

# GEOMETRÍA E INFORMACIÓN

## OPTATIVO

Mariela Adelina Portesi  
Pedro Walter Lamberti  
Steeve Zozor

Facultad de Ciencias Exactas



UNIVERSIDAD  
NACIONAL  
DE LA PLATA





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 epígrafe 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*



# ADVERTENCIA

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 A. 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
- 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 Entropía condicional, información mutua, entropía relativa
- 2-4 Unas identidades y desigualdades
- 2-5 Unos ejemplos y aplicaciones
- 2-6 Entropías y divergencias generalizadas
- 2-7 Entropías cuánticas 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

# Referencias

# 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, 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  $\Omega$  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  $\Omega$  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  $\Omega$ ).

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

#### *Axiomas de Kolmogorov*

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

b)  $p(\Omega) = 1$

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

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:  $0 \leq p(A) \leq 1 \quad \forall A$ ;
- condición de normalización: si  $\Omega = A_1 + \dots + A_N$ , 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:  $0 \leq p(A, B) = p(B, A) \leq \min\{p(A), p(B)\}$ ;



- si  $A$  y  $B$  son mutuamente excluyentes, entonces  $p(A, B) = 0$ ;
- si  $B_1, \dots, B_M$  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

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

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  $A_1, \dots, A_N$  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  $B_1, B_2, \dots$  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  $A_1, \dots, A_N$  sean estadísticamente independientes es que la probabilidad conjunta se factorice como

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

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

(Mandel & Wolf, 1995)

## 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  $P_X(A) = \Pr(X = x \in A)$ . Para indicar que la variable  $X$  sigue la ley de distribución  $p$  se escribe  $X \sim p$ . 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:  $x \in \Omega = \{x_1, x_2, \dots\}$ . A cada uno de los valores  $x_n$  ( $n = 1, 2, \dots$ ) 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:

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

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  $\{0, 1, 2, \dots, N\}$  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  $p = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix}^t$ , 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:  $p_n = \frac{1}{N} \quad \forall n = 1, \dots, N$ , 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  $p = \begin{pmatrix} \frac{1}{N} & \frac{1}{N} & \dots & \frac{1}{N} \end{pmatrix}^t$ . 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  $p_1^\downarrow \geq p_2^\downarrow \geq \dots \geq p_N^\downarrow$ . En el ejemplo del caso con certeza se tiene  $p^\downarrow = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}^t$ , 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  $n = 1, \dots, N - 1$ , 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,  $(0,50 \ 0,40 \ 0,10)^t$  y  $(0,70 \ 0,15 \ 0,15)^t$  no se comparan por mayorización.

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

$$\left(\frac{1}{N} \ \frac{1}{N} \ \cdots \ \frac{1}{N}\right)^t \prec p \prec \left(1 \ 0 \ \cdots \ 0\right)^t.$$

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 acumulativa* (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  $\Omega$  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  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  y  $\lim_{x \rightarrow \infty} F_X(x) = 1$ , 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  $x_1, x_2, \dots$  con probabilidades  $p_1, p_2, \dots$  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  $\mathbf{X} = \{X^1, \dots, X^d\}$ , o simplemente  $X = \begin{pmatrix} X^1 & \dots & X^d \end{pmatrix}^t$ , caracterizado por  $d$ -uplas de variables aleatorias reales, con función densidad de probabilidad conjunta  $p(x^1, \dots, x^d)$ . La ley del vector  $\mathbf{X}$  es una medida de probabilidad sobre  $\mathbb{R}^d$ , con

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

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

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

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:

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

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

Las  $d$  variables aleatorias  $X^1, \dots, X^d$  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:

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

## 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  $\Psi$  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

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

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

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

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

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

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

$$q(y^1, \dots, y^d) = |J_\Phi| p(x^1, \dots, x^d)$$

donde  $J_\Phi = \frac{\partial(\Phi^1, \dots, \Phi^d)}{\partial(y^1, \dots, y^d)}$  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

Una variable aleatoria continua  $X$  tiene asociado un *promedio* o *media* (también llamado *valor esperado* o *de expectación*) que se obtiene pesando cada valor de  $x$  con la probabilidad asociada a ese valor,  $p(x) dx$ , e integrando sobre el rango permitido de  $x$ :

$$E[X] = \langle x \rangle = \int_{\Omega} x p(x) dx \equiv \mu$$

si la integral existe. La *esperanza* de la variable aleatoria  $X$  representa el valor medio que puede tomar entre todos los eventos de una prueba. Una variable aleatoria  $X$  se dice *integrable* cuando  $E[|X|] < \infty$ .

En general, si  $X$  es una variable aleatoria, cualquier función  $f(X)$  también lo es, y su valor de expectación, si existe, está dado por

$$E[f(X)] = \langle f(x) \rangle = \int_{\Omega} f(x) p(x) dx$$

En particular, para el monomio  $f(x) = x^r$  siendo  $r \in \mathbb{N}$ , se obtiene el *r-ésimo momento (ordinario)* de  $X$ :

$$\nu_r \equiv E[X^r] = \langle x^r \rangle = \int_{\Omega} x^r p(x) dx$$

que tiene unidades de  $X^r$ . Se puede incluir el caso  $r = 0$ , que corresponde a la condición de normalización:  $\nu_0 = \int_{\Omega} p(x) dx = 1$ . La media es el primer momento:  $\nu_1 = \langle x \rangle = \mu$ . Es fácil probar que  $\langle x^2 \rangle \geq \langle x \rangle^2$ . Típicamente, los primeros momentos son más relevantes que los de órdenes mayores, para la caracterización de una distribución.

Por ejemplo, para la distribución uniforme  $p(x) = \frac{1}{b-a}$  en el intervalo  $[a, b]$ , resulta:  $\nu_1 = \langle x \rangle = \frac{1}{2}(b+a)$ ,  $\nu_2 = \langle x^2 \rangle = \frac{1}{3}(b^2 + ab + a^2)$ ,  $\dots$ ,  $\nu_r = \frac{b^{r+1} - a^{r+1}}{(r+1)(b-a)}$ .

Cuando una pdf  $p(x)$  tiene soporte (semi)infinito, necesariamente la función  $p$  debe tender a 0 cuando  $|x| \rightarrow \infty$ . Si  $p(x)$  es *de largo alcance*, en el sentido de que no cae a 0 suficientemente rápido con  $x$  para  $x$  grandes, algunos momentos pueden no existir. Por ejemplo, la distribución de probabilidad de Cauchy–Lorentz (o función de Breit–Wigner), dada por  $p(x) = \frac{\gamma}{\pi} \frac{1}{\gamma^2 + (x-x_0)^2}$  para  $x \in (-\infty, \infty)$ , con  $\gamma > 0$  y  $x_0$  fijos, no tiene momentos finitos de orden  $r \geq 1$ .

En el caso de una variable aleatoria discreta  $X$  que toma valores en  $\Omega = \{x_1, \dots, x_N\}$ , la esperanza de la variable viene dada por  $E[X] = \sum_{n=1}^N x_n p(x_n)$ . Consideraremos que el espacio muestral es  $\mathbb{N}$ , luego resulta

$$E[X] = \langle n \rangle = \sum_{n \geq 1} n p_n,$$

que se puede obtener también como  $E[X] = \sum_{j=0}^{\infty} \Pr(X > j)$ . Para una función  $f$  definida sobre el conjunto  $\{0, 1, 2, \dots\}$  se tiene

$$E[f(X)] = \langle f(n) \rangle = \sum_{n \geq 0} f(n) p_n,$$

y se define el  $r$ -ésimo momento (ordinario) de  $n$  como

$$\nu_r \equiv E[X^r] = \langle n^r \rangle = \sum_{n=1}^{\infty} n^r p_n.$$

En el caso de variables discretas sobre  $\mathbb{N}$ , resulta útil introducir el  $r$ -ésimo *momento factorial* de  $n$  mediante

$$\langle n^{(r)} \rangle \equiv \langle n(n-1) \cdots [n - (r-1)] \rangle = \sum_{n=r}^{\infty} n(n-1) \cdots (n-r+1) p_n.$$

Los *momentos centrales* se definen alrededor de  $x = \langle x \rangle$ , como el valor de expectación de potencias de la *desviación*  $\Delta x \equiv x - \langle x \rangle$ :

$$\mu_r \equiv \langle (x - \langle x \rangle)^r \rangle = \int_{\Omega} (x - \langle x \rangle)^r p(x) dx.$$

Se deduce que si la densidad de probabilidad  $p(x)$  es una función simétrica respecto a la media, entonces todos los momentos centrales impares son nulos. Los momentos centrales brindan medidas que caracterizan la distribución:

1. el primer momento central es idénticamente nulo para toda pdf:

$$\mu_1 = \langle x - \langle x \rangle \rangle = 0;$$

2. el segundo momento central se conoce como *varianza*, *dispersión* o también *desviación cuadrática media*:

$$\mu_2 = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 = \text{Var}(X) \equiv \sigma^2, \quad (1)$$

y es una medida del cuadrado del ancho efectivo de una pdf, es no negativo y se anula sólo cuando  $p(x) = \delta(x)$ , esto es, cuando no hay incerteza sobre el resultado. La varianza está bien definida si



$X$  es una variable aleatoria de cuadrado integrable, esto es, cuando  $E[X^2] < \infty$ . El *ancho* de una distribución está dado por la *desviación estándar*  $\sigma = \sqrt{\mu_2}$ , tiene las mismas unidades de  $X$ , y se usa para normalizar los momentos centrales de orden superior. El *ancho relativo* es otra medida que caracteriza la distribución, dado por  $\frac{\sigma}{\langle x \rangle} = \sqrt{\frac{\langle x^2 \rangle}{\langle x \rangle^2} - 1}$  cuando  $\langle x \rangle \neq 0$ ;

3. el tercer momento central permite definir el *coeficiente de asimetría*:

$$\alpha_3 \equiv \frac{\mu_3}{\sigma^3},$$

que resulta adimensional y puede tener signo positivo o negativo, anulándose para distribuciones que son simétricas respecto del valor medio;

4. el cuarto momento central da lugar a la *curtosis*:

$$\alpha_4 \equiv \frac{\mu_4}{\sigma^4},$$

que posibilita diferenciar entre distribuciones altas y angostas (con  $\alpha_4 < 3$ ), de otras bajas y anchas (con  $\alpha_4 > 3$ )

La relación entre los momentos centrales y los momentos ordinarios se obtiene directamente de las definiciones:

$$\mu_r = \int (x - \langle x \rangle)^r p(x) dx = \sum_{s=0}^r \binom{r}{s} (-\langle x \rangle)^{r-s} \int x^s p(x) dx = \sum_{s=0}^r \binom{r}{s} \nu_s (-\nu_1)^{r-s}$$

para cualquier  $r = 1, 2, \dots$ , siendo  $\nu_0 = 1$ . Por ejemplo,  $\mu_2 = \nu_2 - \nu_1^2$  como en la Ec. (1), mientras que  $\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$ .

Dada una variable aleatoria  $X$  con una distribución de probabilidad  $p(x)$ , teniendo en cuenta que los dos primeros momentos dan las características más importantes de la pdf, puede resultar conveniente hacer una transformación de variable aleatoria a la llamada *forma estándar*:  $Y \equiv \frac{X - \langle X \rangle}{\sigma}$ , que entonces tiene media igual a 0 y desviación estándar igual a 1.

Mencionamos algunas propiedades de  $E[X]$  y de  $E[X^2]$ .

Proposición: Sean  $X$  e  $Y$  dos variables aleatorias integrables, y sean  $a, b \in \mathbb{R}$  arbitrarios. Entonces la variable aleatoria  $Z = aX + bY$  es integrable, siendo  $E[Z] = aE[X] + bE[Y]$ .

Proposición: Sean  $X$  e  $Y$  dos variables aleatorias integrables. Si  $X$  e  $Y$  son independientes, entonces  $E[XY] = E[X]E[Y]$ .

Teorema: Sean  $X$  e  $Y$  dos variables aleatorias reales. Las variables  $X$  e  $Y$  son independientes si y sólo si  $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$  para todo par de funciones  $f$  y  $g$  en  $\mathbb{R}$ , continuas y acotadas.

Proposición: Sea  $X$  una variable aleatoria de cuadrado integrable, y sea  $\text{Var}(X) = E[(X - \langle X \rangle)^2] \equiv \sigma^2$  su varianza. Luego:

$$1. \text{Var}(X) = E[X^2] - (E[X])^2$$

$$2. \forall a \in \mathbb{R} : \text{Var}(X + a) = \text{Var}(X), \text{Var}(aX) = a^2 \text{Var}(X)$$

$$3. \text{ Si } Y \text{ es otra variable aleatoria de cuadrado integrable, e independiente de } X, \text{ entonces: } \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

### *Desigualdades de Chebyshev y de Bienaymé–Chebyshev*

Estas desigualdades dan una cota superior a la probabilidad de que una cantidad que fluctúa aleatoriamente exceda cierto valor umbral, aún sin conocer detalladamente la forma de la distribución de probabilidad.

#### Desigualdad de Chebyshev:

Sea  $X$  una variable aleatoria real con función densidad de probabilidad  $p(x)$ . Sea  $g(x) \geq 0 \forall x \in \mathbb{R}$ , con  $g(x) \geq K \forall x \in D \subset \mathbb{R}$ , para algún  $K > 0$ . Entonces por un lado

$$\Pr[g(X) \geq K] = \Pr[X \in D] = \int_D p(x) dx$$

y por otro lado

$$\langle g(X) \rangle = \int_{\mathbb{R}} g(x)p(x) dx \geq \int_D g(x)p(x) dx \geq K \int_D p(x) dx,$$

luego se tiene la desigualdad:

$$\Pr[g(X) \geq K] \leq \frac{\langle g(X) \rangle}{K}. \quad (2)$$

#### Desigualdad de Bienaymé–Chebyshev:

Sea  $X$  una variable aleatoria real de esperanza  $\mu$  y varianza  $\sigma^2$  finita. Entonces,  $\forall \epsilon > 0$  se tiene la desigualdad:

$$\Pr[|X - \mu| > \epsilon] \leq \frac{\sigma^2}{\epsilon^2}.$$

En forma equivalente, se puede plantear la probabilidad de que  $X$  se aparte de su valor medio en más de cierto número  $\eta$  de desviaciones estándar: tomando  $g(x) = (\Delta x)^2 = (x - \mu)^2$  en la Ec. (2), resulta la desigualdad :

$$\Pr[|\Delta X| \geq \eta\sigma] = \Pr[(\Delta X)^2 \geq \eta^2\sigma^2] \leq \frac{1}{\eta^2} \quad (3)$$

Estas relaciones afirman que cuanto más chica es la varianza, más se concentra la variable en torno a su media. Ambas cotas son en general débiles; por ejemplo, la desigualdad (3) indica que la probabilidad de encontrar una fluctuación superior a  $\eta = 3$  desviaciones estándar alrededor de la media, está por debajo de  $1/9$ ; el cálculo para una distribución típica como la Gaussiana ajusta dicha probabilidad por debajo de 0.003.

### *Momentos para varias variables aleatorias*

En el caso de varias variables aleatorias  $X, Y, Z, \dots$  con pdf conjunta  $p(x, y, z, \dots)$  se define el *momento central de orden*  $r, s, t, \dots$  como (Mandel & Wolf, 1995; Cover & Thomas, 2006)

$$\mu_{r,s,t,\dots} \equiv \langle (\Delta x)^r (\Delta y)^s (\Delta z)^t \dots \rangle = \int (x - \langle x \rangle)^r (y - \langle y \rangle)^s (z - \langle z \rangle)^t \dots p(x, y, z, \dots) dx dy dz \dots$$

Por ejemplo, para  $\begin{pmatrix} X \\ Y \end{pmatrix} \sim p(x, y)$  los momentos centrales de orden lineal resultan:  $\mu_{1,0} = \mu_{0,1} = 0$ , y los momentos centrales de orden cuadrático están dados por las varianzas de cada variable y por la llamada covarianza:  $\mu_{2,0} = \sigma_X^2$ ,  $\mu_{0,2} = \sigma_Y^2$ , y  $\mu_{1,1} = \langle \Delta X \Delta Y \rangle$ . Estos últimos se pueden acomodar en una matriz, con propiedades interesantes como veremos a continuación.

Sea  $X^1, \dots, X^d$  un conjunto de  $d$  variables aleatorias. La *covarianza* entre  $X^i$  y  $X^j$  se define como

$$\mu^{ij} \equiv \langle \Delta x^i \Delta x^j \rangle = \mu^{ji}$$

para  $i, j = 1, \dots, d$ . Las  $d(d+1)/2$  cantidades de este tipo se disponen en un arreglo (simétrico) de  $d \times d$ , la *matriz de covarianza*  $\Sigma$ , cuya diagonal son las varianzas  $(\sigma^i)^2$ . Por ejemplo, si  $d = 2$  se tiene

$$\begin{pmatrix} X^1 \\ X^2 \end{pmatrix} \sim p(x^1, x^2) : \quad \Sigma = \begin{pmatrix} (\sigma^1)^2 & \mu^{12} \\ \mu^{21} & (\sigma^2)^2 \end{pmatrix}.$$

Proposición:

$$|\mu^{ij}|^2 \leq \mu^{ii} \mu^{jj}$$

La demostración de esta proposición involucra la desigualdad de Cauchy–Schwarz .....

Se define el *coeficiente de correlación* que es adimensional y toma valores entre  $-1$  (variables completamente anticorrelacionadas) y  $1$  (variables completamente correlacionadas) como:  $\rho^{ij} = \rho^{ji} \equiv \frac{\mu^{ij}}{\sigma^i \sigma^j}$ .

Como ejemplo, dadas  $X^1$  y  $X^2 = aX^1 + b$  que fluctúan en fase ( $a > 0$ ) o al revés ( $a < 0$ ), se tiene  $\Delta x^2 = a\Delta x^1$ , luego  $\rho^{12} = \frac{a}{|a|} = \pm 1$ .

....

### 1.3.2 Funciones generatrices

Se definen un conjunto de funciones que permiten hallar fácilmente los distintos momentos de una distribución de probabilidad. Se llaman *funciones generadoras* o *funciones generatrices*, y están dadas como valores de expectación de funciones de la variable aleatoria (discreta o continua), con un parámetro real o complejo.

La *función generadora de momentos* (MGF, *moment generating function*) se define como

$$M(\xi) \equiv \langle e^{\xi X} \rangle = \int e^{\xi x} p(x) dx, \quad \xi \in \mathbb{R}$$

en el caso de una variable aleatoria continua  $X$  con pdf  $p(x)$ . Se tiene  $M(0) = \int p(x) dx = 1$  (que corresponde a la condición de normalización). Si la variable  $X$  es positiva y se toma  $\xi = -s$  con  $s > 0$ , se interpreta en términos de la transformada de Laplace de la función  $p$ .

Si existe, la MGF posibilita obtener fácilmente los momentos (ordinarios) de  $X$  a distintos órdenes, mediante los coeficientes del desarrollo de  $M$  en serie de potencias de  $\xi$ :

$$M(\xi) = \sum_{r=0}^{\infty} \frac{\xi^r}{r!} \int x^r p(x) dx = 1 + \sum_{r=1}^{\infty} \frac{\nu_r}{r!} \xi^r$$

o, alternativamente, mediante las sucesivas derivadas de  $M$  respecto de  $\xi$  en 0:

$$\nu_r = \left. \frac{d^r M(\xi)}{d\xi^r} \right|_{\xi=0}, \quad r = 1, 2, \dots; \quad \nu_0 \equiv 1.$$

En el caso de una variable aleatoria discreta, suponiendo que el espacio muestral es  $\mathbb{N}$ , se definen dos funciones: la *función generadora de momentos (ordinarios)* (MGF) dada por

$$M(\xi) \equiv \langle e^{\xi N} \rangle = \sum_{n \geq 0} e^{\xi n} p_n,$$

y la *función generadora de momentos factoriales* (FMGF, *factorial moment generating function*) como

$$F(\xi) \equiv \langle (1 + \xi)^N \rangle = \sum_{n \geq 0} (1 + \xi)^n p_n$$

para  $\xi \in \mathbb{R}$  en ambos casos. Se verifica  $M(0) = F(0) = \sum_{n=0}^{\infty} p_n = 1$ . Se muestra simplemente que

$$M(\xi) = \sum_{r=0}^{\infty} \frac{\langle n^r \rangle}{r!} \xi^r,$$

lo que permite obtener los momentos de la distribución para cualquier orden  $r \geq 1$ . Por otro lado, el desarrollo de la FMGF da

$$F(\xi) = \sum_{n=0}^{\infty} \sum_{r=0}^n \binom{n}{r} \xi^r p_n = \sum_{r=0}^{\infty} \sum_{n=r}^{\infty} \frac{n(n-1) \cdots (n-r+1)}{r!} \xi^r p_n = \sum_{r=0}^{\infty} \frac{\langle n^{(r)} \rangle}{r!} \xi^r$$

teniendo en cuenta en las dobles sumas que  $0 \leq r \leq n$ , con  $n$  hasta  $n_{\text{máx}}$  ó  $\infty$ . Se ve entonces que  $F$  permite obtener los momentos factoriales de orden  $r$  arbitrario.

Dada una variable aleatoria a valores naturales, la función  $G(\xi) = \sum_{n=0}^{\infty} p_n \xi^n$ , con  $-1 \leq \xi \leq 1$ , es también una función generatriz. Por ejemplo, si  $G$  admite derivadas primera y segunda en  $\xi = 1$  se obtienen:  $\langle N \rangle = G'(1)$ ,  $\langle N(N-1) \rangle = G''(1)$ ,  $\text{Var}(N) = G''(1) + G'(1) - [G'(1)]^2$ ; además, se obtiene la ley de distribución evaluando derivadas de  $G$  en  $\xi = 0$ :  $p_n = \frac{G^{(n)}(0)}{n!}$ . (François, 2009)

La *función característica* (CF, *characteristic function*) tiene argumento complejo: (Lukacs, 1961)

$$C_X(\xi) \equiv \langle e^{i\xi X} \rangle = \int e^{i\xi x} p(x) dx.$$

La importancia de esta función reside en que siempre existe y está bien definida, dado que es la transformada de Fourier de una función absolutamente integrable (i.e.  $\int |f(x)| dx < \infty$ ) (Golberg, 1961)

Si la pdf  $p(x)$  es de cuadrado integrable, entonces

$$p(x) = \frac{1}{2\pi i} \int e^{-i\xi x} C_X(\xi) d\xi.$$

El requisito para esta importante relación es que  $\int_{-\infty}^{\infty} |p(x)|^2 dx < \infty$ ; sin embargo, aún es válida para distribuciones con una contribución tipo  $\delta$ . Por otro lado los momentos, si existen, se obtienen derivando la función  $C$  tal como expresa la siguiente proposición:

**Proposición:** La variable aleatoria  $X$  admite momento de orden  $r$  si y sólo si la función característica  $C$  es  $r$  veces derivable en  $\xi = 0$ , siendo

$$\langle X^r \rangle = (-i)^r C_X^{(r)}(0).$$

Por ejemplo, en el caso de la distribución de Cauchy–Lorentz resulta

$$C(\xi) = \frac{\gamma}{\pi} \int_{-\infty}^{\infty} \frac{e^{i\xi x}}{\gamma^2 + (x - x_0)^2} dx = e^{-\gamma|\xi|e^{ix_0\xi}}$$

tomando  $\gamma > 0$ . Esta función está definida para todo  $\xi$ , pero no es derivable en  $\xi = 0$ , lo que coincide con el hecho de que no están definidos los momentos para esta pdf.

Para una variable aleatoria compleja  $Z = X + iY$ , usando la noción de transformada de Fourier bidimensional, se define:

$$C_Z(\mu) \equiv \int e^{\mu^* z - \mu z^*} p(z) d^2 z.$$

Resumimos algunas propiedades importantes de la función característica:

1.  $C(0) = 1$
2.  $|C(\xi)| \leq C(0)$
3.  $C(\xi)$  es una función continua en  $\mathbb{R}$  (aún si la pdf  $p(x)$  tiene discontinuidades)
4.  $C(-\xi) = C(\xi)^*$
5.  $C(\xi)$  es definida no negativa, de tal forma que para un conjunto arbitrario de  $N$  números reales  $\xi_1, \dots, \xi_N$  y  $N$  números complejos  $a_1, \dots, a_N$ , se cumple

$$\sum_{i,j=1}^N a_i^* a_j C(\xi_j - \xi_i) \geq 0.$$

6.  $C(\xi) = M(i\xi) = F(e^{i\xi} - 1)$ , si  $M$  y  $F$  existen;  $F(\xi) = M(\ln(1 + \xi))$

**Teorema 1-1.** (Bochner, Goldberg)....

**Proposición:** Sean  $X$  e  $Y$  dos variables aleatorias reales independientes, cuyas funciones características son  $C_X$  y  $C_Y$ . Entonces  $C_{X+Y} = C_X C_Y$ .

Cumulant generating function ....

Extendemos la definición de función característica para un vector aleatorio. ...

....

## 1.4 Algunos ejemplos de distribuciones de probabilidad

introducción...

### 1.4.1 Distribuciones de variable discreta

Variable con certeza

...

Ley de Bernoulli

...

Ley geométrica

...

Distribución 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 realmente 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 desarrollo de la comunicación moderna, por ejemplo a través del telégrafo siguiendo el patente de Moorse en 1840. La idea de asignar un código (punto o barra, más espacio entre letras y entre palabras) a las letras del alfabeto es la semilla de la codificación entrópica, la que se basa precisamente sobre la asignación de un código a símbolos de una fuente (codificación de fuente) según las frecuencias (o probabilidad de aparición) de cada símbolo en una cadena. De hecho, el principio de codificar un mensaje y mandar la versión codificada por un canal de transmisión es mucho más antiguo, a pesar de que no había ninguna formalización matemática ni siquiera explícitamente una noción de información. Entre otros, se puede mencionar el telégrafo óptico de Claude Chappe (1794), experimentos con luces por Guillaume Amontons (en los años 1690 en París), o aún más 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 práctica del esquema de comunicación de Shannon (Shannon, 1948; Shannon & Weaver, 1964), es decir la codificación de la información, potencialmente de manera la más económica que se puede, su transmisión a un "receptor" (por un canal ruidoso) que la interpreta/lee/decodifica. Implícitamente, la noción de información es al menos tan antigua que la humanidad.

