)dx + y^2 \int_{\frac{y}{1-\lambda}}^{\frac{y}{1-\mu}} s^{-3} f(s)ds \\ &\quad -(1-y)^2 \int_{\frac{\lambda}{1-y}}^{\frac{\mu}{1-y}} f(t)dt. \end{aligned} \tag{3.13}$$

El lado derecho de (3.13) es continuo en y, y como $\mu - \lambda \neq 0$, f en el lado izquierdo de dicha ecuación es continua en $(0,1)$; entonces f es diferenciable. Pero si f es diferenciable, entonces el lado izquierdo lo es y por ende f es 2 veces diferenciable, y así sucesivamente.

Diferenciando la ecuación fundamental respecto de x

$$f'(x) - f\left(\frac{y}{1-x}\right) + \frac{y}{1-x} f'\left(\frac{y}{1-x}\right) = f'\left(\frac{x}{1-y}\right),$$

volviendo a derivar dicha función

$$-\frac{1}{1-x} f'\left(\frac{y}{1-x}\right) + \frac{1}{1-x} f'\left(\frac{y}{1-x}\right) + \frac{y}{(1-x)^2} f''\left(\frac{y}{1-x}\right) = \frac{x}{(1-y)^2} f''\left(\frac{x}{1-y}\right),$$

entonces

$$\left(\frac{y}{1-x}\right) f''\left(\frac{y}{1-x}\right) = \frac{1-x}{y}\frac{x}{1-y} f''\left(\frac{x}{1-y}\right).$$

Sean $s := \frac{y}{1-x}, t := \frac{x}{1-y}$ (por (3.12), $s \in (0,1), t \in (0,1)$; y, para dados $s,t \in (0,1)$ existen $(x,y) \in D$ tales que satisfacen esas ecuaciones, por ejemplo se puede tomar $x = \frac{t-ts}{1-ts}$,$y = \frac{s-st}{1-st}$), entonces

$$s(1-s)f''(s) = t(1-t)f''(t) \text{ si } s,t \in (0,1),$$

por tanto $t(1-t)f''(t) = c'$, entonces

$$\begin{aligned} f''(t) &= \frac{c'}{t(1-t)} = \frac{c'}{t} + \frac{c'}{1-t} \\ \Rightarrow f'(t) &= c' \ln t - c' \ln(1-t) + a \\ \Rightarrow f(t) &= c't(\ln t - 1) + c'(1-t)(\ln(1-t) - 1) + at + b' \end{aligned}$$

donde $a, b', c' \in \mathbb{R}$. Sean $c = -c' \ln 2$ y $b = b' - c'$, entonces

$$f(t) = c\left[-t \log t - (1-t)\log(1-t)\right] + at + b.$$

Como $f(t) = f(1-t)$, se sigue que $a = 0$. Usando la ecuación fundamental se llega a que $b = 0$. Dado que $f\left(\frac{1}{2}\right) = 1$, entonces $c = 1$. De aquí que $f(x) = S(x)$ en $(0,1)$. Usando que $f(0) = f(1) = 0$, se concluye que $f(x) = S(x)$ en $[0,1]$.

∎

Usando la caracterización anterior para las funciones de información medibles se obtiene como corolario otra caracterización de la entropía de Shannon.

Corolario 3.4. Tenemos que $\{I_n\}$ es recursiva, 3-simétrica y medible si y sólo si

$$I_n(p_1, \ldots, p_n) = H_n(p_1, \ldots, p_n).$$

Capítulo 4

El mercado de acciones

El objetivo del presente capítulo es mostrar una relación entre la teoría de la información y la teoría de portafolios. Específicamente, veremos que hay una dualidad entre el crecimiento de la razón de la riqueza en el mercado de acciones y la razón de la entropía del mercado[1]. Se encuentra el crecimiento asintótico de la razón de la riqueza para un mercado de acciones ergódico. Además, se da una prueba usando un argumento del "sandwich" de la propiedad asintótica de equipartición (AEP por sus siglas en inglés) para procesos ergódicos que es motivada por la noción de portafolios óptimos para mercados de acciones estacionarios y ergódicos. Asimismo, la demostración del teorema 4.10 que se da es desarrollada a detalle a partir de las ideas generales que aparecen en el libro de Cover. Concluyentemente se menciona brevemente el portafolio universal.

En este capítulo se utiliza la teoría presentada en el capítulo 1. Asimismo, es necesario conocer la ley fuerte de los grandes números, el teorema ergódico y las propiedades esenciales de la medida e integral de Lebesgue.

4.1. Definiciones

Un mercado de acciones es un vector de acciones $\mathbf{X} = (X_1, \ldots, X_m), X_i \geq 0, i = 1, \ldots, m$ donde m es el número de acciones y X_i es el precio relativo que representa la razón del precio al final del día entre el precio al inicio del día. Entonces, típicamente, X_i es cercano a 1. Por ejemplo, $X_i = 1{,}01$ significa que la acción i-ésima se incrementó un 1 % en ese día.

Sea $\mathbf{X} \sim F(\mathbf{x})$, donde $F(\mathbf{x})$ es la distribución conjunta del vector de los precios relativos.

Un portafolio $\mathbf{b} = (b_1, \ldots, b_m), b_i \geq 0, \sum b_i = 1$ es una distribución de la riqueza en las acciones. Donde b_i es la fracción de la riqueza que se invierte en la acción i.

La riqueza relativa es $S = \mathbf{b}^t\mathbf{X}$. Lo que se busca es maximizar S en algún sentido. Como S es una variable aleatoria, existe una controversia acerca de seleccionar la mejor distribución de S. La teoría estándar sugiere considerar el primer y segundo

[1]La razón de la entropía de un proceso estocástico $\{X_i\}$ está definida por $\lim \frac{1}{n} H(X_1, \ldots, X_n)$ cuando el límite existe.

momento de S. Y el objetivo es maximizar la esperanza de S, sujeto a una condición en la varianza.

La media de una variable aleatoria proporciona información acerca del comportamiento asintótico de la suma de variables aleatorias i.i.d. con la misma distribución que la variable inicial. Dado que en el mercado de acciones normalmente se invierte cada día, entonces la riqueza al final de n días es el producto de factores, uno para cada día del mercado. Por tanto, el comportamiento del producto es determinado por el valor esperado del logaritmo. Lo cual motiva las siguientes definiciones.

Supondremos que $\{\log \mathbf{b}^t\mathbf{X} : \mathbf{b} \text{ es un portafolio}\} \subset L^1(\mathbb{P})$. En particular se podría suponer que $\#\Omega < \infty$.

Definición. La razón de crecimiento (o la razón de duplicación) de un portafolio de un mercado de acciones $\mathbf{b}$ está definida como

$$W(\mathbf{b}, F) = \int \log \mathbf{b}^t\mathbf{x} \, \mathrm{dF}(\mathbf{x}) = \mathbb{E}\left(\log \mathbf{b}^t\mathbf{X}\right).$$

Definición. La razón óptima de crecimiento $W^*(F)$ está definida como

$$W^*(F) = \max_b W(\mathbf{b}, F),$$

donde el máximo es tomado sobre todos los posibles portafolios $b_i \geq 0, \sum_i b_i = 1$, siempre que el máximo exista. Asumiremos que el conjunto de vectores que maximiza $W(\mathbf{b}, F)$ es distinto del vacío.

Definición. Un portafolio b^* que alcanza el máximo de $W(\mathbf{b}, F)$ se llama un portafolio log-óptimo.

La definición de la razón de crecimiento es justificada por el siguiente teorema, que muestra que la riqueza crece tanto como 2^{nW^*}.

Teorema 4.1. Sean $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. con distribución $F(\mathbf{x})$. Sea

$$S_n^* = \prod_{i=1}^{n} \mathbf{b}^{*t}\mathbf{X}_i$$

la riqueza después de n días usando el portafolio constante $\mathbf{b}^*$. Entonces

$$\begin{aligned} \frac{1}{n}\log S_n^* &= \frac{1}{n}\sum_{i=1}^{n} \log \mathbf{b}^{*t}\mathbf{X}_i \\ &\xrightarrow{c.s.} W^*. \end{aligned}$$

Demostración.

$$\begin{aligned}\frac{1}{n}\log S_n^* &= \frac{1}{n}\sum_{i=1}^{n}\log \mathbf{b}^{*t}\mathbf{X}_i \\ &\xrightarrow{c.s.} W^*,\end{aligned}$$

por la ley fuerte de los grandes números. Entonces, $S_n^* \approx 2^{nW^*}$.

∎

Ahora veremos algunas propiedades de concavidad y convexidad de la razón de crecimiento.

Lema 4.1. $W(\mathbf{b}, F)$ es cóncavo en $\mathbf{b}$ y lineal en F. $W^*(F)$ es convexa en F.

Demostración. La razón de crecimiento es

$$W(\mathbf{b}, F) = \int \log \mathbf{b}^t\mathbf{x}\ \mathrm{dF}(\mathbf{x}).$$

Dado que la integral es lineal en F, entonces $W(\mathbf{b}, F)$ es lineal en F.

Puesto que

$$\log\left(\lambda\mathbf{b}_1 + (1-\lambda)\,\mathbf{b}_2\right)^t\mathbf{X} \geq \lambda\log\mathbf{b}_1^t\mathbf{X} + (1-\lambda)\log\mathbf{b}_2^t\mathbf{X},$$

porque el logaritmo es cóncavo, se sigue, tomando esperanzas, que $W(\mathbf{b}, F)$ es cóncavo en $\mathbf{b}$.

Finalmente, para probar la convexidad de $W^*(F)$ como función de F, sean F_1 y F_2 dos distribuciones en el mercado de acciones y sean $\mathbf{b}^*(F_1)$ y $\mathbf{b}^*(F_2)$ los correspondientes portafolios óptimos, respectivamente. Sea $\mathbf{b}^*(\lambda F_1 + (1-\lambda)F_2)$ el portafolio log-óptimo correspondiente a $\lambda F_1 + (1 - \lambda F_2)$. Entonces por linealidad de $W(\mathbf{b}, F)$ con respecto a F, se tiene que

$$\begin{aligned}W^*(\lambda F_1 + (1-\lambda)F_2) &= W(\mathbf{b}^*(\lambda F_1 + (1-\lambda)F_2), \lambda F_1 + (1-\lambda)F_2) \\ &= \lambda W(\mathbf{b}^*(\lambda F_1 + (1-\lambda)F_2), F_1) + (1-\lambda) \\ &\quad \times W(\mathbf{b}^*(\lambda F_1 + (1-\lambda)F_2), F_2) \\ &\leq \lambda W(\mathbf{b}^*(F_1), F_1) + (1-\lambda)W^*(\mathbf{b}^*(F_2), F_2),\end{aligned}$$

puesto que $\mathbf{b}^*(F_1)$ maximiza $W(\mathbf{b}, F_1)$ y $\mathbf{b}^*(F_2)$ maximiza $W(\mathbf{b}, F_2)$.

∎

Lema 4.2. El conjunto de los portafolios log-óptimos forman un conjunto convexo.

Demostración. Sean $\mathbf{b}_1^*$ y $\mathbf{b}_2^*$ cualesquiera dos portafolios en el conjunto de los portafolios log-óptimos. Por el Lema 4.1

$$\begin{aligned} W\left(\lambda \mathbf{b}_1^* + (1-\lambda)\mathbf{b}_2^*, F\right) &\geq \lambda W\left(\mathbf{b}_1^*, F\right) + (1-\lambda) W\left(\mathbf{b}_2^*, F\right) \\ &\geq \min\ \left\{W\left(\mathbf{b}_1^*, F\right), W\left(\mathbf{b}_2^*, F\right)\right\}, \end{aligned}$$

por lo tanto dicho conjunto es convexo.

∎

4.2. Caracterización de Kuhn-Tucker del portafolio log-óptimo

En lo siguiente se usarán estas propiedades para caracterizar el portafolio log-óptimo. Sea $\mathcal{B} = \{\mathbf{b} \in \mathbb{R}^m : \mathbf{b}_i \geq 0, \sum_{i=1}^m \mathbf{b}_i = 1\}$ el conjunto de todos los portafolios. El determinar el portafolio $\mathbf{b}^*$ que alcanza $W^*(F)$ es un problema de maximización de la función cóncava $W(\mathbf{b}, F)$ sobre el conjunto convexo $\mathcal{B}$. El máximo podría estar en la frontera. Presentaremos las condiciones de Kuhn-Tucker que caracterizan el máximo y cuya demostración usa el Teorema de Kuhn-Tucker del área de optimización.

Teorema 4.2. El portafolio log-óptimo $\mathbf{b}^*$ para un mercado de acciones $\mathbf{X}$, es decir, el portafolio que maximiza la razón de crecimiento $W\left(\mathbf{b}, F\right)$, satisface las siguientes condiciones necesarias y suficientes

$$\mathbb{E}\left(\frac{X_i}{\mathbf{b}^{*t}\mathbf{X}}\right) \begin{cases} = 1 & \text{si } b_i^* > 0 \\ \leq 1 & \text{si } b_i^* = 0. \end{cases}$$

Demostración. Para usar el Teorema de Kuhn-Tucker primero notemos lo siguiente:

$$b_i \geq 0$$

y

$$\sum_{i=1}^n b_i - 1 = 0.$$

Además

$$\frac{\partial W}{\partial b_i} = \mathbb{E}\frac{X_i}{\mathbf{b}^t\mathbf{X}},$$

pues $\mathbf{b}^t\mathbf{X} \leq \max\{X_i\}$ y X_i tiene esperanza finita.

Por el Teorema de Kuhn-Tucker, el portafolio $\mathbf{b}^*$ es óptimo si existen constantes $\mu_i \geq 0$ y $v \in \mathbb{R}$, tales que

$$-W'(\mathbf{b}^*) - \sum_{i=1}^n \mu_i \mathbf{e}_i + v * (1, 1, \ldots, 1) = 0,$$

donde $\mathbf{e}_i$ es el vector canónico que vale 1 en la i-ésima entrada y 0 en las otras entradas. Y también, $\mu_i \mathbf{b}^{*^t} \mathbf{e}_i = 0$.

Esto quiere decir que

$$\mathbb{E}\frac{X_j}{\mathbf{b}^{*^t}\mathbf{X}} + \mu_j - v = 0$$

y

$$\mu_j b_j^* = 0.$$

Sumando las ecuaciones anteriores y multiplicandolas por b_j^*, obtenemos que:

$$\mathbb{E}\frac{\mathbf{b}^{*^t}\mathbf{X}}{\mathbf{b}^{*^t}\mathbf{X}} + \sum_{j=1}^{n} \mu_j b_j^* - \sum_{j=1}^{n} v b_j^* = 0,$$

entonces

$$v = 1.$$

Ahora,

$$b_j^* > 0 \Rightarrow \mu_j = 0 \Rightarrow \mathbb{E}\frac{X_j}{\mathbf{b}^{*^t}\mathbf{X}} = 1,$$

y

$$b_j^* = 0 \Rightarrow \mu_j \geq 0 \Rightarrow \mathbb{E}\frac{X_j}{\mathbf{b}^{*^t}\mathbf{X}} \leq 1.$$

Que era lo que se quería demostrar.

∎

Éste teorema tiene unas consecuencias inmediatas.

Teorema 4.3. Sea $S^* = \mathbf{b}^{*t}\mathbf{X}$ la riqueza aleatoria que resulta del portafolio log-óptimo $\mathbf{b}^*$. Sea $S = \mathbf{b^t X}$ la riqueza que resulta al utilizar cualquier otro portafolio $\mathbf{b}$. Suponemos que S/S^* y $\ln \frac{S}{S^*}$ son variables aleatorias integrables. Entonces

$$\mathbb{E}\left(\frac{S}{S^*}\right) \leq 1.$$

También, si $\mathbb{E}\,(S/S^*) \leq 1$ para todos los portafolios $\mathbf{b}$, entonces $\mathbb{E}\,(\log S/S^*) \leq 0$ para todo $\mathbf{b}$.

Nota. Éste teorema puede ser enunciado, también, de la siguiente manera

$$\mathbb{E}\ln\frac{S}{S^*} \leq 0, \text{ para todo } S \Leftrightarrow \mathbb{E}\frac{S}{S^*} \leq 1, \text{ para todo } S.$$

Demostración. Usando el teorema anterior, se sigue que para un portafolio log-óptimo $\mathbf{b}^*$

$$\mathbb{E}\left(\frac{X_i}{\mathbf{b}^{*\mathbf{t}}\mathbf{X}}\right) \leq 1$$

para todo i. Multiplicando la ecuación por b_i y sumando sobre todos los i, se obtiene

$$\sum_{i=1}^{m} b_i \mathbb{E}\left(\frac{X_i}{\mathbf{b}^{*\mathbf{t}}\mathbf{X}}\right) \leq \sum_{i=1}^{m} b_i = 1$$

lo cual es equivalente a

$$\mathbb{E}\frac{\mathbf{b}^{\mathbf{t}}\mathbf{X}}{\mathbf{b}^{*\mathbf{t}}\mathbf{X}} = \mathbb{E}\frac{S}{S^*} \leq 1.$$

El regreso se sigue usando la desigualdad de Jensen, puesto que

$$\mathbb{E}\log\frac{S}{S^*} \leq \log\mathbb{E}\frac{S}{S^*} \leq \log 1 = 0.$$

∎

Nota. Considerar las acciones al final del primer día y la inicial distribución de la riqueza es $\mathbf{b}^*$. La proporción de la riqueza en la acción i-ésima al final del primer día es $\frac{b_i^* X_i}{\mathbf{b}^{*t}\mathbf{X}}$ y el valor esperado es

$$\mathbb{E}\left(\frac{b_i^* X_i}{\mathbf{b}^{*t}\mathbf{X}}\right) = b_i^* \mathbb{E}\left(\frac{X_i}{\mathbf{b}^{*t}\mathbf{X}}\right) = b_i^* \cdot 1 = b_i^*$$

entonces el valor esperado de la proporción de la riqueza en la acción i-ésima es la misma al final que al inicio del día.

Ahora consideremos una sucesión i.i.d. de mercado de acciones, es decir, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ son i.i.d. y su distribución es $F(\mathbf{x})$. Sea

$$S_n = \prod_{i=1}^{n} \mathbf{b}_i^t \mathbf{X}_i$$

la riqueza después de n días para un inversionista que usa el portafolio $\mathbf{b}_i$ en el día i. Sea,

$$W^* = \max_{\mathbf{b}} W(\mathbf{b}, F) = \max_b \mathbb{E}\log \mathbf{b}^t\mathbf{X}$$

la razón óptima de crecimiento y sea $\mathbf{b}^*$ el portafolio que alcanza dicha razón.

4.3. Optimalidad asintótica del portafolio log-óptimo

En lo que sigue sólo se permitirán portafolios que dependan del pasado y sean independientes del futuro (a menos que se indique lo contrario).

De la definición de W^*, se sigue que el portafolio log-óptimo maximiza el valor esperado del logaritmo de la riqueza final. Esto se establece en el siguiente lema.

Lema 4.3 Sea S_n^* la riqueza después de n días de un inversor que usa la estrategia log-óptima en un mercado de acciones i.i.d., y sea S_n la riqueza de cualquier otro inversionista usando una estrategia $\mathbf{b}_i$. Entonces

$$\mathbb{E}\log S_n^* = nW^* \geq \mathbb{E}\log S_n.$$

Demostración.

$$\begin{aligned}
\max_{\mathbf{b}_1,\ldots,\mathbf{b}_n} \mathbb{E}\log S_n &= \max_{\mathbf{b}_1,\ldots,\mathbf{b}_n} \mathbb{E}\sum_{i=1}^{n}\log \mathbf{b}_i^t\mathbf{X_i} \\
&= \sum_{i=1}^{n} \max_{\mathbf{b}_i(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1})} \mathbb{E}\log \mathbf{b}_i^t\left(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1}\right)\mathbf{X}_i \\
&= \sum_{i=1}^{n} \mathbb{E}\log \mathbf{b}^{*t}\mathbf{X}_i = nW^*,
\end{aligned}$$

y el máximo es alcanzado por el portafolio constantemente reequilibrado $\mathbf{b}^*$.

∎

Ahora probemos el siguiente resultado, que muestra que S_n^* es mayor que la riqueza de cualquier otro inversor para casi cualquier sucesión del mercado de acciones.

