GEOMETRÍA E INFORMACIÓN

OPTATIVO

Mariela Portesi
Pedro Walter Lamberti
Steeve Zozor

Facultad de Ciencias Exactas

Esto es una dedicatoria
del libro.

Agradecimientos

Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla.

Esto es un epígrafe con texto simulado.
Esto es un epģrafe con texto simulado.
AUTOR DEL EPÍGRAFE, TÍTULO DE LA OBRA

PRÓLOGO

Este libro surge de la experiencia de los autores en el dictado del curso semestral ”Métodos de geometría diferencial en teoría de la información", que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. ...

Los autores

ADVERTENCÍA

Este libro surge de la experiencia de los autores en el dictado del curso semestral ”Métodos de geometría diferencial en teoría de la información", que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. ...

Mariela Portesi
Grenoble, Junio de 2016

Índice

Capítulo 1

Elementos de teoría de probabilidades

Mariela A. Portesi

1-1 Probabilidades

1-2 Variables aleatorias y distribuciones de probabilidad

1-3 Esperanza, momentos y funciones generadoras

0.1. Funciones generatrices 23

1-4 Algunos ejemplos de distribuciones de probabilidad

Capítulo 2

Nociones de teoría de la información

Steeve Zozor

2-1 Introducción

2-2 Entropía como medida de incerteza

2-3 Entropia condicional, información mutua, entropia relativa

2-4 Unas identidades y desigualdades

2-5 Unos ejemplos y aplicaciones

2-6 Entropias y divergencias generalizadas

2-7 Entropias cuanticas discretas

Capítulo 3

Elementos de geometría diferencial

Pedro Walter Lamberti

3-1 Estructuras

Capítulo 4

Geometría de la información

4-1 La Sección 4.1

Capítulo 5

Aplicaciones

5-1 La Sección 5.1

Referencías

CAPÍTULO 1
Elementos de teoría de probabilidades

Mariela A. Portesi

While writing my book I had an argument with Feller.
He asserted that everyone said "random variable"
and I asserted that everyone said "chance variable."
We obviously had to use the same name in our books,
so we decided the issue by a stochastic procedure.
That is, we tossed for it and he won.
J. L. DOOB (CITA DEL LIBRO *Statistical Science*, 1953)

1.1 Probabilidades

introducción...

El concepto de *probabilidad* es importante en situaciones donde el resultado (o *outcome*) de un dado proceso o medición es incierto, cuando la salida de una experiencia no es totalmente previsible. La probabilidad de un evento es una medida que se asocia con cuán probable es el evento o resultado.

Una definición de probabilidad puede obtenerse en base a la enumeración exhaustiva de los resultados posibles de un experimento o proceso, suponiendo que el conjunto de posibilidades es completo en el sentido de que una de ellas debe ser verdad. Si el proceso tiene N resultados distinguibles, mutuamente excluyentes e igualmente probables (esto es, no se prefiere una posibilidad frente a otras), y si n de esos N tienen un dado atributo, la probabilidad asociada a dicho atributo en un dado procesos es $\frac{n}{N}$. Por ejemplo, sorteando un número entre los naturales del 1 al 10, la probabilidad de "obtener un número par" es $\frac{5}{10} = \frac{1}{2}$.

Otra definición de probabilidad se basa en la frecuencia relativa de ocurrencia de un evento. Si en una cantidad N muy grande de procesos independientes cierto atributo aparece n veces, se identifica a la probabilidad asociada a un proceso o ensayo con la frecuencia relativa de ocurrencia $\frac{n}{N}$ del atributo.

Los axiomas de Kolmogorov proveen requisitos suficientes para determinar completamente las propieda-

des de la medida de probabilidad $p(A)$ que se puede asociar a un evento A entre un conjunto de resultados o eventos de un proceso.

Llamemos Ω al *espacio muestral* o espacio fundamental, que es el espacio total de eventos. Por ejemplo, si A es el evento "es un número natural par" y B indica "es un número natural impar", el espacio muestral $\Omega = \{A, B\}$ indica "es un número natural"; en el caso de analizar el tiempo de vida de un aparato, $\Omega = \mathbb{R}$; en el lanzamiento de un dado de 6 caras es Ω es el conjunto de las etiquetas que se asigne a cada una de las caras (los números naturales del 1 al 6, o las letras *a, b, c, d, e, f*, u otro etiquetado). El conjunto de resultados posibles se supone conocido, aún cuando se desconozca de antemano el resultado de una prueba.

Entre los eventos se pueden considerar operaciones análogas a las de la teoría de conjuntos:

- combinación o unión de eventos: $A + B$ se corresponde con $A \cup B$, implicando que se da A, ó B, o ambos;
- intersección de eventos: A, B se corresponde con $A \cap B$, implicando que se dan ambos A y B;
- complemento de un evento: $-A$ se corresponde con $\tilde{A}$ e indica que no se da A.
- eventos disjuntos o mutuamente excluyentes: son aquellos que no se superponen, se anota $A, B = \emptyset$ donde $\emptyset = -\Omega$ denota el evento nulo (evento que no puede ocurrir, es el complemento de Ω).

Las propiedades de la probabilidad de un dado evento quedan determinadas por los siguientes

Axiomas de Kolmogorov

a) $p(A_i) \geq 0 \ \ \forall \, A_i$

b) $p(\Omega) = 1$

c) Si $A_1, A_2, A_3, \ldots$ son eventos mutuamente excluyentes, entonces $p(A_1 + A_2 + A_3 + \cdots) = p(A_1) + p(A_2) + p(A_3) + \cdots$

A partir de estos axiomas se pueden probar varios corolarios y propiedades:

- la probabilidad de un evento seguro o cierto es 1;
- la probabilidad de un evento que no puede ocurrir es 0: $p(\emptyset) = 0$;
- el rango de las probabilidades está acotado: $0 \leq p(A) \leq 1 \ \ \forall \, A$;
- condición de normalización: si $\Omega = A_1 + \cdots + A_N$, con A_i mutuamente excluyentes, entonces $\sum_{i=1}^{N} p(A_i) = 1$;
- si A es subconjunto de B, entonces $p(A) \leq p(B)$.

La *probabilidad conjunta* $p(A, B) = p(B, A)$ es la probabilidad del evento conjunto dado por la composición de los eventos A y B. Se demuestra que

- $p(A, B)$ está acotada: $0 \leq p(A, B) = p(B, A) \leq \text{mín}\{p(A), p(B)\}$;

- si A y B son mutuamente excluyentes, entonces $p(A, B) = 0$;
- si $B_1, \ldots, B_M$ es un conjunto completo de eventos posibles excluyentes entre sí, entonces $\sum_{j=1}^{M} p(A, B_j) = p(A)$.

En el caso de eventos no necesariamente mutuamente excluyentes, se prueba que la *ley de composición* es

$$p(A + B) = p(A) + p(B) - p(A, B) \leq p(A) + p(B),$$

y que para N eventos resulta

$$p(A_1 + \cdots + A_N) \leq p(A_1) + \cdots + p(A_N).$$

La igualdad vale en el caso especial de eventos mutuamente excluyentes (recuperando el tercer axioma de Kolmogorov).

La *probabilidad condicional* de A dado B es la razón entre la probabilidad del evento conjunto y la probabilidad de que se dé B (cuando éste es un evento no nulo):

$$p(A|B) = \frac{p(A, B)}{p(B)}.$$

Es fácil demostrar que esta cantidad toma valores entre 0 y 1, con $p(\Omega|B) = 1$, y que es aditiva para una unión de eventos mutuamente excluyentes referidos al cumplimiento de B. Luego, $p(A|B)$ es una probabilidad. Algunas propiedades interesantes son las siguientes:

- condición de normalización: $\sum_{i=1}^{N} p(A_i|B) = 1$, siendo $A_1, \ldots, A_N$ un conjunto completo de resultados posibles mutuamente excluyentes;
- relación entre probabilidades condicionales inversas: $p(B|A) = \frac{p(B)}{p(A)} p(A|B)$, de donde $p(A|B)$ y $p(B|A)$ coinciden sólo cuando A y B tienen la misma probabilidad;
- *fórmula de Bayes*: si $B_1, B_2, \ldots$ es un conjunto completo de eventos no nulos mutuamente excluyentes, entonces

$$p(B_i|A) = \frac{p(A, B_i)}{p(A)} = \frac{p(A|B_i)p(B_i)}{\sum_j p(A|B_j)p(B_j)}.$$

Dos eventos A y B se dicen *estadísticamente independientes* si la probabilidad condicional de A dado B es igual a la probabilidad incondicional de A: $p(A, B) = p(A)p(B)$. La condición necesaria y suficiente para que N eventos $A_1, \ldots, A_N$ sean estadísticamente independientes es que la probabilidad conjunta se factorice como

$$p(A_1, \ldots, A_N) = p(A_1) \cdots p(A_N).$$

Se deduce que los eventos mutuamente excluyentes no son estadísticamente independientes.

(?, ?)

1.2 Variables aleatorias y distribuciones de probabilidad

En un experimento o un dado proceso, los posibles resultados son típicamente números reales, siendo cada número un evento. Luego los resultados son mutuamente excluyentes. Se considera a esos números como valores de una variable aleatoria X *a valores reales, que puede ser discreta (cuando el espacio muestral es finito o infinito numerable) o continua. La ley de la variable aleatoria X es una medida de probabilidad definida por $P_X(x) = \Pr(X = x)$ o, en general, por $P_X(A) = \Pr(X = x \in A)$. Puede ser útil también considerar variables aleatorias complejas $Z = X + iY$, donde X e Y son variables aleatorias reales.*

1.2.1 Variable aleatoria discreta

Los posibles valores de una variable aleatoria discreta X consisten en un conjunto contable (finito o infinito numerable) de números reales: $x \in \Omega = \{x_1, x_2, \ldots\}$. A cada uno de los valores x_n $(n = 1, 2, \ldots)$ se puede asociar una probabilidad $p_n = p(x_n)$, de modo que se satisface la condición de normalización:

$$\sum_n p_n = 1.$$

La *función (de masa) de probabilidad* es de la forma:

$$p(x) = \begin{cases} \Pr(X = x) & \text{si } x = x_1, x_2, \ldots \\ 0 & \text{en todo otro punto} \end{cases}$$

En la Fig. 1-1 se muestra una representación gráfica de una distribución de probabilidad discreta.

Figura 1-1: Una distribución de probabilidad discreta.

También, se puede caracterizar la ley de la variable discreta X por medio de su *función de repartición*:

$$F_X(x) = \Pr(X \in (-\infty, x]) = \Pr(X \leq x) = \sum_{\forall n\,:\,x_n \leq x} p(x_n)$$

que es una función discontinua, con saltos finitos, y no decreciente.

Sin pérdida de generalidad, el conjunto de valores que toma una variable aleatoria discreta X puede considerarse como $\{0, 1, 2, \ldots, N\}$ para algún N natural, o todo $\mathbb{N}$. Entonces la ley de una variable aleatoria a valores naturales está dada por $\{p_n = \Pr(X = n),\ n \in \mathbb{N}\}$. Luego $\Pr(X \in A) = \sum_{n \in A \cap \mathbb{N}} p_n$, y la función de repartición se calcula como $\Pr(X \leq x) = \sum_{n \leq x} \Pr(X = n)$ que es una función que presenta un salto finito en cada número natural. En general un salto de la función de repartición corresponde a la presencia de una *masa de Dirac* en el entorno del salto.

Un caso especial se tiene cuando un valor x_j es cierto o seguro, y no ocurre ninguno de los otros valores x_i $(i \neq j)$. La forma de la distribución es: $p_n = \delta_{nj}$, donde

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

es el símbolo *delta de Kronecker*. Cuando el espacio muestral es finito de dimensión N, la ley de distribución se puede representar por medio del siguiente vector columna:

$$p = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

con un 1 en el lugar j-ésimo, que también se escribe como $p = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}^t$, donde t indica transposición. La función de repartición resulta una función escalón o de Heaviside: $F(x) = \Theta(x - x_j)$.

Otra situación particular es la de *equiprobabilidad* o *distribución uniforme*. La forma de la distribución es: $p_n = \frac{1}{N} \ \ \forall\ n = 1, \ldots, N$, donde N señala el tamaño del espacio muestral. La ley de distribución se puede representar por medio del siguiente vector columna:

$$p = \begin{pmatrix} 1/N \\ 1/N \\ \vdots \\ 1/N \end{pmatrix}$$

que también se escribe como $p = \begin{pmatrix} \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{pmatrix}^t$. La función de repartición resulta una función escalonada, con saltos de altura $\frac{1}{N}$ para cada n entre 1 y N.

Reordenamiento y relación de mayorización

Para comparar dos distribuciones es útil reordenar el vector de probabilidad permutando sus elementos hasta listarlos de forma descendente. Se anota $p^{\downarrow}$, de modo que $p_1^{\downarrow} \geq p_2^{\downarrow} \geq \ldots \geq p_N^{\downarrow}$. En el ejemplo del caso con certeza se tiene $p^{\downarrow} = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}^t$, mientras que la distribución uniforme no varía.

Se define *mayorización* del siguiente modo, para distribuciones de dimensión N (con sus elementos acomodados en forma decreciente): una distribución p es mayorizada por otra q, y se denota $p \prec q$, si las primeras $N-1$ sumas parciales de $p^{\downarrow}$ y $q^{\downarrow}$ satisfacen $\sum_{i=1}^{n} p_i^{\downarrow} \leq \sum_{i=1}^{n} q_i^{\downarrow}$ para todo $n = 1, \ldots, N-1$, con $\sum_{i=1}^{N} p_i = 1 = \sum_{i=1}^{N} q_i$.

Por ejemplo, $\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}^t \prec \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \end{pmatrix}^t$. Es posible comparar por mayorización distribuciones de distinta dimensionalidad, completando con ceros el vector de probabilidad de menor dimensión. Es importante resaltar que la mayorización provee un *orden parcial* (no total) entre distribuciones, existiendo pares de distribuciones tales que ninguna mayoriza a la otra. Por ejemplo, $\begin{pmatrix} 0{,}50 & 0{,}40 & 0{,}10 \end{pmatrix}^t$ y $\begin{pmatrix} 0{,}70 & 0{,}15 & 0{,}15 \end{pmatrix}^t$ no se comparan por mayorización.

Es interesante notar que la siguiente propiedad es válida para toda distribución p de tamaño N:

$$\begin{pmatrix} \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{pmatrix}^t \prec p \prec \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}^t.$$

En este sentido, los casos particulares de equiprobabilidad y de certeza, se dice que son distribuciones extremas. Notamos que uno implica ignorancia máxima en el resultado de la variable mientras que el otro corresponde a conocimiento completo.

Figura 1-2: Orden parcial por mayorización

1.2.2 Variable aleatoria continua

Los posibles valores de una variable aleatoria continua X son cualesquiera de los números en un dado intervalo de la recta real: $x \in \Omega \subset \mathbb{R}$ que puede ser un intervalo $[x_m, x_M]$ o un subconjunto (semi)infinito. Es conveniente asociar una *función densidad de probabilidad* (comúnmente anotada por su sigla en inglés: pdf por *probability density function*) $p(x)$ que tiene el sentido de que la probabilidad de que X tome valor entre a y b está dada por:

$$\Pr(a \leq X \leq b) = \int_a^b p(x)\, dx,$$

siendo $p(x)\, dx$ la densidad de probabilidad de hallar a la variable con valores en el intervalo infinitesimal entre x y $x + dx$. La condición de normalización se escribe

$$\int_{x_m}^{x_M} p(x)\, dx = 1.$$

En la Fig. 1-3 se muestra una representación gráfica de una función densidad de probabilidad para una variable continua.

Figura 1-3: Una distribución de probabilidad continua.

También, se puede caracterizar la ley de la variable continua X por medio de su *función de repartición* o *función de distribución cumulativa* (CDF por *cumulative distribution function*):

$$F_X(x) = \Pr(X \leq x) = \int_{x_m}^{x} p(t)\, dt$$

que da la probabilidad de que X sea menor o igual que cierto valor x dado (dentro del conjunto Ω de todos los valores posibles de la variable). En forma análoga, $\Pr(X \in A) = \int_A p(x)\, dx$ acumula la densidad de probabilidad en un subconjunto A del espacio muestral. Por la propiedad de la inclusión, se tiene $\Pr(X \leq x_1) \leq \Pr(X \leq x_2)$ siempre que $x_1 \leq x_2$; luego $F_X(x)$ es una función creciente de x, acotada por la unidad, con valores extremos dados por $\lim_{x\to-\infty} F_X(x) = 0$ y $\lim_{x\to\infty} F_X(x) = 1$, tomando $\Omega = \mathbb{R}$. Además la derivada respecto de x es la pdf:

$$\frac{dF_X(x)}{dx} = p(x).$$

De aquí se observa que la densidad de probabilidad $p(x)$ puede no ser una función “ordinaria” cuando $\Pr(X \leq x)$ es discontinua, pero como mucho tiene la singularidad de una distribución *delta de Dirac* cuya

representación integral es:

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx} dt.$$

Un caso especial se tiene cuando la variable aleatoria X toma el valor x_0 con certeza. La forma de la pdf es: $p(x) = \delta(x - x_0)$. Otra situación particular es la distribución uniforme en un intervalo; la pdf es de la forma $p(x) = \frac{1}{b-a} \; \forall \; x \in [a, b]$, donde $[a, b]$ es el espacio muestral.

Usando las funciones delta de Dirac, se puede unificar el tratamiento de las variables aleatorias discretas con las continuas: si una variable aleatoria discreta toma los valores $x_1, x_2, \ldots$ con probabilidades $p_1, p_2, \ldots$ respectivamente, entonces formalmente se puede describir mediante una variable aleatoria continua X con función densidad de probabilidad $p(x) = \sum_j p_j \, \delta(x - x_j)$.

1.2.3 Vector aleatorio

Cuando se trabaja con $d \geq 2$ variables aleatorias es conveniente definir un *vector aleatorio* de dimensión d, y apelar para su estudio a nociones del álgebra lineal y a notación matricial. Se tiene el vector aleatorio d-dimensional $\mathbf{X} = \{X^1, \ldots, X^d\}$, o simplemente $X = \begin{pmatrix} X^1 & \cdots & X^d \end{pmatrix}^t$, caracterizado por d-uplas de variables aleatorias reales, con función densidad de probabilidad conjunta $p(x^1, \ldots, x^d)$. La ley del vector $\mathbf{X}$ es una medida de probabilidad sobre $\mathbb{R}^d$, con

$$P_{\mathbf{X}}(\mathbf{A}) = \Pr(\mathbf{X} \in \mathbf{A}) = \int_{\mathbf{A}} p(x^1, \ldots, x^d) \, dx^1 \ldots dx^d$$

para $\mathbf{A} \subset \mathbf{\Omega}$, siendo la pdf conjunta p una función positiva, definida sobre $\mathbf{\Omega} \subset \mathbb{R}^d$, y tal que se satisface la condición de normalización:

$$\int_{\mathbf{\Omega}} p(x^1, \ldots, x^d) \, dx^1 \ldots dx^d = 1.$$

La *función densidad de probabilidad marginal* que caracteriza a la variable aleatoria X^i es la ley que se obtiene integrando la pdf conjunta sobre todas las variables excepto la i-ésima:

$$p_{X^i}(x^i) = \int_{\mathbf{\Omega}^{(i)}} p(x^1, \ldots, x^d) \, dx^1 \ldots dx^{i-1} dx^{i+1} \ldots dx^d$$

donde $\mathbf{\Omega}^{(i)} \subset \mathbb{R}^{d-1}$ barre el espacio muestral para $X^1, \ldots, X^{i-1}, X^{i+1}, \ldots, X^d$.

Las d variables aleatorias $X^1, \ldots, X^d$ de un vector aleatorio $\mathbf{X}$ se dicen *independientes* si corresponden a eventos mutuamente independientes. Esto se da si y sólo si la pdf conjunta se factoriza en las d pdf marginales:

$$p(x^1, \ldots, x^d) = p_{X^1}(x^1) \cdots p_{X^d}(x^d).$$

1.2.4 Transformación de variables aleatorias

Sea X una variable aleatoria (continua, en general) definida en el intervalo $[x_m, x_M]$ con función densidad de probabilidad $p(x)$. Sea $Y = \Psi(X)$ una función real de X, luego Y toma los valores $y = \Psi(x)$ en el intervalo $[y_m, y_M]$. La función densidad de probabilidad $q(y)$ para la variable aleatoria transformada Y se obtiene de la siguiente manera, dependiendo de la forma de la transformación:

- Si Ψ es inversible, con inversa (única), se tiene $x = \Phi(y)$, con $\Phi = \Psi^{-1}$. A partir de la propiedad de conservación de la probabilidad

  $$|q(y)\, dy| = |p(x)\, dx|$$

  para una correspondencia biunívoca entre x e y, se obtiene la pdf transformada

  $$q(y) = p(x) \left|\frac{dx}{dy}\right| = p\left(\Phi(y)\right)\ |\Phi'(y)| = \frac{p\left(\Phi(y)\right)}{|\Psi'(\Phi(y))|}.$$

  Una forma alternativa de derivar este resultado es partir de la función de repartición:

  $$F_Y(y) = P(Y \leq y) = P(\Psi(X) \leq y) = P(X \leq \Psi^{-1}(y)) = F_X(\Phi(y))$$

  y calcular las derivadas del primer y último términos respecto de la variable transformada y.

- Si la inversa de Ψ es multivaluada, cada valor de y se corresponde con un conjunto de valores de x, digamos $\{x_k = \Phi_k(y),\ k = 1, 2, \ldots\}$. Debido a que estas soluciones son mutuamente excluyentes, las probabilidades se suman, de modo que

  $$q(y) = \sum_k p(x_k) \left|\frac{dx_k}{dy}\right| = \sum_k \frac{p\left(\Phi_k(y)\right)}{|\Psi'(\Phi_k(y))|},$$

  que formalmente se puede expresar como $q(y) = \int p(x)\, \delta(y - \Psi(x)\,)\, dx$, donde se usa la expansión de la función delta en términos de sus ceros: $\delta(y - \Psi(x)\,) = \sum_k \delta(x - x_k)/|\Psi'(x_k)|$.
  Por ejemplo, para la transformación de variables $Y = X^2$ se tiene $Y = \Psi(X) = X^2$ cuyas inversas son $X_1 = \Phi_1(Y) = +\sqrt{Y}$ y $X_2 = \Phi_2(Y) = -\sqrt{Y}$; luego $q(y) = \frac{p(\sqrt{y})}{2\sqrt{y}} + \frac{p(-\sqrt{y})}{|-2\sqrt{y}|}$, para $y > 0$.

Consideramos ahora el caso de un vector aleatorio $\mathbf{X} = \{X^1, \ldots, X^d\}$ con función densidad de probabilidad conjunta $p(x^1, \ldots, x^d)$. Se define otro vector aleatorio $\mathbf{Y} = \{Y^1, \ldots, Y^d\}$, por medio de las transformaciones $Y^j = \Psi^j(X^1, \ldots, X^d),\ j = 1, \ldots, d$. Suponiendo que las funciones Ψ^j tienen inversa (única), se puede escribir $X^j = \Phi^j(Y^1, \ldots, Y^d)$ para cada j. La función densidad de probabilidad conjunta $q(y^1, \ldots, y^d)$ para $\mathbf{Y}$ se puede obtener a partir de la propiedad de conservación de la probabilidad

$$|q(y^1, \ldots, y^d)\ dy^1 \cdots dy^d| = |p(x^1, \ldots, x^d)\ dx^1 \cdots dx^d|.$$

Para una correspondencia biunívoca entre $\mathbf{x}$ e $\mathbf{y}$, se obtiene la pdf transformada

$$q(y^1, \ldots, y^d) = |\mathrm{J}_\Phi|\ p(x^1, \ldots, x^d)$$

donde $\mathrm{J}_\Phi = \frac{\partial(\Phi^1, \ldots, \Phi^d)}{\partial(y^1, \ldots, y^d)}$ es el Jacobiano de la transformación.

Una *variable aleatoria compleja* $Z = X + iY$ puede interpretarse en términos de las dos variables aleatorias reales X e Y. La pdf asociada $P(z) = p(x, y)$ está dada por la función densidad de probabilidad conjunta de las variables reales. La condición de normalización se escribe

$$\int P(z)\, d^2z = 1$$

donde $d^2z = dx\, dy$.

1.3 Esperanza, momentos y funciones generadoras

introducción...

1.3.1 Momentos de una distribución

....

0.1. Funciones generatrices

....

1.4 Algunos ejemplos de distribuciones de probabilidad

introducción...

1.4.1 Distribuciones de variable discreta

Variable con certeza

...

Ley de Bernoulli

...

Ley geométrica

...

Distribuciń binomial

...

Distribución de Poisson

...

Estadística de los números de ocupación de niveles energéticos: distribuciones de Maxwell–Boltzmann, de Fermi–Dirac, y de Bose–Einstein

...

Leyes de los grandes números

1.4.2 Distribuciones de variable continua

Distribución uniforme sobre un intervalo

...

Distribución exponencial

...

Distribución normal o Gaussiana

...

Distribución Gamma

...

Teorema del límite central

...

CAPÍTULO 2
Nociones de teoría de la información

Steeve Zozor

"Deberías llamarla 'entropía', por dos motivos.
En primer lugar su función de incerteza
ha sido usada en la mecánica estadística
bajo ese nombre, y por ello, ya tiene un nombre.
En segundo lugar, y lo que es más importante,
nadie sabe lo que es realmenta la entropía,
por ello, en un debate, siempre llevará la ventaja.
VON NEUMANN TO SHANNON (TRIBUS & MCIRVINE, 1971)

2.1 Introducción

La noción de información encuentra su origen con el desarollo de la comunicación moderna, por ejemplo a través del telegrafo siguiendo el patente de Moorse en 1840. La idea de asignar un código (punto o barra, mas espacio entre letras y entre palabras) a las letras del alfabeto es la semilla de la codificación entropica, la que se basa precisamente sobre la asignación de un código a simbolos de una fuente (codificación de fuente) según las frequencias (o probabilidad de aparición) de cada simbolo en una cadena. De hecho, el principio de codificar un mensage y mandar la versión codificada por un canal de transmisión es mucho mas antiguo, a pesar de que no habia ninguna formalización matematica ni siquiera explicitamente una noción de información. Entre otros, se puede mencionar el telegrafo optico de Claude Chappe (1794), experimentos con luces por Guillaume Amontons (en los años 1690 en Paris), o aún mas antiguamente la transmisión de mensaje con antorchas en la Grecia antigua, con humo por los indios o chiflando en la prehistoria (Montagné, 2008). Cada forma es una instancia practica del esquema de comunicación de Shannon (Shannon, 1948; Shannon & Weaver, 1964), es decir la codificación de la información, potencialmente de manera la mas economica que se puede, su transmisión a un "receptor" (por un canal ruidoso) que la interpreta/lee/decodifica. Implicitamente, la noción de información a lo menos tan antigua que la humanidad.

A pesar de que la idea de codificar y transmitir “información” sea tremendamente antigua, la formalización matematica de la noción de incerteza o falta de información, intimamente vinculado a la noción de información, nació bajo el impulso de C. Shannon y la publicación de su papel seminal, “A mathematical theory of communication” en 1948 (Shannon, 1948), o un año depues en su libre re-titulado “The mathematical theory of communication” reamplanzando el “A” por un “The”. Desde esto años, las herramientas de la dicha teoria de la información dio lugar a muchas aplicaciones especialmente en communicación (ver por ejemplo (Cover & Thomas, 2006; Verdu, 1998; Gallager, 2001, y ref.), pero también en otros campos muy diversos tal como **Completar con ref, Boltzman, von Neumann, Gibbs, Maxwell, Planck...**.

La meta de este capítulo es de describir las ideas y los pasos dando lugar a la definicion de la entropia, como medida de incerteza o (falta de) información. En este capítulo, se empieza con la descripción intuitiva que subtiende a la noción de información contenida en una cadena de simobolo, lo que condujo a la definición de la entropia. Esta defición puede ser deducida también de una seria de propiedades razonables que debería cumplir una medida de incerteza (enfoque axiomatico). Se continua con la descripción de tal noción de entropia, pasando del mundo discreto (simbolos, alfabeta) al mundo continuo, lo que no es trivial ni siquiera intuitivo. Se adelanto presentando el concepto de información compartida entre dos sistemas o variables aleatorias, concepto fundamental en el marco de la transmisión de información o de mensajes. **Seguir.**

2.2 Entropía como medida de incerteza

2.2.1 Entropía de Shannon, propiedades

Un de los primeros trabajos tratando de formalizar la noción de información de una cadena de simbolos es debido a Raph Hartley (Hartley, 1928). En su papel, Hartley definó la información de una secuencia como siendo proporcional a su longitud. Mas precisamente, para simbolos de un alfabeto de cardinal α, exiten α^n cadenas diferentes de longitud n; Se definó la información de tales cadenas como siendo Kn (K dependiente de α). Para ser consistente, dos conjuntos de mismo tamanio $\alpha_1^{n_1} = \alpha_2^{n_2}$ deben llegar a la misma información, así que la información de Hartley es definida como $H = \log(\alpha^n)$ donde la base del logaritmo es arbitraria. Dicho de otra manera, tomando un logaritmo de base 2, esta información es nada mas que los numeros de bits (0-1) necesarios para codificar todas las cadenas de longitud n de simbolos de un alfabeto de cardinal α.

Hartley, equiv de Gibbs de la termostat.

Una debilidad del enfoque de Hartley es que considera implicitamente que en un mensage, cada cadena de longitud dada puede aparecer con la misma frecuencia, o probabilidad $1/\alpha^n$, siendo la información menos el logaritmo de estas probabilidades. A contrario, parece mas lógico considerar que secuencias muy frecuentes no llevan mucha información (se sabe que aparecen), mientras que las que aparencen raramente llevan mas

información (hay mas sorpresa, mas incerteza en observarlas). Volviendo a los simbolos elementales x, vistos como aleatorios (o valores o estados que puede tomar una variable aleatoria), la (falta de) información o incerteza va a ser intimamente vinculada a la probabilidad de aparición de estos simbolos x. Siguiendo la idea de Hartley, la información elemental asociado al estado x va a ser $-\log p(x)$ donde $p(x)$ es la probabilidad de apararición de x. Se define la incerteza asociada a la variable aleatoria como el promedio estadistico sobre todos los estados posibles x (Shannon, 1948; Shannon & Weaver, 1964) [1].

Definición 2-1 (Entropia de Shannon)**.** *Sea X una variable aleatoria definida sobre una alpfabeto discreto $\mathcal{X} = \{x_1, \ldots, x_\alpha\}$ de cardinal $\alpha = |\mathcal{X}| < +\infty$ finito. Sea p_X la distribución de probabilidad de X, i. e., $\forall\, x \in \mathcal{X}, \quad p_X(x) = \Pr[X = x]$. La entropia de Shannon de la variable X es definida por*

$$H(p_X) = H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\, \log p_X(x)$$

con la convención $0 \log 0 = 0$.

La base del logaritmo es arbitrario; si es $\log_2$ el logaritmo de base 2, H es en bits y si se usa el logaritmo natural $\ln$, H es en nats. *En este capítulo, se usara H con el logaritmo correspondiente sin especificar la base. Si es necesario que tenga una base $a(\neq 1)$ dada, se notara la entropia corespondiente H_a y se especifiera la base del logaritmo* $\log_a$. Fijense de que $\log_a x = \frac{\log x}{\log a}$, dando

$$H_a(X) = H_b(X) \log_a b$$

En lo que sigue, aún que, rigorosamente, H sea una función de la distribución de probabilidad p_X y no de la variable X, se usara indistamente la notación $H(p_X)$ tal como $H(X)$ según lo mas conveniente. Ademas, p_X podrá denotar indistamente la distribución de probabilidad, o el vector de probabilidad $p_X \equiv \begin{bmatrix} p_X(x_1) & \cdots & p_X(x_\alpha) \end{bmatrix}^t$.

H tiene propiedades notables, que coresponden a las que se puede exigir de una medida de incerteza (Shannon, 1948; Shannon & Weaver, 1964; Cover & Thomas, 2006; Rioul, 2007; Dembo, Cover & Thomas, 1991; Johnson, 2004):

[P1] *Continuidad:* Vista como una función de α variables $p_i = p_X(x_i)$, H es continua con respeto a los p_i.

[P2] *Invariance bajo una permutación:* Obviamente, la entropia es invariante bajo una permutación de las probabilidades, *i. e.,*,

para cualquier permutación $\sigma : \mathcal{X} \to \mathcal{X}, \quad H(p_{\sigma(X)}) = H(p_X)$ con $p_{\sigma(X)}(x) = p_X(\sigma(x))$

[1] En la misma época que Shannon, independientemente, la noción de información o medidades equivalentes apareciendo por ejemplo en calculo de capacidad de canal aparecen en varios trabajo como el de Calvier (Clavier, 1948), Laplume (Laplume, 1948), Wiener (Wiener, 1948, Cap. III) entre varios otros (ver (Verdu, 1998; Lundheim, 2002; Rioul & Magossi, 2014; Flandrin & Rioul, 2016; Rioul & Flandrin, 2017; Chenciner, 2017)).

lo que se escribe también $H(\sigma(X)) = H(X)$. En particular, denotando $p_X^{\downarrow}$ la distribución de probabilidades obtenida a partir de p_X, clasificando las probabilidades en orden decreciente, $p_X^{\downarrow}(x_1) \geq p_X^{\downarrow}(x_2) \geq \cdots \geq p_X^{\downarrow}(x_\alpha)$,

$$H(p_X^{\downarrow}) = H(p_X)$$

[P3] *Invariance bajo una transformación biyectiva:* La entropia es invariante bajo cualquier transformación biyectiva, *i. e.,*

$$\text{para cualquier función biyectiva } g : \mathcal{X} \to g(\mathcal{X}), \quad H(g(X)) = H(X)$$

A través tal transformación los estados cambian, pero no cambia la distribución de probabilidad vinculada al alfabeto transformado. Tomando el ejemplo de un dado, la incerteza vinculada al dado no debe depender de los simbolos escritos sobre las caras, sean enteras o cualquier letras.