A pesar de que la idea de codificar y transmitir “información” sea tremendamente antigua, la formalización matemática 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 después en su libro re-titulado “The mathematical theory of communication” reemplazando el “A” por un “The”. Desde estos años, las herramientas de la dicha teoría de la información dio lugar a muchas aplicaciones especialmente en comunicación (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 definición de la entropía, 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 símbolos, lo que condujo a la definición de la entropía. Esta definición puede ser deducida también de un conjunto de propiedades “razonables” que debería cumplir una medida de incerteza (enfoque axiomático). Se continuará con la descripción de tal noción de entropía, pasando del mundo discreto (símbolos, alfabeto) al mundo continuo, lo que no es trivial ni siquiera intuitivo. Se adelantará 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 símbolos es debido a Ralph Hartley (Hartley, 1928). En su papel, Hartley definió la información de una secuencia como siendo proporcional a su longitud. Más precisamente, para símbolos de un alfabeto de cardinal  $\alpha$ , existen  $\alpha^n$  cadenas distintas de longitud  $n$ ; Se definió la información de tales cadenas como siendo  $Kn$  ( $K$  dependiente de  $\alpha$ ). Para ser consistente, dos conjuntos de mismo tamaño  $\alpha_1^{n_1} = \alpha_2^{n_2}$  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 más que los números de bits (0-1) necesarios para codificar todas las cadenas de longitud  $n$  de símbolos de un alfabeto de cardinal  $\alpha$ . La información de Hartley es el equivalente de la entropía de Boltzman de la termostática, la famosa formula  $S = k_B \log W$  (Boltzmann, 1896, 1898; Jaynes, 1965).

Una debilidad del enfoque de Hartley es que considera implícitamente que en un mensaje, cada cadena de longitud dada puede aparecer con la misma frecuencia, o probabilidad  $1/\alpha^n$  (en Boltzman, misma probabilidad de cada configuración), siendo la información menos el logaritmo de estas probabilidades. A contrario,



parece más lógico considerar que secuencias muy frecuentes no llevan mucha información (se sabe que aparecen), mientras que las que aparecen raramente llevan más información (hay más sorpresa, más incerteza en observarlas). Volviendo a los símbolos 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 símbolos  $x$ . 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 aparición de  $x$ . Se define la incerteza asociada a la variable aleatoria como el promedio estadístico sobre todos los estados posibles  $x$  (Shannon, 1948; Shannon & Weaver, 1964) <sup>1</sup>.

**Definición 2-1** (Entropía de Shannon). *Sea  $X$  una variable aleatoria definida sobre una alfabeto discreto  $\mathcal{X} = \{x_1, \dots, x_\alpha\}$  de cardinal  $\alpha = |\mathcal{X}| < +\infty$  finito. Sea  $p_X$  la distribución de probabilidad de  $X$ , i. e.,  $\forall x \in \mathcal{X}$ ,  $p_X(x) = \Pr[X = x]$ . La entropía 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  $0 \log 0 = 0$  ( $\lim_{t \rightarrow 0} t \log t = 0$ ).

La base del logaritmo es arbitrario; si es  $\log_2$  el logaritmo de base 2,  $H$  es en binary unidades o bits (se encuentra también la denominación Shannons), si se usa el logaritmo natural  $\ln$ ,  $H$  es en natural unidades o nats, si es el de base 10,  $H$  es en dígitos decimales o dits (se encuentra también las denominaciones bans o Hattleys). En este capítulo, se usará  $H$  con el logaritmo correspondiente sin especificar la base. Si es necesario que tenga una base  $a$  dada, se notará la entropía correspondiente  $H_a$  y se especificará 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, rigurosamente,  $H$  sea una función de la distribución de probabilidad  $p_X$  y no de la variable  $X$ , se usará indistamente la notación  $H(p_X)$  tal como  $H(X)$  según lo más conveniente. Además,  $p_X$  podrá denotar indistamente la distribución de probabilidad, o el vector de probabilidad  $p_X \equiv [p_X(x_1) \ \cdots \ p_X(x_\alpha)]^t$  donde  $\cdot^t$  denota la transpuesta.

$H$  es el equivalente de la entropía de Gibbs en termostática. La letra  $H$  viene del teorema-H debido a... Ludwig Boltzman (Jaynes, 1965).

$H$  tiene propiedades notables que corresponden 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):

---

<sup>1</sup>En la misma época que Shannon, independientemente, la noción de información o medidas equivalentes apareciendo por ejemplo en calculo de capacidad de canal surió en varios trabajos como los de los ingenieros franceses André Clavier (Clavier, 1948) o Jacques Laplume (Laplume, 1948), o en el libro del estadounidense Norbert Wiener (Wiener, 1948, Cap. III) entre varios otros (ver (Verdu, 1998; Lundheim, 2002; Rioul & Magossi, 2014; Flandrin & Rioul, 2016; Rioul & Flandrin, 2017; Chenciner, 2017)).

[P1] *Continuidad*: Vista como una función de  $\alpha$  variables  $t_i = p_X(x_i)$ ,  $H$  es continua con respecto a los  $t_i$ .

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

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

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

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

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

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

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 símbolos escritos sobre las caras, sean enteras o cualesquiera letras.

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

$$H(X) \geq 0$$

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

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

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  $-p_X(x) \log p_X(x) \geq 0$ . Además, la suma de terminos positivos vale cero si y solamente si cada termino de la suma vale cero, dando  $p_X(x) = 0$  o  $p_X(x) = 1$ . Se concluye  $p_X$  siendo una distribución de probabilidad, sumando a 1.

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

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

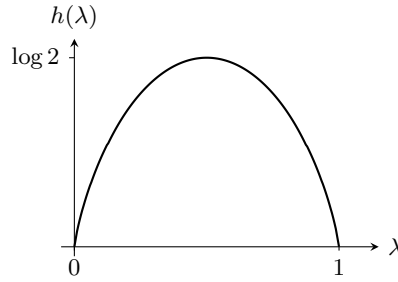
con igualdad si y solamente si existe  $X$  es uniforme sobre  $\mathcal{X}$ , i. e.,

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

En otras palabras, la incerteza es máxima 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 maximización de  $H$  sujeto a  $\sum_x p_X(x) = 1$ , es decir, con la tecnica del

Lagrangiano, notando  $t_i = p_X(x_i)$ , de la minimización de  $\sum_i (-t_i \log t_i + \lambda t_i)$ . Se obtiene sencillamente que  $\log t_i = -\lambda$ , dando la distribución uniforme.

La figura Fig. 2-4 representa la entropía de un sistema a dos estados, de probabilidades  $\lambda$  y  $1 - \lambda$  (ley de Bernoulli de parametro  $\lambda$ ), entropía a veces dicha *entropía binaria*, en función de  $\lambda$ . Esta figura ilustra ambas cotas ( $\lambda = 1$  o  $0$ ,  $\lambda = \frac{1}{2}$ ) así que la invariancia bajo una permutación ( $h(\lambda) = H(\lambda, 1 - \lambda) = H(1 - \lambda, \lambda) = h(1 - \lambda)$ ).



**Figura 2-4:** Entropía 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 entropía, i. e., sean  $X$  definido sobre  $\mathcal{X}$  y  $\tilde{X}$  sobre  $\tilde{\mathcal{X}}$ ,

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

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

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

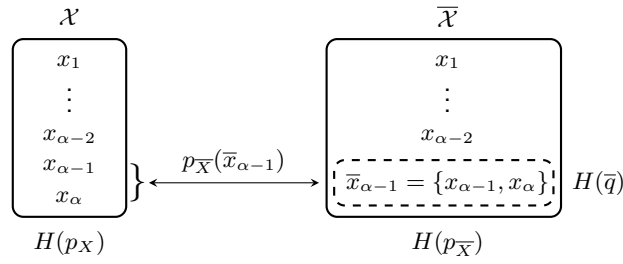
$$\left\{ \begin{array}{l} \bar{\mathcal{X}} = \{x_1, \dots, x_{\alpha-2}, \bar{x}_{\alpha-1}\} \quad \text{con el estado interno} \quad \bar{x}_{\alpha-1} = \{x_{\alpha-1}, x_{\alpha}\}, \\ p_{\bar{X}}(x_i) = p_X(x_i), \quad 1 \leq i \leq \alpha-1 \quad \text{y} \quad p_{\bar{X}}(\bar{x}_{\alpha-1}) = p_X(x_{\alpha-1}) + p_X(x_{\alpha}) \quad \text{distribución sobre } \bar{\mathcal{X}} \\ \bar{q}(x_j) = \frac{p_X(x_j)}{p_X(x_{\alpha-1}) + p_X(x_{\alpha})}, \quad j = \alpha-1, \alpha \quad \text{distribución del estado interno} \end{array} \right.$$

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

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

[P8] *Concavidad:* La entropía es cóncava, en el sentido de que la entropía de una combinación convexa de distribuciones (mezcla) de probabilidades es siempre mayor o igual a la combinación convexa de entropías:

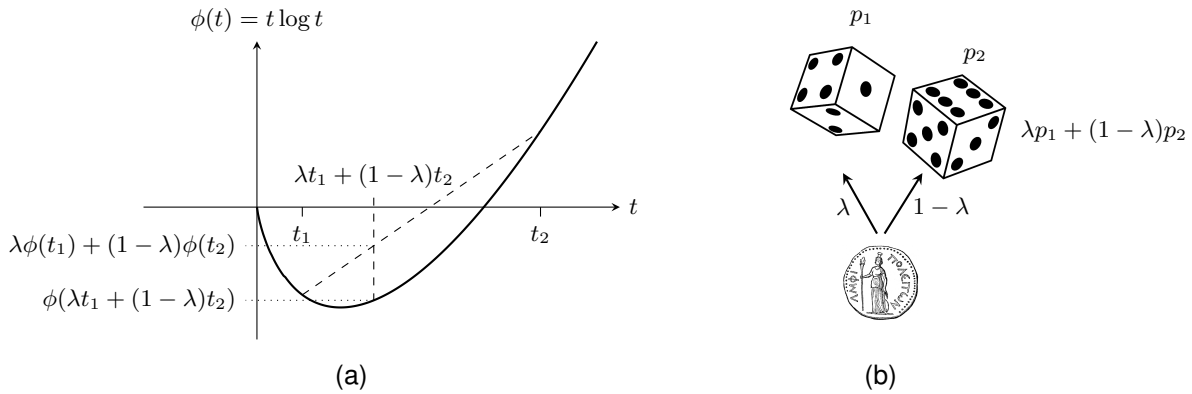
$$\forall \{\lambda_i\}_{i=1}^n, \quad 0 \leq \lambda_i \leq 1, \quad \sum_{i=1}^n \lambda_i = 1 \quad \text{and cualquier conjunto de distribuciones} \quad \{p_i\}_{i=1}^n,$$



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

$$H\left(\sum_{i=1}^n \lambda_i p_i\right) \geq \sum_{i=1}^n \lambda_i H(p_i)$$

Esta desigualdad es conocida como desigualdad de Jensen. Es una consecuencia directa de la convexidad de la función  $\phi : t \mapsto t \log t$ , como ilustrado en la figura Fig. 2-6-(a). La figura Fig. 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  $\lambda$  de elegir el dado izquierdo).



**Figura 2-6:** (a)  $\phi(t) = t \log t$  es convexa: la curva es siempre debajo de sus cuerdas; entonces, cada promedio de  $\phi(t_1)$  y  $\phi(t_2)$  estando en la cuerda juntando estos puntos, queda arriba de la función tomada en el promedio de  $t_1$  y  $t_2$ . Escribiendo eso para (más 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 querer, lo más “concentrado” es una distribución de probabilidad, lo menos hay incerteza, y entonces lo más pequeño debe ser la entropía. Esta propiedad intuitiva se resume a partir de la noción de mayorización:

**Definición 2-2** (Mayorización). Una distribución discreta finita de probabilidad  $p$  mayorizada por una distribución  $q$ , notado  $p \prec q$ , se define como:

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

(las últimas sumas siendo igual a 1). Si los alfabetos de definición de  $p$  y  $q$  son de tamaños diferentes,  $\alpha$  es el tamaño lo más grande y la distribución sobre el alfabeto lo más corto es completada por estados de probabilidad 0 (recuerdense de que no va a cambiar la entropía).

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 consecuencias de esta desigualdad: la distribución cierta mayoriza cualquier distribución y cualquier distribución mayoriza la distribución uniforme (Marshall, Olkin & Arnold, 2011, p. 9, (6)-(8)). Además, de la Schur-concavidad se obtiene que

$$H\left(\left[\frac{1}{\alpha} \quad \dots \quad \frac{1}{\alpha}\right]^t\right) \quad \text{es una función creciente de } \alpha$$

La prueba de la Schur-concavidad se apoya sobre la desigualdad de Schur o Hardy-Littlewood-Pólya o Karamata (Schur, 1923; Hardy, Littlewood & Pólya, 1929; Karamata, 1932; Hardy, Littlewood & Pólya, 1952), (Marshall et al., 2011, Cap. 3, Prop. C.1) o (Bhatia, 1997, Teorema II.3.1):  $t \prec t' \Rightarrow \sum_i \phi(t_i) \leq \sum_i \phi(t'_i)$  para cualquier función  $\phi$  convexa. Sufice considerar  $\phi(t) = t \log t$  para concluir.

En muchos casos, uno tiene que trabajar con varias variables aleatorias. Para simplificar las notaciones, consideramos un 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 variables puede ser vista como una variable  $(X, Y)$  definida sobre el alfabeto  $\mathcal{X} \times \mathcal{Y}$  de cardinal  $\alpha\beta$  tal que se defina naturalmente la entropía para esta variable; tal entropía es llamada *entropía conjunta* de  $X$  y  $Y$ :

**Definición 2-3** (Entropía conjunta). Sean  $X$  e  $Y$  dos variable aleatorias definidas sobre los alfabetos 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.,  $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ ,  $p_{X,Y}(x, y) = \Pr[X = x, Y = y]$ . La entropía 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 entropía de Shannon.

[P10] *Aditividad*: La entropía 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 “ $\Rightarrow$ ” es consecuencia directa de  $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ . Se va a probar en la sección siguiente la recíproca. Esta propiedad se escribe también

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

donde  $\otimes$  es el producto de Kronecker <sup>2</sup> Se generaliza sencillamente a un conjunto de variables aleatorias  $\{X_i\}_{i=1}^n$  (Kronecker product de un conjunto de vectores de probabilidades).

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

$$H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i) \quad i. e., \quad H(p_{X_1, \dots, X_n}) \leq H(p_{X_1} \otimes \dots \otimes p_{X_n}) = \sum_{i=1}^n H(p_{X_i})$$

Dicho de otra manera, las variables aleatorias pueden compartir información, de tal manera que la entropía global sea menor que la suma de cada entropía. De la propiedad anterior, se obtiene la igualdad si y solamente si los  $X_i$  son independientes.

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

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

Es importante notar que existen varios enfoques basados sobre una serie de axiomas, dando lugar a la definición de la entropía tal como definida. Estos axiomas son conocidos como axiomas de Shannon-Khinchin y son la continuidad [P1], la maximalidad [P5], la expansibilidad [P6] y la aditividad [P10]. Existen varios otros conjuntos de axiomas, conduciendo también a la entropía de Shannon (ver (Shannon, 1948, Sec. 6) o (Shannon & Weaver, 1964; Fadeev, 1956, 1958; Khinchin, 1957; Rényi, 1961), entre otros).

Para una serie de variables aleatorias,  $X_1, X_2, \dots$ , representando símbolos, se puede definir una entropía por símbolo como una entropía conjunta dividido por el número de símbolos,  $\frac{H(X_1, \dots, X_n)}{n}$ , así que una tasa de entropía cuando  $n$  va al infinito.

**Definición 2-4** (Tasa de entropía). Sea  $\mathcal{X} = \{X_i\}_{i \in \mathbb{N}^*}$  una serie de variables aleatorias, o proceso estocástico. La tasa de entropía del proceso es definida por

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

Esta cantidad siempre existe porque  $H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i) \leq \sum_{i=1}^n \log \alpha_i \leq n \max_{1 \leq i \leq n} \alpha_i$  donde los  $\alpha_i$  son los cardinales de los alfabetos de definición de los  $X_i$ .

---

<sup>2</sup>  $\begin{bmatrix} p_X(x_1) & \dots & p_X(x_\alpha) \end{bmatrix}^t \otimes \begin{bmatrix} p_Y(y_1) & \dots & p_Y(y_\beta) \end{bmatrix}^t = \begin{bmatrix} p_X(x_1)p_Y(y_1) & \dots & p_X(x_1)p_Y(y_\beta) & \dots & p_X(x_\alpha)p_Y(y_1) & \dots & p_X(x_\alpha)p_Y(y_\beta) \end{bmatrix}$

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 analogía, se puede siempre definir la entropía como en la definición Def. 2-1. Esta extensión resuelta delicada dando de que unas propiedades se perdien. Por ejemplo, la entropía no queda acotada por arriba como se lo puede probar para la distribución de probabilidad  $p(x) \propto \frac{1}{(x+2)(\log(x+2))^2}$ ,  $x \in \mathbb{N}$ , correctamente normalizada ( $\propto$  significa “proporcional a”):  $\frac{\log \log(x+2)}{(x+2)(\log(x+2))^2} \geq 0$  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 entropía de Shannon, usando el operador  $E$  promedio estadístico o esperanza matemática, se puede reescribir la entropía de Shannon como  $H(X) = E[-\log p_X(X)]$ . Con este punto de vista, es fácil extender la definición de la entropía para variables aleatorias continuas admitiendo una densidad de probabilidad. Eso da lugar a lo que es conocido como la *entropía diferencial*:

**Definición 2-5** (Entropía diferencial). Sea  $X$  una variable aleatoria continua definida sobre  $\mathbb{R}^d$  y  $\mathcal{X} \subseteq \mathbb{R}^d$  el soporte de  $p_X(x)$  ( $0 \notin p_X(\mathcal{X})$ ), su densidad (distribución) de probabilidad. La entropía 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  $X = (X_1, \dots, X_d)$ , esta entropía de  $X$  es dicha entropía conjunta de los componentes  $X_i$ .

Como se lo va a ver, la entropía 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. Más allá, 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 este hecho, consideramos una variable aleatoria escalar  $X$  y  $p_X$  su densidad de probabilidad de soporte  $\mathbb{R}$ . 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 figura Fig. 2-7. Se define la variable aleatoria discreta  $X^\delta$  sobre  $\mathcal{X}^\delta$  tal que  $\Pr[X^\delta = x_k] = p_{X^\delta}(x_k) = p_X(x_k)\delta$ . Se puede ver  $X^\delta$  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  $X^\delta$  cuando  $\delta$  tiende a 0. Ahora, es claro de que

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

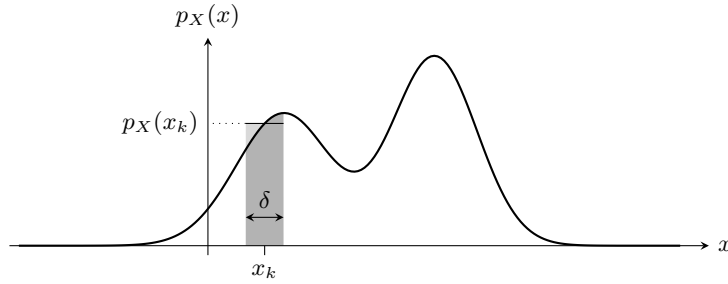
lo que se escribe tambien

$$H(X^\delta) + \log \delta = - \sum_k \left( p_X(x_k) \log p_X(x_k) \right) \delta$$

Entonces, de la intergración de Rieman sale que

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

Dicho de otra manera, la entropía diferencial de  $X$  no es el límite de la entropía de su versión cuantificada: aparece con la entropía el termino “diferencial”  $\log \delta$ .



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

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

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

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

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

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

donde el último termino,  $E[\log |J_g(X)|]$  no vale cero en general. En particular, si  $H$  es invariante bajo una translación,

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



no es invariante por cambio de escala,

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

Esta última relación queda válido para  $a$  matriz invertible. Por esta última relación, se puede ver que, dado  $X$ , cuando  $a$  tiende a 0, la entropía 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 [P4]. Por esta perdida, se quita definitivamente la interpretación de incerteza/información que hubiera podido tener la entropía diferencial.

A veces, se usa lo que es llamado potencia entrópica:

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

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

Por construcción,  $N(X) \geq 0$ . Además, en el caso continuo,  $N(aX + b) = |a|^2 N(X)$  (queda válido para una matriz  $a$  invertible): esta propiedad puede justificar la idea de “potencia”; además  $N(aX + b)$  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 determinista).

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

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

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

donde  $|\cdot|$  denota la norma euclídeana. Eso es ilustrado figura Fig. 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<sup>3</sup> rearreglada simétrico  $p_X^\downarrow$  de  $p_X$  es definida por

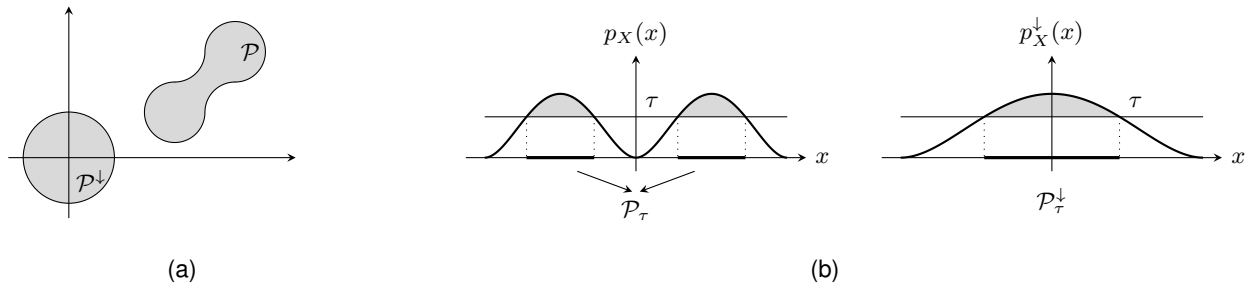
$$p_X^\downarrow(x) = \int_0^{+\infty} \mathbb{1}_{\mathcal{P}_u^\downarrow}(x) du$$

con  $\mathbb{1}_A$  el indicador 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. Más allá, sobre  $\mathcal{P}_{\tau+d\tau} \setminus \mathcal{P}_\tau$  la función  $p_X$  “vale”  $\tau$  y sobre  $\mathcal{P}_{\tau+d\tau}^\downarrow \setminus \mathcal{P}_\tau^\downarrow$  la función  $p_X^\downarrow$  “vale” también  $\tau$ , 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 una prueba más rigurosa). La representación de la definición es conocida como representación en capas de pastel (“layer cake” en ingles). Eso es ilustrado en la figura Fig. 2-8-b

---

<sup>3</sup>Se prueba de que esta función, positiva por definición, suma a 1. Además, por construcción, depende unicamente de  $|x|$  y decrece con  $|x|$ .



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

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

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

Esta propiedad es probada para funciones convexas de la densidad de probabilidad por ejemplo en (Lieb & Loss, 2001) o (Wang & Madiman, 2004, Lema 7.2) <sup>4</sup>, y entonces para el caso particular  $\phi(t) = t \log t$ .

Una pregunta natural es de saber lo que pasa en termino de mayorización en el contexto continuo  $d$ -dimensional. Por eso, se necesita primero de redefinir la noción de mayorización en este contexto:

**Definición 2-8** (Mayorización en el contexto continuo). *Una densidad de probabilidad  $p$  es dicha mayorizada por una distribución  $q$  si:*

$$p \prec q \quad \text{ssi} \quad \int_{\mathcal{B}(0,r)} p^\downarrow(x) dx \leq \int_{\mathcal{B}(0,r)} q^\downarrow(x) dx \quad \forall r > 0, \quad \text{y} \quad \int_{\mathbb{R}^d} p^\downarrow(x) dx = \int_{\mathbb{R}^d} q^\downarrow(x) dx$$

donde  $\mathcal{B}(0,r) = \{x \in \mathbb{R}^d : \|x\| \leq r\}$  es la bola centrada en 0 y de rayo  $r$  (las últimas integrales son obviamente iguales a 1).

La Schur-concavidad [P9] se conserva en el caso continuo, i. e.,

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

La desigualdad inversa es probada para cualquiera función  $\phi$  convexa de la densidad (Chong, 1974) o (Wang & Madiman, 2004, Prop. 7.3), en particular para  $\phi(t) = t \log t$ .

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

<sup>4</sup>En (Lieb & Loss, 2001, Sec. 3.3) lo muestran para  $\phi(p_X)$  donde  $\phi$  es la diferencia de dos funciones monotonas, siendo  $\phi(t) = t \log t$  un caso particular.

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

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

con igualdad si y solamente si  $X$  es uniforme.

b) Si  $\mathcal{X} = \mathbb{R}^d$  y  $X$  tiene una matriz de covarianza dada  $\Sigma_X = E[XX^t]$ , la entropía es también acotada por arriba,

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

con igualdad si y solamente si  $X$  es gaussiana. En particular, la potencia entrópica de la gaussiana vale  $N(X) = |\Sigma_X|^{\frac{1}{d}}$ , dando de nuevo un “sabor” de potencia a  $N$ . Como se lo va a ver en este capítulo, la gaussiana juega un rol central en la teoría de la información.

En ambos casos, estas desigualdades con la distribución maximizante se obtienen resolviendo el problema de maximización de la entropía sujeto a vínculos. Se trata del caso más general en la sección Sec. 2.4.1.

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 entrópica, se puede deducir la desigualdad de Hadamard, desigualdad puramente matricial:  $|R| \leq \prod_i R_{i,i}$  para cualquiera matriz simétrica definida positiva (viene de la propiedad [P11] escrita para una gaussiana de covarianza  $R$  y tomando una exponencial de la desigualdad).