Teorema 4.4 Sean $\mathbf{X}_1,\ldots,\mathbf{X}_n$ una sucesión de vectores de acciones i.i.d. con distribución $F(\mathbf{x})$. Sea $S_n^* = \prod \mathbf{b}^{*t}\mathbf{X}_i$, donde $\mathbf{b}^*$ es el portafolio log-óptimo, y sea $S_n = \prod \mathbf{b}_i^t\mathbf{X}_i$ la riqueza que resulta de usar cualquier otro portafolio. Entonces

$$\limsup_{n\to\infty} \frac{1}{n}\log\frac{S_n}{S_n^*} \overset{c.s.}{\leq} 0.$$

Demostración. Por las condiciones de Kuhn-Tucker, se tiene que

$$\mathbb{E}\frac{S_n}{S_n^*} \leq 1.$$

Por la desigualdad de Markov, se tiene que

$$\mathbb{P}\left(S_n > t_n S_n^*\right) = \mathbb{P}\left(\frac{S_n}{S_n^*} > t_n\right) < \frac{1}{t_n}.$$

Entonces

$$\mathbb{P}\left(\frac{1}{n}\log\frac{S_n}{S_n^*} > \frac{1}{n}\log t_n\right) \leq \frac{1}{t_n}.$$

Tomando $t_n = n^2$ y sumando sobre todo n, se obtiene

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\frac{1}{n}\log\frac{S_n}{S_n^*} > \frac{2\log n}{n}\right) \leq \sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}.$$

Entonces, por el lema de Borel-Cantelli

$$\mathbb{P}\left(\frac{1}{n}\log\frac{S_n}{S_n^*} > \frac{2\log n}{n},\ \text{un número infinito de veces}\right) = 0.$$

Esto implica que para casi toda sucesión del mercado de acciones, existe una N tal que para todo $n > N, \frac{1}{n}\log\frac{S_n}{S_n^*} < \frac{2\log n}{n}$. Entonces

$$\limsup \frac{1}{n}\log\frac{S_n}{S_n^*} \overset{c.s.}{\leq} 0.$$

∎

Este teorema prueba que el portafolio log-óptimo será tan bueno o mejor que cualquier otro portafolio.

4.4. Información indirecta y la razón de crecimiento

El siguiente teorema establece una cota en el decrecimiento de la razón de crecimiento que ocurre cuando se considera la distribución incorrecta.

Teorema 4.5. Sean $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. que se distribuyen $f(\mathbf{x})$. Sea $\mathbf{b}_f^*$ el portafolio log-óptimo correspondientes a la densidad $f(\mathbf{x})$ y sea $\mathbf{b}_g^*$ el portafolio log-óptimo correspondiente para otra densidad $g(\mathbf{x})$. Suponemos que $f(\mathbf{x})\frac{\mathbf{b}_f^{*t}\mathbf{x}}{\mathbf{b}_g^{*t}\mathbf{x}}\frac{g(\mathbf{x})}{f(\mathbf{x})}$ y $f(\mathbf{x})\log\frac{\mathbf{b}_f^{*t}\mathbf{x}}{\mathbf{b}_g^{*t}\mathbf{x}}\frac{g(\mathbf{x})}{f(\mathbf{x})}$ son integrables. Entonces el incremento en la razón de crecimiento usando $\mathbf{b}_f^*$ en lugar de $\mathbf{b}_g^*$ está acotada por

$$\Delta W = W\left(\mathbf{b}_f^*, F\right) - W\left(\mathbf{b}_g^*, F\right) \leq D\left(f \parallel g\right).$$

Demostración. Se tiene que

$$\begin{aligned}
\Delta W &= \int f(\mathbf{x}) \log \mathbf{b}_f^{*t}\mathbf{x} - \int f(\mathbf{x}) \log \mathbf{b}_g^{*t}\mathbf{x} \\
&= \int f(\mathbf{x}) \log \frac{\mathbf{b}_f^{*t}\mathbf{x}}{\mathbf{b}_g^{*t}\mathbf{x}} \\
&= \int f(\mathbf{x}) \log \frac{\mathbf{b}_f^{*t}\mathbf{x}}{\mathbf{b}_g^{*t}\mathbf{x}} \frac{g(\mathbf{x})}{f(\mathbf{x})} \frac{f(\mathbf{x})}{g(\mathbf{x})} \\
&= \int f(\mathbf{x}) \log \frac{\mathbf{b}_f^{*t}\mathbf{x}}{\mathbf{b}_g^{*t}\mathbf{x}} \frac{g(\mathbf{x})}{f(\mathbf{x})} + D(f \parallel g) \\
&\leq \log \int f(\mathbf{x}) \frac{\mathbf{b}_f^{*t}\mathbf{x}}{\mathbf{b}_g^{*t}\mathbf{x}} \frac{g(\mathbf{x})}{f(\mathbf{x})} + D(f \parallel g) \qquad (4.1) \\
&= \log \int g(\mathbf{x}) \frac{\mathbf{b}_f^{*t}\mathbf{x}}{\mathbf{b}_g^{*t}\mathbf{x}} + D(f \parallel g) \\
&\leq \log 1 + D(f \parallel g) \qquad (4.2) \\
&= D(f \parallel g),
\end{aligned}$$

donde 4.1 se sigue de la desigualdad de Jensen; y 4.2 se sigue de las condiciones de Kuhn-Tucker y del hecho de que $\mathbf{b}_g^*$ es un portafolio log-óptimo para g.

∎

Teorema 4.6. El incremento ΔW en la razón de crecimiento debida a la información indirecta Y está acotada por,

$$\Delta W \leq I(\mathbf{X}; Y).$$

Demostración. Dada la información indirecta $Y = y$, el log-óptimo portafolio se obtiene usando la distribución $f(\mathbf{x} \mid Y = y)$. Entonces, condicionando en $Y = y$, se obtiene usando el teorema anterior que

$$\Delta W_{Y=y} \leq D(f(\mathbf{x} \mid Y = y) \parallel f(\mathbf{x})) = \int_{\mathbf{x}} f(\mathbf{x} \mid Y = y) \log \frac{f(\mathbf{x} \mid Y = y)}{f(\mathbf{x})} d\mathbf{x}.$$

Entonces

$$\begin{aligned}
\Delta W &= \int \Delta W_{Y=y} df_Y(y) \leq \int_y f(y) \int_{\mathbf{x}} f(\mathbf{x} \mid Y = y) \log \frac{f(\mathbf{x} \mid Y = y)}{f(\mathbf{x})} \frac{f(y)}{f(y)} d\mathbf{x} dy \\
&= \int_y \int_{\mathbf{x}} f(y) f(\mathbf{x} \mid Y = y) \log \frac{f(\mathbf{x} \mid Y = y)}{f(\mathbf{x})} \frac{f(y)}{f(y)} d\mathbf{x} dy \\
&= \int_y \int_{\mathbf{x}} f(\mathbf{x}, y) \log \frac{f(\mathbf{x}, y)}{f(\mathbf{x}) f(y)} d\mathbf{x} dy \\
&= I(\mathbf{X}; Y).
\end{aligned}$$

Por tanto, el incremento en el radio duplicado está acotado por arriba por la información conjunta entre la información indirecta Y y el mercado de acciones $\mathbf{X}$.

4.5. Inversiones en mercados estacionarios

En esta sección extenderemos los resultados anteriores a mercados que son dependientes del tiempo. Sea $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \ldots$ un proceso estocástico de vectores. Supondremos que $\mathbf{b}_i$ puede depender de $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}$. Sea

$$S_n = \mathbf{b}_1^t \mathbf{X}_1 * \prod_{i=2}^{n} \mathbf{b}_i^t \left(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}\right) \mathbf{X}_i.$$

El objetivo es maximizar $E \log S_n$ sobre todos los portafolios $\{\mathbf{b}_i(\cdot)\}$ dependientes del pasado. Ahora

$$\begin{aligned} \max_{\mathbf{b}_1,\ldots,\mathbf{b}_n} \mathbb{E} \log S_n &= \sum_{i=1}^{n} \max_{\mathbf{b}_i^t(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1})} \mathbb{E} \log \mathbf{b}_i^t \mathbf{X}_i \\ &= \sum_{i=1}^{n} \mathbb{E} \log \mathbf{b}_i^{*t} \mathbf{X}_i, \end{aligned}$$

donde $\mathbf{b}_i^*$ es el portafolio log-óptimo para la distribución condicional de $\mathbf{X}_i$ dado los valores pasados del mercado de acciones, es decir, $\mathbf{b}_i^*\left(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}\right)$ es el portafolio que alcanza el máximo condicional, denotado por

$$\begin{aligned} \max_{\mathbf{b}} \mathbb{E}\left[\log \mathbf{b}^t \mathbf{X}_i \mid (\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}) = (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1})\right] \\ = W^*\left(\mathbf{X}_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}\right). \quad \text{si} \quad i \geq 2. \end{aligned}$$

Tomando la esperanza, escribimos

$$W^*\left(\mathbf{X}_i \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}\right) = \mathbb{E}\left\{\max_b \mathbb{E}\left[\log \mathbf{b}^{*t} \mathbf{X}_i \mid (\mathbf{X}_1, \ldots, \mathbf{X}_{i-1})\right]\right\}$$

donde el máximo es tomado sobre todos los portafolios $\mathbf{b}$. Asumimos la existencia del máximo para todo $w \in \Omega$. Sea

$$W^*\left(\mathbf{X}_1, \ldots, \mathbf{X}_n\right) = \max_{\mathbf{b_1},\ldots,\mathbf{b_n}} \mathbb{E} \log S_n$$

donde el máximo es tomado sobre todos los portafolios. Como $\log S_n^* = \sum_{i=1}^{m} \log \mathbf{b}_i^{*t} \mathbf{X}_i$, entonces se tiene la siguiente regla de la cadena para W^*:

$$W^*\left(\mathbf{X}_1, \ldots, \mathbf{X}_n\right) = W^*\left(X_1\right) + \sum_{i=2}^{n} W^*\left(\mathbf{X}_i \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}\right).$$

Esta regla de la cadena es formalmente la misma que la regla de la cadena para H. De alguna forma, W es el dual de H. En particular, cuando se condiciona se reduce H pero se incrementa W.

Definición. La razón de crecimiento W_∞^* está definida por

$$W_\infty^* = \lim_{n\to\infty} \frac{W^*(\mathbf{X}_1, \ldots, \mathbf{X}_n)}{n}$$

si el límite existe y no está definida en otro caso.

Definiciones. Sea $\{\ldots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots\}$ un mercado estocástico estacionario.

(I) $\overline{\mathbb{F}}_s := \sigma(\mathbf{X}_{-1}, \ldots, \mathbf{X}_{-s})$.

(II) $\mathbb{F}_k := \sigma(\mathbf{X}_0, \ldots, \mathbf{X}_{k-1})$.

(III) $\overline{w}_t^* := \mathbb{E}\left(\log \mathbf{b}_t^* \cdot \mathbf{X}_0 \mid \overline{\mathbb{F}}_t\right)$.

(IV) $w_t^* := \sup_{b\in\mathbb{F}_t} \mathbb{E}(\log(b \cdot X_t) \mid \mathbb{F}_t)$. Aquí se evalúa la función en $w \in \Omega$ y luego se considera el supremo.

(V) $\overline{W}_t^* := \mathbb{E}(\bar{w}_t^*)$.

(VI) b_t^* es un portafolio log-óptimo que corresponde a $\overline{\mathbb{F}}_t$.

(VII) Sea $0 \le t < k$, entonces $\mathbb{F}_t^k := \begin{cases} \sigma(\boldsymbol{X}_0, \ldots, \boldsymbol{X}_{t-1}) & \text{si } 0 \le t < k \\ \sigma(\boldsymbol{X}_{t-k}, \ldots, \boldsymbol{X}_{t-1}) & \text{si } k \le t < \infty. \end{cases}$

(VIII) $\mathbb{F}_t^\infty := \sigma(\ldots, \boldsymbol{X}_{-1}, \boldsymbol{X}_0, \ldots, \boldsymbol{X}_{t-1})$.

(IX) b_t^k es un portafolio log-óptimo que corresponde a $\mathbb{F}_t^k$.

(X) b_t^∞ es un portafolio log-óptimo que corresponde a $\mathbb{F}_t^\infty$.

(XI) $S_n^k := \prod_{t=0}^{n-1} b_t^k \cdot \boldsymbol{X}_t$.

(XII) $S_n^\infty := \prod_{t=0}^{n-1} b_t^\infty \cdot \boldsymbol{X}_t$.

(XIII) $R := \text{máx}_b\, \mathbb{E}(\log b^t \cdot \boldsymbol{X}_0 \mid \boldsymbol{X}_{-1}, \boldsymbol{X}_{-2}, \ldots)$.

(XIV) $R_n := \text{máx}_b\, \mathbb{E}(\log b^t \cdot \boldsymbol{X}_n \mid \boldsymbol{X}_{n-1}, \boldsymbol{X}_{n-2}, \ldots)$.

(XV) $w^*(P) := \sup_b \mathbb{E}(\log b \cdot \mathbf{X})$, donde $\mathbf{X}$ es un vector aleatorio con distribución P.

(XVI) $w(b, P) := \mathbb{E}_p(\log b \cdot \mathbf{X})$.

(XVII) $B^*(Q) :=$ conjunto de portafolios log-óptimos cuando se usa la distribución Q.

Teorema 4.7. Para un mercado estacionario, la razón de crecimiento existe y es igual a

$$W_\infty^* = \lim_{n\to\infty} W^*(\mathbf{X}_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}).$$

Demostración. Demostremos que si $0 \leq t \leq s \leq \infty$, entonces

$$\overline{W}_t^* \leq \overline{W}_s^*.$$

Veamos que

$$\begin{aligned}
\mathbb{E}\left(\log \mathbf{b}_t^* X_0 \mid \overline{\mathbb{F}}_s\right) &\leq \overline{w}_s^* = \mathbb{E}\left(\log \mathbf{b}_s^* X_0 \mid \overline{\mathbb{F}}_s\right) \\
\Rightarrow \overline{w}_t^* = \mathbb{E}\left(\log \mathbf{b}_t^* X_0 \mid \overline{\mathbb{F}}_t\right) &\leq \mathbb{E}\left(w_s^* \mid \overline{\mathbb{F}}_t\right) = \text{ (pues } \overline{\mathbb{F}}_\mathfrak{t} \subset \overline{\mathbb{F}}_\mathfrak{s}) \\
\mathbb{E}\left(\mathbb{E}\left(\overline{w}_s^* \mid \overline{\mathbb{F}}_t\right) \mid \overline{\mathbb{F}}_s\right) &= \mathbb{E}\left(\mathbb{E}\left(\overline{w}_s^* \mid \overline{\mathbb{F}}_s\right) \mid \overline{\mathbb{F}}_t\right) = \\
\mathbb{E}\left(\overline{w}_s^* \mid \overline{\mathbb{F}}_\mathfrak{s}\right) &= \overline{W}_s^* \\
\Rightarrow \overline{W}_t^* &\leq \overline{W}_s^*.
\end{aligned}$$

Como estamos suponiendo que el mercado es estacionario, entonces $W^*\left(\mathbf{X}_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}\right)$ es no decreciente, y por tanto existe el límite cuando $n \to \infty$. Notemos que

$$\frac{W^*\left(\mathbf{X}_1, \ldots, \mathbf{X}_n\right)}{n} = \frac{1}{n} \sum_{i=1}^{n} W^*\left(\mathbf{X}_i \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}\right),$$

por el teorema de Césaro se sigue que

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W^*\left(\mathbf{X}_i \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}\right) &= \\
\lim_{n\to\infty} W^*\left(X_n \mid X_1, \ldots, X_{n-1}\right) &= W_\infty^*.
\end{aligned}$$

∎

A continuación se busca demostrar que para un proceso $\{X_n\}_{n\in\mathbb{Z}}$ estocástico, estacionario y ergódico se cumple que

$$\frac{1}{n} \log S_n^* \xrightarrow{c.s.} W^*.$$

Para esto demostraremos primero los siguientes tres teoremas.

Teorema 4.8. Sea S_n^* la riqueza que resulta de una serie de portafolios condicionales log-óptimos invertidos en un mercado estocástico estacionario $\{\mathbf{X}_i\}$. Sea S_n la riqueza que resulta de usar otros portafolios. Entonces S_n/S_n^* es una supermartiganla positiva con respecto a la sucesión de σ-álgebras generadas por el pasado $X_1, \ldots, X_n$. Entonces

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{S_n}{S_n^*} \leq 0,$$

y existe una variable aleatoria V tal que

$$\begin{aligned}
\frac{S_n}{S_n^*} &\xrightarrow{c.s.} V \\
EV &\leq 1
\end{aligned}$$

y

$$\mathbb{P}\left\{\limsup_n \frac{S_n}{S_n^*} \geq t\right\} \leq \frac{1}{t}.$$

Demostración. Tenemos que S_n/S_n^* es una supermartingala positiva porque

$$\begin{aligned}\mathbb{E}\left[\frac{S_{n+1}(X^{n+1})}{S_{n+1}^*(X^{n+1})} \mid X^n\right] &= \mathbb{E}\left[\frac{\left(\mathbf{b}_{n+1}^t\mathbf{X}_{n+1}\right)S_n\left(X^n\right)}{\left(\mathbf{b}_{n+1}^{*t}\mathbf{X}_{n+1}\right)S_n^*(X^n)} \mid X^n\right] \\ &= \frac{S_n(X^n)}{S_n^*(X^n)}\mathbb{E}\left[\frac{\mathbf{b}_{n+1}^t\mathbf{X}_{n+1}}{\mathbf{b}_{n+1}^{*t}\mathbf{X}_{n+1}} \mid X^n\right] \\ &\leq \frac{S_n(X^n)}{S_n^*(X^n)},\end{aligned}$$

por las condiciones de Kuhn-Tucker aplicadas al portafolio log-óptimo condicional. Entonces, por el teorema de convergencia de martingalas S_n/S_n^* tiene un límite, sea V dicho límite, y $\mathbb{E}V \leq \mathbb{E}\left(S_0/S_0^*\right) = 1$, y en particular $\mathbb{E}\left(S_n/S_n^*\right) \leq 1$, usando la prueba del caso del mercado i.i.d. se obtiene que

$$\limsup_{n\to\infty} \frac{1}{n}\log\frac{S_n}{S_n^*} \leq 0.$$

Finalmente, el resultado

$$\mathbb{P}\left\{\limsup_n \frac{S_n}{S_n^*} \geq t\right\} \leq \frac{1}{t}.$$

se sigue de la desigualdad de Kolmogorov para martingalas positivas.

∎

Teorema 4.9. w^*es una función semi-continua por abajo (respecto a la topología débil).

Demostración. Sean $\mathbf{X}$ un vector aleatorio con distribución P y β un portafolio tal que $\beta^i > 0$ para todo i. Consideremos el conjunto $\mathbb{U} = \left\{u = (u^j)_{1\leq j\leq m} \in \mathbb{R}_+^m : \beta\cdot u = 1\right\}$. Asimismo, sea $U : \Omega \to \mathbb{U}$ con $U = u(\mathbf{X})$ tal que $u(x) = \frac{x}{\beta\cdot x}$. Sea Q la distribución de U, y P la de $\mathbf{X}$, entonces como $\mathbf{X} = (\beta\cdot\mathbf{X})\mathbf{U}$, se sigue que

$$\begin{aligned}\mathbb{E}_p\left(\log(b\cdot\mathbf{X})\right) &= \mathbb{E}_P\left(\log\beta\cdot\mathbf{X}\right) + \mathbb{E}_p\left(\log b\cdot U(\mathbf{X})\right) \\ \Rightarrow w(b,P) &= r(P) + w(b,Q) \\ \Rightarrow w^*\left(P\right) &= r(P) + w^*(Q).\end{aligned}$$

Notemos que $w^*(Q) \geq 0$ pues $w(\beta,Q) \geq 0$. Sean $0 \leq \lambda \leq 1$ y b un portafolio; y $b_\lambda := (1-\lambda)\beta + \lambda b$, $\beta_\lambda = \{b_\lambda : b \in \beta\}$. Consideremos

$$w_\lambda^*\left(Q\right) = \sup_b \mathbb{E}_Q\left(\log b\cdot U\right) = \sup_b \mathbb{E}_Q\left(\log b_\lambda\cdot U\right),$$

es claro que w_λ^* es creciente en λ, pues si $\lambda \geq r$

$$\begin{aligned} b_\lambda \cdot u &= (1-\lambda) + \lambda(b \cdot u) \geq 1 + \lambda(b \cdot u - 1) \\ &\geq 1 + r\,(b \cdot u - 1) = b_r \cdot u. \end{aligned}$$

Además,

$$w^*(Q) = w_1^*(Q) \geq w_\lambda^*(Q) \geq w_0^*(Q) = 0.$$

Si $\lambda > 1$, $\log(b_\lambda \cdot u)$ está acotada por abajo por $\log(1-\lambda)$, y por tanto es semi-continua por abajo en u. Como $w(b_\lambda, Q) = \mathbb{E}_Q(\log b_\lambda \cdot U)$ es semi-continua por abajo , entonces $w_\lambda^*(Q) = \sup_b w(b_\lambda, Q)$ es semi-continua por abajo (dado que es el supremo de funciones semi-continuas por abajo).