[P4] *Positividad:* La entropia es acotada por debajo,

$$H(X) \geq 0$$

con iguladad si y solamente si existe un $x_0 \in \mathcal{X}$ tal que $p_X(x_0) = 1$ y $p_X(x) = 0$ para $x \neq x_0$,

$$H(X) = 0 \quad \text{ssi} \quad X \text{ es deterministico}$$

En otras palabras, cuando X no es aleatoria, *i. e.,* $X = x_0$, no hay incerteza, o la observación no lleva información (se sabe lo que va a salir, sin duda): $H = 0$. La positividad es consecuencia de $p_X(x) \leq 1$, dando $-p_X(x) \log p_X(x) \geq 0$. Ademas, la suma de terminos positive es cero si y solo si cada termino es cero, dando $p_X(x) = 0$ o $p_X(x) = 1$.

[P5] *Maximalidad:* La entropia es acotada por arriba,

$$H(X) \leq \log \alpha$$

con iguladad si y solamente si existe X es uniforme sobre $\mathcal{X}$, *i. e.,*

$$H(X) = \log \alpha \quad \text{ssi} \quad \forall\, x \in \mathcal{X},\ p_X(x) = \frac{1}{\alpha}$$

En otras palabras, la incerteza es maxima cuando cualquier estado x puede aparecer con la misma probabilidad; cada observación lleva una información importante sobre el sistema que genera X. La cota máxima resuelta de la maximisación de H sujeto a $\sum_x p_X(x) = 1$, es decir, con la tecnica del Lagrangiano, notando $p_i = p_X(x_i)$, de la minimización de $\sum_i (-p_i \log p_i + \lambda p_i)$. Se obtiene sencillamente que $\log p_i = -\lambda$, dando la distribución uniforma.

La figura 2-4 representa la entropia de un sistema a dos estados, de probabilidades λ y $1 - \lambda$ (lei de Bernoulli de parametro λ), entropia a veces dicha *entropia binaria*, en función de λ. Esta figura ilustra ambas cotas ($\lambda = 1$ o 1, $\lambda = \frac{1}{2}$) así que la invariancia bajo una permutación ($h(\lambda) = H(\lambda, 1-\lambda) = H(1-\lambda, \lambda) = h(1-\lambda)$).

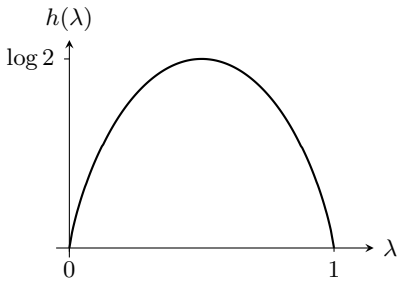


Figura 2-4: Entropia binaria (de una variable de Bernoulli) $h(\lambda) = H(\lambda, 1-\lambda)$ en función de $\lambda \in [0,1]$.

[P6] *Expansibilidad:* Añadir un estado de probabilidad 0 no cambia la entropia, *i. e.,* sean X definido sobre $\mathcal{X}$ y $\widetilde{X}$ sobre $\widetilde{\mathcal{X}}$,

$$\widetilde{\mathcal{X}} = \mathcal{X} \cup \{\widetilde{x}_0\} \quad \text{con} \quad p_{\widetilde{X}}(x) = p_X(x) \quad \text{si} \quad x \in \mathcal{X}, \quad p_{\widetilde{X}}(\widetilde{x}_0) = 0, \qquad \text{entonces} \quad H(p_{\widetilde{X}}) = H(p_X)$$

Esta propiedad es obvia, consecuencia de $\lim_{p\to 0} p \log p = 0$.

[P7] *Recursividad:* Juntar dos estados baja la entropia de una cantidad igual a la entropia interna de los dos estados por la probabilidad de ocurencia de este conjunto de estados, y vice-versa, *i. e.,* sean X definido sobre $\mathcal{X}$ y $\overline{X}$ sobre $\overline{\mathcal{X}}$,

$$\begin{cases} \overline{\mathcal{X}} = \{x_1, \ldots, x_{\alpha-2}, \overline{x}_{\alpha-1}\} \quad \text{con el estado interno} \quad \overline{x}_{\alpha-1} = \{x_{\alpha-1}, x_\alpha\}, \\ p_{\overline{X}}(x_i) = p_X(x_i), \quad 1 \le i \le \alpha - 1 \quad \text{y} \quad p_{\overline{X}}(\overline{x}_{\alpha-1}) = p_X(x_{\alpha-1}) + p(x_\alpha) \quad \text{distribución sobre } \overline{\mathcal{X}} \\ \overline{q}(x_j) = \dfrac{p_X(x_j)}{p_X(x_{\alpha-1}) + p_X(x_\alpha)}, \quad j = \alpha - 1, \alpha \quad \text{distribución del estado interno} \end{cases}$$

$$H(p_X) = H(p_{\overline{X}}) + p_{\overline{X}}(\overline{x}_{\alpha-1})\, H(\overline{q})$$

Esta relación viene de $a \log a + b \log b = (a+b)\left(\frac{a}{a+b} \log\left(\frac{a}{a+b}\right) + \frac{b}{a+b} \log\left(\frac{a}{a+b}\right) - \log(a+b)\right)$ esta ilustrada en la figura 2-5 siguiente.

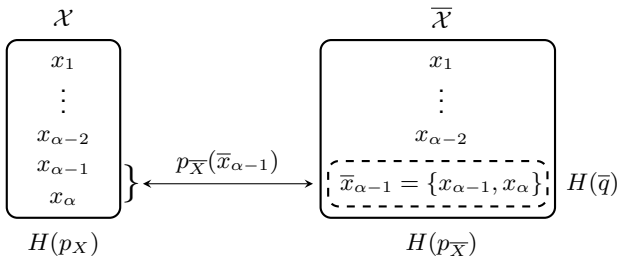


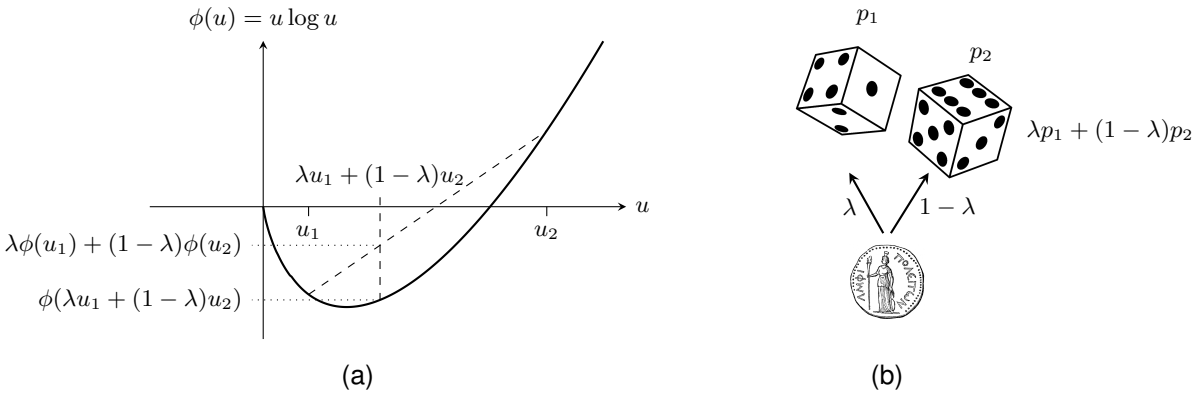
Figura 2-5: Ilustración de la propiedad de recursividad, que cuantifica como decrece la entropia en un conjunto cuando se juntan dos estados, vincluando la entropia total, la entropia despues del la agrupación y la entropia interna a los dos estados juntados.

[P8] *Concavidad:* La entropia es concava, en el sentido de que la entropia de una combinación convexa de distribuciones (mezcla) de probabilidades es siempre major o igual a la combinación convexa de entropias:

$$\forall \{\lambda_i\}_{i=1}^n, \quad 0 \leq \lambda_i \leq 1, \quad \sum_i \lambda_i = 1 \quad \text{and cualquier conjunto de distribuciones} \quad \{p_i\}_{i=1}^n,$$

$$H\left(\sum_i \lambda_i p_i\right) \geq \sum_i \lambda_i H(p_i)$$

Esta desigualdad es conocida como desigualdad de Jensen. Es una consequencia directa de la convexidad de la función $\phi : u \mapsto u \log u$, como ilustrado en la figura 2-6-(a). La figura 2-6-(b) ilustra como se puede obtener una mezcla de distribuciones de dos probabilidad p_1 (dado izquierda) y p_2 (dado derecho) haciendo una elección aleatoria a partir de una moneda en este ejemplo (probabilidad λ de elegir el dado izquierda).



(a) (b)

Figura 2-6: (a) $\phi(u) = u \log u$ es convexa: la curva es siempre debajo de sus cuerdas; entonces, cada promedio de $\phi(u_1)$ y $\phi(u_2)$ estando en la cuerda juntando estos punto, queda arriba de la función tomada en el promedio de u_1 y u_2. Escribiendo eso para (mas de dos puntos) sobre los $\sum_i \lambda_i p_i(x)$ y sumando sobre los x da la desigualdad de Jensen. (b) Ilustración de una distribución de mezcla, acá mezclando p_1 y p_2 a partir de una tercera variable aleatoria (acá de Bernoulli).

[P9] *Schur-concavidad:* Como se lo puede querrer, lo mas "concentrado" es una distribución de probabilidad, lo menos hay incerteza, y entonces lo mas pequeño debe ser la entropia. Esta propiedad intuitiva se resuma a partir de la noción de mayorización:

Definición 2-2 (Mayorización)**.** *Una distribución discreta finita de probabilidad p es dicha mayorizada por una distribución q,*

$$p \prec q \quad \textit{ssi} \quad \sum_{i=1}^k p^{\downarrow}(x_i) \leq \sum_{i=1}^k q^{\downarrow}(x_i), \quad 1 \leq k < \alpha \quad \textit{y} \quad \sum_{i=1}^{\alpha} p^{\downarrow}(x_i) = \sum_{i=1}^{\alpha} q^{\downarrow}(x_i)$$

(las últimas sumas siendo igual a 1). Si los alfabetos de definición de p y q son de tamaños diferentes,

α es el tamaño lo mas grande y la distribución sobre el alfabeto lo mas corto es completada por estados de probabilidad 0 (recuerdense de que no va a cambiar la entropia).

La Schur-concavidad se traduce por la relación

$$p \prec q \quad \Rightarrow \quad H(p) \geq H(q)$$

Fijense de que las cotas sobre H pueden ser vistas como consecuencia de esta desiguldad: la distribución cierta mayoriza cualquier distribución y cualquier distribución mayoriza la distribución uniforme (?, ?, p. 9, (6)-(8)). La prueba de la Schur-concavidad se apoya sobre la desigualdad de Schur (o Hardy-Littlewood-Pólya o Karamata) (?, ?, ?, ?; Hardy, Littlewood & Pólya, 1952), (?, ?, Cap. 3, Prop. C.1) o (?, ?, Teorema II.3.1): $p \prec q \Rightarrow \sum_i \phi(p_i) \leq \sum_i \phi(q_i)$ para cualquier función ϕ convexa. Sufice considerar $\phi(u) = u \log u$ para concluir.

En muchos casos, uno tiene que trabajar con varias variables aleatorias. Para simplificar les notaciones, considera una par de variables X y Y definidas respectivamente sobre los alfabetos $\mathcal{X}$ y $\mathcal{Y}$ de cardinal $\alpha = |\mathcal{X}|$ y $\beta = |\mathcal{Y}|$. Tal par de variable puede ser vista como una variable (X, Y) definida sobre el alfabeto $\mathcal{X} \times \mathcal{Y}$ de cardinal $\alpha\beta$ tal que se definie naturalmente la entropia para esta variable; tal entropia es llamada *entropia conjunta* de X y Y:

Definición 2-3 (Entropia conjunta)**.** *Sean X e Y dos variable aleatorias definidas sobre los alpfabetos discretos $\mathcal{X}$ y $\mathcal{Y}$, de cardinal $\alpha = |\mathcal{X}| < +\infty$ y $\beta = |\mathcal{Y}| < +\infty$ respectivamente. Sea $p_{X,Y}$ la distribución de probabilidad conjunta de X e Y, i. e., $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad p_{X,Y}(x, y) = \Pr[X = x, Y = y]$. La entropia conjunta de Shannon de las variables X e Y es definida por*

$$H(p_{X,Y}) = H(X, Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x, y) \log p_{X,Y}(x, y)$$

con la convención $0 \log 0 = 0$.

A partir de esta definición, aparecen otras propiedades importantes, sino que fundamentales, de la entropia de Shannon.

[P10] *Aditividad:* La entropia conjunta de dos variables aleatorias X e Y independientes se suma, y recíprocamente:

$$X \text{ e } Y \text{ independientes} \quad \Leftrightarrow \quad H(X, Y) = H(X) + H(Y)$$

Dicho de otra manera, para dos variables aleatorias, la incerteza global es la suma de las incertezas de cada variable individual. La propiedad "⇒" es consecuencia directa de $p_{X,Y}(x, y) = p_X(x)p_Y(y)$. Se va a probar en la sección siguiente la reciproca. Esta propiedad se escribe también

$$H(p_X \otimes p_Y) = H(p_X) + H(p_Y)$$

donde $\otimes$ es el producto de Kronecker [2] Se generaliza sencillamente a un conjunto de variables aleatorias $\{X_i\}$ (Kronecker product de un conjunto de vectores de probabilidades).

[P11] *Sub-aditividad:* La entropia conjunta de dos variables aleatorias $\{X_i\}_{i=1}^n$ es siempre menor que la suma de cada entropia individual:

$$H(X_1,\ldots,X_n) \leq \sum_{i=1}^{n} H(X_i) \qquad \textit{i. e.,} \qquad H\left(p_{X_1,,\ldots,X_n}\right) \leq H\left(p_{X_1}\otimes\cdots\otimes p_{X_n}\right) = \sum_{i=1}^{n} H\left(p_{X_i}\right)$$

Dicho de otra manera, variables pueden compartir información, de tal manera de que le entropia global sea menor que la suma. De la propiedad anterior, se obtiene la igualdad ssi los X_i son indepedientes.

[P12] *Super-aditividad:* La entropia conjunta de dos variables aleatorias $\{X_i\}_{i=1}^n$ es siempre major que cualquiera de las entropias individuales

$$H(X_1,\ldots,X_n) \geq \max_{1\leq i\leq n} H(X_i)$$

Es importante notar de que existen varios enfoques basados sobre una series de axiomas, dando lugar a la definición de la entropia tal como definido. Estos axiomas son conocidos como axiomas de Shannon-Khinchin y son la continuidad (propiedad [P1]), la maximalidad (propiedad [P5]), la expansabilidad (propiedad [P6]) y la aditividad (propiedad [P10]). Existen varios otros conjunto de axiomas, conduciendo también a la entropia de Shannon (ver Shannon (Shannon, 1948, Sec. 6) or (Shannon & Weaver, 1964), Rényi (Rényi, 1961), Fadeev (Fadeev, 1956, 1958), Khintchin (Khinchin, 1957) entre otros).

Para una serie de variables aleatorias, $X_1, X_2, \ldots$, representando simbolos, se puede definir una entropia por simbolo como una entropia conjunta divido por numero de simbolos, $\frac{H(X_1,\ldots,x_n)}{n}$, asá que una taza de entropia cuando n va al inifinito.

Definición 2-4 (Taza de entropia)**.** *Sea $\mathcal{X} = \{X_i\}_{i\in\mathbb{N}^*}$ una serie de variable aleatoria. La taza de entropia de esta serie es definida por*

$$\mathcal{H}(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1,\ldots,X_n)}{n}$$

Esta cantidad siempre existe, porque $H(X_1,\ldots,X_n) \leq \sum_i H(X_i) \leq \sum_i \log\alpha_i \leq n\max_i \alpha_i$ donde los α_i son los cardinales de los alfabetos de definición de los X_i.

Se termina esta sub-sección con el caso de variables discretas definidas sobre un alfabeto $\mathcal{X}$ de cardinal infinito $|\mathcal{X}| = +\infty$, por ejemplo $\mathcal{X} : \mathbb{N}$. Por analogia, se puede siempre definir la entropia como en la definición Def. 2-1. Esta extensión resuelta delicada dando de que unas propiedades se perdien. Por ejemplo, la entropia no queda acotada por arriba como se lo puede probar para la distribución de probabilidad $p(x) \propto \frac{1}{(x+2)\left(\log(x+2)\right)^2}$, $x \in \mathbb{N}$, corectamente normalizada ($\propto$ significa "proporcional a"):

---

[2] $\begin{bmatrix} p_X(x_1) & \cdots & p_X(x_\alpha)\end{bmatrix}^t \otimes \begin{bmatrix} p_Y(y_1) & \cdots & p_Y(y_\beta)\end{bmatrix}^t = \begin{bmatrix} p_X(x_1)p_Y(y_1) & \cdots & p_X(x_1)p_Y(y_\beta) & \cdots & p_X(x_\alpha)p_Y(y_1) & \cdots & p_X(x_\alpha)p_Y(y_\beta)\end{bmatrix}^t$

$\frac{\log\log(x+2)}{(x+2)\left(\log(x+2)\right)^2} \geq 0$ y la serie $\sum_x \frac{1}{(x+2)\log(x+2)}$ es divergente, así que la serie $-\sum_x p(x)\log p(x)$ diverge.

2.2.2 Entropía diferencial

Volviendo a la definición Def. 2-1 de la entropia de Shannon, usando el operador E promedio estadistica o esperanza matematica, se puede rescribir la entropia de Shannon como $H(X) = \mathrm{E}\left[-\log p_X(X)\right]$. Con este punto de vista, es facil extender la definición de la entropia para variables aleatorias continuas admitiendo una densidad de probabilidad. Eso da lugar a lo que es conocido como la *entropia diferencial*:

Definición 2-5 (Entropia diferencial)**.** *Sea X una variable aleatoria definida sobre un espacio d-dimensional $\mathcal{X} \subseteq \mathbb{R}^d$ y sea $p_X(x)$ la densidad (distribución) de probabilidad de X, La entropia diferencial de la variable X es definida por*

$$H(p_X) = H(X) = -\int_{\mathcal{X}} p_X(x)\,\log p_X(x)\,dx$$

(con la convención $0\log 0 = 0$, se puede escribir la integración en $\mathbb{R}^d$).

Como en el caso discreto, para $X = (X_1, \ldots, X_d)$, esta entropia de X es dicha entropia conjunta de los componentes X_i.

Como se lo va a ver, la entropia diferencial no tiene la misma significación de incerteza, siendo de que depende no solamente de la distribución de probabilidad, sino que de los estados también. Mas alla, no se la puede ver como límite continua de un caso discreto: a través de tal límite, se va a ver que se llama diferencial, a causa del efecto de la diferencial dx. Para ilustrar eso, considera una variable aleatoria escalar X viviendo sobre $\mathbb{R}$ y p_X su densidad de probabilidad. Sea $\delta > 0$ y sea el alfabeto $\mathcal{X}^\delta = \{x_k\}_{k\in\mathbb{Z}}$ donde los x_k se definen tal que $p_X(x_k)\delta = \int_{k\delta}^{(k+1)\delta} p_X(x)\,dx$, como ilustrado en la figure 2-7. Se define la variable aleatoria discreta X^δ sobre $\mathcal{X}^\delta$ tal que $\Pr[X^\delta = x_k] = p_{X^\delta}(x_k) = p_X(x_k)\delta$. Se puede ver X^δ como la versión cuantificada de X, con $X^\delta = x_k$ cuando $X \in [k\delta, (k+1)\delta)$. Al revés, aún que sea delicado, se puede interpretar X como el "límite" de X^δ cuando δ tiende a 0. Ahora, es claro de que

$$\begin{aligned} H(X^\delta) &= -\sum_k p_{X^\delta}(x_k)\log p_{X^\delta}(x_k) \\ &= -\log\delta - \sum_k \Big(p_X(x_k)\log p_X(x_k)\Big)\,\delta \end{aligned}$$

lo que se escribe tambien

$$H(X^\delta) + \log\delta = -\sum_k \Big(p_X(x_k)\log p_X(x_k)\Big)\,\delta$$

Entonces, de la intergración de Rieman sale que

$$\lim_{\delta\to 0}\big(H(X^\delta) + \log\delta\big) = H(X)$$

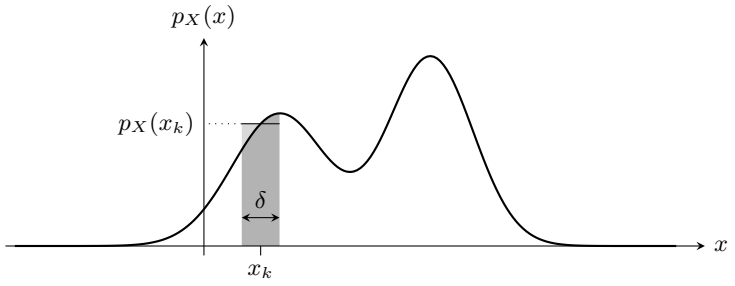


Figura 2-7: Densidad de probabilidad p_X de X, construcción del alfabeto $\mathcal{X}$ donde se define la versión cuantificada X^δ de X con su distribución discreta de probabilidad p_{X^δ}. La superficia en grise oscuro es igual a la superficia definida por el rectangulo en grise claro.

Dicho de otra manera, la entropia diferencial de X no es el limite de la entropia de su versión cuantificada: aparece con la entropia el termino "diferencial" $\log \delta$.

Mas allá de esta notable diferencia entre la entropia y la entropia diferencial, la última depende de los estados, es decir que si $Y = g(X)$ con g biyectiva, no se conserva la entropia, *i. e.,* se pierde la propiedad [P3] del caso discreto:

$$\begin{aligned} H(Y) &= -\int_{\mathbb{R}^d} p_Y(y) \log p_Y(y)\, dy \\ &= -\int_{\mathbb{R}^d} p_Y(g(x)) \log p_Y(g(x)) \,|\, \mathrm{J}_g(x)|\, dx \\ &= -\int_{\mathbb{R}^d} p_Y(g(x)) \Big(\log \left(p_Y(g(x)) \,|\, \mathrm{J}_g(x)|\right) - \log |\nabla^t g(x)| \Big) \,|\, \mathrm{J}_g(x)|\, dx \end{aligned}$$

donde J_g es la matriz de componentes $\frac{\partial g_i}{\partial x_j}$, Jacobiano de la transformación $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ ($g \equiv \Big[g_1(x_1, \ldots, x_d) \quad \cdots \quad g_d(x_1, \ldots, x_d) \Big]^t$) y $|\cdot|$ representa el valor absoluto del determinente de la matriz. Recordandose de que $p_X(x) = p_Y(g(x))|\, \mathrm{J}_g(x)|$, se obtiene

[P'3] Para cualquier biyección $g : \mathbb{R}^d \mapsto \mathbb{R}^d$

$$H(g(X)) = H(X) + \int_{\mathbb{R}^d} p_X(x) \log |\, \mathrm{J}_g(x)|\, dx$$

donde el último termino, $\mathrm{E}\left[\log |\, \mathrm{J}_g(X)|\right]$ no vale cero en general. En particular, si H es invariante bajo un deplazamiento,

$$H(X + \mu) = H(X) \quad \forall\, \mu \in \mathbb{R}^d$$

no es invariante por cambio de escala,

$$H(aX) = H(X) + \log |a| \quad \forall\, a \in \mathbb{R}^*$$

Esta última relación queda valid para a matriz invertible. Por esta última relación, se puede ver que, dado X, cuando a tiende a 0, la entropia de aX tiende a $-\infty$. Es decir que, para a suficientemente pequeño, se puede tener $H(aX) < 0$, así que se pierde también la positividad, propiedad [P4]. Esta perdida definitivamente quita

la interpretación de incerteza/información que hubiera podido tener la entropia diferencial. A veces, se usa lo que es llamado potencia entropica:

Definición 2-6 (Potencia entropica)**.** *Sea X una variable aleatoria d-dimensional. La potencia entropica de X es definida por*

$$N(X) = \frac{1}{2\pi\,\mathrm{e}} \exp\left(\frac{2}{d} H(X)\right)$$

Por construcción, $N(X) \geq 0$. Ademas, en el caso continuo, $N(aX+b) = |a|^2 N(X)$ (queda valida para una matriz a invertible): esta propiedad puede justificar la idea de "potencia"; ademas $N(aX+b)$ tiende naturalmente a cero cuando a tiende a cero. Se recupera así la noción informacional a través de N en este contexto ($aX+b$ "tiende" a b, variable deterministica).

Si se pierde la propiedad de invarianza bajo una biyección, sopredentemente, se conserva la entropia bajo el equivalente continuo del rearreglo.

Definición 2-7 (Rearreglo simetrico)**.** *Sea $\mathcal{P} \subset \mathbb{R}^d$ abierto de volumen finito $|\mathcal{P}| < +\infty$. El rearreglo simetrico $\mathcal{P}^\downarrow$ de $\mathcal{P}$ es la bola centrada en 0 de mismo volumen que $\mathcal{P}$, i. e.,*

$$\mathcal{P}^\downarrow = \left\{ x \in \mathbb{R}^d \;:\; \frac{2\pi^{\frac{d}{2}} |x|^d}{\Gamma\left(\frac{d}{2}\right)} \leq |\mathcal{P}| \right\}$$

donde $|\cdot|$ denota la norma euclideana. Eso es ilustrado figure 2-8-a.
Sea p_X una densidad de probabilidad y sea $\mathcal{P}_t = \{y \;:\; p_X(y) > t\}$ para cualquier $t > 0$, sus conjuntos de niveles. La densidad de probabilidad[3] *rearreglada simetrica $p_X^\downarrow$ de p_X es definida por*

$$p_X^\downarrow(x) = \int_0^{+\infty} \mathbb{1}_{\mathcal{P}_u^\downarrow}(x)\, du$$

con $\mathbb{1}_A$ el indicator del conjunto A, i. e., $\mathbb{1}_A(x) = 1$ si $x \in A$ y cero sino.

Del hecho de que $\forall\, t < \tau \;\Leftrightarrow\; \mathcal{P}_\tau \subseteq \mathcal{P}_t \;\Leftrightarrow\; \mathcal{P}_\tau^\downarrow \subseteq \mathcal{P}_t^\downarrow$ es sencillo ver que si $x \in \mathcal{P}_\tau^\downarrow$, entonces $x \in \mathcal{P}_t^\downarrow$, lo que conduce a $p_X^\downarrow(x) > \tau$ y vice-versa. Mas alla, sobre $\mathcal{P}_{\tau+d\tau} \backslash \mathcal{P}_\tau$ la función p_X "vale" τ y sobre $\mathcal{P}_{\tau+d\tau}^\downarrow \backslash \mathcal{P}_\tau^\downarrow$ la función $p_X^\downarrow$ "vale" tambien τ, lo que da $\int_{\mathcal{P}_\tau^\downarrow} p_X^\downarrow(x)\, dx = \int_{\mathcal{P}_\tau} p_X(x)\, dx$ (ver (Lieb & Loss, 2001; Wang & Madiman, 2004) para une prueba mas rigorosa). La representación de la definición es conocida como representación en capas de pastel (layer cake en ingles). Eso es ilustrado en la figura 2-8-b

[P'2] *invarianza bajo un rearreglo:* Sea p_X densidad de probabilidad sobre un abierto de $\mathbb{R}^d$,

$$H\left(p_X^\downarrow\right) = H(p_X)$$

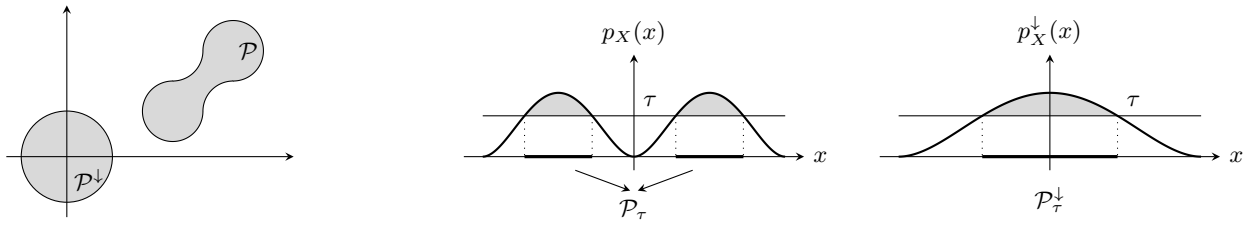
Esta propiedad es probada por ejemplo en (Lieb & Loss, 2001; Wang & Madiman, 2004).

A ver que pasa en termino de mayorizacion ???

Como se lo ha visto, la entropia diferencial no es siempre positiva, como consecuencia de [P'3]. También, la propiedad de cota superior, propiedad [P5] se pierde en general, salvo si se pone vinculos:

---

[3] Se proba de que esta función, positiva por definición, suma a 1. Ademas, por construcción, depende unicamente de $|x|$ y decrece con $|x|$.



(a) (b)

Figura 2-8: (a): Ilustración del rearreglo simetrico $\mathcal{P}^{\downarrow}$ de un conjunto $\mathcal{P}$, siendo la bola centrada en 0 de mismo volumen. (b) Construcción del rearreglo $p_X^{\downarrow}$: dado un τ, se busca $\mathcal{P}_\tau$ y se deduce $P_\tau^{\downarrow}$; dado un x, se busca el mayor t tal que $x \in P_t^{\downarrow}$, este t maximo siendo entonces $p_X^{\downarrow}(x)$; ademas, por construcción, las superficias en grise son iguales.

[P'5] *a*) Si $\mathcal{X}$ es de volumen finito $|\mathcal{X}| < +\infty$, la entropia es acotada por arriba,

$$H(X) \leq \log |\mathcal{X}|$$

con igualdad ssi X es uniforme.

b) Si $\mathcal{X} = \mathbb{R}^d$ y X tiene una matriz de covarianza dada $\Sigma_X = \mathrm{E}\left[XX^t\right]$ donde $\cdot^t$ denota la transpuesta, la entropia es también acotada por arriba,

$$H(X) \leq \frac{d}{2} \log(2\pi\,\mathrm{e}) + \frac{1}{2} \log |\Sigma_X|$$

con igualdad ssi X es gaussiana. En particular, la potencia entropica de la gaussiana vale $N(X) = |\Sigma_X|^{\frac{1}{d}}$, dando de nuevo un "sabor" de potencia a N. Como se o va a ver en este capítulo, la gaussiana juega un rol central en la teoria de la información.

En ambos casos, estas desigualdades con la distribución maximizante se obtienen resolviendo el problema de maximización de la entropia subjeto a vinculos. Se trata del caso más general en la sección 2.4.1.0.

Al final, se conservan las propiedades de concavidad [P8], de aditividad [P10] y de sub-aditividad [P11]. Es interesante de notar que de la desigualdad [P11], puramente entropica, se puede deducir la desigualdad de Hadamard, desigualdad puramente matricial: $|R| \leq \prod_i R_{i,i}$ para cualquier matriz simetrica definida positiva (viene de [P11] escrita para una gaussiana de covarianza R y tomando una exponencial de la desigualdad).

2.3 Entropia condicional, información mutua, entropia relativa

Tratando de un par de variable aleatorias X e Y, una cuestión natural que occure es de cuantificar la incerteza que queda sobre una de las variable cuando se observa la otra. Dicho de otra manera, si se mide $Y = y$, ¿que información lleva sobre X? La respuesta a esta interogación se encuentra en la noción de

entropia condicional. Si uno mide $Y = y$, la descripción estadistica de X conociendo este Y se resuma a la distribución condicional de probabilidad $p_{X|Y} = \frac{p_{X,Y}}{p_Y}$. Con esta restricción, se puede evaluar una incerteza sobre X, sabiendo de que $Y = y$,

$$H(X|Y = y) = H\left(p_{X|Y}(\cdot, y)\right)$$

Entonces, condicionalmente a la variable aleatoria Y, la incerteza va a ser el promedio estadistico sobre todos los estados Y es decir $H(X|Y) = \sum_y p_Y(y)H(X|Y = y)$:

Definición 2-8 (Entropia condicional)**.** *Sean X e Y dos variables aleatorias discretas, la entropia condicional de X sabiendo Y es definida por*

$$H(X|Y) = -\sum_{x,y} p_{X,Y}(x, y) \log p_{X|Y}(x, y)$$

Esta definición se transpone naturalmente a la entropia diferencial:

Definición 2-9 (Entropia diferencial condicional)**.** *Sean X e Y dos variables aleatorias continuas, la entropia condicional de X sabiendo Y es definida por*

$$H(X|Y) = -\int_{\mathbb{R}^d} p_{X,Y}(x, y) \log p_{X|Y}(x, y)\, dx\, dy$$

Si X e Y son indepedientes, $p_{X|Y}$ se reduce a p_X, así que vale cero la entropia condicional:

[P13]

$$X \text{ e } Y \text{ independientes} \quad \Leftrightarrow \quad H(X|Y) = H(X)$$

Esta propiedad vale en ambos casos, discreto como continuo. En el caso discreto, se interpreta como el hecho de que Y no lleva ninguna información sobre X, y intonces ninguna medición de Y va a cambiar la incerteza sobre X.