## 2.3 Entropía condicional, información mutua, entropía relativa

Tratando de un par de variables aleatorias  $X$  e  $Y$ , una cuestión natural que ocurre es de cuantificar la incerteza que queda sobre una de las variables cuando se observa la otra. Dicho de otra manera, si se mide  $Y = y$ , ¿qué información lleva sobre  $X$ ? La respuesta a esta interrogación se encuentra en la noción de entropía condicional. Si uno mide  $Y = y$ , la descripción estadística de  $X$  conociendo este  $Y = y$  se resume 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 que  $Y = y$ ,

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

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

**Definición 2-9** (Entropía condicional). Sean  $X$  e  $Y$  dos variables aleatorias discretas, la entropía 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 entropía diferencial:

**Definición 2-10** (Entropía diferencial condicional). Sean  $X$  e  $Y$  dos variables aleatorias continuas, la entropía 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 independientes,  $p_{X|Y}$  se reduce a  $p_X$ , así que vale cero la entropía condicional:

[P13]

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

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 entonces ninguna medición de  $Y$  va a cambiar la incerteza sobre  $X$ .

Siendo  $H(X|Y = y)$  una entropía, va a heredar de todas las propiedades de la entropía (diferencial). Además, de  $p_{X,Y} = p_{X|Y}p_Y$  se deduce la propiedad siguiente (válida para la entropía como para su extensión diferencial)

[P14] *Regla de cadena*

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

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

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

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

Siendo  $H(X|Y = y)$  una entropía, en el caso discreto esta cantidad es positiva. Entonces, en el caso discreto,  $H(X|Y)$  es positiva, lo que prueba 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 cantidades definen lo que se llama la información mutua entre  $X$  e  $Y$ :

**Definición 2-11** (Información mutua). Sean  $X$  e  $Y$  dos variables aleatorias, la información mutua entre  $X$  e  $Y$  es la cantidad simétrica

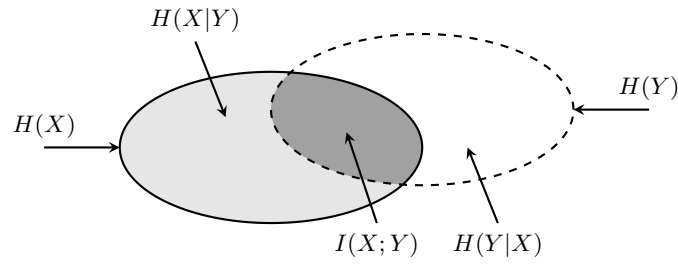
$$I(X; Y) = H(X|Y) - H(X) = H(Y|X) - H(Y) = H(X, Y) - H(X) - H(Y)$$

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

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



**Figura 2-9:** Diagrama de Venn: Ilustración de la definición de la entropía condicional, de la información mutua, y de las relaciones entre cada medida. La superficie del elipse en línea llena (parte grise) representa  $H(X)$  y el interior de la en línea discontinua representa  $H(Y)$ . La parte grise clara representa  $H(X|Y)$  superficie del “conjunto  $H(X)$ ” quitando la parte que pertenece a  $H(Y)$ . La parte blanca representa  $H(Y|X)$  superficie del “conjunto  $H(Y)$ ” quitando la parte que pertenece a  $H(X)$ . La parte en grise oscuro es entonces lo que  $X$  e  $Y$  comparten, es decir  $I(X; Y)$ .

Las diferentes cantidades pueden ser vistas a través de una visión ensemblista, como descrita en la figura Fig. 2-9. Este diagrama es conocido como diagrama de Venn.

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 más general la noción de entropía relativa, conocida también como divergencia de Kullback-Leibler (Kullback & Leibler, 1951; Kullback, 1968; Cover & Thomas, 2006; Rioul, 2007):

**Definición 2-12** (Entropía relativa). *La entropía relativa, o divergencia de una distribución de probabilidad  $q$ , con respecto a una distribución de referencia  $p$ , donde el soporte de  $p$  incluye lo de  $q$  ( $p(x) = 0 \Rightarrow q(x) = 0$ ), es definida como*

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

o, en su forma diferencial

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

(en este último caso, la condición de inclusión del soporte de  $q$  dentro del de  $p$  se formula como de que  $q$  es absolutamente continua con respecto a  $p$ )<sup>5</sup>.

Inicialmente, esta medida fue introducida por Kullback y Leibler en la misma línea que Shannon, interpretando  $\log \left( \frac{q(x)}{p(x)} \right)$  como una información de discriminación entre dos hipótesis de distribuciones  $q$  y  $p$  a partir

<sup>5</sup>Más rigurosamente, en el caso discreto, esta cantidad depende solamente de  $p$  y  $q$  y no de los estados. La condición necesaria es que  $p$  y  $q$  tienen los mismos números de componentes (se completa el vector lo más corto) y si la  $i$ -ésima componente de  $q$  vale cero, entonces la de  $p$  vale cero también. Además, con  $p$  y  $q$  de mismo tamaño, se puede poner en biyección los alfabetos asociados a  $p$  y  $q$ , sin pérdida de generalidad. En el caso continuo, esta razonamiento no vale más, esta cantidad dependiendo de los estados. . .

de la observación  $x$ , la divergencia siendo la información de discriminación promedia. Introdujeron también una versión simétrica, que veremos más adelante.

Esta medida puede ser vista también como una entropía 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  $\alpha$ ,  $D_{kl}(q||p) = \log \alpha - H(q)$ , lo que representa una desviación de la entropía de su valor máximo. La misma interpretación queda en el caso continuo con la ley 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 entropía, cuando se necesitará un logaritmo específicamente de base  $a$ , se notará la divergencia  $D_{kl,a}$ .*

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

$$D_{kl}(q||p) \geq 0 \quad \text{con igualdad ssi } p = q \text{ (c.s.)}$$

donde (c.s.) significa “casi siempre”.

*Demostración.* Existen varias pruebas, pero la más linda puede ser la usando la desigualdad de Jensen <sup>6</sup>: para  $\phi$  estrictamente convexa,  $E[\phi(X)] \geq \phi(E[X])$  con igualdad ssi  $X$  es determinista (casi siempre). Sea  $X$  de distribución o densidad de probabilidad  $p$ . En el caso discreto como diferencial, se escribe la entropía relativa  $D_{kl}(q||p) = E\left[\frac{q(X)}{p(X)} \log\left(\frac{q(X)}{p(X)}\right)\right]$ . Sea  $Y = \frac{q(X)}{p(X)}$  y  $\phi(u) = u \log u$ , función estrictamente convexa. Entonces  $D_{kl}(q||p) = E[\phi(Y)] \geq \phi(E[Y])$ . Pero  $E[Y] = E\left[\frac{q(X)}{p(X)}\right] = \sum_x q(x) = 1$  (y con una integral en el caso diferencial) y  $\phi(1) = 0$ , lo que cierra la prueba. El caso de igualdad apareciendo si y solamente si  $Y$  es determinista, es decir  $\frac{p(X)}{q(X)}$  determinista, es equivalente a  $p(x) \propto q(x)$  (c.s.), i. e.,  $p = q$  (c.s.) porque ambas suman a uno.  $\square$

Esta propiedad, válida en el caso discreto como continuo, tiene consecuencias fijándose de que

$$I(X; Y) = D_{kl}(p_{X,Y} || p_X p_Y)$$

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 entropía*

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

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

---

<sup>6</sup>Se puede usar también la desigualdad  $\sum t_i \log t_i \geq \sum t_i \log t'_i$  una instancia de la desigualdad conocida como desigualdad log-sum, o conocida también como desigualdad de Gibbs (debido a J. W. Gibbs mismo) (Cover & Thomas, 2006; Rioul, 2007; ?, ?).

Fijense que si  $D_{kl}$  es positiva, no es simétrica 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.

Al final, a pesar de que la forma diferencial de  $D_{kl}$  depende de los estados, queda invariante bajo una misma transformación biyectiva sobre ambos  $p$  y  $q$ .

## 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 entropía máxima

En la termodinámica, el estudio de las características macroscópicas (dinámica de las moléculas) es prohibitivo tan el número de moléculas es importante. Por ejemplo, un litro del gas que respiramos contiene  $2,7 \times 10^{22}$  moléculas. De esta constatación se desarrolló la física estadísticas bajo el impulso de Boltzmann (Boltzmann, 1896, 1898), Maxwell (Maxwell, 1867), Gibbs (Gibbs, 1902), Planck (Planck, 2015) entre otros (ver también (Jaynes, 1965)), considerando el sistema macroscópico a través de lo que llamaron ensembles estadísticos: el sistema global (macroscópico) es al equilibrio pero las configuraciones (micro-estados) son fluctuantes. De una forma, se puede asociar a una configuración su frecuencia de ocurrencia (imaginando tener una infinidad de copias del sistema en el mismo estado macroscópico), es decir su probabilidad de ocurrencia. En este marco, la entropía, describiendo la falta de información, juega un rol fundamental. Un sistema sujeto a vínculos, como por ejemplo teniendo una energía dada, debe estar en sus estado lo más desorganizado dados los vínculos. En su marco, se introdujo la noción de entropía termodinámica, pero la misma es tremendamente vinculada a la entropía de Shannon (claramente, identificando las frecuencias a probabilidades de ocurrencia) <sup>7</sup>. En otro terminos, la distribución describiendo los micro-estados debe ser de entropía máxima, dados los vínculos. Por ejemplo, en un gas perfecto, donde las partículas no interactúan (aparte chocándose), la energía es dada por las velocidades (suma de las energías cinéticas individuales). Dada una energía fija, la distribución de las velocidad debe ser de entropía máxima sujeto a la energía dada (nada más que la energía va a “organizar” las configuraciones posibles). Intuitivamente, en un sistema aislado de  $N$  partículas, las configuraciones van a ser equiprobables, precisamente la distribución maximizando la entropía. **En la sección Sec. 2.5.4 se va a desarrollar un poco más este ejemplo.**

De manera general, el problema se formaliza como la búsqueda de la entropía máxima sujeto a vínculos. Si este principio nació en mecánica estadística (ver también (Jaynes, 1957a, 1957b, 1965)), encontró un

---

<sup>7</sup>Ver epígrafe del capítulo. . .

echo en varios campos: en inferencia bayesiana para elegir distribuciones del a priori <sup>8</sup> conociendo unos momentos de la ley (Robert, 2007; Jaynes, 1968, 1982; Csiszàr, 1991), hacer estimación espectral o de procesos estocásticos autoregresivos (Burg, 1967, 1975; Jaynes, 1982) o (Cover & Thomas, 2006, cap. 12), entre otros (Kapur & Kesavan, 1992, & ref.).

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

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

donde los dos primeros vínculos (positividad, normalización) aseguran de que  $p^*$  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  $\lambda = [\lambda_0 \ \lambda_1 \ \cdots \ \lambda_K]^t$  para tener en cuenta los vínculos, el problema variacional consiste a resolver (Gelfand & Fomin, 1963; van Brunt, 2004; Miller, 2000; Cambini & Martein, 2009; Cover & Thomas, 2006)

$$p^* = \operatorname{argm\acute{a}x}_p \int_{\mathcal{X}} (-p(x) \log p(x) + \lambda^t M(x) p(x)) dx$$

donde  $\lambda$  será determinado para satisfacer a los vínculos. De la ecuación de Euler-Lagrange (Gelfand & Fomin, 1963; van Brunt, 2004), esquemáticamente anulando la “derivada” del integrando con respecto a  $p$  (sera realmente un gradiente con respecto a los componentes de  $p$  en el caso discreto), reparametrizando los factores de Lagrange, se obtiene

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

con  $\lambda$  tal que se satisfacen los vínculos de normalización y momentos. Esta distribución cae en la familia conocida como *familia exponencial* donde los  $M_k$  son conocidos como *estadísticas suficientes* y los  $\lambda_k$  *parámetros 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  $\lambda$  tal que se satisfacen todos los vínculos, en particular la de normalización. Por ejemplo, si  $\mathcal{X} = \mathbb{R}$  y  $K = 0$ ,  $p$  debería ser constante (ley uniforme) sobre  $\mathbb{R}$ , lo que no es normalizable. De la misma manera, si  $K = 3$  y  $M_k(x) = x^k$ , tampoco es normalizable la función obtenida <sup>9</sup>. En otros terminos, en este caso, el problema no tiene solución <sup>10</sup>.

---

<sup>8</sup>A partir de una distribución parametrizada por un parámetro  $\theta$ . El enfoque de bayesiano consiste a modelizar  $\theta$  aleatorio, digamos  $\Theta$ , tal que la distribución de observaciones se escribe entonces  $p_{X|\Theta}$ . Inferir  $\theta$  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*  $p_{\Theta}$ . Si se conocen momentos por una razón o una otra, se puede elegir esta distribución como la “menos informativa” posible, *i. e.*, de entropía máxima dados los momentos.

<sup>9</sup>En el enfoque Bayesiano se puede que no sea problemático, si el *a posteriori* es normalizable (Robert, 2007), pero va más allá de la meta de esta sección.

<sup>10</sup>Más precisamente, existen casos en los cuales se puede acotar la entropía por arriba por un  $H^{\sup}$ , tal que  $\sup_p H(p) \leq H^{\sup}$  pero



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

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

*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

$$\begin{aligned} H(p) &= -D_{\text{kl}}(p \| p^*) - \int_{\mathcal{X}} \lambda^t M(x) p(x) dx \\ &= -D_{\text{kl}}(p \| p^*) - \int_{\mathcal{X}} \lambda^t M(x) p^*(x) dx \\ &= -D_{\text{kl}}(p \| p^*) - \int_{\mathcal{X}} p^*(x) \log p^*(x) dx \\ &= -D_{\text{kl}}(p \| p^*) + H(p^*) \end{aligned}$$

porque  $p, p^* \in \mathcal{P}_m$  y  $\lambda^t M = \log p^*$ . La prueba se cierra notando que  $D_{\text{kl}}(p \| p^*) \geq 0$  con igualdad si y solamente si  $p = p^*$ .  $\square$

Este lema prueba que, dados vínculos “razonables”, la entropía 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 entropía máxima es la distribución uniforme de la propiedad [P'5]a de la sección Sec. 2.2.2 en el caso continuo, o propiedad [P5] de la sección Sec. 2.2.1 en el caso discreto.
- Con  $K = 1$ ,  $\mathcal{X} = \mathbb{R}^d$  y  $M(x) = xx^t$  (visto con  $d^2$  vínculos), la distribución de entropía máxima es la distribución gaussiana de la propiedad [P'5]b de la sección Sec. 2.2.2.

## 2.4.2 Desigualdad de la potencia entrópica

Sean  $X$  e  $Y$  dos variables independientes. Si se conoce las relaciones vinculando  $H(X, Y)$ ,  $H(X)$ ,  $H(Y)$ , una pregunta natural concierne la relación que podría tener  $X + Y$  con cada variable en término de

---

no se puede alcanzar esta cota, i. e., es un supremum, no un máximo (Cover & Thomas, 2006, sec. 12.3).

entropía. La respuesta no es trivial, y el resultado general concierne el caso de variables continuas sobre  $\mathbb{R}^d$ . Es conocido como desigualdad de la potencia entrópica (EPI para entropy power inequality en inglés). No vincula las entropías, sino que las potencias entrópicas.

**Teorema 2-2** (Desigualdad de la potencia entrópica). *Sean  $X$  e  $Y$  dos variables  $d$ -dimensionales continuas independientes. Entonces*

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

*con igualdad si y solamente si  $X$  e  $Y$  son gaussianas con matrices de covarianza proporcionales,  $\Sigma_Y \propto \Sigma_X$*

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

1. Sean  $\tilde{X}$  y  $\tilde{Y}$  gaussianas independientes de matrices de covarianza proporcionales y tal que  $H(\tilde{X}) = H(X)$  y  $H(\tilde{Y}) = H(Y)$ . Entonces

$$N(X + Y) \geq N(\tilde{X} + \tilde{Y})$$

con igualdad si y solamente si  $X$  y  $Y$  son gaussianas.

2. *Desigualdad de preservación de covarianza:*

$$\forall 0 \leq \lambda \leq 1, \quad H(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda H(X) + (1-\lambda)H(Y)$$

con igualdad si y solamente si  $X$  e  $Y$  son gaussianas 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)). 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**

Una gracia de la desigualdad de la potencia entrópica es que puede dar lugar a pruebas informacionales de desigualdades matriciales, como por ejemplo la desigualdad de Minkowsky de los determinentes  $|R_1 + R_2|^{\frac{1}{d}} \geq |R_1|^{\frac{1}{d}} + |R_2|^{\frac{1}{d}}$  para cualesquieras matrices  $R_1, R_2$  simétricas definidas positivas, con igualdad si y solamente si  $R_2 \propto R_1$  (viene de  $X$  e  $Y$  gaussianas de covarianza  $R_1$  y  $R_2$ ). Aparece también para acotar la información mutua entre variables y calcular la capacidad de un canal de comunicación como se lo va a ver más adelante (Cover & Thomas, 2006; Dembo et al., 1991; Rioul, 2007; Johnson, 2004).

**En el caso discreto, no hay un resultado general. Existient solamente 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 2-13** (Proceso de Markov). *Una secuencia  $X_1 \mapsto X_2 \mapsto \dots \mapsto X_n$  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*

$$p_{X_{i+1} | X_i, X_{i-1}, \dots} = p_{X_{i+1} | X_i}$$

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

**Teorema 2-3** (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 si y solamente si  $X \mapsto Z \mapsto Y$  es también un proceso de Markov. En particular, es sencillo ver que para cualquiera 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 más 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  $H(X|Y, Z) = H(X|Y)$  (conociendo  $Y$  suffice para caracterizar completamente  $X$ ), lo que da

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

También,

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

Además, 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 nota de que  $H(X|Z) - H(X|Y, Z)$  es la 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, de las dos formas de  $H(X; Y, Z)$  viene

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

La desigualdad del teorema viene de la positividad de  $I(X; Y|Z)$ . Además, se obtiene la igualdad si y solamente si  $I(X; Y|Z) = 0$ , 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.  $\square$

## 2.4.4 Segunda ley de la termodinámica

Tratando de procesos de Markov, aparece el equivalente de la segunda ley de la termodinámica: un sistema aislado evolua hasta llegar su estado lo más desorganizado (ver ej. (Cover & Thomas, 2006)).

**Lema 2-3.** Sea  $X_1 \mapsto X_2 \mapsto \dots \mapsto X_n \mapsto \dots$  un proceso de Markov, con probabilidades de transición  $p_{X_{n+1}|X_n}$  dadas. Estas últimas modelizan 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  $n \geq 1$ ,

$$D_{\text{kl}}(q_{n+1} \| p_{n+1}) \leq D_{\text{kl}}(q_n \| p_n)$$

las distribuciones  $p_n$  y  $q_n$  no se “alejan” (tiende a acercarse);

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

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

la distribución no se aleja de la distribución estacionaria.

- Además, 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 (además, recuerdese de que la distribución uniforme es la de entropía máxima, i. e., de desorganización máxima).

**Demostración.** Escribiendo  $p_{n+1,n}$  y  $q_{n+1,n}$  las distribuciones conjuntas 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  $D_{\text{kl}}(p_{n+1,n} \| q_{n+1,n}) = D_{\text{kl}}(p_{n+1} \| q_{n+1}) + D_{\text{kl}}(p_{n+1|n} \| q_{n+1|n}) = D_{\text{kl}}(p_n \| q_n) + D_{\text{kl}}(p_{n|n+1} \| q_{n|n+1})$ . Además,  $p_{n+1|n} = p_{X_{n+1}|X_n} = q_{n+1|n}$ , conduciendo a  $D_{\text{kl}}(p_{n+1|n} \| q_{n+1|n}) = 0$  con consecuencia de que  $D_{\text{kl}}(p_{n+1} \| q_{n+1}) = D_{\text{kl}}(p_n \| q_n) + D_{\text{kl}}(p_{n|n+1} \| q_{n|n+1})$ .  $p_{n|n+1}$  no es necesariamente igual a  $q_{n|n+1}$ , pero la divergencia siendo no negativa, se obtiene la primera desigualdad. La segunda desigualdad se obtiene tomando  $q_n = p^*$ . Además, si  $p^*$  es uniforme  $p^*(x) = \frac{1}{|\mathcal{X}|}$  dando  $D_{\text{kl}}(p_n \| p^*) = -H(X_n) + \log |\mathcal{X}|$ , llevando a la última desigualdad.  $\square$

## 2.4.5 Principio de incerteza entrópico

**Bourret 58, Leipnik 59 entre otros que ya citamos un par de veces**

## 2.4.6 Un foco sobre la información de Fisher

Si la entropía y las herramientas relacionadas son naturales como medidas de información, no se puede resumir una distribución a una medida escalar. En el marco de la teoría de la estimación, R. Fisher introdujo una noción de información intimamente relacionada al error cuadrático en la estimación de un parámetro a partir de una variable parametrizado por este parámetro (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 2-14** (Matriz información de Fisher paramétrica). Sea  $X$  una variable aleatoria parametrizada por un parámetro  $m$ -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. Suponga que  $p_X$  sea diferenciable en  $\theta$  sobre  $\Theta$ . La matriz de Fisher, de tamaño  $m \times m$ , es definida por

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

donde  $\nabla_\theta = \left[ \cdots \frac{\partial}{\partial \theta_i} \cdots \right]^t$  es el gradiente en  $\theta$ . Es la matriz de covarianza del score paramétrico  $S(X) = \nabla_\theta \log p_X(X; \theta)$  (se prueba de que su promedio es igual a cero), siendo  $\log p_X$  la log-verosimilitud. Bajo condiciones de regularidad, se puede mostrar <sup>11</sup> que  $J_\theta(X)$  vale también menos el promedio de la Hessiana <sup>12</sup>  $\mathcal{H}_\theta$  de  $\log p_X(X; \theta)$ . Nota: a veces se define la información de Fisher como  $\text{Tr}(J)$ , traza de la matriz información de Fisher.

Como para la entropía, 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 notará también  $J_\theta(p_X)$  según la escritura la más conveniente.

Tomando el gradiente en  $x$  en lugar de  $\theta$  da la matriz de información de Fisher no paramétrica,

**Definición 2-15** (Matriz información de Fisher no paramétrica). Sea  $X$  una variable aleatoria de distribución de probabilidad  $p_X$  definida sobre  $\mathcal{X} \subseteq \mathbb{R}^d$  su soporte. Suponga que  $p_X$  sea diferenciable (en  $x$ ). La matriz

---

<sup>11</sup> Es una consecuencia del teorema de la divergencia, suponiendo que los bordes del soporte  $\mathcal{X}$  no dependen de  $\theta$  y que la función score se cancela en estos bordes.

<sup>12</sup> Para  $f : \mathbb{R}^m \mapsto \mathbb{R}$ ,  $\mathcal{H}_\theta f$  es la matriz de componentes  $\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}$ .

de Fisher no paramétrica,  $d \times d$ , es definida por

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

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

Es interesante notar que:

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

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

(queda válido 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 dispersada así que no hay información sobre su posición.

- $J_\theta$  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  $p_k$ ,  $k = 1, \dots, K$  (Cohen, 1968; Frieden, 2004),

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

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 para cualquier  $K$ . En este caso, observando que  $(\nabla \log p)(\nabla \log p)^t p = \frac{(\nabla p)(\nabla p)^t}{p}$ , considerando el gradiente con respecto a  $\theta$  (resp. a  $x$ ) tratando de  $J_\theta$  (resp.  $J$ ), se obtiene  $\sum_k \lambda_k \frac{(\nabla p_k)(\nabla p_k)^t}{p_k} - \frac{(\nabla \sum_k \lambda_k p_k)(\nabla \sum_k \lambda_k p_k)^t}{\sum_k \lambda_k p_k} = \frac{1}{\sum_k \lambda_k p_k} \sum_{k,l} \lambda_k \lambda_l \left( \frac{p_l}{p_k} (\nabla p_k)(\nabla p_k)^t - (\nabla p_k)(\nabla p_l)^t \right)$ , lo que vale, tratando del caso  $K = 2$ ,  $\frac{\lambda_1 \lambda_2}{p_2 p_2 (\lambda_1 p_1 + \lambda_2 p_2)} (p_2 \nabla p_1 - p_1 \nabla p_2)(p_2 \nabla p_1 - p_1 \nabla p_2)^t \geq 0$ . No puede ser idénticamente cero (salvo si  $\lambda_1 \lambda_2 = 0$  o  $p_1 = p_2 \dots$ ) así que se obtiene la desigualdad sobre la matriz de Fisher integrando esta última desigualdad.

Una otra interpretación de  $J$  como información es debido a la desigualdad de Cramér-Rao que la relaciona a la covarianza de estimación <sup>13</sup> (Rao, 1945, 1992; Rao & Wishart, 1947; Cramér, 1946; Rioul, 2007; Cover &

<sup>13</sup>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 más vieja y debido a K. Pearson & L. N. G Filon (Pearson & Filon, 1898) en 1898; luego fue extendida por Edgeworth (Edgeworth, 1908), Fisher (Fisher, 1925) o Doob (Doob, 1936).

Thomas, 2006; Frieden, 2004; Kay, 1993; van den Bos, 2007). Sea  $X$  parametrizada por  $\theta$ . La meta es estimar  $\theta$  a partir de  $X$ . Tal estimador va a ser una función únicamente de  $X$ , lo que se escribe usualmente <sup>14</sup>  $\hat{\theta}(X)$  (la función no depende explícitamente de  $\theta$ ). Las características de la calidad de un estimador es naturalmente su sesgo  $b(\theta) = E[\hat{\theta}(X)] - \theta$  y su matriz de covarianza  $\Sigma_{\hat{\theta}}$  (la varianza da la dispersión alrededor de su promedio). La desigualdad de Cramér-Rao acota por debajo esta covarianza.