Ahora

$$\begin{aligned} b_\lambda \cdot u &\geq \lambda(b \cdot u) \\ \Rightarrow w_\lambda^*(Q) &\leq w^*(Q) \leq w_\lambda^*(Q) - \log\lambda \\ \Rightarrow w_\lambda^*(Q) &\uparrow w^*(Q) \text{ cuando } \lambda \uparrow 1, \end{aligned}$$

entonces $w^*(Q)$ es el supremo de las funciones $w_\lambda^*(Q)$, que son semi-continuas, por ende $w^*(Q)$ es semi-continua por abajo.

∎

Teorema 4.10. Cuando $t \to \infty$ se tiene que $W_t^* \uparrow \overline{W}_\infty^*$.

Demostración. Usaremos la misma notación que la usada en el Teorema 4.9. Sea $\mathbf{X}$ un vector aleatorio con distribución P. Sean $u(x) = \frac{x}{x\cdot\beta}$ y $U = u(\mathbf{X})$. Sean Q,$\overline{Q}_t$ las distribuciones de U y de la probabilidad condicional de Udado $\overline{\mathbb{F}}_t$, respectivamente. Por el teorema de convergencia de martingalas de Lévy

$$w^*\left(\overline{Q}_t\right) = \mathbb{E}_P\left(\log b \cdot U \mid \overline{\mathbb{F}}_t\right) \underset{c.s.}{\to} \mathbb{E}_P\left(\log b \cdot U \mid \overline{\mathbb{F}}_\infty\right) = w^*\left(\overline{Q}_\infty\right),$$

ya que $\log b \cdot U \in L^1$ al ser una función acotada, dado que $\log b \cdot U \leq \text{máx}_j\,(-\log \beta^j)$ se obtiene usando que $b \cdot U \leq \sum_j U^j \leq \text{máx}\left(\frac{1}{\beta^j}\right)$ y que $\beta \cdot U = 1$.

Como w^* es semi-continua por abajo

$$\liminf_{t\to\infty} \overline{w}_t^* = \liminf_{t\to\infty} w^*\left(\overline{Q}_t\right) \overset{c.s.}{\geq} \overline{w}_\infty^* = w^*\left(\overline{Q}_\infty\right).$$

Como w^* está acotada por abajo por cero, entonces por el lema de Fatou

$$\liminf_{t\to\infty} \overline{W}_t^* = \liminf_{t\to\infty} E\left(\overline{w}_t^*\right) \geq \overline{W}_\infty^* = E\left(\overline{w}_\infty^*\right).$$

Asimismo, $\overline{W}_t^* \leq \overline{W}_\infty^*$ entonces

$$\lim_{t\to\infty} E\left(\overline{w}_t^*\right) = E\left(\overline{w}_\infty^*\right),$$

es decir que

$$\lim_{t\to\infty} \mathbb{E}\left(\sup_b \mathbb{E}\left(\log b\cdot U \mid \mathbf{X}_{-1},\ldots,\mathbf{X}_{-t}\right)\right) = \mathbb{E}\left(\sup_b \mathbb{E}\left(\log b\cdot U \mid \mathbf{X}_{-1},\ldots,\mathbf{X}_{-\infty}\right)\right).$$

Luego

$$\begin{aligned} \bar{w}_t^* &= r\left(\overline{P}_t\right) + w^*\left(\overline{Q}_t\right) \\ \Rightarrow \overline{W}_t^* &= \mathbb{E}\left(\log\beta\cdot\mathbf{X}\right) + \mathbb{E}\left(w^*\left(\overline{Q}_t\right)\right) \uparrow \\ &\quad \mathbb{E}\left(\log\beta\cdot\mathbf{X}\right) + \mathbb{E}\left(w^*\left(\overline{Q}_\infty\right)\right) \\ &= \overline{W}_\infty^*. \end{aligned}$$

∎

En la demostración del siguiente teorema se usará el teorema ergódico, que es una generalización de la ley fuerte de los grandes números. Un proceso ergódico está definido en un espacio de probabilidad $(\Omega,\mathbb{F},\mathbb{P})$, en donde se considera una variable aleatoria X. También se tiene una transformación $T:\Omega\to\Omega$ cuyo objetivo es avanzar o retroceder el tiempo. Diremos que la transformación es estacionaria si $\mathbb{P}(TA)=\mathbb{P}(A)$ para todo $A\in\mathbb{F}$. La transformación es ergódica si todo conjunto A tal que $TA=A$, satisface que $\mathbb{P}(A)=0$ o 1. Si T es estacionaria y ergódica, diremos que $X_n(w)=X(T^n w)$ es estacionario y ergódico. Para un proceso estacionario y ergódico con esperanza finita, el teorema ergódico de Birkhoff establece que

$$\frac{1}{n}\sum_{i=1}^{n} X_i(w) \xrightarrow{c.s.} EX.$$

Teorema 4.11. Sea $\{\ldots,\mathbf{X}_{-1},\mathbf{X}_0,\ldots,\mathbf{X}_n,\ldots\}$ un proceso estocástico estacionario y ergódico de vectores aleatorios. Sea S_n^* la riqueza al tiempo n usando la estrategia log-óptima, donde

$$S_n^* = \prod_{i=1}^{n} \mathbf{b}_i^{*t}\left(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1}\right)\mathbf{X}_i.$$

Entonces

$$\frac{1}{n}\log S_n^* \xrightarrow{c.s.} W^*.$$

Demostración. Notemos que

$$n^{-1}\log S_n^k = n^{-1}\log S_k^* + n^{-1}\sum_{t=k}^{n-1}\log\left(b_t^k\cdot\boldsymbol{X}_t\right).$$

Entonces, por el teorema ergódico

$$\lim_{n\to\infty} n^{-1}\log S_n^k \overset{c.s.}{=} W_k^* := \mathbb{E}\left(\log b_k^* \cdot \boldsymbol{X}_k\right).$$

De nuevo, por el teorema ergódico

$$\begin{aligned}\lim_{n\to\infty} n^{-1}\sum_{0\le t<n} \max_b \mathbb{E}\left(\log b^t\cdot \boldsymbol{X}_t \mid \boldsymbol{X}_{t-1},\boldsymbol{X}_{t-2},\ldots\right) &\overset{c.s.}{=}\\ \mathbb{E}\left(\max_b \mathbb{E}\left(\log b^t\cdot \boldsymbol{X}_0 \mid \boldsymbol{X}_{-1},\boldsymbol{X}_{-2},\ldots\right)\right) &= \overline{W}_\infty^*.\end{aligned}$$

Luego

$$\limsup_{n\to\infty}\frac{1}{n}\log\left(\frac{S_n^k}{S_n^*}\right)\le 0,$$

y

$$\limsup_{n\to\infty}\frac{1}{n}\log\left(\frac{S_n^*}{S_n^\infty}\right)\le 0,$$

pues $\mathbb{F}_t\subset\mathbb{F}_t^\infty$ y $\mathbb{F}_t^k\subset\mathbb{F}_t$, respectivamente. Entonces

$$\begin{aligned} w_k^* &= \lim_{n\to\infty} n^{-1}\log S_n^k \le \liminf_{n\to\infty} n^{-1}\log S_n^*\\ &\le \limsup_{n\to\infty} n^{-1}\log S_n^* \le \lim_{n\to\infty} n^{-1}\log S_n^\infty\\ &= \overline{W}_\infty^*.\end{aligned}$$

Luego, $W_k^* = \overline{W}_k^* \uparrow \overline{W}_\infty^* = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n W^*\left(\boldsymbol{X}_i \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_{i-1}\right)$.

∎

A continuación usando unas variables aleatorias se demostrará que la probabilidad de que S_n sea mayor que S_n^* es menor o igual que $\frac{1}{2}$.

Teorema 4.12. Sea S^* la riqueza al final de un periodo de inversiones en un mercado de acciones $\mathbf{X}$ usando un portafolio log-óptimo, y sea S la riqueza que produce cualquier otro portafolio. Sea U^*una variable aleatoria independiente de $\mathbf{X}$ y distribuida uniformemente en $[0,2]$, y sea V cualquier otra variable aleatoria independiente de $\mathbf{X}$ y U tal que $V\ge 0$ y $EV = 1$. Entonces

$$\mathbb{P}\left(VS\ge U^*S^*\right)\le\frac{1}{2}.$$

Demostración. Se tiene que

$$\begin{aligned}\mathbb{P}\left(VS\ge U^*S^*\right) &= \mathbb{P}\left(\frac{VS}{S^*}\ge U^*\right)\\ &= \mathbb{P}\left(W\ge U^*\right),\end{aligned}$$

donde $W = \frac{VS}{S^*}$ es una variable aleatoria no negativa con media

$$\mathbb{E}W = \mathbb{E}(V)\mathbb{E}\left(\frac{S_n}{S_n^*}\right) \leq 1,$$

usando la independencia de V y $\mathbf{X}$, y las condiciones de Kuhn-Tucker.

Sea F la función de distribución de W. Como U^* es uniforme en $[0,2]$

$$\begin{aligned}
\mathbb{P}(W \geq U^*) &= \int_0^2 \mathbb{P}(W > w)\, f_{U^*}(w)dw \\
&= \int_0^2 \mathbb{P}(W > w)\, \frac{1}{2}dw \\
&= \int_0^2 \frac{1-F(w)}{2}dw \\
&\leq \int_0^\infty \frac{1-F(w)}{2}dw \\
&= \frac{1}{2}EW \\
&\leq \frac{1}{2}.
\end{aligned}$$

Entonces

$$\mathbb{P}(VS \geq U^*S^*) = \mathbb{P}(W \geq U^*) \leq \frac{1}{2}.$$

∎

U y V^*serían "justas" aleatorizaciones de la riqueza. El efecto es aleatorizar pequeñas diferencias, de tal forma que sólo desviaciones significantes del radio $\frac{S}{S^*}$ afecten la probabilidad de ganar.

4.6. El teorema de Shannon-McMillan-Breiman

Notaciones.

- $\mathbb{P}\left(X_i \mid X_0^{i-1}\right) := \mathbb{P}(X_i \mid X_0, \ldots, X_{i-1})$.
- $\mathbb{P}^K\left(X_0^{n-1}\right) := \mathbb{P}\left(X_0^{k-1}\right) \prod_{i=k}^{n-1} \mathbb{P}\left(X_i \mid X_{i-k}^{i-1}\right)$ si $n \geq k$.
- $\mathbb{P}(X_0^n) := \mathbb{P}(X_0, \ldots, X_n)$.
- $H^k := \mathbb{E}\{-\log \mathbb{P}(X_0 \mid X_{-1}, \ldots, X_{-k})\}$.
- $H := \lim_{k\to\infty} H^k$.

A continuación demostraremos que para un proceso estacionario, ergódico y finito-valuado con razón de entropía H se cumple que

$$-\frac{1}{n}\log \mathbb{P}(X_0, \ldots, X_{n-1}) \xrightarrow{c.s.} H.$$

Primero demostremos los siguientes lemas.

Lema 4.4. Para un proceso estocástico, estacionario, ergódico y finito-valuado se tiene que con probabilidad 1

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log \frac{\mathbb{P}^k\left(X_0^{n-1}\right)}{\mathbb{P}\left(X_0^{n-1}\right)} &\leq 0, \\ \limsup_{n\to\infty} \frac{1}{n} \log \frac{\mathbb{P}\left(X_0^{n-1}\right)}{\mathbb{P}\left(X_0^{n-1} \mid X_{-\infty}^{-1}\right)} &\leq 0. \end{aligned}$$

Demostración. Sea A el soporte de $\mathbb{P}(X_0^{n-1})$. Entonces

$$\begin{aligned} \mathbb{E}\left(\frac{\mathbb{P}^k\left(X_0^{n-1}\right)}{\mathbb{P}\left(X_0^{n-1}\right)}\right) &= \sum_{x_0^{n-1}\in A} \mathbb{P}\left(x_0^{n-1}\right) \frac{\mathbb{P}^k(x_0^{n-1})}{\mathbb{P}\left(x_0^{n-1}\right)} \\ &= \sum_{x_0^{n-1}\in A} \mathbb{P}^k\left(X_0^{n-1}\right) = \mathbb{P}^k(A) \leq 1. \end{aligned} \tag{4.3}$$

Similarmente, sea $B\left(X_{-\infty}^{-1}\right)$ el soporte de $\mathbb{P}\left(\cdot \mid X_{-\infty}^{-1}\right)$. Entonces, se tiene que

$$\begin{aligned} \mathbb{E}\left\{\frac{\mathbb{P}(X_0^{n-1})}{\mathbb{P}\left(X_0^{n-1} \mid X_{-\infty}^{-1}\right)}\right\} &= \mathbb{E}\left[\mathbb{E}\left\{\frac{\mathbb{P}(X_0^{n-1})}{\mathbb{P}\left(X_0^{n-1} \mid X_{-\infty}^{-1}\right)} \mid X_{-\infty}^{-1}\right\}\right] \\ &= \mathbb{E}\left[\sum_{x^n\in B\left(X_{-\infty}^{-1}\right)} \frac{\mathbb{P}\left(x^n\right)}{\mathbb{P}\left(x^n \mid X_{-\infty}^{-1}\right)} \mathbb{P}\left(x^n \mid X_{-\infty}^{-1}\right)\right] \\ &= \mathbb{E}\left[\sum_{x^n\in B\left(X_{-\infty}^{-1}\right)} \mathbb{P}\left(x^n\right)\right] \leq 1. \end{aligned} \tag{4.4}$$

Por la desigualdad de Markov y 4.3, se tiene que

$$\mathbb{P}\left(\frac{\mathbb{P}^k\left(X_0^{n-1}\right)}{\mathbb{P}(X_0^{n-1})} \geq t_n\right) \leq \frac{1}{t_n}$$

o

$$\mathbb{P}\left(\frac{1}{n} \log \frac{\mathbb{P}^k\left(X_0^{n-1}\right)}{\mathbb{P}\left(X_0^{n-1}\right)} \geq \frac{1}{n} \log t_n\right) \leq \frac{1}{t_n}.$$

Poniendo $t_n = n^2$, vemos que por el lema de Borel-Cantelli que el evento

$$\left\{\frac{1}{n} \log \frac{\mathbb{P}^k\left(X_0^{n-1}\right)}{\mathbb{P}\left(X_0^{n-1}\right)} \geq \frac{1}{n} \log t_n\right\}$$

ocurre sólo un número finito de veces con probabilidad 1. Entonces

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{\mathbb{P}^k\left(X_0^{n-1}\right)}{\mathbb{P}\left(X_0^{n-1}\right)} \overset{c.s.}{\leq} 0.$$

Aplicando el mismo argumento usando 4.4 se obtiene que

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{\mathbb{P}\left(X_0^{n-1}\right)}{\mathbb{P}\left(X_0^{n-1} \mid X_{-\infty}^{-1}\right)} \overset{c.s.}{\leq} 0,$$

lo cual prueba el lema.

∎

Lema 4.5. Para un proceso estocástico, estacionario y ergódico $\{X_n\}$,

$$\begin{aligned} -\frac{1}{n} \log \mathbb{P}^k(X_0^{n-1}) &\overset{c.s.}{\longrightarrow} H^k, \\ -\frac{1}{n} \log \mathbb{P}\left(X_0^{n-1} \mid X_{-\infty}^{-1}\right) &\overset{c.s.}{\longrightarrow} H^\infty. \end{aligned}$$

Demostración. Las funciones $Y_n = f(X_{-\infty}^n)$ del proceso ergódico $\{X_n\}$ son procesos ergódicos. Entonces $\mathbb{P}\left(X_n \mid X_{n-k}^{n-1}\right)$ y $\log \mathbb{P}\left(X_n \mid X_{n-1}, X_{n-2}, \dots, \right)$ son también procesos ergódicos, y

$$\begin{aligned} -\frac{1}{n} \log \mathbb{P}^k\left(X_0^{n-1}\right) &= -\frac{1}{n} \log \mathbb{P}\left(X_0^{k-1}\right) - \frac{1}{n} \sum_{i=k}^{n-1} \log \mathbb{P}\left(X_i \mid X_{i-k}^{i-1}\right) \\ &\overset{c.s.}{\longrightarrow} H^k \end{aligned}$$

por el teorema ergódico. Similarmente, por el teorema ergódico,

$$\begin{aligned} -\frac{1}{n} \log \mathbb{P}\left(X_0^{n-1} \mid X_{-1}, X_{-2}, \dots\right) &= -\frac{1}{n} \sum_{i=k}^{n-1} \log \mathbb{P}\left(X_i \mid X_{i-k}^{i-1}, X_{-1}, X_{-2}, \dots\right) \\ &\overset{c.s.}{\longrightarrow} H^\infty. \end{aligned}$$

∎

Lema 4.6. Para un proceso estocástico, estacionario, finito-valuado y ergódico $\{X_n\}$, se cumple que

$$H^k \downarrow H^\infty$$

y

$$H = H^\infty.$$

Demostración. Por definición y por el hecho de que condicionando no se incrementa la entropía $H^k \downarrow H$. Sólo hace falta demostrar que $H^k \downarrow H^\infty$. El teorema de Levy de convergencia de martingalas para probabilidades condicionales establece que

$$\mathbb{P}\left(x_0 \mid X_{-k}^{-1}\right) \overset{c.s.}{\longrightarrow} \mathbb{P}\left(x_0 \mid X_{-\infty}^{-1}\right)$$

para todo $x_0 \in \mathbb{T}$ donde $\mathbb{T}$ es el rango de X_0. Como $\mathbb{T}$ es finito y $p \log p$ es acotada y continua en p para todo $0 \leq p \leq 1$, por el teorema de convergencia acotada,

$$\begin{aligned}
\lim_{k\to\infty} H^k &= \lim_{k\to\infty} \mathbb{E}\left\{ -\sum_{x_0\in\mathbb{T}} \mathbb{P}\left(x_0 \mid X_{-k}^{-1}\right) \log \mathbb{P}\left(x_0 \mid X_{-k}^{-1}\right) \right\} \\
&= \mathbb{E}\left\{ -\sum_{x_0\in\mathbb{T}} \mathbb{P}\left(x_0 \mid X_{-\infty}^{-1}\right) \log \mathbb{P}\left(x_0 \mid X_{-\infty}^{-1}\right) \right\} \\
&= \mathbb{E}\left(-\log \mathbb{P}\left(X_0 \mid X_{-\infty}^{-1}\right)\right) = H^\infty.
\end{aligned}$$

Entonces $H^k \downarrow H = H^\infty$.

∎

Estamos listos para probar el teorema AEP para procesos ergódicos.