Siendo $H(X|Y = y)$ una entropia, va a heredir de todas las propiedades de la entropia (diferencial). Ademas, de $p_{X,Y} = p_{X|Y}p_Y$ de deduce la propiedad siguiente (valida para la entropia como su extensión diferencial)

[P14] *Regla de cadena*

$$H(X, Y) = H(X|Y) + H(Y)$$

Esta regla, valida en ambos casos, discreto como continuo, se generaliza sencillamente a

$$H(X_1, \ldots, X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_{i-1}, \ldots, X_1)$$

De esta regla de cadena se recupera la propiedad [P13] a partir de la propiedad [P10].

Siendo $H(X|Y = y)$ una entropia, en el caso discreto esta cantidad es positiva. Entonces, en el caso discreto, $H(X|Y)$ es positiva, lo que proba la la super-aditividad [P12].

De la regla de cadena $H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$ aparece que las cantidades $H(X|Y) - H(X)$, $H(Y|X) - H(Y)$ y $H(X, Y) - H(X) - H(Y)$ son todas iguales. Estas canditades definen lo que se llama la información mutua entre X e Y:

Definición 2-10 (Información mutua)**.** *Sean X e Y dos variables aleatorias, la información mutua entre X e Y es la cantida simetrica*

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

En el caso discreto se expresa

$$I(X;Y) = \sum_{x,y} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right)$$

y su forma diferencial, se escribe

$$I(X;Y) = \int_{\mathbb{R}^d} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right) dx\, dy$$

Las diferentes cantitades peden ser vista a través una visión ensemblista, como descrita el la figura 2-9. Este diagrama es conocido como diagrama de Venn.

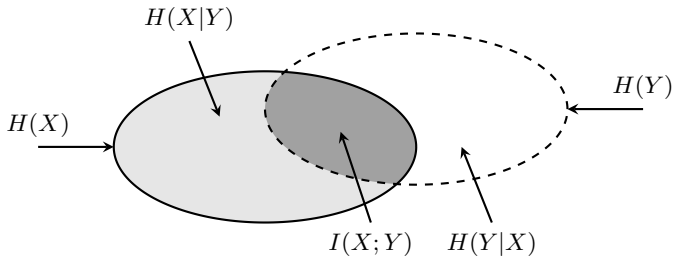


Figura 2-9: Diagrama de Venn: Ilustración de la definición de la entropia condicional, información mutua, y los vinculos entre cada medida. La superficia del elipse en linea llena (parte grise) representa $H(X)$ y el interior de la en linea punteada representa $H(Y)$. La parte grise clara representa $H(X|Y)$ superficia del "conjunto $H(X)$" quitando la parte que partenece a $H(Y)$. La parte blanca representa $H(Y|X)$ superficia del "conjunto $H(Y)$" quitando la parte que partenece a $H(X)$. La parte en grise oscuro es entonces lo que X e Y comparten, es decir $I(X;Y)$.

Como se lo va a probar, I es positiva; representa realmente una información, la compartida entre X e Y. Si de la incerteza de X se quita la incerteza de X una vez que Y es medida, lo que queda tiene la significación de la información que estas variables tienen en común. Para probar la positividad de I, se introduce de manera mas general la noción de entropia relativa, conocida también como divergencia de Kullback-Leibler (Cover & Thomas, 2006; Rioul, 2007; ?, ?): **Buscar ref de Kullback and so on**

Definición 2-11 (Entropia relativa)**.** *La entropia relativa, o divergencia de una distribución de probabilidad q, con respeco a una distribución de referencia p, donde el alfabeto de definición de p inluye lo de q, es definida como*

$$D(q\|p) = \sum_x q(x) \log\left(\frac{q(x)}{p(x)}\right)$$

o, en su forma diferencial

$$D(q\|p) = \int_{\mathbb{R}^d} q(x) \log\left(\frac{q(x)}{p(x)}\right) dx$$

Esta medida se puede ser vista como una entropia de la distribución q, relativamente a una distribución de referencia p. Por ejemplo, en el caso discreto finito, si p es la distribución uniforme sobre un alfabeto de cardinal α, $D(q\|p) = \log\alpha - H(q)$, lo que representa una desviación de la entropia con su valor maximal. La misma interpretación queda en el caso continuo con la lei uniforme (p y q definidas sobre el mismo espacio de volumen finito) o con la gaussiana (p y q dando la misma matriz de covarianza). *Como para la entropia, cuando se necesitará un logaritmo especificamente de base a, se notará la divergencia D_a.*

Lema 2-1 (Positividad de la entropia relativa)**.**

$$D(q\|p) \geq 0 \quad \textit{con igualdad ssi} \quad p = q\ (c.s.)$$

donde $(c.s.)$ *significa "casi siempre".*

Demostración. Existen varias pruebas, pero la mas linda puede ser la usando la desigualdad de Jensen: para ϕ estrictamente convexa, $\mathrm{E}[\phi(X)] \geq \phi(\mathrm{E}[X])$ con igualdad ssi X es deterministica (casi siempre). Sea X de distribución o densidad de probabilidad p. En el caso discreto como diferencial, se escribe la entropia relativa $D(q\|p) = \mathrm{E}\left[\frac{q(X)}{p(X)}\log\left(\frac{q(X)}{p(X)}\right)\right]$. Sea $Y = \frac{q(X)}{p(X)}$ y $\phi(u) = u\log u$, función estrictamente convexa. Entonces $D(q\|p) = \mathrm{E}[\phi(Y)] \geq \phi(\mathrm{E}[Y])$. Con $\mathrm{E}[Y] = \mathrm{E}\left[\frac{q(X)}{p(X)}\right] = \sum_x q(x) = 1$ (y con una integral en el caso diferencial) y $\phi(1) = 0$ sz termina la prueba. El caso de igualdad apareciendo ssi Y es deterministica, es decir $\frac{p(X)}{q(X)}$ deterministica, es equivalente a $p(x) \propto q(x)\ (c.s.)$, *i. e.,* $p = q\ (c.s.)$ porque ambas suman a uno. □

Eso es la desigualdad de Gibbs

Esta propiedad, valide en el caso discreto como continuo, tiene consecuencias, cuando se fije de que

$$I(X;Y) = D\left(p_{X,Y}\|\, p_X p_Y\right)$$

i. e., la información mutua es la divergencia de Kullback-Leibler de la distribución conjunta relativa al producto de las marginales.

[P15] I *es positiva, como medida de independencia:*

$$I(X;Y) \geq 0 \quad \text{con igualdad ssi } X \text{ e } Y \text{ son independientes}$$

[P16] *Condicionar reduce la entropia*

$$H(X|Y) \leq H(X) \quad \text{con igualdad ssi } X \text{ e } Y \text{ son independientes}$$

Esta desigualdad, con la regla de cadena, prueba la sub-aditividad [P11]. Esta reducción vale en promedio, pero el conocimiento de un valor particular puede ser tal que $H(X|Y=y) > H(X)$, *i. e.,* aumentar la entropia! (ver ejemplos en (Rioul, 2007, p. 59))

Fijense que si D es positiva, no es simetrica y tampoco satisface la desigualdad triangular. Por eso, no es una distancia y tiene el nombre de *divergencia*. La distribución de referencia p juega un rol fundamental.

2.4 Unas identidades y desigualdades

Desigualdades de Fano? Rioul p. 78, Cover P. 663, Sanov? Pythagorean? Gene: cf Zyc p60

2.4.1 El principio de entropia máxima

En la termodynamica, el estudio de las caracteristicas macroscopias (dynamica de las moleculas) es prohibitivo tan el numero de moleculas es importante. Por ejemplo, un litro del gas que respiramos contiene $2{,}7 \times 10^{22}$ moleculas. De esta constatación se desarolló la física estadisticas bajo el impulso de Boltzmann (Boltzmann, 1896, 1898), Maxwell (Maxwell, 1867), Gibbs (Gibbs, 1902), Planck (Planck, 2015) entre otros (see also (Jaynes, 1965)), considerando el sistema macroscopico a través de lo que llamaron ensembles esdadisticos: el sistema global (macroscopio) es al equilibrio pero las configuraciones (micro-estados) son flucuantes. Se una forma, se puede asociar una configuración por su frecuencia de ocurrencia (imaginando tener une infinidad de copias del sistema en el mismo estado macroscopio), es decir su probabilidad de occurencia. En este marco, la entropia, describiendo la falta de información, juega un rol fundamental. Un sistema sujeto a vinculos, como por ejemplo teniendo una energia dada, debe estar en sus estado lo mas desorganizado dandos los vinculos. En su marco, se introdujo la noción de entropia termodynamica, pero la misma es tremendamente vinculada a la entropia de Shannon (claramente, identificando las frecuencias a probabilidades de ocurrencia) [4]. En otro terminos, la distribución describiendo los micro-estados debe ser de entropia máxima, dando los vinculos. Por ejemplo, en un gas perfecto, donde las particulas no interactuan (aparte chocandose), la energia es dada por las velocidades (suma de las energias cineticas individuales). Dando una energia fija, la distribución de las velocidad debe ser de entropia máxima sujeto a la energia dada (nada mas que la energia va a "organizar" las configuraciones posibles). Intuitivamente, en un sistema aislado de N particulas, las configuraciones van a ser equiprobables, precisamente la distribución maximizando la entropia. En la sección 2.5.4.0 se va a desarollar un poco mas este ejemplo.

De manera general, el problema se formaliza como la buscada de la entropia máxima sujeto a vinculos. Si este principio nació en mecanica estadistica (ver también (Jaynes, 1957a, 1957b, 1965)), encontró un echo en varios dominio: en inferencia bayesiana para elegir distribuciones del a priori [5] conociendo unos momentos de la lei (Robert, 2007; Jaynes, 1968, 1982), hacer estimación espectral o de procesos estocasticos autoregesivos (Burg, 1967, 1975; Jaynes, 1982) o (Cover & Thomas, 2006, cap. 12), entre otros (Kapur &

[4] Ver epígrafe del capítulo...

[5] A partir de una distribución parametrizada por un parametro θ. El enfoque de bayesiano consiste a modelizar θ aleatorio, digamos Θ, tal que la distribución de observaciones se escribe entonces $p_{X|\Theta}$. Inferir θ a partir de observaciones x consiste a determinar la distribución dicha *a posteriori* $p_{\Theta|X}$. Por eso, hace falta darse una distribución dicha *a priori* p_{Θ}. Si se conocen momentos por una razon o una otra, se puede elegir esta distribución la "menos informativa" posible, *i. e.,* de entropia máxima dados los momentos.

Kesavan, 1992, & ref.).

Sea X variable aleatoria viviendo sobre $\mathcal{X} \subset \mathbb{R}^d$ con K momentos $\mathrm{E}[M_k(X)] = m_k$ fijos, con $M_x : \mathcal{X} \to \mathbb{R}$, el problema de entropia máxima se formula de la manera siguiente en el caso continuo (es el caso discreto, hay que re-emplazar integrales por sumas): sean $M(x) = \begin{bmatrix} 1 & M_1(x) & \cdots & M_K(x) \end{bmatrix}^t$ y $m = \begin{bmatrix} 1 & m_1 & \cdots & m_K \end{bmatrix}^t$, se busca,

$$p^* = \operatorname*{argmáx}_p H(p) \qquad \text{sujeto a} \qquad p \geq 0, \quad \int_{\mathcal{X}} M(x)\, p(x)\, dx = m$$

donde los dos primeros vinculos aseguran de que p^* (positividad, normalización) sea una distribución de probabilidad. En el ejemplo del gas, $K = 1, M_1(x) = \sum_i x_i^2$ (los x_i son las velocidades). Introduciendo factores de Lagrange $\lambda = \begin{bmatrix} \lambda_0 & \lambda_1 & \cdots & \lambda_K \end{bmatrix}^t$ para tener en cuenta los vinculos, el problema variacional consiste a resolver (Gelfand & Fomin, 1963; van Brunt, 2004; Miller, 2000; Cambini & Martein, 2009; Cover & Thomas, 2006)

$$p^* = \operatorname*{argmáx}_p \int_{\mathcal{X}} \left(-p(x) \log p(x) + \lambda^t M(x)\, p(x) \right) dx$$

donde λ será determinado para satisfacer los vinculos. De la ecuación de Euler-Lagrange (Gelfand & Fomin, 1963; van Brunt, 2004), esquematicamente anulando la "derivada" del integrande con respeto a p (sera realmente un gradiente los componentes de p en el caso discreto), reparametrizando los factores de Lagrange, se obtiene

$$p^*(x) = \mathrm{e}^{\lambda^t M(x)}$$

con λ tal que se satisfacen los vinculos de normalización y momentos. Esta distribución cae el la familia conocida como *familia exponencial* donde los M_k son conocidos como *esdadisticas suficientes* y los λ_k *parametros naturales* (Darmois, 1935; Koopman, 1936; Andersen, 1970; Kay, 1993; Lehmann & Casella, 1998; Robert, 2007).

Un problema que puede aparecer es que no se puede determinar λ tal que se satisfacen todos los vinculos, en particular la de normalización. Por ejemplo, si $\mathcal{X} = \mathbb{R}$ y $K = 0$, p deberia ser constante (lei uniforme) sobre... $\mathbb{R}$, lo que no es normalizable [6]. En otros terminos, en este caso, el problema no tiene solución [7].

Existe una prueba informacional de este resultado, saliendo de la solución:

Lema 2-2. *Sea* $\mathcal{P}_m = \left\{ p \geq 0 : \int_{\mathcal{X}} M_k(x)\, p^*(x)\, dx = m \right\}$ *y* $p^* \in \mathcal{P}_m$ *que sea de la forma* $p^* = \mathrm{e}^{\lambda^t M(x)}$. *Entonces*

$$\forall\, p \in \mathcal{P}_m, \quad H(p) \leq H(p^*) \qquad \textit{con igualdad ssi} \quad p = p^*$$

---

[6] En el enfoque aayesiano se puede que no sea problematico, si el a posteriori es normalizable (Robert, 2007), pero va mas allá de la meta de esta sección.

[7] Mas precisamente, existen casos en los cuales se puede acotar la entropia por arriba por un $H^{\sup}$, tal que $\sup_p H(p) \leq H^{\sup}$ pero no se puede alcanzar esta cota, *i. e.,* es un supremum, no un máximo (Cover & Thomas, 2006, sec. 12.3).

Demostración.

$$\begin{aligned} H(p) &= -\int_{\mathcal{X}} p(x)\, \log p(x)\, dx \\ &= -\int_{\mathcal{X}} p(x)\, \log \left(\frac{p(x)}{p^*(x)} \right) dx - int_{\mathcal{X}} p(x)\, \log p^*(x)\, dx \end{aligned}$$

De $\log p^* = \lambda^t M$ se obtiene

$$\begin{aligned} H(p) &= -D\left(p \,\|p^*\right) - \int_{\mathcal{X}} \lambda^t M(x) p(x)\, dx \\ &= -D\left(p \,\|p^*\right) - \int_{\mathcal{X}} \lambda^t M(x)\, p^*(x)\, dx \\ &= -D\left(p \,\|p^*\right) - \int_{\mathcal{X}} p^*(x) \log p^*(x)\, dx \\ &= -D\left(p \,\|p^*\right) + H(p^*) \end{aligned}$$

porque $p, p^* \in \mathcal{P}_m$ y $\lambda^t M = \log p^*$. La prueba se cierra notando que $D \geq 0$ con igualdad si y solamente si $p = p^*$. □

Este lema prueba que, dando vinculos "razonables", la entropia es acotada por arriba, y que se alcanza la cota para una distribución de la familia exponencial. Por ejemplo,

- Con $K = 0$ y $\mathcal{X}$ de volumen finito $|\mathcal{X}| < +\infty$, la distribución de entropia máxima es la distribución uniforme de la propiedad [P'5]*a* sección 2.2.2.0 en el caso continuo, o propiedad [P5] sección 2.2.1.0 en el caso discreto.
- Con $K = 1$, $\mathcal{X} = \mathbb{R}^d$ y $M(x) = xx^t$ (visto con d^2 vinculos), la distribución de entropia máxima es la distribución gausiana de la propiedad [P'5]*b* sección 2.2.2.0.

2.4.2 Desigualdad de la potencia entropica

Sean X e Y dos variables indepedientes. Si se sabe las relaciones entre $H(X,Y)$, $H(X)$, $H(Y)$, una pregunta natural concierna la relación que podrian tener $X + Y$ con cada variable en termino de entropia. La respuesta no es trivial, y el resultado general concierna el caso de variables continuas sobre $\mathbb{R}^d$. Es conocido como desigualdad de la potencia entropica (EPI para entropy power inequality en inglés). No vincula las entropias, sino que las potencias entropicas.

Teorema 2-1 (Desigualdad de la potencia entropica)**.** *Sean X e Y dos variables d-dimensionales continuas indepedientes, entonces*

$$N(X+Y) \geq N(X) + N(Y)$$

con igualdad sii X e Y son gaussianas con matrices de covarianza proporcionales, $\Sigma_Y \propto \Sigma_X$ (siempre verdad en el contexto escalar).

Existen varias formulaciones alternativas a esta desiguladad (Shannon, 1948; Lieb, 1978; Cover & Thomas, 2006; Dembo et al., 1991; Rioul, 2007):

1. Sean $\widetilde{X}$ y $\widetilde{Y}$ gaussianas independientes de matriz de covarianza proporcionales y tal que $H(\widetilde{X}) = H(X)$ y $H(\widetilde{Y}) = H(Y)$. Entonces
$$N(X+Y) \geq N\left(\widetilde{X}+\widetilde{Y}\right)$$
con igualdad sii X y Y son gaussianas.

2. *Desigualdad de preservación de covarianza:*
$$\forall\, 0 \leq \lambda \leq 1, \quad H\left(\sqrt{\lambda}X + \sqrt{1-\lambda}Y\right) \geq \lambda H(X) + (1-\lambda)H(Y)$$
con igualdad el el caso gaussiano con matrices de covarianza proporcionales.

La prueba de esta(s) desigualdad(es) no es trivial. Numeras versiones existen, dadas por ejemplo en las referencias (Blachman, 1965; Stam, 1959; Shannon & Weaver, 1964; Rioul, 2007, 2011, 2017; Cover & Thomas, 2006; Dembo et al., 1991; Lieb, 1978; Verdú & Guo, 2006) (ver tambien teorema 6 de (Lieb, 1975)) entre otros. Como se lo puede ver, la gaussiana juega un rol particular en esta desigualdad, saturandola.

Ver si es corto probar la equivalencia entre las tres formas. Existe una forma, de Madiman, a traver rearreglo

Esta desigualdad se usa para probar otras desigualdades, como por ejemplo la desigualdad de Minkowsky $|R_1 + R_2|^{\frac{1}{d}} \geq$ para cualquier matrices R_1, R_2 simetricas definidas positivas (viene de X e Y gaussianas de covarianza R_1 y R_2). Aparece también para acotar información mutua entre variables y calcular la capacidad de un canal de communicación como se le va a ver (Cover & Thomas, 2006; Dembo et al., 1991; Rioul, 2007; Johnson, 2004).

En el caso discreto, no hay un resultado general. Existent solamento resultados para variables particulares (?, ?, ?).

2.4.3 Desigualdad de procesamiento de datos

Esta desigualdad traduce que procesando datos, no se puede aumentar la información disponible sobre una variable. Se basa sobre una desigualdad que satisface la información mutua aplicada a un proceso de Markov.

Definición 2-12 (Proceso de Markov)**.** *Una secuencia $X_1 \mapsto X_2 \mapsto \ldots \mapsto X_n$ es dicha de Markov si para cualquier $i > 1$,*
$$p_{X_{i-1},X_{i+1}|X_i} = p_{X_{i-1}|X_i} p_{x_{i+1}|X_i}$$
Dicho de otra manera, condicionalmente a X_i, las variables X_{i-1} y X_{i+1} son independientes. Eso es equivalente a
$$p_{X_{i+1}|X_i,X_{i-1},\ldots} = p_{X_{i+1}|X_i}$$

Si i representa un tiempo, significa que la esdadistica de X_{i+1} conociendo todo el pasado se reduce a esa conociendo el pasado inmediato (las probabilidades dichas de transición $p_{X_{i+1}|X_i}$ caracterisan completamente el proceso). Es sencillo fijar de que $X_n \mapsto X_{n-1} \mapsto \ldots \mapsto X_1$ es tambien un proceso de Markov.

Teorema 2-2 (Desigualdad de procesamiento de datos)**.** *Sea $X \mapsto Y \mapsto Z$ un proceso de Markov. Entonces,*

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

con igualdad sii $X \mapsto Z \mapsto Y$ es también un proceso de Markov. En particular, es sencillo ver que para cualquier función g, $X \mapsto Y \mapsto g(Y)$ es un proceso de Markov, lo que da

$$\forall\, g, \quad I(X;Y) \geq I(X;g(Y))$$

La última desigualdad se escribe también $H(X|g(Y)) \geq H(X|Y)$ y significa que procesar Y no aumenta la información que Y da sobre X (la incerteza condicional es mas importante).

Demostración. Por definición de la información mutua, considerando X y la variable conjunta (Y,Z),

$$\begin{aligned} I(X;Y,Z) &= H(X) - H(X|Y,Z) \\ &= H(X) - H(X|Y) + H(X|Y) - H(X|Y,Z) \end{aligned}$$

Por la propiedad que $Z \mapsto Y \mapsto X$ sea también un proceso de Markov, es sencillo probar que $H(X|Y,Z) = H(X|Y)$ (conciendo Y sufice para caracterisar completamente X), lo que da

$$I(X;Y,Z) = I(X;Y)$$

También,

$$\begin{aligned} I(X;Y,Z) &= H(X) - H(X|Z) + H(X|Z) - H(X|Y,Z) \\ &= I(X;Y) + H(X|Z) - H(X|Y,Z) \end{aligned}$$

Ademas, escribiendo $\frac{p_{X|Y,Z}}{p_{X|Z}} = \frac{p_{X|Y,Z}p_{Y|Z}}{p_{X|Z}p_{Y|Z}} = \frac{p_{X,Y|Z}}{p_{X|Z}p_{Y|Z}}$ se muestra que $H(X|Z) - H(X|Y,Z)$ es una divergencia de Kullback-Leibler de $p_{X,Y|Z}$ relativamente a $p_{X|Z}p_{Y|Z}$, o información mutua $I(X;Y|Z)$ entre X e Y, condicionalmente a Z). Entonces

$$I(X;Y) = I(X;Z) + I(X;Y|Z)$$

lo que proba la desigualdad del hecho de que una divergencia sea no negativa. Ademas, se obtiene la igualdad sii $I(X;Y|Z) = 0$, es decir X e Y independientes condicionalmente a Z, lo que es la definición de que $X \mapsto Z \mapsto Y$ sea un proceso de Markov. □

2.4.4 Secunda lei de la termodinamica

Tratando de procesos de Markov, aparece el equivalente de la secunda lei de la termodinamica: un sistema aislado evolua hasta llegar su estado lo mas desorganizado.

Lema 2-3 (ver (Cover & Thomas, 2006)). *Sea* $X_1 \mapsto X_2 \mapsto \cdots \mapsto X_n \mapsto \cdots$ *un proceso de Markov, con probabilidades de transición* $p_{X_{n+1}|X_n}$ *dadas. Estas modelan el sistema, independiente de las condiciones iniciales. Sean dos distribuciones (condiciones) iniciales diferentes* p_1 *y* q_1*, conduciendo a las distribuciones* p_n *y* q_n *para* X_n*. Entonces:*

- *Para cualquier* $n \geq 1$*,*

  $$D(p_{n+1}\|q_{n+1}) \leq D(q_n\|p_n)$$

  las distribuciones p_n *y* q_n *no se "alejan" (tiende a acercarse);*

- *Si* p^* *es una distribución estacionaria,*

  $$D(p_{n+1}\|p^*) \leq D(p_n\|p^*)$$

  la distribución no se aleja de la distribución estacionaria.

- *Ademas, si los* X_n *viven sobre* $\mathcal{X}$ *de cardinal o volumen finito y si* p^* *es uniforme sobre* $\mathcal{X}$*,*

  $$H(X_{n+1}) \geq H(X_n)$$

  el sistema tiende a desorganizarse (y recuerdese de que la distribución uniforme es la de entropia máxima).

Demostración. Escribiendo $p_{n+1,n}$ y $q_{n+1,n}$ las distribuciones conjunta de (X_{n+1}, X_n) para las dos condiciones iniciales, , $p_{n+1|n}$ y $q_{n+1|n}$ las distribuciones condicionales de $X_{n+1}|X_n$ así que, $p_{n|n+1}$ y $q_{n|n+1}$ las distribuciones condicionales de $X_n|X_{n+1}$, se muestra sencillamente que $D(p_{n+1,n}\|q_{n+1,n}) = D(p_{n+1}\|q_{n+1}) + D(p_{n+1|n}\|q_{n+1|n}) = D(p_n\|q_n) + D(p_{n|n+1}\|q_{n|n+1})$. Ademas, $p_{n+1|n} = p_{X_{n+1}|X_n} = q_{n+1|n}$, conduciendo a $D(p_{n+1|n}\|q_{n+1|n}) = 0$ y entonces $D(p_{n+1}\|q_{n+1}) = D(p_n\|q_n) + D(p_{n|n+1}\|q_{n|n+1})$. $p_{n|n+1}$ no es necesariamente igual a $q_{n|n+1}$, pero la divergencia siendo nonnegativa, se obtiene la primera desigualdad. La secunda desigualdad se obtiene tomando $q_n = p^*$. Ademas, si p^* es uniforme $p^*(x) = \frac{1}{|\mathcal{X}|}$ da $D(p_n\|p^*) = -H(X_n) + \log|\mathcal{X}|$, dando la última desigualdad. □

2.4.5 Principio de incerteza entropico

2.4.6 Un foco sobre la información de Fisher

Si la entropia y las heramientas relacionadas son naturales como medida de información, no se puede resumir una distribución a una medida escalar. En el marco de la teoria de la estimación, R. Fisher introdujo una noción de información intimamente relacionada al error cuadratico en la estimación de un parametro a

partir de una variable parametrizado por este parametro (Fisher, 1922, 1925; Kay, 1993; van den Bos, 2007; Cover & Thomas, 2006; Frieden, 2004).

Mencionamos que en esta sección, se usará el logaritmo natural.

Definición 2-13 (Matriz información de Fisher parametrica)**.** *Sea X una variable aleatoria parametrizada por un parametro m-dimensional, $\theta \in \Theta \subseteq \mathbb{R}^m$, de distribución de probabilidad $p_X(\cdot;\theta)$ continua sobre $\mathcal{X} \subseteq \mathbb{R}^d$ su soporte. Asume que p_X sea diferenciable en θ sobre Θ. La matriz de Fisher, de tamanio $m \times m$ es definida por*

$$J_\theta(X) = \mathrm{E}\left[\Big(\nabla_\theta \log p_X(X;\theta)\Big)\Big(\nabla_\theta \log p_X(X;\theta)\Big)^t\right]$$

donde $\nabla_\theta = \left[\cdots \frac{\partial}{\partial\theta_i} \cdots\right]^t$ es el gradiente en θ. Es la matriz de covarianza del score parametrico $S(X) = \nabla_\theta \log p_X(X;\theta)$ (se proba que su promedio es cero), siendo $\log p_X$ la log-verosimilitud. Bajo condiciones de regularidad, se puede mostrar[8] *que $J_\theta(X)$ es también menos el promedio de la Hessiana*[9] *$\mathcal{H}_\theta$ de $\log p_X(X;\theta)$. Nota: a veces se define la información de Fisher como $\mathrm{Tr}(J)$, traza de la matriz información de Fisher.*

Como para la entropia, la matriz de Fisher se escribe generalmente $J_\theta(X)$, a pesar de que no sea función de X pero de la densidad de probabilidad. Se la notara también $J_\theta(p_X)$ según la escritura la mas conveniente.

Tomando el gradiente en x en lugar de θ da la matriz de información de Fisher no parametrica,

Definición 2-14 (Matriz información de Fisher no parametrica)**.** *Sea X una variable aleatoria de distribución de probabilidad p_X definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$ su soporte. Asuma que p_X sea diferenciable (en x). La matriz de Fisher no parametrica, $d \times d$ es definida por*

$$J(X) = \mathrm{E}\left[\Big(\nabla_x \log p_X(X)\Big)\Big(\nabla_x \log p_X(X)\Big)^t\right]$$

Es la matriz de covarianza de la función score $\nabla_x \log p_X(X)$ (se proba que su promedio también es cero) o, bajo condiciones de regularidad, menos el promedio de la Hessiana en x de la log-verosimilitud.

Es interesante notar que:

- Cuando θ es un parametro de posición, $p_X(x;\theta) = p(x-\theta)$, $\nabla_\theta \log p_X = -\nabla_x \log p_X$ y la información parametrica se reduce a la información no parametrica.
- Si X es gaussiano de matriz de covarianza Σ_X, entonces se muestra sencillamente de que $J(X) = \Sigma_X^{-1}$ (o, de una forma, inversa de la dispersión o incerteza en termino de estadisticas de orden 2).
- Es sencillo ver que, por definición $J_\theta(X)$ y $J(X)$ son simetricas y que $J_\theta(X) > 0$ y $J(X) > 0$ donde estas desiguladades significan que las matrices son definidas positivas (los autovalores son positivos). Ademas,

$$\forall\, a \neq 0, \quad J(aX) = \frac{1}{|a|^2} J(X)$$

---

[8] Es una consecuencia del teorema de la divergencia, suponiendo que los bordes del dominio de X no depende de θ y que la función score se cancela en estos bordes.

[9] Para $f : \mathbb{R}^m \mapsto \mathbb{R}$, $\mathcal{H}_\theta f$ es la matriz de componentes $\frac{\partial^2 f}{\partial\theta_i \partial\theta_j}$.

(queda valide para a matriz invertible). Esta relación da a $J(X)$ un sabor de información en el sentido de que, cuando a tiende al infinito, $J(aX)$ tiende a 0; aX tiende a ser muy diepersas así que no hay información sobre su posición.

- J_θ y J son convexas en el sentido de que para cualquier conjunto de $\lambda_k \geq 0, \sum_{k=1}^K \lambda_k = 1$ y cualquier conjunto de distribuciones $p_k,\ k = 1, \ldots, K$ (Cohen, 1968; Frieden, 2004),

$$J_\theta\left(\sum_{k=1}^K \lambda_k p_k\right) < \sum_{k=1}^K \lambda_k\, J_\theta\left(p_k\right) \qquad \text{y} \qquad J\left(\sum_{k=1}^K \lambda_k p_k\right) < \sum_{k=1}^K \lambda_k J\left(p_k\right)$$

donde $A < B$ significa que $B - A$ es definida positiva. La prueba es dada por Cohen en el caso escalar, pero se extiende sin costo adicional en el caso multivariado. Hace falta probarlo para $K = 2$ y, por recurrencia, se extiende por cualquier K. En este caso, observando que $\big(\nabla \log p\big)\big(\nabla \log p\big)^t p = \frac{\big(\nabla p\big)\big(\nabla p\big)^t}{p}$, considerando el gradiente con respeto a θ (resp. x) tratando de J_θ (resp. J), se obtiene $\sum_k \lambda_k \frac{\big(\nabla p_k\big)\big(\nabla p_k\big)^t}{p_k} - \frac{\big(\nabla \sum_k \lambda_k p_k\big)\big(\nabla \sum_k \lambda_k p_k\big)^t}{\sum_k \lambda_k p_k} = \frac{1}{\sum_k \lambda_k p_k} \sum_{k,l} \lambda_k \lambda_l \Big(\frac{p_l}{p_k}\big(\nabla p_k\big)\big(\nabla p_k\big)^t - \big(\nabla p_k\big)\big(\nabla p_l\big)^t\Big)$, lo que vale (tratamos del caso $K = 2$), $\frac{\lambda_1 \lambda_2}{p_2 p_2 (\lambda_1 p_1 + \lambda_2 p_2)}\big(p_2 \nabla p_1 - p_1 \nabla p_2\big)\big(p_2 \nabla p_1 - p_1 \nabla p_2\big)^t \geq 0$. No puede ser identicamente cero (salvo si $\lambda_1 \lambda_2 = 0$ o $p_1 = p_2 \ldots$) así que se obtiene la desigualdad sobre la matriz de Fisher integrando esta desigualdad.

Una otra interpretación de J como información es debido a la desigualdad de Cramér-Rao que la vincula a la covarianza de estimación [10] (Rao, 1945, 1992; Rao & Wishart, 1947; Cramér, 1946; Rioul, 2007; Cover & Thomas, 2006; Frieden, 2004; Kay, 1993; van den Bos, 2007). Sea X parametrizada por θ. La meta es estimar θ a partir de X. Tal estimador va a ser una función unicamente de X, lo que se escribe usualmente [11] $\widehat{\theta}(X)$ (la función no depende explicitamente de θ). Las caractericas de la calidad de un estimator es naturalmente su bias $b(\theta) = \mathrm{E}\left[\widehat{\theta}(X)\right] - \theta$ y su matriz covarianza $\Sigma_{\widehat{\theta}}$ (la varianza da la dispersión alrededor de su promedio). La desigualdad de Cramér-Rao acota por debajo esta covarianza.