**Teorema 2-4** (Desigualdad de Cramér-Rao). *Sea  $X$  parametrizada por  $\theta$ , de densidad de soporte  $\mathcal{X} \subseteq \mathbb{R}^d$  independiente de  $\theta$ , y  $\hat{\theta}(X)$  un estimador de  $\theta$ . Sea  $b(\theta)$  su sesgo y  $\Sigma_{\hat{\theta}}$  su matriz de covarianza. Sea  $J_b(\theta)$  la matriz Jacobiana del sesgo  $b$ . Entonces,*

$$\Sigma_{\hat{\theta}} - (I + J_b(\theta)) J_{\theta}(X)^{-1} (I + J_b(\theta))^t \geq 0$$

En particular, en el caso  $\theta$  escalar,

$$\sigma_{\hat{\theta}}^2 \geq \frac{(1 + b'(\theta))^2}{J_{\theta}(X)}$$

donde  $b'$  es la derivada de  $b$ .

Tomando  $\theta$  parámetro de posición y  $\hat{\theta} = X$ , estimador sin sesgo ( $b = 0$ ), da lo que es conocido como la desigualdad no paramétrica de Cramér-Rao y toma la expresión

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

o, en el caso escalar,

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

Además, en el caso no paramétrico, se alcanza la cota si y solamente si  $X$  es un vector gaussiano.

Esta desigualdad acota la varianza de cualquier estimador, i. e., da la varianza o error mínimo(a) 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  $\theta$ .

*Demostración.* Sea  $S = \nabla_{\theta} \log p_X$  y  $\theta_0 = E[\hat{\theta}(X)] = \theta + b(\theta)$ . Fijandose que  $\nabla_{\theta} \log p_X p_X = \nabla_{\theta} p_X$ , que  $\hat{\theta}$  no es función de  $\theta$ , y que el soporte  $\mathcal{X}$  no depende de  $\theta$ , se obtiene <sup>15</sup>

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

<sup>14</sup>Por ejemplo, si  $\theta$  es un promedio común a los componentes de  $X$ , un estimador podría ser  $\hat{\theta} = \frac{1}{d} \sum_i X_i$ .

<sup>15</sup>Se supone que los integrandos sean  $\theta$ -localmente integrables, tal que se puede invertir derivada en  $\theta$  e integración.

Además, fijándose que  $E[S(X)S(X)^t] = J_\theta(X)$  y  $E\left[\left(\hat{\theta}(X) - \theta_0\right)\left(\hat{\theta}(X) - \theta_0\right)^t\right] = \Sigma_{\hat{\theta}}$ , la desigualdad de Cauchy-Bunyakovsky-Schwarz <sup>16</sup> conduce a

$$\left(u^t (I + J_b(\theta))^t v\right)^2 = E\left[u^t S(X) \left(\hat{\theta}(X) - \theta_0\right)^t v\right]^2 \leq u^t J_\theta(X) u v^t \Sigma_{\hat{\theta}} v$$

La prueba se termina tomando  $u = J_\theta(X)^{-1} (I + J_b(\theta))^t v$  (recordándose que  $J$  es simétrica).

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  $x$ , es 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 desigualdad de Cramér-Rao no paramétrica.  $\square$

En el caso paramétrico, no se puede estudiar el caso de igualdad del hecho de que  $\hat{\theta}$  no es algo dado. Además, aún dado un estimador (explícitamente independiente de  $\theta$ ), no hay garantía de que existe una densidad parametrizada por  $\theta$  que alcanza la cota, o al revés, dada una familia de densidades, tampoco hay garantía que existe un estimador que permite alcanzar la cota (Cover & Thomas, 2006; Kay, 1993).

Fijense de que, nuevamente, la gaussiana juega un rol particular en la desigualdad de Cramér-Rao no paramétrica, permitiendo de 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 Cramér-Rao se deducen desigualdades de Cramér-Rao escalares

$$|\Sigma_{\hat{\theta}}| \geq \frac{|I + J_b(\theta)|^2}{|J_\theta(X)|} \quad \text{y} \quad |\Sigma_X| \geq \frac{1}{|J(X)|}$$

Obviamente, en la segunda, se alcanza la igualdad si y solamente si  $X$  es gaussiano. Además, para una matriz  $A \geq 0$ , existe la “relación determinante-traza”  $|A|^{\frac{1}{d}} \leq \frac{1}{d} \text{Tr}(A)$ , 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

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

En estos casos, se obtiene la igualdad si y solamente si  $X$  es gaussiano (igualdad de la Cramér-Rao no paramétrica) y además de covarianza proporcional a la identidad (igualdad en la relación determinante-traza).

Se notará que, al imagen de las leyes de entropía máxima, la información de Fisher juega también un rol particular en la inferencia bayesiana a través del prior de Jeffrey (Jeffrey, 1946, 1948; Lehmann & Casella, 1998; Robert, 2007) <sup>17</sup>.

---

<sup>16</sup>De hecho, fue probada por Cauchy para sumas en 1821, para integrales por Bunyakovsky en 1859 y más elegantemente por Schwarz en 1888 (Steele, 2004).

<sup>17</sup>Ver nota de pie 8. 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  $\eta$  es unívocamente obtenido con la Fisher en  $\eta$  o por cambio de variables saliendo de  $p_\Theta$ .



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

**Teorema 2-5** (Fisher como curvatura de la entropía relativa). *Sea  $X$  parametrizado por  $\theta_0 \in \Theta$  con  $\Theta$  conteniendo un vecinaje de  $\theta_0$ . Siendo  $D_{kl}(p_X(\cdot; \theta) \| p_X(\cdot; \theta_0))$  función de  $\theta \in \Theta$ , aparece que*

$$D_{kl}(p_X(\cdot; \theta) \| p_X(\cdot; \theta_0)) = \frac{1}{2} (\theta - \theta_0)^t J_{\theta_0}(X) (\theta - \theta_0) + o(\|\theta - \theta_0\|)$$

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

*Demostración.* La relación es consecuencia de un desarrollo de Taylor al orden 2 de la función  $D_{kl}(p_X(\cdot; \theta) \| p_X(\cdot; \theta_0))$  de  $\theta$ , tomada en  $\theta = \theta_0$ . Por propiedad de  $D_{kl}$ , la divergencia es positiva y se cancela cuando  $\theta = \theta_0$ . Entonces, el primer término del desarrollo vale cero y el segundo también,  $D_{kl}$  siendo mínima en  $\theta = \theta_0$ . Además,

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

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

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

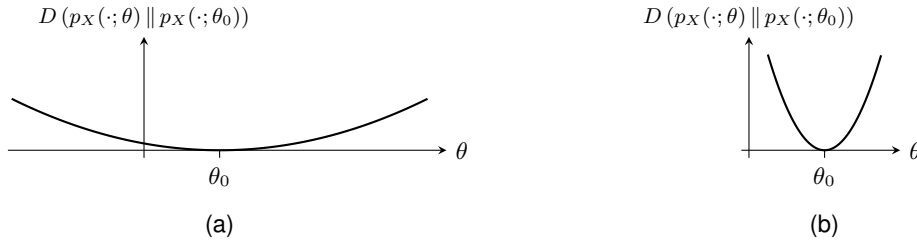
Tomado en  $\theta = \theta_0$  el primer término vale cero. En el segundo se reconoce  $J_{\theta}(X)$ , lo que termina la prueba.  $\square$

Este teorema, ilustrado en la figura Fig. 2-10, relaciona claramente dos objetos viniendo de la teoría de la estimación y de la teoría de la información, mundos a priori diferentes. Como se lo puede ver en la figura, cuando  $J_{\theta}(X)$  tiene pequeños autovalores (figura (a)),  $p_{\theta}$  se “aleja” lentamente de  $\theta_0$  cuando  $\theta$  se aleja de  $\theta_0$ : hay una alta incerteza o pequeña información sobre  $\theta_0$ . Y vice-versa (figura (b)).

Un otro vínculo entre el mundo de la información y el de la estimación aparece a través de la identidad de de Bruijn <sup>18</sup> (Stam, 1959; Cover & Thomas, 2006; Johnson, 2004; Barron, 1984, 1986; Palomar & Verdú, 2006). Esta identidad caracteriza lo que es conocido como canal gaussiano de la figura Fig 2-11-(a), *i. e.*, la

---

<sup>18</sup>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.



**Figura 2-10:** Caso escalar  $\Theta \subseteq \mathbb{R}$  (para la representación) de  $D_{kl}$  en función de  $\theta$ . (a) Caso con  $J_{\theta_0}(X)$  “pequeño” y (b) caso con  $J_{\theta_0}(X)$  “grande”. En el caso (b), la determinación de  $\theta$  usando  $D_{kl}$  va a ser más “sencillo” porque el mínimo es más “picado”.

salida  $Y$  es una versión ruidosa de la entrada. La identidad vincula las variaciones de entropía de salida con respecto al nivel de ruido, y la información de Fisher.

**Teorema 2-6** (Identidad de de Bruijn). Sea  $X$  un vector aleatorio continuo sobre un abierto de  $\mathbb{R}^d$  y admitiendo una matriz de covarianza, y sea  $Y = X + T\mathcal{N}$  donde  $T$  es determinista,  $d \times d'$  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 Fig. 2-11-(a)). Entonces, la entropía 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 parámetro escalar<sup>19</sup>  $\theta$ ,

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

*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  $p(x) = (2\pi)^{-\frac{d}{2}} |T\Sigma_{\mathcal{N}}T^t|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}x^t (T\Sigma_{\mathcal{N}}T^t)^{-1}x\right)$  (el rango máximo de  $T$  asegura que  $T\Sigma_{\mathcal{N}}T^t$  sea invertible). Para una matriz invertible  $R$ , desarrollando  $|R|$  con respecto a su línea  $i$ , se obtiene que  $\frac{\partial |R|}{\partial R_{i,j}} = R_{i,j}^*$  cofactor de  $R_{i,j}$ , dando por la regla de Cramér  $\nabla_R |R| = |R| (R^{-1})^t$  (ver también (Magnus & Neudecker, 1999, cap. 1 & 9)), es decir  $\nabla_R |R|^{-\frac{1}{2}} = -\frac{1}{2}|R|^{-\frac{1}{2}} (R^{-1})^t$ . De  $\frac{\partial |R|^{-\frac{1}{2}}}{\partial T_{i,j}} = \sum_{k,l} \frac{\partial |R|^{-\frac{1}{2}}}{\partial R_{k,l}} \frac{\partial R_{k,l}}{\partial T_{i,j}} = -\frac{1}{2}|R|^{-\frac{1}{2}} \sum_{k,l} (R^{-1})_{l,k} \frac{\partial R_{k,l}}{\partial T_{i,j}}$  con  $R = T\Sigma_{\mathcal{N}}T^t$  (simétrica) y cálculos básicos se obtiene finalmente

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

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

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

<sup>19</sup>Si el parámetro es multivariado, hace falta entender la desigualdad a través de derivas parciales con respecto a los componentes de  $\theta$ .

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 simetría de  $T \Sigma_{\mathcal{N}} T^t$ . Eso significa que

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

dando

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

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

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

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

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

usando la ecuación diferencial en la segunda línea, el hecho de que  $p_Y$  suma a 1 en la tercera línea (su gradiente es cero entonces), y la definición de la matriz de Fisher en la última línea. Usando el teorema de la divergencia (integración 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 términos integrales valen cero, lo que cierra la prueba de la desigualdad general. Además, si  $T = T(\theta)$ , la segunda desigualdad sigue de  $\frac{\partial}{\partial \theta} = \sum_{i,j} \frac{\partial}{\partial T_{i,j}} \frac{\partial T_{i,j}}{\partial \theta} = \text{Tr} \left( \nabla_T \frac{\partial T}{\partial \theta} \right)$ .  $\square$

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

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

se recupera en el caso particular  $T = \sqrt{\theta} I$ . En este caso, la ecuación diferencial satisfecha por la densidad de probabilidad  $p$  es la *ecuación del calor*. Esta desigualdad cuantifica las variaciones de entropías bajo variaciones de “niveles” del ruido del canal de comunicación. De una forma, caracteriza la robustez del canal con respecto 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 más el mundo de la información y el de la estimación. Del lado de la comunicación, consiste a caracterizar la información mutua entre la entrada  $X$  de un canal ruidoso y su salida,  $Y = SX + \mathcal{N}$  donde  $S$  corresponde a un pre-tratamiento antes de la salida.

Eso es ilustrado en la figura Fig. 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 cuadrático promedio  $E \left[ \left\| \hat{X}(Y) - X \right\|^2 \right]$  es la esperanza condicional  $\hat{X}(Y) = E[X|Y]$ . Una característica de un estimador siendo su matriz de covarianza, se notará  $\mathcal{E}(X|Y) = E \left[ (X - E[X|Y]) (X - E[X|Y])^t \right]$  esta matriz. Sorprendentemente, existe también una identidad entre  $I(X; Y)$  y  $\mathcal{E}(X|Y)$ :

**Teorema 2-7** (Identidad de Guo–Shamai–Verdú). Sea  $X$  un vector aleatorio continuo sobre un abierto de  $\mathbb{R}^{d'}$  y admitiendo una matriz de covarianza, y sea  $Y = SX + \mathcal{N}$  donde  $S$  es determinista,  $d \times d'$ , y  $\mathcal{N}$  un vector gaussiano centrado y de covarianza  $\Sigma_{\mathcal{N}}$ , independiente de  $X$  (ver figura Fig. 2-11-(b)). Entonces, la información mutua entre  $X$  e  $Y$  y la matriz de covarianza del estimador de error cuadrático mínimo satisfacen

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

Si  $S = S(\sigma)$  depende de un parámetro escalar  $\sigma$ ,

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

*Demostración.* Notando que  $p_{Y|X}(y, x) = (2\pi)^{-\frac{d}{2}} |\Sigma_{\mathcal{N}}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (y - Sx)^t \Sigma_{\mathcal{N}}^{-1} (y - Sx) \right)$  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

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

(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 covarianza que  $\mathcal{N}$  y de promedio  $Sx$ ), así que

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

La segunda línea viene de la escritura de  $H(Y)$  usando  $p_Y$  como marginales de  $p_{X,Y}$  en  $x$  e 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 e integrando en  $x$  en la

segunda integral; la quinta línea 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 en la segunda ( $p_Y$  sumando a 1, el término se cancela). Además,

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

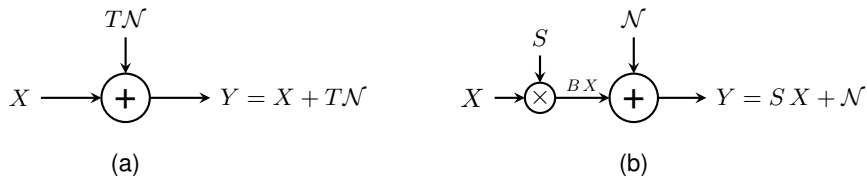
escribiendo  $p_{Y|X}(y, x) p_X(x) = p_{X|Y}(x, y) p_Y(y)$  en la tercera línea. Esta ecuación permite escribir

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

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

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

Existen versiones aún más completas (con gradientes con respecto 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 parámetros (nivel) del ruido. (b) canal gaussiano con un preprocesamiento  $S$  de la entrada.

De la desigualdad de la potencia entrópica y de la identidad de de Bruijn surge una otra desigualdad implicando la potencia entrópica  $N$  y la información de Fisher  $J$ . Esta desigualdad es conocida como desigualdad

de Stam <sup>20</sup> (Cover & Thomas, 2006; Rioul, 2007; Stam, 1959), o a veces “desigualdad isoperimetrica para la entropía” (Wang & Madiman, 2004).

**Teorema 2-8** (Desigualdad de Stam). *Sea  $X$  una variable aleatoria continua sobre  $\mathcal{X} \subseteq \mathbb{R}^d$ . Entonces,*

$$N(X) \operatorname{Tr}(J(X)) \geq d$$

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

*Demostración.* De la desigualdad de la potencia entrópica 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  $\forall \theta > 0$ ,  $\frac{N(X + \sqrt{\theta}\mathcal{N}) - N(X)}{\theta} \geq 1$ . Entonces, tomando el límite  $\theta \rightarrow 0$ , aparece que  $\left. \frac{d}{d\theta} N(X + \sqrt{\theta}\mathcal{N}) \right|_{\theta=0} \geq 1$ . La prueba se cierra con  $\frac{d}{d\theta} N(X + \sqrt{\theta}\mathcal{N}) = \frac{1}{2\pi e} \frac{d}{d\theta} \exp\left(\frac{2}{d} H(X + \sqrt{\theta}\mathcal{N})\right) = \frac{2}{d} N(X + \sqrt{\theta}\mathcal{N}) \frac{d}{d\theta} H(X + \sqrt{\theta}\mathcal{N}) = d N(X + \sqrt{\theta}\mathcal{N}) \operatorname{Tr}(J(X + \sqrt{\theta}\mathcal{N}))$  (por la identidad de de Bruijn). Además, la igualdad se obtiene cuando se alcanza la cota de la desigualdad de la potencia entrópica, es decir cuando  $X$  es gaussiano de varianza proporcional a la del ruido, que es la identidad en este caso.  $\square$

Se puede ver de nuevo el rol central que juega la gaussiana en esta desigualdad. Además, de la desigualdad de Stam se puede deducir también las versiones escalares de la desigualdad de Cramér-Rao. Viene del hecho de que, dada una matriz de covarianza, la entropía  $H(X)$  es máxima cuando  $X$  es gaussiano. Entonces, para cualquier  $X$  de covarianza  $\Sigma_X$ ,  $N(X) \leq |\Sigma_X|^{\frac{1}{d}}$ , dando de la desigualdad de Stam,  $|\Sigma_X|^{\frac{1}{d}} \operatorname{Tr}(J(X)) \geq d$  (y las otras versiones escalares de la relación determinante-traza). Como se lo puede esperar, se obtiene la igualdad si y solamente  $X$  es gaussiano (potencia entrópica alcanzando su cota superior) y de matriz la identidad (desigualdad de Stam se saturando).

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

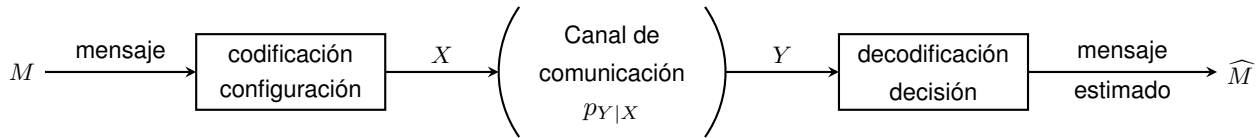
---

<sup>20</sup>Como para 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.

Siguiendo el esquema de comunicación de Shannon, un mensaje que se modeliza como un vector aleatorio <sup>21</sup>  $X$  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 aditivo, 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 cantidad de información que comparten la entrada y la salida del canal. Lo más  $I$  es grande, lo más de información se transmite. Dado el canal, se puede arreglar  $X$  (su distribución) de manera a maximizar  $I(X; Y)$ , es decir la cantidad máxima 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 2-16** (Capacidad de canal). Sea un canal de transmisión,  $X$  su entrada e  $Y$  su salida, como ilustrado figura Fig. 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 codificado (ej. código binario) o puesto en forma (ej. símbolos 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. versión ruidosa de  $X$ ), antes de ser decodificado o usado para tomar una decisión,  $\hat{M}$  siendo la estimación de  $M$  (ej. símbolos estimados a partir de  $Y$ ). Una etapa importante es el vínculo 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 respecto a su entrada.

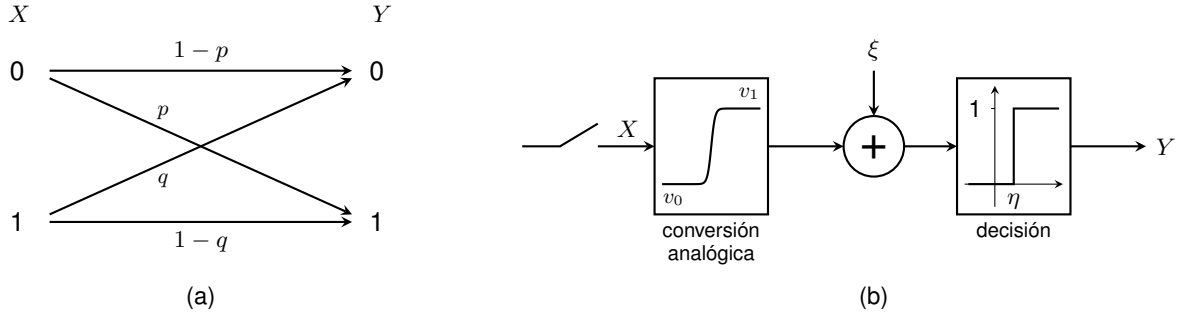
### 2.5.1.1. Canal binario

Suponiendo que el mensaje mandado en un canal es una cadena de símbolos, variables aleatorias independientes, se puede concentrarse sobre cada símbolo. En este marco, un canal de comunicación lo más 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 codificación 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 correctamente o cambia en un 1 (resp. 0), *i. e.*,

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

<sup>21</sup>De punto de vista de un receptor, este mensaje es desconocido. Además, se lo puede ver como una instancia de una clase importante de posibles mensajes, justificando la modelización aleatoria.

$p$  y  $q$  representan errores de comunicación. Tal canal es descrito figura Fig. 2-13-(a). La figura Fig. 2-13-(b) da un esquema “práctico” que podría ser al origen de un tal canal. Cuando  $p = q$ , el canal es conocido como *canal binario simétrico*. Cuando  $p = 0$  y  $q \in (0; 1)$ , el canal es conocido como *canal binario en Z*.



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

En este caso, trabajando con bits, aparece legítimo usar el logaritmo de base 2. Luego, 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

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

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 entropía binaria en bits. Para calcular la capacidad  $C_2$  en bits, hace falta maximizar  $I_2$  con respecto a  $\alpha$ . Diferenciando  $I_2$  en  $\alpha$ , 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,

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

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 probabilidades iguales, y lo mismo



para un 1 en la salida; en otros términos, la salida aparece ser independiente de la entrada. Eso se verifica formalmente con  $\beta = q$ , dando  $p_{Y|X} = p_Y$ , dando una información mutua nula, y entonces una capacidad nula.

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

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

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  $\beta$  “óptimo”), probando de que el extremo es un máximo. Poniendo el  $\alpha^{\text{opt}}$  en la formula de  $I_2(X; Y)$ , luego de muchos cálculos (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  $q \rightarrow 1 - p$ , notando que  $h_2(p) = h_2(1 - p)$  y tomando el límite de esta formula, se recupera que  $C_2 \rightarrow 0$ .

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

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

i. e., la capacidad es menor que 1 bit <sup>22</sup>: 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 además (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

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

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 Fig. 2-14 representa la información mutua  $I(X; Y)$  para unos canales ( $p$  y  $q$  dados) en función de  $\alpha$ . Se nota que la curva es cóncava y tiene un máximo, capacidad del canal. La figura Fig. 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 simético*, 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

---

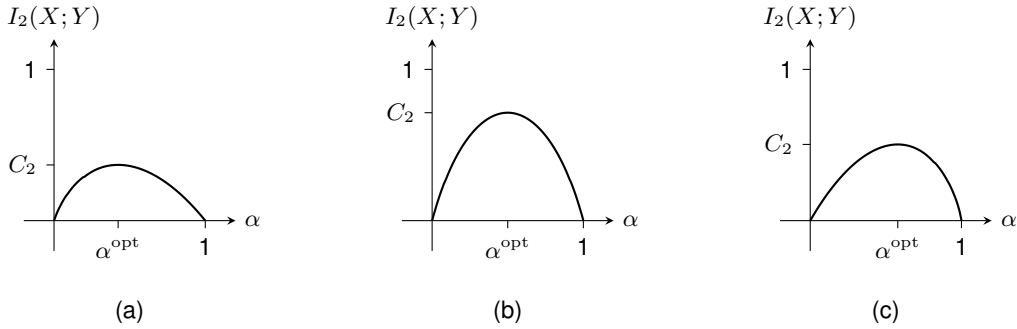
<sup>22</sup>De manera general, de la escritura de  $I$  con entropías condicionales, para  $X$  definido sobre  $\mathcal{X}$  e  $Y$  sobre  $\mathcal{Y}$ , da  $0 \leq C \leq \min(\log |\mathcal{X}|, \log |\mathcal{Y}|)$ . Además,  $p_{Y|X}$  depende solo del canal y no de la entrada, así que para  $p_X = \lambda p_X^{(1)} + (1 - \lambda) p_X^{(2)}$  se obtiene  $p_Y = \lambda p_Y^{(1)} + (1 - \lambda) p_Y^{(2)}$  con  $p_Y^{(i)}$  salida correspondiente a la entrada  $p_X^{(i)}$ . Ahora, de  $I(X; Y) = H(Y) - H(Y|X)$ , el segundo término siendo dependiente solamente del canal, de la concavidad de  $H$  se obtiene de que  $I$  es cóncava con respecto a  $p_X$ .  $p_X$  perteneciendo a un convexo,  $I$  tiene un máximo que es único.

$p = q = \frac{1}{2}$ , y  $C_2 = 0$  (instancia particular de  $q = 1 - p$ ).  $h_2(p)$  es la pérdida en bit para cada bit transmitido. La capacidad  $C_2$  en función de  $p$  es dada figura Fig. 2-15-(b).