Teorema 4.13. (El teorema de Shannon-McMillan-Breiman). Si H es el radio de entropía de un proceso estocástico, estacionario, finito-valuado y ergódico $\{X_n\}$, entonces

$$-\frac{1}{n} \log \mathbb{P}\left(X_0, \ldots, X_{n-1}\right) \xrightarrow{c.s.} H.$$

Demostración. Sabemos por el lema 4.4 que

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \frac{\mathbb{P}^k\left(X_0^{n-1}\right)}{\mathbb{P}(X_0^{n-1})} &\leq 0 \\
\Rightarrow \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}^k\left(X_0^{n-1}\right) &\leq -\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{\mathbb{P}\left(X_0^{n-1}\right)} \\
\Rightarrow \limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{\mathbb{P}\left(X_0^{n-1}\right)} &\leq \limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{\mathbb{P}\left(X_0^{n-1}\right)} = H^k
\end{aligned}$$

para $k = 1, 2, \ldots$. También, por el lema 4.4 se tiene que

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{\mathbb{P}\left(X_0^{n-1}\right)}{\mathbb{P}\left(X_0^{n-1} \mid X_{-\infty}^{-1}\right)} \leq 0,$$

lo cual se puede escribir también como

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mathbb{P}\left(X_0^{n-1}\right)} \geq \lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\mathbb{P}\left(X_0^{n-1} \mid X_{-\infty}^{-1}\right)} = H^\infty,$$

usando el lema 4.4. Por tanto

$$H^\infty \leq \liminf_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(X_0^{n-1}\right) \leq \limsup_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(X_0^{n-1}\right) \leq H^k \text{ para todo } k.$$

Pero, por el lema 4.6, $H^k \to H^\infty = H$. Entonces

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(X_0^n\right) = H.$$

∎

4.7. El portafolio universal

Hasta ahora la forma de encontrar el portafolio log-óptimo considera que se conoce la distribución de los vectores de acciones. Sin embargo, en la práctica a menudo no se conoce la distribución. Por lo que es conveniente no hacer suposiciones estadísticas acerca del mercado de acciones. En este sentido lo que se conoce como el portafolio universal nos indica qué tan bien se comportará un portafolio que seleccionamos sin hacer suposiciones acerca de la distribución del mercado de acciones en comparación con el mejor portafolio, y se busca seleccionar el portafolio que tenga un comportamiento muy similar al mejor. En esta dirección hay trabajos del mismo Cover y de otros autores (ver[10, 9, 6]). En términos generales, se supone que se tiene un mercado de acciones que puede ser representado como una sucesión de vectores $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \mathbb{R}_+^m$, donde x_{ij} es el precio relativo de la acción j en el día i y $\mathbf{x}_i$ es el vector de los precios relativos para todas las acciones en el día i. Primero se supone que se tienen n vectores $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Luego se extiende para el caso que se tiene un número infinito.

Dada esta sucesión de mercado de acciones, ¿qué es lo mejor que se puede hacer? Una estrategia pudieran ser los portafolio óptimos que se obtienen cuando se supone que tenemos un mercado de acciones i.i.d. con distribución conocida (a dichos portafolios se les llama portafolios óptimos constantemente reequilibrados), es por eso que esta estrategia resulta natural. Posteriormente se demuestra que es posible obtener un rendimiento tan bueno como cuando se toma un portafolio óptimo constantemente reequilibrado.

Una forma de analizar este problema es distribuir la riqueza entre un fondo de inversionistas, donde cada uno usa un portafolio óptimo constantemente reequilibrado distinto en base a una distribución distinta. Uno de los inversionistas lo hará exponencialmente mejor que los otros. Se muestra que por un factor de $n^{\frac{m-1}{2}}$ se puede alcanzar la riqueza que alcanza el mejor inversionista.

Otra forma de ver este problema es considerarlo como un juego en contra de un oponente malicioso que puede seleccionar la sucesión de vectores del mercado de acciones. Se define una estrategia de portafolio causal $\hat{\mathbf{b}}_i(\mathbf{x}_{i-1}, \ldots, \mathbf{x}_1)$ que depende sólo de los valores pasados del mercado de acciones. Entonces el oponente, conociendo la estrategia $\hat{\mathbf{b}}_i(\mathbf{x}^{i-1})$, escoge una sucesión de vectores $\mathbf{x}_i$ para hacer que la estrategia funcione tan mal como sea posible relativamente al portafolio óptimo constantemente reequilibrado para dicho mercado de acciones. Sea $\mathbf{b}^*(\mathbf{x}^n)$ un portafolio óptimo constantemente reequilibrado para un mercado de acciones $\mathbf{x}^n$. Notemos que $\mathbf{b}^*(\mathbf{x}^n)$ depende sólo de la distribución empírica de la sucesión, y no del orden en que los vectores ocurren. Al final de n días, un portafolio constantemente reequilibrado $\mathbf{b}$ genera la riqueza:

$$S_n(\mathbf{b}, \mathbf{x}^n) = \prod_{i=1}^{n} \mathbf{b}^t \mathbf{x}_i,$$

y el portafolio óptimo constante reequilibrado $\mathbf{b}^*(\mathbf{x}^n)$ genera la riqueza

$$S_n^*(\mathbf{x}^n) = \max_{\mathbf{b}} \prod_{i=1}^{n} \mathbf{b}^t \mathbf{x}_i,$$

mientras que la estrategia que usa el portafolio $\hat{\mathbf{b}}_i(\mathbf{x}^{i-1})$ genera la riqueza

$$\hat{S}_n(\mathbf{x}^n) = \prod_{i=1}^{n} \hat{\mathbf{b}}_i^t\left(\mathbf{x}^{i-1}\right)\mathbf{x}_i.$$

El objetivo es encontrar una estrategia de portafolio causal

$$\hat{\mathbf{b}}(\cdot) = \left(\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2(\mathbf{x}_1), \ldots, \hat{\mathbf{b}}_i\left(\mathbf{x}^{i-1}\right)\right)$$

que genere la mayor riqueza posible aún en el peor caso en términos de la razón de $\hat{S}_n$ yS_n^*. Se encuentra la estrategia universal óptima y se muestra que esa estrategia para cada mercado de acciones genera la riqueza $\hat{S}_n$ que sólo difiere de un factor $V_n \approx n^{-\frac{m-1}{2}}$ de la riqueza S_n^*. De hecho, la estrategia que se encuentra es:

$$\begin{aligned}
\hat{\mathbf{b}}_1 &= \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right), \\
\hat{\mathbf{b}}_{k+1} &= \frac{\int_\beta \mathbf{b} S_k(\mathbf{b})\, d\mathbf{b}}{\int_\beta S_k(\mathbf{b}) d\mathbf{b}}.
\end{aligned}$$

El portafolio anterior se puede evaluar en base a simulaciones. Si no se tienen bases de datos con acciones reales, se puede generar un mercado de acciones usando un proceso de difusión-saltos (ver [29]).

Capítulo 5

Teoría de la información en redes

En este capítulo se considera la teoría de la información en redes, que es el estudio de los flujos que se pueden alcanzar de manera simultánea en presencia de ruido e interferencia. El objetivo de esta teoría es lograr el flujo óptimo de información en las redes. Cabe mencionar que la teoría del fluido en redes también ha contribuido a responder preguntas en otras áreas como teoría de circuitos y el fluido del agua en las tuberías.

La lectura de este capítulo requiere de los preliminares que se presentaron en el capítulo 1: la entropía, entropía relativa, información mutua, la desigualdad del procesamiento de datos y de Fano. Estas dos desigualdades se utilizan en las demostraciones de las tasas alcanzables para el canal de acceso múltiple y el de emisión, los cuales se presentan en los teoremas 5.3.1 y 5.4.2. Además, introduciremos el concepto de sucesiones conjuntamente típicias que surge en el contexo de la decodificación en los canales, y su teoría se presenta en la sección 5.2.

Demostraremos algunos resultados básicos para demostrar teoremas de los canales de usuarios múltiples y el canal de emisión. Hay varios problemas abiertos en esta área, y todavía no existe una teoría exhaustiva de las redes de información. Inclusive si dicha teoría fuera encontrada, es muy probable que sería muy difícil poder implementarla. Sin embargo, la teoría indicaría a los diseñadores de comunicación qué tan cerca están del modelo óptimo y tal vez sugiera medios para mejorar las tasas de comunicación.

5.1. Introducción

El problema general es el siguiente. Dados mucho remitentes y receptores, y una matriz que es un canal de transición que describe los efectos de la interferencia y el ruido en la red, decidir si las fuentes pueden ser transmitidas por el canal. Este problema involucra la compresión de información y encontrar el conjunto de todas las tasas que se pueden alcanzar (la región de capacidad de la red). Este problema no ha sido todavía resuelto, por lo que en este capítulo consideraremos varios casos especiales. Sin embargo, existe el enfoque de los sistemas MIMO (múltiples-entradas y múltiples-salidas), que usan múltiples antenas en el remitente y receptor, donde

se usan herramientas de la teoría de matrices aleatorias; que evita el problema de no tener una teoría de las tasas simulatáneas.

Ejemplos de redes de comunicación grandes incluyen redes de computadoras, redes de satélites, y el sistema telefónico. Incluso en una sola computadora, hay varios componentes que se comunican entre ellos. Una teoría completa de las redes de información tendría implicaciones en el diseño de redes de computadoras y de comunicación.

Supongamos que m estaciones buscan comunicarse con un satélite a través de un canal de comunicación. A este sistema se le llama *canal de múltiple acceso.* Algunas preguntas que surgen de manera natural y que tienen respuesta son las siguientes. Cómo los remitentes cooperan entre ellos para enviar la información. Qué tasas de comunicación (bits por transmisión) se alcanzan. Cuáles son las limitaciones que la interferencia produce entre los remitentes en la tasa total de comunicación.

En contraste, podemos considerar una red y considerar una estación de televisión que envía información a m televisores. En éste canal sólo se conocen respuestas a las siguientes preguntas en casos especiale:. cómo el remitente codifica la información para los distintos receptores en una sola señal y cuáles son las tasas a las que la información puede ser enviada a diferentes receptores.

Hay otros canales, como el canal de relevos (donde hay una fuente y un destino, pero hay uno o más remitentes que son intermediarios-los recibidores que fungen como relevos sirven para facilitar la comunicación entre la fuente y el receptor), el canal de interferencia (hay dos remitentes y dos receptores con interferencia), y el canal bilateral (hay dos pares de remitente-receptor que se envían información entre ellos). Para todos estos canales, sólo se conoce las respuestas de preguntas acerca de las tasas de comunicación alcanzables y las estrategia de codificación apropiadas.

Estos canales pueden ser considerados como casos especiales de una red general de comunicación que consiste de m nodos que se intentan comunicar entre ellos. A cada instante de tiempo, el i-ésimo nodo manda el símbolo x_i que depende del mensaje que busca mandar y de los símbolos pasados recibidos en el nodo. La transmisión simultánea de los símbolos $(x_1, \ldots, x_m)$ resulta en los símbolos aleatorios recibidos $(Y_1, \ldots, Y_m)$ que se distribuyen con la distribución de probabilidad condicional $\mathbb{P}\left(y^{(1)}, y^{(2)}, \ldots, y^{(m)} \mid x^{(1)}, \ldots, x^{(m)}\right)$. Aquí $\mathbb{P}(\cdot \mid \cdot)$ expresa los efectos del ruido y la interferencia que están presentes en la red. Si $\mathbb{P}(\cdot \mid \cdot)$ toma sólo los valores 0 y 1, la red es determinista.

5.2. Canales gaussianos de usuarios múltiples

A continuación consideraremos ejemplos gaussianos de algunos canales básicos de la teoría de la información en redes. La motivación física de los canales gaussianos radica en las respuestas concretas y de interpretación sencilla. En esta sección se dan sin demostración las ideas clave para establecer las regiones de capacidad de los canales gaussianos de acceso múltiple, emisión, relevos, y bilateral.

El canal discreto, básico, gaussiano y aditivo con ruido blanco con potencia de

entrada P y varianza de ruido N está modelado por

$$Y_i = X_i + Z_i,\ i = 1, 2, \ldots,$$

donde las Z_i son variables aleatorias gaussianas i.i.d. con media 0 y varianza N. La señal $\mathbf{X} = (X_1, \ldots, X_n)$ tiene una restricción de potencia

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq P.$$

La capacidad de Shannon C (el supremo sobre todas las tasas alcanzables) se obtiene al maximizar $I(X;Y)$ sobre todas las variables aleatorias X tales que $EX^2 \leq P$ y está dada por

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \text{ bits por transmisión.}$$

En este capítulo nos restringiremos a los canales discretos; los resultados pueden ser extendidos a los canales gaussianos continuos.

Comenzaremos con unas definiciones.

Definición. Un *canal discreto,* denotado por $(\mathcal{X}, p(y \mid x), \mathcal{Y})$, consiste de dos conjuntos finitos $\mathcal{X}$ y $\mathcal{Y}$ y una colección de funciones de distribución de probabilidad $p(y \mid x)$, una por cada $x \in \mathcal{X}$, tal que para toda x y y, $p(y \mid x) \geq 0$, y para todo x, $\sum_y p(y \mid x) = 1$, con la interpretación de que X es la entrada y Y es la salida del canal.

Definición. Un código (M, n) para un canal gaussiano con restricción de potencia P consiste de lo siguiente:

(I) Un conjunto de índices $\{1, 2, \ldots, M\}$.

(II) Una función codificadora $x : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, que produce las claves $x^n(1), x^n(2), \ldots, x^n(M)$, que satisfacen la restricción de potencia, es decir, para cada clave

$$\sum_{i=1}^{n} x_i^2(w) \leq nP,\ w = 1, 2, \ldots, M.$$

(III) Una función decodificadora

$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}.$$

Definición. La probabilidad condicional de error dado que el índice i fue enviado está definida como

$$\lambda_i = \mathbb{P}\left(g\left(Y^n\right) \neq i \mid X^n = X^n(i)\right) = \sum_{y^n} p(y^n \mid x^n(i)) 1_{\{g(y^n) \neq i\}}.$$

Definición. La máxima probabilidad de error $\lambda^{(n)}$ para un código (M, n) está definida como

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i.$$

Definición. La tasa de un código (M, n) es

$$R = \frac{\log M}{n} \text{ bits por transmisión.}$$

Definición. Una tasa R es alcanzable para un canal gaussiano con restricción de potencia P si existe una sucesión de $\left(2^{nR}, n\right)$ códigos cuyas claves satisfacen la restricción de potencia tal que la máxima probabilidad de error $\lambda^{(n)}$ tiende a cero.

5.2.1. Canal gaussiano de un sólo usuario

En este caso $Y = X + Z$. Seleccionar una tasa $R < \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$. Seleccionar un buen libro de claves $(2^{nR}, n)$ de potencia P (un buen libro de claves es un código tal que tiene una probabilidad máxima de error pequeña). Seleccionar un índice w en el conjunto 2^{nR}. Mandar la palabra $\mathbf{X}(w)$ del libro de claves. El receptor observa $\mathbf{Y} = \mathbf{X}(w) + \mathbf{Z}$ y entonces encuentra el índice $\hat{w}$ del libro de claves que más se parece a $\mathbf{Y}$. Si n es suficientemente grande, la probabilidad de error $\mathbb{P}(w \neq \hat{w})$ será arbitrariamente pequeña.

5.2.2. Canal gaussiano de acceso múltiple con m usuarios

Consideremos m transmisores, cada uno con potencia P. Sea

$$Y = \sum_{i=1}^{m} X_i + Z.$$

Sea

$$C\left(\frac{P}{N}\right) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

la capacidad de un canal gaussiano de un sólo usuario con una razón de señal sobre ruido P/N. La región de tasas alcanzables para el canal gaussiano viene dada por las siguientes ecuaciones

$$\begin{aligned} R_i &< C\left(\frac{P}{N}\right) \\ R_i + R_j &< C\left(\frac{2P}{N}\right) \\ R_i + R_j + R_k &< C\left(\frac{3P}{N}\right) \\ &\vdots \\ \sum_{i=1}^{m} R_i &< C\left(\frac{mP}{N}\right). \end{aligned}$$

Aquí se necesitan m libros de claves, el i-ésimo libro de claves tiene 2^{nR_i} claves de potencia P. Cada uno de los transmisores independientes escogen un código arbitrario de su propio libro de claves. Los usuarios mandan esos vectores simultáneamente. El receptor ve dichos códigos sumados con el ruido gaussiano $\mathbf{Z}$.

La forma óptima de decodificar consiste en buscar los m códigos, uno de cada libro de códigos, tal que la suma vectorial es lo más parecida a $\mathbf{Y}$ en términos de la distancia euclideana. Si $(R_1, \ldots, R_m)$ está en la región de capacidad de arriba, la probabilidad de error tiende a 0 cuando n tiende a infinito.

Nota. Es interesante notar que en este problema la suma de las tasas de los usuarios $C\left(mP/N\right)$ tiende a infinito cuando m tiende a infinito. Entonces, en un sistema con m usuarios de potencia P en presencia del ruido del ambiente N, el que está escuchando recibe una cantidad de información no acotada cuando el número de usuarios crece a infinito. Aparentemente, el incrementar la interferencia cuando el número de emisores $m \to \infty$ no limita la cantidad total de información recibida.

5.2.3. Canal gaussiano de emisión

Supondremos que tenemos un remitente con potencia P y dos receptores distantes, uno con potencia de ruido gaussiano N_1 y el otro con potencia de ruido gaussiano N_2. Sin pérdida de generalidad, supondremos que $N_1 < N_2$. Entonces, el receptor Y_1 recibe menos ruido que el receptor Y_2. El modelo para el canal es $Y_1 = X + Z_1$ y $Y_2 = X + Z_2$, donde Z_1 y Z_2 son variables aleatorias gaussianas arbitrariamente correlacionadas con varianzas N_1 yN_2, respectivamente. El remitente desea enviar mensajes independientes con tasas R_1 y R_2 a los receptores Y_1 y Y_2, respectivamente.

Luego se encuentra que la región de capacidad del canal gaussiano de emisión es

$$\begin{aligned} R_1 &< C\left(\frac{\alpha P}{N_1}\right) \\ R_2 &< C\left(\frac{(1-\alpha)P}{\alpha P + N_2}\right), \end{aligned}$$

donde α puede ser escogido arbitrariamente ($0 \leq \alpha \leq 1$) para equilibrar las tasas R_1 y R_2 como el remitente desee.

Para codificar los mensajes, el transmisor crea dos libros de códigos, uno con potencia αP y tasa R_1, y otro libro de códigos con potencia $\bar{\alpha}P$ y tasa R_2, donde R_1 y R_2 están en la región de capacidad de arriba. Entonces para enviar un índice $w_1 \in \{1, 2, \ldots, 2^{nR_1}\}$ y $w_2 \in \{1, 2, \ldots, 2^{nR_2}\}$ a Y_1 y Y_2, respectivamente, el transmisor toma el código $X(w_1)$ del primer libro de códigos y el código $X(w_2)$ del segundo libro de códigos y realiza la suma. Él envía la suma por el canal.

Los receptores deben decodificar el mensaje. Primero consideremos al peor receptor Y_2. Él solamente considera el segundo libro de códigos para encontrar el código que más se parezca al vector recibido $\mathbf{Y}_2$. Su ratio señal-ruido es $\bar{\alpha}P/(\alpha P + N_2)$, puesto que el mensaje de Y_1 actúa como ruido para Y_2. (Esto puede ser probado).

El mejor receptor Y_1 primero decodifica el código de Y_2, acción que puede realizar debido a que tiene menos ruido N_1. Él resta este código $\hat{\mathbf{X}}_2$ a $\mathbf{Y}_1$. Luego él busca el código que más se parece a $\mathbf{Y}_1 - \hat{\mathbf{X}}_2$ en el primer libro de códigos. La probabilidad de error puede ser tan pequeña como se desee.