Teorema 2-3 (Desigualdad de Cramér-Rao)**.** *Sea X parametrizada por θ, de densidad de soporte $\mathcal{X} \subseteq \mathbb{R}^d$ indendiente de θ y $\widehat{\theta}(X)$ un estimador de θ. Sea $b(\theta)$ su bias y $\Sigma_{\widehat{\theta}}$ su matriz de covarianza. Sea $\mathrm{J}_b(\theta)$ la matriz Jacobiana del bias b. Entonces,*

$$\Sigma_{\widehat{\theta}} - \left(I + \mathrm{J}_b(\theta)\right) J_\theta(X)^{-1} \left(I + \mathrm{J}_b(\theta)\right)^t \geq 0$$

En particular, en el caso θ escalar,

$$\sigma^2_{\widehat{\theta}} \geq \frac{(1 + b'(\theta))^2}{J_\theta(X)}$$

---

[10] De hecho, pareció esta formula también en los papeles de Fréchet y de Darmois (Fréchet, 1943; Darmois, 1945). Como citado por Fréchet, aparece que la primera versión de esta formula es mucho mas vieja y debido a K. Pearson & Filon (Pearson & Filon, 1898) en 1898; luego fue extendido por Edgeworth (Edgeworth, 1908), Fisher (Fisher, 1925) o Doob (Doob, 1936).

[11] Por ejemplo, si θ es un promedio común a los componentes de X, un estimador podría ser $\widehat{\theta} = \frac{1}{d}\sum_i X_i$.

donde b' es la derivada de b.
Tomando θ parametro de posición y $\widehat{\theta} = X$, estimador sin bias ($b = 0$), eso da lo que es conocido como la desigualdad no parametrica de Cramér-Rao y toma la expresión

$$\Sigma_X - J(X)^{-1} \geq 0$$

o, en el caso escalar,

$$\sigma_X^2 \geq \frac{1}{J(X)}$$

Ademas, en el caso no parametrico, se alcanza la cota si y solamente si X es un vector gaussiano.

Esta desigualdad acota la variaza de cualquier estimador, *i. e.,* da la varianza o error mínima que se puede esperar. Esta cota es el inverso de la información de Fisher, *i. e.,* $J_\theta(X)$ caracteriza la información que X tiene sobre θ.

Demostración. Sea $S = \nabla_\theta \log p_X$ y $\theta_0 = \mathrm{E}\left[\widehat{\theta}(X)\right] = \theta + b(\theta)$. Fijandose que $\nabla_\theta \log p_X \, p_X = \nabla_\theta p_X$, que $\widehat{\theta}$ no es función de θ, y que el soporte $\mathcal{X}$ no depende de θ, se obtiene [12]

$$\begin{aligned}
\mathrm{E}\left[S(X)\left(\widehat{\theta}(X) - \theta_0\right)^t\right] &= \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\widehat{\theta}(x)^t \, dx - \left(\int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\, dx\right)\theta_0^t \\
&= \nabla_\theta \int_{\mathbb{R}^d} p_X(x;\theta)\widehat{\theta}(x)^t \, dx - \left(\nabla_\theta \int_{\mathbb{R}^d} p_X(x;\theta)\, dx\right)\theta_0^t \\
&= \nabla_\theta\left(\theta + b(\theta)\right) - \left(\nabla_\theta 1\right)\theta_0^t \\
&= \left(I + \mathrm{J}_b(\theta)\right)^t
\end{aligned}$$

Ademas, fijandose que $\mathrm{E}\left[S(X)S(X)^t\right] = J_\theta(X)$ y $\mathrm{E}\left[\left(\widehat{\theta}(X) - \theta_0\right)\left(\widehat{\theta}(X) - \theta_0\right)^t\right] = \Sigma_{\widehat{\theta}}$, la desigualdad de Cauchy-Bunyakovsky-Schwarz [13] conduce a

$$\left(u^t\left(I + \mathrm{J}_b(\theta)\right)^t v\right)^2 = \mathrm{E}\left[u^t S(X)\left(\widehat{\theta}(X) - \theta_0\right)^t v\right]^2 \leq u^t J_\theta(X) u \, v^t \Sigma_{\widehat{\theta}} v$$

La prueba se termina tomando $u = J_\theta(X)^{-1}\left(I + \mathrm{J}_b(\theta)\right)^t v$ (recordandose que J es simetrica).
Con la elección de u, en la desigualdad de Cauchy-Bunyakovsky-Schwarz, se obtiene la igualdad cuando, $v^t J(X)^{-1} S(x) \propto v^t(x-\theta)$ para cualquier v y x, est decir $\nabla_x p_X(x) \propto J(X)(x-\theta)p_X(x)$, lo que es la ecuación diferencial que satisface (solamente) la gaussiana: en este caso, se verifica a posteriori que $J(X) = \Sigma_X^{-1}$, y entonces que se alcanza la cota de la Cramér-Rao no parametrica. □

En el caso parametrico, no se puede estudiar el caso de igualdad del hecho de que $\widehat{\theta}$ no es algo dado. Ademas, aún dado un estimador (independiente explicitamente de θ), no hay garantia de que existe una

---

[12] Se supone que los integrandes sean θ-localmente integrables, tal que se puede invertir derivada en θ e integración.

[13] De hecho, fue probada por Cauchy para sumas en 1821, para integrales por Bunyakovsky en 1859 y mas elegamente por Schwarz en 1888 (Steele, 2004).

densidad parametrizado por θ que alcanza la cota, o al revés, dado una familia de densidades, tampoco no hay garantia que existe un estimador que permite alcanzar la cota (Cover & Thomas, 2006; Kay, 1993).

Fijense de que, de nuevo, la gaussiana juega un rol particular en la desigualdad de Cramér-Rao no parametrica , permitiendo alcanzar la cota.

Nota: para dos matrices $A \geq 0$ y $B \geq 0$, si $A - B \geq 0$ entonces $|A| \geq |B|$, con igualdad si y solamente si $A = B$ (Magnus & Neudecker, 1999, cap. 1, teorema 25). Entonces, de las desigualdades de Carmér-Rao se deducen desigualdades de Cramér-Rao escalares

$$\left|\Sigma_{\widehat{\theta}}\right| \geq \frac{\left|I + \mathrm{J}_b(\theta)\right|^2}{|J_\theta(X)|} \qquad \text{y} \qquad |\Sigma_X| \geq \frac{1}{|J(X)|}$$

Obviamente, en la secunda, se alcanza la igualdad si y solamente si X es gaussiano. Ademas, para una matriz $A \geq 0$, existe la "relación determinente-traza" $|A|^{\frac{1}{d}} \leq \frac{1}{d}\operatorname{Tr}(A)$, con igualdad si y solamente si $A = I$ (Magnus & Neudecker, 1999, cap. 11, sec. 4), dando otras versiones escalares de la desigualdad de Cramér-Rao, por ejemplo

$$|\Sigma_X|^{\frac{1}{d}} \geq \frac{d}{\operatorname{Tr}(J(X))}, \qquad \operatorname{Tr}(\Sigma_X) \geq \frac{d}{|J(X)|^{\frac{1}{d}}} \qquad \text{o} \qquad \operatorname{Tr}(\Sigma_X) \geq \frac{d^2}{\operatorname{Tr}(J(X))}$$

En estos casos, se obtiene la igualdad si y solamente si X es gaussiana (igualdad de la Cramér-Rao no parametrica) y ademas de covarianza proporcional a la identidad (igualdad en la relación determinente-traza).

Se notará que, al imagen de las leyes de entropia máxima, la información de Fisher juega también un rol particular en la inferencia bayesiana a través del prior de Jeffrey (Jeffrey, 1946; Lehmann & Casella, 1998; Robert, 2007) [14].

Si la desigualdad de Cramér-Rao da a la matriz de Fisher un sabor de información, aparece que J es también relacionado a la entropia relativa (Cover & Thomas, 2006; Frieden, 2004):

Teorema 2-4 (Fisher como curvatura de la entropia relativa)**.** *Sea X parametrizado por $\theta_0 \in \Theta$ con Θ conteniendo un vecinaje de θ_0. Siendo $D\left(p_X(\cdot;\theta) \,\|\, p_X(\cdot;\theta_0)\right)$ función de $\theta \in \Theta$, aparece que*

$$D\left(p_X(\cdot;\theta) \,\|\, p_X(\cdot;\theta_0)\right) = \frac{1}{2}\left(\theta - \theta_0\right)^t J_{\theta_0}(X)\left(\theta - \theta_0\right) + o\left(\|\theta - \theta_0\|\right)$$

donde $o(\cdot)$ es un resto pequeño con respecto a su argumento. En otros terminos, $J_{\theta_0}(X)$ es la curvatura de la entropia relativa en θ_0.

Demostración. La relación es consecuencia de un desarrollo de Taylor al orden 2 de la función $D\left(p_X(\cdot;\theta) \,\|\, p_X(\cdot;\theta_0)\right)$ de θ, tomada en $\theta = \theta_0$. Por propiedad de D, la divergencia es positiva y se cancela cuando $\theta = \theta_0$. Entonces, el primer termino del desarrollo vale cero y el secundo también, D siendo mínima

[14] Ver nota de pie 5. A veces, se toma como distribución a priori $p_\Theta(\theta) \propto |J_\theta(X)|^{\frac{1}{2}}$ por su invarianza por reparametrización $\eta = \eta(\theta)$, *i. e.,* el prior de Jeffrey en η es uniquevocamente obtenido con la Fisher en η o por cambio de variables saliendo de p_Θ.

en $\theta = \theta_0$. Ademas,

$$\begin{aligned}
\nabla_\theta D\left(p_X(\cdot;\theta) \,\|\, p_X(\cdot;\theta_0)\right) &= \nabla_\theta \int_{\mathcal{X}} p_X(x;\theta) \log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right) dx \\
&= \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta) \log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right) dx + \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta)\, dx \\
&= \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta) \log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right) dx + \nabla_\theta \int_{\mathcal{X}} p_X(x;\theta)\, dx \\
&= \int_{\mathcal{X}} \nabla_\theta p_X(x;\theta) \log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right) dx
\end{aligned}$$

la último ecuación como consecuencia de que p_X suma a 1. Entonces,

$$\mathcal{H}_\theta D\left(p_X(\cdot;\theta) \,\|\, p_X(\cdot;\theta_0)\right) = \int_{\mathcal{X}} \mathcal{H}_\theta p_X(x;\theta) \log\left(\frac{p_X(x;\theta)}{p_X(x;\theta_0)}\right) dx + \int_{\mathcal{X}} \frac{\nabla_\theta p_X(x;\theta)\, \nabla_\theta^t p_X(x;\theta)}{p_X(x;\theta)}\, dx$$

Tomado en $\theta = \theta_0$ el primer vale cero. En el secundo se reconoce $J_\theta(X)$, lo que termina la prueba. □

Este teorema, ilsutrado en la figura 2-10, vincula claramente dos objectos viniendo de la teoria de la estimación y la teoria de la información, mundos a priori diferentes. Como se lo puede ver en la figura cuando $J_\theta(X)$ tiene pequeñas autovalores (figura (a)), p_θ se "aleja" lentamente de θ_0 cuando θ se aleja de θ_0: hay una alta incerteza o pequeã información sobre θ_0. Y vice-versa (figura (b)).

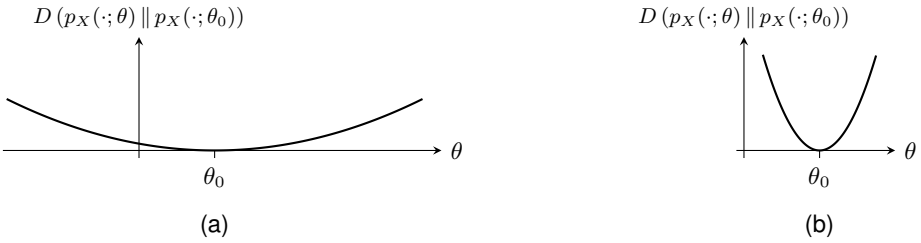


Figura 2-10: Caso escalar $\Theta \subseteq \mathbb{R}$ (para la representación) de D en función de θ. (a) Caso con $J_{\theta_0}(X)$ "pequeño" y (b) caso con $J_{\theta_0}(X)$ "grande". En el caso (b) la determinación de θ usando D va a ser mas "sencillo" porque el mínimo es mas "picado".

Un otro vinculo entre el mundo de la información y la estimación aparece a través de la identidad de de Bruijn [15] (Stam, 1959; Cover & Thomas, 2006; Johnson, 2004; Barron, 1984, 1986; Palomar & Verdú, 2006). Esta identidad caracterisa lo que es conocido como canal gaussiano figura 2-11-(a), *i. e.,* la salida Y es una versión ruidosa de la entrada. La identidad vincula las variaciones de entropia de salida con respeto al nivel de ruido, y la información de Fisher.

[15] A pesar de que tomó este nombre, esta identidad en su primera versión fue publicada por Stam. En su papel (Stam, 1959), menciona que esta identidad fue comunicada al Profesor van Soest por el Profesor de Bruijn.

Teorema 2-5 (Identidad de de Bruijn)**.** *Sea X un vector aleatorio continuo sobre un abierto $\mathbb{R}^d$ y admitiendo una matriz de covarianza, y sea $Y = X + T\mathcal{N}$ donde T es deterministica, $d \times d'$ con $d \leq d'$, de rango máximo, y $\mathcal{N}$ un vector gaussiano centrado y de covarianza $\Sigma_{\mathcal{N}}$, independiente de X (ver figura 2-11-(a)). Entonces, la entropia de Shannon y la información de Fisher de Y satisfacen*

$$\nabla_T H(Y) = J(Y)\, T\, \Sigma_{\mathcal{N}}$$

donde $\nabla_T \cdot$ es la matriz de componentes $\frac{\partial \cdot}{\partial T_{i,j}}$. Si $T = T(\theta)$ depende de un parametro escalar[16] *θ,*

$$\frac{\partial}{\partial \theta} H(Y) = \mathrm{Tr}\left(J(Y)\, T\, \Sigma_{\mathcal{N}}\, \frac{\partial T^t}{\partial \theta} \right)$$

Demostración. La clave de este resultado viene del hecho de que la densidad p de $T\mathcal{N}$ satisface una ecuación diferencial particular. La distribución de $T\mathcal{N}$ se escribe $p(x) = (2\pi)^{-\frac{d}{2}} \left|T\Sigma_{\mathcal{N}}T^t\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}x^t \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} x\right)$ (el rango máximo de T asegura que $T\Sigma_{\mathcal{N}}T^t$ sea invertible). Para una matriz invertible R, desarollando $|R|$ con respecto a su linea i, se obtiene que $\frac{\partial |R|}{\partial R_{i,j}} = R^*_{i,j}$ cofactor de $R_{i,j}$, dando por la regla de Cramér $\nabla_R |R| = |R| \left(R^{-1}\right)^t$ (ver también (Magnus & Neudecker, 1999, cap. 1 & 9)), es decir $\nabla_R |R|^{-\frac{1}{2}} = -\frac{1}{2}|R|^{-\frac{1}{2}} \left(R^{-1}\right)^t$. De $\frac{\partial |R|^{-\frac{1}{2}}}{\partial T_{i,j}} = \sum_{k,l} \frac{\partial |R|^{-\frac{1}{2}}}{\partial R_{k,l}} \frac{\partial R_{k,l}}{\partial T_{i,j}} = -\frac{1}{2}|R|^{-\frac{1}{2}} \sum_{k,l} \left(R^{-1}\right)_{l,k} \frac{\partial R_{k,l}}{\partial T_{i,j}}$ con $R = T\Sigma_{\mathcal{N}}T^t$ (simetrica) y calculos basicos se obtiene finalmente

$$\nabla_T \left|T\Sigma_{\mathcal{N}}T^t\right|^{-\frac{1}{2}} = -\left|T\Sigma_{\mathcal{N}}T^t\right|^{-\frac{1}{2}} \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} T\Sigma_{\mathcal{N}}$$

Ademas, de $\left(T\Sigma_{\mathcal{N}}T^t\right)\left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} = I$ viene $\frac{\partial \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1}}{\partial T_{i,j}} = -\left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} \frac{\partial \left(T\Sigma_{\mathcal{N}}T^t\right)}{\partial T_{i,j}} \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1}$ donde e_i es el vector con 1 en su componente i y cero si no, dando

$$\begin{aligned} \frac{\partial \left(x^t \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} x\right)}{\partial T_{i,j}} &= -x^t \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} \left(e_i e_j^t \Sigma_{\mathcal{N}} T^t + T\Sigma_{\mathcal{N}} e_j e_i^t\right) \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} x \\ &= -2\, e_i^t \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} x x^t \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} T\Sigma_{\mathcal{N}} e_j \end{aligned}$$

usando $x^t A e_k e_l^t B x = e_l^t B x x^t A e_k = e_k^t A^t x x^t B^t e_l$ (escalares comutan y un escalar es igual a su transpuesta) y usando la simetria de $T\Sigma_{\mathcal{N}}T^t$. Eso significa que

$$\nabla_T \left(x^t \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} x\right) = -2 \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} x x^t \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} T\Sigma_{\mathcal{N}},$$

dando

$$\nabla_T p(x) = \left(-\left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} + \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1} x x^t \left(T\Sigma_{\mathcal{N}}T^t\right)^{-1}\right) T\Sigma_{\mathcal{N}}\, p(x)$$

Tomando la Hessiana de p con respeto a x se obtiene sencillamente que p satisface la ecuación diferencial

$$\nabla_T p = \mathcal{H}_x p\, T\, \Sigma_{\mathcal{N}}$$

---

[16] Si el parametro es multivariado, hace falta entender la desigualdad a través de deriva parciales con respeto a los componentes de θ.

Suponiende que se puede intervertir derivadas y integrales (ver (Barron, 1984, 1986) donde se dan condiciones rigorosas), $p_Y(y) = \int_{\mathbb{R}^d} p_X(x)p(y-x)\,dx$ satisface también la ecuación diferencial, y ademas

$$\begin{aligned}
\nabla_T H(Y) &= -\int_{\mathbb{R}^d} \nabla_T\, p_Y(y) \log p_Y(y)\, dy - \int_{\mathbb{R}^d} \nabla_T\, p_Y(y)\, dy \\
&= -\left(\int_{\mathbb{R}^d} \mathcal{H}_y\, p_Y(y) \log p_Y(y)\, dy\right) T\, \Sigma_{\mathcal{N}} - \nabla_T \int_{\mathbb{R}^d} p_Y(y)\, dy \\
&= -\left(\int_{\mathbb{R}^d} \left(\mathcal{H}_y\Big(p_Y(y)\log p_Y(y)\Big) - \mathcal{H}_y\, p_Y(y) - \frac{\nabla_y\, p_Y(y)\, \nabla_y\, p_Y(y)^t}{p_Y(y)}\right) dy\right) T\, \Sigma_{\mathcal{N}} \\
&= -\left(\int_{\mathbb{R}^d} \mathcal{H}_y\Big(p_Y(y)\log p_Y(y)\Big)\, dy - \int_{\mathbb{R}^d} \mathcal{H}_y p_Y(y)\, dy\right) T\, \Sigma_{\mathcal{N}} \;+\; J(Y)\, T\, \Sigma_{\mathcal{N}}
\end{aligned}$$

usando la ecuación diferencial en la secunda linea, el hecho de que p_Y suma a 1 en la tercera linea (su gradiente es cero entonces), y la definición de la matriz de Fisher en la última linea. Usando el teorema de la divergencia (intergración por partes) aplicada respectivamente a los componentes de $\nabla_y p_Y \log p_Y$ y $\nabla_y p_Y$, suponiendo que estos gradientes se cancelan en el borde del dominio de integración, los dos terminos integrales valen cero, lo que cierra la prueba de la desigualdad general. Ademas, si $T = T(\theta)$, la secunda desigualdad sigue de $\frac{\partial \cdot}{\partial \theta} = \sum_{i,j} \frac{\partial \cdot}{\partial T_{i,j}} \frac{\partial T_{i,j}}{\partial \theta} = \mathrm{Tr}\left(\nabla_T \frac{\partial T^t}{\partial \theta}\right)$. □

La versión inicial de la identidad de de Bruijn, con $\Sigma_{\mathcal{N}} = I$,

$$\frac{d}{d\theta} H(X + \sqrt{\theta}\mathcal{N}) = \frac{1}{2}\,\mathrm{Tr}\left(J(X + \sqrt{\theta}\mathcal{N})\right)$$

se recupera en el caso particular $T = \sqrt{\theta} I$. En este caso, la ecuacíon diferencial satisfada por la densidad de probabilidad p es la *ecuación del calor*. Esta desigualdad cuantifica las variaciones de entropias bajo varaciones de "niveles" del ruido del canal de comunicación. De una forma, caracteriza la robustez del canal con respeto al nivel de ruido gaussiano (la gaussiana juega de nuevo un rol central acá).

Existe una otra forma muy similar de esta desigualdad debido a Guo, Shamai, Verdú, Palomar (Guo, Shamai & Verdú, 2005; Palomar & Verdú, 2006). Esta versión vincula aún mas los mundo de la información y el de la estimación. Del lado de la comunicación, consiste a caracterisar la información mutua entre la entrada X de un canal ruidoso y su salida, $Y = SX + \mathcal{N}$ donde S coresponde a un pre-tratamiento antes de la salida, figura 2-11-(b). Del lado de la estimación, uno puede querer estimar X observando solamente Y. Es conocido que el estimador que minimiza el error cuandratico promedio $\mathrm{E}\left[\left\|\widehat{X}(Y) - X\right\|^2\right]$ es la esperanza condicional $\widehat{X}(Y) = \mathrm{E}[X|Y]$. Una caracteristica de un estimador siendo su matriz de covarianza, se notará $\mathcal{E}(X|Y) = \mathrm{E}\left[(X - \mathrm{E}[X|Y])\,(X - \mathrm{E}[X|Y])^2\right]$ esta matriz. Sorpredentemente, existe también una identidad entre $I(X;Y)$ y $\mathcal{E}(X|Y)$:

Teorema 2-6 (Identidad de Guo–Shamai–Verdú)**.** *Sea X un vector aleatorio continuo sobre un abierto $\mathbb{R}^{d'}$ y admitiendo una matriz de covarianza, y sea $Y = SX + \mathcal{N}$ donde S es deterministica, $d \times d'$, y $\mathcal{N}$ un vector gaussiano centrado y de covarianza $\Sigma_{\mathcal{N}}$, independiente de X (ver figura 2-11-(b)). Entonces, la información*

mutua entre X *e* Y *y la matriz de covarianza del estimador de error cuadratico mínimo satisfacen*

$$\nabla_S I(X;Y) = \Sigma_{\mathcal{N}}^{-1} S\, \mathcal{E}(X|Y)$$

Si $S = S($
$sigma)$ *depende de un parametro escalar* σ*,*

$$\frac{\partial}{\partial \sigma} I(X;Y) = \mathrm{Tr}\left(\Sigma_{\mathcal{N}}^{-1}\, S\, \mathcal{E}(X|Y)\, \frac{\partial S^t}{\partial \sigma}\right)$$

Demostración. Notando que $p_{Y|X}(y,x) = (2\pi)^{-\frac{d}{2}} \left|\Sigma_{\mathcal{N}}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-Sx)^t \Sigma_{\mathcal{N}}^{-1}(y-Sx)\right)$ viene $\nabla_S p_{Y|X}(x,y) = p_{Y|X}(x,y)\, \Sigma_{\mathcal{N}}^{-1}(y-Sx)x^t$ (ver unos pasos de la prueba de la identidad de de Bruijn) así que $\nabla_y p_{Y|X}(y,x) = p_{Y|X}(y,x)\, \Sigma_{\mathcal{N}}^{-1}(y-Sx)$, dando

$$\nabla_S p_{Y|X}(y,x) = \nabla_y p_{Y|X}(y,x) x^t \qquad \text{y} \qquad \nabla_S p_{X,Y}(x,y) = \nabla_y p_{X,Y}(x,y) x^t$$

(multiplicando ambos lados por p_X. Ahora, $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\mathcal{N})$ (de la independencia, cuando $X = x$, $Y = Sx + \mathcal{N}$ gaussiana de misma convarianza que $\mathcal{N}$ y de promedio Sx), así que

$$\begin{aligned}
\nabla_S I(X;Y) &= \nabla_S H(Y) \\
&= -\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \nabla_S \Big(p_{X,Y}(x,y)\, \log p_Y(y) \Big)\, dx\, dy \\
&= -\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \nabla_S\, p_{X,Y}(x,y)\, \log p_Y(y)\, dx\, dy - \int_{\mathbb{R}\times\mathbb{R}} p_{X|Y}(x,y)\, \nabla_S\, p_Y(y)\, dx\, dy \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \nabla_y p_{X,Y}(x,y)\, x^t \log p_Y(y)\, dx\, dy - \int_{\mathbb{R}^d} \nabla_S p_Y(y)\, dy \\
&= -\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \nabla_y p_Y(y) x^t p_{X|Y}(x,y)\, dx\, dy \\
&= -\int_{\mathbb{R}^d} \nabla_y p_Y(y)\, \mathrm{E}\left[X^t | Y = y\right]\, dy
\end{aligned}$$

La secunda linea viene de la escritura de $H(Y)$ usando p_Y como marginale de $p_{X,Y}$ en x y intercambiando gradiente e integral (ver pasos de la prueba de la desigualdad de de Bruijn); la tercera de $p_{X,Y}/p_Y = p_{X|Y}$; en la cuarta se usa la ecuación diferencial satisfecha por $p_{X,Y}$ en la primera integral y integrando en x en la secunda integral; la quinta linea se obtiene usando el teorema de la divergencia (intergración por partes) en la integración en y de la primera integral, e intercambiando gradiente e integral el la secunda (p_Y sumando a 1, el termino se cancela). Ademas,

$$\begin{aligned}
\nabla_y p_Y(y) &= \int_{\mathbb{R}^{d'}} \nabla_Y p_{Y|X}(y,x)\, p_X(x)\, dx \\
&= -\Sigma_{\mathcal{N}}^{-1} \int_{\mathbb{R}^{d'}} (y - Sx)\, p_{Y|X}(y,x)\, p_X(x)\, dx \\
&= -\Sigma_{\mathcal{N}}^{-1} \left(y - S \int_{\mathbb{R}^{d'}} x\, p_{X|Y}(x,y)\, dx \right) p_Y(y) \\
&= -\Sigma_{\mathcal{N}}^{-1} \Big(y - S\, \mathrm{E}\left[X | Y = y\right] \Big)\, p_Y(y)
\end{aligned}$$

escribiendo $p_{Y|X}(y,x)\,p_X(x) = p_{X|Y}(x,y)\,p_Y(y)$ en la tercera linea. Esta ecuación permite escribir

$$\begin{aligned}
\nabla_S I(X;Y) &= \Sigma_{\mathcal{N}}^{-1} \int_{\mathbb{R}^d} \Big(y - S\,\mathrm{E}\,[X|Y=y]\Big)\,\mathrm{E}\left[X^t|Y=y\right]\,p_Y(y)\,dy \\
&= \Sigma_{\mathcal{N}}^{-1} \left(\mathrm{E}\left[Y\,\mathrm{E}\left[X^t|Y\right]\right] - S\,\mathrm{E}\left[\mathrm{E}\,[X|Y]\,\mathrm{E}\,[X|Y]^t\right]\right) \\
&= \Sigma_{\mathcal{N}}^{-1} \left(\mathrm{E}\left[YX^t\right] - S\,\mathrm{E}\left[\mathrm{E}\,[X|Y]\,\mathrm{E}\,[X|Y]^t\right]\right) \\
&= \Sigma_{\mathcal{N}}^{-1} S \left(\mathrm{E}\left[XX^t\right] - \mathrm{E}\left[\mathrm{E}\,[X|Y]\,\mathrm{E}\,[X|Y]^t\right]\right)
\end{aligned}$$

la última linea viniendo de $Y = SX + \mathcal{N}$ con $\mathcal{N}$ independiente de X y de promedio 0. La prueba se cierra notando que $\mathrm{E}\,[\mathrm{E}[X|Y]] = \mathrm{E}[X]$ y por la formula de König-Huyggens. La secunda identidad viene de $\frac{\partial \cdot}{\partial \sigma} = \mathrm{Tr}\left(\nabla_S \frac{\partial S^t}{\partial \sigma}\right)$ (ver prueba de la identidad de de Bruijn). □

La primera versión de esta identidad se recupera con $S = \sqrt{s}$, $\Sigma_{\mathcal{N}} = I$ y X de covarianza identidad; s es conocido como relación señale/ruido es este caso.

Existen versiones aún mas completas (con gradientes con respeto a la matriz $\Sigma_{\mathcal{N}}$ por ejemplo) que se pueden consultar en (Johnson, 2004; Palomar & Verdú, 2006; Payaró & Palomar, 2009).

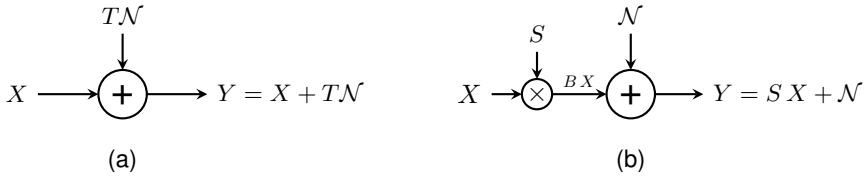


Figura 2-11: Canal de comunicación gaussiano de entrada X. (a) Canal gaussiano usual, donde T maneja los parametros (nivel) del ruido. (b) canal con gaussiano con un preprocesamiento S de la entrada.

De la desigualdad de la potencia entropica y de la identidad de de Bruijn surge una otra desigualdad implicando la potencia entropica N y la información de Fisher J. Esta desigualdad es conocida como desigualdad de Stam [17] (Cover & Thomas, 2006; Rioul, 2007; Stam, 1959), o a veces "desigualdad isoperimetrica para la entropia" (Wang & Madiman, 2004).

Teorema 2-7 (Desigualdad de Stam)**.** *Sea X una variable aleatoria continua sobre $\mathcal{X} \subseteq \mathbb{R}^d$. Entonces,*

$$N(X)\,\mathrm{Tr}\,(J(X)) \;\geq\; d$$

con igualdad si y solamente si X es gaussiano de covarianza proporcional a la identidad.

[17] Como por la identidad de de Bruijn, stam mencionó que esta desigualdad fue comunicada al Profesor van Soest por el Profesor de Bruijn quien da una prueba variacional de la desigualdad.

Demostración. De la desigualdad de la potencia entropica se obtiene $N(X+\sqrt{\theta}\mathcal{N}) \geq N(X) + \theta\,|\Sigma_{\mathcal{N}}|^{\frac{1}{d}}$. Tomando $\Sigma_{\mathcal{N}} = I$, se obtiene $\forall\,\theta > 0,\ \frac{N(X+\sqrt{\theta}\mathcal{N})-N(X)}{\theta} \geq$. Entonces, tomando el limite $\theta \to 0$, aparece que $\frac{d}{d\theta}N(X+\sqrt{\theta}\mathcal{N})\Big|_{\theta=0} \geq 1$. La prueba se cierra con $\frac{d}{d\theta}N(X+\sqrt{\theta}\mathcal{N}) = \frac{1}{2\pi\,\mathrm{e}}\frac{d}{d\theta}\exp\left(\frac{2}{d}H(X+\sqrt{\theta}\mathcal{N})\right) = \frac{2}{d}N(X+\sqrt{\theta}\mathcal{N})\frac{d}{d\theta}H(X+\sqrt{\theta}\mathcal{N}) = dN(X+\sqrt{\theta}\mathcal{N})\,\mathrm{Tr}\left(J(X+\sqrt{\theta}\mathcal{N})\right)$ (por la identitad de de Bruijn). Ademas, la igualdad se obtiene cuando se obtiene la igualdad en la desigualdad de la potencia entropica, es decir cuando X es gaussiano de varianza proporcional a la del ruido, que es la identidad en este caso. □

Se puede ver de nuevo el rol central que juega la gaussiana en esta desigualdad. Ademas, de la desigualdad de Stam se puede deducir tambíen las versiones escalares de la desigualdad Cramér-Rao. Viene del hecho de que, dado una matriz de covarianza, le entropia $H(X)$ es maxima cuando X es gaussiano. Entonces, para cualquier X de covarianza Σ_X, $N(X) \leq |\Sigma_X|^{\frac{1}{d}}$, dando de la desiguldad de Stam, $|\Sigma_X|^{\frac{1}{d}}\,\mathrm{Tr}\,(J(X)) \geq d$ (y las otras versiones escalares de la relación determinente-traza). Como se lo puede esperar, se obtiene la igualdad si y solamente X es gaussiana (potencia entropica alcanzando su cota superior) y de matriz la identidad (desiguladad de Stam se saturando).