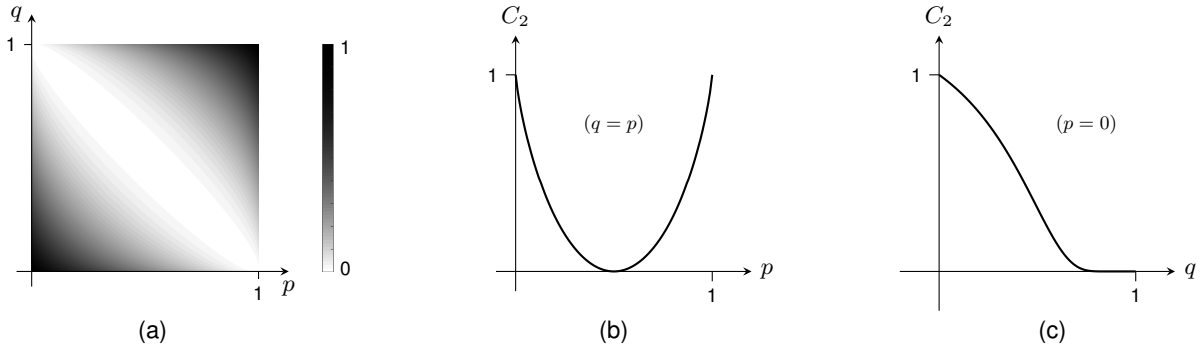
En el caso particular  $p = 0$ , conocido como *canal en Z*, la capacidad es

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

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



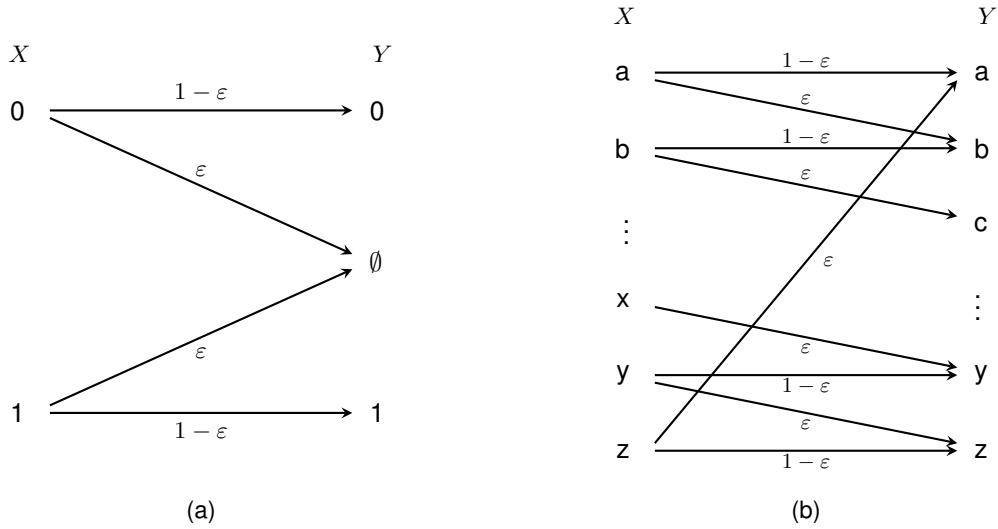
**Figura 2-14:** Información mutua (en bits) entrada-salida  $I_2(X;Y)$  del canal binario en función de  $\alpha = \Pr[X = 0]$ . (a):  $p = 0,4$  y  $q = 0,01$ ; (b):  $p = q = 0,05$  (canal simétrico); (c):  $p = 0$  y  $q = 0,05$  (canal en Z).



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

En (Cover & Thomas, 2006; Rioul, 2007) entre otros, se estudian diversos otros canales discretos, binarios o con más estados. Unos son representados en la figura Fig. 2-16 (ver también (Shannon, 1948; Elias, 1957) o (Arimoto, 1972) para el cálculo numérico de la capacidad en el caso general).

## 2.5.2 Canal de transmisión continuo gaussiano y su capacidad



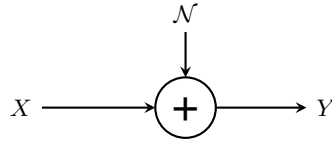
**Figura 2-16:** Ejemplos de canales discretos usuales. (a): canal borrador, donde un 0 (de probabilidad de ocurrencia  $\alpha$ ) o 1 (de probabilidad de ocurrencia  $1 - \alpha$ ) puede transmitirse correctamente o ser borrado/perdido (estado  $\emptyset$ ) con una probabilidad  $\varepsilon$ . Se 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  $\varepsilon$ . De  $I_n(X; Y) = H_n(Y) - H_n(Y|X) = H_n(Y) - h_n(\varepsilon)$  se deduce que  $I_n$  es máxima si  $Y$  es uniforme, lo que es posible si  $X$  es uniforme, dando  $C_n = 1 - h_n(\varepsilon)$ .

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 salida  $Y$  es una versión ruidosa de  $X$ , i. e.,  $Y = X + \xi$  con el ruido  $\xi$  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 Fig. 2-17. Se supone conocida la matriz de covarianza  $\Sigma_{\mathcal{N}}$  del ruido, y se nota  $\Sigma_X$  la de la entrada. En práctica, no se puede mandar un mensaje a una potencia tan alta que se quiere, lo que se traduce por una limitación

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

potencia límite 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  $\text{Tr}(\Sigma_X) \leq P$ . Fijando un  $\Sigma_X$ , la propiedad [P’5]b de la entropía diferencial implica que  $H(Y)$  sea máxima si y solamente si  $Y$  es gaussiana, es decir si y solamente si  $X$  es 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 límite de potencia, hace falta maximizar  $|\Sigma_X + \Sigma_{\mathcal{N}}|$  sujeto a  $\text{Tr} \Sigma_X \leq P$  y  $\Sigma_X \geq 0$  simétrica lo que no es trivial. Se encuentra el enfoque permitiendo solucionar el problema en (Cover & Thomas,



**Figura 2-17:** Canal gaussiano. La entrada  $X$ , modelizada por un vector aleatorio es corrupta 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).

2006, Sec. 9.4). Sea  $U$ , matriz ortogonal ( $UU^t = U^tU = I$ ) de los autovectores de la matriz  $\Sigma_{\mathcal{N}} \geq 0$  simétrica<sup>23</sup>, de columnas  $u_i$  ordenadas tal que los autovalores correspondientes  $\lambda_i^{\mathcal{N}}$  sean en orden creciente, *i. e.*,

$$\Sigma_{\mathcal{N}} = U \text{diag}(\lambda_1^{\mathcal{N}}, \dots, \lambda_d^{\mathcal{N}}) U^t \quad \text{con} \quad 0 \leq \lambda_1^{\mathcal{N}} \leq \dots \leq \lambda_d^{\mathcal{N}}$$

donde  $\text{diag}$  es la matriz diagonal teniendo los  $\lambda_i$  en su diagonal. Sea  $R_X = U^t \Sigma_X U$ . Es sencillo ver que  $|\Sigma_X + \Sigma_{\mathcal{N}}| = |R_X + \Lambda_{\mathcal{N}}|$  (de  $|AB| = |A||B|$ ) y que  $\text{Tr} \Sigma_X = \text{Tr} R_X$  (de  $\text{Tr}(AB) = \text{Tr}(BA)$ ). Entonces, el problema se reduce a maximizar  $|R_X + \Lambda_{\mathcal{N}}|$  sujeto a  $\text{Tr} R_X \leq P$  y  $R_X \geq 0$  simétrica. La desigualdad de Hadamard ya evocada da  $|R_X + \Lambda_{\mathcal{N}}| \leq \prod_i (R_X + \Lambda_{\mathcal{N}})_{i,i} = \prod_i ((R_X)_{i,i} + \lambda_i^{\mathcal{N}})$  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 máximo si los otros términos son nulos). Es decir que la base que diagonaliza  $\Sigma_{\mathcal{N}}$  debe diagonalizar también  $\Sigma_X$ . Sean  $\lambda_i^X$  los términos diagonales de  $R_X$ : queda que maximizar  $\prod_i (\lambda_i^X + \lambda_i^{\mathcal{N}})$  sujeto a  $\sum_i \lambda_i^X \leq P$  y  $\lambda_i^X \geq 0$ . Este problema de optimización sujeto a una desigualdad se resuelva con el enfoque de Karush-Kuhn-Tucker<sup>24</sup> (KKT) (Miller, 2000; Cambini & Martein, 2009), dando  $\lambda_i^X = (\lambda - \lambda_i^{\mathcal{N}})_+$  con  $(\cdot)_+ = \max(\cdot, 0)$  y  $\lambda$  tal que  $\sum_i (\lambda - \lambda_i^{\mathcal{N}})_+ = P$ . En conclusión, la capacidad es dada por

$$C = \frac{1}{2} \log \left( \frac{|\Sigma_{\mathcal{N}} + \Sigma_X|}{|\Sigma_{\mathcal{N}}|} \right) \quad \text{con} \quad \Sigma_X = U \text{diag} \left( (\lambda - \lambda_1^{\mathcal{N}})_+, \dots, (\lambda - \lambda_d^{\mathcal{N}})_+ \right) U^t,$$

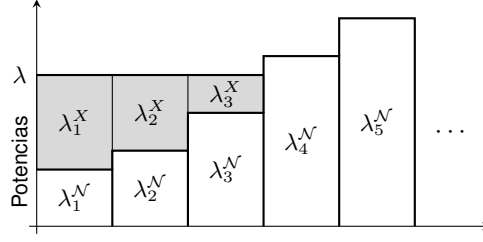
$$\lambda \text{ tal que } \sum_i (\lambda - \lambda_i^{\mathcal{N}})_+ = P$$

alcanzada por  $X$  gaussiano de matriz de covarianza  $\Sigma_X$  así construida.

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

<sup>23</sup>Se recordará de que  $A \geq 0$  significa que  $A$  es definida no negativa.

<sup>24</sup>Se introduce el factor de Lagrange y se maximiza  $\prod_i (\lambda_i^X + \lambda_i^{\mathcal{N}}) + \mu \sum_i \lambda_i^X$ . Eso da  $\lambda_i^X + \lambda_i^{\mathcal{N}} = \lambda$  constante si  $\lambda$  es tal que se satisfaga la positividad de  $\lambda_i^X$ , y  $\lambda_i^X = 0$  sino. En otras palabras,  $\lambda_i^X = (\lambda - \lambda_i^{\mathcal{N}})_+$  con  $\lambda$  el factor de Lagrange después de una reescritura. Queda que maximizar los  $\lambda_i^X$  para maximizar  $|R_X + \Lambda_{\mathcal{N}}|$ , es decir tomar  $\lambda$  lo más grande que se puede, pero satisfaciendo  $\sum_i \lambda_i^X \leq P$ , *i. e.*, alcanzando la igualdad.



**Figura 2-18:** Principio del “water-filling” para obtener los  $\lambda_i^X$  satisfaciendo el vínculo de potencia límite y permitiendo de construir  $\Sigma_X$  a partir de la matriz diagonal de los  $\lambda_i^X$  y la base que diagonaliza la covarianza  $\Sigma_N$  del ruido. La zona en gris representa esquemáticamente  $P$ .

En el caso escalar, se obtiene

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

donde  $\frac{P}{\sigma_N^2}$  es conocido como relación señal-ruido <sup>25</sup>

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 entrópica sin pérdida

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 *fente*, donde los  $X_t$  toman sus valores sobre un alfabeto discreto finito

$$\mathcal{X} = \{x_1, \dots, x_\alpha\} \quad \text{alfabeto fuente}$$

de distribución  $p_X$ . A cada posible secuencia <sup>26</sup>  $s_1 \dots s_n \in \mathcal{X}^n$  de letras de  $\mathcal{X}$ , se quiere asignar un código  $c(s_1 \dots s_n)$  de letras de un alfabeto discreto finito,

$$\mathcal{C} = \{\zeta_1, \dots, \zeta_d\} \quad \text{alfabeto código}$$

El código es dicho *d-ario*. Por ejemplo, se puede asignar un código  $c(x_i) = \zeta_{i,1} \dots \zeta_{i,l_i} \in \mathcal{C}^{l_i}$  a cada símbolo,

<sup>25</sup>Esta formula es muy parecida a la de Shannon, Laplume, o Clavier (Shannon, 1948; Laplume, 1948; Clavier, 1948) (ver también (Cover & Thomas, 2006, Sec. 9.3) o (Rioul, 2007, Sec. 11.2)). De hecho, si se considera símbolos mandados durante  $T$  segundos cada uno (símbolos puestos en forma para dar una señal analógica) usando una banda de transmisión  $B$ , por el teorema de Nyquist  $B = \frac{1}{2T}$  (caso límite). Si el ruido es blanco en la banda  $B$ , de densidad espectral de potencia por unidad de frecuencia igual a  $N_0$ , para un símbolo la relación señal-ruido se escribe  $\frac{P}{N_0 B}$ . Además, se calcula en general la capacidad por unidad de tiempo es decir la capacidad por símbolo dividido por  $T = \frac{1}{2B}$ , i. e.,  $C = B \log \left( 1 + \frac{P}{N_0 B} \right)$  por segundos, lo que es precisamente la capacidad calculada por Shannon. Esta es a veces conocida como formula de Shannon-Hartley.

<sup>26</sup>Por abuso de escritura una cadena de  $n$  símbolos puede ser vista como un  $n$ -uplet.

código llamado *palabras códigos*, y a secuencias  $s_1 \cdots s_n$  la concatenación de las palabras códigos correspondiente a cada símbolo, i. e., el código  $c(s_1) \cdots c(s_n)$ . En el sistema Moore por ejemplo,  $\mathcal{C}$  consiste en un punto, una barra, una espacio entre letras, un espacio entre palabras. En una computadora en general todo se codifica en bits  $\mathcal{C} = \{0, 1\}$ . Más formalmente, sean

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

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 bloques de símbolos de tamaño  $m \geq 1$ :

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

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

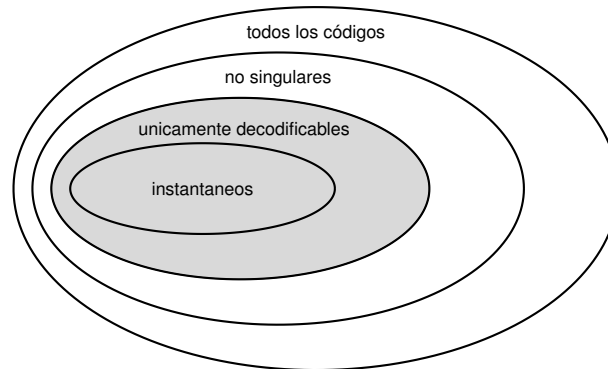
$$\forall n \geq 1, \quad \forall s_1 \cdots s_n \in \mathcal{X}^{nm}, \quad c_m(s_1 \cdots s_n) \equiv c_m(s_1) \cdots c_m(s_n)$$

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

Una manera ingenua de codificar consiste a apoyarse sobre la descomposición de base  $d$  de un entero, i. e., para  $1 \leq i \leq \alpha$ ,  $i-1 = (i_0-1) + (i_1-1)d + \cdots + (i_K-1)d^K$  donde  $K = \lceil \log_d |\mathcal{X}| \rceil$  y  $1 \leq i_k \leq \alpha$ , y de asignar la palabra código  $\zeta_{i_0} \cdots \zeta_{i_K}$  al símbolo  $x_i$ . Haciendo eso, cada palabra código tiene el mismo largo. Pero, es más económico hacer una codificación dicha de largos variables, teniendo en cuenta las probabilidades de aparición de cada  $x_i$ . Implícitamente, es la idea del código de Moore, que asigna un punto o series de puntos o código pequeño a las letras muy frecuentes (ej. un punto para el 'e', dos puntos para el 'i', etc.), y barras o combinaciones largas a las letras que son raras (ej. bara-bara-punto-bara para el 'q' o cinco bars para el '0'). Dicho de otra manera, el código ingenuo sería "eficaz" para  $x_i$  apareciendo con las mismas frecuencias/probabilidades.

En los códigos de largos variables (incluyendo 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 ambigüedad: un código es dicho *descifrable* o *a decodificación única* (o sin perdida) si todas las extensiones son no singulares. Por ejemplo, sean  $\mathcal{X} = \{\alpha, \beta, \gamma, \delta\}$ ,  $\mathcal{C} = \{0, 1\}$  y  $c(\alpha) = 0$ ,  $c(\beta) = 00$ ,  $c(\gamma) = 1$ ,  $c(\delta) = 01$  ( $m = 1$ ). El código es no singular, pero no descifrable. La secuencia 0010 puede provenir de  $\alpha\alpha\gamma\alpha$ , de  $\alpha\delta\alpha$  o de  $\beta\gamma\alpha$ . Obviamente, se requiere en general de un código que sea descifrable. Frecuentemente, se requiere también poder decodificar sobre la marcha, sin esperar de medir toda la secuencia codificada: es lo que se llama *código instantáneo*. Por ejemplo, el código  $c(\alpha) = 00$ ,  $c(\beta) = 10$ ,  $c(\gamma) = 11$ ,  $c(\delta) = 110$  es descifrable, pero no instantáneo. Considera la secuencia 0011011 y marcha sobre ella. 0 no es una palabra código; 00 es y sin ambigüedad proviene de un  $\alpha$  (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  $\gamma$  o de un  $\delta$ ; la letra siguiente siendo un 0, todavía no se puede concluir si 110 vino de  $\gamma$  y algo o  $\delta$ . Al final, con 1101, se sabe que se tuvimos un  $\delta$  porque ninguna palabra código empieza por 01. Al final, sin ambigüedad el antecedente de la secuencia binaria era  $\alpha\delta\gamma$ . Pero se necesitó marchar sobre toda la secuencia antes de decodificar. Obviamente, un código instantáneo 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 están ilustradas en la figura Fig. 2-19 (ver (Cover & Thomas, 2006, cap. 5)).



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

Además de la decodificación sin ambigüedad, una caracterización importante del código es la tasa de codificación <sup>27</sup>

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

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 número de bits promedio del código por símbolo.

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 transmisión. En lo que sigue, nos concentramos en el problema de compresión, sin tener en cuenta el paso de transmisión de mensajes codificados en un canal. Minimizar la tasa es equivalente a minimizar el largo promedio. Además, se puede focalizarse en  $m = 1$ ; todo se extiende sencillamente a  $m > 1$ .

La meta de la compresión es entonces construir un código  $c$ , descifrable, que minimizar el largo promedio

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

---

<sup>27</sup>En (Rioul, 2007) por ejemplo, se define esta tasa suponiendo que cada secuencia fuente es codificado por el mismo número de bits. La tasa es entonces el número de bits por símbolo.

Antes de ir más adelante, hace falta traducir en ecuación el vínculo de que  $c$  sea descifrable. Eso es dado por la desigualdad de Kraft-McMillan (Kraft Jr, 1949; McMillan, 1956; Karush, 1961) <sup>28</sup>

**Teorema 2-9** (Desigualdad de Kraft-McMillan). *Los largos  $l_c(x)$  de las palabras código de un código  $c$  descifrable deben satisfacer la desigualdad*

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

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

**Demostración.** Para cualquier  $k \geq 1$  y cualquiera 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

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

re-escribiendo la segunda suma, agrupando los términos de mismo largos, donde  $\#(m)$  es el número de códigos de  $\mathcal{X}^k$  teniendo el largo  $m$  y  $l_c^{\max} = \max_{x \in \mathcal{X}} l_c(x)$  es el largo mayor.  $c$  siendo descifrable,  $c_k$  debe ser inyectiva, imponiendo  $\#(m) \leq d^m$  (no hay más palabras de largo  $m$  que el cardinal de  $\mathcal{C}^m$ ), dando inmediatamente que necesariamente

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

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

Recíprocamente, 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  $n_\ell$  los números de largos igual a  $\ell = 1, \dots, \ell^{\max} \leq \alpha$ . Consideramos ahora un árbol empezando con una raíz, 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  $\zeta_1, \dots, \zeta_d$ . Estos nudos se dividen cada uno en  $d$  otras ramas, y los nudos de “padre”  $\zeta_i$  se va a asociar las palabras códigos  $\zeta_i \zeta_1, \dots, \zeta_i \zeta_\alpha$ , etc. Este árbol, conocido como árbol de Kraft, es ilustrado en la figura Fig. 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 entonces 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 profundidad de ramas) y de prohibir todas las ramas de padre los nudos seleccionados (líneas punteadas en la figura Fig. 2-20). Estos nudos son llamados *hojas* (no hay ramas). En la capa de “hijos” de profundidad/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 tendríamos  $n_1 d^{-1} + n_2 d^{-2} > 1$ , incompatible con

---

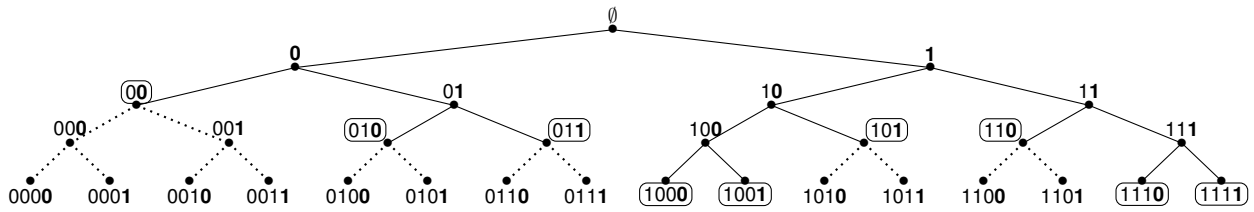
<sup>28</sup>Esta desigualdad fue probada por L. G. Kraft para códigos instantáneos en su tesis de maestría (Kraft Jr, 1949). Luego, fue extendida a los códigos descifrable por B. McMillan (McMillan, 1956) (en una nota de pie de página de su papel, atribua esta observación a J. L. Doob hecha oralmente durante una escuela de verano en Ann Arbor, MI en agosto 1955).



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 (segunda profundidad), y de prohibir que salen de estos nudos nuevas ramas (son entonces hojas en la segunda profundidad), etc. Haciendo así, se asocia un código  $c$  de largos  $l_c(x) = \ell_x$  que aparece libre de prefijo, es decir instantáneo. Entonces, este código es también descifrable.  $\square$

A este punto, se mencionan los hechos siguientes

- Los largos de un código descifrable satisfacen la desigualdad de Kraft-McMillan, pero con el conjunto de largos correspondientes se puede siempre construir un código instantáneo. Claramente, se puede buscar un código de largo promedio mínimo en los códigos instantáneos, sin pérdida de optimalidad (buscar en la clase más amplia de los descifrables no permite bajar el largo promedio).
- En los códigos libres de prefijo, si se fija los números de hojas (última profundidad) borradas contruyendo 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 el número total  $d^{\ell^{\max}}$  de hojas, lo que prueba el teorema para los códigos instantáneos (Kraft Jr, 1949; Karush, 1961).
- El teorema se generaliza obviamente para codificar una fuente (discreta) con un número infinito de estados, tomando el límite  $\alpha \rightarrow \infty$ .
- Si se conocen los largos óptimos, es suficiente para poder construir un código libre de prefijo.



**Figura 2-20:** Árbol de Kraft en el caso binario ( $d = 2$ ). La raíz, de código  $\emptyset$  de largo 0, se divide en dos ramas, de códigos respectivamente 0 y 1 (profundez 1). Cada nudo de esta profundidad se divide en dos ramas (profundez dos), dando cuatro nuevos nudos con los códigos 00 y 01 de padre 0, y 10 y 11 de padre 1. Etc. En cada nudo de esta figura, 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 puede tener nudos “hijos” siendo también una palabra código: se borran las ramas saliendo de un nudo-palabra código (ramas punteadas).

El formalismo dado, se va a ver ahora reaparecer la entropía de Shannon como cota de la codificación de fuente sin pérdida:

**Teorema 2-10** (Cota inferior de códigos descifrables). *Para cualquier código  $c$  descifrable de la fuente  $X$ , su largo promedio es acotado por debajo por la entropía 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

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

El resultado proviene de la positividad de la divergencia de Kullback-Leibler y de la desigualdad de Kraft-McMillan (el argumento del logaritmo siendo menor que 1).  $\square$

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

La entropía aparece también en la cota superior del código óptimo:

**Teorema 2-11** (Cota superior del código descifrable óptimo). *El largo promedio  $L^{\text{opt}}$  del código  $c^{\text{opt}}$  descifrable, de largo promedio mínimo es acotado por arriba por la entropía de Shannon de base  $d$  de  $X$  más un bit (1 símbolo de  $\mathcal{C}$ ),*

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

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

$$\min \sum_{x \in \mathcal{X}} p_X(x) l(x) \quad \text{sujeto a} \quad \sum_{x \in \mathcal{X}} d^{-l(x)} \leq 1$$

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 vínculo, garantizando que el mínimo existe y es único. El problema se resuelva con el enfoque KKT <sup>24</sup>, optimización con vínculos tipo desigualdades (Miller, 2000; Cambini & Martein, 2009), conduciendo a los “largos”

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

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  $c^{\text{sh}}$  descifrable con estos largos. De  $l(x) < -\log_d p_X(x) + 1$  se obtiene

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

$\square$

De

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

se revela el rol fundamental de la entropía en la codificación de fuente sin pérdida. La codificación es a veces dicha *codificación entrópica* y da un rol operacional a la entropía de Shannon. Se notará de la demostración precedente de que aparece un código particular:

**Definición 2-17** (Código de Shannon). *Un código  $c^{\text{sh}}$  de una fuente  $X$ , de largos  $l^{\text{sh}}(x) = \lceil -\log_d p_X(x) \rceil$ , libre de prefijo (construido sobre el árbol de Kraft) es llamado código de Shannon.*

Obviamente, también

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

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  $l^{\text{sh}}(1) = 10$ , y el largo promedio vale  $L(c^{\text{sh}}) = 1,009$ . 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  $\tilde{l}(x) = -\log_d p_X(x)$ , 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 satisfaciendo esta condición es dicha  $d$ -ádica. Sin embargo, el código de Shannon es “competitivo” en el sentido de que:

**Teorema 2-12** (Competitividad del código de Shannon). *Sea  $X$  fuente sobre  $\mathcal{X}$ , de distribución  $p_X$  y  $c^{\text{sh}}$  el código de Shannon asociado sobre el alfabeto código  $\mathcal{C} = \{\zeta_1, \dots, \zeta_d\}$ , de largos  $l^{\text{sh}}(x) = \lceil -\log_d p_X(x) \rceil$ . Para cualquier código  $c$  descifrable y cualquier  $k \geq 1$ ,*

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

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

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

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

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

(añadiendo términos positivos en la suma). La prueba se cierra notando que  $c$  siendo descifrable,  $l_c$  satisface la desigualdad de Kraft-McMillan.  $\square$

Este teorema traduce el hecho de que si  $c^{\text{sh}}$  no es óptimo, tomando cualquier otro código (incluyendo el óptimo), la probabilidad que  $c^{\text{sh}}(X)$  tenga un largo más grande que  $c(X) + k$  decrece exponencialmente con  $k$ . En el ejemplo anterior, si se compara  $c^{\text{sh}}$  y el código óptimo, para  $k = 9$  (caso del código de 1),  $\Pr[l^{\text{sh}}(X) \geq l_c(X) + 9] \leq 0,391\%$ . 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 explícitamente el problema de búsqueda del código óptimo. Números investigadores contruyeron códigos, intentando probar de que eran 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* <sup>29</sup>  $c^{\text{fa}}$  se basa sobre el hecho de que se alcanza la cota mínima para una distribución  $d$ -ádica. El principio es de clasificar los estados de  $\mathcal{X}$  para obtener las probabilidades clasificadas en orden decrecientes ( $p_X^\downarrow$ ). Luego, se divide  $\mathcal{X}$  en  $d$  ensembles a lo más equiprobables que se puede (i. e., de probabilidad a lo más cerca de  $d^{-1}$ ) y de asignar  $\zeta_i$  al conjunto  $i$ . Luego, se repite el proceso a cada sub-conjunto (para tener sub-conjuntos de probabilidades a lo más 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 Fig. 2-21-(a). **Probar/mencionar que también**

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

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  $\zeta_i$  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 profundidad se puede permutar los símbolos asociados a las hojas de esta profundidad sin cambiar el aspecto libre de prefijo y sin que cambien 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) <sup>30</sup>

**Definición 2-18** (Código de Huffman). *Suponemos que existe un  $q \in \mathbb{N}$  tal que <sup>31</sup>  $\alpha = |\mathcal{X}| = d + q(d - 1)$ . El algoritmo de Huffman consiste a construir un arbol donde cada nudo es asociado a un conjunto de símbolos fuente y una letra de  $\mathcal{C}$  de la manera siguiente:*

1. *Clasificar las probabilidades en orden decrecientes: llamamos  $p_i$  las probabilidades rearrregladas y, por cambio de escritura,  $x_i$  el símbolo fuente correspondiente.*

---

<sup>29</sup>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).