5.2.4. Canal gaussiano de relevos

Para el canal de relevos, tenemos un remitente X y un último receptor Y. También está presente el canal de relevos, cuya intención es únicamente ayudar al receptor. El canal gaussiano de relevos está dado por

$$\begin{aligned} Y_1 &= X + Z_1 \\ Y &= X + Z_1 + X_1 + Z_2, \end{aligned}$$

donde Z_1 y Z_2 son variables aleatorias gaussianas e independientes con media cero y varianzas N_1 y N_2, respectivamente. La codificación permitida por el relevo es la sucesión

$$X_{1i} = f_i\left(Y_{11}, Y_{12}, \ldots, Y_{1i-1}\right).$$

El remitente X tiene potencia P y el remitente X_1 tiene potencia P_1. La capacidad es

$$C = \max_{0 \leq \alpha \leq 1} \min \left\{ C\left(\frac{P + P_1 + 2\sqrt{\bar{\alpha}PP_1}}{N_1 + N_2}\right), C\left(\frac{\alpha P}{N_1}\right) \right\},$$

donde $\bar{\alpha} = 1 - \alpha$. Notemos que si

$$\frac{P_1}{N_2} \geq \frac{P}{N_1},$$

se puede ver que $C = C\left(P/N_1\right)$, que es alcanzada por $\alpha = 1$. El canal pareciera no tener ruido después del relevo, y la capacidad $C\left(P/N_1\right)$ de X al relevo se puede alcanzar. Entonces, la tasa $C\left(P/\left(N_1 + N_2\right)\right)$ sin el relevo se incrementa por la presencia del relevo a $C(P/N_1)$. Cuando N_2 es muy grande y $P_1/N_2 \geq P/N_1$, se ve que el incremento en la tasa es de $C\left(P/\left(N_1 + N_2\right)\right) \approx 0$ a $C\left(P/N_1\right)$.

El proceso de codificación y decodificación involucra dos bloques.

Sea $R_1 < C\left(\alpha P/N_1\right)$. Se crean dos libros de códigos. El primer libro de códigos tiene 2^{nR_1} códigos de potencia αP. El segundo tiene 2^{nR_0} códigos de potencia $(1-\alpha)P$. Usaremos códigos de esos libros de códigos sucesivamente para crear la oportunidad de cooperación del relevo. En el primer bloque, empezamos enviando un código del primer libro de códigos. El relevo conoce el índice de este código puesto que $R_1 < C\left(\alpha P/N_1\right)$, pero el receptor final tiene una lista de posibles códigos de tamaño $2^{n(R_1-C(\alpha P/N_1+N_2))}$ (este cálculo de la lista involucra un resultado de listas de códigos). Por lo que el receptor no decodifica dicho código.

En el siguiente bloque, el remitente y el relevo cooperarán para resolver la incertidumbre del receptor acerca del código enviado que está en la lista del receptor. Desafortunadamente, estos no pueden estar seguros de qué lista es ésta pues ellos no conocen la señal recibida Y. Entonces, al azar ellos hacen una partición del primer libro de códigos en 2^{nR_0} celdas con un número igual de códigos en cada celda. El relevo, el receptor, y el remitente están de acuerdo con esta partición. El relevo y el remitente encuentran la celda de la partición en la que el código del primer libro de códigos está y cooperativamente envían el código del segundo libro de códigos con ese índice. Es decir, X y X_1 envían el mismo código designado. El relevo debe escalar dicho código para que el código pueda satisfacer la restricción de potencia P_1. Luego, ellos envían sus códigos simultáneamente.

El remitente también escoge un nuevo código del primer libro de códigos, lo suma al código que se envío de forma cooperativa del segundo libro de códigos, y envía la suma por el canal.

La recepción por el último receptor Y en el segundo bloque involucra primero encontrar el índice del segundo libro de códigos más parecido al código que se envió de forma cooperativa. Luego resta el código de la secuencia recibida y calcula una lista de índices de tamaño 2^{nR_0} que corresponden a todos los códigos del primer libro de códigos que probablemente fueron enviados en el segundo bloque.

Luego el receptor final debe identificar el código del primer libro de códigos que se envió en el primer bloque. Él usa la lista de posibles códigos que pudieron haber sido enviados en el primer bloque y los interseca con la celda de la partición que ha recibido del mensaje enviado en el segundo bloque. Las tasas y potencias han sido elegidas para que sea altamente probable que haya sólo un código en la intersección.

En pocas palabras. En cada nuevo bloque, el remitente y el relevo cooperan para resolver la lista incierta del bloque anterior. En adición, el remitente manda la suma de información nueva de su primer libro de códigos con la enviada del segundo libro de códigos.

5.2.5. Canal gaussiano de interferencia

El canal de interferencia tiene dos remitentes y dos receptores. El remitente 1 desea enviar información al receptor 1. A él no le importa lo que el receptor 2 recibe o entiende; similarmente con el remitente 2 y el receptor 2. Cada canal interfiere con el otro. No es un canal de emisión porque hay sólo un receptor intencional por cada remitente, tampoco es un canal de múltiple acceso porque cada receptor está sólo interesado en lo que va a ser enviado por el correspondiente remitente. El modelo

usado es:

$$\begin{aligned} Y_1 &= X_1 + aX_2 + Z_1 \\ Y_2 &= X_2 + aX_1 + Z_2, \end{aligned}$$

donde Z_1, Z_2 son variables aleatorias gaussianas independientes con media cero y varianza N. Este canal no ha sido resuelto en general incluso en el caso gaussiano. Pero cabe decir que en el caso de mucha interferencia, se puede mostrar que la región de capacidad de este canal es la misma que si se supusiera que no hay interferencia.

Para obtener lo anterior, se generan dos libros de códigos, cada uno con potencia P y tasa $C\left(P/N\right)$. Cada remitente escoge independientemente una palabra de su libro y la manda. Ahora, si la interferencia a satisface que $C\left(a^2P/(P+N)\right) > C(P/N)$, el primer receptor entiende perfectamente el índice del segundo remitente. Él lo encuentra usando la técnica usual de buscar el código que más se parece a la señal recibida. Una vez que encuentra esta señal, él la resta de la señal recibida. Por lo que habrá un canal sin ruido entre él y su remitente. Él entonces busca en el libro de códigos del remitente para encontrar el código que más se parece a la señal.

5.2.6. Canal gaussiano bilateral

El canal bilateral es muy similar al canal de interferencia, con la característica adicional de que el remitente 1 está relacionado con el receptor 2 y el remitente 2 está relacionado con el receptor 1, de tal forma que el remitente 1 puede usar información de símbolos recibidos anteriormente por el receptor 2 para decidir qué enviar después; y similarmente para el remitente 2 y el receptor 1. Este canal introduce otro aspecto fundamental de la teoría de la información en redes: la retroalimentación. La retroalimentación le permite a los remitentes usar información parcial que cada uno tiene acerca del mensaje del otro para cooperar entre ellos.

La región de capacidad del canal bilateral no se conoce en general. Este canal fue concebido primero por Shannon, quien derivo cotas superiores e inferiores para la región. Para canales gaussianos, esas dos cotas coinciden y la región de capacidad es conocida; de hecho, el canal gaussiano bilateral se descompone en dos canales independientes.

Sean P_1 y P_2 las potencias para los remitentes 1 y 2, respectivamente, y sean N_1 y N_2 las varianzas del ruido de los dos canales. Entonces las tasas $R_1 < C(P_1/N_1)$ y $R_2 < C(P_2/N_2)$ pueden ser alcanzadas por las técnicas descritas para el canal de interferencia. En este caso, se generan dos libros de códigos de tasas R_1 y R_2. El remitente 1 envía un código del primer libro de códigos. El receptor 2 recibe la suma de los códigos enviados por los dos remitentes más algo de ruido. Él simplemente resta el código del remitente 2 y obtiene un canal sin ruido del remitente 1 (con sólo el ruido de la varianza N_1). Entonces, el canal gaussiano bilateral se descompone en dos canales gaussianos independientes. Pero este no es el caso en general para los canales bilaterales; en general, existe un intercambio entre los dos remitentes de tal forma que no pueden los dos enviar el mensaje a la tasa óptima al mismo tiempo.

5.3. Sucesiones conjuntamente típicas

En esta sección estudiaremos el teorema conjunto AEP que nos permitirá demostrar algunos teoremas de la teoría de información en redes. Dicho teorema nos permitirá calcular la probabilidad de error para sucesiones conjuntamente típicas que han sido decodificadas para varios de los esquemas de codificación que se consideran en este capítulo. Intuitivamente, dos sucesiones son conjuntamente típicas si la entropía empírica y teórica son parecidas.

Sea $(X_1, \ldots, X_k)$ una colección finita de variables aleatorias discretas con una distribución conjunta fija, $p\left(x^{(1)}, x^{(2)}, \ldots, x^{(k)}\right)$, $\left(x^{(1)}, \ldots, x^{(k)}\right) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$. Sea S un subconjunto ordenado de esas variables aleatorias y consideremos n copias independientes de S. Entonces

$$\mathbb{P}(S = s) = \prod_{i=1}^{n} \mathbb{P}(S_i = s_i), \ s \in \mathcal{S}^n.$$

Por ejemplo, si $S = (X_j, X_l)$, entonces

$$\begin{aligned} \mathbb{P}\left(S = s\right) &= \mathbb{P}\left((\mathbf{X}_j, \mathbf{X}_l) = (\mathbf{x}_j, \mathbf{x}_l)\right) \\ &= \prod_{i=1}^{n} p(x_{ij}, x_{il}). \end{aligned}$$

Por la ley de los grandes números, para cualquier subconjunto S de variables aleatorias

$$-\frac{1}{n} \log p\left(S_1, S_2, \ldots, S_n\right) = -\frac{1}{n} \sum_{i=1}^{n} \log p(S_i) \xrightarrow{c.s.} H(S),$$

$S \subset \left\{X^{(1)}, \ldots, X^{(k)}\right\}$.

Definición. El conjunto $A_\epsilon^{(n)}$ de las n-sucesiones ϵ-típicas $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ está definido por

$$A_\epsilon^{(n)}\left(X^{(1)}, \ldots, X^{(k)}\right) = A_\epsilon^{(n)}$$
$$= \left\{(\mathbf{x}_1, \ldots, \mathbf{x}_k) : \left|-\frac{1}{n} \log p(\mathbf{s}) - H(S)\right| < \epsilon, \forall S \subset \left\{X^{(1)}, \ldots, X^{(k)}\right\}\right\}.$$

Sea $A_\epsilon^{(n)}(S)$ la restricción de $A_\epsilon^{(n)}$ a las coordinadas de S. Entonces, si $S = (X_1, X_2)$, tenemos que

$$\begin{aligned} A_\epsilon^{(n)}(X_1, X_2) = \{&(\mathbf{x}_1, \mathbf{x}_2) : \\ &\left|-\frac{1}{n} \log p\left(\mathbf{x}_1, \mathbf{x}_2\right) - H(X_1, X_2)\right| < \epsilon, \\ &\left|-\frac{1}{n} \log p(\mathbf{x}_1) - H(X_1)\right| < \epsilon, \\ &\left|-\frac{1}{n} \log p(\mathbf{x}_2) - H(X_2)\right| < \epsilon\}. \end{aligned}$$

Definición. Usaremos la notación $a_n \doteq 2^{n(b\pm\epsilon)}$ para indicar que

$$\left|\frac{1}{n}\log a_n - b\right| < \epsilon$$

para n suficientemente grande.

Teorema 5.2.1 Para todo $\epsilon > 0$, y para n suficientemente grande

(I) $\mathbb{P}(A_\epsilon^{(n)}(S)) \geq 1-\epsilon$, $\forall S \subset \{X^{(1)}, \ldots, X^{(k)}\}$.

(II) $\mathbf{s} \in A_\epsilon^{(n)}(S) \Longrightarrow p(\mathbf{s}) \doteq 2^{n(H(S)\pm\epsilon)}$.

(III) $\left|A_\epsilon^{(n)}(S)\right| \doteq 2^{n(H(S)\pm 2\epsilon)}$.

(IV) Sean $S_1, S_2 \subset \left\{X^{(1)}, \ldots, X^{(k)}\right\}$. Si $(\mathbf{s}_1, \mathbf{s}_2) \in A_\epsilon^{(n)}(S_1, S_2)$, entonces

$$p(\mathbf{s}_1 \mid \mathbf{s}_2) \doteq 2^{-n(H(S_1|S_2)\pm 2\epsilon)}.$$

Demostración.

(I) Se sigue de la ley de los grandes números para variables aleatorias en la definición de $A_\epsilon^{(n)}(S)$.

(II) Se sigue directamente de la definición de $A_\epsilon^{(n)}(S)$.

(III) Se sigue de lo siguiente

$$\begin{aligned} 1 &\geq \sum_{\mathbf{s}\in A_\epsilon^{(n)}(S)} p(\mathbf{s}) \\ &\geq \sum_{\mathbf{s}\in A_\epsilon^{(n)}(S)} 2^{-n(H(S)+\epsilon)} \\ &= \left|A_\epsilon^{(n)}(S)\right| 2^{-n(H(S)+\epsilon)}. \end{aligned}$$

Si n es suficientemente grande, se tiene que

$$\begin{aligned} 1-\epsilon &\leq \sum_{\mathbf{s}\in A_\epsilon^{(n)}(S)} p(\mathbf{s}) \\ &\leq \sum_{\mathbf{s}\in A_\epsilon^{(n)}(S)} 2^{-n(H(S)-\epsilon)} \\ &= \left|A_\epsilon^{(n)}(S)\right| 2^{-n(H(S)-\epsilon)}. \end{aligned}$$

De donde, se obtiene que $\left|A_\epsilon^{(n)}(S)\right| \doteq 2^{n(H(S)\pm 2\epsilon)}$ para n suficientemente grande.

(IV) Para $(\mathbf{s}_1, \mathbf{s}_2) \in A_\epsilon^{(n)}(S_1, S_2)$, se tiene que $p(\mathbf{s}_1) \doteq 2^{-n(H(S_1)\pm\epsilon)}$ y $p(\mathbf{s}_1, \mathbf{s}_2) \doteq 2^{-n(H(S_1,S_2)\pm\epsilon)}$. Entonces

$$p(\mathbf{s}_2 \mid \mathbf{s}_1) = \frac{p(\mathbf{s}_1, \mathbf{s}_2)}{p(\mathbf{s}_1)} \doteq 2^{-n(H(S_2|S_1)\pm 2\epsilon)}.$$

∎

El siguiente teorema acota el número de sucesiones condicionalmente típicas para una sucesión típica dada.

Teorema 5.2.2 Sean S_1, S_2 dos subconjuntos de $X^{(1)}, \ldots, X^{(k)}$. Para todo $\epsilon > 0$, definimos $A_\epsilon^{(n)}(S_1 \mid \mathbf{s}_2)$ como el conjunto de las sucesiones $\mathbf{s}_1$ tales que son conjuntamente $\epsilon-$típicas a una sucesión particular $\mathbf{s}_2$. Si $\mathbf{s}_2 \in A_\epsilon^{(n)}(S_2)$, entonces para n suficientemente grande, se tiene que

$$\left|A_\epsilon^{(n)}(S_1 \mid \mathbf{s}_2)\right| \leq 2^{n(H(S_1|S_2)+2\epsilon)}$$

y

$$(1-\epsilon)\, 2^{n(H(S_1|S_2)-2\epsilon)} \leq \sum_{\mathbf{s}_2} p(\mathbf{s}_2) \left|A_\epsilon^{(n)}(S_1 \mid \mathbf{s}_2)\right|.$$

Demostración. Como en la parte 3 del teorema 5.2.1, se tiene que

$$\begin{aligned} 1 &\geq \sum_{\mathbf{s}_1 \in A_\epsilon^{(n)}(S_1|\mathbf{s}_2)} p(\mathbf{s}_1 \mid \mathbf{s}_2) \\ &\geq \sum_{\mathbf{s}_1 \in A_\epsilon^{(n)}(S_1|\mathbf{s}_2)} 2^{-n(H(S_1|S_2)+2\epsilon)} \\ &= \left|A_\epsilon^{(n)}(\mathbf{S}_1 \mid \mathbf{s}_2)\right| 2^{-n(H(S_1|S_2)+2\epsilon)}. \end{aligned}$$

Si n es suficientemente grande, entonces por la propiedad 1 del teorema 5.2.1

$$\begin{aligned} 1-\epsilon &\leq \sum_{\mathbf{s}_2} p(\mathbf{s}_2) \sum_{\mathbf{s}_1 \in A_\epsilon^{(n)}(S_1|\mathbf{s}_2)} p(\mathbf{s}_1 \mid \mathbf{s}_2) \\ &\leq \sum_{\mathbf{s}_2} p(\mathbf{s}_2) \sum_{\mathbf{s}_1 \in A_\epsilon^{(n)}(S_1|\mathbf{s}_2)} 2^{-n(H(S_1|S_2)-2\epsilon)} \\ &= \sum_{\mathbf{s}_2} p(\mathbf{s}_2) \left|A_\epsilon^{(n)}(S_1 \mid \mathbf{s}_2)\right|. \end{aligned}$$

∎

Para calcular la probabilidad de error de decodificación, se necesita saber la probabilidad de que sucesiones que son condicionalmente independientes sean típicamente conjuntas. Sean S_1, S_2 y S_3 tres subconjuntos de $\{X^{(1)}, \ldots, X^{(k)}\}$. Si S_1' y S_2' son condicionalmente independientes dado S_3' pero además tienen las mismas marginales dos a dos de (S_1, S_2, S_3), se tiene la siguiente probabilidad del evento de ser típicamente conjuntas.

Teorema 5.2.3 Sea $A_\epsilon^{(n)}$ el conjunto típico y sea

$$\mathbb{P}(\mathbf{S}_1' = \mathbf{s}_1, \mathbf{S}_2' = \mathbf{s}_2, \mathbf{S}_3' = \mathbf{s}_3) = \prod_{i=1}^{n} p(s_{1i} \mid s_{3i}) p(s_{2i} \mid s_{3i}) p(s_{3i}).$$

Entonces

$$\mathbb{P}\left\{(\mathbf{S}_1', \mathbf{S}_2', \mathbf{S}_3') \in A_\epsilon^{(n)}\right\} \doteq 2^{n(I(S_1;S_2|S_3)\pm 6\epsilon)}.$$

Demostración. Usamos la notación $\doteq$ para evitar calcular la cota inferior y superior por separado. Se tiene que

$$\begin{aligned}
&\mathbb{P}\left\{(\mathbf{S}_1', \mathbf{S}_2', \mathbf{S}_3') \in A_\epsilon^{(n)}\right\} \\
&= \textstyle\sum_{(\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3)\in A_\epsilon^{(n)}} p(\mathbf{s}_3) p(\mathbf{s}_1 \mid \mathbf{s}_3) p(\mathbf{s}_2 \mid \mathbf{s}_3) \\
&\doteq \left|A_\epsilon^{(n)}(S_1, S_2, S_3)\right| 2^{-n(H(S_3)\pm\epsilon)} 2^{-n(H(S_1|S_3)\pm 2\epsilon)} 2^{-n(H(S_2|S_3)\pm 2\epsilon)} \\
&\doteq 2^{n(H(S_1,S_2,S_3)\pm\epsilon)} 2^{-n(H(S_3)\pm\epsilon)} 2^{-n(H(S_1|S_3)\pm 2\epsilon)} 2^{-n(H(S_2|S_3)\pm 2\epsilon)} \\
&\doteq 2^{-n(I(S_1;S_2|S_3)\pm 6\epsilon)}.
\end{aligned}$$

∎

Este teorema se usará para elecciones particulares de S_1, S_2 y S_3 en varias pruebas de este capítulo.

5.4. Canal de acceso múltiple

En este canal hay dos o más remitentes que envían información a un receptor común. Ejemplos de este canal son un satélite receptor con muchas estaciones terrestres independientes, o un conjunto de celulares que se comunican con una estación. En este canal los remitentes deben competir contra el ruido y la interferencia entre cada uno de los receptores.

Definición. Un canal discreto de acceso múltiple y sin memoria consiste de tres alfabetos, $\mathcal{X}_1, \mathcal{X}_2$, y $\mathcal{Y}$, y una matriz de probabilidades de transición $p(y \mid x_1, x_2)$.