Varias otras pruebas de la desigualdad de Stam pueden venir de generalizaciones, por ejemplo debido a Lutwak o Bercher (?, ?, ?). **La sección ZZZ lo va a rapidamente evocar.**

(1) Existe un data proc ineq con Fisher, cf Rioul 07 ou Stam 59 ou Frieden 04; cf aussi si $I_\theta(g(X)) \leq I_\theta(X)$ used in Kagan-Smith 1999 ; (2) ver MinFisher Frieden p. 235, Berchet Vignat 2009, Ernst 2017; cf. travaux rederivant MQ de Frieden-Plastino-Soffer (1999, 2002), Reginato 98, Bickel 81

2.5 Unos ejemplos y aplicaciones

2.5.1 Canal de transmisión y su capacidad

Siguiendo el esquema de comunicación de Shannon, un mensaje que se modelisa como un vector aleatorio [18] X pasa por un canal de comunicación y se recibe un mensaje Y, vector aleatorio. En el trabajo de Shannon, el canal es supuesto a ruido additiva, es decir que se añade un ruido a X. De manera general, para conocer la información de X que se recibe, se calcula la información mutua $I(X;Y)$, es decir la cuantidad de información que comparten la entrada y la salida del canal. Lo mas I es grande, lo mas de información se transmite. Dado el canal, se puede arreglar X (su distribución) de manera a maximizar $I(X;Y)$, es de-

[18] De punto de vista de un receptor, este mensaje es desconocido. Ademas, se lo puede ver como una instancia de una clase importante de posibles mensajes, justificando la modelisación aleatoria.

cir la cantidad maxima que se puede transmitir en este canal. Es lo que es conocido como capacidad del canal (Shannon, 1948, part. II & III) (ver también (Cover & Thomas, 2006; Rioul, 2007) entre otros):

Definición 2-15 (Capacidad de canal)**.** *Sea un canal de transmición, X su entrada e Y su salida, como ilustrado figura 2-12. Sea p_X la distribución de probabilidad de X. La capacidad C del canal es definida por*

$$C = \max_{p_X} I(X;Y)$$

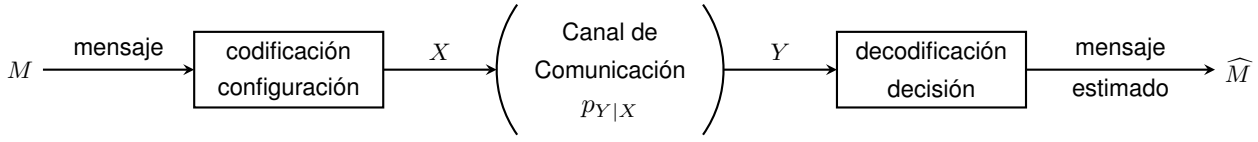


Figura 2-12: Esquema de comunicación de Shannon. En una primera etapa, un mensaje M a transmitir es códificado (ej. código binario) o puesto en forma (ej. simbolos modulando una función para que sea analógica y en una banda frecuencial dada). Sea X este mensaje codificado o puesto en forma. A la recepción, se mide Y (ej. versioón ruidosa de X), antes de ser decodificado o usado para tomar una decisión, $\widehat{M}$ siendo la estimación de M (ej. simbolos estimados a partir de Y). Una etapa importante es el vinculo entre la entrada X y la salida Y del canal, es decir la cantidad de información que tienen en común. La capacidad del canal es la información $I(X;Y)$ máxima con respeto a su entrada.

2.5.1.1. Canal binario

Suponiendo que el mensaje mandado en un canal es una cadena de simbolos, variables aleatorias independientes, se puede concentrarse sobre cada simbolo. En este marco, un canal de comunicación lo mas simple es conocido como *canal binario* (Shannon, 1948, Sec. 15): X es una variable definida sobre $\mathcal{X} = \{0, 1\}$; tal tipo de entrada es natural, pensando a la códificación binaria. La salida Y es también definida sobre $\mathcal{X}$; se puede imaginar medir y tomar una decisión binaria usando la medida. Tal canal es definido por sus probabilidad de transición $p_{Y|X}$, *i. e.,* las probabilidades que un 0 (resp. un 1) se transmite corectamente o cambia en un 1 (resp. 0), *i. e.,*

$$p = \Pr[Y = 1|X = 0] = 1 - \Pr[Y = 0|X = 0] \qquad \text{y} \qquad q = \Pr[Y = 0|X = 1] = 1 - \Pr[Y = 1|X = 1]$$

p y q representan errores de comunicación. Tal canal es descrito figura 2-13-(a). La figura 2-13-(b) da un esquema "físico" que puede se al origen de tal canal. Cuando $p = q$, el canal es conocido como *canal binario simetrico*. Cuando $p = 0$ y $q \in (0\,;\,1)$, el canal es conocido como *canal binario en Z*.

En este caso, trabajando con bits, el logaritmo lo mas lógico es el de base 2. Sea

$$\alpha = \Pr[X = 0]$$

dando la distribución de la entrada. La distribución de la salida va a ser dada a partir de $\beta = \Pr[Y = 0] = \Pr[Y = 0|X = 0]\Pr[X = 0] + \Pr[Y = 0|X = 1]\Pr[X = 1]$ es decir

$$\beta = \Pr[Y = 0] = q + \alpha(1 - p - q)$$

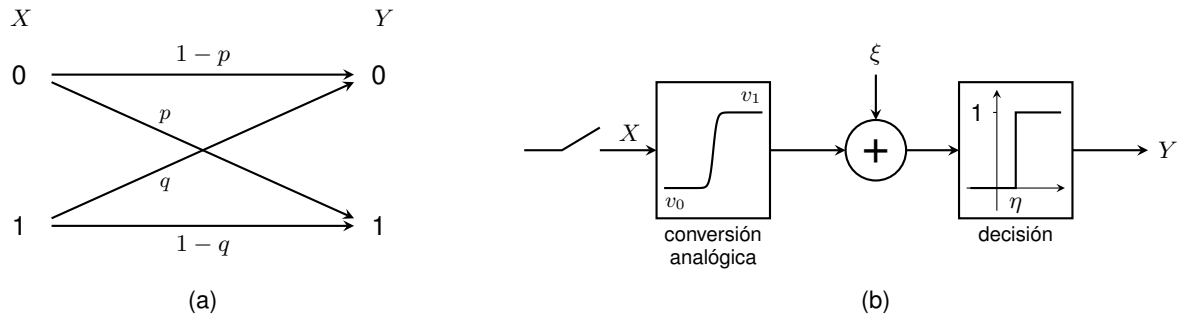


Figura 2-13: (a): canal binario. La entrada X definida sobre $\mathcal{X} = \{0, 1\}$ pasa por este canal e Y definida sobre $\mathcal{X}$ es recibido. Este canal es caracterisado por las probabilidades de transition $p_{Y|X}$. (b): Esquema que puede conducir al canal binario; una variable puede ser la salida de una puerta lógica, con niveles v_0 (nivel bajo, códificando 0) y v_1 (nivel alto, códificando 1). Se puede imaginar que este voltaje es transmitido por un canal añandido un ruido ξ. En la recepción, se toma una decisión, por ejemplo 0 (resp. 1) si la medida es mayor (resp. menor) que $\eta = \frac{v_0+v_1}{2} + \mathrm{E}[\xi]$. En este ejemplo, p y q van a ser caracterisada completamente por la distribución del ruido (y de los deos niveles posibles de la entrada), pero no de la distribución p_X.

La información mutua se escribe $I_2(X;Y) = H_2(Y) - H_2(Y|X) = H_2(Y) - H_2(Y|X=0)\Pr[X=0] + H_2(Y|X=1)\Pr[X=1]$, lo que toma la expresión

$$I_2(X;Y) = h_2(\beta) - \alpha h_2(p) - (1-\alpha)h_2(q)$$

donde $h_2(\lambda) = -\lambda \log_2 \lambda - (1-\lambda)\log_2(1-\lambda)$ es la entropia binaria en bits. Para calcular la capacidad C_2 en bits, hace falta máximizar I_2 con respeto a α. Diferenciando I_2 en α, *i. e.,* $\frac{\partial I_2(X;Y)}{\partial \alpha} = \frac{\partial h_2(\beta)}{\partial \beta}\frac{\partial \beta}{\partial \alpha} - h_2(p) + h_2(q)$, es decir

$$\frac{\partial I_2(X;Y)}{\partial \alpha} = (1-p-q)\log_2\left(\frac{1-\beta}{\beta}\right) - h_2(p) + h_2(q)$$

- Claramente,

  $$q = 1-p \quad \Rightarrow \quad C_2 = 0$$

  Viene del hecho de que para $q = 1-p$, de $h_2(p) = h_2(1-p)$ se deduce que $I_2(X;Y) = 0$ constante. De hecho, en este caso, un 0 en la salida puede venir de un 0 o 1 con probabilida igual, y lo mismo para un 1 en la salida; en otros terminos, la salida aparece independiente de la entrada. Eso se verifica formalmente con $\beta = q$, dando $p_{Y|X} = p_Y$, dando una información mutua cero, y entonces una capacidad cero.

- Si $q \neq 1-p$, la derivada de I_2 con respeto a α se anula para $\beta = \beta^{\mathrm{opt}}$ ($\alpha = \alpha^{\mathrm{opt}}$),

  $$\beta^{\mathrm{opt}} = \frac{1}{1+2^{\frac{h(p)-h(q)}{1-p-q}}} \qquad \text{siendo} \qquad \alpha^{\mathrm{opt}} = \frac{\beta^{\mathrm{opt}} - q}{1-p-q}$$

  y dando un extremo para I_2. A continuación, $\frac{\partial^2 I_2}{\partial \alpha^2} = \frac{(1-p-q)^2}{\beta(1-\beta)} > 0$ (en particular para el β "óptimo"), probando de que el extremo es un máximo. Poniendo el α^{opt} en la formula de $I_2(X;Y)$, luego de muchos calculos (básicos), se obtiene

  $$C_2 = \log_2\left(1+2^{\frac{h_2(p)-h_2(q)}{1-p-q}}\right) - \frac{(1-q)\,h_2(p) - p\,h_2(q)}{1-p-q}$$

Cuando $q \to 1-p$, notando que $h_2(p) = h_2(1-p)$ y tomando el limite de esta formula, se recupera que $C_2 \to 0$.

De $I_2(X;Y) = H_2(Y) - H_2(Y|X) \leq H_2(Y) \leq 1$ bit (Y es binario, de entropia maxima en el caso uniforme), aparece sin calculos que

$$C_2 \leq 1 \text{ bit}$$

i. e., la capacidad es menor que 1 bit [19]: para transmitir información en este canal, hace falta introducir redundancia en el mensaje. Se alcanza $C_2 = 1$ bit si, (i) por un lado $H_2(Y|X) = 0$, es decir $\alpha h_2(p) + (1-\alpha)h_2(q) = 0$ y ademas (ii) $h_2(\beta) = 1$. Estudiando cada caso (ej. con $\alpha = 0$ y $q = 0$ se satisface (i) pero no (ii) porque $\beta = 0$), se obtiene que

$$C_2 = 1 \qquad \Leftrightarrow \qquad \alpha = \frac{1}{2} \quad \text{y} \quad p = q = \frac{1 \pm 1}{2}$$

Para $p = q = 0$ el canal es perfecto, mientras que para $p = q = 1$ el canal es llamado *canal volteando*; en ambos casos, se recupera la entrada (o directamente, o tomando el opuesto) “sin perdida”.

La figura 2-14 representa la información mutua $I(X;Y)$ para unos canales (p y q dados) en función de α. Se nota que la curva es concava y tiene un máximo, capacidad del canal. La figura 2-15 representa la capacidad del canal en función de p y q así que unos casos particulares/cortes.

En el caso particular $p = q$, conocido como *canal simetrico*, la capacidad es

$$C_2 = 1 - h_2(p)$$

(alcanzada con una entrada uniforme). Como visto en el caso general, la capacidad vale 1 bit si y solamente si $h_2(p) = 0$, es decir $p = 0$ o $p = 1$. Al revés, la capacidad es mínima cuando H_2 est máximo, es decir para $p = q = \frac{1}{2}$, y $C_2 = 0$ (instancia particular de $q = 1-p$). $h_2(p)$ es la perdida en bit para cada bit transmitido. La capacidad C_2 en función de p es dada figura 2-15-(b).

En el caso particular $p = 0$, conocido como *canal en Z*, la capacidad es

$$C_2 = \log_2\left(1 + 2^{-\frac{h_2(q)}{1-q}}\right)$$

Se nota en este caso también que la capacidad alcanza 1, su máximo, si y solamente si $q = 0$ (canal perfecto). Al revés, cuando $q \to 1$, $C \to 0$, instancia particular de $q = 1-p$. La capacidad C_2 en función de q es dada figura 2-15-(c).

En (Cover & Thomas, 2006; Rioul, 2007) entre otros, se estudia diversos otros canal discretos, binario o con mas estados. Unos sont representados en la figure 2-16 (ver también (Shannon, 1948; Elias, 1957) o ej. (Arimoto, 1972) para el calculo numerico de la capacidad en el caso general).

---

[19] De manera general, de la escritura de I con entropias condicionales, para X definido sobre $\mathcal{X}$ e Y sobre $\mathcal{B}$, da $0 \leq C \leq \text{mín}(\log|\mathcal{X}|, \log|\mathcal{B}|)$. Ademas, $p_{Y|X}$ depende solo del canal y no de la entrada, así que para $p_X = \lambda p_X^{(1)} + (1-\lambda)p_X^{(2)}$ se obtiene $p_Y = \lambda p_Y^{(1)} + (1-\lambda)p_Y^{(2)}$ con $p_Y^{(i)}$ salida corespondiente a $p_X^{(i)}$ de $I(X;Y) = H(Y) - H(Y|X)$, el secundo termino dependiente solo del canal, de la concavidad de H se obtiene que I es concava con respeto a p_X. p_X parteciendo a un convexo, I tiene un máximo, único.

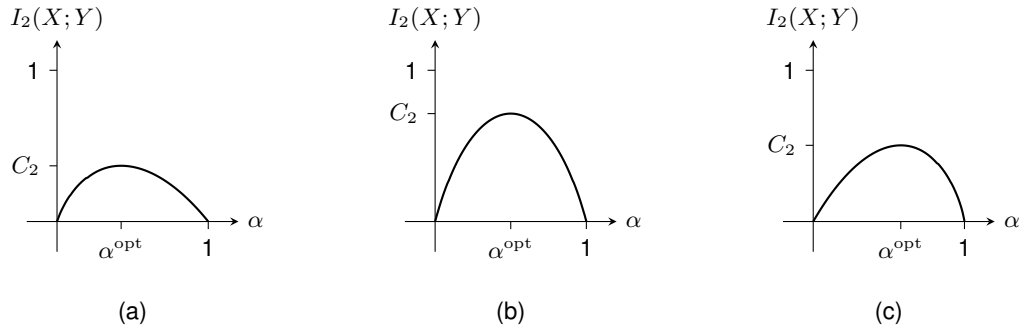


Figura 2-14: Información mutua (en bits) entrada-salida $I_2(X;Y)$ del canal binario en función de $\alpha = \Pr[X = 0]$. (a): $p = 0{,}4$ y $q = 0{,}01$; (b): $p = q = 0{,}05$ (canal simetrico); (c): $p = 0$ y $q = 0{,}05$ (canal en Z).

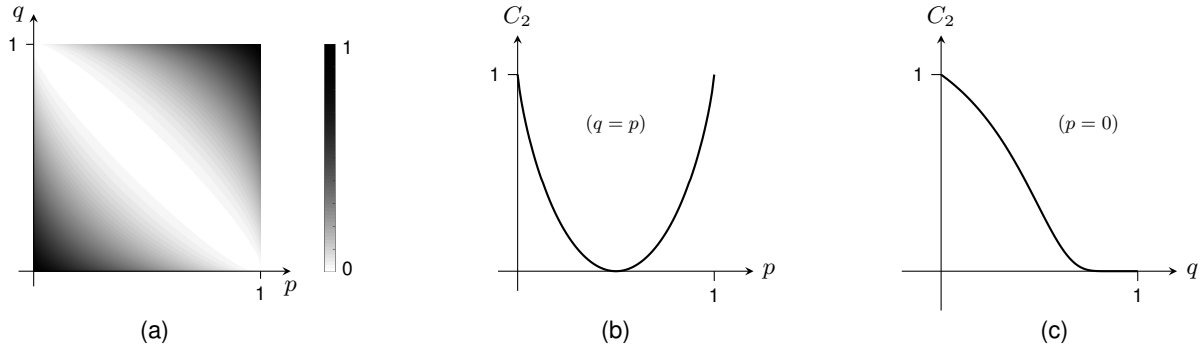


Figura 2-15: Capacidad C_2 del canal binario. (a): en función de p y q. (b): en función de p para el canal simetrico ($p = q$); (c): en función de q para $p = 0$ (canal en Z).

2.5.2 Canal de transmisión continuo gaussiano y su capacidad

Un canal de comunicación continuo relativamente simple es conocido como *canal gaussiano* (Shannon, 1948, Sec. 25), (Cover & Thomas, 2006; Rioul, 2007): X es una variable continua definida sobre $\mathcal{X} \subseteq \mathbb{R}^d$ y la salda Y es una versión ruidosa de X, *i. e.,* $Y = X + \xi$ con el ruido ξ independiente de X; En el canal gaussiano, $\xi \equiv \mathcal{N}$ es un vector gaussiano. Este canal es también definido por su densidad de probabilidad "de transición" $p_{Y|X}$, *i. e.,* por la distribución del ruido. Tal canal es descrito figura 2-17. Se supone conicida la matriz de covarianza $\Sigma_{\mathcal{N}}$ del ruide, y se nota Σ_X la de la entrada. En practica, no se puede mandar un mensaje a una potencia tan alta que se quierre, lo que se traduce por una limitación

$$\operatorname{Tr}(\Sigma_X) \leq P$$

potencia limite permitida por componente (sampleo).

Por definición, la información mutua $I(X;Y)$ entrada-salida es dada por $I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\mathcal{N})$. Maximizar $I(X;Y)$ es equivalente a maximizar $H(Y) = H(X + \mathcal{N})$ sujeto a $\operatorname{Tr}(\Sigma_X) \leq P$.

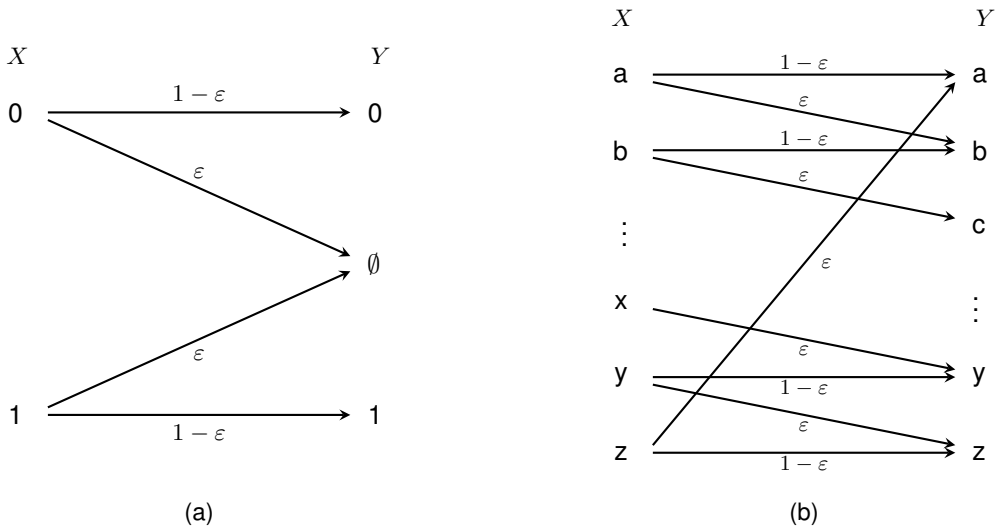


Figura 2-16: Ejemplos de canales discretos usuales. (a): canal borrador, donde un 0 (de probabilidad de ocurrencia α) o 1 (de probabilidad de occurencia $1-\alpha$) puede transmitirse correctamente or ser borado/perdido (estado $\emptyset$) con una probabilidad ε. Se calcula $I_2(X;Y) = (1-\varepsilon)h_2(\alpha)$, dando la capacidad $C_2 = 1-\varepsilon$, alcanzada para una entrada uniforme. (b): canal tipo machina de escribir, donde cada letra de un ensemble de n letras (acá con $n = 26$) se transmite correctamente con una probabilidad $1 - \varepsilon$ o a la letra siguiente (de manera cíclica) con una probabilidad ε. De $I_n(X;Y) = H_n(Y) - H_n(Y|X) = H_(Y) - h_(\varepsilon)$ se deduce que I_n es máxima si Y es uniforme, lo que es posible si X es uniforma, dando $C_n = 1 - h_n(\varepsilon)$.

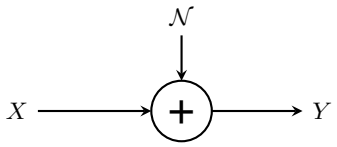


Figura 2-17: Canal gaussiano. La entrada X, modelisada por un vector aleatorio es corrupto aditivamente por un ruido gaussiano $\mathcal{N}$ independiente de X. La salida es entonces $Y = X + \mathcal{N}$ y el canal es completamente descrito por $p_{Y|X}(x,y) = p_{\mathcal{N}}(y-x)$ (obviamente independiente de la distribución de la entrada).

Fijando un Σ_X, la propiedad [P'5]*b* de le entropia diferencial implica que $H(Y)$ sea maximal si y solamente si Y es gaussiana, es decir si y solamente si X est gaussiana, dando $I(X;Y) = \frac{1}{2}\log|\Sigma_X + \Sigma_{\mathcal{N}}| - \frac{1}{2}\log|\Sigma_{\mathcal{N}}|$. Tomando en cuenta el limite de potencia, hace falta maximizar $|\Sigma_X + \Sigma_{\mathcal{N}}|$ sujeto a $\operatorname{Tr}\Sigma_X \leq P$ y $\Sigma_X \geq 0$ simetrica lo que no es trivial. Se encuentra el enfoque en (Cover & Thomas, 2006, Sec. 9.4). Sea U, matriz ortogonal ($UU^t = U^tU = I$) de los autovectores de la matriz $\Sigma_{\mathcal{N}} \geq 0$ simetrica [20], de columnas u_i ordenadas tal que las autovalores corespondientes $\lambda_i^{\mathcal{N}}$ sean en orden crecientes,*i. e.,*

$$\Sigma_{\mathcal{N}} = U\,\mathrm{diag}\left(\lambda_1^{\mathcal{N}},\ldots,\lambda_d^{\mathcal{N}}\right)U^t \qquad \text{con} \qquad 0 \leq \lambda_1^{\mathcal{N}} \leq \cdots \leq \lambda_d^{\mathcal{N}}$$

donde diag es la matriz diagonal teniendo los λ_i en su diagonal, y sea $R_X = U^t\Sigma_X U$. En sencillo ver que

[20] Se recuerda que $A \geq 0$ significa que A es definida no negativa.

$|\Sigma_X + \Sigma_\mathcal{N}| = |R_X + \Lambda_\mathcal{N}|$ (de $|AB| = |A|\,|B|$) y que $\operatorname{Tr}\Sigma_X = \operatorname{Tr} R_X$ (de $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$). Entonces, el problema se reduce a maximizar $|R_X + \Lambda_\mathcal{N}|$ sujeto a $\operatorname{Tr} R_X \leq P$ y $R_X \geq 0$ simetrica. La desigualdad de Hadamard ya evocada da $|R_X + \Lambda_\mathcal{N}| \leq \prod_i (R_X + \Lambda_\mathcal{N})_{i,i} = \prod_i \left((R_X)_{i,i} + \lambda_i^\mathcal{N}\right)$ con igualdad si y solamente si R_X es diagonal: para maximizar $|R_X + \Lambda_\mathcal{N}|$, R_X debe ser diagonal (dada una diagonal, se alcanza el maximo si los otros terminos son nulos). Es decir que la base que diagonaliza $\Sigma_\mathcal{N}$ debe diagonalizar también Σ_X. Sean λ_i^X los terminos diagonales de R_X: queda que maximizar $\prod_i \left(\lambda_i^X + \lambda_i^\mathcal{N}\right)$ sujeto a $\sum_i \lambda_i^X \leq P$ y $\lambda_i^X \geq 0$. Este problema de optimización sujeto a una desigualdad se resolva con el enfoque de Karush-Kuhn-Tucker [21] (KKT) (Miller, 2000; Cambini & Martein, 2009), dando $\lambda_i^X = \left(\lambda - \lambda_i^\mathcal{N}\right)_+$ con $(\cdot)_+ = \max(\cdot, 0)$ y λ tal que $\sum_i \left(\lambda - \lambda_i^\mathcal{N}\right)_+ = P$. En conclusión, la capacidad es dada por

$$C = \frac{1}{2}\log\left(\frac{|\Sigma_\mathcal{N} + \Sigma_X|}{|\Sigma_\mathcal{N}|}\right) \qquad \text{con} \qquad \Sigma_X = U\,\operatorname{diag}\left(\left(\lambda - \lambda_1^\mathcal{N}\right)_+, \ldots, \left(\lambda - \lambda_d^\mathcal{N}\right)_+\right) U^t,$$
$$\lambda \text{ tal que } \sum_i \left(\lambda - \lambda_i^\mathcal{N}\right)_+ = P$$

alcanzada por X gausiano de matriz de covarianza Σ_X así construida.

La última condición se resolva través de lo que es conocido como "llenado de agua" (water-filling en inglés), illustrado figure 2-18. El principio es parecido a tener baras niveles $\lambda_i^\mathcal{N}$ representando las potencias del ruido (en la base que la diagonaliza), y de "llenar con agua" hasta un nivel λ tal que el "volumen" añadido vale P; en cada $\lambda_i^\mathcal{N}$ se ha añadido el λ_i^X (Cover & Thomas, 2006, Sec. 9.4).

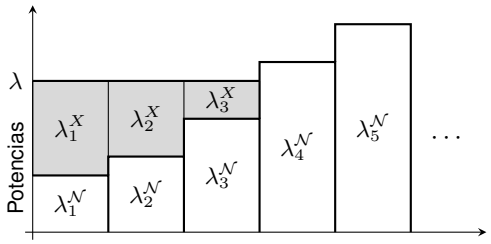


Figura 2-18: Principio del "water-filling" para obtener los λ_i^X satisfaciendo el vinculo de potencia limite y permitiendo de construir Σ_X a partir de la matriz diagonal de los λ_i^X y la base que diagonaliza la covariance $\Sigma_\mathcal{N}$ del ruido. La zone en grise representa esquematicamente P.

En el caso escalar, se obtiene

$$C = \frac{1}{2}\log\left(1 + \frac{P}{\sigma_\mathcal{N}^2}\right)$$

donde $\frac{P}{\sigma_\mathcal{N}^2}$ es conocido como relación señale-ruido [22]

---

[21] Se introduce el factor de Lagrange y se maximiza $\prod_i \left(\lambda_i^X + \lambda_i^\mathcal{N}\right) + \mu \sum_i \lambda_i^X$. Eso da $\lambda_i^X + \lambda_i^\mathcal{N} = \lambda$ constante si esta cantidad es positiva, y cero si no, es decir $\lambda_i^X = \left(\lambda - \lambda_i^\mathcal{N}\right)_+$. Deben ser los mas grande, es decir λ lo mas grande que se puede, pero satisfaciendo $\sum_i \lambda_i^X \leq P$, *i. e.,* alcanzando la igualdad.

[22] Esta formula es muy parecida a la de Shannon, Laplume, Clavier (Shannon, 1948; Laplume, 1948; Clavier, 1948) (ver también (Cover

En (Cover & Thomas, 2006; Rioul, 2007) por ejemplo, se dan otros ejemplos de canal de comunicación en el contexto continuo (entrada X_t siendo una señal/proceso, canal filtrando, canal con feedback, etc.).

2.5.3 Codificación entropica sin perdida

El problema de codificación de fuente puede presentarse de la manera siguiente (Cover & Thomas, 2006, cap. 5) o (Rioul, 2007, cap. 13). Sea un proceso aleatorio $\{X_t\}_{t\in\mathbb{Z}}$, supuesto estacionario, llamado *fuente*, donde los X_t toman sus valores sobre un alfabeto discreto finito

$$\mathcal{X} = \{x_1, \ldots, x_\alpha\} \qquad \text{alfabeto fuente}$$

de distribución p_X. A cada posible secuencia [23] $s_1 \cdots s_n \in \mathcal{X}^n$ de letras de $\mathcal{X}$, se quiere asignar un código $c(s_1 \cdots s_n)$ un alfabeto discreto finito,

$$\mathcal{C} = \{\zeta_1, \ldots, \zeta_d\} \qquad \text{alfabeto código}$$

El código es dicho d-*ario*. Por ejemplo, se puede asignar un código $c(x_i) = \zeta_{i,1} \cdots \zeta_{i,l_i} \in \mathcal{C}^{l_i}$ a cada simbolo, código llamado *palabras códigos*, y a secuencias $s_1 \cdots s_n$ la concatenación de las palabras códigos correspondiente a cada simbolo, *i. e.,* el código $c(s_1) \cdots c(s_n)$. En el sistema Moorse por ejemplo, $\mathcal{C}$ consiste en un punto, barra, espacio letras, espacio palabras. En una computadora en general todo se codifica en bits $\mathcal{C} = \{0\,,\,1\}$. Mas formalmente, sean

$$F_\mathcal{X} = \bigcup_{k=0}^{\infty} \mathcal{X}^k \qquad \text{y} \qquad F_\mathcal{C} = \bigcup_{k=0}^{\infty} \mathcal{C}^k$$

unión de secuencias de k letras de $\mathcal{X}$ y $\mathcal{C}$ respectivamente. Una codificación de fuente consiste en una función de $F_\mathcal{X}$ dentro de $F_\mathcal{C}$. En lo que sigue, nos concentramos en códigos definidos para blocks de simbolos de tamaño $m \geq 1$:

$$\begin{array}{rccl} c_m: & \mathcal{X}^m & \to & F_\mathcal{C} \\ & x & \mapsto & c_m(x) \in \mathcal{C}^{l_{c_m}(x)} \end{array}$$

donde $l_{c_m}(x) \in \mathbb{N}^*$ es el *largo* de la palabra código $c_m(x)$, y

$$\forall\, n \geq 1, \quad \forall\, s_1 \cdots s_n \in \mathcal{X}^{nm}, \quad c_m(s_1 \cdots s_n) \equiv c_m(s_1) \cdots c_m(s_n)$$

& Thomas, 2006, Sec. 9.3) o (Rioul, 2007, Sec. 11.2)). De hecho, si se considera simbolos mandandos durante T secundos (simbolos puesto en forma para dar una señal analogica) usando una banda de transmisión B, por el teorema de Nyquist $B = \frac{1}{2T}$ (caso limite). Si el ruido es blanco en la banda B, de densidad espectral de potencia por unidad de frecuencia igual a N_0, para un simbolo la relación señal-ruido se escribe $\frac{P}{N_0 B}$. Ademas, se calcula en general la capacidad por unidad de tiempo es decir la capacidad por simbolo divido por T, *i. e.,* $C = B \log\left(1 + \frac{P}{N_0 B}\right)$ por secundos, lo que es precisamente la capacidad calculdada por Shannon. Esa es a veces conocida como formula de Shannon-Hartley.

[23] Por abuso de escritura una cadena de n simbolos puede ser vista como un n-uplet.

lo que es llamado *extensión* del codigo. En lo que sigue, se escribirá $c \equiv c_1$.

Una manera ingenua de codificar consiste a apoyarse sobre la descompocisión de base d de un intero, *i. e.,* para $1 \leq i \leq \alpha$, $i-1 = (i_0-1) + (i_1-1)d + \cdots + (i_K-1)d^K$ donde $K = \left\lceil \log_d |\mathcal{X}| \right\rceil$ y $1 \leq i_k \leq \alpha$, y de asignar la palabra código $\zeta_{i_0} \cdots \zeta_{i_k}$ al simbolo x_i. Haciendo eso, cada palabra código tieno el mismo largo. Pero, parece mas economico hacer una codificación dicha de largos variables, teniendo en cuenta las probabilidades de aparición de cada x_i. Implicitamente, es la idea del código de Moorse, que asigna un punto (o series de puntos) a las letras muy frecuentes, y una barra (o combinaciones) a las que son raras. Dicho de otra manera, el código ingenuo sería "bueno" para x_i apareciendo con las mismas frecuencias/probabilidades.

En los códigos de largos variables (incluendo el código ingenuo), volviendo a c_m, existen varios tipos de códigos. Un código es dicho *no singular* si c_m es inyectiva: a cada $x \in \mathcal{X}^m$ corresponde una palabra código única. Esta propiedad es un requisito que parece obvio querer para un código. Pero no es suficiente para poder decodificar un mensaje, compuesta por una secuencia de palabras código. Lo importante en este caso es poder decodificar la secuencia sin ambiguedad: un código est dicho *unicamente decodificable* (o sin perdida) si todas las extensiones son no singulares. Por ejemplo, sean $\mathcal{X} = \{\alpha\,,\,\beta\,,\,\gamma\,,\delta\}$, $\mathcal{C} = \{0\,,\,1\}$ y $c(\alpha) = 0,\ c(\beta) = 0,\ c(\gamma) = 1,\ c(\delta) = 01$ ($m = 1$). El código es no singular, pero no unicamente decodable. La secuencia 0010 puede provenir de $\alpha\alpha\gamma\alpha$, de $\alpha\delta\alpha$ o de $\beta\gamma\alpha$. En general, se requierre de un código que sea unicamente decodificable. A veces, se requiere poder decodificar sobre la marcha, sin esperar de medir toda la secuencia codificada: es lo que se llama *código instantaneo*. Por ejemplo, el código $c(\alpha) = 00,\ c(\beta) = 10,\ c(\gamma) = 11,\ c(\delta) = 110$ es unicamente decodable, pero no instantaneo. Considera la secuencia 0011011 y marcha sobre ella. 0 no es una palabra código; 00 es y sin ambiguedad proviene de un α (no hay otras palabras empezando por 00); luego 1 no es una palabra, y 11 es una palabra código, pero se necesita adelantar para saber si viene de un γ (o de un δ); la letra siguiente siendo un 0, todavía no se puede concluir si 110 vino de γ y algó o δ. Al final, con 1101, se sabe que se tuvo un δ porque ninguna palabra código empieza por 01. Al final, sin ambiguedad el antecedante de la secuencia binaria era $\alpha\delta\gamma$. Pero se necesitó marchar sobre toda la secuencia antes de decodificar. Obviamente, un código instantaneo es tal que ninguna palabra código es prefijo de una otra, *i. e.,* si $c_m(x)$ es una palabra código, las otras palabras código no pueden empezar con $c_m(x)$; el código es también dicho *libre de prefijo*. Estas distinciones estan ilustradas en la figura 2-19 (ver (Cover & Thomas, 2006, cap. 5)).