<sup>30</sup>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 propuso su propio código, andando al revés del enfoque de Fano, y demostró que era óptimo (Stix, 1991).

<sup>31</sup>Si no, se puede elegir  $q = \left\lceil \frac{n-d}{d-1} \right\rceil$ , y completar  $\mathcal{X}$  con  $d + q(d - 1) - \alpha$  símbolos fuente fictivos de probabilidades ceros, lo que no va a cambiar ni la entropía, ni el largo promedio del código aferente.

2. A cada  $x_i$ ,  $n - d + 1 \leq i \leq n$ , asociar un nudo y la letra “hijo”  $\zeta_i$ .
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 símbolos fuente  $\tilde{x}_i = x_i$ ,  $1 \leq i \leq n - d$  de probabilidades respectivas  $\tilde{p}_i = p_i$  y  $\tilde{x}_{n-d+1} = \{x_j, n - d + 1 \leq j \leq n$  de probabilidad  $\tilde{p}_{n-d+1} = p_{n-d+1} + \dots + p_n$ . El último “super-símbolo” fuente es asociado al nudo padre de la etapa 3.
5. Si quedan más de un (super-)símbolo fuente, volver a la etapa 1 con  $p \equiv \tilde{p}$  y  $x \equiv \tilde{x}$ .

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

Se mencionara que a cada etapa, el nuevo conjunto de super-símbolos fuente contiene exactamente  $d - 1$  símbolos menos que a la etapa precedente. Así, con  $n = d + q(d - 1)$  el algoritmo tiene exactamente  $q + 1$  bucles y en cada profundidad no hay ningún nudo vacío en el sentido que o es una hoja, o es un nudo padre/prefijo (quedarán exactamente  $d$  nudos a agregar a la raíz en la última etapa). Por ejemplo, con  $d = 3$  si tuvieramos  $n = 4$ , en la segunda etapa tendríamos 2 estados a juntar, dando un código de largos 2, 2, 2, 1. Empezando la primera etapa con la asociación de 2 estados, es decir 3 teniendo en cuenta un estado fictivo ( $n = 5$ ,  $q = 1$ ) van a quedar 3 estados en la segunda etapa, dando un código de largos 2, 2, 1, 1, es decir de largo promedio más pequeño.

**Teorema 2-13** (Óptimalidad del código de Huffman). *El algoritmo de Huffman da un código  $c^{\text{huf}}$  de largo promedio mínimo en la clase de los códigos descifrable y los libre de prefijo (se recordará que con los largos de códigos descifrable, siempre se puede construir un código libre de prefijo), es decir  $L^{\text{opt}} = L(c^{\text{huf}})$ .*

*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 más sutil. La prueba más general es dada por Huffman (Huffman, 1952) y se consigue también por parte en (Pigeon, 2003). Suponemos que  $q \geq 1$  (sino, el resultado es obvio). Las etapas son

- Sean  $j, k$  dos índices. Si  $c^{\text{opt}}$  es un código óptimo, y  $c$  un código tal que  $l(x_i) = l_i^{\text{opt}}$ ,  $i \neq j, k$ ,  $l_j = l_k^{\text{opt}}$  &  $l_k = l_j^{\text{opt}}$ , se obtiene  $0 \leq L(c) - L^{\text{opt}} = \sum_i p_i (l_i - l_i^{\text{opt}}) = (p_j - p_k) (l_k^{\text{opt}} - l_j^{\text{opt}})$ . Entonces  $p_j > p_k \Rightarrow l_j^{\text{opt}} \leq l_k^{\text{opt}}$ .
- Sea  $m$  el número de símbolos fuente con un código de largo máximo  $l_{\text{máx}}$  y  $m' = \min(m, d)$ . Del punto anterior, los  $m$  símbolos con palabra código de largo máximo son los de probabilidades más pequeñas.
- Como descrito antes, se puede permutar las letras códigos de una profundidad del árbol de Kraft sin cambiar ni el aspecto libre de prefijo, ni el largo promedio. Se puede entonces considerar el código óptimo tal que los  $m'$  símbolos de probabilidades las más 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  $x_i^{(2)} = x_i$ ,  $1 \leq i \leq n-m'$  de probabilidades respectivas  $p(x_i)$  y  $x_{n-m'+1}^{(2)} \equiv \{x_i\}_{i=n-m'+1}^n$  de probabilidad  $p_{n-m'+1} + \dots + p_n$  (se “plegan” las  $m'$  hojas en un super-símbolo). La codificación óptima es entonces una codificación libre de prefijo de  $\mathcal{X}^{(2)}$ , “árbol raíz” del código óptimo, a la cual se añade una letra código  $\zeta_k$  diferente a cada símbolo del super-símbolo  $x_{n-m'+1}^{(2)}$ . La profundidad máxima del código árbol es  $l_{\max} - 1$  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 símbolo de  $x_{n-m'+1}^{(2)}$  al nudo “vacío” de la profundidad  $l_{\max} - 1$ , sin cambiar el aspecto libre de prefijo, pero ganando una letra código sobre un símbolo, i. e., hacer un código libre de prefijo con un largo promedio menor. Sería contradictorio con la optimalidad del código inicial.
- Para codificar  $\mathcal{X}^{(2)}$ , se necesita por lo menos  $\lceil \log_d(n - m' + 1) \rceil$  profundidad en el árbol raíz. En esta profundidad (máxima en el caso optimista), hay  $d^{\lceil \log_d(n-m'+1) \rceil} \geq n - m' + 1$  nudos. En la última profundidad 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 teníamos el entero  $q = \frac{d^k - d}{d - 1} + \frac{m' - 1}{d - 1}$ . La primera fracción  $\frac{d^k - d}{d - 1} = d^{k-1} + \dots + 1$  siendo entera,  $q$  no puede ser entero con  $m' < d$ . En otros términos, necesariamente  $m' = d$ , i. e., los  $d$  símbolos de probabilidad más débiles son el la última profundidad y se puede elegir que compartent el mismo nudo padre.
- Sea  $c^{\text{opt},(1)}$  el código óptimo correspondiente a  $\mathcal{X}$  y  $c^{(2)}$  el código “padre” sobre  $\mathcal{X}^{(2)}$  ( $c^{\text{opt},(1)}$  quitando la última letra código de los símbolos juntados, i. e., con la raíz común de estos). De la misma manera, sea  $c^{\text{opt},(2)}$  un código óptimo sobre  $\mathcal{X}^{(2)}$  y  $c^{(1)}$  el que se obtiene desplegando el super-símbolo  $x_{n-d+1}^{(2)}$  en  $d$  hojas. De  $L^{\text{opt},(1)} = L(c^{(2)}) + p_{n-d+1} + \dots + p_n$  (pasar de  $\mathcal{X}^{(2)}$  a  $\mathcal{X}$  se añade solo una letra palabra a los símbolos del super-símbolo) y  $L(c^{(1)}) = L^{\text{opt},(2)} + p_{n-d+1} + \dots + p_n$  se obtiene  $(L^{\text{opt},(1)} - L(c^{(1)})) + (L^{\text{opt},(2)} - L(c^{(2)})) = 0$ . Cada término entre parentesis siendo positivo, valen necesariamente cero (la suma de términos positivos vale cero si y solamente si todos son nulos). En conclusión,  $c^{(2)}$  padre de  $c^{\text{opt},(1)}$  queda óptimo,  $c^{(2)} \equiv c^{\text{opt},(2)}$  (y  $c^{(1)} \equiv c^{\text{opt},(1)}$ ).
- Notando que  $|\mathcal{X}^{(2)}| = n - (q - 1)(d - 1)$ , el razonamiento se propaga por inducción, pasando de  $c^{\text{opt},(k)}$  a  $c^{\text{opt},(k+1)}$  juntando los  $d$  super-símbolos de probabilidades más débiles, hasta tener un super-símbolo tendiendo todos los símbolos,  $|\mathcal{X}^{(K)}| = 1$ , raíz del árbol.

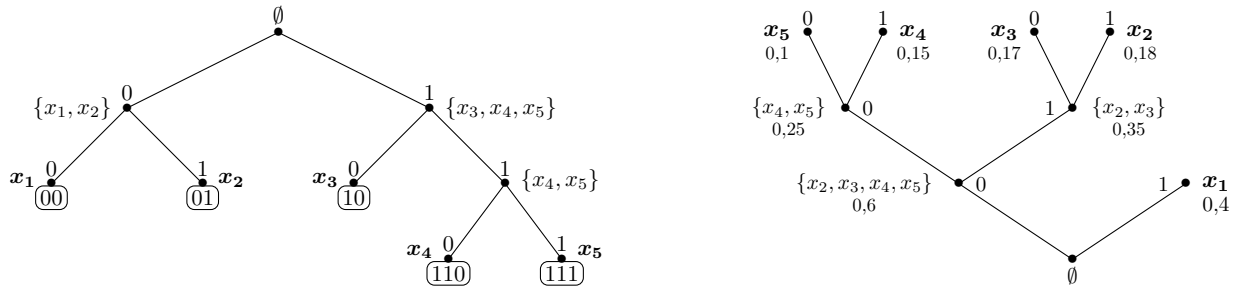
□

De esta prueba, se puede ver que

- Cada profundidad siendo llena, los largos obtenidos van a saturar la desigualdad de Kraft-McMillan.
- Si  $\frac{n-d}{d-1}$  no es entero, en lugar de completar  $\mathcal{X}$  con símbolos fictivos se puede empezar el algoritmo de Huffman juntando los  $n - d - \left\lfloor \frac{n-d}{d-1} \right\rfloor (d - 1) + 1$  símbolos fuentes de probabilidades más débiles en un super-símbolo, y luego hacer el bucle descrito (juntando por super-símbolos de  $d$  símbolos en cada bucle); en este caso, no se satura más 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  $d$  símbolos juntados ( $n$  es necesariamente de la forma  $n = d + q(d - 1) = 2 + q$ ).
- El algoritmo no permite conocer los largos de manera analítica 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 con 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 entropía es  $H_d(X) = \log_d |X|$ , precisamente la incerteza del enfoque de Hartley que corresponde a los números de dits necesarios para codificar (ingenuosamente) la fuente.

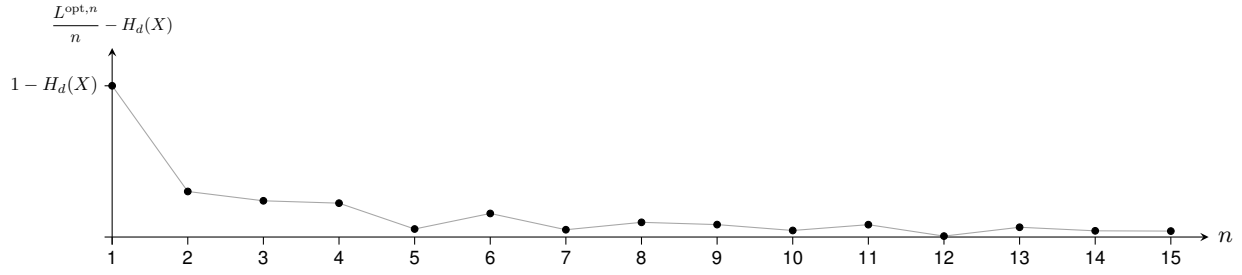


**Figura 2-21:** Construcción de un código binario sobre  $\mathcal{C} = \{0, 1\}$  asociado al vector de probabilidad  $p_X = [0,4 \ 0,18 \ 0,17 \ 0,15 \ 0,1]^t$  sobre el árbol de Kraft. En este caso,  $H_2(X) \approx 2,1514$  (a): Enfoque de Fano, saliendo de la raíz. En cada nudo, se menciona el conjunto de símbolos que va a tener el código correspondiente (en negro cuando es un solo símbolo). Se pasa de una profundidad a la otra dividiendo los conjuntos en sub-conjuntos a lo más equiprobables. 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  $L(c^{\text{fa}}) = 2,25$ . (b): Enfoque de Huffman, saliendo de las hojas. En cada nudo, se menciona el correspondiente (i) conjunto de símbolos, (ii)  $\zeta_i$  de esta profundidad/posición, (iii) la probabilidad asociada al conjunto. Se pasa de una profundidad a la otra juntando los conjuntos menos probables en sobre-conjuntos. En negro son indicados los símbolos simples: van a tener el código agregando los de los nudos yendo de la raíz hasta las hojas. 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  $L^{\text{opt}} = 2,2$ .

Se notará de 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  $H_d(X)$ . El principio es de considerar vectores  $[X_1 \ \dots \ X_n]^t$  viviendo sobre  $\mathcal{X}^n$ , llamado *extensión de orden  $n$  de la fuente*, con un código descifrable (o libre de prefijo) de esta extensión; es llamado *codificación de la extensión de la fuente* pero no es necesariamente una extensión de  $c$ . Así,  $H_d(X_1, \dots, X_n) \leq L^{\text{opt},n} < H_d(X_1, \dots, X_n) + 1$ , es decir, de la independencia,

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

(ver también (Rioul, 2007, cap. 13, teorema de Shannon)). Fijense que si  $\lim_{n \rightarrow \infty} \frac{L^{\text{opt},n}}{n} \rightarrow H(X)$ ,  $\frac{L^{\text{opt},n}}{n}$  no es necesariamente decreciente con respecto a  $n$ . Eso es descrito figura Fig. 2-22. Lo mismo puede ocurrir con el código de Shannon **y lo de Fano**. Además, 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 símbolo de las extensiones  $\mathcal{X}^n$  de orden  $n$  de la fuente  $\mathcal{X}$  y la cota inferior en función de  $n$ . La línea llena en gris sirve como guía. En esta ilustración se usa el ejemplo lo más 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  $X$ . Prácticamente, es usual que no se la tiene. Frecuentemente, es estimada a partir de datos, o, dicho de otra manera, se codifica con una distribución que no es la distribución verdadera de la fuente. Una pregunta que surge es de conocer lo que se pierde usando una distribución no adaptada (o “falsa”). La respuesta general no es obvia, pero tratando del código de Shannon se puede contestar:

**Teorema 2-14** (Código falso de Shannon). *Sea  $c^{\text{sh}}(p)$  el código de Shannon sobre el alfabeto código  $\mathcal{C} = \{\zeta_1, \dots, \zeta_d\}$  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  $L_{c^{\text{sh}}(q)}$  del código  $c^{\text{sh}}(q)$  aplicado a la fuente  $X$  satisface las desigualdades siguientes*

$$H_d(p_X) + D_{\text{kl},d}(p_X \| q) \leq L_{c^{\text{sh}}(q)} < H_d(p_X) + D_{\text{kl},d}(p_X \| q) + 1$$

*Demostración.* Por definición,

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

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

Olvidando el posible extra dit (pensar a la codificación por bloques), este teorema da una interpretación operacional a la entropía relativa, o divergencia de Kullback-Leibler. Esta cantidad cuantifica la pérdida en término 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 codifica la fuente  $X$ , pero suponiendo la distribución  $q$ , se pierde lo que representa la información relativa de  $p_X$  con respecto a la referencia (distribución supuesta)  $q$ .

Existen varios otros modos de codificar símbolos. En particular, con la meta de transmitir los símbolos codificados en un canal de comunicación, a veces no es oportuno de compresar drásticamente el mensaje. Existen por ejemplo codificaciones que permiten una corrección de error en la recepción. Pueden tomar en cuenta las características del canal de transmisión. Estas consideraciones van más allá de la ilustración de



esta sección. El lector puede referirse a (Berlekamp, 1974; Gallager, 1978; Sayood, 2003; Cover & Thomas, 2006; Rioul, 2007) entre otros para tener más 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 Entropías y divergencias generalizadas

A pesar de que la entropía 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, varias nociones informacionales, de tipo entropías o divergencias, aparecieron luego. En esta sección no se desarrollará todos los enfoques ni todas las aplicaciones tan la literatura es importante. La meta es dar los caminos conduciendo a las generalizaciones de la entropía 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 entropía sean asociados nociones de entropías condicionales y relativas.

### 2.6.1 Entropías y propiedades

**Unas primeras generalizaciones particulares** Si la entropía de Shannon fue el punto de salida fundamental en todo el desarrollo de la teoría de la información, un poco más de una decada despues de su papel clave y muy completo, Rényi propuso una medida generalizada (Rényi, 1961). Su punto de vista fue más matemático que físico o ingeniero. Retomó los axiomas de Fadeev (Fadeev, 1956, 1958; Khinchin, 1957) a probabilidades incompletas  $p = [p_1 \ \cdots \ p_n]^t$ ,  $p_i \geq 0$ ,  $w_p = \sum_i p_i \leq 1$ : (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)  $H(\frac{1}{2}) = 1$ , (iv) la aditividad  $H(p \otimes q) = H(p) + H(q)$  donde  $p \otimes q$  es el producto de Kronecker <sup>2</sup>, 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  $p \cup q = [p_1 \ \cdots \ p_n \ q_1 \ \cdots \ q_m]^t$  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 entropía de Shannon. La generalización propuesta por Rényi era de generalizar el axioma (v) reemplazando la media aritmético por una media generalizada (v')  $H^r(p \cup q) = g^{-1} \left( \frac{w_p g(H^r(p)) + w_q g(H^r(q))}{w_p + w_q} \right)$  con  $g$  estrictamente monótona y continua, llamado media *cuasi-aritmética*, o *quasi-lineal*, o de *Kolmogorov-Nagumo*. De las propiedades de la media cuasi-aritmética (Nagumo, 1930; Kolmogorov, 1930, 1991; Hardy et al., 1952), eso es equivalente a buscar una entropía elemental  $H^r(p_i)$  y reemplazar la media aritmética  $\sum_i p_i H^r(p_i)$  por una media de Kolmogorov-Nagumo,  $g^{-1}(\sum_i p_i g(H^r(p_i)))$ . Rényi propuso la función de Kolmogorov-Nagumo  $g_\beta(x) = 2^{(\beta-1)x}$ ,  $\beta > 0$ ,  $\beta \neq 1$ , probando que la entropía que los axiomas (i)-(ii)-(iii)-(iv)-(v') se cumplen y conduce a la entropía de Rényi de un vector de probabilidad  $p$ ,

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

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

$$H_1^r(p) \equiv \lim_{\beta \rightarrow 1} H_\beta^r(p) = H(p) \quad \text{entropía de Shannon}$$

la entropía de Shannon. En su papel, Rényi introdujo una ganancia de información, parecida a una entropía relativa, probando que las solas entropías admisibles son la de Shannon y la que introdujo. Volveremos en la sección siguiente sobre esta entropía relativa, o divergencia de Rényi. Por axiomas, las propiedades [P1] (continuidad), [P2] (invarianza por permutación) y [P10] (aditividad) de la entropía 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 más general.

Unos años después de Rényi, de la famosa escuela matemática checa, J. Havrda & F. Charvát en (Havrda & Charvát, 1967) (ver también (Vajda, 1968, en checo)) volvieron a los axiomas de Khintchin, para extender la entropía de Shannon, *i. e.*, considerando (i) la invarianza por permutación, (ii) la continuidad, (iii) la expansividad, (iv)  $H^{hc}(1) = 0$  y  $H^{hc}(\frac{1}{2}, \frac{1}{2}) = 1$ , pero generalizando la recursividad por (v)  $H^{hc}(p_1, \dots, p_n) = H^{hc}(p_1, \dots, p_{n-2}, p_{n-1} + p_n) + \beta(p_{n-1} + p_n)^\beta H^{hc}\left(\frac{p_{n-1}}{p_{n-1}+p_n}, \frac{p_n}{p_{n-1}+p_n}\right)$ ,  $\beta > 0$ <sup>32</sup>. Con  $\beta = 1$  se recupera la recursividad estándar, 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 entropía (teorema 1)

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

que nombraron  *$\beta$ -entropía structural*. De nuevo, relaxando el axioma (iv), se puede reemplazar en el coeficiente  $2^{1-\beta}$  por  $a^{1-\beta}$ ,  $a > 0$ ,  $a \neq 1$ . De nuevo, para que la entropía de Shannon es un caso particular,

$$H_1^{hc}(p) \equiv \lim_{\beta \rightarrow 1} H_\beta^{hc}(p) = H(p) \quad \text{entropía de Shannon}$$

---

<sup>32</sup>En sus papel, lo imponen para cualquiera par  $(p_i, p_j)$  sin imponer la invarianza por permutación, pero es equivalente a la exposición de este parafo.

(continuidad), [P6] (expansabilidad) de Shannon en este marco. Se probó también que se conserva la propiedad de concavidad con respecto a los  $p_i$  [P8], la de maximalidad [P5] alcanzada para una distribución uniforme (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  $H_\beta^{\text{hc}}$  sea diferente que  $H_\beta^{\text{r}}$ , es sencillo ver que hay un mapa uno-uno entre las dos entropías. Se notara en un marco más general otras propiedades.

Independiente de Havrda & Charvát, todavía en el este, en la escuela húngara, Z. Daróczy en (Daróczy, 1970) defino la entropía  $H^f$  a partir de una *función información*  $f$  satisfaciendo (i)  $f(0) = f(1)$ , (ii)  $f(\frac{1}{2}) = 1$  y la ecuación funcional (ii)  $f(x) + (1-x)f(\frac{y}{1-x}) = f(y) + (1-y)f(\frac{x}{1-y})$  sobre  $\{(x, y) \in [0; 1]^2, x + y \leq 1\}$ , siendo  $H^f(p) = \sum_{i=2}^n s_i f(\frac{p_i}{s_i})$ ,  $s_i = \sum_{j=1}^{i-1} p_j$ . Daróczy mostró que si  $f$  es medible, o continua en 0, o no negativa y acotada, necesariamente  $f(x) = h_2(x) = -x \log_2 x - (1-x) \log_2 (1-x)$ , conduciendo a la entropía 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 entropía de Shannon, propuso extender este axioma (v) por la ecuación funcional  $f_\beta(x) + (1-x)^\beta f_\beta(\frac{y}{1-x}) = f_\beta(y) + (1-y)^\beta f_\beta(\frac{x}{1-y})$ , lo que condujo necesariamente a la entropía (teoremas 2 y 3)

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

es decir nada más que la entropía introducida por Havdra & Charvát. En lo que sigue, se la denota  $H_\beta^{\text{hcd}}$ . Sin embargo, el estudio de Daróczy fue más intensivo que el de Havdra & Charvát. Primero, notó el mapa entre su entropía 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),  $H_\beta^{\text{hcd}}(\frac{1}{2}, \frac{1}{2}) = 1$  (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 determinista y la maximalidad [P5] en el caso uniforme (teorema 6), que incidentalmente  $H_\beta^{\text{hcd}}(\frac{1}{\alpha}, \dots, \frac{1}{\alpha})$  crece con el cardinal  $|\mathcal{X}| = \alpha$ . Muy interesante también es se puede definir una entropía condicional en el mismo modelo que en el caso de Shannon  $H_\beta^{\text{hcd}}(X|Y) = \sum_y [p_{X|Y}(x, y)]^\beta H_\beta^{\text{hcd}}(p_{X|Y}(\cdot, y))$ , que existe una regla de cadena [P14],  $H_\beta^{\text{hcd}}(X, Y) = H_\beta^{\text{hcd}}(Y) + H_\beta^{\text{hcd}}(X|Y)$  y que [P16] condicionar reduce la entropía  $H_\beta^{\text{hcd}}(X|Y) \leq H_\beta^{\text{hcd}}(X)$  (teorema 8). Mostró también que si se pierde la additividad, se obiene para  $X$  e  $Y$  independientes  $H_\beta^{\text{hcd}}(X, Y) = H_\beta^{\text{hcd}}(X) + H_\beta^{\text{hcd}}(Y) + (2^{1-\beta} - 1) H_\beta^{\text{hcd}}(X) H_\beta^{\text{hcd}}(Y)$ . La propiedades de regla de cadena le permitió revisar la caracterización de un canal de transmisión y redefinir una capacidad canal extendidas (capacidad tipo  $\beta$ ; basicamente se usa el mismo enfoque que Shannon, pero usando  $H_\beta^{\text{hcd}}$  en lugar de  $H$ , ver sección 6 del papel).