Definición. Un código $\left(\left(2^{nR_1}, 2^{nR_2}\right), n\right)$ para una canal de acceso múltiple consiste de dos conjuntos de enteros $\mathcal{W}_1 = \{1, 2, \ldots, 2^{nR_1}\}$ y $\mathcal{W}_2 = \{1, 2, \ldots, 2^{nR_2}\}$, llamados los conjuntos de mensajes, de dos funciones codificadoras

$$X_1 : \mathcal{W}_1 \to \mathcal{X}_1^n$$

y

$$X_2 : \mathcal{W}_2 \to \mathcal{X}_2^n,$$

y una función decodificadora

$$g : \mathcal{Y}^n \to \mathcal{W}_1 \times \mathcal{W}_2.$$

Hay dos remitentes y un receptor para este canal. El remitente 1 escoge un índice W_1 de manera uniforme del conjunto $\left\{1, 2, \ldots, 2^{nR_1}\right\}$ y manda el código correspondiente por el canal. El remitente 2 hace lo mismo. Suponiendo que la distribución de los mensajes en el producto $\mathcal{W}_1 \times \mathcal{W}_2$ es uniforme (es decir, los mensajes son independientes y tienen la misma probabilidad), se define la probabilidad promedio de error para el código $\left(\left(2^{nR_1}, 2^{nR_2}\right), n\right)$ como sigue

$$P_e^{(n)} = \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1,w_2)\in \mathcal{W}_1\times\mathcal{W}_2} \mathbb{P}(g(Y^n) \neq (w_1, w_2) \mid (w_1, w_2) \text{ fue enviado}).$$

Usaremos esta notación $\lambda_{(w_1,w_2)} = \mathbb{P}(g(Y^n) \neq (w_1, w_2) \mid (w_1, w_2)$ fue enviado).

Definición. Una tasa par (R_1, R_2) se dice que es alcanzable para el canal de acceso múltiple si existe una sucesión de libros de códigos $\left(\left(2^{nR_1}, 2^{nR_2}\right), n\right)$ tal que $P_e^{(n)} \to 0$.

Definición. La región de capacidad para un canal de acceso múltiple es la cerradura del conjunto de todas las tasas pares $(R_1.R_2)$ que son alcanzables.

Ahora estableceremos la región de capacidad en el siguiente teorema llamado teorema de tasas alcanzables. Dicho teorema es el más importante de la sección 5.3., esencialmente todos los otros resultados de la sección se centran en demostrar este teorema. Primero probaremos la ida, en la sección 5.3.1 probaremos el regreso.

Teorema 5.3.1 La capacidad de un canal de acceso múltiple $(\mathcal{X}_1 \times \mathcal{X}_2, p(y \mid x_1, x_2), \mathcal{Y})$ es la cerradura de la envolvente convexa de todos los (R_1, R_2) que cumplen

$$\begin{aligned} R_1 &< I(X_1; Y \mid X_2), \\ R_2 &< I(X_2; Y \mid X_1), \\ R_1 + R_2 &< I(X_1, X_2 \mid Y) \end{aligned}$$

para una distribución producto $p_1(x_1)p_2(x_2)$ en $\mathcal{X}_1 \times \mathcal{X}_2$.

Demostración. Fijemos $p(x_1, x_2) = p_1(x_1)p_2(x_2)$.

Generación del libro de códigos : Generemos 2^{nR_1} códigos independientes $\mathbf{X}_1(i)$, $i \in \{1, 2, \ldots, 2^{nR_1}\}$, de tamaño n, cada elemento generado i.i.d.$\sim \prod_{i=1}^n p_1(x_{1i})$. Similarmente, generamos 2^{nR_2} códigos independientes $\mathbf{X}_2(j), j \in \{1, 2, \ldots, 2^{nR_2}\}$, cada elemento generado i.i.d.$\sim \prod_{i=1}^n p_2(x_{2i})$. Estos códigos forman el libro de códigos, que es relevado a los remitentes y el receptor.

Codificación: Para enviar el índice i, el remitente 1 envía el código $\mathbf{X}_1(i)$. Similarmente, para enviar j, el remitente 2 envía el código $\mathbf{X}_2(j)$.

decodificación: Sea $A_\epsilon^{(n)}$ el conjunto de las sucesiones $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ típicas. El receptor Y^n escoge el par (i, j) tal que

$$(\mathbf{x}_1(i), \mathbf{x}_2(j), \mathbf{y}) \in A_\epsilon^{(n)}$$

si dicho par (i, j) existe y es único; de otra forma, se dice que hubo un error.

Análisis de la probabilidad de error : Notemos que la probabilidad de generar un libro de códigos cualquiera C es

$$\mathbb{P}(C) = \left(\prod_{w=1}^{2^{nR_1}} \prod_{i=1}^{n} p_1(x_{1i}(w)) \right) \left(\prod_{w=1}^{2^{nR_2}} \prod_{i=1}^{n} p_2(x_{2i}(w)) \right).$$

(Se debe tener n fijo para que lo anterior pueda ser una probabilidad.)

Sea $W = (W_1, W_2)$ con distribución uniforme en $\{1, \ldots, 2^{nR_1}\} \times \{1, 2, \ldots, 2^{nR_2}\}$ y usaremos $\hat{W}(\mathbf{y})$ para denotar la decodificación del mensaje. Sea $E = \{\hat{W}(\mathbf{y}) \neq W\}$ el evento de tener error. Calculemos el promedio de la probabilidad de error, promediada sobre todos los códigos en el libro, y promediada sobre todos los libros de códigos

$$\begin{aligned}
\mathbb{P}(E) &= \sum_C \mathbb{P}(C) \mathbb{P}_e^{(n)}(C) \\
&= \sum_C \mathbb{P}(C) \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1,w_2) \in \mathcal{W}_1 \times \mathcal{W}_2} \lambda_{(w_1,w_2)}(C) \\
&= \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1,w_2) \in \mathcal{W}_1 \times \mathcal{W}_2} \sum_C \mathbb{P}(C) \lambda_{(w_1,w_2)}(C).
\end{aligned}$$

Por la simetría de la construcción del código, el promedio de la probabilidad de error promediada sobre todos los libros de códigos no depende en el índice que haya sido enviado (i.e., $\sum_C \mathbb{P}(C)\lambda_{(w_1,w_2)}(C)$ no depende de (w_1, w_2)). Entonces, podemos asumir sin pérdida de generalidad que el mensaje $W = (1, 1)$ fue enviado, pues

$$\begin{aligned}
\mathbb{P}(E) &= \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1,w_2) \in \mathcal{W}_1 \times \mathcal{W}_2} \sum_C \mathbb{P}(C) \lambda_{(w_1,w_2)}(C) \\
&= \sum_C \mathbb{P}(C) \lambda_{(1,1)}(C) \\
&= \mathbb{P}(E \mid W = (1, 1)).
\end{aligned}$$

Definamos los siguientes eventos

$$E_{ij} = \{(\mathbf{X}_1(i), \mathbf{X}_2(j), \mathbf{Y}) \in A_\epsilon^{(n)}\}.$$

Recordemos que $\mathbf{Y}$ es el resultado de enviar el índice $(1, 1)$ por el canal.

Entonces un error ocurre si cuando se decodifica ocurre E_{11}^c (cuando la palabra código y la secuencia recibida no son típicamente conjuntas) o ocurre $E_2 \bigcup E_3 \bigcup \cdots \bigcup E_{2^{nR}}$

(cuando un código incorrecto es típicamente conjunto con la sucesión recibida). Entonces, denotando a $\mathbb{P}(E \mid W = (1,1))$ por $\mathbb{P}(E)$, tenemos que

$$\begin{aligned}\mathbb{P}(E \mid W = (1,1)) &= \mathbb{P}\left(E_{11}^c \bigcup_{(i,j)\neq(1,1)} E_{(i,j)}\right) \\ &\leq \mathbb{P}(E_{11}^c) + \sum_{i\neq 1, j=1} \mathbb{P}(E_{i1}) + \sum_{i=1, j\neq 1} \mathbb{P}(E_{1j}) \\ &\quad + \sum_{i\neq 1, j\neq 1} \mathbb{P}(E_{ij}),\end{aligned}$$

donde $\mathbb{P}$ denota la probabilidad condicional dado que $(1,1)$ fue enviado. Por el teorema AEP (ver [8]) , $\mathbb{P}(E_{11}^c) \to 0$. Por los teoremas 5.2.1 y 5.2.3, para $i \neq 1$, tenemos que

$$\begin{aligned}\mathbb{P}(E_{i1}) &= \mathbb{P}((\mathbf{X}_1(i), \mathbf{X}_2(1), \mathbf{Y}) \in A_\epsilon^{(n)}) \\ &= \sum_{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) \in A_\epsilon^{(n)}} p(\mathbf{x}_1) p(\mathbf{x}_2, \mathbf{y}) \\ &\leq \left|A_\epsilon^{(n)}\right| 2^{-n(H(X_1)-\epsilon)} 2^{-n(H(X_2,Y)-\epsilon)} \\ &\leq 2^{-n(H(X_1)+H(X_2,Y)-H(X_1,X_2,Y)-3\epsilon)} \\ &= 2^{-n(I(X_1;X_2,Y)-3\epsilon)} \qquad (5.1) \\ &= 2^{-n(I(X_1;Y|X_2)-3\epsilon)}, \qquad (5.2)\end{aligned}$$

donde (5.1) y (5.2) se siguen de la independencia de X_1 y X_2, y de $I(X_1; X_2, Y) = I(X_1; X_2) + I(X_1; Y \mid X_2) = I(X_1; Y \mid X_2)$. Similarmente, para $j \neq 1$,

$$\mathbb{P}(E_{1j}) \leq 2^{-n(I(X_2;Y|X_1)-3\epsilon)},$$

y para $i \neq 1, j \neq 1$,

$$\mathbb{P}(E_{ij}) \leq 2^{-n(I(X_1,X_2;Y)-4\epsilon)}.$$

Se sigue que

$$\begin{aligned}\mathbb{P}(E \mid W = (1,1)) \leq \mathbb{P}(E_{11}^c) &+ 2^{nR_1} 2^{-n(I(X_1;Y|X_2)-3\epsilon)} + 2^{nR_2} 2^{-n(I(X_2;Y|X_1)-3\epsilon)} \\ &+ 2^{n(R_1+R_2)} 2^{-n(I(X_1,X_2;Y)-4\epsilon)}.\end{aligned}$$

Dado que $\epsilon > 0$ es arbitrario, las hipótesis del teorema implican que cada término tiende a 0 cuando $n \to \infty$. Entonces la probabilidad promedio de error, promediada sobre todos los libros de códigos y códigos, tiende a cero cuando las desigualdades del teorema se cumplen.

Para finalizar la demostración, consideremos lo siguiente. Como la probabilidad promedio de error sobre todos los libros de códigos es pequeña, existe al menos un libro de códigos C^* con una probabilidad promedio de error pequeña. Entonces $\mathbb{P}(E \mid C^*) \leq \epsilon$. La determinación de C^* se hace a través de una búsqueda sobre todos los códigos $\left(2^{nR_1}, n\right)$ y $\left(2^{nR_2}, n\right)$. Notemos que

$$\mathbb{P}(E \mid C^*) = \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1,w_2)\in \mathcal{W}_1 \times \mathcal{W}_2} \lambda_{(w_1,w_2)}(C^*),$$

pues (W_1, W_2) tiene distribución uniforme. Así se concluye la ida de la demostración para una distribución fija.

Ahora consideremos un ejemplo de canales de múltiple acceso.

Ejemplo. (Canales independientes, binarios y simétricos). Un canal binario y simétrico, es un canal con entrada binaria y salida binaria, y error de probabilidad p, es decir $p(0 \mid 1) = 1-p, p(0 \mid 1) = p, p(1 \mid 0) = p, p(1 \mid 1) = 1-p$. Asumamos que tenemos dos canales independientes, binarios y simétricos, uno para el remitente 1 y otro para el remitente 2.

Analicemos el primer canal. Acotemos la información mutua por

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) \\ &= H(Y) - \sum p(x)H(Y \mid X = x) \\ &= H(Y) - \sum p(x)H(p) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p), \end{aligned}$$

la última desigualdad se sigue del hecho de que Yes una variable aleatoria binaria. Y se alcanza la igualdad cuando la distribución de X es la uniforme. Entonces, podemos enviar datos a una tasa de $1 - H(p_1)$ por el primer canal y a una tasa de $1 - H(p_2)$ por el segundo canal. Dado que los canales son independientes, no hay interferencia entre los remitentes.

5.4.1. Convexidad de la región de capacidad para el canal de acceso múltiple

Ahora se introducirá una nueva variable aleatoria para considerar la región de capacidad desde otro enfoque. Empezaremos probando que la región de capacidad es convexa.

Teorema 5.3.2 La región de capacidad C de un canal de acceso múltiple es convexa.

Demostración. Consideremos dos sucesiones de códigos, una alcanzando la tasa (R_{11}, R_{21}) y la otra alcanzando la tasa (R_{12}, R_{22}). Para cada bloque de longitud n, asumamos sin pérdida de generalidad que αn es un entero. Consideremos los códigos $(2^{\alpha n R_{11}}, 2^{\alpha n R_{21}}, \alpha n)$ y $(2^{(1-\alpha) n R_{12}}, 2^{(1-\alpha) n R_{22}}, (1-\alpha)n)$, de la primera y segunda sucesión de códigos, respectivamente.

Para enviar el mensaje (M_1, M_2), dividiremos el mensaje. Representamos el mensaje M_1 por los mensajes independientes M_{11} a la tasa αR_{11} y M_{12} a la tasa $(1-\alpha)M_2$. Similarmente, representamos el mensaje M_2 por los mensajes independientes M_{21} a la tasa αR_{21} y M_{22} a la tasa $(1-\alpha)R_{22}$. Entonces, $R_1 = \alpha R_{11} + (1-\alpha)R_{12}$ y $R_2 = \alpha R_{21} + (1\text{-}\alpha)R_{22}$.

Para las primeras αn transmisiones, el remitente $j = 1, 2$ envía su código para M_{j1} del código $(2^{\alpha n R_{11}}, 2^{\alpha n R_{21}}, \alpha n)$ y para el resto de las transmisiones, transmite el código para M_{j2} del código $(2^{(1-\alpha) n R_{12}}, 2^{(1-\alpha) n R_{22}}, (1-\alpha)n)$. Cuando se recibe y^n, el receptor decodifica $y^{\alpha n}$ usando el decodificador del primer código y $y_{\alpha n+1}^n$ usando el decodificador del segundo código.

Por hipótesis, la probabilidad de error de cada decodificador tiende a cero cuando $n \to \infty$. Entonces, la probabilidad de error tiende a cero para el nuevo código y la tasa (R_1, R_2) es alcanzable.

∎

Ahora probaremos un resultado que involucra una propiedad de conjuntos convexos definido por desigualdades lineales. En particular, queremos demostrar que la envolvente convexa de dos conjuntos convexos definidos por restricciones lineales es la región definida por la combinación convexa de las restricciones.

La región de capacidad para una distribución fija $p(x_1)p(x_2)$ está definida por tres informaciones mutuas, $I(X_1; Y \mid X_2), I(X_2; Y \mid X_1)$, y $I(X_1, X_2; Y)$, a las que llamaremos I_1, I_2, y I_3, respectivamente. Para cada $p(x_1)p(x_2)$, hay un vector correspondiente, $\mathbf{I} = (I_1, I_2, I_3)$, y una región de tasas definida por

$$C_{\mathbf{I}} = \{(R_1, R_2) : R_1 \geq 0, R_2 \geq 0, R_1 \leq I_1, R_2 \leq I_2, R_1 + R_2 \leq I_3\}. \tag{5.3}$$

También, para cualquier distribución $p(x_1)p(x_2)$, tenemos que $I(X_2; Y \mid X_1) = H(X_2 \mid X_1) - H(X_2 \mid Y, X_1) = H(X_2) - H(X_2 \mid Y, X_1) = I(X_2; Y, X_1)$ que a su vez es igual a $I(X_2; Y) + I(X_2; X_1 \mid Y) \geq I(X_2; Y)$, y entonces, $I(X_1; Y \mid X_2) + I(X_2; Y \mid X_1) \geq I(X_1; Y \mid X_2) + I(X_2; Y) = I(X_1, X_2; Y)$, tenemos que para todos los vectores $\mathbf{I}$ que $I_1 + I_2 \geq I_3$.

Lema 5.3.1 Sean $\mathbf{I}_1, \mathbf{I}_2 \in \mathbb{R}^3$ dos vectores de información mutua que definen las regiones de tasas C_1 y C_2, respectivamente, como están definidas en el párrafo anterior. Para $0 \leq \lambda \leq 1$, definamos $\mathbf{I}_\lambda = \lambda \mathbf{I}_1 + (1-\lambda)\mathbf{I}_2$, y sea C_3 la región de tasas definida por $\mathbf{I}_\lambda$. Entonces

$$C_3 = \lambda C_1 + (1-\lambda) C_2.$$

Demostración. Demostraremos este teorema en dos partes. Primero mostraremos que cualquier punto en $\lambda C_1 + (1-\lambda)C_2$ satisface la restricción $\mathbf{I}_\lambda$. Pero esto se sigue directamente, pues cualquier punto en C_1 satisface las desigualdades para $\mathbf{I}_1$ y cualquier punto en C_2 satisface las desigualdades para $\mathbf{I}_2$, entonces la combinación $(\lambda, 1-\lambda)$ de esos puntos satisface la combinación $(\lambda, 1-\lambda)$ de dichas restricciones. Entonces, se sigue que

$$\lambda C_1 + (1-\lambda) C_2 \subset C_3.$$

Para probar la otra inclusión, consideremos los puntos extremos de las regiones C_i $i = 1, 2, 3$. No es difícil ver que las regiones de las tasas definidas en (5.3) están siempre en la forma de pentágono, o en el caso extremo cuando $I_3 = I_1 + I_2$, en la

forma de rectángulo. Entonces, la región de capacidad C_1 puede ser definida como el área convexa entre los cinco puntos

$$(0,0),(I_1,0),(I_1,I_3-I_1),(I_3-I_2,I_2),(0,I_2). \tag{5.4}$$

Consideremos la región definida por $\mathbf{I}_\lambda$, la cual está también definida por cinco puntos. Tomemos un punto cualquiera de esos cinco puntos, digamos $\left(I_3^\lambda - I_2^\lambda, I_2^\lambda\right)$. Este punto puede ser escrito como la combinación (λ,1-λ) de los puntos $(I_3^1 - I_2^1, I_2^1)$ y $(I_3^2 - I_2^2, I_2^2)$, y por tanto está en la envolvente convexa de C_1 y C_2, o

$$C_3 \subset \lambda C_1 + (1-\lambda)C_2.$$

Combinando estas dos partes, se sigue el resultado del lema.

∎

En la prueba del teorema, se ha usado implícitamente el hecho de que todas las regiones de las tasas están definidas por cinco puntos extremos (en el peor de los casos, unos puntos son iguales). Los cinco puntos definidos por el vector $\mathbf{I}$ están en la región de las tasas. Si la condición $I_3 \leq I_1 + I_2$ no se satisface, algunos de los puntos en (5.4) pueden estar afuera de la región de las tasas y la demostración ya no sería cierta.

Como consecuencia de la propiedad de convexidad presentada en el lema anterior, se tiene otro resultado de convexidad:

Teorema 5.3.3 La envolvente convexa de la unión de las regiones de las tasas definidas por vectores individuales $\mathbf{I}$ es igual a la región de las tasas definida por la envolvente convexa de los vectores $\mathbf{I}$.

Los argumentos anteriores acerca de las combinaciones convexas en las regiones de las tasas se pueden extender al canal de acceso múltiple para m usuarios. Una prueba de dicho resultado se puede ver en un artículo de Hahn [20].

En el siguiente teorema volvemos a establecer el resultado de la región de capacidad para un canal de acceso múltiple usando una variable aleatoria Q que "comparte el tiempo".

Teorema 5.3.4 El conjunto de tasas alcanzables de un canal discreto sin memoria y de acceso múltiple está dado por la cerradura convexa del conjunto de todos los pares (R_1, R_2) que satisfacen

$$\begin{aligned} R_1 &< I(X_1;Y \mid X_2,Q), \\ R_2 &< I(X_2;Y \mid X_1,Q), \\ R_1+R_2 &< I(X_1,X_2;Y \mid Q), \end{aligned} \tag{5.5}$$

para una elección de la distribución conjunta $p(q)p(x_1 \mid q)p(x_2 \mid q)p(y \mid x_1,x_2)$ con $|\mathcal{Q}| \leq 4$, donde $\mathcal{Q}$ es el rango de Q.

Demostración. Consideremos una tasa $\mathbf{R}$ que satisface las desigualdades de la hipótesis del teorema. Veamos que

$$I(X_1; Y \quad | \quad X_2, Q) = \sum_{q=1}^{m} p(q) I(X_1; Y \mid X_2)_{p_{1q} p_{2q}}, \tag{5.6}$$

donde m es la cardinalidad del soporte de Q, y $p_{iq} = p(x_i \mid q)$ para $i = 1, 2$. Las otras informaciones mutuas se expanden de manera similar.