Ademas de la decodificación sin ambiguedad, una caracterización importante del código es la taza de codificación [24]

$$R_{c_m} = \frac{\log_d \left(\sum_{x \in \mathcal{X}^m} l(x) \Pr[X = x] \right)}{m}$$

donde X representa una secuencia de m variables X_t. El argumento del logaritmo (de base adecuada al cardinal de $\mathcal{C}$) es el *largo promedio* del código. Por ejemplo, para $d = 2$, R_{c_m} es el numero de bits promedio del código por simbolo.

---

[24] En (Rioul, 2007) por ejemplo, se define esta taza suponiendo que cada secuencia fuente es codificado por el mismo numero de bits. La taza es entonces el numero de bits por simbolo.

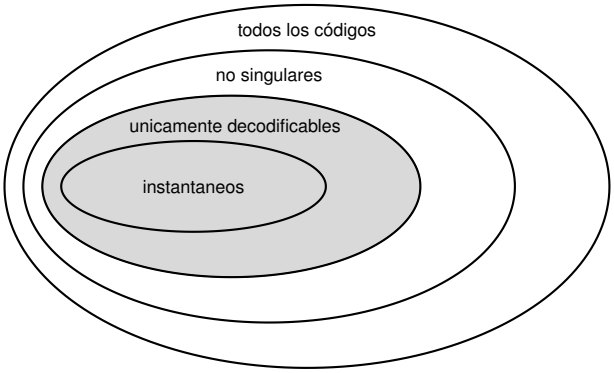


Figura 2-19: Clases de códigos. Los códigos contienen la clase de los no singular. La misma contiene la clase de los códigos unicamente decodificables. Ella contiene los códigos instantaneos. En grise se representan las clases de códigos sin perdida a lo cuales se dedica esta sección.

En general, se quiere minimizar R_{c_m} (compresar el mensaje a mandar), lo que puede ser contradictorio con la necesidad de añadir redundancia para no perder información durante una transmsión. En lo que sigue, nos concentramos el el problema de compresión, sin tener en cuenta el paso de transmisión de mensajes codificados en un canal. Minimizar la taza es equivalente a minimizar el largo promedio. Se puede focalisarse en $m = 1$; todo se extiende sencillamente a $m > 1$.

La meta de la compresión es entonces construir un codigo c, unicamente decodificable, que minimizar el largo promedio

$$L(c) = \sum_{x \in \mathcal{X}} p_X(x)\, l(x)$$

Antes de ir mas adelante, falta traducir en ecuación el vinculo de que c sea unicamente decodificale. Eso es dado por la desigualdad de Kraft-McMillan (Kraft Jr, 1949; McMillan, 1956; Karush, 1961) [25]

Teorema 2-8 (Desigualdad de Kraft-McMillan)**.** *Los largos $l_c(x)$ de la palabras código de un código c unicamente decodable deben satisfacer la desigualdad*

$$\sum_{x \in \mathcal{X}} d^{-l_c(x)} \leq 1$$

Reciprocamente, para cada conjunto de enteros $\{\ell_x\}_{x \in \mathcal{X}}$ satisfaciendo esta desigualdad, es posible de construir un código unicamente decodable con $l_c(x) = \ell_x$.

Demostración. Para cualquier $k \geq 1$ y cualquier cadena $s = s_1 \cdots s_k \in \mathcal{X}^k$, la extensión del código, $c_k(s_1 \cdots s_k) = c(s_1) \cdots c(s_k)$ satisface $l_{c_k}(s) = \sum_i l_c(s_i)$. Entonces

$$\left(\sum_{x \in \mathcal{X}} d^{-l_c(x)} \right)^k = \sum_{\bar{x} \in \mathcal{X}^k} d^{-l_{c_k}(\bar{x})} = \sum_{m=1}^{k\, l_c^{\max}} \#(m)\, d^{-m}$$

[25] Esta desigualdad fue probabada por Kraft para códigos instantaneos en su tesis de maestria (Kraft Jr, 1949). Luego, fue extendiad a los códigos unicamente decodificables por B. McMillan (McMillan, 1956) (en una nota de pie de pagina de su papel, atribua esta observación a J. L. Doob hecho oralemente durante una escuela de verano en Ann Arbor, MI en agosto 1955).

re-escribiendo la secunda suma, agrupando los terminos de mismo largos, donde $\#(m)$ es el numero de códicos de $\mathcal{X}^k$ tiendo el largo m y $l_c^{\max} = \max_{x\in\mathcal{X}} l_c(x)$ largo mayor. c siendo unicamente decodificable, c_k debe ser inyectiva, imponiendo $\#(m) \leq d^m$ (no hay mas palabras de largo m que el cardinal de $\mathcal{C}^m$), dando inmediatemente que necesariamente

$$\forall\, k \in \mathbb{N}^*, \quad \sum_{x\in\mathcal{X}} d^{-l_c(x)} \leq \left(k\, l_c^{\max}\right)^{\frac{1}{k}} \quad \Leftrightarrow \quad \sum_{x\in\mathcal{X}} d^{-l_c(x)} \leq \min_{k\in\mathbb{N}^*} \left(k\, l_c^{\max}\right)^{\frac{1}{k}}$$

Un estudio rapido de $u \mapsto \left(u\, l_c^{\max}\right)^{\frac{1}{u}}$ para $u \geq 1$ y tendiendo en cuenta de que $l_c^{\max} \leq 1$ permite concluir que el mínimo es igual a 1, terminando la parte directa del teorema.

Reciprocamente, sea $\{\ell_x\}_{x\in\mathcal{X}}$ un conjunto de enteros satisfaciendo la desigualdad de Kraft-McMillan. Se puede agrupar los largos iguales y clasificarlos. Sea n_ℓ los numeros de largos igual a $\ell = 1, \ldots, \ell^{\max} \leq \alpha$. Consideramos ahora un arbol empezando con una raiz, correspondiente a un largo 0, que se divide en d ramas, correspondiente a los largos iguales a 1; a cada nudo se asocian las letras $\zeta_1, \ldots, \zeta_d$. Esto nudos se dividen cada uno en d otras ramas, y los nudos de "padre" ζ_i se va a asociar las palabras códigos $\zeta_i\zeta_1, \ldots, \zeta_i\zeta_\alpha$, etc. Este arbol, conocido como arbol de Kraft, es ilustrado en la figura 2-20 para $d = 2$ y $\mathcal{C} = \{0\,,\,1\}$. Claramente, $n_1 \leq d$ si no $n_1\, d^{-1} > 1$ y los largos no podrían satisfacer la desigualdad de Kraft-McMillan. El principio es entonce de asociar a los n_1 (posiblemente igual a 0) largos iguales a 1 unos nudos con las palabras código asociadas de largo 1 (primera profundez de ramas) y de prohibir todos las ramas de padre los nudos seleccionados (lineas punteadas en la figura2-20). Estos nudos son llamados *hojas* (no hay ramas). En la capa de "hijos" de profundez/largos 2, quedan $d^2 - n_1\, d$ nudos (accessibles) que se puede dividir en ramas. Nuevamente, $n_2 \leq d^2 - n_1\, d$ sino se tendría $n_1\, d^{-1} + n_2\, d^{-2} > 1$, incompatible con la desigualdad de Kraft-McMillan. Se puede asociar a los n_2 largos iguales a 2 unos nudos con las palabras código asociadas de largo 2 (secunda profundez), y de prohibir que salen de estos nudos nuevas ramas (son entonces hojas de la secunda profundez), etc. Haciendo así, se asocia un código c de largos $l_c(x) = \ell_x$ que aparece libre de prefijo, es decir instantaneo. Entonces, este código es también unicamente decodable. □

A este punto, se menciona los hechos siguientes

- Los largos de un código unicamente decodable satisfacen la desigualdad de Kraft-McMillan, pero con el conjunto de largos correspondientes se puede siempre construir un código instantaneo. Claramente, se puede buscar un código de largo promedio mínimo en los códigos instantaneo, sin perdida de optimalidad (buscar el la clase mas amplia de los unicamente decodable no permite bajar el largo promedio).

- En los códigos libre de prefijo, si se fija los numeros de hojas (última profundez) borradas contruiando un código, este vale $\sum_{i=1}^{\ell^{\max}} n_i\, d^{\ell^{\max}-i} = \sum_{x\in\mathcal{X}} d^{\ell^{\max}-l_c(x)}$. Es necesariamente menor que los numeros total $d^{\ell^{\max}}$ de hojas, lo que prueba el teorema para los códigos instantaneos (Kraft Jr, 1949; Karush, 1961).

- El teorema se generaliza obviamente para codificar una fuente discreta pero con un número infinito de estados, tomando el límite $\alpha \to \infty$.

- Si se conocen los largos óptimos, es suficiente para poder construir un código libre de prefijo.

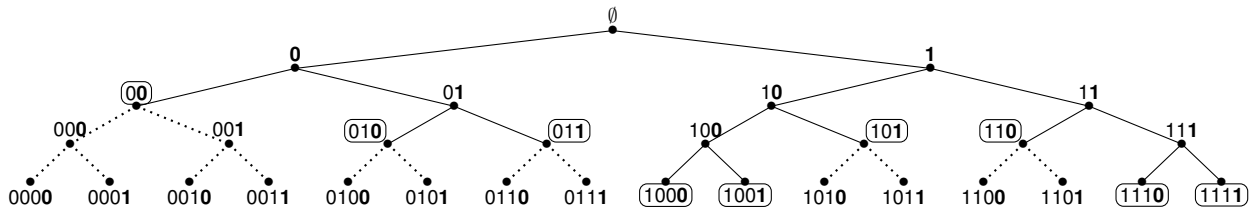


Figura 2-20: Arbol de Kraft en el caso binario ($d = 2$). De la raiz, de código $\emptyset$ de largo 0, se divide en dos ramas, de códigos respectivamente 0 y 1 (profundez 1). Cada nudo de esta profundez se divide de nuevo en dos ramas (profundez dos), dando cuatros nuevos nudos con los códigos 00 y 01 de padre 0, y 10 y 11 de padre 1. Etc. En cada nudo, en el código, se marca en negrita la letra correspondiente al bit añadido al código padre. Para hacer un código libre de prefijo, una vez que un nudo es seleccionado para ser una palabra código (encuadrado en la figura), no se puede tener nudo "hijos" siendo también una palabra código: se boran las ramas saliendo de un nudo-palabra código (ramas punteadas).

El formalismo dado, se va a ver ahora reapacer la entropia de Shannon como cota de la codificación de fuente sin perdida:

Teorema 2-9 (Cota inferior de códigos unicamente decodable)**.** *Para cualquier código c unicamente decodable de la fuente X, su largo promedio esta acotado por debajo por la entropia de Shannon de base d de X,*

$$L(c) = \sum_{x \in \mathcal{X}} p_X(x)\, l_c(x) \;\geq\; H_d(X)$$

Demostración. Sea $q(x) = \frac{d^{-l_c(x)}}{\sum_{x \in \mathcal{X}} d^{-l_c(x)}}$, siendo una distribución de probabilidad por construcción. Escribiendo $l_c(x) = \log_d d^{-l_c(x)}$, se puede expresar el largo promedio de la forma

$$L(c) = -\sum_{x \in \mathcal{X}} p_X(x)\, \log_d d^{-l_c(x)} = -\sum_{x \in \mathcal{X}} p_X(x)\, \log_d q(x) - \log_d \sum_a d^{-l_c(x)}$$

Notando que $-\log_d q = \log_d \left(\frac{p_X}{q}\right) - \log_d p_X$ se obtiene

$$L(c) = H_d(X) + D_d\left(p_X \| \, q\right) - \log_d \sum_{x \in \mathcal{X}} d^{-l_c(x)}$$

El resultado proviene de la positividad de la divergencia de Kullback-Leibler y de la desigualdad de Kraft (el argumento del logaritmo siende menor que 1). □

Este resultado significa que la taza de compresión sin perdida no puede ser mas bajo que el contenido informacional de la fuente. En este sentido, H tiene realmente un sabor de información sobre la fuente X.

La entropia aparece también el la cota superior del código óptimo:

Teorema 2-10 (Cota superior del código unicamente decodable óptimo)**.** *El largo promedio L^{opt} del código c^{opt} unicamente decodable, de largo promedio mínimo esta acotado por arriba por la entropia de Shannon de base d de X mas un dit (1 simbolo de $\mathcal{C}$),*

$$L^{\mathrm{opt}} \;<\; H_d(X) + 1$$

Demostración. Por eso, empezamos por buscar los largos óptimos, solución de la optimización

$$\text{mín} \sum_{x\in\mathcal{X}} p_X(x)\, l(x) \qquad \text{sujeto a} \qquad \sum_{x\in\mathcal{X}} d^{-l(x)} \; 1$$

Sea $q(x) = \frac{d^{-l_c(x)}}{\sum_{x\in\mathcal{X}} d^{-l_c(x)}}$, siendo una distribución de probabilidad por construcción. Escribiendo $l_c(a) = \log_d d^{-l_c(a)}$, se puede expresar el largo promedio de la forma

$$L(c) = -\sum_{x\in\mathcal{X}} p_X(x)\, \log_d d^{-l_c(x)} = -\sum_{x\in\mathcal{X}} p_X(x)\, \log_d q(x) - \log_d \sum_{x\in\mathcal{X}} d^{-l_c(x)}$$

Olvidando que los $l_i \equiv l(x_i)$ son enteros, $L(c)$ es convexa con respecto a los l_i así que el vinculo, garantizando que el mínimo existe y es único. El problema se resuelva con el enfoque KKT [21], optimización con vinculos tipo desigualdades (Miller, 2000; Cambini & Martein, 2009), conduciendo a los "largos"

$$\widetilde{l}(x) = -\log_d p_X(x)$$

Una posibilidad puede ser de tomar la parte entera superior,

$$l(x) = \left\lceil -\log_d p_X(x) \right\rceil$$

Obviamente el conjunto de largos satisface la desigualdad de Kraft-McMillan, así que se puede construir un código c^{sh} unicamente decodable con estos largos. De $l(x) < -\log_d p_X(x) + 1$ se obtiene

$$L^{\text{opt}} \leq L\left(c^{\text{sh}}\right) < H_d(X) + 1$$

□

De

$$H_d(X) \leq L^{\text{opt}} < H_d(X) + 1$$

se revela el rol fundamental de la entropia en la codificación de fuente sin perdida. La codificación es a veces dicha *codificación entropica* y da un rol operacional a la entropia de Shannon. Se notara de la demostración precediente de que aparece un código particular:

Definición 2-16 (Código de Shannon)**.** *Un código c^{sh} de una fuente X, de largos $l^{\text{sh}}(x) = \left\lceil -\log_d p_X(x) \right\rceil$, libre de prefijo (construido sobre el arbol de Kraft) es llamado código de Shannon.*

Obviamente, también

$$H_d(X) \leq L\left(c^{\text{sh}}\right) < H_d(X) + 1$$

Al lo contrario de primer vista, un código de Shannon no es óptimo. Un ejemplo sencillo para verlo consiste a tomar $\mathcal{X} = \mathcal{C} = \{0\,,\,1\}$ y $p_x(0) = 0{,}999 = 1 - p_X(1)$. Los largos de Shannon van a ser $l^{\text{sh}}(0) = 1$ y $l^{\text{sh}}(1) = 10$, de largo promedio $L\left(c^{\text{sh}}\right) = 1{,}009$. Obviamente, un código óptimo es $c(x) = x$ de largos $l(x) = 1$ dando $L^{\text{opt}} = 1$ bit. De hecho, volviendo al problema con largos virtualmente no enteros, el mínimo se alcanza para $\widetilde{l}(x) = -\log_d p_X(x)$, es decir que, los largos siendo enteros, se alcanza la cota mínima del código óptimo si y solamente si $-\log_d p_X(x)$. Una distribución satifaciendo esta condición es dicha d-adica. Sin embargo, el código de Shannon es "competitivo" en el sentido de que:

Teorema 2-11 (Competitividad del código de Shannon)**.** *Sea X fuente sobre $\mathcal{X}$, de distribución p_X y c^{sh} el código de Shannnon asociado sobre el alfabeto código $\mathcal{C} = \{\zeta_1, \ldots, \zeta_d\}$, de largos $l^{\mathrm{sh}}(x) = \left\lceil -\log_d p_X(x) \right\rceil$. Para cualquier código c unicamente decodable y cualquier $k \geq 1$,*

$$\Pr\left[l^{\mathrm{sh}}(X) \geq l_c(X) + k\right] \leq \frac{1}{d^{k-1}}$$

Demostración. Por definición de un código de Shannon, de $a + 1 > \lceil a \rceil \geq b \Rightarrow a \geq b - 1$, se obtiene

$$\begin{aligned}
\Pr\left[l^{\mathrm{sh}}(X) \geq l_c(X) + k\right] &\leq \Pr\left[-\log_d p_X(X) \geq l_c(X) + k - 1\right] \\
&= \Pr\left[p_X(X) \leq d^{-l_c(X)-k+1}\right] \\
&= \sum_{x \in \mathcal{X} : p_X(x) \leq d^{-l_c(X)-k+1}} p_X(x)
\end{aligned}$$

Pero, sumando sobre lo x tal que $p_X(x) \leq d^{-l_c(X)-k+1}$, se obtiene

$$\begin{aligned}
\Pr\left[l^{\mathrm{sh}}(X) \geq l_c(X) + k\right] &\leq d^{1-k} \sum_{x \in \mathcal{X} : p_X(x) \leq d^{-l_c(X)-k+1}} d^{-l_c(X)} \\
\Pr\left[l^{\mathrm{sh}}(X) \geq l_c(X) + k\right] &\leq d^{1-k} \sum_{x \in \mathcal{X}} d^{-l_c(X)}
\end{aligned}$$

(añadiendo terminos positivos en la suma). La prueba se cierra notando que c siendo unicamente decodable, l_c satisface la desigualdad de Kraft-McMillan. □

Este teorema traduce el hecho de que si c^{sh} no es óptimo, tomando cualquier otro código (incluyendo el óptimo), la probabilidad que $c^{\mathrm{sh}}(X)$ tiene un largo mas importante que $c(X) + k$ decrece exponencialmente con k. El el ejemplo anterior, si se compara c^{sh} y el código óptimo, para $k = 9$ (caso del código de 1), $\Pr\left[l^{\mathrm{sh}}(X) \geq l_c(X) + 9\right] \leq 0{,}391\,\%$. De hecho, una palabra código de largo 10 aparece con una probabilidad $0{,}1\,\%$...

En el problema de minimización, el hecho de que los largos deben ser enteros no permite solucionar explicitamente el problema de buscada del código óptimo. Numeros investigadores contruyeron códigos, intentando probar que son óptimos (ver ej. (Shannon, 1948; Shannon & Weaver, 1964; Fano, 1949) por los primeros, y (Cover & Thomas, 2006, & ref.)). El código conocido como *código de Fano*n [26] c^{fa} se basa sobre el hecho de que se alcanza la cota mínima para una distribución d-adica. El principio es de clasificar $\mathcal{X}$ para obtener las probabilidades clasificadas en orden decreciences ($p_X^{\downarrow}$). Luego, se divide $\mathcal{X}$ en d ensembles a lo mas equiprobables que se puede (*i. e.,* de probabilidad a lo mas cerca de d^{-1}) y de asignar ζ_i al conjunto i. Luego, se repite el proceso a cada sub-conjunto (para tener sub-conjuntos de probabilidades a lo mas cerca de d^{-2}) y al subconjunto j del conjunto i se va a asignar le código $\zeta_i\zeta_j$, etc. Eso es ilustrado en la figura 2-21-(a). **Probar/mencionar que también**

$$H(X) \leq L\left(c^{\mathrm{fa}}\right) < H(X) + 1$$

[26] A pesar de que sea diferente del de Shannon y que cada uno fueron hechos independientemente, a veces es conocido como código de Fano-Shannon, o aun Shannon-Fano-Elias (Cover & Thomas, 2006; Krajči, Liu, Mikeš & Moser, 2015).

Fijense de que no hay un único código de Fano o de Shannon (tal como no hay un óptimo único). Por exemple, hacer una permutacion de los ζ_i da los mismos largos y el mismo largo promedio sin cambiar el aspecto libre de prefijo. De la misma manera, en el arbol de Kraft, en cada profundez se puede permutar los simbolos asociados a las hojas de esta profundez sin cambiar el aspecto libre de prefijo y sin que cambie los largos $l(x_i)$ (y entonces con el mismo largo promedio).

Una solución para construir un código óptima fue propuesta por Huffman en 1951-1952 (Huffman, 1952; Pigeon, 2003) [27]

Definición 2-17 (Código de Huffman)**.** *Suponemos que existe un* $q \in \mathbb{N}$ *tal que* [28] $\alpha = |\mathcal{X}| = d + q(d-1)$*. El algoritmo de Huffman consiste a construir un arbol donde cana nudo es asociado a un conjunto de simbolos fuente y una letra de* $\mathcal{C}$ *de la manera siguiente:*

1. *Clasificar las probabilidades en orden decrecientes: llamamos* p_i *las probabilidades rearregladas y, por cambio de escritura,* x_i *el simbolo fuente correspondiente.*
2. *A cada* x_i, $n-d+1 \leq i \leq n$, *associar un nudo y la letra "hijo"* ζ_i*.*
3. *Crear* d *ramas saliendo de un nudo padre hasta los* d *nudos* x_i, $n-d+1 \leq i \leq n$*.*
4. *Crear un nuevo conjunto de simbolos fuente* $\widetilde{x}_i = x_i$, $1 \leq i \leq n-d$ *de probabilidades respectivas* $\widetilde{p}_i = p_i$ *y* $\widetilde{x}_{n-d+1} = \{x_j,\ n-d+1 \leq j \leq n$ *de probabilidad* $\widetilde{p}_{n-d+1} = p_{n-d+1} + \ldots + p_n$*. El útimo "super-simbolo" fuente es asociado al nudo padre de la etapa 3.*
5. *Si quedan mas de un (super-)simbolo fuente, volver a la etapa 1 con* $p \equiv \widetilde{p}$ *y* $x \equiv \widetilde{x}$*.*

Como descrito tratando del código usando el arbol de Kraft, $c^{\mathrm{huf}}(x_i)$ *se construye saliendo de la raiz del arbol así construido, agregando las letras del camino que llega a la hoja* x_i*. Eso es ilustrado en la figura 2-21-(b) en el caso binario.*

Se mencionara que a cada etapa, el nuevo conjunto de super-simbolos fuente contiene exactamente $d-1$ simbolos menos que a la etapa precediente. Así, con $n = d + q(d-1)$ el algoritmo tiene exactamebte $q+1$ buclas y en cada profundes no hay nudo vacio en el sentido que o es una hoja, o es un nudo padre/prefijo (quedaran exactamente d nudos a agregar a la raiz en la última etapa). Por exemplo, con $d = 3$ si tuvieramos $n = 4$, en la secunda etapa tendriamos 2 estados a juntar, dando un código de largos $2, 2, 2, 1$. Empezando la primera etapa con la asociaciación de 2 estados, est decir 3 teniendo en cuenta un estado fictivo ($n = 5$, $q = 1$)

---

[27] De hecho, Huffman fue estudiantes de maestria de Fano, trabajando en el MIT. Su tesis era de probar que el código de Fano era óptimo: Huffman propusó su propio código, andando al revés del enfoque de Fano, y demostró que era óptimo (Stix, 1991).

[28] Si no, se puede elegir $q = \left\lceil \frac{n-d}{d-1} \right\rceil$, y completar $\mathcal{X}$ con $d + q(d-1) - \alpha$ simbolos fuente fictivos de probabilidades ceros, le que no va a cambiar ni la entropia, ni el largo promededio del códico aferente.

van a quedar 3 estados en la secunda etapa, dando un código de largos $2, 2, 1, 1$, es decir de largo promedio mas pequeño.

Teorema 2-12 (Óptimalidad del código de Huffman)**.** *El algoritmo de Huffman da un código* c^{huf} *de largo promedio mínimo en la clase de los códigos unicamente decodables y los libre de prefijo (se recordará que con los largos de códigos unicamente decodable, siempre se puede construir un código libre de prefijo), es decir* $L^{\mathrm{opt}} = L\left(c^{\mathrm{huf}}\right)$.

Demostración. Una prueba es dada por ejemplo en (Cover & Thomas, 2006, Sec. 5.8) en el caso binario, pero la extensión para $d > 2$ es un poco mas subtile. La prueba mas general es dada por Huffman (Huffman, 1952) y se consigue también por parte en (Pigeon, 2003). Suponnemos que $q \geq 1$ (sino, el resultado es obvio). Las etapas son

- Sean j, k dos indices. Si c^{opt} es un código optimo, y c un código tal que $l(x_i) = l_i^{\mathrm{opt}}, \quad i \neq j, k, \quad l_j = l_k^{\mathrm{opt}} \quad \& \quad l_k = l_j^{\mathrm{opt}}$, se obtiene $0 \leq L(c) - L^{\mathrm{opt}} = \sum_i p_i \left(l_i - l_i^{\mathrm{opt}}\right) = (p_j - p_k)\left(l_k^{\mathrm{opt}} - l_j^{\mathrm{opt}}\right)$. Entonces $p_j > p_k \Rightarrow l_j^{\mathrm{opt}} \leq l_k^{\mathrm{opt}}$.

- Sea m el número de simbolos fuente con un código de largo máximo $l_{\mathrm{máx}}$ y $m' = \mathrm{mín}(m, d)$. Del punto anterior, los m simbolos con palabra código de largo máximo son los de probabilidades mas pequeñas.

- Como descrito antes, se puede permutar las letras códigos de una profundez del arbol de Kraft sin cambiar ni el aspecto libre de prefijo, ni el largo promedio. Se puede entonces considera el código óptimo tal que los m' simbolos de probabilidades las mas pequeñas tienen el mismo nudo padre, *i. e.,* solamente la última letra código cambia entre ellos.

- Suponemos que $m' = m < d$. Sea una "super-fuente" $\mathcal{X}^{(2)} = \left\{x_i^{(2)}\right\}_{i=1}^{n-m'+1}$ con $x_i^{(2)} = x_i, \quad 1 \leq i \leq n - m'$ de probabilidades respectivas $p(x_i)$ y $x_{n-m'+1}^{(2)} \equiv \{x_i\}_{i=n-m'+1}^{n}$ de probabilidad $p_{n-m'+1} + \cdots + p_n$ (se "plegan" las m' hojas en un super-simbolo). La códificación óptima es entonces una codificación libre de prefijo de $\mathcal{X}^{(2)}$, "arbol raiz" del código óptimo, a la cual se añade una letra código ζ_k diferente a cada simbolo del super-simbolo $x_{n-m'+1}^{(2)}$. La profundez máxima del código arbol es $l_{\mathrm{máx}} - 1$ y debe ser llena, en el sentido de que no debe tener un nudo que sea ni una hoja, ni un prefijo. En el caso contrario, se podría desplazar un simbolo de $x_{n-m'+1}^{(2)}$ al nudo "vacio" de la profundes $l_{\mathrm{máx}} - 1$, sin cambiar el aspecto libre de prefijo, pero ganando una letra código sobre un simbolo, *i. e.,* hacer un código libre de prefijo con un largo promedio menor. Sería contradictorio con la optimalidad del código inicial.

- Para códificar $\mathcal{X}^{(2)}$, se necesita por lo menos $\lceil \log_d(n - m' + 1) \rceil$ profundez en el arbol raiz. En esta profundez (máxima en el caso optimistico), hay $d^{\lceil \log_d(n-m'+1) \rceil} \geq n - m' + 1$ nudos. En la última profundez pueden ser todos ocupados si y solamente si $d^{\lceil \log_d(n-m'+1) \rceil} = n - m' + 1$. En otras palabras, es posible si y solamente si existe un entero k tal que $n - m' + 1 = d^k$, es decir, con $n = d + q(d-1)$, que teniamos el entero $q = \frac{d^k - d}{d-1} + \frac{m'-1}{d-1}$. La primera fracción $\frac{d^k - d}{d-1} = d^{k-1} + \cdots + 1$ siendo entera, q no puede ser entero con $m' < d$. En otros terminos, necesariamente $m' = d$, *i. e.,* los d simbolos de

probabilidad mas debiles son el la última profundez y se puede elegir que compartent el mismo nudo padre.

- Sea $c^{\mathrm{opt},(1)}$ el código óptimo correspondiente a $\mathcal{X}$ y $c^{(2)}$ el código "padre" sobre $\mathcal{X}^{(2)}$ ($c^{\mathrm{opt},(1)}$ quitando la último letra código de los simbolos juntados, *i. e.,* con la raiz común de estos). De la misma manera, sea $c^{\mathrm{opt},(2)}$ un código óptimo sobre $\mathcal{X}^{(2)}$ y $c^{(1)}$ el que se obtiene deplegando el super-simbolo $x^{(2)}_{n-d+1}$ en d hojas. De $L^{\mathrm{opt},(1)} = L\left(c^{(2)}\right) + p_{n-d+1} + \cdots + p_n$ (pasar de $\mathcal{X}^{(2)}$ a $\mathcal{X}$ se añade solo una letra palabra a los simbolos del super-simbolo) y $L\left(c^{(1)}\right) = L^{\mathrm{opt},(2)} + p_{n-d+1} + \cdots + p_n$ se obiene $\left(L^{\mathrm{opt},(1)} - L\left(c^{(1)}\right)\right) + \left(L^{\mathrm{opt},(2)} - L\left(c^{(2)}\right)\right) = 0$. Cada termino entre parentesis siendo positivos, valen necesariamente cero (la suma de terminos positivos vale cero si y solo si todos son nulos). En conclusión, $c^{(2)}$ padre de $c^{\mathrm{opt},(1)}$ queda óptimo, $c^{(2)} \equiv c^{\mathrm{opt},(2)}$ (y $c^{(1)} \equiv c^{\mathrm{opt},(1)}$).
- Notando que $|\mathcal{X}^{(2)}| = n-(q-1)(d-1)$, el razonamiento se progaga por inducción, pasando de $c^{\mathrm{opt},(k)}$ a $c^{\mathrm{opt},(k+1)}$ juntando los d super-simbolos de probabilidades mas debiles, hasta tener un super-simobolo tendiendo todos los simbolos, $\left|\mathcal{X}^{(K)}\right| = 1$, raiz del arbol.

□

De esta prueba, se puede ver que

- Cada profundez siendo llena, los largos obtenidos van a saturar la desigualdad de Kraft-McMillan.
- Si si $\frac{n-d}{d-1}$ no es entero, en lugar de completar $\mathcal{X}$ con simbolos fictivos se puede empezar el algoritmo de Huffman juntando los $n - d - \left\lfloor \frac{n-d}{d-1} \right\rfloor (d-1) + 1$ simbolos fuentes de probabilidades mas debiles en un super-simbolo, y luego hacer la bucla descrita (juntando por super-simbolos de d simbolos en cada bucla); en este caso, no se satura mas la desigualdad de Kraft-McMillan.
- Obviamente, en el caso binario $d = 2$, no es necesario completar $\mathcal{X}$ por estados fuentes, o empezar con menos de d simbolos juntados (n es necesariamente de la forma $n = d + q(d-1) = 2 + q$).
- El algoritmo no permite conocer los largos de manera analitica en función de p_i, y tampoco el largo promedio. Se los pueden deducir solamente implementando el algoritmo (una vez que es construido). Era el caso también en el enfoque de Fano.

Volviendo al código ingenuo, sería óptimo (y equivalente a los de Fano y de Shannon) para una distribución uniforme. En este contexto, la entropia es $H_d(X) = \log_d |X|$, precisamente la incerteza del enfoque de Hartley que corresponde a los numeros de dit necesarios para condificar (ingenuosamente) la fuente.