**Una clase más general y sus propiedades** Las entropías tipo Havdra-Charvát-Daróczy fueron (re)descubiertos varios otras veces y/o estudiados más 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; Boekke & van der Lubbe, 1980; Ferreri, 1980; Tsallis, 1988; Rathie, 1991; Kaniadakis, 2001;

Beck, 2009, entre otros). Un primer enfoque más general es debido a S. Arimoto en los primeros años de la década 1970 (Arimoto, 1971) y redescubierto y estudiado con más detalle y una década después por Burbea y Rao (Burbea & Rao, 1982) y luego estudiado por Salicrú (Salicrú, 1987). La medida propuesta, llamada  $\phi$ -entropía, es definida por

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

Burbea y Rao asociaron una medida de divergencia a esta entropía. Las  $\phi$ -entropías 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  $\phi$ -entropías se enmarcan en una clase un poco más amplia, llamada  $(h, \phi)$ -entropías (Salicrú, Menéndez, Morales & Pardo, 1993; Menéndez, Morales, Pardo & Salicrú, 1997). Cambiamos acá substancialmente su escritura por razones de homogeneidad con la  $\phi$ -entropía (y las divergencias que se introdujera luego) <sup>33</sup>

**Definición 2-19** ( $(h, \phi)$ -entropía). La  $(h, \phi)$ -entropía de una distribución de probabilidad  $p_X$  definida sobre  $\mathcal{X}$  de cardinal finito  $|\mathcal{X}| = \alpha$  es definida por

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

donde o

- $\phi$  es estrictamente convexa y  $h$  creciente, o
- $\phi$  es estrictamente cóncava y  $h$  decreciente

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

(ver también (Esteban, 1997) para una generalización aún más amplia). Cuando  $h(x) = x$  se recupera la  $\phi$ -entropía, incluyendo la de Shannon y las de Havdra-Charvát-Daróczy. Además, la familia de Rényi cae también en esta familia ( $\phi(x) = -x^{\beta}$  y  $h(x) = \frac{\log x}{1-\beta}$ ) así que todas las entropías evocadas en el párrafo anterior.

Como en el caso de Shannon, para  $X = (X_1, \dots, X_d)$ , la  $(h, \phi)$ -entropía de  $X$  es una  $(h, \phi)$ -entropía conjunta de los  $X_i$ .

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

Además se conserva la Schur-convexidad con una recíproca:

---

<sup>33</sup>En la literatura, no hay el signo  $-$ , y hay que invertir cóncava y convexa.

[P<sub>φ</sub>9] Schur-concavidad:

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

En otros terminos, se obtiene la relación de mayorización si se cumple la relación de orden entrópicas para cualquiera par de funciones entrópicas  $(h, \phi)$ . La Schur-concavidad (y su reciproca) es consecuencia de la desigualdad de Schur (Schur, 1923) o Hardy-Littlewood-Pólya (Hardy et al., 1929, 1952) o Karamata (Karamata, 1932) (ver también (Marshall et al., 2011, Cap. 3, Prop. C.1 & Cap. 4, Prop. B.1) o (Bhatia, 1997, Teorema II.3.1)):  $p \prec q \Rightarrow \sum_i \phi(p_i) \leq \sum_i \phi(q_i)$  para toda función  $\phi$  convexa.

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

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

así que

$$H_{(h,\phi)}\left(\left[\frac{1}{\alpha} \quad \dots \quad \frac{1}{\alpha}\right]^t\right) \quad \text{función creciente de } \alpha$$

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

[P<sub>φ</sub>8] Si  $h$  es cóncava, entonces  $H_{(h,\phi)}(p)$  es cóncava con respecto a  $p$ . Eso es una consecuencia de la concavidad de  $\phi$  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 entropía de Rényi es cóncava, pero se prueba 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 cóncava (Bengtsson & Życzkowski, 2006, p. 57).

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

[P<sub>φ</sub>7] Sean  $X$  definido sobre  $\mathcal{X}$  y  $\bar{X}$  sobre  $\bar{\mathcal{X}}$ ,

$$\left\{ \begin{array}{l} \bar{\mathcal{X}} = \{x_1, \dots, x_{\alpha-2}, \bar{x}_{\alpha-1}\} \quad \text{con el estado interno} \quad \bar{x}_{\alpha-1} = \{x_{\alpha-1}, x_{\alpha}\}, \\ p_{\bar{X}}(x_i) = p_X(x_i), \quad 1 \leq i \leq \alpha-1 \quad \text{y} \quad p_{\bar{X}}(\bar{x}_{\alpha-1}) = p_X(x_{\alpha-1}) + p(x_{\alpha}) \quad \text{distribución sobre } \bar{\mathcal{X}} \\ \bar{q}(x_j) = \frac{p_X(x_j)}{p_X(x_{\alpha-1}) + p_X(x_{\alpha})}, \quad j = \alpha-1, \alpha \quad \text{distribución del estado interno} \end{array} \right.$$

$$H_{(h,\phi)}(p_X) \geq H_{(h,\phi)}(p_{\bar{X}})$$

Esta desigualdad es consecuencia de la desigualdad de Petrović (Kuczma, 2009, 43, Teorema 8.7.1),  $\phi(a+b) \geq \phi(a) + \phi(b)$  para  $\phi$  convexa y que se cancela en 0 (y la conversa en el caso cóncavo), conjuntamente con  $h$  creciente (resp. decreciente). A parte en el caso de Shannon y el de Havdra-Charvát-Daróczy, no hay un vínculo inmediato entre  $H_{(h,\phi)}(p_X)$  y  $H_{(h,\phi)}(p_{\bar{X}})$ .

Se conserva la superaditividad [P12]. De hecho, si  $\phi$  es convexa (resp. cóncava) 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,  $\phi(p_{X,Y}(x_i, y_j)) =$

$\phi(p_{X|Y}(x_i, y_j)p_Y(y_j)) \leq p_{X|Y}(x_i, y_j)\phi(p_Y(y_j))$ , i. e.,  $\sum_{i,j} \phi(p_{X,Y}(x_i, y_j)) \leq \sum_{i,j} p_{X|Y}(x_i, y_j)\phi(p_Y(y_j)) = \sum_i \phi(p_Y(y_j))$  (resp. desigualdad reversa). Se cierra la prueba con la crecencia (resp. decrecencia) de  $h$ .

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

**Teorema 2-15.** 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).

$$H_{(h,\phi)}(p_{X,Y}) \leq H_{(h,\phi)}(p_X \otimes p_Y) \quad \forall p_{X,Y} \quad \Longleftrightarrow \quad \phi(x) = x \log x$$

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

*Demostración.* La recíproca de este teorema es nada más que la propiedad [P11] con el hecho de que  $h$  es creciente en este caso.

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

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

**Etapas 1:** Sea el vector de probabilidad

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

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

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

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

$$C_{a,\alpha,b,\beta} = [-1 + \max\{ab, \alpha\beta, 1-a\beta, 1-\alpha b\}, \min\{ab, \alpha\beta, 1-a\beta, 1-\alpha b\}]$$

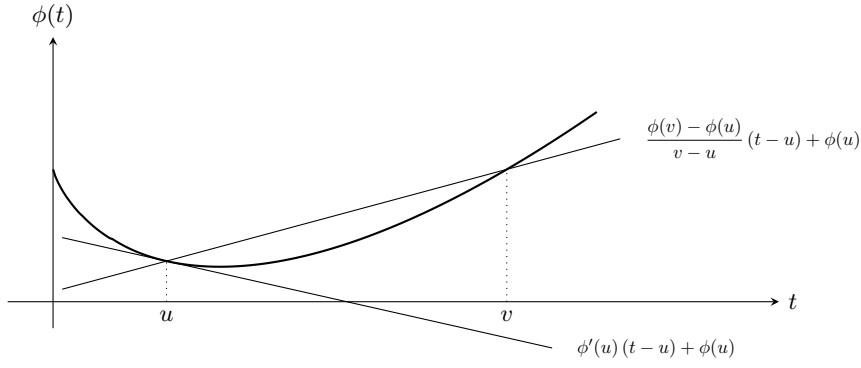
Ahora, si  $\phi$  es convexa (resp. cóncava)

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

i. e., la variación (cuerda) es mayor que la derivada en  $b$ , como ilustrado figura Fig. 2-23 (desigualdad reversa para  $\phi$  cóncava).

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  $c \in \overset{\circ}{C}_{a,\alpha,b,\beta}$  donde  $\overset{\circ}{\phantom{x}}$  denota el interior de un conjunto, se obtiene para  $\phi$  convexa,

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



**Figura 2-23:**  $\phi$  estrictamente convexa: la variación (cuerda)  $\frac{\phi(v) - \phi(u)}{v - u}$  es mayor que la derivada  $\phi'(u)$ . Aplicado a dos distribuciones  $p$  y  $q$ , de componentes  $p_i$  y  $q_i$ , con  $u = p_i$  y  $v = q_i$  y sumando, se obtiene  $H_\phi(q) - H_\phi(p) \geq \sum_i (p_i - q_i) \phi'(p_i)$  con  $H_\phi \equiv H_{(\text{id}, \phi)}$ ,  $\text{id}$  siendo la identidad.

(para  $\phi$  cóncava se reemplaza  $H_\phi$  por  $-H_{-\phi}$  con la igualdad inversa), donde

$$g(a, \alpha, b, \beta, c) = \phi'(ab + c) + \phi'(\alpha\beta + c) - \phi'(a\beta - c) - \phi'(\alpha b - c). \quad (4)$$

Si existe  $(s, u, t, v) \in \hat{D}$  tal que  $g(s, u, t, v, 0) \neq 0$ . De la continuidad de  $\phi'$ , la función  $g$  es continua, y entonces exista un vecinaje  $V_0 \subset \hat{C}_{s, u, t, v}$  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 concluir que, de la desigualdad debido a la concavidad de  $\phi$  (resp. convexidad),  $H_\phi(p_{X,Y})$  puede ser mayor (resp. menor) que  $H_\phi(p_X \otimes p_Y)$ , y entonces, con la crecencia (resp. decrecencia) de  $h$  que si  $g(a, \alpha, b, \beta, 0)$  no es idénticamente cero sobre  $\hat{D}$ ,  $H_{(h, \phi)}$  no puede ser subaditiva (conjunta vs product of marginales).

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

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

así que no se puede usar el argumento de la etapa 1 para concluir. Sin embargo, se puede solucionar esta ecuación funcional, siguiendo (Daróczy & Járαι, 1979, § 6) donde una ecuación funcional muy similar es estudiada. Por eso, se fija  $(a, b) \in (0, 1)^2$ , se deriva la identidad precedente con respecto a  $\alpha$  se multiplica el resultado por  $\alpha$  para obtener

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

Eso significa de que  $x \phi''(x)$  es constante sobre  $x \in (0, (1 - a) \max\{b, 1 - b\})$ , y para cualquiera par  $(a, b) \in (0, 1)^2$ . Entonces,  $x \phi''(x)$  es constante sobre  $x \in (0, 1)$ , es decir que  $\phi$  es necesariamente de la forma  $\phi(x) = \lambda x \ln x + \mu x + \nu$ . Debido a la continuidad de  $\phi$ , queda válido 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  $\mu$  en  $\nu$  sin cambiar el valor de entropía). Además, la constante  $\nu$  no altera la concavidad (resp. convexidad) de  $\phi$ , así que se la puede trasladar en la función  $h$  (sin cambiar la monotonidad). Para que  $\phi$  sea convexa (resp. cóncava) hace falta tener  $\lambda > 0$  (resp.  $\lambda < 0$ ) así que, sin pérdida de generalidad,  $\lambda$  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 subaditiva (propiedad [P11]), cualquiera función creciente de  $H$  va obviamente quedar subaditiva, lo que cierra la prueba.  $\square$

Al revés, a partir de  $p_{XY} = \frac{1}{2} \begin{bmatrix} 1 & 0 \end{bmatrix}^t \otimes \begin{bmatrix} 1 & 0 \end{bmatrix}^t + \frac{1}{2} \begin{bmatrix} 0 & 1 \end{bmatrix}^t \otimes \begin{bmatrix} 0 & 1 \end{bmatrix}^t$  se obtiene  $p_X = p_Y = \frac{1}{2} \begin{bmatrix} 1 & 1 \end{bmatrix}^t$  y entonces (i)  $H_{(h,\phi)}(p_{XY}) = h(-2\phi(\frac{1}{2}))$ ,  $H_{(h,\phi)}(p_X \otimes p_Y) = h(-4\phi(\frac{1}{4}))$  y  $H_{(h,\phi)}(p_X) + H_{(h,\phi)}(p_Y) = 2h(-2\phi(\frac{1}{2}))$ , así que, en este ejemplo  $H_{(h,\phi)}(p_{XY}) > H_{(h,\phi)}(p_X \otimes p_Y)$  (consecuencia de la Schur-convavidad) y  $H_{(h,\phi)}(p_{XY}) > H_{(h,\phi)}(p_X) + H_{(h,\phi)}(p_Y)$ : Tampoco las  $(h, \phi)$ -entropía no son super-additivas.

La definición de entropías generalizadas condicionales aparece mucho más problematico. Por ejemplo, si se define a la Shannon, es decir definiendo  $H_{(h,\phi)}(X|Y)$  tomando  $\sum_{y \in \mathcal{Y}} p_Y(y) H_{(h,\phi)}(p_{X|Y}(\cdot, y))$  se pierde la regla de cadena [P14]. Como se lo ha visto, en el marco de la entropía de Havdra-Charvát-Daróczy, se conserva la regla de cadena si se reemplaza  $p_Y$  por su potencia  $p_Y^\beta$ . 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 ejemplo se pierde el diagrama de Venn aparte si se define la entropía condicional a partir de la regla de cadena. Pero en este caso, si la superaditividad garantiza la positividad de la entropía 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.

Como en el caso de Shannon, se puede extender la generalización de la entropía al caso de vectores aleatorios admitiendo una densidad de probabilidad, reemplazando la suma por una integración.

**Definición 2-20** ( $(h, \phi)$ -entropía 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  $(h, \phi)$ -entropía diferencial de la variable  $X$  es definida por*

$$H_{(h,\phi)}(p_X) = H_{(h,\phi)}(X) = h\left(-\int_{\mathcal{X}} \phi(p_X(x)) dx\right)$$

con  $h$  y  $\phi$  cumpliendo los requisitos de la definición discreta 2-19 (de  $\phi(0)$ , se puede escribir la integración en  $\mathbb{R}^d$ ).

De nuevo para  $X = (X_1, \dots, X_d)$ , la  $(h, \phi)$ -entropía diferencial de  $X$  es una  $(h, \phi)$ -entropía diferencial conjunta de los  $X_i$ .

La versión diferencial de la  $(h, \phi)$ -entropía comparte obviamente las mismas debilidades del caso particular de Shannon: se pierden la propiedad de invarianza por transformación biyectiva [P3], i. e., independencia de los estados, la positividad [P4], la de cota superior [P5] (salvo si se pone vínculos, ver más adelante), en adición de las que ya la versión discreta perdió.



Sin embargo, se conservan unas propiedades, y entre otros si  $h$  es cóncava, la  $(h, \phi)$ -entropía diferencia el cóncava [P <sub>$\phi$</sub> 8]. Más sorprendentemente a primer vista, se conserva la  $(h, \phi)$ -entropía diferencia bajo un rearreglo [P'2],

$$H_{(h,\phi)}(p_X^\downarrow) = H_{(h,\phi)}(p_X)$$

De hecho, como evocado en el caso de Shannon, eso fue probado entre otros en (Lieb & Loss, 2001) o (Wang & Madiman, 2004, Lema 7.2) <sup>34</sup>.

Se probó en (Chong, 1974) o (Wang & Madiman, 2004, Prop. 7.3) que se conserva la Schur-concavidad [P9] para las  $\phi$ -entropías. Entonces, de  $h$  creciente (para  $\phi$  cóncava desigualdad reversa para la integral, pero  $h$  es decreciente), se generaliza a las  $(h, \phi)$ -entropías, *i. e.*,

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

**Quid subaditividad ssi fct creciente de Shannon?**

## 2.6.2 Divergencias y propiedades

### Primer generalizaciones, saliendo de Shannon y Kullback-Leibler - divergencia de Jensen-Shannon

Como se lo ha visto tratando de la entropía relativa, la divergencia de Kullback-Leibler no define una distancia entre distribuciones de probabilidades, siendo no simétrica entre otros. Un primer paso para recuperar la simetría sin perder la positividad de esta medida informacional fue simetrizarla, definiendo lo que es conocido como *J-divergencia* (Kullback & Leibler, 1951; Kullback, 1968; Lin, 1991) <sup>35</sup>,

$$D_J(q||p) = D_{kl}(p||q) + D_{kl}(p||q)$$

Esta versión simetrizada de la divergencia queda naturalmente positiva, pero sufre todavía de unas debilidades de  $D_{kl}$ . Esta bien definida siempre que el soporte de  $p$  es incluido en lo de  $q$  y vice-versa. Además, no cumple tampoco la desigualdad triangular. A pesar de sus debilidades, se usó bastante en problemas de discriminación, debido a su positividad con igualdad si y solamente si  $p = q$  (propiedad herida del hecho de que la suma de términos positivos es nula si y solo si cada uno vale cero).

Unas décadas después, Lin introdujo lo que llamó *K-divergencia* directada,  $K(p, q) = D_{kl}(p||\frac{p+q}{2})$ , su

---

<sup>34</sup>Recuerdese que en (Lieb & Loss, 2001, Sec. 3.3) lo muestran para  $\phi$  d es la diferencia de dos funciones monótonas, siendo una función convexa un caso particular.

<sup>35</sup>Esta expresión aparece en (Jeffrey, 1946, Ec. (1)) o (Jeffrey, 1948), antes de la introducción de la divergencia de Kullback-Leibler en el campo de la estimación Bayesiana, Jeffrey siendo citado por Kullback y Leibler.

versión simetrizada, antes de generalizarla bajo la terminología de *divergencia de Jensen* (Lin, 1991) <sup>36</sup>.

$$\begin{aligned} D_{js}^{\pi}(p_1, p_2) &= \pi_1 D_{kl}(p_1 \| \pi_1 p_1 + \pi_2 p_2) + \pi_2 D_{kl}(p_2 \| \pi_1 p_1 + \pi_2 p_2) \\ &= H(\pi_1 p_1 + \pi_2 p_2) - \pi_1 H(p_1) - \pi_2 H(p_2) \quad \pi = [\pi_1 \quad \pi_2], \quad 0 \leq \pi_1 = 1 - \pi_2 \leq 1 \end{aligned}$$

$D_{js}^{\pi}$  heride obviamente de  $D_{kl}$  su positividad con igualdad si y solamente si  $p_1 = p_2$ . La misma propiedad puede ser visto a través de la desigualdad de Jensen, dando este nombre a la medida. Además, se quita el problema de definición, siendo de que el soporte de  $\pi_1 p_1 + \pi_2 p_2$  siempre contiene el de  $p_1$  y el de  $p_2$ . No es simétrica en general, pero se obtiene esta propiedad cuando  $\pi = \pi_u \equiv [\frac{1}{2} \quad \frac{1}{2}]^t$ . Además, en este caso, a pesar de que la divergencia no cumpla la desigualdad triangular, aparece que  $\left(J_{js}^{\pi_u}(p_1, p_2)\right)^s$ ,  $0 < s \leq \frac{1}{2}$  es una métrica (Osán, Bussandri & Lamberti, 2018) o (Endres & Schindelin, 2003; Österreicher & Vajda, 2003; Kafka, Österreicher & Vincze, 1991, para  $s = \frac{1}{2}$ ). Si puede parecer más lógico definir tal divergencia con a priori/proporciones  $\pi_i$  iguales, de hecho la versión no simétrica, con pesos  $\pi_i$  vuelve natural en el marco de la discriminación donde apareció implícitamente esta cantidad. En particular cuando estamos frente a dos hipótesis  $i = 1, 2$  o clases, a las cuales la distribución de las observaciones es  $p_i$ , con probabilidad a priori  $\pi_i$ . A partir de observaciones  $x$  hay que elegir si eran sorteando de  $p_1$  o  $p_2$  (distribuciones de sampleos, *i. e.*, condicionalmente a la hipótesis). El enfoque Bayesiano más natural consiste maximizar la probabilidad a posteriori (probabilidad de estar en hipótesis  $i$  condicionalmente a la observación), y se prueba que la probabilidad de error es dada por  $P_e = \sum_x \min(\pi_1 p_1(x), \pi_2 p_2(x))$  (o con una integral en el caso continuo) (Kay, 1993). Probó Lin de que

$$\frac{1}{4} \left( H_2(\pi) - D_{js}^{\pi}(p_1, p_2) \right)^2 \leq P_e \leq \frac{1}{2} \left( H_2(\pi) - D_{js}^{\pi}(p_1, p_2) \right)$$

con el logaritmo de base 2 en la definición de  $D_{js}^{\pi}$ , lo que da naturalmente un rol operacional a esta divergencia. Incidentalmente, de esta desigualdad es inmediato ver de que  $D_{js}^{\pi}(p_1, p_2) \leq H_2(\pi) - 2P_e \leq H_2(\pi)$ :

$$0 \leq D_{js}^{\pi}(p_1, p_2) \leq \log(2)$$

(cota igual a 1 usando el logaritmo de base 2).

Un otro vínculo natural entre la divergencia de Jensen-Shannon y las medidas informacionales a la Shannon viene todavía del campo de la clasificación. Si unos datos pueden provenir de una distribución  $p_i$ ,  $i = 1, 2$ , con una probabilidad  $\pi_i$ , la variable aleatoria  $X$  dada por los datos tiene la distribución de mezcla  $p = \sum_i \pi_i p_i$  como ilustrado figura Fig. 2-6-(b). Sea  $Z$  la variable aleatoria binaria sobre  $\{1, 2\}$  tal que  $\Pr[Z = i] = \pi_i$ , variable de selección entre las distribuciones  $p_i$  (ej. la moneda de la figura). Por definición de la entropía condicional,  $H(X|Z) = \sum_i \pi_i H(X|Z = i) = \sum_i \pi_i H(p_i)$ . De  $D_{js}^{\pi}(p_1, p_2) = H(p) - \sum_i \pi_i H(p_i)$  viene  $D_{js}^{\pi}(p_1, p_2) = H(X) - H(X|Z)$ , es decir

$$D_{js}^{\pi}(p_1, p_2) = I(X; Z)$$

---

<sup>36</sup>De hecho, apareció implícitamente en varios trabajos anteriores, por ejemplo en mecánica cuántica (Holevo, 1973, 2011) o en reconocimiento de patrones (Wong & You, 1985)

La divergencia de Jensen-Shannon mide la información mutua entre la observación  $X$  y la variable de selección  $Z$ , justificando aun más su uso natural en problemas de clasificación o selección de modelos. Incidentalmente, de  $I(X; Z) = H(Z) - H(Z|X) \leq H(Z) \leq \log(2)$  ( $Z$  siendo discreta) se recupera la cota mayor de  $D_{js}^\pi$ , y una cota aun más fina

$$D_{js}^\pi(p_1, p_2) \leq H(\pi)$$

Se encuentran otras desigualdades implicando  $D_{js}^\pi$  y  $D_J$  o  $D_{js}^\pi$  y la distancia  $L^1$  entre distribuciones o divergencia de variación total en (Lin, 1991).

Más allá, en el campo de la clasificación, se puede tratar de más de dos clases, dando lugar a la generalización de la divergencia de Jensen-Shannon a  $n$  distribuciones de probabilidad y  $\pi$  un  $n$ -componentes vector de probabilidad,

$$D_{js}^\pi(p_1, \dots, p_n) = H\left(\sum_i \pi_i p_i\right) - \sum_i \pi_i H(p_i)$$

De la desigualdad de Jensen, esta cantidad queda positiva con igualdad si y solo si todos los  $p_i$  son iguales. Se conserva una cota superior

$$D_{js}^\pi(p_1, \dots, p_n) \leq H(\pi) \leq \log(n)$$

así que  $D_{js}^\pi(p_1, p_2) = I(X; Z)$  con  $X$  de distribución la mezcla  $\sum_i \pi_i p_i$  y  $Z$  definida sobre  $\{1, \dots, n\}$  variable de selección de distribución  $\pi$ .

**convexidad?**

**Clase de Burbea-Rao o  $f$ -Jensen (0) Jensen generalizada caso  $H(\phi)$  Ref. Nielsen ArXiv 2010, Nielsen Boltz 2011 -> cualquier función cóncava  $H^\phi$  de la distribución de proba., en particular  $h(\phi)$**

**Se ilustra a que corresponde esta cantidad con respecto a  $f$  en la figura Fig. 2-24 más adelante.**

**Clase de Csiszár y versión simetrizada (1) Extension a la Renyi,**

**(2) a la HC/D/T, Cressie Reads, Cressie Pardo, Vajda;**

**(3) generalization Csiszar (et cf Burbea Rao aussi avec csiszar), Ali-Silvey, Ben-Tal**

**(4) generalization Csiszar Vajda, et voir avec  $h(\phi)$  avant même Salicru, Ben-Tal**

**BenTal livre cap. 17 p. 255**

**Csiszar 1995 dans appli**

**autres de Csiszar 2012 versikon Bregman;**

**gupta Sharma 1976**

**BoeLub79, Vajda72, Salicru94 Orsak et Paris; Voir OesVaj en termes de metrique pour les Arimoto, Vajda 2009, Kafka 91 Bas89, Bas13, application a le test d'adequation Pardo 99; MenMor97:5, Cf Pardo 2006 et ref.**

**La clase de las divergencias de Bregman** Estas divergencias fueron introducidos en el campo de la programación lineal convexa, para resolver problemas de minimización convexa <sup>37</sup> (Bregman, 1967), pero con aplicaciones en varios campos (Basseville, 1989, 2013, y ref.):

**Definición 2-21** (Divergencia de Bregman). Sea  $f : \Omega \subset \mathbb{R}^m \mapsto \mathbb{R}$  convexa y de clase  $C^1$  sobre  $\Omega$ , un cerrado convexo de  $\mathbb{R}^d$ . La divergencia de Bregman de un punto  $v \in \Omega$  relativamente a un punto  $u \in \Omega$  es definida por

$$B_f(v||u) = f(v) - f(u) - (v - u)^t \nabla f(u)$$

Dicho de otra manera,  $B_f$  corresponde al desarrollo de Taylor al orden 1 de  $f$  en la referencia  $u$ . Se ilustra a que corresponde esta cantidad con respecto a  $f$  en la figura Fig. 2-24 más adelante.