Una tasa $\mathbf{R}_{p_1, p_2}$ que satisface las desigualdades de las hipótesis del teorema para una distribución producto $p_{1q}(x_1) p_{2q}(x_2)$ se denotará como $\mathbf{R}_q$. Específicamente, sea $\mathbf{R}_q = (R_{1q}, R_{2q})$ una tasa par que satisface que

$$\begin{aligned} R_{1q} &< I(X_1; Y \mid X_2)_{P_{1q}(x_1) P_{2q}(x_2),} \\ R_{2q} &< I(X_2; Y \mid X_1)_{p_{1q}(x_1) p_{2q}(x_2),} \\ R_{1q} + R_{2q} &< I(X_1, X_2; Y)_{p_{1q}(x_1) p_{2q}(x_2).} \end{aligned} \tag{5.7}$$

Entonces por el teorema 5.3.1, $\mathbf{R}_q$ es alcanzable. Entonces como $\mathbf{R}$ satisface (5.5) y los lados derechos de las desigualdades se pueden desarrollar como en (5.6), por el teorema 5.3.3 existe un conjunto de vectores $\mathbf{R}_q$ que satisfacen (5.7) tales que

$$\mathbf{R} = \sum_{q=1}^{m} p(q) \mathbf{R}_q.$$

Como una combinación convexa de tasas alcanzables es alcanzable, $\mathbf{R}$ es una tasa alcanzable.

El regreso del teorema se probará en la siguiente sección.

∎

La prueba de la convexidad de la región de capacidad muestra que cualquier combinación convexa de tasas alcanzables es alcanzable también. Este proceso se puede continuar, y de esta forma tomar combinaciones convexas de más puntos. Una pregunta que surge es si la región de capacidad se incrementará. El siguiente teorema de Carathéodory que forma parte de la teoría de conjuntos convexos nos dice que esto no es posible. Su demostración puede consultarse en el libro de Grünbaum ([19]).

Teorema 5.3.5 Cualquier punto en la cerradura convexa de un conjunto compacto A en un espacio euclidiano de dimensión d puede ser representado como una combinación convexa de $d + 1$ o menos puntos en el conjunto A.

Este teorema nos permite enfocarnos únicamente a cierta combinación finita y convexa cuando se calcule la región de capacidad. En el canal de acceso múltiple, las desigualdades definen un conjunto compacto y conexo en tres dimensiones. Entonces, todos los puntos en la cerradura pueden ser definidos como la combinación

convexa de a lo más cuatro puntos. Entonces, podemos restringir la cardinalidad del rango de Q a que sea a lo más 4 en el teorema 5.3.4.

Ahora probaremos el regreso del teorema del canal de acceso múltiple.

Demostración (Regresos de los teoremas 5.3.1 y 5.3.4). Debemos mostrar que dada una sucesión de libros de códigos $\left(\left(2^{nR_1}, 2^{nR_2}\right), n\right)$ tales que $P_e^n \to 0$, entonces las tasas cumplen que

$$\begin{aligned} R_1 &\leq I(X_1; Y \mid X_2, Q), \\ R_2 &\leq I(X_2; Y \mid X_1, Q), \\ R_1 + R_2 &\leq I(X_1, X_2; Y \mid Q) \end{aligned}$$

para una variable aleatoria Q definida en $\{1, 2, 3, 4\}$ y con distribución conjunta $p(q)p(x_1 \mid q)p(x_2 \mid q)p(y \mid x_1, x_2)$. Fijemos n. Consideremos un bloque dado de longitud n. La distribución conjunta en $\mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{X}_1^n \times \mathcal{X}_2^n \times \mathcal{Y}^n$ está bien definida. Lo único aleatorio se debe a la elección aleatoria uniforme de los índices W_1 y W_2, y el ruido que produce el canal. La distribución conjunta es

$$p(w_1, w_2, x_1^n, x_2^n, y^n) = \frac{1}{2^{nR_1}} \frac{1}{2^{nR_2}} p(x_1^n \mid w_1) p(x_2^n \mid w_2) \prod_{i=1}^{n} p(y_i \mid x_{1i}, x_{2i}),$$

donde $p(x_1^n \mid w_1)$ es 1 o 0, dependiendo de si $x_1^n = \mathbf{x}_1(w_1)$, el código corresponde a w_1, o no, similarmente, para $p(x_2^n \mid w_2)$. La información mutua que sigue se calcula en base a esta distribución.

Por la construcción del libro de códigos, es posible estimar (W_1, W_2) de una sucesión recibida Y^n con probabilidad de error pequeña. Entonces, la entropía condicional de (W_1, W_2) dado Y^n debe ser pequeña. Por la desigualdad de Fano

$$H(W_1, W_2 \mid Y^n) \leq n(R_1 + R_2)P_e^n + H(P_e^n) \triangleq n\epsilon_n.$$

Es claro que $\epsilon_n \to 0$ cuando $P_e^n \to 0$. Entonces tenemos que

$$\begin{aligned} H(W_1 \mid Y^n) &\leq H(W_1, W_2 \mid Y^n) \leq n\epsilon_n, \\ H(W_2 \mid Y^n) &\leq H(W_1, W_2 \mid Y^n) \leq n\epsilon_n. \end{aligned}$$

Ahora podemos acotar la tasa R_1 por

$$\begin{aligned}
nR_1 &= H(W_1) \\
&= I(W_1; Y^n) + H(W_1 \mid Y^n) \\
&\overset{(a)}{\leq} I(W_1; Y^n) + n\epsilon_n \\
&\overset{(b)}{\leq} I(X_1^n(W_1); Y^n) + n\epsilon_n \\
&= H(X_1^n(W_1)) - H(X_1^n(W_1) \mid Y^n) + n\epsilon_n \\
&\overset{(c)}{\leq} H(X_1^n(W_1) \mid X_2^n(W_2)) - H(X_1^n(W_1) \mid Y^n, X_2^n(W_2)) + n\epsilon_n \\
&= I(X_1^n(W_1); Y^n \mid X_2^n(W_2)) + n\epsilon_n \\
&\overset{(d)}{=} H(Y^n \mid X_2^n(W_2)) - \sum_{i=1}^{n} H(Y_i \mid Y^{i-1}, X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n \\
&\overset{(e)}{=} H(Y^n \mid X_2^n(W_2)) - \sum_{i=1}^{n} H(Y_i \mid X_{1i}, X_{2i}) + n\epsilon_n \\
&\overset{(f)}{\leq} \sum_{i=1}^{n} H(Y_i \mid X_2^n(W_2)) - \sum_{i=1}^{n} H(Y_i \mid X_{1i}, X_{2i}) + n\epsilon_n \\
&\overset{(g)}{\leq} \sum_{i=1}^{n} H(Y_i \mid X_{2i}) - \sum_{i=1}^{n} H(Y_i \mid X_{1i}, X_{2i}) + n\epsilon_n \\
&= \sum_{i=1}^{n} I(X_{1i}; Y_i \mid X_{2i}) + n\epsilon_n,
\end{aligned}$$

donde (a) se sigue de la desigualdad de Fano; (b) se sigue de la desigualdad de procesamiento de datos[1]; (c) se sigue del hecho de que W_1 y W_2 son independientes, y por ende también lo son $X_1^n(W_1)$ y $X_2^n(W_2)$, y entonces $H(X_1^n(W_1) \mid X_2^n(W_2)) = H(W_1^n(W_1))$, y $H(X_1^n(W_1) \mid Y^n, X_2^n(W_2)) \leq H(X_1^n(W_1) \mid Y^n)$; (d) se sigue de la regla de la cadena; (e) se sigue del hecho de que Y_i sólo depende X_{1i} y X_{2i} porque el canal no tiene memoria; (f) se sigue de la regla de la cadena y quitando el condicionamiento; (g) se sigue de quitar condicionamientos.

Entonces, se tiene que

$$R_1 \leq \frac{1}{n} \sum_{i=1}^{n} I(X_{1i}; Y_i \mid X_{2i}) + \epsilon_n. \tag{5.8}$$

Similarmente, se tiene que

$$R_2 \leq \frac{1}{n} \sum_{i=1}^{n} I(X_{2i}; Y_i \mid X_{1i}) + \epsilon_n. \tag{5.9}$$

[1]Si $X \to Y \to Z$ es una cadena de Markov entonces $I(X;Y) \geq I(X;Z)$.

Para acotar la suma de las tasas, se tiene que

$$\begin{aligned} n(R_1+R_2) &= H(W_1,W_2) \\ &= I(W_1,W_2;Y^n)+H(W_1,W_2\mid Y^n) \\ &\overset{(a)}{\leq} I(W_1,W_2;Y^n)+n\epsilon_n \\ &\overset{(b)}{\leq} I(X_1^n(W_1),X_2^n(W_2);Y^n)+n\epsilon_n \\ &= H(Y^n)-H(Y^n\mid X_1^n(W_1),X_2^n(W_2))+n\epsilon_n \\ &\overset{(c)}{=} H(Y^n)-\sum_{i=1}^n H(Y_i\mid Y^{i-1},X_1^n(W_1),X_2^n(W_2))+n\epsilon_n \\ &\overset{(d)}{=} H(Y^n)-\sum_{i=1}^n H(Y_i\mid X_{1i},X_{2i})+n\epsilon_n \\ &\overset{(e)}{\leq} \sum_{i=1}^n H(Y_i)-\sum_{i=1}^n H(Y_i\mid X_{1i},X_{2i})+n\epsilon_n \\ &= \sum_{i=1}^n I(X_{1i},X_{2i};Y_i)+n\epsilon_n, \end{aligned} \tag{5.10}$$

donde (a) se sigue de la desigualdad de Fano; (b) se sigue de la desigualdad de procesamiento de datos; (c) se sigue de la regla de la cadena; (d) se sigue del hecho de que Y_i sólo depende de X_{1i} y X_{2i}; (e) se sigue de la regla de la cadena y quitando condicionamientos.

Entonces, se tiene que

$$R_1+R_2 \leq \frac{1}{n}\sum_{i=1}^n I(X_{1i},X_{2i};Y_i)+\epsilon_n.$$

Ahora podemos escribir las ecuaciones (5.8),(5.9) y (5.10) con una nueva variable aleatoria Q, donde $Q=i\in\{1,2,\ldots,n\}$ con probabilidad $1/n$. Dichas ecuaciones se transforman en

$$\begin{aligned} R_1 &\leq \frac{1}{n}\sum_{i=1}^n I(X_{1q};Y_q\mid X_{2q},Q=i)+\epsilon_n \\ &= I(X_{1Q};Y_Q\mid X_{2Q},Q)+\epsilon_n \\ &= I(X_1;Y\mid X_2,Q)+\epsilon_n, \end{aligned}$$

donde $X_1\triangleq X_{1Q}, X_2\triangleq X_{2Q}$, y $Y\triangleq Y_Q$ son variables aleatorias cuyas distribuciones dependen de Q en la misma manera que las distribuciones de X_{1i}, X_{2i} y Y_i dependen de i. Como W_1 y W_2 son independientes, también $X_{1i}(W_1)$ y $X_{2i}(W_2)$ son independientes, y entonces

$$\begin{aligned} \mathbb{P}(X_{1i}(W_1) &= x_1, X_{2i}(W_2)=x_2) \\ &\triangleq \mathbb{P}(X_{1Q}=x_1\mid Q=i)\mathbb{P}(X_{2Q}=x_2\mid Q=i). \end{aligned}$$

Entonces, tomando el límite cuando $n \to \infty$, $P_e^n \to 0$, y entonces

$$\begin{aligned} R_1 &\leq I(X_1; Y \mid X_2, Q), \\ R_2 &\leq I(X_2; Y \mid X_1, Q), \\ R_1 + R_2 &\leq I(X_1, X_2; Y \mid Q) \end{aligned}$$

para una distribución conjunta $p(q)p(x_1 \mid q)p(x_2 \mid q)p(y \mid x_1, x_2)$. Por el teorema de Carathéodory, la región no cambia si limitamos la cardinalidad de $\mathcal{Q}$ a 4.

∎

5.4.2. Canales de acceso múltiple con m usuarios

En esta sección se generaliza el resultado que se obtuvo en la sección anterior para m usuarios, $m \geq 2$. Se envían de manera independiente los índices $1, 2, \ldots, m$ a través del canal de los remitentes $1, 2, \ldots, m$, respectivamente. Los códigos, tasas, y tasas alcanzables están definidas en la misma manera que en el caso cuando $m = 2$.

Sea $S \subset \{1, \ldots, M\}$. Sea S^c el complemento de S. Sea $R(S) = \sum_{i \in S} R_i$, y sea $X(S) = \{X_i; i \in S\}$. Entonces se tiene el siguiente teorema.

Teorema 5.3.6 La región de capacidad para el canal de acceso múltiple para m usuarios es la cerradura de la envolvente convexa de las tasas de vectores que satisfacen

$$R(S) \leq I(X(S); Y \mid X(S^c)) \text{ para todo } S \subset \{1, 2, \ldots, m\}$$

para una distribución producto $p_1(x_1) \cdots p_m(x_m)$.

Demostración. La prueba no tiene ideas nuevas. Ahora habrá $2^m - 1$ términos en la probabilidad de error en la demostración de la ida del teorema 5.3.1 y un número igual de desigualdades en la demostración del regreso. La generalización es sencilla y no se hará.

∎

Comentarios Finales. Asociadas con algunos de los nodos en la red hay fuentes de información estocásticas, que están para ser comunicadas con otros de los nodos en la red. Si las fuentes son independientes, los mensajes enviados por los nodos son también independientes. Sin embargo, para tener mayor generalidad, se permitirán que las fuentes sean dependientes. ¿Cómo uno pudiera tomar ventaja de la dependencia para reducir la cantidad de información transmitida? Dada la distribución de probabilidad de las fuentes a lo largo del canal y la función de transición del canal ($\mathbb{P}(y^n \mid x^n)$), ¿se pueden transmitir esas fuentes por el canal y recuperar las fuentes en los destinos con la distorsión apropiada?

El problema de codificar la fuente (compresión de la información) cuando los canales no presentan ruido y están sin interferencia, se reduce a encontrar el conjunto

de tasas asociadas con cada fuente tal que las fuentes requeridas puedan decodificar el mensaje en el destino final con una pequeña probabilidad de error. El caso más simple para la codificación de una fuente distribuida es el problema de codificación de fuente de Slepian-Wolf, donde se tienen dos fuentes que deben ser codificadas por separado, pero decodificadas juntas en un mismo nodo. Extensiones de esta teoría cuando sólo se necesita que una de las dos fuentes sea recuperada en el destino han sido consideradas por Cover.

5.5. El canal de emisión

El canal de emisión es un canal de comunicación en donde hay un remitente y dos o más receptores. El problema básico es encontrar el conjunto de tasas que se pueden alcanzar simultáneamente en este canal. Antes de analizar dicho canal, presentemos unos ejemplos.

Ejemplo. (TV y radio) El ejemplo más simple de un canal de emisión es una estación de televisión o la radio. En este ejemplo, normalmente, la estación busca enviar la misma información a todos que estén conectados a dicho canal; la capacidad es $\text{máx}_{p(x)}\, min_i I(X_i; Y_i)$, que puede ser menor que la capacidad del peor receptor. Un problema sería enviar la información de tal manera que los mejores receptores reciban más información, lo que produce una mejor imagen o sonido, mientras que los peores receptores reciban la información básica. Ahora que las estaciones de televisión tienen servicios como la alta definición (HDTV), pudiera ser necesario codificar la información para que los peores receptores reciban la señal de TV normal mientras que los mejores receptores reciban información extra para la señal de alta definición.

Ejemplo. (Una clase) Una clase en un salón es información que se comunica a los estudiantes en la clase. Los estudiantes reciben distintas cantidades de información, debido a las diferencias entre los mismos. Algunos de los estudiantes reciben más información; otros reciben sólo un poco. En una situación ideal, el profesor pudiera impartir su clase de tal forma que los estudiantes buenos reciban más información y los estudiantes malos reciban al menos la información indispensable. Sin embargo, una clase que no se ha preparado correctamente toma como preferencia a los peores estudiantes. Esta situación es otro ejemplo de un canal de emisión.

Ejemplo. El canal de emisión más simple consiste de dos canales independientes que van a dos receptores. Aquí se puede enviar información de manera independiente por los dos canales, y se puede alcanzar una tasa R_1 para el receptor 1 y una tasa R_2 para el receptor 2 si $R_1 < C_1$ y $R_2 < C_2$..

Ejemplo. (Superposición) Para mostrar la idea de superposición, consideremos el ejemplo de una persona que puede hablar español e inglés. Hay dos receptores:

uno que entiende solamente español y el otro sólo entiende inglés. Asumamos por simplicidad que el vocabulario de cada lenguaje es de 2^{20} palabras y que el hablante habla a un tasa de 1 palabra por segundo en cada lenguaje. Entonces él puede transmitir 20 bits de información por segundo al receptor 1 hablándole todo el tiempo; en este caso, no envía información al receptor 2. Similarmente, puede enviar 20 bits por segundo al receptor 2 sin enviar información al receptor 1. Entonces, el puede alcanzar una tasa de $R_1 + R_2 = 20$ usando este método. Sin embargo pudiera hacerlo mejor.

Notemos que el receptor que sólo entiende ingles, puede distinguir cuando la palabra es en español. Similarmente, para el receptor que entiende el español. El hablante puede usar esta información; por ejemplo, si la proporción del tiempo en que usa cada lenguaje es del 50 %, entonces en una sucesión de 100 palabras, él hablará 50 en español y 50 en inglés. Ahora, hay muchas manera de ordenar las palabras en español y inglés; de hecho, hay cerca de $\binom{100}{50} \approx 2^{100H(1/2)}$ maneras de ordenar las palabras. Escogiendo una de esas maneras se envía información a ambos receptores. Este método permite al hablante enviar información a una tasa de 10 bits por segundo al receptor mexicano, y a 10 bits por segundo al receptor inglés, y 1 bit por segundo de información común a los dos receptores, lo que es un total de 21 bits por segundo. Este es un ejemplo de superposición de información.

Los resultados del canal de emisión se pueden aplicar al caso en que no se conoce la distribución de un canal de un sólo usuario. En este caso, el objetivo es conseguir al menos el mínimo de información cuando el canal es malo y conseguir información extra cuando el canal es bueno. Se pueden usar los mismos argumentos de superposición como en el canal de emisión para encontrar tasas a las que se puede enviar información.

Definición. Un canal de emisión consiste de un alfabeto de entrada $\mathcal{X}$ y dos alfabetos de salida, $\mathcal{Y}_1$ y $\mathcal{Y}_2$, y una función de probabilidad de transición $p(y_1, y_2 \mid x)$. Se dirá que el canal de emisión no tiene memoria si $p(y_1^n, y_2^n \mid x^n) = \prod_{i=1}^n p(y_{1i}, y_{2i} \mid x_i)$.

Definimos los códigos, la probabilidad de error, tasas alcanzables, y regiones de capacidad para el canal de emisión como se definieron para el canal de acceso múltiple. Un código $\left(\left(2^{nR_1}, 2^{nR_2}\right), n\right)$ para un canal de emisión con información independiente consiste de un codificador

$$X : \left(\left\{1, 2, \ldots, 2^{nR_1}\right\} \times \left\{1, 2, \ldots, 2^{nR_2}\right\}\right) \to \mathcal{X}^n,$$

y dos decodificadores

$$g_1 : \mathcal{Y}_1^n \to \{1, 2, \ldots, 2^{nR_1}\}$$

y

$$g_2 : \mathcal{Y}_2^n \to \{1, 2, \ldots, 2^{nR_2}\}.$$

Definimos la probabilidad promedio de error como la probabilidad de que el mensaje decodificado no sea igual al mensaje transmitido; esto es

$$P_e^n = \mathbb{P}(g_1(Y_1^n) \neq W_1 \text{ o } g_2(Y_2^n) \neq W_2),$$

donde (W_1, W_2) tiene distribución uniforme en $2^{nR_1} \times 2^{nR_2}$.