Se notara que, tratando de una fuente $\{X_t\}_{t\in\mathbb{Z}}$ de variables independientes, se puede codificar la fuente con un largo promedio arbitrariamente cerca de $H(X)$. El principio es de considerar vectores $\begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^t$ viviendo sobre $\mathcal{X}^n$, llamado extensión de orden n de la fuente, con un código unicamente decodable (o libre de prefijo) de esta extensión (es llamado *codificación de la extensión* (no es necesariamente una extension de c). Así, $H_d(X_1, \ldots, X_n) \leq L^{\mathrm{opt},n} < H_d(X_1, \ldots, X_n) + 1$, es decir, de la independencia,

$$H_d(X) \leq \frac{L^{\mathrm{opt},n}}{n} < H_d(X) + \frac{1}{n} \quad \text{por simbolo}$$

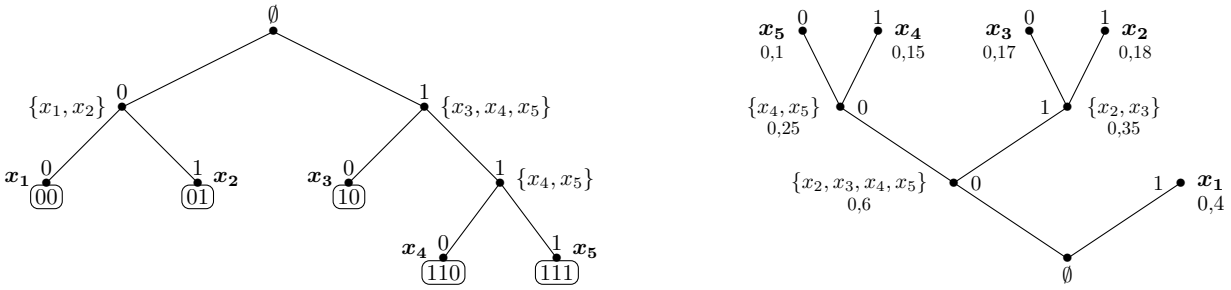


Figura 2-21: Construcción de un código binario sobre $\mathcal{C} = \{0\,,\,1\}$ asociado al vector de probabilidad $p_X = \begin{bmatrix}0{,}4 & 0{,}18 & 0{,}17 & 0{,}15 & 0{,}1\end{bmatrix}^t$ sobre el arbol de Kraft. En este caso, $H_2(X) \approx 2{,}1514$ (a): enfoque de Fano, saliendo de la raiz. En cada nudo, se menciona el conjunto de simbolos que van a tener el código correspondiente (en negro cuando es un solo simbolo). Se pasa de una profundez a la otra dividiendo los conjunto en sub-conjuntos a lo mas equiprobables. En esta construcción da el código $c^{\text{fa}}(x_1) = 00,\ c^{\text{fa}}(x_2) = 01,\ c^{\text{fa}}(x_3) = 10,\ c^{\text{fa}}(x_4) = 110,\ c^{\text{fa}}(x_5) = 111$ de largo promedio $L\left(c^{\text{fa}}\right) = 2{,}25$. (b): enfoque de Huffman, saliendo de las hojas. En cada nudo, se menciona el correspondiente (i) conjunto de simbolos, (ii) ζ_i de esta profundez/posición, (iii) la probabilidad asociada al conjunto. Se pasa de una profundez a la otra juntando los conjuntos menos probables en sobre-conjuntos. En negro son indicados los solo simbolos simples: van a tener el código agregando los de los nudos yendo de la raiz hasta las hojas. En esta construcción da el código $c^{\text{huf}}(x_1) = 1,\ c^{\text{huf}}(x_2) = 011,\ c^{\text{huf}}(x_3) = 010,\ c^{\text{huf}}(x_4) = 001,\ c^{\text{huf}}(x_5) = 000$ de largo promedio $L^{\text{opt}} = 2{,}2$.

(ver también (Rioul, 2007, cap. 13, teorema de Shannon)). Fijense que si $\lim\limits_{n\to\infty} \dfrac{L^{\text{opt},n}}{n} \to H(X)$, $\frac{L^{\text{opt},n}}{n}$ no es necesariamente decreciente con respeto a n. Eso es descrito figura 2-22. Lo mismo puede ocurir con el código de Shannon **y lo de Fano**. Ademas, el cardinal del alfabeto extendido $\mathcal{X}^n$ crece exponencialmente con n, lo que no permite elegir un n muy grande.

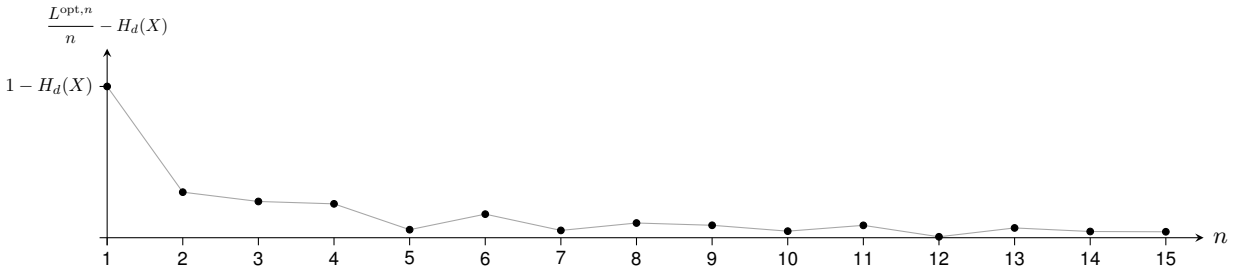


Figura 2-22: $\frac{L^{\text{opt},n}}{n} - H_d(X)$ (puntos), diferencia entre el largo promedio óptimo por simbolo de las extensiones $\mathcal{X}^n$ de orden n de la fuente $\mathcal{X}$ y la cota inferior en función de n. La linea llena en grise sirve como guia. En esta ilustración se usa el ejemplo lo mas simple con $d = 2$ y $p = [0{,}33 \quad 0{,}67]^t$.

Para codificar una fuente, que se haga el código óptimo o de Shannon, hace falta usar la distribución de probabilidad de la fuente X. Practicamente, es usual que no se la tiene. Frecuentemente, es estimada a partir de datos, o, dicho de otra manera, se códifica con una distribución que no es la distribución de la fuenta. Una pregunta que surge es de saber que se pierde usando una distribución no adaptaba (o “falsa”). La respuesta

general no es obvia, pero tratando del c/'odigo de Shannon se puede contestar:

Teorema 2-13 (Código falso de Shannon)**.** *Sea $c^{\mathrm{sh}}(p)$ el código de Shannnon sobre el alfabeto código $\mathcal{C} = \{\zeta_1, \ldots, \zeta_d\}$ asociado a la distribución p. Sea X fuente sobre $\mathcal{X}$, de distribución p_X y q una distribución cualquiera (ej. estimada de p_X presupuesta. . .). Entonces el largo promedio $L_{c^{\mathrm{sh}}(q)}$ del código $c^{\mathrm{sh}}(q)$ aplicado a la fuente X satisface las desigualdades siguientes*

$$H_d(p_X) + D_d\left(p_X \| \, q\right) \;\leq\; L_{c^{\mathrm{sh}}(q)} \;<\; H_d(p_X) + D_d\left(p_X \| \, q\right) + 1$$

Demostración. Por definición,

$$L_{c^{\mathrm{sh}}}(q) = \sum_{x \in X} p_X(x) \left\lceil -\log_d q(x) \right\rceil$$

La desigualdad viene de $a \leq \lceil a \rceil < a + 1$ y escribiendo $-p_X \log_d q = -p_X \log_d p_X + p_X \log\left(\frac{p_X}{q}\right)$. □

Olvidando el posible extra dit (penser a la códification block), este teorema da una interpretación operacional a la entropia relativa, o divergencia de Kullback-Leibler. Esta cuantifica la perdida en termino de largo promedio codificando con una distribución falsa. Dicho de otra manera, usando q en lugar de p_X, se usa la información de p_X porque se códifica la fuente X, pero suponiendo la distribucíon q, se piedre lo que representa la información relativa o de p_X con respeto a la referencia (distribución supuesta) q.

Existent varios otros modos de codificar simbolos. En particular, con la meta de transmitir los simbolos codificados en un canal de comunicación, a veces no es oportuno de compresar drasticamente. Existen por ejemplo codificaciones que permiten una corrección de error en la recepción. Pueden tomar en cuenta las caracteristicas del canal de transmisión. Estos van mas alla de la ilustración de esta sección. Ver (Berlekamp, 1974; Gallager, 1978; Sayood, 2003; Cover & Thomas, 2006; Rioul, 2007) entre otros para tener mas detalles sobre varios esquemas de codificación/compresión.

2.5.4 Gas perfecto

En el marco del gas perfecto

Va donner un lien avec Boltzmann

Feder Merhav IT'94 et lien avec discrimination; Vacisek en test de Gaussianite et cf plus loin avec generaloses Go75 etc

2.6 Entropias y divergencias generalizadas

A pesar de que la entropia de Shannon y sus cantidades asociadas demostraron sus potencias tan de un punto de vista descriptivo que en termino de aplicaciones en la transmisión de la información y la compresión, varios nociones informacionales, de tipo entropias o divergencias, aparecieron luego. En esta sección no se desarollará todos los enfoques ni todas las aplicaciones tan la literatura es importante. La meta es dar los caminos conduciendo a las generalizaciones de la entropia de Shannon por un lado, y de la divergencia de Kullback-Leibler por el otro lado. No son siempre vinculados, a pesar de que sea desirable que a cada entropia sean asociados nociones de entropias condicionales y relativas.

2.6.1 Entropias y propiedades

Si la entropia de Shannon fue el punto de salida fundamental en todo el desarollo de la teoria de la información, un poco mas de una decada despues de su papel clave y muy completo, Rényi propuso una medida generalizada (Rényi, 1961). Su punto de vista fue mas matematico que fisico o ingeniero. Retomó los axiomas de Fadeev (Fadeev, 1956, 1958; Khinchin, 1957) a probabilidades incompletas $p = \begin{bmatrix} p_1 & \cdots & p_n \end{bmatrix}^t, \quad p_i \geq 0, \quad w_p = \sum_i p_i \leq 1$: (i) la invarianza de $H(p)$ por permutación de os p_i, (ii) la continuidad de la incerteza elemental $H(p_i)$ (p_i visto como probabilidad incompleta), (iii) $H\left(\frac{1}{2}\right) = 1$, (iv) la aditividad $H(p \otimes q) = H(p) + H(q)$ donde $p \otimes q$ es el producto de Kronecker [2], *i. e.,* probabilidad conjunta de dos variables independientes, y consideró en lugar de la recursividad un axioma dicho de valor promedio, axioma muy parecido a la recursividad. Para p y q probabilidades incompletas tales que $p \cup q = \begin{bmatrix} p_1 & \cdots & p_n & q_1 & \cdots & q_m \end{bmatrix}^t$ sea incompleta ($w_p + w_q \leq 1$), el axioma (v) es $H(p \cup q) = \frac{w_p\, H(p) + w_q\, H(q)}{w_p + w_q}$. Demostró que con (v) en lugar de la recursividad, el conjunto de axiomas conduce de nuevo a la entropia de Shannon. La generalización propuesta por Rényi era de generalizar el axioma (v) remplazando la media aritmético por una media generalizada (v') $H^{\mathrm{r}}(p \cup q) = g^{-1}\left(\frac{w_p\, g\left(H^{\mathrm{r}}(p)\right) + w_q\, g\left(H^{\mathrm{r}}(q)\right)}{w_p + w_q}\right)$ con g estrictamente monotona y continua, llamado media *cuasi-aritmética, o quasi-lineal, o de Kolmogorov-Nagumo*. De las propiedades de la media cuasi-aritmetica (Nagumo, 1930; Kolmogorov, 1930, 1991; Hardy et al., 1952), eso es equivalente a buscar una entropia elemental $H^{\mathrm{r}}(p_i)$ y remplazar la media aritmética $\sum_i p_i H^{\mathrm{r}}(p_i)$ por una media de Kolmogorov-Nagumo, $g^{-1}\left(\sum_i p_i g\big(H^{\mathrm{r}}(p_i)\big)\right)$. Rényi propusó la función de Kolmogorov-Nagumo $g_\beta(x) = 2^{(\beta-1)x}, \quad \beta > 0, \quad \beta \neq 1$, probando que la entropia que los axiomas (i)-(ii)-(iii)-(iv)-(v') se cumplen y conduce a la entropia de Rényi de un vector de probabilidad p,

$$H^{\mathrm{r}}_\beta(p) = \frac{1}{1-\beta} \log_2 \left(\sum_{i=1}^{n} p_i^\beta \right)$$

Relaxando el axioma (iii), se puede elegir $g_\beta(x) = a^{(\beta-1)x}, \quad a > 0, \quad a \neq 1$; el logaritmo será de la base a cualquiera; En lo que sigue, usaremog $\log$ sin precisar la elección de base. Rényi nombró esta medida de incerteza *entropia de orden* β. Notablemente,

$$H^{\mathrm{r}}_1(p) \equiv \lim_{\beta \to 1} H^{\mathrm{r}}_\beta(p) = H(p) \quad \text{entropia de Shannon}$$

la entropia de Shannon. En su papel, Rényi introdujo una ganancia de información, parecida a una entropia

relativa, probando que las solas entropias admisibles son la de Shannon y la que introdujo. Volveremos en la sección siguiente sobre esta entropia relativa, o divergencia de Rényi. Por axiomas, las propiedades [P1] (continuidad), [P2] (invarianza por permutación) y [P10] (additividad) de la entropia de Shannon se conservan entonces en el marco de Rényi y se pierde [P7] (recursividad), todavía por axiomas. Veremos luego la otras que se conservan o modifican en un marco mas general.

Unos años después de Rényi, de la famosa escuela matematica checa, J. Havrda & F. Charvát en (Havrda & Charvát, 1967) (ver también (Vajda, 1968, en checo)) volvieron a los axiomas de Khintchin, para extender la entopia de Shannon, *i. e.,* considerando (i) la invarianza por permutación, (ii) la continuidad, (iii) la expensividad, (iv) $H^{\mathrm{hc}}(1) = 0$ y $H^{\mathrm{hc}}\left(\frac{1}{2}, \frac{1}{2}\right) = 1$, pero generalisando la recursividad por (v) $H^{\mathrm{hc}}(p_1, \ldots, p_n) = H^{\mathrm{hc}}(p_1, \ldots, p_{n-2}, p_{n-1} + p_n) + \beta(p_{n-1} + p_n)^{\beta} H^{\mathrm{hc}}\left(\frac{p_{n-1}}{p_{n-1}+p_p}, \frac{p_n}{p_{n-1}+p_p}\right), \quad \beta > 0$ [29]. Con $\beta = 1$ se recupera la recursividad estandar, pero con $\beta \neq 1$ eso permite dar un peso diferente a la incerteza del estado interno *i. e.,* probabilidades que se juntan (la describen como clasificación refinada). Estos axiomas conducen necesariamente a la entropia (teorema 1)

$$H^{\mathrm{hc}}_{\beta}(p) = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_i p_i^{\beta}\right)$$

que nombraron β-*entropia structural*. De nuevo, relaxando el axioma (iv), se puede remplazar en el coeficient $2^{1-\beta}$ por $a^{1-\beta}, \quad a > 0, \quad a \neq 1$. De nuevo, parae que la entropia de Shannon es un caso particular,

$$H^{\mathrm{hc}}_{1}(p) \equiv \lim_{\beta \to 1} H^{\mathrm{hc}}_{\beta}(p) = H(p) \quad \text{entropia de Shannon}$$

(continuidad), [P6] (expansabilidad) de Shannon en este marco. Se probó también que se conserva la propiedad de concavavidad con respeto a los p_i [P8], la de maximalidad [P5] alcanzada para una distribución uniforma (teorema 2). Aun que no aparece así en el papel, satisface la propiedad de Schur-concavidad [P9] (teorema 3). A pesar de que mencionan que H^{hc}_{β} sea diferente que H^{r}_{β}, es sencillo ver que hay un mapa uno-uno entre las dos entropias. Se notara en un marco mas general otras propiedades.

Independiente de Havrda & Charvát, todavía en el este, en la escuela húgara, Z. Daróczy en (Daróczy, 1970) defino la entropia H^f a partir de una *función información* f satifaciendo (i) $f(0) = f(1)$, (ii) $f\left(\frac{1}{2}\right) = 1$ y la ecuación funcional (ii) $f(x) + (1-x) f\left(\frac{y}{1-x}\right) = f(y) + (1-y) f\left(\frac{x}{1-y}\right)$ sobre $\{(x, y) \in [0; 1)^2, \quad x + y \leq 1\}$, siendo $H^f(p) = \sum_{i=2}^{n} s_i f\left(\frac{p_i}{s_i}\right), \quad s_i = \sum_{j=1}^{i-1} p_j$. Daróczy mostró que si f es medible, o continua en 0, o no negatiba y acotada, necesariamente $f(x) = h_2(x) = -x \log_2 x - (1-x) \log_2 (1-x)$, conduciendo a la entropia de Shannon (teorema 1; ver también (Lee, 1964; Tverberg, 1958; Kendall, 1964)). En otros terminos, su axioma (v) es alternativa a la recursividad. Para extender la entropia de Shannon, propuso extender este axioma (v) por la ecuación funcional $f_\beta(x) + (1-x)^{\beta} f_\beta\left(\frac{y}{1-x}\right) = f_\beta(y) + (1-y)^{\beta} f_\beta\left(\frac{x}{1-y}\right)$, lo que condujo necesariamente a la entropia (teoremas 2 y 3)

$$H^{\mathrm{d}}_{\beta}(p) = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_i p_i^{\beta}\right)$$

---

[29] En sus papel, lo imponen para cualquier pars p_i, p_j sin imponer la invarianza por permutación, pero es equivalente a la exposición de este parafo.

es decir nada mas que la entropia introducida por Havdra & Charvát. En lo que sigue, se la denota H_β^{hcd}. Sin embargo, el estudio de Daróczy fue mas intensivo que el de Havdra & Charvát. Primero, notó el mapa entre su entropia y la de Rényi. Adicionalmente a Havdra-Charvát probaron que se conserva la propiedad [P2] (invarianza por permutación, que no era un axioma en su enfoque), $H_\beta^{\mathrm{hcd}}\left(\frac{1}{2},\frac{1}{2}\right)=1$ (lo llama normalización), la expansividad [P6], una additividad extendida, una recursividad extendida precisamente del modelo de Havrda-Charvát (teorema 4). Probó también [P4], positividad alcanzado en el caso deterministico y la máximalidad [P5] en el caso uniforme (teorema 6), que incidentalmente $H_\beta^{\mathrm{hcd}}\left(\frac{1}{\alpha},\ldots,\frac{1}{\alpha}\right)$ crece con el cardinal $|\mathcal{X}|=\alpha$. Muy interesante también es se puede definir una entropia condicional en el mismo modelo que en el caso de Shannon $H_\beta^{\mathrm{hcd}}(X|Y)=\sum_y\left[p_{X|Y}(x,y)\right]^\beta H_\beta^{\mathrm{hcd}}(p_{X|Y}(\cdot,y))$, que existe una regla de cadena [P14], $H_\beta^{\mathrm{hcd}}(X,Y)=H_\beta^{\mathrm{hcd}}(Y)+H_\beta^{\mathrm{hcd}}(X|Y)$ y que [P16] condicionar reduce la entropia $H_\beta^{\mathrm{hcd}}(X|Y)\leq H_\beta^{\mathrm{hcd}}(X)$ (teorema 8). Mostró también que si se pierde la additividad, se obiene para X e Y independientes $H_\beta^{\mathrm{hcd}}(X,Y)=H_\beta^{\mathrm{hcd}}(X)+H_\beta^{\mathrm{hcd}}(Y)+\left(2^{1-\beta}-1\right)H_\beta^{\mathrm{hcd}}(X)H_\beta^{\mathrm{hcd}}(Y)$. La propiedades de regla de cadena le permitió revisitar la caracterisación de un canal de transmisión y redefinir una capacidad canal extendidas (capacidad tipo β; basicamente se usa el mismo enfoque que Shannon, pero usando H_β^{hcd} en lugar de H, ver sección 6 del papel).

Estas entropia fueron (re)descubiertos varios otras veces y/o estudiados mas detenidamente en varios campos y varios extensiones fueron introducidas (Varma, 1966; Onicescu, 1966; Kapur, 1967; Vajda, 1968; Lindhard & Nielsen, 1971; Arimoto, 1971; Burg, 1972; Aczél & Daróczy, 1975; Sharma & Mittal, 1975, 1975; Sharma & Taneja, 1975; Mittal, 1975; Boekee & van der Lubbe, 1980; Ferreri, 1980; Tsallis, 1988; Rathie, 1991; Kaniadakis, 2001, entre otros). Un primer enfoque mas general es debido a S. Arimoto en los primeros años de la decada 1970 (Arimoto, 1971) y rediscubierto y estudiado con mas detaller y una decada despues por Burbea y Rao (Burbea & Rao, 1982) y luego estudiado por Salicrú (Salicrú, 1987). La medida propuesta, llamada ϕ-entropia, es definida por

$$H_\phi(p)=-\sum_i\phi(p_i)\qquad\text{con}\qquad\phi\text{ convexa}$$

Burbea y Rao asociaron una medida de divegencia a esta entropia. Las ϕ-entropias contienen Shannon como caso particular ($\phi(x)=x\log x$), así que la clase de Havdra-Charvát-Daróczy ($\phi(x)=\frac{x-x^\beta}{2^{1-\beta}-1}$) como mencionado, pero no la clase de Rényi. De hecho, las ϕ-entropias se enmarcan en una clase un poco mas amplia, llamada (h,ϕ)-entropias (Salicrú, Menéndez, Morales & Pardo, 1993; Menéndez, Morales, Pardo & Salicrú, 1997),

Definición 2-18 ((h,ϕ)-entropia)**.** *La (h,ϕ)-entropia de una distribución de probabilidad p_X definida sobre $\mathcal{X}$ de cardinal finito $|\mathcal{X}|=\alpha$ es definida por*

$$H_{(h,\phi)}(X)=H_{(h,\phi)}\left(p_X\right)=h\left(\sum_{x\in\mathcal{X}}\phi\left(p_X(x)\right)\right)$$

donde o

- *ϕ es concava y h creciente, o*

- ϕ *es convexa y* h *decreciente*

Frecuentemente, se supone adicionalmente que ϕ *y* h *son de clase* C^2*, que* $\phi(0) = 0$ *(incerteza elemental asociada a un estado de probabilidad nula vale cero) y, sin perdida de generalidad, que* $h(\phi(1)) = 0$*.*

(ver también (Esteban, 1997) para una generalización aún mas amplia). Cuándo $h(x) = x$ se recupera la ϕ-entropia, incluyendo la de Shannon y las de Havdra-Charvát-Daróczy. Ademas, la familia de Rényi cae tambíen en esta familia ($\phi(x) = x^\beta$ y $h(x) = \frac{\log x}{1-\beta}$) así que todas las entropias evocadas en el parafo anterior.

Obviamente, de las propiedades de la entropia de Shannon, se conservan [P1] (continuidad), [P2] (invariaza por permutación), [P3] (invarianza por transformación biyectiva de X), [P6] (expansabilidad, debido a $\phi(0) = 0$).

Ademas se conserva la Schur convavidad con una reciproca

$$\forall\,(h,\phi), \quad p \prec q \;\Rightarrow\; H_{(h,\phi)}(p) \geq H_{(h,\phi)}(q)$$

$$\text{Reciprocamente, si} \quad \forall\,(h,\phi), \quad H_{(h,\phi)}(p) \geq H_{(h,\phi)}(q) \quad \text{entonces} \quad p \prec q$$

[P$_\phi$9] En otros terminos, se obtiene la relación de mayorisación si se cumple la relación de ordre entropicas para cualquier par de funciones entropicas (h,ϕ). La Schur-concavidad (y su reciproca) es consecuencia de la desigualdad de Schur (?, ?) o Hardy-Littlewood-Pólya (?, ?; Hardy et al., 1952) o Karamata (?, ?) (ver también (?, ?, Cap. 3, Prop. C.1 & Cap. 4, Prop. B.1) o (?, ?, Teorema II.3.1)): $p \prec q \;\Rightarrow\; \sum_i \phi(p_i) \leq \sum_i \phi(q_i)$ para toda función ϕ convexa.

Como consecuencia, se conserva la positividad [P4] gracia a $\phi(0) = 0$ y $h(\phi(1)) = 0$ (alcanzado en el caso determinístico) y la maximalidad [P5] (caso uniforme),

$$0 \leq H_{(h,\phi)}(X) \leq h\left(\alpha\,\phi\left(\frac{1}{\alpha}\right)\right)$$

Con respeto a la concavidad [P8], no se conserva en general:

[P$_\phi$8] Si h est concava, entonces $H_{(h,\phi)}$ es concava. Eso es una consecuencia de la concavidad de ϕ y decrecencia de h (resp. convexidad/crecencia) conjuntamente a la concavidad de h. La reciproca no es verdad. Por ejemplo, se puede ver que si $\beta < 1$, la entropia de Rényi es concava, pero se proba que existe un $\beta^*(\alpha) > 1$ tal que para cualquier $\beta \leq \beta^*(\alpha)$ se conserva la concavidad, a pesar de que h no sea necesariamente concava (Bengtsson & Życzkowski, 2006, p. 57).

Se pierde la propiedad de recursividad [P7], pero se puede vincular la entropia total con la obtenida juntando dos estados por una desigualdad:

[P$_\phi$7] Sean X definido sobre $\mathcal{X}$ y $\overline{X}$ sobre $\overline{X}$,

$$\begin{cases} \overline{\mathcal{X}} = \{x_1, \ldots, x_{\alpha-2}, \overline{x}_{\alpha-1}\} \quad \text{con el estado interno} \quad \overline{x}_{\alpha-1} = \{x_{\alpha-1}, x_\alpha\}, \\ p_{\overline{X}}(x_i) = p_X(x_i), \quad 1 \leq i \leq \alpha-1 \quad \text{y} \quad p_{\overline{X}}(\overline{x}_{\alpha-1}) = p_X(x_{\alpha-1}) + p(x_\alpha) \quad \text{distribución sobre } \overline{\mathcal{X}} \\ \overline{q}(x_j) = \dfrac{p_X(x_j)}{p_X(x_{\alpha-1}) + p_X(x_\alpha)}, \quad j = \alpha-1, \alpha \quad \text{distribución del estado interno} \end{cases}$$

$$H(p_X) \geq H(p_{\overline{X}})$$

Esta desigualdad es una de la desigualdad de Petrović (Kuczma, 2009, 43, Teorema 8.7.1), $\phi(a+b) \leq \phi(a) + \phi(b)$ para ϕ concava cancelando en 0 (y la conversa en el caso convexo), conjuntamente con h creciente (resp. decreciente). A parte en el caso de Shannon y el de Havdra-Charvát-Daróczy, no hay un vinculo inmediato entre $H(p_X)$ y $H(p_{\overline{X}})$.

Se conserva la superadditividad [P12]. De hecho, si ϕ es convexa (resp. concava) con $\phi(0) = 0$, $\forall\, 0 \leq a \leq 1,\ \phi(au) = \phi(au + (1-a)0) \leq a\phi(u)$ (resp. desigualdad reversa). Entonces, $\phi\left(p_{X,Y}(x_i, y_j)\right) = \phi\left(p_{X|Y}(x_i, y_j) p_Y(y_j)\right) \leq p_{X|Y}(x_i, y_j)\phi\left(p_Y(y_j)\right)$, *i. e.,* $\sum_{i,j} \phi\left(p_{X,Y}(x_i, y_j)\right) \leq \sum_{i,j} p_{X|Y}(x_i, y_j)\phi\left(p_Y(y_j)\right) = \sum_i \phi\left(p_Y(y_j)\right)$ (resp. desigualdad reversa). Se cierra la prueba con la decrecencia (resp. crecencia) de h.

Sin embargo, en general, se pierden las propiedades [P10] (additividad), y la propidad [P11] (subadditividad). En particular, se conserva solamente en el caso Shannon:

Teorema 2-14. *Sea $p_{X,Y}$ distribución conjunta de variables aleatorias discretas X y Y y p_x y p_Y las de X y de Y (marginales) cualquieras.*

$$\forall\, p_{X,Y} \quad H_{(h,\phi)}\left(p_{X,Y}\right) \leq H_{(h,\phi)}\left(p_X \otimes p_Y\right) \qquad \Leftrightarrow \qquad \phi(x) = -x \log x$$

i. e., $H_{(h,\phi)}$ es una función creciente de la entropia de Shannon

Demostración. La reciproca de este teorema es nada mas que la propidad [P11] con el hecho de que h es decreciente en este caso.

A continuación, la parte directa se demuestra en dos etapas:

- Con un caso particular sobre $\mathcal{X}$ e $\mathcal{Y}$ de cardenal 3 cada unos se proba de que la desigualdad no se puede cumplir, salvo si la función entropica ϕ' satisface a una ecuación funcional.
- la sola solución admisible de esta ecuación se recuce a $\phi(x) = -x \ln x$.

Etapa 1: Sea el vector de probabildad

$$p_{X,Y} = p_X \otimes p_Y - c \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \qquad \text{con} \qquad p_X = \begin{bmatrix} a \\ \alpha \\ 1-a-\alpha \end{bmatrix} \quad \text{y} \quad p_Y = \begin{bmatrix} b \\ \beta \\ 1-b-\beta \end{bmatrix}$$

donde $(a, \alpha, b, \beta) \in D$,

$$D = \{a, \alpha, b, \beta : \quad 0 < a, b < 1 \quad \wedge \quad 0 < \alpha \leq 1 - a \quad \wedge \quad 0 < \beta \leq 1 - b\}$$

y $c \in C_{a,\alpha,b,\beta}$,

$$C_{a,\alpha,b,\beta} = \left[-1 + \text{máx}\left\{ab, \alpha\beta, 1-a\beta, 1-\alpha b\right\},\ \text{mín}\left\{ab, \alpha\beta, 1-a\beta, 1-\alpha b\right\}\right]$$

Ahora, si ϕ es concava (resp. convexa)

$$\forall\, u, v \quad \phi(v) - \phi(u) \leq (v-u)\,\phi'(u),$$

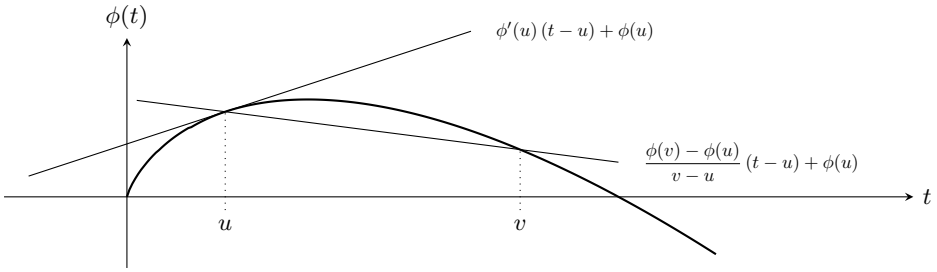


Figura 2-23: ϕ concava: la variación (cuerda) $\frac{\phi(v)-\phi(u)}{v-u}$ es menore que la derivada $\phi'(u)$. Aplicado a dos distribuciones p y q, de componentes p_i y q_i, se obtiene $H_\phi(q)-H_\phi(p)\le\sum_i(q_i-p_i)\phi'(q_i)$ con $H_\phi\equiv H_{(\mathrm{id},\phi)}$, id siendo la identidad.

i. e., la variación (cuerda) es menore que la derivada en b, como ilustrado figura **??** (desigualdad reversa para ϕ convexa). Aplicamos esta desigualdad a $u=p_{X,Y}(x,y)$ y $v=p_X(x)p_Y(y)$ y sumamos en x,y, para $(a,b)\in(0\,,\,1)^2$ (para que $C_{a,\alpha,b,\beta}$ no sea reducido a $\{0\}$), y $c\in\mathring{C}_{a,\alpha,b,\beta}$ donde $\mathring{}$ denota el interior de un conjunto, se obtiene para ϕ concava,

$$H_\phi\left(p_X\otimes p_Y\right)-H_\phi\left(p_{X,Y}\right)\;\le\;c\,g(a,\alpha,b,\beta,c),$$

(y la reversa para ϕ convexa), donde

$$g(a,\alpha,b,\beta,c)=\phi'\big(ab-c\big)+\phi'\big(\alpha\beta-c\big)-\phi'\big(a\beta+c\big)-\phi'\big(\alpha b+c\big).\tag{1}$$

Si existe $(s,u,t,v)\in\mathring{D}$ tal que $g(s,u,t,v,0)\neq0$. De la continuidad de ϕ', la función g es continua, y entonces exista un vecinaje $V_0\subset\mathring{C}_{s,u,t,v}$ de 0 tal que la función $c\mapsto g(x,u,y,v,c)$ tiene un signo constante sobre V_0. Eso permite concluir que $c\mapsto c\,g(x,u,y,v,c)$ no tiene un signo constante sobre V_0, y entonces de concluire que, de la desigualdad dedibo a la concavidad de ϕ (resp. convexidad), $H_\phi\left(p_{X,Y}\right)$ puede ser major (resp. menor) que $H_\phi\left(p_X\otimes p_Y\right)$, y entonces, con la crecencia (resp. decrecencia) de h que si $g(a,\alpha,b,\beta,0)$ no es identicamente cero sobre $\mathring{D}$, $H_{(h,\phi)}$ no puede ser subadditiva (conjunta vs product of marginales).