Esta definición fue generalizada a funciones actuando sobre espacios más generales (ej. actuando sobre matrices o operadores en espacios de Hilbert de dimensión infinita) (Petz, 2007). En lo que nos concierna en este capítulo, tratando posiblemente de densidad de probabilidad, nos interesa a funciones de funciones (Frigyik, Srivastava & Gupta, 2008; Nielsen & Nock, 2017):

**Definición 2-22** (Divergencia de Bregman funcional). Sea  $f : \Omega \mapsto \mathbb{R}$  convexa y de clase  $C^1$  sobre  $\Omega$ , un cerrado convexo de un espacio de Banach. La divergencia de Bregman de un “punto” (una función)  $v \in \Omega$  relativamente a un “punto” ( $u \in \Omega$ ) es definida por

$$B_f(v||u) = f(v) - f(u) - \lim_{t \rightarrow 0} \frac{f(u + t(v - u)) - f(u)}{t}$$

El último de esta formula es conocida como derivada de Gâteaux (o derivada direccional) de  $f$  en  $u$  en la dirección  $v - u$  (siendo  $u$  una función) <sup>38</sup>.

En el caso de que  $\Omega \subset \mathbb{R}^d$  se recupera sencillamente la definición original.

Para  $(h, \phi)$ -entropías discretas cóncavas (ej. con  $h$  cóncava), se puede entonces asociar una divergencia de Bregman

$$\begin{aligned} D_{(h,\phi)}^b(q||p) &\equiv B_{-H_{(h,\phi)}}(q||p) = H_{(h,\phi)}(p) - H_{(h,\phi)}(q) - (p - q)^t \nabla H_{(h,\phi)}(p) \\ &= H_{(h,\phi)}(p) - H_{(h,\phi)}(q) - h'(H_\phi(p))(q - p)^t \phi'(p) \end{aligned}$$

Cuando  $h \equiv \text{id}$ , se notará  $D_\phi^b$  y es equivalente a salir de la definición inicial con  $\Omega = [0; 1]$ ,  $u$  y  $v = q(y_i)$   $i$ -ésima componente de  $p$  y  $q$  respectivamente, y sumar la divergencia obtenida sobre  $i$ .

---

<sup>37</sup>Aún que aparece en una revista de matematica y física matematica, una gracia del papel de Bregman es que toma el ejemplo de maximización de la entropía de Shannon sujeto a momentos...

<sup>38</sup>De hecho, en la extensión de Frigyik et al. (Frigyik et al., 2008), se usa la derivada de Féchet, que es más general. Viene de un límite idéntica independientemente de la dirección. Entonces, si una función tiene una derivada de Fréchet, tiene necesariamente derivadas de Gâteaux, pero no es reciproca. Esta sutileza va más allá de la meta de esta sección.

En el caso continuo, para las  $(h, \phi)$ -entropías, se obtiene

$$D_{(h, \phi)}^b(q \| p) = H_{(h, \phi)}(p) - H_{(h, \phi)}(q) - h'(H_{\phi}(p)) \int_{\mathcal{X}} (q(x) - p(x)) \phi'(p(x)) dx$$

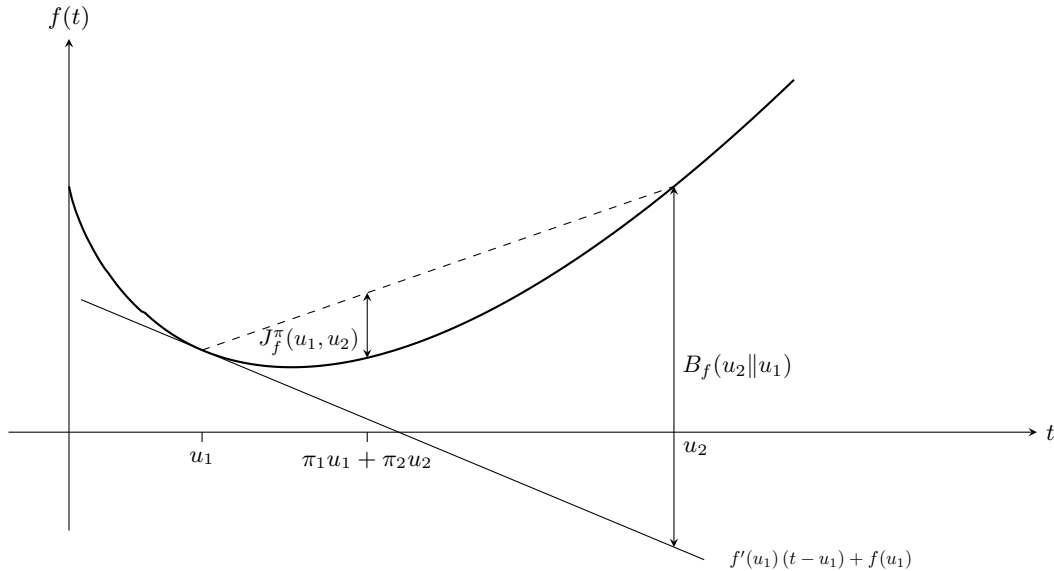
De nuevo, cuando  $h \equiv \text{id}$ , se notará  $D_{\phi}^b$  y es equivalente a salir de la definición inicial  $u = p(x)$ ,  $v = q(x)$  y sumar la divergencia obtenida sobre  $\mathcal{X}$ .

Notablemente, cuando  $\phi(x) = x \log x$  se recupera de nuevo la divergencia de Kullback-Leibler: esta última pertenece simultaneamente a la clase de Csiszar y a la de Bregman y es la sola en este caso (Csiszàr, 1991). **Sin embargo, se puede ver que si las  $f$ -Jensen se escriben como combinaciones convexas de divergencias de Bregman**

$$J_f^{\pi}(p_1, p_2) = \pi_1 B_f(p_1 \| \pi_1 p_1 + \pi_2 p_2) + \pi_2 B_f(p_2 \| \pi_1 p_1 + \pi_2 p_2)$$

**(Zhang, 2004; Nielsen & Boltz, 2011; Nielsen & Nock, 2017)**

La figura Fig. 2-24 ilustra a que corresponden  $D_f$  y  $J_f$  con respecto a la función convexa  $f$ .



**Figura 2-24:**  $f$  estrictamente convexa. Las cantidad positiva marcada por la dupla-flecha representan respectivamente la divergencia de  $f$ -Jensen  $J_f^{\pi}(u_1, u_2)$ , diferencia entre la combinación convexa de los  $f(u_i)$  y  $f$  de la combinación convexa de los  $u_i$ , y la divergencia de Bregman  $B_f(u_2 \| u_1)$  diferencia entre el valor en  $u_2$  (punto de evaluación) y la tangente en  $u_1$  (punto referencia). Para  $J_f^{\pi}$ , se toma como referencia  $\pi_1 u_1 + \pi_2 u_2$ , se calcula  $D_f$  en los  $u_i$  y se toma la combinación convexa.

La divergencia de Bregman tiene las propiedades siguientes

1. Positividad:

$$B_f(q \| p) \geq 0 \quad \text{con igualdad si y solamente si} \quad p = q$$

Esta propiedad es la consecuencia directa de la convexidad estricta de  $f$ , como ilustrado **figura Fig. 2-23.**

2.  $B_f(q||p)$  es convexa con respecto a  $q$ , pero no necesariamente con respecto a  $p$ . Es también consecuencia directa de la convexidad de  $f$ .
3. Pensando a  $B_f$  con respecto a  $f$ , es lineal en el sentido de que  $B_{\lambda f_1 + \lambda_2 f_2} = \lambda_1 B_{f_1} + \lambda_2 B_{f_2}$  (con  $f_i$  convexas y  $\lambda_i \geq 0$ ).

Ver (Frigyik et al., 2008; Nielsen & Boltz, 2011; Nielsen & Nock, 2017) para tener más propiedades.

Se notará que la clase de las divergencias de Bregman contiene el cuadrado de la distancia  $L^2$  ( $f(p) = \int p^2$ ), el cuadrado de la distancia  $L^1$  ( $f(p) = \left(\int p\right)^2$ ), la distancia de Itakura-Saito cuando  $\phi(u) = -\log u$  (asociado a la entropía de Burg), entre otros.

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

### 2.6.3 ¿Como se generalizan las identidades y desigualdades?

**Principio de entropía máxima** Si este principio nació en el marco de la termodinámica o física, con la entropía de Shannon (Boltzman), tratando de las nociones generalizadas de incertidumbre, vuelve natural preguntarse sobre la extensión de este problema en el marco general. **Tal estudio fue hecho en varios trabajos (?, ?) nous, Kesavan, Kagan 63.**

El problema se formaliza como en el caso Shannon, buscando la entropía máxima sujeto a vínculos: sea  $X$  variable aleatoria viviendo sobre  $\mathcal{X} \subset \mathbb{R}^d$  con  $K$  momentos  $E[M_k(X)] = m_k$  fijos, con  $M_x : \mathcal{X} \rightarrow \mathbb{R}$ , el problema de  $(h, \phi)$ -entropía máxima se formula de la manera siguiente en el caso continuo (es el caso discreto, hay que reemplazar integrales por sumas): sean  $M(x) = \begin{bmatrix} 1 & M_1(x) & \dots & M_K(x) \end{bmatrix}^t$  y  $m = \begin{bmatrix} 1 & m_1 & \dots & m_K \end{bmatrix}^t$ , se busca,

$$p^* = \operatorname{argm\acute{a}x}_p H_{(h,\phi)}(p) \quad \text{sujeto a} \quad p \geq 0, \quad \int_{\mathcal{X}} M(x) p(x) dx = m$$

donde los dos primeros vínculos aseguran de que  $p^*$  (positividad, normalización) sea una distribución de probabilidad. Si  $\phi$  es convexa (resp. cóncava),  $h$  es creciente (resp. decreciente) así que maximizar  $H_{(h,\phi)}$  es equivalente a maximizar  $H_\phi$  (resp.  $H_{-\phi}$ ). Sin pérdida de generalidad, se puede considerar la situación  $\phi$  convexa. Como en el caso de Shannon, introduciendo factores de Lagrange  $\lambda = \begin{bmatrix} \lambda_0 & \lambda_1 & \dots & \lambda_K \end{bmatrix}^t$  para tener en cuenta los vínculos, el problema variacional consiste a resolver (Gelfand & Fomin, 1963; van Brunt, 2004; Miller, 2000; Cambini & Martein, 2009; Cover & Thomas, 2006)

$$p^* = \operatorname{argm\acute{a}x}_p \int_{\mathcal{X}} (-\phi(p(x)) + \lambda^t M(x) p(x)) dx$$

donde  $\lambda$  será determinado para satisfacer los vínculos. De nuevo, de la ecuación de Euler-Lagrange (Gelfand & Fomin, 1963; van Brunt, 2004) se obtiene la ecuación  $-\phi'(p(x)) + \lambda^t M(x) = 0$ . La función entrópica  $\phi$  es

cóncava y de clase  $C^2$ , así que  $\phi'$  es continua decreciente, y de la monotonía es invertible. Entonces,

$$p^*(x) = \phi'^{-1}(\lambda^t M(x))$$

con  $\lambda$  tal que se satisfacen los vínculos de normalización y momentos. Si el resultado no es positivo en  $\mathcal{X}$ , de las condiciones KKT,  $p^*(x) = \left( \phi'^{-1}(\lambda^t M(x)) \right)_+$ . Estas distribuciones no caen en general en la familia exponencial. De una forma, usando entropías generales permite escaparse de esta familia.

Como en el caso de Shannon, queda obviamente el hecho de que no se puede determinar  $\lambda$  tal que se satisfacen todos los vínculos (y en particular la de normalización).

Tal como en el caso Shannon, existe una prueba informacional:

**Lema 2-4.** 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 satisfaga  $\phi'(p^*(x)) = \lambda^t M(x)$ . Entonces

$$\forall p \in \mathcal{P}_m, \quad H_{(h,\phi)}(p) \leq H_{(h,\phi)}(p^*) \quad \text{con igualdad ssi } p = p^*$$

*Demostración.* Sin pérdida de generalidad, consideramos  $\phi$  convexa. Calculando la divergencia de Bregman asociado a  $\phi$  de  $p$  relativamente a  $p^*$  da

$$\begin{aligned} D_{\phi}^b(p \| p^*) &= H_{\phi}(p^*) - H_{\phi}(p) - \int_{\mathcal{X}} (p(x) - p^*(x)) \phi'(p^*(x)) dx \\ &= H_{\phi}(p^*) - H_{\phi}(p) - \lambda^t \int_{\mathcal{X}} (p(x) - p^*(x)) M(x) dx \\ &= H_{\phi}(p^*) - H_{\phi}(p) \end{aligned}$$

siendo  $p$  y  $p^*$  en  $\mathcal{P}_m$ . El resultado proviene entonces de la positividad de la divergencia de Bregman, con igualdad si y solamente si  $p = p^*$  conjuntamente a la creencia de  $h$ .  $\square$

Este lema prueba que, dando vínculos “razonables”, la  $(h, \phi)$ -entropía es acotada por arriba, y que se alcanza la cota. Por ejemplo,

- Con  $K = 0$  y  $\mathcal{X}$  de volumen finito  $|\mathcal{X}| < +\infty$ , la distribución de  $(h, \phi)$ -entropía máxima es la distribución uniforme en el caso discreto tal como en el caso continuo.
- **Con  $K = 1$ ,  $\mathcal{X} = \mathbb{R}^d$  y  $M(x) = xx^t$  (visto con  $d^2$  vínculos), y  $\phi(u) = u^\beta$  (Rényi o Havrda-Charvát-Daróczy), la distribución de entropía máxima es Student; Costa and son on. Gausiana se recupera caso límite.**

**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 cuantificación 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*

*ἀγεωμέμετρος 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*, aquí 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 Kähler, 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*  $\cup$  e *intersección*  $\cap$  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 través de la idea de una **topología** sobre  $X$ .

**Definición 3-23.** Una **Topología**  $\tau$  sobre el conjunto  $X$  es una familia de subconjuntos de  $X$  que cumple con las siguientes condiciones:

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

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

3. La unión de una colección arbitraria - finita o no- de elementos de  $\tau$ , pertenece a  $\tau$ :

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

**Definición 3-24.** Al par  $(X, \tau)$  lo llamaremos **Espacio Topológico**. Los conjuntos que están en  $\tau$  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  $\tau$ , 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  $p = (p_1, \dots, p_n) \in \mathbb{R}^n$  y de radio  $r > 0$  es el conjunto

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

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 3-25.** 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 3-26.** Sea  $f : X \rightarrow Y$  una función entre dos espacios topológicos  $(X, \tau)$  e  $(Y, \omega)$ .  $f$  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 3-27.** Un **homomorfismo**  $\Psi$  entre dos espacios topológicos  $(X, \tau)$  e  $(Y, \omega)$  es una función

$$\Psi : X \rightarrow Y \subseteq Y$$

biyectiva, continua y con inversa continua.

**Definición 3-28.** Una **sucesión** en un conjunto  $X$  es una aplicación  $s : \mathbb{N} \rightarrow X$  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 3-29.** Sea  $(X, \tau)$  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  $(X, \tau)$  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 3-30.** Un **Espacio Métrico** en un conjunto  $X$  munido de una función  $d : X \times X \rightarrow \mathbb{R}_+$  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  $d : X \times X \rightarrow \mathbb{R}^0$  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 características 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 3-31.** 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  $\mathbb{R}^n$ :

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

tal que

$$\phi : U \rightarrow V$$

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

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

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

la colección de números reales  $(p_1, \dots, p_n)$  se llaman las coordenadas de  $p$  de acuerdo a la carta  $(U, \phi)$ . Podría suceder que un mismo punto  $p$  pertenezca a más de una carta, digamos  $(U_1, \phi_1)$  y  $(U_2, \phi_2)$ . En ese caso hablaremos de un cambio de coordenadas:

$$\psi_2 \circ \phi_1^{-1} : \phi_1(U_1 \cap U_2) \rightarrow \psi_2(U_1 \cap U_2) \quad (5)$$

Si denotamos por  $(p_1, \dots, p_n)$  a las coordenadas correspondientes a la carta  $(U_1, \phi_1)$  y por  $(\tilde{p}_1, \dots, \tilde{p}_n)$  a las correspondientes a la carta  $(U_2, \psi_2)$ , entonces las funciones  $\tilde{p}_i = \tilde{p}_i(p_1, \dots, p_n)$  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:

$$\mathbb{S}^n = \{(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1} \text{ tales que } x_1^2 + \dots + x_{n+1}^2 = 1\}$$

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.*

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 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 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) <sup>39</sup> .

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 simulado, (Arial 10, interlineado de 1,5 líneas,

[illegible]

**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 4-25:** Eso es una figura, con su leyenda sobre varias líneas para ver como queda en el texto. Eso es una figura, con su leyenda sobre varias líneas 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 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 simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm).

Ejemplo con respeto al capítulo 1

Para ver que las referencias de capítulos andan: 1; 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 epígrafe 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 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 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) <sup>40</sup> .

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 simulado, (Arial 10, interlineado de 1,5 líneas,

[illegible]

**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 5-26:** Eso es una figura, con su leyenda sobre varias líneas para ver como queda en el texto. Eso es una figura, con su leyenda sobre varias líneas 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 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 simulado, (Arial 10, interlineado de 1,5 líneas, sangría en primera línea de 0,5cm).

Ejemplo con respeto al capítulo 1

Para ver que las referencias de capítulos andan: 1; 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).

# EPÍLOLOGO

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*



# Referencias

- 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.
- Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4), 349–369.
- Basseville, M. (2013). Divergence measures for statistical data processing – an annotated bibliography. *Signal Processing*, 93(4), 621–633.
- Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics*, 50(4), 495–510.
- 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.
- Bhatia, R. (1997). *Matrix Analysis*. New-York: Springer Verlag.
- 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.

Barth.

- Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problem in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3), 200–217.
- 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.
- Chong, K. M. (1974). Some extensions of a theorem of Hardy, Littlewood and Pólya and their applications. *Journal canadien de mathématiques*, 26, 1321–1340.
- 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.
- Csiszár, I. (1991). Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4), 2031–2066.
- Darmois, G. (1935). Sur les lois de probabilités à estimation exhaustive. *Comptes rendus de l'Académie 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.
- Daróczy, Z. & Járαι, A. (1979). On the measurable solution of a functional equation arising in information theory. *Acta Mathematica Academiae Scientiarum Hungaricae*, 34(1-2), 105–116.
- 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.
- Endres, D. & Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7), 1858–1860.
- 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). *Foundations 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.
- François, O. (2009). Notes de cours de probabilités appliquées. Ensimag.
- 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.
- Frigyik, B. A., Srivastava, S., & Gupta, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54(11), 5130–5139.
- 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.
- Golberg, R. R. (1961). *Fourier Transforms*. Cambridge University Press.
- 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.
- Hardy, G. H., Littlewood, J. E., & Pólya, G. (1929). Some simple inequalities satisfied by convex functions. *Messenger of Mathematics*, 58, 145–152.
- 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  $\alpha$ -entropy. *Kybernetika*, 3(1), 30–35.
- Holevo, A. (2011). *Probabilistic and statistical aspects of quantum theory* (2nd ed.), volume 1 of *Quaderni Monographs*. Pisa: Edizioni Della Normale.
- Holevo, A. S. (1973). Bounds for the quantity of information transmitted by a quantum communication channel. *Problems of Information Transmission*, 9(3), 177–183.
- 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.
- Jeffrey, H. (1948). *Theory of Probability* (2nd ed.). Oxford: Clarendon.
- Johnson, O. (2004). *Information Theory and The Central Limit Theorem*. London: Imperial college Press.
- Kafka, P., Öesterreicher, F., & Vincze, I. (1991). On powers of  $f$ -divergences defining a distance. *Studia Scientiarum Mathematicarum Hungarica*, 24(4), 415–422.
- Kaniadakis, G. (2001). Non-linear kinetics underlying generalized statistics. *Physica A*, 296(3-4), 405–425.
- Kapur, J. N. (1967). Generalized entropy of order  $\alpha$  and type  $\beta$ . *The Mathematical Seminar*, 4, 78–94.
- Kapur, J. N. & Kesavan, H. K. (1992). *Entropy Optimization Principle with Applications*. San Diego: Academic Press.
- Karamata, J. (1932). Sur une inégalité relative aux fonctions convexes. *Publications Mathématiques de l'Université de Belgrade*, 1, 145–148.
- 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.
- Krajčič, 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.
- Kullback, S. (1968). *Information Theory and Statistics*. Dover Publications.
- Kullback, S. & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86.
- Laplume, J. (1948). Sur le nombre de signaux discernables en présence de bruit erratique dans un système de transmission à bande passante limitée. *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 subadditivity 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.
- Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1), 145–151.
- Lindhard, J. & Nielsen, V. (1971). Studies in statistical mechanics. *Det Kongelige Danske Videnskabernes Selskab Matematisk-fysiske Meddelelser*, 38(9), 1–42.
- Lukacs, E. (1961). Recent developments in the theory of characteristic functions. In *Proceeding of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 2: Contributions to Probability Theory, (pp. 307–335). University of California Press, Berkeley, CA.
- Lundheim, L. (2002). On Shannon and “Shannon's formula”. *Teletronikk*, 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.

- Mandel, L. & Wolf, E. (1995). *Optical coherence and quantum optics*. Cambridge University Press.
- Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of Majorization and Its Applications* (2nd ed.). New-York: Springer Verlag.
- 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).  $(h, \phi)$ -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. *Kybernetika*, 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.
- Nielsen, F. & Boltz, S. (2011). The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8), 5455–5466.
- Nielsen, F. & Nock, R. (2017). Generalizing skew Jensen divergences and Bregman divergences with comparative convexity. *IEEE Signal Processing Letters*, 24(8), 1123–1127.
- Onicescu, O. (1966). Energie informationnelle. *Comptes rendus de l'académie des Sciences. série 1, mathématiques*, 263(3), 841–842.
- Osán, T. M., Bussandri, D. G., & Lamberti, P. W. (2018). Monoparametric family of metrics derived from classical Jensen-Shannon divergence. *Physica A*, 495, 336–344.
- Österreicher, F. & Vajda, I. (2003). A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3), 639–653.
- 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.
- Petz, D. (2007). Bregman divergence as relative operator entropy. *Acta Mathematica Hungarica*, 116(1-2), 127–131.
- 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 *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Contributions to the Theory of Statistics, (pp. 547–561). University of California Press, Berkeley, CA.
- 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. *Qüestió*, 11(3), 3–12.
- Salicrú, M., Menéndez, M. L., Morales, D., & Pardo, L. (1993). Asymptotic distribution of  $(h, \phi)$ -entropies. *Communications in Statistics – Theory and Methods*, 22(7), 2015–2031.
- Sayood, K. (Ed.). (2003). *Lossless Compression Handbook*. San Diego, CA: Academic Press.
- Schur, I. (1923). Über eine klasse von mittelbildungen mit anwendungen auf die determinantentheorie. *Sitzungsberichte der Berliner Mathematischen Gesellschaft*, 22, 9–20.
- 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  $(\alpha, \beta)$  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. *Infor-*

*mation 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  $\alpha$ . *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 and the Machine* (2nd ed.). Cambridge, MA: MIT Press.
- Wong, A. K. C. & You, M. (1985). Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(5), 599–609.
- Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation*, 16(1), 159–195.

# 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 Adelina

Obtuvo el título de Licenciada en Física en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata, y el grado de Doctora en Física en la misma casa de altos estudios. Es Investigador Independiente del Consejo Nacional de Investigaciones Científicas y Técnicas, con lugar de trabajo en el Instituto de Física La Plata. Su especialidad es la teoría y geometría de la información en mecánica cuántica. Posee cargo docente de Profesor Adjunto en el Departamento de Matemática de la Facultad de Ciencias Exactas de la UNLP, desempeñándose desde 2013 como integrante del Equipo Coordinador de la asignatura Análisis Matemático II (CiBEx). cursos de grado avanzados y de posgrado en la Facultad de Ciencias Exactas de la UNLP y en la Facultad de Matemática, Astronomía, Física y Computación de la Universidad Nacional de Córdoba. También ha participado en el dictado del curso de grado “Probabilidades” como Profesor Visitante de la Université Grenoble-Alpes en Francia.

## Zozor, Steeve

Nació en 1972 en Colmar, Francia. Obtuvo el título de Ingeniero, de Licenciada, el grado de Doctor y la “Habilitation à diriger de Recherches”, respectivamente en 1995, 1999 y 2012, ambos del Instituto Nacional Politécnico de Grenoble (Grenoble INP), Francia. En 2001, paso varios meses en el Laboratorio de Procesamiento de Señales de la Escuela Politécnica Federal de Lausanne (EPFL), Suiza como postdoctorante. Pasó un año en el Instituto de Física de La Plata (IFLP) de la Universidad Nacional de La Plata (UNLP), Argentina (2012-2013) así que varios estancias desde 2010 como profesor visitante. En 2001 ingresó al Centro Nacional de la Investigación Científica (CNRS), equivalente Francés del CONICET, como “Chargé de Recherche” (cargado de investigación) y es “Directeur de Recherches” (director de investigación) desde 2017, ambos en el Laboratorio de Imágenes, Palabras, Señales y Automática de Grenoble (GIPSA-Lab), Francia. Desde 2015 es editor asociado de la revista IEEE Signal Processing Letters. Sus temas de investigación incluyen el procesamiento no lineal de señales, el estudio del efecto de resonancia estocástica, el estudio de procesamiento de datos en contextos  $\alpha$ -estables y/o de distribuciones de probabilidad elípticas, la teoría de la información

(medidas informacionales generalizadas clásicas y cuánticas) con aplicaciones en procesamiento de datos, mecánica cuántica o ingeniería biomédica. Es a cargo de docencia en varias escuelas de Grenoble-INP de matemática para el ingeniero, probabilidades aplicadas, procesamiento estadístico de señales, métodos bayesianos. Da regularmente un mini-curso sobre los básicos de la teoría de la información en la Facultad de Ciencias Exactas de la UNLP.