Definición. Una tasa par (R_1, R_2) es alcanzable para el canal de emisión si existe una sucesión de códigos $\left(\left(2^{nR_1}, 2^{nR_2}\right), n\right)$ tales que $P_e^n \to 0$.

Ahora definiremos las tasas para el caso en que se tiene información común que se envía a ambos receptores. Un código $\left(\left(2^{nR_0}, 2^{nR_1}, 2^{nR_2}\right), n\right)$ para un canal de emisión con información en común consiste de un codificador

$$X : \left(\left\{1, 2, \ldots, 2^{nR_0}\right\} \times \left\{1, 2, \ldots, 2^{nR_1}\right\} \times \{1, 2, \ldots, 2^{nR_2}\}\right) \to \mathcal{X}^n,$$

y dos decodificadores

$$g_1; \mathcal{Y}_1^n \to \{1, 2, \ldots, 2^{nR_0}\} \times \{1, 2, \ldots, 2^{nR_1}\}$$

y

$$g_2; \mathcal{Y}_2^n \to \{1, 2, \ldots, 2^{nR_0}\} \times \{1, 2, \ldots, 2^{nR_2}\}.$$

Asumiendo que la distribución de (W_0, W_1, W_2) es uniforme, definimos la probabilidad de error como la probabilidad de que el mensaje decodificado no sea igual al mensaje transmitido

$$P_e^n = \mathbb{P}(g_1(Y_1^n) \neq (W_0, W_1) \text{ o } g_2(Y_2^n) \neq (W_0, W_2)).$$

Definición. Una tasa triple (R_0, R_1, R_2) es alcanzable para el canal de emisión con información común si existe una sucesión de códigos $\left(\left(2^{nR_0}, 2^{nR_1}, 2^{nR_2}\right), n\right)$ tales que $P_e^n \to 0$.

Definición. La región de capacidad del canal de emisión es la cerradura del conjunto de todas las tasas alcanzables.

Notemos que un error para el receptor Y_1^n depende solamente de la distribución $p(x^n, y_1^n)$ y no de la distribución conjunta $p(x^n, y_1^n, y_2^n)$. Entonces, tenemos el siguiente teorema:

Teorema 5.4.1 La región de capacidad para un canal de emisión depende solamente de las distribuciones marginales $p(y_1 \mid x)$ y $p(y_2 \mid x)$.

Demostración. Sea $((2^{nR_1}, 2^{nR_2}), n)$ un código dado, y sean

$$\begin{aligned} P_1^n &= \mathbb{P}(\hat{W}_1(\mathbf{Y}_1) \neq W_1), \\ P_2^n &= \mathbb{P}(\hat{W}_2(\mathbf{Y}_2) \neq W_2), \\ P^n &= \mathbb{P}((\hat{W}_1, \hat{W}_2) \neq (W_1, W_2)). \end{aligned}$$

Entonces es claro que

$$\max\{P_1^n, P_2^n\} \leq P^n \leq P_1^n + P_2^n.$$

Como P_1^n sólo depende de $p(y_1 \mid x)$ y P_2^n sólo depende $p(y_2 \mid x)$, entonces el hecho de que $\max\{P_1^n, P_2^n\}$ y $P_1^n + P_2^n$ tiendan a cero sólo depende de las distribuciones marginales $p(y_1 \mid x)$ y $p(y_2 \mid x)$.

∎

5.5.1. Canales de emisión distorsionados

Definición. Un canal de emisión es físicamente distorsionado si $p(y_1, y_2 \mid x) = p(y_1 \mid x)p(y_2 \mid y_1)$.

Definición. Un canal de emisión es estocásticamente distorsionado si sus distribuciones condicionales marginales son las mismas que las de un canal de emisión físicamente distorsionado; es decir, si existe una distribución $p'(y_2 \mid y_1)$ tal que

$$p(y_2 \mid x) = \sum_{y_1} p(y_1 \mid x)p'(y_2 \mid y_1).$$

Dado que la capacidad de un canal de emisión depende solamente de las distribuciones marginales condicionales, la región de capacidad de un canal de emisión estocásticamente distorsionado es la misma que la correspondiente al canal de emisión físicamente distorsionado. Por lo que casi en todo lo siguiente se asumirá que el canal es físicamente distorsionado.

Ahora consideraremos la situación en que se envía información de manera independiente a través de un canal de emisión distorsionados a una tasa de R_1 a Y_1 y una tasa de R_2 a Y_2.

Teorema 5.4.2 La región de capacidad cuando se envía información de manera independiente a través de un canal de emisión distorsionado $X \to Y_1 \to Y_2$ es la envolvente convexa de la cerradura de todos los (R_1, R_2) que cumplen

$$\begin{aligned} R_2 &\leq I(U; Y_2), \\ R_1 &\leq I(X; Y_1 \mid U) \end{aligned}$$

para una distribución conjunta $p(u)p(x \mid u)p(y_1, y_2 \mid x)$, donde la variable aleatoria auxiliar U tiene cardinalidad acotada por $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}_1|, |\mathcal{Y}_2|\}$.

Demostración. (La cota de la cardinalidad para la variable aleatoria U se demuestran usando métodos del teoría de conjuntos convexos y los argumentos no serán dados aquí. Se sugiere ver [3, 34]) Fijemos $p(u)$ y $p(x \mid u)$.

Generación aleatoria del libro de códigos: Generemos 2^{nR_2} códigos independientes de longitud n, $\mathbf{U}(w_2), w_2 \in \{1, 2, \ldots, 2^{nR_2}\}$, se distribuye con densidad

$\prod_{i=1}^{n} p(u_i)$. Para cada código $\mathbf{U}(w_2)$, generemos 2^{nR_1} códigos independientes $\mathbf{X}(w_1, w_2)$ con distribución $\prod_{i=1}^{n} p(x_i \mid u_i(w_2))$. Aquí $\mathbf{u}(i)$ desarrolla el papel de ser la nube de códigos que pueden entender Y_1 y Y_2, mientras que $\mathbf{x}(i, j)$ es el código j-ésimo en la nube i-ésima.

Codificación: Para enviar el par (W_1, W_2), enviaremos el código correspondiente $\mathbf{X}(W_1, W_2)$.

decodificación: El receptor 2 determina el único $\hat{\hat{W}}_2$ tal que $\left(\mathbf{U}(\hat{\hat{W}}_2), \mathbf{Y}_2\right) \in A_\epsilon^n$. Si no hay ninguno o hay más que uno, se dice que hay un error.

El receptor 1 busca el único $\left(\hat{W}_1, \hat{W}_2\right)$ tal que $\left(\mathbf{U}(\hat{W}_2), \mathbf{X}(\hat{W}_1, \hat{W}_2), \mathbf{Y}_1\right) \in A_\epsilon^n$. Si no hay ninguno o hay más que uno, se dice que hay un error.

Análisis de la probabilidad de error: Por la simetría de la generación del código, la probabilidad de error no depende en qué código fue enviado. Entonces, sin pérdida de generalidad, podemos asumir que el par $(W_1, W_2) = (1, 1)$ fue enviado. Sea $P(\cdot)$ la probabilidad condicional de un evento dado que $(1, 1)$ fue enviado.

Dado que esencialmente tenemos un canal de un solo usuario de U a Y_2, podremos decodificar U con una probabilidad de error pequeña si $R_2 < I(U; Y_2)$. Para probar esto, definamos los eventos

$$E_{Yi} = \left\{(\mathbf{U}(i), \mathbf{Y}_2) \in A_\epsilon^n\right\}.$$

Entonces la probabilidad de error del receptor 2 es

$$\begin{aligned} P_e^n(2) &= P\left(E_{Y1}^c \bigcup \bigcup_{i \neq 1} E_{Yi}\right) \\ &\leq P(E_{Y1}^c) + \sum_{i \neq 1} P(E_{Yi}) \\ &\leq \epsilon + 2^{nR_2} 2^{-n(I(U;Y_2)-2\epsilon)} \qquad (5.11) \\ &\leq 2\epsilon \end{aligned}$$

si n es suficientemente grande y $R_2 < I(U; Y_2)$, donde (5.11) se sigue del AEP (teorema 5.2.1). Similarmente, para decodificar el mensaje del receptor 1, se definen los eventos

$$\begin{aligned} \tilde{E}_{Yi} &= \left\{(\mathbf{U}(i), \mathbf{Y}_1) \in A_\epsilon^n\right\}, \\ \tilde{E}_{Yij} &= \left\{(\mathbf{U}(i), \mathbf{X}(i, j), \mathbf{Y}_1) \in A_\epsilon^n\right\}, \end{aligned}$$

donde la tilde se refiere a eventos que están definidos para el receptor 1. Entonces podemos acotar la de probabilidad de error por

$$\begin{aligned} P_e^n(1) &= P\left(\tilde{E}_{Y1}^c \bigcup \tilde{E}_{Y11}^c \bigcup \bigcup_{i \neq 1} \tilde{E}_{Yi} \bigcup \bigcup_{j \neq 1} \tilde{E}_{Y1j}\right) \\ &\leq P(\tilde{E}_{Y1}^c) + P(\tilde{E}_{Y11}^c) + \sum_{i \neq 1} P(\widetilde{E}_{Yi}) + \sum_{j \neq 1} P(\tilde{E}_{Y1j}). \qquad (5.12) \end{aligned}$$

Por los mismos argumentos para el receptor 2, se tiene que $P(\tilde{E}_{Yi}) \leq 2^{-n(I(U;Y_1)-3\epsilon)}$. Entonces, el tercer término de (5.12) se va a cero si $R_2 < I(U;Y_1)$. Pero por la desigualdad de procesamiento de datos y la distorsión del canal, $I(U;Y_1) \geq I(U;Y_2)$, y entonces las condiciones del teorema implican que el tercer término de (5.12) se va a 0. Para el cuarto término de (5.12) veamos que

$$\begin{aligned}
P(\tilde{E}_{Y1j}) &= P\left((\mathbf{U}(1), \mathbf{X}(1,j), \mathbf{Y}_1) \in A_\epsilon^n\right) \\
&= \sum_{(\mathbf{U},\mathbf{X},\mathbf{Y}_1)\in A_\epsilon^n} P\left((\mathbf{U}(1), \mathbf{X}(1,j), \mathbf{Y}_1)\right) \\
&= \sum_{(\mathbf{U},\mathbf{X},\mathbf{Y}_1)\in A_\epsilon^n} P(\mathbf{U}(1))P(\mathbf{X}(1.j) \mid \mathbf{U}(1))P(\mathbf{Y}_1 \mid \mathbf{U}(1)) \\
&\leq \sum_{(\mathbf{U},\mathbf{X},\mathbf{Y}_1)\in A_\epsilon^n} 2^{-n(H(U)-\epsilon)}2^{-n(H(X|U)-\epsilon)}2^{-n(H(Y_1|U)-\epsilon)} \\
&\leq 2^{n(H(U,X,Y_1)+\epsilon)}2^{-n(H(U)-\epsilon)}2^{-n(H(X|U)-\epsilon)}2^{-n(H(Y_1|U)-\epsilon)} \\
&= 2^{-n(I(X;Y_1|U)-4\epsilon)}.
\end{aligned}$$

Entonces, si $R_1 < I(X;Y_1 \mid U)$, el cuarto término de (5.12) se va a 0. Entonces, podemos acotar la probabilidad de error por

$$\begin{aligned}
P_e^n(1) &\leq \epsilon + \epsilon + 2^{nR_2}2^{-n(I(U;Y_1)-3\epsilon)} + 2^{nR_1}2^{-n(I(X;Y_1|U)-4\epsilon)} \\
&\leq 4\epsilon
\end{aligned}$$

si n es suficientemente grande y $R_2 < I(U;Y_1)$ y $R_1 < I(X;Y_1 \mid U)$. La cota de arriba muestra que podemos decodificar el mensaje con una probabilidad de error que tiende a 0. Entonces, existe una sucesión de códigos $\left(\left(2^{nR_1}, 2^{nR_2}\right), n\right)$ con probabilidad de error que tiende a 0. Con esto se completa la prueba de ida del teorema. Para la demostración del regreso ver [17].

∎

Problemas abiertos. Sobre el canal de emisión hay algunos problemas abiertos, para poderlos establecer primero presentaremos unas definiciones necesarias.

Un canal de emisión sin memoria $p(y_1, y_2 \mid x)$ que satisface $I(U;Y_1) \geq I(U;Y_2)$ para toda densidad $p(u,x)$ es conocido como *menos ruidoso*. En este caso decimos que el receptor 1 recibe menos ruido que el receptor 2. La región de capacidad cuando $R_0 = 0$ (ver la definición cuando se tiene información común) para este canal es el conjunto de los pares (R_1, R_2) tales que

$$\begin{aligned}
R_1 &\leq I(X;Y_1 \mid U), \\
R_2 &\leq I(U;Y_2)
\end{aligned}$$

para una función de distribución $p(u,x)$, donde $|\mathcal{U}| \leq \text{mín}\left\{|\mathcal{X}|, |\mathcal{Y}_1|, |\mathcal{Y}_2|\right\} + 1$.

Un canal de emisión sin memoria $p(y_1, y_2 \mid x)$ que satisface $I(X;Y_1) \geq I(X;Y_2)$ para todo $p(x)$ se dice que es *más capaz*. En este caso decimos que el receptor 1 es

más capaz que el receptor 2. La región de capacidad cuando $R_0 = 0$ es el conjunto de los pares (R_1, R_2) tales que

$$\begin{aligned} R_1 &\leq I(X;Y_1 \mid U), \\ R_2 &\leq I(U;Y_2), \\ R_1 + R_2 &\leq I(X;Y_1) \end{aligned}$$

para una función de distribución $p(u,x)$, donde $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}_1|, |\mathcal{Y}_2|\} + 1$.

Fácilmente se puede ver que si un canal es degradado entonces es menos ruidoso, y que si es menos ruidoso entonces es más capaz. El recíproco para las dos afirmaciones anteriores no son necesariamente ciertas.

Estos canales fueron introducidos por Körner y Marton en 1977 (ver [24]), que además establecieron la región de capacidad para el canal menos ruidoso. La región de capacidad para el canal más capaz fue establecida por El Gamal en 1979 (ver [15]).

La región de capacidad no es conocida en general para el canal de emisión sin memoria menos ruidoso con $k > 3$ receptores, y para el canal de emisión sin memoria más capaz con $k > 2$ receptores. La región de capacidad para el canal de emisión sin memoria menos ruidoso con 3 receptores fue probado por Wang y Nair en 2010 (ver [33]).

Bibliografía

[1] Aczél J. y Daróczy Z.: *On Measures of Information and Their Characterizations,* Academic Press, New York, 1975. VI, VI

[2] Algoet P. H. y Cover T. M.: Asymptotic Optimality and Asymptotic Equipartition Properties of Log-Optimum Investment, *Annals of Probability*, Vol. 16, No. 2, 1988, 876-898. VI

[3] Ahlswede R. y Körner, J.: Source coding with side information and a converse for degraded broadcast channels. *IEEE Transactions on Information Theory*, 1975, Vol. 21, pp. 629-637. VII, 96

[4] Arndt C.: *Information Measures. Information and its Description in Science and Engineering*, Springer, Berlin, 2004. V, V, V, V

[5] Barron A. R.: The Strong Ergodic Theorem for Densities: Generalized Shannon-McMillan-Breiman Theorem, *Annals of Probability*, Vol. 13, No. 4, 1985, 1292-1303.

[6] Bean A. J. y Singer A. C.: Factor Graphs for Universal Portfolios, *Signals, Systems and Computers*, 2009, pp. 1375-1379. VI, 68

[7] Bekenstein J. D.: Black holes and information theory, *Contemporary Phys*ics, Vol. 45, 2003, pp. 31-43. V

[8] Cover T. M. y Thomas J. A.: *Elements of Information Theory*, 2nd ed., John Wiley y Sons, Inc.,United States, 2006. V, V, VI, VI, VII, 84

[9] Cover T. y Ordentlich E.: Performance of Universal Portfolios in the Stock Market, *Information Theory. Proceedings. IEEE International Symposium on*, 2000, p. 232. VI, 68

[10] Cover T. y Julian D.: Performace of Universal Portfolios with Side Information, *IEEE Transactions on Information Theory,* 1996,Vol. 42, no.2, pp. 348-363. VI, 68

[11] Csiszár I.: Axiomatic Characterizations of Information Measures. *Entropy*, 10, 2008, pp. 261-273; DOI: 10.3390/E10030261. V

[12] Daróczy Z.: On the Shannon Measure of Information (Hungarian), *Magyar Tud. Akad. Mat. Fiz. Ostz. Közl.*, 1969, Vol. 19, pp. 9-24.

[13] Daróczy Z. y Kátai I.: Additive zahlentheoretische Funktionen und das Mass der Information, *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.*, 1970, Vol. 13, pp. 83-88.

[14] Díaz M.: Análisis de la eficiencia espectral Ergódica asintótica de sistemas MIMO con correlación de Kronecker, Tesis de licenciatura, Universidad de Guadalajara, 2011. VI

[15] El Gamal A.: The capacity of a class of a broadcast channels. *IEEE Transactions on Information Theory*, 1979, Vol. 25(2), pp. 166-169. VII, 99

[16] Faddeev D. K.: On the Concept of Entropy of a Finite Probabilistic Scheme (Russian), *Uspehi Mat. Nauk (N.S.)*, 1956, Vol.11 , No. 1 (67), pp. 227-231. V, VI

[17] Gallager R. G.: A simple derivation of the coding theorem and some applications, *IEEE Transactions on Information Theory*, 1974, IT-11, pp. 3-18. VII, 98

[18] Gray R. M.: *Entropy and Information Theory*, Springer, 2011.

[19] Grünbaum B.: *Convex Polytopes*, Interscience, New York, 1967. 88

[20] Han T.S.: The capacity region of a general multiple acces channel with certain correlated sources, *Information Control*, 1979, Vol. 40, pp. 37-60. VII, 87

[21] Hartley R.VL.L.: Transmission of information, *Bell System Tech. J.*, 1928, Vol. 7, pp. 535-563. IV

[22] Khinchin A. J.: The Concept of Entropy in the Theory of Probability (Russian), *Usephi Mat. Nauk*, 1953, Vol. 8, No. 3 (55), pp. 3-20. V, VI

[23] Kipp M. R.: *Large Scale Linear and Integer Optimization: A Unified Approach*, Kluwer Academic Publishers, Massachusetts, 2004.

[24] Körner J. y Marton K.: Comparison of two noisy channels. *In I. Csiszár and P. Elias (eds.) Topics in Information Theory (Colloquia Mathematica Societatis János Bolyai, Keszthely, Hungary 1975)*, pp. 411-423. VII, 99

[25] Lee P. M.: On the Axioms of Information Theory, *Ann. Math. Statist.*, 1964, Vol. 35, pp. 415-418. VI

[26] Martinez O. y Reyes-Valdés H. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proceedings of the American Academy of Sciences*, 2007, 105: 9709-9714. v

[27] Nair C.: Capacity regions of two new classes of two-receiver broadcast channels. *IEEE Transactions on Information Theory*, 2010, Vol. 56(9), pp. 4207-4214. vii

[28] Salvador R., Martínez A., Pomarol-Clotet E., Sarró S., Suckling J. y Bullmore E.: Frequency based mutual information measures between cluster of brain regions in functional magnetic resonance imaging, *NeuroImage,* 2007,Vol.35, pp. 83-88. v

[29] Seydel R. U.: *Tools For Computational Finance*, 3rd ed. Springer, 2006. 69

[30] Shannon C.E.: A mathematical theory of communication, *Bell Syst. Tech. J.,* 1948,Vol. 27, pp. 379-423,623-656. iv

[31] Shannon C.E.: Coding theorems for a discrete source with a fidelity criterion. *In IRE National Convention Record*, 1959,Part 4, pp. 142-163 iv, vii

[32] Twichpongtorn P.: Information Theory and Stock Market, University of Illinois at Chicago, 2011. vi

[33] Wang Z.V. y Nair C.: The capacity region of a class of broadcast channels with a sequence of less noisy receivers. *In Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, 2010, pp. 595-598. vii, 99

[34] Wyner A. D. y Ziv J.: The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 1973, Vol. 22, pp. 1-10. vii, 96