Etapa 2. Si $g(a,\alpha,b,\beta,0)=0$ sobre $\mathring{D}$, entonces ϕ' satisface la ecuación funcional

$$\phi'\big(ab\big)+\phi'\big(\alpha\beta\big)-\phi'\big(a\beta\big)-\phi'\big(\alpha b\big)=0,$$

así que no se puede usar el argumento de la etata 1 para concluir. Sin embargo, se puede solucionar esta ecuación funcional, siguiendo (?, ?, § 6) donde una ecuación funcional muy similar es estudiada. Por eso, se fija $(a,b)\in(0\,,\,1)^2$, se deriva la identidad precediente con respeto a α se multiplica el resultado por α para obtener

$$\alpha\beta\,\phi''(\alpha\beta)=\alpha b\,\phi''(\alpha b)\qquad\text{for}\qquad(\alpha,\beta)\in(0\,,\,1-a)\times(0\,,\,1-b).$$

Eso significa de que $x\,\phi''(x)$ es constante sobre $x\in(0\,,\,(1-a)\,\mathrm{máx}\{b,1-b\})$, y para cualquier par $(a,b)\in(0\,,\,1)^2$. Entonces, $x\,\phi''(x)$ es constante sobre $x\in(0\,,\,1)$, es decir que ϕ es necesariamebte de la forma $\phi(x)=-\lambda\,x\ln x+\mu x+\nu$. Debido a la continuidad de ϕ, queda valide sobre el cerrado $[0\,,\,1]$. De que se aplica a un vector de probabilidad, sumando a uno, se puede reducir el problema a $\mu=0$ (poniendo μ en ν sin

cambiar el valor de entropia). Ademas, la constante ν no altera la concavidad (resp. convexidad) de ϕ, así que se la puede translatar en la función h (sin cambia la monotonicidad). Para que ϕ sea concava (resp. convexa) hace falta tener $\lambda > 0$ (resp. $\lambda < 0$) así que, sin perdida de generalidad, λ puede ser puesta también en h. Tomar $\phi(x) = -x \ln x$ con h creciente o $\phi(x) = x \ln x$ con h decreciente es completamente equivalente, así que se puede fijar $\phi(x) = -x \ln x$ satisfaciendo la ecuación funcional, y h creciente.

$H_\phi = H$ siendo subadditiva (propidad [P11]), cualquier función queda subadditiva, lo que cierra la prueba. □

La definición de entropias generalizadas condicionales aparece mucho mas problematico. Por ejemplo, si se define a la Shannon, es decir definiendo $H_{(h,\phi)}(X|Y)$ tomando $\sum_{y\in\mathcal{Y}} p_Y(y) H_{(h,\phi)}\left(p_{X|Y}(\cdot, y)\right)$ se pierde la regla de cadena [P14]. Como se lo ha visto, en el marco de la entropia de Havdra-Charvát-Daróczy, se conserva la regla de cadena si se remplaza p_Y por su potencia p_Y^β. Sin embargo, generalizar este esquema en el caso general falla (la gracia en Havdra-Charvát-Daróczy viene de la propiedad de morfismo de la exponencial y del logaritmo). Como consecuencia, generalizar la noción se vuelve problematico también. Por ejemple se pierde el diagrama de Venn aparte si se define la entropia condicional a partir de la regla de cadena. Pero en este caso, si la superadditividad garantiza la positividad de la entropia condicional, se pierde la propiedad [P13] por perdida de la additividad, y por consecuencia la propiedad de positividad/independencia [P15] de una información mutua construida sobre un modelo diagrama de Venn. Veremos en la sección siguiente que un tercero camino puede ser usar divergencia. **ver con detalles [P16] (condiconar)**

versiones diferenciales y ver [P'3] (biyeccion), [P'2] (invarianza por rearreglo (?, ?))

2.6.2 Divergencias y propiedades

(1) Extension a la Renyi, (2) a la HC/D/T, Cressie Reads, Cressie Pardo, Vajda; (3) cf Burbea Rao: (4) generalization Czizar Vajda, et voir avec h phi avant meme Salicru. Cf aussi Bregmann, autres de Csizsar 2012 versikon Bregman; gupta Sharma 1976 BoeLub79, Vajda72, Salicru94 Orsak et Paris; application a le test d'adequation Pardo 99; MenMor97:5, Cf Pardo 2006 et ref.

FriSri08 pour Bregman; reapparition Fisher comme courbure, cf Varma, Jizba, MenMor97...

Aplicaciones poner acá MaxEnt nous, Kesavan, Kagan 63 A.M.

On the theory of Fisher's amount of information Sov. Math. Dokl., 4 (1963), pp. 991-993, etc, la codificación a la Renyi (Cambell, Hooda 2001, Bercher)

y la cuantificacion fina; EPI generalizada por Madiman, etc. Lutwak, Bercher etc., Kagan; Boeke 77 An extension of the Fisher information measure I. Csiszár, P. Elias (Eds.), Topics in Information Theory, North-Holland, Berlin/New York (1977), pp. 113-123 o Hammad o Vajda 73 o Ferentinos81 en el marco Fisher; Kesavan gene MaxEnt

Revisite capacite a la Daroczy? codage; parler de la quantification fine et HCD

2.7 Entropias cuanticas discretas

Mas alla caso de informaciones a partir de medida; caso infinito, continuo queda en discusiones

CAPÍTULO 3
Elementos de geometría diferencial

Pedro Walter Lamberti

ἀγεωμέμέτρητος μηδεις εισιτω
Que no ingrese nadie que no sepa geometría.
FRASE GRABADA EN LA ENTRADA DE LA ACADEMIA DE PLATÓN

3.1 Estructuras

Una de las nociones más elementales de la matemática es la de *Conjunto*. Un conjunto es una colección de elementos perfectamente caracterizados. Los elementos pueden ser de cualquier tipo: números, funciones, personas, autos, etc. El enfoque matemático moderno es ir montando estructuras de distinta naturaleza sobre un dado conjunto. En este capítulo comenzaremos con la noción de Espacio Topológico y llegaremos al concepto de Variedad Riemanniana. Este procedimiento ha mostrado ser de utilidad en el marco de la física, que es nuestro principal ámbito de interés. El mapa de ruta de las de las distintas estructuras que veremos en este capítulo es el siguiente:

- Espacio Topológico
- Espacio Métrico
- Variedad Topológica
- Estructura Diferenciable (Variedad Diferenciable)
- Estructura Afin (Noción de paralelismo)
- Estructura métrica (Finsler y Riemann)

Si bien existe una estructura intermedia entre la topológica y la diferenciable, que se conoce como *estructura lineal a trozos*, aquá prescindiremos de su estudio. A su vez, hay otras estructuras matemáticas que son usadas en el marco de las teorías físicas. Se destacan la estructura de producto interno sobre un conjunto complejo, la cual conduce a la noción de espacio de Hilbert, de fundamental importancia en Mecánica Cuántica; la estructura simpléctica, útil en Mecánica Clásica y la estructura de Khaler, de relevancia en teoría de cuerdas.

Comenzaremos con la noción de espacio topológico.

3.1.1 Espacio Topológico

Un conjunto arbitrario X está desprovisto de toda estructura que permita definir nociones tales como la *convergencia* de una sucesión de elementos de X, la *proximidad* de dos elementos de X, etc. En principio se dispone sólo de las operaciones elementales de *unión* $\bigcup$ e *intersección* $\bigcap$ de subconjuntos. Estas operaciones también pueden realizarse entre distintos conjuntos. Denotaremos con $\emptyset$ al conjunto vacío. Surge entonces el desafío entonces de construir alguna estructura matemática definida sobre X que permita definir, de manera precisa las nociones de proximidad, continuidad, convergencia, etc. Esto se logra a traveés de la idea de una **topología** sobre X.

Definición 3-19. *Una* **Topología** τ *sobre el conjunto* X *es una familia de subconjuntos de* X *que cumple con las siguientes condiciones:*

1. X *y* $\emptyset$ *están en* τ*:* $X,\ \emptyset \in \tau$
2. *La intersección de cualquier colección finita de elementos de* τ *está en* τ*:*

$$A_i \in \tau,\ \forall\, i = 1, \ldots, n \ \Rightarrow \bigcap_{i=1}^{n} A_i \in \tau$$

3. *La unión de una colección arbitraria - finita o no- de elementos de* τ*, pertenece a* τ*:*

$$A_\alpha \in \tau \ \Rightarrow \bigcup_{\alpha} A_\alpha \in \tau$$

Definición 3-20. *Al par* (X, τ) *lo llamaremos* **Espacio Topológico**. *Los conjuntos que están en* τ *se llaman abiertos.*

Ejemplos:

- **Topología trivial**. Es la que consta de sólo dos elementos, el conjunto vacío y el conjunto total $X : \tau = \{\emptyset, X\}$.
- **Topología discreta**. Es la que en todo subconjunto de X está en τ, es decir $\tau = \mathcal{P}(X)$ donde $\mathcal{P}(X)$ representa a las partes de X

- En los cursos elementales de análisis matemático hemos estudiado en $\mathbb{R}^n$, es decir el conjunto de $n-tuplas$ de números reales, la noción de bolas abiertas. Más precisamente, una bola abierta en $\mathbb{R}^n$ centrada en el punto $p=(p_1,\ldots,p_n)\in\mathbb{R}^n$ y de radio $r>0$ es el conjunto

$$\mathcal{B}_{r,p}=\{(x_1,\ldots,x_n)\quad \text{tal que}\quad 0\leq\sqrt{\sum_i(x_i-p_i)^2}<r\}$$

  La colección de todas las bolas abiertas en $\mathbb{R}^n$ constituyen una topología para $\mathbb{R}^n$. Se conoce como la **topología usual** de $\mathbb{R}^n$

Definición 3-21. *Un* **entorno** *de un punto $x\in X$ es un conjunto U que contiene a x y tal que existe un abierto V contenido en $U:\ x\in V\subseteq U\ \ con\ \ U\in\tau$.*

Definición 3-22. *Sea $f:X\to Y$ una función entre dos espacios topológicos (X,τ) e (Y,ω). f es una* **función continua** *en $x\in X$ sii dado cualquier entorno abierto $U\subset Y$ de $f(x)$, existe un entorno de x, $V\subset X$ tal que $f(V)\subset U$.*

Definición 3-23. *Un* **homomorfismo** *Ψ entre dos espacios topológicos (X,τ) e (Y,ω) es una función*

$$\Psi:X\to V\subseteq Y$$

biyectiva, continua y con inversa continua.

Definición 3-24. *Una* **sucesión** *en un conjunto X es una aplicación $s:\mathbb{N}\to X$ donde $\mathbb{N}$ es el conjunto de los números naturales. Denotaremos a la sucesión por $\{x_n\}$ donde $n\in\mathbb{N}$.*

En un espacio topológico podemos introducir la noción de convergencia de una sucesión. Obsérvese que ésto es posible gracias a que disponemos de la noción de conjunto abierto.

Definición 3-25. *Sea (X,τ) un espacio topológico y $\{x_n\},\ n\in\mathbb{N}$ una sucesión en X. Diremos que x es el* **límite** *de x_n si para todo entorno V de x, existe un $n_0\in\mathbb{N}$ tal que $\forall\, n\geq n_0$ se tiene que $x_n\in V$.*

Los límites de las sucesiones no tienen porque ser únicos. Una condición que debe cumplir el espacio topológico (X,τ) para que las sucesiones tengan un único límite es que dados dos puntos distintos $x\neq y$,con $x,y\in X$ existen entornos disjuntos de x e y.

A los espacios topológicos que satisfacen con esta condición se los llama espacios de Hausdorff o espacio T_2.

3.1.2 Espacios métricos

En el tercer ejemplo de espacio topológico, usamos la noción de métrica euclídea para definir las bolas abiertas en $\mathbb{R}$. El disponer de una métrica no es algo que ocurre en todo conjunto. Eso motiva la siguiente definición:

Definición 3-26. *Un* **Espacio Métrico** *en un conjunto X munido de una función $d : X \times X \to \mathbb{R}_+$ tal que se cumplen las condiciones:*

1. $d(x, y) \geq 0 \forall x, y \in X$ *y la igualdad se cumple sii* $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, y) \leq d(x, z) + d(z, y) \quad \forall\, x, y, z \in X$

La última condición se conoce como *desigualdad triangular*. Mas adelante en este libro veremos funciones $d : X \times X \to \mathbb{R}^0$ que no satisfacen ni la condición 2 ni la condición 3, pero que sin embargo sirven para medir cuán separados están dos puntos de X. En ese caso diremos que d es una distancia en X.

3.1.3 Variedad Topológica

Nuestra experiencia cotidiana de percibir que estamos inmersos en un espacio de 3 dimensiones, en el cual podemos medir ángulos y determinar distancias entre dos puntos, ha hecho que usemos estas caracteristicas de nuestro habitat, como motivación de la defición de ciertas estructuras matemáticas en espacios abstractos.

En primer lugar, con la noción de una variedad topológica buscaremos simular en un conjunto cualquiera, la noción de cercanía y dimensionalidad que tenemos en $\mathbb{R}^n$.

Definición 3-27. *Una* **Variedad Topológica n-dimensional** *es un espacio topológico $\mathcal{M}$ tal que es localmente euclídeo, es decir que para cada $x \in \mathcal{M}$ existe un entorno abierto U de x, homeomorfo a un abierto V de $\mathbb{R}^n$:*

$$\phi : U \subseteq \mathcal{M} \to \mathbb{R}^n$$

tal que

$$\phi : U \to V$$

y ϕ es un homeomorfismo. También pediremos que $\mathcal{M}$, como espacio topológico, sea un espacio Hausdorff.

A los pares (U, ϕ) se llaman cartas sobre $\mathcal{M}$. Se supone que la colección de todas las cartas cumbren completamente a $\mathcal{M}$. Las cartas permiten asignar *coordenadas* a $\mathcal{M}$:

$$\text{Si} \quad p \in U \subseteq \mathcal{M} \quad \text{entonces} \quad \phi : p \to (p_1, \ldots, p_n) \in \mathbb{R}^n$$

la colección de números reales $(p_1, \ldots, p_n)$ se llaman las coordenadas de p de acuerdo a la carta (U, ϕ). Podría suceder que un mismo punto p pertenezca a más de una carta, digamos (U_1, ϕ_1) y (U_2, ϕ_2). En ese caso hablaremos de un cambio de coordenadas:

$$\psi_2 \circ \phi_1^{-1} : \phi_1(U_1 \cap U_2) \to \psi_2(U_1 \cap U_2) \tag{2}$$

Si denotamos por $(p_1, \ldots, p_n)$ a las coordenadas correspondientes a la carta (U_1, ϕ_1) y por $(\tilde{p}_1, \ldots, \tilde{p}_n)$ a las correspondientes a la carta (U_2, ψ_2), entonces las funciones $\tilde{p}_i = \tilde{p}_i(p_1, \ldots, p_n)$ son funciones continuas, y dan el cambio de coordenadas. Estas funciones son invertibles con inversa continua.

Ejemplos de variedades topológicas son:

- $\mathbb{R}^n$. En este caso hay una carta coordenada global que cubre toda la variedad y donde el homeomorfismo es la identidad.
- $\mathbb{S}^n$, la esfera de dimensión n. Está definida como el conjunto:

$$\mathbb{S}^n = \left\{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \quad \text{tales que} x_1^2 + \cdots + x_{n+1}^2 = 1\right\}$$

  Se debe observar que al definir $\mathbb{S}^n$ no estamos pensando que está inmersa en $\mathbb{R}^n$.

CAPÍTULO 4
Geometría de la información

Esto es un epígrafe con texto simulado.
Esto es un epgrafe con texto simulado.
AUTOR DEL EPÍGRAFE, TÍTULO DE LA OBRA

4.1 La Sección 4.1

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm. Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simula- do, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm) [30] .

> Esto es un ejemplo de cita de mas de 40 palabras. Esto es un ejemplo de cita de mas de 40 palabras. Esto es un ejemplo de cita de mas de 40 palabras. Esto es un ejemplo de cita de mas de 40 palabras. Esto es un ejemplo de cita de mas de 40 palabras.

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm. Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simula- do, (Arial 10, interlineado de 1,5 líneas,

[30] Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas.

Tabla 4-1: Eso es un ejemplo de tabla

Título (negrita)	Título (negrita)	Título (negrita)
A	Texto simulado (normal)	Texto simulado (normal)
B	Texto simulado (normal)	Texto simulado (normal)

Fuente: Eso sería el fuente de la tabla

sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm).

Figura 4-24: Eso es una figura, con su leyenda sobre varias lineas para ver como queda en el texto. Eso es una figura, con su leyenda sobre varias lineas para ver como queda en el texto.

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm. Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simula- do, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm).

Ejemplo con respeto al capitulo **??**

Para ver que las referencias de capitulos andan: **??**; que las de secciones también **??**, de subsecciones **??**, de subsubsecciones **??**, de figuras **??**, y de tablas **??**.

Eso es una cita, para ver como queda (Cover & Thomas, 2006; Amari & Nagaoka, 2000).

CAPÍTULO 5
Aplicaciones

Esto es un epígrafe con texto simulado.
Esto es un epgrafe con texto simulado.
AUTOR DEL EPÍGRAFE, TÍTULO DE LA OBRA

5.1 La Sección 5.1

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm. Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simula- do, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm) [31] .

> Esto es un ejemplo de cita de mas de 40 palabras. Esto es un ejemplo de cita de mas de 40 palabras. Esto es un ejemplo de cita de mas de 40 palabras. Esto es un ejemplo de cita de mas de 40 palabras. Esto es un ejemplo de cita de mas de 40 palabras.

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm. Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simula- do, (Arial 10, interlineado de 1,5 líneas,

[31] Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas. Eso es una footnote sobre varias lineas.

Tabla 5-2: Eso es un ejemplo de tabla

Título (negrita)	Título (negrita)	Título (negrita)
A	Texto simulado (normal)	Texto simulado (normal)
B	Texto simulado (normal)	Texto simulado (normal)

Fuente: Eso sería el fuente de la tabla

sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm).

Figura 5-25: Eso es una figura, con su leyenda sobre varias lineas para ver como queda en el texto. Eso es una figura, con su leyenda sobre varias lineas para ver como queda en el texto.

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm. Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simula- do, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm).

Ejemplo con respeto al capitulo **??**

Para ver que las referencias de capitulos andan: **??**; que las de secciones también **??**, de subsecciones **??**, de subsubsecciones **??**, de figuras **??**, y de tablas **??**.

Eso es una cita, para ver como queda (Cover & Thomas, 2006; Amari & Nagaoka, 2000).

EPILÓLOGO

Este libro surge de la experiencia de los autores en el dictado del curso semestral ”Métodos de geometría diferencial en teoría de la información", que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. ...

Los autores

Referencías

Aczél, J. & Daróczy, Z. (1975). *On Measures of Information and Their Characterizations*. New-York: Academic Press.

Amari, S.-I. & Nagaoka, H. (2000). *Methods of Information Geometry*. Rhode Island: Oxford University Press.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, *65*(331), 1248–1255.

Arimoto, S. (1971). Information-theoretical considerations on estimation problems. *Information and control*, *19*(3), 181–194.

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20.

Barron, A. R. (1984). Monotonic central limit theorem for densities. Technical report no. 50, Department of Statistics, Stanford University.

Barron, A. R. (1986). Entropy and the central limit theorem. *The Annals of Probability*, *14*(1), 336–342.

Bengtsson, I. & Życzkowski, K. (2006). *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge: Cambridge University Press.

Berlekamp, E. R. (Ed.). (1974). *Key Papers in the Development of Coding Theory*. IEEE Press.

Blachman, N. M. (1965). The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, *11*(2), 267–271.

Boekee, D. E. & van der Lubbe, J. C. A. (1980). The R-norm information measure. *Information and Control*, *45*(2), 136–155.

Boltzmann, L. (1896). *vorlesungen über Gastheorie - I*. Leipzig, Germany: Verlag von Johann Ambrosius Barth.

Boltzmann, L. (1898). *vorlesungen über Gastheorie - II*. Leipzig, Germany: Verlag von Johann Ambrosius Barth.

Burbea, J. & Rao, C. R. (1982). On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, *28*(3), 489–495.

Burg, J. P. (1967). Maximum entropy spectral analysis. In *Proceedings of 37th Meeting, Society of Exploration Geophysics,*, Oklahoma City, Oklahoma.

Burg, J. P. (1972). The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, *37*(2), 375–376.

Burg, J. P. (1975). *Maximum entropy spectral analysis*. PhD thesis, Department of Geophysics, Stanford University, Stanford University, Stanford, CA.

Cambini, A. & Martein, L. (2009). *Generalized Convexity and Optimization: Theory and Applications*. Heidelberg: Springer Verlag.

Chenciner, A. (2017). La force d'une idée simple. *Gazette de la Société de Mathématiques Française*, *152*, 16–22.

Clavier, A. G. (1948). Evaluation of transmission efficiency according to Hartley's expression of information content. *Technical Journal of the International Telephone and Telegraph Corporation and Associate Companies*, *25*(4), 414–420.

Cohen, M. (1968). The Fisher information and convexity. *IEEE Transactions on Information Theory*, *14*(4), 591–592.

Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Cramér, H. (1946). *Mathematical Methods of Statistics*. New-York: Princeton University Press.

Darmois, G. (1935). Sur les lois de probabilités à estimation exhaustive. *Comptes rendus de l'Acadéie des Sciences*, *200*, 1265–1966.

Darmois, G. (1945). Sur les limites de la dispersion de certaines estimations. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *13*(1/4), 9–15.

Daróczy, Z. (1970). Generalized information functions. *Information and Control*, *16*(1), 36–51.

Dembo, A., Cover, T. M., & Thomas, J. A. (1991). Information theoretic inequalities. *IEEE Transactions on Information Theory*, *37*(6), 1501–1518.

Doob, J. L. (1936). Statistical estimation. *Transactions of the American Mathematical Society*, *39*(3), 410–421.

Edgeworth, F. Y. (1908). On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, *71*(3, 6 & 7), 381–397, 499–512 & 499–512.

Elias, P. (1957). List decoding for noisy channels. Technical Report 335, Research Laboratory of Electronics, MIT, MIT, Cambridge, MA.

Esteban, M. D. (1997). A general class of entropy statistics. *Applications of Mathematics*, *42*(3), 161–169.

Fadeev, D. K. (1956). On the concept of entropy of a finite probabilistic scheme (russian). *Uspekhi Matematicheskikh Nauk*, *11*(1(67)), 227–231.

Fadeev, D. K. (1958). *Fundations in Information Theory*, chapter On the concept of entropy of a finite probabilistic scheme (English traduction). New-York: McGraw-Hill.

Fano, R. M. (1949). The transmission of information. Technical Report 65, Research Laboratory of Electronics, MIT, MIT, Cambridge, MA.

Ferreri, C. (1980). Hypoentropy and related heterogeneity, divergence and information measures. *Statistica*, *2*, 155–167.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, *222*(594-604), 309–368.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725.

Flandrin, P. & Rioul, O. (2016). Laplume, sous le masque.

Fréchet, M. (1943). Sur l'extension de certaines evaluations statistiques au cas de petits echantillons. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *11*(3/4), 182–205.

Frieden, B. R. (2004). *Science from Fisher Information: A Unification*. Cambridge, UK: Cambridge University Press.

Gallager, R. (1978). Variations on a theme by Huffman. *IEEE Transactions on Information Theory*, *24*(6), 668–674.

Gallager, R. (2001). Claude E. Shannon: a retrospective on his life, work, and impact. *IEEE Transactions on Information Theory*, *47*(7), 2681–2695.

Gelfand, I. M. & Fomin, S. V. (1963). *Calculus of Variations*. Englewood Cliff, NJ, USA: Prentice Hall.

Gibbs, J. W. (1902). *Elementary Principle in Statistical Mechanics*. Cambridge, USA: University Press - John Wilson and son.

Guo, D., Shamai, S., & Verdú, S. (2005). Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, *51*(4), 1261–1282.

Hardy, G., Littlewood, J. E., & Pólya, G. (1952). *Inequalities* (2nd ed.). Cambridge, UK: Cambridge University Press.

Hartley, R. V. L. (1928). Transmission of informations. *The Bell System Technical Journal*, *7*(3), 535–563.

Havrda, J. & Charvát, F. (1967). Quantification method of classification processes: Concept of structural α-entropy. *Kybernetika*, *3*(1), 30–35.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), 1098–1101.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical Review*, *108*(2), 171–190.

Jaynes, E. T. (1965). Gibbs vs Boltzmann entropies. *American Journal of Physics*, *33*(5), 391–398.

Jaynes, E. T. (1968). Prior probabilities. *IEEE transactions on systems science and cybernetics*, *4*(3), 227–241.

Jaynes, E. T. (1982). On the rational of maximum-entropy methods. *Proceedings of the IEEE*, *70*(9), 939–952.

Jeffrey (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A*, *186*(1007), 453–461.

Johnson, O. (2004). *Information Theory and The Central Limit Theorem*. London: Imperial college Press.

Kaniadakis, G. (2001). Non-linear kinetics underlying generalized statistics. *Physica A*, *296*(3-4), 405–425.

Kapur, J. N. (1967). Generalized entropy of order α and type β. *The Mathematical Seminar*, *4*, 78–94.

Kapur, J. N. & Kesavan, H. K. (1992). *Entropy Optimization Principle with Applications*. San Diego: Academic Press.

Karush, J. (1961). A simple proof of an inequality of McMillan. *IEEE Transactions on Information Theory*, *7*(2), 118–118.

Kay, S. M. (1993). *Fundamentals for Statistical Signal Processing: Estimation Theory*. vol. 1. Upper Saddle River, NJ: Prentice Hall.

Kendall, D. G. (1964). Functional equations in information theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *2*(3), 225–229.

Khinchin, A. I. (1957). *Mathematical foundations of information theory*. New-York: Dover Publications.

Kolmogorov, A. N. (1930). Sur la notion de la moyenne. *Atti della Reale Accademia Nazionale dei Lincei*, *12*, 388–391.

Kolmogorov, A. N. (1991). On the notion of mean. In V. M. Tikhomirov (Ed.), *Selected Works of A. N. Kolmogorov*, volume I: Mathematics and Mechanics (pp. 144–146). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, *39*(3), 399–399.

Kraft Jr, L. G. (1949). A device for quantizing, grouping, and coding amplitude-modulated pulses. Master's thesis, Department of Electrical Engineering, MIT, Massachusetts Institute of Technology.

Krajči, S., Liu, C.-F., Mikeš, L., & Moser, S. M. (2015). Performance analysis of Fano coding. In *2015 IEEE International Symposium on Information Theory (ISIT)*, (pp. 1746–1750)., Hong-Kong, China.

Kuczma, M. (2009). *An Introduction to the Theory of Functional Equations and Inequalities: Cauchy's Equation and Jensen's Inequality* (2nd ed.). Basel: Birkhäuser.

Laplume, J. (1948). Sur le nombre de signaux discernables en présence de bruit erratique dans un système de transmission à bande passance limitéee. *Comptes Rendus de l'Academie des Sciences*, *226*, 1348–1349. Séance du 26 avril.

Lee, P. M. (1964). On the axioms of information theory. *The Annals of Mathematical Statistics*, *35*(1), 415–418.

Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). New-York: Springer-Verlag.

Lieb, E. H. (1975). Some convexity and subadditvity properties of entropy. *Bulletin of the American Mathematical Society*, *81*(1), 1–13.

Lieb, E. H. (1978). Proof of an entropy conjecture of Wehrl. *Communications in Mathematical Physics*, *62*(1), 35–41.

Lieb, E. H. & Loss, M. (2001). *Analysis* (2nd ed.). Providence, Rhode Island: American Mathematical Society.

Lindhard, J. & Nielsen, V. (1971). Studies in statistical mechanics. *Det Kongelige Danske Videnskabernes Selskab Matematisk-fysiske Meddelelser*, *38*(9), 1–42.

Lundheim, L. (2002). On Shannon and "Shannon's formula". *Telektronikk*, *98*(1), 20–29.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (3rd ed.). New-York: John Wiley & Sons.

Maxwell, J. C. (1867). On the dynamical theory of gases. *Philosophical Transactions of the Royal Society of London*, *157*, 49–88.

McMillan, B. (1956). Two inequalities implied by unique decipherability. *IEEE Transactions on Information Theory*, *2*(4), 115–116.

Menéndez, M. L., Morales, D., Pardo, L., & Salicrú, M. (1997). (h, ϕ)-entropy differential metric. *Applications of Mathematics*, *42*(1-2), 81–98.

Miller, R. E. (2000). *Optimization: Foundations and Applications*. New-York: John Wiley & Sons, inc.

Mittal, D. P. (1975). On additive and non-additive entropies. *Kybernetica*, *11*(4), 271–276.

Montagné, J.-C. B. (2008). *Transmissions. L'histoire des moyens de communication à distance depuis l'Antiquité jusqu'au milieu du xxe siècle*. Bagneux, JCB Montagné.

Nagumo, M. (1930). Über eine klasse der mittelwerte. *Japanese journal of mathematics: transactions and abstracts*, *7*, 71–79.

Onicescu, O. (1966). Energie informationnelle. *Comptes rendus de l'académie des Sciences. série 1, mathématiques*, *263*(3), 841–842.

Palomar, D. P. & Verdú, S. (2006). Gradient of mutual information in linear vector Gaussian channels. *IEEE Transactions on Information Theory*, *52*(1), 141–154.

Payaró, M. & Palomar, D. P. (2009). Hessian and concavity of mutual information differential entropy, and entropy power in linear vector gaussian channels. *IEEE Transactions on Information Theory*, *55*(8), 3613–3628.

Pearson, K. & Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. IV. on the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London A*, *191*, 229–311.

Pigeon, S. (2003). Huffman coding. In K. Sayood (Ed.), *Lossless Compression Handbook* chapter 4, (pp. 79–99). San Diego, CA: Academic Press.

Planck, M. (2015). *Eight Lectures on Theoretical Physics*. New-York: Columbia University Press.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, *37*(3), 81–91.

Rao, C. R. (1992). Information and the accuracy attainable in the estimation of statistical parameters. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Foundations and Basic Theory*, volume I (pp. 235–247). New York: Springer.

Rao, C. R. & Wishart, J. (1947). Minimum variance and the estimation of several parameters. *Mathematical Proceedings of the Cambridge Philosophical Society*, *43*(2), 280–283.

Rathie, P. N. (1991). Unified (r, s)-entropy and its bivariate measures. *Information Sciences*, *54*(1-2), 23–39.

Rényi, A. (1961). On measures of entropy and information. *in Proceeding of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 547–561.

Rioul, O. (2007). *Théorie de l'information et du codage*. Paris: Lavoisier.

Rioul, O. (2011). Information theoretic proofs of entropy power inequalities. *IEEE Transactions on Information Theory*, *57*(1), 33–55.

Rioul, O. (2017). Yet another proof of the entropy power inequality. *IEEE Transactions on Information Theory*, *63*(6), 3595–3599.

Rioul, O. & Flandrin, P. (2017). Le dessein de laplume. In *Colloque GRETSI*, Juan-les-Pins, France.

Rioul, O. & Magossi, J. (2014). On Shannon's formula and Hartley's rule: Beyond the mathematical coincidence. *Entropy*, *16*(12), 4892–4910.

Robert, C. P. (2007). *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation* (2nd ed.). New-York: Springer.

Salicrú, M. (1987). Funciones de entropía asociada a medidas de Csiszár. *Qüestiió*, *11*(3), 3–12.

Salicrú, M., Menéndez, M. L., Morales, D., & Pardo, L. (1993). Asymptotic distribution of (h, ϕ)-entropies. *Communications in Statistics – Theory and Methods*, *22*(7), 2015–2031.

Sayood, K. (Ed.). (2003). *Lossless Compression Handbook*. San Diego, CA: Academic Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(4), 623–656.

Shannon, C. E. & Weaver, W. (1964). *The Mathematical Theory of Communication*. Urbana, USA: The University of Illinois Press.

Sharma, B. D. & Mittal, D. P. (1975). New non-additive measures of entropy for discrete probability distributions. *Journal of Mathematical Sciences*, *10*, 28 –40.

Sharma, B. D. & Taneja, I. J. (1975). Entropy of type (α, β) and other generalized measures in information theory. *Metrika*, *22*(1), 205–215.

Stam, A. J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, *2*(2), 101–112.

Steele, J. M. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge: Cambridge University Press.

Stix, G. (1991). Profile: Davis a. Huffman. *Scientific American*, *265*(3), 54–58.

Tribus, M. & McIrvine, E. C. (1971). Energy and information. *Scientific American*, *225*(3), 179–188.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, *52*(1-2), 479–487.

Tverberg, H. (1958). A new derivation of the information function. *Mathematica Scandinavica*, *6*, 297–298.

Vajda, I. (1968). Axioms for a-entropy of a generalized probability scheme. *Kybernetika*, *4*(2), 105–112.

van Brunt, B. (2004). *The Calculus of Variations*. New-York: Springer Verlag.

van den Bos, A. (2007). *Parameter Estimation for Scientists and Engineers*. Hoboken, New Jersey: John Wiley & Sons.

Varma, R. S. (1966). Generalization of Rényi's entropy of order α. *Journal of Mathematical Sciences*, *1*, 34–48.

Verdu, S. (1998). Fifty years of Shannon theory. *IEEE Transactions on Information Theory*, *44*(6), 2057–2078.

Verdú, S. & Guo, D. (2006). A simple proof of the entropy-power inequality. *IEEE Transactions on Information Theory*, *52*(5), 2165–2166.

Wang, L. & Madiman, M. (2004). Beyond the entropy power inequality via rearrangements. *IEEE Transactions on Information Theory*, *60*(9), 5116–5137.

Wiener, N. (1948). *Cybernetics: or Control and Communication in the Animal end the Machine* (2nd ed.). Cambridge, MA: MIT Press.

Los autores

Lamberti, Pedro Walter

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea).

Portesi, Mariela

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea).

Zozor, Steeve

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea).